Introduction

For this project, I hope to investigate how various predictors influence the price ($) of an Airbnb reservation (per night). Airbnb was selected for three reasons: (a) Accessibility to data (b) Applicability in a business context (c) Suitability for performing several evaluation models covered in this course. For Airbnb, my report will generate insights on pricing, and offer data-driven suggestions on how to improve the customer experience (new features).

The original dataset was extracted from Kaggle and included information across all major cities in the United States. In order to improve the accuracy of our investigation (and reduce the dataset size), I decided to reduce this down to only Seattle-area data. This dataset includes 19 columns and 3,817 rows (records). Predictors include:

- Whether host is a super host, or identity verified, amenities, accommodates, room type, extra people, instant bookable, cancellation policy, review_scores_rating, cleaning_fee, guests_included, minimum_nights, maximum_nights, neighbourhood_group_cleansed, bathrooms, bedrooms, beds

The goal is finding out which predictors affect listing price the most. Then, I will build several models (multiple linear regression, K-means clustering, classification tree) to extract insights on the predictors' relationship with Airbnb listing prices. The first step was examining the 18 predictors, analyzing them against price individually and looking at correlation values. In the second part of the project, I decided which predictors (strongest) to ultimately use in the final models and business analysis.

Step 1: Results from individual predictor analysis:

Predictors: “host_is_superhost”, “host_identity_verified”

- Correlation between ‘price’ and ‘host_is_superhost’: 0.0128673 (Weak Positive)
- Correlation between ‘price’ and ‘host_identity_verified’: 0.003438852 (Weak Positive)

The initial setup in data mining and analytics is to clean the data and convert it into acceptable and meaningful format. Consider the example of converting the column values having “t” with the meaning of “True” and “f” signifying “False”. It is easier for the computer to comprehend “1” (versus “t”) and “0” (instead of “f”) to perform computational analysis on that column. Similarly, I did modifications on columns that describe whether each:

1) Host is a superhost: Superhosts are experienced hosts who provide a shining example for other hosts, and extraordinary experiences for their guests. An individual may or may not be a Superhost and the response will have only “t” or “f” type of response.
2) Host identity is verified: Hosts IDs are checked to ensure they are legitimate individuals listing their property and help to avoid fraudulent incidents. A host identity may not be verified and the possibilities of response is restricted to “t” or “f”.

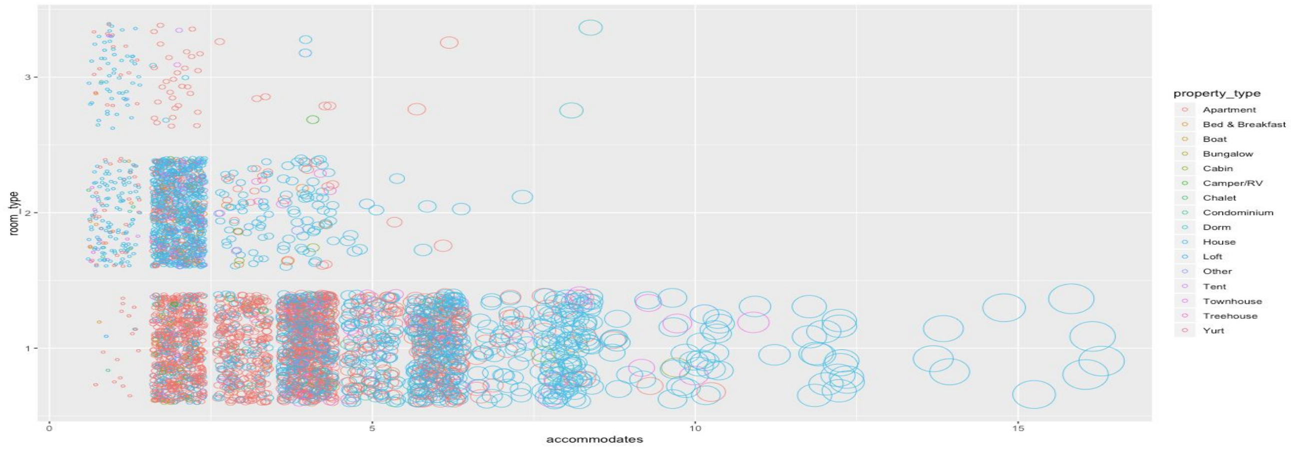
These values are modified to a 1-0 binary format as part of data cleaning and preprocessing. Another interesting aspect of Airbnb dataset involved a set of values that described what all amenities each property provides for. These descriptions (T.V., Wi-Fi, AC, etc.) were broken down using text mining.

