Scraping By

2020.09

HOUSING RESEARCH COLLABORATIVE

Authors

Riley Iwamoto – M.A.P., Consultant, Living Daylight
Iris Gao – B.Sc Mathematical Sciences, University of British Columbia
Tori Lin - M.U.D., University of Hong Kong, MCRP, University of British Columbia

Acknowledgements

Foremost, we acknowledge that the Metro Vancouver housing market we study is an imposition on the ancestral, traditional, and unceded territories of the Musqueam People, the Hwlitsum, Katzie, Kwantlen, Kwikwetlem, Matsqui, Qayqayt, Semiahmoo, Squamish, Tsawwassen, and Tsleil-Waututh First Nations.

We would like to thank those whose contributions and consultation made this study possible

Penny Gurstein – Director, Housing Research Collaborative; Professor, School of Community and Regional Planning, University of British Columbia
Raymond Ng – Professor at the Faculty of Data Science, UBC
Elvin Wyly – Professor at the Department of Geography, UBC

Andy Yan –Director of City Program, Continuing Studies at Simon Fraser University
Aaron Cheng - Planning Assistant at City of Vancouver
Craig Jones – Coordinator at Housing Research Collaborative, UBC
Louie Dinh – Quantitative Rhetoric
Jay Feng – Co-Founder, Interview Query
Jocelyn Lee – Data Scientist at Infineon Technologies
Kyle Crowder – Blumstein-Jordan Professor of Sociology at University of Washington
Suzanne Milbourne – Manager of Programs at Canada Mortgage and Housing Corporation
The Ha – IT Director, Department of Statistics, UBC
Tim Thomas – Postdoctoral Fellow at University of California Berkley
UBC Legal Counsel

UNC Behavioral Research Ethics Board Approval: H19-00880

Contents

Executive Summary

Vancouver is famously expensive to live in, with rents and homelessness dramatically rising in the previous decade (HRC 2018). In the context of the ongoing COVID-19 pandemic, there is a need to track changes in the rental market to understand the extent of financial strain renters may be experiencing. However, federally conducted surveys and census operate at coarse annual and five year intervals and, while providing useful summaries on how much renters are already paying, offer comparatively less detail regarding how much one can expect to pay when moving into a unit today. While condominiums – part of the secondary rental market - have received considerable attention, data on the rest of significant secondary market is comparatively more challenging to collect and thus scarce. As the non-condominium secondary market accounts for nearly one third of Vancouver's estimated rental stock, this lack of detail makes it difficult to estimate the realities of supply and where renters are finding relative affordability. Boeing and Waddell (2017) have previously explored the potential of collecting data directly from online rental market activity in a paradigm referred to as Voluntary Geographic Information (VGI) which is purported to address some of these information gaps in conventional sources.

Based on a desire for the City of Surrey to understand its "hidden rental market", students at the University of British Columbia (UBC) developed a tool to scrape VGI from popular classified sites to help the municipality understand what conventional data does not capture (Lee et al. 2018). The results prompted the question of what the role of VGI may be in enhancing public information in the whole Census Metropolitan Area (CMA) of Metro Vancouver, the topics of affordability, and aspects of the market beyond.

Expanding the scope and functionality of this tool, the Housing Research Collaborative (HRC) gathered over 17,000 points of data from one of the oldest and most dominant classified ad websites for the Metro Vancouver region over five months in late 2019 to early 2020. Five months of data were used to produce a cursory snapshot of the rental market, mapping the price and spatial distribution of rental housing across the Metro Vancouver CMA. The results of the data were also compared to existing government and private sources to serve conversations around the cost of living and the relative value of VGI. This study was also developmental, iterating the tool and manual processing methods as the nature of the data and the novel needs of this kind of research became better understood through study. The improved state of the tool, documentation of the processing methods, and outline of future developments are as integral products of this study as the market information itself.

The results from this early analysis are promising but should not be taken as conclusive arguments about the reality of renting in Vancouver, requiring some assumptions and intersections with other data to be made. The variety of topics that can be further explored, either with sustained collection or software development and analysis, also show potential value. Data derived from classified ads appears to have a complementary role to play with existing sources of data, offering provocative insights into questions of supply, composition, affordability, trends, and behaviors of tenants and landlords in an online market.

Key Findings

The potential role of VGI in issues of rental housing and affordability.

- VGI can offer a very agile source of vacant rental housing data that stands to complement occupied unit data collected by government sources, rounding out estimates of inventory composition, location, and affordability pressures.
- While collection is fast and automated, processing is currently highly manual and intensive. While the tools make the collection of large volumes of VGI possible in near real-time, the messy Big Data nature of classified ads mean that there is currently significant manual overhead in processing and cleaning the data.
- There are many qualifiers, caveats, drawbacks and necessary assumptions to make when drawing inferences from VGI. Researchers and users of the information must be very careful interpreting the meaning of the data.
- Realization of VGI's potential will require much further methodological development and data collection from multiple platforms.

What does this data say about renting in Metro Vancouver?

- The highest densities of ads, particularly in the one and two-bedroom range, were typically found on the downtown Vancouver peninsula and other central areas, principally along major transit corridors. Three-bedroom units were relatively more dispersed with a still noticeable concentration in the downtown peninsula. Dwellings with four or more bedrooms were nearly absent from the downtown core and, although typically found in detached houses, were more prevalent in some neighbourhoods than others of comparable form.
- The volume of ads by bedroom type appears to show a significantly different composition than counts of the Primary Market in the Canada Mortgage and Housing Corporation's (CMHC) Rental Market Survey, including a higher proportion of three or more bedroom units (14.6% vs 3.6%). One-bedroom units, while still a large share of the collected ads (26.3%), represented much less than their share of the primary market as counted by the CMHC (60.4%). Although one-bedroom units still accounted for nearly twice the volume of three or more bedroom units, they offered less than half the bedrooms (15.8% vs 31.7%) during the study period. This may inform conversations about how the region's stock of dwellings in the non-condominium secondary market is accommodating new renters.
- Affordability for a median renter household is poor, with most two-bedroom units further west than Surrey requiring more than 30% of the model household's monthly income for rent alone.
- On a per-bedroom basis, ads in the private rooms category were found to offer some of the most affordable opportunities for renters and insight on how people find affordability in Vancouver via shared spaces and transactions that do not involve new leases.

SECTION 1

Introduction
Rental Data Landscape and VGI

1.1 An Overview of Volunteer Geographic Information

The data collection method in this study is generally referred to as Volunteer Geographic Information. VGI builds sets of geographic data from voluntarily provided, user-generated content, often from the internet and GPS-enabled personal devices, rather than rigorously administered professional survey. Used in both physical and human geography, the typical tradeoff can be simplified as one of gaining a great volume of data, which may be logistically insurmountable under conventional research, at the expense of rigor and accuracy.

The specific technique applied in this study is referred to as web 'scraping', where a scripted tool simulates a user on a web browser and parses the information on the webpages it visits – classified ads for rental housing in this case - into meaningful research variables such as rent, number of bedrooms, and location. Scraping classified ads is a form of so-called "Big Data" wherein users are not explicitly tailoring their information for the purposes of the research but are instead conducting their own business in a publicly open forum which is then observed; in planning terms, it may thought of as a digital form of behavioral observation of people in public spaces like parks (Gehl and Svarre 2013). Because of this, there is no obligation for users to provide all details which might be of interest to the research, nor be entirely accurate in the information they do submit. This creates an issue of veracity in the data (Laney 2001) which, in turn, forms a variety of error trends which must be acknowledged and addressed if it is to produce meaningful analysis.

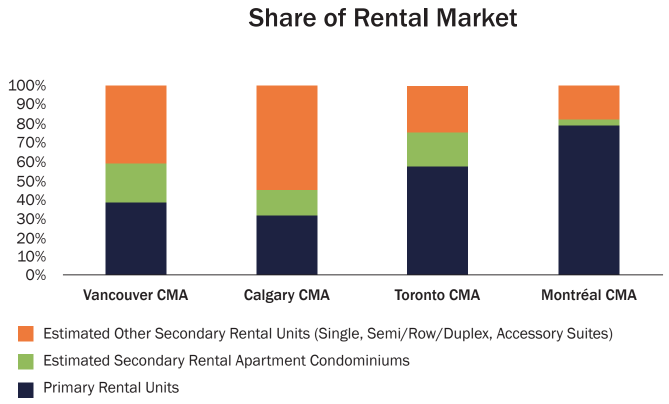
In 2014, researchers at the University of California, Berkeley deployed these methods to major rental markets in the United States, collecting around eleven million listings over three months (Boeing 2017). The resulting publication became the primary reference for this study. Noting the difficulty of conventional survey and census to capture the small private transactions that compose a great deal of rental markets, Boeing and Waddell (2017) highlight the dominance of online classified ads, particularly Craigslist.org (Craigslist), as the public's platform of choice for these transactions. While Craigslist has dominated the rental classified market not long after its founding in the 1990's, these novel forms of data collection have yet to be embraced by the planning profession today (Boeing 2017).

1.2 The Current Landscape of Canadian Rental Data

Why scrape rental classified ads? This fundamental question interrogates the value of VGI's ability to form an understanding of rental housing in Metro Vancouver and other Metropolises. Answering this requires understanding its relationship to established sources of rental market information.

Rental Market Survey and Canadian Census

In Canada, two federal agencies serve as primary sources of insight on the rental market and the circumstances of renting households: the Canada Mortgage and Housing Corporation's (CMHC) annual Rental Market Surveys (RMS) as well as the Canadian Census administered by Statistics Canada. The former is conducted largely by telephone survey and the latter is a component of the long-form Canadian Census sent to households every 5 years. These agencies also conduct supplementary studies. More recently, Statistics Canada has also published the Canadian Housing Survey which is generally concerned with qualitatively complex topics around how housing relates to Canadians' quality of life. The insights offered are used by local administration, the public, and the media alike in crafting policy and discussing the cost of living in Canada.



> The secondary rental market includes condominium apartments, laneway houses, and secondary suites, among other types. The secondary rental market is particularly important in the Vancouver CMA compared to the Toronto and Montreal CMAs, with secondary rental market units making up approximately 60 per cent of all rental units.

Figure 1: Relative proportions of Rental Market. The significance of the secondary market depends on the CMA in question.
Source: CMHC 2016 Rental Market Survey

The RMS is, for many, a primary source of data on the current state of the rental universe. It is relatively robust for the primary rental market, composed of larger apartment buildings of three units or more, but less so for the secondary market, which is composed of private landlords owning a few units in a range of structures including detached houses to condominiums. The secondary market is estimated to be larger by about a 60/40 proportion in Vancouver (fig. 1). The dedicated Secondary Rental Market Survey was apparently halted after 2016 in favour of a consolidated survey which tends to offer detail chiefly on condominiums, estimated to be one-third of the Metro Vancouver secondary market. Despite the significance of this market, logistical challenges and a general shift away from landline telephones – the directories of which CMHC traditionally based its surveys upon – has compelled the CMHC to divert its efforts and explore alternative sources, such as the Labour Force Survey, to cover the intervening market blind spot (CMHC, personal communication, March 25, 2020). As such, there is little detail published about suites in detached, duplex, and accessory dwellings which are estimated to account for forty per cent of Metro Vancouver's rental universe. This missing detail is expected to have variable relevance depending on which region is being looked at, with non-condominium secondary units counting for a small portion of the Montreal CMA's market yet more than half of Calgary CMA's.

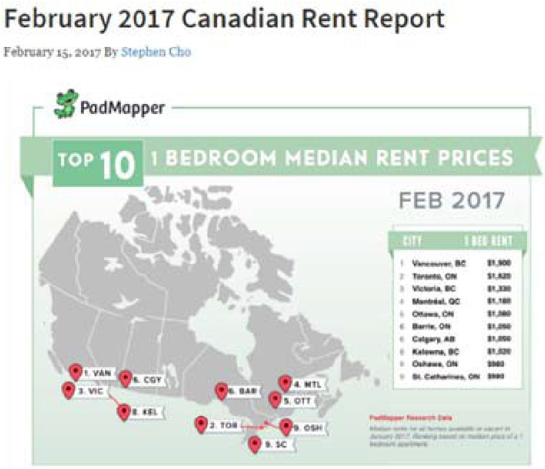
The Canadian Census provides a fine spatial resolution of information at a wide five-year interval, comparable to the United States Census Bureau's American Community Survey which was identified by Boeing and Waddell (2017) as unable to provide a responsive pulse from the rental market. Since the Canadian Census samples households, rather than landlords and building managers, it is possible that it covers more of the non-condominium secondary market. Further, certain factors for renting households are limited to the long form of the census which, in 2016, was issued to a sample of 25% of households. The five-year interval presents an issue in the relevance of the data, potentially leaving policymakers and the public uncertain about the rental realities of intervening years.

When it comes to the question of how much one will need to pay for a certain unit today, these sources may not deliver an accurate picture. This is a question that prices collected from classified ads are likely better suited to answer.

While rigorous in collection and analysis, the design of the RMS and Census does limit the kind of information they can offer. Most importantly, their analyses are focused on rents of occupied units: the rents people are currently paying. This sample would include more long-standing relationships between tenants and landlords on leases that have renewed many times. In British Columbia, rent increases on renewed leases are limited by set percentages every year. The maximum rate these rents could increase is knowable and some may have increased less than the maximum since the original lease many years, even decades, ago. These leases likely represent a more affordable portion of rents within the averages of the Census and RMS, which will gradually rise as rent on renewed leases may increase and newer, likely higher priced, leases are signed. A recent comparison of the rent difference between vacant and occupied units has been released (CMHC 2020b) but is also limited to the primary market, with rents for vacant bachelor and three or more bedroom units deemed unpublishable in Metro Vancouver. When it comes to the question of how much one will need to pay for a certain unit today, these sources may not deliver an accurate picture. This is a question that prices collected from classified ads are likely better suited to answer.

Other Ad-Based Sources

Other sources of data are monthly reports published by private rental classified sites Padmapper.com and Rentals.ca. These are noteworthy because they draw from data of similar nature to this study. Padmapper's blog currently posts median rental prices for one- and two-bedroom suites in major Canadian cities on a monthly basis, offering an estimate of what it may cost to begin renting in Vancouver. These reports are based on ads made on the Padmapper platform as well as its partner entities. Rentals.ca's reports follow a similar format and are presumed to apply similar methods. For legal reasons, however, Padmapper is prohibited from gathering data from competitor site Craigslist which, as mentioned, captures the largest share of rental classified ads in the United States. This dominant share is expected to be true for Canada as well, raising a question about how representative these monthly reports are of the total rental market. To be fair, one should not assume that Craigslist's larger share is entirely representative of the whole market either. An immediate question that follows is how comparable these monthly medians are to those derived from the larger platform.



February 2017 Canadian Rent Report

February 15, 2017 By Stephen Cho

PadMapper

TOP 10 1 BEDROOM MEDIAN RENT PRICES

FEB 2017

	CITY	1 BED RENT
1	Vancouver, BC	$1,900
2	Toronto, ON	$1,620
3	Victoria, BC	$1,330
4	Montréal, QC	$1,180
5	Ottawa, ON	$1,080
6	Barrie, ON	$1,050
6	Calgary, AB	$1,050
8	Kelowna, BC	$1,020
9	Oshawa, ON	$980
9	St. Catharines, ON	$980

Padmapper Research Data

Figure 2: Capture of Padmapper Report

Padmapper's reports do not distinguish price figures at any smaller scale than whole municipalities, or Census Subdivisions (CSD). This may be due to several factors, including a sample size that would make monthly median rent estimates unreliable at finer geographic units. Without knowing the exact methods of the monthly reports, it is difficult to know for certain. In any case, it begs the question of how median prices map across a city – from Yaletown on the downtown peninsula to Southeast Vancouver or any rent gradients that may exist across other municipalities.

City Observatory, a Portland-based think-tank which focuses on cities and data, has published cautionary articles about the limits of ad-based median summaries from Zumper – Padmapper's U.S. counterpart (2018). These cautions point to spatial biases in cities where opportunities in more expensive neighbourhoods are more likely to be posted on Zumper's platform than more affordable neighbourhoods. Another caution focuses on how relatively overpriced ads may be over-represented versus attractively priced ones for comparable units since the former are more likely to persist on the platform over several months; the more attractively priced ads are more likely to be filled and taken down within days.

The preliminary results of this study will be compared to both the CMHC RMS conducted in October, 2019 and the Padmapper Blog results from a comparable timeframe. Given the different nature of these sources, these comparisons will mean different things.

1.3 Scope and Methods

Scope

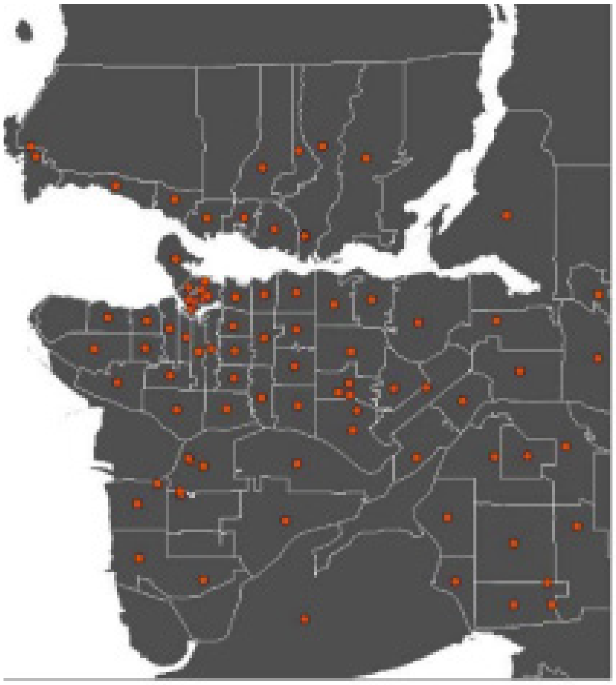


Figure 3: Allocation of low-accuracy results within Forward Sortation Areas

Scraping began in October 2019 upon the Vancouver.craigslist.org subdomain. This mostly returned results from the Vancouver Census Metropolitan Area (CMA) although, given the wider capture of the source, many results from the Abbotsford CMA and elsewhere were also found in the dataset and later omitted from analysis. Initially focusing on the "apartments/ housing for rent" category of the website, the "private rooms & shares" category was included in the latter half of the study when its considerable volume of ads was noted and their value reconsidered. The HTML content, excluding photographs, of each page was scraped and parsed into common table data.

Analysis herein covers October 2019 to February 2020, although collection continues through 2020. As many comparative figures were drawn from other sources, including the 2016 Canadian Census, one must mind that changes in the intervening four years will compromise the relevance of these comparisons until the 2021 Census.

Due to an apparent approximation of geolocations on a basis of postal codes in many ads, Forward Sortation Areas (FSA) - aggregations of postal codes - became the primary geographic unit of analysis. Low accuracy postings are expected to properly join with the correct FSA (fig. 3). With five months of data across all types, most FSA's contained at least thirty samples. Assuming an adequate sample volume and distribution from the study period, this makes FSA's most useful for summary figures. When unit types are isolated to evaluate price for a specific bedroom count, however, most types find very few samples in most FSA's and would need a longer data collection period to do so. Even then, some areas exhibit such a low population, that they may simply need to be omitted from long-term analysis or merged with others.

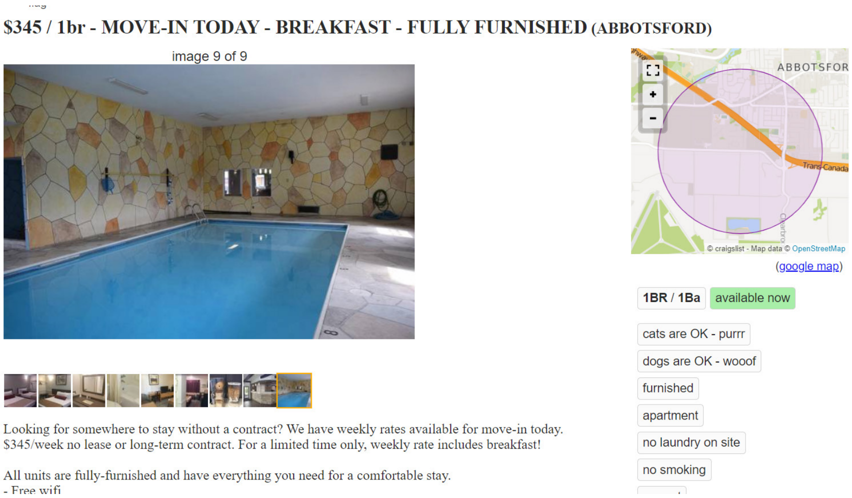
High Level Workflow

At a high-level, the study involved configuring and running the scraper for daily data collection, deduplicating, cleaning and preparing the table data for Geographic Information Systems (GIS) spatial analysis, and the analysis itself. Deduplication and cleaning – the removal of repeat and otherwise ineligible postings – was manually performed.

Initial operation of the tool commenced under the configuration inherited from the previous study. As the nature of unusable ads became better understood, the criteria and techniques for doing so were iteratively changed throughout the study. Some variables were dropped, some added, and the tool was reconfigured to collect raw html archives of the ads so any unforeseeable change to the scraper or the structure of the ads would not compromise the ability of the study to derive useful information or direct different research needs to the data. Specific procedures are documented in Appendix D.

Aspatial analysis involved discussion around the distribution of price by bedroom type. Spatial analysis centred around the question of how rental opportunities and price varied across the region. It involved basic summary figures of price, segmented by bedroom types, at various units of geography: Canada Post Forward Sortation Areas, CMHC Survey Zones, and Census Subdivisions depending on the comparison or analysis in question. Information from other sources, such as the RMS and Padmapper's rental market surveys, was brought in where relevant to observe any differences and discuss the possible sources of that difference.

Definition of a Usable Ad.



$345 / 1br - MOVE-IN TODAY - BREAKFAST - FULLY FURNISHED (ABBOTSFORD)

image 9 of 9

Looking for somewhere to stay without a contract? We have weekly rates available for move-in today. $345/week no lease or long-term contract. For a limited time only, weekly rate includes breakfast!

All units are fully-furnished and have everything you need for a comfortable stay.
- Free wifi

1BR / 1Ba available now

cats are OK - purrr
dogs are OK - wooof
furnished
apartment
no laundry on site
no smoking

Figure 4: Hotel advertising short term rentals in apartments category.

To pursue the highest veracity in the dataset possible, a definition of an acceptable ad was developed to direct the deduplication and cleaning procedures. A usable ad needed to...

- ***Be unique:*** Duplicate sets were reduced to one unique ad.
- ***Be in the Vancouver CMA:*** Ads that fell outside a polygon representing the legal CMA were clipped out of analysis. Ads located in the legal CMA but over water were retained, acknowledging the spatial error.
- ***Be a long-term residential rental opportunity in a fixed structure:*** The duration of a lease was rarely declared in an ad, but if a monthly rate was declared, the ad was retained. Prices determined to be weekly or daily rates were not retained. Ads for renting boats, yurts, parking spaces, land, or mobile homes were not retained. Any ad that was otherwise not a rental opportunity at all was not retained.

- ***Be advertised with a realistic monetary price:*** Ads with 'lure' prices of 1$ were not retained if a true price could not be found in the title or description. Ads with prices that involve a partial or full in-kind request such as a land swap, maintenance or health care work, were not retained.
- ***Have essential variables populated:*** Price, bedroom type, and geolocation.

A more detailed discussion of the usability criteria and the nature of the raw data can be found in Appendix D.

Classification

Units were classified by their number of bedrooms which, on Craigslist, range from one to eight although bedrooms not a required field and may be left null by the user. This null value is sometimes specified for bachelor and studio suites where the sleeping space is shared with other living functions in an open format; no distinct bedroom is defined from other living spaces with walls. Initially, these values were adjusted to bachelor, studio, and loft based on the title contents as there appeared to be consistent price distinctions between them. These types were later merged to better compare with CMHC data and collapse relatively small classes: studio and bachelor units were combined while lofts were merged with one-bedroom ads. Results from the private rooms and shares category, a separate section of Craigslist, were classed as "room" to distinguish their mixed nature of mainly very small and minimal subordinate units in detached houses as well as openings of constituent bedrooms in already occupied units.

