Evaluating Rental Prices in Utrecht City

1. Introduction

1.1. Background

Nowadays rental houses is a business continuously growing in most of Europe, with a higher rate in cities where expats' market is predominant, thanks to presence of mainstream Universities and High Tech hubs.
Utrecht, the fourth largest city of the Netherlands located in the very center of the mainland, surely belongs to this category.
For this reason, with the purpose of getting - all expats and/or any person going to move - acquainted with the current scenario of rentals in the area, a well detailed analysis is carried out.

1.2. Purpose

This project's purpose is to give a clear overview of rental prices in Utrecht area and evaluate mutual influence of multiple variables such as neighborhoods, house surface and typology.

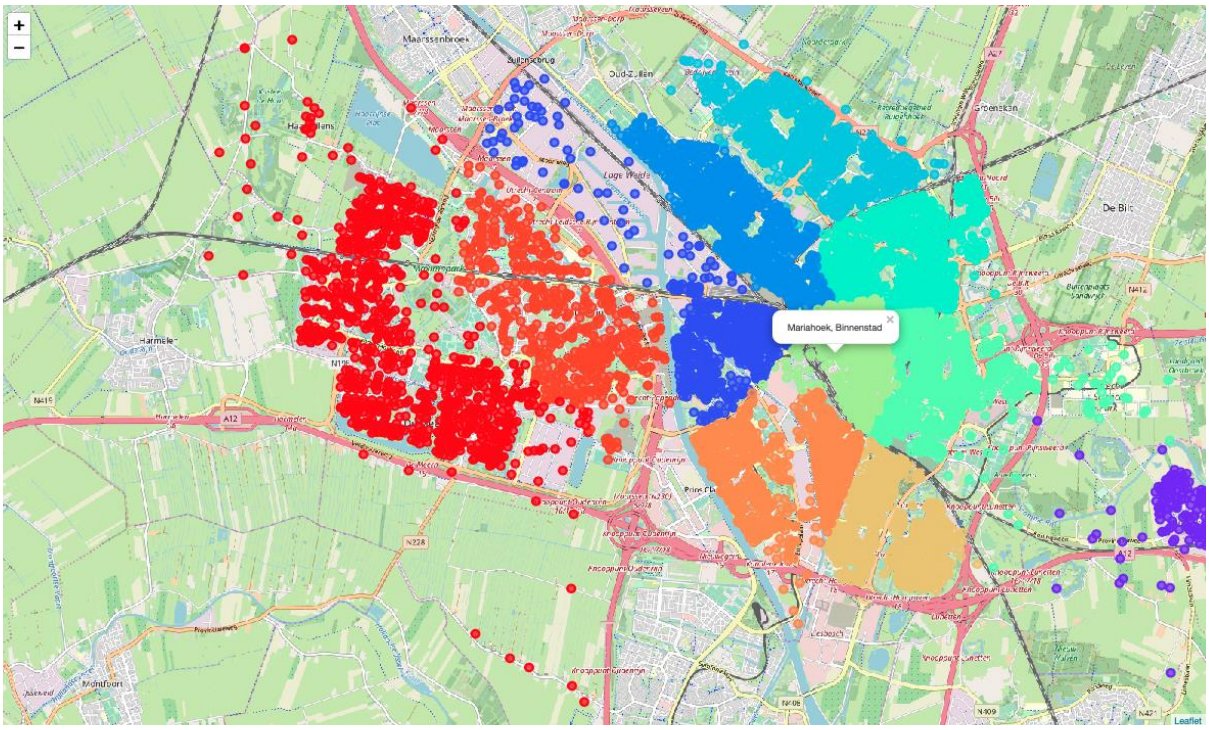
2. Data

2.1. Data source

Most dataset are quite recent and have been recovered from Kaggler here. Also, Foursquare API is used to get information related to the neighborhoods of the houses listed in the csv.

