IBM Data Science Professional Certificate

Predicting property prices of Lisbon's boroughs using Foursquare data

Luis Domingues

Course 9: Capstone

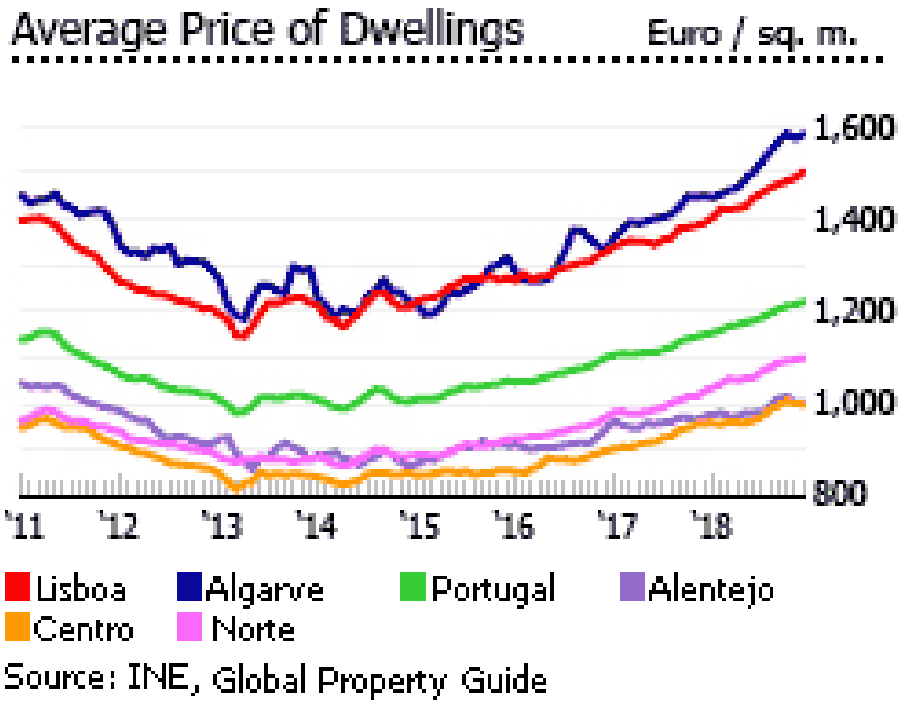
25/May/2019

Overview

- Motivation
- Methodology
  - Data Collection
  - Methods and Algorithms
- Results
  - Understanding the Data
  - Multiple Linear Regression
- Conclusions

Motivation

- Property prices have been steadily rising in Portugal since 2014, and in particular in Lisbon [1].
- Property is not only needed to live but is also an investment instrument.
- Being able to predict property prices is very useful.



Methodology

Data Collection

- Data was collected from sources [2-4]:

Lisbon boroughs' info
Names
Location (lat., long.)
Property prices
Population
Area

Lisbon venues' info
Name
Location (lat., long.)
Category

(1007 venues returned)

Reduced categories used due to high number of Foursquare categories (182) -> too sparse

Reduced categories
Arts, entertainment and nightlife
Food and drink
Supermarkets and groceries
Shopping
Historic sites and museums
Hotels and accomodation
Athletics and sports
Transport
Public buildings
Health and education buildings
Outdoors

Methods and Algorithms

The main goal of the work is to predict property prices (€/m^2) from a set of features.

Methodology used:

1. Understand the data

   Methods: Descriptive statistics and k-means clustering

2. Determine factors to be used to predict property prices

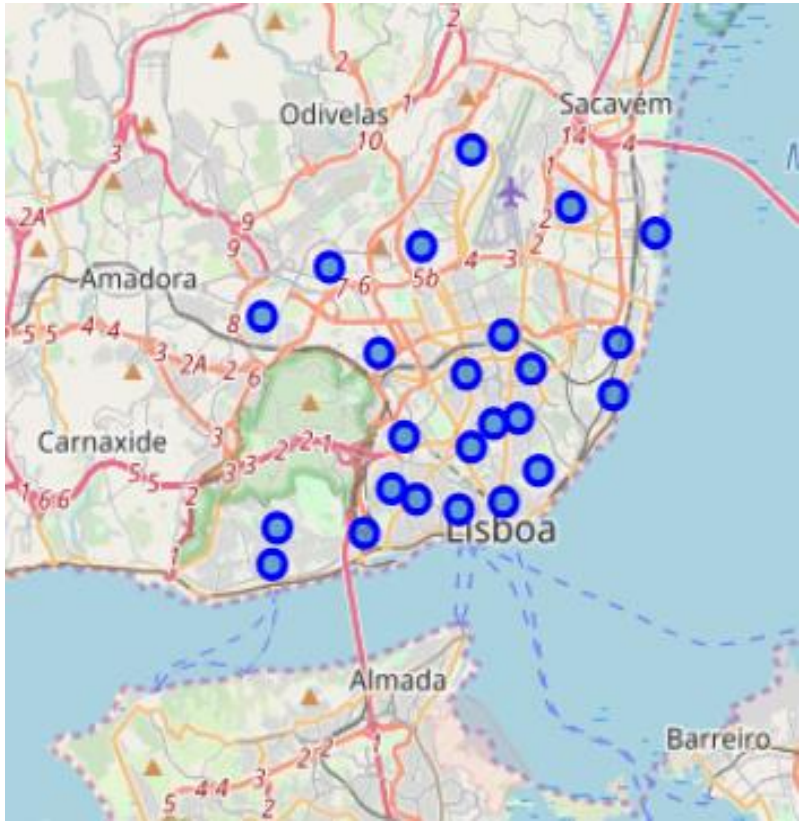
   Methods: Correlation factors and multiple linear regression