Craigslist provides a housing type field when creating an ad, offering a variety of descriptions for either the structure the unit is a part of or the unit itself. It was noted that the default value, "apartment", was dominant across ads, including those evidently of different structure according to their description. This suggests that some users do not specify type from this default, over-representing what is classed as an apartment. This field was left unused for this project but may, in future study, be used to make results comparable with Census and CMHC structural classes.

In order to compare affordability across all ads which could accommodate various household sizes, an "effective beds", and price per bed (PPB) field were created to show how much a prospective renter would need to pay for housing in various arrangements. Renters with dependents could consider an additional bedroom as part of their costs. A price per square foot measure has been used by others but, as detailed in section 2.3, such an approach was deemed unsuitable for this study.

SECTION 2

The Potential of VGI in the Rental Data Landscape

This study's attentions are equally placed on the nature and value of VGI as they are about what the data suggests about Metro Vancouver. This section discusses the utility, limits, and challenges of this novel form of rental housing research.

2.1 What Can Rental Ads Show?

Core Questions of Supply and Affordability

With a near real-time account of geolocated rental opportunities as they become available, VGI has potential for a level of detail not possible in conventional methods. This detail comes, however, with many caveats and qualifiers.

The basic question to probe rental ads to answer is "What is the current asking price for x kind of rental unit across the region?" as well as follow-up questions of distributions and trends over time. This stands in complement to the occupied unit rents reported by Census and RMS and, when the two are held together, a clear picture of the market pressure - usually upward – rents are subject to in a given area.

While total inventory is difficult to assess with less than half of a year of data, the proportional share of different unit types have proven to be interestingly divergent.

VGI can offer insights into the profile of supply, in terms of distribution, of the non-condominium secondary market which is estimated to be very significant, yet currently less understood than other segments. One caveat to this potential is that the dataset does not exclude condominiums or apartments in the primary market; without an estimated breakdown of market type, the picture drawn from Craigslist is a mix of both primary and secondary markets. One of this study's analytical products will compare the reported composition of the primary rental market according to the RMS with the presumed wider market diversity of classified ads. While total inventory is difficult to assess with less than half of a year of data, the proportional share of different unit types have proven to be interestingly divergent.

Over long periods, with effective cleaning procedures, the inclusion of other platforms, and carefully chosen assumptions, a dataset may help approximate the composition of a region's rental stock.

Although a single month of Craigslist ads in the Metro Vancouver region does not seem to generate enough samples to confidently summarize the cost or inventory at a neighbourhood level or finer geography, acute events may be observable. Increases in volumes of ads over short periods, for example, may assist investigations into changes in turnover rates; the loss of supply followed by a resurgence due to redevelopment of a large land assembly may cause noticeable changes in the activity of one month to the next. Other phenomena we may see in some metropolitan areas are surges in vacancies once pandemic-related moratoria on evictions are lifted.

A focus on ads that continue to be reposted for multiple months may offer some profile on housing that remains vacant and on the market in a city with a famously low vacancy rate. The details of such ads may offer insight into why they may remain empty.

Although short and medium term rentals were eliminated for this study's analysis, future research may use such ads to better understand the nature of that market, such as the response of hotels and motels to AirBnB disruptions to their traditional market.

Other Questions

Of particular interest are the kinds of research this data opens up beyond essential supply and affordability questions. One may be interested, for example, in mapping pet-friendliness of units and estimating a price premium, if any, that accompanies such units or if certain language is used in accompanying ads. If basement suites were isolated, is there a significant price difference between them and above grade equivalents? Although short and medium term rentals were eliminated for this study's analysis, future research may use such ads to better understand the nature of that market, such as the response of hotels and motels to AirBnB disruptions to their traditional market.

There are apparent typological distinctions among units in the market which may not be represented in conventional surveys. As mentioned in section 1.3, both the bachelor and one-bedroom categories include types of ads that are nominally separable. Though technically similar, and while some ads used them interchangeably, the terms ”bachelor” and “studio” appear to carry different connotations. The language around the former appears to take a more budget-oriented character and is often marketed to students. Studios, on the other hand, are often advertised with a more cosmopolitan, young professional profile. Lofts describe units that are often converted from industrial spaces and are often found in Vancouver's Gastown district. There is nothing consistently true about a loft unit's construction that makes it clearly different from studio or one-bedroom units but they are typically much larger, more expensive, and are accompanied with language of luxury. Because they are very few, lofts were merged into the one-bedroom category but certainly added to the higher end

2.2 What Can Rental Ads not Show?

"One of the key interrogations of VGI is if rental ads can provide enough data per unit of geography – census tracts, for example - to deliver on the promise of both high spatial and temporal resolution. Do an adequate number of ads fall within each area to constitute a meaningful summary or represent the rental inventory?

Craigslist is a Large yet Incomplete Share of the Marketplace

The preceding section mentioned significant qualifiers and caveats to the promise of rental market VGI. The most basic qualifier lies in how representative the dataset is of the realities of renting in a region; Craigslist, while perhaps the single largest source of rental market data points, alone may not be an adequate share of the rental market to offer conclusive information, especially considering how each platform may exhibit its own user-selected biases. This study's predecessor found that another classified site in Canada, Kijiji, contributed another 10% of results to their earlier research, although cross-posted ads between platforms may constitute a measure of duplication that would need to be explored in any multi-platform scraping endeavour. Each additional online platform uses a different technical format and would require a customized tool to capture its postings, and additional deduplication for cross-posted ads.

Craigslist, while perhaps the single largest source of rental market data points, alone may not be an adequate share of the rental market to offer conclusive information, especially considering how each platform may exhibit its own user-selected biases.

One must be careful not to exclude less visible online and offline sources as well. SFU researcher Andy Yan (A. Yan, personal communication, April 24, 2020) has cautioned that the contemporary dominance of online classifieds is still not a full picture of the market. Yan notes that some market segments conduct business within ethnic, religious, or neighbouring communities, for example, via less public website, newspapers, bulletin boards, posters or other offline media. While expected to be smaller in volume, there is likely a unique profile of users and market behavior oriented around their community that would be missed if one were to presume all activity is conducted on Craigslist. Some community-specific opportunities were noted in the private room category, where some of the lowest priced offerings were specific to female Punjabi tenants in areas with amenity access to Sikh places of worship. Although no other groups posted notable volumes of ads on Craigslist, it is entirely likely they do so off-platform. There are, of course, transactions made among personal networks which absolve both landlords and tenants from posting ads or house hunting at all. Dedicated assessment of all major rental market transaction platforms should be performed to assess what kind of niches each fills and how representative Craigslist may be of the whole market.

Ad Distribution Affects the Significance of Analysis Unevenly

One of the key questions to be asked of VGI is if rental ads can provide enough samples per unit – census tracts, for example - of geography to reliably deliver on the promise of both high spatial and temporal resolution. Do an adequate number of ads fall within each area of analysis to constitute, for example a meaningful median? What is the minimum sample size for such a thing? Do we aggregate a longer period of time, a larger area, or try to include more sources of data to satisfy these needs? Confounding this further, another risk lies in areas that may have a significant amount of households but simply do not have many rented units; they may never exhibit a significant volume of rental ads and that may be more important than how much median rent is.

Dense areas such as downtown Vancouver will produce a high sample volume for any type in one month where a rural FSA may take five years to turn over enough data points for a defensible median or average.

It was found that, when the dataset was broken down into bedroom types, some types were so few in many Forward Sortation Areas (FSA) that a median was calculated off of a small handful of units, compromising its usefulness of that median. This issue manifests unevenly. Dense areas such as downtown Vancouver will produce a high sample volume for any type in one month where a rural FSA may take five years to turn over enough data points for a defensible median or average.

Vacant vs. Occupied Units

What makes this kind of data imperfectly comparable to survey-based methods like those conducted by the CMHC RMS is that these are new leases or, more accurately, proposed terms of lease for vacant units. This data does not capture stable, sustained or recurring leases between tenants and landlords which may compose a significant - and lower-priced - segment of the rental housing universe. One interesting exception of this lies in the private rooms category. Often postings in this category seek to replace roommates under existing leases. These existing leases may have been established recently or many years prior, giving some prospective tenants access to older, lower market prices and the protections of annual rent increase restrictions. Considering that prices in this category are among the lowest in absolute and per-bed terms, this may offer a glimpse into the ways the region's more affordable options are circulated in the market. This phenomenon will be discussed further in the results section.

In the regulatory context of the province of British Columbia, increases in rent are restrained to certain percentage gains for renewing leases but not for new agreements. Therefore, figures from scraping ads in the Metro Vancouver region might be seen as reflective of market pressures on rents, typically upward, when a unit is released from these restraints. Overall rents being paid will lag behind these upward pressures, and these differences may be analyzed over the long term to estimate how long it takes a market to effectively turn over. In regions where such rent increases are not regulated, such as the Calgary Metropolitan Region (CMR), occupied unit rents match vacant unit rents within the primary market much more closely than in Metro Vancouver, even exceeding them in some survey zones (CMHC 2020b). It may be more reasonable, then, to assume that rents reported in the Census may better match ads on Craigslist for that year. The minority proportion of the primary market, however, in both Metro Vancouver and the CMR (CMHC 2018) should be reiterated and the question of how the larger secondary market may differ remains.

Ad Volume and Distribution: Supply or Turnover

One must take caution when assigning meaning to the density and coverage of data points. This again stems from the fact that these listings represent eligible vacancies in rental housing. A greater point density might indicate a concentration of rental units such as a downtown core; ad density can indicate supply density. However, a concentration of ads may simply indicate a higher turnover rate than other areas of comparable rental composition, either ephemerally or in a more sustained way. Another scenario which may register an increase in listings is a surge in offerings in a small area as new product comes online - a trend that may fade longitudinally as new units are occupied. More than one factor may be responsible for a high density of listings in any given area; likely multiple trends and events are in play at any one time. Analysis of ad title, descriptions, relative ad density over time, and even visually investigating the physical character of an area with site visits or aerial imagery may help disentangle these factors.

Ad density can indicate supply density. However, a concentration of ads may simply indicate a higher turnover rate than other areas of comparable rental composition, either ephemerally or in a more sustained way.

Fixed-Location

A key component of conventional survey-based data is the fixed-location aspect, where selected properties are resampled on a regular basis. This allows them to isolate changes in rent while controlling for a possibly great deal of variation brought on by quality, structural (age, size, etc.), or locational factors. While hypothetically possible with VGI, doing similarly in a scraping paradigm would present significant challenges and rely on each unit to periodically be advertised with a high certainty that it is, in fact, the same unit being advertised with possibly different language and owners each time.

2.3 Operational Challenges

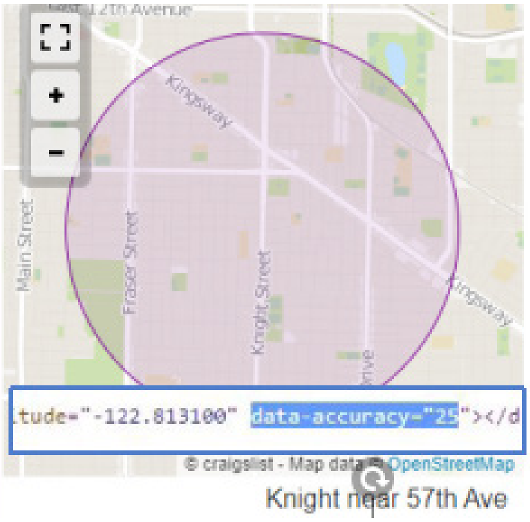
Whether in realizing the potential of VGI or mitigating its risks, a significant measure of care and effort must be exercised in conducting such study. While it is relatively easy to scrape a large volume of data, it was found to have limited veracity out-of-the-box. Deduplication and cleaning can address many of the discovered complications in the data - some from the nature of the platform, others from the habits of users. The apartments category (excluding private rooms) in January, for example, yielded 5,834 raw results which were reduced by 35% once apparent duplicates and otherwise ineligible ads were deleted. Given the many forms of user-generated issues, it is likely some erroneous ads remain. Many of the caveats of user-sourced big data of this kind stem from the often divergent objectives of the platform, user, and researcher.

The apartments category (excluding private rooms) in January, for example, yielded 5,834 raw results which were reduced by 35% once apparent duplicates and otherwise ineligible ads were deleted.

How Craigslist Complicates the Data

One of the challenges introduced by the platform stems from their own iterative website development. Changes to the html structure of the website and its ads can cause scraping software to fail to collect the targeted variables or even navigate the website at all. If the tool is operating in an unsupervised automated basis, it may miss a sizable volume of data before researchers catch the change. Certain strategies can be enacted to archive the html pages and detach the parsing operations to separate scripts. This makes the tool robust to changes in the html structure as the parsing script may be adapted and run on the archived ads even if the originals have been deleted from the website. This also makes the tool adaptable to different research questions, which may require different parsing algorithms to produce novel variables.

A significant challenge to the geolocation accuracy appears to have been introduced to Craigslist shortly before this study commenced. The default geolocation of an ad seems to be based upon the mandatory postal code field when creating an ad. This specifies latitude and longitude at locations sometimes kilometers away from the actual property. This approach appears to be designed to protect user and property privacy. While users can still place a pin on a map to specify a relatively precise location of the advertised unit, it is now an optional refinement. As many ads demonstrating this coarse geolocation also specify a street address, it is likely that they are less concerned about obscuring the unit's location and are either not aware of the more precise geolocation option or simply elect not to bother since they have already typed out the address (fig. 5). This creates a variable precision among ads, represented in the html code as "location accuracy" attribute. When discovered, the tool was adapted to collect this as a variable that could be used to refine analysis.



```
tude="-122.813100" data-accuracy="25"></d
```

Knight near 57th Ave

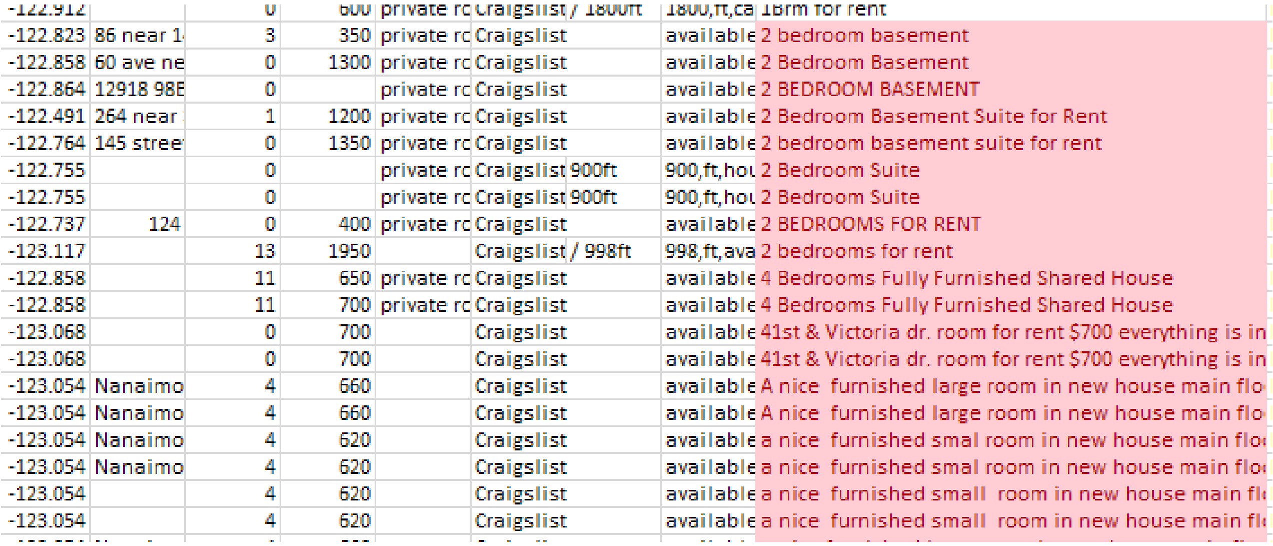
Figure 5: Screen capture of approximated property location and code example. More precise address voluntarily defined by user in text field.

Low accuracy results are seemingly centred on FSAs, with a few exceptions (fig. 3), which do not conform to common delineations of census geography. Users of this data are presented with a trade-off in selecting different geographies; FSAs will collect more accurate summary figures at the expense of comparability with data gathered on, for example, census tracts. Alternatively, users may either omit the significant proportion of low-accuracy data points or accept the error of analysis when operating on Census geography, including CMHC survey zones.

The phenomena of low accuracy postings produces stacks of multiple listings over a single point, ambiguating actual distributions of ads, visually reducing the apparent volume of ads, and creating visual anomalies in interpolated products where these postal code centres appear to be erroneously dense. Omitting these results creates more sensible interpolation among data points but significantly reduces the volume and coverage of points interpreted.

How Users Complicate the Data

When posting an ad, there are few enforceable obligations the user must conform to that would ensure all ads contribute to a reliable picture of the renting reality. The result is a body of ads that can distort the total picture of the dataset. Some may even satisfy Craigslist's criteria for a proper ad in the category but still be inappropriate for a dataset that needs to be comparable to conventional survey data. The most common behaviour that can distort the results is duplicate, or repeat, posting which would obviously over-represent both the price and spatial distribution of ads. Users often alter each

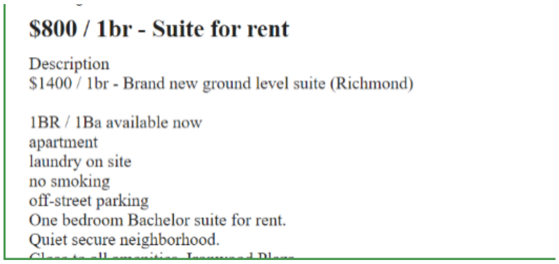


-122.912		0	600	private rc	Craigslist	/ 1800ft	1800,ft,ca	1brm for rent
-122.823	86 near 1	3	350	private rc	Craigslist		available	2 bedroom basement
-122.858	60 ave ne	0	1300	private rc	Craigslist		available	2 Bedroom Basement
-122.864	12918 98E	0		private rc	Craigslist		available	2 BEDROOM BASEMENT
-122.491	264 near	1	1200	private rc	Craigslist		available	2 Bedroom Basement Suite for Rent
-122.764	145 stree	0	1350	private rc	Craigslist		available	2 bedroom basement suite for rent
-122.755		0		private rc	Craigslist	900ft	900,ft,hou	2 Bedroom Suite
-122.755		0		private rc	Craigslist	900ft	900,ft,hou	2 Bedroom Suite
-122.737	124	0	400	private rc	Craigslist		available	2 BEDROOMS FOR RENT
-123.117		13	1950		Craigslist	/ 998ft	998,ft,ava	2 bedrooms for rent
-122.858		11	650	private rc	Craigslist		available	4 Bedrooms Fully Furnished Shared House
-122.858		11	700	private rc	Craigslist		available	4 Bedrooms Fully Furnished Shared House
-123.068		0	700		Craigslist		available	41st & Victoria dr. room for rent $700 everything is in
-123.068		0	700		Craigslist		available	41st & Victoria dr. room for rent $700 everything is in
-123.054	Nanaimo	4	660		Craigslist		available	A nice furnished large room in new house main flo
-123.054	Nanaimo	4	660		Craigslist		available	A nice furnished large room in new house main flo
-123.054	Nanaimo	4	620		Craigslist		available	a nice furnished smal room in new house main flo
-123.054	Nanaimo	4	620		Craigslist		available	a nice furnished smal room in new house main flo
-123.054		4	620		Craigslist		available	a nice furnished small room in new house main fl
-123.054		4	620		Craigslist		available	a nice furnished small room in new house main fl

Figure 6: Distinguishing actual duplicate ads from simply common titles requires manual scrutiny.

duplicate slightly to mislead Craigslist's measures to reduce abuse of repeat posting so detection typically required a human analyst to scrutinize two or more variables. Others post short-term rentals that cite daily or weekly rates that would skew results down if processed as a monthly rate. Some offer cleaning services or are not advertising anything at all, choosing arbitrary values in any variable. Because the criteria for determining an unusable ad require auditing multiple variables concurrently, including the long-form description of the ad, this is currently a manually performed process and forms the largest time and labour expense of this kind of data collection.

City Observatory (2018) points out a potential bias in using classified data where median or mean estimates may over-represent more expensive offerings that spend more time on the market than more attractive offerings which are closed soon after being posted. To address this issue, our deduplication procedure favours the latest instance of a reposted ad with the assumption that, in cases where the price is adjusted from one repost to another, the latest offer is the price the unit is rented for. There is an accepted trade-off in selecting the latest instance which will introduce some error into temporal analysis of ads: the original posting date will have been lost and evaluations of posting volumes will be biased to later dates. This policy does have exception where one instance may exhibit more detail, such as square footage which may be absent in other instances, or a better location accuracy than other repeat postings. In such cases, richer dimensionality was favoured over representation of the assumed final price.



$800 / 1br - Suite for rent

Description
$1400 / 1br - Brand new ground level suite (Richmond)

1BR / 1Ba available now
apartment
laundry on site
no smoking
off-street parking
One bedroom Bachelor suite for rent.
Quiet secure neighborhood.
Close to all amenities, Ironwood Plaza

Figure 7: One rent posted in the title. Another rent detailed in the ad itself.

Users often creatively misrepresent prices to stand out among other ads with unusually low prices, often $1, declared in the appropriate field, with actual prices declared in the title or description. Others will post multiple units in an ad, populating the price with the lowest of the set. Some ads post sincere prices but will also ask a supplementary value be provided by their tenants, such as in-home care or other work-in kind. Some homestay ads for international students will offer alternate prices for meal plans.

Other sources of error can be found from ads which are not offerings of rental housing. These can include cleaning services advertising to tenants and landlords within the rental pages of Craigslist and are easy to detect by prices too low to be a realistic rent. Harder to detect are those posted by users seeking places to rent and will post their budget for rent which will not necessarily stand out from prices for places to rent. Others, still, are not ads at all but are information, sometimes warning of fraudulent or abusive individuals who may post on the platform. These sometimes post unusual prices like $1234 or $666 and can be discovered by searching for terms such as “warning”, “scam”, etc. in the description or title.

Besides the characteristics of the unit itself, the type of building it is constructed within is an important variable to much housing research. These types include detached house, duplex, condominium, apartment, etc. One challenge in comparing classified ad data with other sources is the mismatch of the types of building available, even between the RMS and Census. The flag for ‘apartment’ as a type was the most common in the dataset but was found to be applied to many ads that would not conform, for example, the CMHC’s definition of a primary market apartment as a purpose-built structure with three or more units. It is possible that many such “apartment” type ads were selected by the posting user simply because it is an ad in the “apartments” category of Craigslist, or that it is the default class and users do not specify it appropriately.

Square footage can be another generally important variable in comparing dwelling units, particularly in normalizing price among dwelling units of different bedroom sizes where more bedrooms usually means more square footage. Boeing and Waddell (2017), for example, looked at rent/sqft figures which allowed them to use all units to assess price in various metro regions. In the data collected for Metro Vancouver, however, square footage was undeclared for many (40.2%) ads in the dataset. Rather than discard nearly half of the results, a price per bed figure, where bachelor, studio, and loft are considered to have 1 bedroom, was evaluated for certain analytical products.

In addition to the systematic ambiguation of geolocations, users will create other forms of locational error. Some ads were for properties well outside of the Vancouver CMA, including Victoria and at least one internationally located vacation rental. Others will have placed a pin for a precise location, but mistakenly done so over a body of water for a property that was not a boat. Some ads do both and result in point data in the middle of Hudson’s Bay. These errors are easy to correct by clipping the dataset to the appropriate CMA polygon, but it does reduce the volume of data slightly.

The above are some of the most common sources of error that may be grounds to clean an ad from the dataset, and many others abound. It should be noted that what is a “usable” ad depends on one’s research question and what may be sources of error to one study may actually be the object of another, such as one deliberately interested in the nature of deviance in online classified ads. Indeed, formally interrogating these behaviors may actually assist future studies in automating ways of detecting and filtering results.