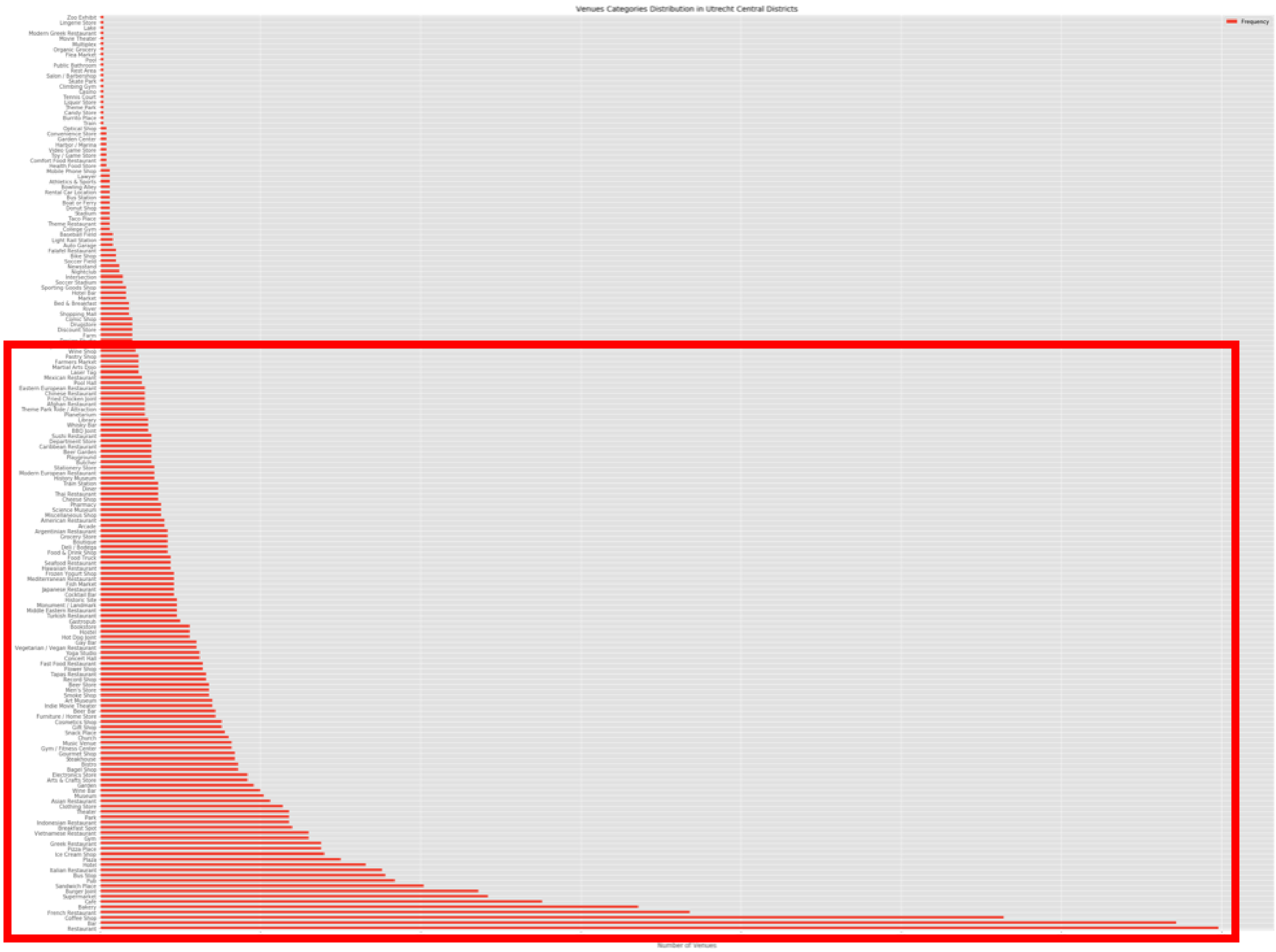
2.2. Data Wrangling

Data will be reported as dataframe from Jupyter Notebooks and then manipulated in order to get expected results.
Firstly, a plot of Utrecht map with Folium of dataset *Utrecht_postcode_v1* is performed in order to have a clear perspective of Utrecht geospatial coordinates and its districts' distribution.



