Group Presentation: Airbnb Listings in NYC

Data Analysis by Ethan Highet, Ningning Wu, Liangfu Li, Xiangyu Song, Veena Otari

Introduction:

In the bustling cityscape of New York, Airbnb has become an essential part of the accommodation, catering to a variety of options for hosts and guests. Understanding the dynamics of Airbnb listings is vital for hosts to optimise their offerings and for guests to find their ideal stay.

This project delves into the NYC Airbnb market, aiming to unveil crucial patterns. We'll explore questions such as the distribution of listing prices, availability in different neighbourhoods, correlations between listing popularity and prices across boroughs, and average pricing by the number of bedrooms.

We used “New York Airbnb Open Data 2024” from Kaggle for this project; the dataset contains several key attributes including price, location, number of bedrooms, and other features. Given the limited background information and industry-specific knowledge, our goal is to provide a general data analysis, serving as a foundation for more in-depth analysis to aid decision-making for Airbnb New York and the hosts in the city.

Here’s the GitHub link to our R script for this data analysis.

Business Questions:

1. What is the overall distribution of listing prices in New York?
2. What is the listing popularity of New York's boroughs?
3. Does the listing popularity of New York's boroughs, correlate with the price in those boroughs?
4. How does the number of bedrooms impact the Airbnb listing prices in New York?

Let’s begin with loading and cleaning the data,

```
# Loading pre-installed libraries
library(readr)
library(tidyverse)
library(dplyr)
library(ggplot2)

# Setting work directory
setwd("C:\\Users\\DELL\\Downloads\\MBusAn\\R programming")

# Reading data from CSV
data <- read_csv("new_york_listings_2024.csv")

# Creating a dataframe
data_df <- data.frame(data)
```

```
## Data Cleaning

# Removing "name" column
data_df <- data_df %>% select(-name)

# Adding surrogate key
data_df$s_id <- seq(1, nrow(data_df))

# Specify columns to be processed
columns_to_process <- c("rating", "bedrooms", "baths")

# Loop through specified columns
for (col in columns_to_process) {
  # Check if the current column contains character data
  if (is.character(data_df[[col]])) {
    # Convert 'Studio' to 0, and replace 'No rating' and 'New' with NA in character columns
    data_df[[col]] <- ifelse(data_df[[col]] == "Studio", 0, data_df[[col]])
    data_df[[col]] <- ifelse(data_df[[col]] == "No rating"|data_df[[col]] == "New", NA,
data_df[[col]])
    # Convert the column to numeric
    data_df[[col]] <- as.numeric(as.character(data_df[[col]]))
  }
}

str(data_df)
```

Once we have the data ready for analysis, we can start with answering the business questions.

1. What is the overall distribution of listing prices in New York?

Key Findings:

Examining Airbnb listing prices using numerous indicators reveals a complex picture. The average price per night is $187.77, reflecting a central trend. The median price is $125, while the mode, the most typical price, is $100. Extremes range from $10 per night to $100,000 per night, demonstrating the variety of options. The spread is further highlighted by the quantile distribution, which has a median value of $125. The interquartile range (IQR) is $119, representing the middle 50% of prices. This concise analysis provides a detailed overview of Airbnb pricing patterns, providing useful insights for both hosts and guests as they navigate the various terrain of listing rates.

```
1. The average price per night for all the listings.
mean(data_df$price, na.rm = TRUE)
[1] 187.7766

→ The average price per night is $187.77

2. The median price per night for all the listings
median(data_df$price, na.rm = TRUE)
[1] 125

→ The median price per night is $125

3. The standard deviation of listing price
sd(data_df$price)
```

```
[1] 1022.797
```

→ ***The standard deviation is 1022.797***

```
4. The mode price (most common price) per night for all the listings
get_mode <- function(v) {
  unique_value <- unique(v)
  unique_value[which.max(tabulate(match(v, unique_value)))]
}
get_mode(data_df$price)
[1] 100
```

