Dillon Fouts & Tyler Noga Final Project

Air BnB Prices and their Neighborhoods

1. Introduction

Airbnb has revolutionized the hospitality industry by providing a platform for people to book short-term accommodations worldwide. With over 7 million listings available as of 2021, there is a wealth of data available to analyze the characteristics and patterns of Airbnb properties. In this project, we aim to explore the relationship between the price of an Airbnb listing and various features of the property. Using Airbnb listings data, we will examine factors such as location, type of accommodation, demographic data, and amenities offered to gain insights into how these factors influence pricing. By analyzing this data, we hope to provide useful information for both Airbnb hosts and guests looking to optimize their experiences on the platform.

2. Data

This project uses three primary sources of data: Airbnb listing data, United States demographic data, and state income data. We horizontally merged all 3 data frames to give extra insights into the price per night of an Airbnb beyond the scope of just the Airbnb listing.

2.1 Airbnb data

To make sense of the vast amounts of data we collected for this project, we had to roll up our sleeves and get our hands dirty with some serious data cleaning! We first started by reading in the Airbnb listings data, which was gathered from here, and stored as a file named "airbnb-listings.csv". This file was separated by semicolons, so we used the read.csv function to load it into R using sep = ";".

Next, we had to select the columns we needed for our analysis. We only wanted to include data that was relevant to our research question, so we used the subset function to select the following columns: "City", "State", "Property.Type", "Room.Type", "Accommodates", "Bathrooms", "Bedrooms", "Beds", "Bed.Type", "Amenities", "Price", "Number.of.Reviews", and "Review.Scores.Rating". We left out any columns that contained missing data or were not useful for our research.

After selecting the columns, we had to clean the data to make sure it was complete and accurate. We removed any incomplete cases using the complete.cases function. This made sure that we only included Airbnb listings with complete information, which was necessary for our analysis. We also inspected our data to check for any inconsistencies or other missing data.

To further clean the data, we removed any listings with missing price data. This helped us to ensure that we had accurate pricing information for each listing. The resulting dataset, named airbnb_cleaned, contained 134,545 Airbnb listings from various cities across the United States.

2.2 Demographic data

We also collected demographic data for major US cities using this dataset. We read this data into R using the read.csv function and selected the columns that were relevant to our analysis, including "City", "Median.Age", "Total.Population", "Race", and "Count". We then used the dplyr package to calculate summary statistics for each group. This included the majority race, mean total population, and mean median age for each city. The resulting dataset, named demographics_cleaned, provided us with demographic data for 2,891 major US cities. Once the data was cleaned, we then horizontally merged the demographic data set with our original Airbnb data set forming "merged_df".

2.3 Web scraping state income data

Next, we scraped income data on the 50 US states from here. We gathered the state and median single income for all 50 states using the xml_find_all() function. Once the nodes were scraped we created a data frame named "state_incomes_df." The file state_income.R contains the code to scrape, transform, and write to a csv file (state_income.csv) that will be used in our merged_df.R file to later be horizontally merged. This data allows us to compare income levels with Airbnb listing prices across different states.

2.4 Amentities

Finally, we cleaned our data about the amenities offered by Airbnb listings using the dataset "amenities.xlsx" which was created from our original data frame. We then selected the columns that were relevant to our analysis, including "Price" and "Count". Finally, we merged this data with the airbnb_cleaned dataset based on the price variable.

In conclusion, cleaning the data was a crucial step in our project. By selecting relevant columns, removing incomplete cases, and calculating summary statistics, we were able to create accurate and complete datasets that were ready for further analysis. This helped us to gain valuable insights into the relationship between Airbnb listing prices and various features of the properties. We choose to utilize different R scripts when reading and cleaning data to make reading our code easier. Our final data frame was written to a csv file named "merged_df.csv" by the file "final_project_merged_dataframe.R". The file is then read into our main script ("tnog_drfouts_projectCheckin.R") using read.csv for analysis. Found below in table 1 is an overview of "merged_df" consisting of 102,890 observations with 18 variables.

Table 1: Data Dictionary of Merged Data Frame

Column Name	Data Type	Description
State	Chr	The state where the Airbnb listing is located
City	Chr	The city where the Airbnb listing is located
Property.Type	Chr	The type of property, such as apartment, house, or condo
Room.Type	Chr	The type of room, such as entire home, private room, or shared room

Accommodates	Int	The maximum number of guests the Airbnb listing can accommodate
Bathrooms	num	The number of bathrooms in the Airbnb listing
Bedrooms	int	The number of bedrooms in the Airbnb listing
Beds	int	The number of beds in the Airbnb listing
Bed.Type	chr	The type of bed, such as queen, king, or sofa bed
Price	int	The price per night of the Airbnb listing
Number.of.Reviews	int	The total number of reviews for the Airbnb listing
Review.Scores.Rating	int	The average rating score of the Airbnb listing based on reviews
Race	chr	The majority race for the city where the Airbnb listing is located
Total_Count	int	The total number of people from the majority race in the city where the Airbnb listing is located
Total.Population	num	The total population of the city where the Airbnb listing is located
Median.Age	num	The median age of the population in the city where the Airbnb listing is located
Amenities.Offered	int	The number of amenities offered by the Airbnb listing, such as wifi, air conditioning, or pool
State.Incomes	num	The median income of the state where the Airbnb listing is located

