STAT-S670

Exploratory Data Analysis

Final Project Report

Analyzing Airbnb Listings Prices

Team Airbnb London

Meet Palod
Nakul Havaldar
Sai Teja Burla
Saurabh Damle

1. Introduction

In the landscape of modern travel, Airbnb stands as a beacon of innovation, redefining the hospitality sector by offering unique accommodation experiences worldwide. Understanding the dynamic interplay of factors that contribute to Airbnb pricing is necessary in navigating this evolving market. Analyzing these factors not only sheds light on the essence of what shapes the cost of a stay but also holds the key to unlocking strategic insights for hosts and helps guests to make informed decisions. As the travel industry continues to transform, dissecting these elements becomes an essential compass for both hosts seeking to optimize revenue and guests aiming to find the perfect blend of value and experience in their accommodations.

2. Dataset

The dataset utilized in this project is sourced from the research study, "Determinants of Airbnb prices in European cities: A spatial econometrics approach"[1] and is hosted on Kaggle[2]. The data was collected through a systematic web-scraping experiment. The experiment targeted Airbnb listings in ten major European cities. Prices were collected four to six weeks ahead of the travel dates, encompassing the total accommodation cost, including reservation and cleaning fees. Each city yielded two datasets: one capturing weekday offers (Tuesday to Thursday) and another capturing weekend offers (Friday to Sunday). For the purpose of this project, data from only one of the ten cities (London) has been analyzed.

The data for London Airbnbs is split into two csv files.

1. **london_weekdays.csv** - This file contains data for weekdays. It has 4615 rows and 20 columns.
2. **london_weekends.csv** - This file contains data for weekends. It has 5380 rows and 20 columns.

The columns that are used in this project are described in detail as follows:

Column Name	Column Description	Datatype
Id	Unique ID for all the entries	Numeric
realSum	The total price of the Airbnb listing	Numeric
room_type	The type of Airbnb room being offered (e.g. private, entire home/apt, etc.)	Categorical

room_shared	Whether the listing is shared or not	Boolean
room_private	Whether the room is private or not	Booleant
person_capacity	The maximum number of people that can stay in the room	Numeric
dist	The distance of the listing from the city center	Numeric
lng	The longitude of the listing	Numeric
lat	The latitude of the listing	Numeric

3. Research Goals

The main goal is to identify the various factors that affect the price of the Airbnb listings, and the following are the sub-questions that we are planning to answer:

1. Do the Airbnb prices differ between weekdays and weekends, and if so, how?
2. How do factors like distance from the city center, room type, room size, etc. affect pricing of these Airbnbs?

4. Methodology

We began by using a correlation plot to guide our selection of variables for exploration, identifying three prominent choices: person_capacity, room_type, and room_privacy. However, considering the similarity between room type and room privacy, we opted to not use room_privacy and decided to use dist based on intuitive reasoning and the fact that it is a continuous variable.

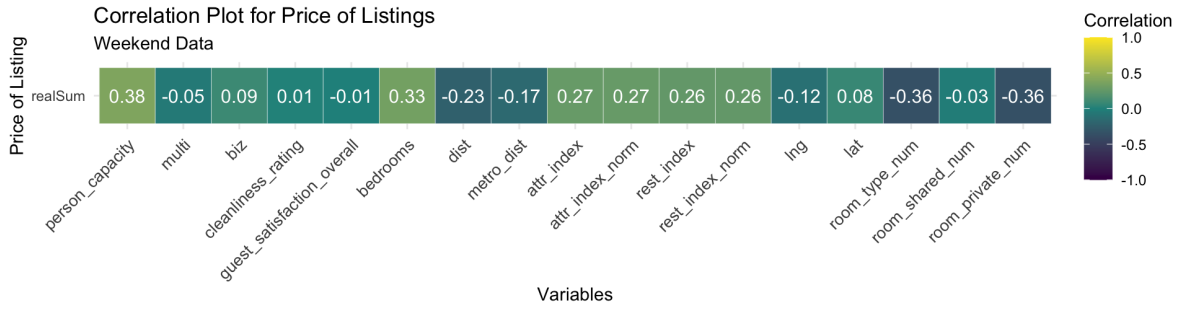

Fig. 4.1: Correlation Plot for Price of the Listings for Weekend Data

4.1 Cleaning the Dataset

We did the following cleaning for each of the four variables in question:

1. **realSum** -> Used boxplot to detect outliers and filtered the data to include listings with price more than 1000 EUR.(less than 2% data)
2. **dist** -> Used boxplot to detect outliers and filtered the data to include listings with distance more than 13 km.(less than 2% data)
3. **room_type** -> The barplot that was used revealed that one of the categories, 'Shared Room', accounted for less than 0.5% of the data, prompting its removal as it lacked sufficient data for meaningful analysis.
4. **person_capacity** -> The barplot that was used showed uneven distribution among the six capacities. Due to limited data for capacities 5 and 6, we combined them into a single category labeled as 5 for further analysis. In the report, any mention of room capacity 5 include rooms with 5 or 6 people capacity.

Please refer to the appendix to see the plots.

4.2 Exploratory Data Analysis

In this phase we tried to find out the trends and relationships between the three variables mentioned before to better understand how we could make our final model to predict prices of the Airbnb's.

General distribution of the prices with respect to room types:

For weekdays, the average price and median price is 360 Euros and 256 Euros respectively, and 75% percent of listings are between 167 and 435 euros. For weekends, the average price and median price is 364 Euros and 268 Euros respectively, and 75% percent of listings are between 175 and 438 Euros.