Then, with reference to Utrecht central districts (Binnenstad, Noordoost, Zuidwest, Zuid, Noordoost) the dependent variable Rental Price is evaluated. Way of working adopted is Machine Learning method typically used for continuous values (i.e. **Multiple Linear Regression**), rental price vs following independent variables x_i:

- Influence of neighborhoods
    - the foursquare API is applied to the dataset *rental_central.csv*;
    - venues frequency is plotted with a bar chart


Venues Categories Distribution in Utrecht Central Districts
Frequency
Zoo Exhibit
Lingerie Store
Lake
Modern Greek Restaurant
Movie Theater
Multiplex
Organic Grocery
Flea Market
Pool
Public Bathroom
Rest Area
Salon / Barbershop
Skate Park
Climbing Gym
Casino
Tennis Court
Liquor Store
Theme Park
Candy Store
Burrito Place
Train
Optical Shop
Convenience Store
Garden Center
Harbor / Marina
Video Game Store
Toy / Game Store
Comfort Food Restaurant
Health Food Store
Mobile Phone Shop
Lawyer
Athletics & Sports
Bowling Alley
Rental Car Location
Bus Station
Boat or Ferry
Donut Shop
Stadium
Taco Place
Theme Restaurant
College Gym
Baseball Field
Light Rail Station
Auto Garage
Falafel Restaurant
Bike Shop
Soccer Field
Newsstand
Nightclub
Intersection
Soccer Stadium
Sporting Goods Shop
Hotel Bar
Market
Bed & Breakfast
River
Shopping Mall
Comic Shop
Drugstore
Discount Store
Farm
Wine Shop
Pastry Shop
Farmers Market
Martial Arts Dojo
Laser Tag
Mexican Restaurant
Pool Hall
Eastern European Restaurant
Chinese Restaurant
Fried Chicken Joint
Afghan Restaurant
Theme Park Ride / Attraction
Planetarium
Library
Whisky Bar
BBQ Joint
Sushi Restaurant
Department Store
Caribbean Restaurant
Beer Garden
Playground
Butcher
Stationery Store
Modern European Restaurant
History Museum
Tram Station
Diner
Thai Restaurant
Cheese Shop
Pharmacy
Science Museum
Miscellaneous Shop
American Restaurant
Arcade
Argentinian Restaurant
Grocery Store
Boutique
Deli / Bodega
Food & Drink Shop
Food Truck
Seafood Restaurant
Hawaiian Restaurant
Frozen Yogurt Shop
Mediterranean Restaurant
Fish Market
Japanese Restaurant
Cocktail Bar
Historic Site
Monument / Landmark
Middle Eastern Restaurant
Turkish Restaurant
Gastropub
Bookstore
Hostel
Hot Dog Joint
Gay Bar
Vegetarian / Vegan Restaurant
Yoga Studio
Concert Hall
Fast Food Restaurant
Flower Shop
Tapas Restaurant
Record Shop
Beer Store
Men's Store
Smoke Shop
Art Museum
Indie Movie Theater
Beer Bar
Furniture / Home Store
Cosmetics Shop
Gift Shop
Snack Place
Church
Music Venue
Gym / Fitness Center
Gourmet Shop
Steakhouse
Bistro
Bagel Shop
Electronics Store
Arts & Crafts Store
Garden
Wine Bar
Museum
Asian Restaurant
Clothing Store
Theater
Park
Indonesian Restaurant
Breakfast Spot
Vietnamese Restaurant
Gym
Greek Restaurant
Pizza Place
Ice Cream Shop
Plaza
Hotel
Italian Restaurant
Bus Stop
Pub
Sandwich Place
Burger Joint
Supermarket
Café
Bakery
French Restaurant
Coffee Shop
Bar
Restaurant
Number of Venues

  - given the amount of data (more than 5000 entries) the 10% of the entire dataset having less impact is not considered (data area out of red rectangule in previous picture). For the remaining, macro groups context based are generated

- Size of apartment

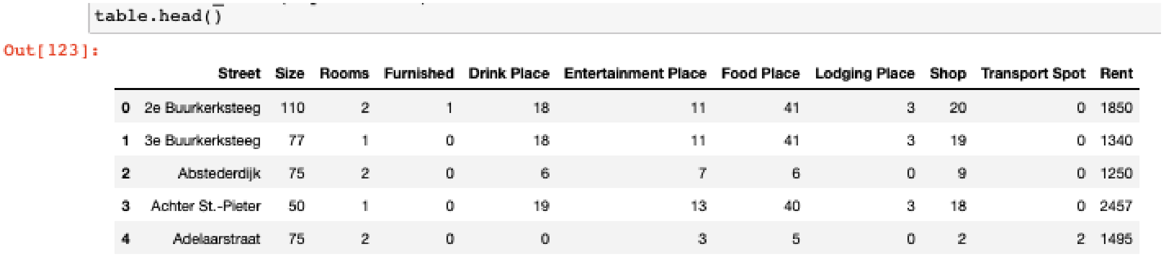
- Type of apartment

- Furnished

After the effect of each variable is evaluated, the x_i dataset will be splitted in training and testing data, the xi variable will be inserted in a Pipeline and the fitting model completely developed/evaluated.

