BOARD INFINITY

R PROGRAMMING

PROJECT REPORT: DATA CLEANING AND ANALYSIS OF AIRBNB LISTINGS

Name: Akula Dilipkumar

Email Address: akuladilipkumar99@gmail.com

College: Lovely Professional University Punjab

INTRODUCTION:

This report outlines the process of data cleaning and exploratory data analysis (EDA) performed on a dataset of Airbnb listings. The goal was to prepare the data for further analysis and visualization, focusing on understanding the relationships between various features, particularly price and availability.

DATA IMPORT AND INITIAL SETUP:

The dataset was read from a CSV file using the read.csv function, and essential libraries for data manipulation and visualization were loaded:

```
library(dplyr)
library(tidyr)
library(ggplot2)
library(plotly)
library(reshape2)
library(corrplot)
```

DATA CLEANING STEPS:

1. Data Frame Creation

A copy of the original data was created for manipulation: df <- data

2. Column Removal

Unnecessary columns (`id`, `name`, `host_id`, `host_name`) were removed to simplify the dataset:

df <- df %>% select(-id, -name, -host_id, -host_name)

3. Missing Value Analysis

Missing values in each column were checked and counted:

missing_values <- sapply(df, function(x) sum(is.na(x)))

4. Outlier Removal

Outliers in the `price` and `minimum_nights` columns were identified and removed using z-scores:

```
df <- df %>%
  mutate(z_price = abs(scale(price)),
         z_min_nights = abs(scale(minimum_nights)))

df_clean <- df %>%
  filter(z_price < 3, price > 3, z_min_nights < 3)
```

5. Conversion of Numeric to Categorical Variables

Categorical variables were created for `minimum_nights` and `calculated_host_listings_count` to facilitate analysis:

```
df_clean <- df_clean %>%
  mutate(minimum_nights_group = case_when(
    minimum_nights == 1 ~ "one night",
    minimum_nights == 2 ~ "two nights",
    minimum_nights == 3 ~ "three nights",
    minimum_nights == 4 ~ "four nights",
    minimum_nights > 4 ~ "five nights or more",
    TRUE ~ "Others"
  ))
```

6. Column Cleanup

Unused columns were removed to streamline the data frame:

df_clean <- df_clean %>% select(-z_price, -z_min_nights, -minimum_nights, -last_review, -neighbourhood, -calculated_host_listings_count, -reviews_per_month)

Exploratory Data Analysis

1. Correlation Analysis

A correlation matrix was computed for numeric columns, excluding latitude and longitude, and visualized using a heatmap:

```
df_cor <- numeric_columns %>%
  select(-latitude, -longitude) %>%
  cor()
corrplot(df_cor, method = "color", type = "upper",
         tl.col = "black", tl.srt = 45,
         addCoef.col = "black", number.cex = 0.7)
```

2. Density Map Visualization

A density map of price distribution across geographical coordinates was generated using Plotly:

```
fig <- plot_ly(
  data = df_clean,
  lat = ~latitude,
  lon = ~longitude,
  z = ~price,
  type = "densitymapbox",
  radius = 2
)
```

3. Summary Statistics

Summary statistics were calculated by `neighbourhood_group`:

```
summary_stats <- df_clean %>%
```

```
group_by(neighbourhood_group) %>%
summarise(
  price_count = n(),
  price_mean = mean(price, na.rm = TRUE),
  price_median = median(price, na.rm = TRUE),
  number_of_reviews_mean = mean(number_of_reviews, na.rm = TRUE),
  number_of_reviews_median = median(number_of_reviews, na.rm = TRUE),
  availability_365_mean = mean(availability_365, na.rm = TRUE),
  availability_365_median = median(availability_365, na.rm = TRUE)
)
```

4. Boxplots

Several boxplots were created to visualize the relationship between price and various categorical variables:

- **Price by Neighbourhood Group and Room Type:**

```
ggplot(df_clean, aes(x = neighbourhood_group, y = price, fill = room_type)) +
  geom_boxplot() +
  theme(legend.position = "top") +
  labs(title = "Boxplot of Price by Neighbourhood Group and Room Type")
```

- **Availability by Neighbourhood Group and Room Type:**

```
ggplot(df_clean, aes(x = neighbourhood_group, y = availability_365, fill = room_type)) +
geom_boxplot() +
theme(legend.position = "top") +
labs(title = "Boxplot of Availability by Neighbourhood Group and Room Type")
```

- **Availability by Minimum Nights Group:**
- **Price by Minimum Nights Group:**

```
ggplot(df_clean, aes(x = minimum_nights_group, y = availability_365)) +
geom_boxplot() +
labs(title = "Boxplot of Availability by Minimum Nights Group")
```

- **Price by Minimum Nights Group:**

```
ggplot(df_clean, aes(x = minimum_nights_group, y = price)) +
 geom_boxplot() +
 labs(title = "Boxplot of Price by Minimum Nights Group")
```

5. Pairplots

A scatterplot matrix (pairplot) was created to visualize relationships among numeric variables:

```
numeric_df_clean <- df_clean %>% select_if(is.numeric) %>% select(-latitude, -longitude)

pairs(numeric_df_clean)
```

FINAL MODEL DATA FRAME:

The final model data frame was prepared by removing latitude and longitude, ensuring it contains relevant features for further analysis:

```
df_model <- df_clean %>%
 select(-latitude, -longitude)
```

CONCLUSION:

The data cleaning and exploratory analysis revealed insights into the dataset's structure and relationships among features. The removal of outliers and conversion of numeric to categorical variables improved the dataset's integrity. Visualizations such as correlation heatmaps, density maps, and boxplots provided a clearer understanding of the data's dynamics, setting the stage for deeper analyses or predictive modeling in subsequent projects.

This report serves as a foundation for further investigations into the factors influencing Airbnb listing prices and availability in the dataset.