Multivariate analysis of Airbnb listings in Vienna, Austria

Elveren, Cihan Premoli, Tommaso Sangiovanni, Luca
Temu, Honest

Abstract

This project investigates the determinants of Airbnb prices in Vienna using advanced statistical techniques on a rich dataset of listings. The analysis begins with robust regression to handle outliers and identify key predictors, followed by bootstrapping to ensure result stability and accurate confidence intervals. Next, random mixed models capture neighborhood-specific effects, which increases prices in central districts like Innere Stadt but is weaker in suburban areas. Finally, clustering techniques group neighborhoods with similar price patterns, offering insights into spatial dynamics and market segmentation. This combined approach reveals structural and spatial heterogeneity in Vienna's Airbnb market.

1 Exploratory Data Analysis

The dataset that we chose comes entirely from the website *insideairbnb.com*, a website that contains Airbnb datasets for various cities around the world. We decided to analyze the Airbnbs of the city of Vienna, Austria, and performed some data cleaning on the dataset, which originally contained around 14,000 observations. The first thing that we did was dropping the columns that we didn't need for our objective; then we added 3 more columns by using the *OpenStreetMap* API, and we performed some transformations on some other columns. The columns that we ended up with are the following:

Column	Description
ID	Unique identifier of the Airbnb apartment
host_id	Unique identifier of the Airbnb host
price_dollars	Price per night, in dollars
latitude	Latitude of the Airbnb
longitude	Longitude of the Airbnb
dist_stephansdom_km	Distance, in km, from Stephansdom (the main cathedral)
dist_schonbrunn_km	Distance, in km, from the Schönbrunn palace
dist_train_station_km	Distance, in km, from the main train station
neighbourhood	Neighbourhood in which the Airbnb is located
room_type	Entire home/apartment or private room
accomodates	Number of people that can be accomodated
bathrooms	Number of bathrooms
cleaning_service	Presence of a cleaning service
air_conditioning	Presence of air conditioning
self_checkin	Possibility of doing a self check-in
host_acceptance_rate	Percentage of guests accepted by the host
host_listings_count	Number of Airbnbs owned by the host
number_of_reviews	Total number of reviews submitted
apt_age_days	Days since the Airbnb was made available on the website for the first time
review_scores_rating	Average rating score, up to 5
reviews_per_month	Number of reviews submitted every month

Table 1: Variables of our dataset

Then, we proceedeed at removing outliers, in order to make the assumptions for the models that we are going to create more robust. We noticed the presence of a high number of outliers, therefore we used more than one method to try to remove them: starting from a basic linear model, we removed the outliers based on the *Cook's distance*, the *absolute standardized residuals*, the *leverage*, the *studentized residuals* and the *z-scores*. The dataset that we ended with is the one we used for all our calculations, and it is made of 7931 observations and 21 columns.
Here we can see how the variables related to price and review scores are distributed:

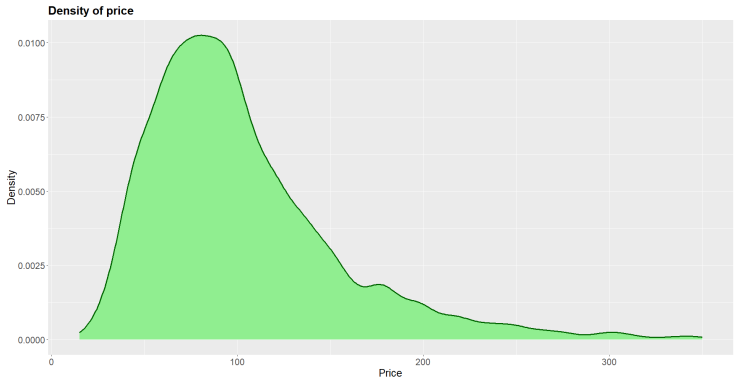


Figure 1: Density of price per night, in dollars

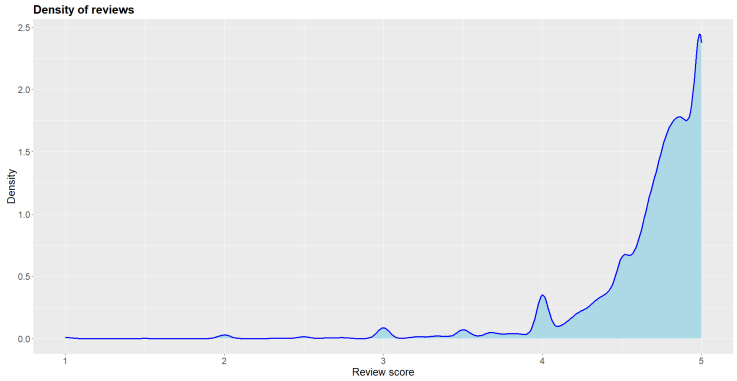


Figure 2: Density of reviews scores

We can also notice that older Airbnbs tend to have higher review scores, and if we look at the map divided by neighbourhood we can see that listings in the city center tend to be on the Airbnb website for more time compared to the ones outside the city center:

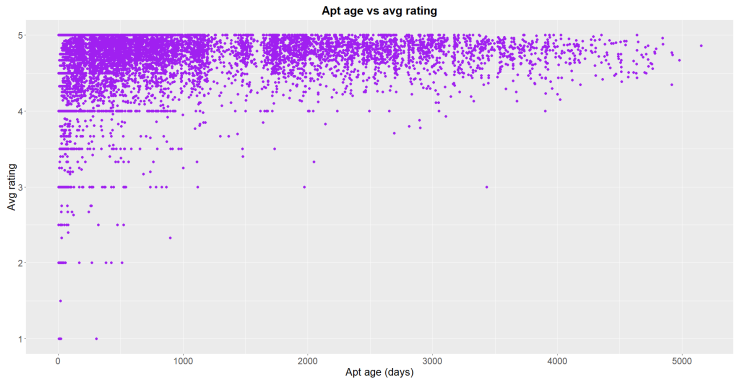


Figure 3: Apartment age vs. Review score

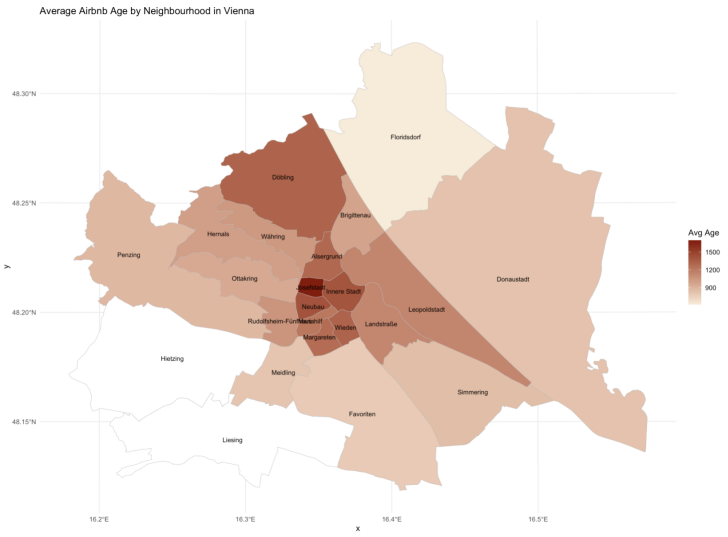


Figure 4: Map of average Airbnb age, by neighbourhood

By looking at the room type, we can clearly see that most of the Airbnb listings in Vienna are entire homes or apartments and not so many are private rooms, with no particular distinction among neighbourhoods:

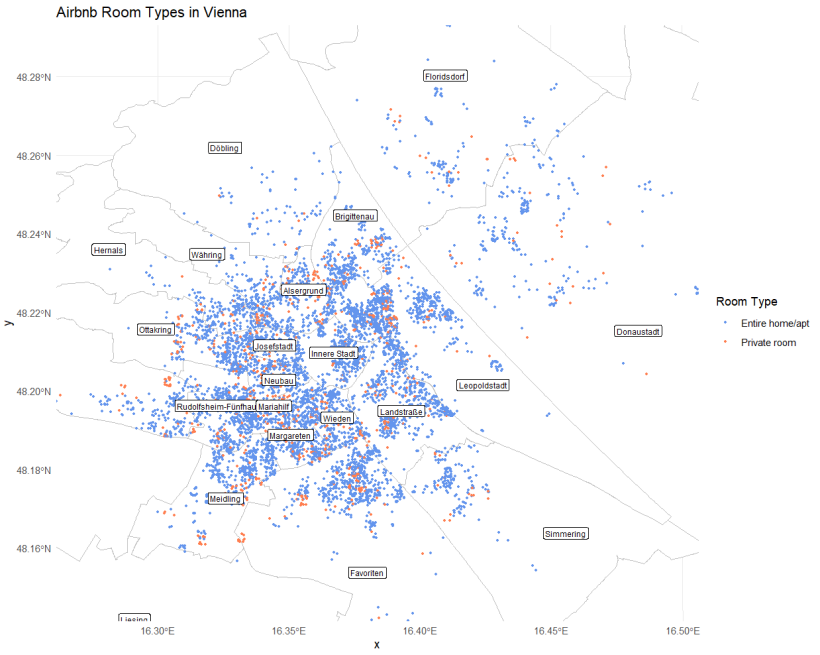


Figure 5: Airbnb listing, by room type

If we instead focus on the various neighbourhoods (there are 21 distinct neighbourhoods) we can analyze which ones have the most Airbnbs, and which have the lowest amount:

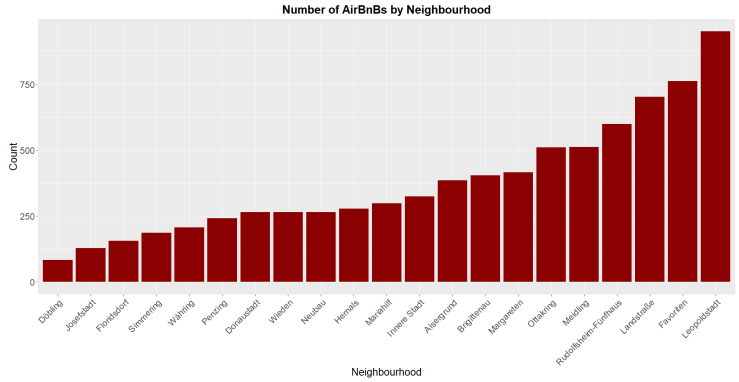


Figure 6: Number of Airbnbs in each neighbourhood

Or also the average price per night in each neighbourhood:

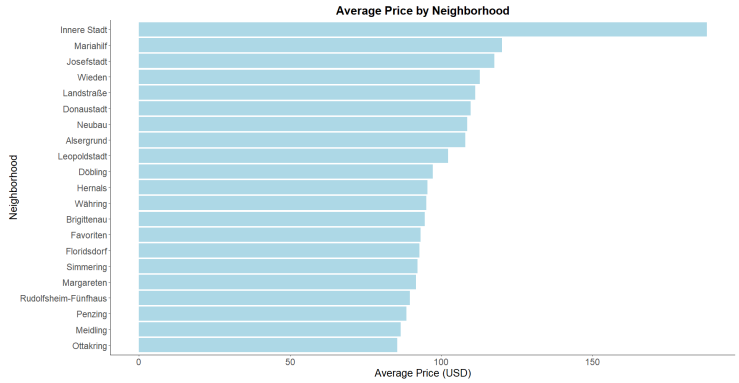


Figure 7: Average price by neighbourhood

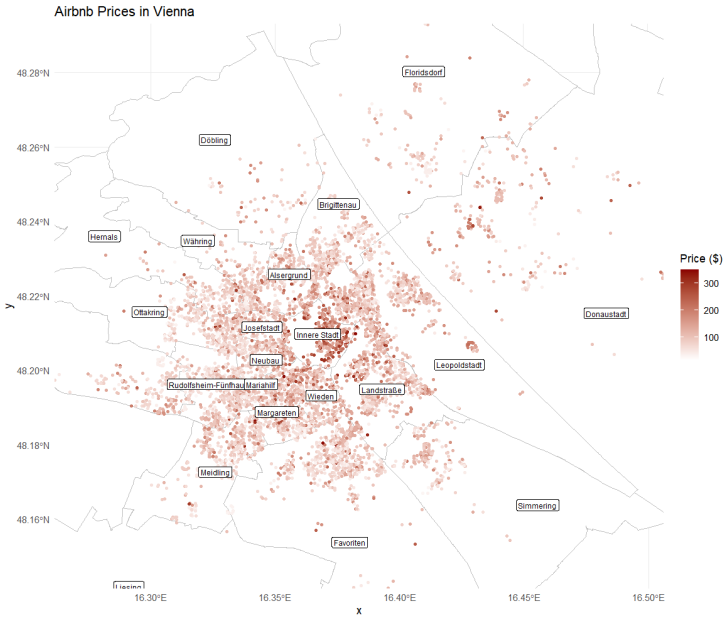
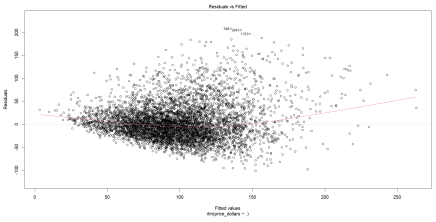


Figure 8: Map of average price by neighbourhood

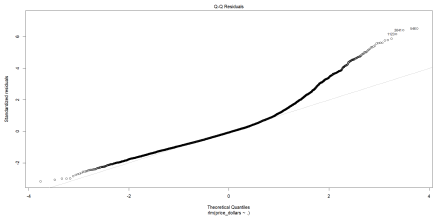
2 Robust Inference

One of the main variables that we can see in the dataset is the *price_dollars* variable, which expresses the cost, in dollars, of one night at a given Airbnb. This variable is quite probably the one that guests are more interested in: when choosing a place to stay in a new city, the price per night is what influences us the most; therefore, what we wanted to do is we wanted to analyze how each of the other variables influences the final price, and if it is possible to predict the price per night by looking at the other features of the Airbnb. To do so, we split the data into training and testing (75% for training and 25% for testing), and we used all the variables, except for the *neighbourhoods* and *apartment type*, which are not numerical; we used different regression models, and compared the results.