This suggests that prices on weekends are slightly higher than prices on weekdays. This could be attributed to various factors such as increased demand for accommodations during weekends as most people take short trips on the weekends.

From fig 4.2.1, it can be seen that the number of reservations on weekends is higher.The number of rooms booked for the type Private room is the highest and the majority of the houses can be seen in the range of 100 to 750 euros range.

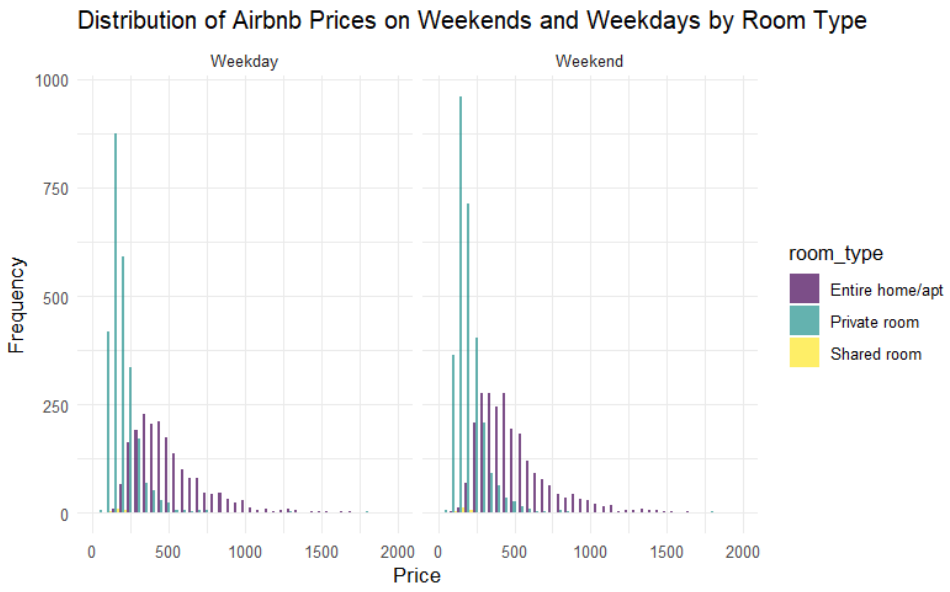

Fig. 4.2.1 : Distribution of Airbnb Prices on Weekends and Weekdays by Room Type

Relationship of Average Price by Room Type:

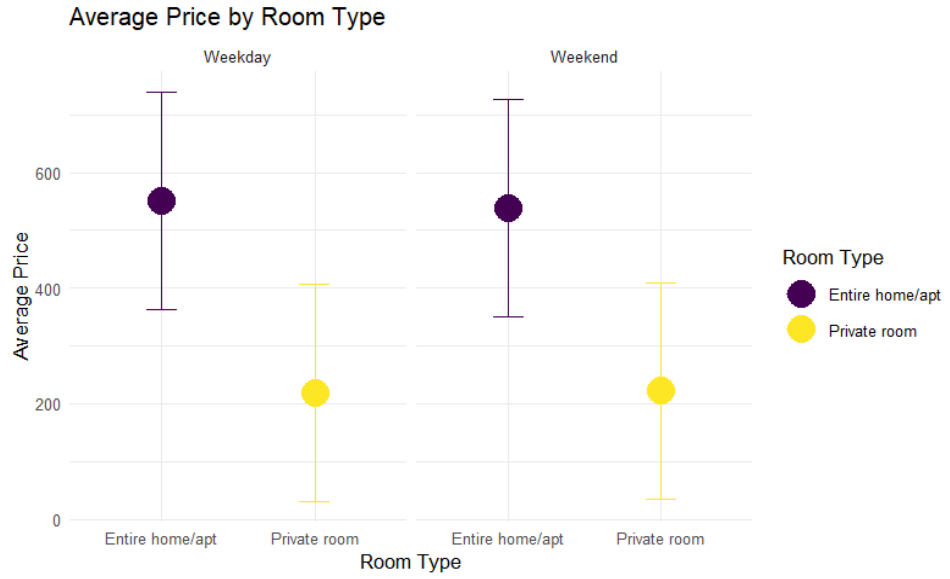

Fig 4.2.2 : Plot of the average price of a room by room type

The plot 4.2.2 tries to capture the relationship of how the average price looks like for the variable of interest of type of rooms. The average price all over London for all the 3 room types looks the same. For an entire home

people are listing on an average of 500 euros whereas a private room type airbnb on an average is listed for 200 euros for weekend and weekday. The data is pretty large for entire home type and private room type apartments therefore on an average people will be paying 500 euros and 200 euros respectively for the room types here.

Relationship between average price, room type and person capacity:

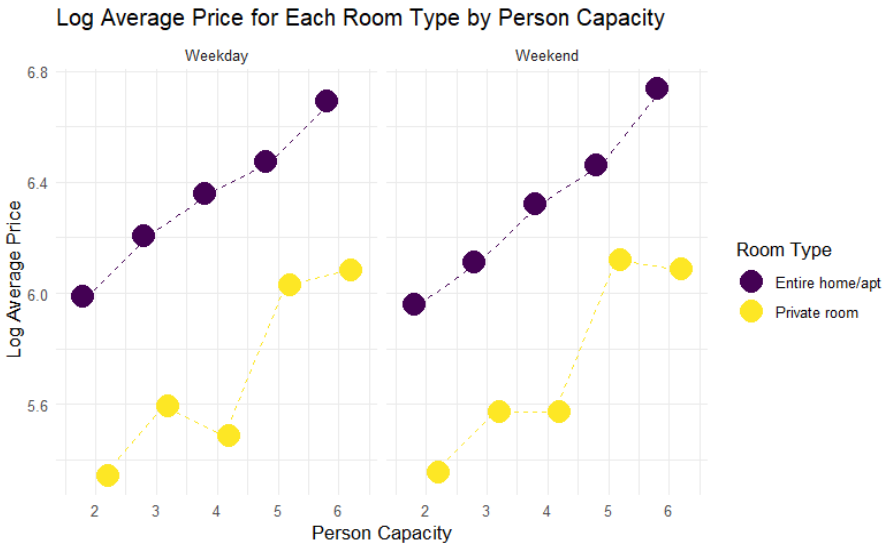

Fig. 4.2.3 : Relationship between log average price of each room type and person capacity

