FINAL PROJECT

AIRBNB DATA ANALYSIS

STATISTICS 501

Joanna Jeneralczuk

- Kien Le, Tony Nguyen -

I. INTRODUCTION

For this project, we have decided to take a dataset from Kaggle - a platform for Data Science with a lot of open and interesting datasets to explore. We have come through a dataset of New York City's Airbnb dataset. The dataset has 16 columns in total and approximately 50,000 rows of listing through Airbnb API. The New York City Airbnb Open Data dataset from Kaggle contains information on Airbnb listings in New York City as of 2019. The dataset includes 16 columns, each representing a different aspect of the listing, such as the listing ID, the host ID, the name of the listing, the neighborhood group, the neighborhood, the latitude and longitude coordinates, the room type, the price, the minimum number of nights, the number of reviews, the last review date, the reviews per month, the calculated host listings count, and the availability of the listing in the next 365 days.

We have come up with our *research question* which is: "How will the price per night for a listing in Aribnb be affected by other factors?" and also "How will the average price between neighborhood groups in NYC differ?"

In this project, by conducting *hypothesis tests for the difference between* means, we can identify significant variations in average prices between different neighborhood groups, and specific neighborhoods. Applying *ANOVA techniques* allows us to explore how the average prices of Airbnb listings differ across neighborhood groups or specific neighborhoods. Finally, through *linear regression analysis*, we can model the relationship between the price (dependent variable) and independent variables such as room type,

minimum nights, host listings count, and more. This enables us to quantify the impact of these variables on the price and gain insights into how different factors influence listing prices. Hosts can use this information to optimize pricing strategies, while potential guests can assess the value they receive for a given price.

II. DATA

The dataset includes 16 columns, each representing a different aspect of the listing. Here's a brief overview of the variables:

- listing_id: The unique identifier for each listing.
- host_id: The unique identifier for each host.
- name: The name or title of the listing.
- neighbourhood_group: The larger area or borough in which the listing is located (e.g., Manhattan, Brooklyn, etc.).
- neighborhood: The specific neighborhood within the borough where the listing is situated.
- latitude: The latitude coordinate of the listing's location.
- longitude: The longitude coordinate of the listing's location.
- room_type: The type of room being listed (e.g., entire home/apartment, private room, shared room).
- price: The price per night for the listing.
- minimum_nights: The minimum number of nights required to book the listing.
- number_of_reviews: The total number of reviews received for the listing.
- last_review: The date of the last review.
- reviews_per_month: The average number of reviews per month.
- calculated_host_listings_count: The total number of listings managed by the host.
- availability_365: The number of days the listing is available for booking within the next 365 days.

The NYC dataset contains information on Airbnb listings in New York City as of 2019. In terms of dataset units, each row in the dataset represents a single listing on Airbnb and, therefore, one unit of the dataset corresponds to one Airbnb listing in New York City. We decided to choose our response variable as “Price” and the key explanatory variable

initially could be neighbourhood_group, neighborhood, room_type, minimum_nights required for booking, or the number of reviews for that listing.

