Seattle Airbnb Data

By: Group 1 (Nikita Adhikari, Aabha Desai, Rong He, Shraddha Kadam, Vineela Reddy Karnatakam, Lakshit Rajput

Executive Summary

Airbnb is a large company that is in more than 191 countries. It is only continuing to grow in the future. The business success of airbnb is impressive. This project revealed some secrets behind the success. More guests and hosts have used Airbnb to travel in a personalized way. This project allowed different methods of data visualization to express these interesting insights efficiently.

- **INDEX**

- **Data Description**

The Seattle Airbnb dataset provides 4107 observations with 72 features/variables including host, property types, neighborhood, prices and review scores. This dataset solely represents Airbnb details of the Seattle listings.

- **Data Cleaning**

After analyzing the dataset, we decided to keep only the variables that would help us with our analysis.

We used Python for further data cleaning and analysis:

We read the data and stored it into a dataframe. For data exploration, we used various functions like head(), tail(), describe(), and info() to get a better understanding of the data-

```
In [4]: df.describe()
```

Out[4]:

	id	host_id	host_total_listings_count	latitude	longitude	accommodates	bathrooms	bedrooms	beds	price	...
count	4.197000e+03	4.197000e+03	4193.000000	4197.000000	4197.000000	4197.000000	0.0	3619.000000	4177.000000	4197.000000	...
mean	2.750170e+07	8.494779e+07	133.539232	47.624849	-122.334593	3.659757	NaN	1.568942	1.864017	127.375745	...
std	1.514280e+07	1.036887e+08	365.787002	0.046859	0.033056	2.254967	NaN	0.906574	1.367798	111.628517	...
min	2.318000e+03	2.536000e+03	0.000000	47.496210	-122.418760	0.000000	NaN	1.000000	0.000000	0.000000	...
25%	1.490220e+07	8.607308e+06	1.000000	47.604570	-122.356340	2.000000	NaN	1.000000	1.000000	75.000000	...
50%	2.792874e+07	3.983446e+07	2.000000	47.620930	-122.334570	3.000000	NaN	1.000000	1.000000	100.000000	...
75%	4.224708e+07	1.187390e+08	10.000000	47.660090	-122.310960	4.000000	NaN	2.000000	2.000000	145.000000	...
max	4.825671e+07	3.888968e+08	1398.000000	47.733530	-122.245540	16.000000	NaN	7.000000	19.000000	3212.000000	...

8 rows × 26 columns

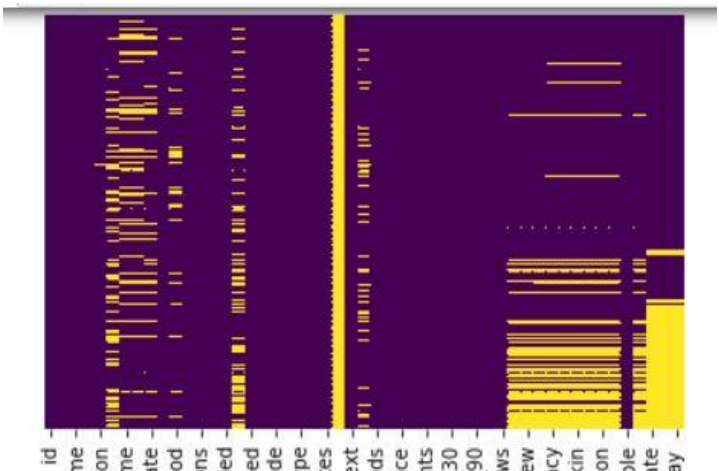
```
In [5]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4197 entries, 0 to 4196
Data columns (total 51 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   id                            4197 non-null   int64
 1   host_id                       4197 non-null   int64
 2   host_name                     4193 non-null   object
 3   host_since                    4193 non-null   object
 4   host_location                 4187 non-null   object
 5   host_about                    3103 non-null   object
 6   host_response_time            3523 non-null   object
```

We identified the columns having missing values-

```
In [6]: df.isnull().sum()
```

```
Out[6]: id                              0
        host_id                         0
        host_name                       4
        host_since                      4
        host_location                  10
        host_about                   1094
        host_response_time            674
        host_response_rate            674
        host_acceptance_rate          498
        host_is_superhost               4
        host_neighbourhood            387
        host_total_listings_count       4
        host_verifications              0
        host_has_profile_pic            4
        host_identity_verified          4
        neighbourhood                1197
        neighbourhood_cleansed          0
        neighbourhood_group_cleansed    0
        latitude                        0
        longitude                       0
        property_type                   0
        room_type                       0
        accommodates                    0
```



Imputation of missing values:

Handling missing values of categorical features:

We replaced the missing values with MODE (value that appears most frequently)

```
#since it has string values, we will consider mode
df['host_name'].mode(dropna=True)
```

```
0    Kia
dtype: object
```

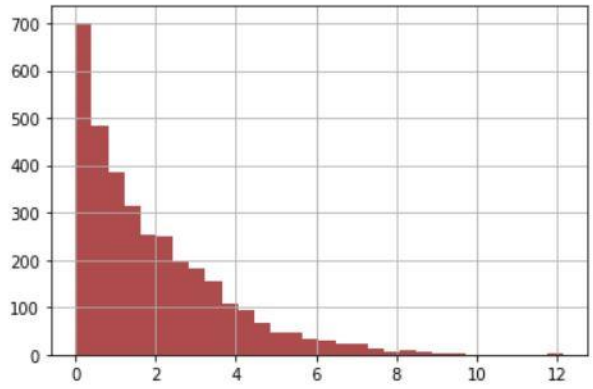
```
#replacing with Kia
df['host_name'].replace(np.NaN, 'Kia',inplace = True)
```

Handling missing values of numerical features:

We checked the distribution of values and found out that the graph was right skewed and hence replaced the missing values with median.

```
df['reviews_per_month'].dropna().hist(bins=30, color='darkred',alpha=0.7)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x294af777470>
```




```
df['reviews_per_month'].median()
```

```
1.41
```

```
df['reviews_per_month'].replace(np.NaN, 1.41,inplace = True)
```

After imputing all of the missing values, we converted the data frame into a csv and used it for visualizations in Tableau.

