i. Introduction

On Kaggle, an Airbnb price dataset was proposed. This dataset can be used for finding out the determinants of price. I offer three initial questions: 1) How is price distributed? 2) How does the price pattern differ among places? 3) What would affect the room price? Different visualisation techniques will answer them. Furthermore, follow-up questions will be proposed, including: 4) How do the expensive rooms differ from the cheap rooms regarding the distance to the metro station and city centre? 5) Why can some rooms far from the city centres and far from the metro stations sell at a high price? 6) Is the reason why rooms in less desirable locations command high prices in London because they have the capacity to accommodate more people? R studio and ggplot2, along with some other libraries, will be used. The charts would be designed critically based on visualisation theory.

Section ii will introduce the data and the initial transformation. Section iii will present the questions and the fitness of the data. Section iv describes the software, the process of proposing new questions and the visualisation strategies. Section v will provide the charts, analysis for each question, a critical discussion of each graph, and the new question. Section vi will reflect on the process.

Word count: 2903 (not including bibliography and appendix)

ii. Data Description

The dataset (*Airbnb Prices in European Cities*, n.d. cited from Gyódi & Nawaro, 2021) covers the Airbnb price data in Amsterdam, Athens, Barcelona, Berlin, Budapest, Lisbon, London, Paris, Rome, and Vienna. For each city, there are one file for weekday data and one file for weekend data. Before analysing the data, the data were merged into one file by the R code in "mergeData.R." These data were saved into "airbnb.csv" after the null value was checked.

realSum	Q	Dimension	price
room_type	N	Dimension	Type of the room
room_shared	N	Dimension	If the room is shared
person_capacity	O	Dimension	The number of guests who can live
host_is_superhost	N	Dimension	If the host is a super host
multi	N	Dimension	If this data is for multiple rooms
biz	N	Dimension	If the data is for business purposes
cleanliness_rating	N	Measure	Rating for cleanliness
guest_satisfaction_overall	Q	Measure	Rating for satisfaction
bedrooms	Q	Dimension	Number of bedrooms
dist	Q	Dimension	Distance to the city centre
metro_dist	Q	Dimension	Distance to metro station
attr_index	Q	Dimension	Description not given
attr_index_room	Q	Dimension	Description not given
rest_index	Q	Dimension	Description not given
rest_index_room	Q	Dimension	Description not given
lng	Q	Dimension	Geolocation
lat	Q	Dimension	Geolocation
Location (after merging)	N	Dimension	City

Time (after merging)	N	Dimension	Weekend or weekday

iii. Questions

Three initial questions are the following: 1) How is price distributed? 2) How does the price pattern differ among the different places? 3) What would affect the room price?

Our data is suitable for answering the initial questions because price, location and other attributes are given.

After answering those questions, questions 4, 5 and 6 are generated (figure 3.1): 4) Why are some rooms far from the city centre and far away from the metro stations that can still sell at a high price? 5) Why can some rooms far from the city centres and far from the metro stations sell at a high price? 6) Is the reason why rooms in less desirable locations command high prices in London because they have the capacity to accommodate more people?

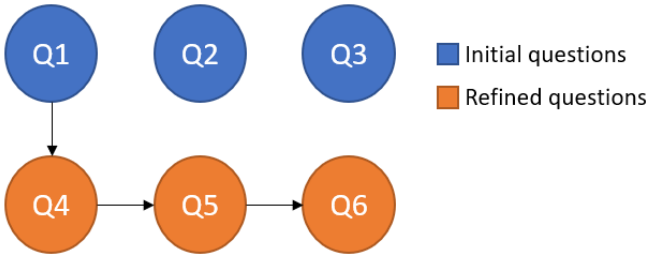


Figure 3.1. Question-Generating Process

iv. Methods

To visualise the information, the software used in this project is R, R Studio, dplyr, ggplot2, corrplot, and tidyr. The coding skill for this project comes from the R graphics cookbook (Chang, n.d.) and the w3school R tutorial (R Tutorial, n.d.).

The visualisation knowledge was learned from COMP3021 (Zhou & Xu, 2023), Mackinlay (1986) and Oetting (2022).

I first answered all the initial questions. Then, I intuitively generated question 4. After answering question 4, the outliers surprised me, so I proposed question 5. Then to dig deep into the data of London, question 6 was formed.

The remaining of this section will introduce each question's data transformation and visual encoding strategies. The critical discussion of visualisation will be explained in the result section. The chosen graphs are in the Analysis section, and the candidate graphs are in the Appendix.

Question 1

For question 1, I first tried a scatter plot and a histogram and eventually decided to use two histograms for the cheap rooms and expensive rooms, respectively. I transferred the “realSum” into “priceGroup” with two possible values “low” for the rooms whose “realSum" are lower than 2500 and “high” for those higher than 2500. In the chart, the price is encoded into position x and the count of rooms is encoded into position y.

Question 2

I chose a boxplot with a logarithmic "realSum" scale for this question after comparing it with the linear scale box plot and the scatter plot. The data are grouped by location. Locations are mapped into position y, and prices are mapped into position x. The degree of price dispersion is implicitly encoded into the length of each box.

Question 3

The sorted bar chart was chosen after comparing it with the sampled scatter plot matrix and the correlation heat map matrix. In the sorted bar chart, the variable name is encoded into position y, and the correlation coefficient is encoded into length and implicitly into position x.
To get the correct data, the correlation coefficient was first calculated. Then, before transforming the data, the four attributes (i.e., attr_index_room, rest_index, rest_index_norm, and attr_index) with no information on Kaggle were removed. After that, I converted "room_type", "room_shared", "room_private", and "host_is_superhost" to quantitative values. Sequentially, the coefficient matrix is calculated, and the column related to realSum is selected and converted to the long format with two columns: the variable name and their correlation coefficient to price. Finally, the data is sorted by the correlation coefficient and plotted.