In the plot 4.2.3, a log scale is used to incorporate all the high ranges and bring them down to a scale where visualization is easy. For an entire home we can see on weekends and weekdays it follows a similar upward trend when the number of people starts increasing. The prices start from around 600 euros for a 1 person room and go all the way up till 700 euros for 6 people capacity. For Private room type for a weekday it can be said that it starts from around 530 euros on an average and increases up till 600 euros. For the weekdays the price for a private room drops for a capacity of 4 people and sees a steep increase for person capacity of 5. We think it might be due to the fact that there would be more families in that zone of 4 people and the demand would be more so for competitive advantage. Maybe the prices are lesser there. A similar trend for private rooms can be seen for the weekends but there is not a dip in prices for 4 person capacity but it remains same as for 3 people capacity type. Then on both kinds of days (weekends and weekdays) the average log price for a private room increases to around 600 Euros.

Relationship with respect to the distance:

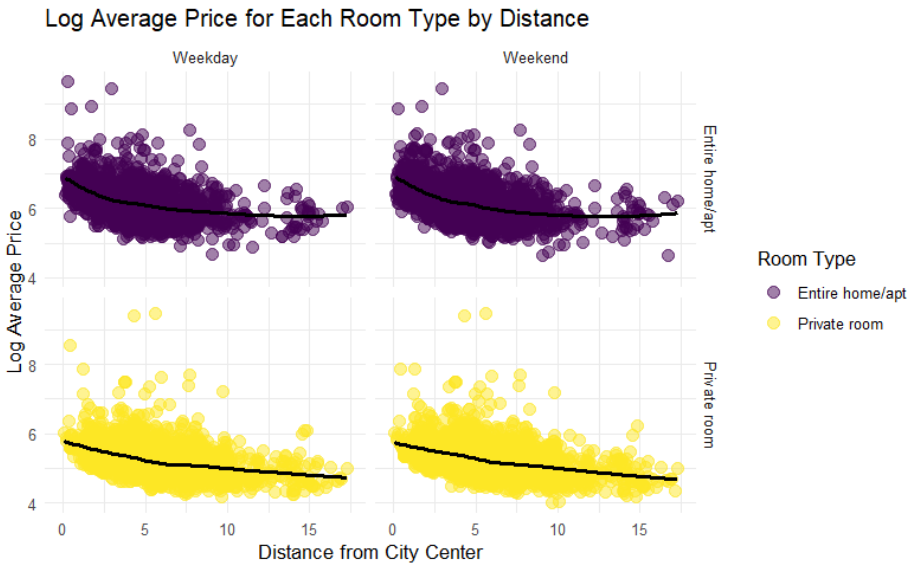

Fig. 4.2.4 : Relationship between Log Avg. price of room type and distance

The average and median distance of listing from the city center is 5.33km and 4.92km respectively, while 75% percent of listings are within 3.55 and 6.84 km radius.

Fig 4.2.4 shows the relationship between the log average prices and how they vary with respect to the distance from the city center. In both the kinds of days (weekends and weekdays) it can be seen that the trend is decreasing as the distance from the city center increases. So as one moves aways from the city center the average prices of the rooms keep on decreasing. For an entire home/apt it can be seen that the prices start from 700 euros and go all the way down to 600 euros as you move away from the city. For a private room type it starts from around 600 euros and would go until 400 euros as you move to the farthest on an average.

Relationship between price distance and person capacity:

The main observations from the below plot can be seen that in an overview the average price is decreasing when the distance increases from the city center. The same trend can be seen for entire home apt and private room house types. So taking into account all the person capacity it can be seen the trend of prices going down as we move away from the city center.