Converting dataframe to csv

```
df.to_csv('Seattle_AirBnB_cleaned.csv')
```

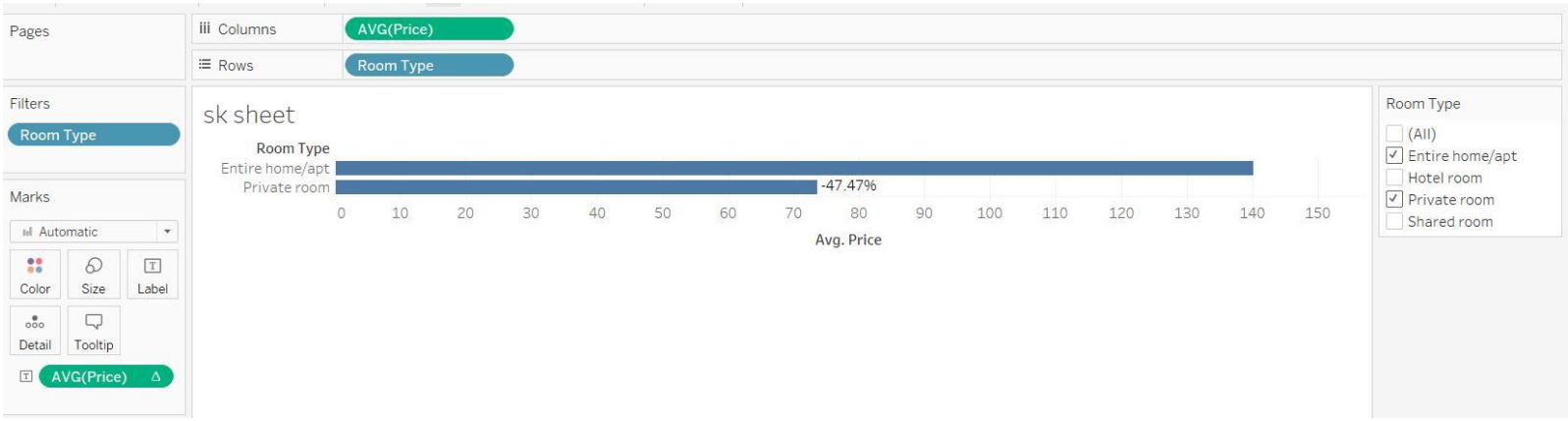
- ## General Introduction

  Seattle is home to the biggest industry and tech titans such as Amazon, Microsoft, Expedia, and Starbucks. With numerous festivals, ecological diversity and urban attractions, Seattle is a place worth visiting and a prime spot for Airbnb business. In this report we have analyzed listing data sources from insideairbnb.com to better understand features that impact Airbnb listings.

- ## Insights and findings

Hypothesis 1: Entire Apartment/ Home costs more than Private Room

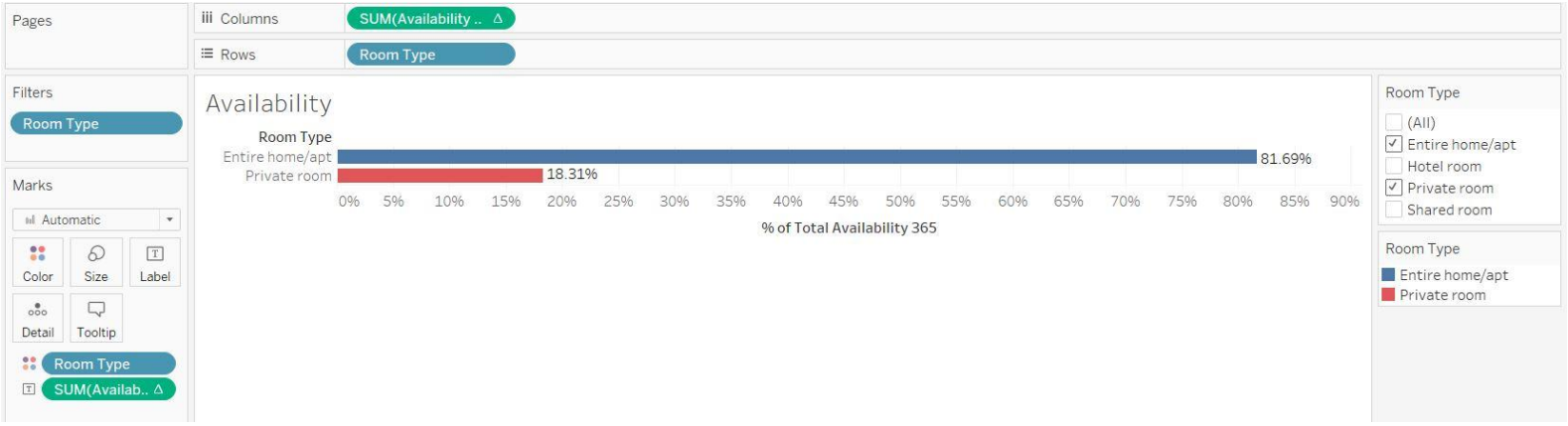
In order to come up with insights related to the hypothesis, we first compared the average price of both room types by plotting a bar chart and using quick table calculation to view the difference between their average prices. Results showed that the average price of the entire apartment / home is 47.47% greater than the average price of the private room. This finding gives us an overall idea that the entire apartment / home is more pricey.



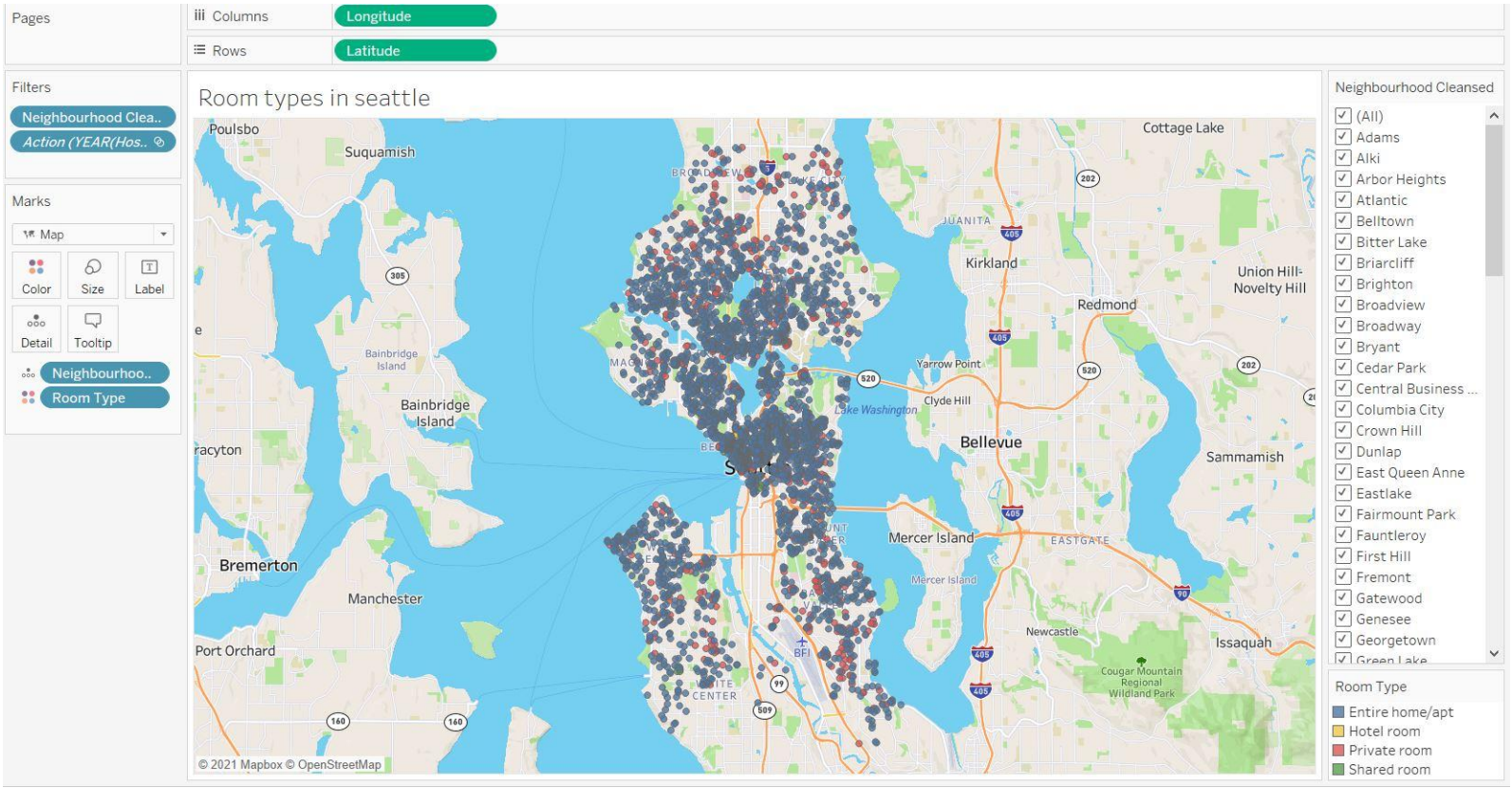
To get a clear view, we visualized the running sum of average price of all room types from the years 2008 to 2021 with the help of a circle chart and using pages to get an animated visualization of the changes occurring over time.

