Project Preliminary Report

Trinath Sai Subhash Reddy Pittala, Uma Maheswara R Meleti, Hemanth Vasireddy

2023-03-09

Problem

The purpose of this project is to perform Data Analysis on the data from this Kaggle Link to get meaningful insights on the data which would in turn help to set prices and help people to select their destination with further ease.

Description

Each major city has its own dataset for weekend and weekdays Variables included in dataset:

- Host ID (Id)
- Total price of listing (realSum)
- Room type: private, shared, entire home, apt (room_type)
- Whether or not room is shared (room_shared)
- Max number of people allowed in property (person_capacity)
- Whether or not host is superbost (host_is_superhost)
- Whether or not it is multiple rooms (multi)
- Whether for business or family use (biz)
- Distance from city center (dist)
- Distance from nearest metro (metro_dist)
- Latitude and longitude (lat lng)
- Guest satisfaction (guest_satisfaction_overall)
- Cleanliness (cleanliness_rating)
- Total quantity of bedrooms available among all properties for single host (bedrooms)

Questions we can answer with the dataset:

- Price Forecasting: use pricing, room type, amenities to predict potential rental prices given other hotel attributes.
- Hotspots: use listing location in relation to business and tourism centers and correlating this with pricing to determine where Airbnb rentals would be most profitable
- Customer sentiment analysis: analyze customer comments and satisfaction ratings to evaluate listing on overall customer experience and use it to optimize hosts' services to improve user satisfaction ratings.

How can this information be used:

- Data can help travelers find accommodation that meets their needs without going over budget.
- Can help hosts set competitive pricing and optimize listings to get more bookings.
- Help investors evaluate value in investing in real estate in different european cities based on pricing trends.

Exploratory Data Analysis

We analysed some variables of the dataset.

```
##  [1] "./archive/amsterdam_weekdays.csv" "./archive/amsterdam_weekends.csv"
##  [3] "./archive/athens_weekdays.csv"    "./archive/athens_weekends.csv"
##  [5] "./archive/barcelona_weekdays.csv" "./archive/barcelona_weekends.csv"
##  [7] "./archive/berlin_weekdays.csv"    "./archive/berlin_weekends.csv"
##  [9] "./archive/budapest_weekdays.csv"  "./archive/budapest_weekends.csv"
## [11] "./archive/lisbon_weekdays.csv"    "./archive/lisbon_weekends.csv"
## [13] "./archive/london_weekdays.csv"    "./archive/london_weekends.csv"
## [15] "./archive/paris_weekdays.csv"     "./archive/paris_weekends.csv"
## [17] "./archive/rome_weekdays.csv"      "./archive/rome_weekends.csv"
## [19] "./archive/vienna_weekdays.csv"    "./archive/vienna_weekends.csv"
```

Adding the data from all 20 .csv files to a table and removing outliers.

```r
# Get a list of all the csv files in the directory
file_list <- list.files(path = my_dir, pattern = "*.csv", full.names = TRUE)

# Initialize an empty list to store the data frames
df_list <- list()

# Loop through each file and read it into a data frame
# after removing outliers
for (i in seq_along(file_list)) {
    df <- read.csv(file_list[i])

    # Add a new column with the city_day
    df$city_day <- gsub("\\.csv", "", basename(file_list[i]))

    iqr_var1 <- IQR(df$realSum)

    # Calculate the upper and lower bounds for each
    # variable
    upper_var1 <- quantile(df$realSum, 0.75) + 1.5 * iqr_var1
    lower_var1 <- quantile(df$realSum, 0.25) - 1.5 * iqr_var1

    # Filter the data based on the upper and lower bounds
    # for each variable
    filtered_data <- filter(df, realSum > lower_var1 & realSum <
        upper_var1)

    # Append the data frame to the list
    df_list[[i]] <- filtered_data
}


# Combine all the data frames into a single dataset
my_data <- bind_rows(df_list)
# Removing the .csv ext
my_data$city_day <- gsub("\\.csv", "", my_data$city_day)
```

Percentage of Outliers.