3. Predict property prices of Lisbon's boroughs

   Methods: Multiple linear regression

Results

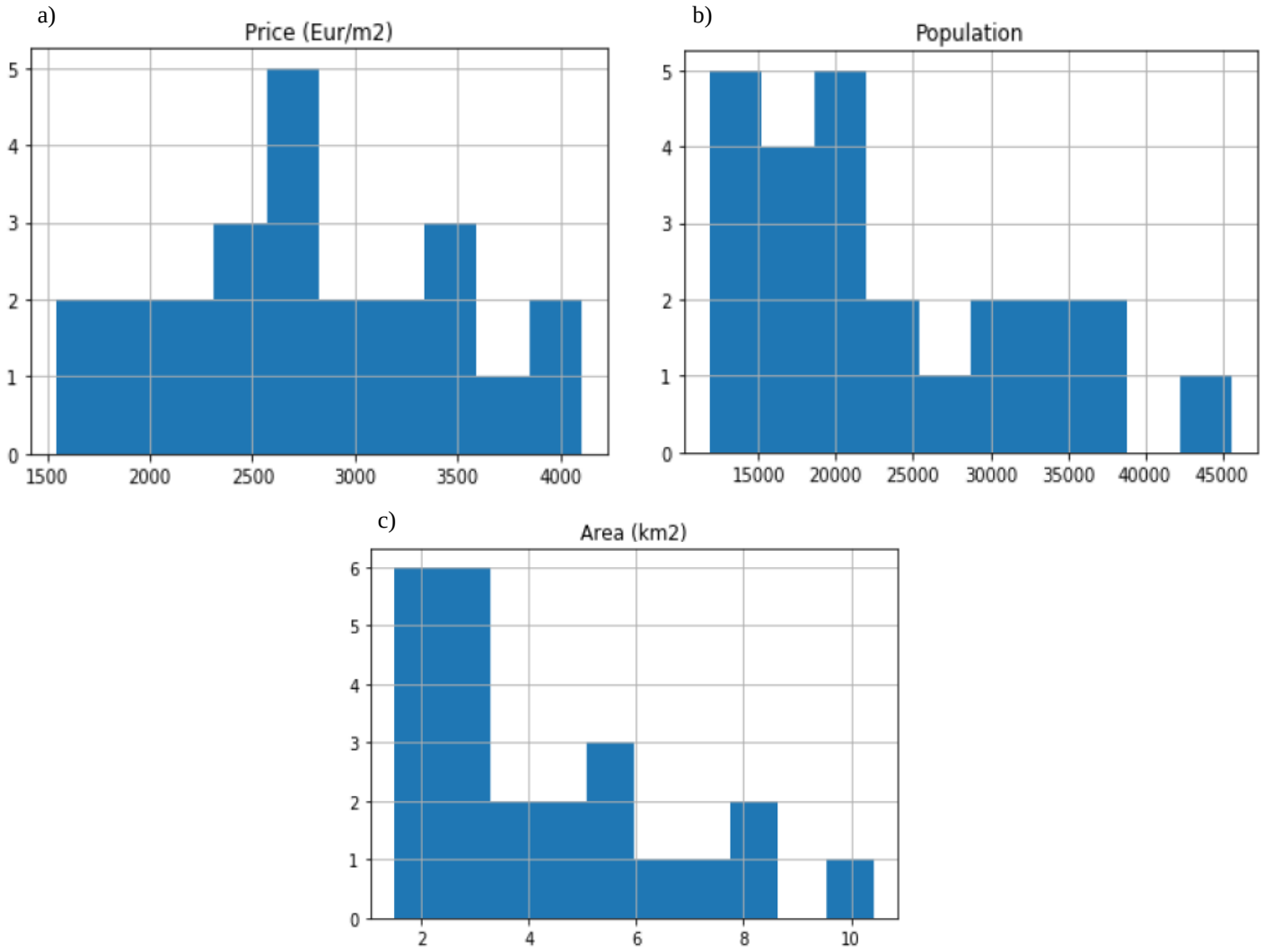
Understanding the Data - I



Statistics	Price (€/m²)	Population	Population dens. (inhab./km²)	Area (km²)
mean	2771.75	23029	6608	4.17
std	667.60	9519	3193	2.40
min	1543.00	11836	1584	1.49
25%	2381.00	15430	4548	2.39
50%	2741.00	20578	5769	3.19
75%	3155.75	31693	7704	5.37
max	4105.00	45605	14860	10.43

- Differences between maximum and minimum are quite significant for all features presented

Understanding the Data - II



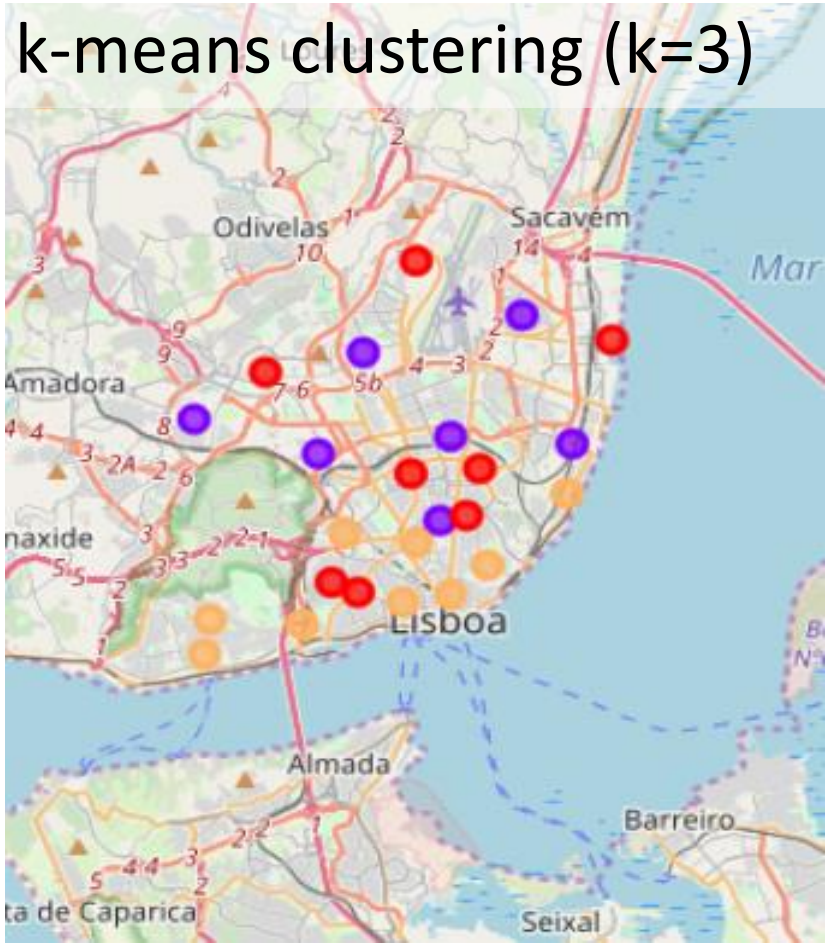
- None of the variables seem to follow a Normal distribution.
- For price the mean (2771.75 €/m^2) corresponds to the bin with more observations.
- For population and area the first three bins (with lowest values) have the highest number of observations.
- Price is differently distributed than population and area.

Understanding the Data - III

Category	Number of venues
Shopping	80
Food and drink	604
Arts, entertainment and nightlife	113
Athletics and sports	30
Transport	13
Outdoors	22
Hotels and accomodation	90
Public buildings	19
Historic sites and museums	11
Health and education buildings	4
Supermarkets and groceries	21
Shopping	80
TOTAL	1007

- The category with highest number of venues is *'Food and drink'*, with *ca.* 60% of retrieved venues.
- *'Food and drink'*, *'Arts, entertainment and nightlife'* and *'Hotels and accomodation'* account for *ca.* 80% of total venues retrieved by Foursquare.
- Foursquare seems biased towards venues in these categories.