For further analysis, we checked the availability of these two room types by again using a bar chart, filtered for room types and utilized quick table calculation to see the percentage of total availability which turned out to be 81.69% for the entire apartment / home and 18.31% for private rooms.

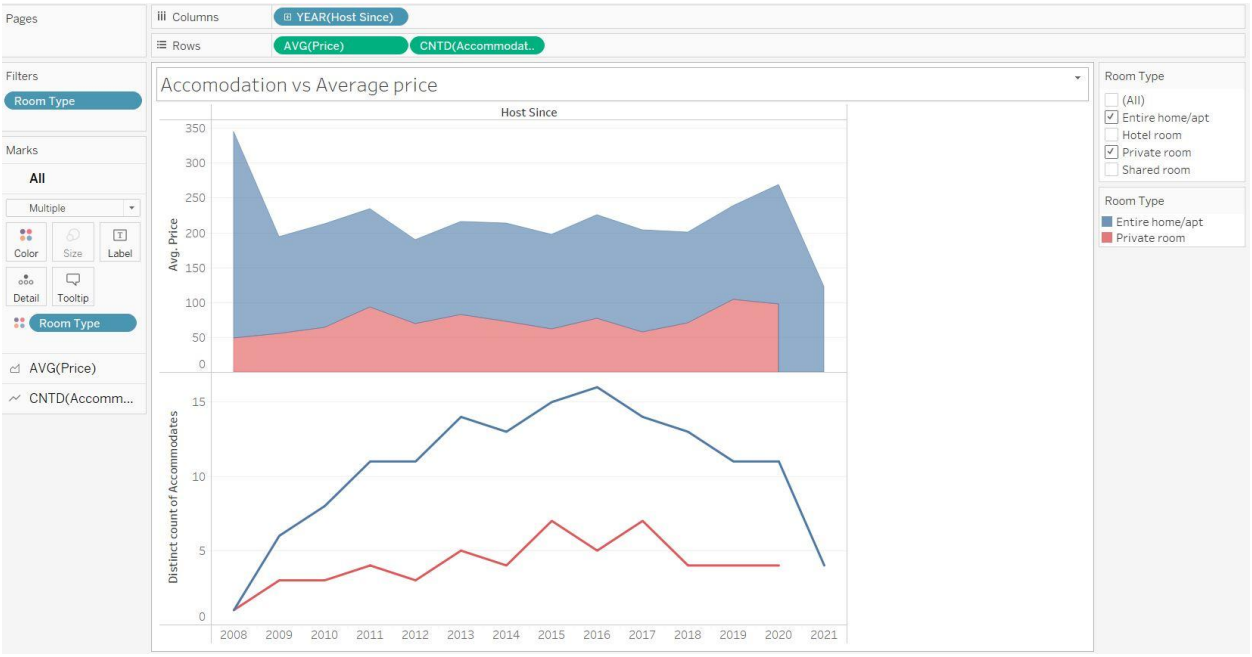
This chart tells us that the entire apartment / home has more availability than a private room.



Furthermore, we also looked into listings of the two room types by utilizing map chart and graphically visualizing the spread of the different room types. Here in this chart we could see that there are maximum listings of the entire apartment / home.

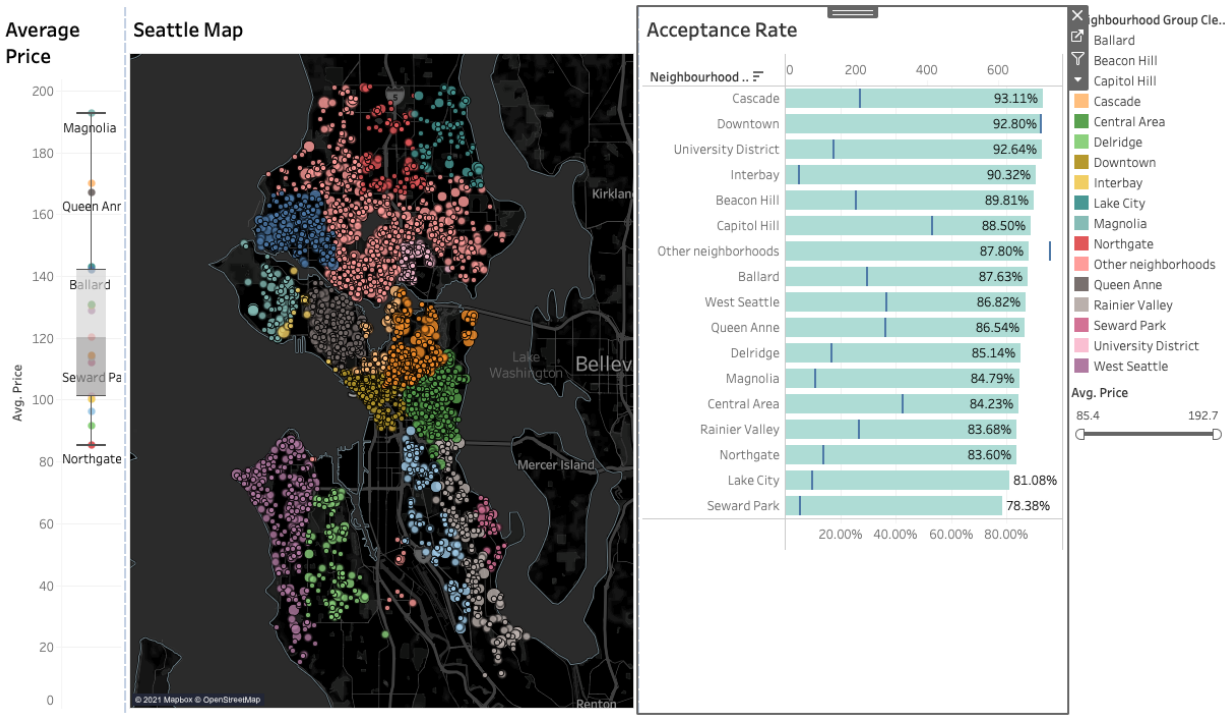


Thereafter, we explored accommodation and compared it with the average price of both of the room types from the years 2008 to 2021 by plotting an area chart for the average price and line chart for accommodation. An interesting insight was found that, although the price of the entire apartment / home is more, it allows a maximum number of guests (max 16) than that allowed by a private room (max 7).