1 https://public.opendatasoft.com/explore/dataset/airbnb-listings

2 https://public.opendatasoft.com/explore/dataset/us-cities-demographics

3 https://www.justice.gov/States/Median/Income

3. Analysis

Our analysis of the Airbnb data aimed to gain insights into the relationship between the price of the listing and the various features of the property, such as the location, type of accommodation, reviews/ratings, demographic data, and amenities offered. Specifically, we investigated the following research questions:

1. What is the relationship between the price of an Airbnb listing and the location of the property, as determined by the state and city? (Average price, top 5 cities, bottom 5 cities)
2. Does the type of accommodation and room type have an impact on the price of the Airbnb listing?
3. How do the number of reviews and the rating score affect the price of an Airbnb?
4. Are there any demographic factors, such as race, median age, and total population, that are associated with the price of an Airbnb listing?

3.1 Relationship between Price and Location

To investigate the relationship between the price of an Airbnb listing and the location of the property, we first looked at the median income of each state. We found that there was a positive correlation between the median income of a state and the average price per night of an Airbnb listing located in that state. States with higher median incomes tended to have higher average prices for Airbnb listings, while states with lower median incomes tended to have lower average prices. We then utilized the mean() function to calculate the average price per night for the entire data set. We discovered that the average cost per night was $164.78. We were also interested in finding the top 5, and bottom 5 cities with the highest/lowest average cost per night. We did this by using the aggregate() function and grouping our data by cities. We then sorted highest to lowest, and lowest to highest selecting the top 5 records' using the head() function, resulting in the 5 highest, and 5 lowest average cost per night cities. Shown in figure 1 are the top 5 highest costing cities in our dataset.

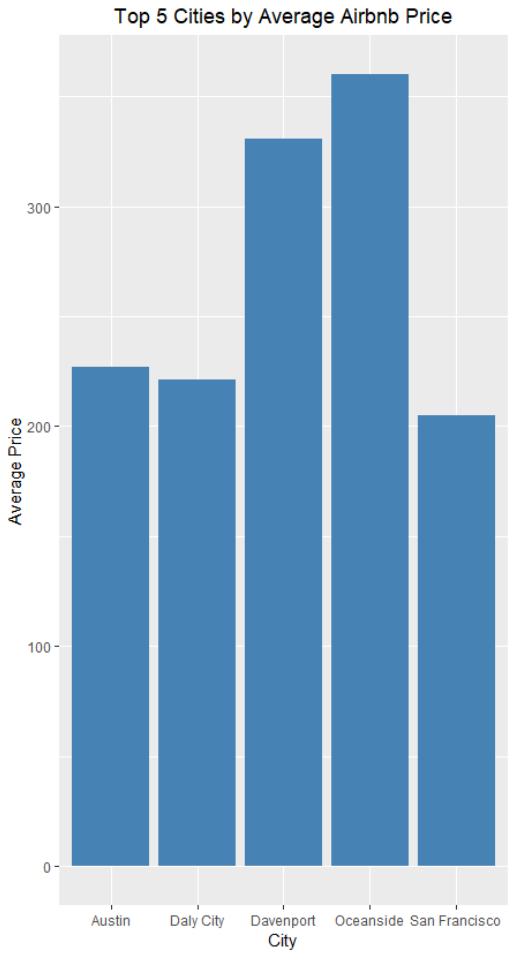

Figure 1 Top 5 Cities Based on Average Airbnb Price

We then looked at the relationship between the price of an Airbnb listing and the city where it is located. We found that there were significant differences in the average price per night of Airbnb listings across different cities in the United States. For example, Oceanside, CA had the highest average price per night, while smaller cities like Tulsa, Oklahoma had much lower average prices.