→ ***The mode price per night is $100***

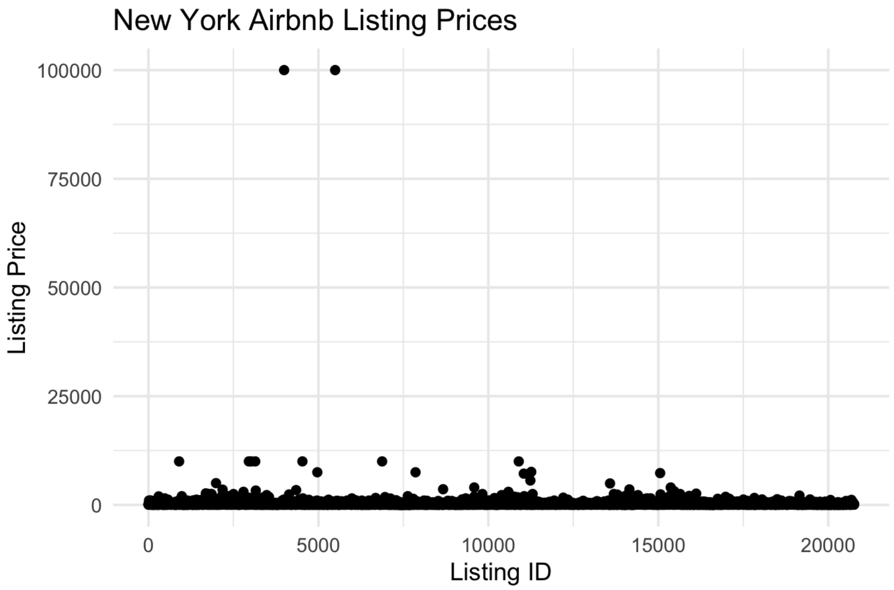
```
5. The quantile distribution of the price
quantile(data_df$price, na.rm = FALSE)
    0%    25%    50%    75%   100% 
    10     80    125    199 100000 
```

→ ***The quantile distribution of price per night is $100000***

```
6. The interquartile range
IQR(data_df$price)
[1] 119
```