Predictors: "accommodates",”room_type” ,”extra_people”,

- Correlation between ‘price’ and ‘**accommodate**s’: **0.1484314** (Moderate Positive)
- Correlation between ‘price’ and ‘room_type’: -0.0836181 (Weak Negative)
- Correlation between ‘price’ and ‘extra_people’: 0.0911211 (Weak Positive)

Thus, according to the correlation analysis between the variables and price, we can tell that the number of residents a property can accommodate has the most obvious positive correlation. The more people stay in the Airbnb, the higher the total price will be. Room type doesn't affect the price that much.

Next, I created a plot to show how many people can be accommodated in each type of room and property. By looking at this scatter plot, people can save time if they have a bigger group. They can just go looking for an entire house instead of other properties or room types.

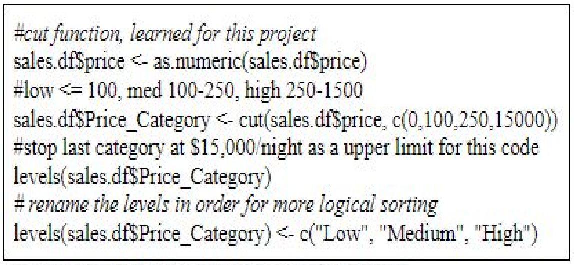


Predictors: "instant_bookable", "cancellation_policy", "review_scores_rating"

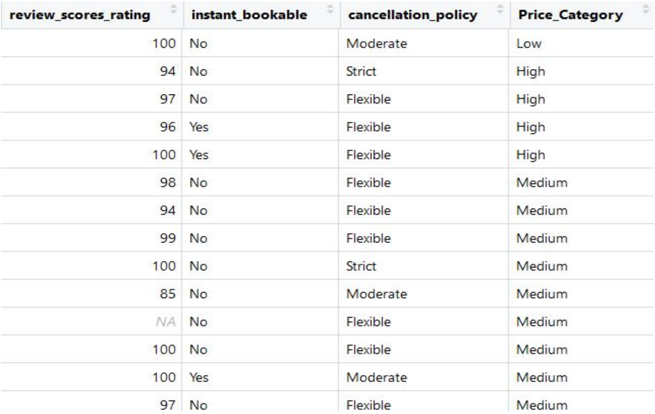
- Correlation between 'price' and 'instant_bookable': 0.01300163 (Weak Positive)
- Correlation between 'price' and '**cancellation_policy**': **0.1070552 (Moderate Positive)**
- Correlation between 'price' and 'review_scores_rating': 0.01256239 (Weak Positive)

Taking the correlation between the Airbnb being instantly bookable and the price, it seems that this factor is not correlated and may wish to be dropped from the model as it could produce overfitting. Similarly, the review scores rating is even less of a factor when we look at the correlation between price the review scores. The cancellation policy correlation with price is much higher than the other 2 predictors, and may have some weight in determining the price.

For preprocessing, data was releveled and a new column, ‘Price_Category’, was added to the data frame to help describe the price-level of the data, allowing easier sorting and grouping by price. This decision was made because operating at the grain of exact price/night for these particular predictors (i.e. the cancellation policy) will produce issues with accuracy. In other words, whether a host has a strict cancellation policy will not be a strong predictor of price, as confirmed through a confusion matrix, where the accuracy was extremely low. Hence, through the implementation of the cut function, the ‘Price_Category’ column separated price into three levels: “Low” being up to $100 a night, “Medium” being between $100 and $250 and “High” being greater than $250. This level of grain is more appropriate in the initial exploration of price against these 3 particular predictors. A code snippet and screenshot can be observed below:



```
#cut function, learned for this project
sales.df$price <- as.numeric(sales.df$price)
#low <= 100, med 100-250, high 250-1500
sales.df$Price_Category <- cut(sales.df$price, c(0,100,250,15000))
#stop last category at $15,000/night as a upper limit for this code
levels(sales.df$Price_Category)
# rename the levels in order for more logical sorting
levels(sales.df$Price_Category) <- c("Low", "Medium", "High")
```



review_scores_rating	instant_bookable	cancellation_policy	Price_Category
100	No	Moderate	Low
94	No	Strict	High
97	No	Flexible	High
96	Yes	Flexible	High
100	Yes	Flexible	High
98	No	Flexible	Medium
94	No	Flexible	Medium
99	No	Flexible	Medium
100	No	Strict	Medium
85	No	Moderate	Medium
NA	No	Flexible	Medium
100	No	Flexible	Medium
100	Yes	Moderate	Medium
97	No	Flexible	Medium

