AIRBNB-YELP DATA ANALYSIS

INFX 573 – GROUP 3

AVANTI CHANDE

DHRUV PAREKH

GOSUDDIN SIDDIQI

TABLE OF CONTENTS

ABSTRACT

In this project we wanted to find if there is any relation between the housing listed on the Airbnb website with the amenities such as food, pharmacies, nightlife and tourist must see attractions listed on Yelp. The data used in this project was publicly available on Airbnb and Yelp. From our analysis we found that there is no correlation between the review scores of the house listings and the amenities present within one mile walking distance of the listing. From further analysis we found that, the review scores may be related to the apartment type offered and the cancellation policy of the listing. Further, we performed a sentiment analysis of the reviews of the listing and concluded that there was positive feedback from the users for all the zip codes. We also found interesting results from our text mining analysis.

INTRODUCTION

David enrolled for MSIM course at the University of Washington and he preferred to stay at a homely apartment which was close to the University. Being an avid food lover, he also wanted restaurants which would offer different cuisines and delicious food around that apartment. Adam was on a vacation and wanted to explore the beautiful city of Seattle. He preferred a low cost dorm but wouldn't mind spending more on nightlife and adventurous activities in the city. Each of these individuals, came with a different reason to the city of Seattle. Each of them had an inclination to a different set of amenities- nightlife, restaurants, airport, adventure activities etc. All of this was available in Seattle.

This prompted us to research whether there was a correlation of available housing options with the amenities surrounding it. Which of these amenities has a high influence on a home buyer's choice? Do home buyers prefer one amenity over the other? Understanding these preferences is essential for the real estate business to grow.

RESEARCH QUESTIONS

The research is centered to determine preference of amenities while renting or buying a housing option in Seattle. Amenities means features such as nightlife, restaurants, coffee shops, financial institutions, educational institutions, activities etc. The housing options considered are that of Airbnb Seattle, for the year 2016. Our research questions are as follows:

- Do Airbnb units within walking distance(1 mile) of more than one type of restaurant rent more frequently than those outside of walking distance, adjusted for city, price, etc?
- Is there an association between the prices of the Airbnb units and the availability of the amenities available around it?

DATA COLLECTION/ WRANGLING

For the research we collected data from two publicly available sources of data:

- **Yelp**: The amenities' data was extracted from the Yelp website. We selected Yelp for three reasons. First, it offers a variety of options in terms of amenities such as bars, banks, breweries, food, nightlife, tourist's attractions, educational institutions etc. Second, it provides in depth details of different parameters such as price, location, reviews etc. which was crucial for our analysis. Third, it provides APIs which can be leveraged to extract different parameters for the analysis. To narrow our scope, we have centered our research on 4 amenities: pharmacies, food, tourists' must see places and nightlife. To extract the details of each amenity we used two Yelp APIs - API 2.0 and Fusion API. From the API 2.0 we extracted details such as geospatial information, neighbourhood, address, reviews, number of reviews etc. However, this API did not provide the price associated with each listing. Hence, we used Fusion API from which we extracted the price information for each of the listing.

- **Airbnb**: The details pertaining to housing were extracted from Airbnb website. There were different datasets. Airbnb provides housing data for the several city across the world, that primarily consists of cities in Europe and North American region. For the scope of this project, we intend to perform our analysis on data available for the city of Seattle. We chose Airbnb because it is one of the leading organizations in the real estate industry. Also, it's data is publicly available for the city of Seattle. Most importantly, it provided housing details such as geospatial information, reviews of houses, house types, number of bedrooms, reviews of property managers and other parameters which were crucial in our analysis.

We had around 92 features from our Airbnb dataset. However, we were not provided with any kind of dictionary for this dataset. It was left to use to interpret various features. Apart from the features already available, we created features that would be relevant and useful for our study. We computed number of all type of amenities for each of the listings present. Using the listing dataset we constructed the feature that counted the number of times a particular listing was rented. This was done by examining how many times the availability flag switched or toggled from "t" to "f". After constructing such features and eliminating features irrelevant in our study we narrowed down to 12-15 features for our study.

ASSUMPTIONS AND LIMITATIONS

Before delving into the Analysis, we would first like to mention our assumptions and limitations. They are as follows:

ASSUMPTIONS

For our analysis we made the following assumptions:

- Walking distance is considered as Haversine Distance and not the actual walking distance.

- Euclidean distance would not have been a great choice for our analysis. The results from this function would not have given a true representation of the actual distance. We choose to go with Haversine Distance that takes into account the radius of curvature of the earth.

- Amenities with 3 star rating and above were considered as good amenities.