We started by using two *robust* linear models: the *Huber model* and the *bisquare model*, and first we checked some assumptions:



(a) Residuals vs Fitted values, Huber robust model.



(b) QQ plot, Huber robust model.

Figure 9: Diagnostics for the Huber robust model.

We checked for multicollinearity in the model by looking at the Variance Infaltion Factor (VIF), and the values were all below 5, so we can assume there's no multicollinearity among the variables:

Variable	**VIF**
dist_stephansdom_km	1.62
dist_schonbrunn_km	1.27
dist_train_station_km	1.90
room_type	1.13
accomodates	1.18
bathrooms	1.08
cleaning_service	1.02
air_conditioning	1.03
self_checkin	1.23
host_acceptance_rate	1.24
host_listings_count	1.15
number_of_reviews	2.64
apt_age_days	2.20
review_scores_rating	1.10
reviews_per_month	1.94

Table 2: VIF values of robust Huber model

We then looked at the regression summary, and this is the result:

```
Call: rlm(formula = price_dollars ~ ., data = regr_trainset, psi = psi.huber)
Residuals:
     Min       1Q   Median       3Q      Max
-101.097  -21.561   -2.133   21.467  208.040

Coefficients:
                       Value    Std. Error t value
(Intercept)              0.9277   6.1147     0.1517
dist_stephansdom_km     -7.1438   0.3600   -19.8428
dist_schonbrunn_km       1.0397   0.1964     5.2947
dist_train_station_km    0.7263   0.3165     2.2946
room_typePrivate room  -20.4680   1.5260   -13.4129
accomodates             11.0188   0.2636    41.7964
bathrooms               12.2605   0.9572    12.8091
cleaning_service         3.2847   1.3913     2.3609
air_conditioning        25.0856   1.1459    21.8911
self_checkin            -3.3396   1.0275    -3.2501
host_acceptance_rate    11.5061   2.6103     4.4079
host_listings_count      0.0244   0.0083     2.9405
number_of_reviews        0.0048   0.0086     0.5638
apt_age_days            -0.0029   0.0006    -4.8197
review_scores_rating    12.0196   1.1218    10.7149
reviews_per_month       -4.8660   0.3304   -14.7260

Residual standard error: 31.96 on 5934 degrees of freedom
```

Figure 10: Summary of the Huber robust regression

If we calculate the p-value for each variable and look at how much significant each variable is, we can see that almost all of the variables all very significant in determining the price of each night in the Airbnbs. The only less significant variables are the distance from the main train station, the number of reviews of each Airbnb and the presence of a cleaning service:

Variable	P-Value	Significance
(Intercept)	0.6892	
dist_stephansdom_km	0.0000	***
dist_schonbrunn_km	7.4e-08	***
dist_train_station_km	0.0614	.
room_typePrivate_room	0.0000	***
accomodates	0.0000	***
bathrooms	0.0000	***
cleaning_service	0.0077	**
air_conditioning	0.0000	***
self_checkin	0.0002	***
host_acceptance_rate	0.0000	***
host_listings_count	0.0006	***
number_of_reviews	0.2566	
apt_age_days	3.1e-08	***
review_scores_rating	0.0000	***
reviews_per_month	0.0000	***

Figure 11: P-Value and significance level of Huber regression variables

This is instead a plot showing the presence of outliers and leverage points after the robust regression:

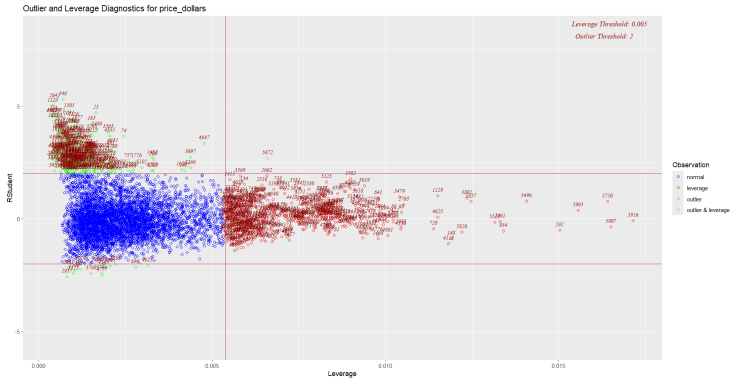


Figure 12: Outliers and leverage points plot

The results of the regression model can be better seen by looking at the difference between the actual values of price per night vs. the values predicted by our model, by plotting them:

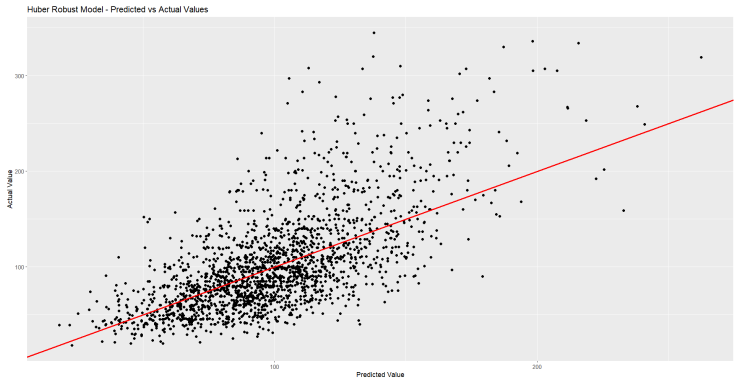


Figure 13: Predicted vs actual values, Huber robust model

Next, we did the same things but with another type of robust model: the bisquare robust model; however, the results are practically the same, with a slightly difference in the residual standard error between the two models:

Model	**R-Squared**	**RMSE**
Huber robust model	0.43	39.18
Bisquare robust model	0.41	39.77

Table 3: Comparison of Huber and bisquare regression results

We then wanted to see how the prediction would change if we used other

non robust models; so, we ran the same problem with three other regression models: *Random Forest Regression*, *XGBoost* and *Support Vector Regression*, and the overall results were better. If we look at the R-Squared, the random forest model is the best one, with XGBoost and SVR slightly worse; the robust models were the ones that performed worse. Instead, by looking at the RMSE, the XGBoost is the best one, and the other models have similar results to each other.