SECTION 3

What does the Data Say about Vancouver: Results from Oct-Feb

3.1 Aspatial Analysis

3.1.1 Total Ad Volume

	October	November	December	January	February	Total
Bachelor/Studio	87	95	16	149	122	469
1br	987	894	101	1493	999	4474
2br	1068	979	111	1409	1199	4766
3br	392	290	45	450	369	1546
4+br	204	211	24	255	231	925
Private room*	8	16	159	2235	2405	4823
Total	2746	2485	456	5991	5325	17003**

Table 1: Breakdown of ad volume by month.

** Private rooms were not categorically scraped until January. Low counts found in previous months were either scraped in January or mis-posted to the apartments category and corrected in data cleaning.*

***Some ads were created in September resulting in a lower count in this table than elsewhere.*

Market activity in the “apartments” category – not including private rooms - demonstrated an ad volume between 2,500 and 2,700 in the autumn months, with a sharp decline below 500 in December, a peak of 3,800 in January, then a seeming return to the high 2,000’s in February. Analysis was truncated at February as data collected from March onward may show pattern disruption from the COVID-19 pandemic, although establishment of a pre-pandemic annual pattern – if any existed – will require procurement of archived data this study.

Without other years or a longer frame to consider, the apparent low volume in December was the only outstanding temporal event. This may be explained by an aversion to engage in housing hunts during the holiday season, to show or view units, and the travel many people are doing during the month. Some of the accounted low volume, either originally posted in December or earlier, was reposted through to early 2020 and would have been counted either in January or February due to the deduplication policy described in section 2.3.

Interestingly, the private rooms category had the most postings of any type in the months it was deliberately scraped. Assuming it was equally dominant in other months, it may likely have accounted for over ten thousand ads in the five-month period, bringing the grand total well over twenty thousand. As mentioned previously, it is composed of a mix of very small single bedroom units and ads where single bedrooms in larger units – roommate replacement – are posted. This suggests that many individuals are finding accommodation without turning over leases while paying rents closer to what is reported in the Census and RMS; a kind of sub-unit housing market.

> **Interestingly, the private rooms category had the most postings of any type in the months it was deliberately scraped. Assuming it was equally dominant in other months, it may likely have accounted for over ten thousand ads in the five-month period.**

3.1.2 Volume by Bedroom Count

Comparison between the volume of ads and reported proportion of units in the RMS primary market can be seen in table 2. Comparing columns three and four, there is a strong difference between the shares of bachelor, one bedroom, and two- or more bedroom units. Although a breakdown of condominium rental units by bedroom type is not provided in the RMS, it may be assumed that the proportion of three- or more bedroom units is not much higher than the primary market, if at all.

The 2,503 ads for units of three or more bedrooms is more than half of the 4,146 accounted units of that size range in the primary market according to the 2019 RMS (CMHC 2019). Three-bedroom ads alone more than triple the representation of three or more bedroom units in the RMS count. This stands converse to the proportion of one-bedroom ads being less than half of the accounted proportion of units in the RMS, about one third versus a 60% majority.

Multiple explanations for this difference are possible, the simplest being the different composition of the non-condominium secondary market

	Count	Proportion of Ads	Proportion of Whole Unit Ads	CMHC share of Primary Market (2019 RMS Table 3.1.3)	Total Bed-rooms	Proportion of Bedrooms
Bachelor	475	2.8%	3.9%	11.0%	475	1.7%
1br	4,518	26.3%	36.6%	60.4%	4,518	15.8%
2br	4,839	28.2%	39.2%	24.9%	9,678	33.9%
3br (3+ col. 3)	1,565	9.1%	12.7%	3.6%	4,695	16.6%
4+br	938	5.5%	7.6%	-	4,342	15.1%
Private room *	4,823	28.1%	-	-	4,823	16.9%
Total	17,158		12,335		28,807	

Table 2: Breakdown of ad volume by bedroom count vs CMHC Primary Market proportions.
**Private Rooms were not categorically scraped until January. Actual counts expected to be ~2x higher.*

wherein units in detached houses offer more bedrooms, even when partitioned into multiple units. Conversely, in multi-family structures, one-bedroom units are common while very few three-bedroom units are known to be built, let alone four or more besides rare penthouse cases. Another explanation may point to higher turnover during the study period in three or more bedroom units and fewer single bedroom units turning over, although this departs severely from the RMS reported turnover rate for the primary market wherein that type demonstrated a slightly lower rate than others. Turnover rates would need to vary drastically from month to month, or be consistently higher in the secondary market, to support this explanation, provided all counts are reasonably accurate.

When reviewing the number of bedrooms each type accounts for, one can approximate the number of people housed in a transaction and in volume. This can vary when considering couples or multiple children sharing single bedrooms as well as households using bedrooms as guestrooms, offices, or other functions besides sleeping at least one person. Assuming, for simplicity, that one bedroom sleeps one person, the above data suggests that, of the nearly twenty nine thousand rooms captured in ads from October to February, one bedroom and bachelor suites only accounted for seventeen percent of people housed

These observations carry an assumption that the breakdown of ads by bedroom type can approximate the composition of rental stock in Vancouver or at least represent a typical annual behavior pattern in the market itself– an assumption which should be taken with caution for a 5 month timeframe based on one source.

despite being 30% of the ad volume. Two-bedroom units accounted for a leading proportion of ads and bedrooms while composing a significant, yet inferior inventory share to one-bedroom units in the RMS. As will be seen later, two-bedroom units appear to be reasonably present in both dense central areas as well as those dominated by detached and duplex housing, whereas one-bedroom and three- or more bedroom units exhibit a reciprocal presence in such areas. This may reflect a larger presence two-bedroom units in the secondary market.

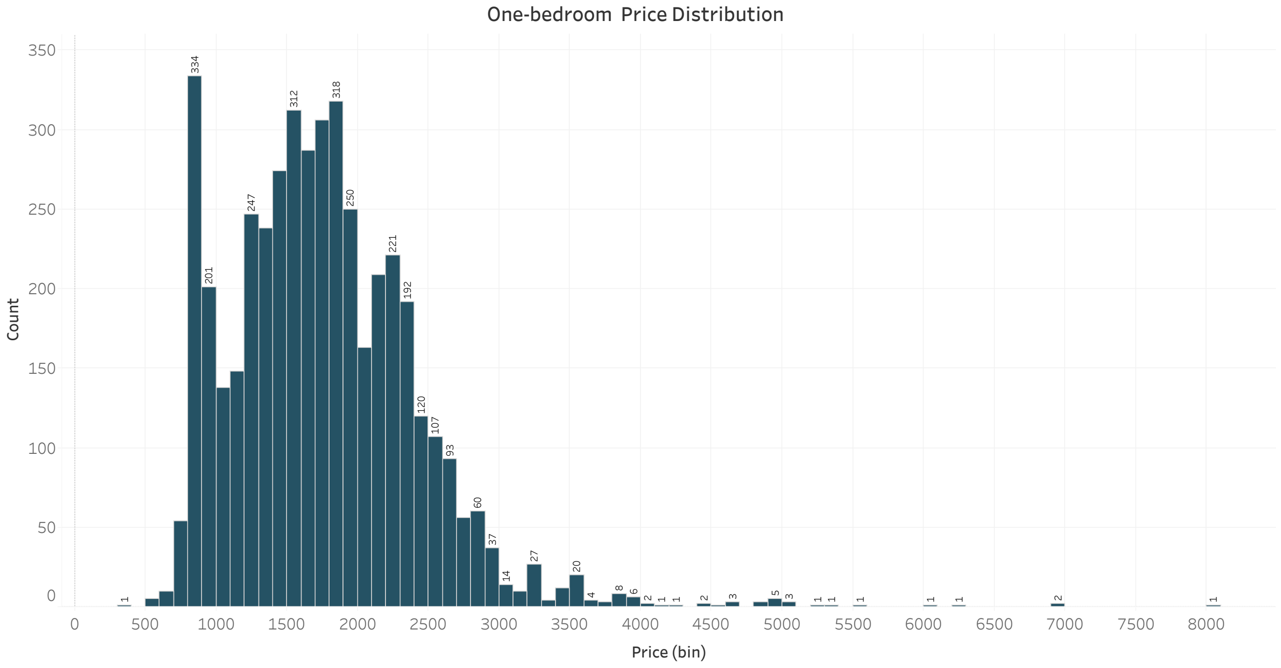
These observations carry an assumption that the breakdown of ads by bedroom type can approximate the composition of rental stock in Vancouver or at least represent a typical annual behavior pattern in the market itself– an assumption which should be taken with caution for a 5 month timeframe based on one source.

3.1.3 Breakdown of Prices

	Mean Price	Median Price	Q_1 Price	Q_3 Price	Price Skewness	Mean PPB	Median PPB	Q_1 PPB	Q_3 PPB	PPB Skewness
Bachelor	1,532	1,500	1,250	1,800	0.89	1,532	1,500	1,250	1,800	0.89
1br	1,736	1,695	1,295	2,100	1.37	1,736	1,695	1,295	2,100	1.37
2br	2,405	2,100	1,600	2,755	6.12	1,202	1,050	800	1,376	6.12
3br	3,052	2,500	2,150	3,300	4.76	1,017	833	717	1,100	4.76
4+br	4,452	3,500	2,800	4,950	4.63	974	800	650	1,094	4.02
Private Room	770	750	600	900	0.92	770	750	600	900	0.92

Table 3: Prices and price per bed of bedroom types.

All bedroom types exhibited a right-tailed distribution (fig. 8), likely representing a small amount of units that, for any combination of reasons - large floorspace, new construction, location, or luxury - can exceed most other ads by a factor of multiples. Resultantly, all averages were higher than their corresponding median, demonstrating a considerable premium for luxury units while most units land in a group towards the lower end of each price range.



The trend of count of Price for Price (bin). The view is filtered on Price (bin), which keeps non-Null values only.

Figure 8: Distribution of one-bedroom unit prices, exhibiting right skewness and a median rent of $2,100.

All types except one-bedroom were found to be unimodally priced. There may be a sub-group in the lower $750 to $950 dollar range, perhaps the very small units that are technically one bedroom but sized and priced such that most are found in the private rooms category, that composes this disruption to an otherwise simple distribution.

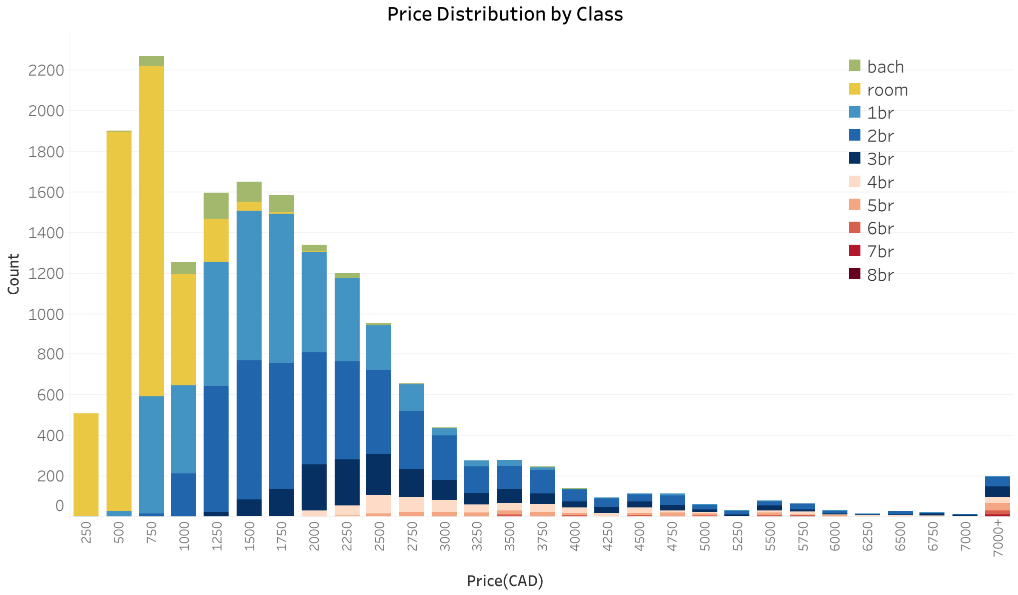


Figure 9: Stacked Histogram showing distribution of unit prices by bedroom type.

Prices generally increased with bedroom count *(fig. 9).* Conversely, price per bed decreases as unit size increases *(fig. 10),* suggesting a premium for exclusive space and amenities. One can see, excepting private room listings, how each price and price per bed bracket is composed in an overall reversion of each class' place on the list.

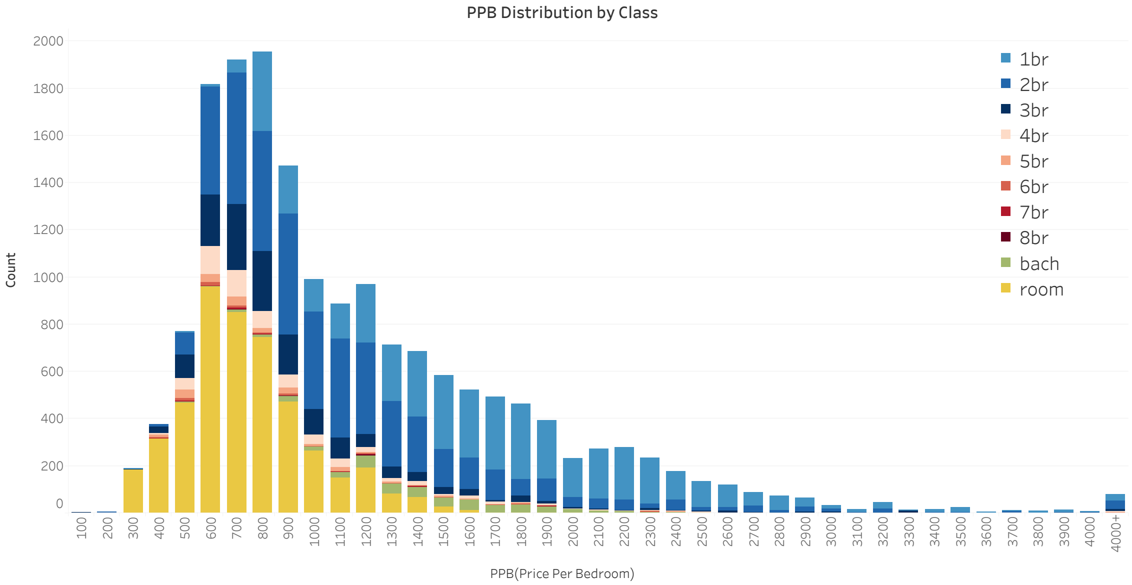


Figure 10: Stacked Histogram showing Price per Bed. Private room and larger units compose lower end of range while one and two bedroom units fill out higher price range per bed.

3.1.4 Price per Bed: Relative Affordability in Small and Large Places

Based on the same median renting figure used in section 4.3.1 (CMHC 2019), a median two earner household where both earners made equal income would require a PPB of $764.32 (2019 dollars) for a two bedroom unit with utilities included to be spending less than 30% of their before-tax income on household expenses. The private rooms category, although ads are not for whole two-bedroom units, is the only one where median PPB falls below this level, followed by units with four or more bedrooms at $800 PPB. Independent of location, these two categories appear to be where renters willing to live modestly or share a dwelling can find it more affordably.

One possible factor in the lower PPB of the private rooms category lies in the partial turnover of a unit for ads it applies to. Since the lease is sustained in the name of at least one staying tenant, essentially counting the unit as occupied under a lease that may have been counted in previous RMS or Census, the advertised share of rent is likely to reflect older and thus lower rents in a region where new rents have been consistently increasing. These scenarios are perhaps more likely to occur in units with more bedrooms that are more likely, themselves, to be shared among income-earning adults who may part ways, leaving some in the unit while others depart. If true, this relates private rooms to units with three, four, or more bedrooms that also exhibit the next lowest PPB in table 3. If rental units with many bedrooms inhabited by non-census family households are consistently circulated on a partial basis through private room ads, the number of units of these sizes would be under-represented in the apartments category as they rarely completely turn over at once; there may, therefore, be an even higher proportion of units with four or more bedrooms in the secondary rental market than represented in table 2.

By analysis of ad description, and reference to the photographic content of any standing ads during analysis, some private room listings were identified as standalone suites. Technically these were one-bedroom suites, yet very small and comprised of little more than a modest bedroom, bathroom, and a small hallway with which to access a dedicated entrance. Being so structurally minimal, and often recommended for students in their description, these units may help explain the general price point of group. These descriptive characteristics were only peripherally noted during dataset processing, however, and would require dedicated scrutiny to verify the real proportion of ads fitting this description.

3.2 Geography

Sample size and Distribution by Bedroom Type and PPB

Where are the rental units on Craigslist located?

In general, ads were distributed more heavily in dense, central areas such as downtowns and transit station areas, with downtown Vancouver exhibiting a particularly high concentration of ads, particularly in lower bedroom types and becoming nearly absent of four- or more bedroom ads. This is sensible and, at first glance, it appears that there is a gradual spread from dense nodes and corridors as one moves up from one-bedroom units to four and more. This not only suggests a pattern in the built form but, as per the note about ad volume per geographic unit in section 1.3, begins to indicate how summary figures in FSA's is variably reliable.

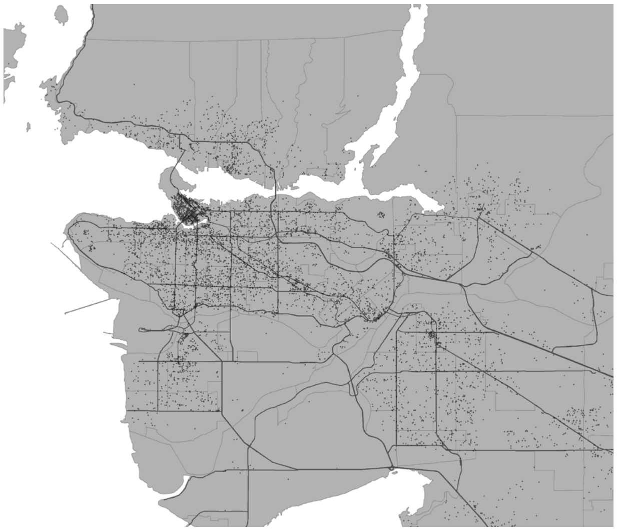


Figure 11: General distribution of all ad types across Vancouver CMA.

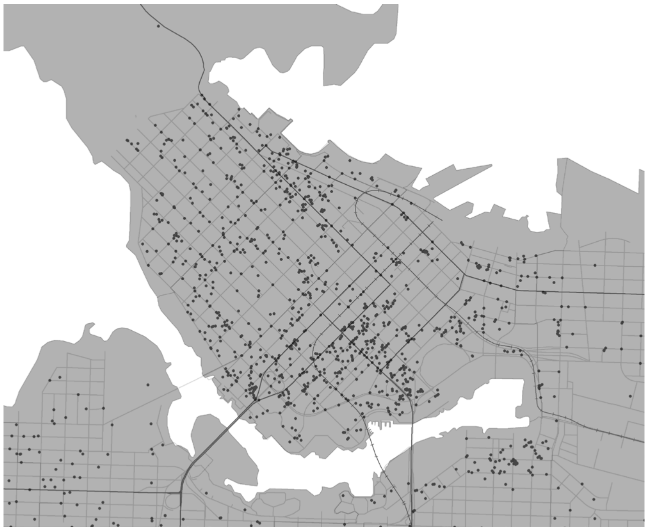


Figure 12: General distribution of all ad types across City of Vancouver

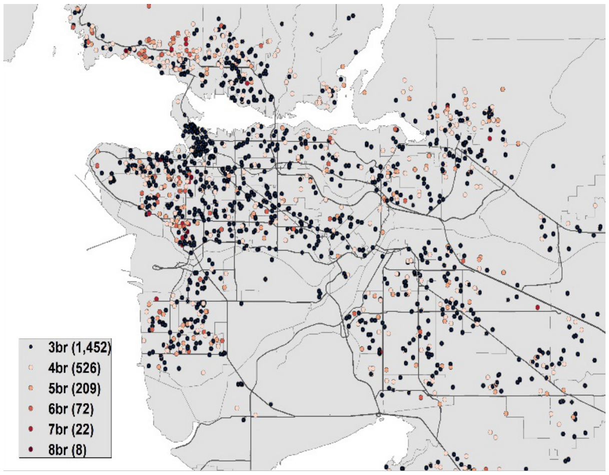


Figure 13: Distribution of large bedroom types across Vancouver CMA

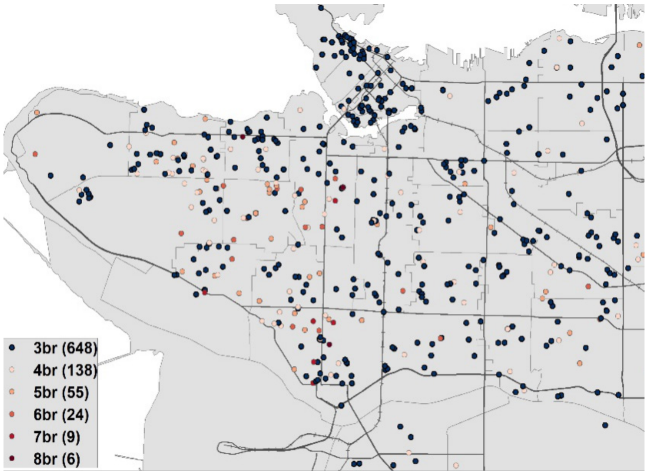


Figure 14: Distribution of large bedroom types across City of Vancouver

Other patterns of note include the relative concentration of three bedroom units in Coal Harbor and Yaletown versus the central West End, which seems be somewhat absent of that type despite, as will be discussed in the following section, being one of the more affordable districts in the downtown peninsula.

As discussed in section 3.1.3, those willing to share space for more affordable accommodation may find lower prices per bedroom in large houses with four or more bedrooms. This class of units appears west of Granville Street in the neighbourhoods most proximate to the University of British Columbia. It may be that this is an artefact of the student housing market. In other neighbourhoods dominated by detached dwellings, a relatively even visual distribution of listings appears.