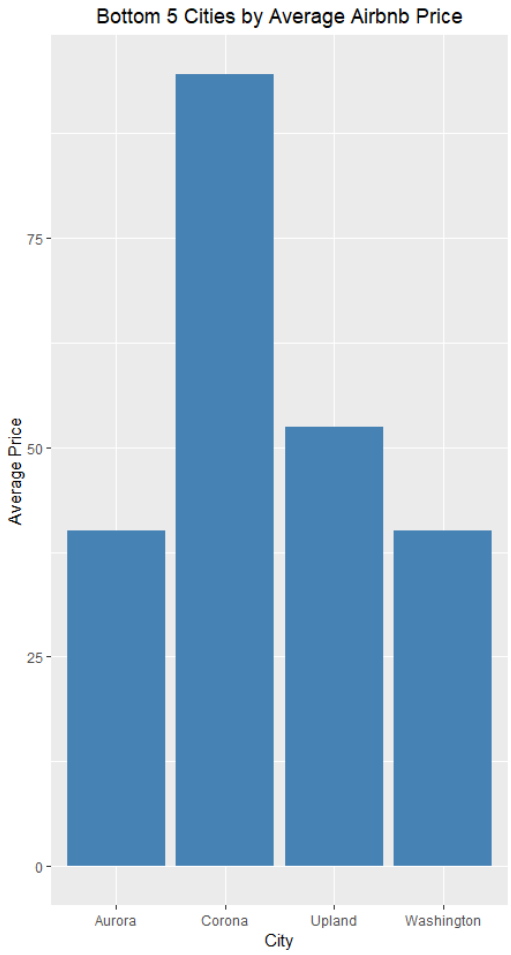

Figure 2 Bottom 5 Cities Based on Average Airbnb Price

3.2 Impact of Accommodation and Room Type on Price

Next, we investigated the impact of the type of accommodation and room type on the price of an Airbnb listing. We utilized linear regression to create models for both Room.Type and Property.Type. Our findings suggest that entire homes/apartments had the highest average price per night ($212.26), followed by private rooms ($91.86), and shared rooms ($59.47). This model, however, shouldn't be used as a standalone model to predict the price of an Airbnb as it has a low R-Squared value of only 0.1798, suggesting that there may be other variables affecting price. As a rule of thumb an R-Squared value below 0.5 may indicate that the model has poor explanatory power, however we feel as though this model is a good baseline to show that renting an entire Airbnb vs only renting out a private room seems to be logical.

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           212.2577     0.5116  414.93   <2e-16
Room.TypePrivate room -120.3953     0.8476 -142.04   <2e-16
Room.TypeShared room  -152.7902     2.2571  -67.69   <2e-16
```

Figure 3 Linear Regression on Price and Room.Type

```
Multiple R-squared:  0.1798,    Adjusted R-squared:  0.1798
```

Figure 4 Room.Type R-Squared Value

We then plotted our Room.Type model to check the assumptions of a linear regression model. We did not see any patterns in our residual plots, and our model seemed to follow all assumptions of linear regression. The following graph shows our residuals vs our fitted values. Note that the price at the different levels: entire house, private room, and shared room, seem to all be randomized and don't follow any potential pattern.

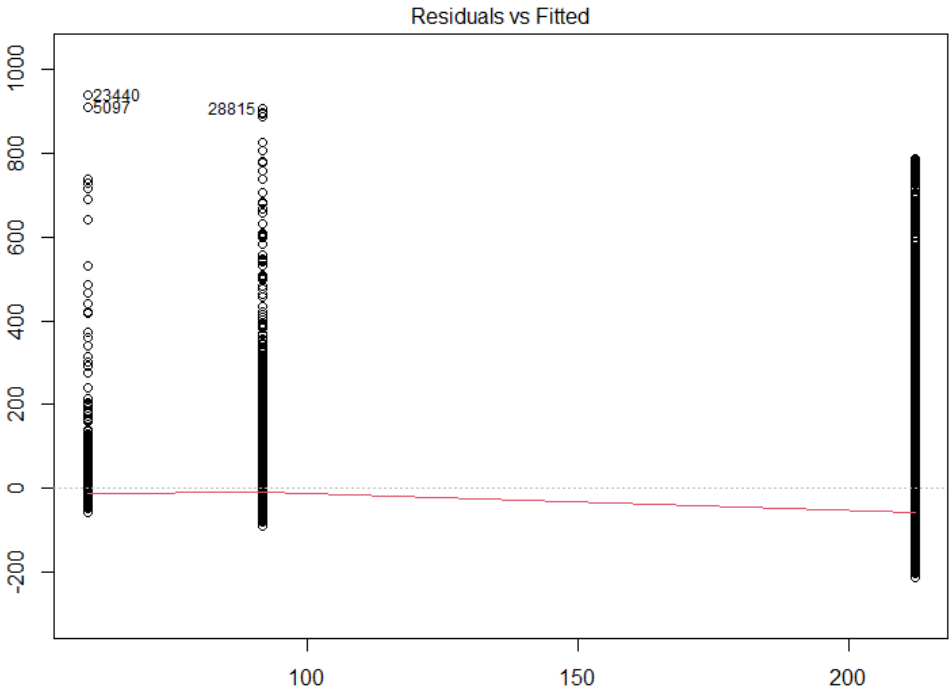

Figure 5 Room.Type Residuals

In addition, we discovered that certain property types had a higher cost per night than others. "Apartment" was the intercept of our model at $154.30 per night. Some of the more expensive property types consist of: boat ($222.79), castle ($254.53), vacation home ($348.64), and train ($534.50). Some cheaper options are: hostel ($50.24), dorm ($54.21), and tipi ($80.33). **As an important note**, not all values in the model were significant predictors of price. Many of the property types had P values greater than .05 which should be removed from the model one by one starting with the largest value first. Once all non-significant predictors are removed, the model will be parsimonious. However, like our previous model, this model also had a very low R-Squared value of only 0.0168. We felt that since this model represented such a low coverage in the fluctuations of price that we would skip creating the final parsimonious model for Property.Type, and instead focus on our additional research questions.