Question 4

"juxtapositioned"(Gleicher, 2018) scatter plots and two boxplot-barChart pairs for "dist" and "metro_dist" were compared. The latter (figure 4.4) was chosen, and position x of the boxplot outliers is jittered. Two boxplots show the condition of distance to the city centre and to the metro station of the rooms with different prices. The bar charts show the percentage of outliers. The box plots map the price to position x and the distance to position y. The bar charts map the price to position x and the percentage of outliers to length and implicitly to position y.
To transform the data, rooms are grouped into four different price ranges. The rooms whose price were lower than the 25% percentile was labelled as "low"; the rooms whose price were between 25% to 50% were labelled as "mid_low"; the rooms whose price were between 50% to 75% were labelled as "mid_high", and the rest were labelled as "high".
To jitter the position of the outliers, a new column was created to determine if the room was an outlier, and those outliers were plotted using the geom_point function.

Question 5

To answer this question, I made one hypothesis: 1) The distribution of the city in the outliers is different from the ordinary data. This hypothesis would be assessed using visualisation. The chosen chart was a "superpositioned" (Gleicher, 2018) bar Chart. The colour was used for a different group, the location was encoded in position y, and the percentage was encoded in position x.
The data transformation pross is the following: 1) filter the rooms whose "priceRange" is "high"; 2) create a new column called "isHighOutlier", to indicate if the room is far from both city centres and metro stations; 3) group the data by "isHighOutlier" and calculate the percentage of each city.

Question 6

The graph used for this question is also a bar chart. The difference is that position x is for "percon_capacity", position y is for "percentage", and the colour is for "isHighOutlier".
To transform the data, I chose the data whose location is in London and whose "priceRange" is "high" (i.e., more expensive than 75% of the rooms). Secondly, a new column, "isHighOutlier", is created in the same way as in question 5. Then, data were transformed into relative number to make the different groups share the same y-axis. Finally, the data were summarised for each "person_capacity" and "isHighOutlier" group.

v. Results

This section will show the chosen chart, the data analysis, and the critical discuss of visual techniques. Follow-up questions will be introduced if they exist. The charts that were considered but not chosen will be shown in the Appendix.

Question 1

As shown in Figure 4.1, most of the house is cheaper than one thousand, but there are some expensive rooms.
The advantage of using Figure 4.1 is that this chart is not overplotted as the scatter plot (see appendix) does. This chart can show information about the expensive rooms hidden in a single histogram. The disadvantage is that the data is aggregated, which makes it less expressive, but it can still answer question 1.
The effectiveness of this chart is better than the others. Because the variables are encoded into position and length of the bars.
A new question: How do expensive rooms differ from cheap ones?

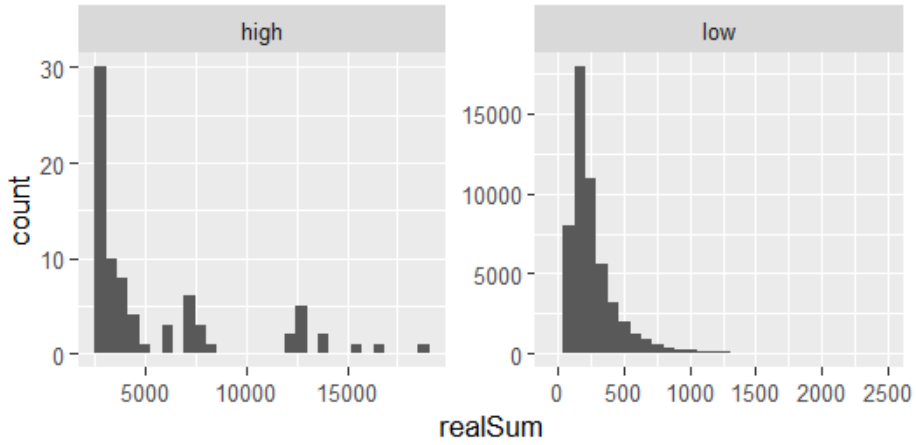


Figure 4.1. The number of rooms. Price Higher than 2500 (left) and Lower than 2500 (right)

Question 2

Figure 4.2 illustrates the price distribution of different locations. As the box plot shows, the average price in Amsterdam is the highest, whereas the average price in Athens is the lowest. The price difference in London is the greatest, although no low-price outlier exists in London.

Additionally, the outliers in London are generally more expensive than the other cities, but in Paris and Athens, more expensive records exist. The lowest price appears in Budapest, although the mean price is not the lowest.

This chart is more expressive than a bar chart (appendix), as the distribution of prices in each city is also displayed. This chart is less overplotted than the scatter plot (appendix). The x scale is logarithmic so that the outliers can be displayed, but as a cost, the logarithmic scale makes it less expressive. However, the alternative way, which is to hide the outliers, would also make the chart less expressive and make it challenging to find interesting information. Additionally, colour is used to encode a report's count, making it more expressive.

This encoding scheme is more effective than the other combination because positions are used for the most critical variable: location and price. The colour is used for encoding some information which is less important for the question.

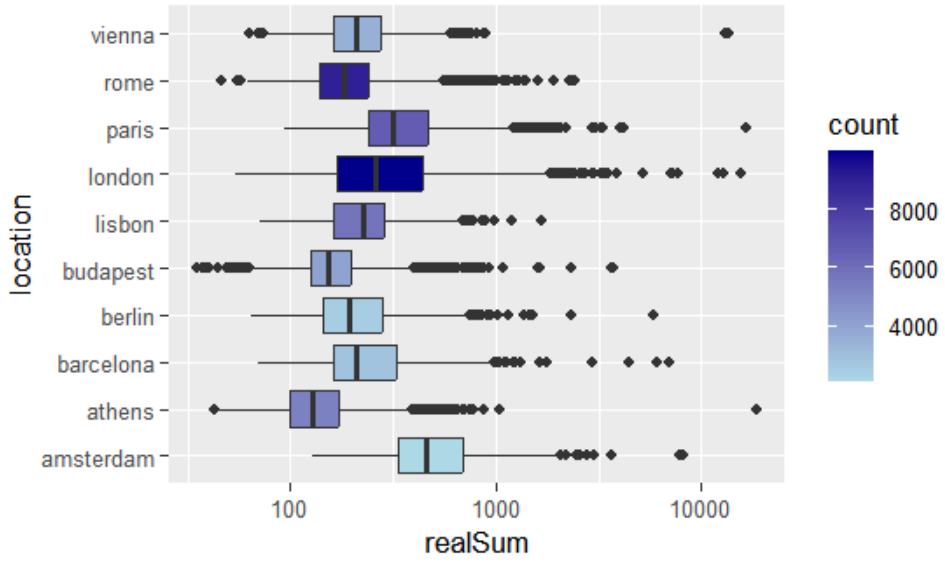


Figure 4.2. Price Distribution Box Plot for 10 Locations with Count and Logarithmic Scale.

