Multivariate Statistical Analysis of Airbnb Listings in European Cities

Introduction, Design and Primary Questions

A self-guided journey to Western Europe is one of the top items on my wish list. An exciting and fun part of such a trip is the research and planning stage. As accommodations are often essential in shaping travel experiences, exploring Airbnb listings in different cities would be an ideal starting point. Airbnb listings, with their diverse characteristics, often mirror the unique aspects of each city. With this in mind, I decide to investigate Airbnb listing information in three European cities of interest: Amsterdam, Berlin, and Paris. Beyond enriching my own experience, I hope my exploration could also help other travelers make informed choices and improve their stays.

This project intends to offer guidance to various types of travelers and provide actionable insights to Airbnb hosts on enhancing guest experiences. Below are the four main questions I will analyze using four different multivariate techniques:

1. How can information about Airbnb listings be presented more efficiently, and which features significantly explain variation among listings? Principal Components Analysis will help me to reduce the dimensionality of the data and address this question.
2. Can Airbnb listings reflect cities? I will sample an equal, small number of listings from each city and perform Cluster Analysis to explore this.
3. What specific characteristics distinguish Airbnb listings across different cities? Discriminant Analysis will help me to identify the variables that differentiate the listings in Amsterdam, Berlin, and Paris.
4. The term 'Superhost' refers to highly rated and reliable hosts on Airbnb. Is Superhost status an important factor to consider when choosing listings? I will run a brief MANOVA to answer this question.

Data

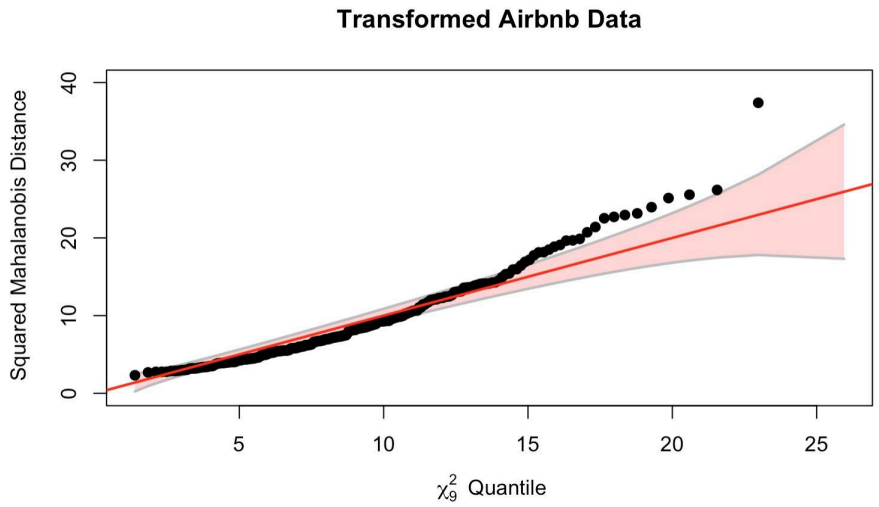
My data were sampled from a merged dataset available on Kaggle. The original dataset contains 41,714 Airbnb listings in 9 popular cities in Europe. My analysis focuses on the listings in Amsterdam, Berlin, and Paris. There are 11 variables in my dataset, including 6 continuous variables, 3 discrete variables and 2 categorical variables. They are grouped as follows:

- Continuous variables:
    - Price

    - Guest Satisfaction (A score ranging from 0 to 100)
    - City Center Distance (km)
    - Metro Distance (km)
    - Normalised Attraction Index
    - Normalised Restaurant Index
- Discrete variables:
    - Person Capacity
    - Bedrooms
    - Cleanliness Rating (Ten-point scale)
- Categorical variables:
    - City (Amsterdam, Berlin, Paris)
    - Superhost: binary (True / False)

I assessed univariate normality for each numeric variable using boxplots, normal quantile plots, and histograms. The results showed that *Price*, *City Center Distance* and *Metro Distance* were right-skewed, whereas *Guest satisfaction* was left-skewed. Given these observations, I took the log of *Price*, *City Center Distance* and *Metro Distance*, and transformed *Guest satisfaction* by subtracting 0.1 and then applying the logit function. With these transformations, the distributions of each variable were closer to normality. Then I made a chi-square quantile plot for the transformed data to check for multivariate normality.

Figure 1: Chi-Square Quantile Plot for Transformed Airbnb Data



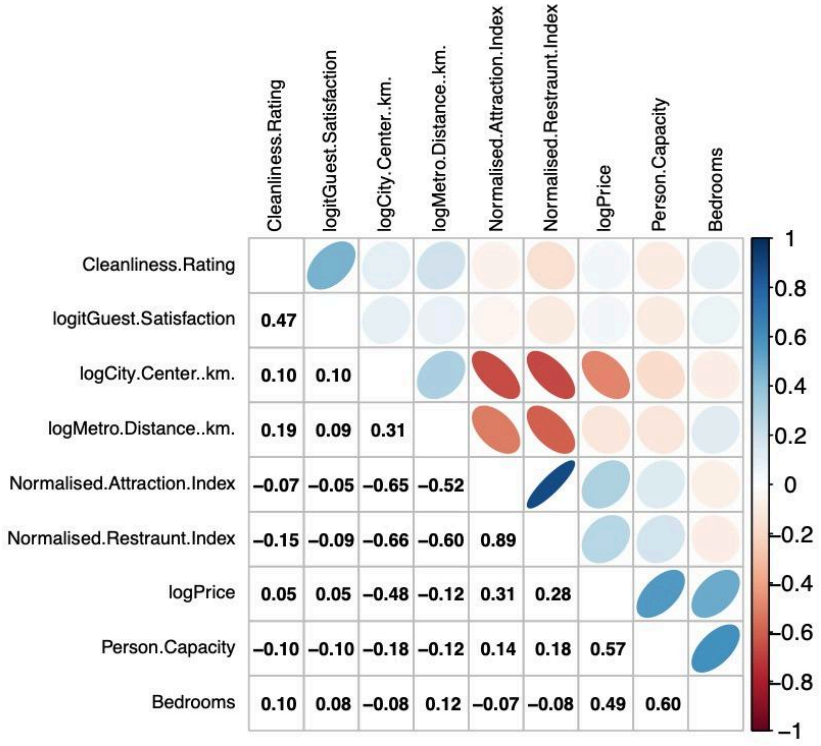
The chi-square quantile plot (Figure 1) indicates that there is no evidence of significant deviation from multivariate normality; though the distribution is not perfectly normal, it is close enough to what is expected.

Multivariate Analysis

Part 1: Principal Components Analysis

I will start my principal components analysis (PCA) by computing the correlation matrix. As is shown in Figure 2, there are high correlations between several pairs of variables, including a strong positive correlation between *Attraction Index* & *Restaurant Index*, *Price* & *Person Capacity*, and strong negative correlations between *City Center Distance* & *Attraction / Restaurant Index*. We also find that there is a moderate positive correlation between *Cleanliness Rating* & *Guest Satisfaction*, as well as a moderate negative correlation between *Metro Distance* & *Attraction Index*. These high correlations suggest that PCA should work well with this data.

Figure 2: Correlation matrix



Below are the results of Principal Components Analysis.