General Median Rent by FSA

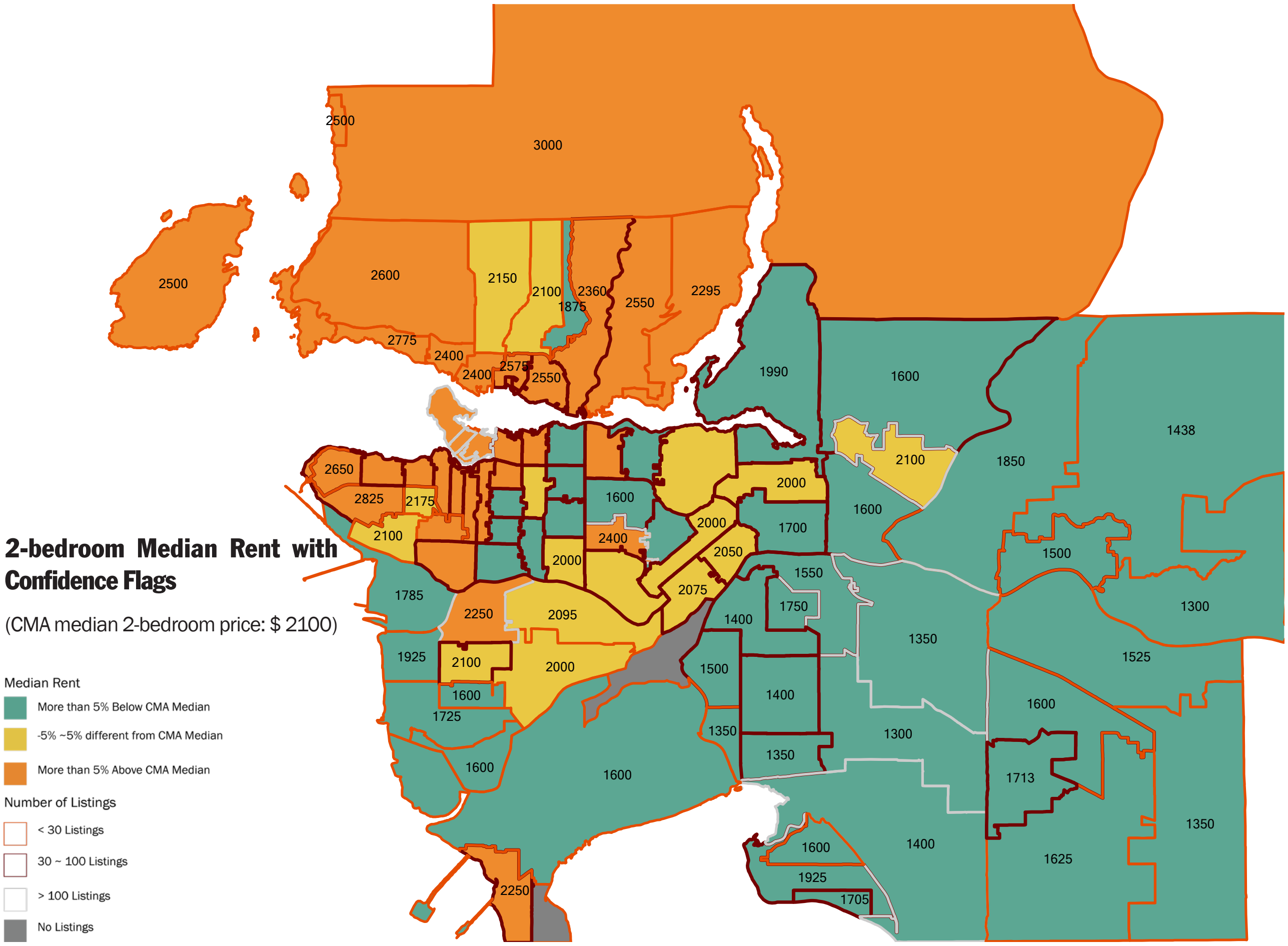


Figure 15: Median rent of two-bedroom units across Vancouver CMA with confidence flags. CMA median two-bedroom price: 2100

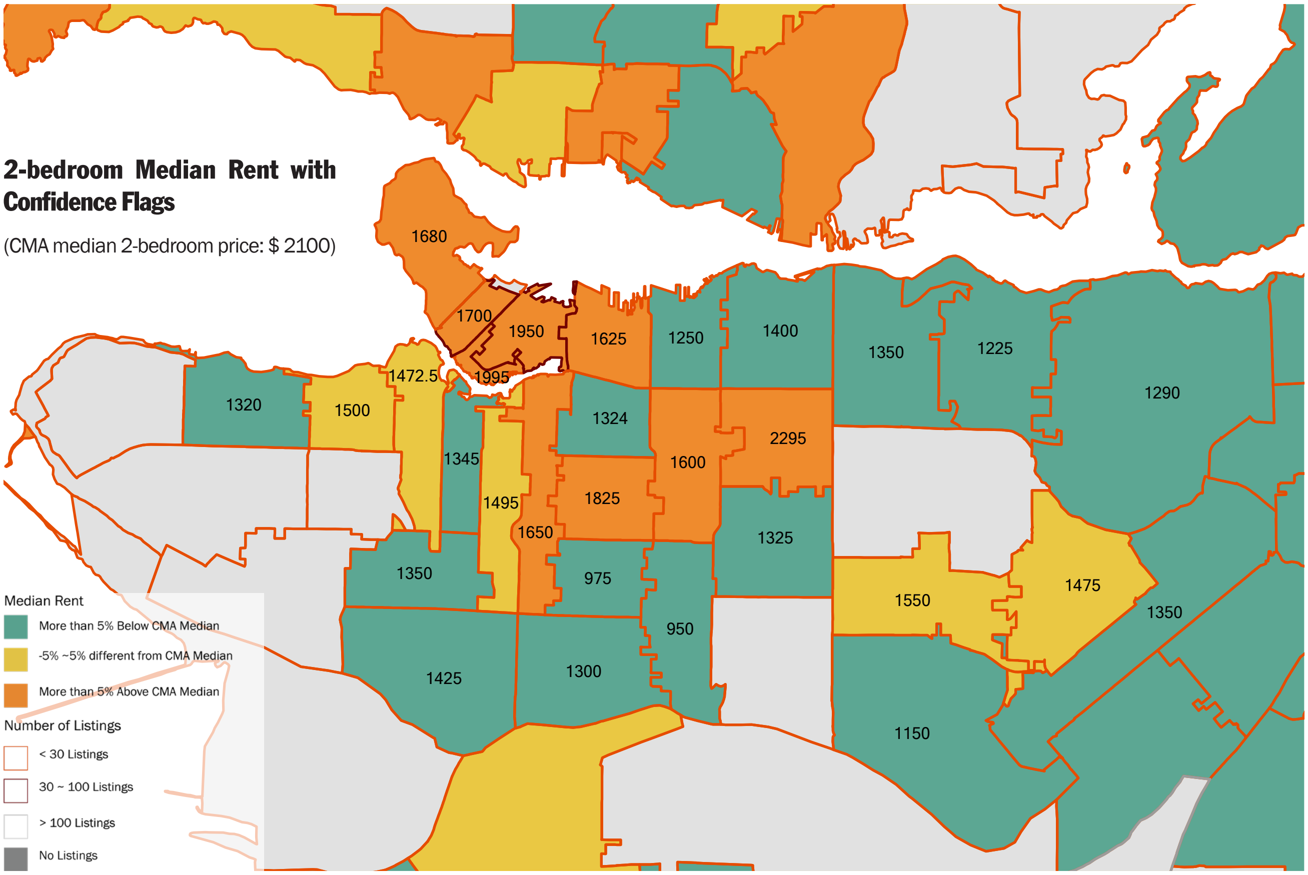


Figure 16: Median rent of bachelor units across City of Vancouver and Burnaby with confidence flags. Nearly all outlines indicate low ad volume and confidence.
CMA median bachelor price: 1500

The maps below show two matters: The median price of a unit type per FSA through polygon colour, and a cautionary flag for low sample counts in the outline colour of each FSA. One can observe both the patterns of price as well as how reliable individual figures are.

As bachelor/studio units are generally found in low numbers in both the dataset's numeric tables and the CMHC's data (2020a), this type unsurprisingly fails to muster more than thirty data points in most FSA's. Meanwhile, two-bedroom units are abundant in much more of the region's FSA's, enabling more confident analysis. In any case, a general pattern of price follows most types, with most of the south and eastern parts of the lower mainland exhibiting median rents at or below the CMA median, transitioning through middle-belt cities like New Westminster and Richmond to regions more expensive than the regional median in central, western, and north Vancouver.

3.3 Comparison

3.3.1 Affordability: 2br Prices vs Median renter income

The first comparison focuses on affordability, juxtaposing median rents with renter incomes. Rent results were compared to a CMHC table of median renter household incomes by CMA (2019). The 2017 median before-tax income for renter households was adjusted to 2019 dollars and common thresholds of low-income cut-off were calculated to measure the affordability of each FSA's median rent. Using the average CMA household size of 2.2 (Statistics Canada 2017), a household size of two income earning adults seeking a two-bedroom unit was selected as the model scenario. Common compositions covered by this might cover two non-family roommates or a couple with one child.

	Monthly Income	30% Cut-Off	50% Cut-Off	70% Cut-Off
Household	**$5,095.48**	$1,528.64	$2,547.48	$3,566.83
Individual (PPB)	**$2,547.74**	$764.32	$1,273.87	$1,783.42

Table 4: Model median renter household and individual income with housing expense cut-off's in 2019 dollars. Source: Real Median Total Household Income - Before Taxes (CMHC 2019)

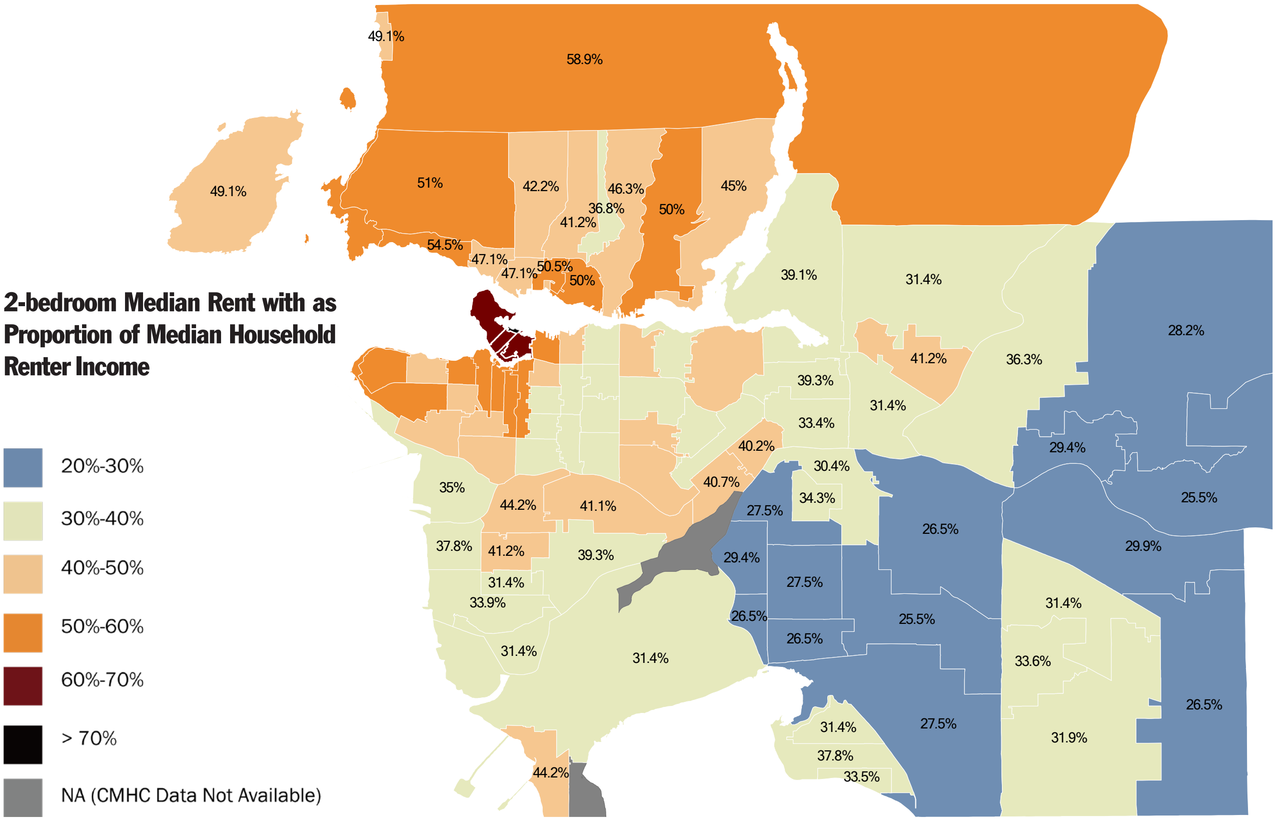


Figure 17: Two-bedroom median FSA rent as proportion of CMHC median household renter income. Vancouver CMA. Source: Real Median Total Household Income - Before Taxes (CMHC 2019)

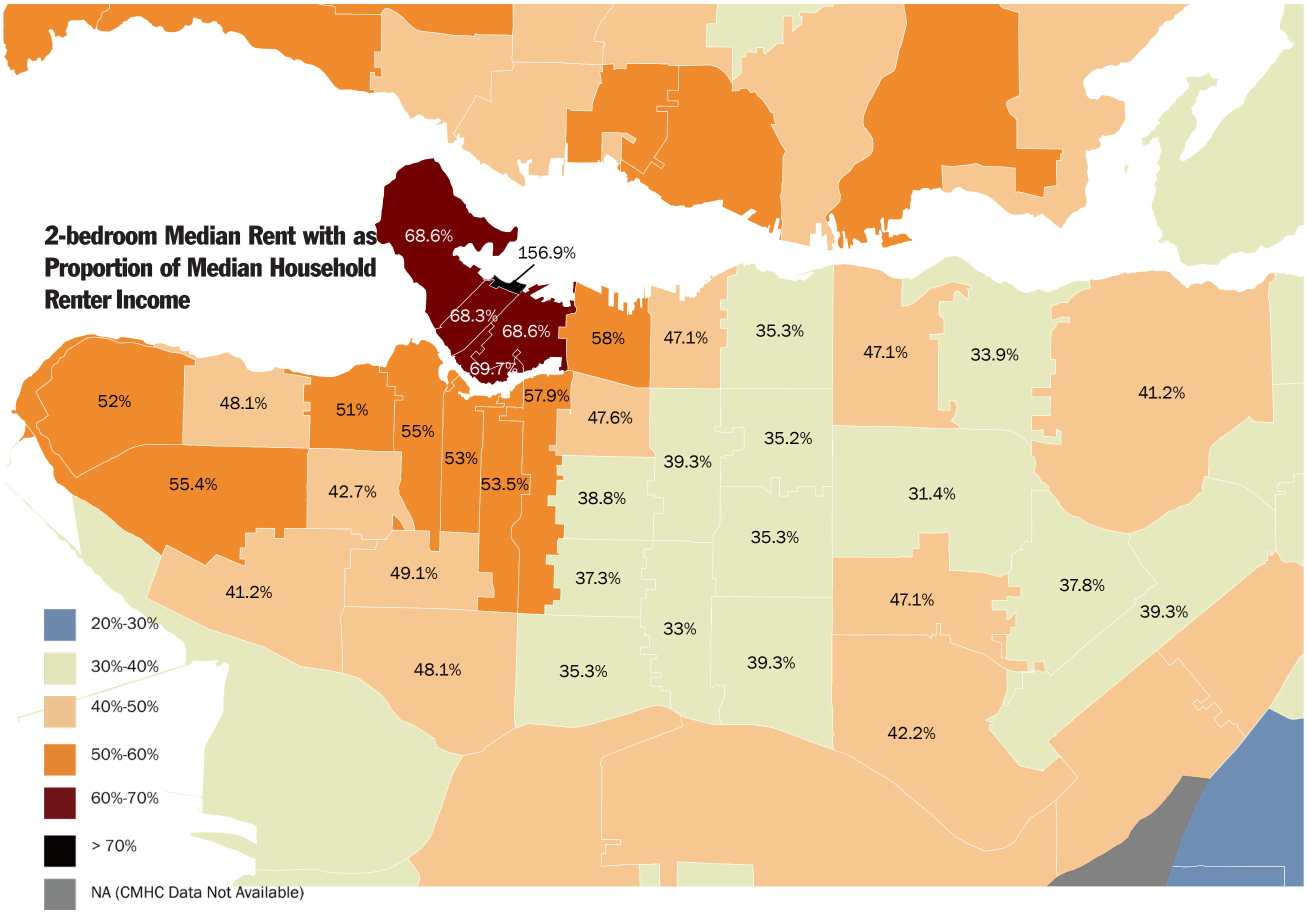


Figure 18: Two-bedroom median FSA rent as proportion of CMHC median household renter income. City of Vancouver. Source: Real Median Total Household Income - Before Taxes (CMHC 2019)

The data suggests that only median two-bedroom prices in some Surrey, Maple Ridge, and Langley FSA's will cost less than 30% of the household's monthly income. As one moves west into Burnaby and Richmond, the proportion of rent to income climbs consistently over 30%. Only a few FSA's in south and east Vancouver exhibit median rents between 30 and 40% while all others are more expensive. The model household would consistently be spending more than half of its monthly income on rent west of Main Street and north of the Georgia Straight. The most expensive FSA's, Yaletown and Coal Harbor, reach 70% and 158% respectively of the model renter household's income. It should be added that utilities are not necessarily included in these rents, therefore some may actually cost more per month.

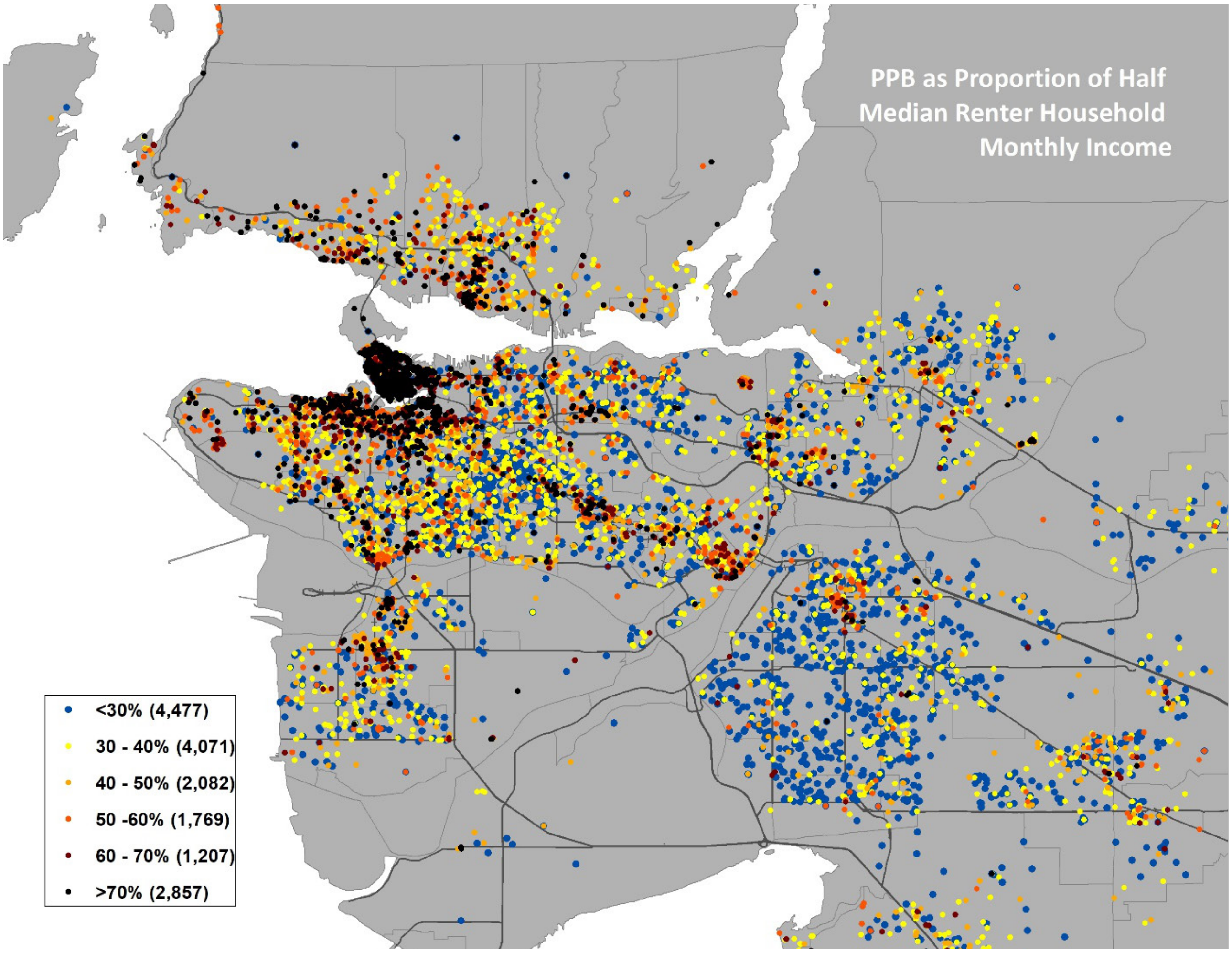


Figure 19: Price per bed as proportion of half CMHC median household renter income. Counts in brackets. Vancouver CMA.

Source: Real Median Total Household Income - Before Taxes (CMHC 2019)

If the median renter household income is assumed to house two equal income-earning adults, what do their opportunities as individuals look like across the region? Figure 19 shows the PPB of each posting relative to that model individual's monthly income – $2,548 in 2019 dollars – where the affordable opportunities are, and how many opportunities there are in volume across the Metro Vancouver region. In the frame of this map, there were nearly 4,500 opportunities for the model renter to spend less than 30% of their before-tax income on rent, a large visible share of which dominated Surrey's less central areas. This stands in contrast with nearly 12,000 opportunities priced over 30% of the model monthly income, one quarter of which would demand more than 70%.

3.3.2 Market Segmentation Vs. Padmapper's new leases

The figures here compare the medians reported by Padmapper's February 2020 Rental Market Report (2020) with medians from Craigslist data for the two municipalities in the Vancouver CMA ranked by Padmapper: Vancouver and Burnaby. To be careful, this is a single month's snapshot. Any conclusions from this kind of analysis will likely require a more longitudinal look than available here, but it does offer an introduction to how comparable the sources are and what discrepancies might mean.

Differences between the two sources for the same month may suggest a mix of factors regarding sample size and user self-selection bias. City Observatory (2018) cautions that sample size and composition can cause median estimates to make prices appear more volatile than they are, with divergence of trends between bedroom types being an indicator of such issues. If, over many months, the difference between Padmapper's rental reports and data from Craigslist vary wildly, including fluctuations between positive and negative differences, this may confirm City Observatory's critique in the context of the Vancouver market. This does assume Craigslist's data pool is large enough to anchor a representative median of the full market, and that is not yet certain. A large sample of the market should change relatively smoothly unless the whole market does, in fact, change erratically. If monthly differences between the platforms sustain a relatively stable direction and magnitude, it may suggest a degree of price-point segmentation or, due to a high prevalence of cross-posted ads on both platforms, that Padmapper simply hosts a subset of the ads available on Craigslist.

	One Bedroom			Two Bedroom		
	Padmapper Feb 2020	CL Feb 2020	CL – Five Months	Padmapper Feb 2020	CL Feb 2020	CL Five Months
Burnaby	$1,760	$1,550 -10% (101)	$1,500 -15% (399)	$2,350	$2,100 -11% (127)	$2,100 -11% (483)
Vancouver	$2,150	$2,100 -2% (468)	$2,050 -5% (2242)	$2,990	$2,995 0% (521)	$2,750 -8% (1974)

Table 5: Comparison of median rents from Padmapper Rental Market Report vs. Craigslist data for February and the full five month study period. Differences between sources also shown as percentages of Padmapper figure. Sample size shown in brackets.

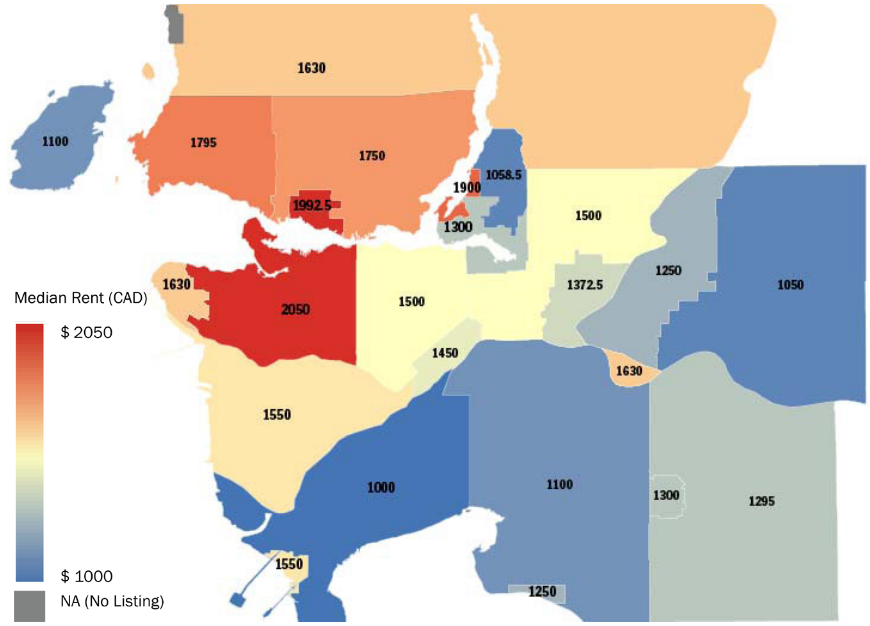


Figure 20: Median one-bedroom rent by Census Subdivision.

