NYC AIRBNB DATA ANALYSIS

Lisa Siefker

PROJECT 1

AU24 GTDA 5401

Introduction

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

- Neighborhood Group (5 unique groups)
- 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

The analysis will attempt to answer the following question:

Which Airbnb is the best deal (i.e., the rating is above average and lowest price) in each neighborhood? The analysis will only consider Airbnbs that meet my family’s criteria: entire home/apt, at least 2 beds, at least 1 bathroom, minimum nights must be less than 4.

Methodology

To answer the research question, the first step is to clean the data and make sure it is in a tidy format. I will review the data structure and determine whether any variables need to be converted to a different data type. I will also remove any variables that aren’t relevant to the analysis to remove clutter from the output. Next, I will filter the data to remove rows that don’t meet my family’s criteria. Lastly, I will explore and analyze ratings, price and neighborhood information using tables, histograms, boxplots, scatterplots, and statistical summary data to determine which Airbnbs are the best deal (i.e., the rating is above average and the price is below average) in each neighborhood.

Results & Interpretation

First, the data is converted to a tibble for cleaner output and ease of use with tidyverse functions. The structure is examined to identify variables that need to be converted to a different data type. Next, the data is cleaned and filtered. Columns that are not relevant to the analysis are removed, spelling of a variable is corrected, numeric data is converted from character to numeric, and rows with missing data are removed. Finally, the data is filtered so that only Airbnbs that minimally meet my family’s criteria are left in the tibble. These steps reduced the tibble from 20,758 rows to 574 rows.

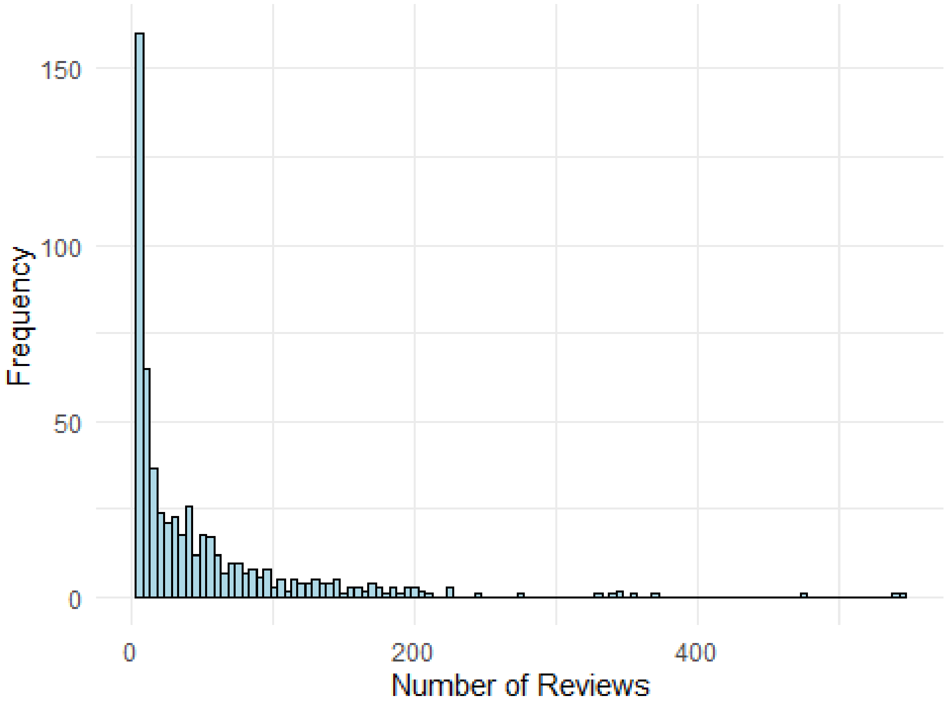
```
#check dimensions of new tibble
dim(data4)

## [1] 574  12
```