Question 3

Figure 4.3 provides the information that can answer the question. As the figure shows, the most attribute that would affect the room price is the number of bedrooms, followed by the longitude and the latitude. Other attributes, including "person_capacity", "room_private", and "room_type", would also affect the price. However, "metro_dis", "host_is_superhost", "multi", "dist", "room_shard_n", and "biz" is less related to the price. Surprisingly, "cleanliness_rating" and "guest_satisfaction_overall" almost do not affect the price.

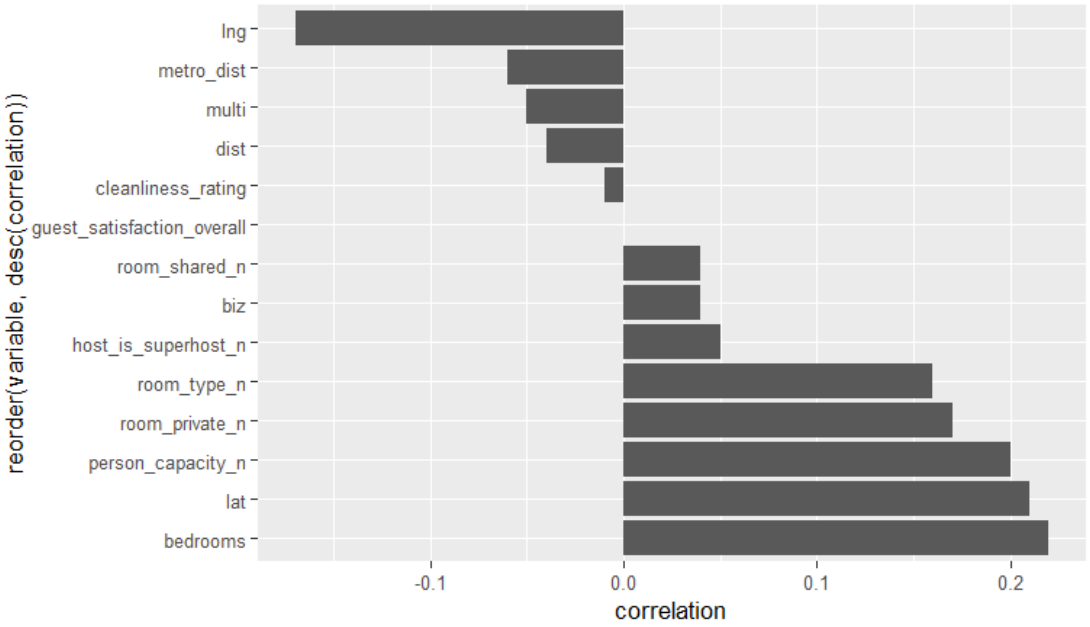


Figure 4.3. Correlation Coefficient of realSum with Other Attribute Bar Chart

This chart is better than the scatter plot matrix with sampled data and the correlation heat map (in the appendix). The scatter plot matrix is either overplotted or becoming less accurate due to sampling. The correlation heat map encodes quantitative data into colour, which is not accurate according to Mackinlay (1986). Another problem with both is that only the first row in these two unchosen graphs is relevant to the question, so space is wasted on unimportant information. However, in Figure 4.3, the most effective and accurate visual channels are used. Although there is less information than the correlation scatter plot matrix, the bar chart is suitable for the question.

Question 4

Figure 4.4 compares the condition of the distance from the room with a different price to the city centre and metro station. Most rooms share similar distances distribution regardless of price. In contrast, the distribution of the outliers and the percentage of the number of outliers of each price group differs.

The rooms in the lower price groups will have outliers with further distance than the higher price groups. The percentage of the outlier rooms far from metro stations is similar among different price groups. In contrast, the number represents the distance from the rooms to the city centre is different between different price group. It is more likely to have outliers far from city centres in the low-price group than in the high-price group.

The multiple views used for this question will be better than the others. Rather than the "juxtapositioned" scatter plots (see the appendix), Figure 4.4 illustrates the percentage and distance. It is also better than histograms. Because if the histograms are "superpositioned", they are overplotted, as there are four price groups. On the other hand, finding the difference between four histograms is like comparing the shapes if "juxtapositioned", which is ineffective.

By contrast, Figure 4.4 is better, because it encodes statistical information on distances into position, and the distribution of the outliers into density (implicitly). Position and density are more effective than shapes for quantitative data (Mackinlay, 1986). Additionally, Figure 4.4 is less

overplotted.

A new question: Why can some rooms far from the city centres and far from the metro stations sell at a high price?