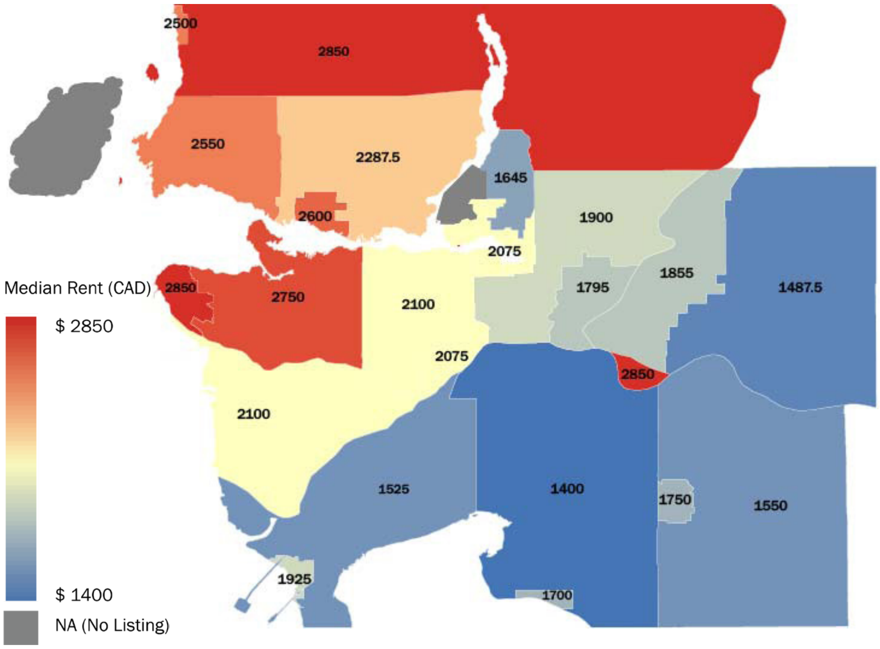


Figure 21: Median two-bedroom rent by Census Subdivision.

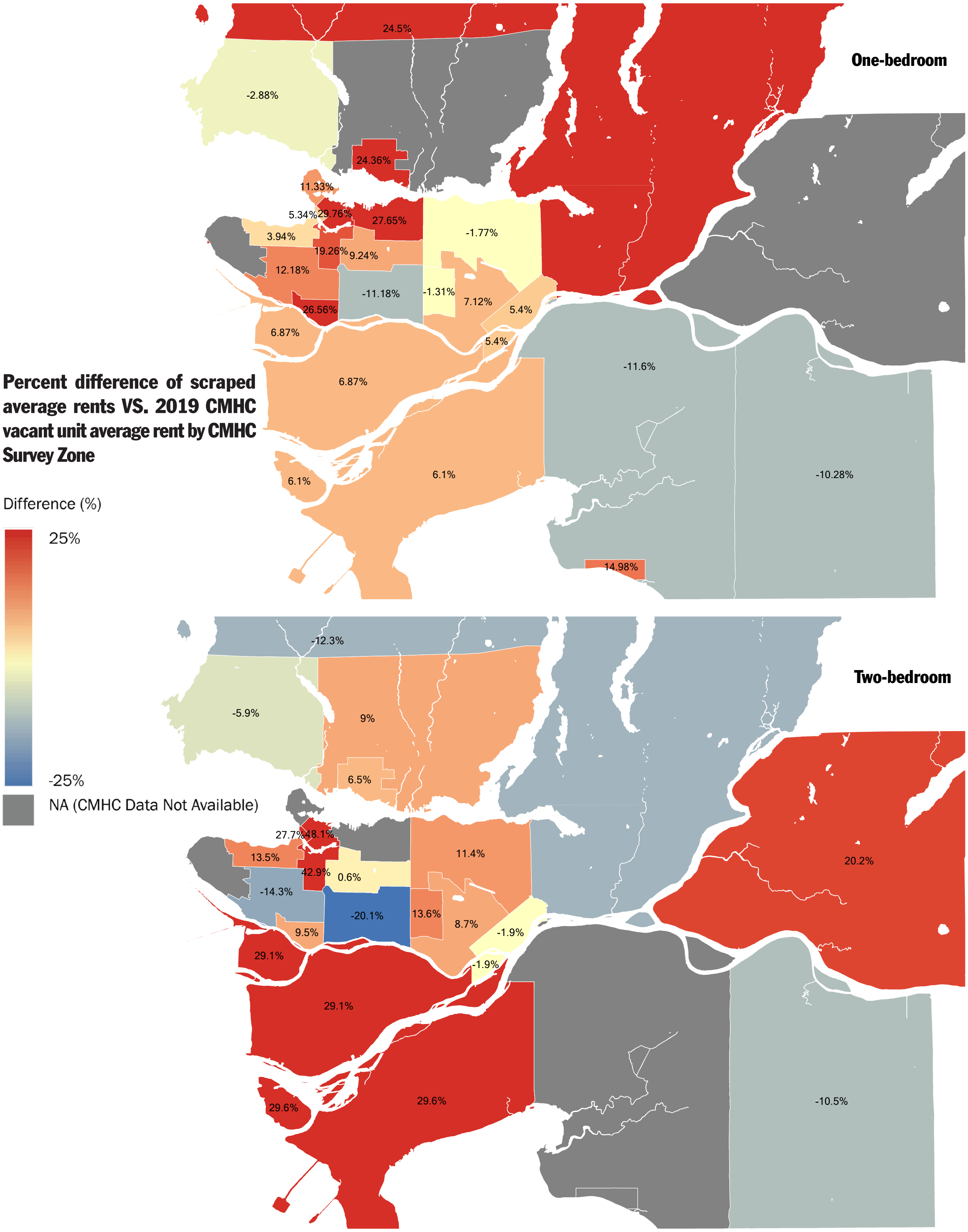
In general, the Craigslist data consistently registered lower medians than Padmapper's report, excepting one $5 premium on two-bedroom units in the February-to-February comparison. Craigslist median rents for the City of Burnaby in February were about ten percent lower than reported by Padmapper for both one- and two-bedroom units yet the two platforms tracked somewhat closer for the City of Vancouver. One explanation may be that the two platforms may sample the same areas and submarkets of Vancouver proportionally while ads on Padmapper focus on Burnaby's more expensive areas than Craigslist. Another may be the aforementioned question of sample size. A third may be the inclusion of short-term rentals in Padmapper's estimates. Without knowing Padmapper's sample size and method, it is difficult to tell. There may be other factors at play this difference which, at this single month timeframe, should be taken as speculative.

3.3.3 “Moving Penalty” Comparison

The figures below compare the average rent of one- and two- bedroom units on FSA geography to the primary market average rents of occupied and vacant units according to the CMHC (2020b). They explore the cost a household may incur to move from an average-priced unit in the Primary market to a similar unit in the same CMHC survey zone as posted on Craigslist’s average. This is referred to as the “moving penalty” by Vancouver-based data scientist, Jens von Bergmann, who has reviewed the CMHC rents (Mountain Math 2018). It is important to highlight the operating meaning of a vacant unit in this analysis: a unit that is vacant at the time of data collection or occupied but will be available for new tenant in a short time. It is also important to remember that the Craigslist data represents an unassessed mix of primary and secondary markets, so some of the comparative differences may be explained by the prevalence of the secondary market not factored into the CMHC figures.

CL Prices vs. Vacant Unit Averages

Do the prices found on Craigslist match CMHC vacant averages? This may suggest how the secondary market changes the affordability picture. There is an irregular pattern of difference between the two averages across the region. Some zones exhibit higher average prices in the Craigslist dataset, others with lower averages with only some zones exhibiting similar differences for both types. Downtown Vancouver, for example, showed a consistent premium of 30% and 48% for one- and two- bedroom types respectively on Craigslist while southeast Vancouver reported 11% and 20% lower average prices than the CMHC figures for the same respective types. Kerrisdale in the southwest of the city exhibited a curious 12% premium for one-bedroom units yet a 14% lower two-bedroom average. Generally, the differences in price may reflect differences of quality and age between the primary and secondary markets and how they may vary by locale. Marpole, characterized by many walk-up apartment buildings, exhibits one of the smallest differences in price between the CMHC’s vacant and occupied one-bedroom averages, yet one of the most drastic premiums when moving to Craigslist. This relationship eases considerably in two-bedroom unit comparisons. Given how right-tailed the distribution of prices for each bedroom type are in the scraped dataset, the extreme difference in average prices downtown are not too surprising considering the apparent proportion of luxury units in the secondary market there.



Percent difference of scraped average rents VS. 2019 CMHC vacant unit average rent by CMHC Survey Zone

One-bedroom

Two-bedroom

Figure 22 and 23: Percent difference of scraped average rents versus 2019 CMHC vacant (primary market) unit average rent by CMHC Survey Zone. Survey zones in grey did not have CMHC averages published.

Top: one-bedroom. Bottom: two-bedroom.

Source: Average Rents - Vacant and Occupied (CMHC 2020b)

Craigslist Prices vs. Occupied Unit Averages

In all cases except one-bedroom units on the University Endowment Lands, the Craigslist average was found to be substantially higher than the CMHC occupied unit average. The moving penalty was roughly observed to be larger in more central, higher-priced zones, with larger penalties for two-bedroom units than one. While moving penalties in more outlying cities could expect to demand several hundred dollars more for new leases in their area, two-bedroom households could face up to a $1,548 premium to move from an average existing lease in a purpose-built downtown apartment to a two-bedroom unit elsewhere in the downtown core. Steeper moving penalties may reflect areas experiencing more price-pressure than others. Southeast Vancouver, in addition to a lower Craigslist average price than the CMHC's for both vacant room types, also exhibited the lowest moving penalties for two-bedroom units and a disruption to the generally higher penalties from the exurban periphery towards the core. This may be due to any combination of lower rent increases in the area overall in recent years, lower relative prices in the secondary market which only influence the Craigslist averages, or other factors.

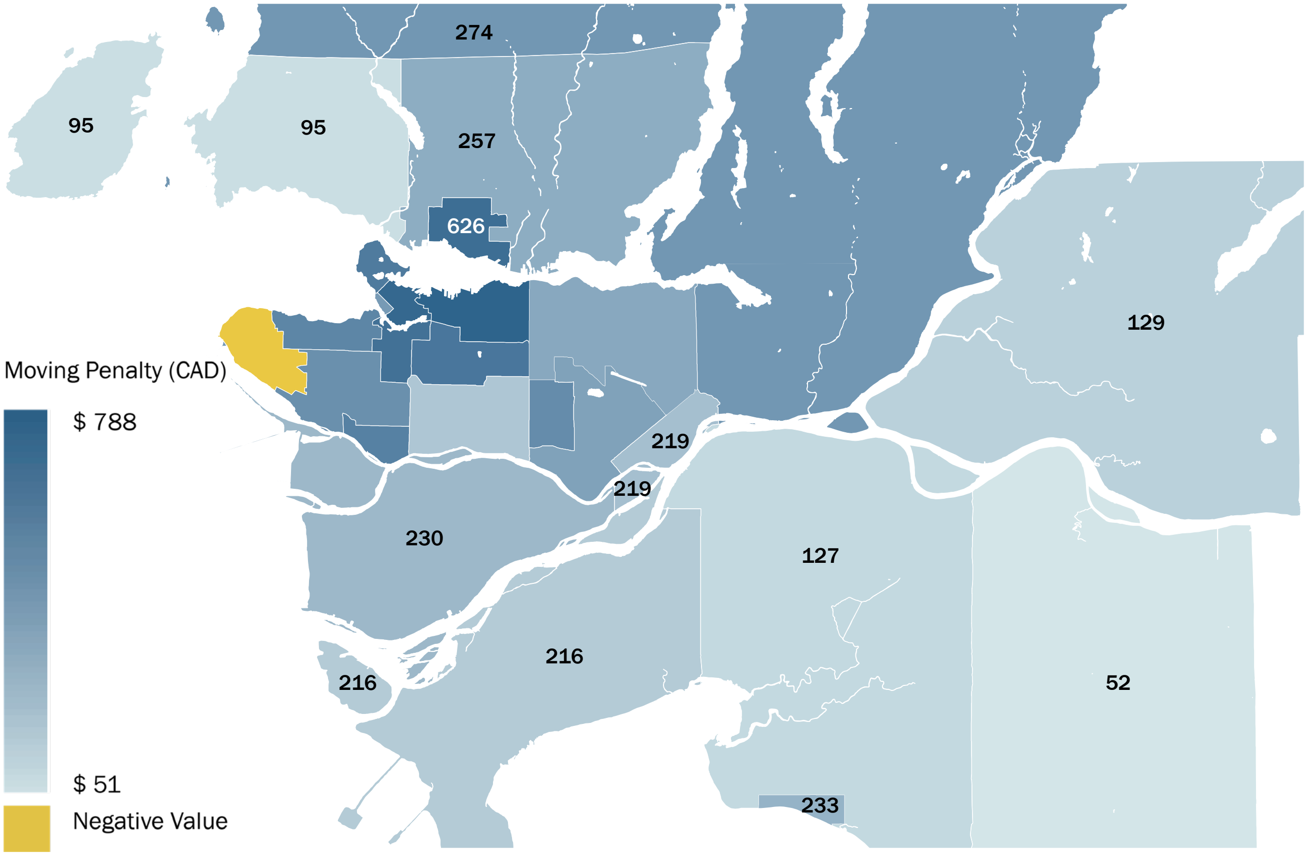


Figure 24: Moving penalty – Price premium of average Craigslist one-bedroom rents vs CMHC occupied primary market unit. Vancouver CMA

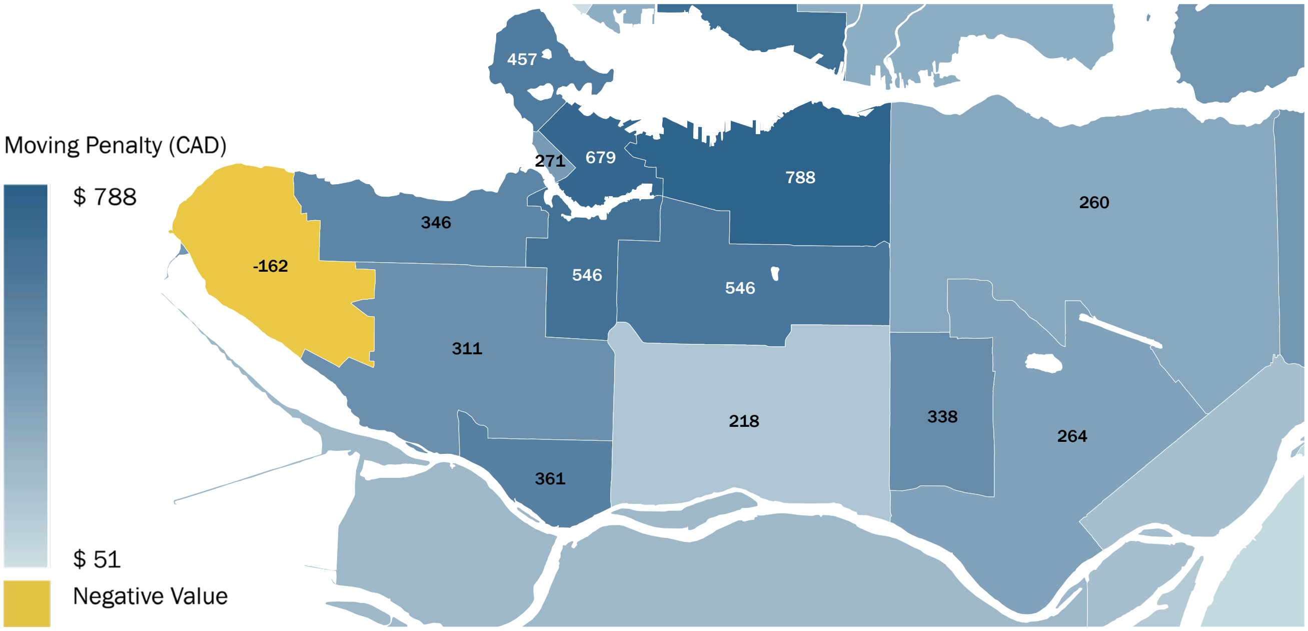


Figure 25: Moving penalty - Average one-bedroom rents vs CMHC occupied primary market unit. City of Vancouver.

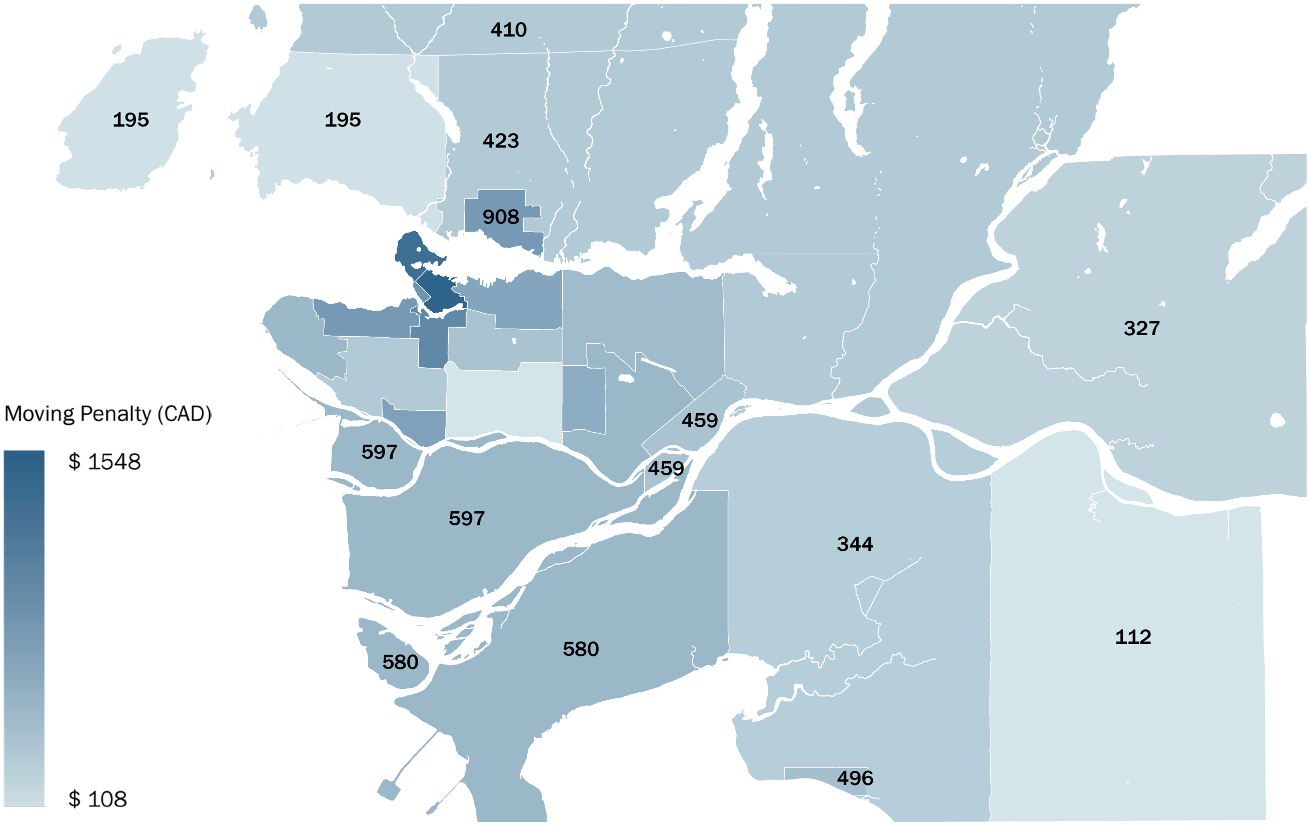


Figure 26: Moving penalty - Average two-bedroom rents vs CMHC occupied primary market unit. Vancouver CMA

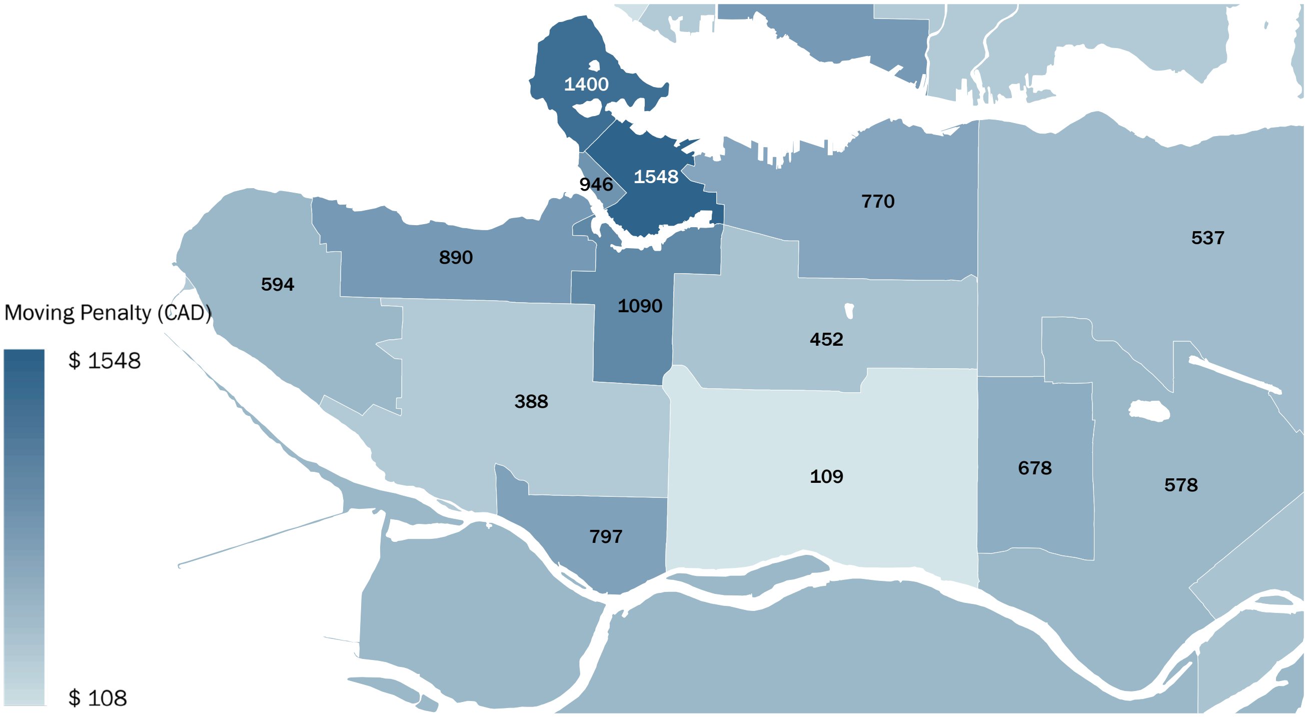


Figure 27: Moving penalty - Average two-bedroom rents vs CMHC occupied primary market unit. City of Vancouver.

SECTION 4

Closing Words and Further Research

As in most matters, and especially from this exploratory first step, more research and development is needed. While the body of data is relatively short and the depth of analysis wide and shallow, the aim was to point out the kinds of questions this kind of data can begin to address. We have outlined some of the more immediate ways these methods can be enriched and streamlined to produce stronger understanding of rental markets through VGI.

The research can be furthered is expansion in several dimensions. Sustaining collection and analysis of the data over time will enable meaningful long-term trends to be characterized, including any impacts the COVID-19 pandemic has had on Vancouver's rental market. Acquiring historical classified data will help understand how Vancouver's affordability crisis has affected the rental universe. Collection over more platforms such as Kijiji and Facebook Marketplace may help draw more confident pictures of the realities of renting, including how the market may segment among platforms with unique biases. Affordability and housing are not topics relevant to Vancouver alone, so expanding this research to other metropolitan areas is desirable.

