Group Project

Chandima Attanayake, Ryan Boyle, Evan Avelluto, Nicole Minor, Hannah Orazem

2024-07-10

**Data Cleaning:**

# Load necessary libraries  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(lubridate)

##   
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':  
##   
## date, intersect, setdiff, union

# Load the dataset  
airbnb\_data <- read.csv("/Users/chandimaattanayake/Downloads/Airbnb.csv")  
  
# 1. Remove Duplicates  
airbnb\_data <- airbnb\_data %>%  
 distinct()  
  
# 2. Handle Missing Data  
  
# Replace missing 'name' and 'host\_name' with 'Unknown'  
  
airbnb\_data$name <- ifelse(airbnb\_data$name == "", "UNKNOWN", airbnb\_data$name)   
airbnb\_data$host\_name <- ifelse(airbnb\_data$host\_name == "", "UNKNOWN", airbnb\_data$host\_name)  
  
# Replace missing 'last\_review' with 'N/A'  
airbnb\_data$last\_review <- ifelse(is.na(airbnb\_data$last\_review), "N/A", airbnb\_data$last\_review)  
  
# Replace missing 'reviews\_per\_month' with 0  
airbnb\_data$reviews\_per\_month <- ifelse(is.na(airbnb\_data$reviews\_per\_month), 0, airbnb\_data$reviews\_per\_month)  
  
# 3. Correct Inconsistent Data  
# Standardize text: Proper case for 'name' and Upper case for 'neighbourhood\_group'  
airbnb\_data$name <- tolower(airbnb\_data$name)  
airbnb\_data$name <- tools::toTitleCase(airbnb\_data$name)  
  
airbnb\_data$neighbourhood\_group <- toupper(airbnb\_data$neighbourhood\_group)  
  
# 4. Convert Data Types  
# Convert 'last\_review' to Date type  
airbnb\_data$last\_review <- as.Date(airbnb\_data$last\_review, format="%m/%d/%Y")  
  
# Convert 'price', 'minimum\_nights', 'number\_of\_reviews', 'reviews\_per\_month', 'calculated\_host\_listings\_count', and 'availability\_365' to numeric  
airbnb\_data$price <- as.numeric(airbnb\_data$price)  
airbnb\_data$minimum\_nights <- as.numeric(airbnb\_data$minimum\_nights)  
airbnb\_data$number\_of\_reviews <- as.numeric(airbnb\_data$number\_of\_reviews)  
airbnb\_data$reviews\_per\_month <- as.numeric(airbnb\_data$reviews\_per\_month)  
airbnb\_data$calculated\_host\_listings\_count <- as.numeric(airbnb\_data$calculated\_host\_listings\_count)  
airbnb\_data$availability\_365 <- as.numeric(airbnb\_data$availability\_365)  
  
# 5. Filter and Sort Data  
# Example: Filter listings with price > 100  
filtered\_data <- airbnb\_data %>%  
 filter(price > 100)  
  
# Example: Sort data by 'price'  
sorted\_data <- airbnb\_data %>%  
 arrange(price)  
  
# 6. Handle Outliers  
# Identify outliers in 'price' using IQR  
Q1 <- quantile(airbnb\_data$price, 0.25)  
Q3 <- quantile(airbnb\_data$price, 0.75)  
IQR <- Q3 - Q1  
outliers <- airbnb\_data %>%  
 filter(price < (Q1 - 1.5 \* IQR) | price > (Q3 + 1.5 \* IQR))  
  
# Optionally remove outliers  
airbnb\_data <- airbnb\_data %>%  
 filter(price >= (Q1 - 1.5 \* IQR) & price <= (Q3 + 1.5 \* IQR))  
  
# Save the cleaned data  
write.csv(airbnb\_data, "/Users/chandimaattanayake/Downloads/Airbnb\_Cleaned4.csv", row.names = FALSE)

**Data Analysis:**

Bullet Point 1: What insights can we glean about different hosts and areas?