Model	**R-Squared**	**RMSE**
Huber Robust	0.43	39.18
Bisquare Robust	0.41	39.77
Random Forest	0.63	31.32
XGBoost	0.62	65.02
Support Vector Regression	0.57	34.02

Table 4: Comparison of Huber and bisquare regression results

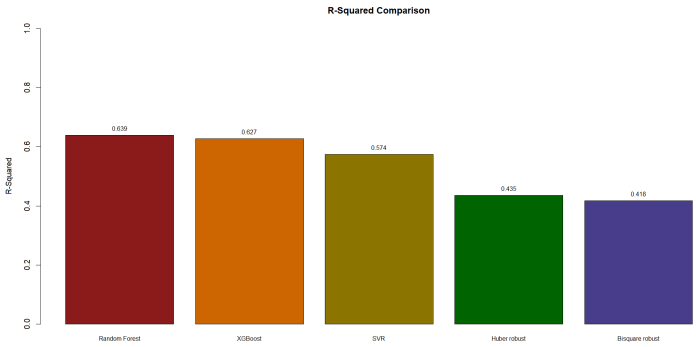


Figure 14: Comparison of R-Squared

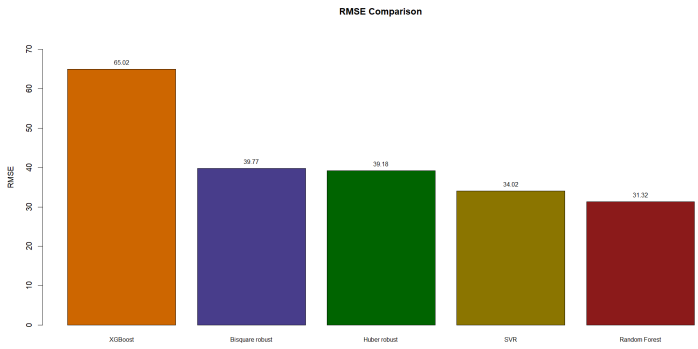


Figure 15: Comparison of RMSE

It is interesting to look at the variable importance in the random forest model, and compare it to the significance level of the Huber robust regression:

Variable	**Overall Score**
accommodates	163.80
dist_stephansdom_km	121.55
air_conditioning	97.24
host_listings_count	85.26
reviews_per_month	80.78
review_scores_rating	66.65
dist_schonbrunn_km	65.52
room_type	64.49
dist_train_station_km	61.08
apt_age_days	56.42
number_of_reviews	53.70
host_acceptance_rate	53.32
bathrooms	46.03
self_checkin	28.34
cleaning_service	22.81

Table 5: Variable importance in the Random Forest model

Although the models aren't particularly accurate at predicting the price per night of the Airbnbs, something very useful that we obtained from the regression is the capacity of determining the importance of each variable in the final price, which can be very useful for tourists when choosing which Airbnb to book.

3 Bootstrapping Methods

In the context of analyzing Airbnb listing prices, ensuring the reliability and robustness of model estimates is paramount. Traditional regression models often rely on strong assumptions about the data distribution, such as normality and homoscedasticity, which may not hold for real-world datasets. To overcome these limitations, we implemented bootstrap resampling methods, including both parametric and nonparametric approaches. Bootstrapping allows us to estimate confidence intervals (CIs) for model coefficients by resampling from the data, providing a flexible way to quantify uncertainty without strict distributional assumptions.

3.1 Parametric vs Nonparametric Bootstrapping

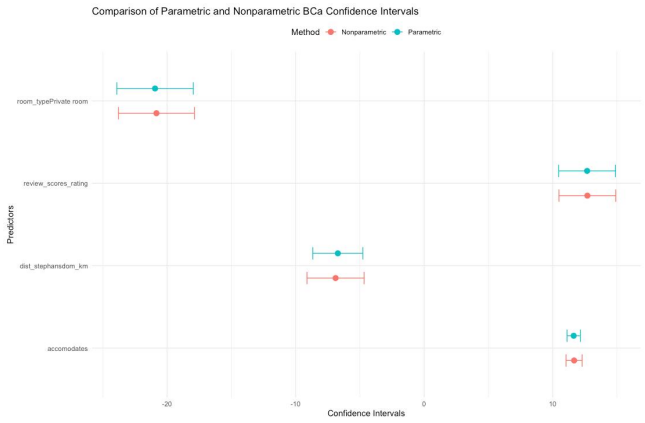
The parametric bootstrap generates resampled datasets by assuming that the model residuals follow a specific distribution (e.g., normal), while the nonparametric bootstrap resamples directly from the observed data. By comparing these two methods, we assess the consistency and stability of our coefficient estimates. The key advantage of this dual approach lies in its ability to confirm whether the results are robust across methods, thereby enhancing the model's credibility.

3.2 Selected Predictors and Research Scope

For this study, we focused on four significant predictors that are both statistically and logically relevant for understanding Airbnb prices:

- Accommodates: The number of guests a listing can host.
- Distance to Stephansdom (dist_stephansdom_km): Proximity to a central landmark, where increasing distance is expected to decrease value.
- Room Type (Private Room): Listings classified as private rooms compared to entire apartments, expected to have a lower price.
- Review Scores Rating: A measure of customer satisfaction, where higher ratings are expected to increase price.

To evaluate these predictors, we derived Bias-Corrected and Accelerated (BCa) confidence intervals using 1,000 bootstrap resamples for both parametric and nonparametric methods. This analysis enables us to verify whether the estimated coefficients remain stable and whether the predictors align with real-world expectations.

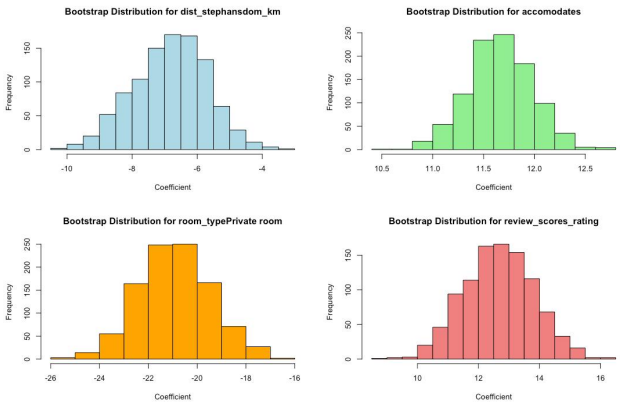


3.3 Bootstrap Distributions and Stability Analysis

To further analyze the robustness of the estimates, we generated bootstrap distributions for the four selected predictors using 1,000 bootstrap samples. These distributions provide insights into the central tendency and variability of the coefficients. As illustrated in the histograms, the distributions are approximately symmetric and centered around their respective coefficient estimates. The lack of substantial skewness or outliers confirms the stability of the predictors under resampling.