The process of cleaning and correcting the dataset involves intersectionally scrutinizing multiple variables as well as language in the description. This currently stands as a labour-intensive manual process and optimizing it to clean a month's worth of data has proven to be an art form in software use. Appendix D details standing procedures, specifics of the complications encountered in the data, and further insights on extending research for those wishing to continue and iterate this work. If the scope of study is to expand as described in the previous paragraph, however, sophisticated algorithm development and natural language processing may be necessary to avoid exorbitant manual processing, and Appendix C may prove useful in defining the needs of automation as well as manual processing.

Spatial analysis is compromised by varying locational accuracy among ads. Developing an address-to-geolocation disambiguation process on ads may reliably create a higher-accuracy location for ads with declared addresses. The volume of low-accuracy results can be reduced and spatial analysis will likely benefit.

To improve comparability with CMHC and Census data, information in the type, description, and title fields could possibly be processed into reliable building types that would allow researchers to assess which CMHC market category an ad may fit into or, alternately, the structure types defined by the Canadian Census.

Appendices

Appendix A: References

Boeing, G., & Waddell, P. (2017). New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings. Journal of Planning Education and Research, 37(4), 457–476.

City Observatory. (2018, May 30). Caveat Rentor. https://cityobservatory.org/caveat_rentor/ Gehl, J., & Svarre, B. (2013). How to Study Public Life. Washington: Island Press.

Laney, D. 2001. 3D Data Management: Controlling Data Volume, Velocity, and Variety. Report 949. META Group. Mountain Math. (2018, November 28). https://doodles.mountainmath.ca/blog/2018/11/28/moving-penalty/ Padmapper (2020, various dates). https://blog.padmapper.com/canadian-rent-trends

CMHC (2018). 2016 Rental Market Report. Retrieved from: https://www.cmhc-schl.gc.ca/en/data-and- research/publications-and-reports/rental-market-reports-major-centres

CMHC (2019). Real Median Total Household Income (Before Taxes). Retrieved from: https://www.cmhc-schl.gc.ca/en/data-and-research/data-tables/real-median-after-tax-household-income-renter-households

CMHC (2020a). 2019 Rental Market Report. Retrieved from: https://www.cmhc-schl.gc.ca/en/data-and- research/publications-and-reports/rental-market-reports-major-centres

CMHC (2020b). Average Rents (Vacant and Occupied). Retrieved from: https://www.cmhc-schl.gc.ca/en/data-and- research/data-tables/average-apartment-rents-vacant-occupied

Lee, J., Fink, A., Park, H., Jiang, Z. (2018). Uncovering the Hidden Universe of Rental Units in Surrey. Retrieved from https://dsi.ubc.ca/data-science-social-good-2018

Statistics Canada (2017). Census of Population, 2016: Vancouver [Census Metropolitan Area]. Catalogue no. 98-316- X2016001. Retrieved from: https://www12.statcan.gc.ca/census-recensement/2016/dp- pd/prof/index.cfm?Lang=E

Appendix B: Additional Maps and Tables

Zone	Name	Mean PPB	Mean Price	1-bedroom Mean PPB	1-bedroom Mean Price	2-bedroom Mean PPB	2-bedroom Mean Price
T01	West End/Stanley Park	1,593	2,125	2,019	2,019	1,843	3,686
T02	English Bay	1,718	1,999	1,936	1,936	1,642	3,283
T03	Downtown	2,012	2,835	2,357	2,357	2,045	4,090
T04	South Granville/Oak	1,585	2,353	2,037	2,037	1,552	3,105
T05	Kitsilano/Point Grey	1,462	2,424	1,920	1,920	1,513	3,025
T06	Westside/Kerrisdale	1,199	2,369	1,795	1,795	1,342	2,685
T07	Marpole	1,075	1,822	1,516	1,516	1,163	2,326
T08	Mount Pleasant/Renfrew Heights	1,154	1,565	1,840	1,840	1,115	2,229
T09	East Hastings	1,294	1,736	2,030	2,030	1,213	2,426
T10	Southeast Vancouver	921	1,389	1,520	1,520	977	1,954
T11	University Endowment Lands	1,462	2,826	1,666	1,666	1,472	2,944
T12	Central Park/Metrotown	1,132	1,621	1,582	1,582	1,131	2,262
T13	Southeast Burnaby	967	1,491	1,340	1,340	965	1,929
T14	North Burnaby	1,026	1,578	1,507	1,507	1,070	2,140
T15	New Westminster	1,095	1,474	1,412	1,412	1,022	2,043
T16	North Vancouver CY	1,397	2,240	2,007	2,007	1,301	2,601
T17	North Vancouver DM	1,265	2,383	1,810	1,810	1,196	2,393
T18	West Vancouver	1,464	4,299	1,846	1,846	1,387	2,773
T19	Richmond	1,043	1,797	1,479	1,479	1,052	2,104
T20	Delta	832	1,558	1,176	1,176	917	1,834
T21	Surrey	812	1,362	1,145	1,145	779	1,557
T22	White Rock	1,021	2,068	1,312	1,312	943	1,887
T23	Langley City and Langley DM	906	1,491	1,259	1,259	831	1,662
T24	Tri-Cities	971	1,737	1,419	1,419	954	1,909
T25	Maple Ridge/Pitt Meadows	801	1,682	1,086	1,086	778	1,557

Zone (Cont.)	Name	Median PPB	Median Price	1-bedroom Median PPB	1-bedroom Median Price	2-bedroom Median PPB	2-bedroom Median Price
T01	West End/Stanley Park	1650	1,863	1,990	1,990	1,600	3,200
T02	English Bay	1750	1,850	1,850	1,850	2,925	1,463
T03	Downtown	1995	2,400	2,300	2,300	1,750	3,500
T04	South Granville/Oak	1500	2,175	2,000	2,000	1,475	2,950
T05	Kitsilano/Point Grey	1300	1,950	1,800	1,800	1,300	2,600
T06	Westside/Kerrisdale	1000	1,660	1,690	1,690	1,240	2,480
T07	Marpole	975	1,378	1,458	1,458	1,050	2,100
T08	Mount Pleasant/Renfrew Heights	950	1,313	1,700	1,700	1,000	2,000
T09	East Hastings	1100	1,500	1,800	1,800	1,125	2,250
T10	Southeast Vancouver	800	1,100	1,500	1,500	900	1,800
T11	University Endowment Lands	1400	2,550	1,680	1,680	1,413	2,825
T12	Central Park/Metrotown	1000	1,575	1,645	1,645	1,183	2,365
T13	Southeast Burnaby	850	1,400	1,350	1,350	975	1,950
T14	North Burnaby	900	1,500	1,500	1,500	1,090	2,180
T15	New Westminster	1050	1,450	1,450	1,450	1,025	2,050
T16	North Vancouver CY	1300	2,200	1,993	1,993	1,300	2,600
T17	North Vancouver DM	1125	1,950	1,750	1,750	1,148	2,295
T18	West Vancouver	1300	3,300	1,788	1,788	1,275	2,550
T19	Richmond	950	1,750	1,550	1,550	1,050	2,100
T20	Delta	750	1,400	988	988	775	1,550
T21	Surrey	750	1,300	1,100	1,100	700	1,400
T22	White Rock	975	1,700	1,250	1,250	853	1,705
T23	Langley City and Langley DM	850	1,435	1,295	1,295	825	1,650
T24	Tri-Cities	867	1,600	1,400	1,400	950	1,900
T25	Maple Ridge/Pitt Meadows	750	1,500	1,050	1,050	750	1,500

Zone (Cont.)	Name	CMHC Mean Price (Vacant 1-bedroom)	CMHC Mean Price (Occupied 1-bedroom)	CMHC Mean Price (Vacant 2-bedroom)	CMHC Mean Price (Occupied 2-bedroom)	Percentage Difference of sample VS. CMHC Vacant Mean Price (1-bedroom)	Percentage Difference of sample VS. CMHC Vacant Mean Price (2-bedroom)
T01	West End/Stanley Park	1,814	1,562	NA	2,286	11%	NA
T02	English Bay	1,838	1,665	2,571	2,337	5%	28%
T03	Downtown	1,816	1,678	2,762	2,542	30%	48%
T04	South Granville/Oak	1,708	1,491	2,172	2,015	19%	43%
T05	Kitsilano/Point Grey	1,847	1,574	2,666	2,135	4%	13%
T06	Westside/Kerrisdale	1,600	1,484	3,134	2,297	12%	-14%
T07	Marpole	1,198	1,155	2,125	1,529	27%	9%
T08	Mount Pleasant/Renfrew Heights	1,684	1,294	2,216	1,777	9%	1%
T09	East Hastings	1,590	1,242	NA	1,656	28%	NA
T10	Southeast Vancouver	1,711	1,302	2,444	1,845	-11%	-20%
T11	University Endowment Lands	NA	1,828	NA	2,350	NA	NA
T12	Central Park/Metrotown	1,603	1,244	1,991	1,584	-1%	14%
T13	Southeast Burnaby	1,251	1,076	1,774	1,351	7%	9%
T14	North Burnaby	1,534	1,247	1,920	1,603	-2%	11%
T15	New Westminster	1,340	1,193	2,083	1,584	5%	-2%
T16	North Vancouver CY	1,614	1,381	2,443	1,693	24%	6%
T17	North Vancouver DM	NA	1,553	2,195	1,970	NA	9%
T18	West Vancouver	1,901	1,751	2,946	2,578	-3%	-6%
T19	Richmond	1,384	1,249	1,629	1,507	7%	29%
T20	Delta	1,108	960	1,415	1,254	6%	30%
T21	Surrey	1,295	1,018	NA	1,213	-12%	NA
T22	White Rock	1,141	1,079	NA	1,391	15%	NA
T23	Langley City and Langley DM	1,403	1,207	1,857	1,550	-10%	-10%
T24	Tri-Cities	1,140	1,145	2,176	1,499	24%	-12%
T25	Maple Ridge/Pitt Meadows	NA	957	NA	0.2	NA	20%

CSD	Median PPB	Bachelor/ studio Median Price	1-bedroom Median Price	2-bedroom Median Price	3+ bedroom Median Price	1-bedroom Median Price (Feb.)	2-bedroom Median Price (Feb.)
City of New Westminster	1,050	1,315	1,450	2,075	2,900	1,450	2,120
City of North Vancouver	1,300	1,331	1,993	2,600	3,525	2,050	2,900
City of Pitt Meadows	800	NA	1,250	1,855	2,400	NA	1,655
City of Port Coquitlam	875	1,000	1,373	1,795	2,498	1,175	1,550
City of Port Moody	900	760	1,300	2,075	2,998	1,350	1,825
City of Richmond	950	1,300	1,550	2,100	2,600	1,480	2,150
City of Surrey	750	1,150	1,100	1,400	2,395	1,100	1,450
City of Vancouver	1,300	1,678	2,050	2,750	3,700	2,100	2,995
City of White Rock	975	985	1,250	1,700	3,150	1,600	1,800
District of North Vancouver	1,121	1,510	1,750	2,288	3,950	1,800	2,400
District of West Vancouver	1,300	1,500	1,795	2,550	5,500	1,800	2,450
Electoral Area A	1,367	NA	1,630	2,850	5,100	1,893	2,850
Township of Langley	800	NA	1,295	1,550	2,400	1,200	1,550
Tsawwassen First Nation	913	NA	1,550	1,925	2,500	NA	1,925
Village of Anmore	888	NA	1,059	1,645	5,150	1,125	1,450
Village of Belcarra	1,300	NA	1,900	NA	6,500	NA	NA
Village of Lions Bay	1,250	NA	NA	2,500	NA	NA	2,500
Bowen Island Municipality	850	NA	1,100	NA	NA	NA	NA
City of Burnaby	900	1,350	1,500	2,100	2,600	1,550	2,100
City of Coquitlam	850	1,150	1,500	1,900	2,700	1,630	1,900
City of Delta	750	1,360	1,000	1,525	2,400	1,100	1,688
City of Langley	875	925	1,300	1,750	2,275	1,305	1,690
City of Maple Ridge	750	1,025	1,050	1,488	2,400	1,100	1,350

FSA	Median Price	Median PPB	Bachelor Median Price	Bachelor Median PPB	1-bedroom Median Price	1-bedroom Median PPB	2-bedroom Median Price	2-bedroom Median PPB	3+bedroom Median Price	3+bedroom Median PPB
V0N	1,100	1,100	NA	NA	1,100	1,100	2,500	1,250	NA	NA
V1M	1,250	800	NA	NA	1,300	1,300	1,525	763	3,300	933
V2W	1,820	713	NA	NA	1,100	1,100	1,300	650	2,700	699
V2X	1,425	750	1,050	1,050	1,050	1,050	1,500	750	2,200	657
V2Y	1,400	850	NA	NA	1,300	1,300	1,600	800	2,300	742
V2Z	1,613	775	NA	NA	1,200	1,200	1,625	813	2,625	734
V3A	1,475	875	925	925	1,300	1,300	1,713	857	2,275	687
V3B	1,750	967	1,000	1,000	1,438	1,438	2,100	1,050	2,500	750
V3C	1,560	837	NA	NA	1,350	1,350	1,600	800	2,550	680
V3E	1,700	750	NA	NA	1,200	1,200	1,600	800	2,900	725
V3H	1,650	900	760	760	1,275	1,275	1,990	995	3,275	942
V3J	1,650	949	1,350	1,350	1,650	1,650	2,000	1,000	2,600	794
V3K	1,400	767	1,100	1,100	1,375	1,375	1,700	850	2,300	667
V3L	1,550	1,050	1,295	1,295	1,515	1,515	2,050	1,025	3,095	950
V3M	1,400	1,000	1,330	1,330	1,400	1,400	2,075	2,075	2,765	769
V3N	1,600	985	1,350	1,350	1,475	1,475	2,000	1,000	2,600	833
V3R	1,400	750	850	850	1,000	1,000	1,550	775	2,250	667
V3S	1,200	700	1,000	1,000	1,025	1,025	1,300	650	2,300	650
V3T	1,350	875	1,200	1,200	1,363	1,363	1,750	875	2,000	667
V3V	989	667	765	765	989	989	1,400	700	2,300	640
V3W	900	650	898	898	900	900	1,400	700	2,300	650
V3X	1,250	700	NA	NA	904	904	1,350	675	2,700	733
V3Y	1,850	800	NA	NA	1,250	1,250	1,850	925	2,400	800
V3Z	1,300	750	1,188	1,188	1,000	1,000	1,400	700	2,400	733
V4A	2,100	850	NA	NA	1,100	1,100	1,925	963	2,750	800
V4B	1,700	975	985	985	1,250	1,250	1,705	853	3,150	925
V4C	1,300	700	NA	NA	1,000	1,000	1,500	750	2,200	629
V4E	975	700	NA	NA	963	963	1,350	675	2,700	688
V4K	1,800	733	NA	NA	890	890	1,600	800	NA	NA
V4L	1,265	1,133	NA	NA	1,265	1,265	NA	NA	2,350	684
V4M	1,800	900	1,360	1,360	900	900	2,250	1,125	2,000	667
V4N	1,200	749	838	838	1,100	1,100	1,350	675	2,700	800
V4P	1,550	825	NA	NA	1,250	1,250	1,600	800	2,500	707
V4R	1,400	767	1,000	1,000	1,050	1,050	1,438	719	2,800	825
V4W	1,200	750	NA	NA	950	950	1,350	675	2,450	763

FSA	Median Price	Median PPB	Bachelor Median Price	Bachelor Median PPB	1-bedroom Median Price	1-bedroom Median PPB	2-bedroom Median Price	2-bedroom Median PPB	3+bedroom Median Price	3+bedroom Median PPB
V5A	1,500	850	1,290	1,290	1,600	1,600	2,100	1,050	2,175	619
V5B	990	800	1,225	1,225	1,425	1,425	1,725	863	2,500	733
V5C	1,800	1,143	1,350	1,350	1,500	1,500	2,400	1,200	3,550	892
V5E	1,200	825	1,475	1,475	1,250	1,250	1,925	963	2,800	767
V5G	1,225	793	NA	NA	1,300	1,300	1,600	800	3,300	825
V5H	1,625	1,125	1,550	1,550	1,645	1,645	2,400	1,200	2,600	700
V5J	1,300	800	1,150	1,150	1,350	1,350	2,150	1,075	2,800	933
V5K	1,250	825	1,400	1,400	1,400	1,400	1,800	900	2,600	742
V5L	1,463	1,014	1,250	1,250	1,700	1,700	2,400	1,200	2,795	867
V5M	1,000	800	2,295	2,295	1,395	1,395	1,795	898	3,400	1,133
V5N	1,065	900	1,600	1,600	1,500	1,500	2,000	1,000	2,100	659
V5P	1,000	750	950	950	1,200	1,200	1,680	840	2,943	918
V5R	900	775	1,325	1,325	1,615	1,615	1,800	900	2,400	733
V5S	1,650	878	NA	NA	1,663	1,663	2,000	1,000	2,500	767
V5T	1,800	1,395	1,324	1,324	1,950	1,950	2,425	1,213	2,840	800
V5V	1,200	900	1,825	1,825	1,350	1,350	1,975	988	3,900	1,267
V5W	1,100	845	975	975	1,475	1,475	1,900	950	2,720	886
V5X	1,300	848	1,300	1,300	1,680	1,680	1,800	900	2,675	850
V5Y	2,180	1,525	1,650	1,650	2,200	2,200	2,950	1,475	2,575	842
V5Z	1,900	1,250	1,495	1,495	1,900	1,900	2,725	1,363	4,500	1,354
V6A	1,750	1,480	1,625	1,625	1,798	1,798	2,955	1,478	3,500	1,167
V6B	2,400	2,000	1,950	1,950	2,333	2,333	3,495	1,748	3,795	1,092
V6C	7,500	3,750	NA	NA	2,300	2,300	7,995	3,998	5,800	1,933
V6E	2,175	1,695	1,700	1,700	2,100	2,100	3,480	1,740	18,500	5,417
V6G	2,016	1,695	1,680	1,680	1,975	1,975	3,495	1,748	4,500	1,500
V6H	1,775	1,400	1,345	1,345	1,750	1,750	2,700	1,350	4,995	1,625
V6J	2,080	1,498	1,473	1,473	1,825	1,825	2,800	1,400	4,800	1,286
V6K	1,950	1,374	1,500	1,500	1,850	1,850	2,600	1,300	5,650	1,434
V6L	1,965	1,000	NA	NA	1,950	1,950	2,175	1,088	4,500	1,498
V6M	1,500	1,000	1,350	1,350	1,800	1,800	2,500	1,250	5,650	1,090
V6N	1,450	950	NA	NA	1,650	1,650	2,100	1,050	3,648	1,017
V6P	1,450	1,000	1,425	1,425	1,500	1,500	2,450	1,225	3,600	1,100
V6R	1,998	1,150	1,320	1,320	1,825	1,825	2,450	1,225	3,850	979

FSA	Median Price	Median PPB	Bachelor Median Price	Bachelor Median PPB	1-bedroom Median Price	1-bedroom Median PPB	2-bedroom Median Price	2-bedroom Median PPB	3+bedroom Median Price	3+bedroom Median PPB
V6S	2,000	1,275	NA	NA	1,500	1,500	2,825	1,413	4,300	1,163
V6T	2,300	1,333	NA	NA	1,680	1,680	2,650	1,325	3,700	1,103
V6V	850	733	NA	NA	1,200	1,200	2,095	1,048	7,750	1,817
V6W	1,750	1,148	1,280	1,280	1,698	1,698	2,000	1,000	2,400	767
V6X	1,900	1,100	1,500	1,500	1,750	1,750	2,250	1,125	2,400	2,400
V6Y	1,750	975	1,300	1,300	1,575	1,575	2,100	1,050	2,650	850
V6Z	2,600	2,166	1,995	1,995	2,400	2,400	3,550	1,775	2,550	800
V7A	1,490	750	875	875	1,400	1,400	1,600	800	7,200	2,390
V7B	1,368	922	NA	NA	NA	NA	1,785	893	2,650	675
V7C	1,600	960	1,250	1,250	1,325	1,325	1,925	963	NA	NA
V7E	1,600	800	NA	NA	1,425	1,425	1,725	863	3,150	884
V7G	2,925	1,075	NA	NA	2,500	2,500	2,295	1,148	2,650	753
V7H	2,213	1,075	NA	NA	2,075	2,075	2,550	1,275	4,200	1,050
V7J	1,800	1,100	1,950	1,950	1,800	1,800	2,360	1,180	3,525	1,050
V7K	1,795	1,000	1,500	1,500	1,600	1,600	1,875	938	3,300	1,067
V7L	2,250	1,258	1,250	1,250	1,990	1,990	2,550	1,275	3,840	988
V7M	2,150	1,350	1,713	1,713	1,990	1,990	2,575	1,288	3,525	1,117
V7N	1,725	983	1,200	1,200	1,650	1,650	2,100	1,050	3,400	1,100
V7P	1,850	1,400	1,553	1,553	1,750	1,750	2,400	1,200	3,550	987
V7R	2,545	1,100	1,000	1,000	1,638	1,638	2,150	1,075	4,400	1,447
V7S	4,098	1,340	NA	NA	2,375	2,375	2,600	1,300	4,300	1,125
V7T	2,525	1,216	1,920	1,920	1,800	1,800	2,400	1,200	6,600	1,450
V7V	3,250	1,375	1,475	1,475	1,800	1,800	2,775	1,388	4,350	1,216
V7W	3,500	1,329	1,600	1,600	1,450	1,450	3,000	1,500	5,450	1,299

FSA (Cont.)	3-bedroom Median Price	3-bedroom Median PPB	4+bedroom Median Price	4+bedroom Median PPB	Difference of Median Price between CMA and FSA (Bachelor)	Difference of Median Price between CMA and FSA (1-bedroom)	Difference of Median Price between CMA and FSA (2-bedroom)
V0N	NA	NA	NA	NA	NA	-35%	19%
V1M	3,050	1,017	3,500	875	NA	-24%	-27%
V2W	2,225	742	2,850	680	NA	-35%	-38%
V2X	2,000	667	2,650	625	-30%	-38%	-29%
V2Y	2,300	767	2,650	663	NA	-24%	-24%
V2Z	2,400	800	3,000	625	NA	-29%	-23%
V3A	2,150	717	2,800	588	-38%	-24%	-18%
V3B	2,498	833	2,970	633	-33%	-15%	0%
V3C	2,200	733	3,000	633	NA	-21%	-24%
V3E	2,443	815	3,320	725	NA	-29%	-24%
V3H	3,500	825	4,200	943	-49%	-25%	-5%
V3J	2,350	784	3,550	794	-10%	-3%	-5%
V3K	2,100	700	2,650	632	-27%	-19%	-19%
V3L	2,850	950	3,498	875	-14%	-11%	-2%
V3M	2,750	917	2,950	600	-11%	-18%	-1%
V3N	2,550	850	3,375	678	-10%	-13%	-5%
V3R	2,000	667	2,600	592	-43%	-41%	-26%
V3S	2,200	733	2,775	619	-33%	-40%	-38%
V3T	2,000	667	2,600	638	-20%	-20%	-17%
V3V	2,000	667	2,630	581	-49%	-42%	-33%
V3W	2,100	700	2,500	625	-40%	-47%	-33%
V3X	2,250	750	2,825	707	NA	-47%	-36%
V3Y	2,400	800	2,588	647	NA	-26%	-12%
V3Z	2,250	750	2,900	700	-21%	-41%	-33%
V4A	2,400	800	3,250	750	NA	-35%	-8%
V4B	2,900	967	3,700	790	-34%	-26%	-19%
V4C	2,000	667	2,600	598	NA	-41%	-29%
V4E	2,700	900	2,600	650	NA	-43%	-36%
V4K	2,150	717	2,600	650	NA	-48%	-24%
V4L	2,000	667	NA	NA	NA	-26%	NA
V4M	2,500	833	3,700	753	-9%	-47%	7%
V4N	2,250	750	2,800	660	-44%	-35%	-36%
V4P	2,600	867	3,200	800	NA	-26%	-24%
V4R	2,300	767	2,700	625	-33%	-38%	-32%
V4W	1,700	567	5,000	917	NA	-44%	-36%
V5A	2,200	733	3,300	750	-14%	-6%	0%
V5B	2,475	825	4,350	984	-18%	-16%	-18%
V5C	2,500	833	2,850	713	-10%	-12%	14%
V5E	2,398	799	3,500	825	-2%	-26%	-8%
V5G	2,100	700	2,900	688	NA	-24%	-24%
V5H	2,800	933	4,000	1,000	3%	-3%	14%
V5J	2,375	792	3,075	650	-23%	-21%	2%
V5K	2,630	877	3,000	750	-7%	-18%	-14%
V5L	3,400	1,133	NA	NA	-17%	0%	14%
V5M	2,000	667	2,500	625	53%	-18%	-15%
V5N	2,943	981	3,145	787	7%	-12%	-5%