III. ANALYSIS

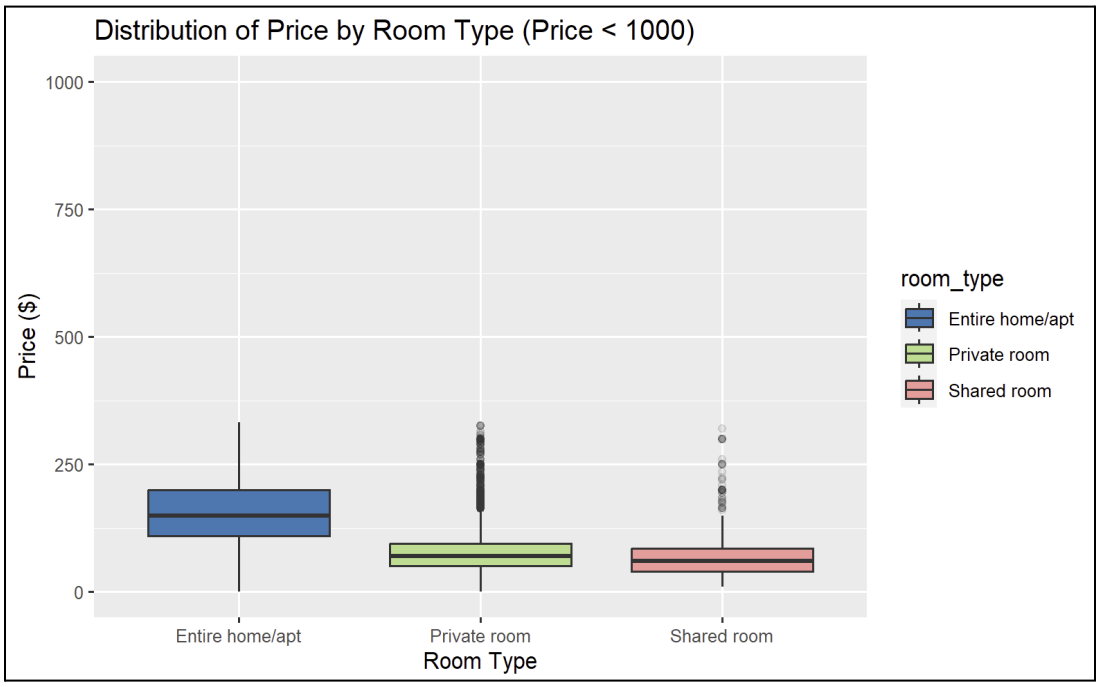
In terms of analysis for our project, we will solely use R to perform statistical testing and modeling of the data. This section consists of 4 sections which are: Preliminary Analysis of the Dataset, Hypothesis test of the difference between 2 means (Manhattan & Brooklyn), ANOVA, and Simple Linear Regression (Price as dependent variable).

1. Preliminary Analysis of the Dataset

We decided to take a look into our summary of statistics of our dataset and specifically the "Price" variable which we choose to work with for our research question. Based on the R output for summary (*Appendix code 1*), the population mean is $122.1

After that, we generated a distribution graph for "Price" with R (*see code 2 Appendix*). As we can see, our graph is a bit right-skewed. The peak density of price focuses on the range of $70 - $100 per night on Airbnb listings in New York.

We are also interested in seeing how the Price is distributed based on our column "Room Type" (the type of room listed on Airbnb). We ended up generating a box plot for Price based on Room Type below by R (*see Appendix code 3*).



2. Hypothesis test of the difference between 2 means (Manhattan & Brooklyn)

We have come up with our Null hypothesis and Alternative hypothesis:

H0: The true difference in means is equal to 0

Ha: The true difference in means is not equal to 0

We began by filtering out our dataset into two separate datasets that only contain Price for Manhattan and the other for Brooklyn neighborhood (*see Appendix, code 4*). We performed t-test and the output will be included in our Appendix (*see code 5*).

The test statistic (t) is 47.064 with degrees of freedom (df) equal to 29577. The p-value is reported as less than 2.2e-16, indicating that the p-value is extremely small and significantly less than the conventional threshold of 0.05. This implies strong evidence to reject the null hypothesis.

The alternative hypothesis states that the true difference in means is not equal to 0, which aligns with the finding of a significant p-value. Therefore, we can conclude that there is a statistically significant difference between the mean prices of listings in Manhattan and Brooklyn.

The 95 percent confidence interval for the mean difference between the two groups is calculated to be (33.84393, 36.78538). This interval suggests that we can be 95 percent confident that the true mean difference in prices between Manhattan and Brooklyn lies within this range.

The sample estimates indicate that the mean price of listings in Manhattan (mean of x) is 144.8019, while the mean price of listings in Brooklyn (mean of y) is 109.4872.

Overall, these results did show a substantial and statistically significant difference in the average prices of Airbnb listings between Manhattan and Brooklyn.

3. ANOVA:

Firstly, we use analysis of variance (ANOVA) to investigate if there are significant differences in the average price among different neighborhood groups.

Secondly, we use analysis of variance to see if there are significant variations in the average number of reviews per month across different room types.

Strong evidence indicates significant variations in both the average number of reviews per month across different room types and the average price among different neighborhood groups. The p-values for both analyses are extremely small ($p < 1.79e-10$ and $p < 2e-16$, respectively), indicating that these differences in averages are highly unlikely to occur by chance.

Additionally, the F-values of 22.45 and 1507, along with their associated p-values, confirm that the variation in both reviews and price explained by the respective variables

(room type and neighborhood group) is significantly larger than the variation observed within the groups (residuals) (*see Appendix, code 6 & 7*).