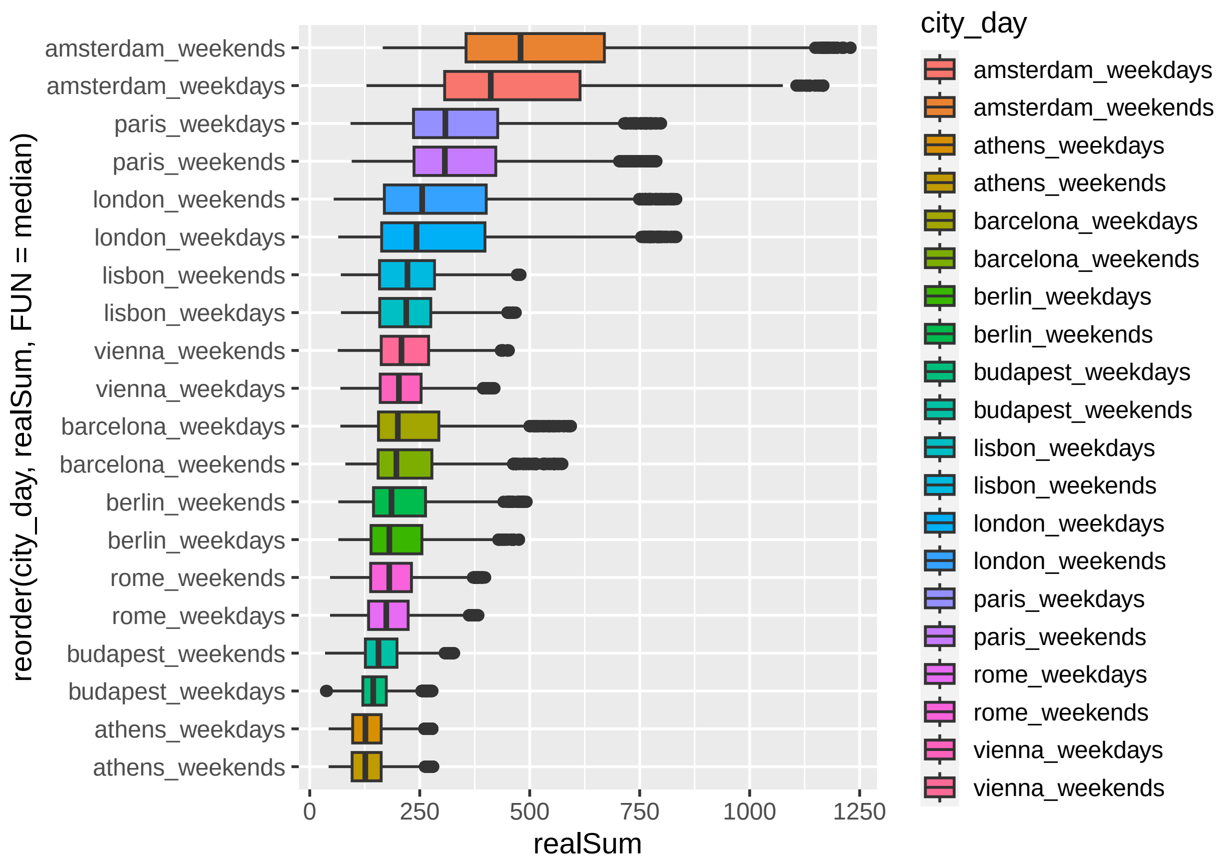
```
# Create empty table
outliers_table <- data.frame(City_day = character(), Data_Length = numeric(),
    Percent_Outliers = numeric(), stringsAsFactors = FALSE)

# Loop through city_data and fill in table
for (city_day in unique(my_data$city_day)) {
    x = my_data[my_data$city_day == city_day, ]$realSum
    q1 <- quantile(x, 0.25)
    q3 <- quantile(x, 0.75)
    iqr <- IQR(x)
    upper_bound <- q3 + 1.5 * iqr
    lower_bound <- q1 - 1.5 * iqr
    x_no_outliers <- x[x >= lower_bound & x <= upper_bound]
    percent_outliers <- ((length(x) - length(x_no_outliers))/length(x)) *
        100

    # Add row to table
    outliers_table <- rbind(outliers_table, data.frame(City_day = city_day,
        Data_Length = length(x), Percent_Outliers = percent_outliers))
}

# Print table
outliers_table
```

```
##              City_day Data_Length Percent_Outliers
## 1  amsterdam_weekdays        1047        1.2416428
## 2  amsterdam_weekends         922        2.4945770
## 3     athens_weekdays        2500        2.3600000
## 4     athens_weekends        2485        1.5291751
## 5  barcelona_weekdays        1438        3.4770515
## 6  barcelona_weekends        1175        8.6808511
## 7     berlin_weekdays        1203        1.6625104
## 8     berlin_weekends        1126        2.4866785
## 9   budapest_weekdays        1951        2.9215787
## 10  budapest_weekends        1840        1.5217391
## 11    lisbon_weekdays        2761        0.8330315
## 12    lisbon_weekends        2805        0.1426025
## 13    london_weekdays        4367        1.6487291
## 14    london_weekends        5082        1.8890201
## 15     paris_weekdays        2938        2.4166099
## 16     paris_weekends        3367        2.5839026
## 17      rome_weekdays        4266        1.1954993
## 18      rome_weekends        4308        1.8337976
## 19    vienna_weekdays        1664        1.5625000
## 20    vienna_weekends        1725        0.8695652
```