```
Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      154.3001     0.5935 259.987  < 2e-16 ***
Property.TypeBed & Breakfast     -31.2008     4.8512  -6.432 1.27e-10 ***
Property.TypeBoat                 68.4894    14.4649   4.735 2.19e-06 ***
Property.TypeBoutique hotel      -16.7053    11.4039  -1.465 0.142958
Property.TypeBungalow            -21.2769     5.5698  -3.820 0.000133 ***
Property.TypeCabin               -31.9001    10.5164  -3.033 0.002419 **
Property.TypeCamper/RV           -40.0098    10.3459  -3.867 0.000110 ***
Property.TypeCasa particular     -74.3001   140.8688  -0.527 0.597888
Property.TypeCastle              100.2332    36.3767   2.755 0.005863 **
Property.TypeCave                 56.9499    70.4363   0.809 0.418787
Property.TypeChalet              -70.0501    49.8077  -1.406 0.159605
Property.TypeCondominium          46.6102     2.2253  20.945  < 2e-16 ***
Property.TypeDorm               -100.0899     7.9717 -12.556  < 2e-16 ***
Property.TypeEarth House         -23.8556    46.9596  -0.508 0.611452
Property.TypeEntire Floor         -2.4177    34.1706  -0.071 0.943593
Property.TypeGuest suite         -37.5924    17.4825  -2.150 0.031535 *
Property.TypeGuesthouse          -29.2408     4.9881  -5.862 4.58e-09 ***
Property.TypeHostel             -104.0604    12.8199  -8.117 4.83e-16 ***
Property.TypeHouse                25.3128     0.9699  26.099  < 2e-16 ***
Property.TypeHut                 -88.3001    46.9596  -1.880 0.060064 .
Property.TypeIn-law              -63.9001    31.5045  -2.028 0.042534 *
```

*Figure 6 Property.Type Linear Regression (**Non-Parsimonious**)*

```
Multiple R-squared:  0.0168,    Adjusted R-squared:  0.01645
```

Figure 7 Property.Type R-Squared Value

And lastly, we checked the number of bedrooms, bathrooms, beds, and the number of people the Airbnb can accommodate. We ran a linear regression model and found that all variables are significant predictors of price (alpha < .05). However, we found the value of beds to be unexpected and illogical (we wouldn't expect the price of the Airbnb to be $7.52 less for every 1 increase in bed).

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.6504     0.8374   7.942 2.01e-15 ***
Bedrooms      33.3999     0.6486  51.496  < 2e-16 ***
Beds          -7.5202     0.4753 -15.823  < 2e-16 ***
Bathrooms     36.1924     0.7424  48.749  < 2e-16 ***
Accommodates  23.3135     0.2942  79.247  < 2e-16 ***
```

Figure 8 Amenities Linear Regression

```
                Price      Beds Accommodates  Bedrooms Bathrooms
Price        1.0000000 0.4722592    0.5599249 0.5375353 0.4501822
Beds         0.4722592 1.0000000    0.8164268 0.7123119 0.5472548
Accommodates 0.5599249 0.8164268    1.0000000 0.7461497 0.5388516
Bedrooms     0.5375353 0.7123119    0.7461497 1.0000000 0.6113313
Bathrooms    0.4501822 0.5472548    0.5388516 0.6113313 1.0000000
```

Figure 9 Amenities Correlation

Because of this illogical result we checked the correlations between our variables. We found that "Accommodates" and "Beds" are extremely correlated (value > .80). Because of the strong correlation to each other, and its effect on the price, we removed "Beds" from our linear regression model and re-ran the regression. We found this model to be a better predictor of price than our previous models, with a

higher R-Squared value of 0.36. One interesting thing about this result is that Bathrooms was a bigger increase in price than bedrooms.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.3833     0.8371    8.82   <2e-16 ***
Bedrooms      31.2764     0.6355   49.21   <2e-16 ***
Bathrooms     34.6017     0.7365   46.98   <2e-16 ***
Accommodates  20.5360     0.2361   86.97   <2e-16 ***
```

Figure 10 Amenities Final Linear Regression