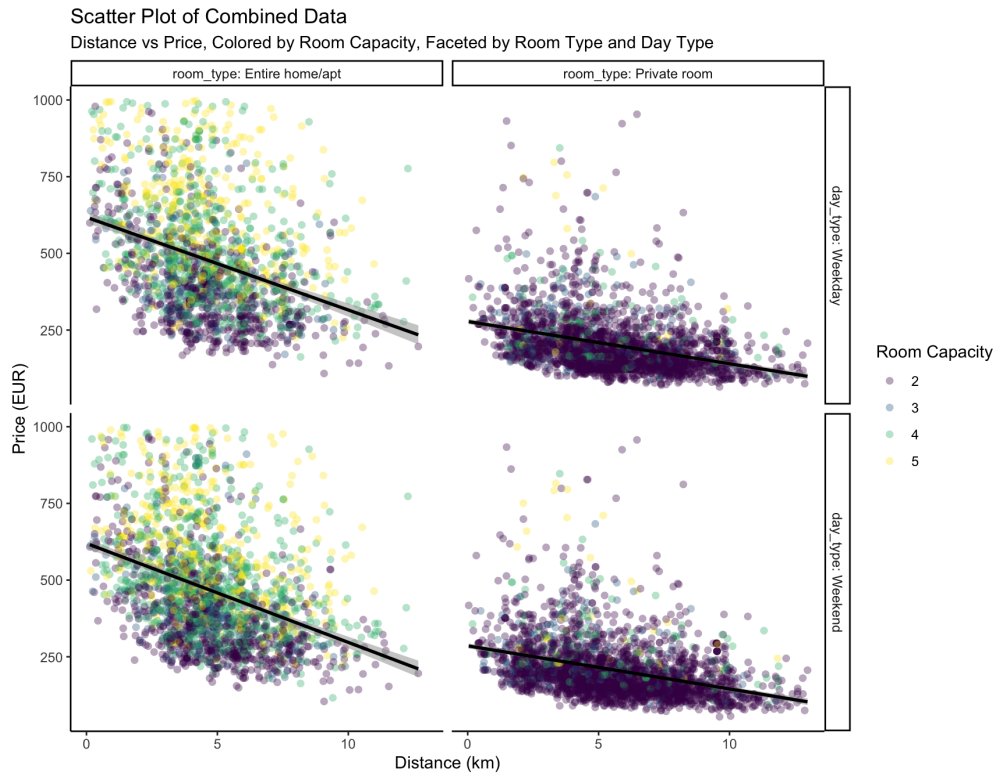

Fig. 4.2.5: Relationship between price, room type. Distance and person capacity

Added to the multivariate analysis of the variables of interest, exploration of weekday vs weekend prices for same listings has been analyzed in Fig 4.2.6. It reveals a notable number of listings maintaining consistent prices throughout the week. However, among listings showing price fluctuations, the count of listings with higher prices during weekends seem to be nearly double that of those with higher weekday prices.

Upon the analysis and comparison with actual tourist map of London, we found that the places where the weekend prices are higher and the difference in the weekend and weekday data was around 300 euros, those places are situated in the central part of the city near the River Thames have major tourist attractions like tower bridge, the London eye and the HMtower of london.

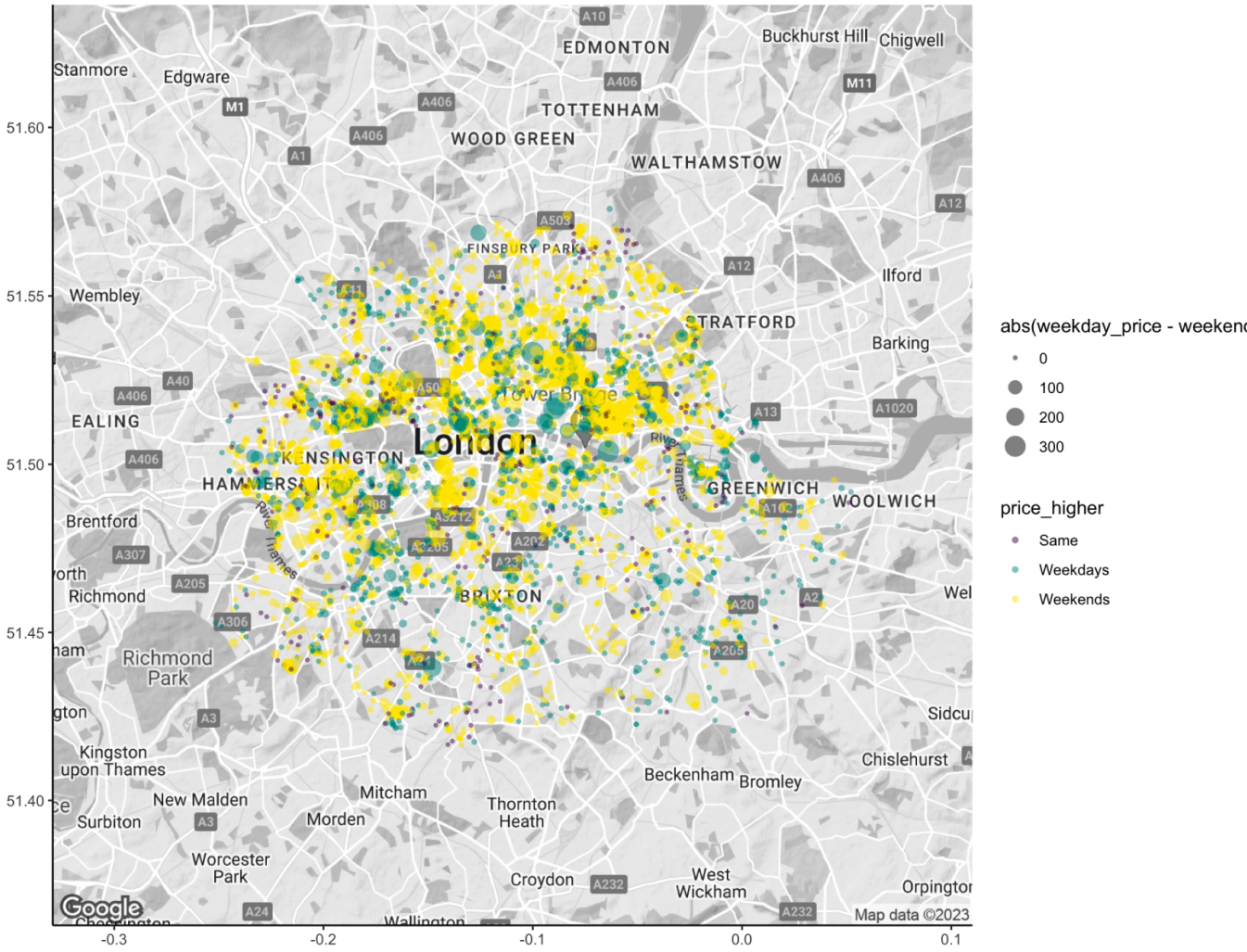

Fig 4.2.6 : Variation in Airbnb Prices between Weekdays and Weekends

While we can notice clusters of higher weekday prices just to the northeast of greenwich,after doing some research about the geography of London, we found that it is located in the neighborhood of “Canary Wharf” which is a business complex, so it makes sense that the weekday prices are higher than the weekend near such commercial locations.