So, from the above insight we can give a recommendation that if the customer has a group of less than or equal to 7 people and is trying to book AirBnB in Seattle, then they should look for a private room, otherwise they should go ahead with the entire apartment / home.

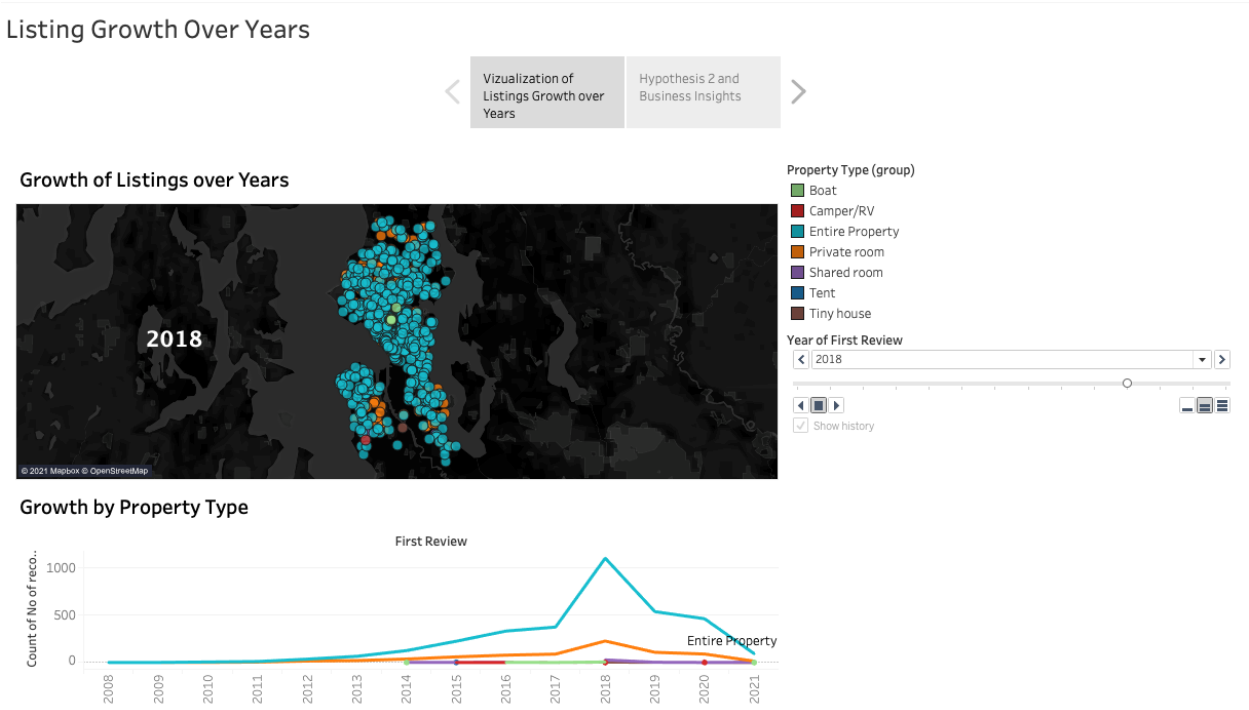
Additionally, for customers, it is easier to have a quick view of something while making a decision. When it comes to the Seattle data, if there is a general view of a chart showing location, with prices and the host acceptance, it would be easier for someone trying to get an Airbnb. That is what this chart is below:



This dashboard provides a general understanding of the different locations. Box and whisker plot, bar chart and maps were used to create this dashboard. Customers could hover over or even click on a specific region to figure out the number of bedrooms the property offers, the neighbourhood it is in, host acceptance rate, and even its average price.

Hypothesis 2: Airbnb has constant upward growth.

To look into this hypothesis, we analyzed the growth aspect of Airbnb. The chart below mentions the listings gro

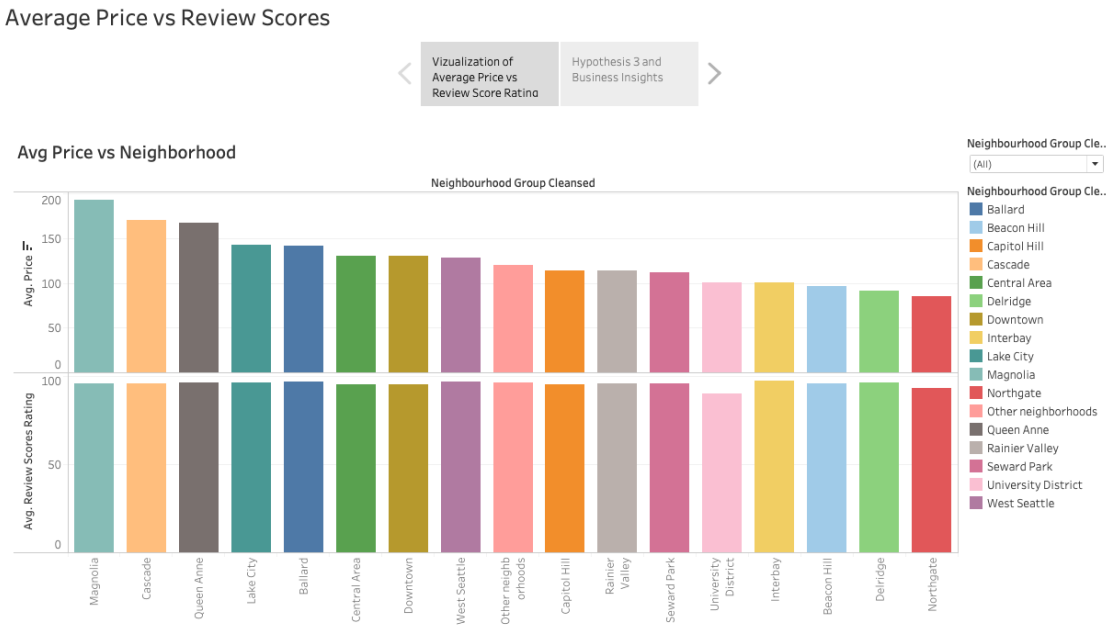


wth since 2008.

