August 9, 2021

The University of Texas at Austin
McCombs School of Business

Amsterdam Airbnb Insights and Visualizations

How can hosts make sure to get the best out of their listing?

Avery Shepherd
Jacob Pammer
Kaushik Kumaran
Surya Pamireddygari
Vishu Agarwal

The Problem

airbnb

Steps to a solution

1) Data Cleansing
2) Exploratory Data Analysis
3) Models
4) Business Impact

Data: Amsterdam AirBnB Listings

Listing data from the July 2020-July 2021
Sourced from AirBnB's open data source
16724 rows of data with 16 variables

Variables	Description
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 for latitude and longitude.
Calculated Host Listings Count	The number of listings the host has in the region

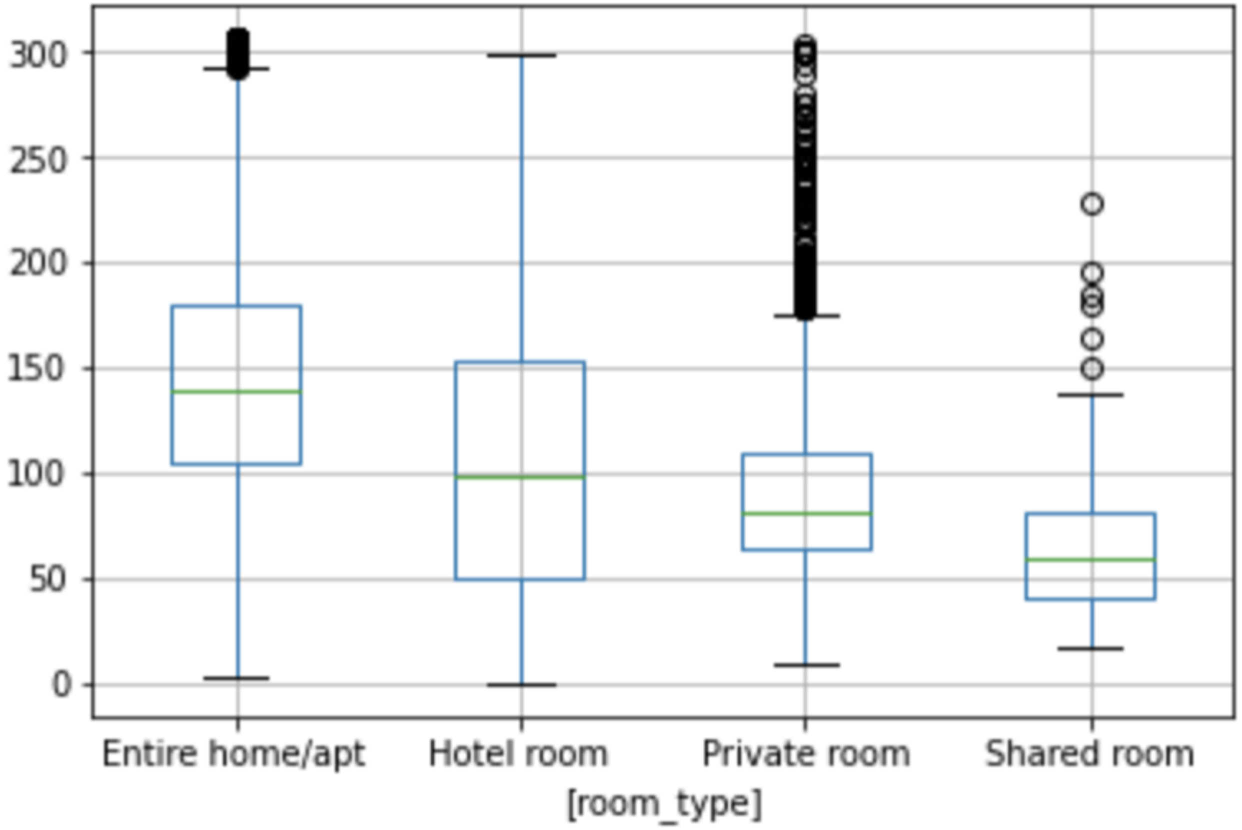
Variables	Description
Longitude	Uses the World Geodetic System (WGS84) projection for latitude and longitude.
Room Type	Entire apt, Private room, Shared room/Hotel
Price	Daily price in local currency
Minimum Nights	Minimum number of night stay for the listing (calendar rules may be different)
Reviews 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.
Last Review	The date of the last/newest review
Number of Reviews	The number of reviews the listing has