3. Exploratory Data Analysis

Result of Data Wrangling is a single table containing all pertinent information per each (house in the) street



```
table.head()
```

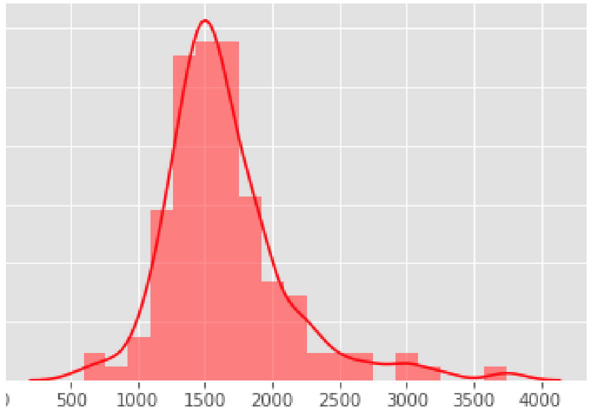
Out[123]:

	Street	Size	Rooms	Furnished	Drink Place	Entertainment Place	Food Place	Lodging Place	Shop	Transport Spot	Rent
0	2e Buurkerksteeg	110	2	1	18	11	41	3	20	0	1850
1	3e Buurkerksteeg	77	1	0	18	11	41	3	19	0	1340
2	Abstederdijk	75	2	0	6	7	6	0	9	0	1250
3	Achter St.-Pieter	50	1	0	19	13	40	3	18	0	2457
4	Adelaarstraat	75	2	0	0	3	5	0	2	2	1495

In the dataset most of houses are apartments, houses are few and contribution of Villa or Studio is negligible. Therefore, since this variable is not producing significative variation on price, it is not included in the model.

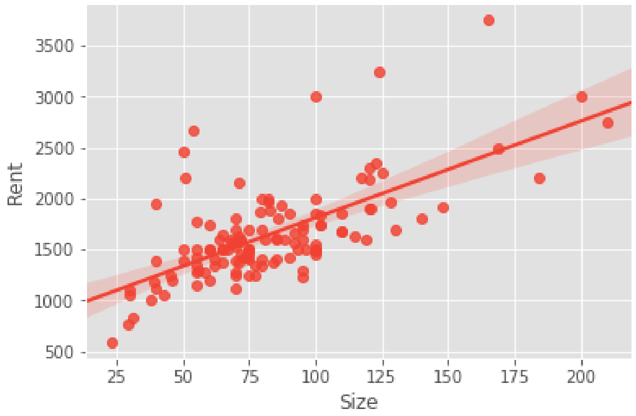
As the main dataframe table containing all dependent variables and the one to be predicted, the effect of each x_i on the y is separately evaluated:

- Rental Prices Distribution with Histogram



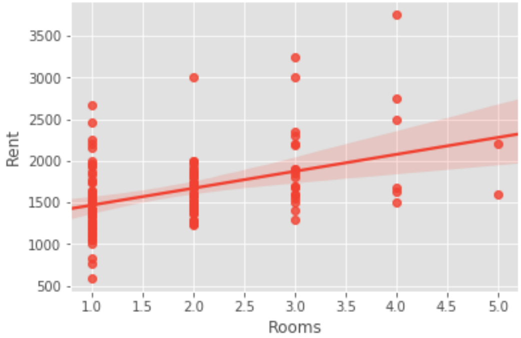
Most of the data prices is comprised within 1000 and 2500 euros.

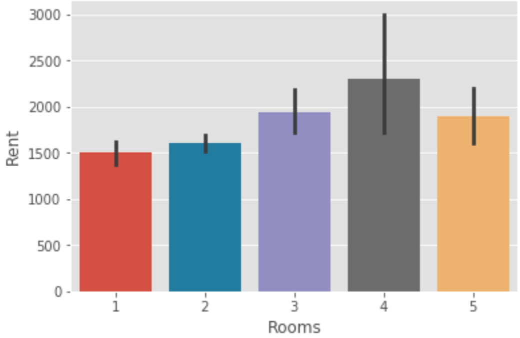
- Regression plot for Price vs House surface



Main data population is mostly concentrated for houses between 50 and 125 sqm. In this range the linear regression curve fits well, while standard deviation increases out of this range (smaller and bigger houses).