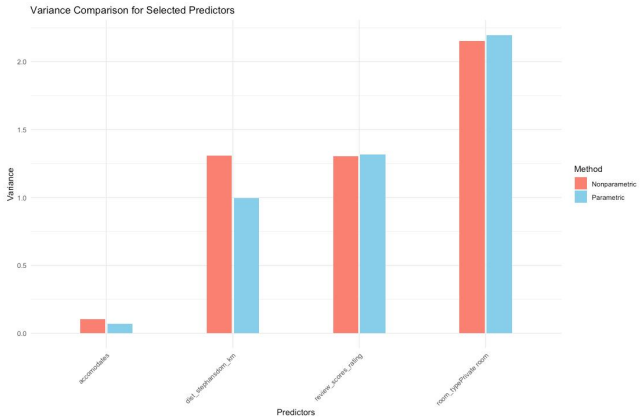
- Accommodates: The bootstrap distribution is centered around a positive coefficient, confirming that larger capacity listings increase the price.
- Distance to Stephansdom: A consistently negative distribution highlights the expected trend that increasing distance from the city center reduces price.
- Room Type (Private Room): The negative values align with the understanding that private rooms are priced lower than entire apartments.
- Review Scores Rating: A strong positive distribution underscores the importance of customer satisfaction ratings in driving higher prices.

These results reinforce the logical relationships between the predictors and the response variable, validating their inclusion in the model.



3.4 Variance Comparison and BCa Confidence Intervals

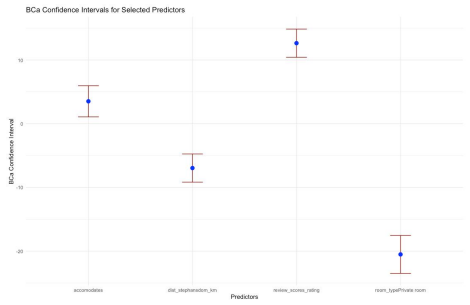
To quantify the variability in the coefficient estimates, we compared the variances derived from parametric and nonparametric bootstrapping methods. The bar plot reveals that for predictors like Accommodates and Review Scores Rating, the variance remains relatively low and consistent across methods. However, slight differences are observed for predictors such as Distance to Stephansdom and Room Type (Private Room). These differences are expected due to the nature of the methods—parametric bootstrap relies on the assumption of normally distributed residuals, whereas nonparametric bootstrap is fully data-driven.



3.5 Final BCa Confidence Interval Analysis

To ensure robust conclusions, Bias-Corrected and Accelerated (BCa) confidence intervals were derived for the selected predictors. BCa intervals adjust for potential bias and skewness, providing more accurate and reliable intervals. The results demonstrate that the intervals for Accommodates and Review Scores

Rating are strictly positive, supporting their significant positive effect on price. Conversely, the intervals for Distance to Stephansdom and Room Type (Private Room) remain consistently negative, reflecting their expected impact.



The bootstrapping analysis conducted in this study highlights the robustness of the regression model estimates by comparing parametric and nonparametric methods. Both approaches yielded consistent results for the selected predictors—Accommodates, Distance to Stephansdom, Room Type (Private Room), and Review Scores Rating—indicating that the findings are stable regardless of the underlying assumptions. The overlap in the BCa confidence intervals, as depicted in Figure 1, further demonstrates this agreement, reinforcing the reliability of the model. The predictors analyzed align well with practical and economic expectations. For instance, Accommodates has a clear positive effect, as higher guest capacity directly increases the potential value of a property. On the other hand, Distance to Stephansdom exhibits a negative effect, confirming the intuitive relationship between proximity to central locations and pricing: properties further from the center are less desirable and priced lower. Similarly, Room Type (Private Room) has a negative coefficient, which reflects the general price difference between private rooms and entire apartments. Lastly, Review Scores Rating has a strong positive effect, underlining the importance of guest satisfaction and reputation in influencing pricing decisions. The comparison of variance between the two bootstrapping approaches, illustrated in Figure 2, provides additional insights into the stability of the coefficient estimates. While slight differences in variance are observed—particularly for Distance to Stephansdom—these differences are minimal. The parametric bootstrap, which assumes normality in the residuals, results in slightly narrower intervals under ideal conditions. Meanwhile, the nonparametric approach captures more data-driven variations, producing slightly wider intervals for some predictors. Nonetheless, the BCa intervals offer an advantage by correcting for bias and skewness in the resampling process, as seen in Figure 3. This adjustment ensures more accurate confidence intervals, which is especially valuable when working with real-world data that often deviates from perfect distributional assumptions. The histograms of the bootstrap distributions, displayed in Figure 4, further confirm the stability of the coefficient estimates. All distributions exhibit near-symmetry around their central values, reflecting the reliability of

the regression results. For example, the distribution of the Accommodates predictor is tightly concentrated, while Distance to Stephansdom and Room Type show slightly wider spreads, corresponding to the observed variances. In conclusion, the bootstrapping analysis demonstrates the robustness of the selected predictors and their contributions to understanding Airbnb pricing. The consistency across methods, supported by BCa intervals and variance comparisons, provides strong evidence that the findings are not artifacts of specific assumptions or sampling variability. These results emphasize the practical relevance of predictors such as location, room type, and guest ratings in determining listing prices. Future extensions of this work could incorporate more granular predictors, such as seasonal trends or neighborhood-specific effects, while exploring advanced modeling techniques like mixed-effects models for improved accuracy.

4 Random Mixed Model

The following chapter aims to provide an analysis of random mixed models, which are highly efficient statistical models that combine the approaches of traditional linear models with the ability to model random effects. In the context of our study, the model includes fixed effects to examine the variable of interest and random effects to capture the variability between different neighbourhoods in the city of Vienna. Such models are used when the data have a hierarchical structure (also known as 'grouped'), i.e. when observations are grouped into districts, and may not be independent. In this way, the dependency between observations is taken into account, thus improving the accuracy of the results.

Fixed effects refer to the variables of each flat in the dataset, and are parameters associated with the entire data population. In contrast, random effects are useful for explaining the excess variability of the dependent variable, thus capturing the price variance associated with neighbourhood-specific characteristics. This approach makes it possible to explore whether the price per night varies according to the area in which the Airbnb is located. It will be possible to observe how some areas of the city are more influenced by factors such as proximity to points of interest than others. The random effects follow a normal distribution, so that, on average, the clustering effect, the influence of the neighbourhood, does not affect the result, as the average of these effects is zero. For this reason, a statistical model that incorporates both fixed effects and random effects is called a mixed-effects model. Before applying this model statistically, certain assumptions need to be defined. Firstly, in MEMs, a clustering factor is added to the linear model, which is the same for all observations within the same group, but varies between the different groups. The specific effects of each group (in this case, neighbourhoods) are captured by the model, making the interpretation clearer and more direct. In addition, the observations may not be independent of each other and, therefore, may be correlated. The model takes into account the fact that there are common characteristics for flats located in the same district. Finally, the dependent variable may not follow a normal distribution.