Percent of Outliers is below 8 percent and seperate Analysis will be made on that.

Boxplot of Price Vs City

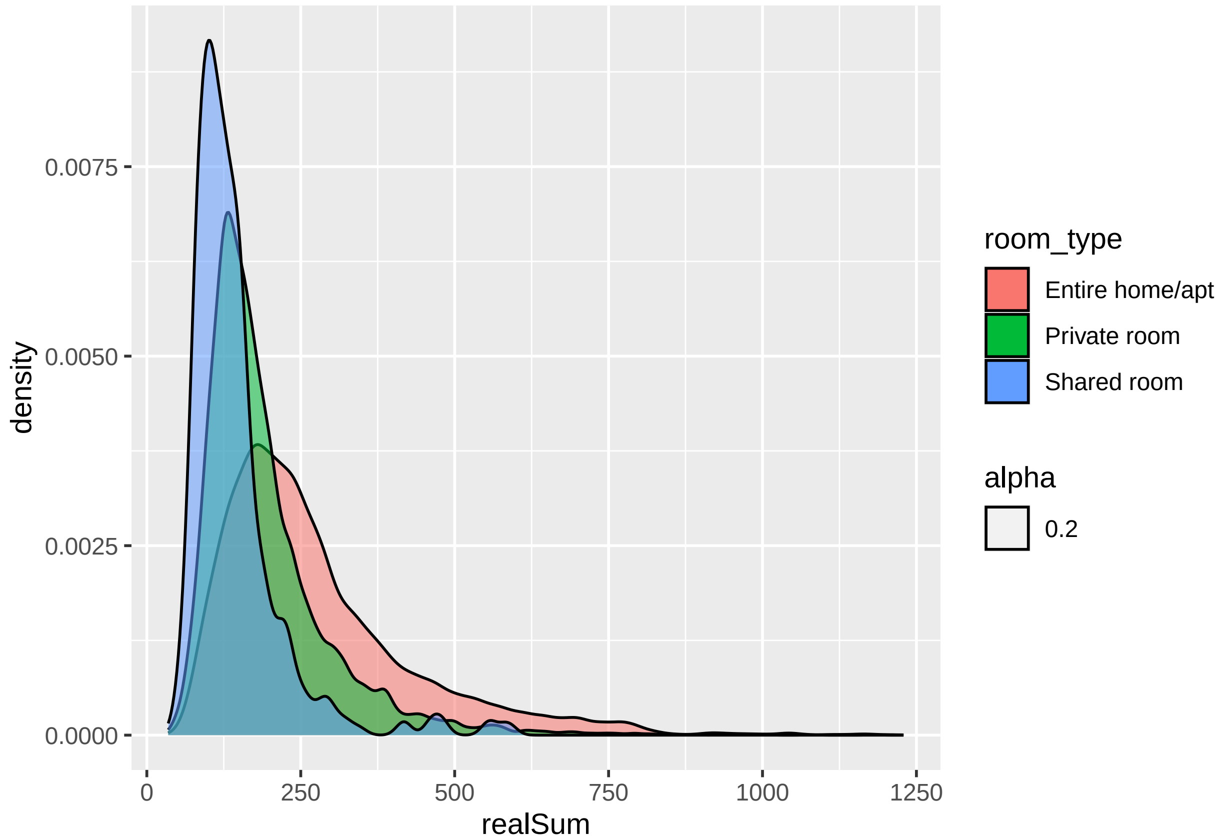
```
ggplot(my_data, aes(x = reorder(city_day, realSum, FUN = median),
    y = realSum, fill = city_day)) + geom_boxplot() + coord_flip() +
    theme(legend.key.height = unit(0.5, "cm"), legend.key.size = unit(1,
        "lines"))
```



The highest prices in europe are found in amsterdam.