This dashboard provides a dynamic view of the increase in the first review left by the customers. Maps and line charts were being used in this dashboard. Looking at the trend of Airbnb growth, rather than having an upward trend, 2018 was the turning point for Airbnb. According to the article posted by Vox, in 2018, airbnb expanded to more locations. That same year, rather than focusing on profits, they also started investing more on efforts such as strong airbnb security, tech and administrative cost, and acquisitions (Molla, 2020). Their rising expenses overtook the revenue they were creating. Also, Airbnb's revenue took a massive hit from quarantines and lockdowns throughout 2020. As it can be seen from the dashboard, when it comes to the most popular property type, the entire property was the most popular. Private room was next in line. There are some things that can be inferred from the dynamic graph growth by property type. Year of first review determines the new property added. It is being inferred that the new review means brand new listings arising. First reviews are higher in 2018 and that was the year Airbnb expanded to 1000 new experiences/destinations.

Hypothesis 3: If the review scores are better, the price of an Airbnb would be more.

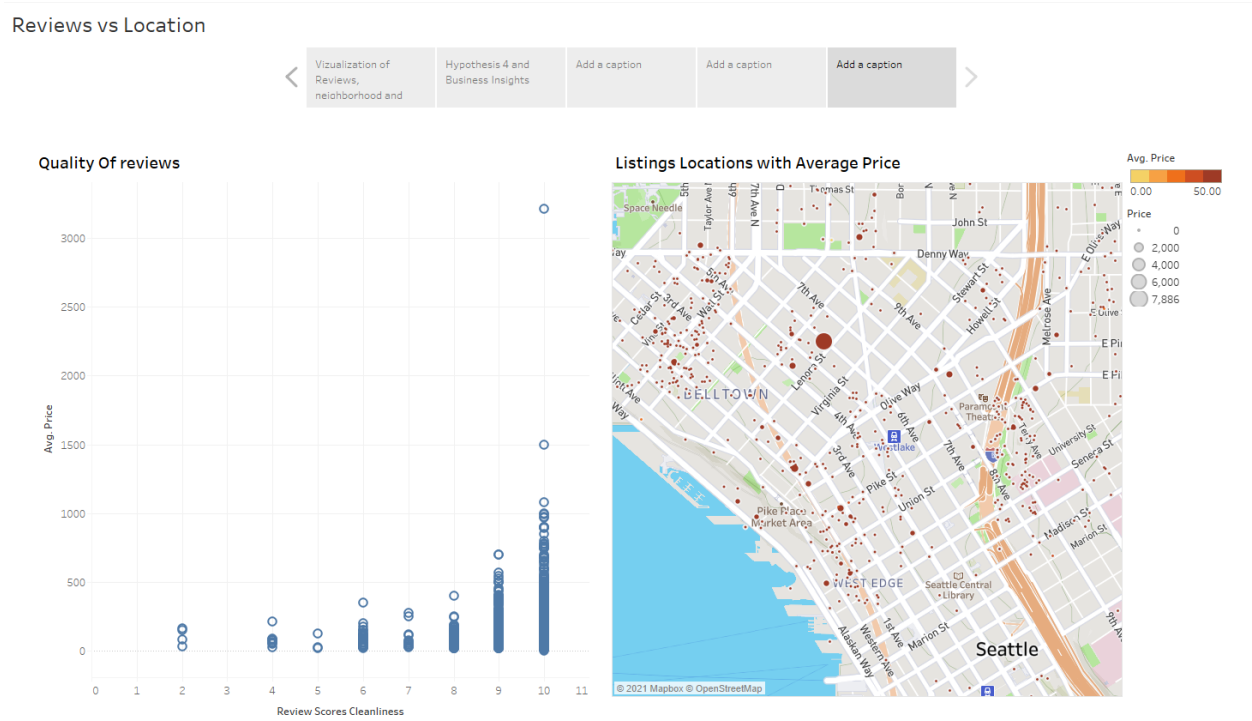
It was important to understand if higher prices correlated with good reviews. The chart below is being considered to look into this hypothesis.



The bar chart Price vs. Neighbourhood proves that higher review scores does not mean higher price. If looked into the average price for neighbourhood group Magnolia, it has a higher price than Interbay. Interbay has higher reviews, it has lower prices compared to Magnolia. It can be seen from the chart itself that there is no positive or negative relationship. There is no specific trend in the relationship between review and price. If the neighbourhood group has a higher price, it could have a lower rating and if it has a higher rating, it could have a lower price as well.

Hypothesis4: Average price is directly related to the cleanliness score or review score.

To validate this hypothesis, a scatter plot between the Avg Price and Review score cleanliness is shown below

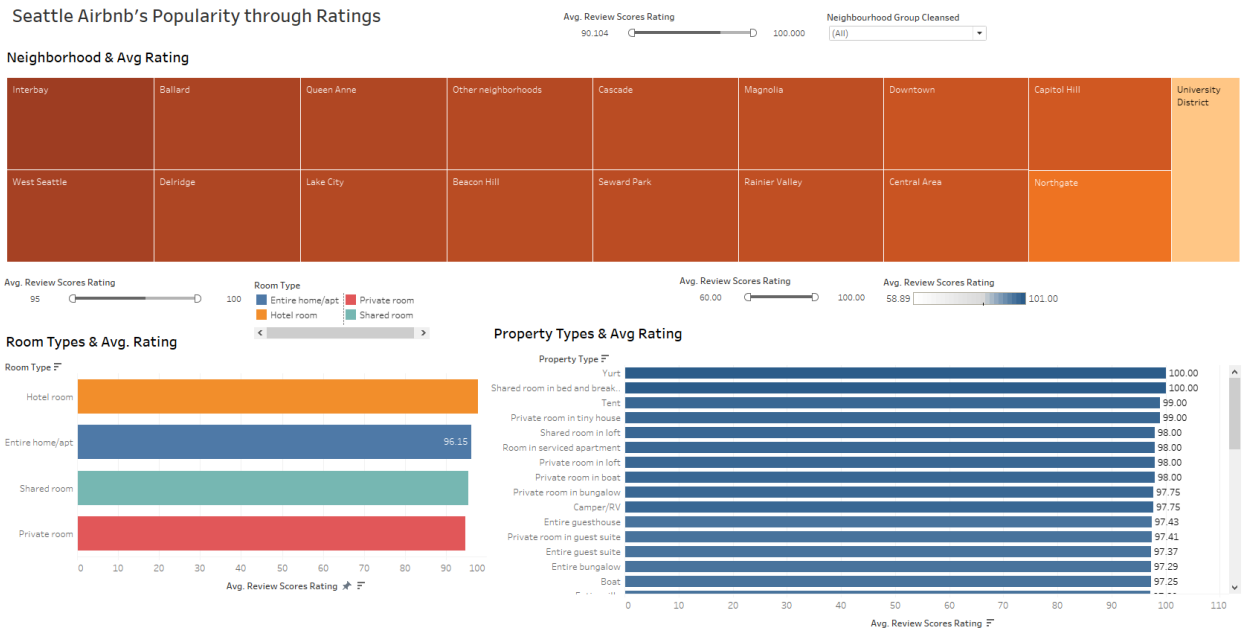


As the average price increases the cleanliness score also increases means the hotel which has a greater price that has a review score around 10. So from the graph we can say that our hypothesis is correct.

Hypothesis 5: Neighbourhood near popular Tourists Places such as Bell Town, Queen Ann have Highest Rating.

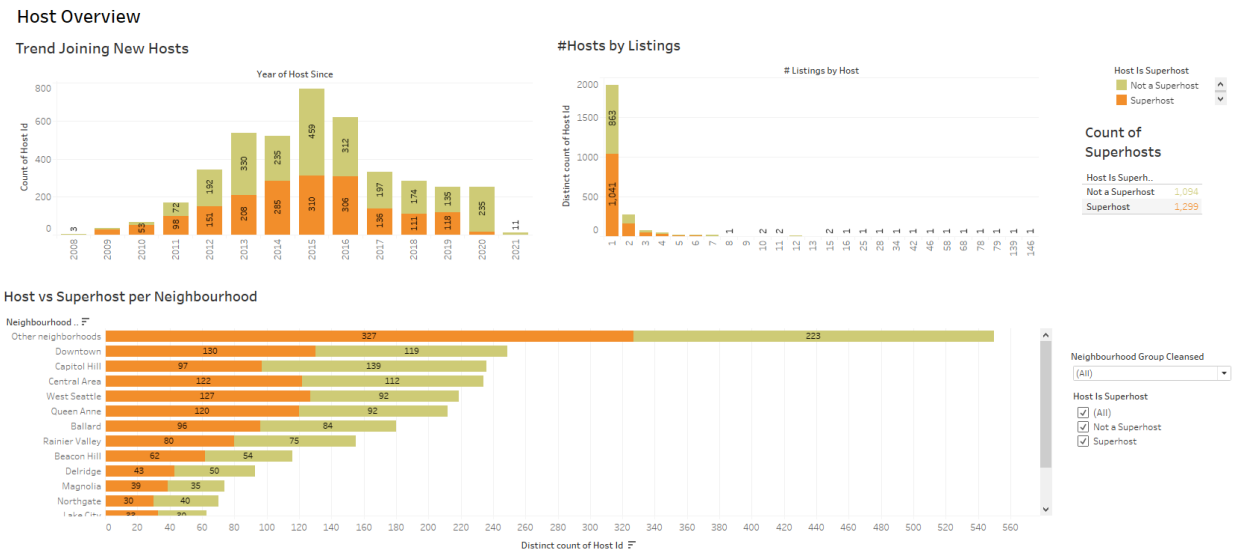
To validate this hypothesis, a heatmap of average rating and neighborhood was plotted along with the bar plots of Room Type vs. Average Rating and Property Type vs. Average Rating shown

below.



Seattle's popular tourist places like space needle, seattle center are located near neighborhoods such as Bell Town and Queen Ann. We expected to see higher average ratings in these neighborhoods but through our visualization analysis, it was found that the neighborhoods such as Interbay, West Seattle, and Ballard have higher average ratings. We also found that among the various property and room types Yurts, Private rooms in tiny houses and Hotel Rooms have the highest rating.

Hypothesis 6: More than 75% of Hosts are Superhosts.



Superhosts are experienced hosts who provide a shining example for other hosts, and provide extraordinary experiences for their guests. Superhosts are supposed to have 4.8% overall rating out of 5, at least 10+ stays in the previous year, less than 1% cancellation rate and 90% & above response rate as per Airbnb Superhosts Program. Superhosts get several benefits such as significant increase in their earnings as they get extra 20% on top of usual bonus and attract more guests.
In this visualization we explored the trend of joining new hosts, host vs superhost in each neighborhood and no of hosts per listing. We found that only 54% are superhosts as against our hypothesis of having more than 75% superhosts. So many hosts have the scope of working in a direction to become superhosts. We also found that there is a rise in the number of hosts through 2015 followed by a decline. Neighborhoods like downtown and capitol hill have more superhosts than the other neighborhoods. We also found that 1041 superhosts out of total 1299 superhosts have only one listing on airbnb.

Hypothesis 7: Superhosts have only 1 listing, which helps them get better ratings.

To validate this hypothesis we explored review ratings for hosts for various categories like Accuracy rating which tells how accurate the listing description is, Check-in rating which gives idea about the smooth check-in process, Cleanliness rating, Location rating, communication rating and rating for value for money. We found that overall superhosts have better ratings in all these categories as compared to normal hosts.

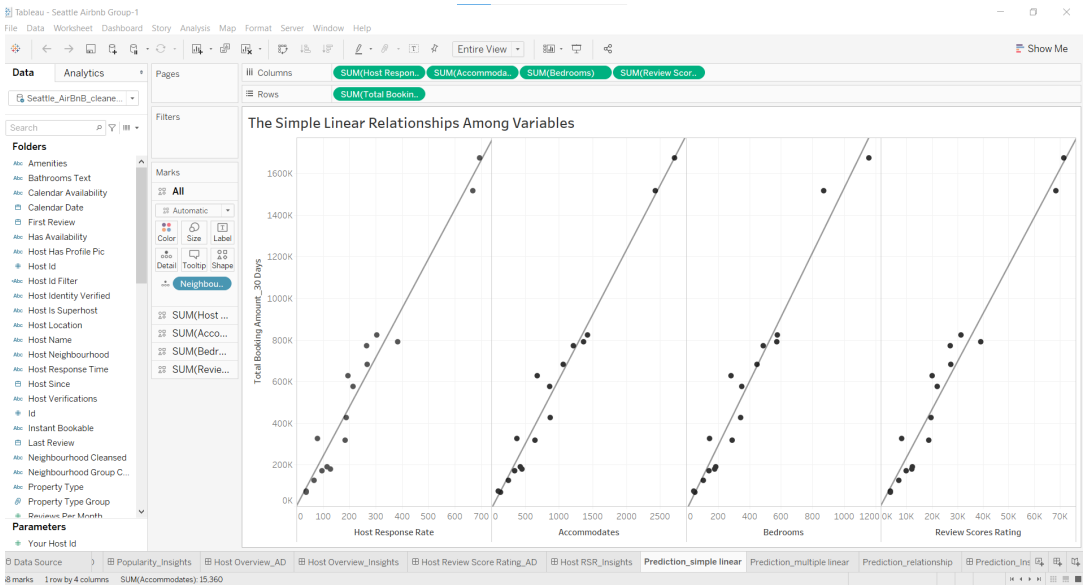
We also plotted average response rate vs. review score rating to validate if all the superhosts have more than 90% of response rate. We found that except one or two superhosts all have more than 90% rating.
We found that the average listing per superhost is just one, whereas it is 2 per normal hosts. This proves our hypothesis to be true.

Hypothesis 8: The total booking amount in 30 days can be predicted by other independent variables like Review Score Rating or Bedrooms.

To confirm this hypothesis, the model of multiple linear regression is the good solution. Despite being an excellent tool to efficiently visualize the data, Tableau can also be used to create and verify linear regression models used for predictive analytics.

