
airbnb


# Team Introduction

# Random Forest Gump

# Outline

PROBLEM STATEMENT

DATA CLEANING AND TREATMENT

EXPLORATORY DATA ANALYSIS

MODEL SELECTION

SHINY APP

# Problem Statement

→ **Each listing is UNIQUE.**

→ **Predicting the price for an Airbnb host's new listing in Berlin, Germany.**

→ **What additional features a host can offer to command higher prices?**

**The following questions will drive this project -**

*What are the features that affect pricing of the listing?*

*Based on the features, can we determine a fair price for a new listing that fits into its specific market environment and competitors in Berlin?*


- Most popular city in Germany
- Size – 891km$^2$
- Average 25 homes/ km$^2$

# Outline

PROBLEM STATEMENT

DATA CLEANING AND TREATMENT

EXPLORATORY DATA ANALYSIS

MODEL SELECTION

SHINY APP

# Data Cleaning & Treatment

**Step 1 - Removing unwanted variables.**

- Original Data Frame has 22522 observations and 96 variables
- Based on business understanding, irrelevant parameters were removed
- Using DataQualityReport, identified columns with Null values (More than 40% Null values) and dropped them.

# Data Cleaning & Treatment

**Step 1 - Removing unwanted variables.**

**Step 2 - Treatment of Numerical and Categorical Variables**

- Numerical Variables - All numerical variables are appropriately classified into interval (numeric) class.
- Categorical Variables - All categorical variables are appropriately classified into nominal and ordinal class, as necessary.
  → Missing values treated using Mode.
- Date Variables - All date variables are formatted in 'DD-MM-YYYY' format.

# Data Cleaning & Treatment

**Step 1 - Removing unwanted variables.**

**Step 2 – Treatment of Numerical and Categorical Variables**

**Step 3 – Outliers Treatment**

- Capped the values above 2 Standard Deviations of Mean
- Floored the values below 2 Standard Deviations of Mean