→ ***The inter-quantile range is 119***

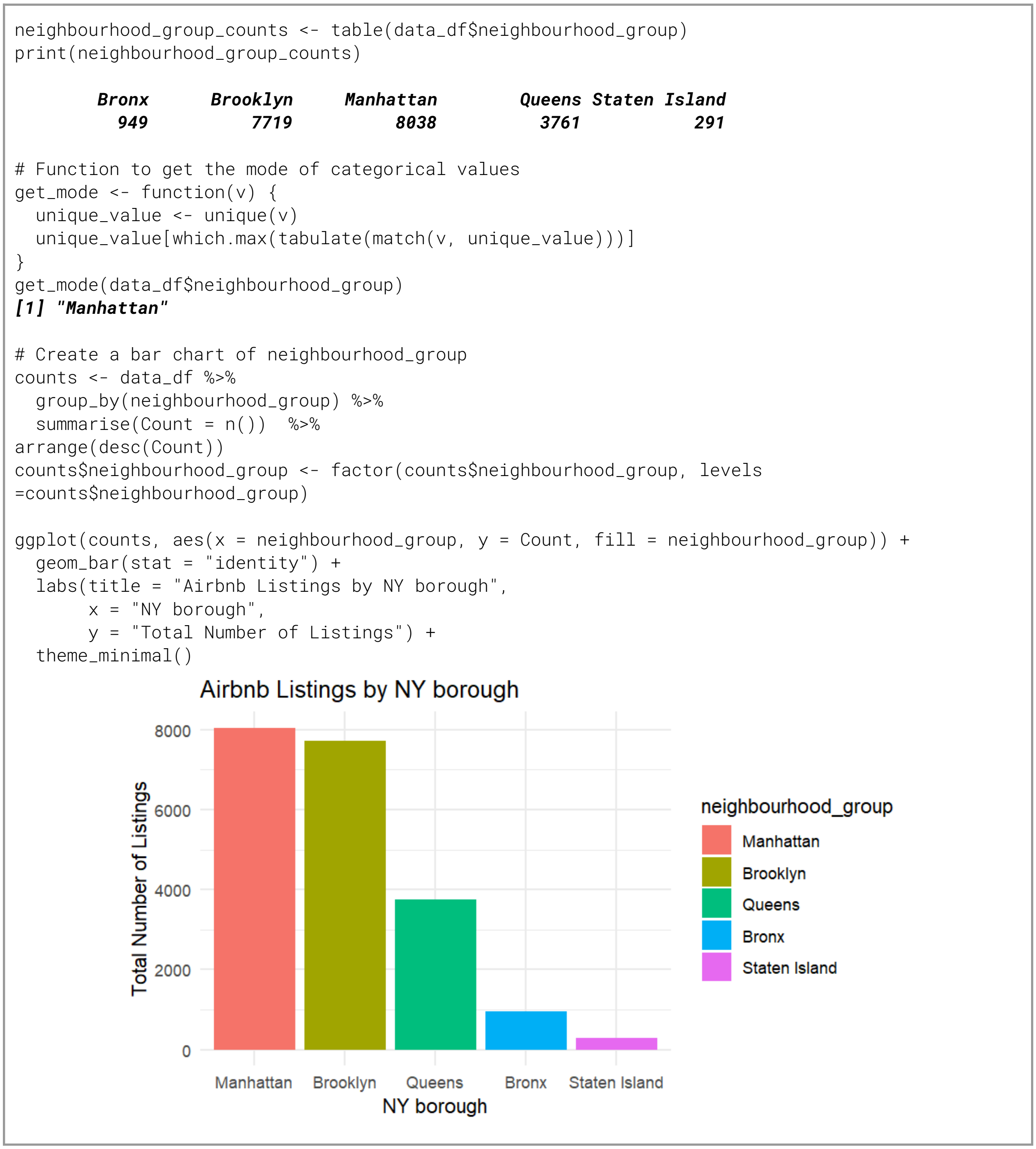
```
7. Visualization of the distribution of listing prices using a scatter plot
ggplot(data = data_df, aes(x = s_id, y = price)) +
  geom_point() +
  labs(title = "New York Airbnb Listing Prices",
       x = "Listing ID",
       y = "Listing Price") +
  theme_minimal()
```



2. How many listings are available per neighbourhood group?

Key Findings:

The data demonstrates how Airbnb listings are distributed among several New York City neighbourhoods. Manhattan has 8038 listings, making it the most common neighbourhood group. Brooklyn has 7719 listings, while Queens, the Bronx, and Staten Island have 3761, 949, and 291 listings, respectively. The brief report provides an overview of the number of Airbnb listings in different areas of the city, with Manhattan serving as the primary focus.