Next, I investigate the number of reviews that each Airbnb received. I only want to include Airbnbs that have received at least one review. The summary statistics show that the minimum number of reviews in the data set is three, so there is no need to further filter the data based on number of reviews. The histogram and summary statistics show that the data is strongly right skewed. 25% of Airbnbs in the data set received at least 3 and less than 7 reviews.

```
#Explore number of reviews - make sure no rows with 0 reviews
summary(data4$number_of_reviews)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    3.00    7.00   23.00   48.98   61.00  544.00
```



Next, the analysis examines the difference between the ‘neighborhood’ and ‘neighborhood_group’ variables in the filtered dataset. There are 73 unique neighborhoods, and those neighborhoods are grouped into five ‘neighborhood_groups’, which I determine are the five boroughs of New York City (Manhattan, Brooklyn, Queens, Staten Island, and the Bronx). The large number of unique neighborhoods will make it unwieldy to complete a grouped analysis based on the ‘neighborhood’ variable. Therefore, further analysis will focus on the ‘neighborhood_group’ variable, which I rename ‘borough’ for clarity.

In the filtered dataset, Brooklyn and Manhattan encompass the majority of the remaining Airbinbs at 166 (28.9%) and 366 (63.8%) respectively. The Bronx and Staten Island together comprise only 6 (1.0%) of the Airbnbs remaining in the filtered dataset.

```
#explore neighborhood and neighborhood_group variables
#determine number of unique neighborhoods and unique neighborhood_groups
print(length(unique(data4$neighborhood)))

## [1] 74

print(length(unique(data4$borough)))

## [1] 5

#what percentage of Airbnbs are in each borough in the filtered dataet?
print(filtered_result)

##        Category Count Percentage
## 1         Bronx     4        0.7
## 2      Brooklyn   166       28.9
## 3     Manhattan   366       63.8
## 4        Queens    36        6.3
## 5 Staten Island     2        0.3
```

I compare the number of Airbnbs in each borough in the filtered dataset to the number of Airbnbs in each borough in the original dataset to determine whether my filters skewed the data toward a particular borough. Airbnbs in Manhattan were more likely to meet my filter criteria than those in other boroughs. Manhattan Airbnbs made up 38.7% of the original dataset, and my filters increased the percentage of Manhattan Airbnbs to 63.8%.