1. We can see which areas are the most expensive on average

DF = **read.csv**("C:/Users/boyle/Downloads/Airbnb\_Cleaned.csv")  
  
**library**(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

**library**(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.4.1

avg\_price\_by\_group <- DF **%>%** **group\_by**(neighbourhood\_group) **%>%** **summarise**(average\_price = **mean**(price, na.rm = TRUE)) **%>%** **arrange**(**desc**(average\_price))  
  
**print**(avg\_price\_by\_group, caption = "Neighbourhood Groups Ranked by Average Price")

## # A tibble: 5 × 2  
## neighbourhood\_group average\_price  
## <chr> <dbl>  
## 1 MANHATTAN 146.   
## 2 BROOKLYN 106.   
## 3 STATEN ISLAND 89.2  
## 4 QUEENS 88.9  
## 5 BRONX 77.4

Summary:

As you can see, the most expensive area on average is Manhattan, and the least expensive is the Bronx.

We can also see which hosts have the most number of reviews:

top\_hosts <- DF **%>%** **group\_by**(host\_id, host\_name, neighbourhood\_group) **%>%** **summarise**(total\_reviews = **sum**(number\_of\_reviews, na.rm = TRUE), .groups = 'drop') **%>%** **arrange**(**desc**(total\_reviews)) **%>%** **slice\_head**(n = 5)  
  
**print**(top\_hosts, caption = "Top 5 Hosts with Most Number of Reviews")

## # A tibble: 5 × 4  
## host\_id host\_name neighbourhood\_group total\_reviews  
## <int> <chr> <chr> <int>  
## 1 37312959 Maya QUEENS 2273  
## 2 344035 Brooklyn& Breakfast -Len- BROOKLYN 2205  
## 3 26432133 Danielle QUEENS 2017  
## 4 35524316 Yasu & Akiko MANHATTAN 1971  
## 5 40176101 Brady BROOKLYN 1818

Summary:

Here we can see the top 5 hosts with the most number of reviews, as well as their neighborhood group, and total number of reviews.

Bullet point 2: Is the number of reviews a statistically significant predictor of the price?

linear\_model <- **lm**(price **~** number\_of\_reviews, data = DF)  
  
**summary**(linear\_model)

##   
## Call:  
## lm(formula = price ~ number\_of\_reviews, data = DF)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -120.97 -53.97 -20.47 39.03 222.28   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 120.967357 0.359545 336.445 < 2e-16 \*\*\*  
## number\_of\_reviews -0.041643 0.007015 -5.936 2.94e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 68.12 on 45921 degrees of freedom  
## Multiple R-squared: 0.0007667, Adjusted R-squared: 0.000745   
## F-statistic: 35.24 on 1 and 45921 DF, p-value: 2.943e-09

Summary:

This linear model helps us determine if the number of reviews is a significant predictor of price statistically. The p-value of 2.943e-09 is well below 0.05, and this shows that there is a strong statistical relationship between the number of reviews and price. However, due to the low r^2 value of 0.0007667 this means that even though there is a strong relationship, the number of reviews is not a significant predictor when is comes to the price variable.

Bullet Point 3: Who are the busiest hosts, and why might this be the case?

Who are the busiest hosts?

busy\_hosts <- DF **%>%**  
 **group\_by**(host\_id, host\_name) **%>%**  
 **summarise**(  
 avg\_reviews\_per\_month = **mean**(reviews\_per\_month, na.rm = TRUE)  
 ) **%>%**  
 **arrange**(**desc**(avg\_reviews\_per\_month)) **%>%**  
 **top\_n**(10, avg\_reviews\_per\_month)

## `summarise()` has grouped output by 'host\_id'. You can override using the  
## `.groups` argument.

**print**(busy\_hosts)

## # A tibble: 35,389 × 3  
## # Groups: host\_id [35,389]  
## host\_id host\_name avg\_reviews\_per\_month  
## <int> <chr> <dbl>  
## 1 228415932 Louann 20.9  
## 2 156684502 Nalicia 18.1  
## 3 217379941 Brent 15.8  
## 4 47621202 Dona 14.0  
## 5 244361589 Row NYC 14.0  
## 6 26432133 Danielle 13.6  
## 7 256290334 Aisling 13.4  
## 8 257832461 Stephanie 13.3  
## 9 111841534 Malini 13.2  
## 10 27287203 Ben 13.1  
## # ℹ 35,379 more rows

Summary:

Based on the number of reviews per month, These are the top 10 hosts that could be considered the busiest.