Table 1: Importance of Components

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     1.7890349 1.4127784 1.1968551 0.85377417 0.72127370
## Proportion of Variance 0.3556273 0.2217714 0.1591625 0.08099226 0.05780397
## Cumulative Proportion  0.3556273 0.5773987 0.7365612 0.81755346 0.87535743
##                           Comp.6     Comp.7     Comp.8     Comp.9
## Standard deviation     0.64243568 0.6006611 0.49962766 0.31406668
## Proportion of Variance 0.04585818 0.0400882 0.02773642 0.01095976
## Cumulative Proportion  0.92121561 0.9613038 0.98904024 1.00000000
```

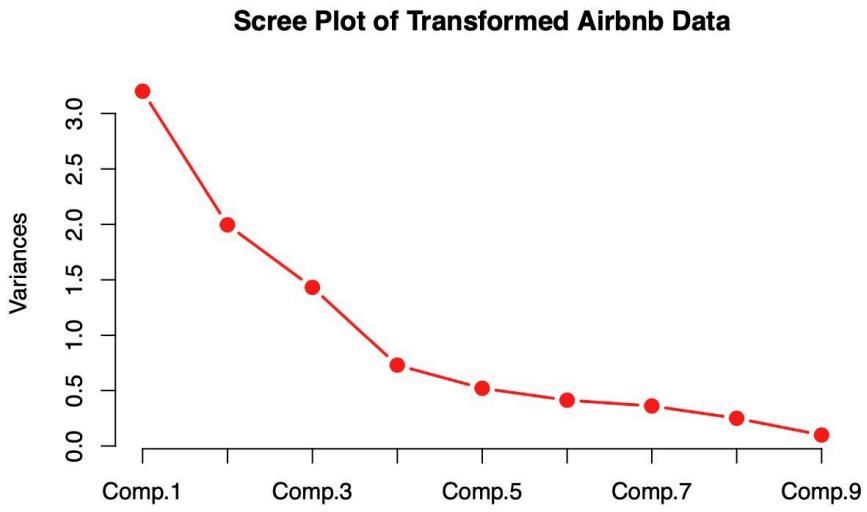
Table 1 shows the total variance explained by a given number of principal components. If we use the 80% threshold for the total variance explained method, it suggests retaining 4 principal components (4 explain 81.8% of the total variance).

Table 2: Eigenvalues

```
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
##   3.20   2.00   1.43   0.73   0.52   0.41   0.36   0.25   0.10
```

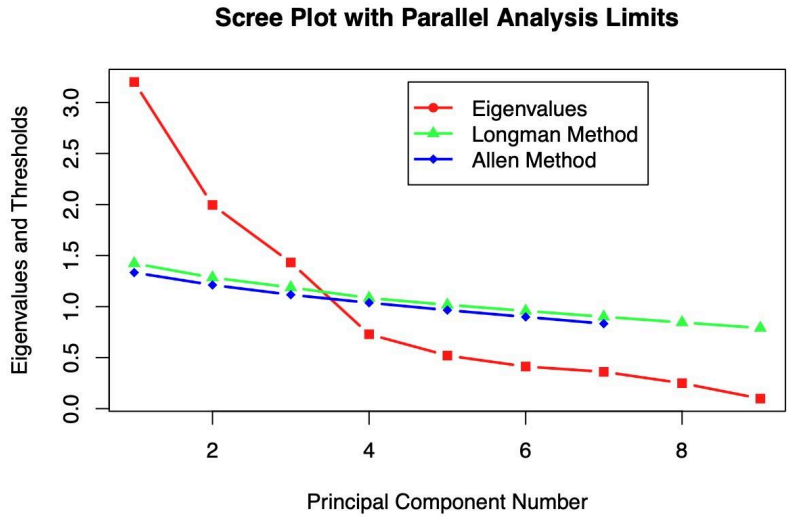
The eigenvalue > 1 criteria would argue for 3 components.

Figure 3: Scree plot



According to the Scree Plot (Figure 3), there is an elbow at 4, which would argue for retaining 3 components.

Figure 4: Scree Plot with Parallel Analysis Limits



From parallel analysis, the first three eigenvalues are larger than the corresponding Longman / Allen threshold. This suggests keeping 3 components.

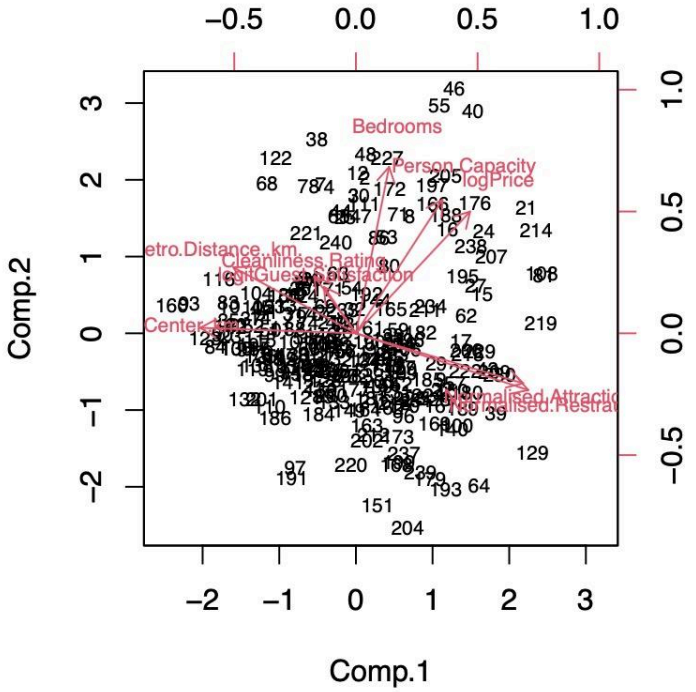
Based on a comprehensive analysis with all the methods above, I decide to retain 3 principal components. With 3 principal components, approximately 73.7% of the total variance in the data could be explained, a level that I find considerably acceptable.

Table 3: Loadings