Following our analysis, we sought to quantify the extent of price variation between weekdays and weekends. For this purpose the ratio of average prices of weekdays by weekends within each category was calculated and the observations were as follows:

- For Entire Home/Apt listings, on average, the weekend prices were 1 or 2% more than the weekday prices.
- For Private Room listings, on average, the weekend prices were 3 - 8% more than the weekday prices.

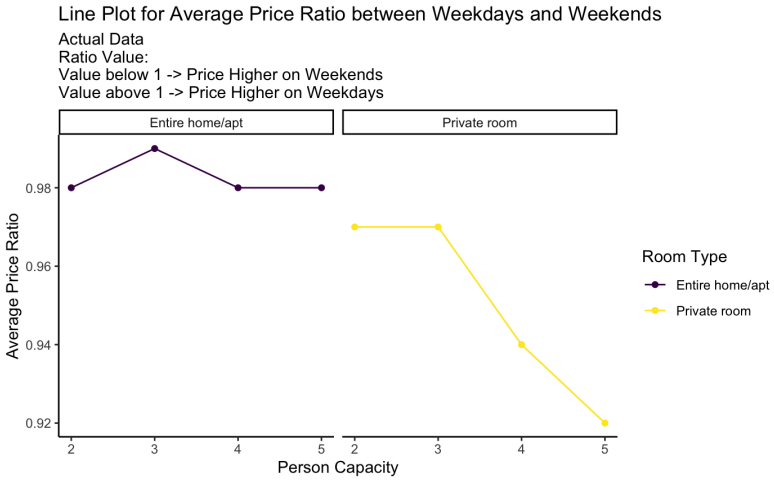

Fig 4.2.7 : Line Plot for Average Price Ratio between Weekdays and Weekends

Considering the findings from our exploratory analysis, we moved forward to the phase of experimenting with various models.

4.3 Models

After performing exploratory data analysis on the selected variables and seeing their trends we realized that a linear model might not be the best fit for replicating the trends in our data so we decided to use GAM (Generalized Additive Model) instead.

Model 1:
To begin with, we decided to go with a simple GAM of price v/s distance and the person capacity.
Formula -> realSum ~ s(dist) + person_capacity
This model's R squared values: Weekdays -> 46.3% and Weekends -> 45.4%
Inferences:

1. For weekdays and weekends we see similar trends in terms of the predicted price values.
2. As the distance increases price decreases and as the room capacity increases price increases.
3. For the following cases we see this model is not performing as good:
    a. Room Type - Entire Home/Apt with Person Capacity - 2
    b. Room Type - Private Room with Person Capacity - 3, 4, 5 (Probably overfitting because of lesser data compared to other categories)

So due to the above inferences even though the model is kind of preserving the trends of the actual data it could be modified to perform better by maybe using a more complex model which is exactly what we explored next.

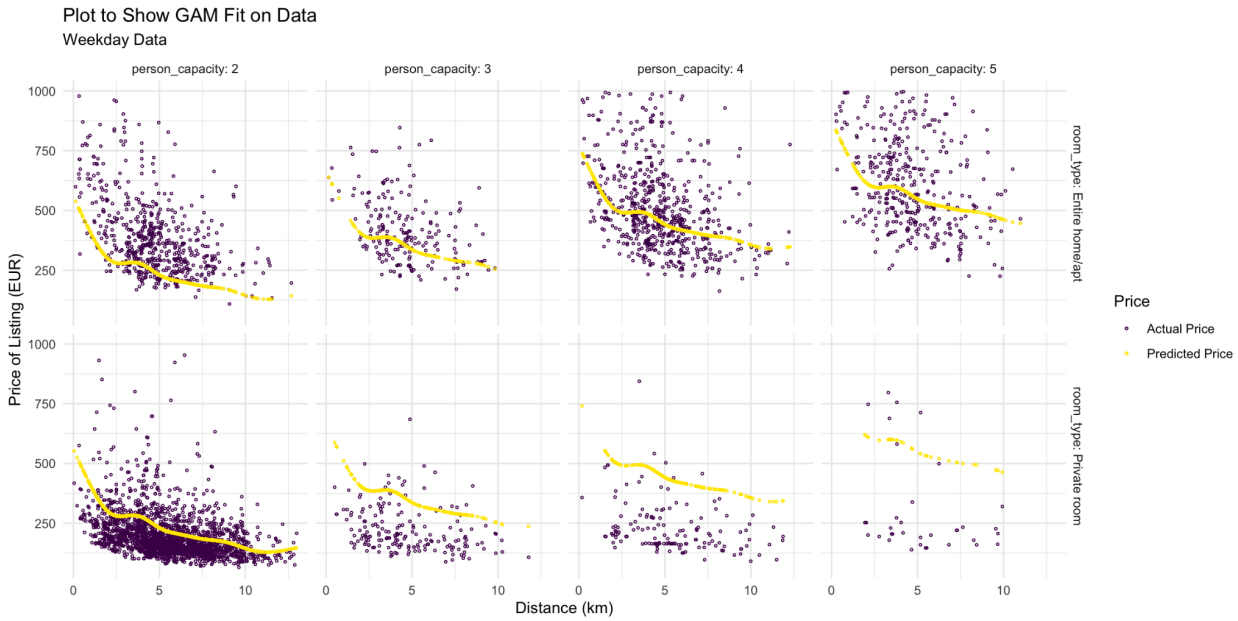

Fig 4.3.1 : Plot on Weekday Data for Actual and Predicted Prices vs Distance for Gam Model - 1