```
Multiple R-squared:   0.36,     Adjusted R-squared:   0.36
```

Figure 11 Amenities R-Squared Value

3.3 How reviews/ ratings affect price

Next, we looked to see how reviews and ratings affected the price of Airbnb's. We wanted to answer the question "Are Airbnb's with more reviews and better scores more likely to be expensive?" Our hypothesis is that higher rated Airbnb's would be priced higher than lower rated ones.

To Answer the question, we ran a linear regression to check if they were significant predictors of price. Both Review.Scores.Rating and Number.of.Reviews were significant predictors (alpha<0.05) so we continued with our analysis. We also noted that the model had a very low R-Squared value of 0.01339 indicating a poor representation of change in price. In addition, it seemed that for every increase in the number of reviews you could expect a $0.29 cent decrease in the price of for the Airbnb. We speculate this may be due to people only leaving reviews if they had a negative experience, however this is an oversimplification and shouldn't be taken as anything other than pure speculation.

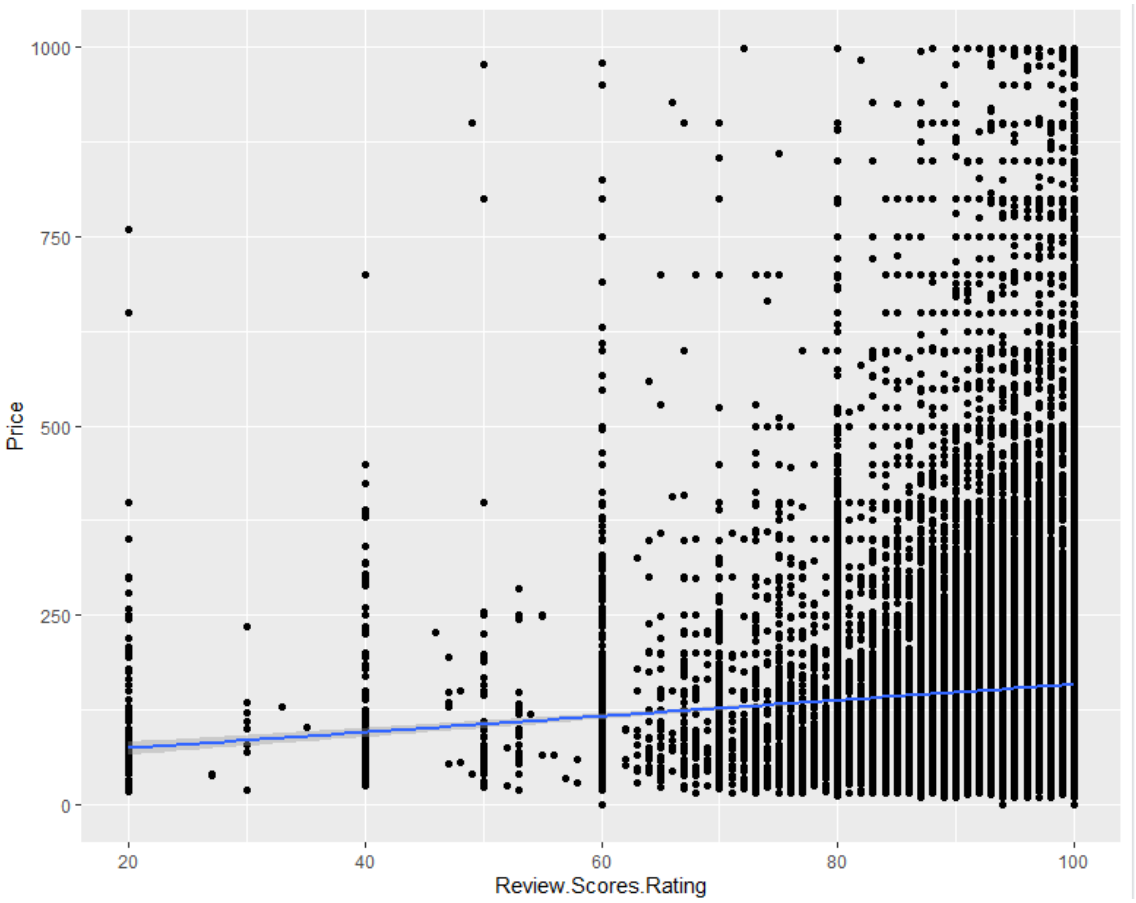

Figure 12 Review.Scores.Rating Scatterplot

Figure 12 illustrates the positive relation between review ratings and price. For every 1-point increase in the review score, you can expect that the price of the Airbnb to cost $1.09 more. We also found that the only Airbnb's with a price of $999 (the maximum allowed in our dataset) to be those with scores above 70, but mainly in the 90-100 range. We would expect this because if people are rating your Airbnb highly that would suggest to others that it is a good place to stay or that it is a quality Airbnb resulting in a higher price.

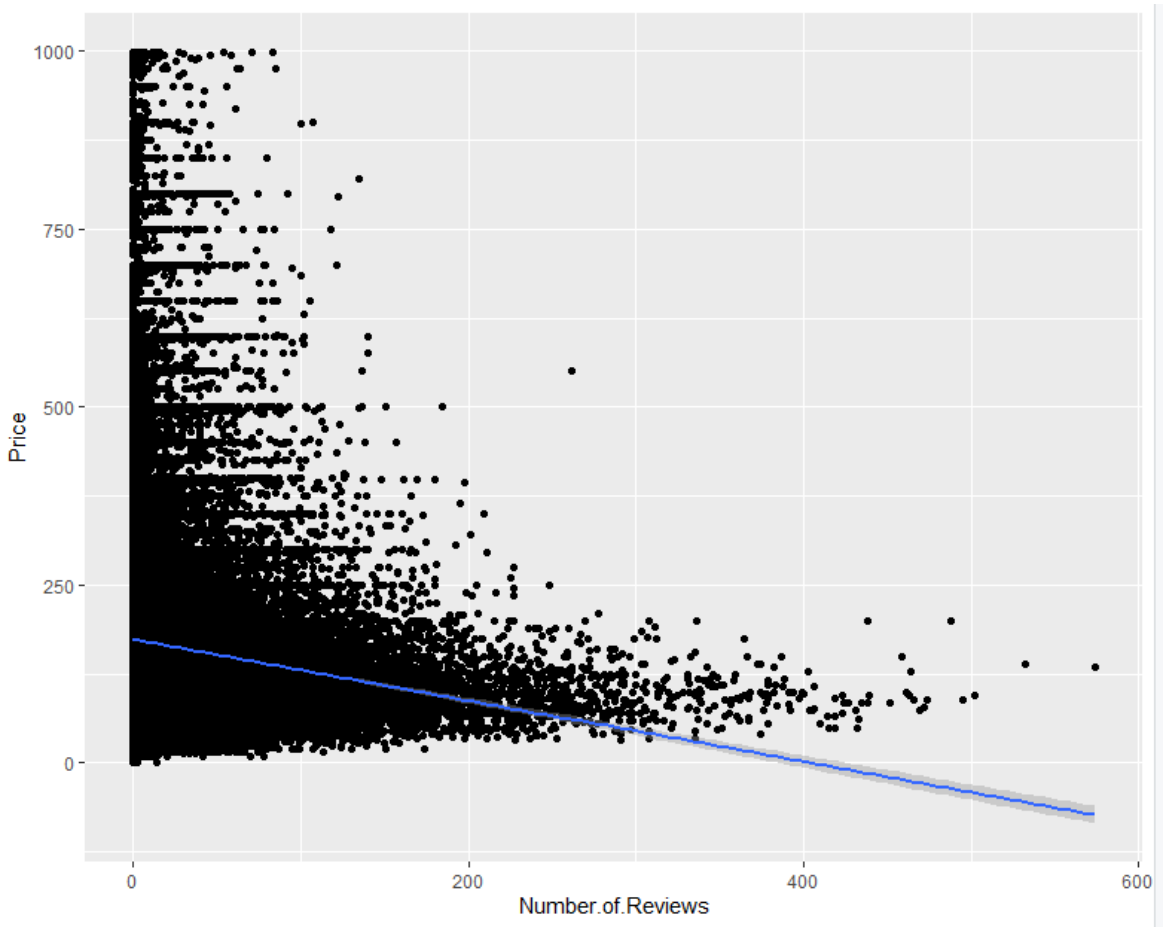

Figure 13 Number.of.Reviews Scatterplot

Figure 13 illustrates the weak downward trend between the price and number of reviews. A "theory" we came up with when interpreting this graph is that not as many people are renting $999 / night Airbnb's, so there are less reviews. We used the mindset: instead of thinking of this graphic as "the more reviews it has, the lower the price", think of it as "the higher the price, the less reviews it will have." This thinking is further supported by our original hypothesis that many people only leave bad reviews when having a negative experience, like when staying at a very cheap Airbnb. This logic seems to follow that the cheaper Airbnb is, the more reviews it will have. Another possible explanation is that Airbnb's have "sweet-spot pricing". This term refers to the idea that there is a "sweet spot" for the price of an Airbnb, and thus, there are a lot more $150-$200 Airbnb's than there are $999 Airbnb's and thus, there would be more reviews for $150-$200 Airbnb's. Again, this is an oversimplification, and these suggestions should be considered hesitantly. To accurately explain the graph, further research would be needed to find the exact cause for this relationship (such as only leaving bad reviews, sweet-spot pricing, etc.).

3.4 Association between Demographic Factors and Price

Finally, we explored whether there were any demographic factors associated with the price of an Airbnb listing. Specifically, we found that Total.Population, Median.Age, Race, and State.Incomes are all significant predictors of price (alpha <0.05). However, for Race, Hispanic and Latino were not significant predictors of price and were removed from our model. Furthermore, we discovered that these variables exhibit a weak negative correlation with the

Airbnb price, indicating minimal to no influence on the price. Total.Population, Median.Age, and State.Incomes all had a negative correlation of less than 10% (-0.10). We can't say for sure why these values are negative, but we assume it has something to do with the locational demographics (though this should be considered hesitantly). In addition, we were also interested in seeing the average price of Airbnb's grouped by the majority race of that city. We wanted to answer the question "Are certain race 'dominated' cities more expensive than others?" We answered this question using the dpylr package and grouping by race. This resulted in four bins: Asian ($86.8), Black or African American ($185), Hispanic or Latino ($82.6), and White ($163).

	Race	avg_price
	<chr>	<dbl>
1	Asian	86.8
2	Black or African-American	185.
3	Hispanic or Latino	82.6
4	White	163.

Figure 14 Average Price Grouped by Race

These findings suggest that if an Airbnb's city's majority race is Black or African American it would have the highest cost, and if it was in a Hispanic or Latino it would be the cheapest. Again, these results are generalized, and only cover the scope of our data (which is limited).

Our final research question addresses the number of listings in each city covered by our data. Specifically, were there more listings in states where median salary is in the top half of the United States (top 25)? To do so, we grouped by city, and counted the number of listings in each city/state and compared it to our state_incomes column.

	City	Sta
	<chr>	<ch
1	Los Angeles	CA
2	New York	NY
3	San Francisco	CA
4	Washington	DC
5	Austin	TX
6	San Diego	CA
7	Chicago	IL
8	Seattle	WA
9	New Orleans	LA
10	Boston	MA

Figure 15 Top 10 Cities by Number of Listings

Our findings from our data suggest that the top 10 cities with the most Airbnb listings are usually in the top 25 of state income earners. For example, CA was the state with the most listings and is ranked 14th in our state_incomes data at $65,895/ year median income. In addition: MA (5th @ $75,077), WA (6th @ $74,398), DC (7th @ $74,266), NY (18th @ $63,548), and IL (19th @ $61,456) are all in the top 25 state

earners, with a majority in the top 10. The only state that didn't meet these criteria was TX at 32nd with a median income of $55,441.

To verify these results, we ran a correlation between the number of listings in each city and state_incomes. We found a significantly strong correlation of 0.997 between these variables indicating that if a city has more listings than another, its state income is expected to be higher. We then plotted the number of listings on a scatterplot with the price of the Airbnb. Our findings suggest the more listings there were in a city, the more "unique" prices there were. This can be seen by the 2 lines on the right in the graph. These 2 lines represent Los Angeles and New York City which both nearly had 20,000 listings in our data. The furthest left bars represent cities with less listings. Here the price tended to hover below $250. This graphic is super generalized but is aimed at showing the relation of price to the number of listings.

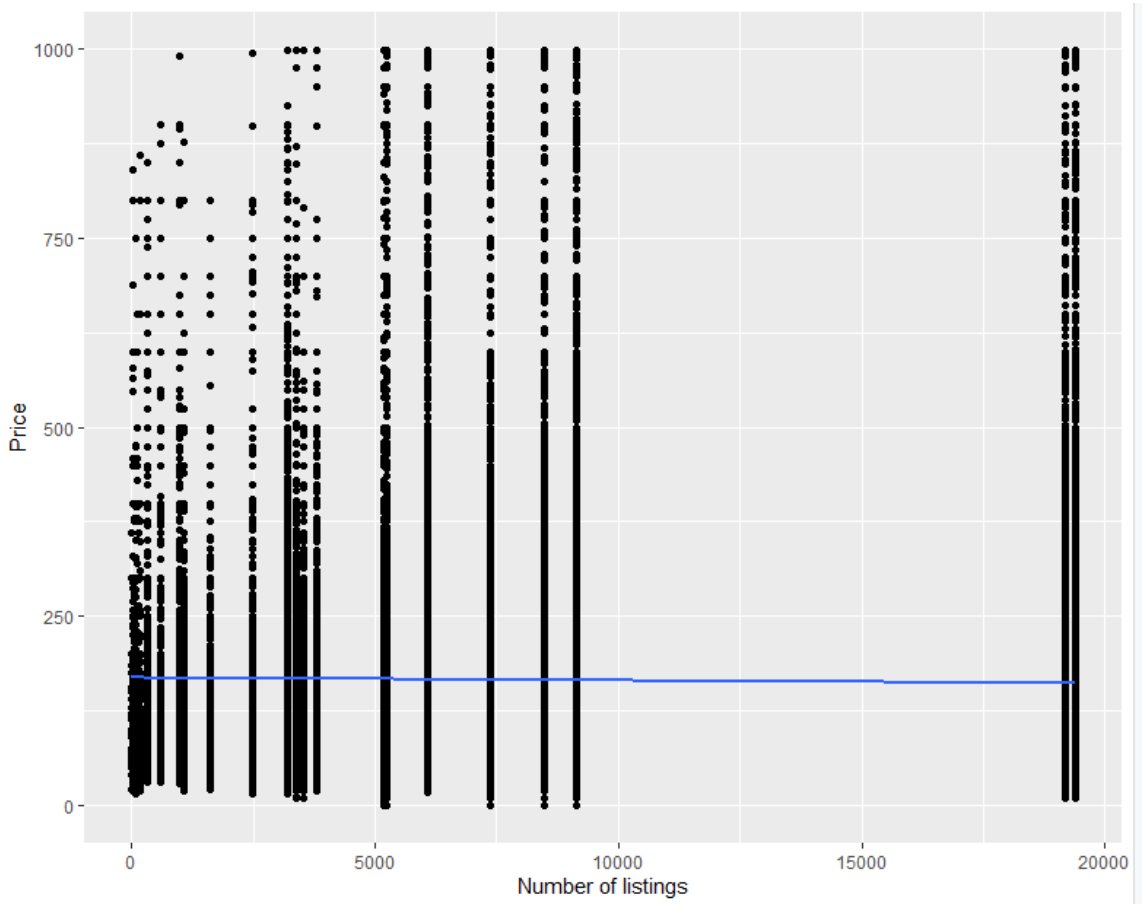

Figure 16 How number of listings effects price

4. Conclusion

In this project, we analyzed four aspects of Airbnb's pricing: the effect location has on price, the amenities offered, do reviews/ ratings matter, and how demographics affect the price (such as median salary, race, etc.). In summary, from the analysis questions presented in our proposal, we found the following results.

1. What is the relationship between the price of an Airbnb listing and the location of the property, as determined by the state and city? (Average price, top 5 cities, bottom 5 cities)
    a. We found that there was a positive correlation between the median income of a state and the average price per night of an Airbnb listing located in that state. States with higher median incomes tended to have higher average prices for Airbnb listings, while states with lower median incomes tended to have lower average prices
    b. Average price per night of data set was $164.78 / night
    c. Top 5 cities by average price
        i. Oceanside, CA @ $360 / night average
        ii. Davenport, CA
        iii. Austin, TX
        iv. Daly City, CA
        v. San Fransico, CA
    d. Bottom 5 cities by average price
        i. Corona, CA @ $39 / night average
        ii. Washington, MD
        iii. Aurora, CO
        iv. Upland, CA
        v. Corona, NY
2. Does the type of accommodation, room type, or amenities offered have an impact on the price of the Airbnb listing?
    a. Yes accommodation, room type, property type, and the amenities offered are all significant predictors of price (alpha < 0.05). We found that amenities such as number of bedrooms, bathrooms, and the number of people to be accommodated have the greatest "explanatory" power when predicting price. That model had an R-Squared value of .36 suggesting a 36% explanation for the changes in price.
    b. The average cost to rent an entire home is $212.26, followed by private rooms $91.86, and shared rooms $59.47
    c. In addition, for every 1 increase:
        i. In the number of bedrooms, you can expect a $31.28 increase in price
        ii. In the number of bathrooms, you can expect a $34.60 increase in price
        iii. In the number people it can accommodate, you can expect a $20.54 increase in price
        iv. For example: A 4-bedroom, 2-bathroom house that can accommodate 6 guests would roughly cost $324.91 per night.
3. How do the number of reviews and the rating score affect the price of an Airbnb?