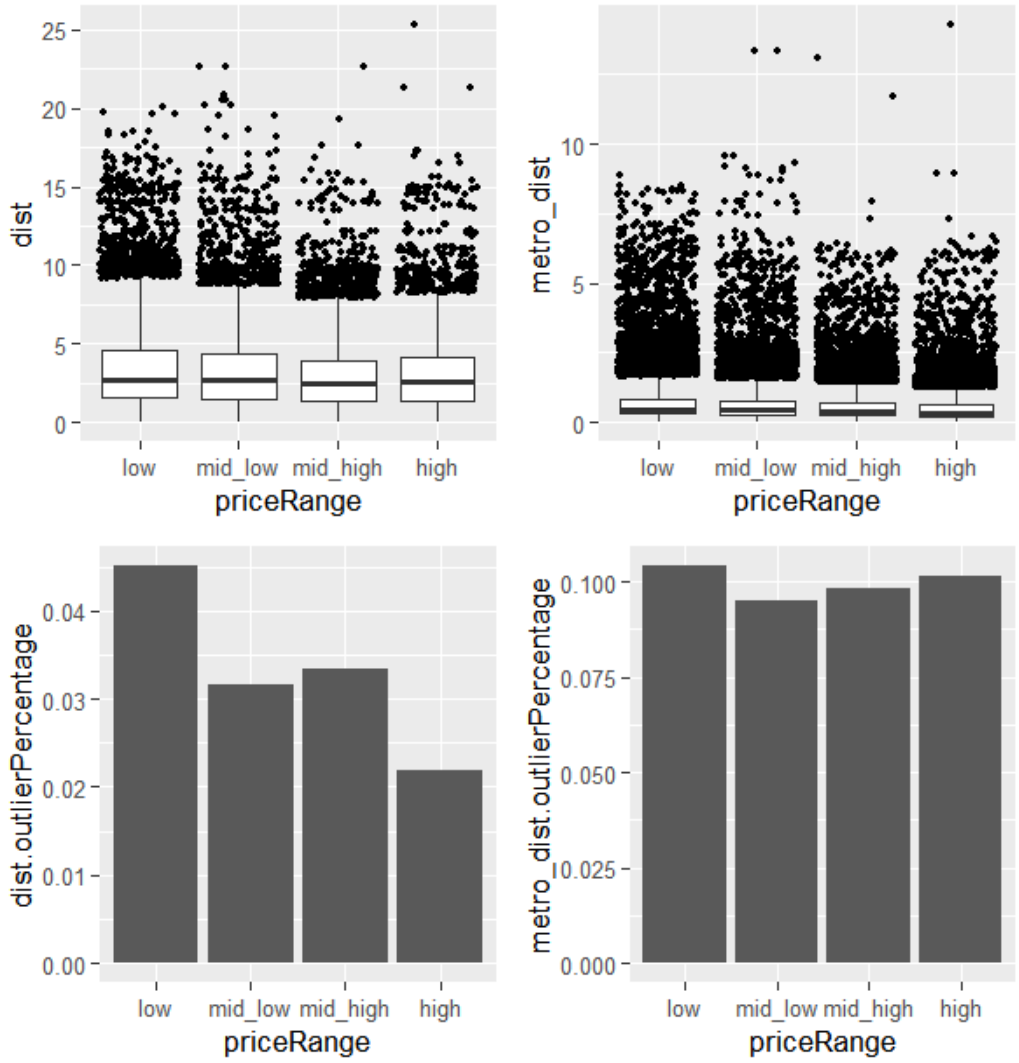


Figure 4.4. Room Distance and Metro Distance by Price Range with Outliers.

Question 5

I hypothesis that “The distribution of the city in the outliers is different from the ordinary data”. As Figure 4.5 shows, the rooms with a long distance but still sold at a high price are only in Vienna, London, and Berlin.

This chart is better for comparing different groups than the pie chart, as the length can be aligned and less colour is used. It is also better than the ring chart (not in appendix) because the area of each fan would make the chart less expressive. Figure 4.5 encodes an important quantitative variable into position x. Nominal data, “isHighOutlier”, is encoded into colour. The reason for not using colour for location and not using position y for isHighOutlier is that there are too many locations. The colour channel will be ineffective with too many colours (Zhou & Xu, 2023). The width of the red bar is the same so that the graph does not show more information than it has. This makes it more expressive (Mackinlay, 1986).

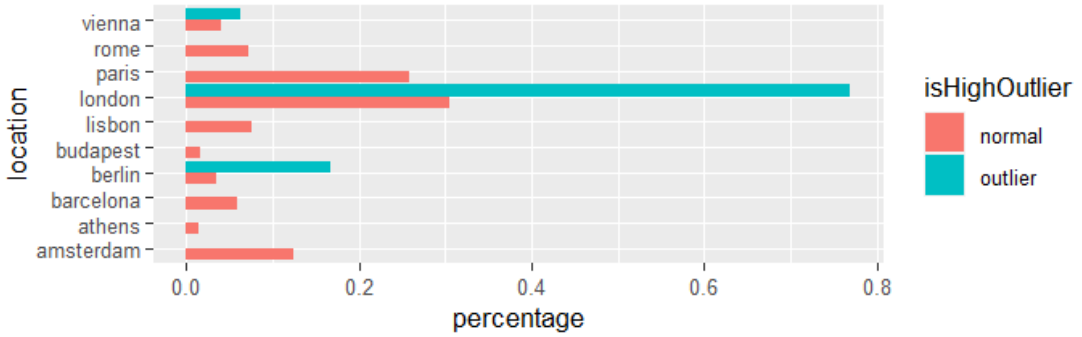


Figure 4.5. A Bar Chart of City Distribution of the Expensive Rooms. Long vs Short Distance from City Center and Metro Station. The rooms whose price is lower than 75% of the average price of all rooms are filtered out. The normal part is the short-distance group. The blue one is the long-distance group (i.e., outliers).

Transferring data into percentages is because the number of rooms in the outlier group is much smaller than that for the regular group.

A new question: Is the reason rooms in a bad location still command a high price in London because they can accommodate more people?

Question 6

Figure 4.6 illustrates the ratio of the room with different "person_capacity" for the outliers and the regular group, whose prices are high, and the location is in London. The rooms in the regular group are likely to have a more petite person capacity than those in the outlier group.

