Airbnb Project

Lessa Fleming, Ashley Silverstein, Jennifer Yun

IST 652 Final Project Report

Introduction

This project focused on two Airbnb data sets from Kaggle. The first one, called 'AB_US.cvs', focuses on Airbnb data collected throughout different cities in the United States of America. The other data set focused on Airbnb prices in European cities. Each data set has an average of 15 columns. For this project, our primary interests are daily rates, locations, and customer satisfaction ratings or reviews. We wanted to learn about correlations between location, price, and reviews. At the beginning of the project, there was a third data set to pull in that had train station data for Europe that was to be used to compare distance in the Europe data set, but this was unsuccessful in finding a join that would work for this data set and removed this from the list.

Describe the Data

The first data set AB_Us did not need manipulation before it was loaded, while the second data set had a large folder with multiple different EU data. This data set was extremely large so for time constraints a sample was created with three of the files in the data folder. This was labeled Combined Europe for a sampling of the data for analysis.

The data was read into Jupyter and then separated into two data frames. The data frame called 'df' uses the 'AB_US' dataset, while 'df2' uses the 'Combined Europe' dataset.

For the AB_US (df1) dataset, this focused on 9 columns of data. Those were the following: 'name', 'host_id', 'latitude', 'longitude', 'room_type', 'price', 'number_of_reviews', 'availability_365', and 'city'.

The Combined Europe (df2) dataset uses 7 columns: 'realSum', 'room_type', 'dist', 'metro_dist', 'lng', 'lat', and 'guest_satisfaction_overall'. To keep the longitude and latitude information consistent, these were renamed from 'lng' and 'lat' to 'longitude' and 'latitude' respectively.

Fields	**Description**	**Example**
Name	Description of the Airbnb property	"Bright, Modern Garden Unit - 1BR/1BTH"
Price	Daily Rate	202

Room type	What type of accommodations are provided	Private room
Longitude	Longitude of the property (US), possibly used to join with the European dataset	-122.43317
Latitude	Latitude of the property (US), possibly used to join with the European dataset	37.77028
Number of reviews	Number of reviews (for US Airbnb only)	383
Availability_365	Number of days the property is available in a year	128
City	City of the property	San Francisco
realSum	Total price of the listing (US$)	194.0336981
room_type	Type of rooms/accommodation	“Private Room”, “Entire home/apt”, “Shared room”
guest_satisfaction_overall	Satisfaction rating (~100)	20~100
metro_dist	Distance to the metro station	2.539380003
Dist	Distance from the city center	5.022963798
Ing	Longitude (Europe)	4.90569
lat	Latitude (Europe)	52.41772

Program

The data sets were first attempted as a dictionary but after further review it was determined that using a data frame would be the best method for analysis. After reading in the data the data was cleaned. The cleaning was a simple change to the columns to only include the columns that were required for analysis and renaming some of those columns for better understanding.

Describing the data using the “Describe” function showed the means, and quartiles for each data point in the data. This can be used to help determine where most of the data lies in the dataset. This step helped to identify where the outliers are in the data. The describe function also to gain a better feel for the data and initial results to help interpret further analysis. Scatter plots were created using the DF US data set to calculate the number of reviews per city. From there the relationship between price and availability was viewed which helps to see if the price influences the availability.

Correlation analysis used the Pearson method which measures linear correlations. This provides a grid style product. The numbers are analyzed and the closer to absolute 1, means how close the relationship is. If the number is positive and close to 1 then the correlation is stronger the further from 1 you get the weaker the correlation is between those two points of data. If the correlation is negative then the relationship is weak, the same as with the positive correlation the further negative the number goes the weaker the correlation.