- After rounds of discussions within the team and with classmates, we agreed to consider amenities with 3 stars or above as good amenities for our study. Only amenities listed on the Yelp website were taken into consideration. Our study focussed only on the amenities mentioned on Yelp

LIMITATIONS

Our study has the following limitations:

- Data of the year 2016 was only extracted for both Airbnb

- We choose to analysis the Airbnb data that was available for the year 2016. The data is available for the years 2014 and 2015 as well. Since we had no information through Yelp, as to when a particular amenity was established, we choose the listings for 2016.

- Statistical significance of other factors that affect the review score rating is not taken into consideration

ANALYSIS

During our preliminary analysis we figured out two outcome variables:

- Review score rating: it is the rating received by a listing from a user
- Frequency with which an apartment has been rented

Based on the frequency with which an apartment gets rented, we found that the most popular Airbnb listings were located in Capitol Hill, Portage Bay and First Hill as shown in the image below.

host_neighbourhood <fctr>	avg_times_rented <dbl>
Capitol Hill	49.0
Portage Bay	43.0
First Hill	34.0
Minor	34.0
Ballard	33.0
	29.0
Eastlake	27.0
Belltown	25.8
University District	24.0
Central Business District	21.0

Based on the average rating, we found that the neighbourhood of Alki, Anaheim and Arbor Heights were highly rated as shown in the image below.

host_neighbourhood <fctr>	avg_rating <dbl>
	100
Alki	100
Anaheim	100
Arbor Heights	100
Atlantic	100
Ballard	100
Belltown	100
Bitter Lake	100
Broadview	100
Bryant	100

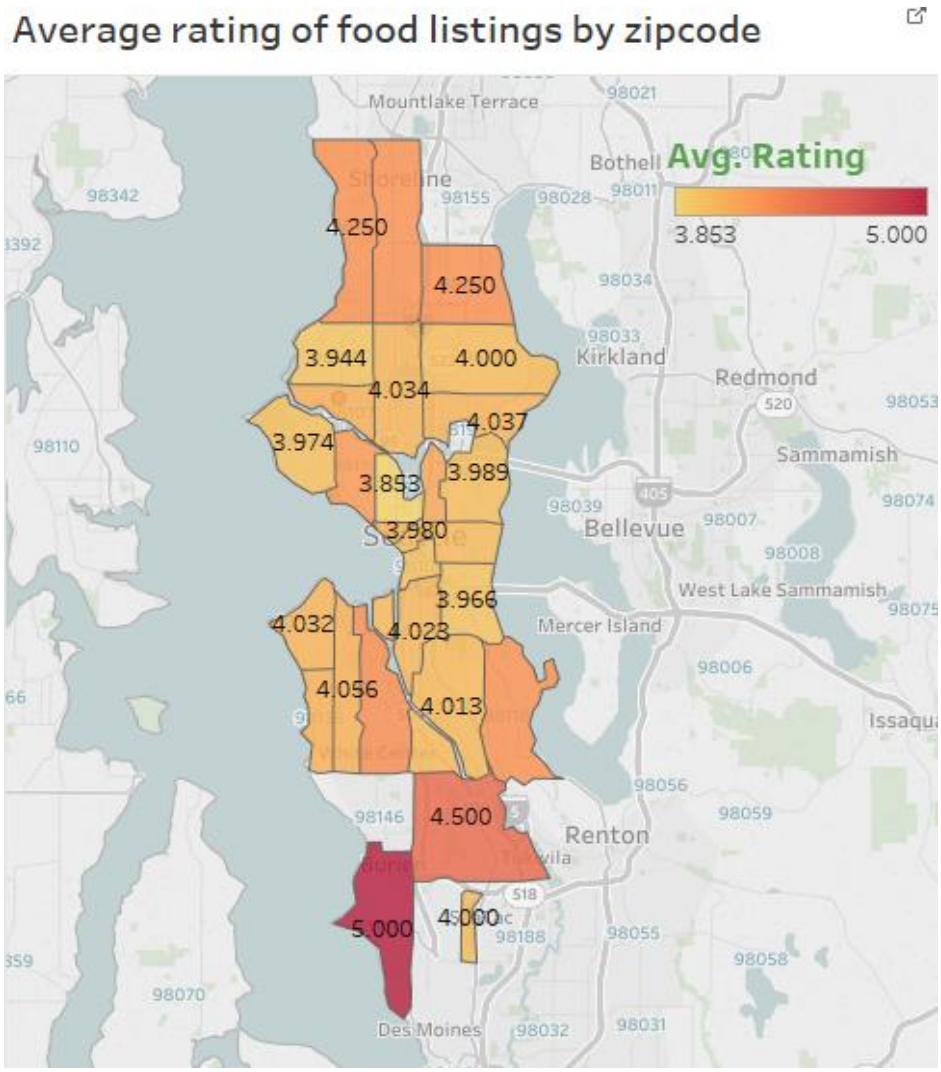
Similarly, based on the average rating, we found that the neighbourhood of Central Business District, Lower Queen Anne and Mount Baker had the lowest ratings as shown in the image below:

host_neighbourhood <fctr>	avg_rating <dbl>
Central Business District	40
Lower Queen Anne	40
Mount Baker	40
South Lake Union	40
Stevens	20

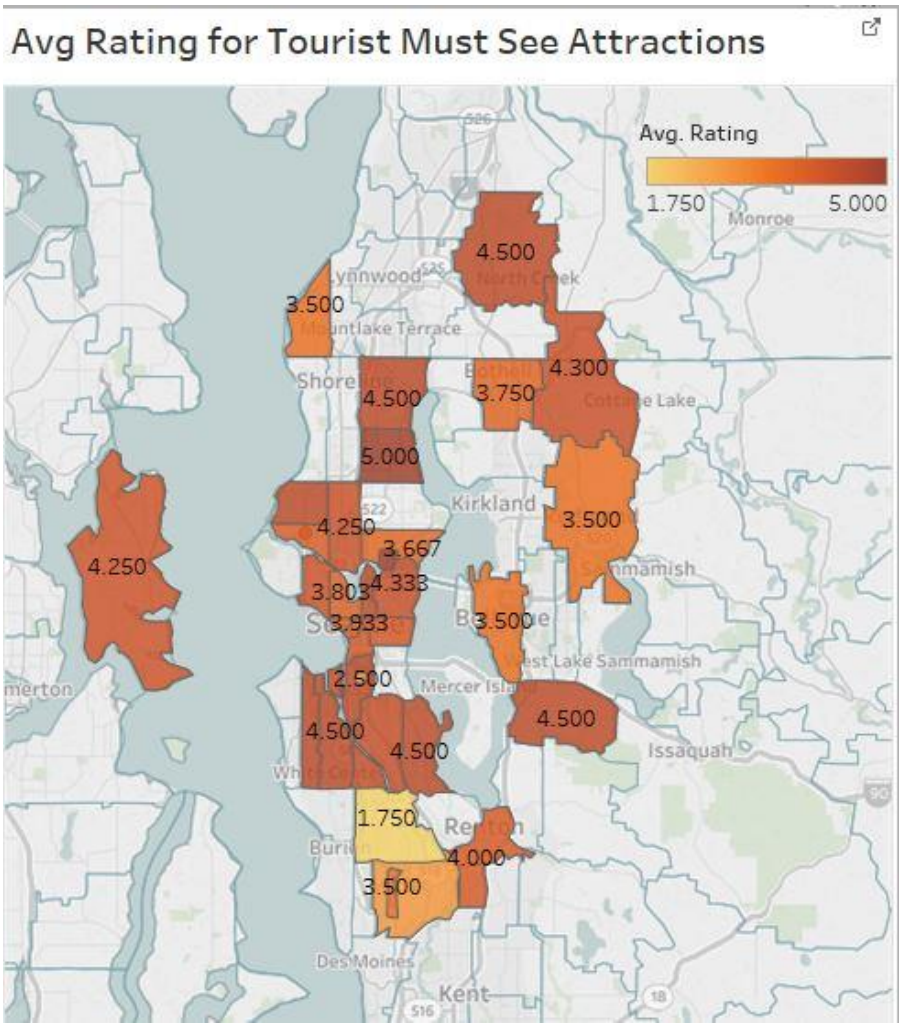
We extended our exploratory data analysis to the amenities we considered.
The areas with most number of nightlife options are the ones with zip code 98122, 98101 and 98103. Below is the screenshot of the same.

	zip	count_nightlife
	<int>	<int>
1	98122	102
2	98101	77
3	98103	77
4	98104	70
5	98107	60
6	98109	58
7	98121	41
8	98105	31
9	98102	28
10	98115	28

On the base of average rating, we found that Burien had the highest rating. This was validated with the fact that there is a Farmer's Market there.