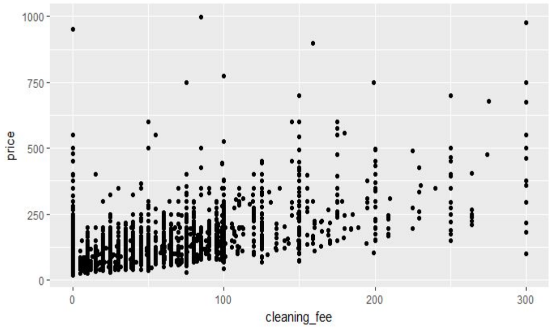
Predictors: "cleaning_fee", "guests_included", "minimum_nights", "maximum_nights"

- Correlation between ‘price’ and ‘**cleaning_fee**’: **0.5858448** (Strong Positive)
- Correlation between ‘price’ and ‘**guests_included**’: **0.3993583** (Strong Positive)
- Correlation between ‘price’ and ‘min. nights’: 0.01771233 (Weak Positive)
- Correlation between ‘price’ and ‘max. nights’: -0.003729496 (Weak Negative)

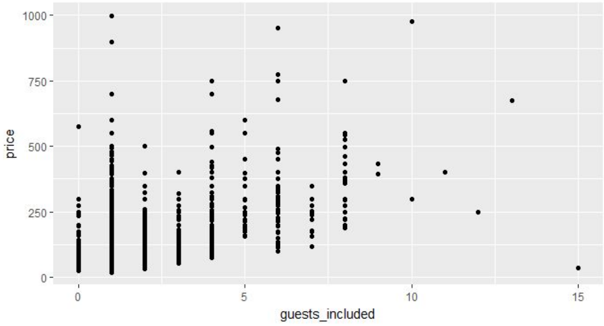
For cleaning, the dollar sign was removed from the price column and cleaning fee columns. Because the raw dataset contains blanks for certain cleaning fee rows (which should be interpreted as these properties having no cleaning fees), I replaced these blanks with 0's. Afterwards, the correlation between these individual predictors and price was inspected.

<u>Plots Inspecting Correlation:</u>

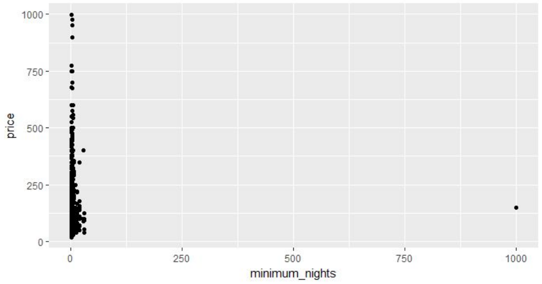
Price, Cleaning Fee ($)



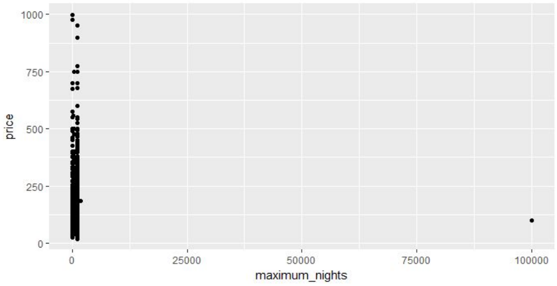
Price, Number of Guests Included in Stay



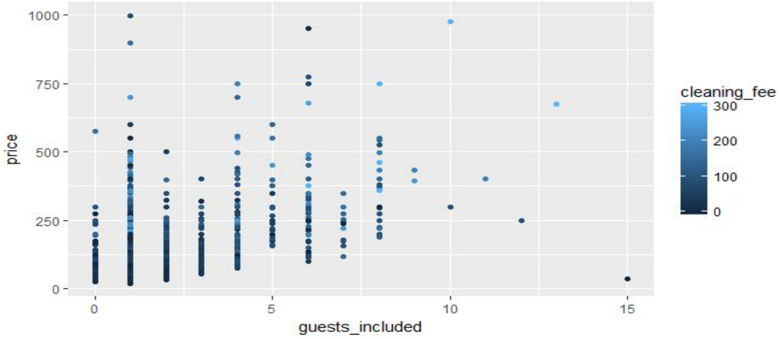
Price, Min. Nights Property is Available



Price, Max. Nights Property is Available



Already we can see that min and max nights (that a place is available to be rented) do not seem to have a strong correlation with the price (varies greatly even if the nights are the same). Below is a plot for guests included, versus price, but this time with cleaning fee visualized (as the key):



Predictors: “neighbourhood_group_cleansed”, “bathrooms”, “bedrooms”, “beds”

- Correlation between price, beds (# available)**: 0.596**（Strong Positive)
- Correlation between price, bathrooms: **0.524**（Strong Positive)
- Correlation between price, bedrooms: **0.634**（Strong Positive)
- Correlation between price, neighbourhood_group (in Seattle): -0.0192 (Weak Negative)

It seems that correlation between location and price is relatively small in the Seattle area (very weak negative correlation). This was surprising for me, as I expected neighborhood regions (i.e downtown Seattle, Bellevue) to have a much stronger impact on price. This means Airbnb prices in the Seattle area are fairly homogeneous, perhaps due to 2 reasons: Seattle being a Satellite city (meaning highway transportation is required to get to many amenities anyways) so absolute location matters less, or that housing in the Seattle area tends to be similar (in value/size). Regardless, I omitted this predictor and considered the other 2 predictors (beds, bathrooms) later in the consolidation stage. For cleaning, the neighborhood column contains character descriptions (names) of 18 places (i.e. Capitol Hill, Northgate). So, I converted the rows to numeric, exchanging names for 0-18 values. The other 3 predictors are already numeric, and did not require additional processing.

Step 2: Consolidating Results

After examining the individual correlation values for all of the predictors against price, I decided to set a threshold of correlation > 0.1 in strength, to proceed with the consolidated models. This means that we eliminated 11 predictors (attributes), leaving us with strongest 7:

- Beds
- Bathrooms
- Bedrooms
- Cleaning_fee
- Guests included
- Cancellation policy
- Accommodates

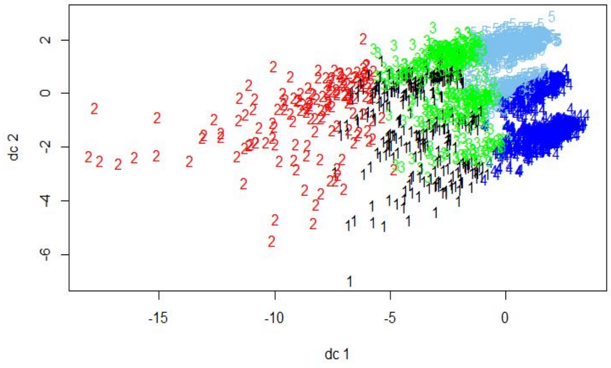
I will now proceed with 3 methods: the first, being a k-means model on these predictors to determine groupings for listings, the second being a multiple linear regression using these predictors (to determine a price model), and the third being a classification tree (for hosts).

Method 1: K-Means

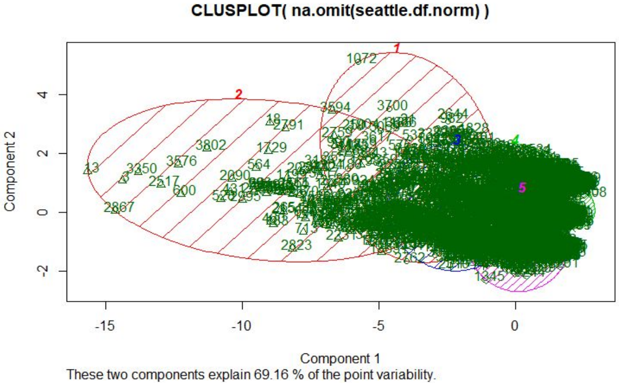
Applying unsupervised learning on our 7 strongest predictors, I wish to extract general business insights for similarities between groups (clusters). Procedurally, I first normalized the data for our predictors (to make sure they fall under a common range) and set the seed as 1. After iterating through several values for k (the number of groups ranging from 3 to 10), I determined (k) = 5 to be the optimal elbow point. For plotting, I called on library(stats), library(fpc), and library(cluster). Below is the output of our k-means and summary statistics:

withinss: within-cluster sum of squares, one component per cluster.
2512.956 2434.750 2444.166 2834.636 2453.428

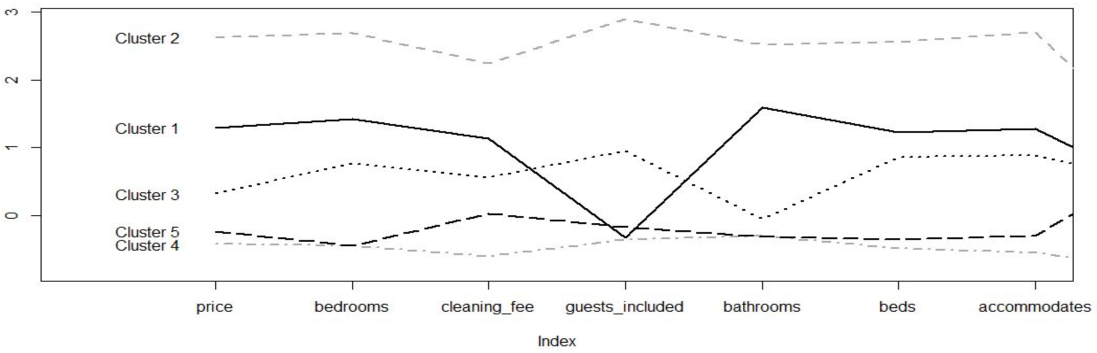
Cluster Centers (each color is a cluster):



Clusterplot: Each Cluster is Circled



Profiling Centroids (our 5 clusters, for the 7 predictors)



Interpreting these results, each cluster has around 650-800 properties (a relatively good split). From the cluster center plot, we can observe that cluster 2 has the widest distribution (points are the most scattered), suggesting more variance between these kind of properties. There is some significant overlap between points in clusters 3, 4, and 5. This suggests most of the predictors for these clusters have similar values, although there should be differences in 1-2 key predictor areas. Overlap to a certain extent makes sense, because on a fundamental level, these are all properties which must have at least 1 bed, 1 bathroom (examples of predictors) for guests to use.

Turning to the profiling of the clusters' centroids, we can examine details more closely:

- Cluster 2 has (on average) the highest values for all 7 predictors (including guests included) and price. This group is akin to a renting a party house, or an entire building - suitable for catering large groups of people
- Cluster 1 still has relatively high price and cleaning fees, although it has fewer guests included in the booking (listing) records. These listings are likely not booked for accommodating too many guests, but rather for private use. As such, the high prices here represent a cluster of 'luxury' listings.
- Cluster 3 is similar to cluster 1, but lower prices and more guests accommodated. This could be a 'premium' or 'family' listing group.
- Cluster 4 has affordable prices, but also an average cleaning fee. There are minimal additional guests, and likely only 1 bedroom, bed, and bathroom. This represents a 'typical' (like a 2 or 3 star hotel) Airbnb listing, selected for being economical.
- Cluster 5 is similar to cluster 4, but has no cleaning fee and even fewer accommodations. This is the 'cheapest' group - similar to a hostel/couchsurfing, instead of a private room

Using these groups (party house, luxury, family premium, regular, budget) Airbnb can classify customers through targeted advertising, understanding the typical prices and behaviors of demographics using its service in the Seattle region. Certain groups, say, the family/premium might be an opportunity for Airbnb to push more extensively in the Holiday season as a differentiator from competitors like hotels.

Method 2: Multiple Linear Regression

Applying supervised learning, I now hope to determine the estimated price of a property (numerical value) based on our 7 strongest predictors. This requires running a multiple linear regression to generate our pricing model. I set the seed to 1, and assumed a 70% split between training and 30% for the validation set. A summary of our coefficients (training data) is below:

```
                   Estimate   Std. Error t   value          Pr(>|t|)
(Intercept)         11.0516     4.4043        2.51           0.012 *
bedrooms           20.2283      2.4705       8.19           0.00000000000000041 ***
cleaning_fee        0.4250      0.0352       12.07           < 0.0000000000000002 ***
guests_included    4.6636       1.1124       4.19           0.00002850314837230 ***
bathrooms          24.9916      2.8529       8.76           < 0.0000000000000002 ***
beds              -3.6407       2.3057      -1.58           0.114
accommodates       12.9618      1.4348       9.03           < 0.0000000000000002 ***
cancellation_policy -2.3327      1.6986       -1.37           0.170
```

<u>Equation</u>:

Y(price) = 11.0516 + (20.2283 Bedrooms) + (0.4250 Cleaning Fee) + (4.6636 Guests Included) + (24.9916 Bathrooms) + (-3.6407 Beds) + (12.9618 Accommodations) - (2.3327 Cancellation Policy)

<u>Accuracy</u>: testing on validation set

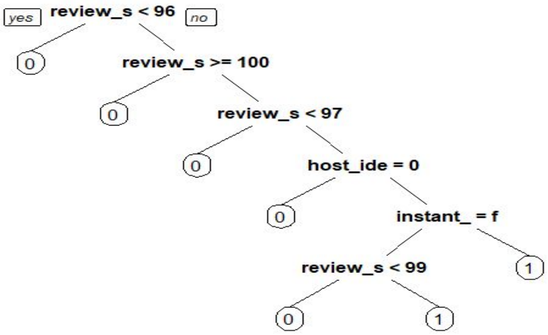
```
           ME     RMSE   MAE   MPE     MAPE
Test set   -1.59   53.8    37.4  -16.8     35.6
```

Using the pricing model generated, I am able to predict how much a listing in the Seattle region should cost, by entering values for the 7 predictors. We see that the accuracy of our linear model is good (i.e. mean error being single digit), which improves the confidence of our results. This could be useful for Airbnb, by letting potential hosts enter these values into the app to give them an estimate of how much their property should be listed at. This makes it easier for the host to determine a fair price (less research needed), and to quickly decide whether using the service is worthwhile. From a customer's perspective, having this information would also allow them to know if a listing is being priced too high, or if a property is a great deal. Hence, information from this predictive model could improve the user experience when integrated into Airbnb's service.

Method 3: Decision Tree, Classification

While I dropped the super-host attribute for our price prediction, it was crucial to understand why the label is important to Airbnb business and how predicting it can be helpful to Airbnb's growth potential as well as its appeal to the customers.

The idea of being a superhost goes in line with making the customer's stay comfortable and peaceful. Airbnb operating in the hospitality industry uses a customer satisfaction index as a metric to measure the performance of the hosts affiliated to them. It utilises the number of trips hosted, quality of the customer review, and the tendency of the hosts to respond on the website, all to measure the value of the service being provided. If this is above a certain benchmark during the quarterly checks, the system automatically awards the host this title of Superhost. In order to keep up with the rest of the hosts, it makes them competitive and in turn ensures that customers can choose a quality service experience. This allows a system of checks and balances for maintaining both – customer satisfaction index and performance measures.



Goal: Superhost or not. As per our analysis, the host will be a superhost if:

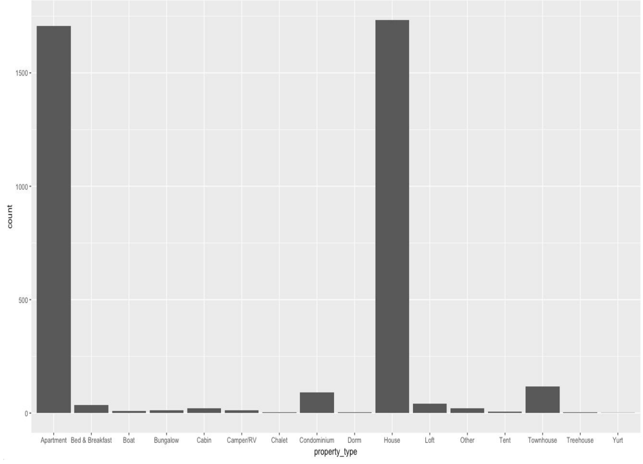
- The host identity has been verified and it has an instant bookability feature and the review score is above 97.
- The host identity has been verified and it does not have an instant bookability feature but the review score is above 99.

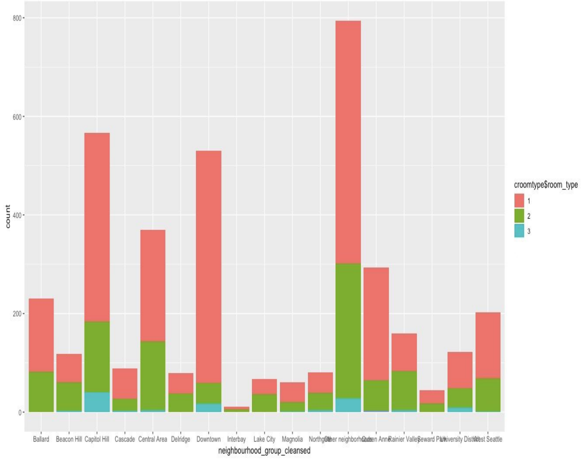
From the perspective of predictability, the above conditions need to be met for a host to be a superhost. Interestingly, price is neither a contributor nor an affected factor by the superhost type. On the contrary, the average price of properties where host is not a superhost is actually greater than the properties where host = superhost.

Data Visualization & Business Takeaways

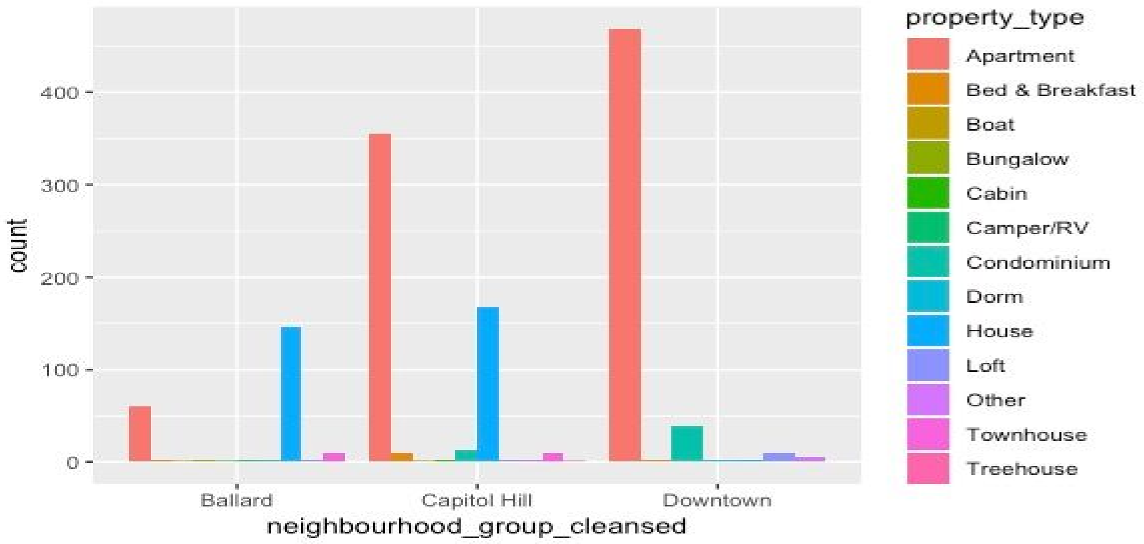
In this section, I will explore some shadow questions Airbnb may want to answer using this dataset. Through visualizations and plots, I will extract additional information about properties. First, we'll take a look at property type (apartment versus houses) and examine prices/demand projection for popular neighborhoods below:

Different property-type count in Seattle   Property count in different neighborhoods



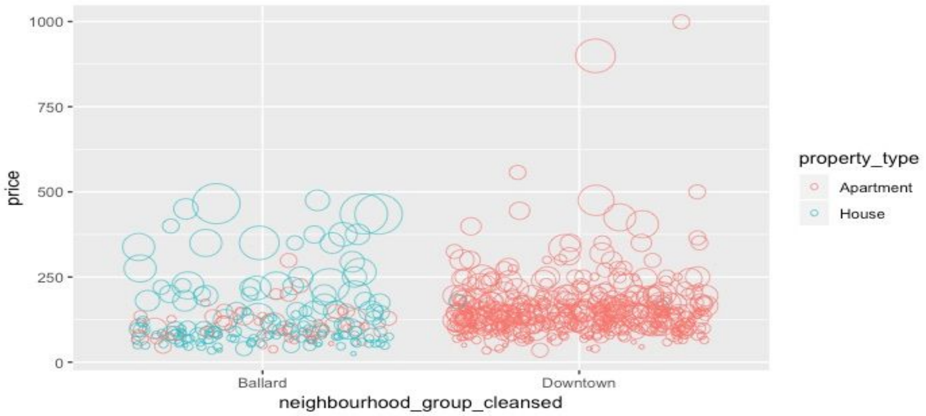


Density of each property per neighbourhood



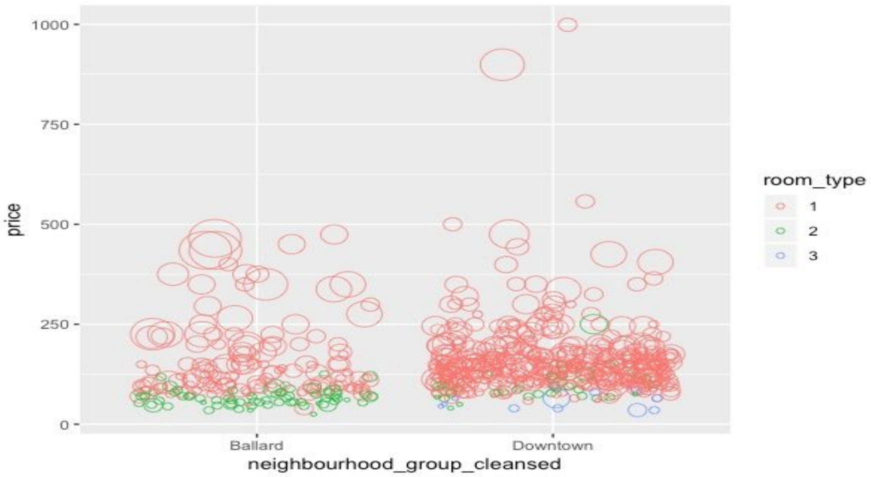
From the charts above, we can see that listings for downtown and Capitol hill are the most numerous. Additionally, apartments and houses are the most common type of properties with around 1750 counts each in Seattle area. Some unique property types like “Cabin”, ”Boat”, “Treehouse” can be found as well. We can discover the density of each property type in different neighbourhoods (extracting three areas as samples) to better inform the customer of what to expect from that area and the host to have a better idea of what this market is like.

Price for houses or apartments in Ballard vs Downtown



By plotting the price against property type, we can pick certain areas to explore (i.e. Downtown vs Ballard). The customer will have a better idea of what to pay for what kind of property in a certain area. In this example, we see that if you pay $250 to stay a night in seattle, you can get either an apartment downtown or a house in Ballard for a medium size group.

Price for Room_type Ballard vs Downtown (size based on accommodates)

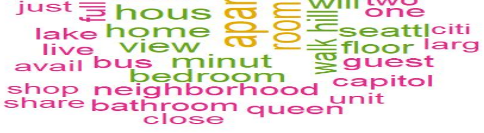


With the help of this chart, customers can develop a better projection on what location they can book for their group. A group of three, for instance, can book a shared room in downtown area with the lowest price, or they can pay slightly more to book an entire place in Ballard.

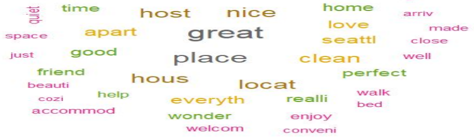
Text analysis

Word cloud 1 (below) displays the most frequently used words to share property description by the hosts. The occurrence can be segregated into 3 parts such as: property type, features, and surroundings. The words give a sense of the of these properties, their surroundings and potential preferences of what people look for in a holiday home, such as scenic views, quiet atmosphere, or access to bus services.

Word cloud 1: Description of properties



Word cloud 2: Customer reviews



Word cloud 2 (above) shows the result of text mining operation performed on the customer's reviews that gives an overview of their feedback. This in turn helps develop a sentiment analysis summary of the fact that Seattle-based travellers tend to have a comfortable and joyful experience at their host properties, and thus are more likely to choose Airbnb, or even recommend it to peers. There are almost no negatively connotated words in this word cloud.

Conclusion

Although I did gain valuable predictions and insights in this project, the process was not easy. In particular, I faced several obstacles: One was when I was creating the word cloud. Initially I wanted to examine the reviews with the lowest review scores to understand customers' pain points. To do this, I needed to merge the original dataframe with the review dataframe by listing_id. This was hard since both datasets are huge (thousands of rows long, over 1GB merged file). The generated cloud was not meaningful for our original goal of finding pain points since it still did not generate a sufficient number of negative words. It seems that people are being "nice", saying the positive words through comments, despite potentially giving low scores.

In this project, I used 7 predictors, which I selected as the most correlated ones to predict Airbnb listing price per night. This prediction model can be used for hosts setting listing prices and also for customers knowing whether the prices are fair. For preparation, I cleaned the huge dataset and selected strongest variables. For techniques learned, I used k-means for clustering, linear regression models to form a prediction equation, a classification tree, as well as text mining sentiment analysis and data visualization to draw further business insights.

Work Cited

"Cut Function." Function | R Documentation, www.rdocumentation.org/packages/base/versions/3.5.1/topics/cut.

Deshpande, Bala. "Market Price and Cost Modeling: One Application of k-Means Clustering." *Analytics Made Accessible: for Small and Medium Business and Beyond*, www.simafore.com/blog/bid/137814/Market-price-and-cost-modeling-one-application-of-k-means-clustering.

Kaggle, www.kaggle.com . Accessed 6 Nov. 2018.