Worldwide Short-Term Rental Market Analysis

Connor James

Ryerson University

Dr. Erdem

Student ID: 501007638

March 28, 2022

Ryerson University

Table of Contents

Research Questions

1. How does the price of short-term rental homes in eight different worldwide markets compare to each other in 2021 on the Airbnb platform?
2. How can we better understand the weight each variable has against the price? Which variables influence the price the most? Which variables influence the price the least?
3. What patterns can be found using text mining in the unstructured description text of the listings and how can it be related back to the price?
4. What will the price of short-term rental market Airbnb homes be in the future?

Literature Review

For the last fourteen years, Airbnb has been growing exponentially in the short-term rental market. When a company grows as fast as Airbnb has, they're bound to be the subject of several academic journals and papers. Studies related to the vacation rental company have been written on topics varying about the company's economic impact to their legal issues, as well as papers using predictive analytics to predict the price. The goal of this report is to extend previous analyses done by several different credited sources with recent data for eight different cities around the world. When an individual wants to buy a short-term rental home or list their property as a short-term rental, there are many variables associated with the property to consider. For example, previous studies have found that in Valencia, Spain, location played a large role in the overall cost of a rental unit. Accommodation prices increase incrementally by 1.3% per kilometer from the tourist area, however, at the same time, accommodation prices decrease incrementally as the distance from the coastline increases. (Perez-Sanchez 2018). This may be true for certain markets, however, the need to be close to tourist-heavy areas may vary based on what the city has to offer. Other studies have also found that the hotel prices around each listing and the number of competing AirBnbs in the same neighbourhood can have both complementary and detrimental effects on the price (Mango 2018, Önder 2018). Further investigation should be done around other cities around the world as a comparison tool. Also, with the world shut down for large periods of time in the last two years, it's worth investigating again as the priorities of the consumer may have changed. Other studies have found that the number of reviews and the

review rating scores play an important role in the price of a rental unit (Chen 2017). This plays into the report that the reputation of the host as well as the hosts' policies related to the property. While these definitely play a role, this report will combine the community perception variables such as the number of reviews with more tangible variables such as the number of bedrooms and number of bathrooms. What's missing from the majority of these academic journals is the comparison between different markets. These studies tend to grab data for one central location such as Nashville or New York City; they don't look at understanding how/if the markets can be different from each other. One of the research questions this report will focus on is how the price of short-term rental homes in eight different worldwide markets compares to each other in 2021 on the Airbnb platform. Given that the datasets use active listing data from the past year (January 2021 to January 2022), the data that will be used to train my models will be different from any academic papers currently published. Analysis of Airbnb data has been done using a diverse amount of models, such as OLS, GLM & Linear Regression. Previous reports have attempted to predict the price using a simple linear regression model, however, they've all been conducted with historical data; this report will be using updated data from 2021-2022 to further the investigation. There have been a few contradicting points of view related to text mining the description as one study found using text mining on the description that the location-specific characteristics are not significant in the majority of the cases except for ski with a significantly higher price (Falk 2019); however, similar studies, found that location commands the highest price (Chew 2017). This might be caused by the location each of these reports was completed in

as one was done in a bustling tourist city of London, England and the other was done for the entire country of Switzerland. This report will cover cities all over the world to allow for a deeper comparison of locations that may differ by culture and economic standards; this sentiment was echoed by (Zhang 2017), who found that hosts lower their prices to attract more guests in cities where Airbnb is not as prominent. This report aims to answer that question by text mining the description variable for multiple different cities. In summary, although there have been previous studies conducted around predicting the prices for Airbnb listings, the findings should be extended as prices can vary from location to location. Furthermore, the COVID-19 pandemic has changed the mindset of how society operates in regard to public health and safety measures. Further investigation would be beneficial around the pandemic's impact on the market and where the market is headed in the future.

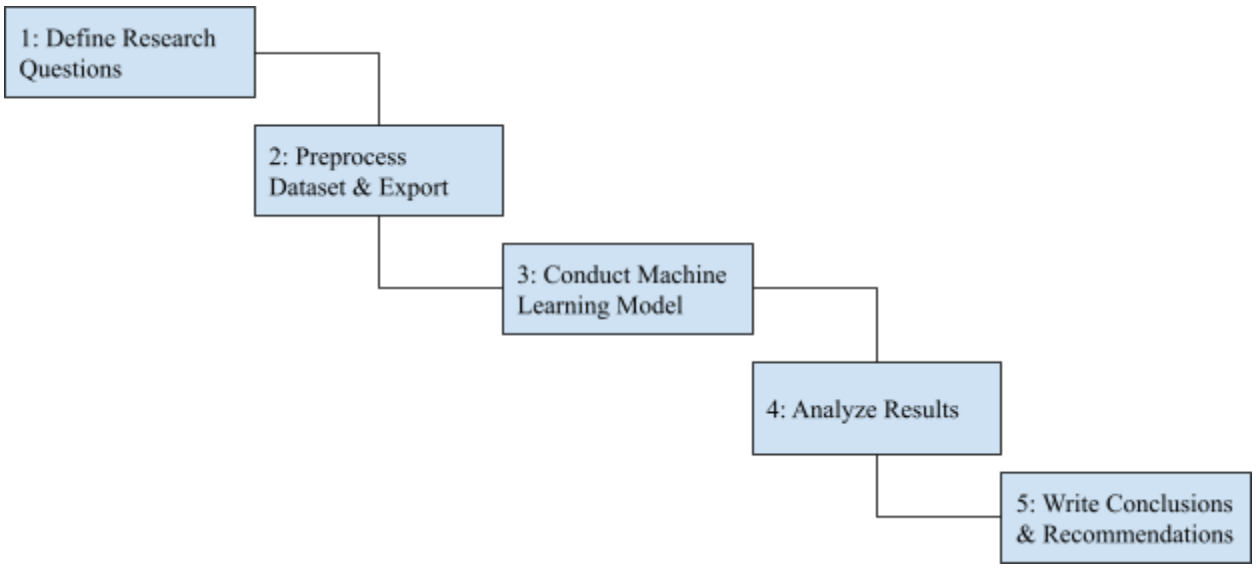
Github Repository

All code and results will be uploaded to the following Github repository:

https://github.com/cjames99/CIND820-capstone

Methodology

Overall Methodology Flowchart



Step 1: After determining a topic and finding a dataset(s), I need to search through the data to better understand its variables and come up with 3 to 6 research questions.

Step 2: Each dataset will need to be cleaned for outliers and null values to be ready for the machine learning model.

Step 3: Machine learning models will be conducted on the dataset.

Step 4: The results from the models will be analyzed to uncover key differences in the data.

Step 5: Conclusion and recommendations will be given based on the results found in Step 4.

Before I can conduct regression models and text mining analysis, I need to understand the data inside the eight different datasets I have by following standard data preprocessing techniques. I'll need to check for any correlation between the variables as well as identify any outliers. I'll also want to figure out key metrics for each variable such as their mean, median, maximum, minimum, standard deviation, etc. Given I want to predict the price variable, I'll use boxplots to identify which values are in the IQR and remove the outliers. After that, I'll use variable graphs such as scatter plots & stacked bar plots to understand the patterns within the values for each variable. Furthermore, I'll need to examine each variable closely to identify if I want to remove it entirely or fill in any null values with the median. From here, I can compare the price, the accommodation, the average number of reviews plus several other variables between cities to better understand the market and derive comparisons/conclusions between them. After I've compared the prices, I'll next take a look at the description variable. Here I'll be using the Latent Dirichlet Allocation (LDA) model to text mine the description to identify any recurring themes. The process starts off by creating a corpus of the description and removing stop words, making all characters lowercase, and using lemmatization. From there, the clean corpus will be indexed for each unique word into a document term matrix. After that, the LDA model is run to uncover topics rooted in the corpus. Once the LDA model is run on all datasets, the topics will be displayed on their respective intertopic distance maps and compared to each other. Adding on the text mining analysis, word clouds will be created for each dataset to highlight the word frequency in the description and used as another comparison tool. The next

models that will be used are three different regression models: Ridge, Lasso & ElasticNet. Individual alphas will be used to determine the model's parameters yielding the best results. RMSE, MSE, Adjusted R^2 & R^2 will be calculated for each regression model and each dataset. Residual error plots will also be created to further examine the efficiency of each model's results as well as using cross-validation for different k-values. These models will also be run before and after standardization to better understand the effect (improvements) standardization will have on the model results. Finally, the XGBoost model will be run to better understand the weight of each variable in the dataset. After examining the initial yield, the XGBoost model will be rerun using the top ten features with the most weight to see any improvements in the error rates. Similar to previous steps, the results for each dataset will be compared to each other to draw any conclusions about the rental markets for each city.

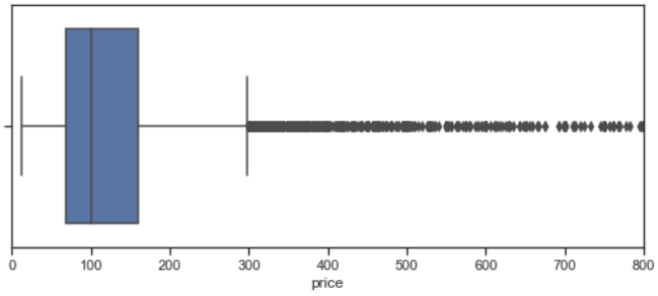
Analysis

Each dataset contains the same 36 variables (13 Qualitative, 26 Quantitative). The list of variables can be seen below.

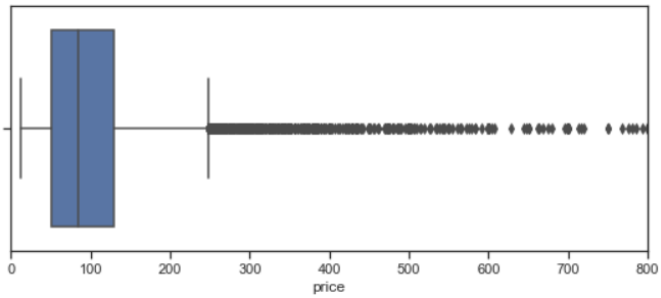
1. description: text used to describe the details of the listing
2. host_id: integer used to identify each unique host
3. host_location: text used to describe where the host is currently located
4. host_response_time: text used to describe how long it takes for a host to respond
5. host_acceptance_rate: percentage on how often host accepts renters
6. host_is_superhost: boolean on if a host is a super host
7. host_total_listings_count: integer of total listings the host has
8. neighbourhood: text on the properties neighbourhood
9. latitude: numerical coordinates of the property
10. longitude: numerical coordinates of the property
11. property_type: text describing what the property is
12. room_type: text describing what the room is
13. accommodates: integer stating how many people the unit occupies
14. Accommodation Bin: text grouping the accommodates field
15. bathrooms_text: text describing the bathroom
16. private_shared: text stating if the bathroom is private or shared
17. bathrooms: integer stating the number of bathrooms
18. bedrooms: integer stating the number of bedrooms
19. beds: integer stating the number of beds
20. amenities: array of amenities the rental property offers
21. price: integer of the rental price
22. minimum_nights: integer of the minimum consecutive nights a renter can stay
23. maximum_nights: integer of the maximum consecutive nights a renter can stay
24. availability_30: integer describing the availability for the next 30 days
25. availability_60: integer describing the availability for the next 60 days
26. availability_90: integer describing the availability for the next 90 days
27. availability_365: integer describing the availability for the next 365 days
28. number_of_reviews: integer stating the number of reviews
29. review_scores_ratings: integer of the review score based on ratings
30. review_scores_accuracy: integer of the review score based on ratings
31. review_scores_cleanliness: integer of the review score based on cleanliness
32. review_scores_checkin: integer of the review score based on check-in
33. review_scores_communication: integer of the review score based on communication
34. review_scores_location: integer of the review score based on location
35. review_scores_value: integer of the review score based on the value
36. instant_bookable: boolean on if the rental property can be booked immediately

Given that the price variable is our target; the boxplots in Figure 1 were used to identify and remove any outliers on the price.

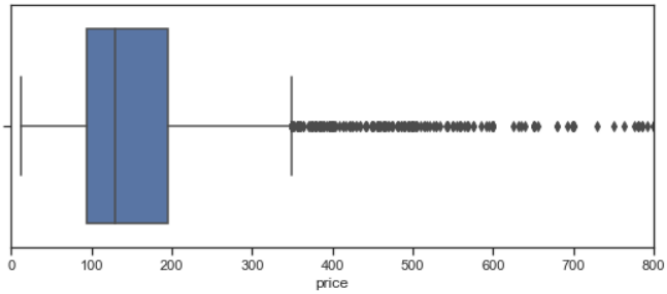
Figure 1: Box Plots: Price for each City



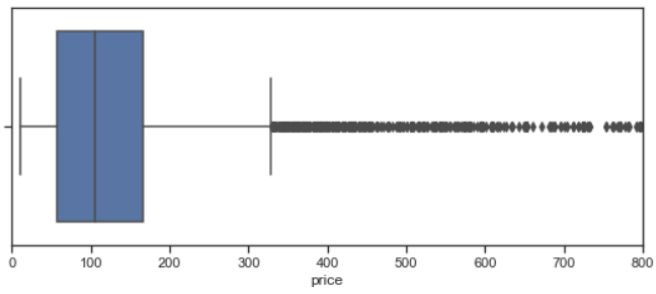
For Toronto, any price over $300 was considered an outlier.



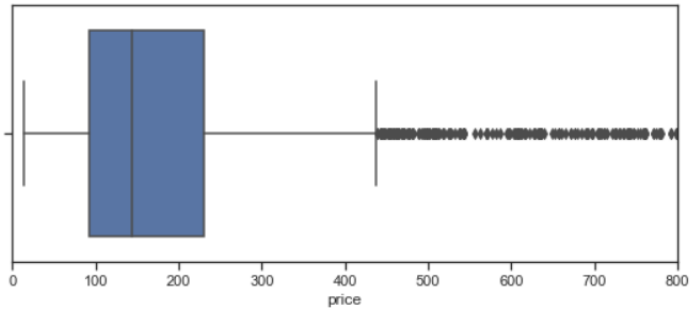
For Montreal, any price over $250 was considered an outlier.



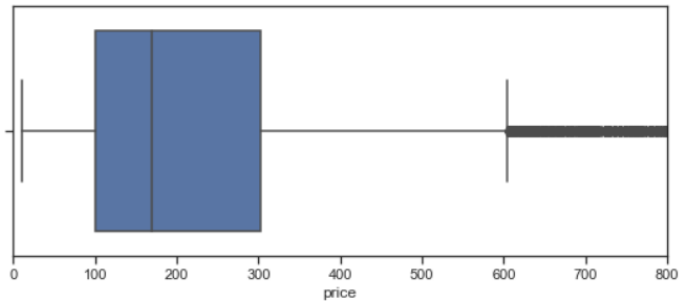
For Vancouver, any price over $350 was considered an outlier.



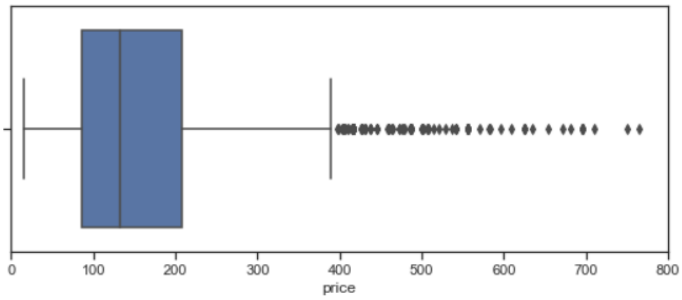
For Barcelona, any price over $320 was considered an outlier.



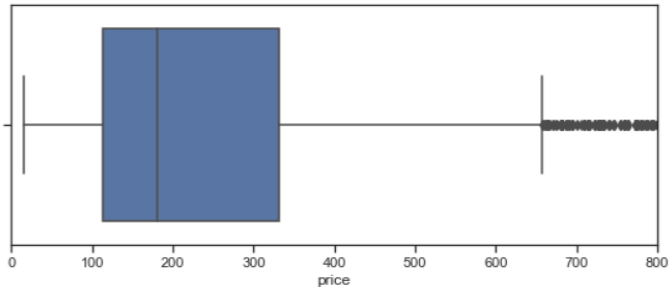
For Chicago, any price over $450 was considered an outlier.



For LA, any price over $600 was considered an outlier



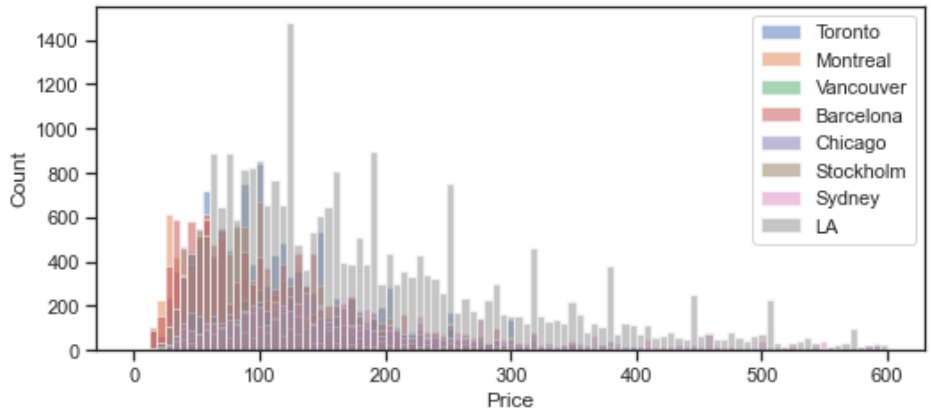
For Stockholm, any price over $400 was considered an outlier



For Sydney, any price over $650 was considered an outlier

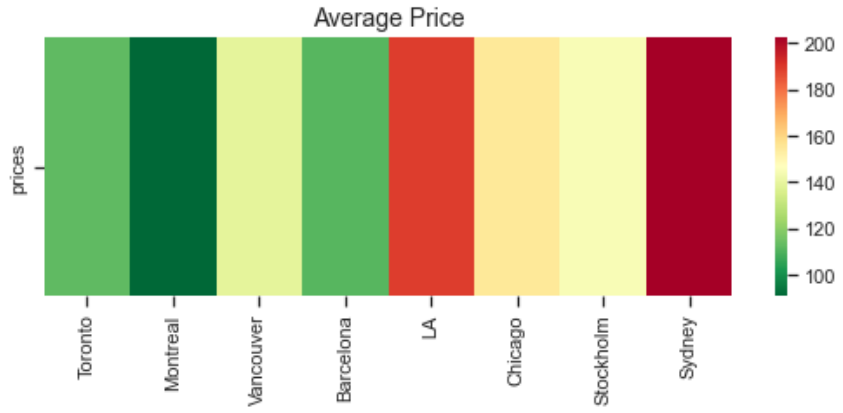
After removing outliers, we can see in Figure 2 that all prices for all datasets follow an approximately normal distribution. This will help in later steps with the model accuracy.

Figure 2: Histogram: Count of Price for each City after Standardization



Without considering any other variables, we can see in Figure 3 that the cheapest short-term rental properties by the city are in the following order.