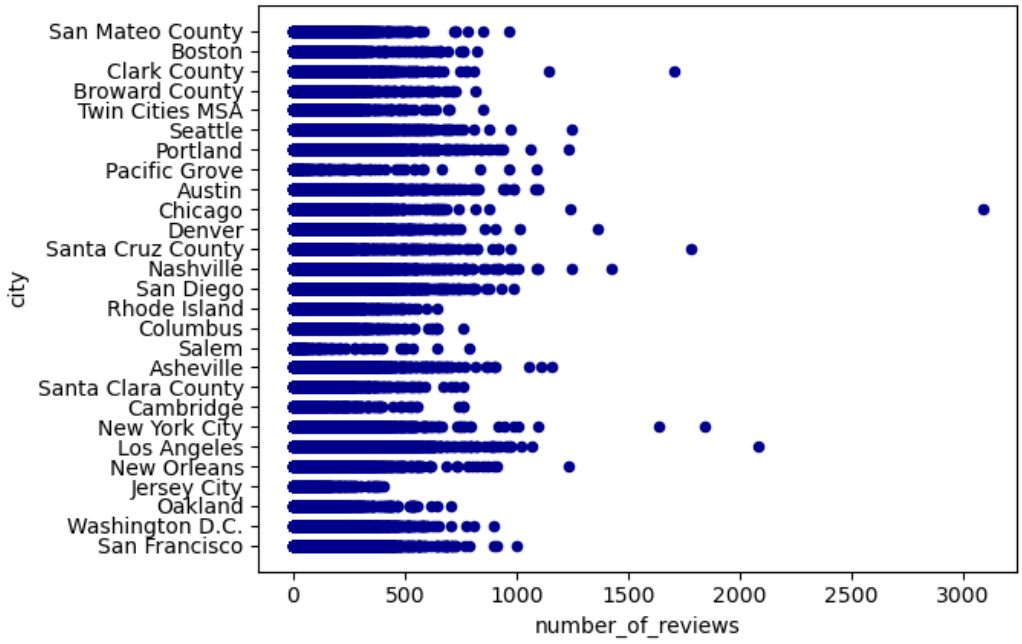
The “group by” function was used to group certain points of data, and then sort that data by the means of other points within that. In the US data set this was done by using the number of reviews, price, and availability and then grouped these by availability and then price. The Europe dataset used this function as well to group by the realSum and guest_satisfaction_overall.

Two pivot tables per data set were created to further analyze the data. These pivot tables compared room type, price and the locations. Finally, our last step was to see the average of the pivot tables. Once the pivot tables were created the averages using the “mean” function were created to see the averages within those tables to see the averages per housing type.

Analyzing the data

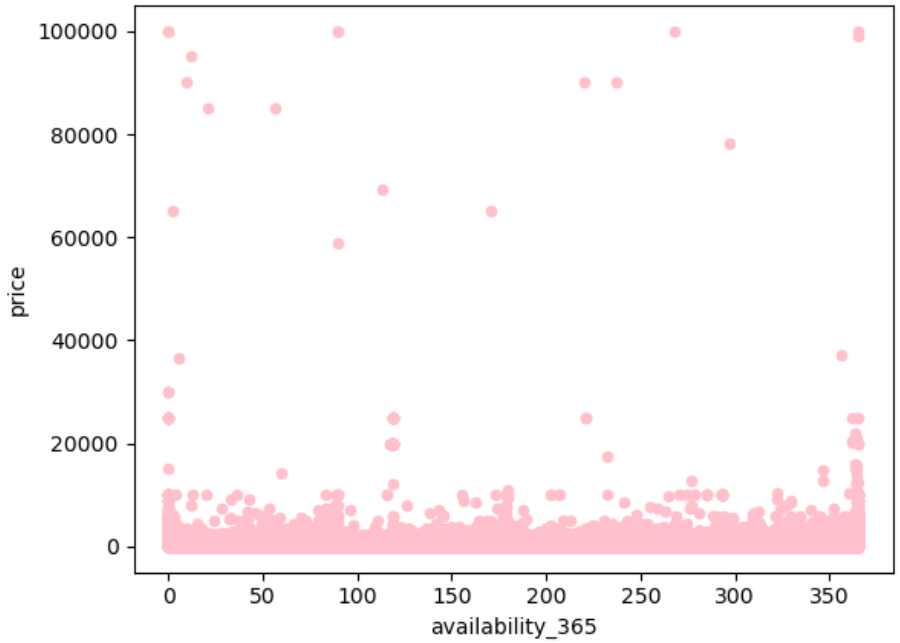
This Scatter plot was used in the US dataset to see the number of reviews per city. This helped to see where the most reviews were located per city as well as the outliers in the reviews.