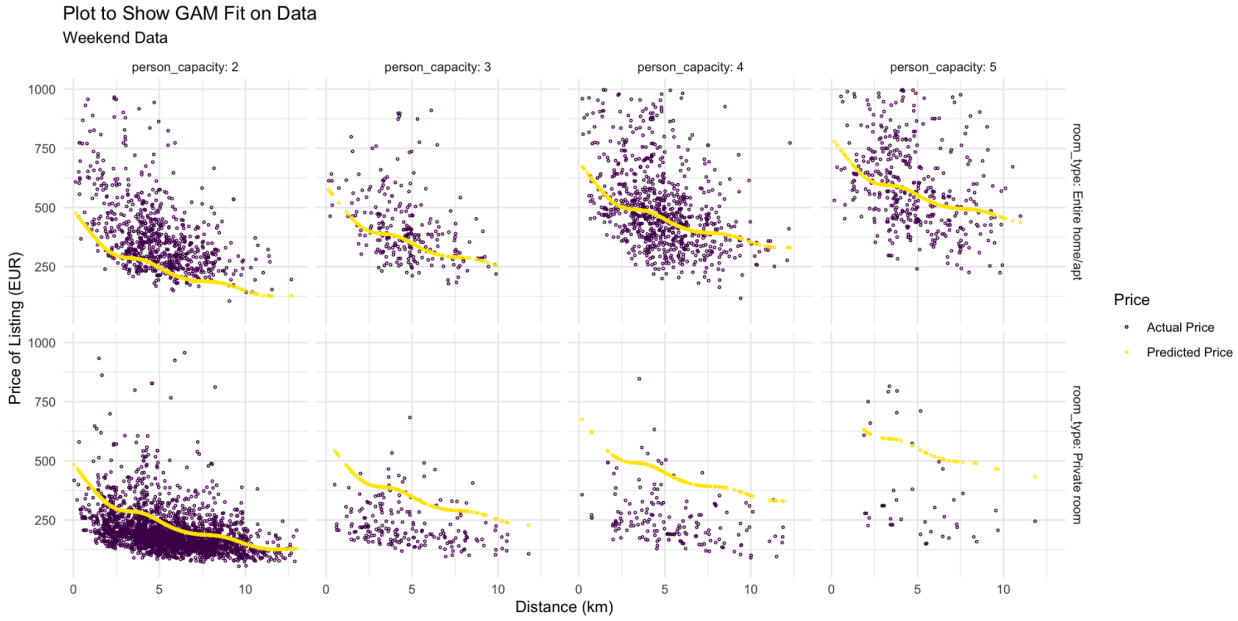

Fig 4.3.2 : Plot on Weekend Data for Actual and Predicted Prices vs Distance for Gam Model - 1

Model 2:

Next, we decided to use a more complicated GAM which is price vs distance with interaction between Room_Type and Person_Capacity.

Formula -> realSum ~ s(dist) + room_type * person_capacity

This model's R squared values: Weekdays -> 62.6% and Weekends -> 60.4%

Inferences:

1. For weekdays and weekends we see similar trends in terms of the predicted price values as it was seen in the previous model as well.
2. As the distance increases price decreases and as the room capacity increases price increases similar to the actual data trends.
3. R squared value is better than the previous model and we can see that this model is capturing the trends of the actual data better much better and it seems like it is not overfitting as well.

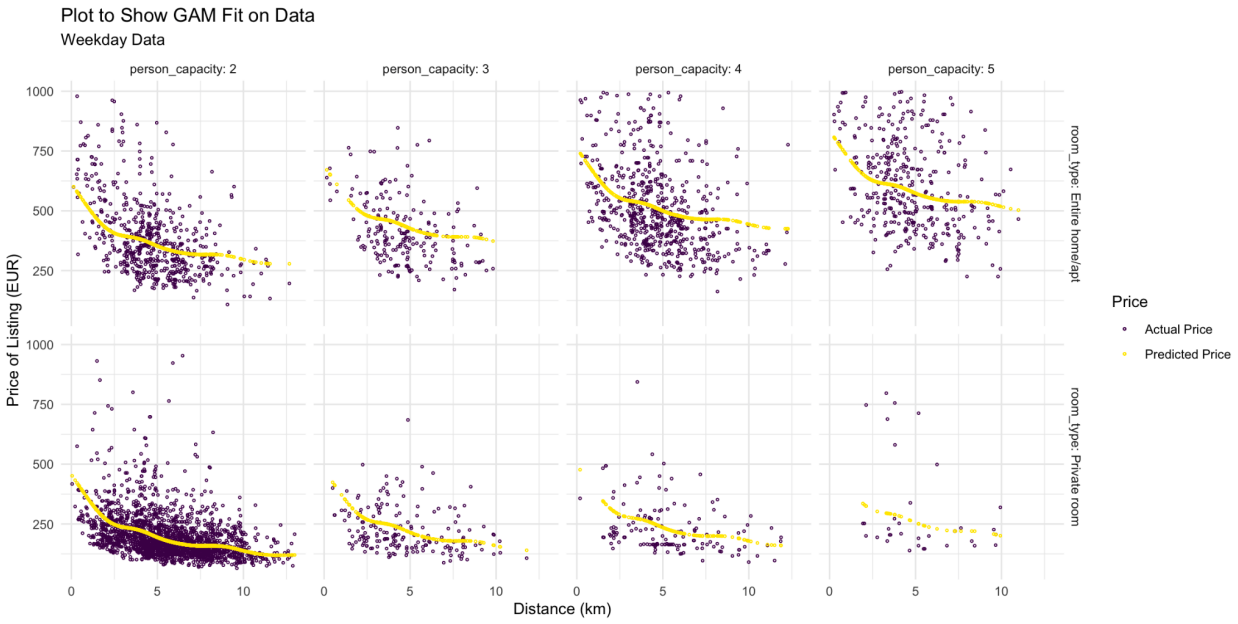

Fig 4.3.3 : Plot on Weekday Data for Actual and Predicted Prices vs Distance for Gam Model - 2