Why might this be the case?

To find this, lets look at what variables are the most statistically significant to higher average number of reviews per month.

lm\_model <- **lm**(reviews\_per\_month **~** price **+** room\_type **+** neighbourhood\_group **+** availability\_365, data = DF)  
  
*# ANOVA analysis*  
anova\_result <- **anova**(lm\_model)  
  
*# Print ANOVA table*  
**cat**("ANOVA Analysis for Linear Regression Model:**\n**")

## ANOVA Analysis for Linear Regression Model:

**cat**("--------------------------------------------**\n**")

## --------------------------------------------

**print**(anova\_result)

## Analysis of Variance Table  
##   
## Response: reviews\_per\_month  
## Df Sum Sq Mean Sq F value Pr(>F)   
## price 1 268 268.22 107.6250 <2e-16 \*\*\*  
## room\_type 2 9 4.62 1.8519 0.157   
## neighbourhood\_group 4 1599 399.63 160.3551 <2e-16 \*\*\*  
## availability\_365 1 3126 3125.84 1254.2630 <2e-16 \*\*\*  
## Residuals 45914 114426 2.49   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Summary:

Based on the ANOVA Analysis, we can see that variables such as price, neighborhood group, and availability are statistically significant factors related to the number of reviews per month for hosts. We can infer that the top hosts have prices, availability, and location that cause to place reviews and thus keep them busy.

data <- read.csv("~/Downloads/Airbnb\_Cleaned.csv")  
  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

#Room Type Regression  
data$Room\_Type\_Encoded <- ifelse(data$room\_type == "Private room", 1, ifelse(data$room\_type == "Shared room", 2, NA))  
  
room\_data <- data[!is.na(data$Room\_Type\_Encoded), ]  
  
model\_room\_type <- lm(price ~ Room\_Type\_Encoded, data = room\_data)  
  
summary(model\_room\_type)

##   
## Call:  
## lm(formula = price ~ Room\_Type\_Encoded, data = room\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -79.019 -29.019 -9.019 15.981 260.707   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 98.744 1.344 73.49 <2e-16 \*\*\*  
## Room\_Type\_Encoded -19.725 1.254 -15.72 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 41.26 on 23132 degrees of freedom  
## Multiple R-squared: 0.01058, Adjusted R-squared: 0.01053   
## F-statistic: 247.3 on 1 and 23132 DF, p-value: < 2.2e-16

Summary: The average price for a private room is about $19.73 lower than the price of a shared room. The relationship between room type and price is statistically significant due to the p-value. The population of this data is skewed significantly, with private room data having around 23,000 data points, while shared room data has about 200 data points. The R-squared value is very low, at 1.06%. This indicates that there are many variables missing in the model that could explain the variance in the price.

#Neighborhood Regression  
library(dplyr)  
  
data$Neighbourhood\_Group\_Encoded <- ifelse(data$neighbourhood\_group == "MANHATTAN", 1,   
 ifelse(data$neighbourhood\_group == "BROOKLYN", 2, NA))  
  
neighbourhood\_data <- data[!is.na(data$Neighbourhood\_Group\_Encoded), ]  
  
model\_neighbourhood\_group <- lm(price ~ Neighbourhood\_Group\_Encoded, data = neighbourhood\_data)  
  
summary(model\_neighbourhood\_group)

##   
## Call:  
## lm(formula = price ~ Neighbourhood\_Group\_Encoded, data = neighbourhood\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -145.95 -50.70 -15.70 44.05 227.30   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 186.206 1.055 176.43 <2e-16 \*\*\*  
## Neighbourhood\_Group\_Encoded -40.253 0.668 -60.26 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 65.89 on 38919 degrees of freedom  
## Multiple R-squared: 0.08535, Adjusted R-squared: 0.08532   
## F-statistic: 3632 on 1 and 38919 DF, p-value: < 2.2e-16

Summary Manhattan Airbnbs are $40.25 more expensive than Brooklyn Airbnbs.