FSA (Cont.)	3-bedroom Median Price	3-bedroom Median PPB	4+bedroom Median Price	4+bedroom Median PPB	Difference of Median Price between CMA and FSA (Bachelor)	Difference of Median Price between CMA and FSA (1-bedroom)	Difference of Median Price between CMA and FSA (2-bedroom)
V5P	2,200	733	3,800	800	-37%	-29%	-20%
V5R	2,400	800	3,598	0	-12%	-5%	-14%
V5S	2,400	800	3,650	730	NA	-2%	-5%
V5T	3,800	1,267	4,050	908	-12%	15%	15%
V5V	2,690	897	3,900	875	22%	-21%	-6%
V5W	2,550	850	4,250	963	-35%	-13%	-10%
V5X	2,448	816	3,950	963	-13%	-1%	-14%
V5Y	4,500	1,500	4,250	846	10%	29%	40%
V5Z	3,500	1,167	3,825	884	0%	12%	30%
V6A	3,825	1,275	3,795	949	8%	6%	41%
V6B	5,800	1,933	NA	NA	30%	37%	66%
V6C	19,000	6,333	18,000	4,500	NA	35%	281%
V6E	4,500	1,500	NA	NA	13%	24%	66%
V6G	4,750	1,583	7,000	1,750	12%	16%	66%
V6H	4,500	1,500	6,150	1,193	-10%	3%	29%
V6J	3,430	1,144	7,375	1,525	-2%	7%	33%
V6K	4,500	1,500	6,145	1,237	0%	9%	24%
V6L	3,000	1,000	5,800	1,100	NA	15%	4%
V6M	3,240	1,080	4,975	935	-10%	6%	19%
V6N	3,300	1,100	5,190	1,080	NA	-3%	0%
V6P	2,800	933	5,550	1,100	-5%	-12%	17%
V6R	3,550	1,184	4,650	1,125	-12%	7%	17%
V6S	3,500	1,167	4,100	1,025	NA	-12%	35%
V6T	4,000	1,333	11,500	2,300	NA	-1%	26%
V6V	2,350	784	3,880	700	NA	-29%	0%
V6W	2,400	800	NA	NA	-15%	0%	-5%
V6X	2,600	867	3,500	833	0%	3%	7%
V6Y	2,440	814	3,000	740	-13%	-7%	0%
V6Z	6,900	2,300	12,975	2,945	33%	41%	69%
V7A	2,150	717	2,990	650	-42%	-18%	-24%
V7B	NA	NA	NA	NA	NA	NA	-15%
V7C	2,590	863	4,350	938	-17%	-22%	-8%
V7E	2,500	833	2,950	700	NA	-16%	-18%
V7G	3,048	1,016	4,395	1,050	NA	47%	9%
V7H	2,900	967	4,200	1,050	NA	22%	21%
V7J	3,200	1,067	3,950	988	30%	6%	12%
V7K	3,750	1,250	4,000	967	0%	-6%	-11%
V7L	3,400	1,133	3,800	850	-17%	17%	21%
V7M	3,300	1,100	5,375	1,344	14%	17%	23%
V7N	3,100	1,033	4,500	900	-20%	-3%	0%
V7P	4,400	1,467	5,200	1,040	4%	3%	14%
V7R	3,595	1,199	4,550	1,100	-33%	-4%	2%
V7S	2,850	950	7,525	1,459	NA	40%	24%
V7T	3,695	1,232	5,000	1,200	28%	6%	14%
V7V	4,000	1,333	5,650	1,180	-2%	6%	32%
V7W	3,850	1,283	5,850	1,313	7%	-15%	43%

FSA (Cont.)	# of Units (Total)	# of Bachelor Units	# of 1-bedroom Units	# of 2-bedroom Units	# of 3+bedroom Units	2-bedroom Median Price / Median CMA Household Income
V0N	3	0	1	1	0	49%
V1M	49	0	11	12	7	30%
V2W	61	0	10	8	35	26%
V2X	120	1	23	41	34	30%
V2Y	150	0	36	47	28	32%
V2Z	26	0	1	9	8	32%
V3A	168	1	52	62	24	34%
V3B	263	9	44	112	43	41%
V3C	146	0	42	40	36	32%
V3E	170	0	16	48	74	32%
V3H	112	1	18	31	30	39%
V3J	306	4	94	81	42	39%
V3K	179	5	32	51	35	34%
V3L	161	8	55	42	13	40%
V3M	308	13	119	76	16	41%
V3N	235	4	71	81	24	39%
V3R	267	5	57	88	52	31%
V3S	270	3	60	106	41	26%
V3T	537	20	178	147	48	35%
V3V	248	2	38	64	35	28%
V3W	321	2	43	75	50	28%
V3X	244	0	51	97	35	27%
V3Y	29	0	3	7	11	36%
V3Z	423	2	94	152	84	28%
V4A	127	0	20	38	56	38%
V4B	142	2	37	52	38	34%
V4C	191	0	23	52	46	30%
V4E	25	0	6	2	8	27%
V4K	64	0	16	9	30	32%
V4L	6	0	3	0	1	NA
V4M	114	5	30	34	27	44%
V4N	317	2	84	106	61	27%
V4P	43	0	11	13	13	32%
V4R	37	1	7	8	14	28%
V4W	24	0	5	7	6	27%
V5A	175	7	24	39	35	41%
V5B	166	2	26	38	14	34%
V5C	260	2	79	98	22	47%
V5E	126	3	23	28	19	38%
V5G	152	0	25	43	17	32%
V5H	334	9	111	106	16	47%
V5J	156	2	30	35	28	42%
V5K	185	8	25	47	23	36%
V5L	140	7	33	30	13	47%
V5M	151	1	22	35	8	35%
V5N	278	3	59	55	20	39%
V5P	241	1	25	67	31	33%
V5R	397	5	62	66	31	36%

FSA (Cont.)	# of Units (Total)	# of Bachelor Units	# of 1-bedroom Units	# of 2-bedroom Units	# of 3+bedroom Units	2-bedroom Median Price / Median CMA Household Income
V5S	164	0	24	62	25	39%
V5T	253	22	100	48	17	48%
V5V	151	2	27	36	14	39%
V5W	222	6	28	53	32	37%
V5X	321	5	51	86	34	36%
V5Y	323	5	126	95	18	58%
V5Z	207	9	42	66	20	54%
V6A	147	19	46	42	5	58%
V6B	998	71	542	326	31	69%
V6C	25	0	1	17	2	158%
V6E	657	37	288	193	15	69%
V6G	448	23	161	118	31	69%
V6H	159	11	53	43	15	53%
V6J	203	10	70	61	24	55%
V6K	306	10	91	84	37	51%
V6L	88	0	10	22	22	43%
V6M	143	8	27	25	28	49%
V6N	105	0	16	18	21	41%
V6P	346	8	66	67	64	48%
V6R	204	4	34	48	53	48%
V6S	127	0	26	44	22	56%
V6T	53	0	13	23	4	52%
V6V	37	0	3	1	13	41%
V6W	31	2	14	8	3	39%
V6X	317	3	93	126	30	44%
V6Y	231	5	61	67	47	41%
V6Z	490	25	240	149	25	70%
V7A	80	2	17	11	31	32%
V7B	2	0	0	1	0	35%
V7C	115	2	28	23	36	38%
V7E	121	0	20	19	42	34%
V7G	38	0	5	8	21	45%
V7H	56	0	12	15	14	50%
V7J	114	2	25	37	13	47%
V7K	41	3	9	18	6	37%
V7L	238	7	61	97	38	50%
V7M	166	4	50	66	15	51%
V7N	109	4	19	17	24	41%
V7P	108	14	29	27	15	47%
V7R	72	1	8	18	30	42%
V7S	74	0	4	11	42	51%
V7T	94	3	15	24	30	47%
V7V	98	3	17	24	46	55%
V7W	78	5	11	9	44	59%

Contents

1 Introduction

The goal of the data analyst of HRC's Craiglist project is to enable real time collection and storage of Vancouver's housing rental postings in order to study the housing market over time. The main source of data for this project are the rental advertisements on https://vancouver.craigslist.org. To capture as many postings as possible, a scraper has been deployed and is scheduled to crawl rental postings from the Craigslist website on a daily basis. The scraper is deployed on a virtual machine (VM) on UBC's HRC server. It is imple- mented mainly based on Scrapy framework, which is specifically designed for web crawling and data extraction. In order to implement Scrapy, the rest of the code package is written in Python.

2 Functionality of the Scraper

Within the Scrapy framework, spiders are the building blocks of the crawling process. Each Scrapy spider has a customized pipeline of what to crawl and how to parse the crawled items. This section explains the scraper's important features and how the Scrapy spiders divide the scraping tasks.

2.1 Web mode and Archive mode

The scraper has two modes: web mode and archive mode, each executed by multiple dedicated spiders (described in section 2.4). In web mode, the scraper crawls Craiglists' apartment rental site (https://vancouver.craigslist.org/ search/apa)and single room rental site (https://vancouver.craigslist.org/ search/roo), then stores the HTML file of each advertisement locally. The product is a collection of all the scraped advertisements in HTML format. In archive mode, the scraper crawls the HTML files mentioned above into a CSV file. This process is completely offline since it only involves scraping local files.

The purpose of separating direct crawling into two modes is to archive as much data as possible from the raw HTML. Since spiders can only crawl designated fields and the rest will be lost, archiving the original HTML content enables researchers to extract unused information from old data. Furthermore, this structure enables the scraper to be resilient to changes of HTML structure of the Craiglist website, which potentially causes contamination in the data.

2.2 Daily Scraping Limit

The scraper's web scraping volume is restricted to 1000 pages per day in order to moderate the effect on Craigslist's infrastructure and lower the risk the research is exposed to. The particular 1000 pages limit is set in response to the following clause from Craigslist's Term of Use:

> Requesting, viewing, or accessing more than 1,000 pages of CL in any 24-hour period - $0.25 per page during the 24 hour period after the first 1,000 pages

In order to minimize the load on Craigslist's server and thus the risk to re- search, each web mode spider has been configured to close after the crawling a certain number of pages, which is defined by the parameter CLOSESPI- DER ITEMCOUNT in a spider's custom settings. The daily limit does not apply to archive mode spiders since they crawl locally and do not involve ac- cessing Craigslist's server.

2.3 Scrapy DeltaFetch

Deduplication is an essential part of the collection process. Craiglist advertise- ments can be duplicated in two ways such as repeat posting, where the author posts the same advertisement multiple times after tweaking some minor details; and repeat scraping, where the scraper crawls the same posting URL more than once. A solution to minimize repeat scraping is DeltaFetch, a Scrapy middle- ware that can filter out items (in this case url's) seen on previous crawls, so that the scraper ignores those duplicates. More importantly, by avoiding re- peat scraping, DeltaFetch reduces the scraping volume by a significant amount, which conserves storage and the daily scraping limit.

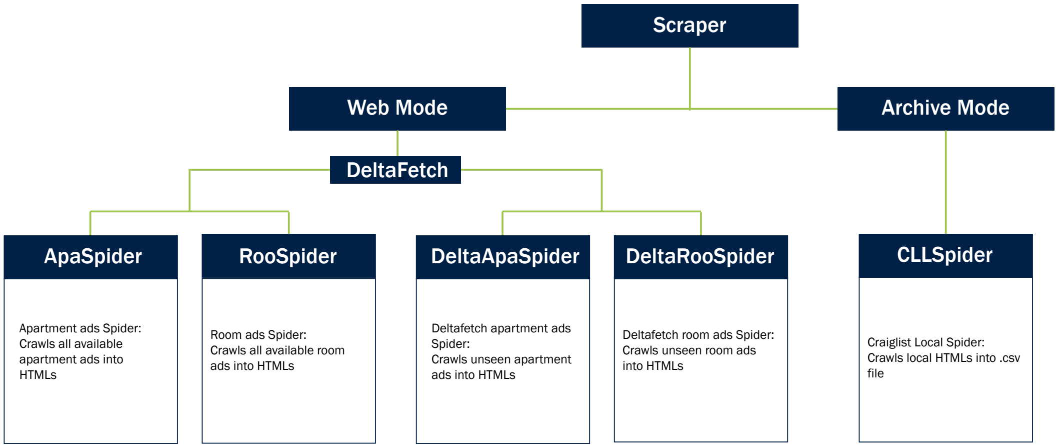

Scraper
Web Mode
Archive Mode
DeltaFetch
ApaSpider
Apartment ads Spider: Crawls all available apartment ads into HTMLs
RooSpider
Room ads Spider: Crawls all available room ads into HTMLs
DeltaApaSpider
Deltafetch apartment ads Spider: Crawls unseen apartment ads into HTMLs
DeltaRooSpider
Deltafetch room ads Spider: Crawls unseen room ads into HTMLs
CLLSpider
Craiglist Local Spider: Crawls local HTMLs into .csv file

Figure 1: Hierarchy of Spiders

The data is currently organized in a monthly order. As an example, January data in this project refers to all postings that are available from the apartment rental and room rental categories on Craiglist throughout January. In order to achieve deduplication in a monthly scope, DeltaFetch is disabled and reset on the first day of each month, so that all available postings posted from the previous month are captured, and is re-enabled throughout the month to avoid duplicated ads. Section 3.1 describes in details how this process is embedded into the spiders.

2.4 Scrapy Spiders

As mentioned in the beginning of section 2, each Scrapy spider is in charge of a specific crawling process. After initiating the scraping process, each spider executes a specific crawling task. Therefore, it is crucial to understand the hierarchy of the spiders as illustrated by Figure 1. Each web-mode spider is im- plemented to scrape either apartment rental ads or room rental ads (specified by starting URL), and is either DeltaFetch enabled or not.

The web mode spiders inherit the class of scrapy.spiders.Crawlspider, which initiates its crawl from the site specified by a parameter start urls, parse all the available HTML listings according to its pre-defined parsing functions and finds the next page of the website using rules specified by the parameter rules. The parsing function simply requires the spider to write out an HTML file locally, which is a verbatim copy of the original webpage that is left to be interpreted. The output of web mode spiders becomes the input of archive mode spiders.

An archive mode spider belongs to the scrapy.spiders.Spider class, which is a simpler spider than the Crawlspider mentioned above since it is not required to access the web. During archive mode, the scraper feeds into the archive spider a list of paths to the HTML files scraped by web mode spiders, as the parameter start urls, and the archive spider then parses the the designated data fields, such as price, address, and location, from the HTML files into a CSV file.

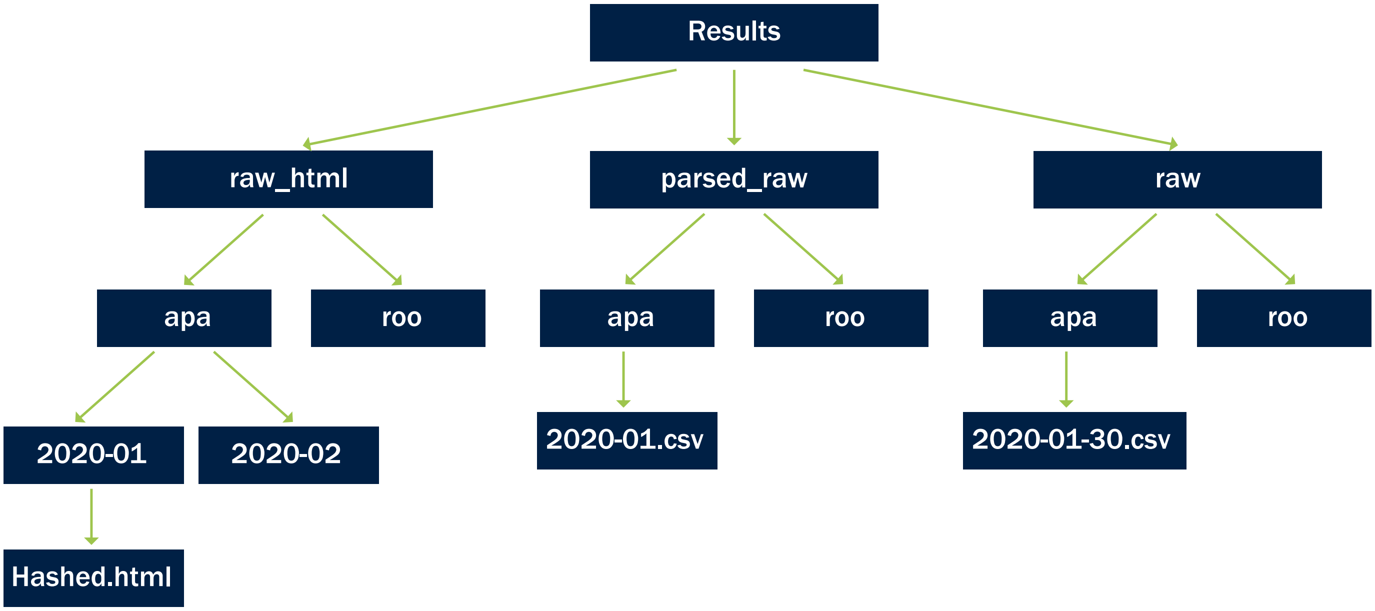


Figure 2: File Sotrage Structure

3 Operation of the Scraper

3.1 Schedule

The scraper has been deployed to the VM and is currently scheduled to crawl from Craiglist daily. On the first day of each month, the scraper activates the ApaSpider and the RooSpider (as described in Figure 1) to obtain all the available apartment rental postings and single room rental postings from Craiglist. Every day starting from the second day of each month, the scraper uses DeltaApaSpider and the DeltaRooSpider to capture apartment rentals and room rentals avoiding duplicates. The resulting monthly data contains all apartment/room postings that are available on Craiglist throughout the month. Archive mode is activated on the last day of each month, and processes the collection from the previous month into a interpretable CSV format.

3.2 Operating and Testing on the Server

The code package rental crawler is stored on the VM under: /home/sysad- min/pyenv/hrc2019. The user can start a remote session as sysadmin on the VM through SSH or the eduCloud web portal. The script that starts the scraper is scheduler.py. To run the scraper, open Terminal and type in the following command:

```
cd pyenv/hrc2019/rental_crawlerpython3-m
```

```
scheduler
```

In order for the scraper to continuously operate and produce information logs, the terminal window must be left open and unchanged, which is not interrupted by users' starting or ending a remote session, but requires the sysadmin user to stay logged in. To terminate the scraper, type in Ctrl + C under the same terminal window and scraping will be wrapped up properly.

In order to test the individual spiders without changing the scheduler, there is a script run spiders.py dedicated to test the spiders once they have been modified. The script can run a specific spider with two parameters: mode and dir/directory. For example, to test the functionality of the CLLSpider which is the archive spider, type the following command in a Terminal window:

```
cd pyenv/hrc2019/rental_crawler

python3 -m run_spiders --mode=archive --dir
=2020 -02
```

The command above instructs the CLLSpider to scrape from the folder:

/results/raw html/2020-02. For details about running the test script, please refer to the documentation within the script: /rental crawler/run spiders.py

4 File Structure

All data is stored on the VM in the results folder: /home/sysadmin/pyen- v/hrc2019/ results. The output of web-mode spiders is stored in the raw html folder, while the archive spider scrapes data in raw html and stores the output in parsed raw. Figure 2 illustrates the hierarchy structure of the files.

The data files are first separated into HTML files(in raw html) and CSV files (in parsed raw). Within each type, they are then separated into apa (i.e. apart- ment rental postings) and roo (i.e. room rental postings) files. Moreover, within each ad type, data is organized into a monthly manner.

The raw folder stores the results of a regular spider CLSpider which is used for confirming the results from web mode and archive mode. The spider can be tested using run spiders.py mentioned in section 3.2.

5 Code Documentation

For the documentation of the code package rental crawler, please refer to the README.md in the *rental crawler* folder

Appendix D:Extended Discussion on Data Veracity and Manual Processing Procedures

Outline

This appendix serves as an extended methodological discussion for the benefit of others seeking to reproduce or continue this work. While Appendix C outlines operation of the automated scraping tools for data collection, the manual processing of the collected data is addressed herein. HRC's observations regarding the disorganized nature of this kind of data are offered alongside how they were addressed.

First are some targeted discussions about the various challenges in classifying the unbridled data that comes from the online market. Second is a high - level discussion of the three stages to improving the veracity of the data: Deduplication, cleaning, and empty attribute population. Examples of problematic ads are provided in brief discussion of the kinds of issues that can be encountered in VGI. Third is a description of software - specific techniques that were used to conduct the three aforementioned stages efficiently.

1 – Classification Challenges

1.1 Challenges in Defining Type: Unit vs. Structure

One of the more immediate complications of big data in unit type classification. Unlike formal surveys, including the CMHC's RMS and the Canadian Census, the classified ad market does not conform to firm typologies. The concept of type can be broken into two qualities for each given dwelling: the first being the format of the unit itself in terms of layout and bedroom count, the second being the type of structure the unit exists in, such as a detached house, duplex, condominium, etc. This is not necessarily adhered to in market activity and there is inconsistency in what one user may report as type from the next.

Craigslist does provide a selection of structural types: house, apartment, condo, flat, cottage/cabin, in - law, loft, townhouse, manufactured, assisted living, and land. These do not, however, align with the typologies used by the Canadian Census or CMHC and are subject to interpretations of terms such as "flat" and "apartment" that may differ by geography or individual. Perhaps because the "apartment" selection is the default value in a classified category that namesakes it, due to user indiscretion or apathy, that type is found on many ads which would not be considered a purpose - built apartment in the CMHC's primary market.

Further complicating the matter are terms which describe the unit but neither satisfy bedroom count or structural definitions. Some users will use terms such as "basement", for example, which are important details for certain research questions but specify neither bedroom count nor anything about the structure the unit is within the basement level of. "Loft" is offered by Craigslist as a selectable type but may not be selected by the user who will opt to instead describe it in the body of the ad. Loft may also describe a room in a unit that is not a loft but instead, like a den, detailed as a storage space or an accessory room that may be used as an optional bedroom in excess of the specified number of bedrooms.

The result is a collection of terms that are not necessarily comprehensive, mutually exclusive, or mandatory for users to report as an analytically valid typology demands. This study's predecessor (Lee et al. 2018) reported difficulties in training a classification model for structural types. This study found that units aimed to house one person, or couple can take on materially different shapes, sizes, and prices. A significant amount of manual work and development in interpretation methods would need to be done to establish a typology that could align with, for example, the Canadian Census' structural types or estimate which units belong in the CMHC's primary vs. secondary rental market.

For the purposes of this study's cursory analysis, structural types were not classified from the raw data. The bedroom attribute, where not filled by the poster, was derived in - faith from the title or description. Most bedroom values are a numeric count of the bedrooms in a unit, excepting nominal sub - types in the one - bedroom class, discussed in following sections. Data points where no bedroom value could be determined were deleted from enumeration and analysis.