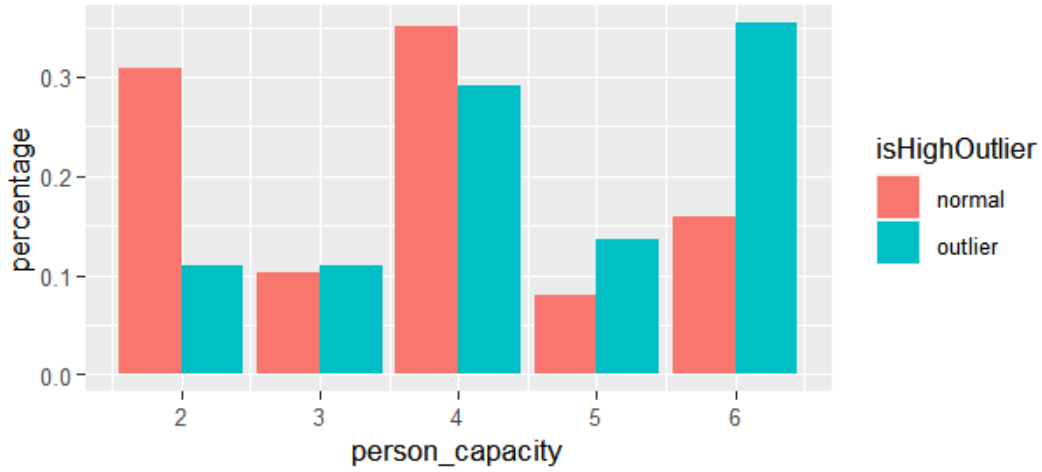


Figure 4.6. London rooms within high price range rooms. Long vs Short Distance from City Center and Metro Station.

It is better than a pie chart or ring chart because of the same reason in question 5. The pie chart encodes quantitative value in to colour hue. In contrast, figure 4.6 effectively use position for quantitative value and colour hue for nominal value.

The reason to use relative numbers is due to the significant difference between the size of the regular group and of the outlier group.

vi. Reflection

By doing this coursework, I gain experience in every stage of information visualisation. I gathered data from Kaggle, proposed initial questions, transformed the data using R and library, chose the chart and visual encoding based fundamental information visualisation knowledge, and improved those decisions after I saw the view.

When trying to find a dataset, I learned it is not so hard to find one on Kaggle. In the future, I will be more confident in finding a data set for this kind of coursework or just for fun.

When transforming and visualising, it is essential to focus on the questions rather than make the chart as expressive and effective as possible. For example, figure 4.3 is less expressive than a scatter plot matrix, but it is better for the question than the latter chart.

Additionally, when I review my R code, it is not reusable enough. In a big project, this might not be productive. So, in the future, I will periodically refactor my R code into functions if possible.

Furthermore, it is important to review the code. Because when I review it, I typed a wrong variable name (figure 6.1) which make the figure 4.3 wrong. It is better to review the code before analysing to avoid drafting report for the wrong chart.

Finally, I tasted what it means by "User Task" (Zhou & Xu, 2023), which is to redo the visulisation stages for many times after I saw the chart.

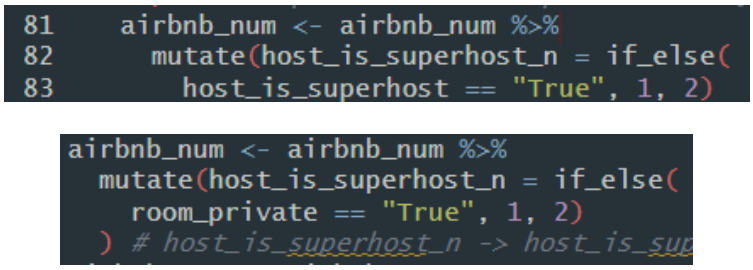


Figure 6.1. Top: incorrect code; Bottom: correct code.

Bibliography

Gleicher, M. (2018). Considerations for Visualizing Comparison. *IEEE Transactions on Visualization and Computer Graphics*, *24*(1), 413–423. https://doi.org/10.1109/TVCG.2017.2744199

Gyódi, K., & Nawaro, Ł. (2021). *Determinants of Airbnb prices in European cities: A spatial econometrics approach (Supplementary Material) [Data set]*. Zenodo. https://doi.org/10.5281/zenodo.4446043

Holtz, Y. (n.d.). *Ggplot2 Piechart*. Retrieved 21 April 2023, from https://www.r-graph-gallery.com/piechart-ggplot2.html

Mackinlay, J. (1986). Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics, 5(2)*, 110–141. https://doi.org/10.1145/22949.22950

Oetting, J. (2022, August 11). *14 Best Types of Charts and Graphs for Data Visualization [+ Guide]*. https://blog.hubspot.com/marketing/types-of-graphs-for-data-visualization *R Tutorial*. (n.d.). Retrieved 18 April 2023, from https://www.w3schools.com/r/

Peter. (2017, May 23). *Answer to 'apply jittering to outliers data in a boxplot with ggplot2'*. Stack Overflow. https://stackoverflow.com/a/44144640

TheDevastator. (2021). *Airbnb Prices in European Cities*. https://www.kaggle.com/datasets/thedevastator/airbnb-prices-in-european-cities

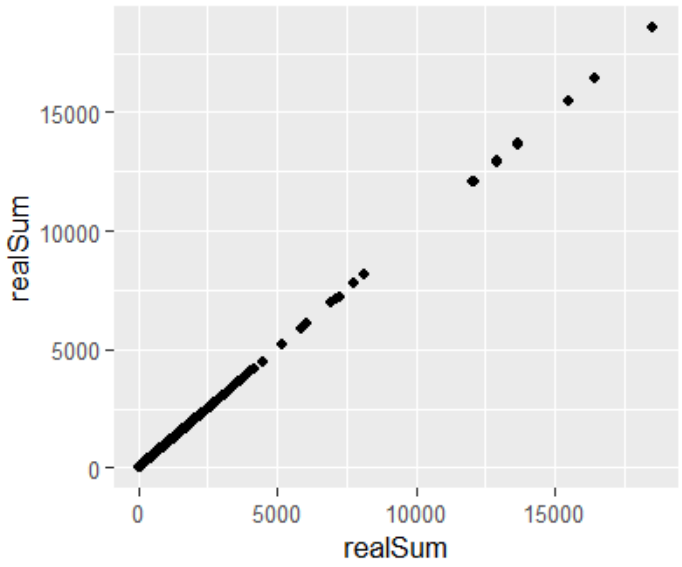
Zhou, K., & Xu, K. (2023). *COMP3021 (G53FIV): Fundamentals of Information Visualization*.

