Vienna **Airbnb** Price Analysis

Honest Albert Temu
Cihan Elveren
Tommaso Premoli
Luca Sangiovanni

Dataset Description

Distance from Stephansdom	Distance from Schönbrunn castle	Distance from the central station	Neighbourhood

Room type	Accomodates	Bathrooms
Cleaning service	Air conditioning	Self check-in

Host acceptance rate	Host listings count	Number of reviews
Age of the flat	Review scores rating	Reviews per month

Data Cleaning

- Cook's Distance
- Absolute standardized residuals
- High leverage points
- Studentized residuals
- Z - score

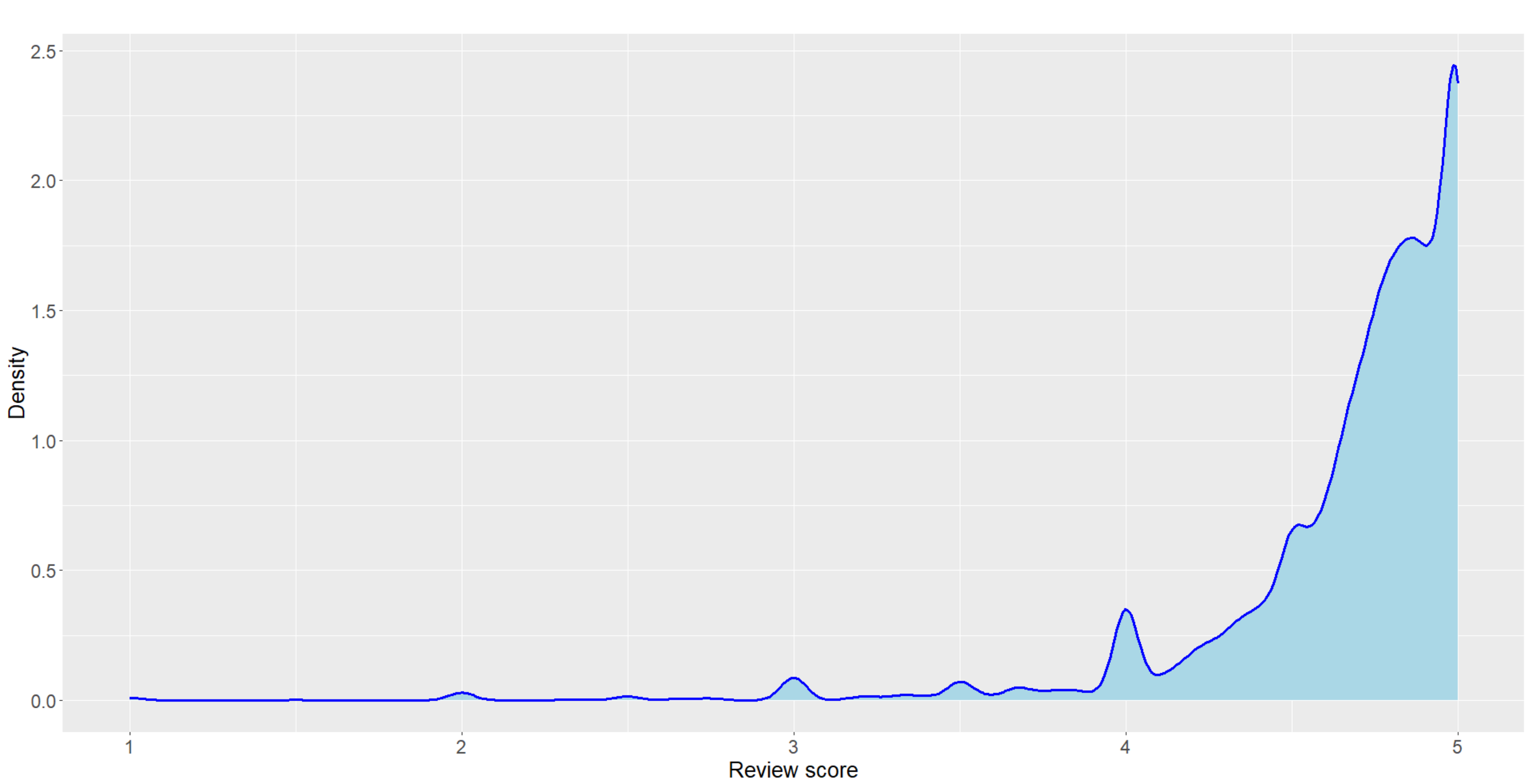
14,396

Total observations in the original dataset

7,931

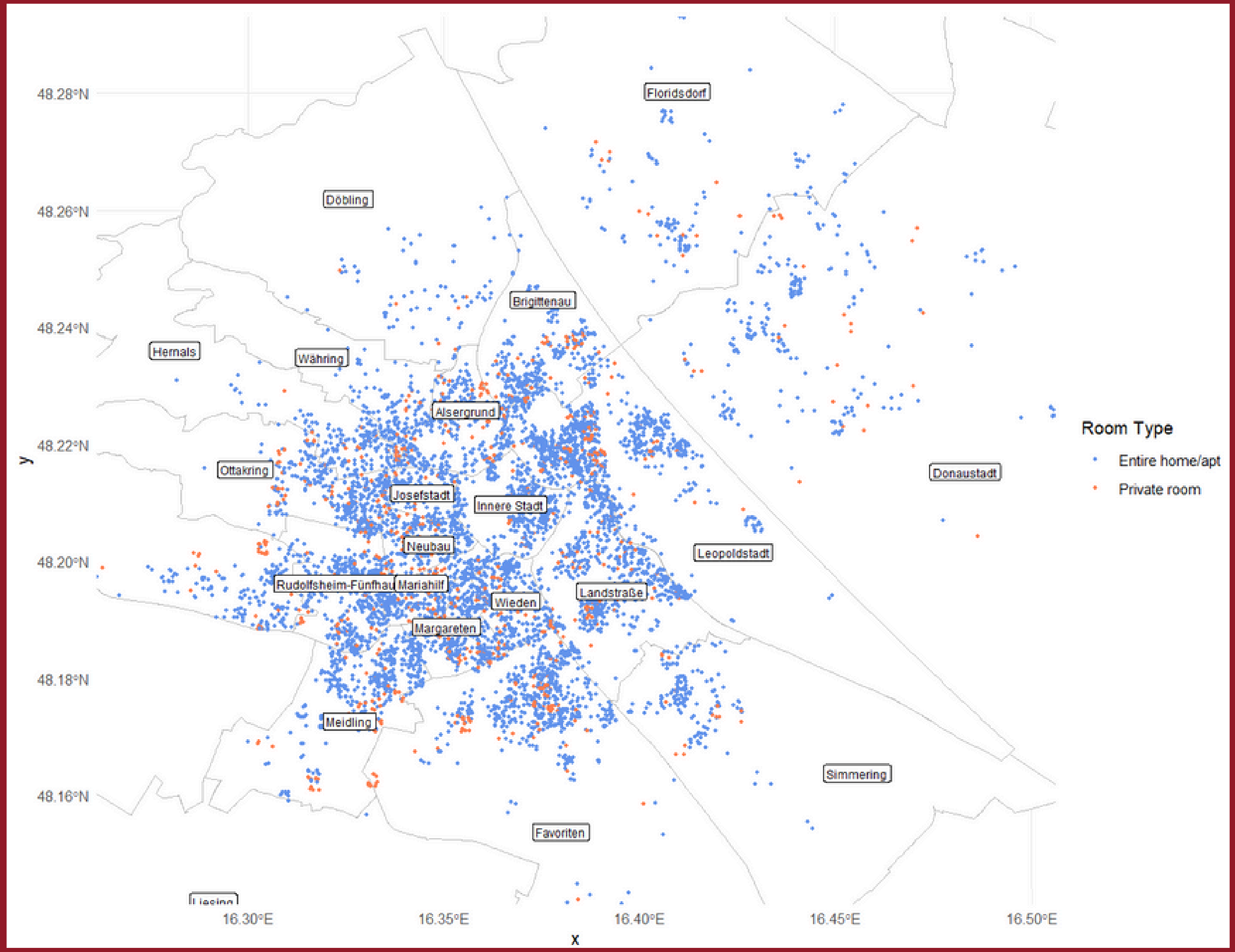
Total observations after outlier removal

Exploratory Data Analysis

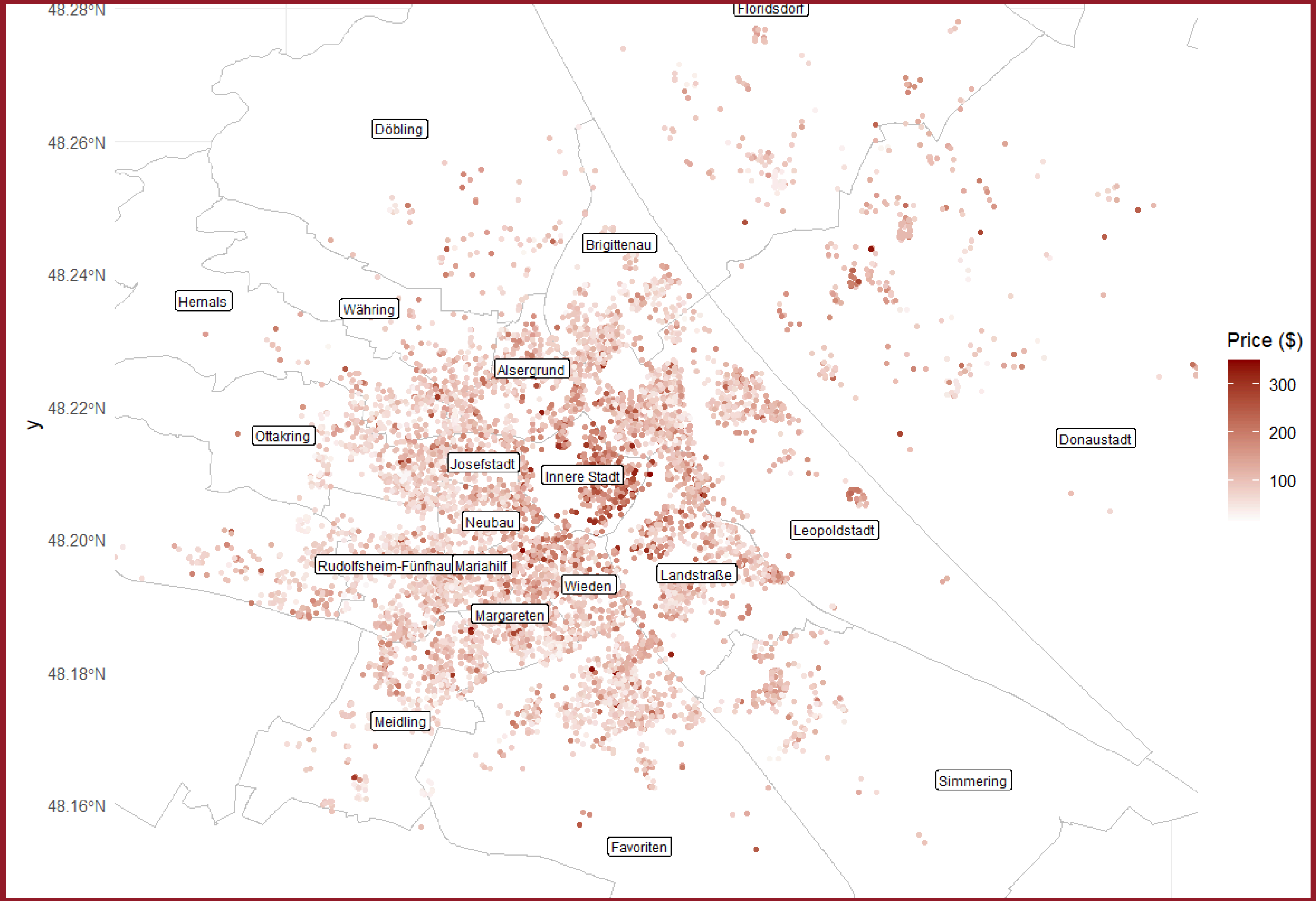


As we can see, listings tend to have high review scores, on average.

If we look at room type, we can see that most of the listings are entire rooms or apartments, and just a few are private rooms.

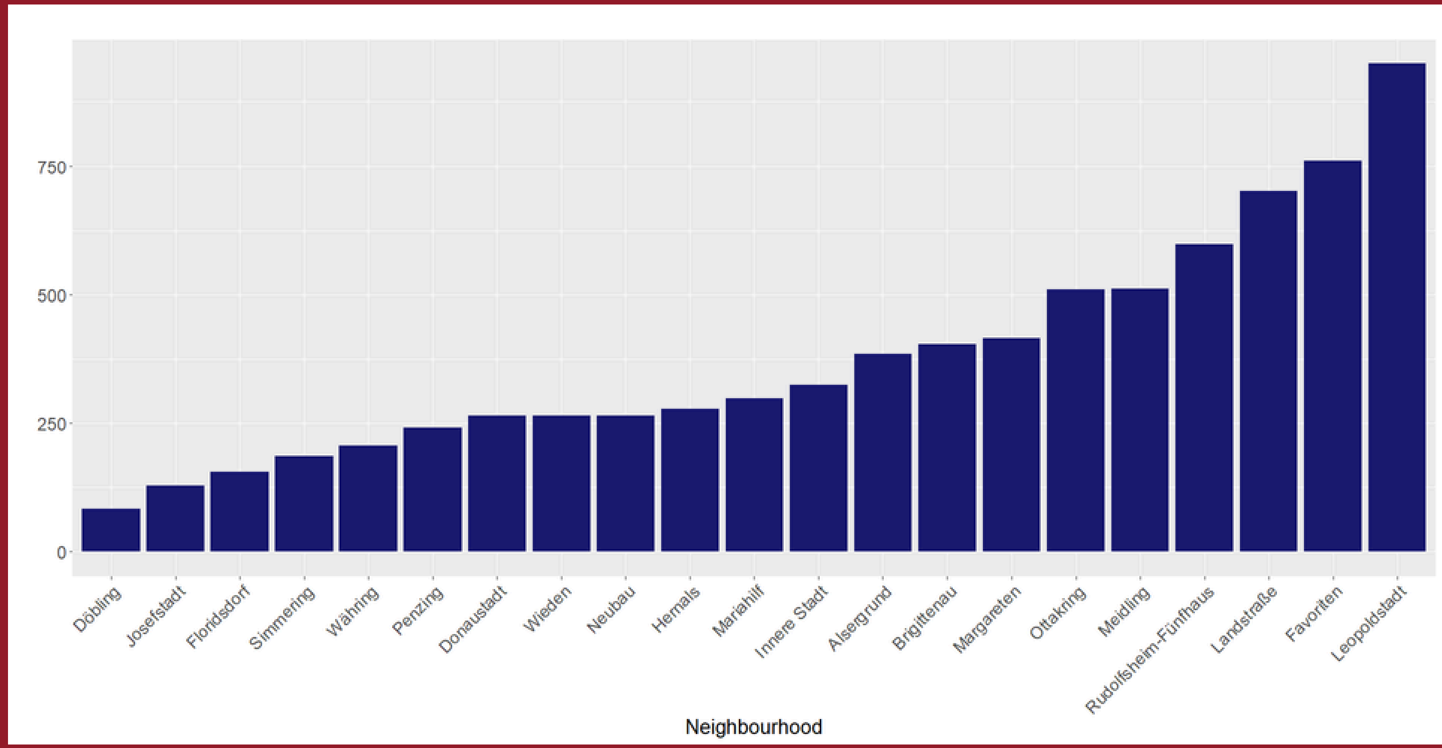


Exploratory Data Analysis



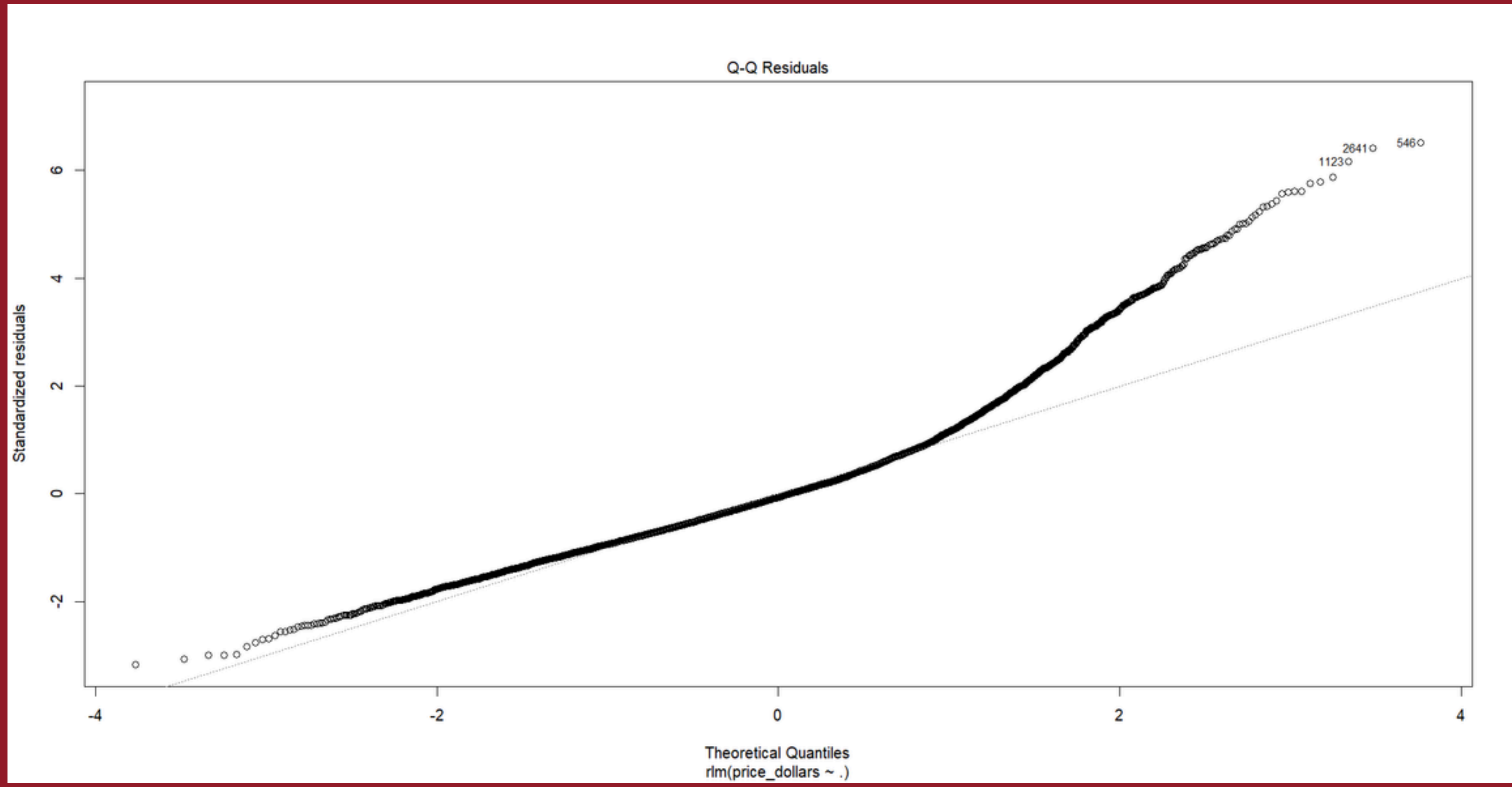
Airbnbs in the city center, of course, tend to have higher prices.

If we instead look at the number of listings in each neighbourhood, we can see that some nieghbouhoods have way more Airbnbs than others.



Robust Regression

By looking at the QQ plot, we can assume normality



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

If we look at the Variance Infalation Factor (VIF), we can assume there is no multicollinearity among the variables

Response variable: price_dollars

In order to improve the model, we split the data into training and testing

75% Training set

25% Testing set

Robust Regression

Almost all of the variables are significant in determining the price per night of an Airbnb

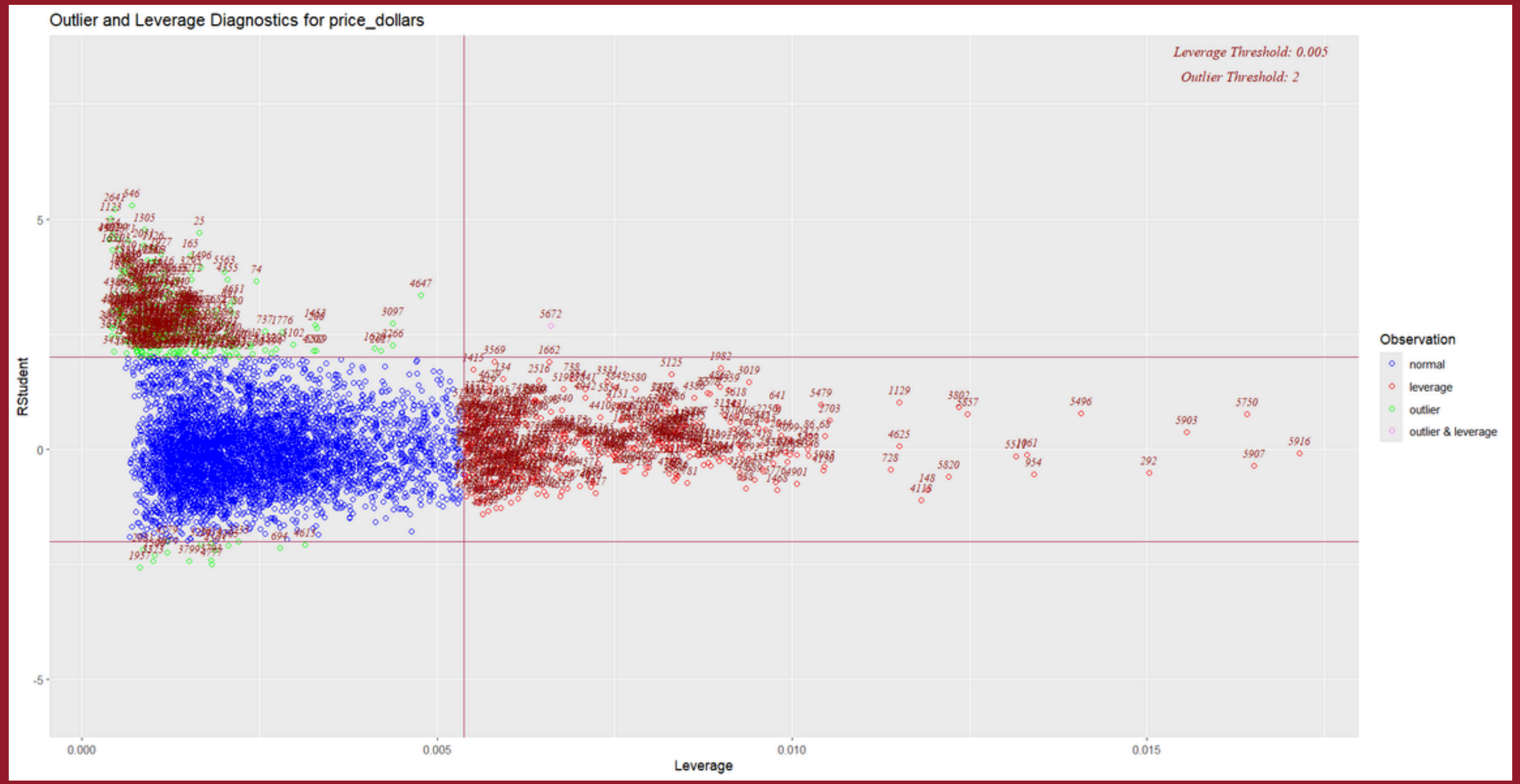
```
Call: rlm(formula = price_dollars ~ ., data = regr_trainset, psi = psi.huber)
Residuals:
     Min       1Q   Median       3Q      Max
-101.097  -21.561   -2.133   21.467  208.040
```

Coefficients:

	Value	Std. Error	t value	P-Value	Significance
(Intercept)	0.9277	6.1147	0.1517	0.6892	
dist_stephansdom_km	-7.1438	0.3600	-19.8428	0.0000	***
dist_schonbrunn_km	1.0397	0.1964	5.2947	7.4e-08	***
dist_train_station_km	0.7263	0.3165	2.2946	0.0614	.
room_typePrivate room	-20.4680	1.5260	-13.4129	0.0000	***
accomodates	11.0188	0.2636	41.7964	0.0000	***
bathrooms	12.2605	0.9572	12.8091	0.0000	***
cleaning_service	3.2847	1.3913	2.3609	0.0077	**
air_conditioning	25.0856	1.1459	21.8911	0.0000	***
self_checkin	-3.3396	1.0275	-3.2501	0.0002	***
host_acceptance_rate	11.5061	2.6103	4.4079	0.0000	***
host_listings_count	0.0244	0.0083	2.9405	0.0006	***
number_of_reviews	0.0048	0.0086	0.5638	0.2566	
apt_age_days	-0.0029	0.0006	-4.8197	3.1e-08	***
review_scores_rating	12.0196	1.1218	10.7149	0.0000	***
reviews_per_month	-4.8660	0.3304	-14.7260	0.0000	***

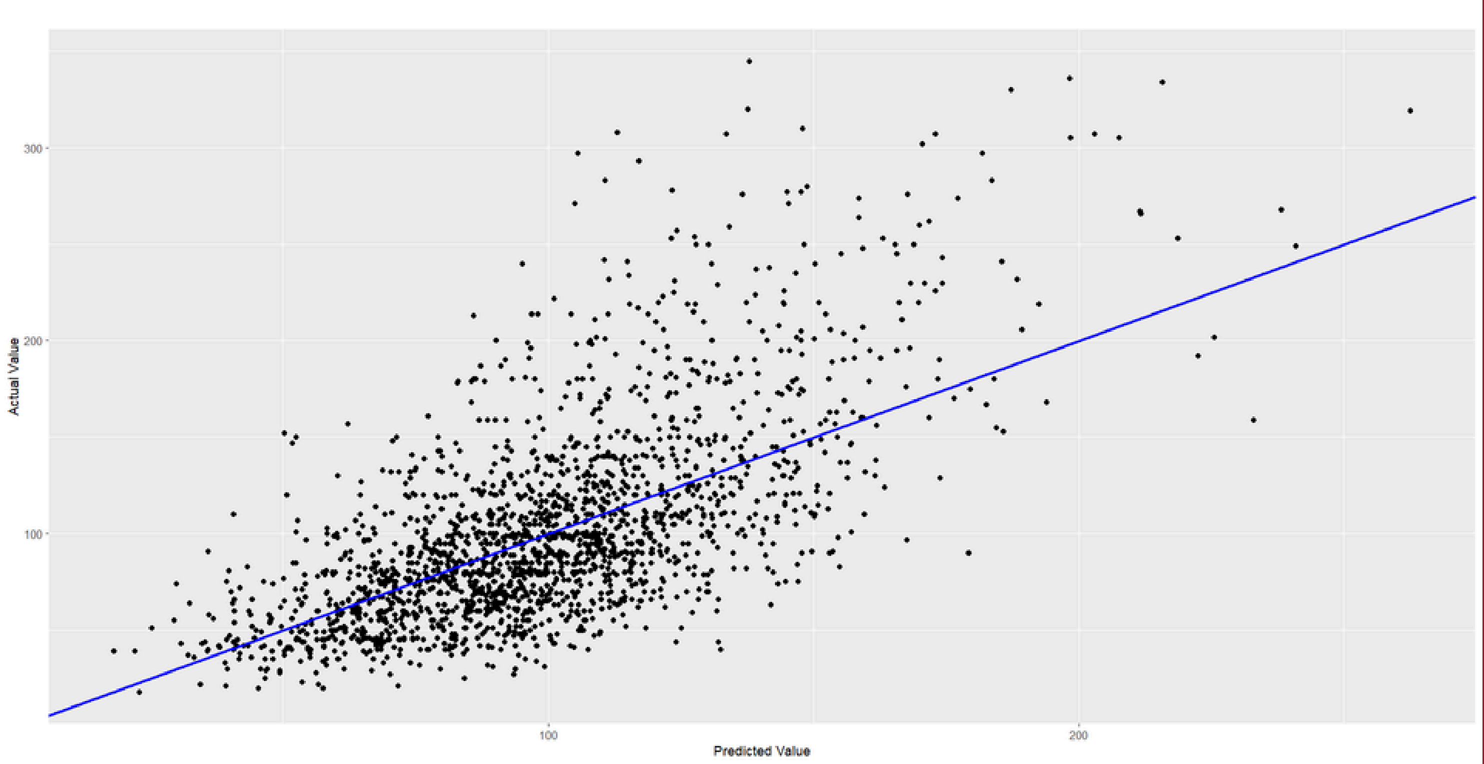
Residual standard error: 31.96 on 5934 degrees of freedom

Even after using various techniques for outliers removal, there are still a lot of points that can be considered outliers



Robust Regression

Predicted vs actual values in the Huber robust model:

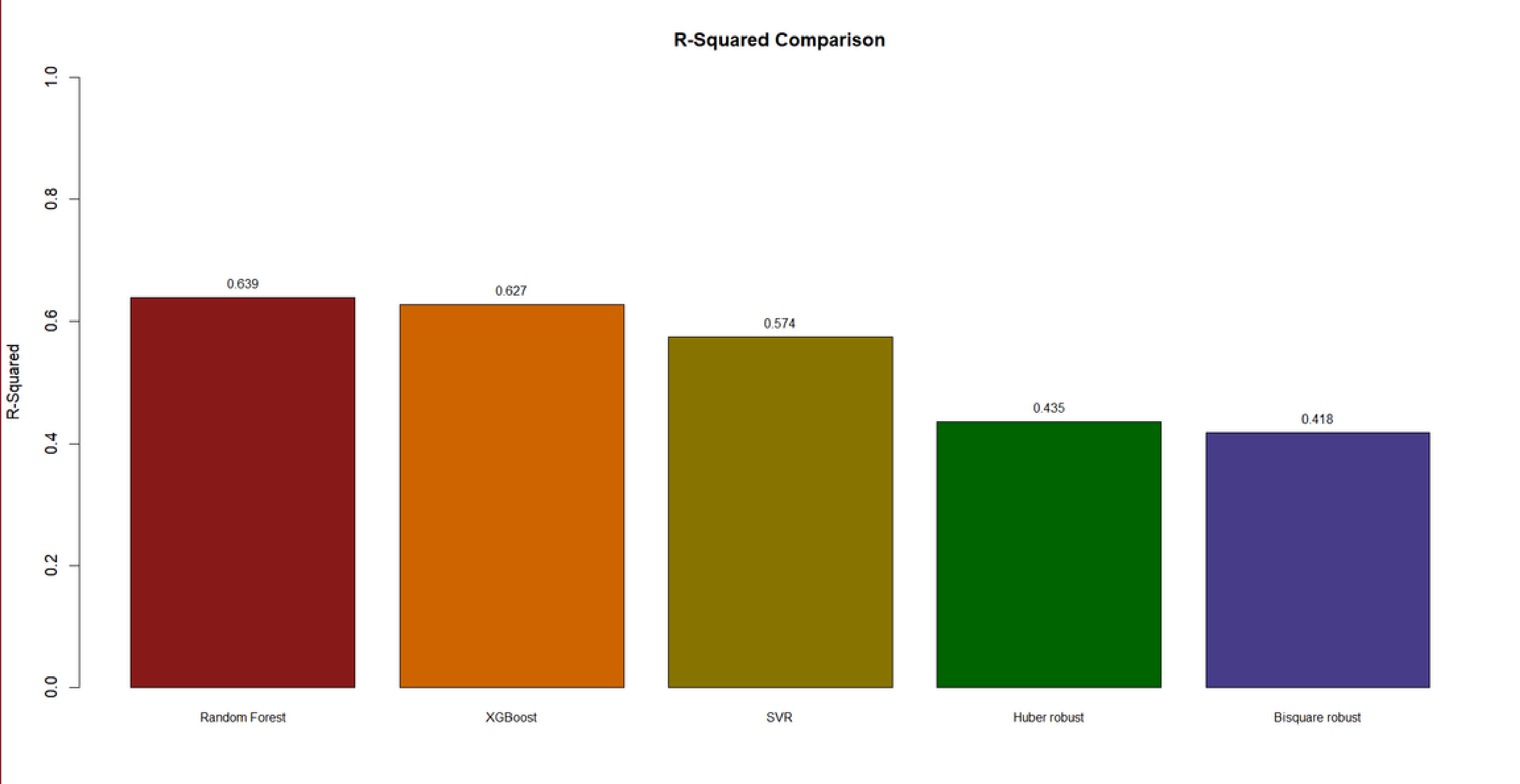


Variable	Overall Score
accommodates	163.80
dist_stephansdom_km	121.55
air_conditioning	97.24
host_listings_count	85.26
reviews_per_month	80.78
review_scores_rating	66.65
dist_schonbrunn_km	65.52
room_type	64.49
dist_train_station_km	61.08
apt_age_days	56.42
number_of_reviews	53.70
host_acceptance_rate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

This is the variable importance according to the Random Forest model. It is interesting to see the differences and similarites with the output of the Huber robust model

We tried to run a prediction using other regression methods, and these are the results:

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



Bootstrap

Methods used:

Parametric

Non parametric

1000 resamples

(R = 1000)

Most 4 used predictors:

accomodates

dist_stephansdom_km

room_typePrivateRoom

review_scores_rating

Average confidence intervals:

accomodates	room_typePrivateRoom
[11.04, 12.29]	[-23.91, -17.87]

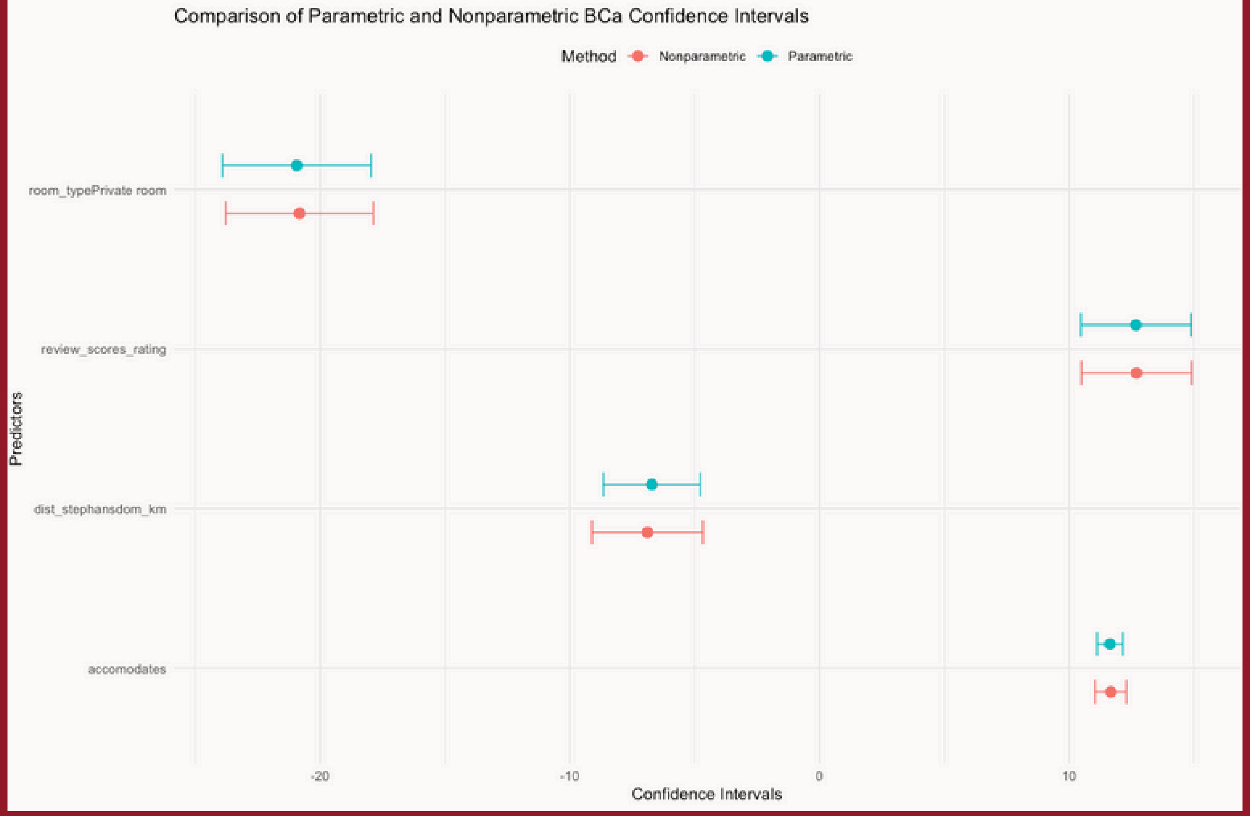
General Perspective: narrow and consistent confidence intervals confirm the robustness of our model.

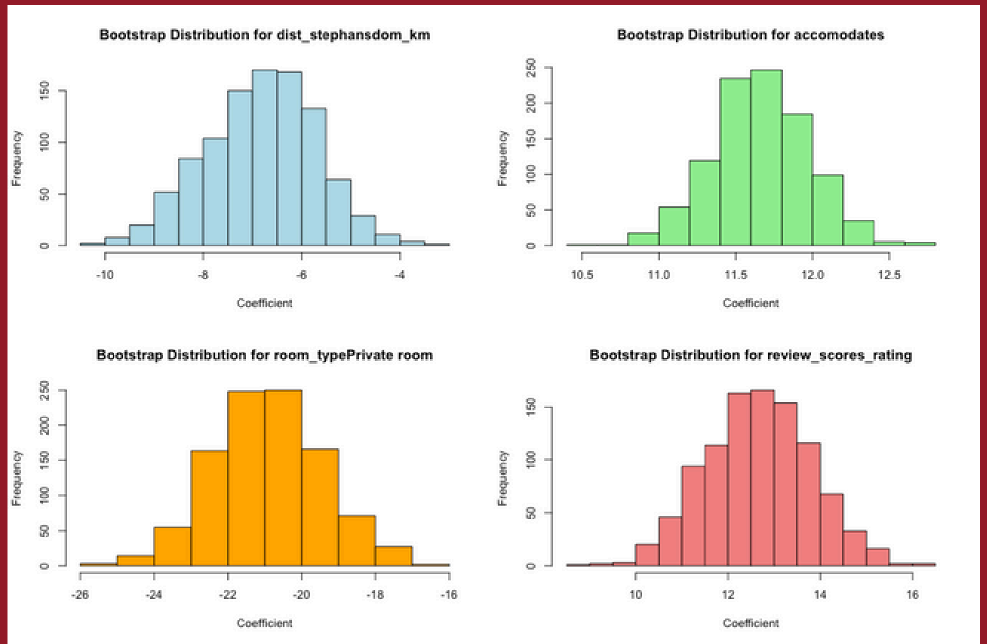
Bootstrap

- Why Both Methods?
  - Parametric: Assumes specific distribution of data.
  - Nonparametric: Distribution-free, relies on resampling.
    - Additionally, the effects of the 4 predictors align logically with expectations, confirming the model’s practical relevance
- **Consistency** between methods strengthens the model's reliability.

- Why BCa Confidence Intervals?
  - Adjust for **bias** and **skewness** in the sampling process.
  - Provide more accurate and reliable confidence intervals.

- The signs of the coefficients (positive or negative effects) for the 4 predictors align with real-world logic:





<u>Bootstrap Resampling</u>: 1000 samples.

<u>Symmetrical Distributions</u>: Stable & consistent results.

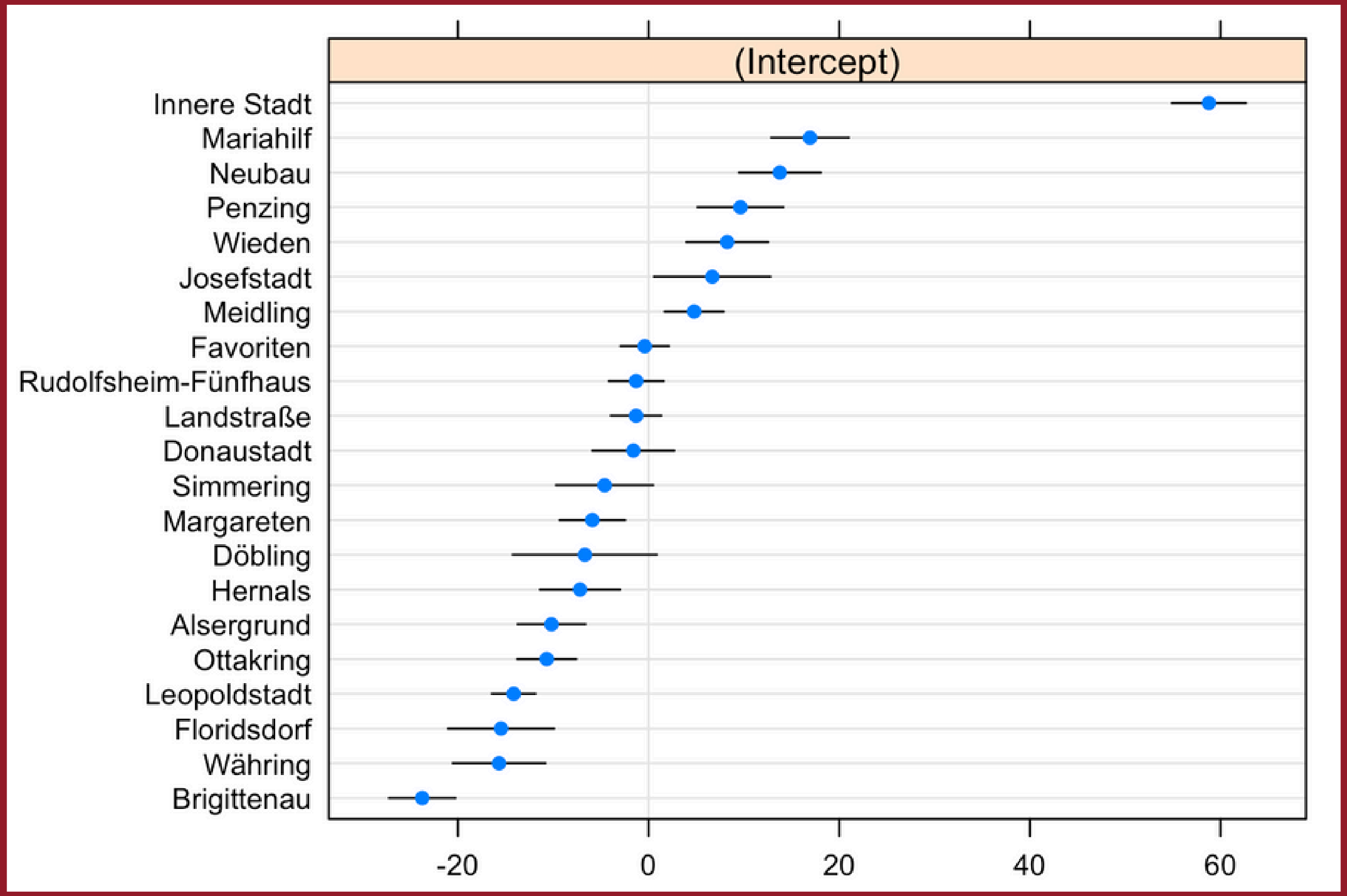
<u>Reliable Estimates</u>: Minimal variation in key predictors.

<u>Reinforced Robustness</u>: Model outputs are reproducible

Random Intercept Model

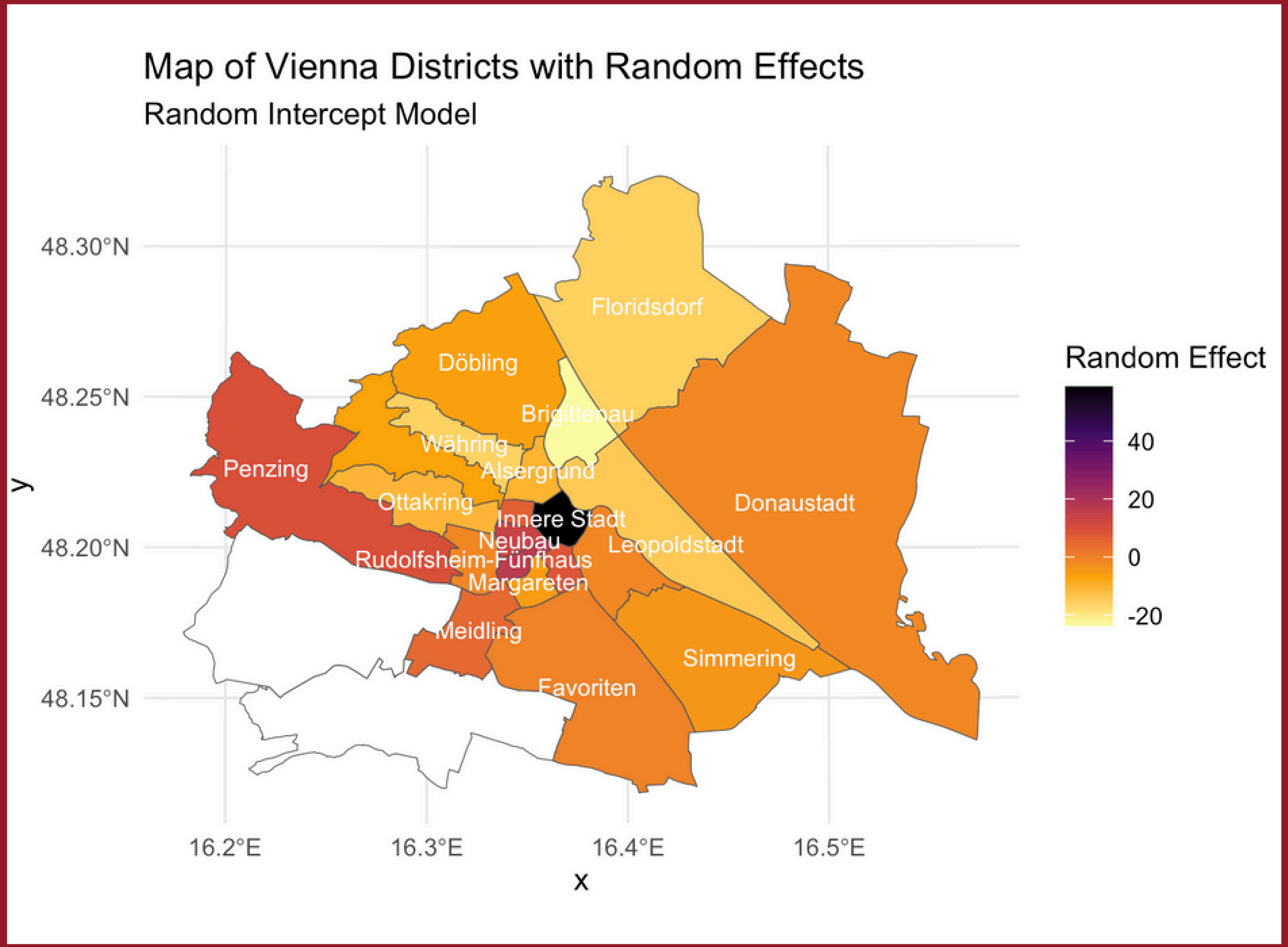
Variability among neighbourhoods

Significant differences in average prices between neighbourhoods.



Importance of location

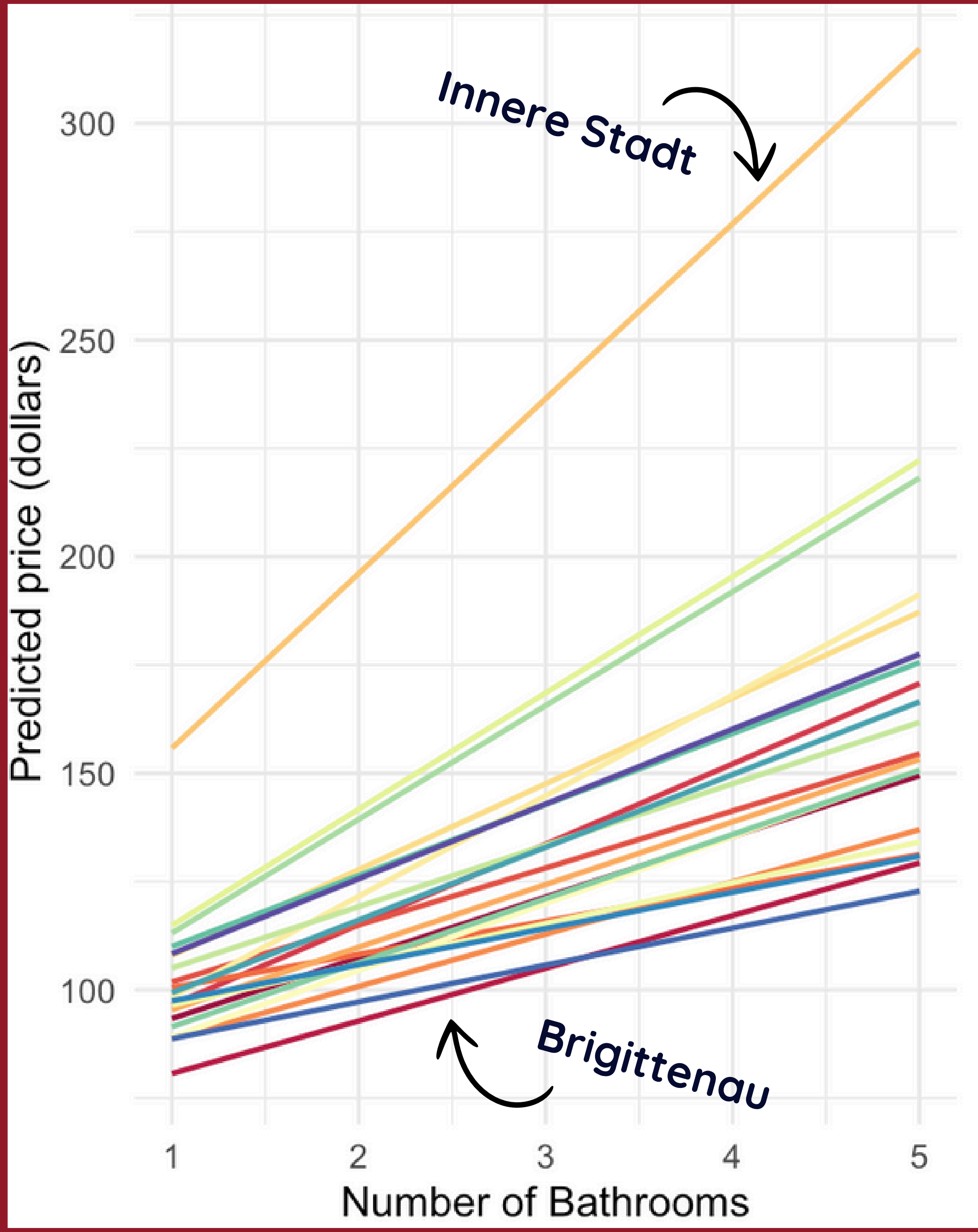
Central neighbourhoods show higher prices than suburban ones.



Random Intercept and Slopes Model

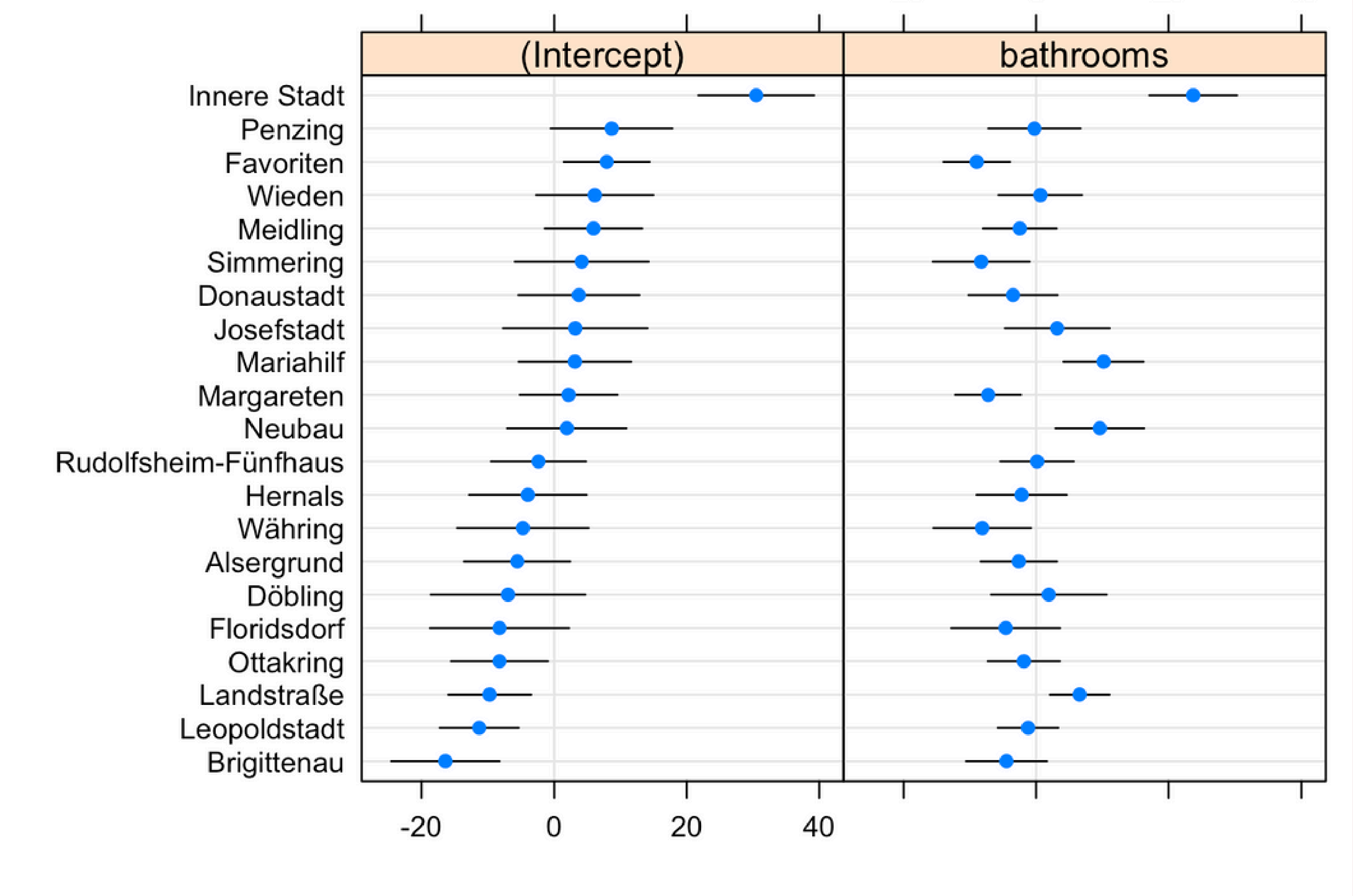
Variation in prices

Central districts such as Innere Stadt show higher average prices, while peripheral ones has lower prices .

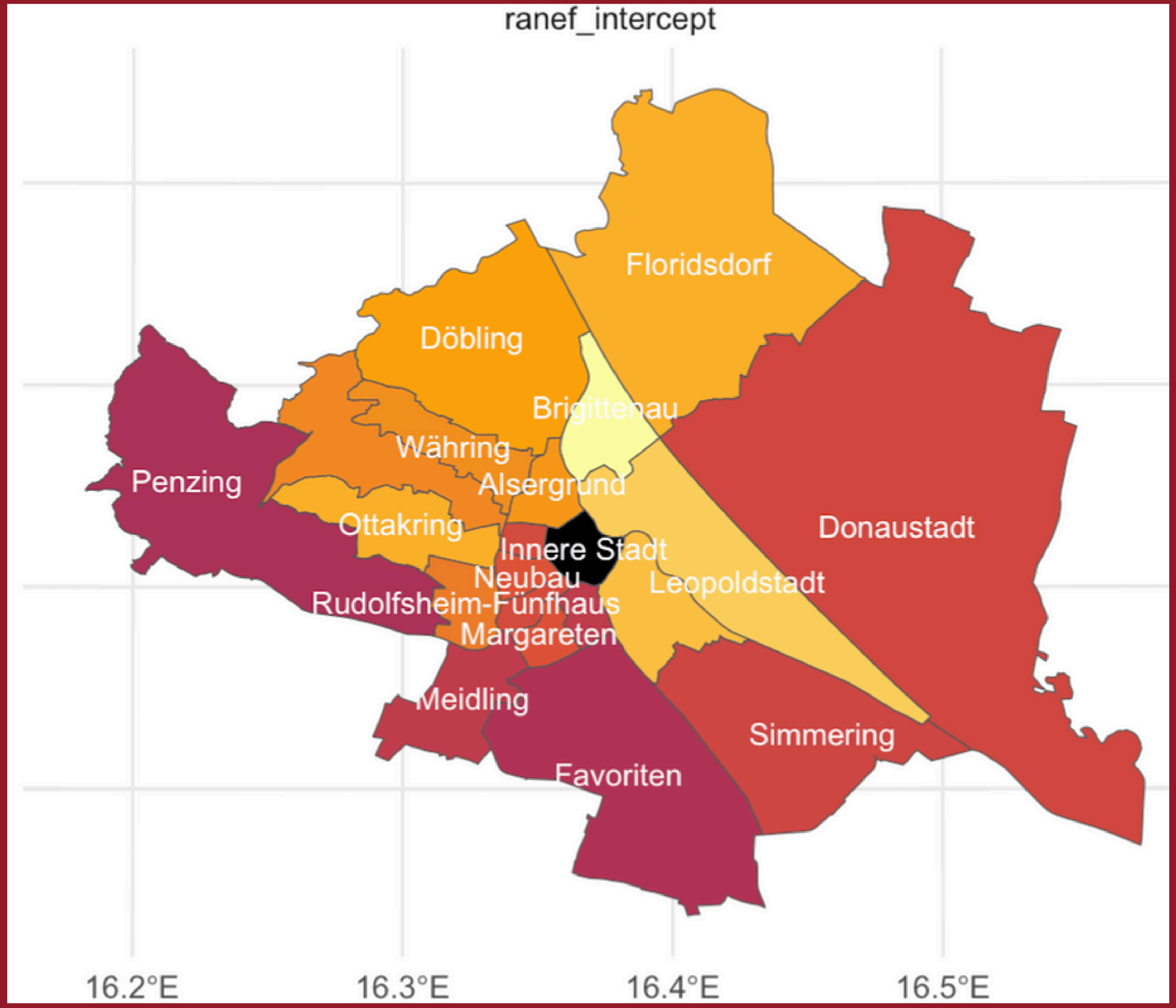


Comfort value

In prestigious districts like Innere Stadt, bathrooms indicate comfort and size, raising prices significantly (+23.7). In suburban districts like Simmering, the number of bathrooms has little influence on price, reflecting a lower demand for additional comfort and space.

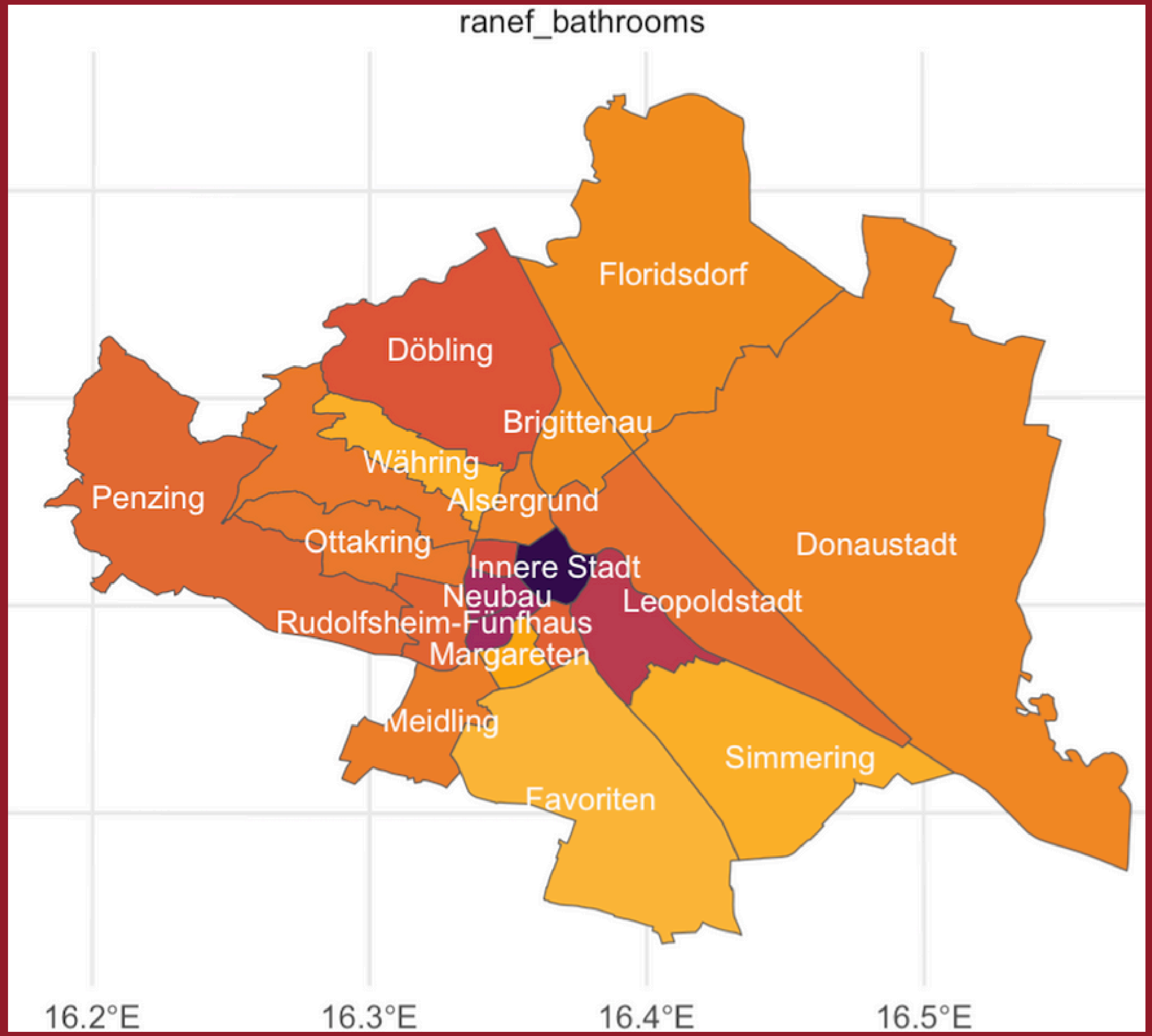


Random Intercept and Slopes Model

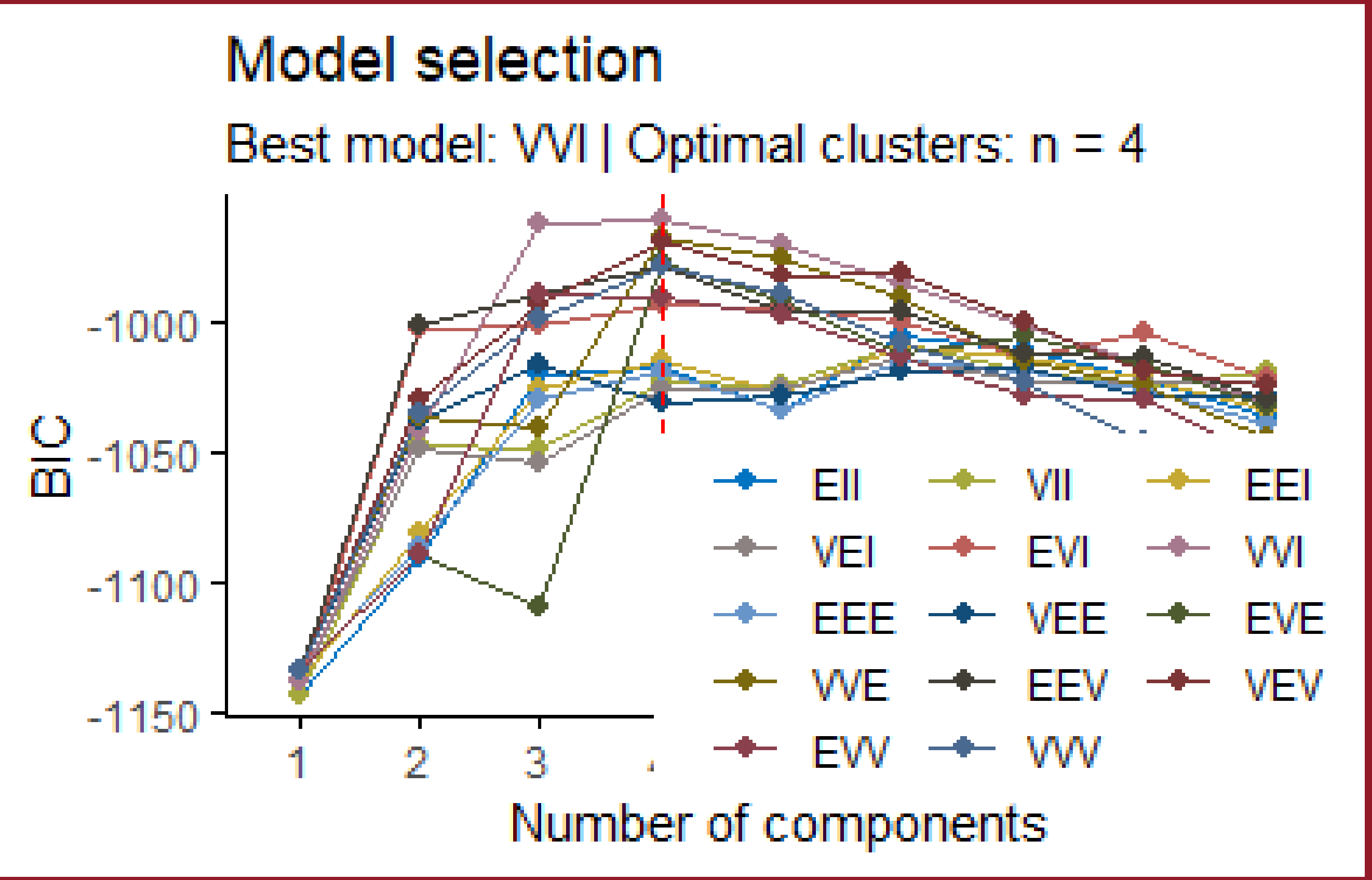


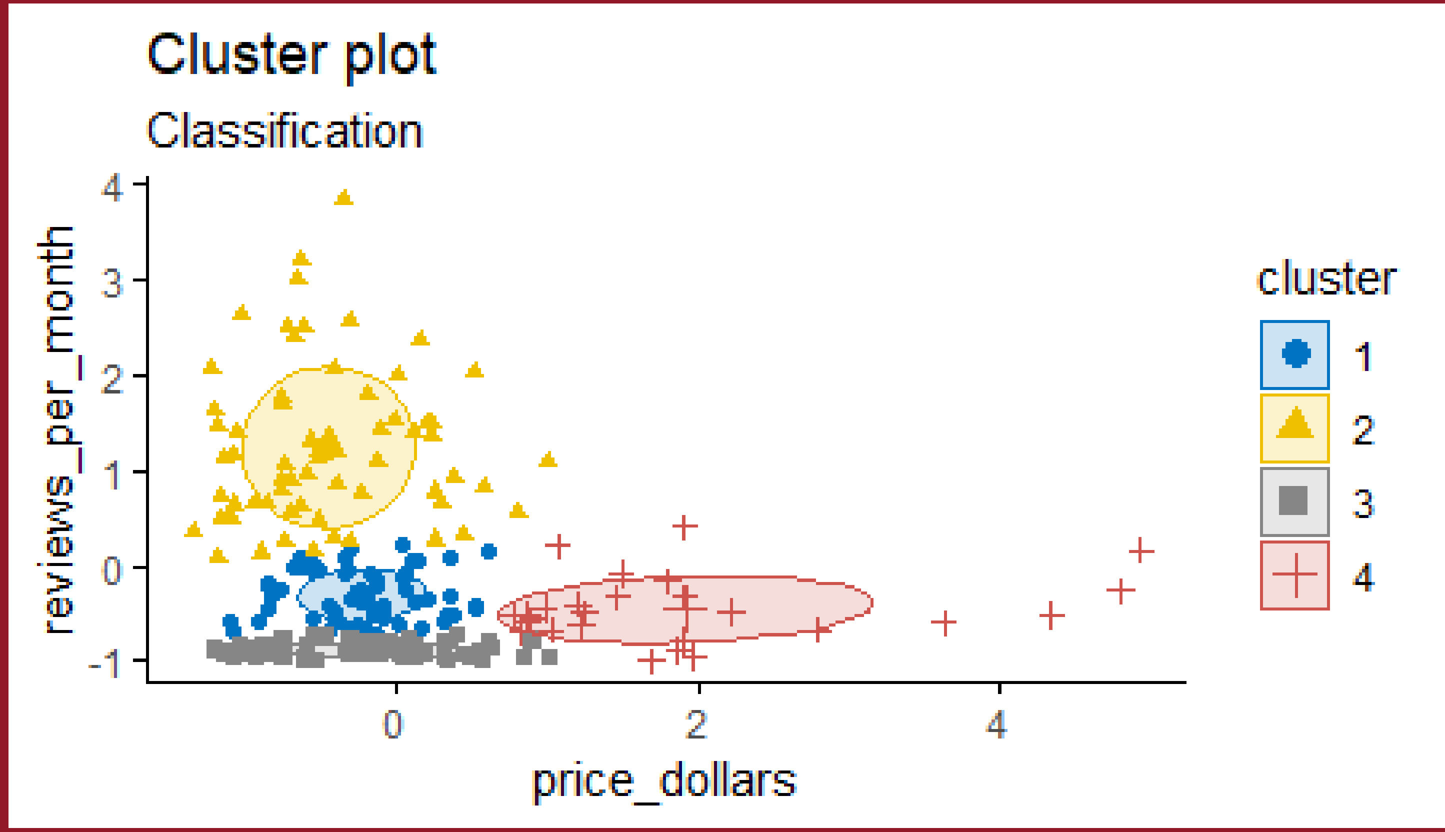
How the average price of a flat in that neighbourhood differs from the average value.

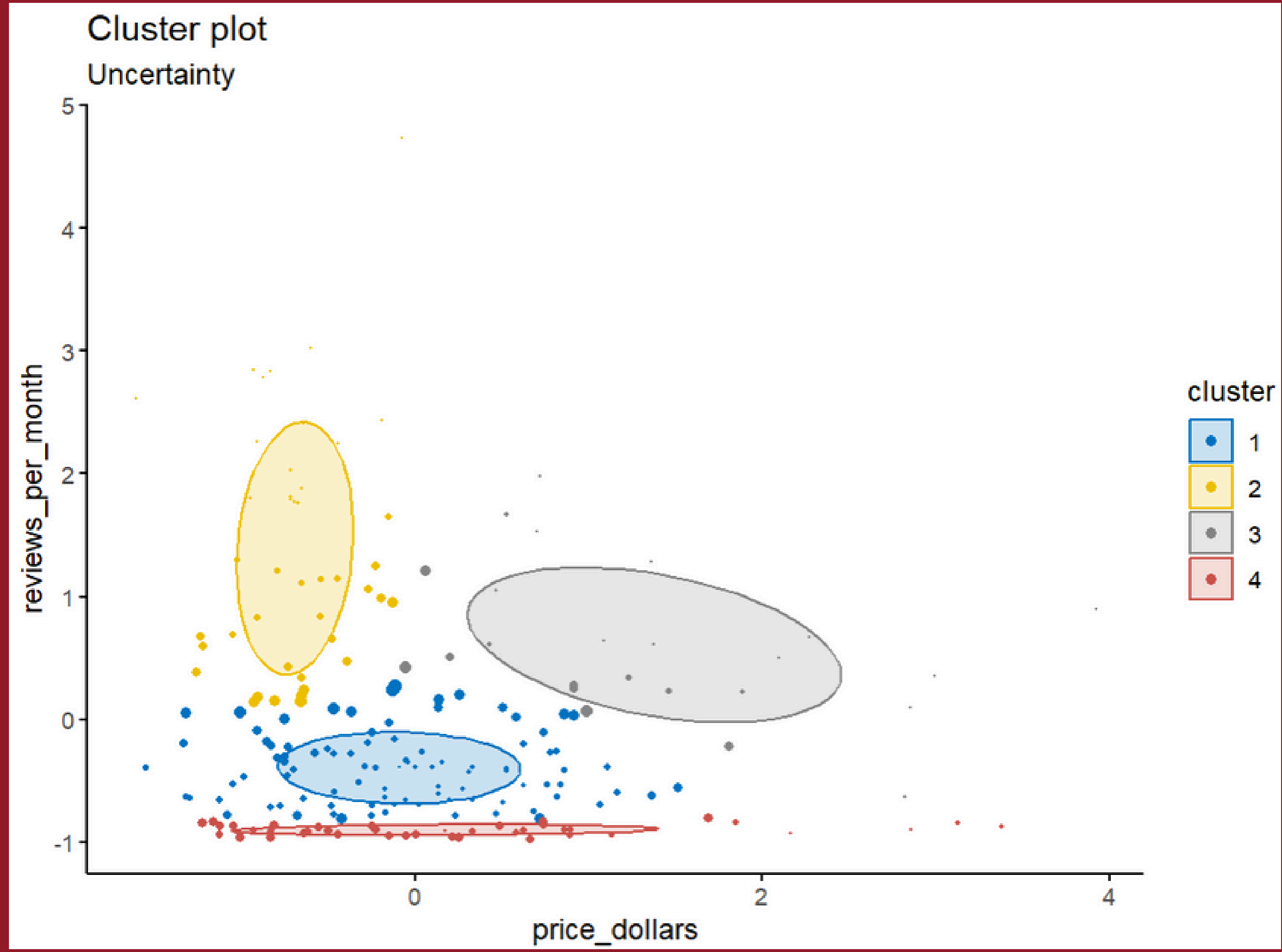
How much the impact of the number of bathrooms on prices varies between districts.



Model Based Clustering



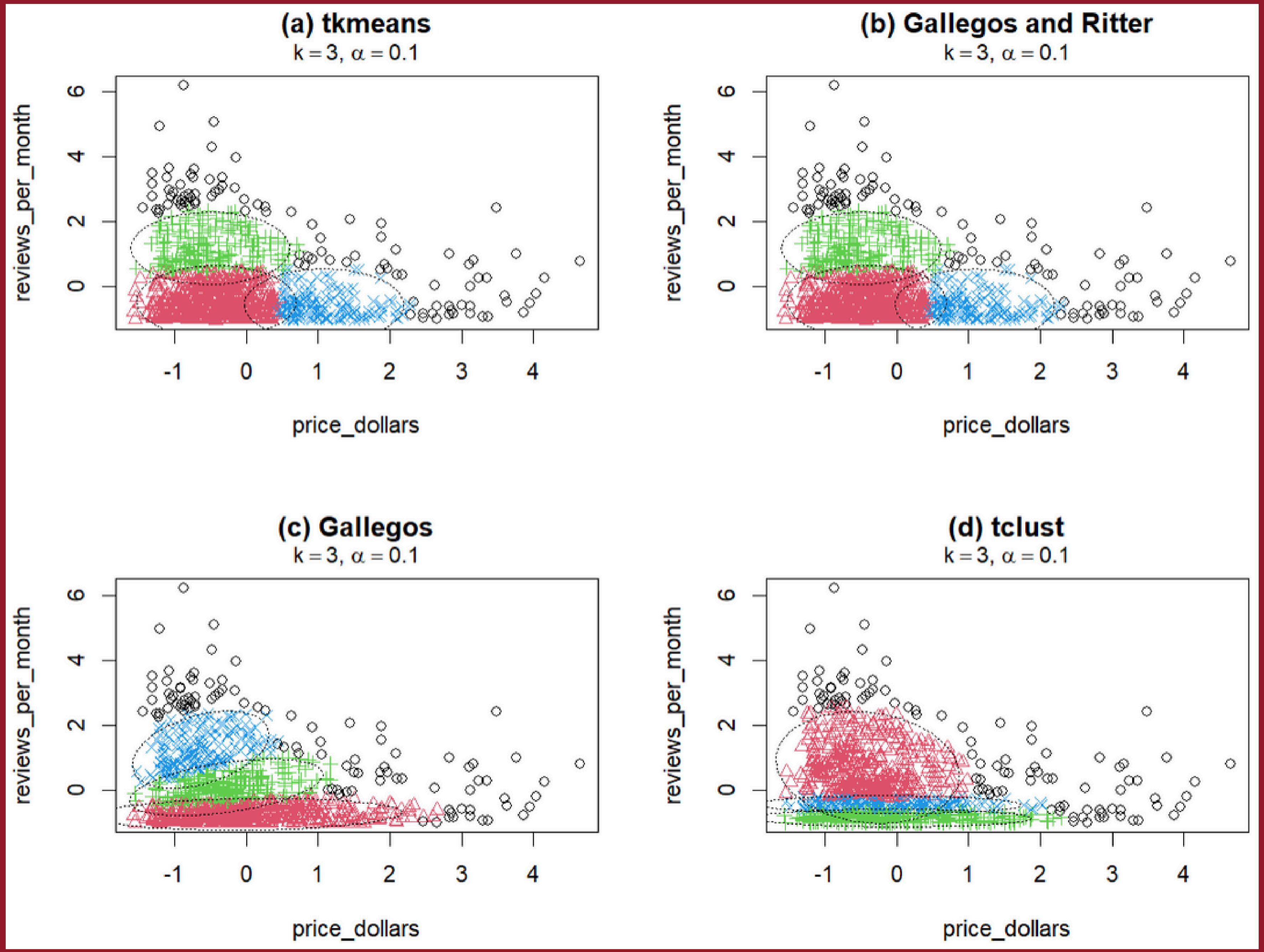




This method allows for the estimation of both the number of clusters and their structure using techniques like expectation-maximization.

Note that, in the uncertainty plot, larger symbols indicate the more uncertain observations.

Robust Clustering



Outlying data can heavily influence standard clustering methods. Hence the development of feasible robust model-based clustering approaches. Instead of trying to "fit" noisy data, a proportion α of the most outlying observations is trimmed.

Tkmeans is a simple modification of the K-Means algorithm with minimal computational overhead.

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.

- Market Segmentation: Ensures reliable grouping of customer data, even with missing or anomalous entries.
- Cybersecurity: Helps in anomaly detection for identifying malicious activities in network data.

Final remarks

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 detecting 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.

Thank you