```
## Loadings:
##                            Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## logPrice                    0.33   0.44   0.02   0.20   0.14   0.62   0.24
## Person.Capacity             0.25   0.48  -0.25  -0.30  -0.13  -0.19   0.51
## Cleanliness.Rating         -0.12   0.21   0.65   0.08  -0.70   0.11  -0.01
## logitGuest.Satisfaction    -0.09   0.17   0.67  -0.28   0.64  -0.12   0.08
## Bedrooms                    0.10   0.60  -0.11  -0.10  -0.03  -0.34  -0.66
## logCity.Center..km.        -0.45   0.02  -0.09  -0.52  -0.18  -0.13   0.33
## logMetro.Distance..km.     -0.35   0.24  -0.02   0.71   0.11  -0.44   0.32
## Normalised.Attraction.Index  0.48  -0.18   0.20   0.07  -0.10  -0.36   0.14
## Normalised.Restraunt.Index   0.49  -0.21   0.12  -0.03  -0.06  -0.32   0.10
##                            Comp.8 Comp.9
## logPrice                    0.44   0.08
## Person.Capacity            -0.48  -0.08
## Cleanliness.Rating         -0.11   0.04
## logitGuest.Satisfaction    -0.09  -0.01
## Bedrooms                    0.22   0.02
## logCity.Center..km.         0.60   0.08
## logMetro.Distance..km.      0.09   0.09
## Normalised.Attraction.Index  0.35  -0.64
## Normalised.Restraunt.Index   0.13   0.75
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## SS loadings      1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00   1.00
## Proportion Var   0.11   0.11   0.11   0.11   0.11   0.11   0.11   0.11   0.11
## Cumulative Var   0.11   0.22   0.33   0.44   0.56   0.67   0.78   0.89   1.00
```

Table 3 indicates that the dominant loadings for Component 1 are on *City Center Distance*, *Attraction Index* and *Restaurant Index*, suggesting that it is picking up ‘location and local amenities’. A high score on this component could indicate a listing that is closer to the city center, with better access to attractions and restaurants. Component 2 is mostly *Bedrooms*, *Person Capacity* and *Price*, capturing ‘accommodation capacity and pricing’. It highlights the positive relationship between the size and cost of accommodations. Component 3 is highly correlated with *Cleanliness Rating* and *Guest Satisfaction*, suggesting that listings with higher cleanliness ratings also tend to have higher guest satisfaction scores; it effectively reflects 'customer experience'.

Figure 5: Principal Components Biplot (Score Plot) for the first two components



In the biplot above, although most observations are clustered around the center, we could still observe some trends from the vectors pointing to different directions. For instance, the vectors for *Price* and *Capacity* are pointing in the upper right direction, indicating that Airbnb listings clustered in the upper right corner have higher price and capacity; these listings would be more appropriate for families instead of solo backpackers. Another example is the vector for *City Center Distance* and the vectors for *Attraction* & *Restaurant index*, which point to the left and right respectively, indicating that Airbnb listings clustered in the right side are closer to city center and have higher attraction and restaurant index; these listings are likely to be more popular among tourists.

From the analysis above, we could conclude that it is effective to use principal components on this data. Through PCA, the dimensionality of the data is reduced from 9 to 3. The three principal components are all meaningful and interpretable, explaining 73.7% of the total variance in the data. With dimensionality reduction, we could discover patterns more efficiently. For example, the biplot allows us to identify listings with specific features, which is helpful to tourists with different needs and preferences. In addition, the loadings and variance explained by each principal component also offers valuable insights. Notably, the fact that Comp.1 (the component capturing location and local amenities) explains the most variation in the data and the positive correlation between *Cleanliness rating* and *Guest satisfaction* are all critical findings for Airbnb hosts.

Part 2: Cluster Analysis

For this part, I sampled 30 Airbnb listings from each city. My goal is to determine the number of potential clusters within the sample and to see if the clusters we get can reflect cities. Based on the transformed data, I standardized the variables. Since some distance measures are scale variant, standardization could ensure that no single variable disproportionately influences the results.

Hierarchical Clustering

For hierarchical cluster analysis, I tested various distance metrics and agglomeration methods. Euclidean distance measures the straight-line distance between observations and gives equal weight to each of the variables, which is appropriate for the data as we want to consider Airbnb listing features equally. Thus, I primarily used Euclidean distance. For agglomeration methods, complete linkage tends to produce more distinctly separated groups as it defines the distance between clusters as the maximum distance between any observation in one cluster to any in another; while Ward's method joins observations together by minimizing internal sums of squares. Comparing the outputs, it turns out that the dendrogram using Ward's method (as shown in Figure 7) achieves a relatively better clustering structure.

Figure 6: Dendrogram with Euclidean distance and Complete Linkage

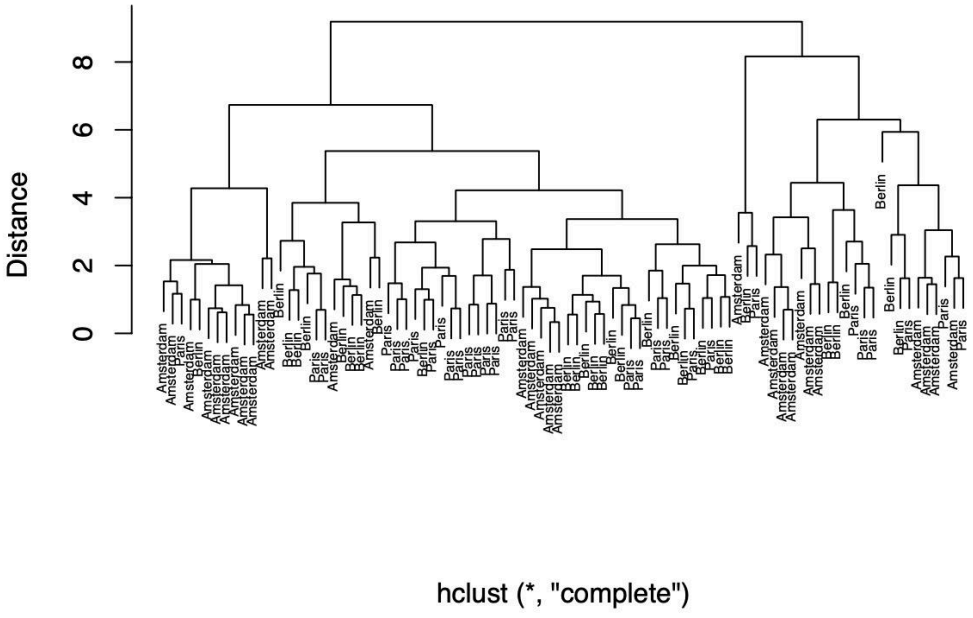
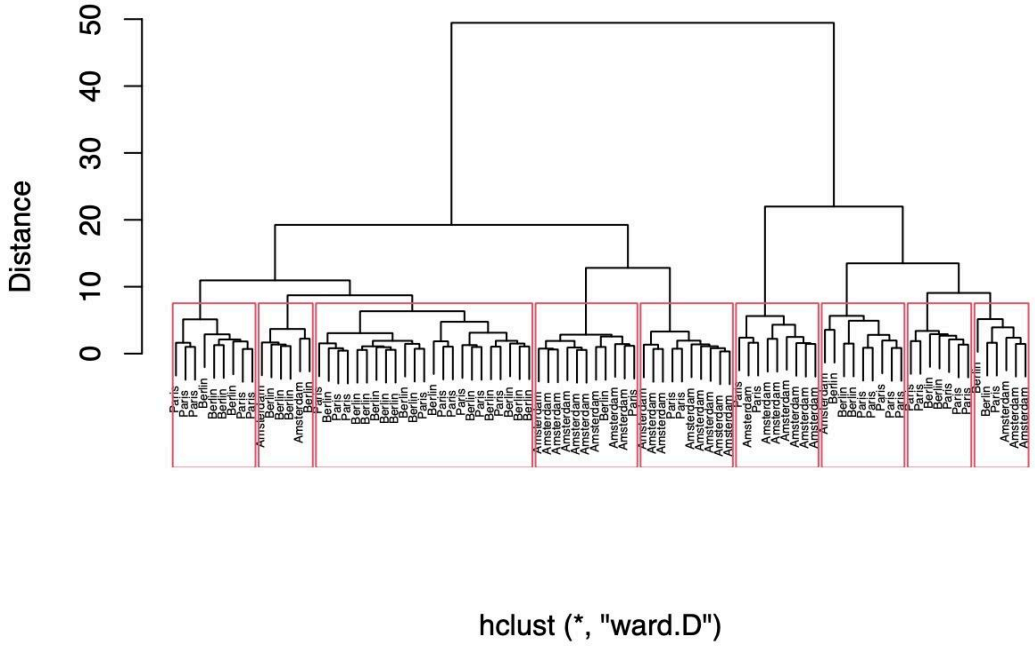


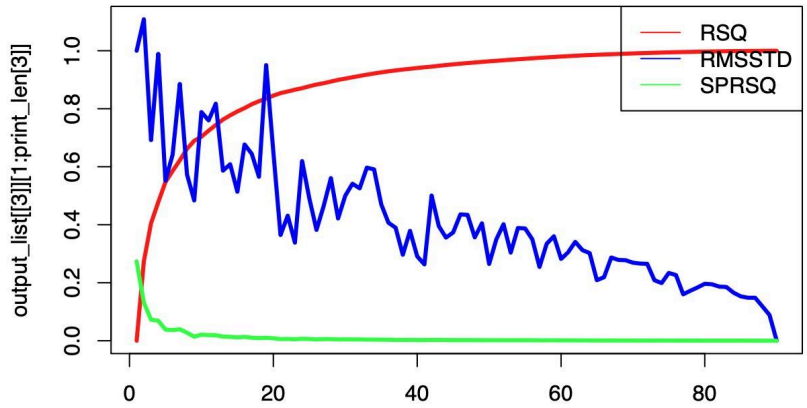
Figure 7: Dendrogram with Euclidean distance and Ward's method



To evaluate how many clusters are present in the dendrogram (Figure 7), I examined Root-mean-square standard deviation (RMSSTD), Semi-Partial R-Squared (SPRSQ), and R-squared (RSQ). Figure 8 indicates that RMSSTD has local minima at 3, 5 or 6, and 9; SPRSQ has elbows at 3, 5 and 9 clusters; R-squared has change points at 5, 9 and 15. So I would say 3 or 5 or 9 groups. It is hard to determine the exact number of clusters in the data; there may be three main groups with the possibility of further division into subgroups within these main ones.

I added 9 red rectangles in the dendrogram above just for easier interpretation.

Figure 8: RMSSTD, SPRSQ, RSQ



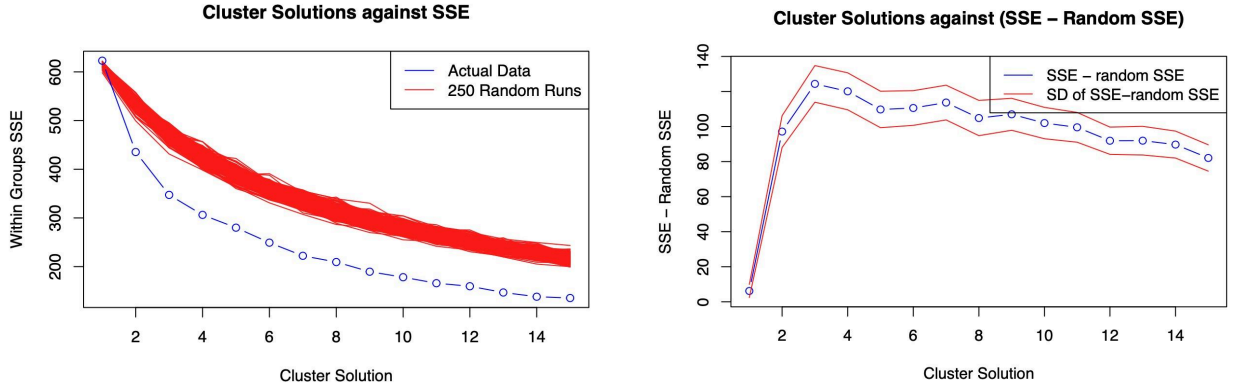
Looking into the dendrogram (Figure 7), we notice that while it doesn't perfectly segregate the Airbnb listings from Amsterdam, Berlin, and Paris into three distinct clusters, there is still a tendency that listings from the same city are grouped together. One observation is that almost all small groups at the bottom of the dendrogram are formed by listings from the same city - listings clustered at the initial stage share the most similarity. It also shows that most listings in the middle three red rectangles are from Amsterdam, and most listings in the left three red rectangles are from Berlin. Listings from Paris are more scattered, it might be a bit harder to tell some listings from Berlin and Paris apart.

K-Means Clustering

For comparison, I tested k-means clustering and plotted within cluster sum of squares vs. number of clusters (k) to further explore the number of clusters presented in the data. Results are shown below in Figure 9.

Figure 9: K-Means Clustering Result