# Outline

PROBLEM STATEMENT

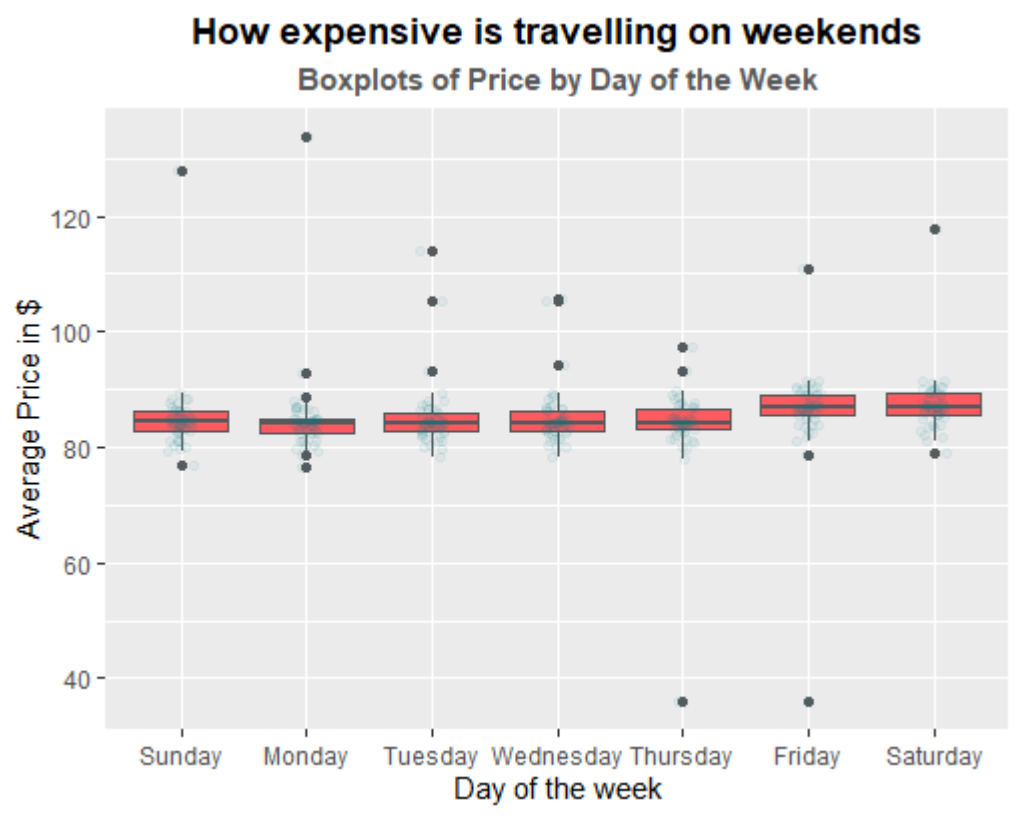
DATA CLEANING AND TREATMENT

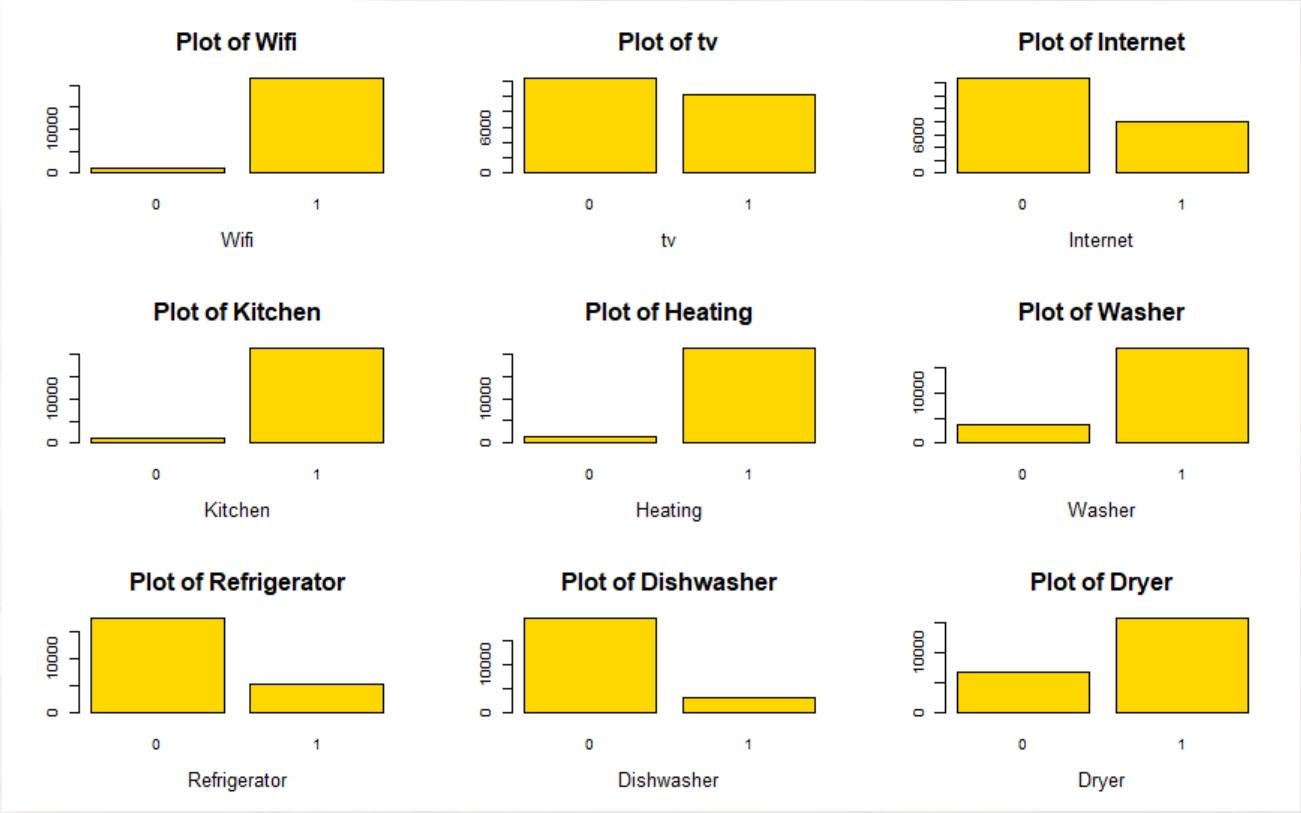
EXPLORATORY DATA ANALYSIS

MODEL SELECTION

SHINY APP

# Exploratory Data Analysis

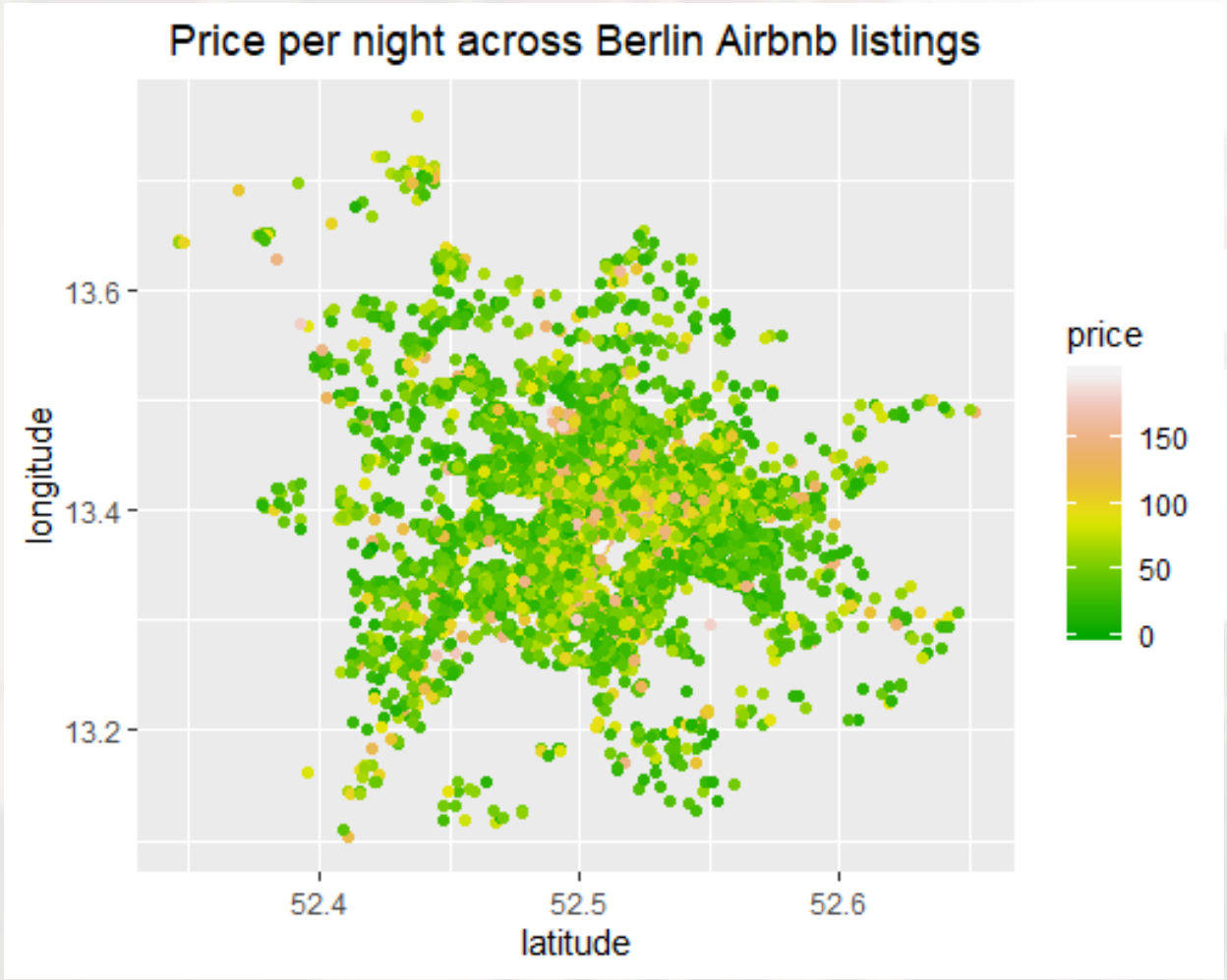


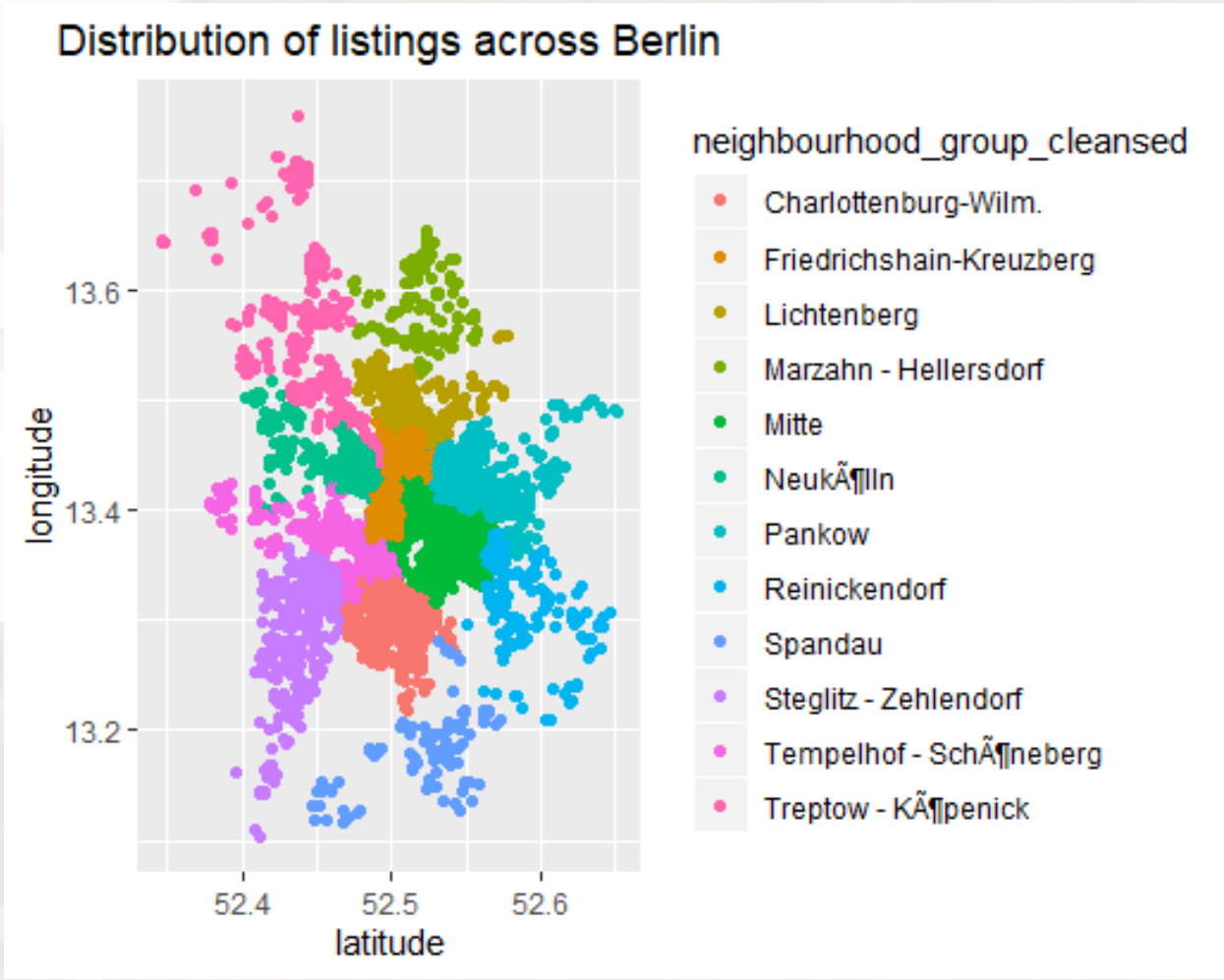




- Price are higher on Fridays and Saturdays compared to other days of the week

- Analysis of amenities in existing Airbnb listings at Berlin.
- TV and Internet has significant values for comparison

# Price per night and Neighbourhoods of Berlin



# Text Analysis


Transit/ Connectivity


House Rules

# Outline

PROBLEM STATEMENT

DATA CLEANING AND TREATMENT

EXPLORATORY DATA ANALYSIS

MODEL SELECTION

SHINY APP