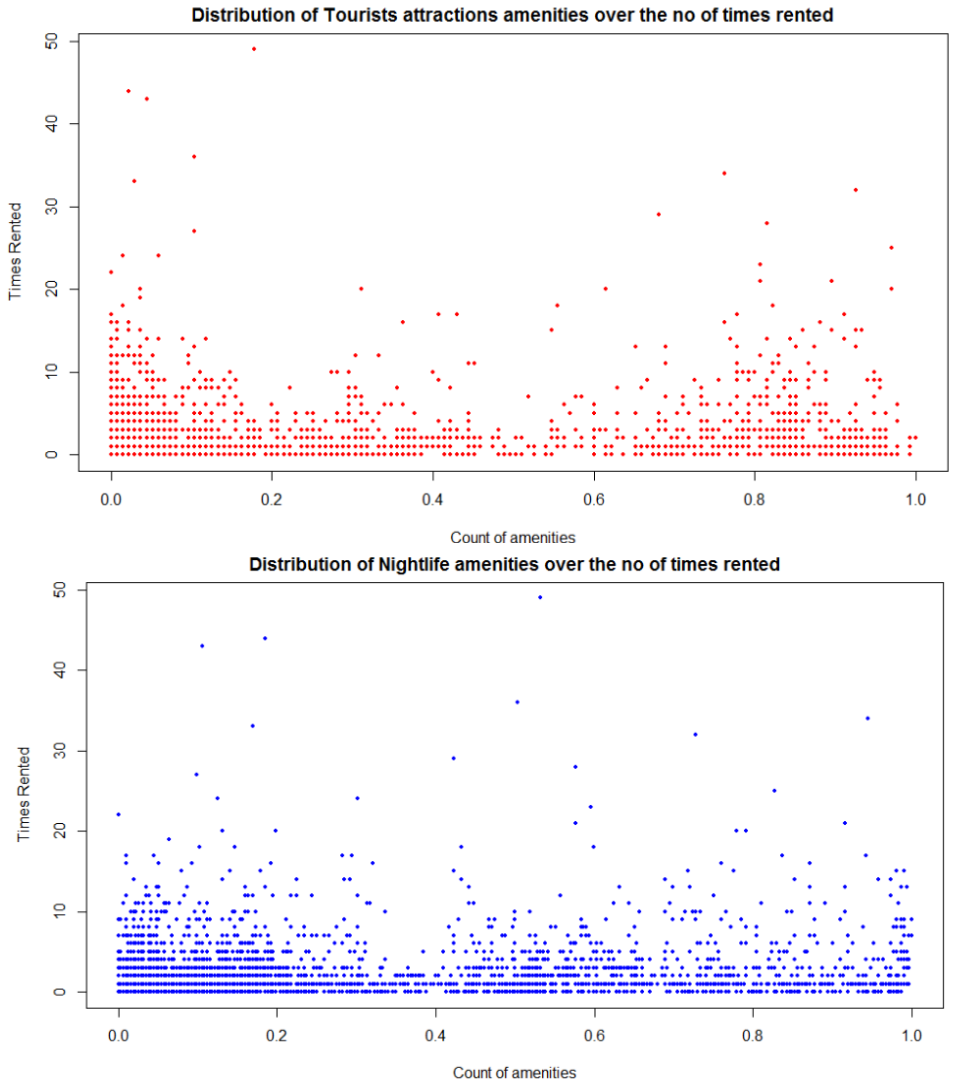
We also found that Bainbridge Island, Rainier Valley, Factoria, Downtown are some of the highly rated areas for tourist attractions.



Similarly, for pharmacies we found that the average rating was highest in the areas of 98008, 98177 and 98020 as shown below.

	zip	mean_rating
	<int>	<dbl>
1	98008	4.500
2	98177	4.500
3	98020	4.125
4	98119	4.125
5	98370	4.100
6	98070	4.000
7	98077	4.000
8	98110	4.000
9	98199	4.000
10	98027	3.900

An obvious test to explore our research question was to test for linear correlation relationship between number of any kind of amenities and number of times a particular listing was rented or review score rating. To our surprise, we observed a random distribution that was flat.



Multiple Linear Regression also did not provide any of strong statistically significant results.