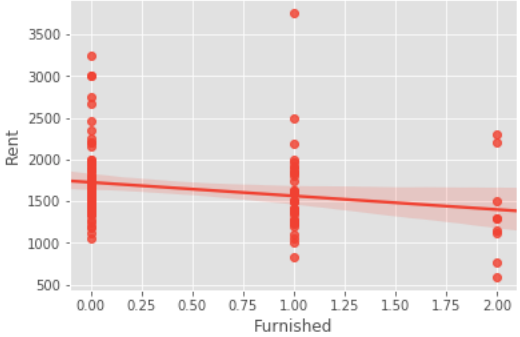
- Barplot and Regression Plot for price vs House room number

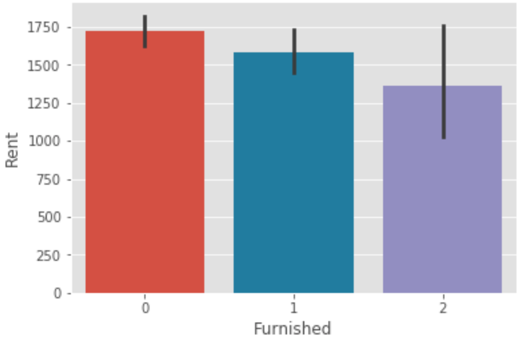




This graph shows linear regression might be a good fit for the model, but potentially improved with higher order polynomial. Moreover, as the number of rooms increases over the average value 1500 euros, the price increase as well -with higher dispersion of data (std deviation), probably due to lower number of data available in the population.

- Barplot and Regression Plot for price vs House furnishing

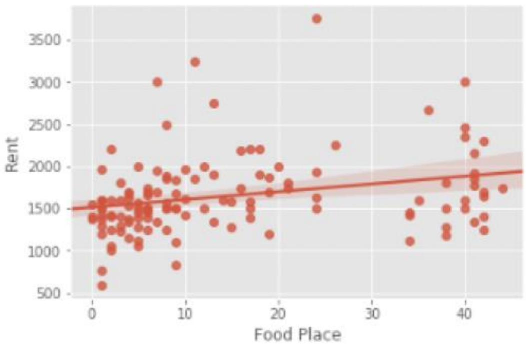


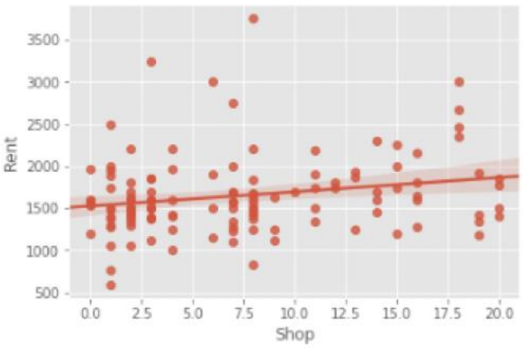


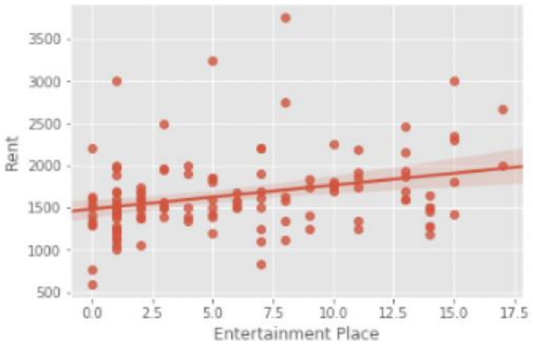
where:

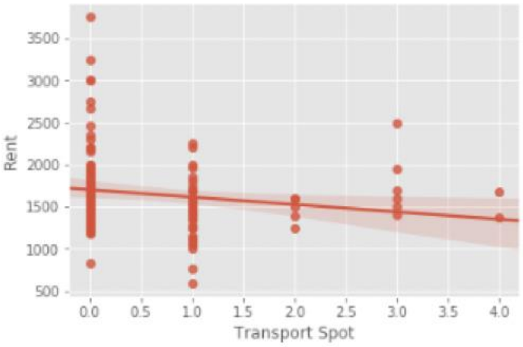
- Furnished apartments are indexed 0
- Upholstered apartments are indexed 1
- Shell apartments are indexed 2