Figure 3:Heatmap: Average Price per City

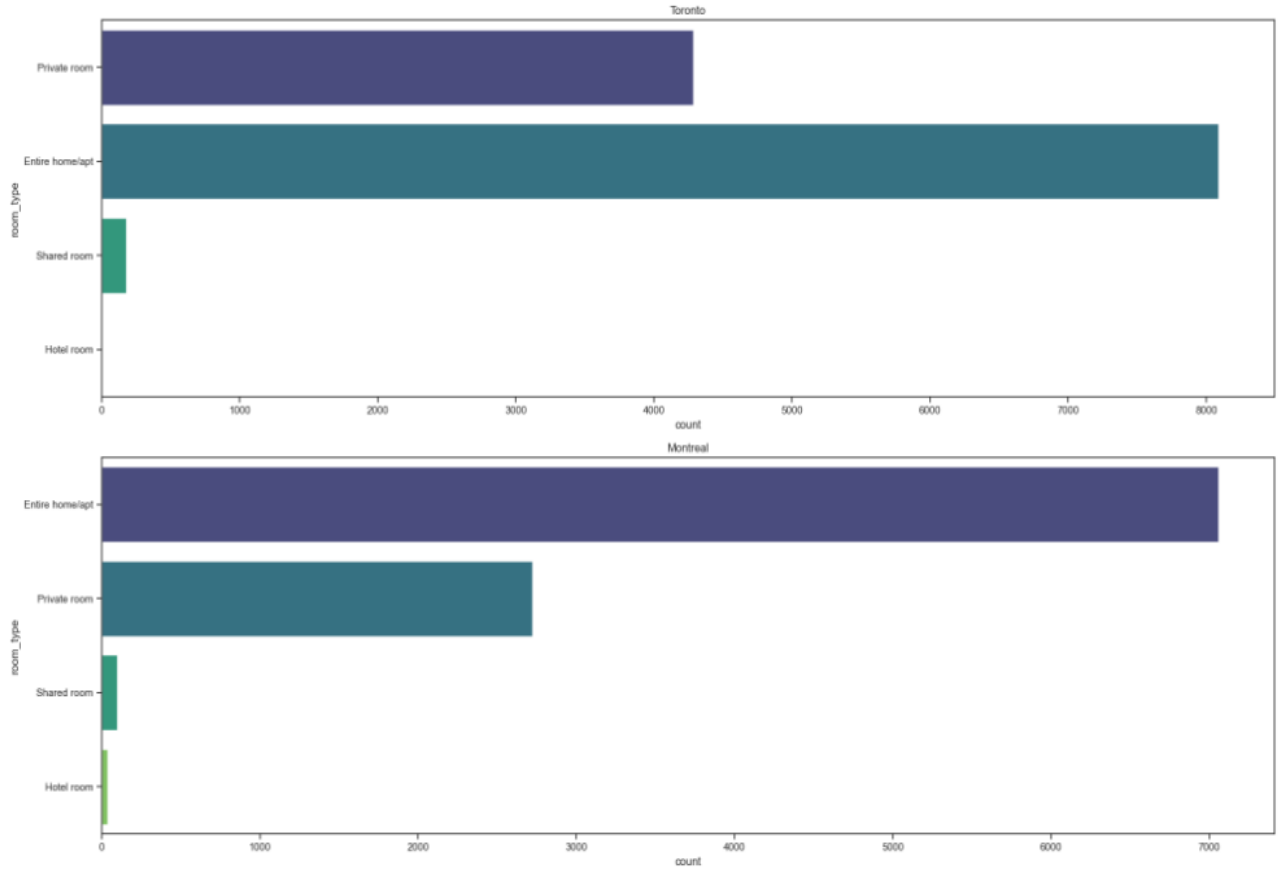


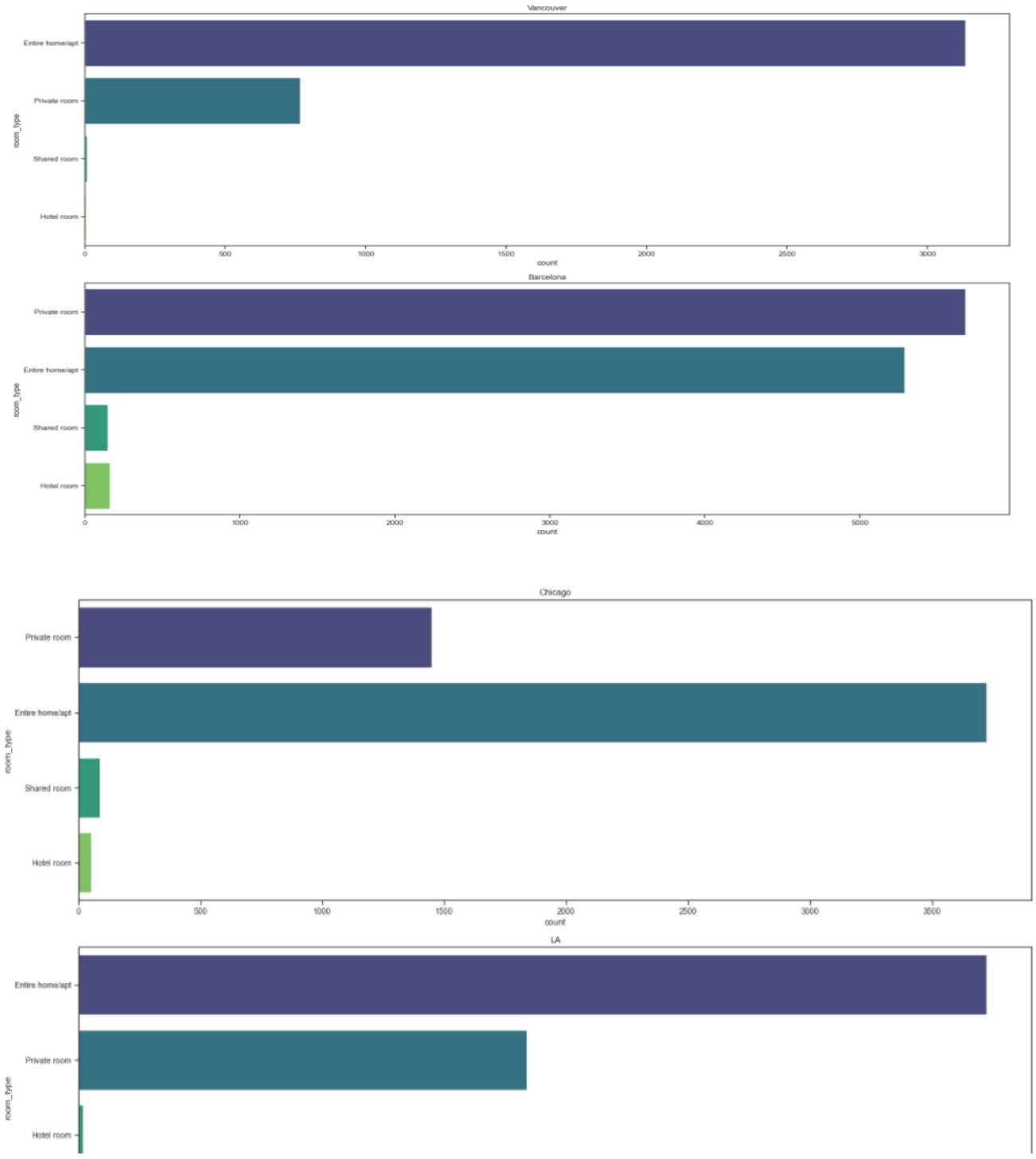
1. Montreal: $90.77
2. Barcelona: $111.08
3. Toronto: $112.34
4. Vancouver: $138.93
5. Stockholm: $144.75
6. Chicago: $154.93
7. LA: $188.88
8. Sydney: $202.76

Each rental property has been classified into the following bins: Entire home/apt, Private room, Shared room & Hotel room. Figure 4 below tells a story that in all cities, almost all available properties are either an entire home/apartment or a private room.

Figure 4: Horizontal Bar Chart: Room Type Count by City

Order: Toronto, Montreal, Vancouver, Barcelona, Chicago, LA, Stockholm, Sydney




Vancouver
Entire home/apt
Private room
Shared room
Hotel room
room_type
0
500
1000
1500
2000
2500
3000
count
Barcelona
Private room
Entire home/apt
Shared room
Hotel room
room_type
0
1000
2000
3000
4000
5000
count
Chicago
Private room
Entire home/apt
Shared room
Hotel room
room_type
0
500
1000
1500
2000
2500
3000
3500
count
LA
Entire home/apt
Private room
Hotel room
room_type

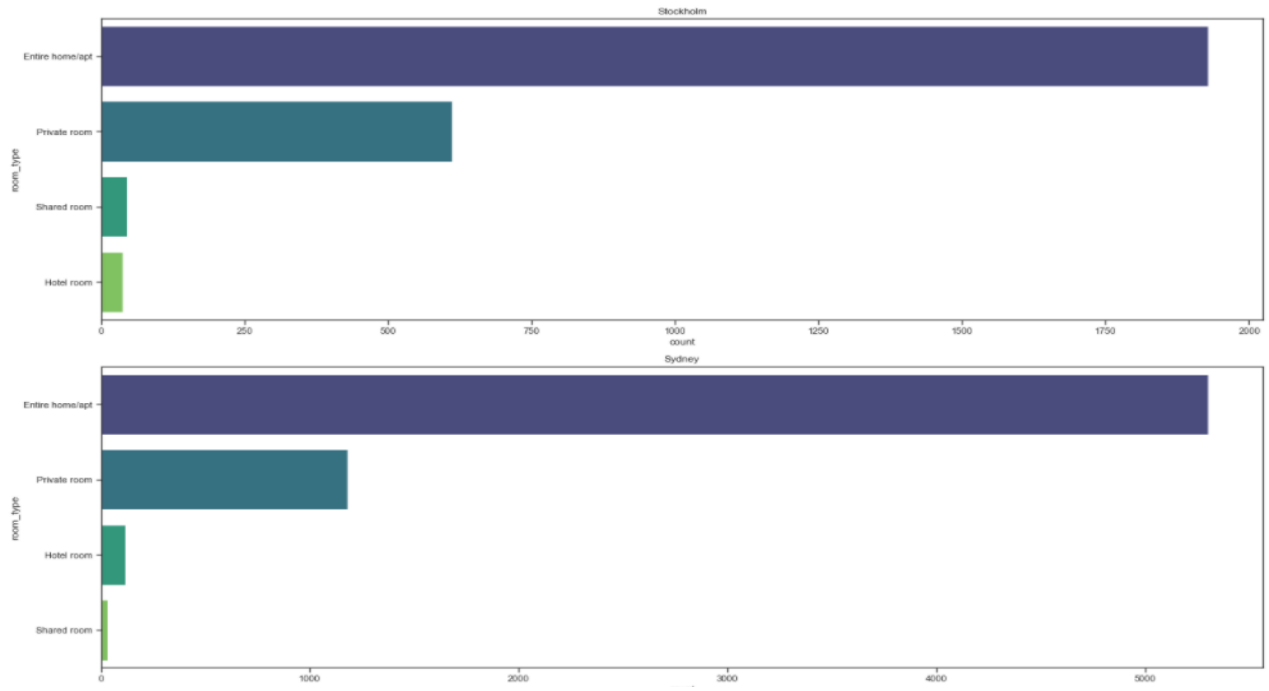


Figure 5 below allows for the conclusion that the price will vary significantly based on the room type selected. Having a shared room is by far the cheapest option. This makes sense as staying in a shared room, you're giving up your privacy and a shared room is unlikely to have several different amenities which will add to the overall cost. This also follows the cadence of the average price per city established earlier on in this report with Montreal being the cheapest and Sydney being the most expensive. Hotel rooms are by far the least frequent room type in the dataset with <100 in the data for each city. Most hotels are likely to have their own platform for users to use which makes sense why there is a lack of data for these room types. Only boutique hotels are on Airbnb for the most part. Private rooms and Entire homes/apartments allow renters

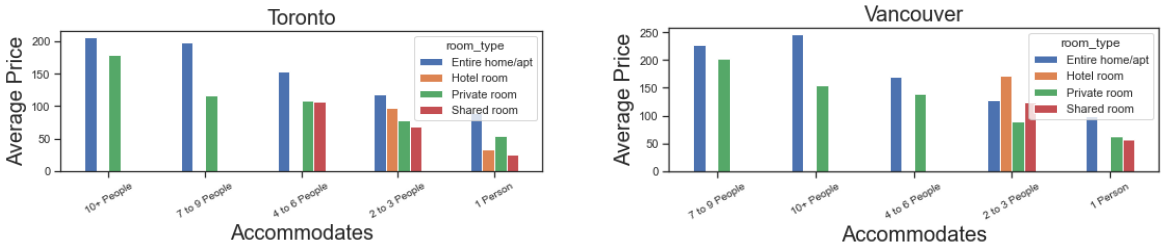
a sense of privacy. Given that an entire home/apartment gives more privacy, is a larger space, and likely has more amenities, it makes sense that this room type is far more expensive than the three other bins. This bin is where we see the largest gap between cities. For example, the average entire home in LA or Sydney is twice as expensive as an entire home in Montreal.

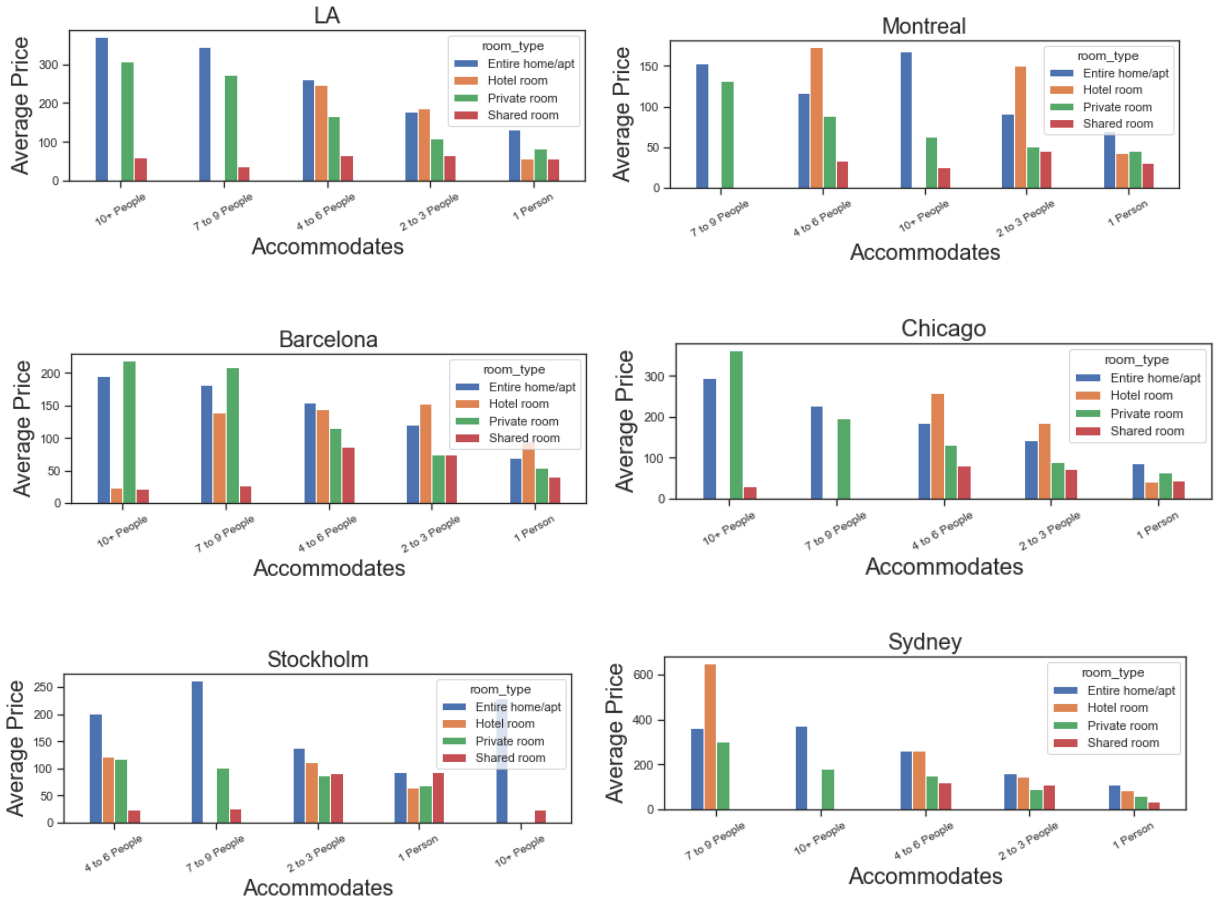
Figure 5: Heatmap: Average Price by Room Type for each City

	Toronto	Montreal	Vancouver	Barcelona	LA	Chicago	Stockholm	Sydney
Entire home/apt	135.700000	106.130000	151.720000	152.800000	232.840000	181.770000	167.190000	229.130000
Private room	71.400000	52.340000	86.980000	72.320000	109.560000	91.400000	81.440000	92.030000
Shared room	40.150000	35.430000	73.750000	54.490000	60.260000	60.980000	76.020000	71.250000
Hotel room	51.920000	136.970000	171.000000	139.180000	190.020000	158.600000	105.770000	161.720000

Adding to the investigation above, the accommodation variable was added to the mix. Each stacked bar chart in Figure 6 below showcases the linear relationship between the accommodation bins. As the accommodation bin size increases, the price increases. Logically, this makes sense; as the more people that are staying at a unit, the more space, beds & bathrooms will be required.

Figure 6: Stacked Bar Chart: Average Price vs Accommodation Bin by Room Type





After sifting through the scatter plots and bar charts for each city in Figures 7 to 14, the number of bedrooms tends to drop off significantly after ~4 and the bathrooms tend to drop off at ~3. As I did for the price variable, the outliers for both bathrooms & bedrooms will be removed to not throw off the model accuracy in a later step. It's also important to call out the significance of each variable. Per the scatterplots above, the number of bedrooms doesn't impact the price very much. There can be expensive 1 bedroom homes and expensive 4 bedroom homes. What is

significant is the bathrooms. Properties with 0 bathrooms tend to have significantly lower prices than properties that do. This makes sense as a bathroom is an essential amenity that most people would want. Bathrooms that are also shared tend to have lower prices than those that are private. This also makes sense as if you have a shared bathroom you're giving up a sense of privacy.