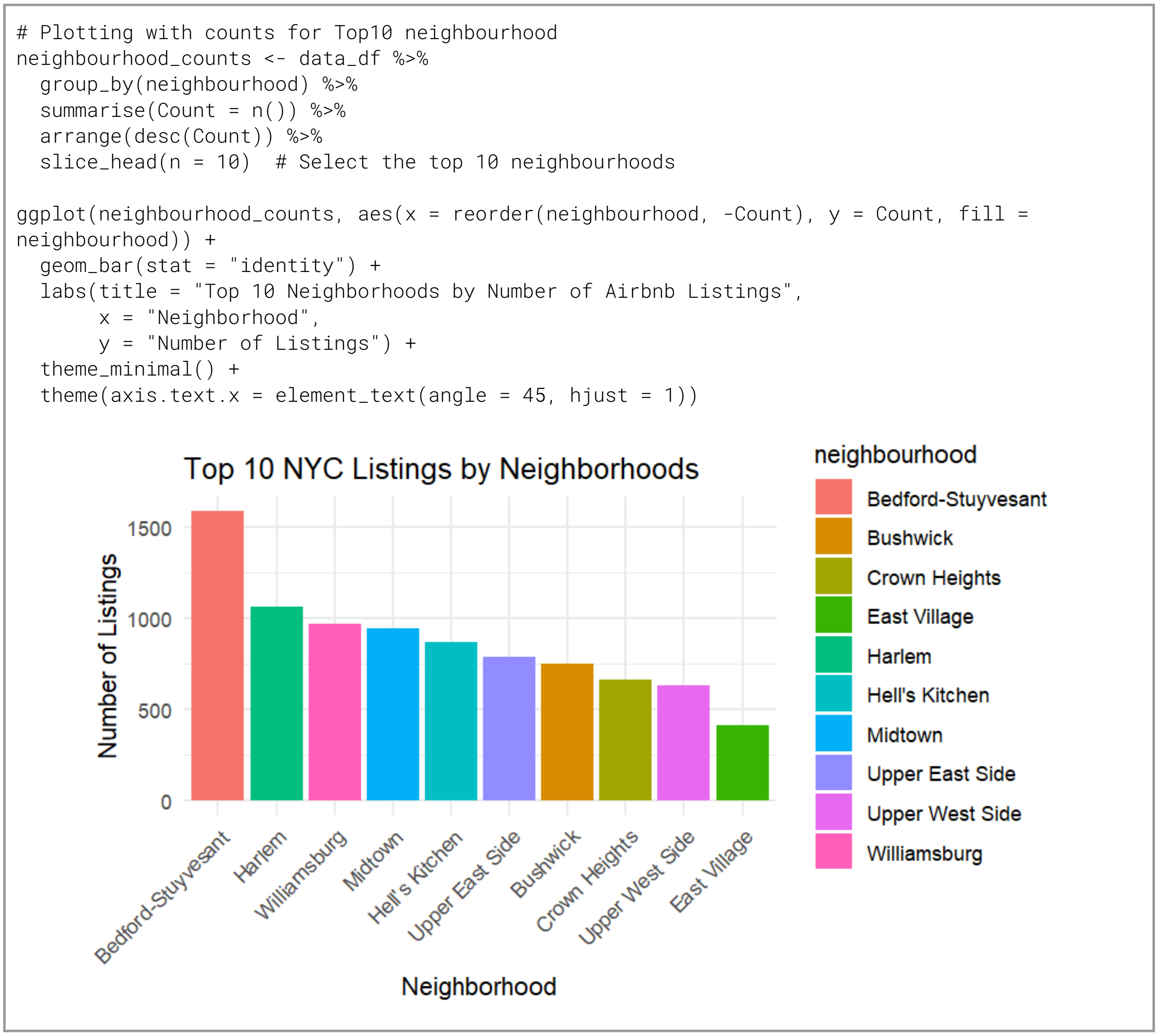
```
neighbourhood_group_counts <- table(data_df$neighbourhood_group)
print(neighbourhood_group_counts)

        Bronx      Brooklyn     Manhattan        Queens Staten Island
          949          7719          8038          3761           291

# Function to get the mode of categorical values
get_mode <- function(v) {
  unique_value <- unique(v)
  unique_value[which.max(tabulate(match(v, unique_value)))]
}
get_mode(data_df$neighbourhood_group)
[1] "Manhattan"

# Create a bar chart of neighbourhood_group
counts <- data_df %>%
  group_by(neighbourhood_group) %>%
  summarise(Count = n())  %>%
arrange(desc(Count))
counts$neighbourhood_group <- factor(counts$neighbourhood_group, levels
=counts$neighbourhood_group)

ggplot(counts, aes(x = neighbourhood_group, y = Count, fill = neighbourhood_group)) +
  geom_bar(stat = "identity") +
  labs(title = "Airbnb Listings by NY borough",
       x = "NY borough",
       y = "Total Number of Listings") +
  theme_minimal()
```



```r
# Plotting with counts for Top10 neighbourhood
neighbourhood_counts <- data_df %>%
  group_by(neighbourhood) %>%
  summarise(Count = n()) %>%
  arrange(desc(Count)) %>%
  slice_head(n = 10)  # Select the top 10 neighbourhoods

ggplot(neighbourhood_counts, aes(x = reorder(neighbourhood, -Count), y = Count, fill =