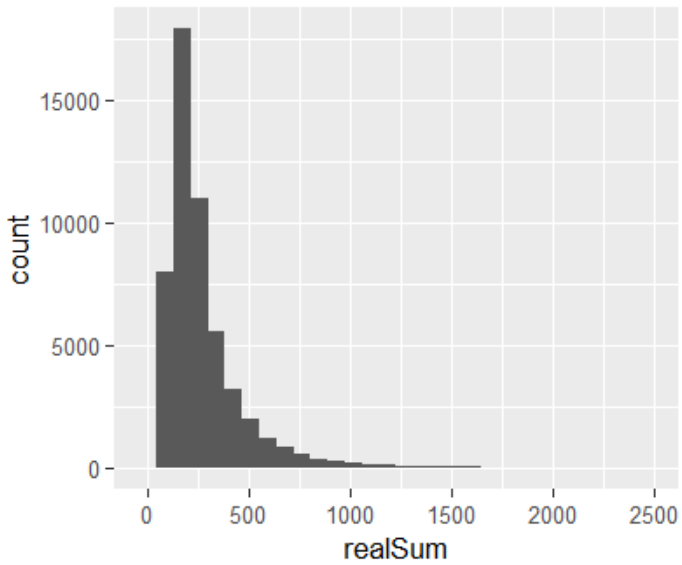
Chang, W. (n.d.). *R Graphics Cookbook, 2nd edition*. Retrieved 18 April 2023, from https://r-graphics.org/index.html

Appendix

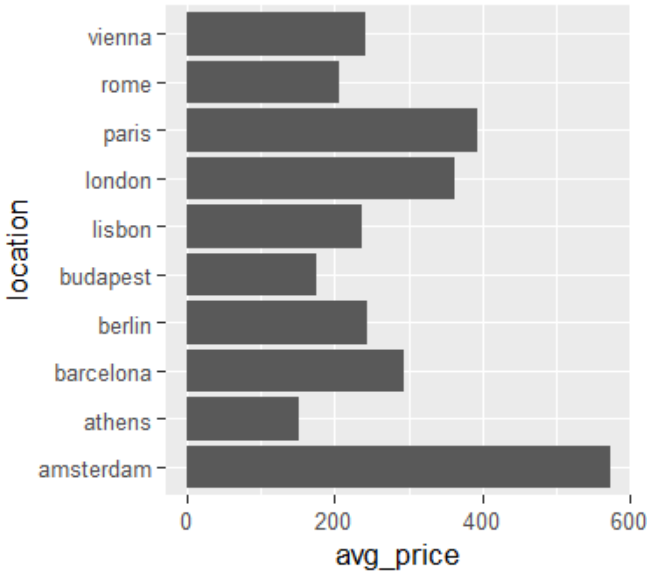
The following graphs were considered for the questions but were not chosen. Note that there is no candidate chart for question 6.