First of all, the linear regression allowed us to observe that there are general trends between the variables, such as the fact that the distance to Stephansdom has an inverse relationship to the price per night. In other words, the further away the flat is from Vienna Cathedral, the lower its price will be. The graph below shows this trend

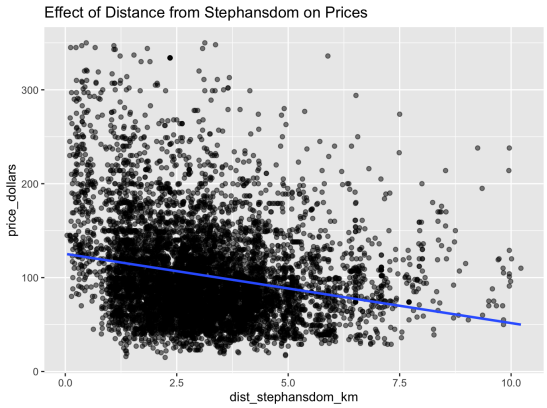


Figure 16: Relationship Between Distance from Stephansdom and Price per Night

Nevertheless, the objective of this chapter is to observe whether, within different neighbourhoods, there are significant variations in flat prices due to neighbourhood-specific factors. The analysis aims to capture random effects, i.e. variability between groups (in this case, neighbourhoods). This highlights how each neighbourhood can have a unique impact on flat prices, justifying the use of mixed models. There are three main types of mixed models: the random intercept model, the random intercept and slope model and the random slope model, the latter of which is less widely used and, therefore, will not be implemented.

In the random intercept model, each district has a different intercept, while the slope remains fixed, so all lines are parallel. This implies that each district can have a different base level for the price per night. By statistically processing the code on R, the following output is obtained.

```
Random effects:
 Groups        Name        Variance Std.Dev.
 neighbourhood (Intercept)  299.4   17.30
 Residual                  1301.9   36.08
Number of obs: 7931, groups:  neighbourhood, 21

Fixed effects:
                        Estimate Std. Error t value
(Intercept)           -1.352e+01  7.539e+00  -1.794
dist_stephansdom_km   -8.037e+00  9.087e-01  -8.844
dist_schonbrunn_km     2.181e+00  6.585e-01   3.311
dist_train_station_km  2.902e+00  9.683e-01   2.997
room_typePrivate room -1.959e+01  1.412e+00 -13.879
accomodates            1.170e+01  2.444e-01  47.871
bathrooms              1.643e+01  8.809e-01  18.654
cleaning_service       2.470e+00  1.276e+00   1.936
air_conditioning       2.543e+01  1.057e+00  24.067
self_checkin          -4.606e-01  9.459e-01  -0.487
host_acceptance_rate   1.062e+01  2.414e+00   4.398
host_listings_count    1.402e-02  7.714e-03   1.818
number_of_reviews     -8.943e-03  7.964e-03  -1.123
apt_age_days          -2.474e-03  5.618e-04  -4.404
review_scores_rating   1.237e+01  1.054e+00  11.734
reviews_per_month     -5.551e+00  3.102e-01 -17.892
```

Figure 17: Random Intercept Model Output

The first important point to note in this output is the estimated variance between the groups. The variance of the random intercept is 299.4, while the residual variance is 1301.9. The intra-group variance indicates that there is a significant variation in the average flat price between the different neighbourhoods in Vienna, which is not explained by the other independent variables. In other words,

the neighbourhoods do not have the same average price level and show significant differences. The residual variance represents the unexplained part of the model, which is the difference between the observed and predicted values for each flat. The fraction of variance explained by random effects (PVRE) is 0.1869, which means that 18.69% of the variability of flat prices is explained by the variability between neighbourhoods. Next, the fixed effects of the independent variables can be assessed. For example, the variable 'distance to Stephansdom' has a negative effect on price, with an estimate of -8.037, indicating that as the distance to Stephansdom increases, the price tends to decrease. To observe the random effects, two graphs are presented: a dotplot and a density map.

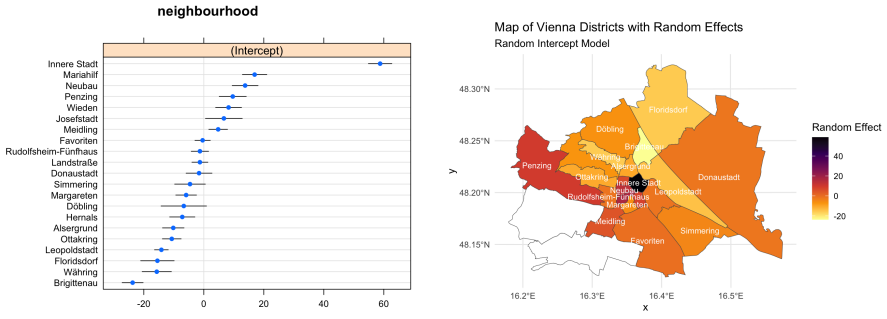


Figure 18: Random Intercept Model Dotplot - Random Effects

The graph shows the distribution of random effects for each district and how far the average price of a district deviates from the overall average, which is 0. As can be seen, the Innere Stadt district, located in the historic centre, has an extremely high value of 58.76, indicating that this district has a significantly higher price than the average. Other central districts such as Mariahilf and Neubau show high random effects. In contrast, more suburban districts such as Brigittenau, Floridsdorf and Währing have negative random effects, suggesting that their average prices are lower than the overall average. This variability between districts cannot be explained by fixed variables. Finally, the geographical map allows for a better visualisation of the neighbourhoods and their variability. Thus, there is a difference of more than $80 in average prices between the neighbourhood with the highest intercept and the one with the lowest intercept. This demonstrates differences between neighbourhoods in terms of location, tourist attractiveness, and availability of services.

```
Random effects:
 Groups        Name        Variance Std.Dev. Corr
 neighbourhood (Intercept)  120.25  10.966
               bathrooms     68.93   8.303   0.28
 Residual                  1288.59  35.897
Number of obs: 7931, groups:  neighbourhood, 21

Fixed effects:
                            Estimate Std. Error t value
(Intercept)               -1.284e+01  6.818e+00  -1.884
dist_stephansdom_km       -7.254e+00  8.610e-01  -8.426
dist_schonbrunn_km         2.106e+00  6.065e-01   3.472
dist_train_station_km      2.146e+00  9.118e-01   2.354
room_typePrivate room     -1.888e+01  1.408e+00 -13.411
accomodates                1.174e+01  2.435e-01  48.200
bathrooms                  1.665e+01  2.052e+00   8.116
cleaning_service           2.469e+00  1.271e+00   1.943
air_conditioning           2.545e+01  1.051e+00  24.203
self_checkin              -4.483e-01  9.427e-01  -0.476
host_acceptance_rate       1.027e+01  2.405e+00   4.272
host_listings_count        1.321e-02  7.692e-03   1.718
number_of_reviews         -9.723e-03  7.929e-03  -1.226
apt_age_days              -2.421e-03  5.596e-04  -4.326
review_scores_rating       1.235e+01  1.050e+00  11.771
reviews_per_month         -5.519e+00  3.089e-01 -17.867
```

We proceed with the analysis using a random intercept and slope model, which allows us to examine random effects both at the neighbourhood level and in relation to the number of bathrooms, a variable that can be interpreted as indicative of flat size and comfort. This methodology makes it possible to capture variations in average prices between neighbourhoods and to understand how the impact of the number of bathrooms on prices may differ according to the local context. In the model, the variable `bathrooms` was selected, as it is significant in determining the price of a flat. The variance of the intercept is 120.25, indicating a significant difference in average prices between neighbourhoods underlining the importance of geographical location in flat prices..

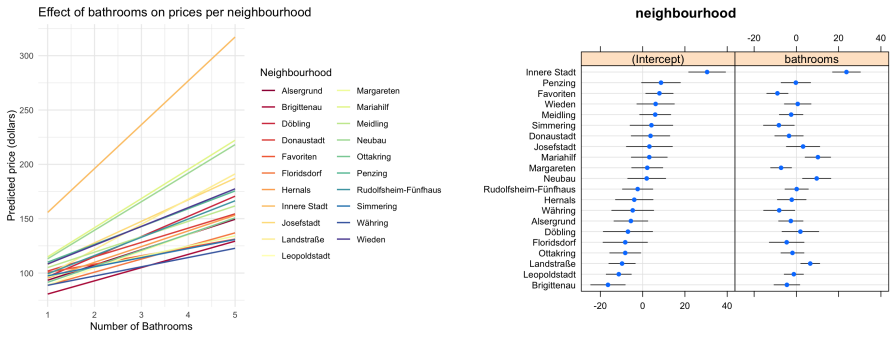


Figure 19: Random Intercept and Slopes Model Dotplot - Random Effects

The dotplot illustrates the random effects for the intercept (baseline neighbourhood influence on prices) and bathrooms. Innere Stadt, in the historic city center, stands out with significantly higher baseline prices, followed by the central districts of Favoriten, Wieden, and Meidling. In contrast, outer districts like Brigittenau, Leopoldstadt, and Floridsdorf have notably lower baseline prices. In the random intercept and slope model, the slope for the number of bathrooms varies significantly between Vienna's districts, reflecting different market dynamics. In central and prestigious districts, such as Innere Stadt, Mariahilf, and Neubau, the number of bathrooms has a strongly positive impact on prices, indicative of the high demand for spacious and comfortable flats in these areas. In contrast, in suburban or residential neighbourhoods, such as Favoriten, Simmering, and Währing, the effect is weak or even negative, suggesting that the number of bathrooms plays a less relevant role in price determination. This re-

sult highlights how the value attributed to comfort, represented by bathrooms, is closely linked to geographical location and neighbourhood profile.

To further clarify, the density map with the values is shown.

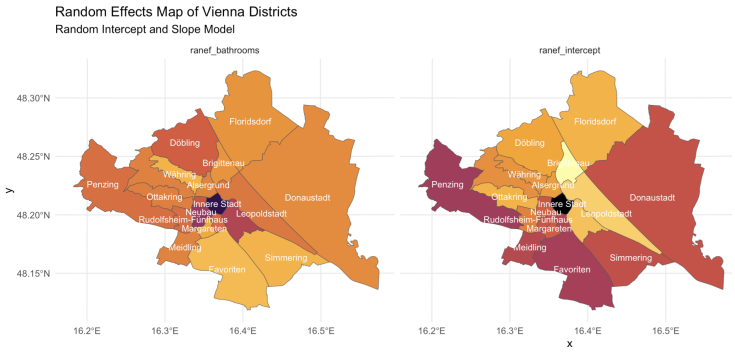


Figure 20: Random Effects Map of Vienna Districts - Random Intercept and Slope Model

To conclude this chapter, we compare the two models by using the `anova` method.

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)	
fit_lmm_rand_intercept	18	79496	79621	-39730	79460				
fit_lmm_rand_int_and_slope	20	79438	79578	-39699	79398	61.459	2	4.511e-14	***

Figure 21: Comparison between Models

The random intercept and slopes model has a lower AIC (Akaike Information Criterion) value, indicating that, despite the larger number of parameters, it fits the data better. The BIC (Bayesian Information Criterion) is also lower, suggesting a preference for this model. The log-likelihood, which measures the probability that the model generated the observed data, is higher in the model with intercept and random slopes, confirming that it is the best fit. Furthermore, the lower deviance of the model with intercept and random slopes suggests a better fit. The very low p-value indicates that the random intercept and slopes model is significantly better than the random intercept. In conclusion, the Random Intercept and Slope is preferred because of its better fit to the data, its statistical significance and its greater flexibility.

5 Model Based Clustering

The traditional clustering methods, such as hierarchical clustering and k-means clustering, are heuristic and are not based on formal models. Furthermore, k-means algorithm is commonly randomly initialized, so different runs of k-means will often yield different results. Additionally, k-means requires the user to

specify the optimal number of clusters. An alternative is model-based clustering, which consider the data as coming from a distribution that is mixture of two or more clusters (Fraley and Raftery 2002, Fraley et al. (2012)). Unlike k-means, the model-based clustering uses a soft assignment, where each data point has a probability of belonging to each cluster.

5.1 Estimating model parameters

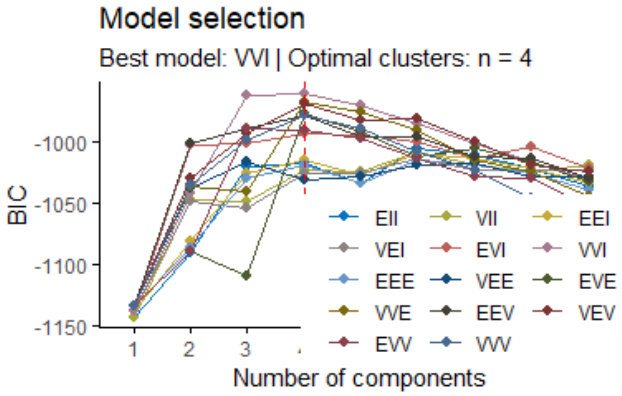
The model parameters can be estimated using the *Expectation-Maximization* (EM) algorithm initialized by hierarchical model-based clustering. The available model options, in mclust package, are represented by identifiers including: EII, VII, EEI, VEI, EVI, VVI, EEE, EEV, VEV and VVV.

```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust VVI (diagonal, varying volume and shape) model with 4 components:

 log-likelihood   n df       BIC       ICL
      -440.8218 200 19 -982.3117 -1061.157

Clustering table:
 1  2  3  4
92 43 23 42
```

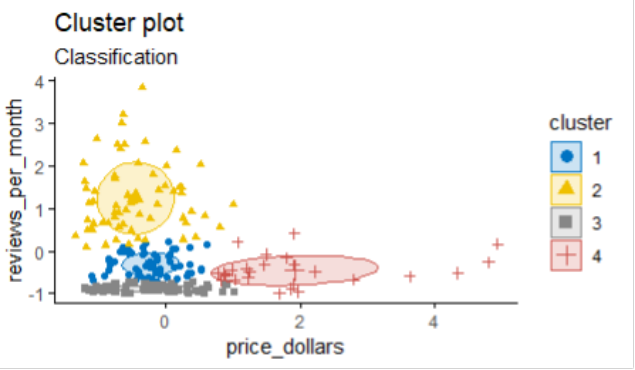
For this data, it can be seen that model-based clustering selected a model with four components (i.e. clusters). The optimal selected model name is VVI model. That is the four components are diagonal, with varying volume and shape.