Data Cleaning

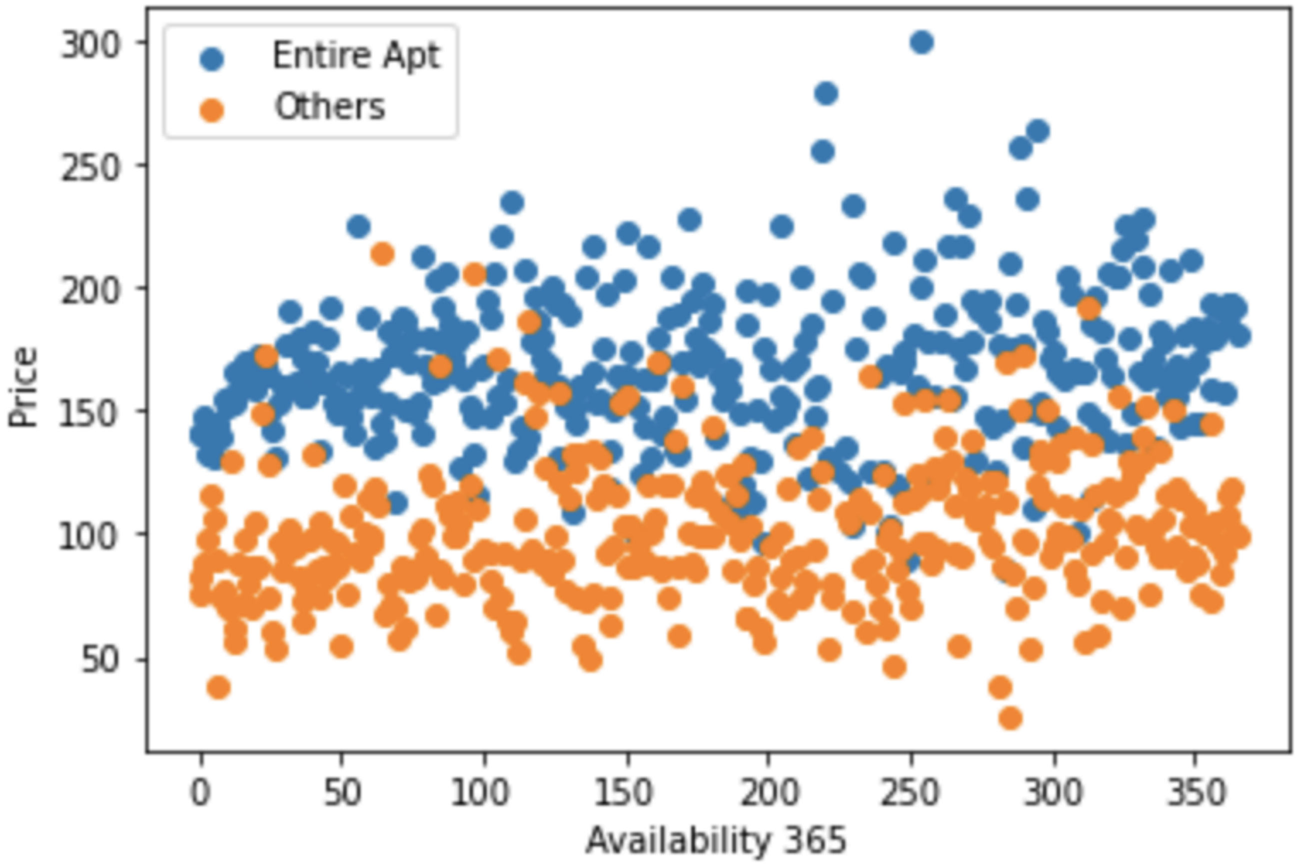
Variable	Problem	Treatment
Reviews per month	Missing values	Replace with 0
Neighborhood Group	Whole column is blank	Remove the column
Price	Presence of outliers	IQR outlier treatment Lower bound : Q1 – 1.5 * IQR Upper bound : Q3 + 1.5 * IQR