Figure 7: Scatter Plot/Bar Chart: Toronto Bathrooms/Bedrooms

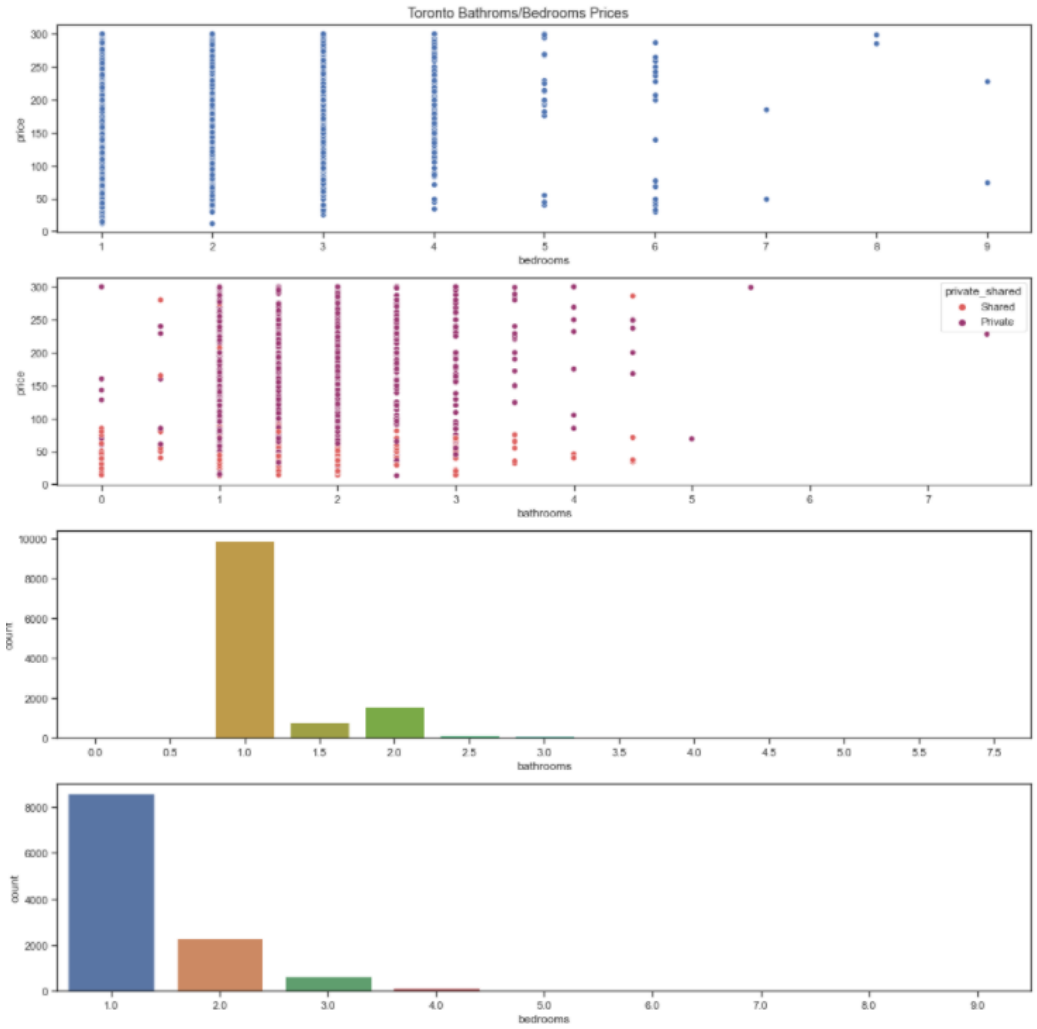


Figure 8:Scatter Plot/Bar Chart: Montreal Bathrooms/Bedrooms

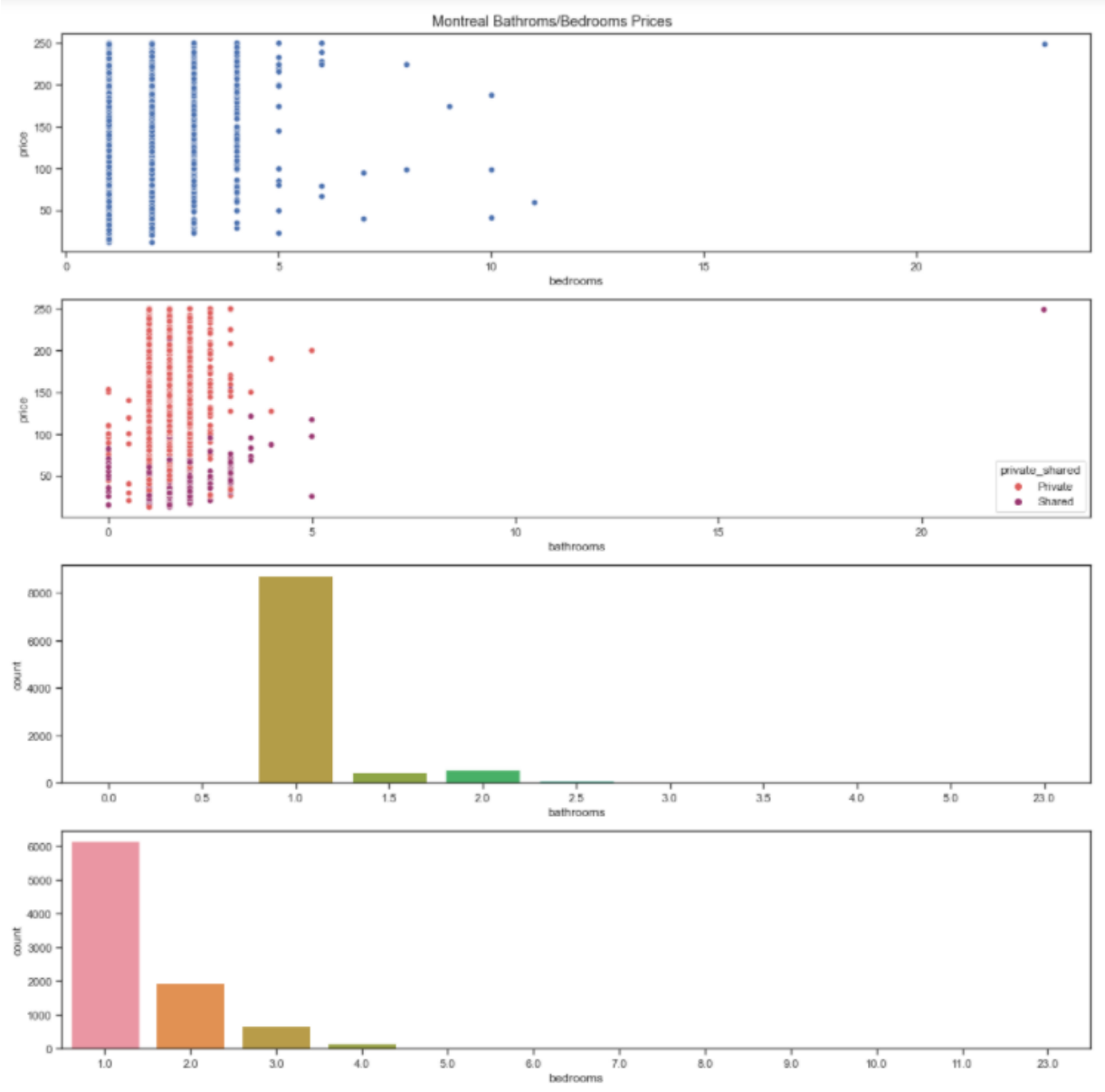


Figure 9: Scatter Plot/Bar Chart: Vancouver Bathrooms/Bedrooms

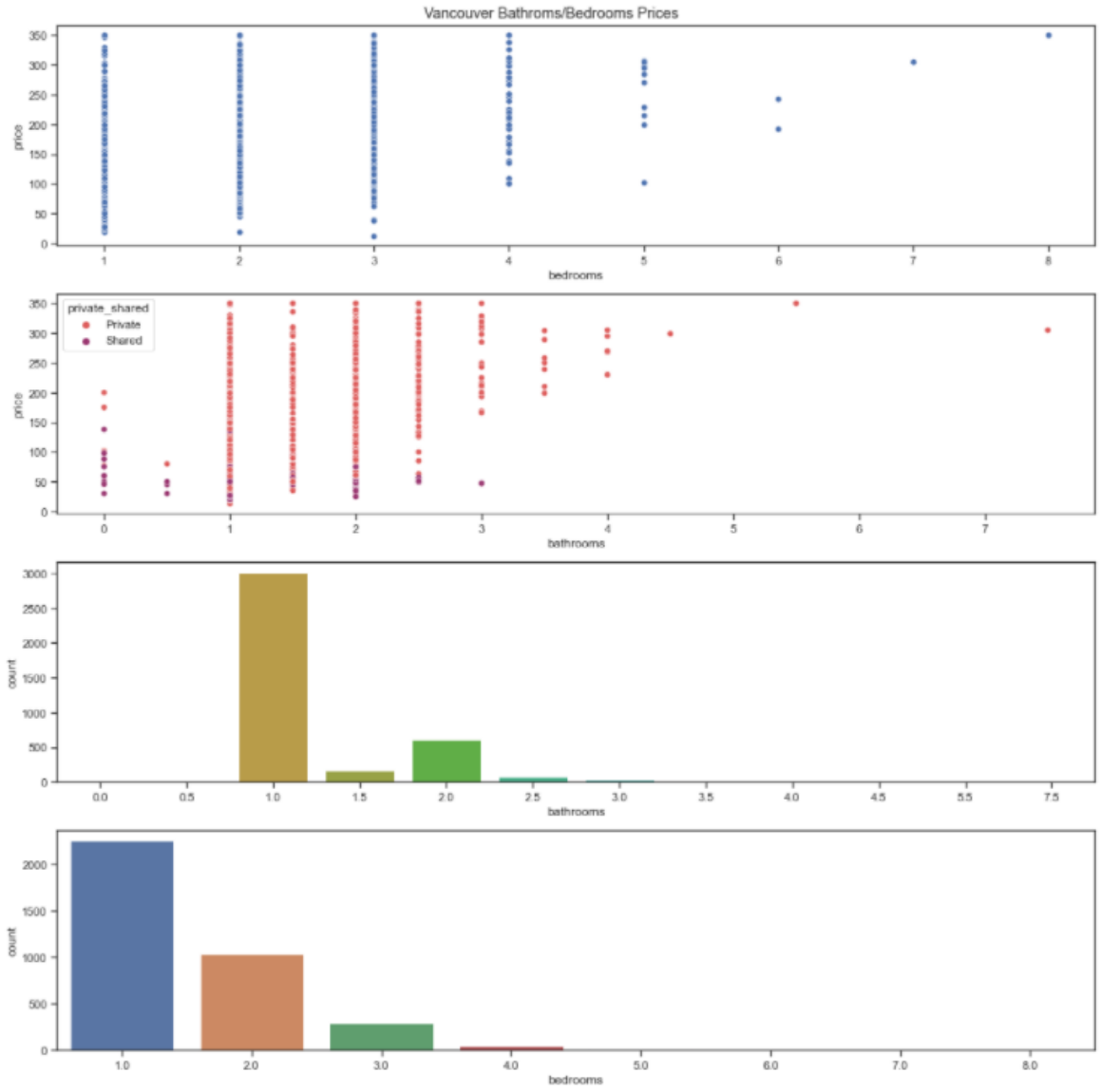


Figure 10: Scatter Plot/Bar Chart: Barcelona Bathrooms/Bedrooms

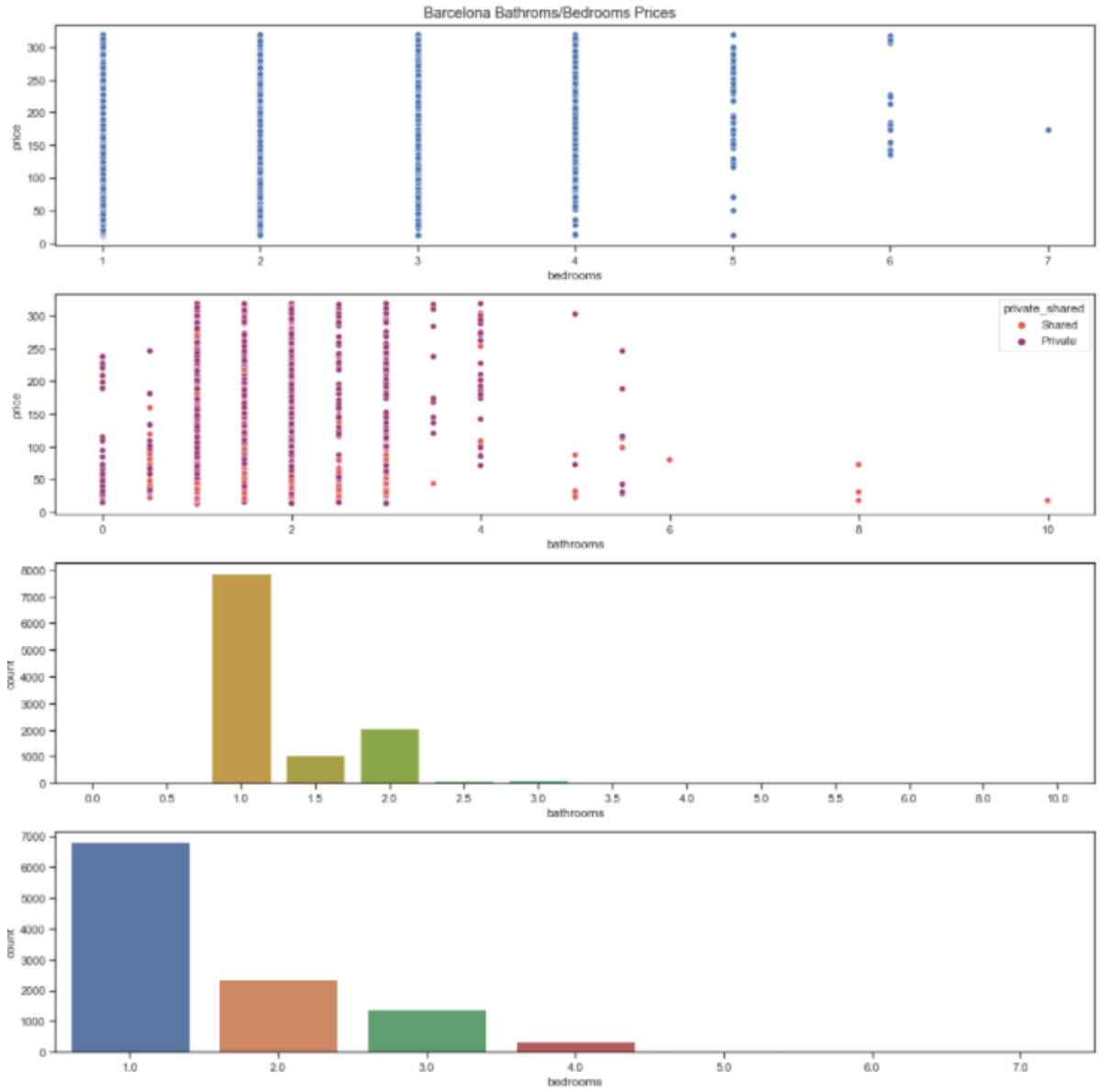


Figure 11:Scatter Plot/Bar Chart: LA Bathrooms/Bedrooms

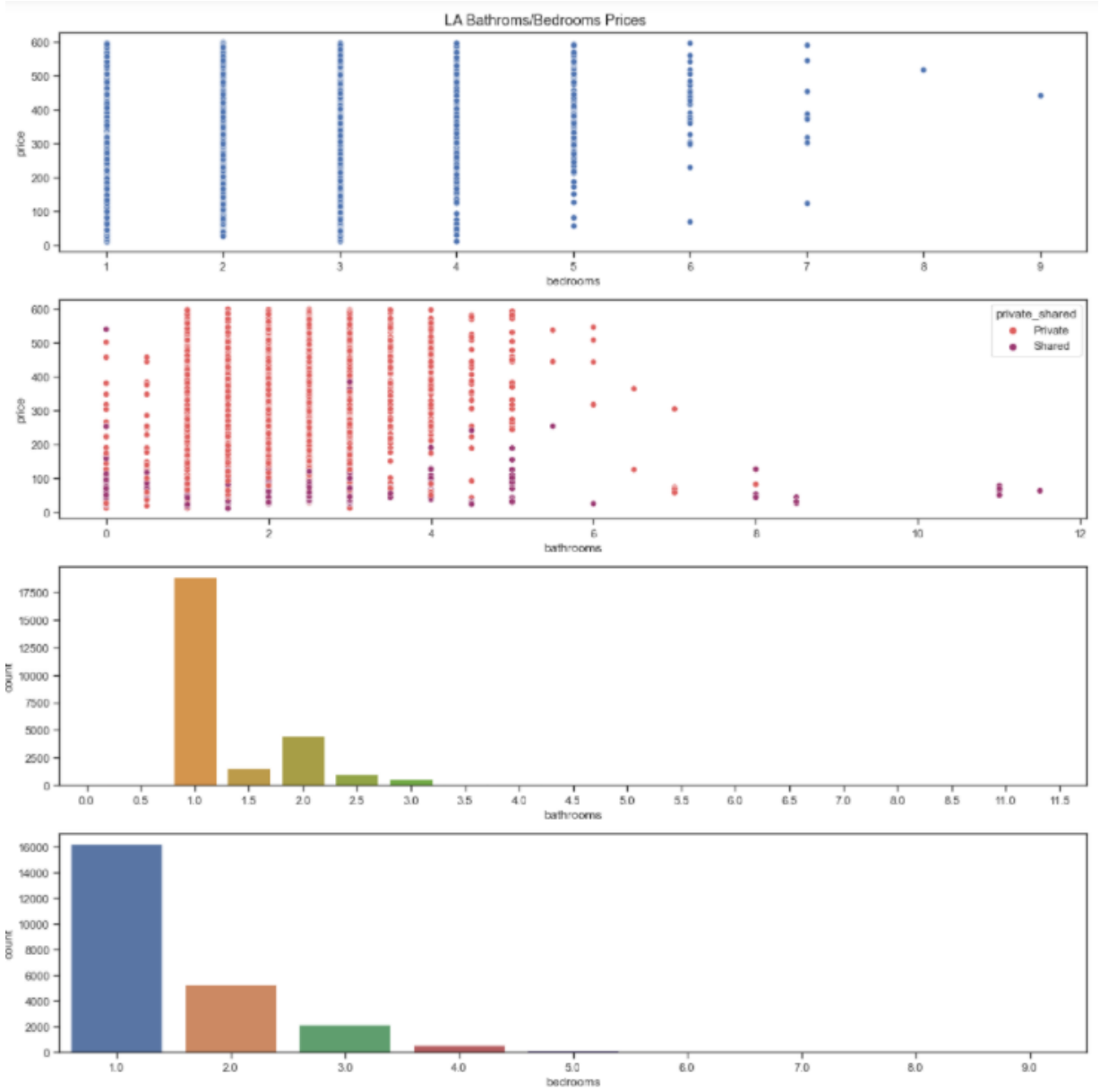


Figure 12: Scatter Plot/Bar Chart: Chicago Bathrooms/Bedrooms

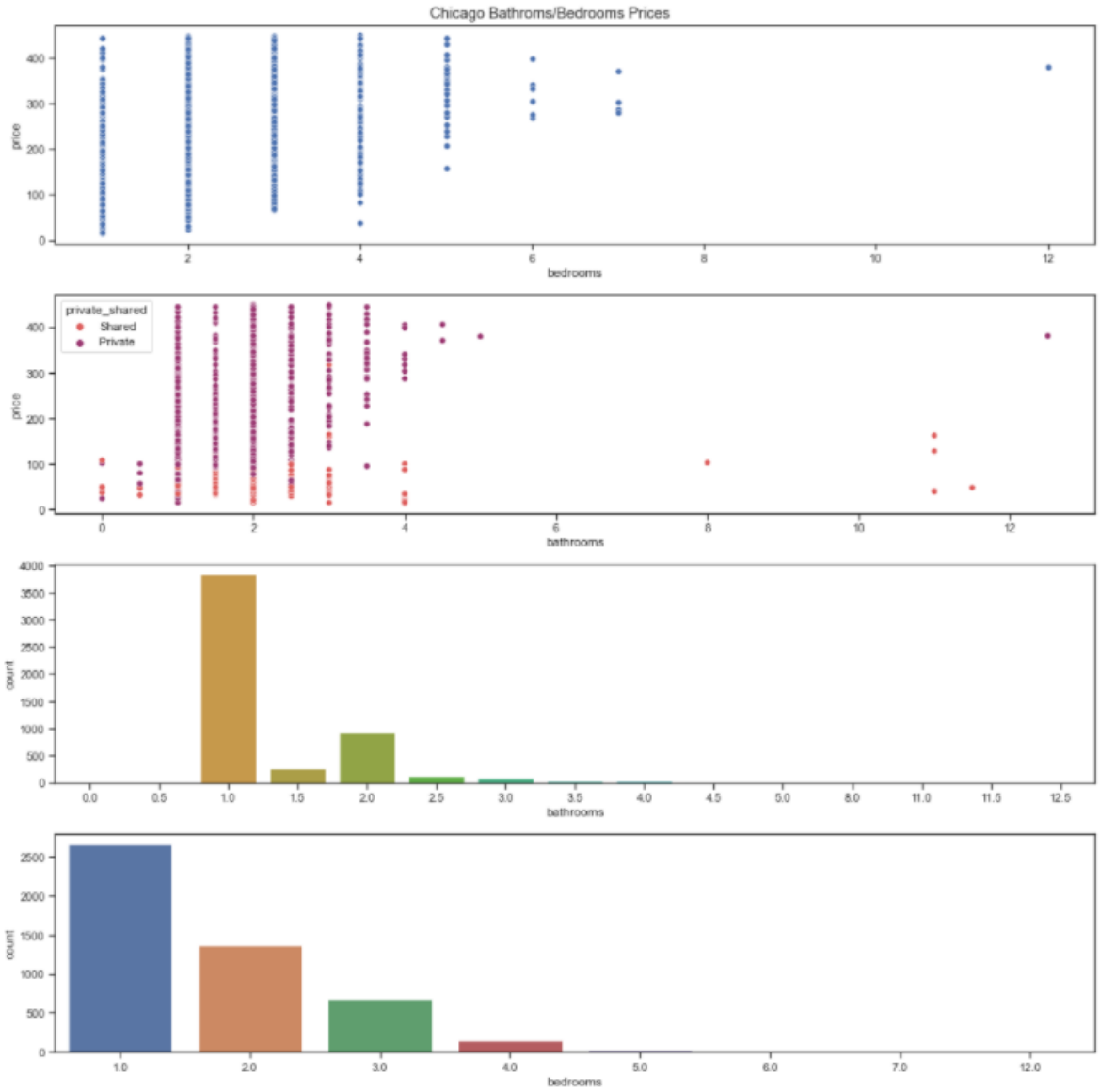


Figure 13: Scatter Plot/Bar Chart: Stockholm Bathrooms/Bedrooms

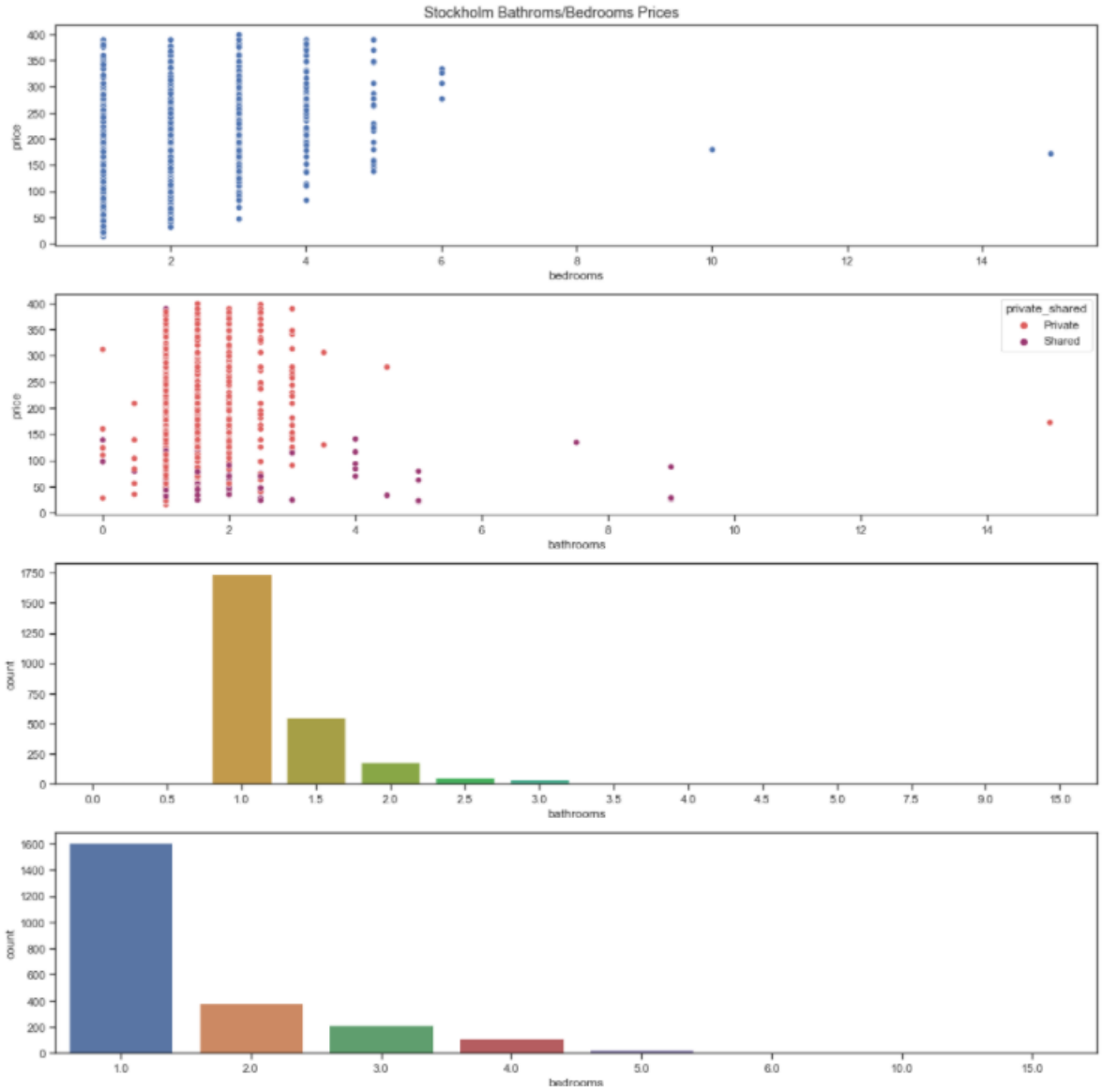


Figure 14: Scatter Plot/Bar Chart: Sydney Bathrooms/Bedrooms

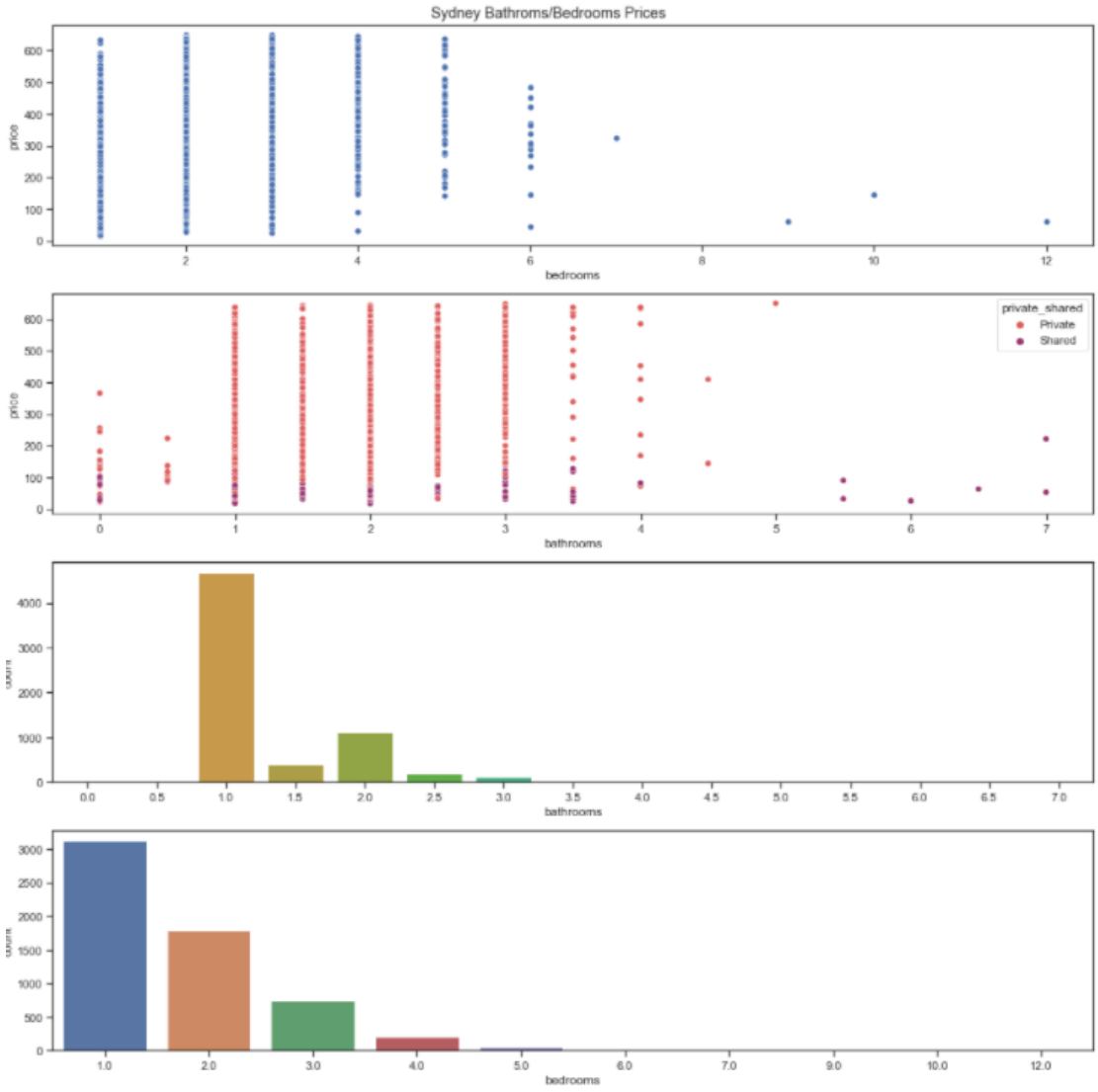


Figure 15 below shows the insignificance of the average review score variable on the price. A large majority of properties have between a 4/5 star score. There's no distinguishable trend based on location. There can be cheap 4/5 star properties and there can be expensive 4/5 star properties. The inverse can also be said around 0-star properties. The attitude of the host and the cleanliness of the rental property plays a large role in the review score, which is not measured in the given datasets.

Figure 15: Scatter Plot: Average Price by Average Review Score

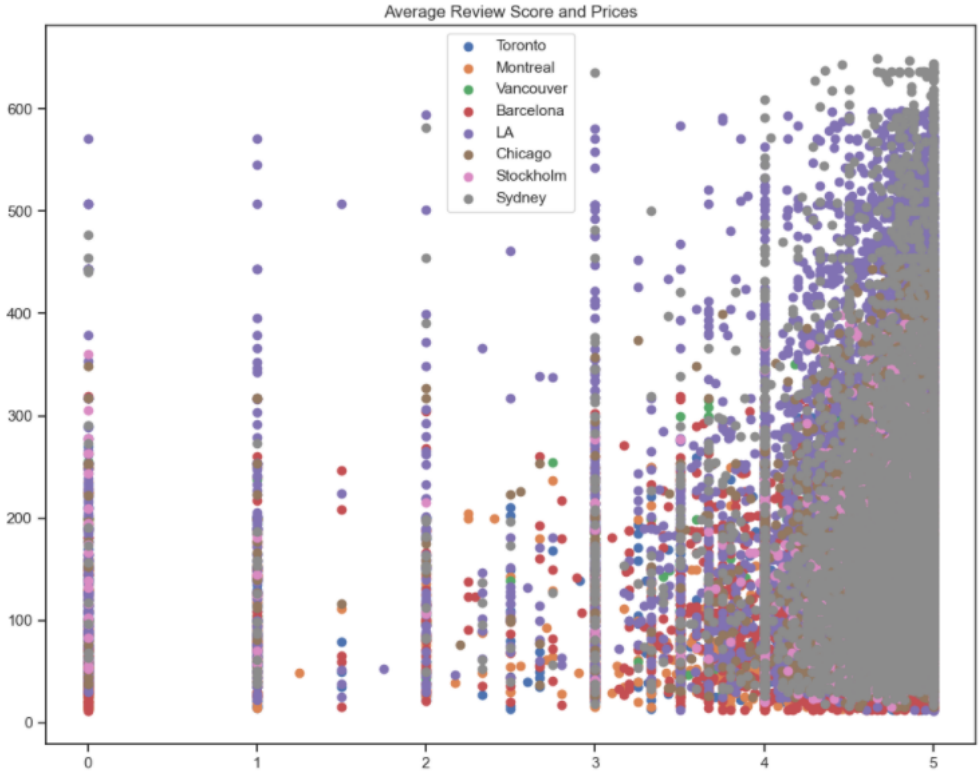
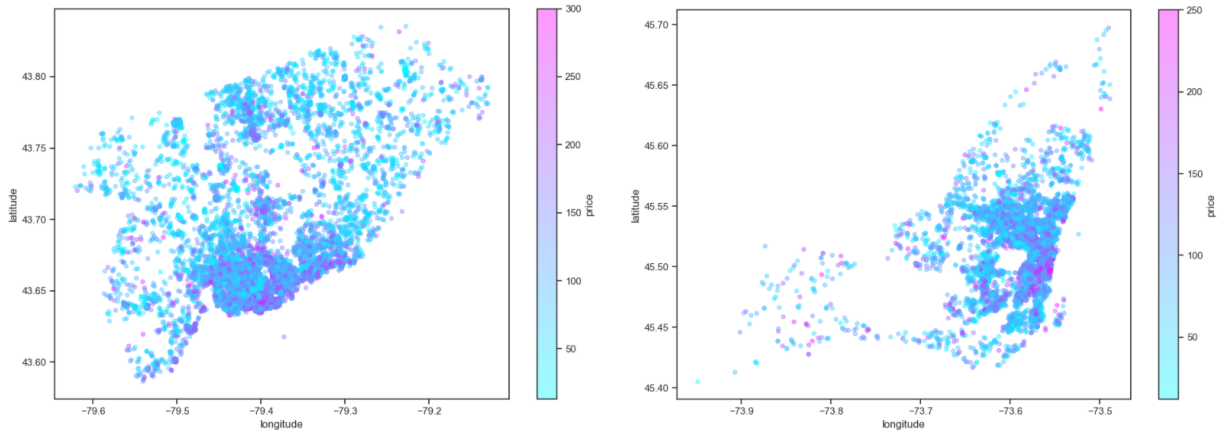
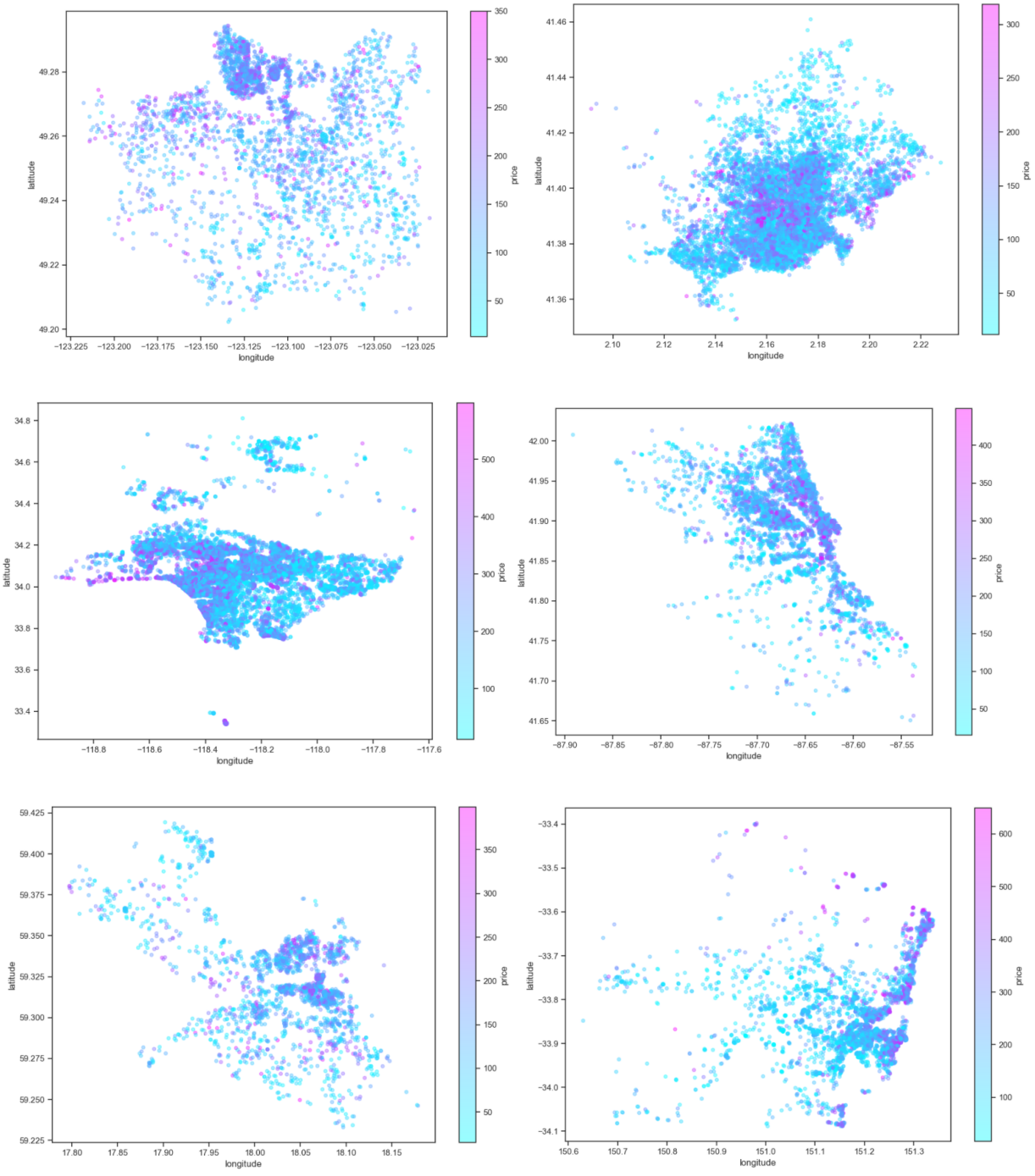