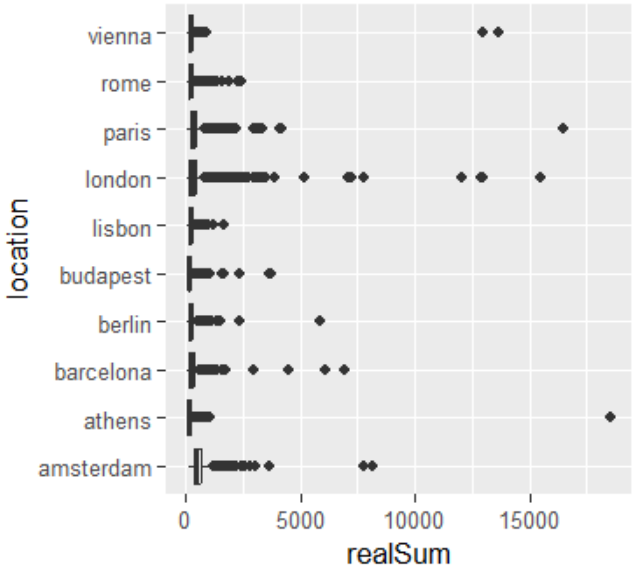
A.1 Candidate charts for question 1

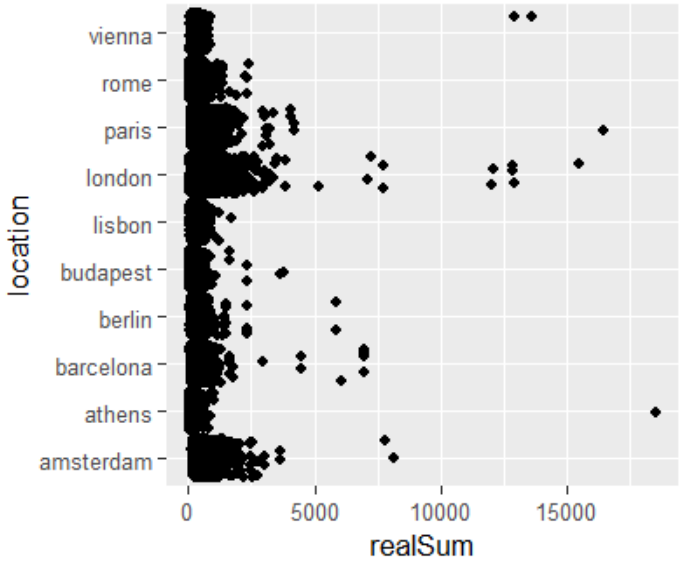




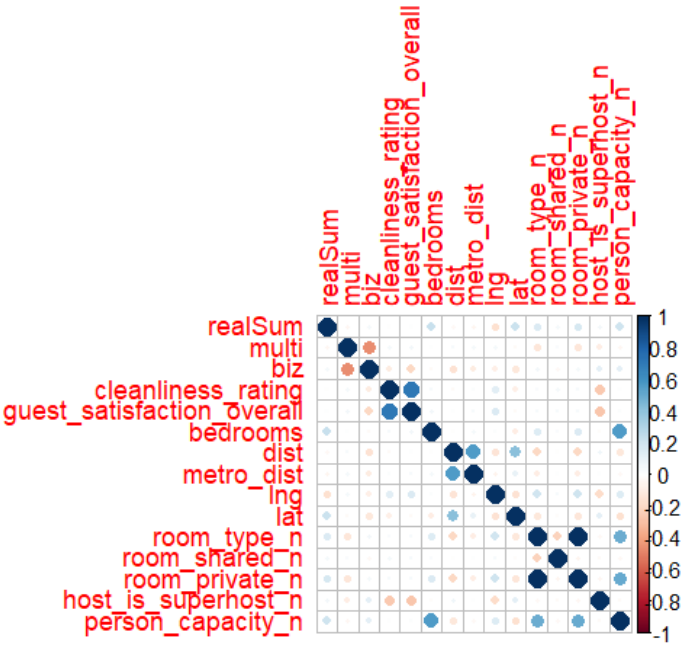
A.2 Candidate charts for question 2

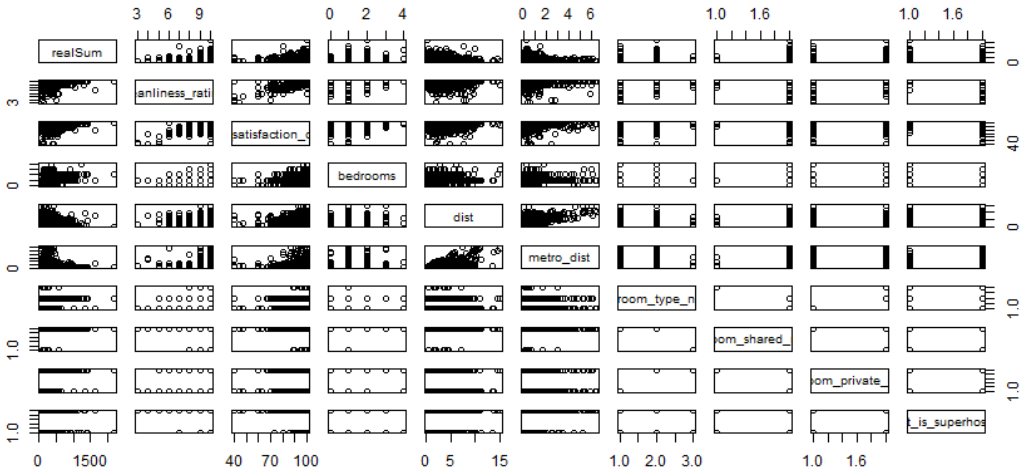




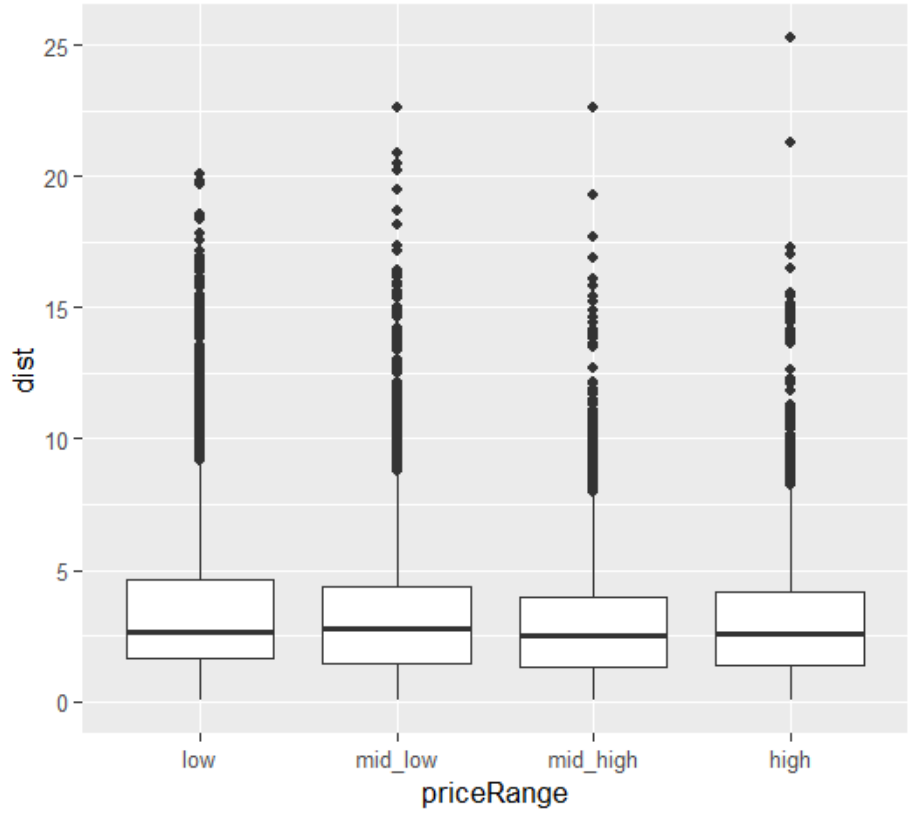