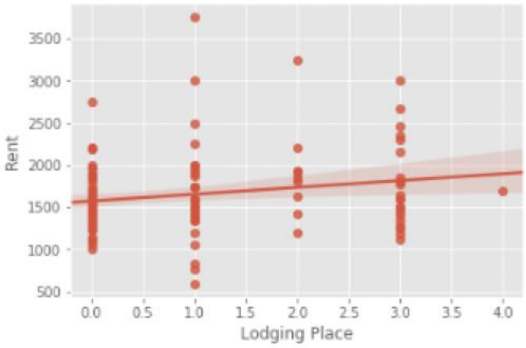
- Regression plot for Price vs Neighborhood spots









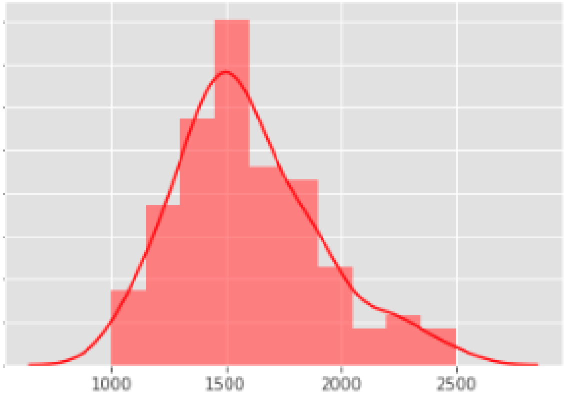


From the graph it can be noticed the price generally increase as the services are more available, even though the slope of interpolating line is small. This suggests these variables may not be good predictors for the price.

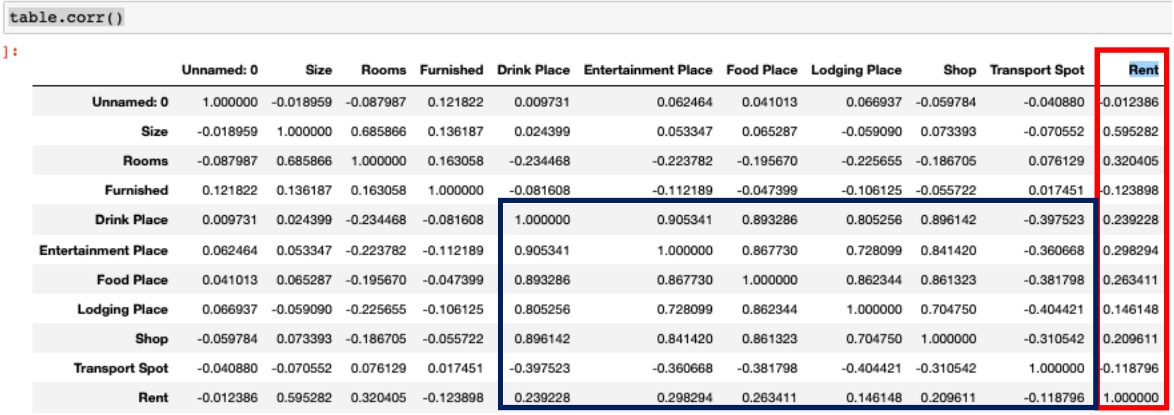
3.1. Results

All the significative data determine an interpolation line with slope comprised between 1500 and 2000.

For this reason in the machine learning Regression Algorithm, the analysis will be done focusing in the price range within 1000 and 2500 euros.



Also, as shown from Pearson coefficient evaluation in the considered dataset of Utrecht neighborhoods:



```
table.corr()
```

]:

	Unnamed: 0	Size	Rooms	Furnished	Drink Place	Entertainment Place	Food Place	Lodging Place	Shop	Transport Spot	Rent
Unnamed: 0	1.000000	-0.018959	-0.087987	0.121822	0.009731	0.062464	0.041013	0.066937	-0.059784	-0.040880	-0.012386
Size	-0.018959	1.000000	0.685866	0.136187	0.024399	0.053347	0.065287	-0.059090	0.073393	-0.070552	0.595282
Rooms	-0.087987	0.685866	1.000000	0.163058	-0.234468	-0.223782	-0.195670	-0.225655	-0.186705	0.076129	0.320405
Furnished	0.121822	0.136187	0.163058	1.000000	-0.081608	-0.112189	-0.047399	-0.106125	-0.055722	0.017451	-0.123898
Drink Place	0.009731	0.024399	-0.234468	-0.081608	1.000000	0.905341	0.893286	0.805256	0.896142	-0.397523	0.239228
Entertainment Place	0.062464	0.053347	-0.223782	-0.112189	0.905341	1.000000	0.867730	0.728099	0.841420	-0.360668	0.298294
Food Place	0.041013	0.065287	-0.195670	-0.047399	0.893286	0.867730	1.000000	0.862344	0.861323	-0.381798	0.263411
Lodging Place	0.066937	-0.059090	-0.225655	-0.106125	0.805256	0.728099	0.862344	1.000000	0.704750	-0.404421	0.146148
Shop	-0.059784	0.073393	-0.186705	-0.055722	0.896142	0.841420	0.861323	0.704750	1.000000	-0.310542	0.209611
Transport Spot	-0.040880	-0.070552	0.076129	0.017451	-0.397523	-0.360668	-0.381798	-0.404421	-0.310542	1.000000	-0.118796
Rent	-0.012386	0.595282	0.320405	-0.123898	0.239228	0.298294	0.263411	0.146148	0.209611	-0.118796	1.000000

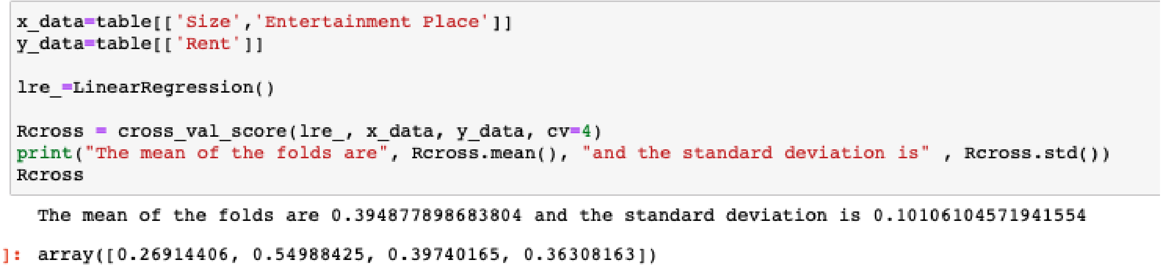
- The main- predictor for the Rental Price of the house seems to be the area of house itself and number of so-called Entertainment Places group in the neighborhoods (red rectangle).
- Neighborhood quantities seems to be in general not a great predictor of rental prices (Pearson coefficient values < 0.3). However, they seem to be mutually connected to each other, with Entertainment Place number of places may eventually be considered as dependent variable of the others (blue rectangle)

Therefore the predictive model for rental prices will be built basing on these 2 variables, seeking the best fitting function.

4. Model Development

4.1. Cross Validation Score

As the title says, the selected dataset is cross validated via division in 4 groups of the xi and evaluation of the R squared coefficient for accuracy



```
x_data=table[['Size','Entertainment Place']]
y_data=table[['Rent']]

lre_=LinearRegression()

Rcross = cross_val_score(lre_, x_data, y_data, cv=4)
print("The mean of the folds are", Rcross.mean(), "and the standard deviation is" , Rcross.std())
Rcross
```

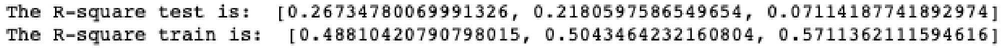
```
  The mean of the folds are 0.394877898683804 and the standard deviation is 0.10106104571941554

]: array([0.26914406, 0.54988425, 0.39740165, 0.36308163])
```

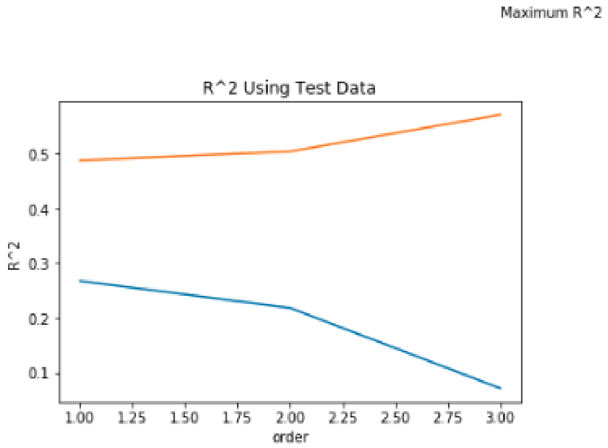
4.2. Model Evaluation

Model is developed with prediction purposes, splitting then the dataset into training and testing value (70-30) and comparing how the model works with the 2 Dataset.