Density plot of Price vs Room type

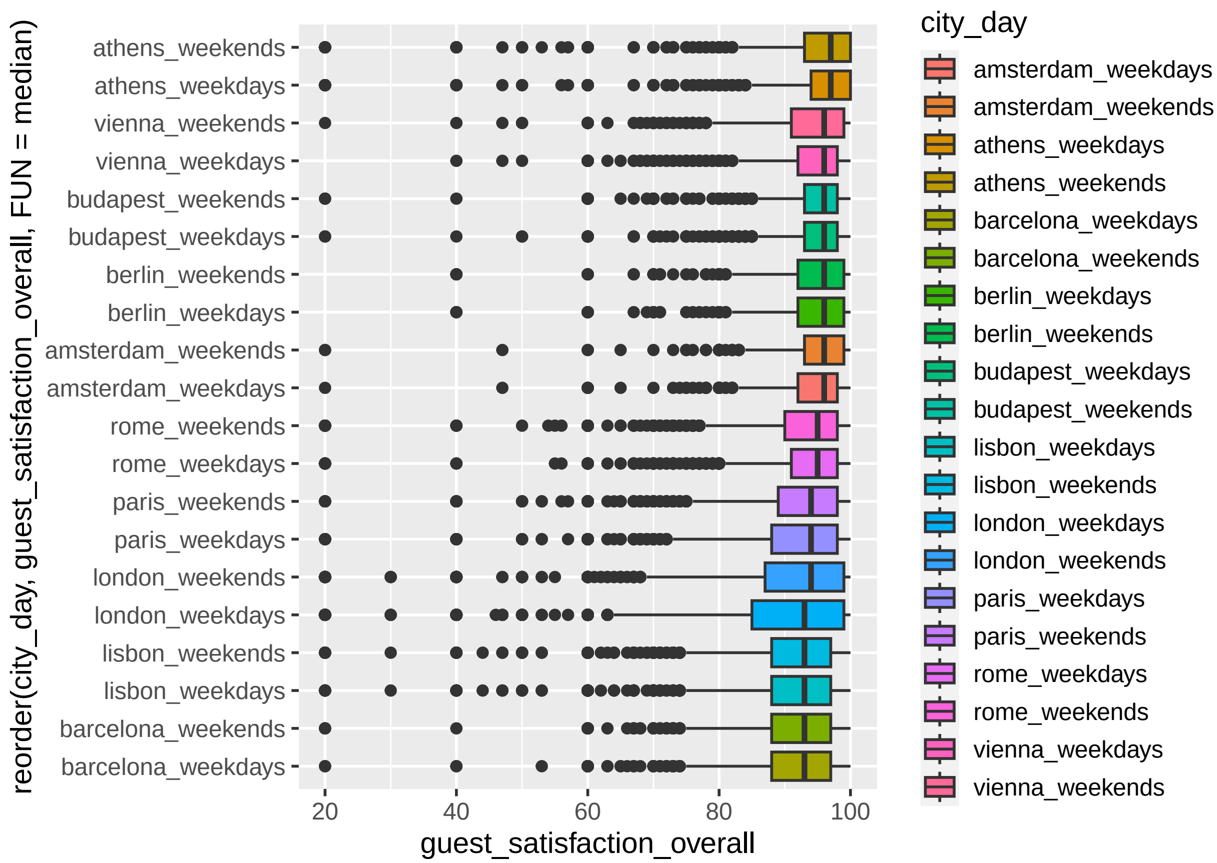
```
ggplot(my_data, aes(x = realSum, group = room_type, fill = room_type,
    alpha = 0.2)) + geom_density()
```



The prices of entire home are high comparitively

Boxplot of City vs Guest Satisfaction

```
ggplot(my_data, aes(x = reorder(city_day, guest_satisfaction_overall,
    FUN = median), y = guest_satisfaction_overall, fill = city_day)) +
    geom_boxplot() + coord_flip() + theme(legend.key.height = unit(0.5,
    "cm"), legend.key.size = unit(1, "lines"))
```



This plot shows there is no major difference in Guest Satisfaction vs City.

