AIRBNB LISTINGS & REVIEWS

Hypothesis Testing, Regression Models & More

Authors: Aditi Poddar, Kristina Cinova, Sylvia Lam

MS Marketing Intelligence, University of San Francisco

MSMI 603: Applied Statistics

Professor Matthew Meister

1st December, 2023

Index

1. Predicting Airbnb Rental Price of a Property and Decisions

1.1. Problem Statement (A)

‘How much should I charge per night to rent this entire property?’

1.2. Approach Overview

The analysis commenced with a thorough examination of the 'SF Listings Data', encompassing a comprehensive array of Airbnb properties in San Francisco. This dataset includes essential attributes of each listing, such as descriptions, host neighborhoods, superhost status, type of property, number of beds and baths, range of amenities, pricing, reviews, and additional pertinent details.

Data Observation and Variable Identification:
The initial phase involved a detailed observation of the dataset to discern the specific factors that significantly influence the pricing of properties. The investigation led to the identification of key determinants of price, notably the property's location, the number of beds and baths, and the variety of amenities provided.

Methodology - Multiple Regression Analysis:
To accurately forecast property prices, we employed Multiple Regression Analysis. This statistical method allows for the consideration of multiple independent variables to predict a dependent variable, in this case, the price.

Model Development:
Two distinct regression models were constructed to estimate prices:

A. **Model A**: Price as a function of Location, Number of Beds, and Number of Baths.

B. **Model B**: Price as a function of Location, Number of Beds, Number of Baths, and Amenities.

Hypothesis Formulation:
To validate the effectiveness of these models, the following hypotheses were proposed:

A. **Null Hypothesis**: There is no significant difference in price prediction capabilities between Model I and Model II.

B. **Alternative Hypothesis**: Model II, which includes amenities as a variable, will yield different price predictions compared to Model I, thereby indicating the influence of amenities in determining property prices.

1.3. Methodological Implementation: Price Prediction Using Multiple Regression

Model A

1.3.1. Overview

This section outlines the technical steps implemented in R for developing a regression model to predict Airbnb prices in San Francisco based on specific variables: beds, baths, and location.

1.3.2. Data Preparation and Analysis

1. Library Utilization and Data Aggregation:
   - The `dplyr` library was used for data manipulation.
   - Four separate quarter datasets (`sf_bnb_listing 1.csv` to `sf_bnb_listing 4.csv`) were read and then combined into a single dataframe, `sf_bnb_listings`, to facilitate comprehensive analysis.

2. Data Cleaning and Transformation:
   - The Airbnb listings data (`sf_bnb_listings.csv`) was read into the variable `airbnb`.
   - The `price` variable was cleaned and converted from a character string to a numeric value, removing any dollar signs and commas.
   - The `host_neighbourhood` variable was transformed into categorical variables, `inner_richmond` and `outer_richmond`, to represent whether a listing is in the Inner or Outer Richmond neighborhood.

1.3.3. Model Development

- Two separate linear models (`model1` and `model2`) were constructed using the `lm()` function. Each model predicts the price based on the number of bedrooms, number of bathrooms, and the location (Inner or Outer Richmond).
- The first model (`model1`) predicts prices for properties in the Inner Richmond area, and the second model (`model2`) for those in the Outer Richmond area.

1.3.4. Results

1. Price Prediction:

- Predictions were made for a hypothetical property with 2 bedrooms and 2 bathrooms in both Inner and Outer Richmond.
- The predicted price for a property in Inner Richmond was $440.47, and in Outer Richmond, it was $409.27.

2. Average Price Calculation:

- An average price of $424.87 was calculated for properties with similar characteristics in both neighborhoods

MODEL B

1.3.5. Overview

This section outlines the technical steps implemented in R for developing a regression model to predict Airbnb prices in San Francisco based on specific variables: beds, baths, location and amenities.

1.3.6. Data Preparation and Transformation

1. Data Aggregation: The Airbnb listings data was loaded from sf_bnb_listing.csv into the variable bnb.

2. Data Cleaning and Transformation

- Amenities Analysis: The amenities column was cleaned to remove special characters and transformed to lowercase for uniformity.
- A comprehensive list of all amenities was created and the frequency of each amenity was counted. This data was then organized into a dataframe, amenities_data, sorted in descending order based on occurrence.
- Creating Dummy Variables:A set of top amenities was selected for analysis, along with additional amenities such as 'washer' and 'tv'.
- For each amenity in this set, a new column was created in the dataset, indicating the presence (1) or absence (0) of that amenity in each listing.

3. Further Data Cleaning:

- The price column was converted from a character string to a numeric value.
- The host_neighbourhood was categorized into 'inner_richmond' and 'outer_richmond'.

- The bathrooms_text was processed to extract numeric values for the number of bathrooms.
- The columns 'cooking basics' and 'hot water' were checked for their existence and renamed to 'cooking_basics' and 'hot_water' for consistency.

1.3.7. Model Development for Price Prediction

A linear regression model (model1) was developed using the lm() function to predict prices. This model includes the number of bedrooms, bathrooms, location (Inner or Outer Richmond), and selected amenities as predictors.

1.3.8. Results

1. Price Prediction for Inner Richmond Area:
A prediction was made for a hypothetical property in the Inner Richmond area, considering 2 bedrooms, 2 bathrooms, and the presence of selected amenities.
The predicted price for this configuration in Inner Richmond was $366.41.

2. Price Prediction for Outer Richmond Area:
Similarly, a prediction was made for a property in the Outer Richmond area with the same characteristics.
The predicted price for this configuration in Outer Richmond was $378.30.

3. Average Price Calculation:
An average price of $372.36 was calculated for properties with similar characteristics in both neighborhoods.

This extended approach showcases how incorporating amenities into the regression model can provide a more nuanced understanding of factors influencing Airbnb prices in specific neighborhoods.

1.4. Recommendation

We advise adopting Model B for price prediction, where the price is determined by factors including beds, baths, location, and amenities. Our comparison of the Adjusted R Square values for both Model A and Model B, which are 0.008883 and 0.02816 respectively, supports this recommendation. Based on this analysis, we conclude that the appropriate pricing for the property should be set at **$372.36** per night.

1.5. Problem Statement (B)

'Should we list this property on Airbnb or Zillow?'

1.6. Approach Overview

From the results obtained by the models in Q1, we compared the price predicted in the model with the data in Zillow's online dashboard[1]. The dashboard shows the average monthly price of properties in the rental market in San Francisco, depending on the number of bedrooms of the listed properties.

1.7. Methodology - Mean Comparison

We identified the variable for comparison would be the number of bedrooms. Since the target property is a two-bedroom condo, we looked up the average monthly rent of two-bedroom apartments in the dashboard. After that, we divided the monthly rent by 30 to obtain the rent per night.

1.7.1. Results

The average monthly rent of two-bedroom properties as shown on the dashboard is $3895. The rent per night would be $129.83 in this case. Compared to the predicted price per night of the property if listed on Airbnb, which is $313.52, we could conclude that it would be more profitable if the property is listed on Airbnb instead of Zillow.

1.9. Limitations and Recommendations

We considered selecting 15 random properties with similar listing details, such as the number of bedrooms, amenities and location, to build a new dataset to analyze this problem statistically with R. However, considering the possibility of having sampling errors arose from this approach, as well as the comparability of the two datasets (Airbnb Listings with 7418 data points after filtering the listings in San Francisco while we only randomly selected 15 properties on Zillow), it would be more accurate to analyze the problem with the dashboard available on Zillow's website.

For homeowners who seek to rent their properties in short-term, it would be more profitable to list the properties on Airbnb since the rent per night would be higher than that on Zillow.

---

[1] https://www.zillow.com/rental-manager/market-trends/san-francisco-ca/?bedrooms=2

2. Percent of full-time properties listed on Airbnb

2.1. Problem statement

‘What percent of properties listed on Airbnb in San Francisco would you classify as “full time rental properties”? In other words, properties in which the owners never reside.’

