**Amsterdam AirBnB Insights**

*How can AirBnB hosts get the most out of their listing?*

**Description of Project Goals**

The dataset that we used for this project has data on AirBnB listings in Amsterdam and was sourced from publicly available information from the Airbnb site. It has 16724 rows of data with 16 variables, described below:

* ID: Airbnb's unique identifier for the listing
* Name: Name of the listing
* Host ID: Airbnb's unique identifier for the host/user
* Host Name: Name of the host. Usually just the first name(s)
* Neighborhood Group: Aggregation of small neighborhoods
* Neighborhood: Neighborhood corresponding to the listing
* Latitude: Uses the World Geodetic System (WGS84) projection
* Longitude: Uses the World Geodetic System (WGS84) projection
* Room Type: Entire apt, Private room, Shared room/Hotel
* Price: Daily price in local currency
* Minimum Nights: Minimum number of night stay for the listing
* Reviews per Month: The number of reviews the listing has over the lifetime of the listing
* Availability 365: The availability of the listing 365 days in the future as determined by the calendar
* Calculated Host Listings: The number of listings the host has in the region
* Last Review: The date of the last/newest review
* Number of Reviews: The number of reviews the listing has

We came to this dataset with the main goal of predicting price. Additionally, while we were doing exploratory analysis we found a rule in Amsterdam that when renting out private rooms, owners cannot rent out more than 40% of the total space. With this rule, we wanted to find potential law breakers.

**Importance of Problem**

The importance of pricing is clear; someone putting a listing on Airbnb needs to properly price their listing in order to reduce potential loss in revenue (1) due to underpricing or (2) due to overpricing which could lead to reduced bookings. The second goal, finding potential law breakers, is important because the Airbnb market should only have rule abiding listings.

**Exploratory Analysis**

The first, and what seems to be the most important insight from our exploratory analysis is that the price of a listling is largely driven by the room type of the listing (fig 1). In order to explore the data better, we split the dataset into two: entire hope/apt listings and all others (hotel room, private room, and shared apartment).

When plotting the price against the availability of the listing, there did not seem to be any trend, even faceting on room type (fig 2).

For entire apartments, minimum nights at a listing seem to have a positive association with price, while for rooms there is no clear trend (fig 3).

When plotting mean price by neighborhood (faceting on room type) we can see that no matter what the room type is, there is a trend that listings in the same neighborhood have similar pricing (figs 4 & 5). This led us to want to answer a question; why are some neighborhoods priced more than others? This led us to research some possible reasons, and we hypothesized two plausible reasons for the same.

Fig. 6 & 7 show heatmap for **mean price per neighborhood** and **safety score per neighborhood** respectively. Green regions on both the heat maps correspond to higher mean price and higher safety while red regions correspond to lower mean price and lower safety. We can see from fig. 6 that **centrally located neighborhoods have higher mean prices**, which is generally the case for any given city. Additionally, if we look at the overlap of both the heat maps (fig. 8), it's clearly visible that **neighborhoods with higher safety scores have higher mean prices** which logically makes sense because people would not be willing to pay much for a listing in “not so safe” locality.

On another note, Amsterdam became the first city to issue new regulations to allow home sharing, in Feb’14. For Amsterdam, it was a matter of making sure that the historic city did not become “Venice, or Florence, or ‘Disneyland’”; that it wasn’t overrun by visitors and that locals weren’t crowded out. Hence, they laid out a myriad of rules to control the tourist influx into the city. One such rule is -

“When renting out Private rooms, owners cannot rent out more than 40% of the total space (in sq. meters)”

This was one of the rules that we could vet through our data and identify any potentially illegal listings. We checked for hosts with multiple private room listings in the same neighborhood since this could potentially mean hosts are separately listing out multiple rooms of the same apartment thus rendering the listings illegal. Fig. 9 shows a table containing information around hosts, neighborhood, room type and listing count; we can clearly identify a few hosts who have a lot of private room listings in the same neighborhood which is a bit unlikely and phishy.