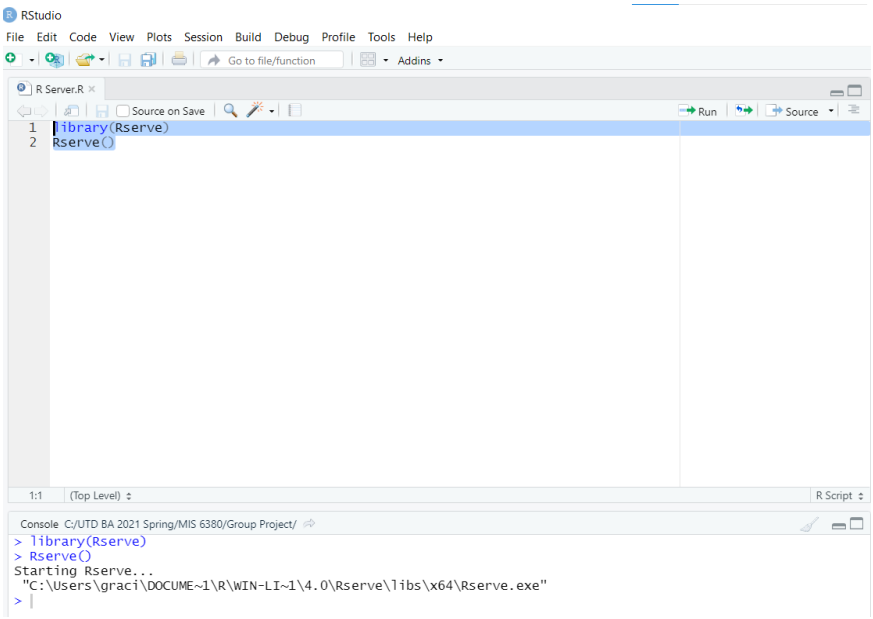
Tableau can develop the single regression model by itself. At first, we will analyze the relationships among these variables via scatter plots. The new measure "Total Booking Amount_30 Days" is created by "Creating Calculated Field" as below.

Then we could make scatter plots with the trend line as below. We could find that the "Total Booking Amount _30 Days" had great linear relationships with other variables. So these variables are good predictors of total booking amount.

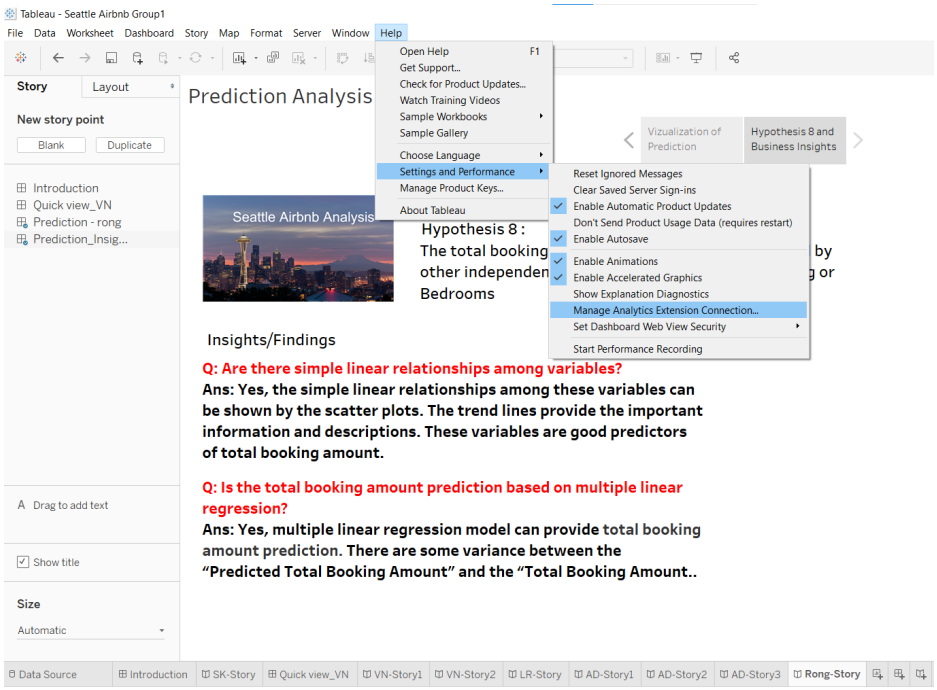


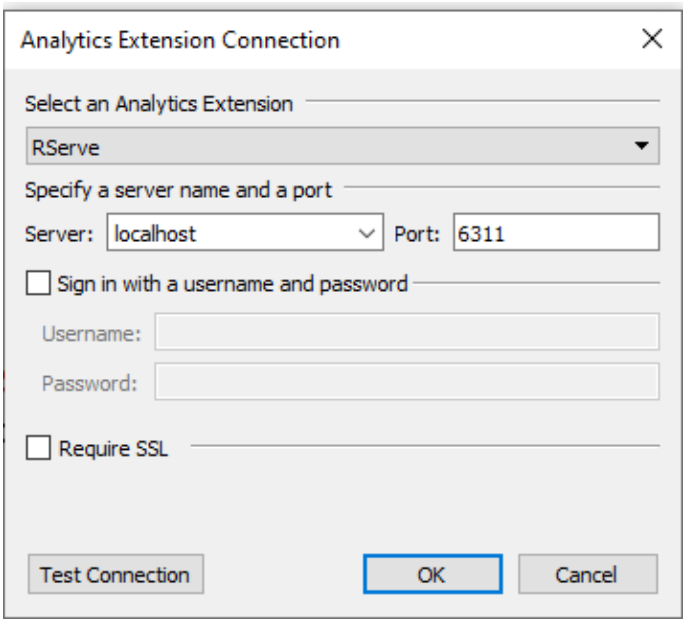
Secondly, we considered that Tableau integrates with external statistical languages R and uses the multiple linear regression model.

Open R Studio and run below commands



In Tableau, go to Help -> Settings and Performance -> Manage analytics extension connection -> Select Rserve as analytics connection and localhost as server. This will connect R to Tableau.