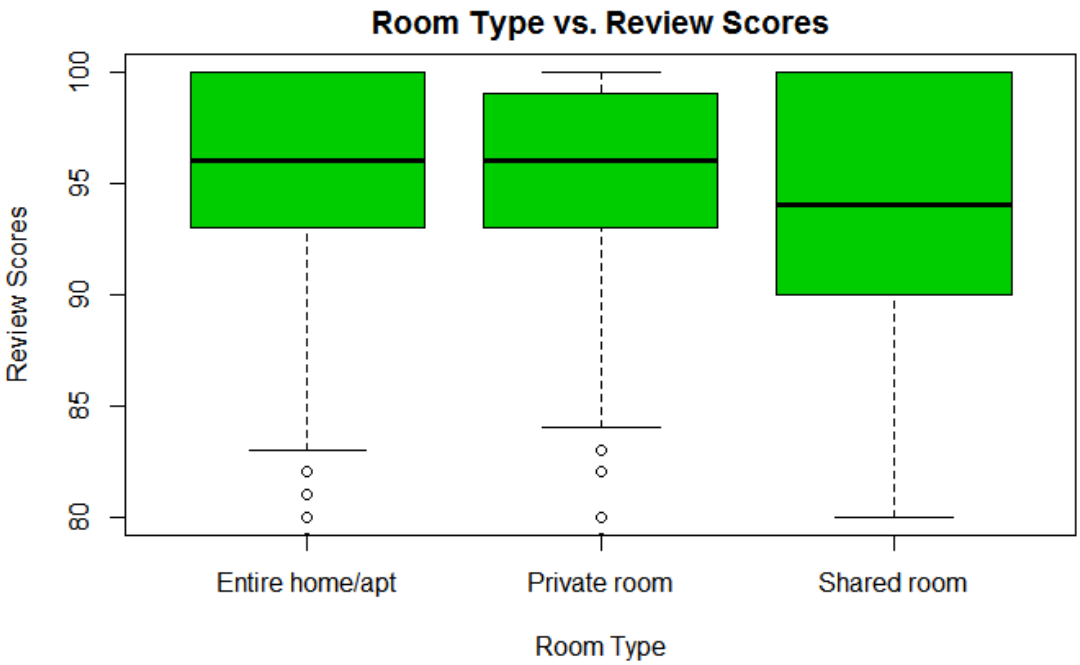
```
Call:
lm(formula = number_of_times_rented ~ Number_good_Nightlife +
    Number_good_food + Numer_Good_Tourist_Must_see + Number_good_pharmacy,
    data = listings_new)

Residuals:
   Min     1Q Median     3Q    Max
-4.247 -2.010 -1.101  0.737 46.511

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.118466   0.118740  17.841   <2e-16 ***
Number_good_Nightlife     -0.003564   0.005730  -0.622   0.5340
Number_good_food           0.006823   0.005098   1.338   0.1809
Numer_Good_Tourist_Must_see 0.010681   0.006829   1.564   0.1179
Number_good_pharmacy      -0.044396   0.021421  -2.073   0.0383 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.537 on 3155 degrees of freedom
Multiple R-squared:  0.02882,   Adjusted R-squared:  0.02759
F-statistic:  23.4 on 4 and 3155 DF,  p-value: < 2.2e-16
```

Since we were dealing with Geospatial Analysis, we wanted to consider Spatial Correlation between different covariates. So we performed Spatial Autocorrelation between listings coordinates and number of any kind amenities. The two listings that are adjacent to each other are bound to be clustered with an index of 0.4, when we take into consideration the number of food amenities. This is understandable because we are considering a mile radius and the food amenities present in adjacent listing's vicinity would overlap. Now, if we consider the correlation with respect to how frequently the listing was rented, the spatial distribution was random, this implies that the two listing which are close to each other does not exhibit a pattern on how many times it was rented in a spatial distribution. Thus could possibly explain why we did not observe any linear correlation between these variables.