**Solutions and Insights**

*Random Forests*

As discussed above, the model is divided based on the room type: Entire House/Apartment and Private room. Trying to predict the values based on a single tree could possibly lead to overfitting the data and high variance error for test RMSE. Hence, the random forests model was considered as it is an ensemble model and gives the flexibility of selecting the number of variables in each tree, which can reduce variance of irrelevant features.

Initially, the model was run including all the features and training and test RMSE were computed based on cross validation of the number of trees used in the random forest and the number of features used in each tree. The training and test RMSE were found to be similar and hence overfitting of the data can be ruled out.

The test RMSE for the model of the entire House/Apartment is 51 dollars, while for the model of private room it is 40 dollars. The corresponding R-squared values are evaluated to 10.39 % and 22.85% respectively. The variable importance is plotted in figure 11 (Similar for both the models)

Based on figure 11, it can be inferred that only top five features may be relevant for the model. With the parameters for the best test RMSE, the models were rerun with only top 5 features. The test RMSE for the model of the entire House/Apartment is 52 dollars, while for the model of private room it is 42 dollars. The corresponding R-squared values are evaluated to 9.16 % and 15.93% respectively.

With just top five features, the test RMSE is approximately similar to the model with all the features included. However, the R-squared value dipped by 5% for the model of private room. The residual errors as a function of price is shown in figure 12. It can be inferred that the model is over predicting the lower prices and under predicting the higher prices

*Linear Regression*

Important Predictors: Room Type, Neighbourhood, Number of reviews, Minimum nights

Key Model Features:Train RMSE - 51.51, Test RMSE - 51.53, Cross Validation RMSE - 51.53

Interpretation: Higher the minimum nights, lesser the price. The prices are very low for shared rooms. Neighbourhood centrum west has higher prices

*Ridge Regression*

Important Predictors**:** Split to two models based on room type

Other important features - Neighbourhood, Number of reviews, Minimum nights

Key Model Features:

Entire Home/Apt (fig 13): Train RMSE - 53.68, Test RMSE - 54.02, Cross Validation RMSE - 53.07 (alpha = 0.89)

Private Room (fig 14):Train RMSE - 39.96, Test RMSE - 42.55, Cross Validation RMSE - 38.64 (alpha = 0.95)

*K - Nearest Neighbors*

To obtain the ideal “k” value, a plot between Price RMSE and K was created to identify the k-value corresponding to the least RMSE (fig. 10). RMSE was obtained by averaging the RMSE obtained from a 3-fold cross validation.

Predictors used for modeling: Neighborhood, Room type, Latitude, Longitude

All other predictors present in the data did not yield any improvement in the RMSE. This model predicted the price with an RMSE of 50 for k = 35.

*Business Impact*

| **Metric** | **Calculation** |
| --- | --- |
| Predicted price of the listing | x = y\_pred - RMSE  (removing RMSE for conservative approach) |
| $x aggregated for listings with  their actual price < x  (potentially underpriced listings) | $11,865  (underpriced listings in test data - revenue) |
| # Potentially underpriced listings | 654  (unique low priced listings in test data) |
| Profit per underpriced listing per day | $18.14 |
| Average occupancy days per underpriced listing per month | 18  (from average reviews per month and minimum nights) |
| Average occupancy days per underpriced listing per year | 216 |
| Profit per year per underpriced listing | $3919  (occupancy days per year \* profit per listing) |
| Assuming 50 % is the true profit | $1959 (Removing some additional costs) |

**The model drives ~$2000 per year per listing if prices are optimized**

**Figures**

Figure 1

Chart, box and whisker chart

Description automatically generated

Figure 2

Chart, scatter chart

Description automatically generated

Figure 3

Chart, scatter chart

Description automatically generated

Figure 4: Mean Price by Neighborhood for Entire Apt.

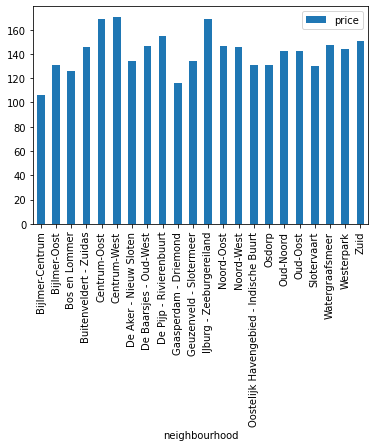


Figure 5: Mean Price by Neighborhood for rooms

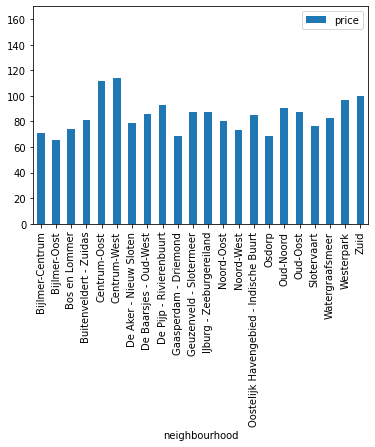


Figure 6 : Neighborhood heat map by mean price

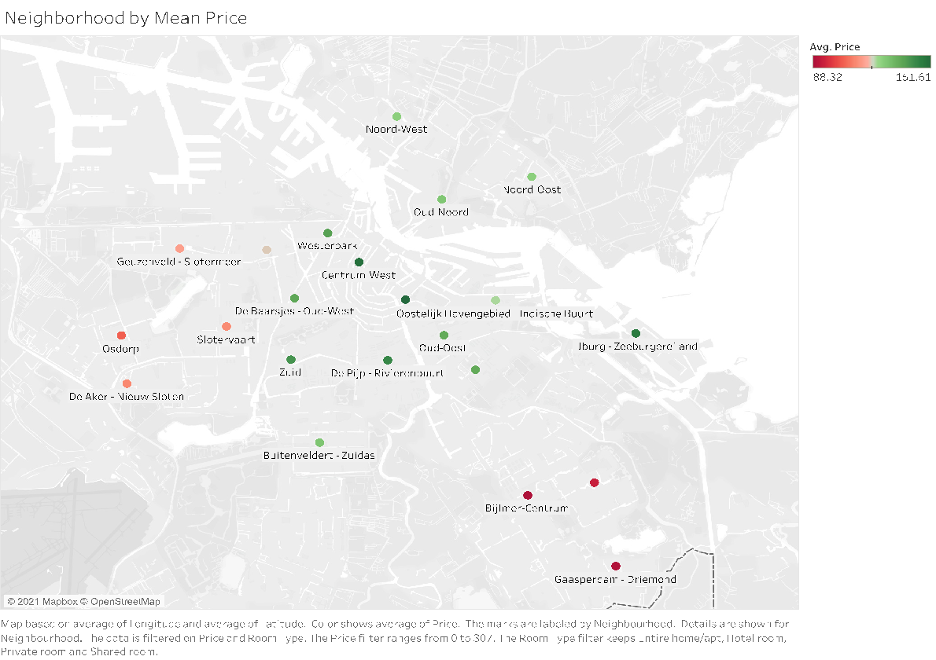


Figure 7 : Neighborhood heat map by safety score

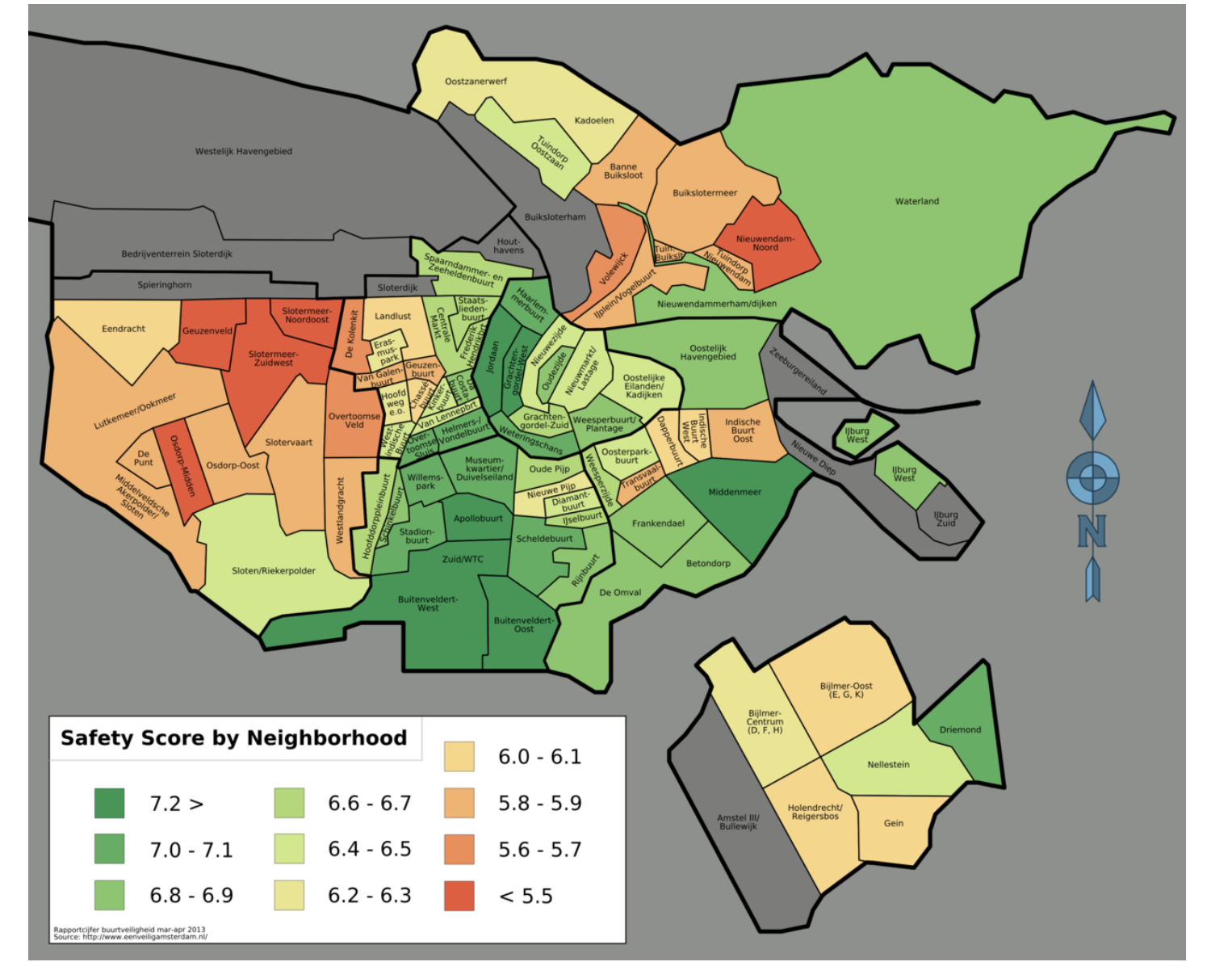


Figure 8 : Overlap of neighborhood heat maps by mean price and safety

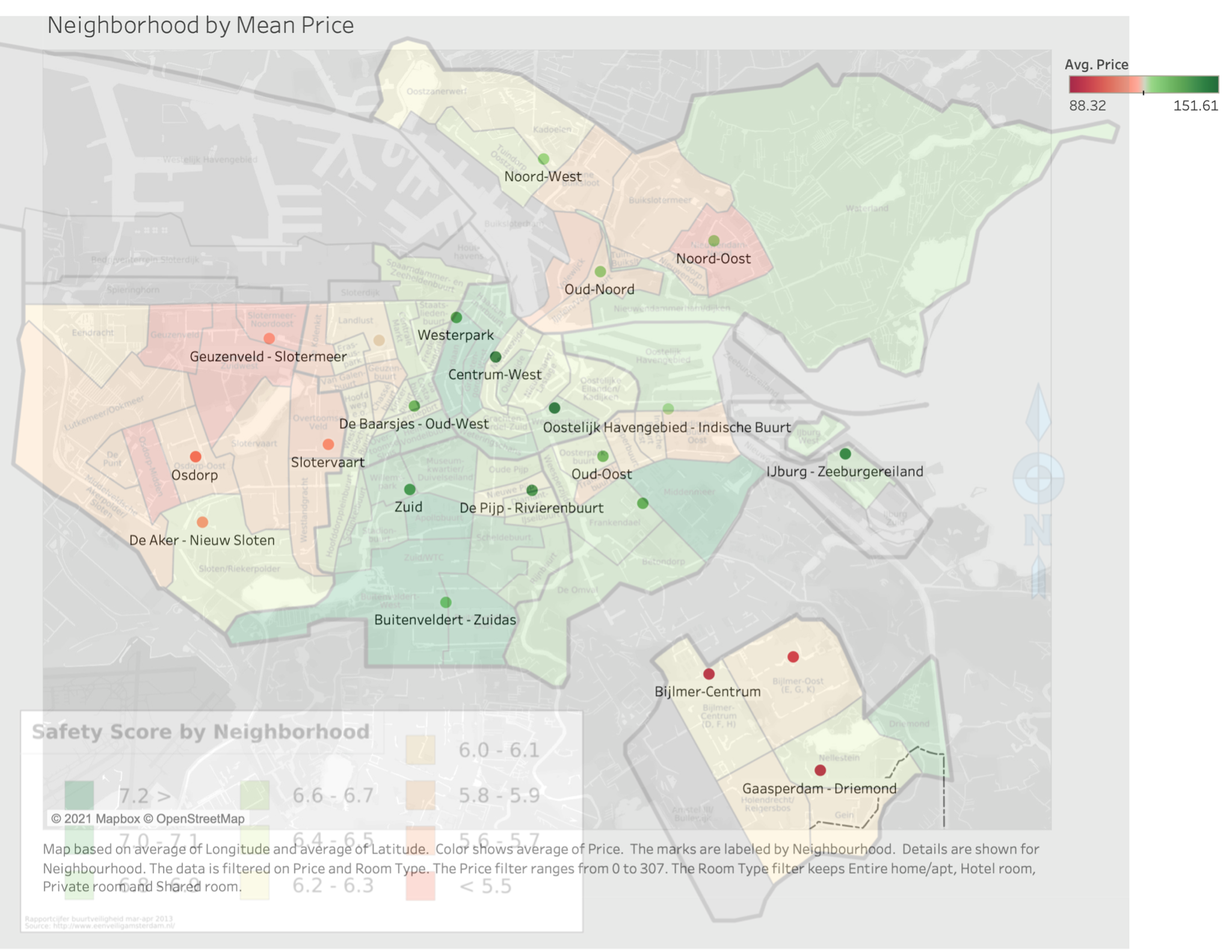


Figure 9 : Table identifying potentially illegal listings

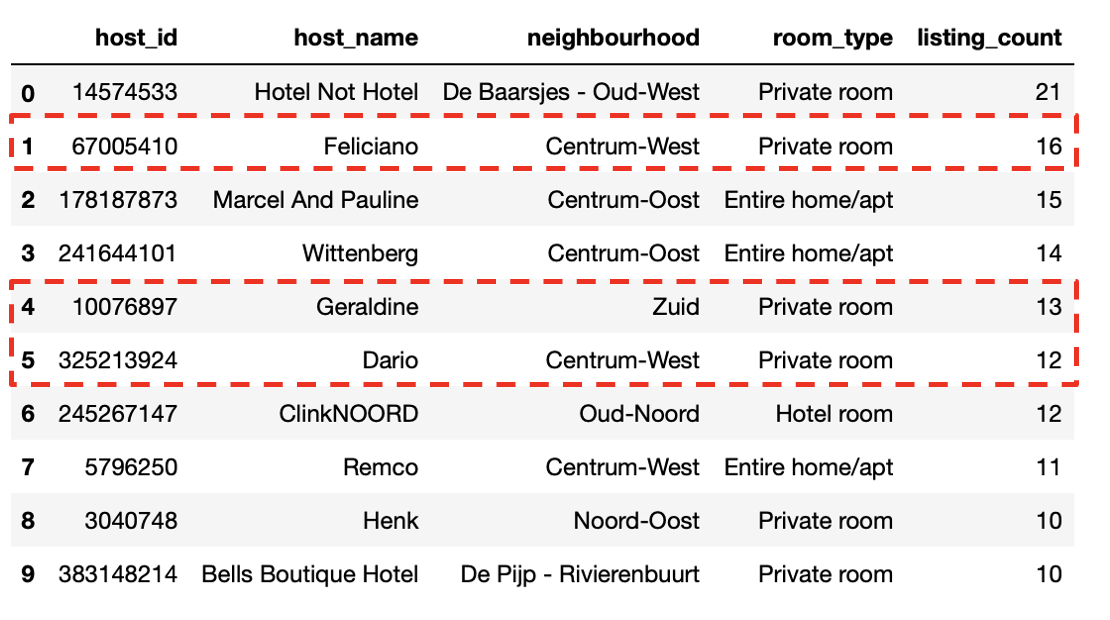


Figure 10 : KNN “K” vs RMSE

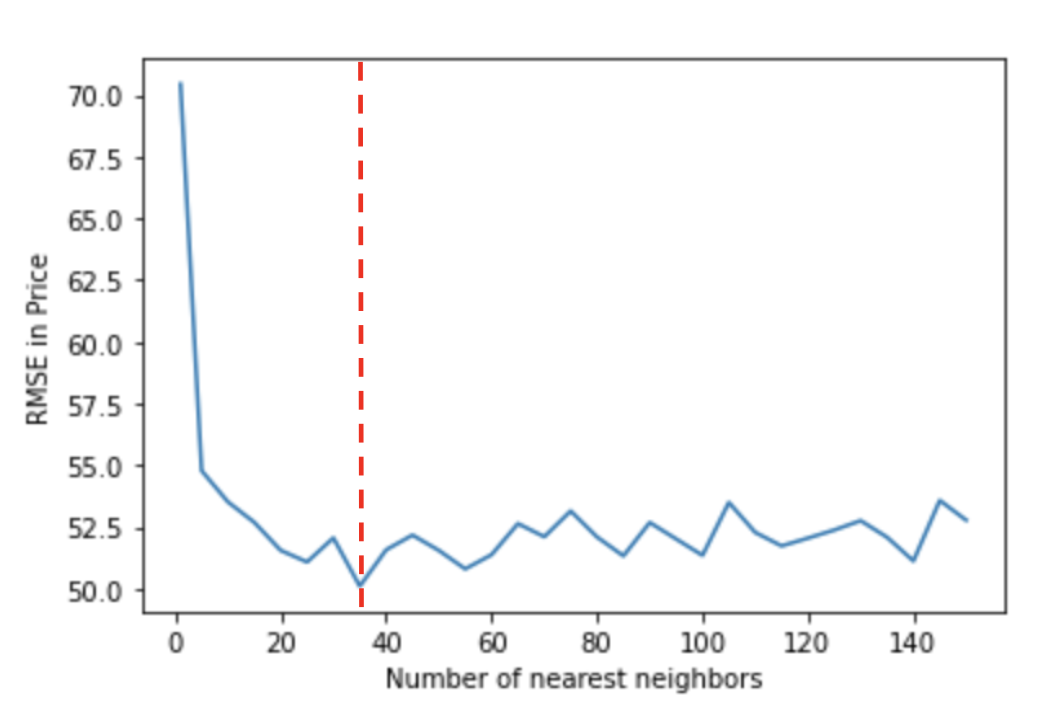


Figure 11

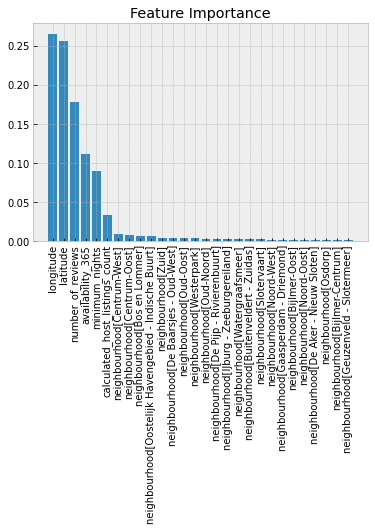


Figure 12

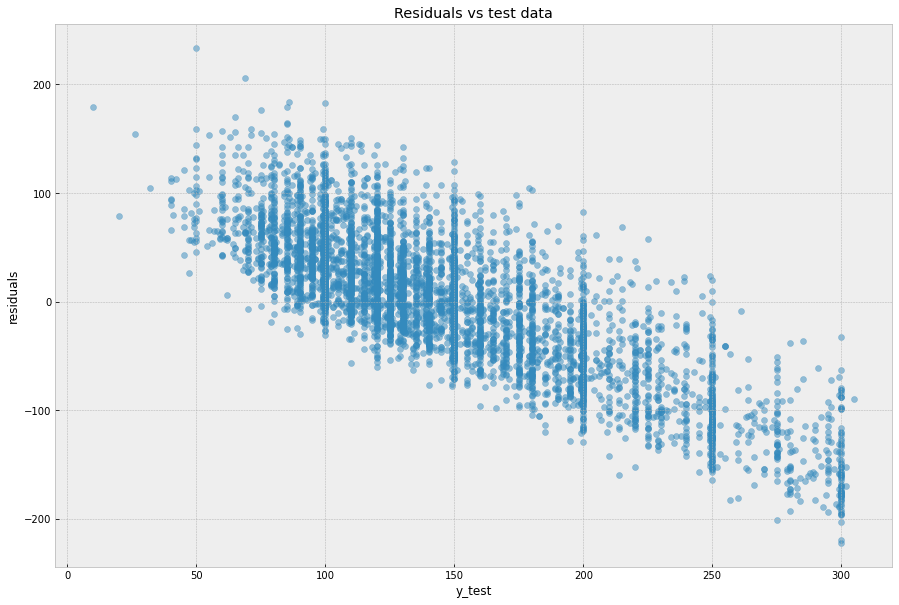


Figure 13

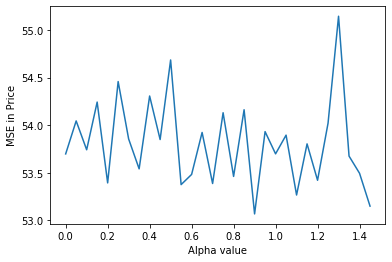
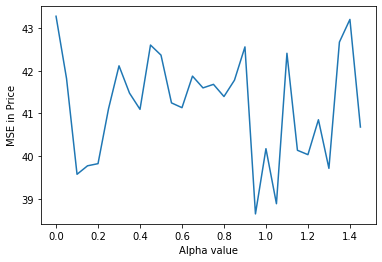
****

Figure 14



**Sources**

AirBnB data source:

<http://insideairbnb.com/get-the-data.html>

Safety data source: <https://commons.wikimedia.org/wiki/File:Amsterdam-unsafety-map.png>

Amsterdam Room Restrictions:

<https://www.forbes.com/sites/ceciliarodriguez/2018/05/17/amsterdam-overwhelmed-by-tourists-gets-tough-on-airbnb-tourists-taxes-too-much-partying/?sh=a9c2b3b2be59>