2.2. Approach Overview

In the analysis to determine the percentage of full-time rental properties listed on Airbnb in San Francisco, we utilized the 'sflistings' dataset. Our approach involved applying specific criteria to identify properties that are likely used exclusively for rental purposes.

2.2.1. Data Observation and Criteria Identification

In examining this dataset, we focused on pinpointing what proportion of Airbnb listings are dedicated to full-time rental purposes. Our criteria for defining a full-time rental property include: (1) listing as an entire home, (2) a high availability index (availability exceeding 300 days annually), and (3) hosts managing more than one listing.

Further, an alternative filter was applied, based on an assumption from Airbnb data: approximately 50% of guests leave reviews. Therefore, we estimated actual guest visits as twice the number of received reviews. This secondary criterion involves (1) receiving at least five reviews in the past year, and (2) being available for rentals with a minimum duration of 30 days.

2.2.2. Methodology

The analysis was conducted using the R programming language, specifically employing the 'dplyr' package. The data was filtered to meet the above criteria, and calculations were performed to determine the percentage of properties meeting these definitions of full-time rentals.

2.2.3. Hypothesis Formulation

- Null Hypothesis (H0): There is no significant difference in the characteristics of full-time rental properties and other Airbnb listings in San Francisco.
- Alternative Hypothesis (H1): Full-time rental properties on Airbnb in San Francisco exhibit distinct characteristics compared to other listings, as determined by the established criteria.

2.2.4. Data Importation and Preparation

The Airbnb listings data is imported from a CSV file into R for analysis.

1. *Method 1: Applying the First Set of Filters:*

The first method filters the dataset based on primary criteria:

- Selection of entire homes or apartments.
- Application of a high availability filter (listings available more than 300 days a year).
- Filtering for listings managed by hosts with multiple listings.

The outcome of this method is a calculated percentage of 9.423025%, representing the proportion of listings that meet these criteria and are likely full-time rentals.

2. *Method 2: Implementing the Second Set of Filters:*

The second method uses a different filtering approach. Based on an assumption from Airbnb data: approximately 50% of guests leave reviews. Therefore, we estimated actual guest visits as twice the number of received reviews:

- Filtering listings based on a minimum number of 5 reviews received in the last 12 months.
- Applying a minimum stay requirement filter as 30 days or more.

This method results in a separate percentage of 3.181451%, reflecting the proportion of listings that align with these criteria.

2.2.5. Limitations and recommendations

From a statistical standpoint, the higher percentage from the first method might be seen as more representative of the overall market, capturing a wider spectrum of rental scenarios.
Conversely, the lower percentage from the second method indicates a more refined subset of the market, potentially offering a more precise estimate of properties predominantly used for full-time rentals.

Recommendations in this context depend on the specific objective of the analysis. If the goal is to understand the full scope of potential full-time rentals, including those that might also be used for short-term rentals, the higher percentage (9.423025%) is more informative.
However, if the focus is on identifying a more exclusive segment of the market, specifically properties more consistently used for rentals, the lower percentage (3.181451%) provides a more targeted insight.

3. Cheapest and Most Expensive Neighborhoods

3.1. Problem Statement

Where are the cheapest and most expensive neighborhoods to rent an Airbnb in the US and Canada?

3.2. Approach Overview

Our primary objective was to compile a thorough list highlighting the cheapest and most expensive neighborhoods within the dataset. In order to obtain a deeper understanding of the question, we also looked into the relationship between the neighborhood price and ratings.

3.2.1. Identification of the cheapest and most expensive neighborhoods

Data Observation and Variable Identification

We began by refining our dataset to include only Airbnb listings in the United States and Canada to ensure our subsequent analyses are directly relevant. We identified the crucial variables for the analysis would be (1) Country, (2) City, (3) Neighborhood and (4) Price.

Methodology - Mean Comparison

We grouped the listings based on Country, City, and Neighborhood, to ensure all neighborhoods in the United States and Canada were included. Then, we calculated the mean price per night for each neighborhood. By comparing the mean prices, we could identify the cheapest and most expensive neighborhoods.

After that, we used the head(10) and tail(10) functions to pinpoint the top ten neighborhoods associated with the lowest mean prices, as well as that with the highest mean prices respectively. Alternatively, arranging the results in descending order would also allow us to obtain the most expensive neighborhoods as well.

Results

The table below shows the top ten cheapest neighborhoods for Airbnb listings in the United States and Canada.

Ranking	Country	City	Neighborhood	Mean Price ($)
1	Canada	Portland	Pleasant Valley/Powellhurst-Gilbert	40

2	United States	New York City	New Dorp	40
3	United States	Los Angeles	Desert View Highlands	40.3
4	United States	Los Angeles	Watts	44.1
5	Canada	New Brunswick	Blissville	45
6	United States	San Mateo	Colma	49.2
7	United States	Oakland	Hegenberger	51
8	United States	Los Angeles	Historic South-Central	51.9
9	United States	Chicago	West Englewood	53.5
10	Canada	Portland	Centennial/Pleasant Valley	54

The table below shows the top ten most expensive neighborhoods for Airbnb listings in the United States and Canada.

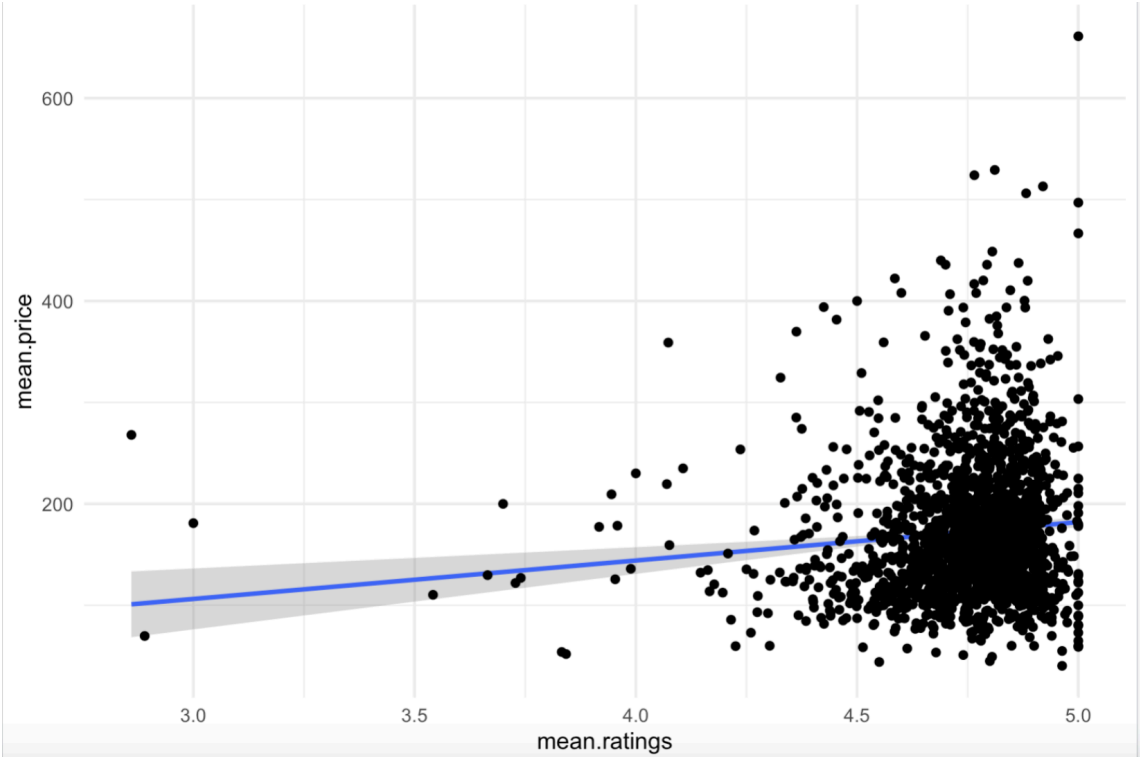
Ranking	**Country**	**City**	**Neighborhood**	**Mean Price ($)**
1	United States	San Diego	Eastlake Woods	661
2	United States	New York City	Fort Wadsworth	650
3	United States	Rhode Island	New Shoreham	529
4	United States	Oakland	Hiller Highlands	524
5	United States	Los Angeles	Aliso and Wood Regional Park	513
6	United States	Los Angeles	Hidden Hills	506
7	United States	Austin	UT Austin	500
8	Canada	Portland	Healy Heights/Southwest Hills	499
9	United States	New York City	Hollis Hills	497
10	United States	San Diego	Horton Plaza	485

3.2.2. Investigation of the relationship between the neighborhood price and ratings

Variable Identification and Data Visualization

In order to investigate if there is a relationship between the neighborhood price and ratings, we aggregated the variable for ratings in the data to obtain the mean ratings of all the listings. A new variable named "mean.ratings" was created as a result. The dependent variable is price, while the independent variable is the mean ratings.

The scatterplot below visualizes the relationship between these two variables.



Hypothesis Formulation

To test the relationship between the neighborhood price and ratings, the following hypotheses were proposed:

A. **Null Hypothesis:** There is no relationship between the neighborhood price and ratings.
B. **Alternative Hypothesis:** There is a relationship between the neighborhood price and ratings.

Methodology - Regression

The relationship between the neighborhood price and ratings was tested with the linear regression model.

Results

It shows that the p-value is 0.0000142, which means the null hypothesis is not true that there is indeed a relationship between neighborhood price and ratings. Ratings are a significant predictor of neighborhood price. Also, it shows the slope is 37.9, meaning that when the rating goes up by 1 score, the price will increase by $37.9 as a result.

3.3. Conclusion and Recommendations

The cheapest neighborhood is Pleasant Valley/Powellhurst-Gilbert in Portland, Canada with the mean price of $40 per night. The most expensive neighborhood is Eastlake Woods in San Diego, United States with the mean price of $661 per night. Through testing the relationship between the neighborhood price and ratings in the two analyses, we could confidently conclude that there is a relationship between these two variables, with ratings being a significant predictor of price of the Airbnb listings.

With that, homeowners in Airbnb in the United States and Canada can have a better understanding of the price per night of their properties listed on Airbnb is significantly affected by the neighborhood. They should also adjust the price of the listings according to the neighborhoods. On the other hand, the analysis for this question shows great opportunities for businesses and potential renters who are looking for economical accommodations.

4. Influence of Crime Rate on Airbnb Prices in San Francisco

4.1. Problem Statement

‘Does the crime rate affect the price of Airbnbs in San Francisco?’

4.2. Approach Overview

The analysis aims to determine whether there is a correlation between the crime rate in a neighborhood and the price of Airbnb listings in that area.

The study begins with an examination of Airbnb listings in San Francisco, correlating these with neighborhood crime data. This analysis seeks to understand how the external factor of crime rate influences Airbnb pricing.

4.2.1. Hypothesis Formulation

Null Hypothesis (H0): There is no statistically significant correlation between the crime rate in a neighborhood and the price of Airbnb listings in that area.
Alternative Hypothesis (H1): There is a statistically significant correlation between the crime rate in a neighborhood and the price of Airbnb listings in that area.

4.3. Data Preparation and Analysis

4.3.1. Data Loading and Transformation:

- Airbnb listing data was loaded from sf_bnb_listing.csv.
- The price field was cleaned and converted from a character string to a numeric value for accurate analysis.

4.3.2. Calculation of Average Prices:

- Using the dplyr package, the Airbnb data was grouped by host_neighbourhood.
- The average price for listings in each neighborhood was calculated, omitting any missing values.

4.3.3. Integration with Crime Data:

- Crime data was loaded from revisedcrimedata.csv.
- The crime data was then merged with the average Airbnb prices based on the host neighborhood to create a combined dataset, bnb_crime.

4.4. Statistical Analysis and Model Development

4.4.1. Linear Regression Model

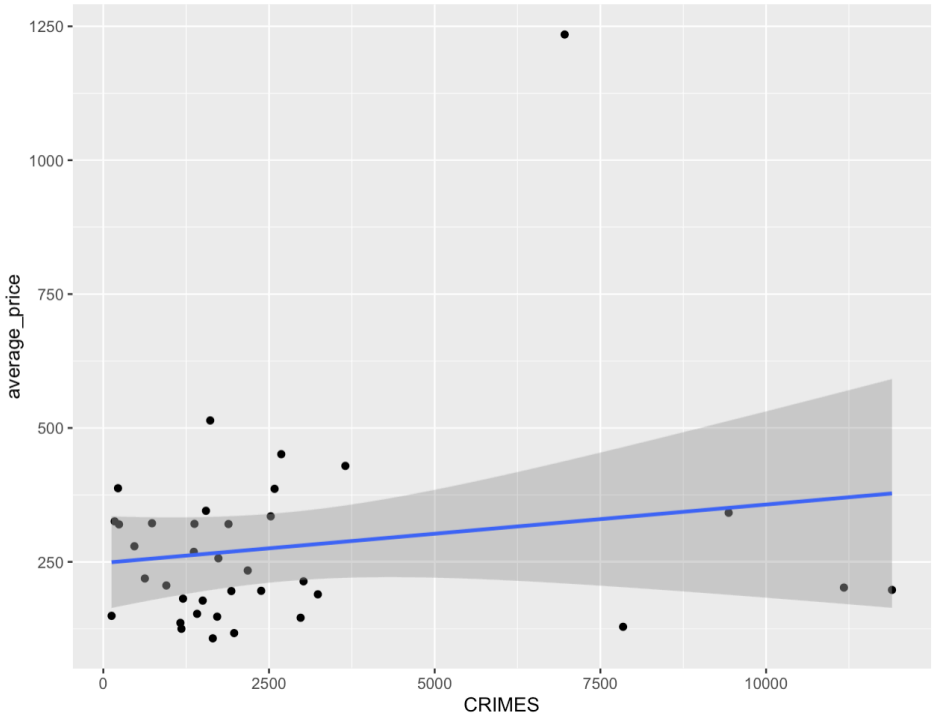
- A linear regression model was constructed to assess the relationship between average Airbnb prices (average_price) and crime rates (CRIMES).
- The summary() function was used to analyze the model, focusing on coefficients, significance levels, and overall model fit.

4.4.2. Correlation Analysis

- Pearson correlation was calculated between the average price and crime rates to quantify their linear relationship.
- The correlation value was printed for interpretation.

4.4.3. Visualization

- A scatter plot was created using ggplot, displaying the relationship between crime rates and average Airbnb prices.
- A linear model was superimposed on the plot to visually assess the trend.



4.5. Conclusion

Coefficient of 'CRIMES': The coefficient for the CRIMES variable is 0.01090, which suggests that for every additional crime reported in an area, the average Airbnb price increases by approximately 0.0109 units. However, this increase is very small and may not be practically significant.

P-value of 'CRIMES': The p-value associated with the CRIMES coefficient is 0.324, which is much higher than the conventional alpha level of 0.05. This high p-value indicates that the relationship between crime rate and Airbnb prices is not statistically significant.

Correlation Analysis: The Pearson correlation coefficient is 0.166863, suggesting a weak positive correlation between crime rates and average Airbnb prices. However, this correlation is not strong enough to imply a meaningful or significant relationship.

The findings suggest that there is no significant statistical relationship between the number of crimes in an area and the average price per night of Airbnb listings in that area. The weak correlation coefficient reinforces the conclusion that crime rate is not a strong predictor of Airbnb pricing in the studied dataset. Therefore, it would be prudent to consider other variables alongside or instead of crime rates when analyzing factors that affect Airbnb pricing.

5. Appendix

A: Data

B: Code