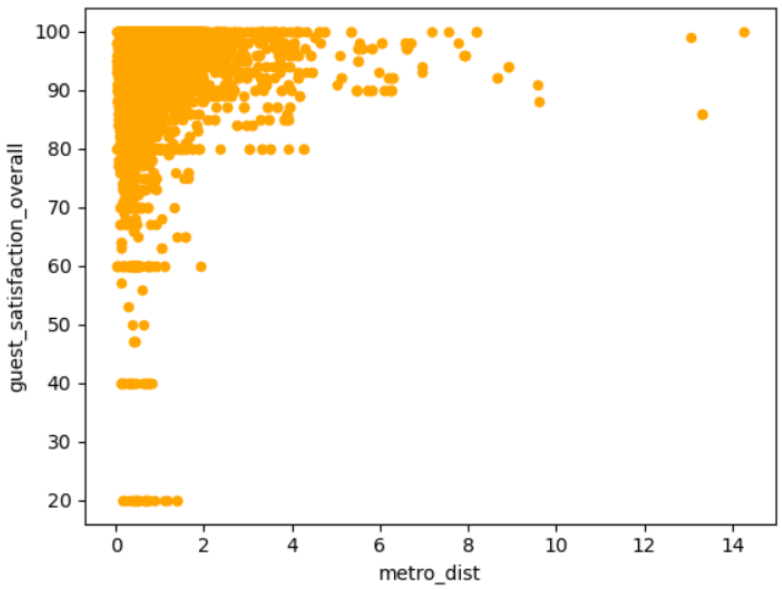
```
ScatPlot = df_clean.plot.scatter(x='number_of_reviews',y='city',c='DarkBlue')
```



This was another scatter plot for the US dataset. This shows the availability compared to the price. This helped to show the extreme values in price per Airbnb.

```
ScatPlot = df_clean.plot.scatter(x='availability_365', y='price', c='Pink')
```





The scatterplot above was used to compare the review rating to the metro distance. This shows again where there are outliers as well as where the ratings are clustered.

```
ScatPlot = df2_clean.plot.scatter(x='metro_dist',y='guest_satisfaction_overall',c='orange')
```

Using the pivot tables, helped to compare the type of housing favored depending on the location. For the US data location compared cities and number of reviews, while the European location was based on distance to the metro and guest satisfaction.

room_type number_of_reviews	Entire home/apt	Hotel room	Private room	Shared room
0	390.449019	1439.427083	369.990484	106.942590
1	293.510660	1265.011905	160.032293	124.373984
2	271.446247	588.025641	119.579863	80.729323
3	270.448254	765.241379	138.529918	81.000000
4	283.553101	349.565217	119.923836	72.298507
...	...	...	...	...
1705	0.000000	0.000000	226.000000	0.000000
1784	164.000000	0.000000	0.000000	0.000000
1842	234.000000	0.000000	0.000000	0.000000
2084	116.000000	0.000000	0.000000	0.000000
3091	0.000000	173.000000	0.000000	0.000000

room_type city	Entire home/apt	Hotel room	Private room	Shared room
Asheville	179.196071	442.125000	136.542416	65.428571
Austin	339.170272	786.750000	176.553579	71.055046
Boston	235.588608	546.333333	98.942263	59.375000
Broward County	334.289677	210.000000	190.506276	50.672897
Cambridge	238.016393	0.000000	108.802920	128.200000
Chicago	210.926607	141.440000	90.709251	45.847222
Clark County	286.644628	1754.788845	772.248985	172.784615
Columbus	144.178314	49.142857	63.471503	35.400000

room_type metro_dist	Entire home/apt	Private room	Shared room
0.011376	0.000000	254.978031	0.0
0.011382	0.000000	254.978031	0.0
0.012994	0.000000	196.895292	0.0
0.013004	0.000000	196.895292	0.0
0.014944	0.000000	381.921009	0.0
...	...	...	...
9.598773	201.692063	0.000000	0.0
13.069964	227.166495	0.000000	0.0
13.314108	155.417407	0.000000	0.0
13.314115	160.091614	0.000000	0.0
14.273577	359.680284	0.000000	0.0

10050 rows × 3 columns

room_type guest_satisfaction_overall	Entire home/apt	Private room	Shared room
20	225.774598	327.252405	0.000000
40	364.781275	200.659223	0.000000
47	173.177419	576.476929	0.000000
50	132.167881	0.000000	0.000000
53	337.933763	0.000000	0.000000
56	111.077262	0.000000	0.000000
57	115.998407	0.000000	0.000000
60	369.006814	250.385233	81.097504
...	...	...	...
93	339.391015	223.119957	152.215816
94	329.399985	218.163188	142.830361
95	375.346267	241.706034	159.068161
96	394.149468	248.286897	206.568352
97	353.126491	234.143548	145.757377
98	371.292744	236.077717	226.431669
99	337.649935	247.036867	81.331214
100	385.897272	250.507418	301.544826

Results

For the scatter plot with the blue colored plots, most of the reviews are less than 1000 per city. There are some outliers where the review goes over 1000 and even some in the 3000 range, these would be considered the outliers in our data. For the pink scatter plot, most of the prices for the BnB's are under the 10,000-dollar range. There are outliers that reach the 100,000-dollar range scattered within the availability. You can also see that the price when the BnB is offered 365 days a year has an increase in price with a spike to the 20,000-dollar area. While the other availability times see spikes as well like 75 days area the 365 is the highest in the grouping of the spike. The third plot was to help to see where the satisfaction rating was compared to the metro distance. In this you can see that the higher reviews are

clustered where the metro is closer to the Airbnb. There are some outliers in this where there are some higher satisfactions away from the metro, but in the other analysis it was determined that these were entire houses.

The correlation analysis was done for each data set. This was done to see where there may have been some connections in the data. For the US Dataset there are a few negative correlations and some positive there was only one very strong correlation and that was between availability and host ID. The Correlation between latitude and longitude was ignored as those would be known to have a direct strong correlation as they are two points that are generally grouped together. The Europe correlation had some strong correlations between the metro distance and the latitude, which does relate to the question being asked. There is also another strong connection between latitude and realSum which is the price.

Creating a pivot table using the European data set, it is determined that private rooms are favored closer to metro stations while whole houses/apartments are favored further away from metro stations. For the US Data comparing prices of each room type for each city. This shows that location impacted price rather than type of Airbnb.

Conclusions about data

From the data and the outcomes, there are several points that can be concluded about Airbnb data analyzed. Based on the findings, it was found for the Europe data that if an Airbnb is closer to the metro, then a private room is booked. The further away from the metro an Airbnb is, it is more likely that the entire home or apartment will be rented out.

Based on the analysis, city impacts the number of times an Airbnb will be rented. There is a possibility that the more popular city is more likely to have more bookings and thus more reviews. On the other hand, the availability of each property shows to impacts the number of reviews as well as the price. Looking at the outliers showed that some propertied had high availability and no price set. No conclusion can be made on this as there is no data available to state why there is not price set.

Based on the questions asked reviews, satisfaction, availability, and price point influence rentals. The more reviews a property has, the higher satisfaction there was and on the opposite side, the less reviews a property had, the lower the satisfaction was. The availability of a property impacted the number of reviews and could be attributed to the more availability the more renters could enter reviews on the property. Price had an impact on the number of reviews as well, properties with lower price points were more likely to be rented and therefor would also allow for more individuals to leave reviews.

Role of each member of the group

We decided on the data set together. The bulk of our project was done together via zoom calls. Jennifer worked on 20% of the programming data set, Lessa worked on 70 % of the program data set, and Ashley worked on 10% of the programming data set. Jennifer worked on 80% of the PowerPoint. Lessa and Ashley worked on the remaining 20% of the PowerPoint. The responsibilities of presenting were evenly distributed. Ashley worked on 30% of the written project write-up, Jennifer worked on 30% of the written

and Lessa worked on 40% of the written. We worked on the memo completely together. We were constantly checking and adding to each other's work.

Sources

Airbnb Prices in European Cities. (n.d.). Retrieved June 8, 2023, from

https://www.kaggle.com/datasets/thedevastator/airbnb-prices-in-european-cities

U.S. Airbnb Open Data. (n.d.). Retrieved June 8, 2023, from

https://www.kaggle.com/datasets/kritikseth/us-airbnb-open-data