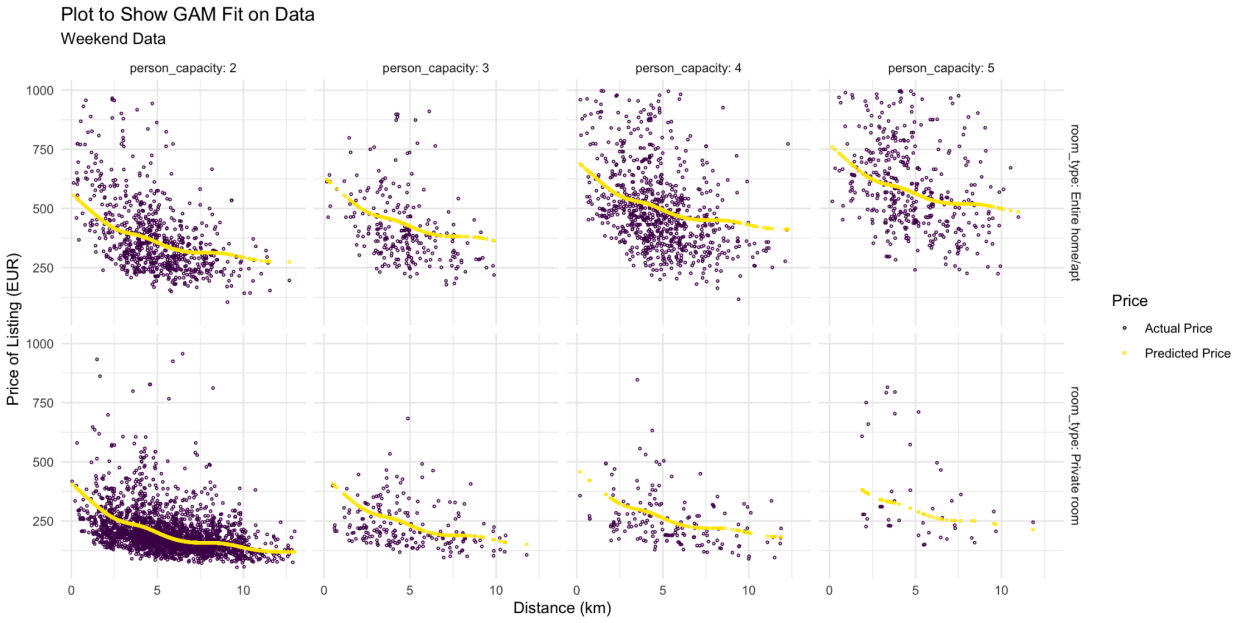

Fig 4.3.4 : Plot on Weekend Data for Actual and Predicted Prices vs Distance for Gam Model - 2

After that, like we did in our exploratory data analysis, we tried to quantify the model's performance by using synthetic data and making predictions for that data using the above model and then finding the ratio of average weekday and weekend prices for each of the variable types. Notably, the model demonstrated a consistent trend which is kind of similar to the one we saw for our actual data, indicating a considerable preservation of the trends of the actual data. This would act as a proof that the model is doing fairly well and is probably not overfitting.

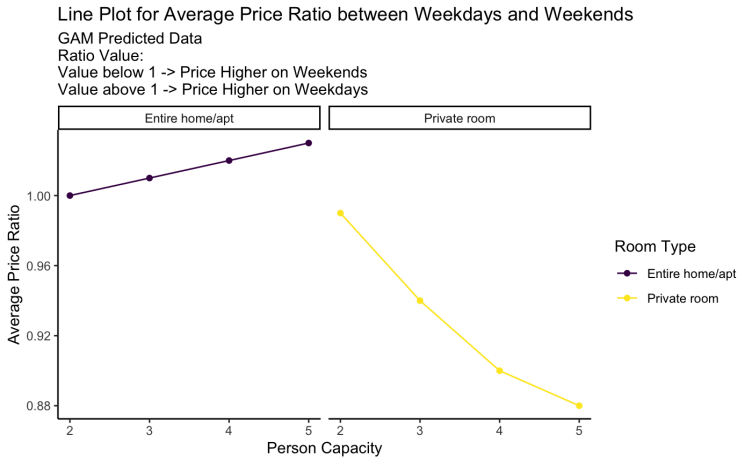

Fig 4.3.5 : Line Plot for Average Price Ratio between Weekdays and Weekends for Predicted Data

5 Conclusion

In summary, our analysis commenced with identifying the top three variables that affect the price of a listing, distance, room type(private room/ entire apartment), and capacity of listing through a correlation plot and intuition. Subsequently, we conducted pre-processing steps and proceeded with multivariate analysis, revealing intriguing trends. Notably, the average weekend prices exhibited an approximate 1-2% increase for the entire home/apt and a 3-8% surge for private rooms compared to weekdays. Both weekday and weekend trends showcased consistent price dynamics: prices decrease with greater distance, larger rooms demand higher prices, and entire homes are marginally more expensive than private rooms. Moving to modeling, our initial attempt with a simple GAM using two variables yielded decent but not optimal results. Consequently, we adopted a more complex GAM incorporating all three variables and their interactions, resulting in our most effective model which preserved the trends of our actual data and seemingly did not overfit the data as well.