4. **Simple Linear Regression:**

We use simple linear regression to model the relationship between the number of reviews or minimum nights required or availability_365 with the price of listings (try to create a scatterplot and cover as many columns in the dataset as possible) and also insert a regression line (use method = lm) (*see Appendix, code 8*).

According to the output for the linear regression, here are some key findings:

- The intercept term (-1.711e+04) represents the estimated price when all the predictor variables are zero.

- The neighborhood groups have significant effects on the price. Compared to the reference group, "neighbourhood_groupBrooklyn" has a negative effect on price, while "neighbourhood_groupManhattan" and "neighbourhood_groupQueens" have positive effects. "neighbourhood_groupStaten Island" has a strong negative effect on price.

- Latitude and longitude also have significant effects on the price. Higher latitude values have a negative effect, while higher longitude values have a positive effect.

- The type of room ("room_typePrivate room" and "room_typeShared room") has a significant effect on the price. Both room types have negative effects compared to the reference group.

- The number of reviews, availability throughout the year, and the count of listings by the host all have significant effects on the price. As the number of reviews and availability increase, the price tends to decrease. On the other hand, as the count of listings by the host increases, the price tends to increase.

- The adjusted R-squared value of 0.4825 indicates that the predictors explain approximately 48.25% of the variability in the price after accounting for the degrees of freedom.

Then, predict the price of a listing (use output regression and then predictions from INFO 248) (*see Appendix, code 9*).

The RMSE value obtained is 48.73927. This value represents the average difference between the actual prices of the listings in the testing set and the predicted prices by the model. A lower RMSE indicates better predictive performance, as it represents smaller prediction errors. In this case, the RMSE of 48.73927 suggests that the model, on average, predicts the price of a listing with an error of approximately $48.74.

IV. CONCLUSION

In conclusion, our analysis of the Airbnb data revealed several significant findings related to the factors influencing price and popularity.

Firstly, we found that the neighborhood group has a significant effect on the average price of listings, indicating that different areas in New York City exhibit varying price levels. Specifically, Manhattan and Brooklyn showed a statistically significant difference in mean prices, with Manhattan having higher average prices than Brooklyn.

Secondly, the room type was found to significantly influence the number of reviews per month, suggesting that different types of rooms attract different levels of attention from guests.

Additionally, our linear regression analysis showed that variables such as neighborhood group, location coordinates, minimum nights, room type, number of reviews, availability, and host listings count collectively explain approximately 48.26% of the variance in price. This indicates that these factors play a substantial role in determining the listing price.

Overall, this analysis provides valuable insights into the factors influencing price and popularity in the Airbnb market.

CODE APPENDIX

Code 1:

```
```{r}
summary(data_no_OL)
```
```

```
   neighbourhood_group           room_type         price        minimum_nights     number_of_reviews reviews_per_month
 Bronx        :  776    Entire home/apt:18758   Min.    :  0.0   Min.    :   1.000   Min.    :  0.00    Min.    : 0.000
 Brooklyn     :15370    Private room   :15858   1st Qu.: 70.0   1st Qu.:   1.000   1st Qu.:  1.00    1st Qu.: 0.040
 Manhattan    :14927    Shared room    :  605   Median :100.0   Median :   2.000   Median :  5.00    Median : 0.330
 Queens       : 3903                           Mean   :122.1   Mean   :   5.842   Mean   : 22.86    Mean   : 1.039
 Staten Island:  245                           3rd Qu.:160.0   3rd Qu.:   4.000   3rd Qu.: 23.00    3rd Qu.: 1.460
                                               Max.   :332.0   Max.   :1250.000   Max.   :629.00    Max.   :20.940
 calculated_host_listings_count availability_365
 Min.    :  1.000               Min.    :  0.00
 1st Qu.:  1.000               1st Qu.:  0.00
 Median :  1.000               Median : 11.00
 Mean   :  1.304               Mean   : 85.14
 3rd Qu.:  1.000               3rd Qu.:156.00
 Max.   :327.000               Max.   :365.00
```

Code 2:

```
```{r}
# Create a histogram plot of the price distribution
ggplot(data_no_OL, aes(x = price)) +
  geom_histogram(fill = color_palette[4], color = "white", bins = 10) +
  scale_fill_manual(values = color_palette) +
  labs(x = "Price", y = "Count", title = "Distribution of Price") +
  theme_minimal()
```
```

Code 3:

```
```{r echo=FALSE, message = FALSE}
#plot distribution of Room Types in NYC Airbnb Listings with price smaller than $1000
ggplot(data = data_no_OL, aes(x=room_type, y=price, fill=room_type)) +
  geom_boxplot(outlier.alpha=0.1) +
  labs(title="Distribution of Price by Room Type (Price < 1000)", x="Room Type", y="Price ($)") +
  scale_fill_manual(values=c("#1f78b4", "#b2df8a", "#fb9a99")) +
  ylim(0, 1000)
```
```

Code 4:

```
```{r}
manhattan_data <- subset(data_no_OL, neighbourhood_group == "Manhattan")
brooklyn_data <- subset(data_no_OL, neighbourhood_group == "Brooklyn")
```
```

Code 5:

```
```{r}
# Perform t-test to compare the means of price in Manhattan and Brooklyn
t_test_result <- t.test(manhattan_data$price, brooklyn_data$price)
t_test_result
```
```

```
        Welch Two Sample t-test

data:  manhattan_data$price and brooklyn_data$price
t = 47.064, df = 29577, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 33.84393 36.78538
sample estimates:
mean of x mean of y
 144.8019  109.4872
```

Code 6:

```
```{r echo=FALSE, message = FALSE}
model <- aov(price ~ neighbourhood_group, data = data)
summary(model)
```
```

```
                     Df    Sum Sq Mean Sq F value Pr(>F)
neighbourhood_group     4  24729714 6182429    1507 <2e-16 ***
Residuals           45913 188323312    4102
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Code 7:

```{r echo=FALSE, message = FALSE}
model <- aov(reviews_per_month ~ room_type, data = data)
summary(model)
```

```
                Df Sum Sq Mean Sq F value   Pr(>F)
room_type        2    114   57.23   22.45 1.79e-10 ***
Residuals    48892 124629    2.55
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Code 8:

```{r echo=FALSE, message = FALSE}
model <- lm(price ~ neighbourhood_group + latitude + longitude + minimum_nights + room_type +
number_of_reviews + availability_365 + calculated_host_listings_count, data = data_no_OL)
summary(model)
```

```
Call:
lm(formula = price ~ neighbourhood_group + latitude + longitude +
    minimum_nights + room_type + number_of_reviews + availability_365 +
    calculated_host_listings_count, data = data_no_OL)

Residuals:
     Min       1Q   Median       3Q      Max
-175.376  -30.553   -7.956   20.769  268.226

Coefficients:
                                 Estimate Std. Error  t value Pr(>|t|)
(Intercept)                    -1.711e+04  7.009e+02  -24.409  < 2e-16 ***
neighbourhood_groupBrooklyn    -8.116e+00  1.913e+00   -4.242 2.22e-05 ***
neighbourhood_groupManhattan    2.222e+01  1.728e+00   12.857  < 2e-16 ***
neighbourhood_groupQueens       4.958e+00  1.836e+00    2.701  0.00692 **
neighbourhood_groupStaten Island -7.837e+01  3.640e+00  -21.528  < 2e-16 ***
latitude                       -7.523e+01  6.889e+00  -10.921  < 2e-16 ***
longitude                      -2.748e+02  7.880e+00  -34.872  < 2e-16 ***
minimum_nights                 -2.134e-01  1.183e-02  -18.034  < 2e-16 ***
room_typePrivate room          -7.563e+01  4.768e-01 -158.622  < 2e-16 ***
room_typeShared room           -1.009e+02  1.497e+00  -67.396  < 2e-16 ***
number_of_reviews              -5.044e-02  5.203e-03   -9.693  < 2e-16 ***
availability_365                5.770e-02  1.900e-03   30.365  < 2e-16 ***
calculated_host_listings_count  9.309e-02  7.844e-03   11.867  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49 on 45905 degrees of freedom
Multiple R-squared:  0.4826,    Adjusted R-squared:  0.4825
F-statistic:  3569 on 12 and 45905 DF,  p-value: < 2.2e-16
```

Code 9:

```{r echo=FALSE, message = FALSE}
set.seed(123)
split <- sample.split(data_no_OL, SplitRatio = 0.8)
train <- subset(data_no_OL, split == TRUE)
test <- subset(data_no_OL, split == FALSE)

# Use the model to predict prices on the testing set
predictions <- predict(model, newdata = test)

# Calculate the root mean squared error (RMSE) to evaluate the performance of the model
rmse <- sqrt(mean((test$price - predictions)^2))
print(rmse)
```

```
[1] 48.73927
```

```
```