Based on the R-squared value, it is slightly higher compared to the Room Type model, with this model having an R-squared value of 8.5%.

It still needs to be said that more variables would increase the R-squared value.

In addition, the p-value for the coefficient of Manhattan is statistically significant since it is very low.

#Heat Map  
  
library(ggplot2)  
library(dplyr)  
library(viridis)

## Loading required package: viridisLite

ggplot(data, aes(x = longitude, y = latitude)) +  
 geom\_point(aes(color = price)) +  
 scale\_color\_viridis(option = "C") +  
 theme\_minimal() +  
 labs(title = "Heatmap of Prices by Latitude and Longitude",  
 x = "Longitude",  
 y = "Latitude",  
 color = "Price")

A map of the heat map

Description automatically generated

Looking at the Heatmap, its pretty interesting to see that a map of New York City was outlined. When looking at around -74 longitude, 40.76 latitude, you can see a white, blank rectangle area…thats Central Park! All the other blank spots you see is the body of water that surrounds New York. Pretty cool!

**Report:**

In recent years the creation of Airbnb has revolutionized the travel industry. They’ve provided easy access to homestays for any type of vacation for any duration, making it easier for people to get the most out of their trips. We were provided the task to sift through relevant information for the area around New York City and provide important insights that can be utilized to further the success and continued performance of Airbnb.

Upon cleaning and reviewing the given Airbnb data for New York City, certain trends were uncovered. First, we focused on the specific hosts and areas that were most popular among consumers and how they favored based on price and reviews. According to the data Manhattan and Brooklyn are the most expensive areas to stay and rent out Airbnb’s. Whereas the Bronx and Queens make up the least expensive. This could be due to location; the Bronx and Queens are farther away from the more popular tourist destinations, which would mean more travel time to and from their intended destination. Manhattan is central to Times Square and Central Park making it more popular.

Based on the data we can also see which destinations were the most reviewed and what city they reside in. The number of reviews can also be utilized to demonstrate a relationship with overall price. Although there is a strong relationship between the two factors, number of reviews is not an effective indicator of price. Further, we can determine the 10 busiest hosts and based on the ANOVA Analysis, we can see that variables such as price, neighborhood group, and availability are statistically significant factors related to the number of reviews per month for hosts. We can infer that the top hosts have prices, availability, and location that cause to place reviews and thus keep them busy.

Next, while utilizing linear regression models, when comparing shared and private room prices we can see that the average price for a private room is about $19.73 lower than the price of a shared room. Although, the population of this data is skewed significantly, with private room data having around 23,000 data points, while shared room data has about 200 data points, creating a large variance. On a side-by-side analysis, we can start to compare Manhattan and Brooklyn and see how the price fluctuates. On average, Airbnbs are $40.25 more expensive in Manhattan than those in Brooklyn, which as previously noted, could be due to their overall proximity to the main tourist attractions. To further demonstrate the price fluctuation, we have created a heat map that shows the prices of Airbnbs in or around New York City as shown above.

In conclusion this information can be utilized to better understand what potential customers are looking for when they book a vacation spot. Airbnb can see that location is a very important factor in choosing where to stay when visiting New York City, whether it be due to ease of access to popular spots or other unknown reasons. They can work to provide helpful insight to customers allowing them access to popular spots to stay, as well as what is around those areas that people also enjoy. This could help to boost customer satisfaction and continue to grow the company.