Hotel Grader

CMPT 353
Summer 2023
Simon Fraser University

Qingrui Li (301400099)
Weifeng Wu (301385627)
Alex Cho (301409492)

Abstract/Introduction/Problem

A spur-of-the-moment trip is an exciting decision, but deciding where to travel can be a headache. The budget at hand can impose many limitations. Among them, hotels can take up a large portion of the budget. In most cases, people choose to use the internet or some apps to find information, but this is not only time-consuming but may not provide as many details as desired.

The idea is to help people get the best price faster based on the requirements of the hotel itself (e.g., the number of rooms) and the points of interest (POI) around the hotel. With this in mind, we started the project, which includes the following questions:

1. What factors affect the price of a hotel?
2. Is the price of the hotel in question high simply because it is in a desirable location?
3. Does the hotel's location contribute more to the hotel price than other factors?
4. Are the estimated prices reliable?

This project uses Vancouver as an example and takes into account the fact that people hesitate to make sudden decisions and are more likely to consider walking or using public transportation when traveling.

Data and Cleaning

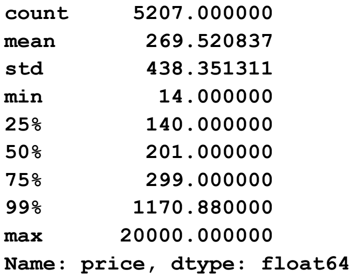
Our data is sourced from the Airbnb site and Overpass.eu for relevant location data in Vancouver. Our challenge is in trying to format the data to a uniform standard to put in our statistical models.

For the Overpass.eu dataset:

1. **Retaining useful features**:
   - ‘id’: The POI id
   - ‘lat’: The latitudinal positions of POIs
   - ‘lon’: The longitudinal positions of POIs
2. **Challenges**:
   - Since the Overpass data does not give consistent and properly formatted data, we had to clean it ourselves.
   - Some examples of these challenges are that the data did not provide explicit longitudinal and latitudinal data. Rather it was hidden further down the hierarchy. For example, some queries explicitly state the lat and lon but some place the lat in lon in a new ‘central’ tag.
3. **Consistency**:
   - Since the Overpass data is given through queries given by the website we were able to separate all the data in separate categories.

For the Airbnb dataset:

1. **Retaining useful features**:
    - ‘id’: The hotel id helps identify hotels.
    - ‘name’: The name of the hotel with its rating and number of rooms.
    - ‘latitude’: The position of hotels.
    - ‘longitude’: The position of hotels.
    - ‘price’: The price per day for the hotel.
    - ‘minimum_nights’: Minimum number of days of residence for customers.
    - ‘room_type’: Type of hotel: one-room rental or whole house/apartment rental.
    - ‘reviews_per_month’: Average reviews by customers per month.
2. **Read all the datasets grabbed from Overpass.eu that have been cleaned.**
3. **Filtering the data to be used:**
    - Select data for ‘room_type’ that only rents the entire house/apartment. While traveling, we consider some privacy as well as convenience of traveling, so we just focus on renting the whole house/apartment.
    - Grab the rate and number of rooms from the ‘name’ and add new columns, ‘rate’ and ‘num_bedroom’. Then, delete the data when ‘num_bedroom’ is empty.
    - Remove some extraordinarily high prices (price > 1170.88). For some of the unusually high-priced houses, most customers will not choose them while traveling.



```
count     5207.000000
mean       269.520837
std        438.351311
min         14.000000
25%        140.000000
50%        201.000000
75%        299.000000
99%       1170.880000
max      20000.000000
Name: price, dtype: float64
```

    - Delete minimum nights larger than 365. For the minimum nights, since we do not consider some long-term rentals, any minimum nights greater than 365 days are deleted.
    - Change the ‘reviews_per_month’ to 0 when it is empty. When ‘reviews_per_month’ is empty, it proves that no one reviewed the hotel during the time being counted so the average can be 0.
    - Calculate the number of SkyTrains within 1000m of each hotel, and the distance from the nearest SkyTrain to the hotel. Create two new columns, ‘num_skytrain’ and ‘dis_skytrain’.
    - Calculate the number of bus stops, markets, parks, and restaurants within 500m of each hotel, and the distance from the nearest bus stop, market, park, and restaurant to the hotel. Create two new columns for

each place as before. Since we wanted to travel easily and without straining ourselves, we decided to keep the distance from hotels to nearby buildings to less than 500m. However, due to the limitation of the number of Skytrain stations, we have to keep it within 1000m.

- Estimating ‘rate’ scores when it is New or NAN. Since some of the hotels are new and have not yet been rated, their number is relatively high. Deleting these hotels directly would result in a significant loss of important information. Therefore, we opt to extrapolate the 'rate' numbers for these unrated hotels based on existing data using the k-nearest neighbors (KNN) method.

Techniques

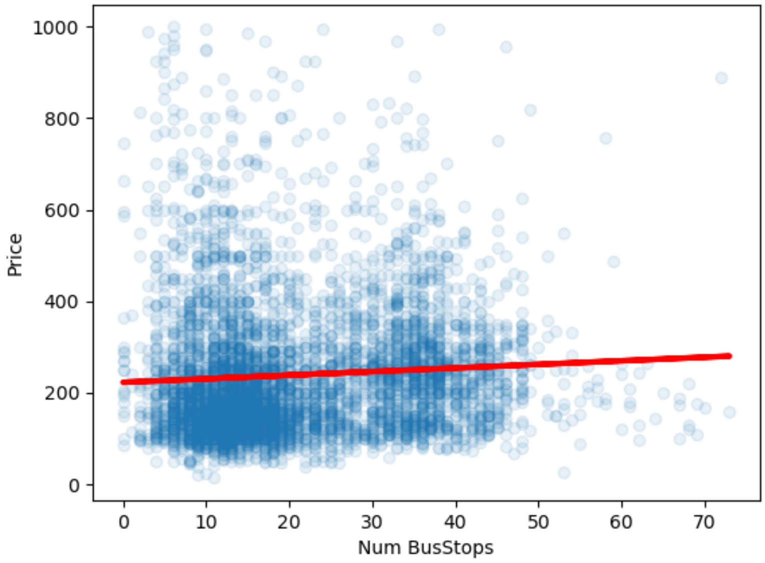
Our project uses ML and statistical analysis techniques from Python libraries.
In our exploratory analysis, we use linear regression to understand the relationship between variables and hotel prices. We also use the ANOVA test and the Tukey test to determine whether the hotels’ prices are different for different numbers of bedrooms.