```
## [1] "Airbnb Listings in Cluster  1"
##  [1] Amsterdam Amsterdam Amsterdam Amsterdam Amsterdam Amsterdam Amsterdam
##  [8] Berlin    Berlin    Berlin    Berlin    Berlin    Berlin    Paris
## [15] Paris     Paris     Paris     Paris     Paris     Paris     Paris
## [22] Paris     Paris     Paris     Paris
## Levels: Amsterdam Berlin Paris
## [1] " "
## [1] "Airbnb Listings in Cluster  2"
##  [1] Amsterdam Amsterdam Berlin    Berlin    Berlin    Berlin    Berlin
##  [8] Berlin    Berlin    Berlin    Berlin    Berlin    Berlin    Berlin
## [15] Berlin    Berlin    Berlin    Berlin    Berlin    Berlin    Berlin
## [22] Berlin    Berlin    Berlin    Paris     Paris     Paris     Paris
## [29] Paris     Paris     Paris     Paris     Paris     Paris     Paris
## [36] Paris     Paris     Paris     Paris     Paris     Paris
## Levels: Amsterdam Berlin Paris
## [1] " "
## [1] "Airbnb Listings in Cluster  3"
##  [1] Amsterdam Amsterdam Amsterdam Amsterdam Amsterdam Amsterdam Amsterdam
##  [8] Amsterdam Amsterdam Amsterdam Amsterdam Amsterdam Amsterdam Amsterdam
## [15] Amsterdam Amsterdam Amsterdam Amsterdam Amsterdam Amsterdam Amsterdam
## [22] Berlin    Berlin    Paris
## Levels: Amsterdam Berlin Paris
## [1] " "
```



It is shown that within groups SSE has an elbow at 3, SSE - Random SSE also has the flat spot at 3. Compared to the results (3 or 5 or 9 clusters) from hierarchical clustering, k-means clustering gives a more explicit outcome of 3 clusters.

K-means clustering output indicates that Airbnb listings in Cluster 3 are mostly from Amsterdam, Cluster 1 and 2 appears to be a mix of listings from Berlin and Paris. This result is sharing some similarity with the dendrogram in hierarchical clustering.

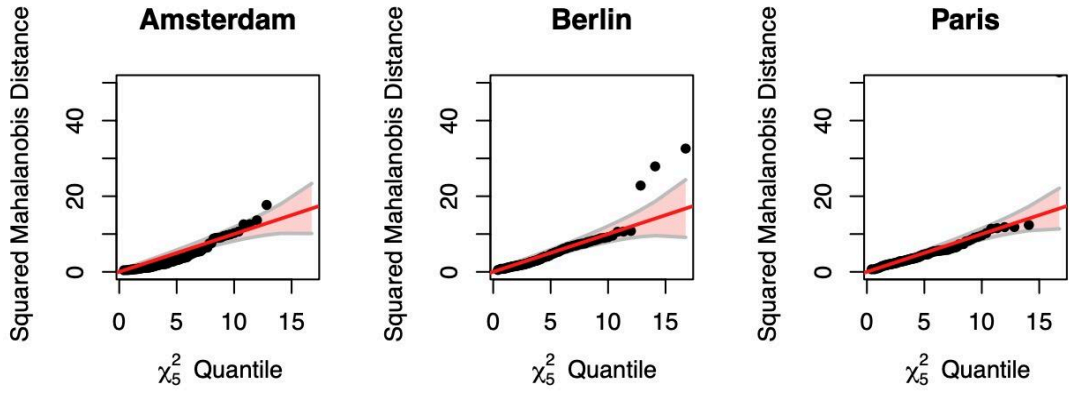
The final clusters from hierarchical clustering and k-means clustering are similar. These clusters do not perfectly reflect cities, but there is still a tendency that listings from the same city are grouped together. For example, in k-means clustering output, Cluster 3 is mostly from Amsterdam; it is also shown in the dendrogram that small groups at the bottom are mostly formed by listings from the same city. Another finding from the dendrogram in hierarchical cluster analysis is that further division into subgroups within the three main clusters is possible, implying the presence of additional layers of similarity - these could relate to specific neighborhood characteristics or property types within the listings. Despite these detailed distinctions, the three clusters capture a broad level of similarity, suggesting that diverse characteristics distinguish Airbnb listings across different cities.

This lays the groundwork for further exploration into “what variables distinguish Airbnb listings across different cities” through discriminant analysis.

Part 3: Discriminant Analysis

In this part, I evaluated the multivariate normality within each city group and the similarity of covariance matrices for the transformed data to ensure that the assumptions of discriminant analysis (DA) are satisfied.

Figure 10: Chi-Square Quantile Plots



The chi-square quantile plots (Figure 10) shows that the multivariate normality assumption is reasonably met within each city group.

Table 4: Box's M test

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  airbnb2_trans[, 2:6]
## Chi-Sq (approx.) = 147.61, df = 30, p-value < 2.2e-16
```