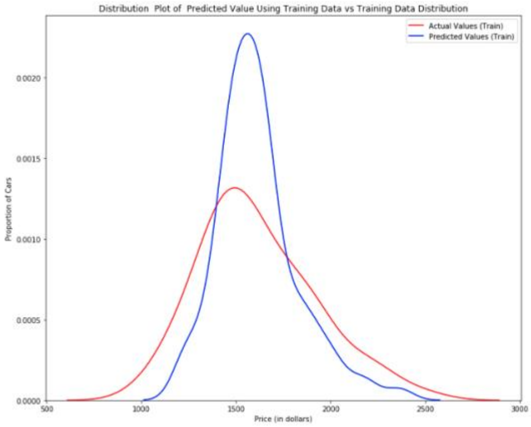
Therefore, in order to choose the best fitting order polynomial function for the dataset, an analysis of the R squared value obtained vs different order polynomial functions:

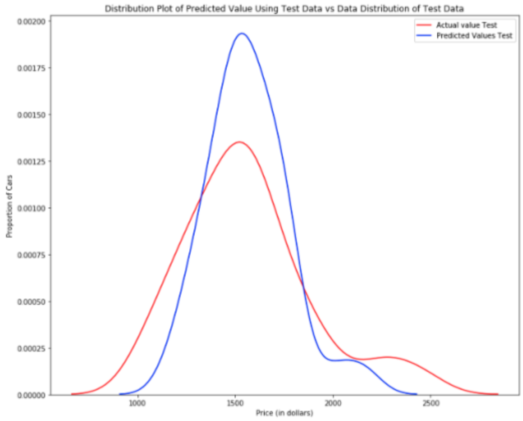


```
The R-square test is:  [0.26734780069991326, 0.21805975865496 54, 0.07114187741892974]
The R-square train is:  [0.48810420790798015, 0.5043464232160804, 0.5711362111594616]
```



The graph shows for train dataset, an improvement of the accuracy when the order increases, while opposite behavior for testing data. The best value of compromise seems therefore to be n=1, the Multiple Linear Regression is chosen.

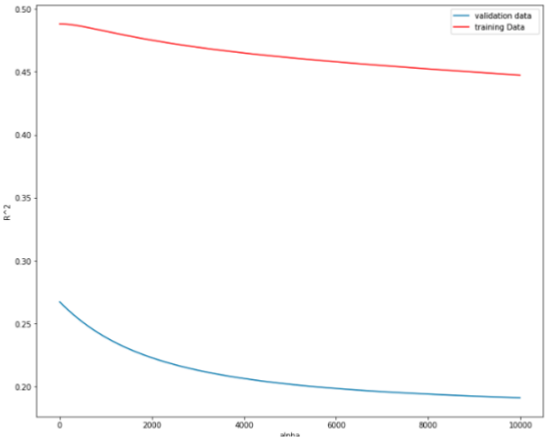




The model shows similar (conservative) behavior for both training and testing data.

4.2.1. Ridge Regression

In order to evaluate introduction of hyperparameter Alpha -as an attenuator of the effect of 2 xi coefficients – a Ridge Regression is carried out. Similarly to previous graph, accuracy of the model is plotted against continuous values of alpha



Best value is alpha = 1 with
R squared train = 0.488
R squared test = 0.267

5. Conclusion

Given the available dataset, the information have been organized in a dataframe having all main variables of interest and using also information of neighborhoods through Foursquare API. Then, each variable have been evaluated separately, in order to investigate its effect on house rental price.

Finally predictive model is built. It works well for training dataset (70% of the entire dataset), while does not fit test values.

This result is strictly connected with the scattering of the data (see Pearson coefficient values pag 8) and the best structure of the model is the one obtained via Ridge Regression on degree 1 polynomial function with the 2 variable Size of the houses and number of Entertainment places in the neighborhoods.

Summing up, in Utrecht cities rental price mainly depend on house dimensions, poorly on amount of neighborhoods (maybe because in the Netherlands all houses are concentrated in building districts), while the neighborhoods are influenced to each other.