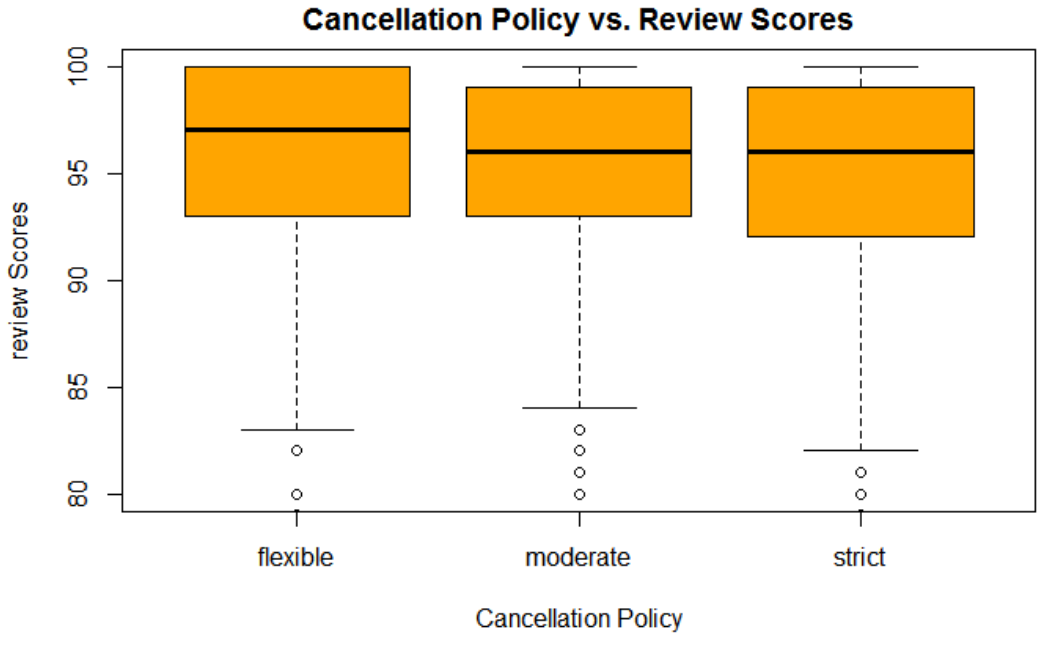
FURTHER ANALYSIS

As we got unexpected results from our analysis, we decided to delve deeper into the question "What other factors might have an effect on the review_scores of a listing? We first decided to look at the categorical variables. The apartment type and the cancellation policy. Both these categorical variables had 3 unique categories:

Apartment/Room type: Full apartment, Private room, Shared room
Cancellation Policy: Flexible, Moderate, Strict.

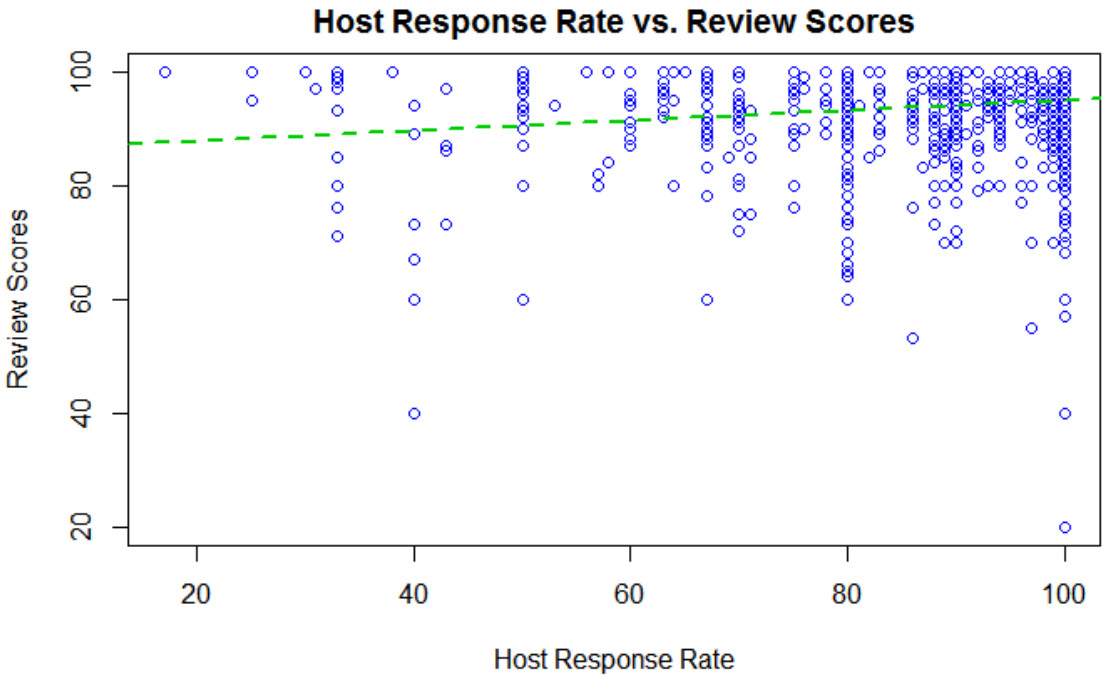


room_type <fctr>	avg_rating <dbl>
Entire home/apt	94.47765
Private room	94.80294
Shared room	93.22826