    a. There is a positive relation between review ratings and price. For every 1-point increase in the review score, you can expect that the price of the Airbnb to cost $1.09 more. We also found that the only Airbnb's with a price of $999 (the maximum allowed in our dataset) to be those with scores above 70, but mainly in the 90-100 range.

b. We also found that there was a negative relation between the number of reviews and the price. This result was surprising, and we devolved some theories that may need further research:
    i. Less people are renting $999 / night Airbnb and therefore there are less reviews for expensive listings.
    ii. In general, people only leave reviews when they have a negative experience, so there may be more reviews for Airbnb's with worse experiences for the customer, resulting in a lower average price.
    iii. There is a "sweet spot" pricing when it comes to Airbnb's. Throughout our research we found that the typical Airbnb was around $150-$250 / night. Because this is the "typical" Airbnb there are more guests renting at this price point, and therefore more reviews. Because the price is "lower" and there are more reviews, it results in a negative relationship between price and number of reviews.

4. Are there any demographic factors, such as race, median age, and total population, that are associated with the price of an Airbnb listing?
    a. We found that the total population, median age, race, and State.Incomes are all significant predictors of price (alpha <0.05). However, for Race, Hispanic and Latino were not significant predictors of price. In addition, these variables had a very low explanatory power (R-Squared value of ~.01).
    b. The average price of an Airbnb in a city with the majority race of:
        i. Black or African American: $185.00 / night
        ii. White: $163.00 / night
        iii. Asian $ 86.80 / night
        iv. Hispanic or Latino $ 82.60 / night
    c. Finally, we also found that the majority of cities with the most listings were among the top 25 state income earners.
        i. CA, MA, WA, DC, NY and IL were all in the top 25 state income earners and had the greatest number of listings in our data
        ii. The more listings a city had, the more "unique" prices there were for Airbnb's. If a city had ~20,000 listings you could expect the price to range anywhere from $40-$999 / night, whereas if a city had less than 1,000 listings you could expect the price to be ~$250 or less.

Overall, our analysis suggests that the price of an Airbnb listing is influenced by a variety of factors, including the location of the property, the type of accommodation / room type, demographic factors, reviews/ratings, and the amenities offered. However, it is important to note that our analysis is limited by the scope of our data, which only includes Airbnb listings in the United States. Therefore, our results may not be generalizable to other countries or regions. Additionally, it is important to exercise caution when interpreting our results and not over-generalize the findings beyond the scope of our data. We hope our analysis is useful to any potential Airbnb users who may want to use Airbnb in the future to better understand what affects an Airbnb's price.