Table 5: Log Determinants of Covariance Matrices

```
# log determinants
log(det(cov(airbnb2_trans[airbnb2_trans$City=="Amsterdam", 2:6])))
```

```
## [1] 5.763852
```

```
log(det(cov(airbnb2_trans[airbnb2_trans$City=="Berlin", 2:6])))
```

```
## [1] 4.367991
```

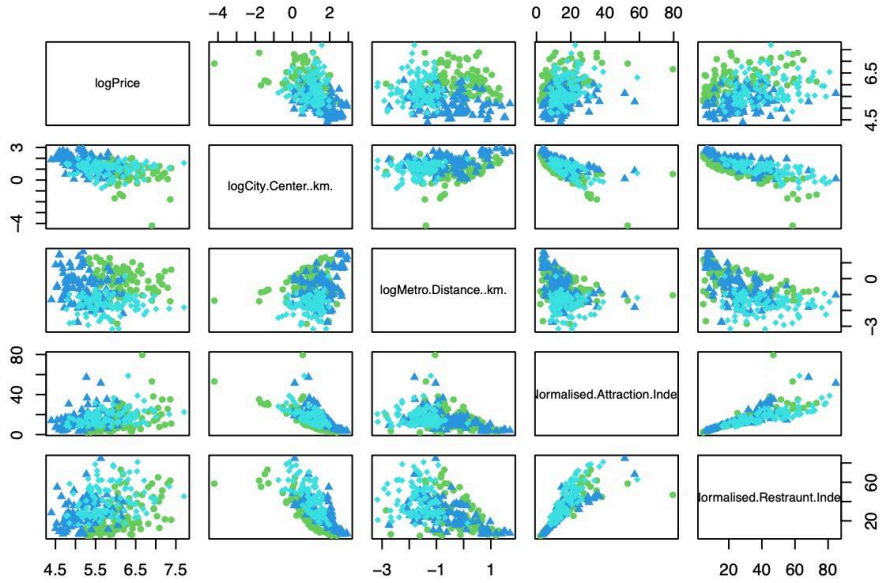
```
log(det(cov(airbnb2_trans[airbnb2_trans$City=="Paris", 2:6])))
```

```
## [1] 4.461481
```

To check the similarity of covariance matrices, I conducted Box's M test, computed log determinants and made a matrix plot. Although the Box's M test rejects the null hypothesis, the log determinants of the covariance matrices (5.76, 4.37, and 4.46) are close to each other,

suggesting that the covariance matrices are similar. Besides, the matrix plot (Figure 11) also shows that the multivariate footprints are relatively similar between groups. Therefore, it is appropriate to conclude that the assumptions are satisfied for the transformed data.

Figure 11: Matrix plot



Since the assumptions – multivariate normality within each group and similarity of covariance matrices – are met, then linear discriminant analysis (LDA) would be more appropriate for my data. I will first use stepwise discriminant analysis to select out more effective predictors, and then perform LDA.

Stepwise Discriminant Analysis

Table 6: Stepwise DA output

```
##  'stepwise classification', using 300-fold cross-validated correctness rate of method lda'.

## 300 observations of 5 variables in 3 classes; direction: both

## stop criterion: improvement less than 5%.

## correctness rate: 0.57667;  in: "logMetro.Distance..km.";  variables (1): logMetro.Distance..km.
## correctness rate: 0.73333;  in: "logPrice";  variables (2): logMetro.Distance..km., logPrice
##
##  hr.elapsed min.elapsed sec.elapsed
##         0.0         0.0         6.3

## method      : lda
## final model : City ~ logPrice + logMetro.Distance..km.
## <environment: 0x7fce163c4198>
##
## correctness rate = 0.7333
```

The stepwise DA output shows that *logMetroDistance* and *logPrice* are included in this process, indicating that the model with these two predictors is the best.

Linear Discriminant Analysis

I proceed with linear discriminant analysis using these two variables, *logMetroDistance* and *logPrice*. Table 7 shows the summary output of LDA.

Table 7: LDA summary output

```
## Call:
## lda(airbnb2_trans[, c(2, 4)], grouping = airbnb2_trans$City)
##
## Prior probabilities of groups:
## Amsterdam    Berlin     Paris
## 0.3333333 0.3333333 0.3333333
##
## Group means:
##           logPrice logMetro.Distance..km.
## Amsterdam 6.152743             -0.1270319
## Berlin    5.214531             -0.6195830
## Paris     5.805893             -1.6319912
##
## Coefficients of linear discriminants:
##                               LD1        LD2
## logPrice               -1.4152249  1.3704916
## logMetro.Distance..km. -0.9774272 -0.7841492
##
## Proportion of trace:
##    LD1    LD2
## 0.6481 0.3519
```

As shown in Table 8.1 and 8.2, the regular classification result reveals an accuracy rate of 74%, the classification accuracy with cross validation is 73%. This indicates that the model has a moderately high level of discriminating proficiency.