neighbourhood)) +
  geom_bar(stat = "identity") +
  labs(title = "Top 10 Neighborhoods by Number of Airbnb Listings",
       x = "Neighborhood",
       y = "Number of Listings") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

3. Does listing popularity of New York's boroughs, correlate with the price in those boroughs?
Key Findings:

The coefficient of correlation is 0.9224997, which is very close to 1. This indicates a very strong positive linear relationship between listing popularity and average price in New York boroughs. Places like Manhattan and Brooklyn have a lot of listings and they also cost more to stay in.

The linear regression model has an intercept of approximately 103.7 and a slope of approximately 0.01252. This means when the number of listings goes up a little, the price goes up a little too.

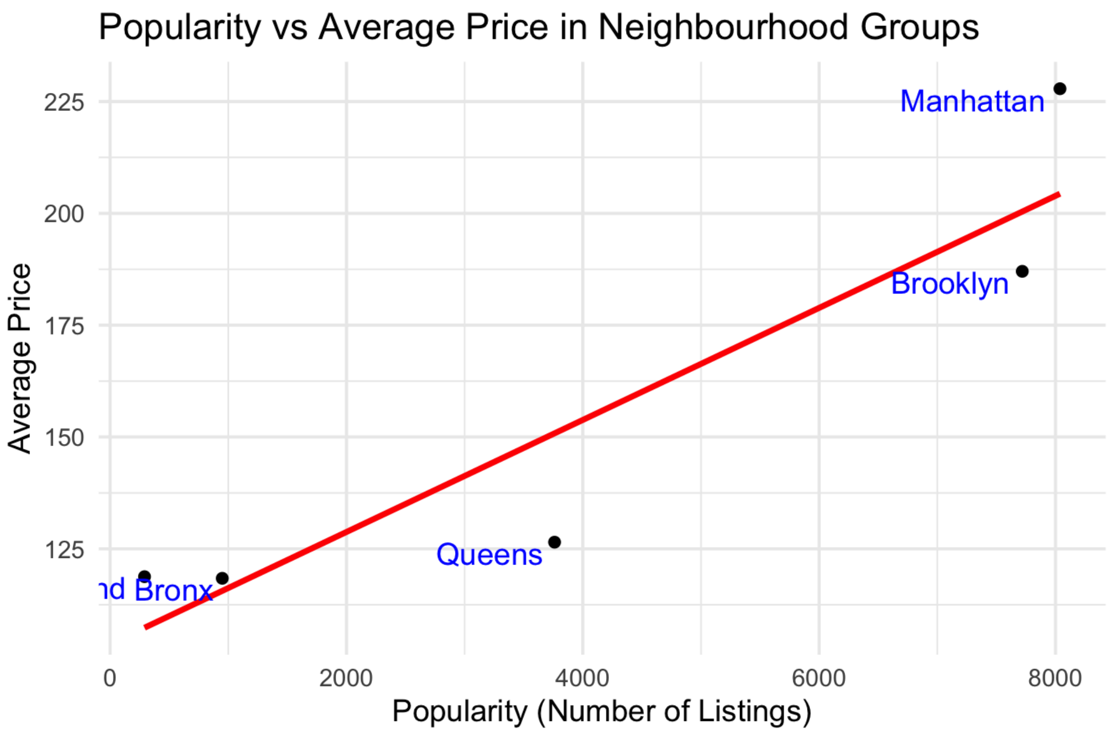
The p-value for the slope coefficient (popularity) is 0.02560, suggesting that the relationship between listing popularity and average price is statistically significant.

The multiple R-squared value is 0.851, indicating that approximately 85.1% of the variability in average price is explained by the number of listings.

However, it's important to note that correlation does not imply causation, and other factors may also influence the average price.