EXPLORATORY DATA ANALYSIS

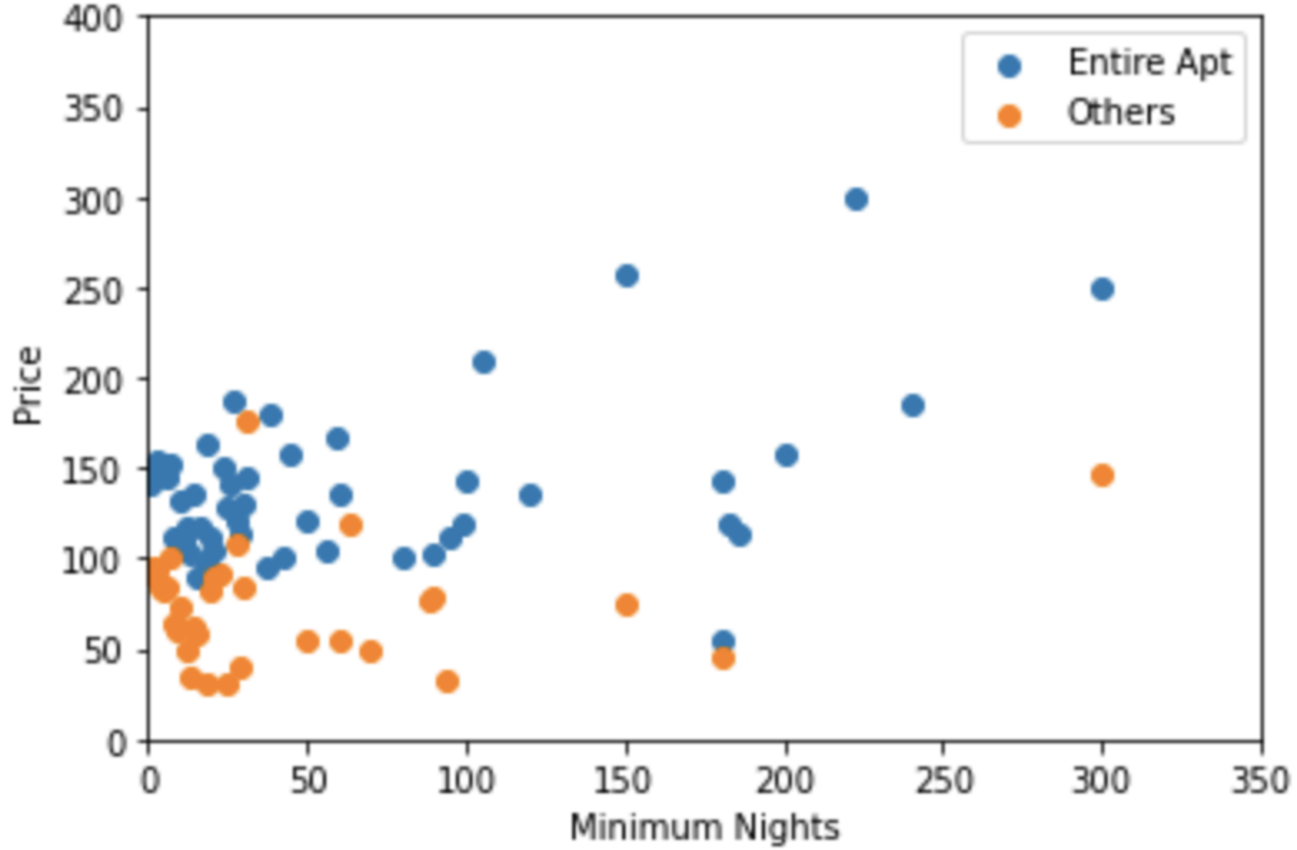
Listing Price seems to be largely driven by room type



Future Availability does not seem to affect the price

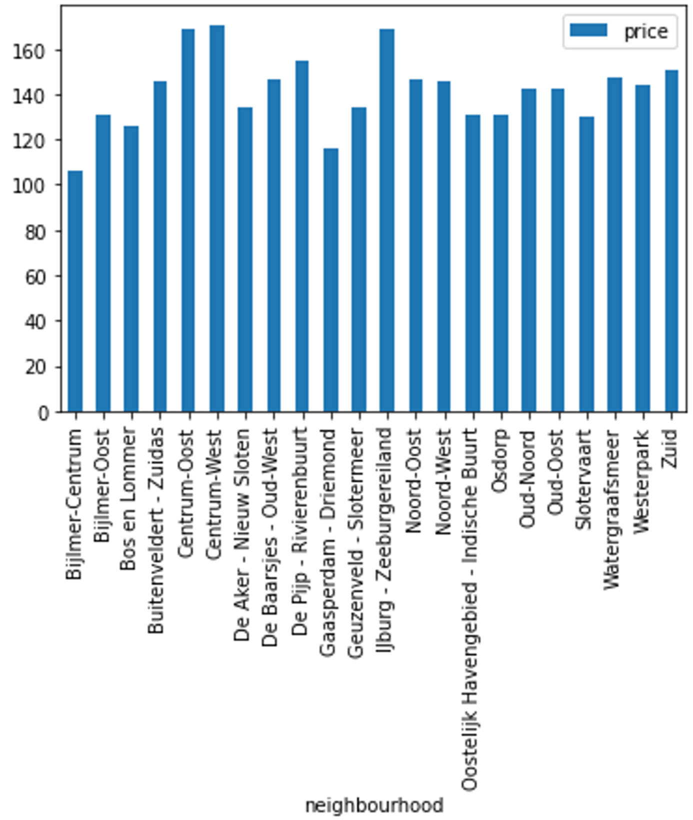


For entire apartments, minimum nights at a listing seem to have a positive association with price