In our predictive analysis, we trained various machine learning models to predict hotel prices by using Linear Regression, K-nearest neighbor Regression, Random Forest Regression, and Natural Network Regression.

Results and visualizations

1. **Bus station**
   a. **Number of Nearby Bus Stops (500m distance)**
   When traveling or moving to a new location for the first time or a repeat tourist. We believe that having public transport nearby to the temporarily chosen residence may be a factor of some kind to the decisions owners have made to price their homes.

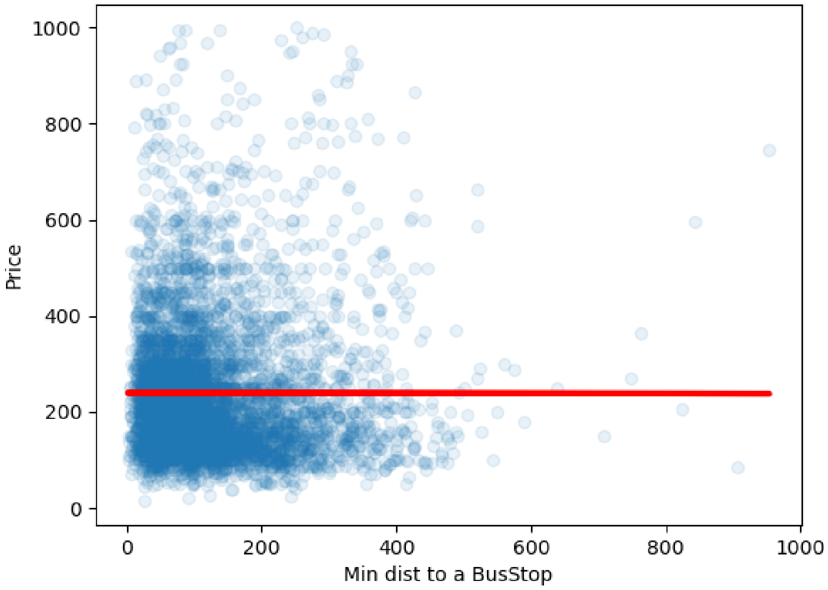
   For our examination, we used linear regression to find any linear relationship between the price of the hotel compared to the number of nearby bus stops. Using Linear regression our null hypothesis is that the 2 values have no relationship to each other. From our results, we got a P-value of 1.953206063912966e-06. Extremely small P-value so we can be confident that there is a kind of relationship between the number of bus stops nearby and the price. However our r-value is 0.06639540412142014, this means that this model doesn’t explain much of the variation of the data but it is significant.



b. **Shortest Distance to 1 bus stop**

We also had a hunch that the distance to the closest bus stop may play a bigger role in the price since we speculate that the shorter walking distance from your temporary place of residence may affect the price of the location.

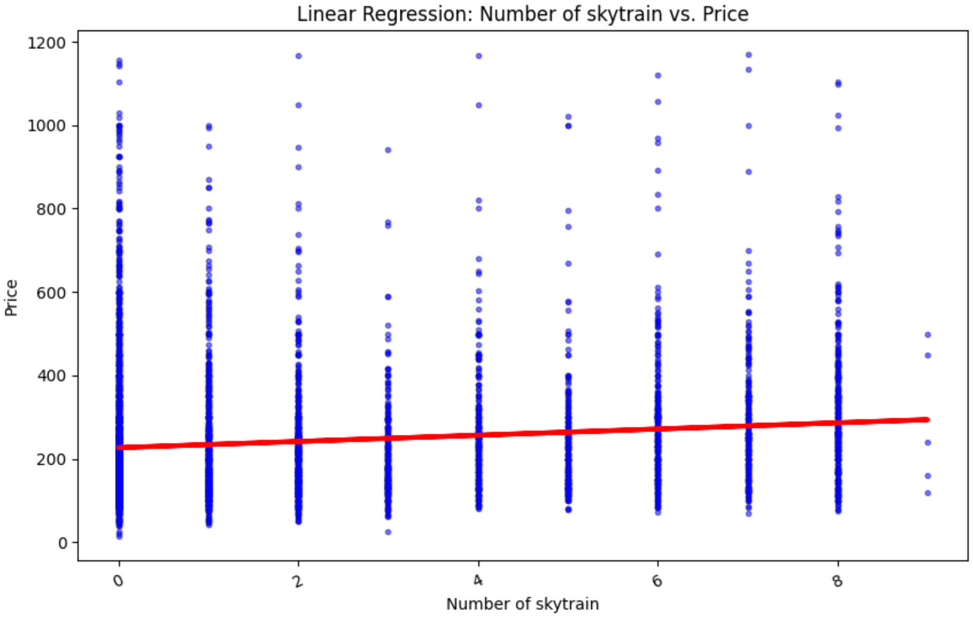
We are using the same method we used for the previous examination. From our results, we got a P-value of 0.9281105057409202. Extremely large P-value so we can be confident that there is no linear relationship between the number of bus stops nearby and the price. Also, our r-value is -0.0012603417513804538, which means that this model is not significant in explanation. With this, we are confident that there is no linear relationship between distance and the price of a bus stop.