Figure 17 below shows the price variable based on location using latitude/longitude coordinates. With the exception of Barcelona & LA, all other cities have a distinct cluster where the majority of rental properties are located. This cluster is also where a majority of the higher-priced units are located. As the distance away from the cluster increases, the price decreases. Without looking at a map or any external knowledge of the city, we can conclude that this cluster is likely the downtown core/ heavy tourist area for each city.

Figure 17: Scatter Plot: Location by Price (Using Latitude/Longitude)

Order: Toronto, Montreal, Vancouver, Barcelona, LA, Chicago, Stockholm, Sydney

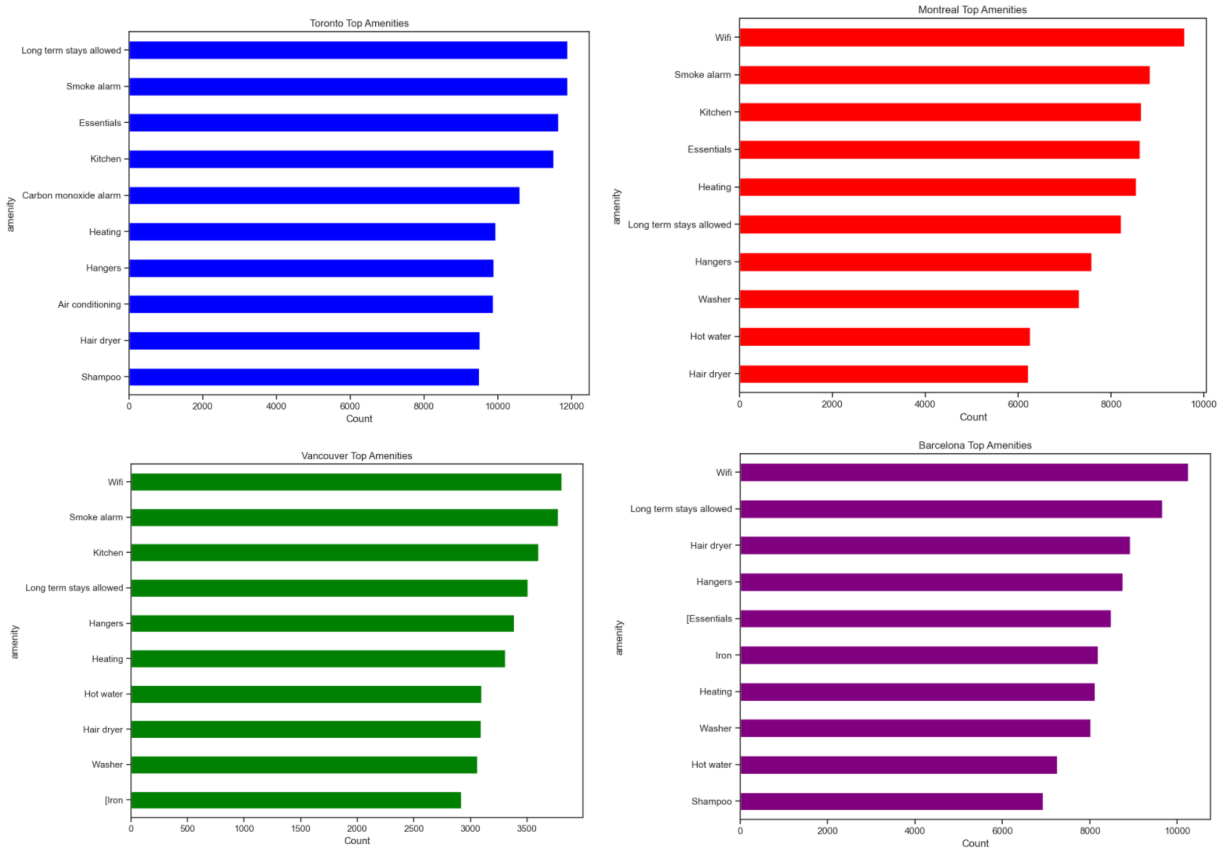


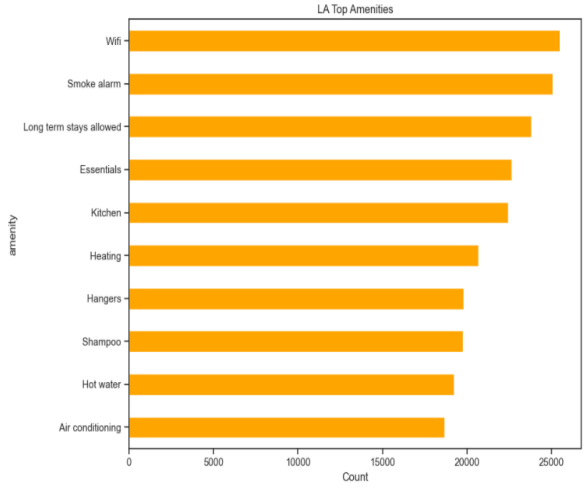

latitude
longitude
price
49.28
49.26
49.24
49.22
49.20
−123.225
−123.200
−123.175
−123.150
−123.125
−123.100
−123.075
−123.050
−123.025
350
300
250
200
150
100
50
41.46
41.44
41.42
41.40
41.38
41.36
2.10
2.12
2.14
2.16
2.18
2.20
2.22
300
250
200
150
100
50
34.8
34.6
34.4
34.2
34.0
33.8
33.6
33.4
−118.8
−118.6
−118.4
−118.2
−118.0
−117.8
−117.6
500
400
300
200
100
42.00
41.95
41.90
41.85
41.80
41.75
41.70
41.65
−87.90
−87.85
−87.80
−87.75
−87.70
−87.65
−87.60
−87.55
400
350
300
250
200
150
100
50
59.425
59.400
59.375
59.350
59.325
59.300
59.275
59.250
59.225
17.80
17.85
17.90
17.95
18.00
18.05
18.10
18.15
350
300
250
200
150
100
50
−33.4
−33.5
−33.6
−33.7
−33.8
−33.9
−34.0
−34.1
150.6
150.7
150.8
150.9
151.0
151.1
151.2
151.3
600
500
400
300
200
100

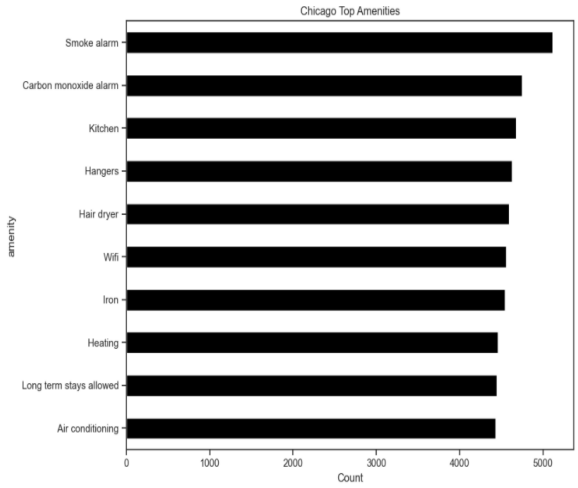
Stripping out the punctuation from the amenities array and making the count of unique values its own data frame, we can see the top ten amenities for each dataset in Figure 18. For all cities, the top ten amenities are very similar with basic amenities such as wi-fi, heating, shampoo, long-term stays all appearing near the top.