Table 8.1 Classification – Raw Results

```
(ctraw <- table(airbnb2_trans$City, predict(airbnb.disc)$class))
```

```
##
##             Amsterdam Berlin Paris
##   Amsterdam        83      8     9
##   Berlin            8     66    26
##   Paris             8     18    74
```

```
# total percent correct
round(sum(diag(prop.table(ctraw))), 2)
```

```
## [1] 0.74
```

Table 8.2 Classification – Cross Validated Results

```
airbnb.discCV <- lda(airbnb2_trans[, c(2,4)], grouping = airbnb2_trans$City, CV = TRUE)
(ctCV <- table(airbnb2_trans$City, airbnb.discCV$class))
```

```
##
##             Amsterdam Berlin Paris
##   Amsterdam        83      8     9
##   Berlin            9     65    26
##   Paris             9     19    72
```

```
# total percent correct
round(sum(diag(prop.table(ctCV))),2)
```

```
## [1] 0.73
```

Table 9: Multivariate Wilk's Lambda test

```
##                     Df   Wilks approx F num Df den Df    Pr(>F)
## airbnb2_trans$City   2 0.38661   90.028      4    592 < 2.2e-16 ***
## Residuals          297
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

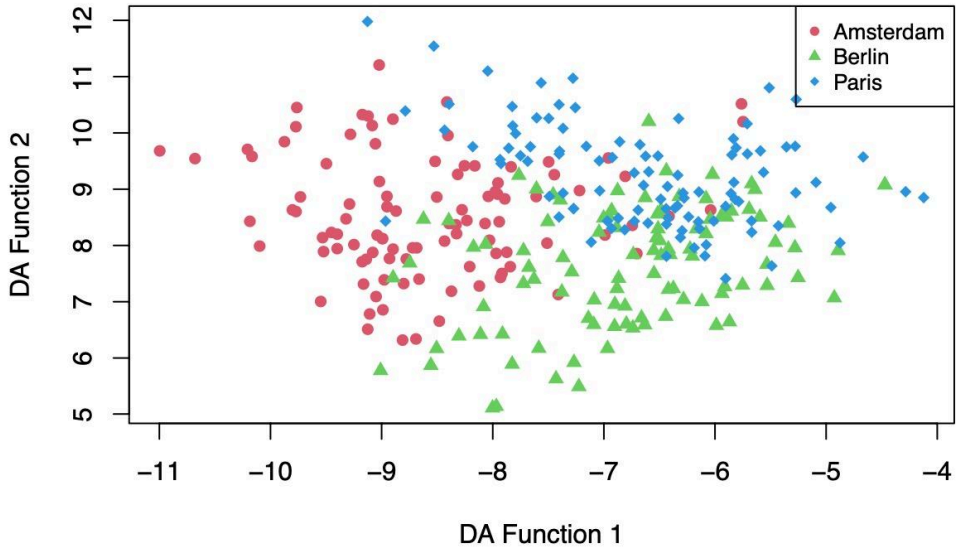
Table 9 shows the results of multivariate Wilk's Lambda test. The p-value in the test is close to 0, indicating there is statistically significant evidence that the multivariate group means are different.

Table 10: Standardized Discriminant Coefficients