6 Limitations and Future Work

We had to remove some data points as outliers to make the model results more logical and interpretable. Using more robust outlier handling methodologies could have prevented this data loss. The current results are based on a limited set of three variables. It's important to acknowledge that numerous other factors could potentially influence Airbnb prices. Therefore, exploring more complex models could have made the results even better. Additionally, the analysis is confined to a single city, London. While valuable insights have been derived, it's plausible that certain trends might remain undiscovered due to this limited scope.

Expanding the scope of variables considered in the pricing model presents an exciting direction for future research. Incorporating additional factors such as neighborhood characteristics and seasonal variations could offer a more comprehensive understanding of Airbnb pricing dynamics. Furthermore, broadening the analysis to encompass multiple cities could unveil trends and patterns that might not be evident within a single-city study. Comparative analyses across various urban landscapes could uncover regional nuances and industry trends, enriching the overall understanding of Airbnb pricing strategies and their determinants.

References

[1] Gyódi, K., & Nawaro, Ł. (2021). Determinants of Airbnb prices in European cities: A spatial econometrics approach. Tourism Management, 86, 104319.
[2] Airbnb Prices in European Cities. (2023, February 20). Kaggle. https://www.kaggle.com/datasets/thedevastator/airbnb-prices-in-european-cities
[3] Visit London. https://www.visitlondon.com/things-to-do/london-attractions-map

Appendix

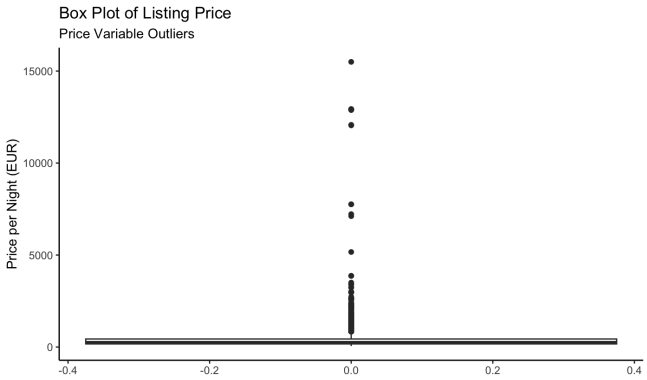

Fig. 1: Box Plot for Outlier Detection in Price of the Listings

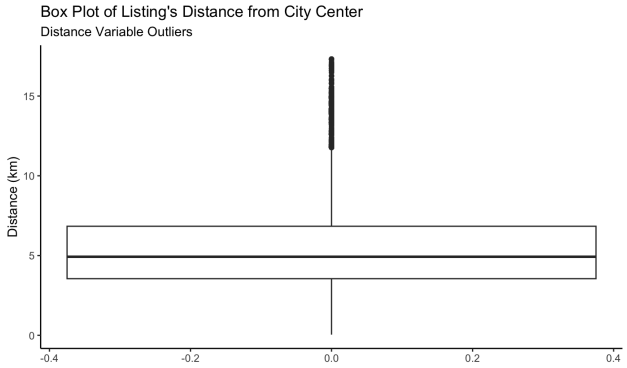

Fig. 2: Box Plot for Outlier Detection in Distance

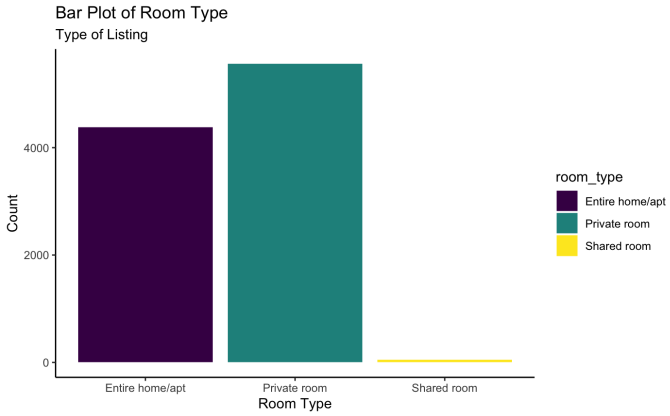

Fig. 3: Bar Plot for visualizing Room Type

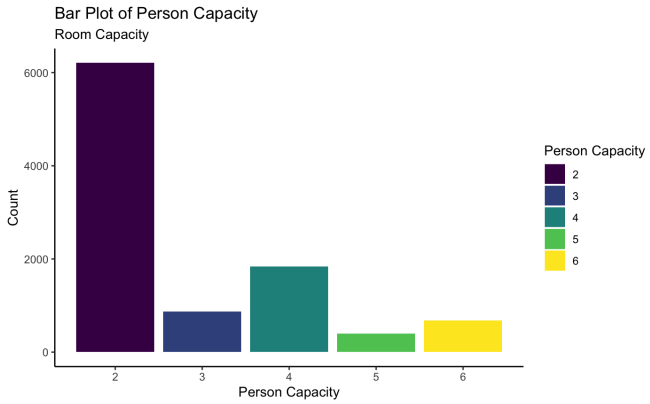

Fig. 4: Bar Plot for visualizing Person Capacity

We had also tried out a LOESS Model which is price vs interaction between distance, room_type and person_capacity. The results are as follows: