NYC AIRBNB DATA ANALYSIS

Lisa Siefker

Statistical Analysis: Jiwon Hong

PROJECT 2

AU24 GTDA 5401

INTRODUCTION

Part I of this project focused on a subset of New York City Airbnb data that included at least two bedrooms, one bathroom, and minimum nights less than 4. In that subset, the analysis showed:

- Price per night was right skewed across all boroughs
- The most expensive outliers were located in Manhattan
- The median price was fairly consistent across all boroughs
- Customer ratings were consistently above four stars on a five star scale across all boroughs
- Customer ratings were consistently above four stars across all price points

Part II of the project asked a consultant the following research questions regarding the entire Airbnb dataset:

1. Can you predict an Airbnb rating based on price per night and neighborhood/borough?

2. Can you predict the price per night of an Airbnb based on neighborhood/borough, and number of bedrooms?

3. Is there a significant difference in average price between Manhattan and Brooklyn Airbnbs that have 2 bedrooms and 1 bathroom?

4. Is there a significant difference in average price between Airbnbs that have average reviews that are more than four stars and Airbnbs that have average reviews that are less than 4 stars?

DATA DESCRIPTION

The project focused on the "Airbnb in NYC" dataset posted on Kaggle (https://www.kaggle.com/datasets/vrindakallu/new-york-dataset). The data summarizes Airbnb listing activity in New York City as of January 5, 2024. There are 20,759 rows of data comprised of the following key variables:

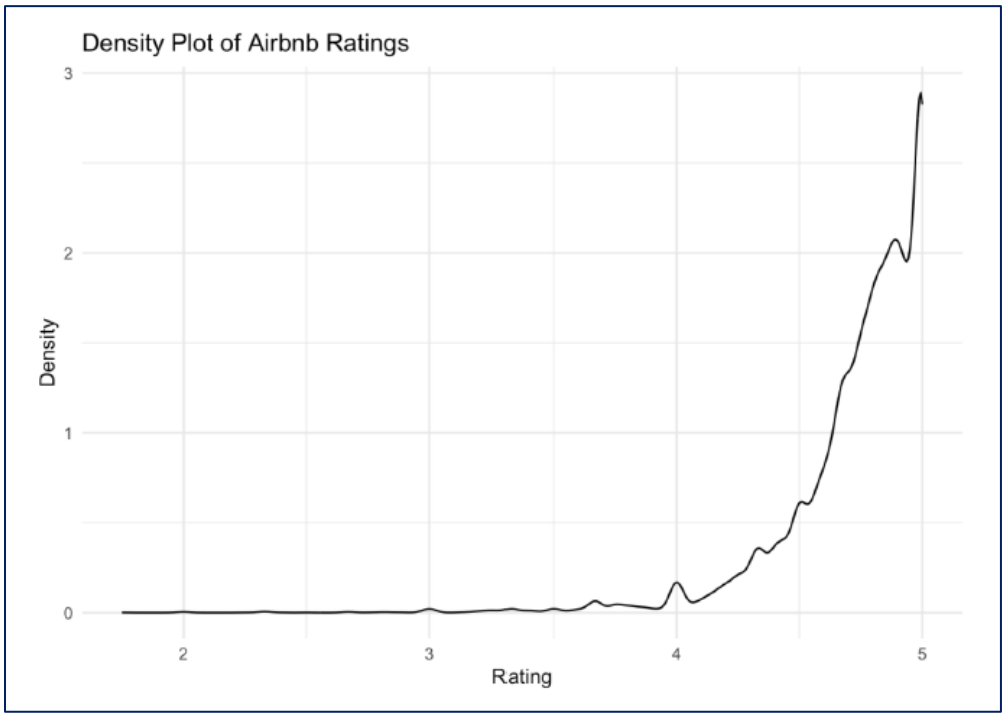
• Neighborhood Group (5 boroughs of NYC)
• Neighborhood (219 unique neighborhoods)
• Type of Listing (e.g., private/shared room)
• Price per night
• Minimum Nights (minimum nights required for a reservation)
• Number of Reviews (total number of reviews)
• Availability_365 (number of days listing is available each year)
• Rating (average total rating)
• Number of Bedrooms
• Number of Beds
• Number of Bathrooms

METHODOLOGY & RESULTS

After preprocessing the data, the consultant used linear regression, log transformation, and two sample t-tests to answer the research questions. Although these models were not the most optimal, they were chosen due to ease of understanding.

Question 1: Can you predict an Airbnb rating based on price per night and neighborhood/borough?

A density plot illustrates that the ratings data is very left-skewed, but the ratings data was not transformed for simplicity of analysis.



Linear regression was used to answer Question 1. Linear regression models the relationship between a dependent/response variable (rating) and one or more independent variables/predictors by minimizing the squared differences between the observed and predicted values.

```
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             4.712e+00  1.740e-02 270.795  < 2e-16
## neighbourhood_groupBrooklyn             2.291e-02  1.828e-02   1.253  0.21024
## neighbourhood_groupManhattan           -5.608e-02  1.810e-02  -3.099  0.00195
## neighbourhood_groupQueens               1.238e-02  1.902e-02   0.651  0.51497
## neighbourhood_groupStaten Island        3.337e-02  3.401e-02   0.981  0.32649
## price                                   9.043e-05  1.301e-04   0.695  0.48717
## neighbourhood_groupBrooklyn:price       1.542e-04  1.329e-04   1.160  0.24608
## neighbourhood_groupManhattan:price      8.223e-05  1.311e-04   0.627  0.53054
## neighbourhood_groupQueens:price         9.616e-06  1.379e-04   0.070  0.94441
## neighbourhood_groupStaten Island:price  1.651e-04  2.312e-04   0.714  0.47514
""
```

The p-values in the model summary show that only the Manhattan neighborhood group is a significant predictor of rating. This suggests that Airbnbs in Manhattan tend to have slightly lower ratings as compared to the baseline group (the Bronx). The baseline group serves as the reference category that other groups are compared against. By default, the baseline group is the first level of a categorical variable in alphabetical order. No other neighborhoods, or the price variable, showed statistical significance as predictors in the model.

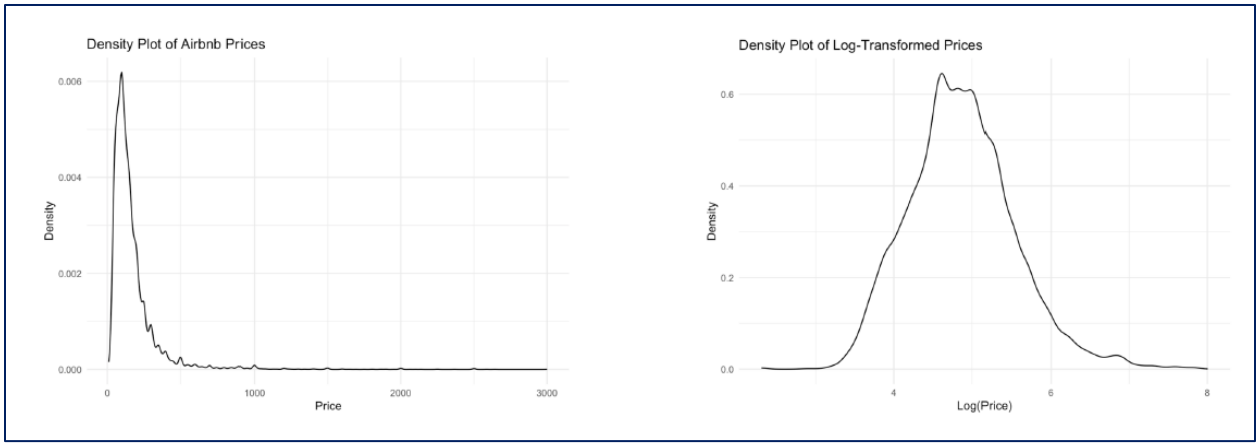
```
##
## Residual standard error: 0.2901 on 16979 degrees of freedom
## Multiple R-squared:  0.02674,    Adjusted R-squared:  0.02623
## F-statistic: 51.84 on 9 and 16979 DF,  p-value: < 2.2e-16
```

The “Adjusted R-squared” value indicates the proportion of variability in ratings explained by the model. The summary output indicated that the model only explains 2.6% of the variance in rating after accounting for the number of predictors in the model. This is very low, suggesting that neighborhood and price do not have a strong linear association with rating and the model’s prediction ability is weak.

Key Insight: While the model can technically predict ratings, its explanatory power is very weak. Most of the variation in Airbnb ratings is not captured by price, neighborhood/borough, or their interaction.

Question 2: Can you predict the price per night of an Airbnb based on neighborhood/borough and number of bedrooms?

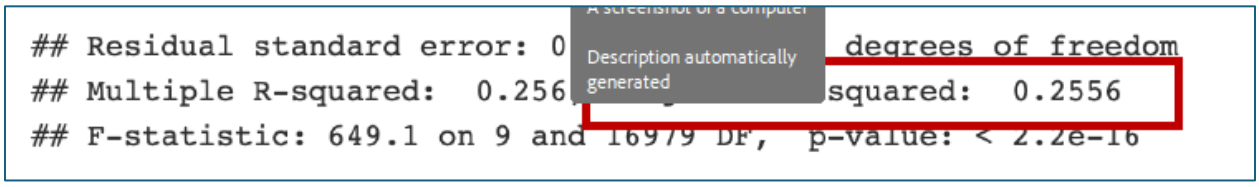
A density plot of price data showed a strong right-skew. A logarithmic transformation was applied to address skew, stabilize the variance, and improve model performance.