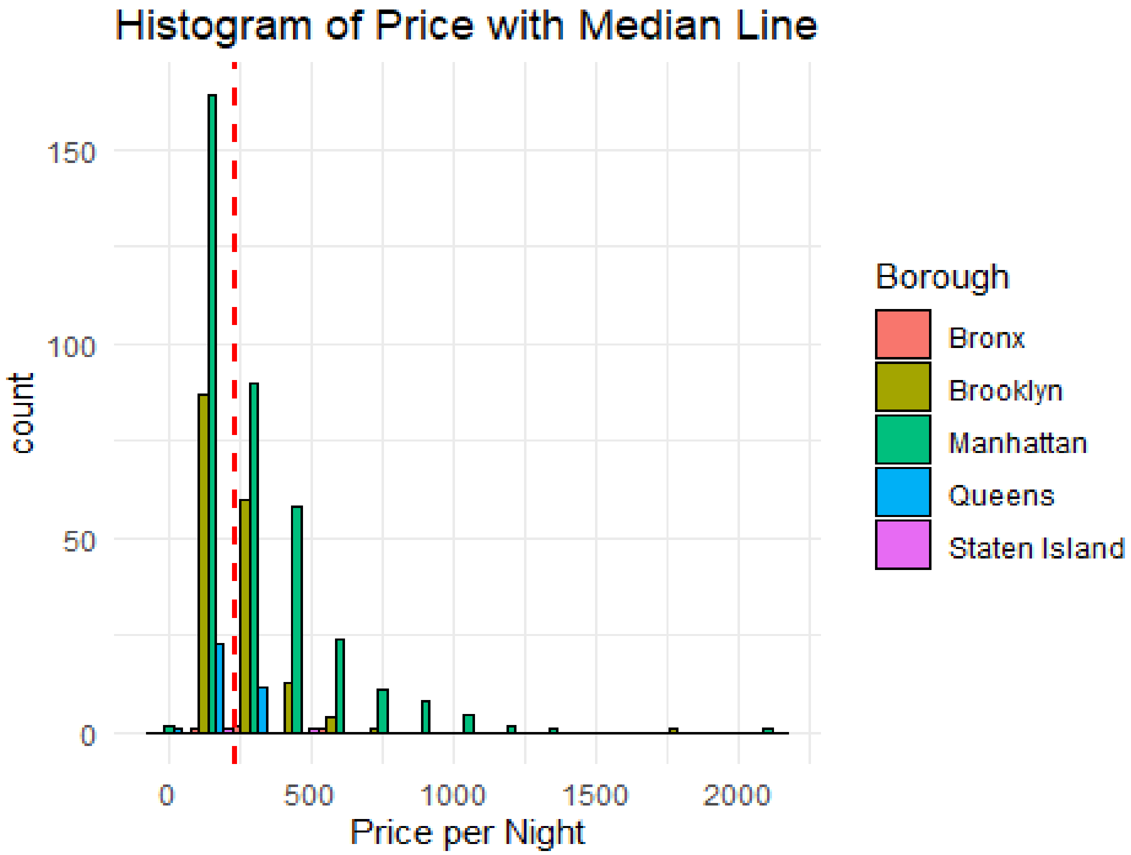
```
#what percentage of Airbnbs are in each borough in the original dataet?
print(result_original)

##        Category Count Percentage
## 1         Bronx   949        4.6
## 2      Brooklyn  7719       37.2
## 3     Manhattan  8038       38.7
## 4        Queens  3761       18.1
## 5 Staten Island   291        1.4
```

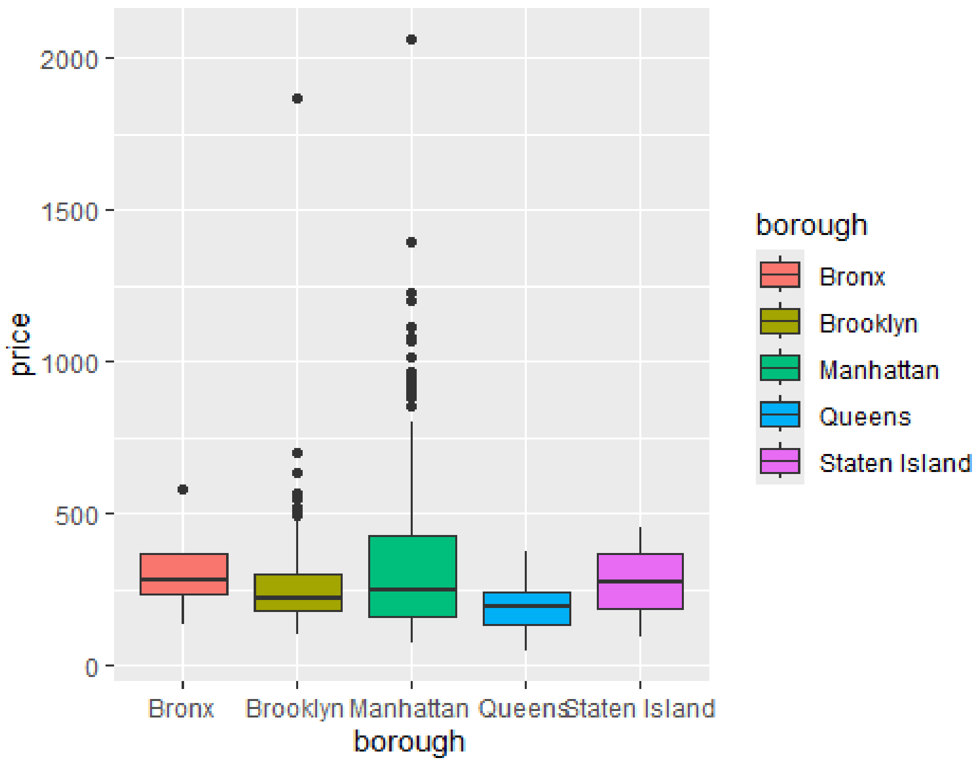
The below summary statistics and histogram show that the price per night of the filtered data is right skewed, with a median price of $235.50/night across all boroughs. Manhattan's Airbnbs are distributed across most price points, and Manhattan is the borough where the most expensive outliers (>$1,000/night) are located. Airbnbs priced below and above the mean are available in all 5 boroughs.

```
#explore price
print(summary(data4$price))

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    49.0   164.2   235.5   299.2   356.8  2064.0
```