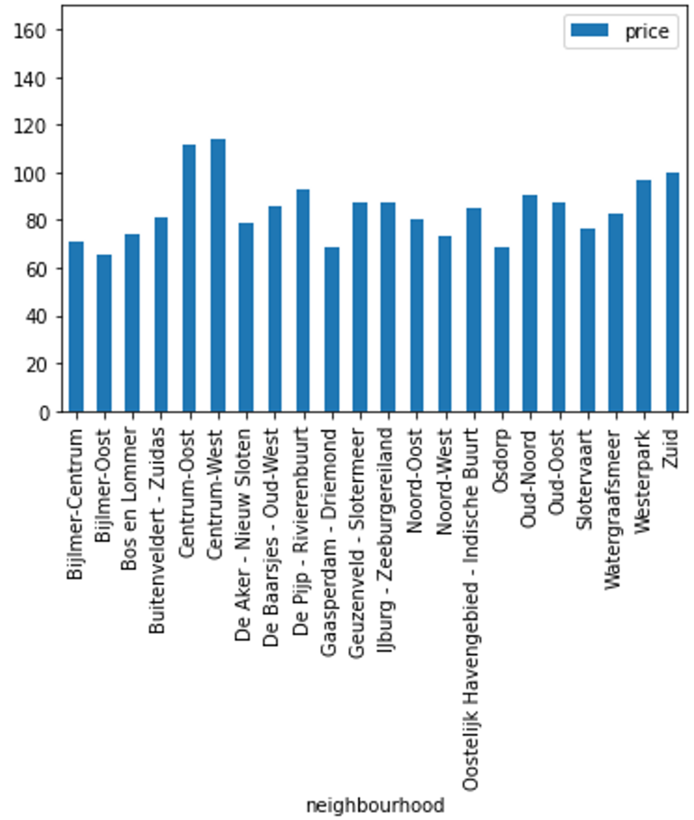


Prices are affected by neighborhood, across homes and rooms

Mean Price by Neighborhood for Entire Apt.



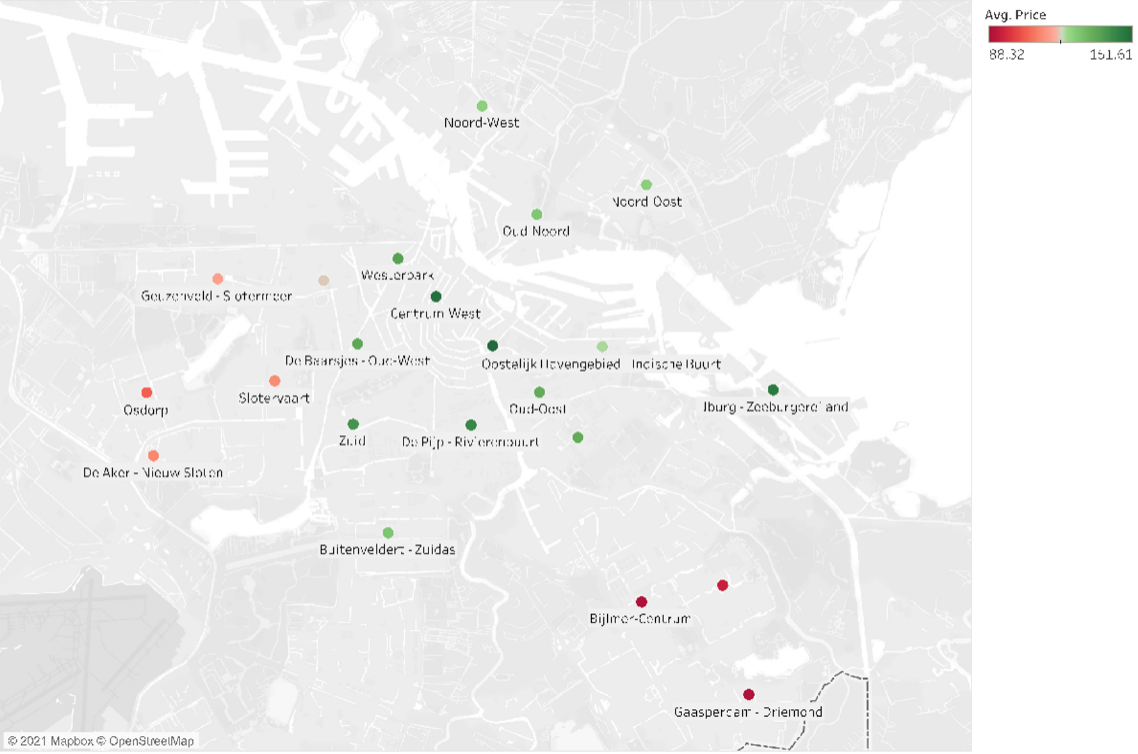
Mean Price by Neighborhood for Private Room



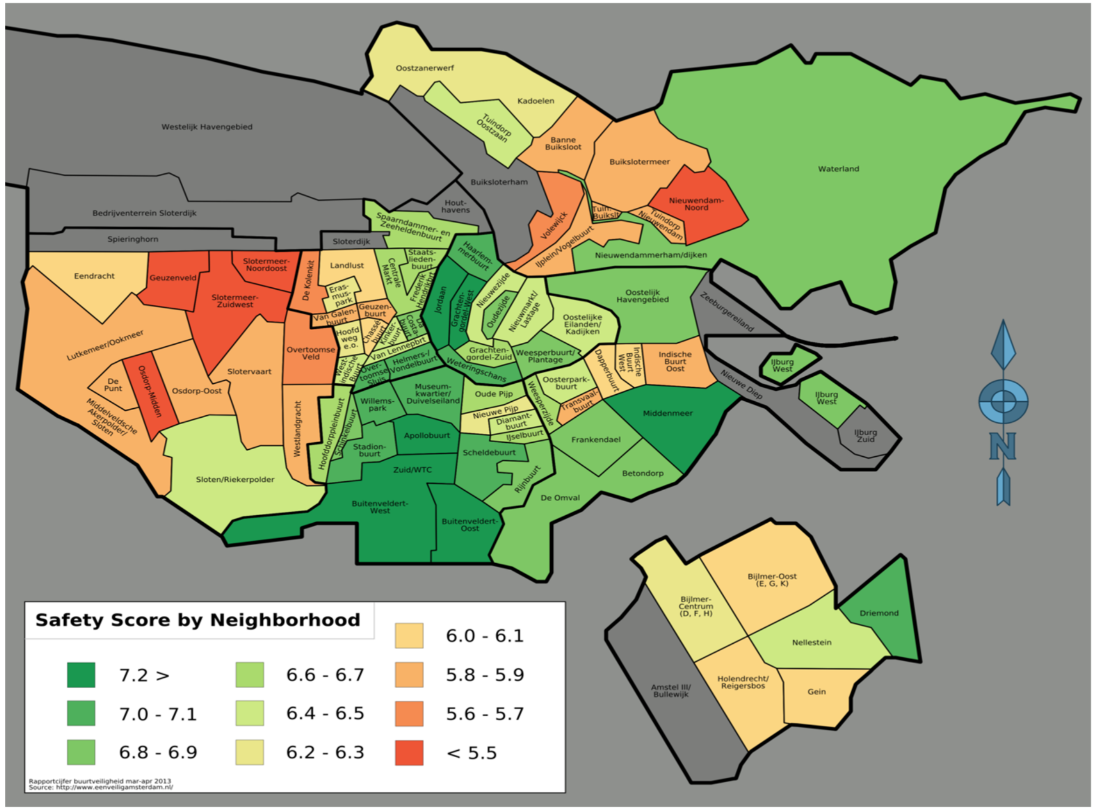
Explaining neighborhood price differentiation with safety

Mean Price Heatmap by Neighborhood

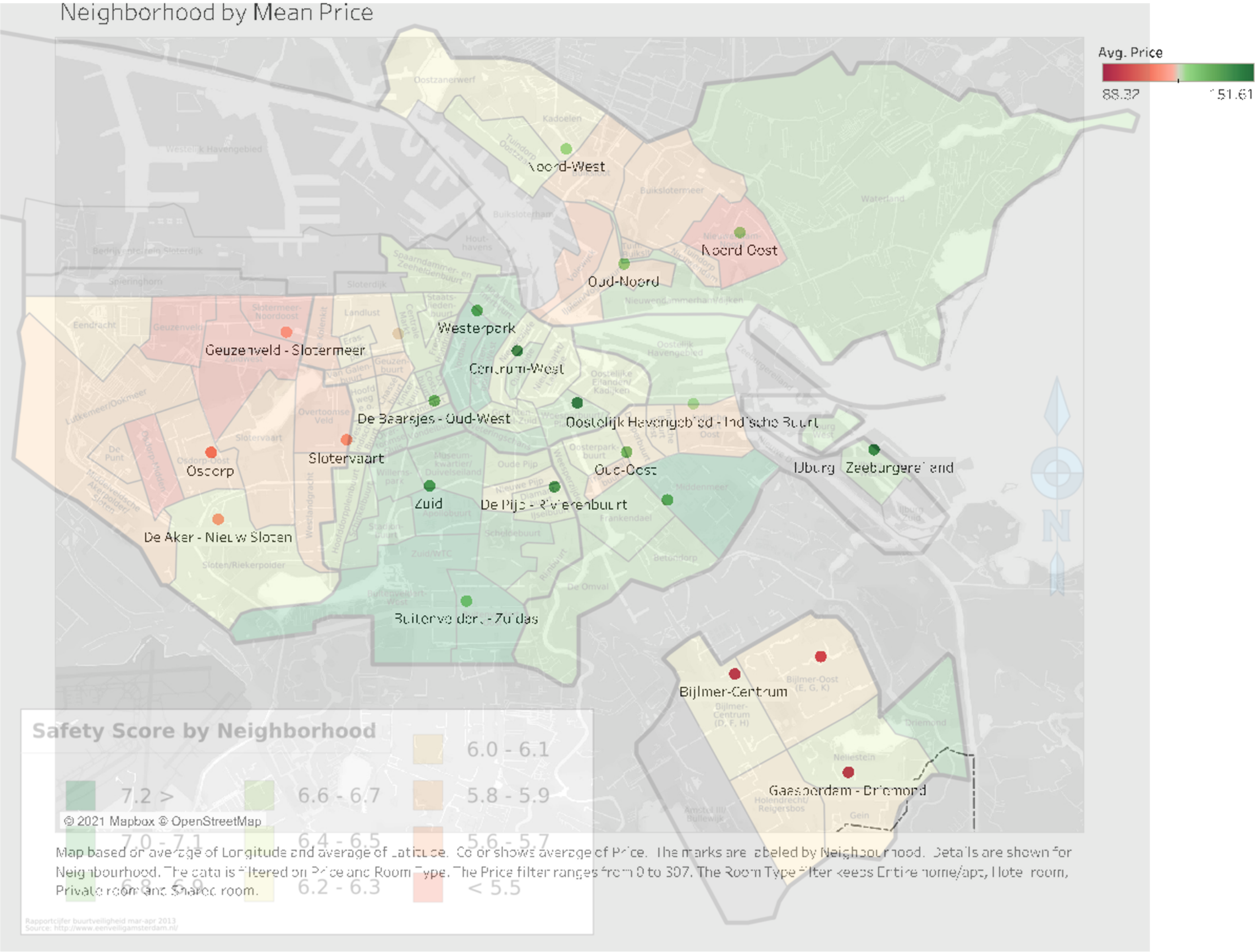
Neighborhood by Mean Price



Safety Heatmap by Neighborhood



Neighborhoods with higher safety score seem to have higher mean price



Potential Illegal Listings

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

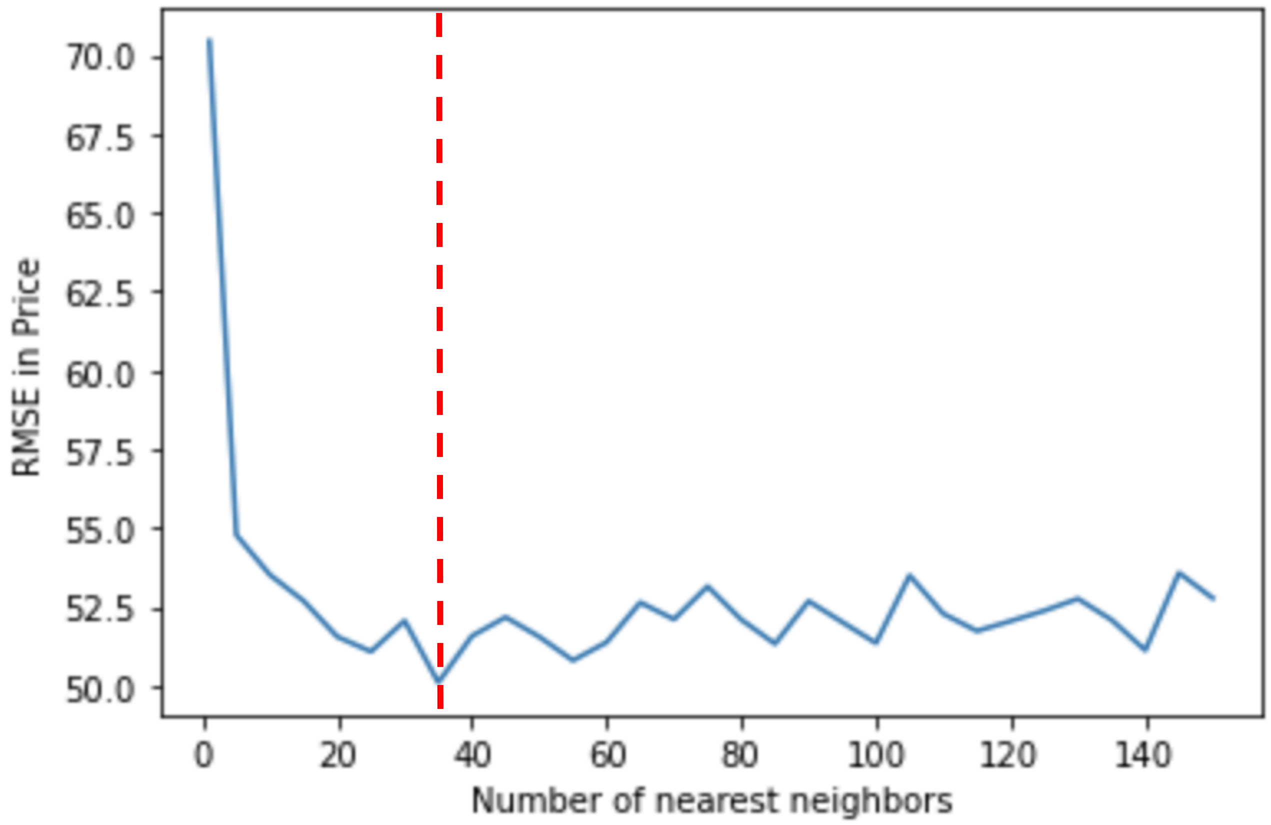
Potential illegal listings on Amsterdam Airbnb

	host_id	host_name	neighbourhood	room_type	listing_count
0	14574533	Hotel Not Hotel	De Baarsjes - Oud-West	Private room	21
1	67005410	Feliciano	Centrum-West	Private room	16
2	178187873	Marcel And Pauline	Centrum-Oost	Entire home/apt	15
3	241644101	Wittenberg	Centrum-Oost	Entire home/apt	14
4	10076897	Geraldine	Zuid	Private room	13
5	325213924	Dario	Centrum-West	Private room	12
6	245267147	ClinkNOORD	Oud-Noord	Hotel room	12
7	5796250	Remco	Centrum-West	Entire home/apt	11
8	3040748	Henk	Noord-Oost	Private room	10
9	383148214	Bells Boutique Hotel	De Pijp - Rivierenbuurt	Private room	10

MODELING

k - nearest neighbors

Running a k-fold cross validation yields least Price RMSE of 50.1 at k = 35



Predictors used for modeling :