1.2 - One Bed Types: The Typological Vocabulary of Online Rental Housing Classified Market

While ads representing two or more bedrooms were simply be classified according to the number of bedrooms in the unit advertised, when dealing with units possessing fewer than two bedrooms there is a variety of terms and subtypes that exhibit a degree of semantic overlap on the market. These include bachelor, studio, loft, one - bedroom, private room ads. Because the latter is not always a complete unit, the former two are often described as having zero bedrooms, and a couple may share a bed and the quarters it furnishes, these ads might accurately be described as one bed opportunities. As such, they were treated as having one effective bedroom for analysis involving the price per bed figure. This section will discuss the complications of ads describing one bed units, give a brief profile of each subtype, and note the apparent overlaps and ambiguities even among these terms.

One Bedroom: A "One bedroom" unit refers to suites with a distinct bedroom separated by walls from other basic living spaces such as kitchen, living room, and washroom. It is one of the most common distinct types on Craigslist during the study period and found to generally be larger and priced higher than other one bed types, excepting lofts.

Bachelor and Studio Units: Some ads will either not cite a bedroom count or will declare zero bedrooms. Scrutiny of titles and descriptions often indicate that these represent studio or bachelor units where, presumably, the poster felt a zero - bedroom value would best reflect that the sleeping area is not architecturally distinct from the rest of the unit, instead in an open concept with other spaces like a kitchen. So - called "micro - suites" were included in this class. Some ads use either or both the terms "bachelor" and "studio" in their title and description. Notably, the term "studio" demonstrates different definitions throughout the dataset, including composite use with the term "loft" in some ads, and a Wikipedia article on the term declares geographic variability as well as a general ambiguity to what a studio can be interpreted as. When derived in-faith according to their descriptions and classified separately, there appeared to be a price distinction between studio and bachelor units although the actual volume of each was small relative to other classes. There appears to be an unspoken consensus that studios are generally larger and more amenity-complete while bachelor units are marketed to more budget-oriented tenants. In all cases, the rooms attribute field was populated as “bachelor” to collapse the two small classes and establish alignment with the CMHC use of the term.

Private Room: “Private room” refers to ads where a single bedroom is on offer and is typically posted in the “Rooms and Shares” section Craigslist. These units were classified as “room” in the rooms field. There are some nuances to this class that bear meaning for analysis.

Some private room ads, typically marketed to students, will use the term “bachelor” where the opportunity is a small yet independent unit within a principal structure, typically not much more than a small bedroom, washroom, exclusive entrance, and access to laundry. The term “studio”, notably, was not commonly found to describe these private room ads in the same way, despite the technical similarities with bachelor units. Complicating this use of terminology further is the apparent tendency for these private room units to have wall-separated bedrooms, making them more technically comparable to the one bedroom yet more fiscally comparable to bachelor units.

Other private room ads represent housing opportunities where landlords or existing tenants are looking for someone to fill a bedroom that is part of, and has access to, a larger two-or-more-bedroom unit. These include instances where the poster is looking to replace a departed roommate, or homestay opportunities for international students. In these cases, it may be improper to describe the opportunity as a unit and the term bachelor is seldom found in the ads. Some of the most affordable options include

While not perfectly comparable to other unit types, private rooms posted under the appropriate domain on CL were also gathered from January 2020 onward as they do represent genuine opportunities for housing for those seeking affordability and willing to share or minimize material amenity for it. These ads also represent a circulation of housing in larger units that do not involve new market leases. Prospective tenants, then, may be entering into shares of older and more affordable leases.

Lofts: The term “loft” generally indicates a space directly under the roof, often open to lower floors. Multiple uses of the term were found in the data. As a unit, a loft can be a converted industrial space with large floor area and high ceiling. Despite technically conforming to single bedroom or studio layouts, Craigslist offers “loft” as an option in their type field, parameterizing it as a description of a building rather than the unit itself. Very few two-or-more-bedroom units with high prices and apparent luxury profiles have been listed as lofts but were classified according to their bedroom count for this study. Some research questions may find it advisable to detach lofts as a distinct class due to their exceptional floorspace and prices. For this study, one bed lofts were included in the one-bedroom class. Any efforts to automate population of the rooms field as a loft must be careful, however. When a unit is listed as having a loft, on the other hand, it generally refers to a small space within a house or other unit that may have a sloped ceiling – the underside of the roof – which may serve as an office or guest room. Sometimes the ceiling is too low and the room windowless, thus only be suitable for storage.

Other terms found to describe one bed units

- Executive & Presidential: These terms are common in hotel and commercial space markets, yet some were used to describe luxury suites in the Craigslist data, often as premium studios. When not determined to be short - term rental opportunities, these were classified as bachelor.
- Junior & Micro: Very uncommon terms in the data. “Junior” appears to draw from the same budget - orientation vocabulary as “bachelor” and micro - suites tend to represent creatively minimal dwellings in trendy areas. Both terms were classified as bachelor type.

2 - Deduplication, Cleaning, and Attribute Population

Processing the raw data was broken into three procedural stages to improve its veracity: Deduplication, cleaning, attribute population. This section outlines the need, objectives, and policies of each stage to offer future research a guide based on the lessons learned about the nature of the data. Many include a rationale that outlines how the nature of the data informed the guideline and how exceptions to a policy may emerge. There is currently no apparent necessary order that these stages be performed, but generally deduplication can eliminate unneeded attribute population on ads that would turn out to be duplicates afterwards. It is advisable, in any case, that one establish an ordered system in order to address all issues. It is also advised that excess spaces, carriage returns, and erroneously parsed HTML characters such as slashes be removed for visual tidiness and compatibility with a GIS workflow before any extensive processing be done.

To support this discussion, the following attributes were collected from each ad:

Date	A timestamp string indicating the year, month, day, hour, minute, second, and time zone the ad was posted.
Title	The title of the ad as posted.
Description	The long-form body of text authored by the posting user.
Domain	Specifies a jurisdictional subsection, usually a municipality, of the Vancouver metro region the ad is posted to.
Latitude and Longitude	Basic geolocation in decimal degrees.
Location Accuracy	A numeric variable ranging from five to twenty-five, indicating whether the latitude and longitude specify an exact or approximated location based upon a postal code. Values of five indicate exact locations placed by the posting user, while values of twenty-five indicate approximated locations. The data did yield values between five and twenty-five, but it is not known what these intermediate values signify.
Location	A nominal location specified by the user. It can name a municipality, district, or street address according to the user's discretion. Some ads returned blank values for this field, indicating that it is optional to the authoring user.
Map Address	Another optional field that users can populate an address underneath the map on the ad's webpage.
Price	Asking price for the advertised housing.
Number of Images	A count of the number of images, usually photographs of the property, provided by the user.
Rooms	A nominal classification variable. Classifies each ad based on its type: private room, bachelor, or one to eight-bedroom units.
SQFT	Specifies the areal size of the housing if provided by the user.
Source	Cites the data source. In this study, all data was derived from Craigslist.
Tags	Captures a variety of options selected by the user by drop down menus and check boxes upon ad creation. Options include, but are not limited to, pet friendliness, availability dates, amenities such as laundry or parking, number of bathrooms, etc.
URL	The full URL of the ad's webpage or hashed archive in local storage.

Most variables were used to assist in processing and only price, latitude, longitude, SQFT, rooms, and date were retained for analysis after processing.

2.1 Deduplication

A duplicate ad is one that represents a rental opportunity already posted; duplicates do not need to be verbatim copies in this case. If left unaddressed, a duplicate can over-represent the particular rental opportunity in the dataset and distort analysis. Therefore, it is critical to ensure duplicates are reduced to a single ad per unique rental opportunity.

Duplicate ads can exist for two reasons: repeat scraping and repeat posting.

Repeat scraping occurs when the tool mistakenly scrapes the same ad more than once. This is most easily identified and a straightforward matter to correct. Repeat postings are more common and complicated to identify occurring when a user wants to improve the visibility of their ad to prospective renters. This may be done to "bump" the ad to the first page of results one sees when first opening the website. Others may adjust the price or other content to improve responses to the ad. Some will flood multiples in a short period of time to take up more than one result on directory pages. As users will often alter details to avoid automated duplicate detection on the website, no single variable can reliably be used to identify all duplicates. Conversely, building managers may use copied language to advertise multiple similar yet unique units. Systematic elimination of these ads may under-represent their volume in this case. An analyst may frequently need to refer to multiple variables and rely on their discretion in determining what to eliminate.

- Begin deduplicating by URL. This will catch any duplicates from multiple scrapes.
- Deduplicating by title alone may create false positives as some unique ads will be posted with simple titles such as "house for rent".
- Deduplicating by location alone is unreliable. Where the poster manually places pins to specify a location, there is a nearly inevitable variance in even carefully placed pins, especially since the geolocation uses six digits of significant precision. When low-accuracy locations are derived from postal codes, duplicates may share a geolocation with other unique ads.
- Multiple postings may be made by larger building managers and landlords, each for a unique unit within a building or across multiple buildings managed by the same user, using similar language in titles and descriptions. Often the rent, square footage, and bedroom count may indicate variance, but not always. Given how multi-unit towers gain an economy of scale by repeating floorplates, unique units may exhibit identical layouts and thus square footages. If being posted by the same individual or business, the proposed rent will be the same as well. These can be problematic and the analyst may use their discretion in judging whether the ads are for unique dwellings or altered duplicates for the same dwelling. Sometimes unit or floor numbers will be listed in the description, helping distinguish unique units that would otherwise appear as duplicates. Some brokers will list an internal serial number for the posting. If the building is declared or determined - through visual inspection in situ or via aerial imagery - to be newly constructed, it is possible that one owner recently purchased several

identical units and is using the same ad as a template to release them on the market at once. Although it is time - intensive, if the building's address is listed, the analyst may elect to explore it in Google Earth to ascertain the likelihood that it may have multiple units within.

- Rent and square footage are not completely reliable indicators on their own and can be carelessly typed in by users. Sometimes a repost within a month will change the asking rent relative to the original. Declared square footage may also change from the original posting to a repost for various reasons, perhaps to improve the perceived value of the unit by including previously uncounted area such as balconies or shared laundry rooms in a house. If the dataset is sorted according to these attributes, the duplicates will not appear next to one another. In cases where a stated address is in a multi - unit structure, such as above, it may be best to treat these near - duplicates as unique units and investigate according to the analyst's discretion.
- Occasionally a posting (such as a $3,295 loft on Manitoba Street posted at least four times in December 2019) will deliberately alter each duplicate such that neither title, tags, nor description will be perfectly identical. This makes flagging the duplicates difficult using automated processes yet straightforward to manual audit.
- Multiple rooms can be posted in a single private room ad with different prices. As long as the rooms were apparently at the same address, the ad was retained. Posters often list the lowest ad in the price field to attract more views. In such cases, the price field was edited to reflect the average price of the rooms. Retaining these as single ads will underrepresent the full volume of offerings scraped, thus this policy could be modified in the future to break the ad up into unique data points for each constituent room offered.
- Another phenomenon in the private room category is the incidence of identical ads where a poster is genuinely offering more than one room but choosing to re - use the contents verbatim of one ad to do so. If the description or title indicates this to be the case, these were retained as the ad volume would more accurately reflect the volume of opportunities.
- Sometimes an ad is reposted with a lower price than the original, presumably to either improve the appeal of an offering that did not elicit satisfactory responses through the month it was posted. A reposted price may go up if responses were overwhelming or if the poster believes they can capture prospective tenants choosing/forced to move on short notice and will "take what they can get". When choosing which post to retain, location accuracy took priority. If the location accuracy is different on one duplicate, the more exact posting was retained. The secondary rule prioritized which, if any, duplicates contain more filled variables. Sometimes, for example, square footage is omitted in one or more duplicates. The final criteria was to retain the youngest version and delete the earlier instances.
- Sometimes an ad will be posted in both the private rooms and apartments sections of Craigslist, or "cross - posted". As they are scraped separately, this duplication will only be detected when the spreadsheets are merged into a complete dataset for the analyzed time period. These may also be priced in the low 100's - 1000$ range and a highlight of cells with "shared" in the contents of the description field can help identify such ads.

2.2 - Cleaning

Besides duplicates, many ads will not meet the study's requirements for a usable ad and need to be eliminated from analysis. Some are not a genuine offer of long - term rental housing, others are but exhibit idiosyncrasies that either warrant correction if correct information can be found in the title or description or deletion if not. In addition to distorting the volume and composition of postings, and because so many of these ads exhibit unrealistic prices and sometimes square footage, leaving them in the dataset unedited will create extreme outliers.

- Postings below 1000$ will generally require more scrutiny. Although many will be usable – particularly in the private rooms category - most prices that seem "too good to be true" will be found here.
- Stated rental prices not likely to be sincere asking prices for long - term rental, often finding overlap with private bedroom listings. Typically, these are 1$ but may also drift to other low values. Some are legitimate ads with the real rent listed in the title or description which will need to be manually edited to the price field. Some are legitimate ads but no rent is cited, either offering the place for free or making it variable on childcare or other household labour in - kind of rent. These should be deleted from quantitative analysis but can be noted in more qualitative discussion.
- Others are not proper ads for rental housing but people selling other services (cleaning) or calling out other users for fraud, poor - faith in market interactions, invasions of privacy, abuse, or other offences. Cleaners may be advertising a per - job rate of, for example, $30, while "warning" ads can exhibit peculiar prices to the single digits ($1234, $666). Such ads can also be found by conducting a search operation for the following words: "scam", "fraud", "warning", "slum", "illegal", "criminal", and "slumlord"
- Short term listings in general are outside of the scope of what this study was concerned with. Rates only listed for daily or weekly bases were deleted from processed results. If an ad included a longer - term monthly rate, it was retained with the proper rate entered in the price field. Although a 12 - month lease is considered a standard for long - term leases in British Columbia, often times the involvement of a lease is not mentioned and month - to - month terms still imply the opportunity for long - term tenure. "Medium - term" listings were retained if a monthly price figure is stated. Apparent sub - lets for a finite period, where a signatory tenant may be leaving town briefly, was admitted to analysis as they represent a long - term rent but should be noted in studies more concerned with market behaviour.
- Weekly or daily listings can usually be identified in the title or description and are priced in the 50 - 1000$ range. These are often hotels, motels, or hostels renting out their rooms, some of which also offer monthly rates without leases (a common clue).
- Targeting the word "weekly" in the description is a relatively reliable conditional term to sift out these ads,

though one should review the language of the ad as sometimes it is used to describe the periodicity of services like garbage pickup or cleaning services in usable ads. Searching for the term “hotel” is not entirely reliable either; in fact, most apparent hotel ads did not explicitly claim to be so. Some ads for non ‑ hotel units will use the word to describe the style of amenities ‑ often in condominium buildings ‑ or even the quality of their concierge “trained by the Fairmont Pacific Rim Hotel”.

- Some hotel ads declare monthly rates for their rooms and were also deleted. Although these will accurately factor into analysis of other monthly ads, this study elected to focus on dwellings intended for permanent residence.
- At least one such ad offered monthly rates – a 5 bedroom listing by Vancouver Luxury
- Rentals – was posted both in September and the following January with no intermediate repostings. This may suggest that it was successfully rented out for a short term and brought back to the market, either by design or a prematurely terminated lease.
- There are also non ‑ hotel ads for private dwellings or rooms that wish to charge a weekly or daily rate. If a monthly rate is also stated, it was entered into the price field. If no monthly rate was offered, it was deleted.
- At least one ad in Jan 2020, posted by a management company named MakeYourselfAtHome, was advertising for a Jan ‑ Apr term, and described a complex of pricing for the following year in four ‑ month brackets that would vary with the time of year. Since the rent would distort averages if considered as a per ‑ month figure, it was deleted. These kinds of ads are not evidently common.
- Renters seeking dwellings sometimes post here, stating the price they are willing to pay.
- These ads were deleted. These can be found with “want”, “need”, “seeking”, “In search of”. Discretion, however, should be exercised over systematically deleting ads with these terms, as they can also be found in valid ads, specifically in private rooms, where a landlord is specifying desired tenant qualities.
- Some types of ad allowed by Craigslist were deleted from the set. These included rental of mobile home rentals, recreational vehicle pads with utility connections, camper vans, farms, barns, yurts, and horse stalls.
- Occasionally a small landlord may post multiple, apparently detached, dwellings in a single ad, applying the rate of only one dwelling to the ad and listing the constituent properties in the description. One in December of 2019 listed several single rooms across multiple detached properties. Although each constituent dwelling could be disambiguated manually and separate entries created using the rent and unit type on offer, only one geolocation is declared by the ad. As each dwelling would needs to be manually geolocated, provided a street address, it was determined that the labour overhead was untenable and such ads were not retained.
- Managers of multi ‑ unit buildings and purpose ‑ built apartment buildings may also aggregate opportunities within the same structure into a general ad, often submitting the price and type of the most affordable and smallest unit to the scraped fields and then breaking down the multitude of offers in the description. These

cases were retained and treated as a single offering, introducing underestimation bias. Revisions to this method may create unique line items for the separate opportunities in the building, depending on the quality of documentation in the description field or supplemental resources such as the management company's website.

- Some private room listings will cite a different rate for a single occupant vs. a couple.
- The single occupant rate was applied in these cases.
- Many ads in the higher price range of the private rooms category are possibly full unit ads. This appeared to become prevalent at the 1500$ point and such postings should be scrutinized, have their rooms field populated appropriately, and be checked for cross - posting duplication. Using search commands to isolate bedroom counts in the description or title are not very useful as many usable private room ads are searching for room - mates in multi - room units.
- Any "Free rent" ads were deleted.
- At least one instance was posted by a "Rent - to - Own" agency. The description disclaimed that the property advertised was not itself for rent or sale but an example of the kinds of properties in their portfolio. Future cases may exhibit a property actually on offer and all necessary fields may be populated. However, the so - called "rent credits" likely distort the asking rent of the place. In all cases, therefore, these were deleted.
- Some places may be posted with alternative furnished/unfurnished prices, either as
- options within a single ad or two separate ads. The unfurnished price was used in this study.
- At the larger bedroom range (5+), sometimes a per - room rate will be posted, often coming from the private room section of Craigslist. These were checked against their descriptions as these low rates can distort the price per bed estimation of these larger unit types which have a relatively small data point volume.
- Extreme square footage fields were audited. Outliers in a rent/sqft calculation can help find ads where a per - room or per - floor rate is listed in a house or multi - unit structure but the area or bedroom count are listed as a total for the whole structure. For example, a $980 per month rate for a single room advertised in a five - bedroom 2000sqft house may either process as a 2000sqft bedroom or a $980 five - bedroom house depending on how it is classified.

2.3 – Attribute Population

The third way veracity can be degraded is in missing or erroneous core attributes. Craigslist users are only required to submit a minimum of details when creating an ad and not required to verify the accuracy of such details nor format them in a machine - readable way. Alternatively, errors in the scraper or a change in the HTML structure of the website may cause a failure in capturing an attribute correctly, if at all. Capturing the title and description text from each ad enables the human analyst to infer the nature of the error and edit the correct value into the core attributes, recovering data points that would otherwise require omission from analysis. Because this process relates to the challenges of cleaning unusable ads, some examples have already been outlined above.

In a general sense, where prices are unrealistic for the size of the unit – such as the single bedroom price listed as a five ‑ bedroom house in the previous section – an analyst may be able to correct the price or rooms attribute based on title and description rather than discard the ad completely.

- Users commonly neglect to enter square footage into the purposive field when creating
- an ad. As noted elsewhere in this report, nearly 40% of ads after deduplication and cleaning were without square footage values. A noticeable proportion, however, will declare it in the description or body of text, so future studies may find it necessary to populate the SQFT field from these fields.
- A similar issue exists for bedrooms, another elective field, to a lesser extent. As this
- detail is as important to renters and landlords as it is this study, it was rare to find an ad with no mention of bedrooms in the title or description. Some cases listed dens as optional bedrooms and perhaps have declined to specify an exact number on that basis. Ads which offered multiple units were also found to not specify a bedroom count. An extreme few would be so minimal that only the required fields were entered, including a very brief description.
- Because "bachelor" and "studio" are not selectable types nor bedroom counts when posting an ad, they must be identified by the analyst. The same process will apply if one wishes to distinguish other non ‑ standard types such as basement or penthouse suites, for example.
- Many bachelor and studio suites are posted as having no bedrooms, as per the above
- point, others as one bedroom. If using a software search function for the type of a unit, one must be aware of the multiple meanings of the term and be mindful of false ‑ positive results. Previously discussed issues note how "loft" and "hotel" can confound searches. For another example, if searching for the term "studio", one must mind that some descriptions list amenities such as "dance studios" or nearby points of interest that will catch such formatting without actually indicating a studio unit for rent.

3 – Processing Strategy and Techniques

The volume of raw data and multi ‑ variate plurality of error therein make the above processing a complex task, whether performed manually or when coding for automation. If left unsorted, or only sorted on a single variable, an analyst would likely need to go over the entirety of thousands of data points multiple times to deduplicate, clean, and populate a dataset such as this. Because of the manifold sources of error, evaluating a single ad on all possibilities at once risks missing some. Analysts are unlikely to identify duplicates if they are dozens of lines apart. Redundancies in effort may occur where an analyst may populate an ad's attributes from the description only to delete it as a duplicate later.

A fundamental strategy to efficiently performing this processing manually in large tabular datasets is organizing the data such that the analyst can notice potential duplicates with a quick pass of the eyes. This involves sorting the table entries for each ad in various criteria such that potentially duplicated ads are adjacent or near one another in the spreadsheet. To this end, "Conditional Formatting, "Find and Replace", and custom sorting are tools in Microsoft Excel for efficiently cleaning, formatting, and deduplicating the dataset. It was found to be useful to have duplicate values of certain attributes visually flagged as coloured cells to obviate relevant patterns to the analyst once the similar ads are sorted on a basis of that flag.

Conditional Formatting: Conditional formatting uses various rules to change the appearance or order of selected cells. For this study's purposes, "Duplicate Values" and "Text that Contains" under the "Highlight Cell Rules group" are most useful in identifying duplicate values or capturing mention of certain terms of interest such as "basement" "short - term" or "fraud". The table can then be sorted by the cell colour of a specific attribute to place all highlighted content adjacently to user navigation through the table. The conditional formatting rules are persistent so, for example, when a set of duplicates is reduced to a single ad, now unique, that ad will lose the formatting colour that flagged it as a copy.

Find and Replace: Find and Replace can be used to remove unwanted characters, such as "/ ", excess spaces or carriage returns that were captured from the original HTML code and may visually clutter the dataset. The same function can be used to find keywords that can flag an ineligible ad in ways that sorting does not obviate. A third application of this tool can serve the exploration of novel classification schemes or linguistic analysis by highlighting the prevalence of certain words or tags.

Concatenation: As mentioned, many of challenges in addressing duplicate ads is in their multivariate nature and exceptions to any rule that might otherwise be simple to automate; flagging and sorting by a single attribute may not adequately reveal all true duplicates or sort out unique offerings with shared values. Creating new columns that are concatenated strings of other attributes can help sort out some of the more likely duplicates from unique offerings that simply share a price, location, or title. This essentially reveals where multiple attributes are coincidentally common and, hypothetically, more likely duplicates. One useful combination includes latitude, longitude, square footage, and room count. One can then conditionally format a colour flag to duplicates in this field and sort the dataset first by cell colour of the concatenated column, then value of the concatenated column, then by date.

One should note that, due to the approximation of geolocations where accuracy values are 25, many unique ads will be collected with duplicates when latitude and longitude are leading components of concatenated fields. It was found that conditionally flagging duplicates in the title field, arranged beside the concatenated column, helped analysts notice more likely duplicates in the table. It is advised that price not be a component of the concatenated field as many reposted ads involve a price change and would be separated by different prices by the sorting operation.

A different issue occurs when users reposted ads with more accurately specified locations – location accuracy values of five or seven. While they are almost always more accurate than the approximated locations with larger accuracy values, the pins are manually “eyeballed” thus it is unlikely two placements will yield the same precise latitude and longitude. What this means for deduplication detection is that if the data is sorted according to location as described above, then duplicates may reside farther away from one another in table form, requiring the analyst to search farther up and down the data to notice the set. It was found that a second pass with a different sorting pattern is required to capture duplicates with higher locational accuracy: title by duplicate – flagged colour, title by value, concatenated location/sqft, date.