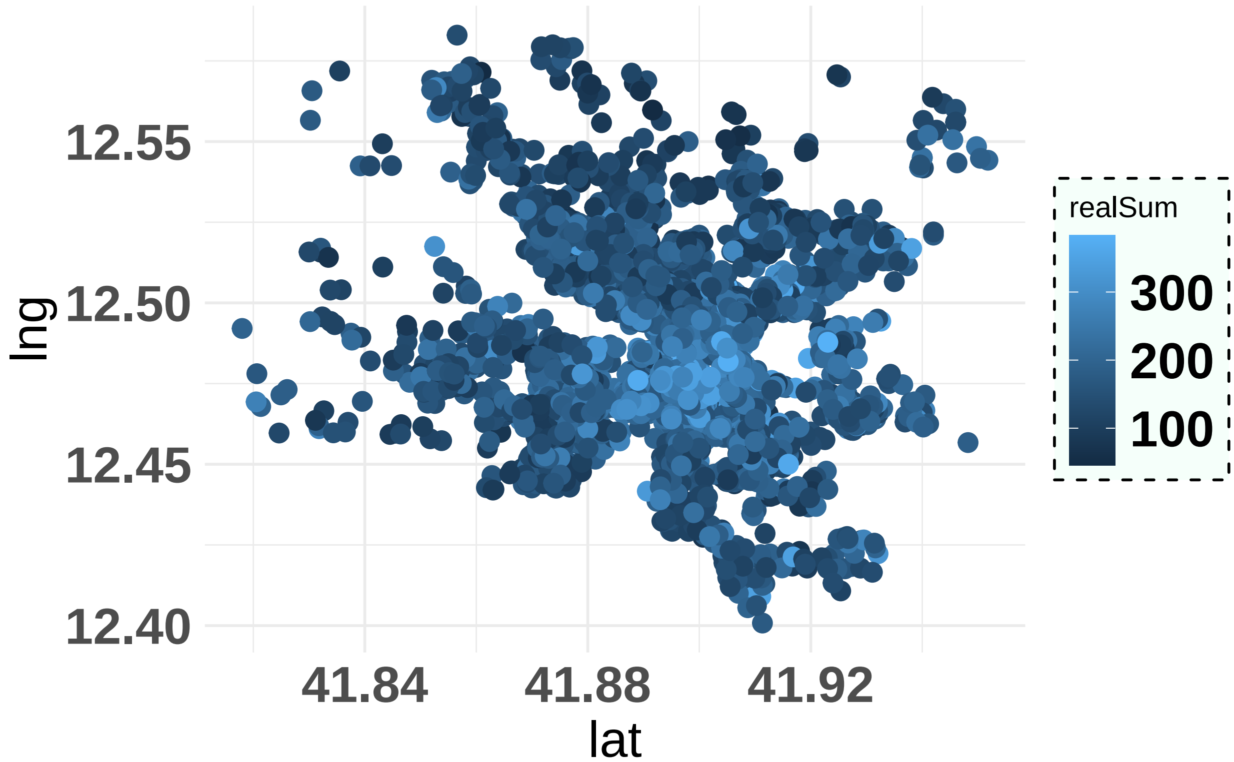
Scatterplot of Price vs Guest Satisfaction filtered by city

```
ggplot(my_data, aes(x = realSum, y = guest_satisfaction_overall,
    color = city_day)) + geom_point() + xlab("Price") + ylab("Guest Satisfaction Overall") +
    scale_color_discrete(name = "City-Day")
```

This plot implies there are good cheaper Airbnb at most cities which give higher guest satisfaction rating

Scatterplot of Prices in Rome w.r.t Latitude and Longitude during weekdays

```
tema <- theme(plot.title = element_text(size = 23, hjust = 0.5),
    axis.text.x = element_text(size = 19, face = "bold"), axis.text.y = element_text(size = 19,
        face = "bold"), axis.title.x = element_text(size = 19),
    axis.title.y = element_text(size = 19), legend.text = element_text(colour = "black",
        size = 19, face = "bold"), legend.background = element_rect(fill = "#F5FFFA",
        size = 0.5, linetype = "dashed", colour = "black"))

rome_data <- my_data %>%
    subset(city_day == "rome_weekdays")

ggplot(data = rome_data, mapping = aes(x = lat, y = lng)) + theme_minimal() +
    scale_fill_identity() + geom_point(mapping = aes(color = realSum),
    size = 3) + ggtitle("") + tema
```



This plot is within expectations of game theory, which suggests similar types of establishments (price and hospitality) tend be in clusters.