```
##                              LD1   LD2
## logPrice                   -0.91  0.88
## logMetro.Distance..km. -1.00 -0.80
```

Table 10 presents the standardized discriminant coefficients. The magnitudes of standardized coefficients for the two predictors are very similar in both DA functions. Specifically, *logMetroDistance* has a slightly larger coefficient in the first DA function, while *logPrice* exhibits a greater coefficient in the second DA function. Given these observations, it is reasonable to conclude that both *logMetroDistance* and *logPrice* are key discriminators of great significance among city groups.

Figure 12: Score plot



The score plot (Figure 12) indicates that the three groups are relatively well separated by the two DA functions. Airbnb listings in Amsterdam are clearly separated from those in Paris in the direction of DA function 1, and listings in Berlin are differentiated from the other two cities in the direction of DA function 2.

Figure 13: Partition Plot

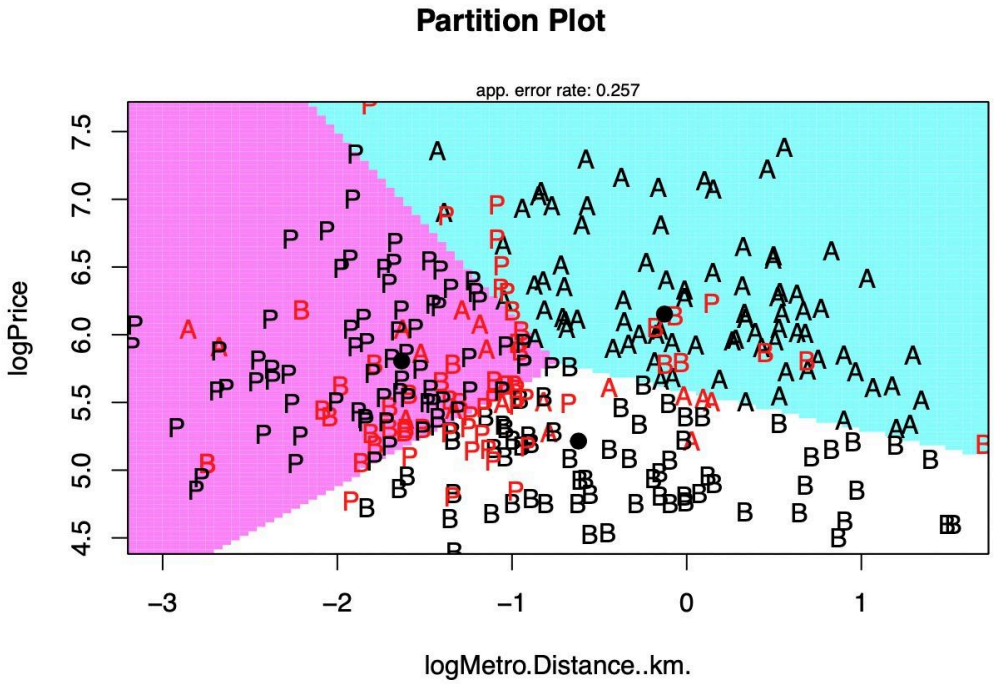


Figure 13 is the partition plot created using the two discriminating variables *logMetroDistance* and *logPrice*. The plot shows the regions assigned to each city group, with an approximate error rate of 25.7%.

From this plot, we can infer that Amsterdam's Airbnb market is characterized by higher prices, with listings typically located further from metro stations. This may indicate that listings in Amsterdam are more appropriate for travelers who value more luxurious accommodations or are traveling as part of larger groups where costs can be shared, as well as those who prefer to use alternative modes of transport such as car rentals. In contrast, Berlin's Airbnb listings appear to be more economically priced and exhibit a wider range in distance from metro stations. This cost-effectiveness could make Berlin a preferred destination for budget-conscious travelers. Airbnb listings in Paris tend to cluster closer to metro stations and span a variable range of prices. This proximity to metro stations implies enhanced accessibility and convenience, which would potentially improve traveler's urban experience in Paris through public transit networks.

Part 4: MANOVA

To further explore the relationships between variables, I will run a brief MANOVA.

In this part, I include the categorical variable *Superhost* and create a new variable *Comb* that combines different levels of *City* and *Superhost*. *Superhost* is binary, with two levels True and False, indicating whether the host is a superhost or not. As defined by Airbnb's official website, superhosts are those who have met a series of criteria in the previous year; generally, they are highly rated, experienced, reliable, and responsive hosts.

For a two-way MANOVA, I will use *City* and *Superhost* as my categorical predictors; and *Price*, *Guest Satisfaction*, *Cleanliness Rating*, *City Center Distance*, *Metro Distance*, *Attraction Index* and *Restaurant Index* as my continuous response variables.

Figure 14: Interaction Plots

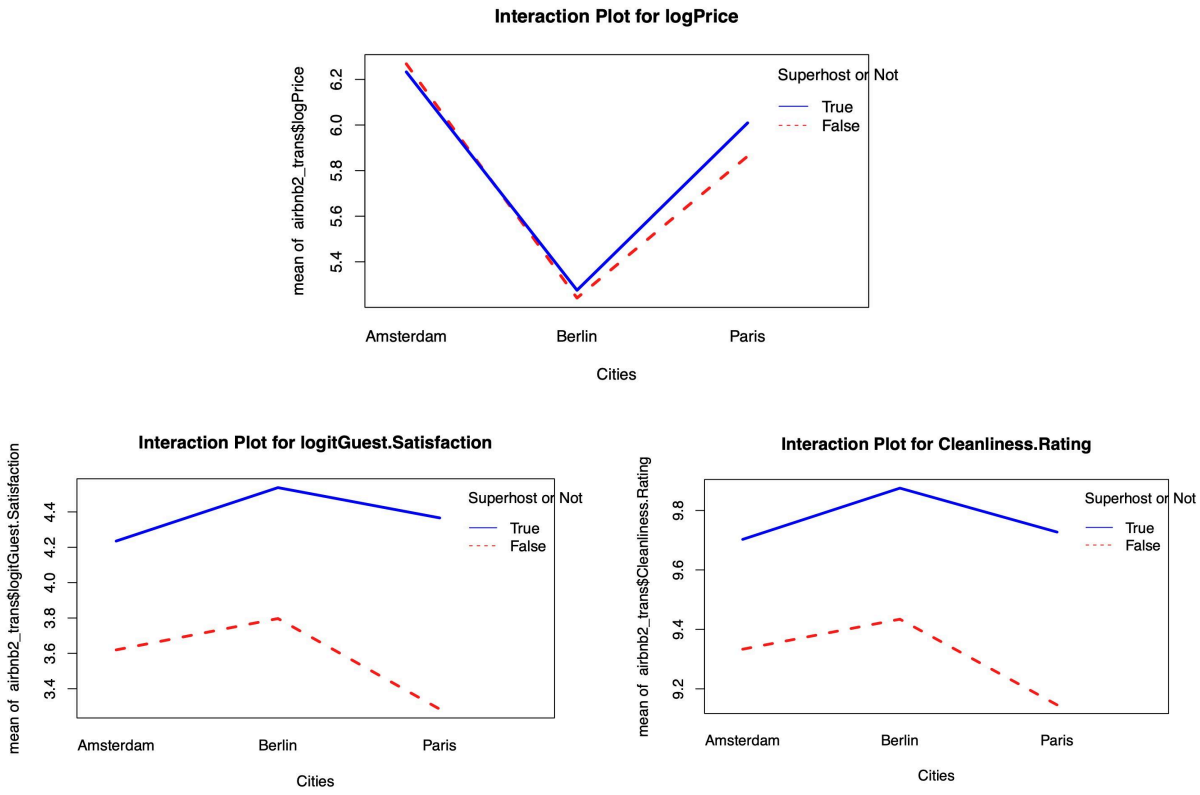


Figure 14 displays three interaction plots that I find interesting among all the plots created. *City* has a main effect on *Price*, the average price is clearly lower in Berlin than in Amsterdam and Paris; while *Superhost* does not appear to affect *Price*. However, superhost status has main effects on *Guest Satisfaction* and *Cleanliness Rating*, guest satisfaction score and cleanliness

rating are generally higher for superhosts. Besides, *Cleanliness Rating* and *Guest Satisfaction* appears to be higher on average in Berlin. None of the plots indicate an apparent interaction between *City* and *Superhost*. We will run MANOVA to further validate these relationships.

Univariate Results

Table 11: Univariate Results for two-way MANOVA by *City* and *Superhost* with interaction term

Term: (Intercept)

Sum of squares and products for the hypothesis:

	logPrice	logitGuest.Satisfaction	Cleanliness.Rating	logCity.Center..km.	logMetro.Distance..km.	Normalised.Attraction.Index	Normalised.Restraunt.Index
logPrice	6090.6212	4161.6452	9988.367	1140.2619	-850.1791	17038.662	34604.504
logitGuest.Satisfaction	4161.6452	2843.6001	6824.926	779.1267	-580.9167	11642.304	23644.824
Cleanliness.Rating	9988.3671	6824.9262	16380.509	1869.9824	-1394.2586	27942.701	56749.956
logCity.Center..km.	1140.2619	779.1267	1869.982	213.4753	-159.1671	3189.911	6478.518
logMetro.Distance..km.	-850.1791	-580.9167	-1394.259	-159.1671	118.6750	-2378.397	-4830.382
Normalised.Attraction.Index	17038.6618	11642.3043	27942.701	3189.9106	-2378.3968	47666.073	96806.946
Normalised.Restraunt.Index	34604.5038	23644.8242	56749.956	6478.5177	-4830.3817	96806.946	196609.122

Term: City

Sum of squares and products for the hypothesis:

	logPrice	logitGuest.Satisfaction	Cleanliness.Rating	logCity.Center..km.	logMetro.Distance..km.	Normalised.Attraction.Index	Normalised.Restraunt.Index
logPrice	40.755309	-10.528528	-6.089759	-34.860716	15.305513	22.111769	29.02273
logitGuest.Satisfaction	-10.528528	3.712229	2.268025	8.573111	2.678337	-27.542780	-90.53201
Cleanliness.Rating	-6.089759	2.268025	1.396457	4.906047	2.356872	-18.589508	-62.47648
logCity.Center..km.	-34.860716	8.573111	4.906047	30.007300	-15.983343	-9.396059	11.37600
logMetro.Distance..km.	15.305513	2.678337	2.356872	-15.983343	50.074723	-137.600010	-544.05977
Normalised.Attraction.Index	22.111769	-27.542780	-18.589508	-9.396059	-137.600010	492.247579	1842.42402
Normalised.Restraunt.Index	29.022725	-90.532008	-62.476483	11.376004	-544.059768	1842.424022	6968.60256

Term: Superhost

Sum of squares and products for the hypothesis:

	logPrice	logitGuest.Satisfaction	Cleanliness.Rating	logCity.Center..km.	logMetro.Distance..km.	Normalised.Attraction.Index	Normalised.Restraunt.Index
logPrice	0.10457890	1.762964	1.0064435	0.2672299	-0.07301922	-1.909041	-0.6772261
logitGuest.Satisfaction	1.76296449	29.719607	16.9663693	4.5048944	-1.23093944	-32.182124	-11.4165046
Cleanliness.Rating	1.00644354	16.966369	9.6857838	2.5717602	-0.70272036	-18.372174	-6.5174695
logCity.Center..km.	0.26722994	4.504894	2.5717602	0.6828513	-0.18658565	-4.878162	-1.7305124
logMetro.Distance..km.	-0.07301922	-1.230939	-0.7027204	-0.1865856	0.05098358	1.332933	0.4728537
Normalised.Attraction.Index	-1.90904078	-32.182124	-18.3721743	-4.8781624	1.33293302	34.848681	12.3624571
Normalised.Restraunt.Index	-0.67722605	-11.416505	-6.5174695	-1.7305124	0.47285369	12.362457	4.3855417

Term: City:Superhost

Sum of squares and products for the hypothesis:

	logPrice	logitGuest.Satisfaction	Cleanliness.Rating	logCity.Center..km.	logMetro.Distance..km.	Normalised.Attraction.Index	Normalised.Restraunt.Index
logPrice	0.2296819	0.5801599	0.2664989	0.2276312	0.4548379	-0.6324191	3.248453
logitGuest.Satisfaction	0.5801599	1.4984349	0.6790955	0.5126858	1.0638568	0.2515529	8.457530
Cleanliness.Rating	0.2664989	0.6790955	0.3102864	0.2529066	0.5124410	-0.4009789	3.814555
logCity.Center..km.	0.2276312	0.5126858	0.2529066	0.3432143	0.6113206	-4.1178202	2.743320
logMetro.Distance..km.	0.4548379	1.0638568	0.5124410	0.6113206	1.1198534	-6.0176138	5.782984
Normalised.Attraction.Index	-0.6324191	0.2515529	-0.4009789	-4.1178202	-6.0176138	105.3620158	5.187860
Normalised.Restraunt.Index	3.2484525	8.4575304	3.8145553	2.7433202	5.7829840	5.1878600	47.871165

F-tests

	logPrice	logitGuest.Satisfaction	Cleanliness.Rating	logCity.Center..km.	logMetro.Distance..km.	Normalised.Attraction.Index	Normalised.Restraunt.Index
(Intercept)	23457.15	392.64	17556.39	206.37	175.22	251.70	749.61
City	78.48	1.03	0.75	58.02	36.97	5.20	13.28
Superhost	0.40	4.10	10.38	0.66	0.08	0.18	0.02
City:Superhost	0.44	0.41	0.17	0.66	0.83	1.11	0.09

p-values

	logPrice	logitGuest.Satisfaction	Cleanliness.Rating	logCity.Center..km.	logMetro.Distance..km.	Normalised.Attraction.Index	Normalised.Restraunt.Index
(Intercept)	< 2.22e-16	< 2.22e-16	< 2.22e-16	< 2.22e-16	< 2.22e-16	< 2.22e-16	< 2.22e-16
City	< 2.22e-16	0.3121332	0.4740453	3.5798e-13	4.7783e-15	0.0233204	2.9964e-06
Superhost	0.5261554	0.0174682	0.0014161	0.5175491	0.7840004	0.8320156	0.8972022
City:Superhost	0.6429879	0.5205486	0.8468885	0.4159558	0.4385088	0.2923548	0.9128077

There are differences in univariate means for *Price*, *City Center Distance*, *Metro Distance*, *Attraction Index* and *Restaurant Index* between three cities. Based on whether or not the host is a superhost, there are significant differences in univariate means for *Guest Satisfaction* and *Cleanliness Rating*. The p-values for the interaction term are large in all univariate tests, indicating that there is no significant interaction effect between *City* and *Superhost*.

Multivariate Results

Table 12: Multivariate Results for two-way MANOVA by *City* and *Superhost* with interaction term