After transformation, linear regression was used to model the relationship between log transformed price and neighborhood or number of bedrooms. The model summary showed that neighborhood groups Brooklyn and Manhattan and the bedroom variable were significant predictors of price. This suggests that prices in Brooklyn and Manhattan are significantly higher than the baseline neighborhood group, the Bronx. In addition, each additional bedroom increased the log transformed price by 0.373.

```
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      4.031291   0.041854  96.318  < 2e-16
## neighbourhood_groupBrooklyn      0.354022   0.043988   8.048 8.96e-16
## neighbourhood_groupManhattan     0.663751   0.043692  15.191  < 2e-16
## neighbourhood_groupQueens        0.029052   0.046513   0.625   0.5322
## neighbourhood_groupStaten Island 0.039361   0.078394   0.502   0.6156
## bedrooms                         0.373212   0.028226  13.222  < 2e-16
## neighbourhood_groupBrooklyn:bedrooms     -0.046424   0.029354  -1.582   0.1138
## neighbourhood_groupManhattan:bedrooms    -0.059068   0.029532  -2.000   0.0455
## neighbourhood_groupQueens:bedrooms        0.037047   0.031000   1.195   0.2321
## neighbourhood_groupStaten Island:bedrooms 0.001739   0.047514   0.037   0.9708

## Residual standard error: 0.5872 on 16979 degrees of freedom
## Multiple R-squared:  0.256,  Adjusted R-squared:  0.2556
## F-statistic: 649.1 on 9 and 16979 DF,  p-value: < 2.2e-16
```



The “Adjusted R-squared” value indicates that about 25.6% of the variance in log transformed price is explained by the model, which is a moderate level of explanatory power. The predictions for actual prices are somewhat imprecise.

Key Insight: Airbnbs in Brooklyn and Manhattan tend to have significantly higher prices than the baseline group, and additional bedrooms lead to higher prices overall.

Question 3: Is there a significant difference in average price between Manhattan and Brooklyn Airbnbs that have 2 bedrooms and 1 bathroom?

A two-sample t-test was used to answer Question 3. Two sample t-tests are used to compare the means of two groups to determine if there is significant difference between them. Welch's t-test was used to adjust for the unequal sample sizes and variance. The Welch t-test is more robust and reliable than a traditional t-test under those conditions.

The two-sample t-test compared average price between Manhattan and Brooklyn Airbnbs with two bedrooms and one bathroom. The results of the t-test show a very small p-value, which indicates high statistical significance. The 95% confidence interval (-38.866, -15.955) does not include 0. Therefore, the difference in average prices in Brooklyn and Manhattan is significantly different than zero.

```
##
##  Welch Two Sample t-test
##
## data:  price by neighbourhood_group
## t = -4.6941, df = 1275.4, p-value = 2.967e-06
## alternative hypothesis: true difference in means between group Brooklyn and group Manhattan is not equal to 0
## 95 percent confidence interval:
##  -38.86596 -15.95453
## sample estimates:
##  mean in group Brooklyn mean in group Manhattan
##                189.4693                216.8795
```

Answer: There is a significant difference in average price between Airbnbs in Manhattan and Brooklyn. On average, Airbnbs are $27.41 more expensive in Manhattan than those in Brooklyn.

Question 4: Is there a significant difference in average price between Airbnbs that have average reviews that are more than four stars and Airbnbs that have average reviews that are less than 4 stars?

A two-sample Welch t-test was used to answer Question 4 because the variances for each group were different. The results of the test show a p-value very close to 0, which indicates high statistical significance. The 95% confidence interval (33.984, 53.492) does not include 0. Therefore, the difference is average price is significantly different than zero.

```
##
##  Welch Two Sample t-test
##
## data:  price by review
## t = 8.8128, df = 445.05, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Above 4 and group Below 4 is not equal to 0
## 95 percent confidence interval:
##  33.98426 53.49195
## sample estimates:
## mean in group Above 4 mean in group Below 4
##              169.7985              126.0604
```

Answer: There is a significant difference in average price between Airbnbs that have average reviews more than four stars and Airbnbs that have average reviews that are less than four stars. On average, Airbnbs with reviews above four stars are $43.74 more expensive than those with reviews before four stars.

Conclusion

Linear regression, log transformation and two-sample t-tests were chosen to analyze the data due for simplicity and ease of understanding. The models and test showed that:

- Price and neighborhood do not strongly influence Airbnb ratings based on the model
- Manhattan Airbnbs have slightly lower ratings than the reference group, but the difference is minimal.
- Airbnbs in Brooklyn and Manhattan tend to have significantly higher prices than the baseline group.
- Additional bedrooms lead to higher prices overall.
- There is a significant difference in average price between Manhattan and Brooklyn Airbnbs with 2 bedrooms and 1 bathroom. Manhattan Airbnbs are, on average, $27.41 more expensive than those in Brooklyn.
- There is a significant difference in average price between Airbnbs that have average reviews above 4 stars and those below 4 stars. Airbnbs with reviews above four stars are, on average, $43.74 mor expensive than those with reviews below four stars.