```
# 1. Calculate popularity (here using the number of listings as proxy) and average_price and
create a new data frame
analysis_df <- data_df %>%
  group_by(neighbourhood_group) %>%
  summarise(
    popularity = n(),
    average_price = mean(price, na.rm = TRUE)
  )

# 2. Visualize the relationship between popularity and average price
ggplot(analysis_df, aes(x = popularity, y = average_price)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  geom_text(aes(label = neighbourhood_group, group = neighbourhood_group), vjust = 1, color =
"blue", hjust = 1.1, check_overlap = TRUE) +
  labs(title = "Popularity vs Average Price in Neighbourhood Groups",
       x = "Popularity (Number of Listings)",
       y = "Average Price") +
  theme_minimal()
```



```
# 3. Calculate and display the correlation coefficient and linear regression model
cor_r <- cor(analysis_df$popularity, analysis_df$average_price)
cat("Coefficient of Correlation:", cor_r, "\n")

[1] Coefficient of Correlation: 0.9307469

linear_m <- lm(average_price ~ popularity, data = analysis_df)
summary(linear_m)
```

```
[2] Call:
lm(formula = average_price ~ popularity, data = analysis_df)

Residuals:
      1       2       3       4       5 
  2.804 -13.357  23.468 -24.332  11.417 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1.037e+02  1.597e+01   6.494  0.00741 **
popularity  1.252e-02  3.026e-03   4.139  0.02560 * 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.06 on 3 degrees of freedom
Multiple R-squared:  0.851,    Adjusted R-squared:  0.8013 
F-statistic: 17.13 on 1 and 3 DF,  p-value: 0.0256
```

4. How does the number of bedrooms impact the Airbnb listing prices in New York?

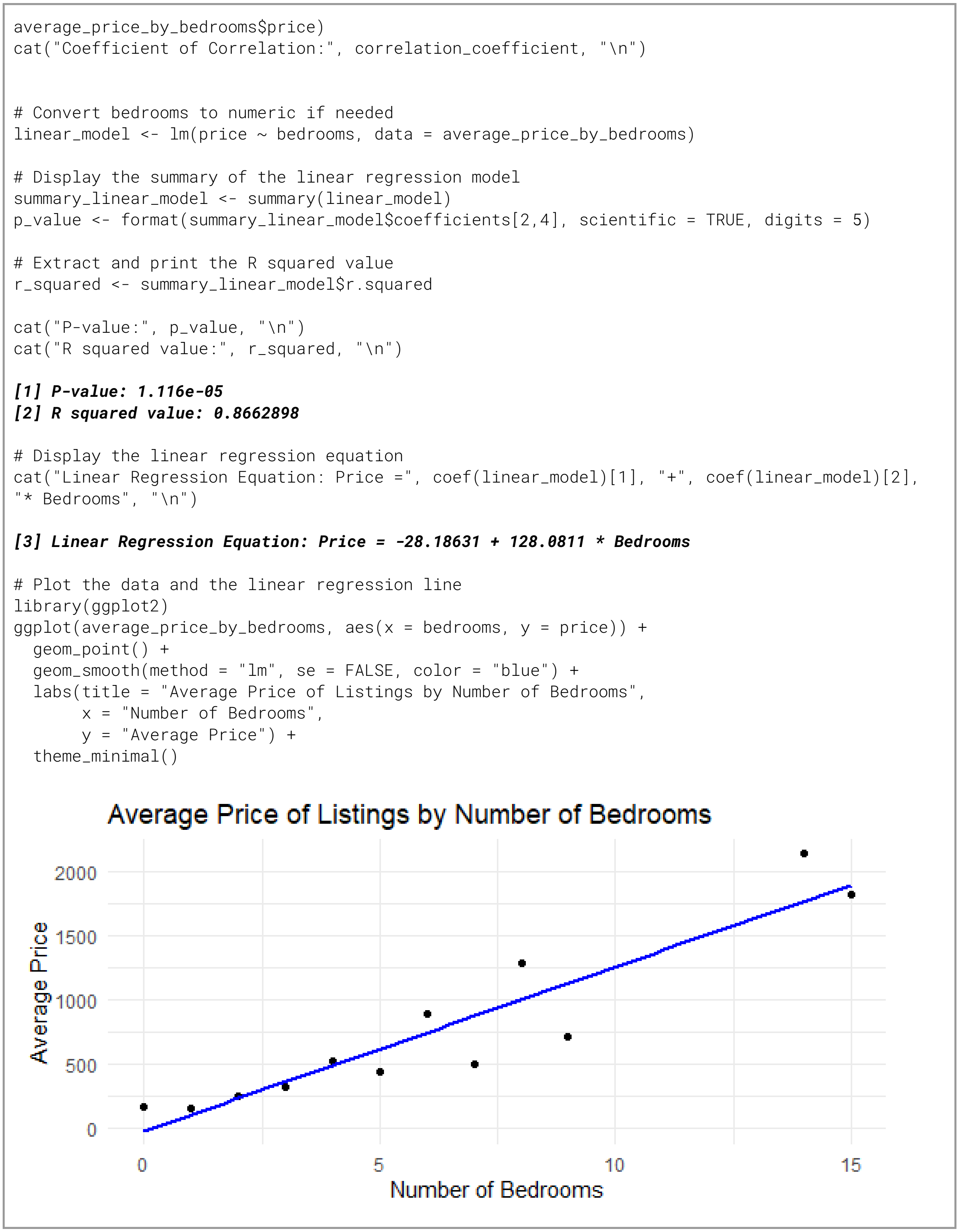
Key Findings:

The analysis of Airbnb listings by bedroom count reveals a simple trend: as the number of bedrooms increases, so do average prices. Notably, homes with 6, 8, 9, 14, and 15 bedrooms attract higher prices. The significant positive correlation coefficient of 0.9307 confirms this association and indicates a persistent increasing trend. According to the linear regression model, each additional bedroom results in an average price increase of $128.08, which is a highly significant finding (p-value = 1.116e-05). The model accounts for 86.63% of the variation in pricing dependent on the number of bedrooms. The scatter plot provides visual confirmation of the beneficial link. This information is essential for both hosts and guests, assisting with pricing plans and preferences.