# Data Modelling

**Model 1 - Multiple Linear Regression**

- Using all variables, we created a Multiple Linear Regression to predict price of the house
- Divided the dataset into Train (80%) and Test (20%)
- RMSE on Test dataset = 22.896

# Data Modelling

**Model 1 - Multiple Linear Regression**

**Model 2 - Random Forest Regression**

- Using RandomForest with Number of Trees = 500, extracted important variables that influence the price of the property.
- Using these variables, built future models for prediction.
- Important Variables selected through Random Forest are RoomType, CleaningFee, Bedrooms, Neighbourhood_Group, Beds, Bathrooms, Cancellation Policy, TV and Internet.
- RMSE on Test dataset = 22.621

# Data Modelling

**Model 1 - Multiple Linear Regression**

**Model 2 – Variable Selection using Random Forest Regression**

**Model 3 - Gradient Boosting Regression**

- Built a Gradient Boosting model (GBM) with Number of Trees = 1000.
- Used Cross Validation test with 10 folds to test the RMSE on the test dataset.
- RMSE on test dataset = 22.579
- Selected **Gradient Boosting** model based on lowest test RMSE

# Outline

PROBLEM STATEMENT

DATA CLEANING AND TREATMENT

EXPLORATORY DATA ANALYSIS

MODEL SELECTION

SHINY APP

# References

**R shiny layout:**
https://shiny.rstudio.com/articles/layout-guide.html
https://www.rdocumentation.org/packages/shinybootstrap2/versions/0.2.1/topics/fluidPage

**Word Clouds:**
https://datascienceplus.com/building-wordclouds-in-r/

**ggplot:**
https://ggplot2.tidyverse.org/reference/scale_gradient.html

# THANK YOU