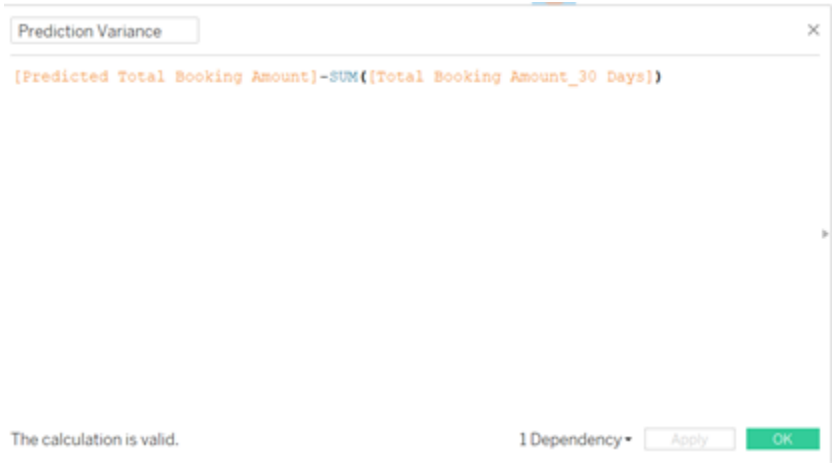




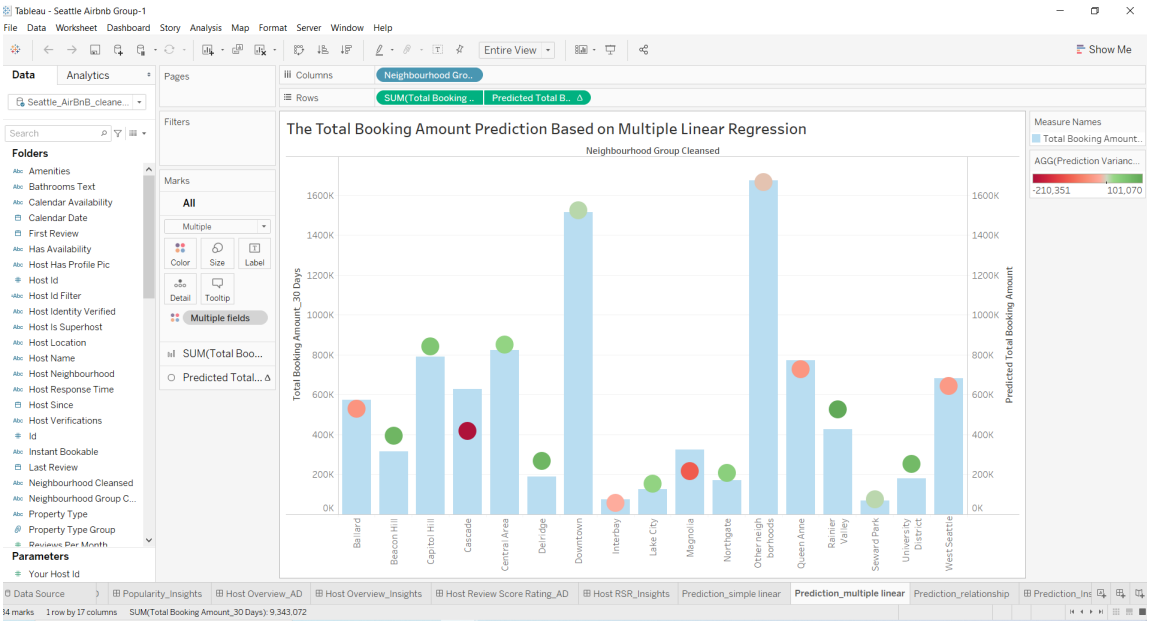
A new calculated field called "Predicted Total Booking Amount" is set. Tableau's SCRIPT_REAL function could be used to embed R code in Tableau's calculation.

We have used the above variables “Host Response Rate”, “Accommodates”, “Bedrooms”, and “Review Score Rating” to predict Total Booking Amount.

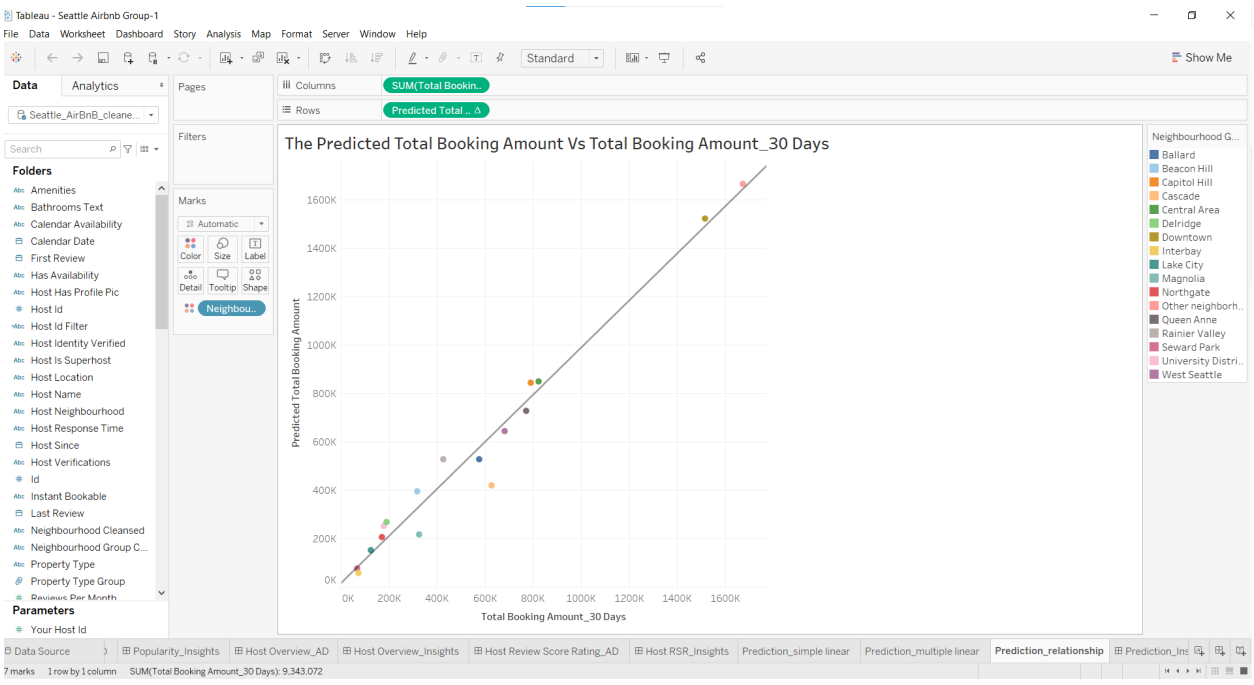
Moreover, we added another new calculated field called “Prediction Variance”.



In the sheet, we can see that based on different “Neighborhood Group”, there are some variance between the “Predicted Total Booking Amount” which are red - green dots and the “Total Booking Amount_30 Days” which are blue bars because of the accuracy of multiple linear regression model.



Finally, we also used the scatter plot to check the relationship between “Total Booking Amount_30 Days” and “Predicted Total Booking Amount”. There is a straight line which shows that the predicted values have a positive relationship with the true values. The total booking amount in 30 days could be predicted by other independent variables like “Host Response Rate”, “Accommodates”, “Bedrooms”, and “Review Score Rating”.



- ## Conclusion

In this project, Seattle's Airbnb data was analyzed using complex and dynamic visualizations through a business intelligence tool called Tableau. To support Tableau visualization, R was also slightly incorporated. When it came to the data cleansing aspect, Python was used fully. Some of the insights included factors such as specific patterns, relationship between two attributes, and even prediction of future bookings. From looking at average price and room type, listing growth over years, average price vs. review score rating, to host review score rating, this project led to a deeper understanding of the Seattle Airbnb listings.

Moving forward, some other scope for analysis would be looking at Airbnb dataset with their latest investments into security and how it is affecting their profit, providing pricing recommendations to owners, and maybe even recommending a new place to invest in.

- ## Citations

https://www.vox.com/2020/2/12/21134477/airbnb-loss-profit-ipo-safety-tech-marketing

https://insights.ehotelier.com/insights/2018/03/14/airbnbs-millennials-strategy-turning-point/

https://sprintmilestone.com/blog/2017/10/17/230/

https://www.youtube.com/watch?v=dgArfzvuAlw