```
Multivariate Tests: (Intercept)
                 Df test stat approx F num Df den Df     Pr(>F)
Pillai            1   0.99404 6861.022      7    288 < 2.22e-16 ***
Wilks             1   0.00596 6861.022      7    288 < 2.22e-16 ***
Hotelling-Lawley  1 166.76096 6861.022      7    288 < 2.22e-16 ***
Roy               1 166.76096 6861.022      7    288 < 2.22e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Multivariate Tests: City
                 Df test stat approx F num Df den Df     Pr(>F)
Pillai            2 0.7966743 27.33364     14    578 < 2.22e-16 ***
Wilks             2 0.3232510 31.22147     14    576 < 2.22e-16 ***
Hotelling-Lawley  2 1.7225737 35.31276     14    574 < 2.22e-16 ***
Roy               2 1.4702347 60.69969      7    289 < 2.22e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Multivariate Tests: Superhost
                 Df test stat approx F num Df den Df   Pr(>F)
Pillai            1 0.0484761 2.096053      7    288 0.043975 *
Wilks             1 0.9515239 2.096053      7    288 0.043975 *
Hotelling-Lawley  1 0.0509457 2.096053      7    288 0.043975 *
Roy               1 0.0509457 2.096053      7    288 0.043975 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Multivariate Tests: City:Superhost
                 Df test stat  approx F num Df den Df  Pr(>F)
Pillai            2 0.0312201 0.6546923     14    578 0.81883
Wilks             2 0.9689231 0.6545927     14    576 0.81892
Hotelling-Lawley  2 0.0319258 0.6544781     14    574 0.81902
Roy               2 0.0263051 1.0860254      7    289 0.37217
```

Overall, there are differences in multivariate means across three cities (all multivariate statistics are significant). The multivariate means are also statistically significantly different based on whether or not the host is a superhost. None of the multivariate tests suggest there is a significant interaction effect between *City* and *Superhost*.

From two-way MANOVA results, we find that the differences in *Price*, *City Center Distance*, *Metro Distance*, *Attraction Index* and *Restaurant Index* among Airbnb listings can be explained by *Cities*. The interaction plot shows that the average price is lower in Berlin than in Amsterdam and Paris, which is consistent with the result of Discriminant Analysis. Another interesting finding from the univariate results is that *Guest Satisfaction* and *Cleanliness Rating* are significantly different based on whether or not the host is a superhost, while *Price* is not; this suggests that superhosts may have better service quality but not necessarily charge higher prices.

Discussion and Conclusion

The four multivariate techniques provide diverse perspectives for analyzing Airbnb listing information across three European cities.

Principal Components Analysis efficiently reduces the dimensionality of Airbnb listing data. The biplot created using the first two principal components enables easier identification of listings with specific features, thereby guiding tourists with different needs and preferences, such as families and solo backpackers. PCA also brings valuable insights to Airbnb hosts. The component accounting for the most variation in the data is ‘location and local amenities’, indicating that proximity to city center, tourist attractions and a high density of restaurants significantly contributes to a listing's popularity. Additionally, given the positive correlation between *Cleanliness Rating* and *Guest Satisfaction*, maintaining a clean environment can be a key strategy for hosts looking to make good impressions on guests.

Cluster Analysis indicates that Airbnb listings are reflecting cities to some extent. Although listings might not form perfect city clusters, both k-means and hierarchical clustering shows that listings from the same city tend to be grouped together. This lays a solid foundation for discriminant analysis.

Discriminant Analysis allows me to identify the variables that distinguish listings in the three cities. Stepwise DA highlights *Metro Distance* and *Price* as significant discriminators. The partition plot based on LDA with these two variables clearly portrays the characteristics of Airbnb listings in Amsterdam, Berlin, and Paris. Amsterdam's Airbnb market features higher prices and listings generally located farther from metro stations, suggesting that it is better suited for larger groups and those preferring alternative transport. Berlin offers more budget-friendly options with varied metro accessibility. Airbnbs in Paris are generally closer to metro stations, offering convenient access to the city via public transit.

Lastly, MANOVA further confirms the differences in multivariate means across the city groups. One interesting finding from MANOVA is the significant difference in *Guest Satisfaction* and *Cleanliness Rating* depending on whether the host is a *Superhost*. Given that superhosts may provide better service without necessarily charging higher prices, it would be sensible to filter Airbnb listings by Superhost status.