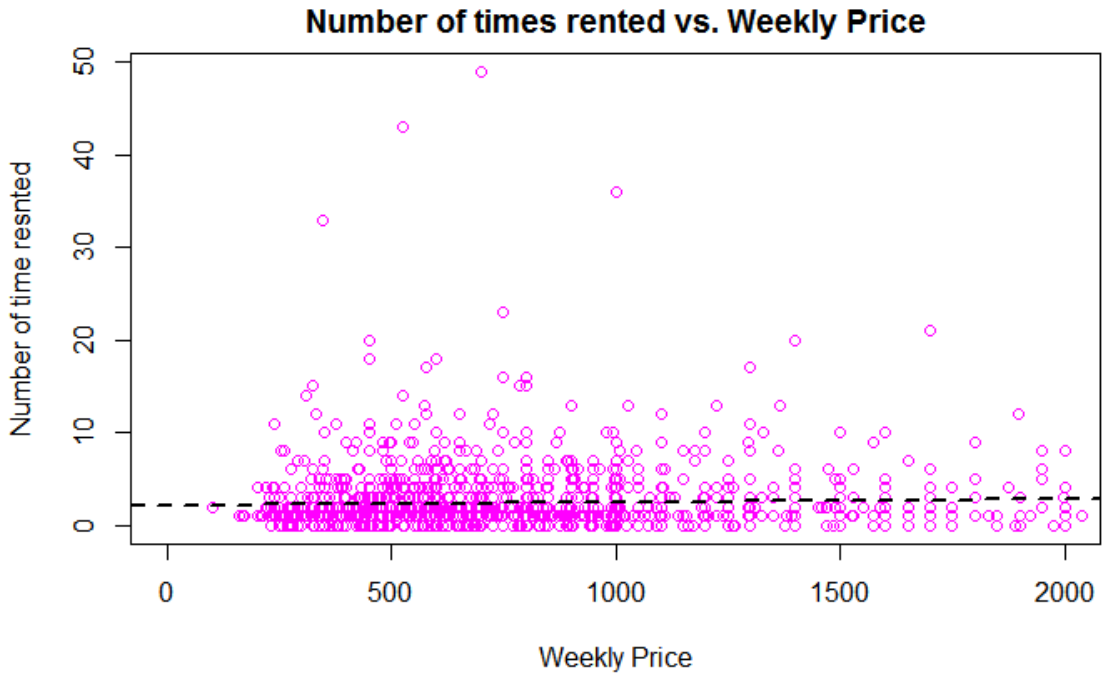


cancellation_policy <fctr>	avg_rating <dbl>
flexible	94.81026
moderate	94.67196
strict	94.25139

As evident in the graph, there is a slightly lower review score for a stricter cancellation policy as it is for flexible and moderate. Similarly, there is a slightly lower review score for a shared room than that for a private room or an entire apartment. So these could be possible predictor for the review score of a listing. Further, we also tried to examine the host_response_rate variable. It describes the amount of times a host responds to a user request out of 100.




```
Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                           85.89650    1.02388   83.89   <2e-16 ***
as.numeric(listings_new$host_response_rate)  0.09077    0.01068    8.50   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.44 on 2850 degrees of freedom
  (308 observations deleted due to missingness)
Multiple R-squared:  0.02472,   Adjusted R-squared:  0.02438
F-statistic: 72.24 on 1 and 2850 DF,  p-value: < 2.2e-16
```




```
Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        9.419e+01  2.815e-01 334.573   <2e-16 ***
as.numeric(listings_new$weekly_price) 5.407e-04  2.965e-04   1.824   0.0683 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.646 on 1785 degrees of freedom
  (1373 observations deleted due to missingness)
Multiple R-squared:  0.00186,   Adjusted R-squared:  0.001301
F-statistic: 3.326 on 1 and 1785 DF,  p-value: 0.06835
```

The trend followed a slightly upward trend with a statistical significant result which suggested that as the host response increases, the review score also increases. The response rate explains about 2% of the variance in the review score data. We also tried to analyse association between the number of times a listing was rented and its review score. We did not find a significant relationship as evident below. In addition, as our review scores were all clustered above 90, we discretized the review scores into grades.

Review Score	Review Grade
100	A
>98	B
>96	C
>94	D
>92	E
<92	F

After the discretization, we again delved deeper into the data to find possible correlations between the number of times a listing was rented and the review grade. We also tried to find possible associations between the review grade. As we can see, the patterns are fairly uniform

for all the grades, with some outliers giving us a direction to look further into the listing parameters. We did not go ahead and remove the outliers for this very reason. They gave us a direction to look particularly into those listings.

THE SURVEY

We were so intrigued with our results, we did a small oral survey. We asked 10 people who've used Airbnb. What makes them rent it? The most frequent response was – Price! They argued that if I wanted food around me, I'd have booked a hotel. There's a reason people book Airbnb, and that is Price. Another frequent response we got was - Why walking distance? The respondents explained that in a city like Seattle where public transport is so good, they don't feel there's a need of anything being at walking distance. This gave us a further explanation of the fact that our findings were right.

TEXT MINING AND SENTIMENT ANALYSIS

A part of Airbnb dataset comprised of reviews for the listings. There were 84,636 reviews. We explored them by preparing Word Clouds, finding Text Mining associations and performed some sentimental analysis as well. Considering the number of reviews, it was difficult to perform the text mining operations on all reviews at once. Instead, we iterated the text mining process by breaking our reviews dataset by grouping them into according to the zipcodes.

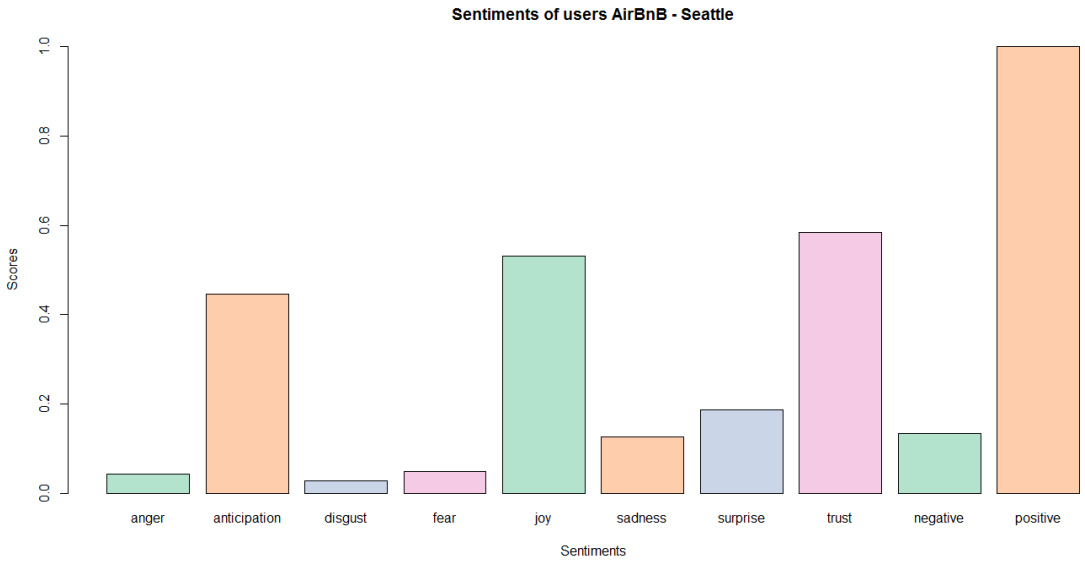
A sample of the word cloud from our study is shown below:

While performing our study we found several interesting text associations that differed across zipcodes.

98119 - Queen Ann Hill “Access” -> “networkpassword”, “downtown”	98101 - Downtown “Walk” -> “distance”, “pike”, “market”, “needle”
98105 - University District “stay” -> “comfortable”, “beds”	98121 - BellTown “Location”,”place” -> “pike”, “market”, “needle”, “walk”

For example for zip code 98119 - Queen Ann Hill, we found associations between “access” and “networkpassword” and “downtown”. Users often tend to mention about access to WiFi networks and accessibility to Downtown from Queen Ann Hill. There was associations between words “walk” and “pike”, “market” and “needle”for the zip code 98101 - Downtown. Thus, people probably often mentioned about the walking distance from listings to places like Pike Place Market and Space Needle. We observed same pattern in the zip code 98121 - Belltown which is adjacent to Downtown. Patterns from reviews in University District - 98105 were intriguing. Users who stayed in this part of Seattle often mentioned about “stay” and “comfortable” and “beds”. This could probably mean that users often mentioned about comfortable the beds were during their stay.

We observed a general trend of positive feedback across all zip codes. The following bar plot shows a normalized for the general trend of sentiments.



FUTURE PROSPECTS

We wish to keep working with this dataset. We wish to use predictive modelling against another outcome variable, for example price and also consider certain other independent variables like availability of transport. We would also like to explore other Airbnb listings of cities like Boston. We would further like to extend our study to country and continent levels.

CONCLUSION AND LESSONS LEARNT

There were a good amount of ups and downs throughout the project in terms of findings. However, we learnt a lot in the process. We spent 90% of our time wrangling and cleaning our data and that was the most fun part. We learnt that there should never be any assumptions made about the data. If we have to, we should explicitly state them. Finally, we realized we loved data for a reason – it often proves us wrong, and that's what keeps us going.

REFERENCES

Airbnb. (2016). *Inside Airbnb*. Retrieved from insideairbnb.com: http://insideairbnb.com/get-the-data.html

Spatial Autocorrelation and Moran's I in GIS. (2017, March). Retrieved from gisgeography.com: http://gisgeography.com/spatial-autocorrelation-moran-i-gis/

Yelp. (2016). Retrieved from yelp.com: https://www.yelp.com/developers

Spatial Autocorrelation. Retrieved from YouTube:https://www.youtube.com/watch?v=M9ecMxVG6jQ