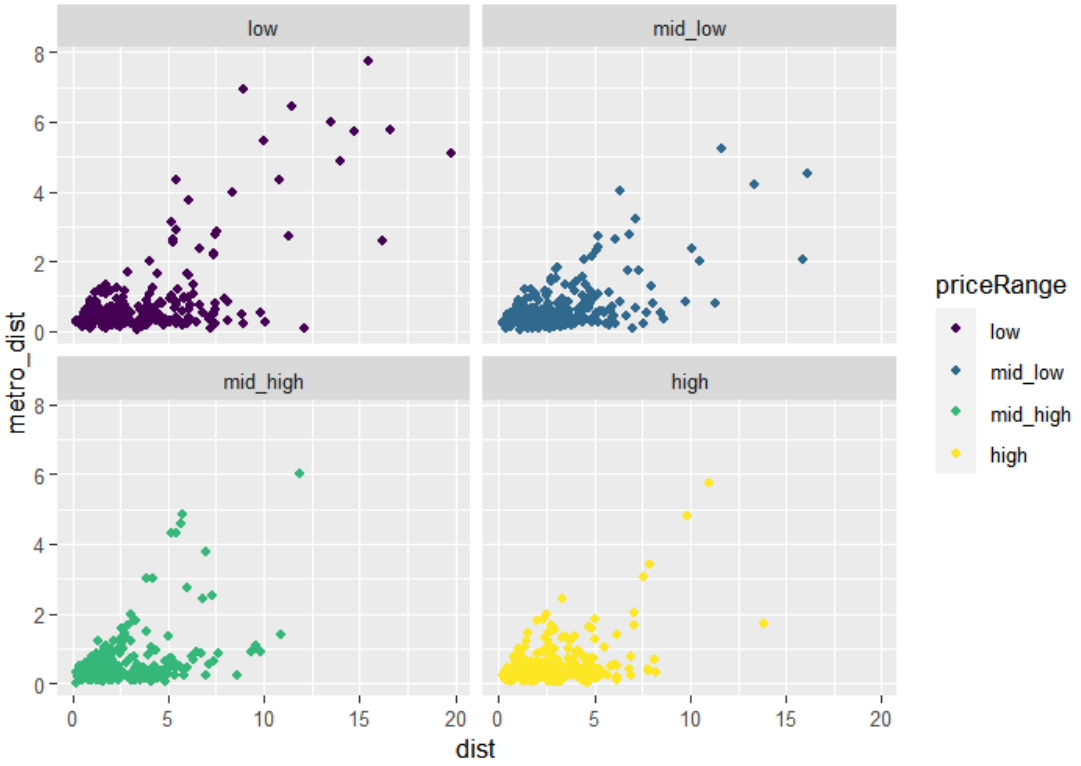
A.3 Candidate charts for question 3



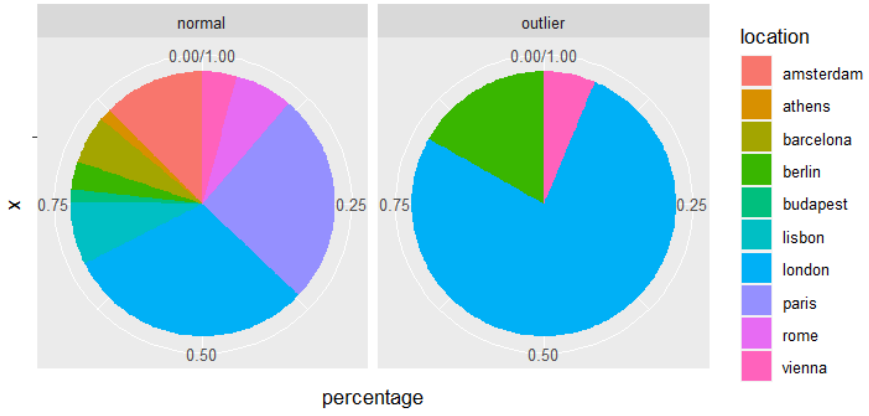


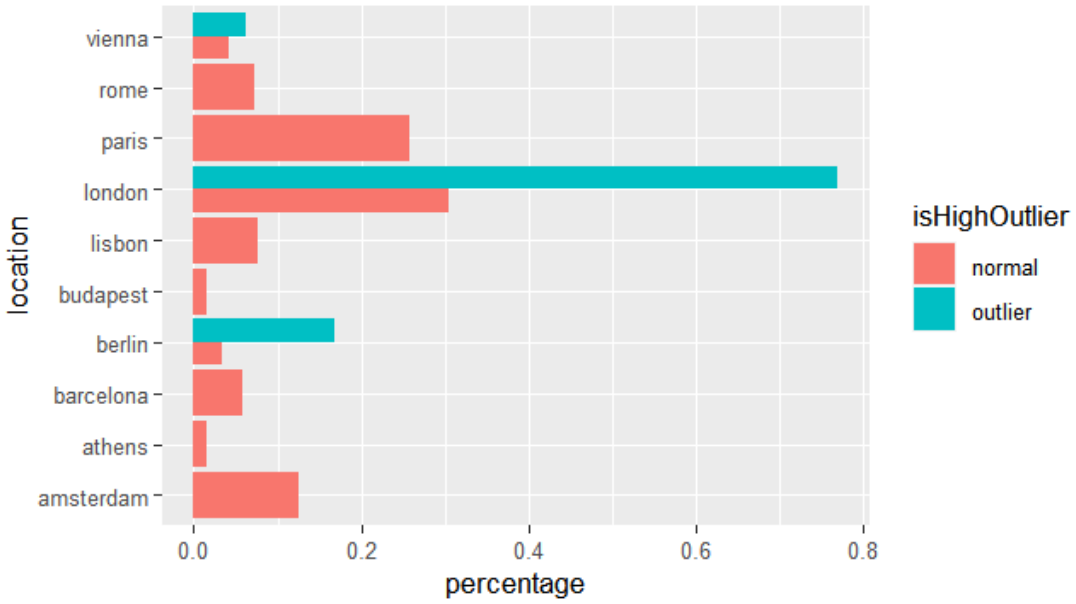
A.4 Candidate charts for question 4





A.5 Candidate charts for question 5





A.6 A candidate chart for question 6