2. **Skytrain (1000m distance)**

In our study, we analyzed the relationship between Airbnb prices and the number of nearby Skytrain stations. Considering that Airbnb's target audience is usually short-term travelers who may not have access to local transportation or a driver's license, the presence of nearby skytrain stations becomes a crucial factor, indicating the convenience of mobility around the Airbnb property. Based on this background, we assume the following hypothesis: the number of skytrain stations has no significant impact on Airbnb prices.

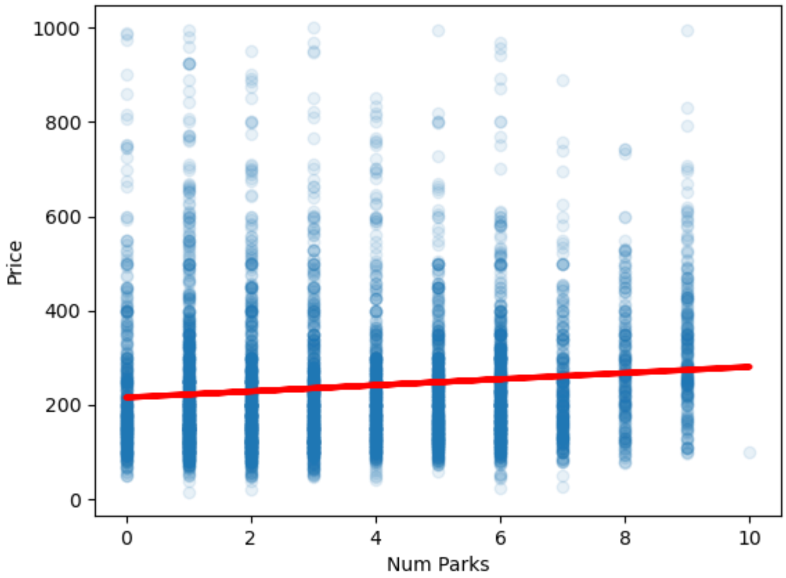
We visualized the relationship between prices and the number of skytrain stations through data visualization. The plot showed a weak linear relationship, indicating that as the number of skytrain stations increases, the corresponding prices tend to rise. However, drawing conclusions solely from the plot is not sufficient. Therefore, we decided to conduct a hypothesis test to assess the significance of this relationship. Through the hypothesis test, we calculated a p-value of 6.973069933083462e-20, which is smaller than 0.05. As a result, we rejected the null hypothesis and arrived at the following conclusion: the number of skytrain stations has a significant impact on prices, and this impact follows a linear relationship.



3. **Park**
   a. **Number of Nearby Park (500m distance)**

Although parks may not be stated in the text to factor in determining prices of homes. It may have an effect aesthetically [1] In this (review of the impact of urban parks and green spaces on residence prices in the environmental health context) states that urban parks do play a role in real estate prices. However, in our case of hotels and rentals from Airbnb, we want to see if it does show in our research.

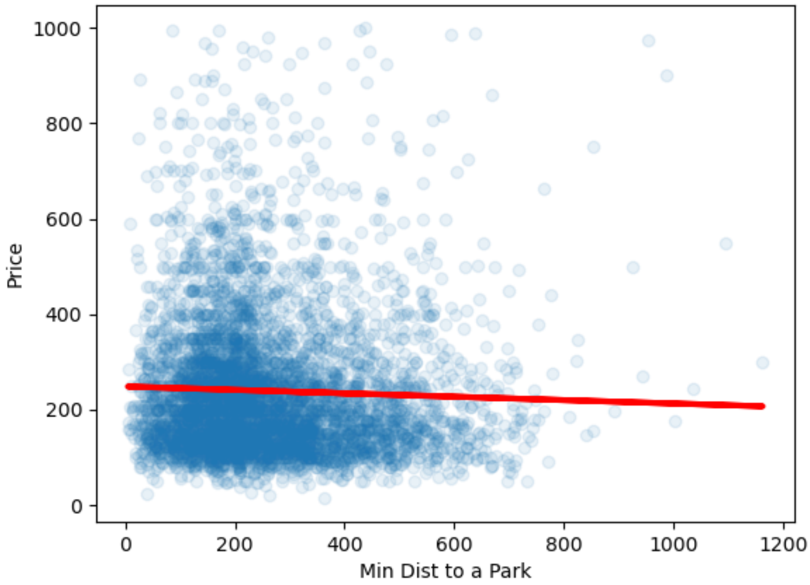
For our examination, we used linear regression again to find any linear relationship with the price of the location compared to the number of nearby parks. Using Linear regression our null hypothesis is that the 2 values have no relationship to each other. From our results, we got a P-value of 5.970796753161142e-16. Extremely small P-value so we can be confident that there is a kind of relationship between the number of bus stops nearby and the price. However our r-value is 0.112650431427431, this means that this model doesn't explain much of the variation of the data but it is significant.



b. **Shortest Distance to 1 Park**

Since parks do play some kind of role in determining the price of living spaces we are also curious if distance plays any role in pricing as well. We are speculating that the closer the park to the residence the higher the price will be.

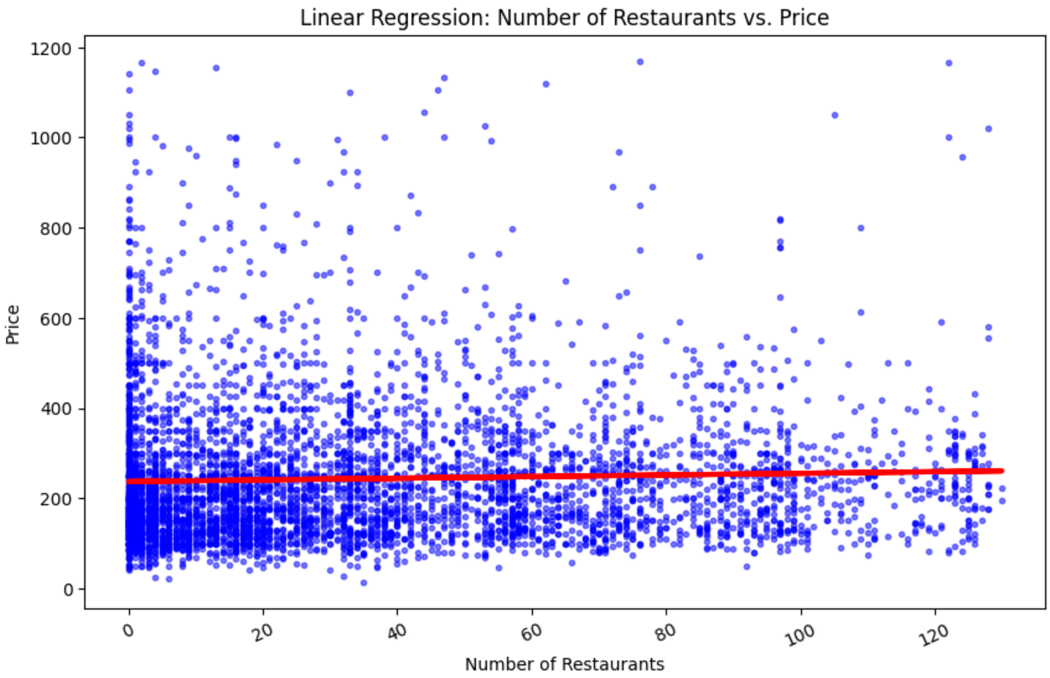
For our examination, we are using the same method as before. Using Linear regression our null hypothesis is that the 2 values have no relationship to each other. From our results, we got a P-value of 0.008427909326658815. Small P-value so we can be confident that there is a kind of relationship between the closest park and the price. However our r-value is -0.036788629545623355, this means that this model doesn't explain much of the variation of the data but it is significant.



4. **Restaurant**

   In our analysis, we investigated the relationship between restaurant availability and hotel prices. We considered the significance of restaurants as a factor in influencing price trends. The data visualization showed a weak linear relationship between restaurant quantity and hotel prices. As the number of restaurants increased, there was a slight tendency for prices to rise. However, drawing definitive conclusions solely from the plot would be premature. Therefore, we performed a hypothesis test to evaluate the significance of this relationship.

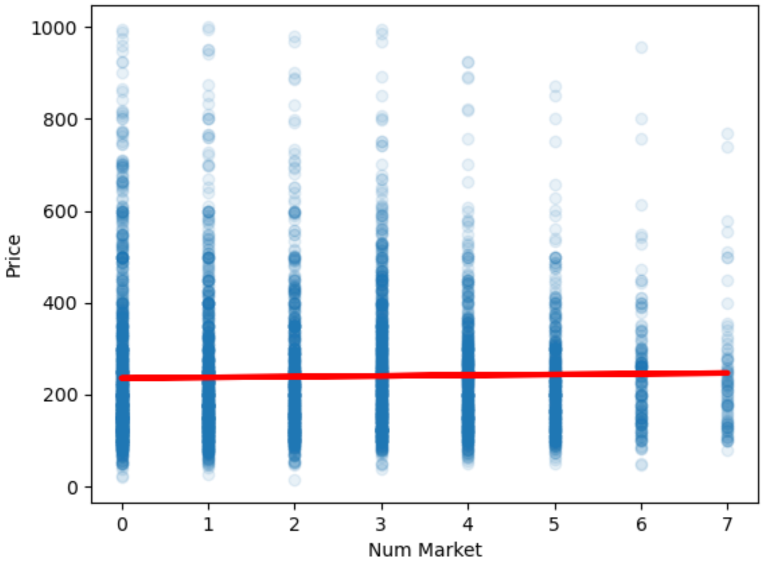
   The hypothesis test yielded a p-value of 0.004745527834520076, which is considerably smaller than the commonly used significance level (typically set at 0.05). As a result, we rejected the null hypothesis and reached the following conclusion: the number of restaurants has a statistically significant impact on hotel prices, and this impact exhibits a weak positive correlation. Combining the results from data visualization and hypothesis testing, we found a statistically meaningful association between restaurant availability and hotel prices. The correlation coefficient of 0.039336 further supports this weak positive relationship.



5. **Market**
   a. **Number of Nearby Markets (500m distance)**

For markets, we speculate that locations that are near markets or supermarkets do play a role in the cost of hotels. Since traveling long distances to get any food could be a deterrent for anyone who wants to live in temporary housing. So we are curious as to what extent markets have a role in the pricing.

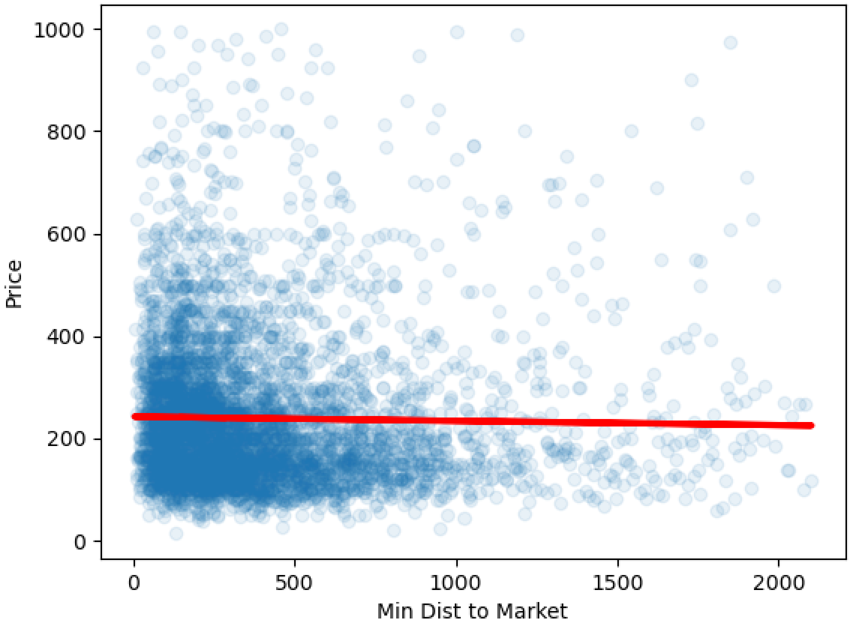
For our examination, we used linear regression to find any linear relationship with the price of the hotel compared to the number of nearby markets. Using Linear regression our null hypothesis is that the 2 values have no relationship to each other. From our results, we got a P-value of 0.1421403084521235. A large P-value so we can be confident that there is no linear relationship between the number of markets nearby and the price. Additionally, our r-value is 0.020502902981866154, this means that this model doesn't reflect the relationship at all.



b. **Shortest Distance to 1 Market**

As stated before we are curious if distance has any part in pricing.

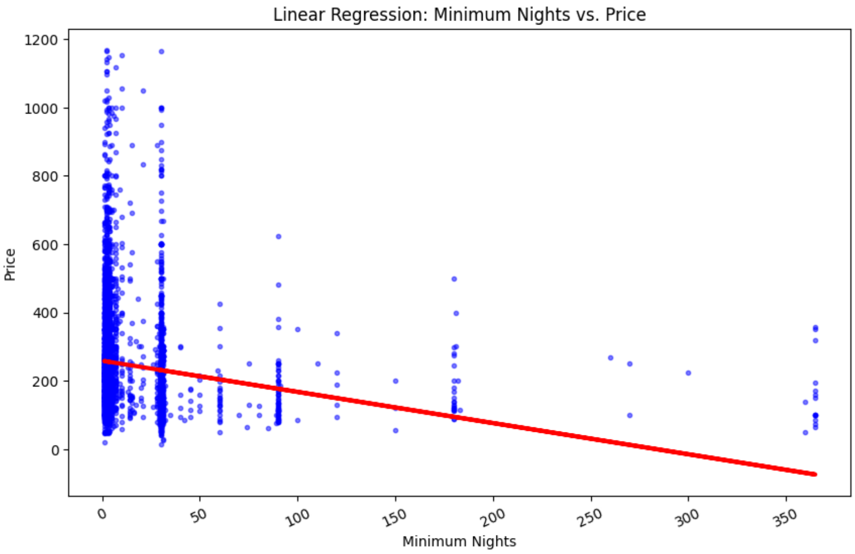
For our examination, we used linear regression to find any linear relationship with the price of the hotel compared to the distance to a single market. Using Linear regression our null hypothesis is that the 2 values have no relationship to each other. From our results, we got a P-value of 0.15744194702697212. Large P-value so we can be confident that there is no linear relationship between the distance and the price. Also, our r-value is -0.019746839284314237, this means that this model does not reflect the data.



6. **Minimum_nights**

   In this part of our analysis, we focused on exploring the relationship between the minimum nights required for booking and the predicted prices. The minimum nights refer to the minimum duration a property can be booked for. From the data visualization, we observed a negative linear relationship between the minimum night and the predicted prices. As the minimum nights required for booking increased, the linear regression predicted lower prices.

   To validate the significance of this relationship, we conducted a hypothesis test. The obtained p-value was 1.6426045980424842e-36, which is significantly smaller than the commonly used significance level (usually set at 0.05). Consequently, we rejected the null hypothesis, leading us to conclude that the number of minimum nights has a statistically significant impact on the predicted prices. Additionally, the correlation coefficient of -0.17451 reinforces this negative relationship. The correlation coefficient value indicates that there is a weak negative
   correlation between the minimum night and the predicted prices.
   Moreover, Since the correlation coefficient between the minimum number of days and the price is negative, when the minimum number of days is large enough, there is a negative rent price, as demonstrated in the picture, which is a limitation in this analysis

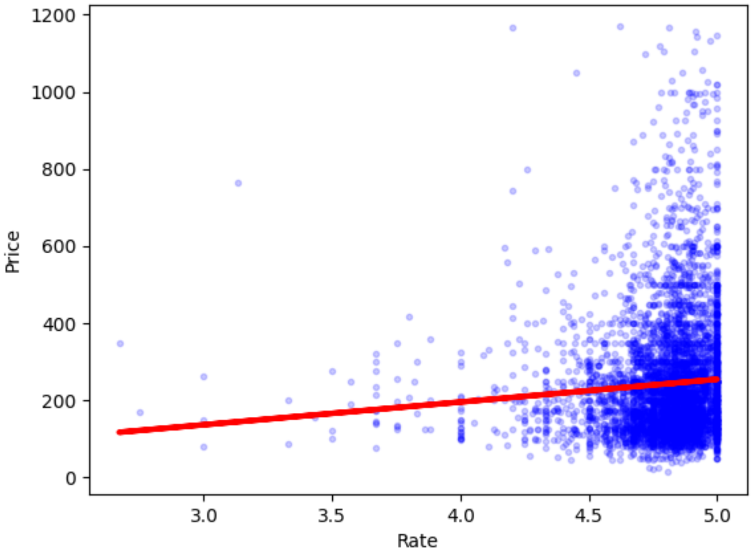


7. **Rate**

   In many cases, the higher a hotel's rating is, the more satisfied the customer is with certain parts of the hotel, such as the environment, service, and so on. Accordingly, we suspect that the hotel will adjust its prices based on the ratings given by these customers.

To prove this, we set the null hypothesis that price and rates have no linear relationship. The alternative hypothesis is that there is a linear relationship between price and rates. We draw a scatterplot of rates and prices. Then, we used linear regression to model the relationship between price and rates and use the slope and intercept we obtained to draw the best-fit-line. From the graph below, we can observe that the price increases as the rates increase.

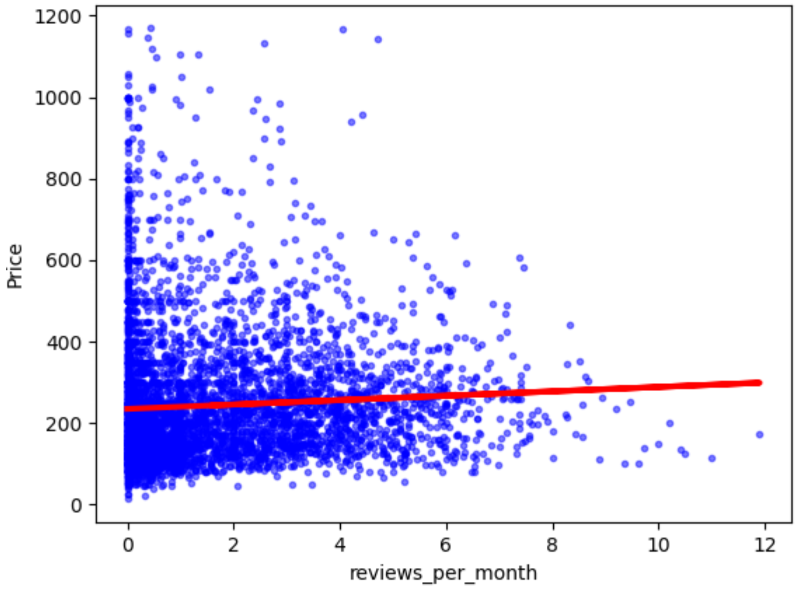
On the rigorous side, we analyze the p-value obtained in the linear regression. Since the p-value is 6.62e-08 <0.05, we reject the null hypothesis. Therefore, we can conclude that there is a linear relationship between the price and rates. However, our r-value is 0.0056, which means that this model doesn't explain much of the variation of the data but it is significant



8. **Number of reviews per month**
By analyzing the average number of reviews per month, we can approximate the popularity of the hotel. To a large extent, the popularity of the hotel will have some influence on the price.

Therefore, the null hypothesis is that price and number of reviews per month have no linear relationship. The alternative hypothesis is that there is a linear relationship between price and the number of reviews per month. We draw the graph and model linear regression on prices and the number of reviews per month as before. From the graph below, we can observe that the price increases as the number of reviews per month increases.

We also analyze the p-value obtained in the linear regression. Since the p-value is 5.57e-06 <0.05, we reject the null hypothesis. Thus, we can confidently conclude that there is a linear relationship between the price and the number of reviews per month. However, our r-value is 0.0040, which means that this model doesn't explain much of the variation of the data but it is significant

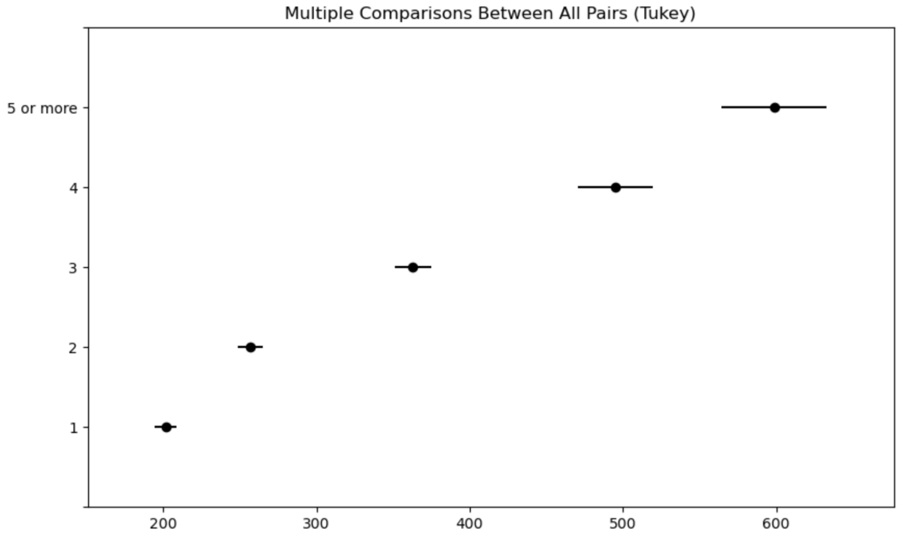


9. **Number of Bedrooms**

   First, we categorize the prices into different groups based on the number of bedrooms: '1 bedroom', '2 bedrooms', '3 bedrooms', '4 bedrooms', and '5 or more bedrooms’. Due to the limited number of bedrooms larger than 5, we combine them into a single group during testing.

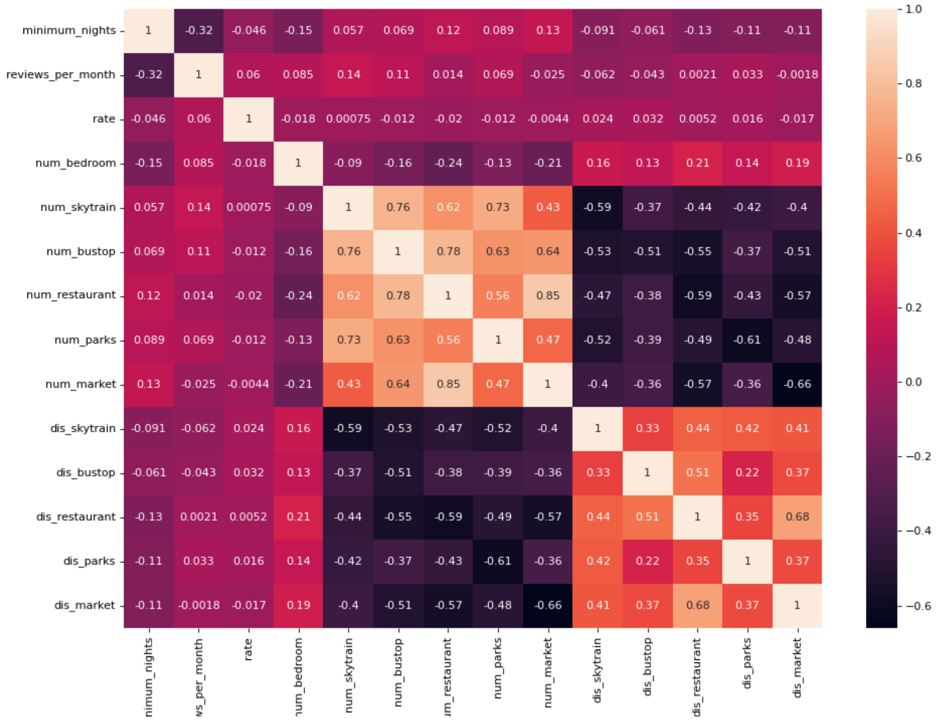
   Then, we use the ANOVA test to determine if the prices vary among the different groups. Our null hypothesis assumes that the prices are the same across all groups, while the alternative hypothesis suggests that at least one group has different prices from the others. The p-value obtained from the Anova test is 5.47e-302, which is less than the significance level of 0.05. Consequently, we reject the null hypothesis, leading us to confidently conclude that there are indeed differences in prices among the various groups. However, by solely performing the Anova test, we cannot identify which specific group exhibits distinct pricing.

   Finally, we conducted a post-hoc analysis to identify the groups where differences existed. The results are presented in the code and the graph below. As it turns out, all the groups show significant differences from one another. Thus, we can confidently affirm that the prices vary according to the number of bedrooms. Besides, there is an observable trend where the price increases with an increasing number of bedrooms.



Correlation:

In order to improve the accuracy of the model, we found that there is multicollinearity between the variables that affect prices. For this reason, we drew a correlation graph to visualize the correlation between each variable. After careful analysis, we found that the correlation coefficients of some of the variables were greater than or equal to 0.6, indicating a strong interdependence between them.



	nimum_nights	ws_per_month	rate	num_bedroom	num_skytrain	num_bustop	m_restaurant	num_parks	num_market	dis_skytrain	dis_bustop	dis_restaurant	dis_parks	dis_market
minimum_nights	1	-0.32	-0.046	-0.15	0.057	0.069	0.12	0.089	0.13	-0.091	-0.061	-0.13	-0.11	-0.11
reviews_per_month	-0.32	1	0.06	0.085	0.14	0.11	0.014	0.069	-0.025	-0.062	-0.043	0.0021	0.033	-0.0018
rate	-0.046	0.06	1	-0.018	0.00075	-0.012	-0.02	-0.012	-0.0044	0.024	0.032	0.0052	0.016	-0.017
num_bedroom	-0.15	0.085	-0.018	1	-0.09	-0.16	-0.24	-0.13	-0.21	0.16	0.13	0.21	0.14	0.19
num_skytrain	0.057	0.14	0.00075	-0.09	1	0.76	0.62	0.73	0.43	-0.59	-0.37	-0.44	-0.42	-0.4
num_bustop	0.069	0.11	-0.012	-0.16	0.76	1	0.78	0.63	0.64	-0.53	-0.51	-0.55	-0.37	-0.51
num_restaurant	0.12	0.014	-0.02	-0.24	0.62	0.78	1	0.56	0.85	-0.47	-0.38	-0.59	-0.43	-0.57
num_parks	0.089	0.069	-0.012	-0.13	0.73	0.63	0.56	1	0.47	-0.52	-0.39	-0.49	-0.61	-0.48
num_market	0.13	-0.025	-0.0044	-0.21	0.43	0.64	0.85	0.47	1	-0.4	-0.36	-0.57	-0.36	-0.66
dis_skytrain	-0.091	-0.062	0.024	0.16	-0.59	-0.53	-0.47	-0.52	-0.4	1	0.33	0.44	0.42	0.41
dis_bustop	-0.061	-0.043	0.032	0.13	-0.37	-0.51	-0.38	-0.39	-0.36	0.33	1	0.51	0.22	0.37
dis_restaurant	-0.13	0.0021	0.0052	0.21	-0.44	-0.55	-0.59	-0.49	-0.57	0.44	0.51	1	0.35	0.68
dis_parks	-0.11	0.033	0.016	0.14	-0.42	-0.37	-0.43	-0.61	-0.36	0.42	0.22	0.35	1	0.37
dis_market	-0.11	-0.0018	-0.017	0.19	-0.4	-0.51	-0.57	-0.48	-0.66	0.41	0.37	0.68	0.37	1

Given this finding, we chose to exclude these highly correlated variables from the model. Finally, we included the following variables to predict prices: “Reviews_per_month”, “rate”, “num_bedroom”, “dis_skytrain”, “dis_bustop”, “dis_restaurant”, “dis_parks”, and “dis_maket”.

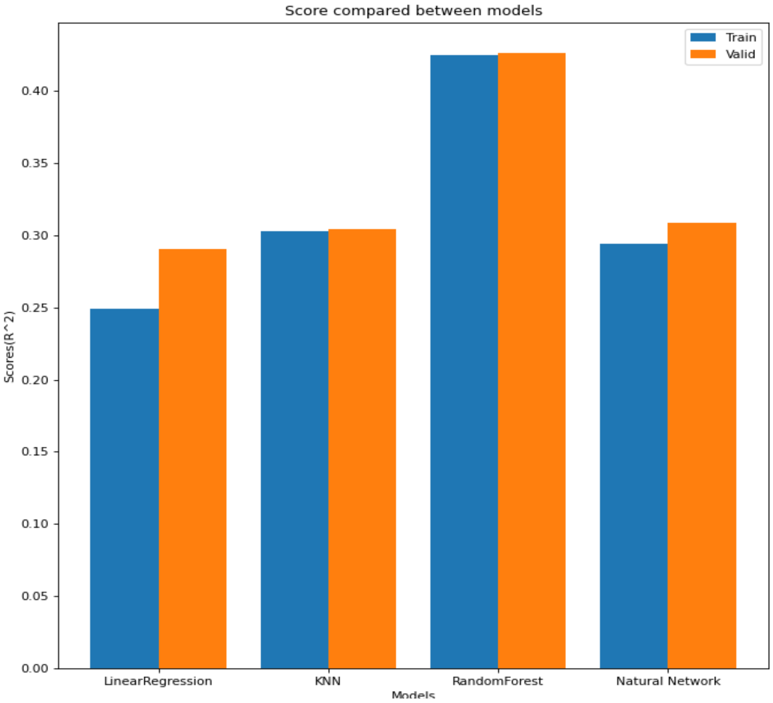
Prediction:

We randomly separate the hotel data into training and validation data. We use the training data to create and train the model and the validation data to test its suitability. As we aim to predict a numerical value, we will utilize Regression Machine Learning.

Method:

1. Linear Regression: The simple way to do the linear relationship with price and other variables.
2. K-nearest neighbors regressor: We use MinMaxScalar to scale each feature 0-1 and then set the n_neighbors=40.
3. Random Forest Regressor: We set n_estimators=500, and then set the max_depth=6, min_samples_leaf=15.
4. Neutral Network Regressor: We use the random_state as 1 and max_iter=600.

	Linear Regression:	K-nearest neighbors regressor:	Random Forest Regressor:	Neutral Network Regressor:
Training Data	0.249	0.303	0.425	0.294
Validation Data	0.290	0.304	0.426	0.309



By analyzing the graph and table, we observed that the Random Forest Regressor achieved the highest scores in both the Training Data and Validation Data. The difference in score between the Training and Validation Data is approximately 0.01, which is good. The Neural Network Regressor, Linear Regression, and K-nearest Neighbors Regressor all performed poorly.

Conclusion

In conclusion, our findings for the correlation between prices and a select few factors have proven to be fruitful. For POIs, we have found that the quantity of local POIs (1km) does have some kind of weak linear relationship to price. Namely, parks, bus stops, skytrains, and restaurants. Markets, not so much. As for the distances to the nearest market and bus stop, we found no linear relationship between the price and distance. For distance to the nearest park, we found a weak linear relationship. As for building features we suspected and confirmed the relationship between it and the price. For online reputation scores, we also found a high correlation as well. For our training model, we used the random forest regression model since it was the best choice, however, it only predicts the correct hotel price with 42.6% accuracy. This may be due to the fact that we have ignored some of the factors that have a greater impact on the price.

Limitations

Some limitation we have for our project is the vast number of data that can contribute to the price of the hotel. Such as ratings of the surrounding POIs, the age of the hotel, neighborhood scores, and much more. But we are not doing a case-by-case hotel analysis, since the data hunting would take far too long. Rather we are analyzing a set of factors that is easily obtainable that may or may not contribute to the price of the hotel. "listing.csv" only contains data for Vancouver and not other cities such as Richmond, which may result in some information affecting hotel prices not being included.

Sources:

[1] Chen, K., Lin, H., You, S., & Han, Y. (2022, September 7). *Review of the impact of urban parks and green spaces on residence prices in the environmental health context*. Frontiers in public health. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9490231/#:~:text=Their%20results%20revealed%20that%20the,weighted%20regression%20(GWR)%20models.

Project Experience Summary

Qingrui Li

- Cleaned Airbnb data(listing.csv), calculate the number of certain places and the minimum distances from the hotel to certain places, and then combined all data into hotel_listing.csv.
- Analyzed the relationship between rate, number of reviews per month, and number of bedrooms with prices.
- Created the correlation plot between variables.
- Trained machine learning models to predict prices based on variables.
- Completed my assigned part of the group project report.

Weifeng Wu

- Extracted and cleaned Bus Stop/ Skytrain/ Restaurant Data from Overpass.eu
- Analyzed the relationship between the number of nearby SkyTrain stations, the number of nearby restaurants, and minimum nights with prices.
- Completed my assigned part of the group project report.

Alex Cho

- Learned how to extract and clean Parks/ Markets Data from Overpass.eu
- Analyzed the linear relationship between the prices and the number of neighboring parks, markets, and bus stops.
- Analyzed the potential relationship between the minimum distances and the price.
- Learned how to use the python libraries verbosely and learned about the relationships between pricing and locality of POIs
- Completed my assigned part of the group project report.