5.2 Visualizing model-based clustering

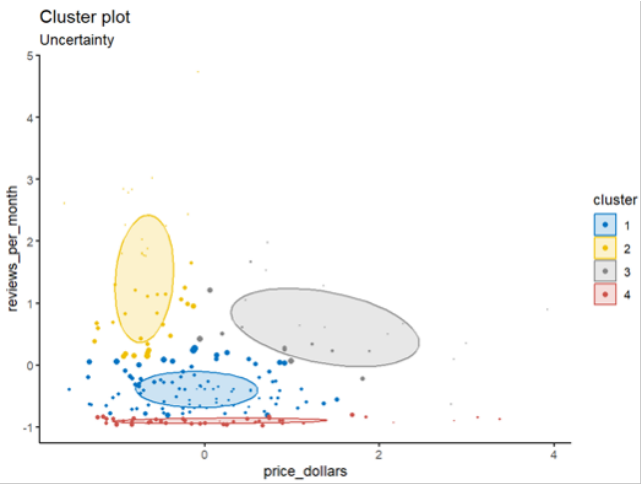
Model-based clustering results can be drawn using the base function plot function in mclust package. Here we'll use the function *fviz_mclust()* [in *factoextra* package] to create beautiful plots based on ggplot2. In the situation, where the data contain more than two variables, *fviz_mclust()* uses a principal component analysis to reduce the dimensionality of the data. However, it's also possible to plot the data using only two variables of interest as we did in our case with the price in dollars and number of reviews per month variables. The BIC plot indicates the optimal number of clusters. The chosen model (e.g., four clusters) corresponds to the highest BIC value, balancing model complexity and goodness of fit.



The classification plot below displays the assignment of data points to clusters based on the model. It helps visualize how well-separated and distinct the clusters are. Clear boundaries between clusters suggest good classification.



And finally the uncertainty graph below represents the uncertainty in cluster assignments for each data point. Low uncertainty indicates high confidence in cluster assignments, while high uncertainty suggests overlapping clusters or ambiguous points.

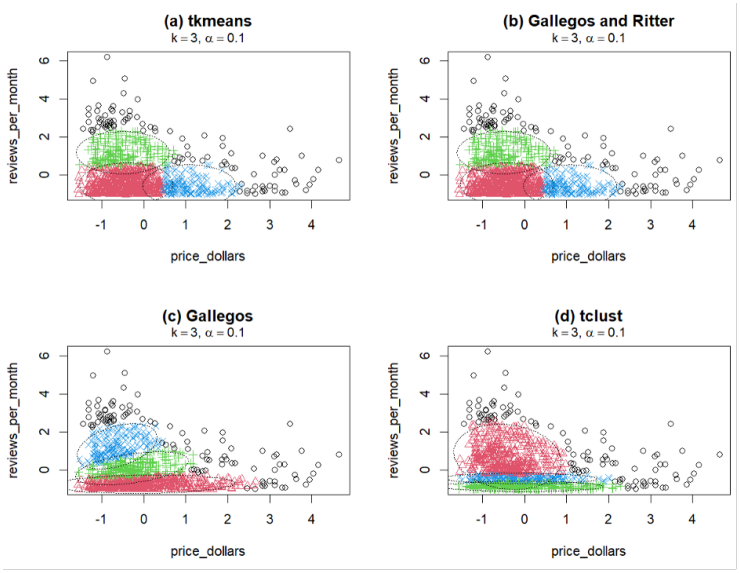


6 Robust Clustering

Robust clustering is a specialized area of clustering techniques aimed at addressing challenges posed by noise, outliers, and other data irregularities. Traditional clustering methods, such as k-means and hierarchical clustering, often assume well-behaved data distributions and can be sensitive to anomalies, leading to misleading or suboptimal cluster assignments. Robust clustering methods enhance the reliability and accuracy of clustering results, particularly in real-world datasets that frequently exhibit complexities. Two prominent robust clustering techniques are Trimmed K-Means and TCLUST (Trimmed Cluster).

1. **Trimmed K-Means** : Trimmed K-Means is a robust extension of the traditional K-Means algorithm. It works by excluding a fixed proportion of the most distant points (outliers) from the clustering process. Some of the Key Features include Outlier Trimming where a fixed proportion of data points, denoted by a parameter α $(0 < \alpha < 1)$, is trimmed based on their distance from cluster centroids. These points are considered outliers and are not included in the cluster assignment.

2. **TCLUST** (Trimmed Cluster Analysis) : TCLUST extends the concept of robust clustering by combining trimming and model-based clustering approaches. It is particularly effective for data with clusters of varying shapes, sizes, and densities. The key advantage that this method has over the tkmeans is that it handles clusters with varying shapes and sizes effectively.



It is mainly used in Image and Video Processing where it aids in identifying patterns in noisy or incomplete visual data. Also in the Bioinformatics fields where it can be used for gene expression analysis and identification of subtypes in disease studies, where datasets are inherently noisy. In Market Segmentation as well where it ensures reliable grouping of customer data, even with missing or anomalous entries. And finally in Cybersecurity where it helps in anomaly detection for identifying malicious activities in network data.

Some of its challenges facing include Scalability vs. Robustness where achieving a balance between computational efficiency and robustness remains a challenge for extremely large datasets. The Interpretability also tends to become an issue where many robust methods involve complex computations, making the resulting clusters less interpretable. Lastly the combination of Dynamic and Streaming Data: Adapting robust clustering methods to dynamic datasets is an area of active research.

7 Conclusion

The analysis of Vienna's Airbnb data using a comprehensive array of statistical methods, including robust inference, bootstrapping, random mixed models, model-based clustering, and robust clustering, provided insightful results. Each technique contributed unique perspectives, allowing for a nuanced understanding of the data. Robust inference and bootstrapping enhanced the reliability of parameter estimates and provided distribution-free confidence intervals, mitigating sensitivity to outliers. Random mixed models captured complex hierarchical relationships, while model-based clustering identified patterns in host and property characteristics. Robust clustering complemented this by detect-

ing structures resistant to noise. Together, these approaches yielded a robust and multidimensional analysis, offering valuable insights for policymakers and stakeholders in Vienna's short-term rental market.