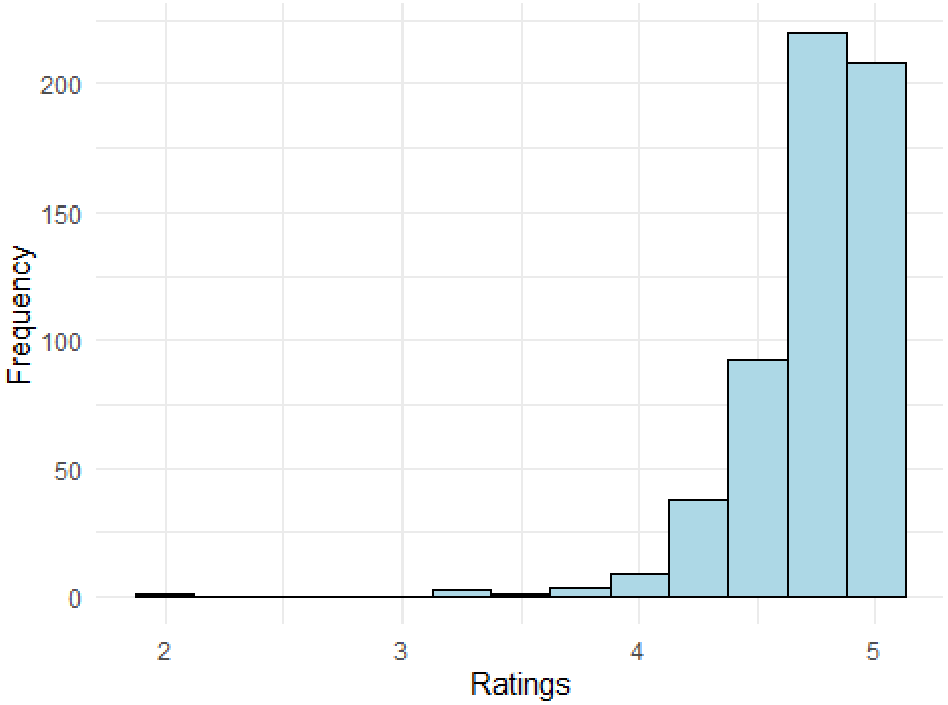
The below boxplot shows that the median price is fairly consistent across boroughs. Manhattan has the largest spread and the largest outliers at the top of the price range, as also observed in the histogram.

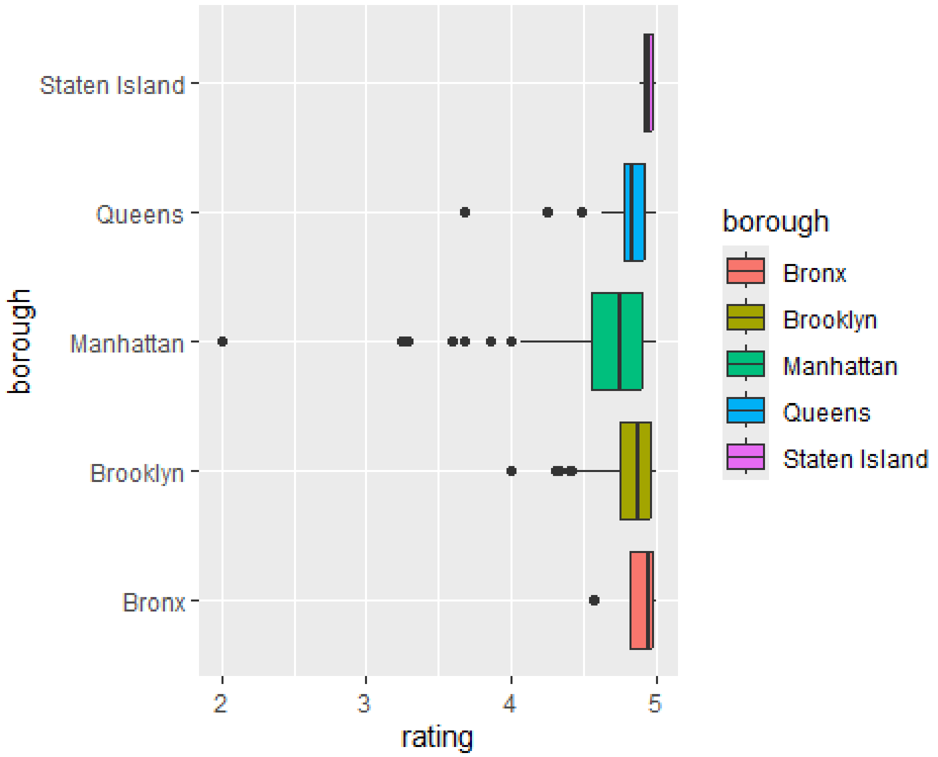


The below summary statistics, histogram and boxplot show that the median rating for all Airbnbs in the dataset is very high: 4.8 out of 5 stars. The data is strongly left skewed. The median rating is slightly higher in Manhattan with a larger spread and more outliers in the 2 and 3 star range. However, Manhattan's sample is much larger than the other neighborhoods. Bronx and Staten Island had very small sample sizes, and the ratings for those Airbnbs were consistently high.

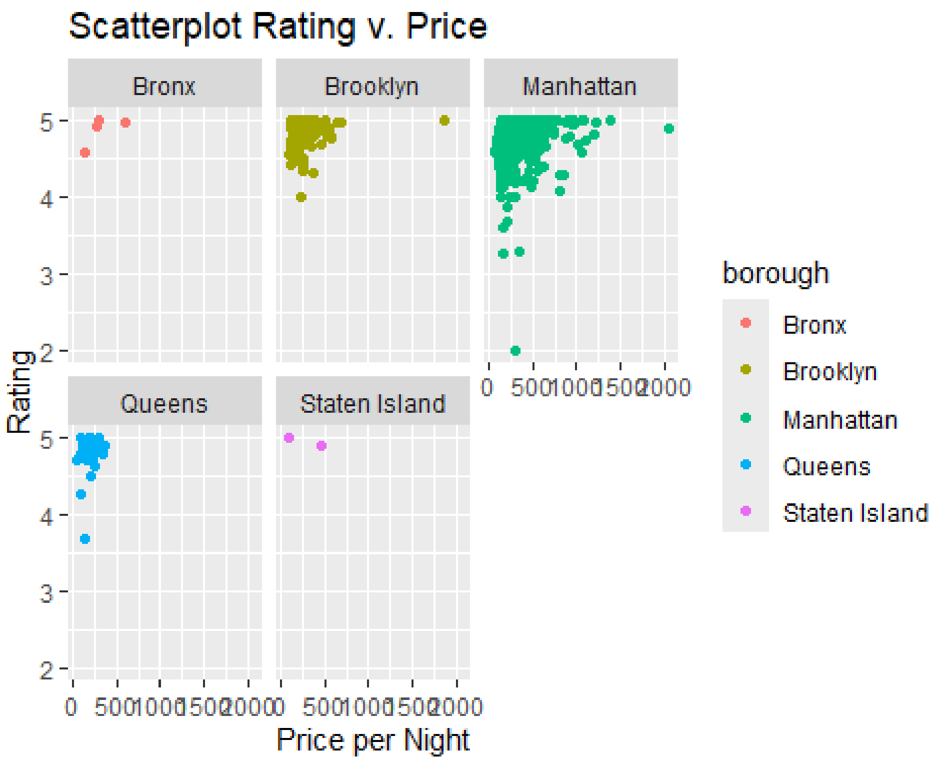
```
summary(data4$rating)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   2.000   4.620   4.800   4.735   4.920   5.000
```





The scatterplot of rating v. price shows that ratings were consistently high across price points. Queens and the Bronx had the most Airbnbs at the lower end of the price range, and almost all were rated above 4 stars.



After confirming that price did not have an impact on ratings, I filtered out Airbnbs above the median price ($235.50/night) and Airbnbs that were below the median rating (4.8). This filter reduced the dataset from 574 Airbnbs to 140 Airbnbs. Next, I grouped the filtered dataset by borough and identified the Airbnb with the lowest price in each borough. No Airbnbs in the Bronx met the below-average-price and above-average-rating criteria, but the remaining 4 boroughs each returned an Airbnb priced around $100/night. The Staten Island Airbnb has more bedrooms/baths than the others, and the Queens Airbnb has the most availability.

```
dim(data5)

## [1] 140  12

#Identify best (above average rating and lowest price) deal for each borough
## # A tibble: 4 × 6
##   borough       price availability_365 rating bedrooms baths
##   <chr>         <dbl>            <int>  <dbl>    <dbl> <dbl>
## 1 Brooklyn        107               36   5           1   1
## 2 Manhattan        97              154   4.76        1   1
## 3 Queens           87              333   4.78        1   1
## 4 Staten Island    95               67   5           3   1.5
```

Because of the range of the difference in number of bedrooms and baths in the Staten Island results, I want to investigate the average price while also considering the number of bedrooms and bathrooms. The below analysis confirms that all four Airbnbs returned as the best deal in each borough are below the median price when considering number of bedrooms/baths.

For example, the Manhattan apartment identified as the lowest priced Airbnb with an above average rating is $97/night, well below the $155/night average price of a 1 bedroom/1 bath rental in Manhattan with an above average rating.

```
#find average price grouped by number of bedrooms and baths in each borough


## # A tibble: 17 × 4
## # Groups:   borough, bedrooms [11]
##    borough       bedrooms baths mean_price
##    <chr>            <dbl> <dbl>      <dbl>
##  1 Brooklyn             0   1          199
##  2 Brooklyn             1   1          191.
##  3 Brooklyn             1   1.5        156.
##  4 Brooklyn             1   2          196
##  5 Brooklyn             2   1          172.
##  6 Brooklyn             2   1.5        168
##  7 Brooklyn             2   2          186
##  8 Brooklyn             3   1          206
##  9 Brooklyn             3   1.5        183
## 10 Manhattan            0   1          151.
## 11 Manhattan            1   1          155.
## 12 Manhattan            2   1          201.
## 13 Queens               1   1          134.
## 14 Queens               2   1          171.
## 15 Queens               2   2          170.
## 16 Queens               3   2          114
## 17 Staten Island        3   1.5         95
```

Summary

This analysis of the NYC Airbnb dataset identified four Airbnbs that are the best deal (i.e., rating is above average and lowest price) in each borough after filtering the Airbnbs according to my family's criteria. Interestingly, the analysis showed that price did not have a relationship to rating for the Airbnbs in the filtered dataset.

In addition, although sample sizes for the Bronx and Staten Island were quite small, Airbnbs below and above the median price were available in all five boroughs. The most expensive Airbnbs were located in Manhattan, but otherwise price did not have a relationship to borough.