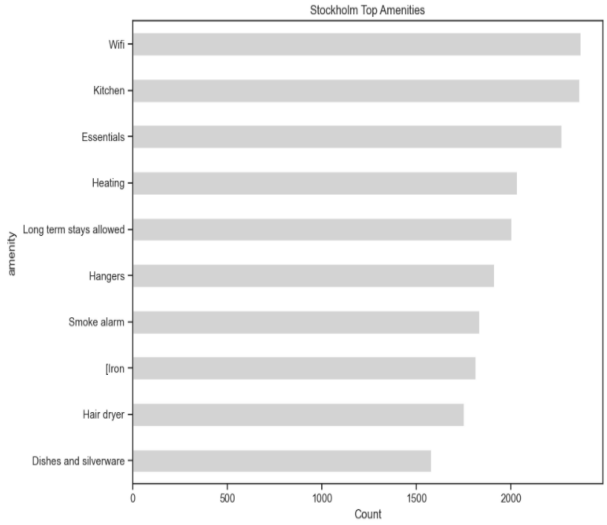
Figure 18: Bar Chart: Count of Amenities

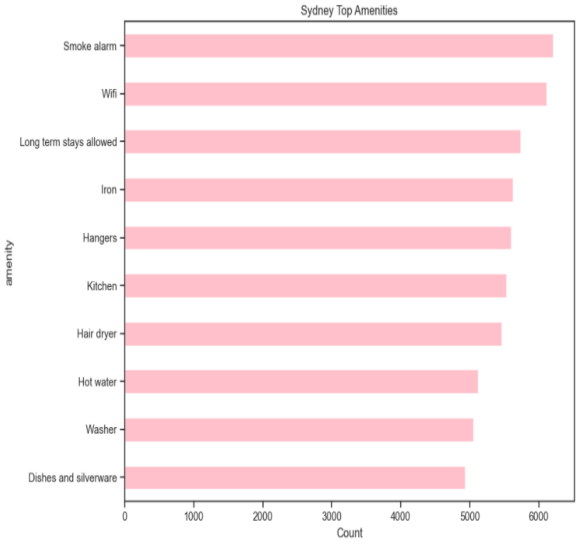
Order: Toronto, Montreal, Vancouver, Barcelona, LA, Chicago, Stockholm, Sydney




LA Top Amenities
Wifi
Smoke alarm
Long term stays allowed
Essentials
Kitchen
Heating
Hangers
Shampoo
Hot water
Air conditioning
amenity
0
5000
10000
15000
20000
25000
Count


Chicago Top Amenities
Smoke alarm
Carbon monoxide alarm
Kitchen
Hangers
Hair dryer
Wifi
Iron
Heating
Long term stays allowed
Air conditioning
amenity
0
1000
2000
3000
4000
5000
Count


Stockholm Top Amenities
Wifi
Kitchen
Essentials
Heating
Long term stays allowed
Hangers
Smoke alarm
[Iron
Hair dryer
Dishes and silverware
amenity
0
500
1000
1500
2000
Count


Sydney Top Amenities
Smoke alarm
Wifi
Long term stays allowed
Iron
Hangers
Kitchen
Hair dryer
Hot water
Washer
Dishes and silverware
amenity
0
1000
2000
3000
4000
5000
6000
Count

Interpretation of Results

Before Standardization

Using the tables below, with the exception of LA and Montreal, the R^2 and Adjusted R^2 are between 0.45-0.55. This means that ~50% of the variance of the price can be explained by the variance of the other independent variables. Montreal has a much lower R^2 of around 0.4, whereas LA has the highest R^2value at 0.6. When comparing the Ridge, Lasso & ElasticNet results with each other in Tables 1,2 & 3; the RMSE, MSE R^2 , and Adjusted R^2 are all very similar. The Lasso model in Table 2 yields marginally better results which could be partially due to the fact that there are several irrelevant variables and little correlation between them.

Table 1:Ridge (Before Standardization)

City	RMSE	MSE	R^2	Adjusted R^2
Toronto	43.4412	1887.1381	0.5085	0.5049
Montreal	40.9295	1675.2300	0.3946	0.3889
Vancouver	50.9795	2598.9142	0.4674	0.4547
Barcelona	47.5123	2257.4281	0.4773	0.4731
LA	76.3892	5835.3099	0.6065	0.6052
Chicago	64.5120	4161.8041	0.4755	0.4662
Stockholm	57.1378	3264.7340	0.5394	0.5227
Sydney	89.4478	8000.9198	0.5506	0.5443

Table 2: Lasso (Before Standardization)

City	RMSE	MSE	R^2	Adjusted R^2
Toronto	43.4051	1884.0072	0.5093	0.5057
Montreal	40.9271	1674.5693	0.3948	0.3892
Vancouver	51.0268	2603.7344	0.4664	0.4537
Barcelona	47.5086	2257.0758	0.4774	0.4732
LA	76.4325	5841.9347	0.6061	0.6048
Chicago	64.5231	4163.2391	0.4752	0.4660
Stockholm	57.1583	3267.0703	0.5391	0.5228
Sydney	89.4083	7993.8391	0.5510	0.5447

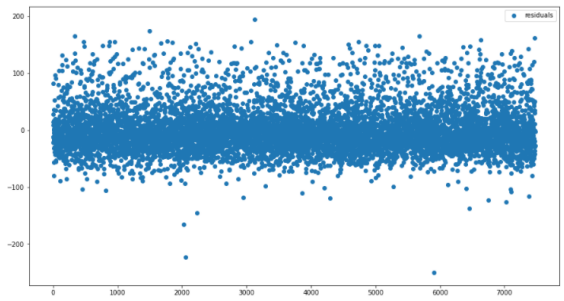
Table 3: Elastic Net (Before Standardization)

City	RMSE	MSE	R^2	Adjusted R^2
Toronto	43.4144	1884.8138	0.5091	0.5055
Montreal	40.9271	1675.0279	0.3947	0.3891
Vancouver	50.9262	2593.4827	0.4685	0.4558
Barcelona	47.5142	2257.6061	0.4773	0.4731
LA	76.4453	5843.8846	0.6060	0.6046
Chicago	64.5234	4163.2786	0.4752	0.4660
Stockholm	57.0864	3258.8580	0.5403	0.5235
Sydney	89.4174	7995.4723	0.5509	0.5446

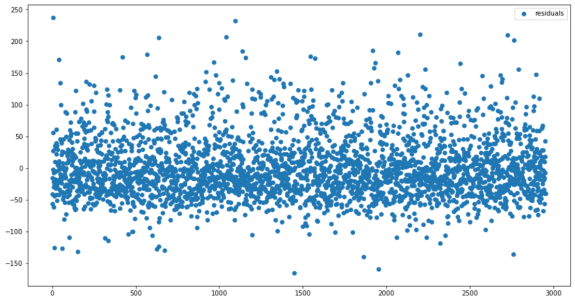
By examining the error residual graphs for each dataset in Figure 19, we can see that each dataset has residuals that follow a fairly random pattern. The residuals are approximately evenly split between positive and negative.

Figure 19: Scatterplot: Error Residuals

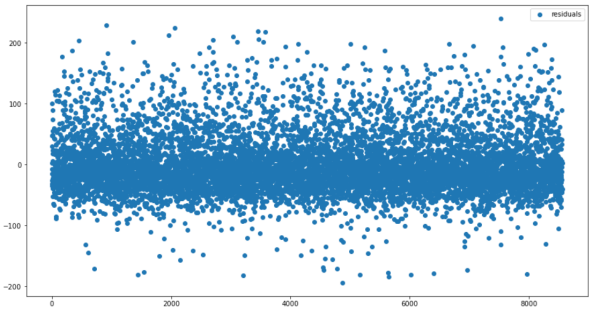
Montreal



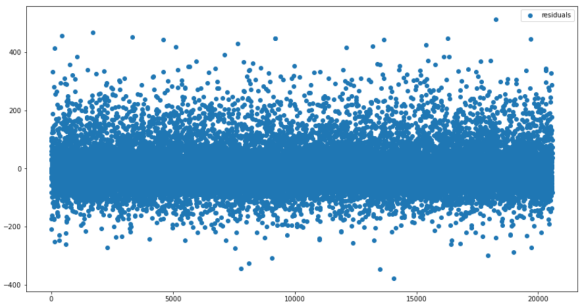
Vancouver



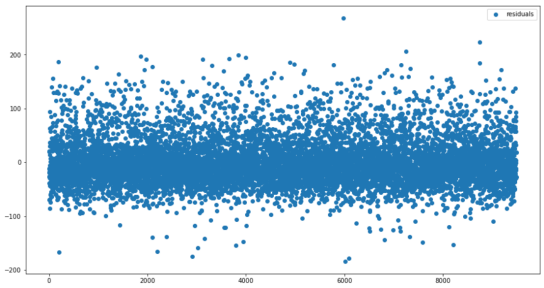
Barcelona



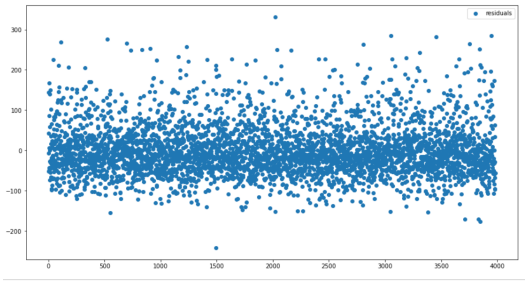
LA



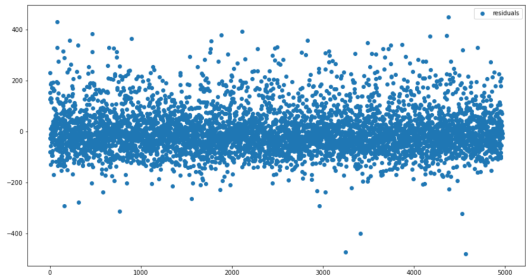
Toronto



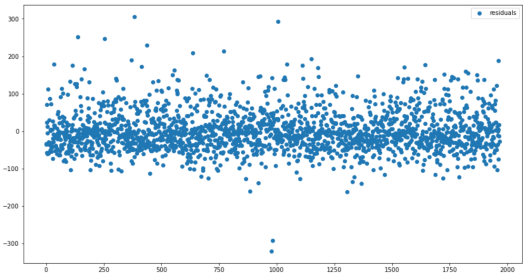
Chicago



Sydney



Stockholm



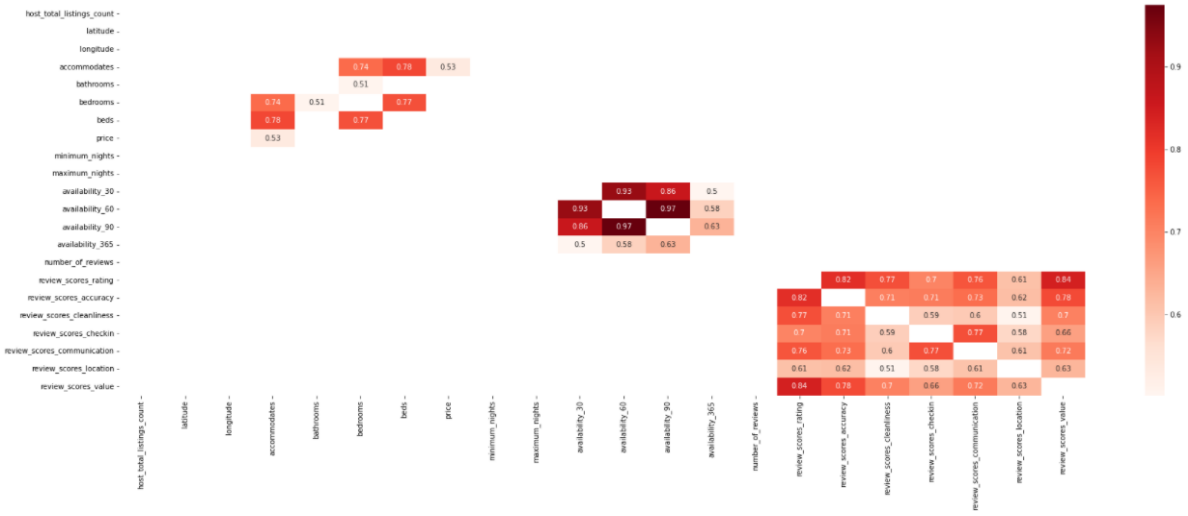
After Standardization

Using spearman's correlation matrix, I was able to determine the highly correlated variables for each dataset. Figure 20 below showcases variables that have a positive/negative correlation that's greater than 0.5. From the matrices below, we can see that specific groups of variables are highly correlated with each other. All of the variables relating to review score (number of reviews, value, communication, check-in, cleanliness) were all positively correlated as well as all variables related to availability (availability in 30,60,90,365). Furthermore, price, accommodates, beds and bedrooms were highly correlated with each other. This makes sense logically as the

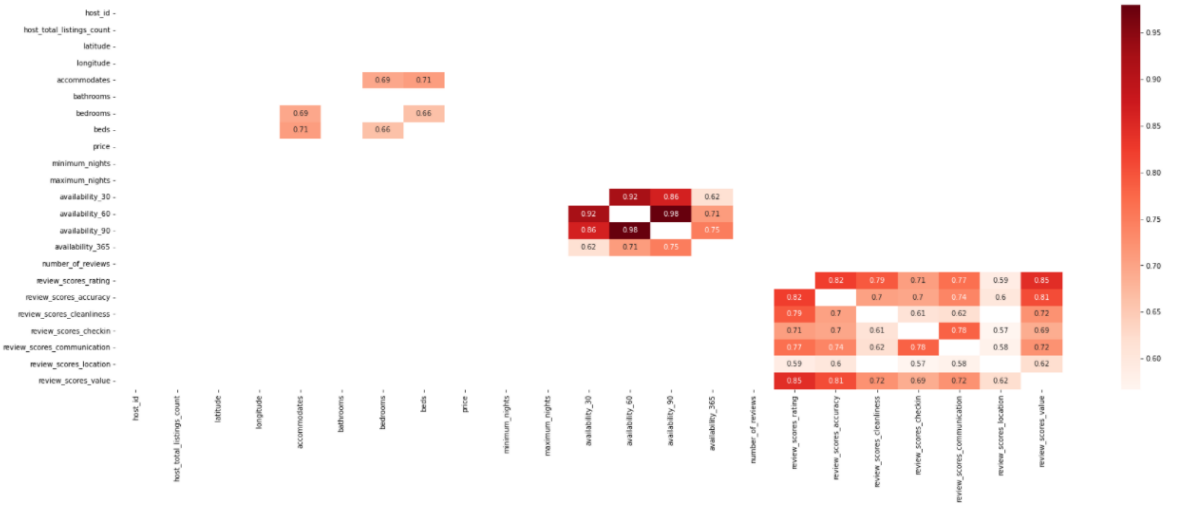
more people a rental property accommodates the more beds and space that will be needed, which would drive up the price.

Figure 20: Correlation Matrices

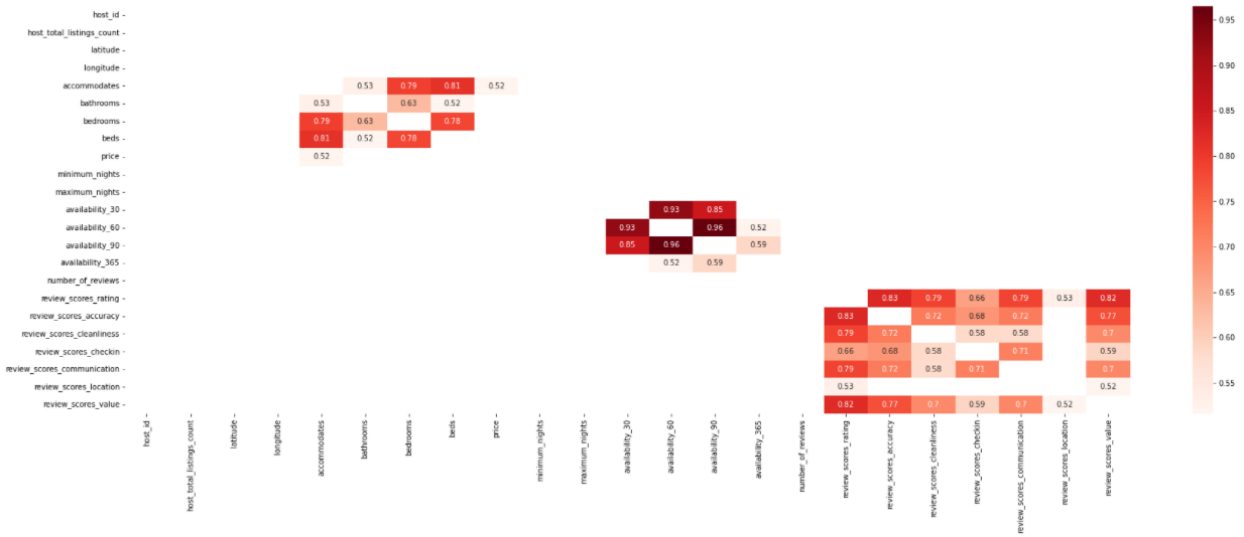
Toronto



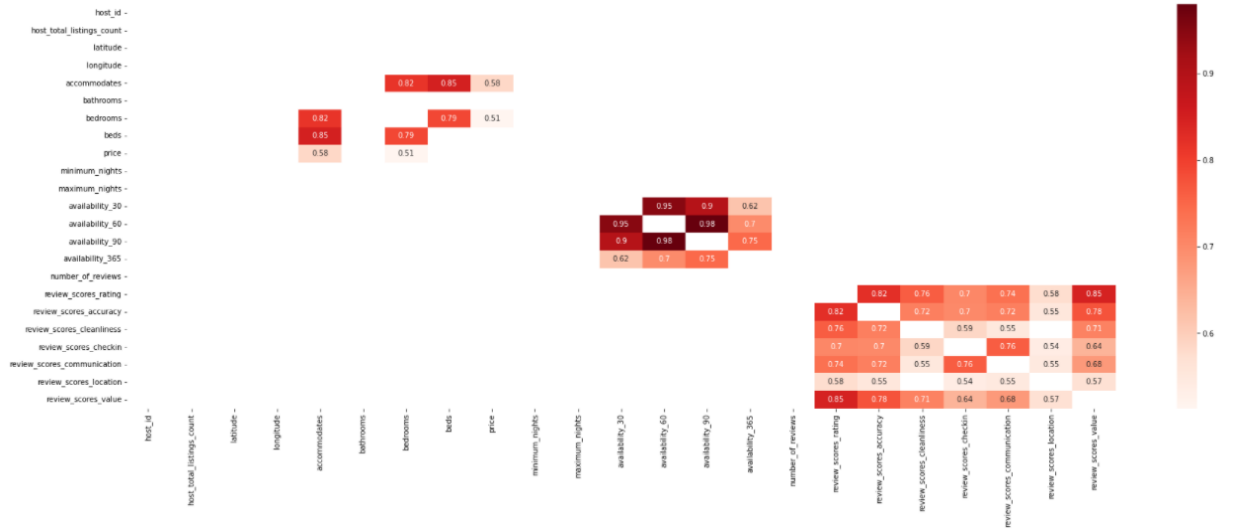
Montreal



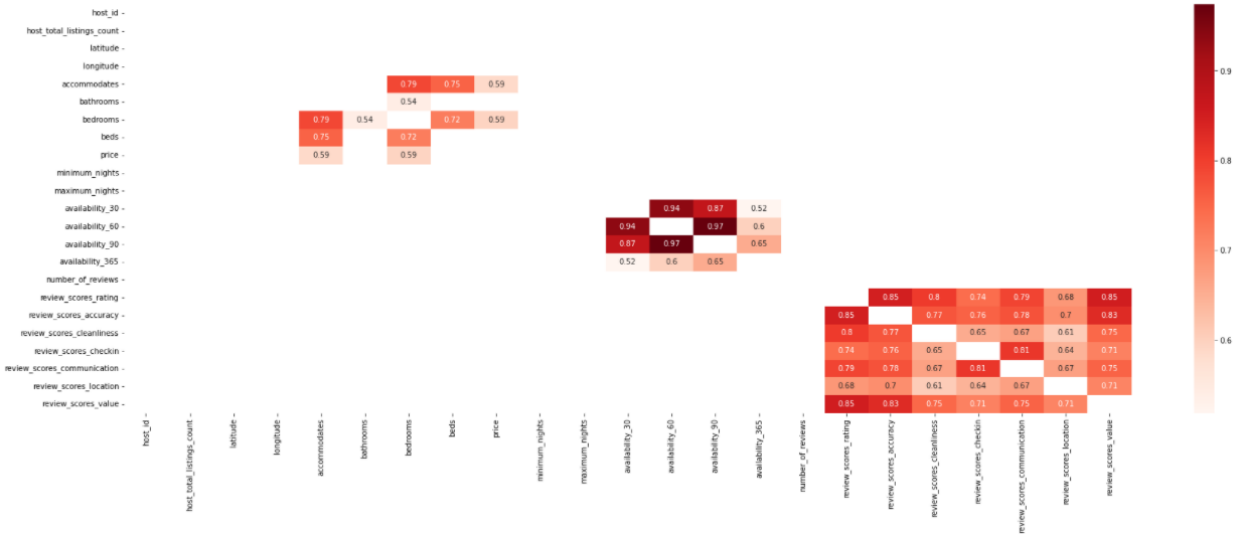
Vancouver



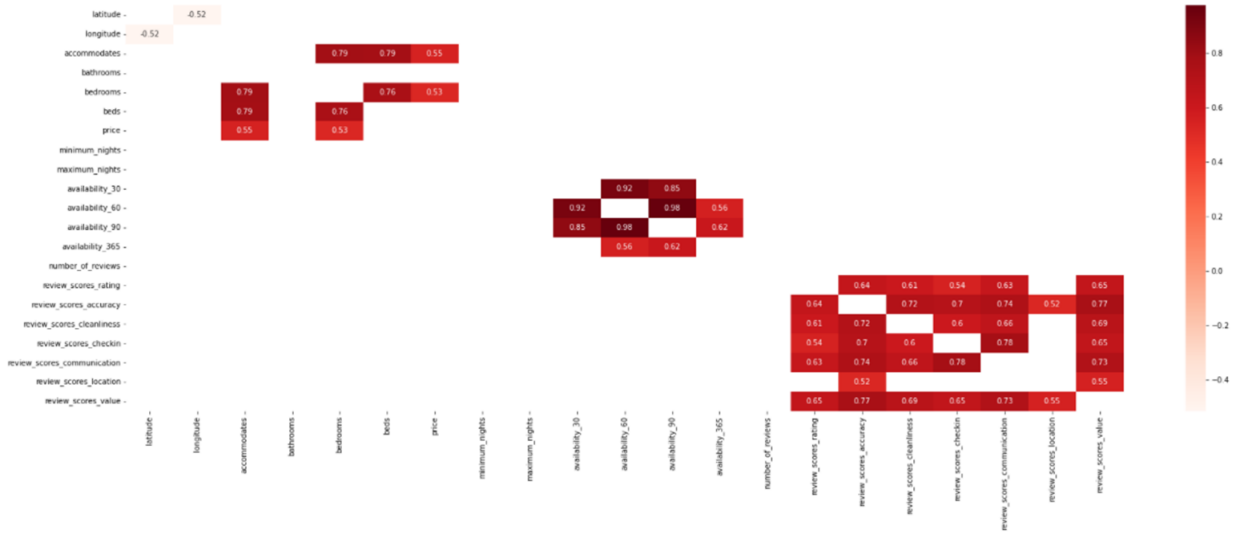
Barcelona



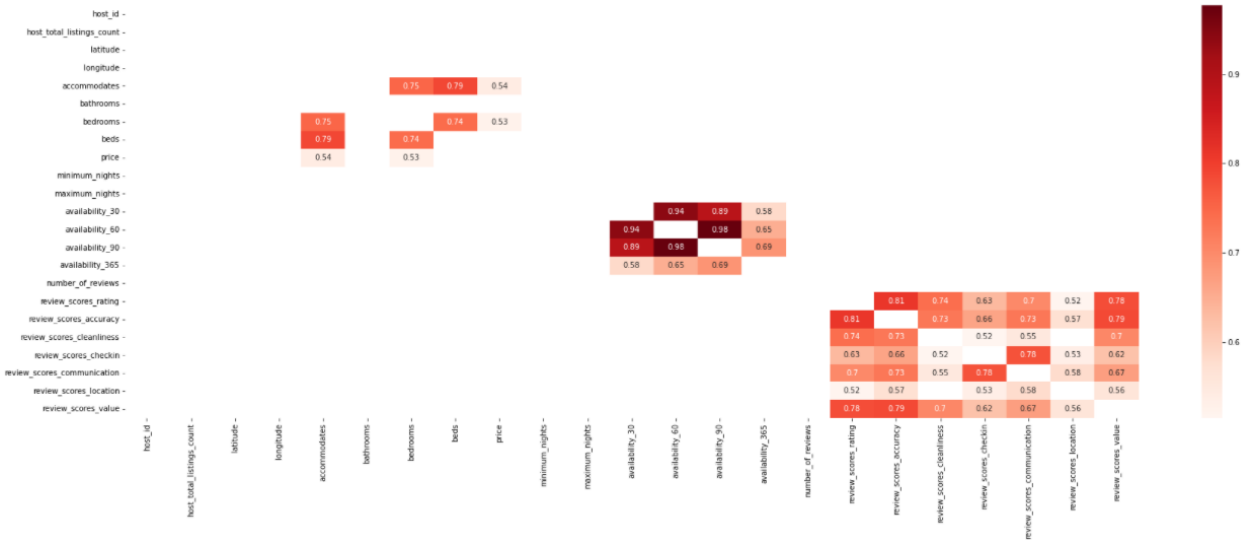
LA


host_id
host_total_listings_count
latitude
longitude
accommodates
bathrooms
bedrooms
beds
price
minimum_nights
maximum_nights
availability_30
availability_60
availability_90
availability_365
number_of_reviews
review_scores_rating
review_scores_accuracy
review_scores_cleanliness
review_scores_checkin
review_scores_communication
review_scores_location
review_scores_value

Chicago


latitude
longitude
accommodates
bathrooms
bedrooms
beds
price
minimum_nights
maximum_nights
availability_30
availability_60
availability_90
availability_365
number_of_reviews
review_scores_rating
review_scores_accuracy
review_scores_cleanliness
review_scores_checkin
review_scores_communication
review_scores_location
review_scores_value

Stockholm



Sydney

The following highly correlated variables were removed from each dataset:

beds, review_scores_value, availability_30 , availability_60, availability_90,

review_scores_location,review_scores_communication,review_scores_checkin,review_scores_cleanliness

After removing several of the highly correlated variables and running the regression models again, we can see that from the results in Tables 4, 5 & 6, half of the datasets outperformed the results compiled before standardization, meanwhile half of the datasets performed worse. Montreal, Vancouver, Barcelona and Sydney all performed marginally better, while Toronto, LA, Chicago and Stockholm performed marginally worse.

Table 4:Ridge (After Standardization)

City	RMSE	MSE	R^2	Adjusted R^2	Better or Worse?
Toronto	43.7700	1915.82	0.5018	0.4997	Worse
Montreal	40.6527	1652.6478	0.3969	0.3937	Better
Vancouver	48.3348	2336.2550	0.4944	0.4877	Better
Barcelona	50.2465	2524.7171	0.4687	0.4663	Better
LA	79.2138	6274.8273	0.5805	0.5797	Worse
Chicago	63.2408	3999.2926	0.4507	0.4453	Worse
Stockholm	58.3211	3401.3511	0.4835	0.4730	Worse
Sydney	94.8287	8992.4819	0.5431	0.5394	Better

Table 5: Lasso (After Standardization)

City	RMSE	MSE	R^2	Adjusted R^2	Better or Worse?
Toronto	43.7666	1915.5167	0.5019	0.4998	Worse
Montreal	40.6847	165.2506	0.3959	0.3927	Better
Vancouver	48.3172	234.547	0.4948	0.4881	Better

Barcelona	50.2411	2524.1771	0.4688	0.4665	Better
LA	79.2411	6279.1613	0.5802	0.5794	Worse
Chicago	63.1506	3988.0003	0.4523	0.4469	Worse
Stockholm	58.3361	3403.1011	0.4832	0.4727	Worse
Sydney	94.8008	8987.1837	0.5433	0.5397	Better

Table 6: Elastic Net (After Standardization)

City	RMSE	MSE	R^2	Adjusted R^2	Better or Worse?
Toronto	43.7700	1915.8162	0.5018	0.4997	Worse
Montreal	40.6610	1653.3181	0.3966	0.3935	Better
Vancouver	48.3254	2335.3505	0.4946	0.4878	Better
Barcelona	50.2454	2524.600	0.4687	0.4663	Better
LA	79.2432	6279.4836	0.5802	0.5794	Worse
Chicago	63.1161	3983.6367	0.4529	0.4475	Worse
Stockholm	58.3411	3403.6889	0.4831	0.4727	Worse
Sydney	94.8302	8992.7648	0.5431	0.5395	Better

Efficiency/Effectiveness/Stability of Linear Regression Algorithms

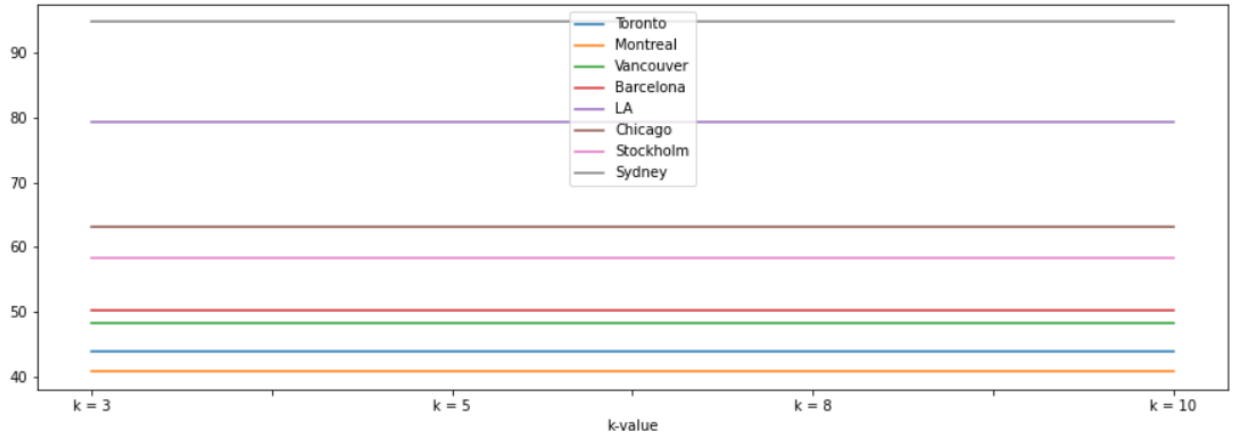
Given the Lasso regression model was slightly better than the other models, cross-validation was conducted for k-values of 3, 5, 8 & 10 exclusively for the Lasso model. From the table and line chart below, we can see that the RMSE is stable for each k-value. As seen in Table 7, the values

are essentially the same. The difference in value which is not noticed in Table 7 below doesn't occur until the fifth or sixth decimal point.

Table 7 : Cross-Validation of Lasso Regression Model

City	RMSE k = 3	RMSE k = 5	RMSE k = 8	RMSE k = 10
Toronto	43.7666	43.7666	43.7666	43.7666
Montreal	40.6847	40.6847	40.6847	40.6847
Vancouver	48.3172	48.3172	48.3172	48.3172
Barcelona	50.2411	50.2411	50.2411	50.2411
LA	79.2411	79.2411	79.2411	79.2411
Chicago	63.1506	63.1506	63.1506	63.1506
Stockholm	58.3361	58.3361	58.3361	58.3361
Sydney	94.8008	94.8008	94.8008	94.8008

Figure 21: Lasso RMSE Line Chart



Variable Weight

The XGBoost algorithm was run on the entire dataset for each city to better understand the weight of each variable. Examining the results in Table 8 below, we can see that the R^2 values are higher than those found in the regression models conducted earlier.

Table 8: XGB Boost (Using best-fit parameters)

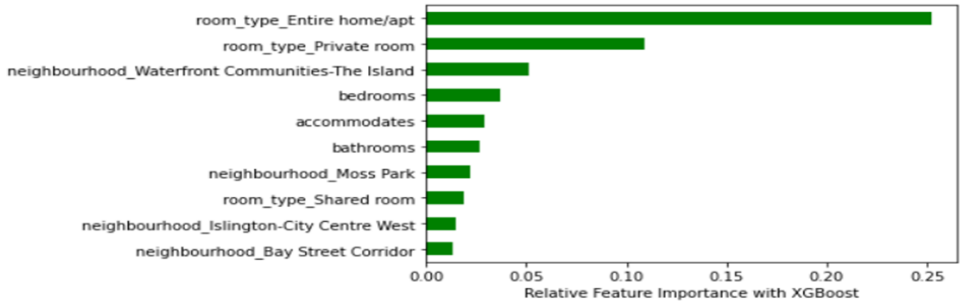
City	RMSE	MSE	R^2	Adjusted R^2
Toronto	38.38	1457.48	0.6203	0.6168
Montreal	35.47	1257.87	0.554	0.5488
Vancouver	45.58	2077.34	0.5281	0.5139
Barcelona	41.86	1751.9	0.6299	0.6261
LA	68.3	4665.18	0.6838	0.6825
Chicago	58.17	3383.85	0.5861	0.5769
Stockholm	48.95	2396.51	0.6276	0.6240
Sydney	79.19	6270.55	0.6527	0.6465

All datasets have very similar variables weighted heavily in terms of their effect on the price. From the bar charts in Figure 22, Entire home/apartment & bedrooms have the most weight. Entire home/apartment is the top-weighted variable for Toronto, Montreal, Barcelona, Chicago, whereas bedrooms are the top-weighted variable for Vancouver, LA, Stockholm, and Sydney. Some other common themes are that private rooms are in the top 2 for every dataset except LA

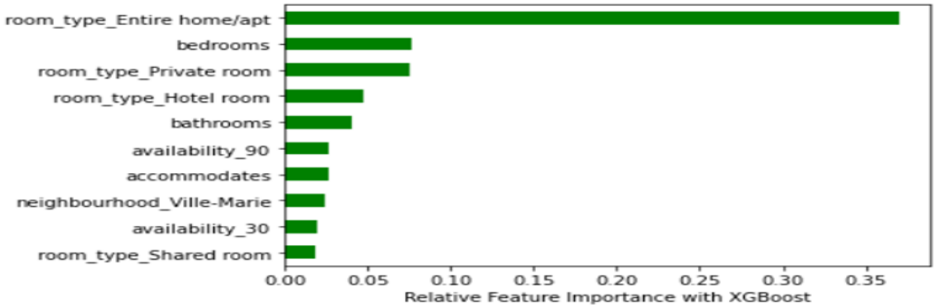
and are in the top 5 for every dataset. Toronto and Chicago both have a neighbourhood in the top 5. Beds, Accommodates, Shared room, and Hotel room all appear in the top 5 throughout each dataset. All of these variables are positively correlated with each other so it makes sense that they would all be weighted high against the price.

Figure 22: Bar Chart: Top 10 Amenities for each City

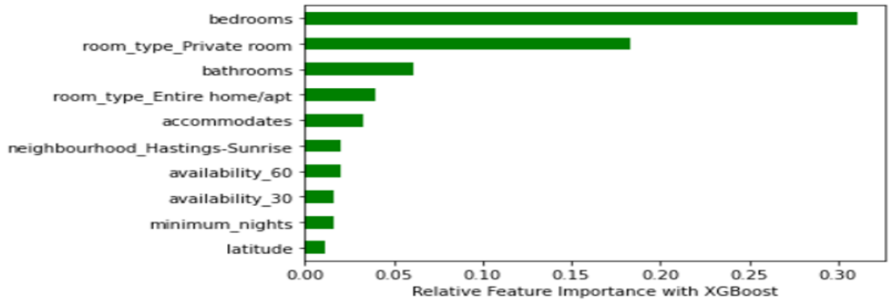
Toronto



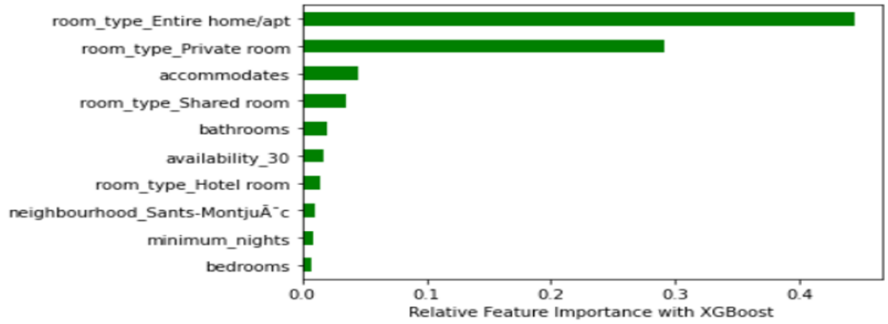
Montreal



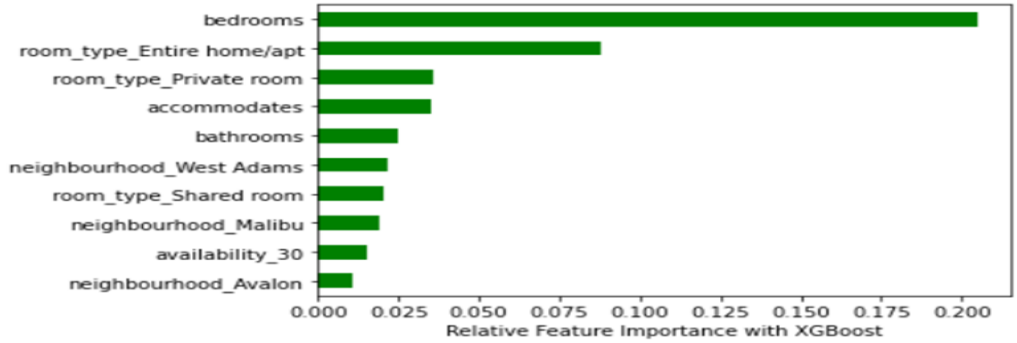
Vancouver



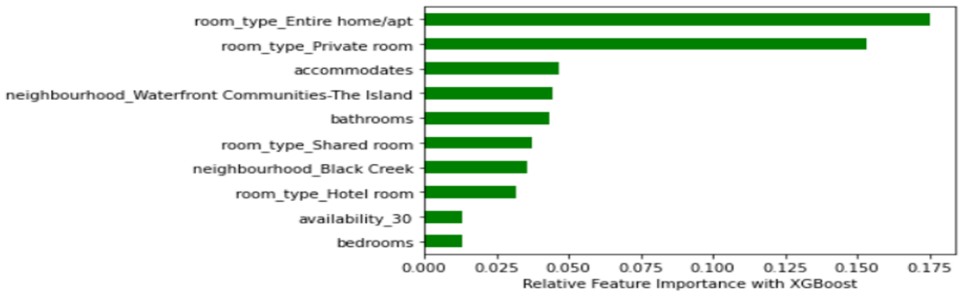
Barcelona



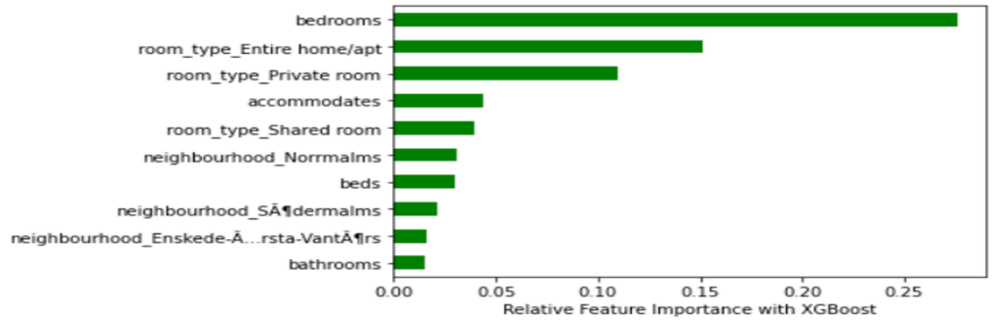
LA



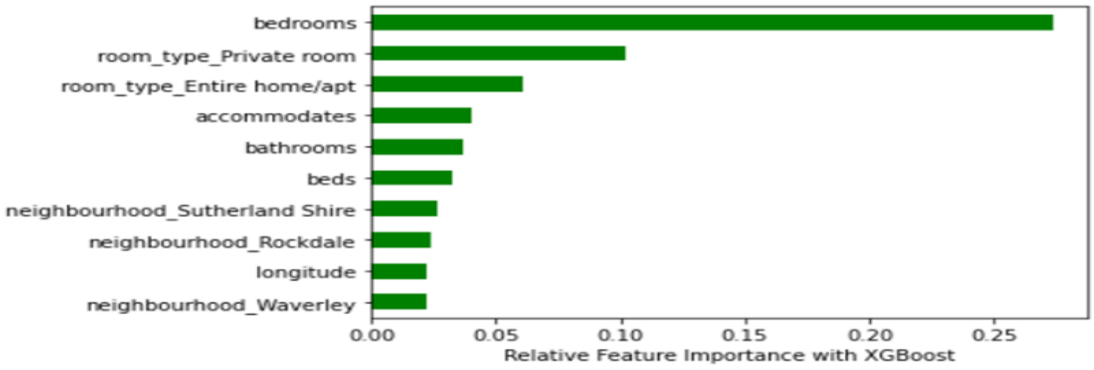
Chicago



Stockholm



Sydney



Using the results found in Figure 22, the XGBoost algorithm was run again for each dataset. This time, instead of running on all of the variables, variables not in the top ten in terms of weight on the price were removed. After running the XGBoost algorithm again with the top ten features, there were no noticeable gains in any of the RMSE/MSE/R^2/Adjusted R^2 values. The results in Table 9 were significantly worse in most cases with the R^2 value dropping by ~0.10.

Table 9: XGBoost Feature Engineering (Top 10 Features)

City	RMSE	MSE	R^2	Adjusted R^2	Better or Worse?
Toronto	45.13	2036.56	0.4885	0.4875	Worse
Montreal	39.7	1575.91	0.4494	0.448	Worse
Vancouver	46.99	2208.42	0.5114	0.5082	Worse
Barcelona	45.01	2025.49	0.538	0.537	Worse
LA	76.57	5862.97	0.6008	0.6005	Worse
Chicago	64.21	4161.93	0.496	0.4872	Worse
Stockholm	59.52	3542.79	0.4864	0.4815	Worse
Sydney	89.89	8080.85	0.5764	0.5848	Worse

Table 10: Efficiency/Effectiveness/Stability of XGBoost Algorithm

City	RMSE n=100	R^2 n=100	RMSE n=150	R^2 n=150	RMSE n= 200	R^2 n=200
Toronto	38.7	0.5959	38.48	0.6007	38.38	0.6203
Montreal	36.47	0.5238	36.11	0.533	35.47	0.554
Vancouver	44.78	0.5349	44.56	0.5395	44.49	0.541
Barcelona	38.62	0.6586	38.26	0.665	38.08	0.6682
LA	66.94	0.6965	66.41	0.7014	66.08	0.7043
Chicago	53.96	0.6327	52.8	0.6483	52.4	0.6537
Stockholm	49.09	0.6132	49.03	0.6142	49.13	0.6125
Sydney	79.9	0.6464	79.42	0.6506	79.19	0.6527

Table 10 shows the effect the number of estimators has on the XGBoost model for each city. For all datasets, as n increases, so does R^2. The RMSE on the other hand decreases. Inside the code, the number of estimators, learning rate, max depth, gamma, and column samplings by tree were all compared with each other to find the best parameters to fit each dataset. The values to be compared are as follows : n_estimators: 100, 150, 200, learning_rate: 0.01, 0.05, 0.1, max_depth: 3, 4, 5, 6, 7, colsample_bytree: 0.6, 0.7, 1, gamma: 0.0, 0.1, 0.2

In terms of efficiency, the runtime to test out the best parameters depends on the size of the dataset, but will normally take around two minutes to run, whereas running the XGBoost model itself will take milliseconds. To measure the XGBoost's model effectiveness, k-folds

cross-validation was run for k values of three, five, and eight. K-folds cross-validation is where all entries in the training dataset are used for both training and validation. Table 11 and Figure 23 show the results yielded for the train dataset whereas Table 12 and Figure 24 show the results for the test dataset. In both cases, the RMSE values are fairly stable; the RMSE values for the are higher for the test dataset compared to the training dataset

Table 11: Cross-Validation Train RMSE Values

City	Train RMSE k = 3	Train RMSE k = 5	Train RMSE k = 8
Toronto	25.80	23.31	25.87
Montreal	20.46	21.59	22.13
Vancouver	36.19	38.52	37.91
Barcelona	22.69	24.28	25.07
LA	51.84	53.03	53.60
Chicago	22.64	25.29	25.94
Stockholm	15.69	17.42	15.54
Sydney	44.71	47.16	48.18

Figure 23: Train RMSE Line Chart

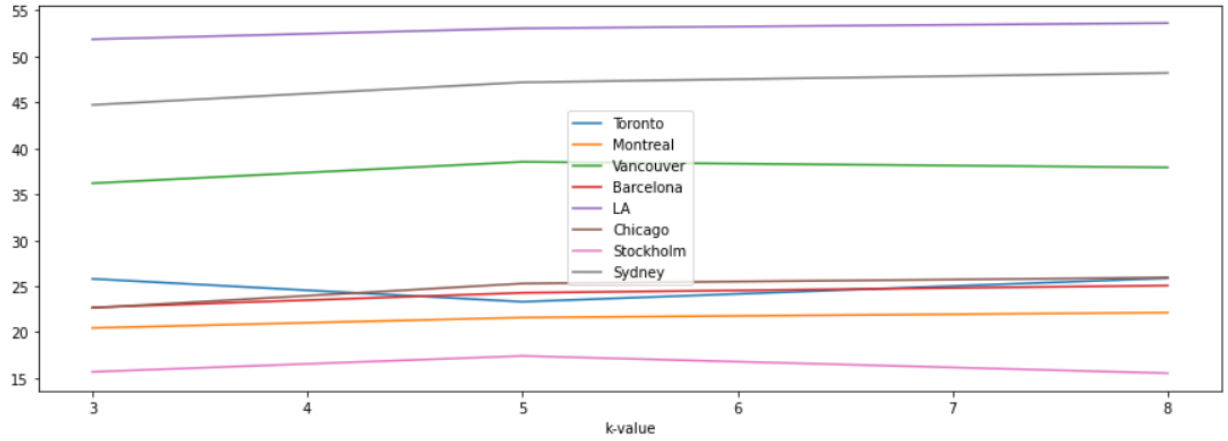
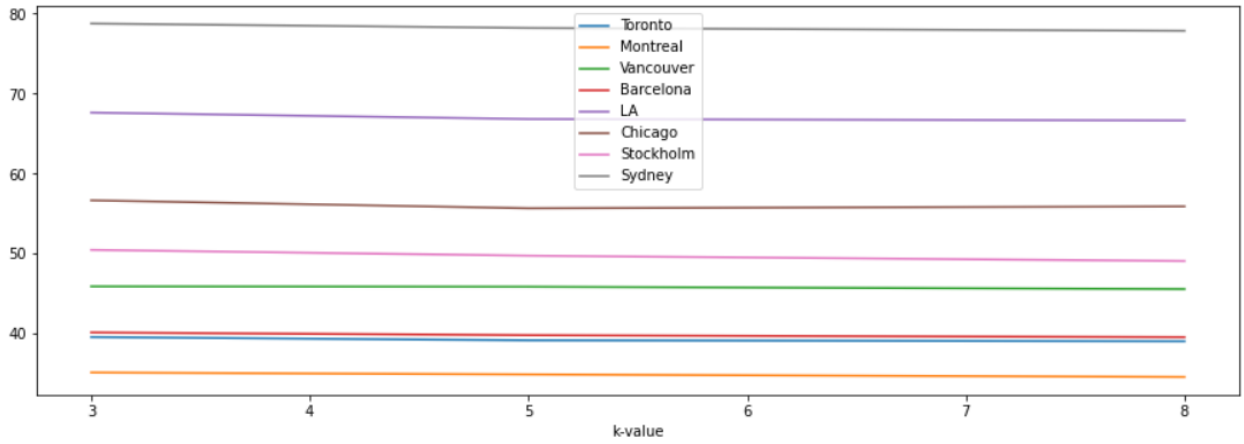


Table 12: Cross-Validation Test RMSE Values

City	Test RMSE k = 3	TestRMSE k = 5	Test RMSE k = 8
Toronto	39.59	39.19	39.07
Montreal	35.17	34.93	34.61
Vancouver	45.92	45.88	45.59
Barcelona	40.17	39.83	39.58
LA	67.63	66.82	66.65
Chicago	56.67	55.68	55.91
Stockholm	50.48	49.76	49.09
Sydney	78.75	78.20	77.84

Figure 24: Test RMSE Line Chart



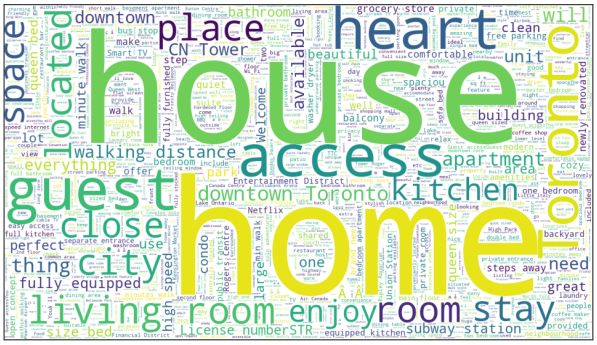
Text Mining

The word clouds in Figure 25 highlight the most frequent words in the description. The larger the word, the higher that word's frequency. Common words/phrases in all eight cities appear to be around the city name as well as the dwelling type. Home, House & Apartment all appeared very frequently, and in certain datasets such as Toronto, they appear the most. Given the description is a place where the lister would be describing what type of dwelling the home is in the description, this makes sense logically. Words like kitchen, shower, living room also appear often, allowing for the assumption that most Air Bnb's have basic amenities that a guest would need for their short-term stay. Another theme that can be derived from the word clouds is words related to a location such as the heart of the city, downtown, subway, walking distance & access. Using this data one can conclude that a large portion of Air Bnb’s are in a location in major tourist areas of the city or close to transportation allowing people to get to these tourist areas. Given the eight

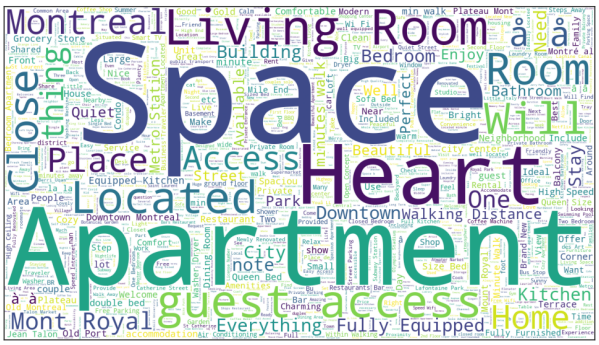
cities for this report are all very large cities, this also makes sense logically. They all have tourist areas and the people traveling to these cities using Air Bnb's are likely not from the area and are traveling to the city for leisure. Both Chicago and Sydney have the license number prominent in the description whereas other cities do not. This is likely due to Airbnb regulations imposed on those two cities that hosts must include this information in their listing.

Figure 25: Word Clouds for each City

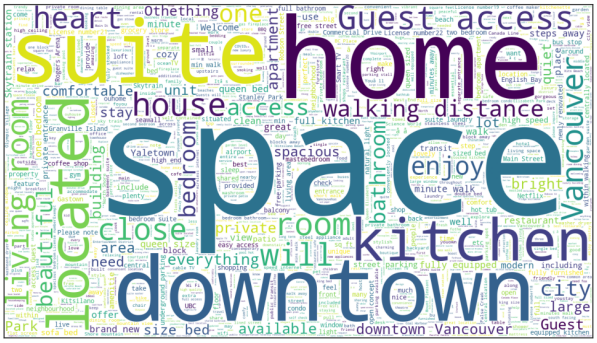
Toronto



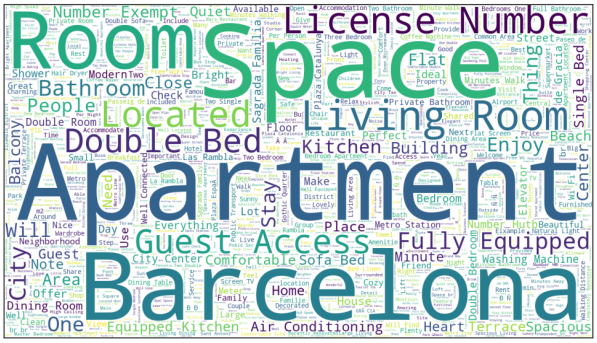
Montreal



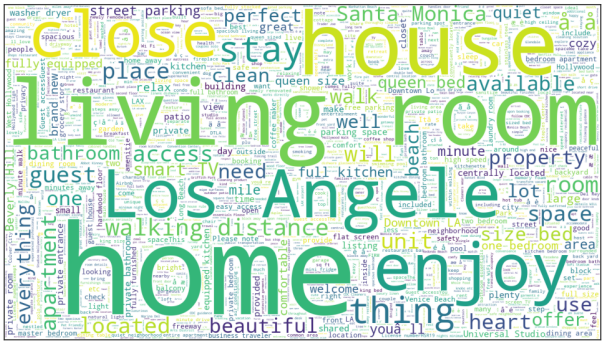
Vancouver



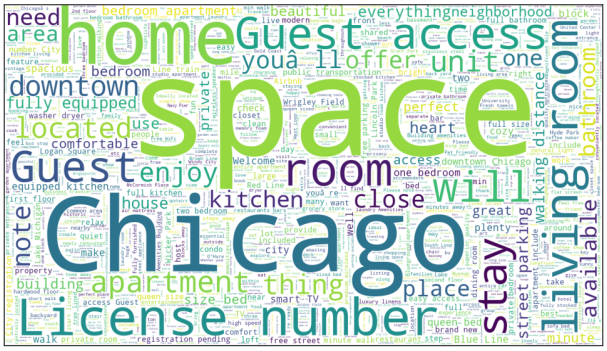
Barcelona



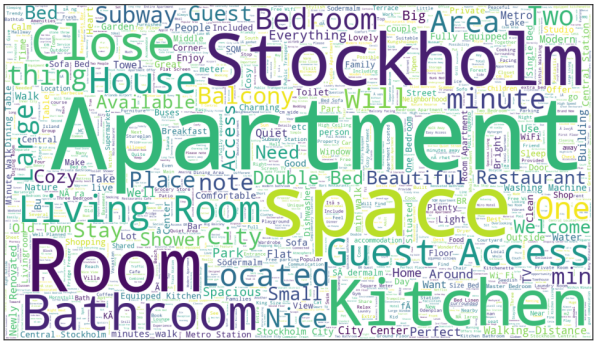
LA



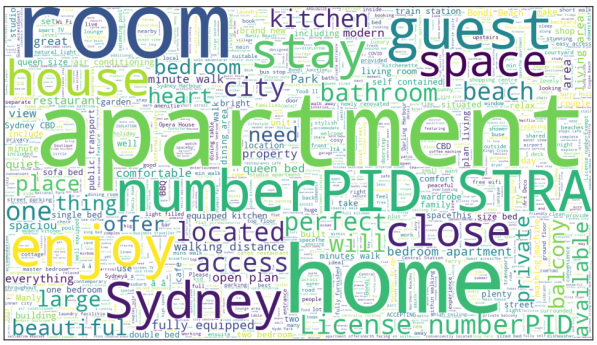
Chicago



Stockholm



Sydney



Note: All word clouds images can be found in the word cloud folder in my Github repository

Intertopic Distance Maps

The intertopic distance maps in Figure 26 group the most prominent words into 3 topics. The table below highlights the breakdown of the percentage of the total words in each topic.

Budget: Contains words such as room, bed, apartment. The focus is on the basic amenities most homes have.

Location: Contains words such as walking distance, subway, bus, downtown. Focus around travel and being close to the tourist-heavy areas.

Luxury: Contains words such as modern, unique, premium. The focus is around high-end features that most homes don't have.

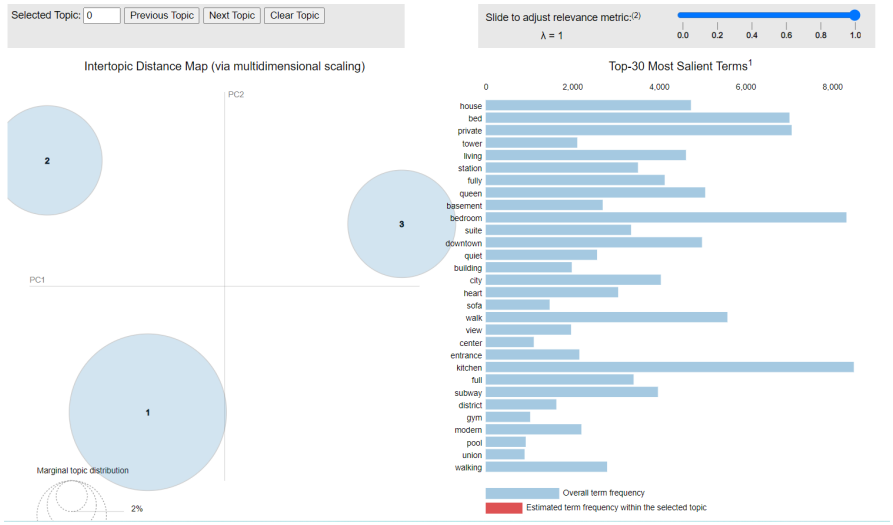
Per Table 13, Canadian cities tend to have a higher percentage of homes in the luxury bin compared to other worldwide cities. For the non-Canadian cities, the Budget and Location groups cover almost the dataset in its entirety with the exception of LA. When exploring the latitude/longitude coordinates in the EDA stage, LA didn't have a specific cluster where properties tended to cluster so it makes sense why the Location percentage isn't as high. From the intertopic distance maps, we can infer that the majority of Airbnb renters are looking for a budget property in an area near the heart of the city center.

Table 13: Topic Breakdown from LDA Analysis

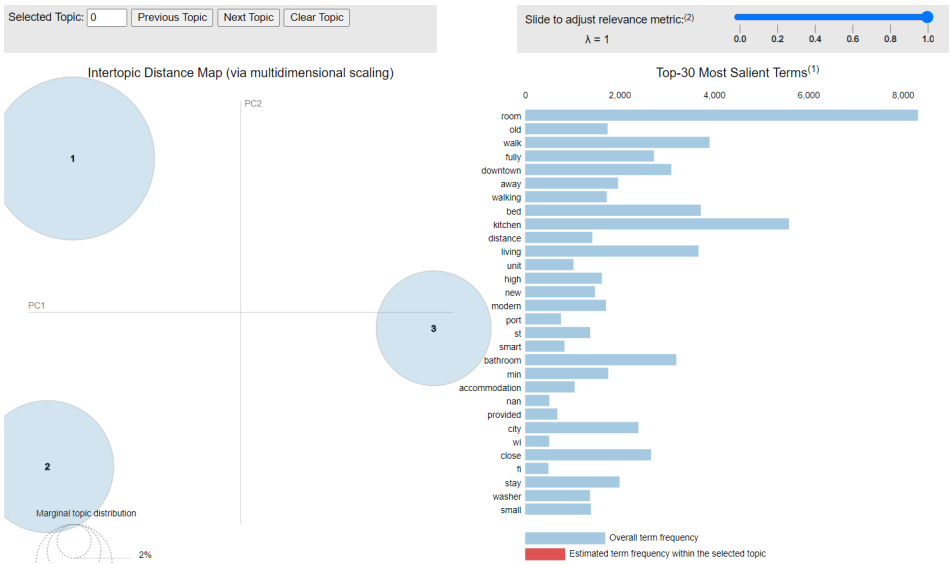
City	Budget	Location	Luxury
Toronto	51%	25%	24%
Montreal	47%	30%	23%
Vancouver	47%	33%	20%
Barcelona	45%	41%	14%
LA	68%	18%	14%
Chicago	46%	40%	14%
Stockholm	50%	41%	9%
Sydney	48%	45%	7%

Figure 26: Intertopic Distance Maps

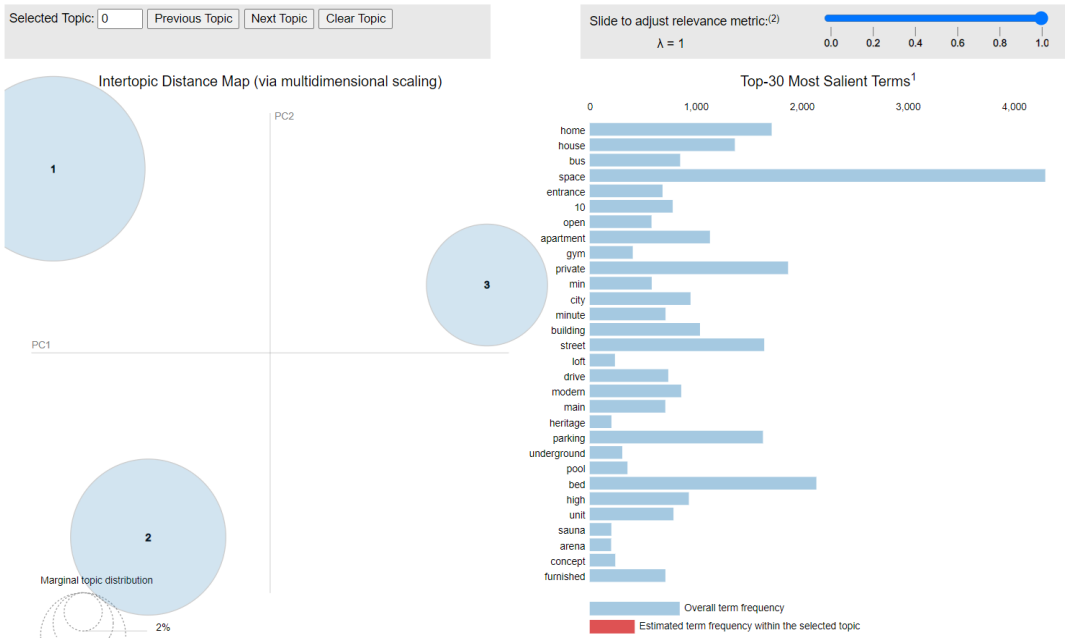
Toronto



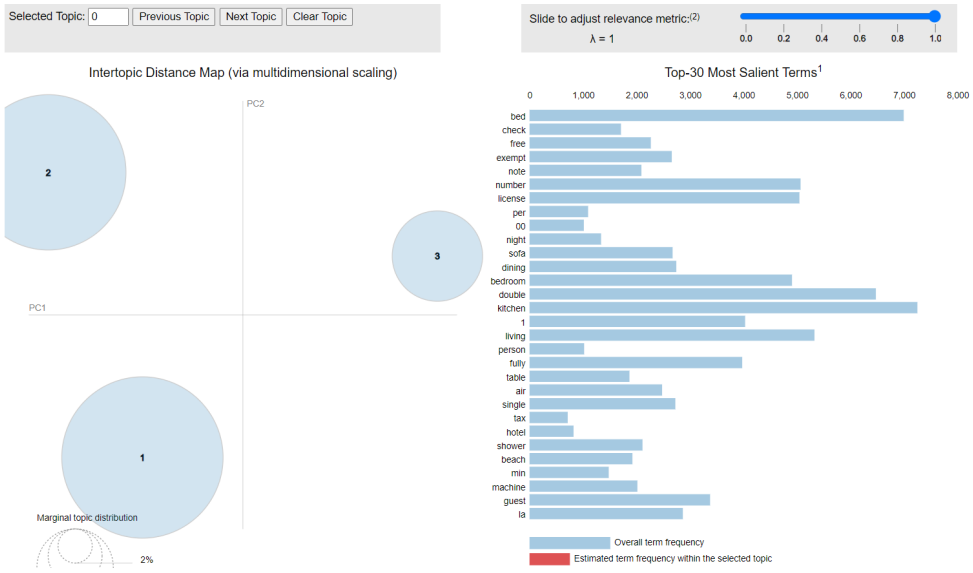
Montreal



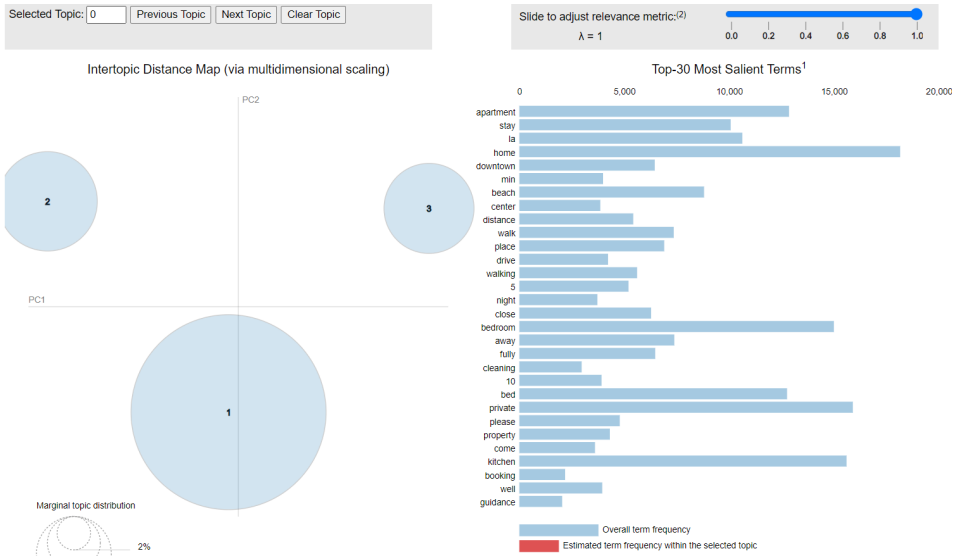
Vancouver



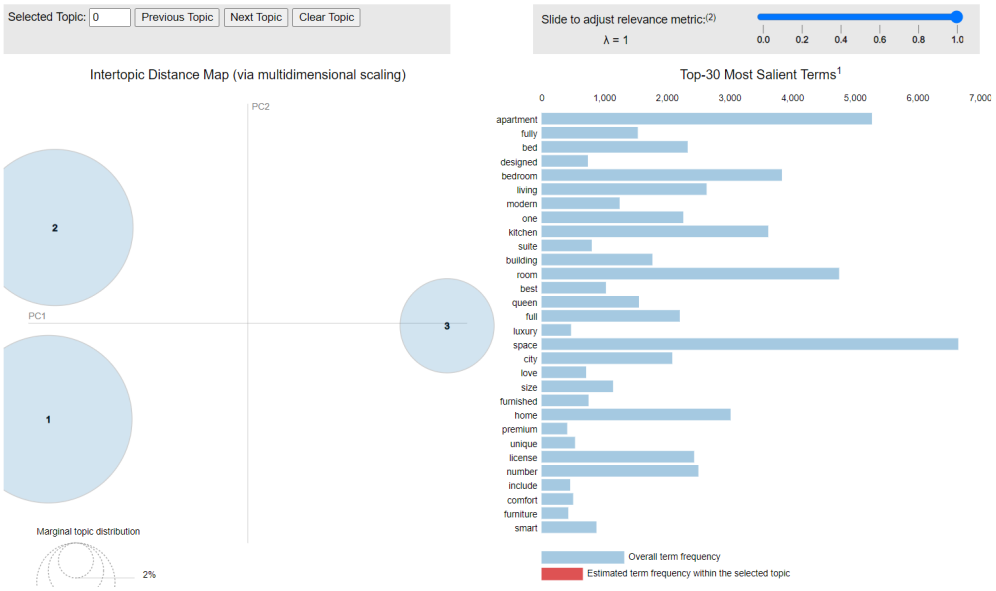
Barcelona



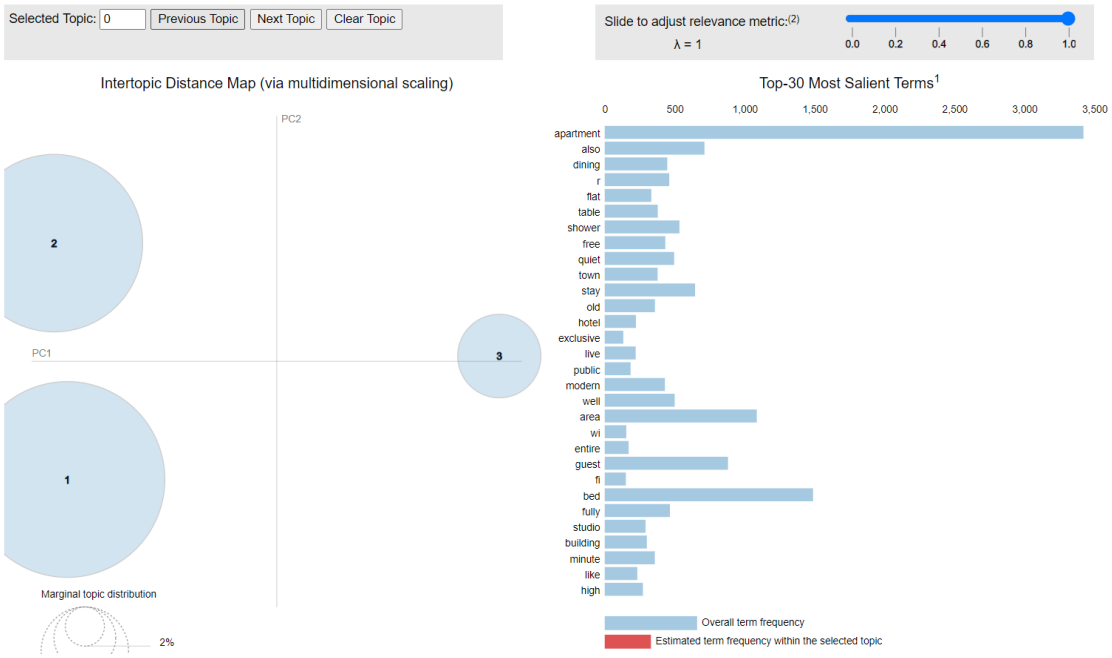
LA


Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
2
3
1
Marginal topic distribution
2%
Top-30 Most Salient Terms1
0 5,000 10,000 15,000 20,000
apartment
stay
la
home
downtown
min
beach
center
distance
walk
place
drive
walking
5
night
close
bedroom
away
fully
cleaning
10
bed
private
please
property
come
kitchen
booking
well
guidance
Overall term frequency
Estimated term frequency within the selected topic

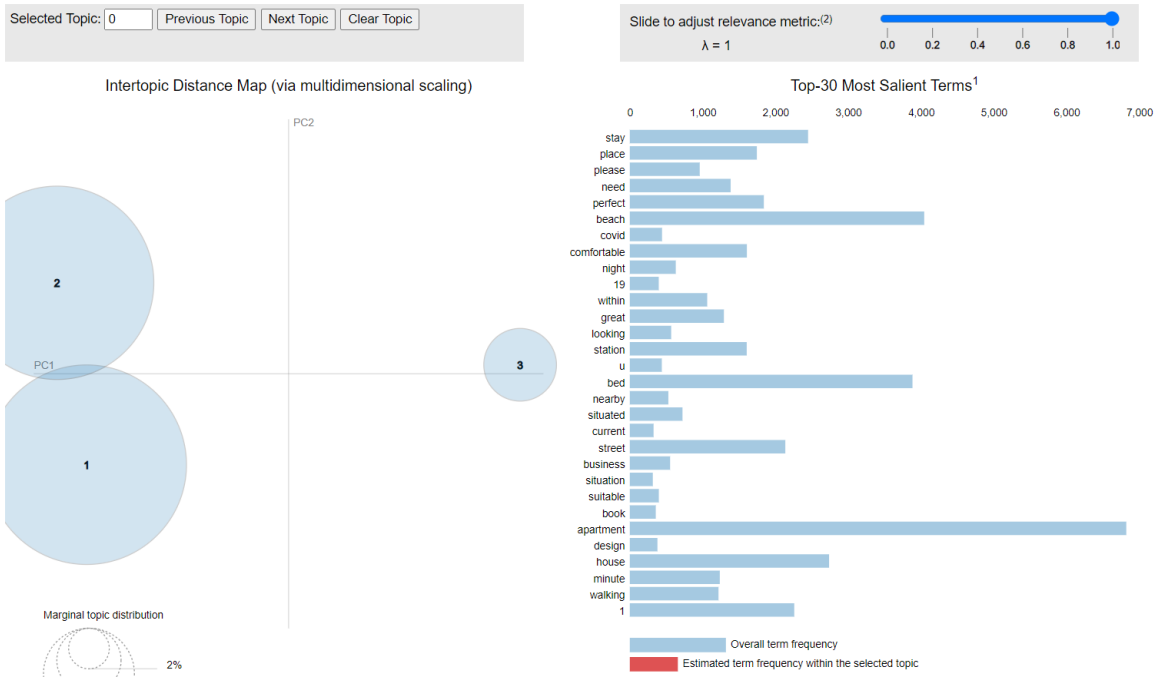
Chicago


Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
2
3
1
Marginal topic distribution
2%
Top-30 Most Salient Terms1
0 1,000 2,000 3,000 4,000 5,000 6,000 7,000
apartment
fully
bed
designed
bedroom
living
modern
one
kitchen
suite
building
room
best
queen
full
luxury
space
city
love
size
furnished
home
premium
unique
license
number
include
comfort
furniture
smart
Overall term frequency
Estimated term frequency within the selected topic

Stockholm



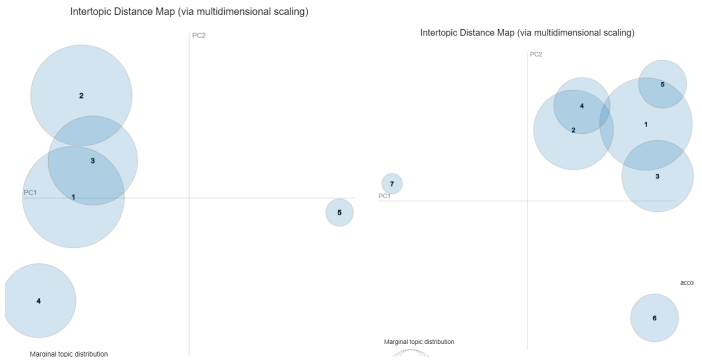
Sydney



Efficiency/Effectiveness/Stability of LDA Algorithm

The LDA algorithm allows for a different number of topics and passes to be completed. Three topics were ultimately chosen as it was the largest number of distinct topics with no overlap. Examples of five and seven topics can be seen below in Figure 27. As the number of passes increases, so does the runtime, with each pass taking approximately one second to complete. Ten passes were used as a baseline as the results seemed to be very similar if the passes were increased after ten.

Figure 27: Different k-value Intetopic Distance Maps



Shortcomings/ Next Steps

One of the shortcomings in regards to data related to specific availability dates. All datasets contain raw data around the rental property itself (beds, bathrooms, etc) as well as the host (id, acceptance rate, etc), but there's no data around availability other than how many days the home is available in 30,60,90 & 365-day intervals. If data around the specific days the rental property was used were available, seasonality could be measured against the price variable given it's a

known fact in the travel industry that there are peak and low points of the year. Having more historical data would also benefit with model accuracy as data, especially given the effect the COVID-19 pandemic has had on the travel industry the last two years. Due to time constraints, only eight cities could be used for this project, however, Insider Airbnb offers data on several other cities such as San Francisco, Athens, and London to name a few. The results and models could be easily extended to more cities to provide deeper comparisons between more worldwide cities. Additionally, in the exploratory data analysis stage, I was able to extract the unique values from the amenities array into their own data frame. If there was more time, the logical next steps would be to add these values back into the full data frame and measure their impact on the price with the regression models used in previous steps. Another next step would be to create a dashboard in a program such as Tableau or Power BI to give users more freedom to play with the raw data and make their own findings. Another aspect that could be extended would be running different models on the dataset such as a random forest to see if there are any improvements there. Standardization didn't yield any improvements for Toronto, LA, Chicago and Stockholm, so I'd be curious to see if using another model would improve the results there. Feature Engineering on the XGBoost model also did not yield any noticeable improvements as well; it would be worthwhile to extend this machine learning technique to the top fifteen or top twenty as a comparison to see if the results improve or not.

Conclusion

After a thorough analysis, we can conclude that trends for worldwide rental markets for these 8 cities are very similar, however, each city has its own small unique attributes. Using exploratory data analysis, we were able to extract similarities related to location. In all cities, there's a central area where Air Bnb properties tend to cluster, and as the distance away from this cluster increases, the price decreases. Property types tend to be in very similar proportions related to the rental properties available with private rooms and entire homes/apartments taking up the vast majority of those offered. Running the Latent Dirichlet Allocation algorithm on the description of each listing, we were able to uncover topics and frequent words used to describe each listing. The city name and location-based buzzwords such as "walking distance" or "public transportation". This further substantiates the findings in the EDA stage around a central cluster of listings, which is the tourist-heavy/downtown area of each city. Additionally, almost all of the description words can be grouped into the "Budget" or "Location" topic bucket, leading to the conclusion that the majority of AirBnb users are looking for a budget option in a tourist area. Running and comparing the Ridge, Lasso, and ElasticNet regression models, ~50% of the variance of the price can be explained by the variance of the other independent variables. The Lasso model performed marginally better and after standardization, there were slight improvements in Montreal, Vancouver, Barcelona and Sydney, while in Toronto, LA, Chicago and Stockholm, the performance was marginally worse. Using the XGBoost algorithm, we were

able to find the most weighted variables for each dataset. Each city has its own unique set of five variables, however, themes around the room type (Entire Home/Apartment, Private room) and bedrooms are prevalent in all cities as top weighted variables. Removing all variables except the top ten highly weighted variables and re-running the XGBoost model yielded little to zero gains in the RMSE/R^2. These conclusions hopefully will help aid and guide user's who are looking for Air Bnb's in the short-term; given the unpredictable nature of the travel industry due to the COVID-19 pandemic, it's worth extending/monitoring this project as more data becomes readily available.

References

Chen, Y., & Xie, K. (2017). Consumer Valuation of Airbnb listings: A hedonic pricing approach. *International Journal of Contemporary Hospitality Management*, *29*(9), 2405–2424. https://doi.org/10.1108/ijchm-10-2016-0606

Chew, J. (2017, October 21). *Improving Airbnb yield prediction with text mining*. Medium. Retrieved February 8, 2022, from https://towardsdatascience.com/improving-airbnb-yield-prediction-with-text-mining-9472c0181731

Falk, M., Larpin, B., & Scaglione, M. (2019). The role of specific attributes in determining prices of Airbnb listings in rural and urban locations. *International Journal of Hospitality Management*, *83*, 132–140. https://doi.org/10.1016/j.ijhm.2019.04.023

Magno, F., Cassia, F., & Ugolini, M. M. (2018). Accommodation prices on Airbnb: Effects of host experience and market demand. *The TQM Journal*, *30*(5), 608–620. https://doi.org/10.1108/tqm-12-2017-0164

Önder, I., Weismayer, C., & Gunter, U. (2018). Spatial price dependencies between the traditional accommodation sector and the sharing economy. *Tourism Economics*, *25*(8), 1150–1166. https://doi.org/10.1177/1354816618805860

Perez-Sanchez, V., Serrano-Estrada, L., Marti, P., & Mora-Garcia, R.-T. (2018). The what, where, and why of Airbnb price determinants. *Sustainability*, *10*(12), 4596. https://doi.org/10.3390/su10124596

Zhang, Z., Chen, R., Han, L., & Yang, L. (2017). Key factors affecting the price of Airbnb listings: A geographically weighted approach. *Sustainability*, *9*(9), 1635. https://doi.org/10.3390/su9091635