- Neighborhood
- Room Type
- Latitude
- Longitude

Random Forest

Random Forest yields a RMSE score of 52 for Entire Apt and 42 for Private Room

Important Predictors

Longitude
Latitude
Number of reviews
availability_365
Minimum nights

Remarks

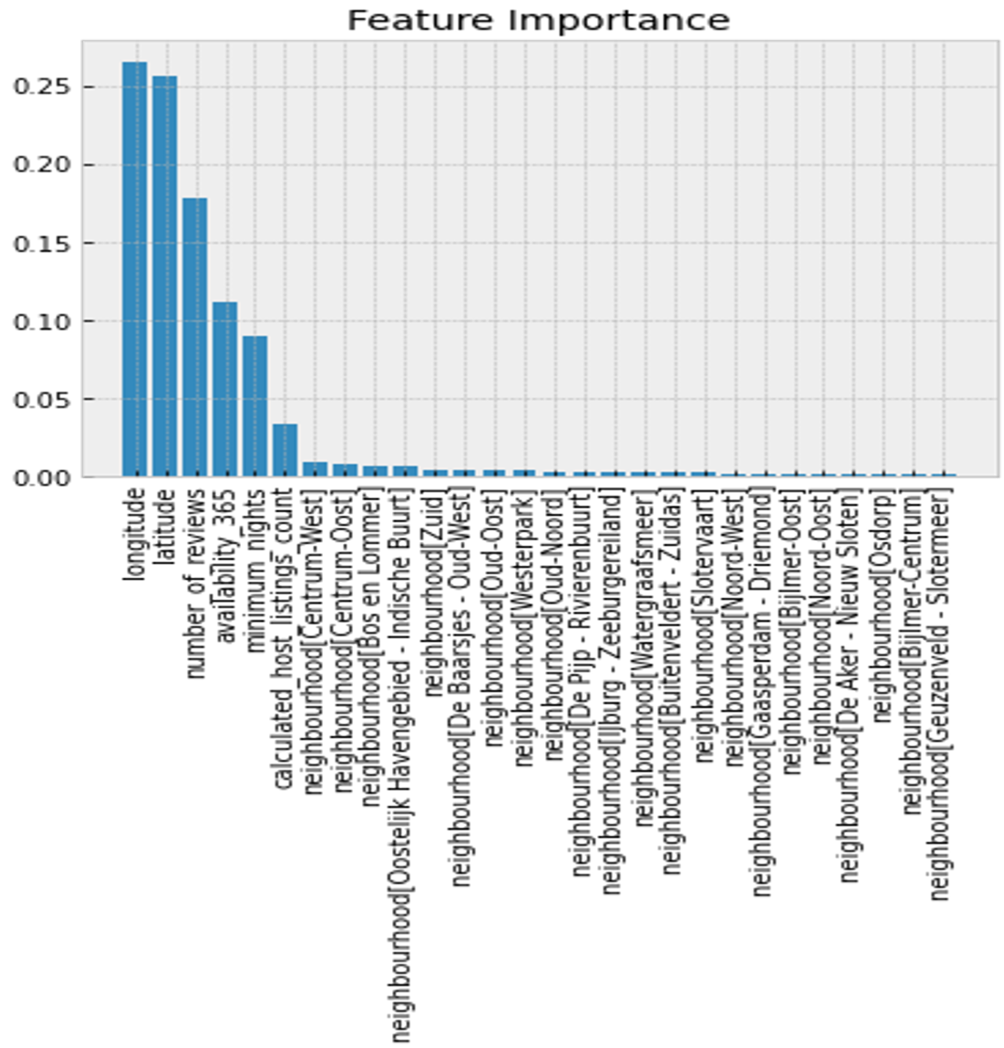
- The prediction is better for private room type than entire apartment/house. However the private room has lower prices
- The model is under estimating at higher prices and over estimating at lower prices

Key Model Features

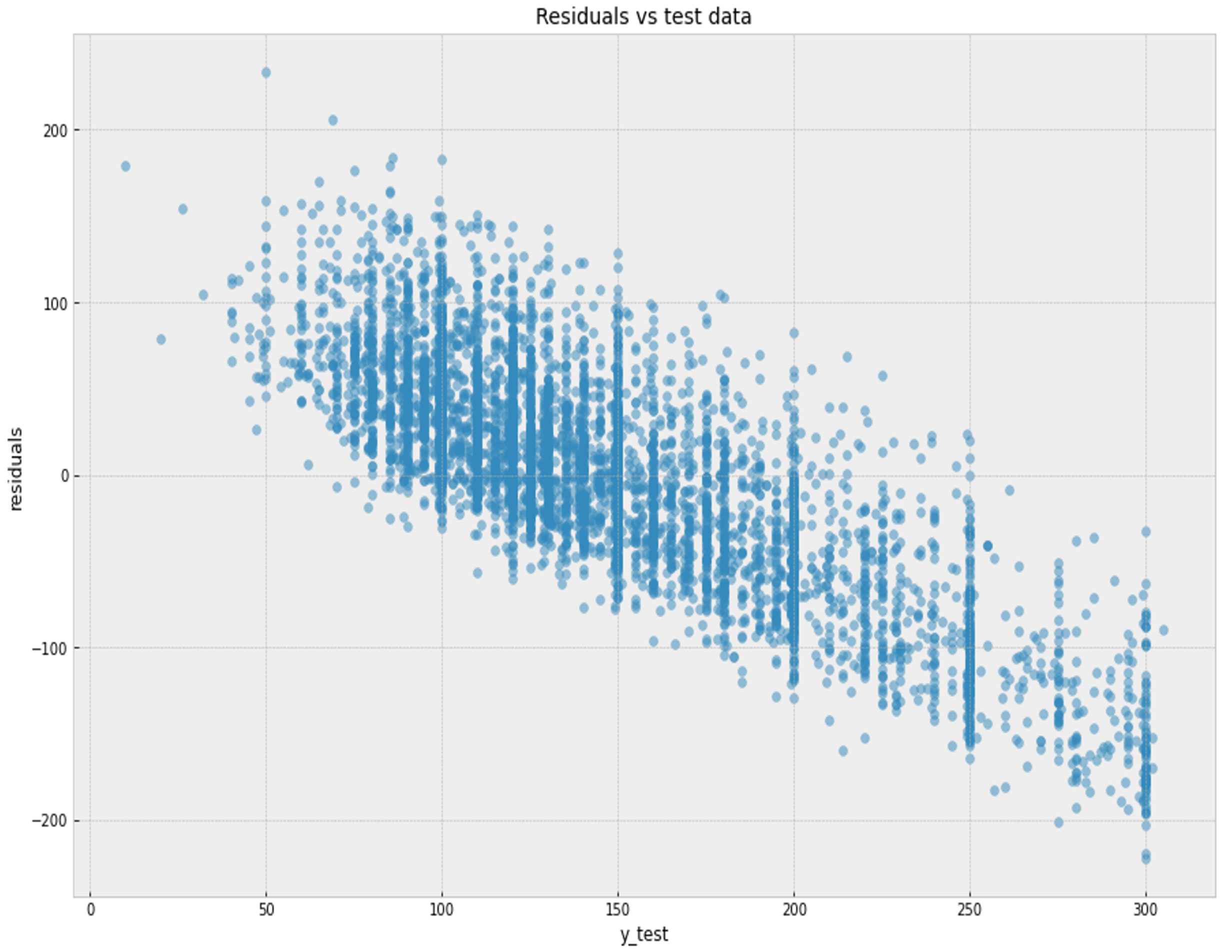
Entire Home/Apt
Train RMSE - 50
Test RMSE - 52

Private Room
Train RMSE - 39
Test RMSE - 42

Important Predictors



Plot of Residual error Vs Prices



Linear Regression

Linear Regression yields a RMSE score of 51.5

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

Remarks

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

Ridge Regression

Ridge Regression yields a RMSE score of 53 for Entire Apt and 42 for Private Room

Important Predictors

Split the model based on room type
Neighbourhood
Number of reviews
Minimum nights

Remarks

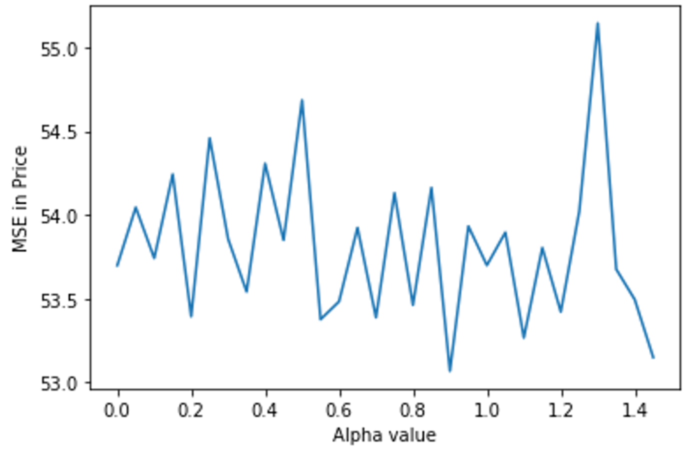
- Provide a grid for alpha values
- Choose the model with least cross validation rmse
- Repeat the same for entire home/apt and private room

Key Model Features

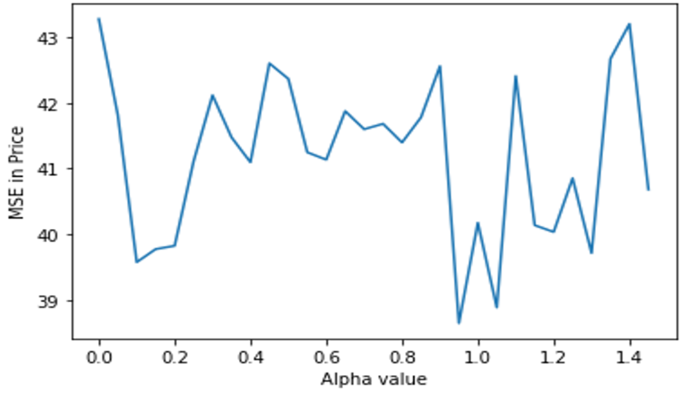
Entire Home/Apt
Train RMSE - 53.68
Test RMSE - 54.02
Cross Validation RMSE - 53.07
(alpha = 0.89)

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

Entire Home / Apt



Private Room



BUSINESS IMPACT

Modeling 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

QUESTIONS?