```
# Convert bedrooms to numeric if needed
data_df$bedrooms <- as.numeric(as.character(data_df$bedrooms))

# Convert bedrooms to numeric if needed
average_price_by_bedrooms <- aggregate(price ~ bedrooms, data = data_df, FUN = mean, na.rm =
TRUE)
print(average_price_by_bedrooms)
   bedrooms     price
1         0  159.0705
2         1  150.4934
3         2  242.0966
4         3  317.2920
5         4  518.5507
6         5  435.4821
7         6  883.3103
8         7  494.0000
9         8 1281.8000
10        9  706.6667
11       14 2136.0000
12       15 1815.0000

# Convert bedrooms to numeric if needed
correlation_coefficient <- cor(average_price_by_bedrooms$bedrooms,
```



```
average_price_by_bedrooms$price)
cat("Coefficient of Correlation:", correlation_coefficient, "\n")


# Convert bedrooms to numeric if needed
linear_model <- lm(price ~ bedrooms, data = average_price_by_bedrooms)

# Display the summary of the linear regression model
summary_linear_model <- summary(linear_model)
p_value <- format(summary_linear_model$coefficients[2,4], scientific = TRUE, digits = 5)

# Extract and print the R squared value
r_squared <- summary_linear_model$r.squared

cat("P-value:", p_value, "\n")
cat("R squared value:", r_squared, "\n")

[1] P-value: 1.116e-05
[2] R squared value: 0.8662898

# Display the linear regression equation
cat("Linear Regression Equation: Price =", coef(linear_model)[1], "+", coef(linear_model)[2],
"* Bedrooms", "\n")

[3] Linear Regression Equation: Price = -28.18631 + 128.0811 * Bedrooms

# Plot the data and the linear regression line
library(ggplot2)
ggplot(average_price_by_bedrooms, aes(x = bedrooms, y = price)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Average Price of Listings by Number of Bedrooms",
       x = "Number of Bedrooms",
       y = "Average Price") +
  theme_minimal()
```