Understanding the Data - IV

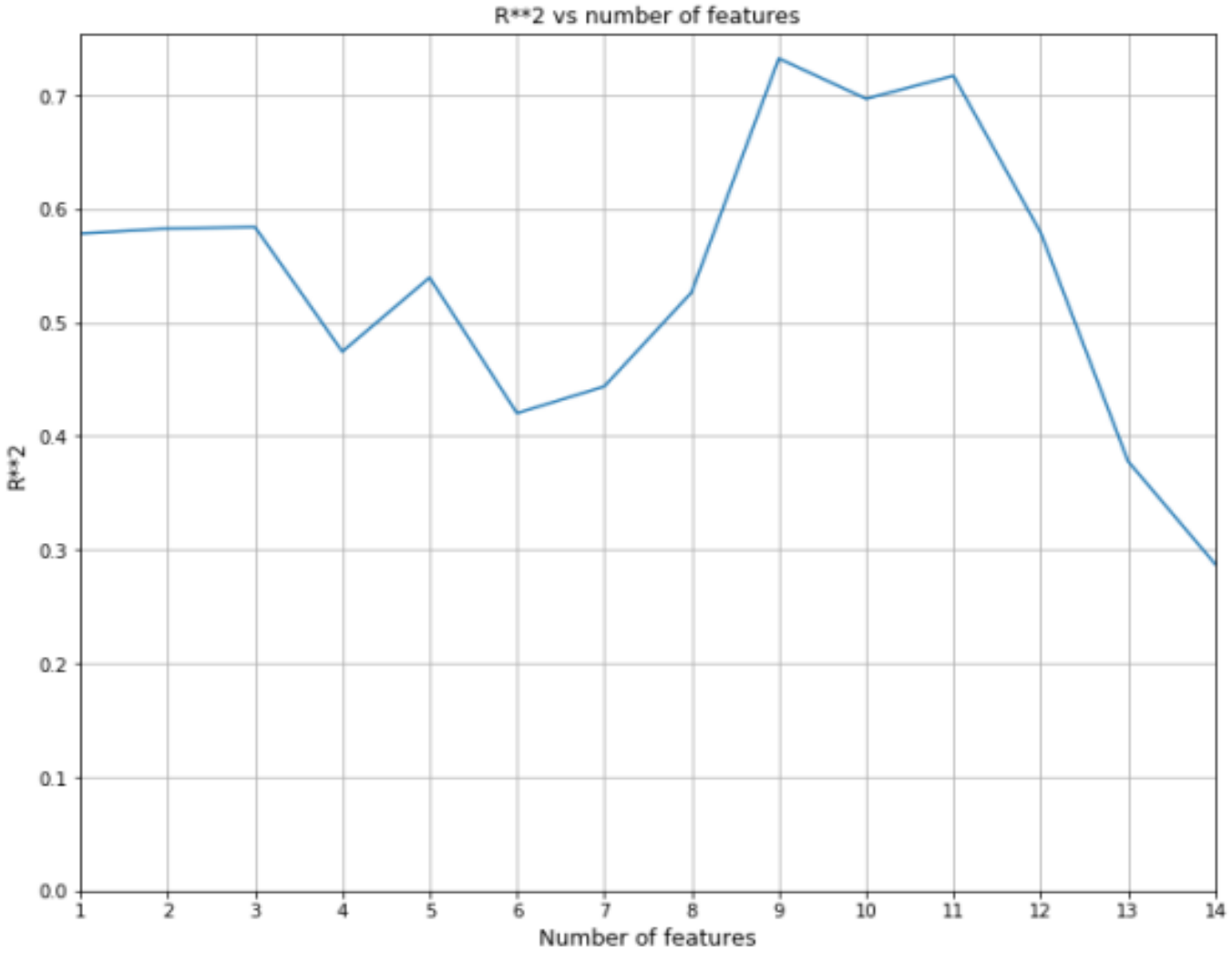


Cluster 1; Cluster 2; Cluster 3;

Feature	Cluster 1	Cluster 2	Cluster 3
Price (€/m^2)	2864.0	2384.8	2990.7
Population (inhab.)	21837	35811	14147
Population dens. (inhab./km^2)	7833	7078	5153
Area (km^2)	3.273	5.939	3.588
Food and drink	28.00	16.29	29.56
Arts, entertainment and nightlife	3.63	2.29	7.56
Hotels and accomodation	3.13	1.29	6.22

- Cluster 2 has the lowest property price, while clusters 1 and 3 have similar prices.
- *'Food and drink'*, *'Arts, entertainment and nightlife'* and *'Hotels and accomodation' correlate very well with price.*

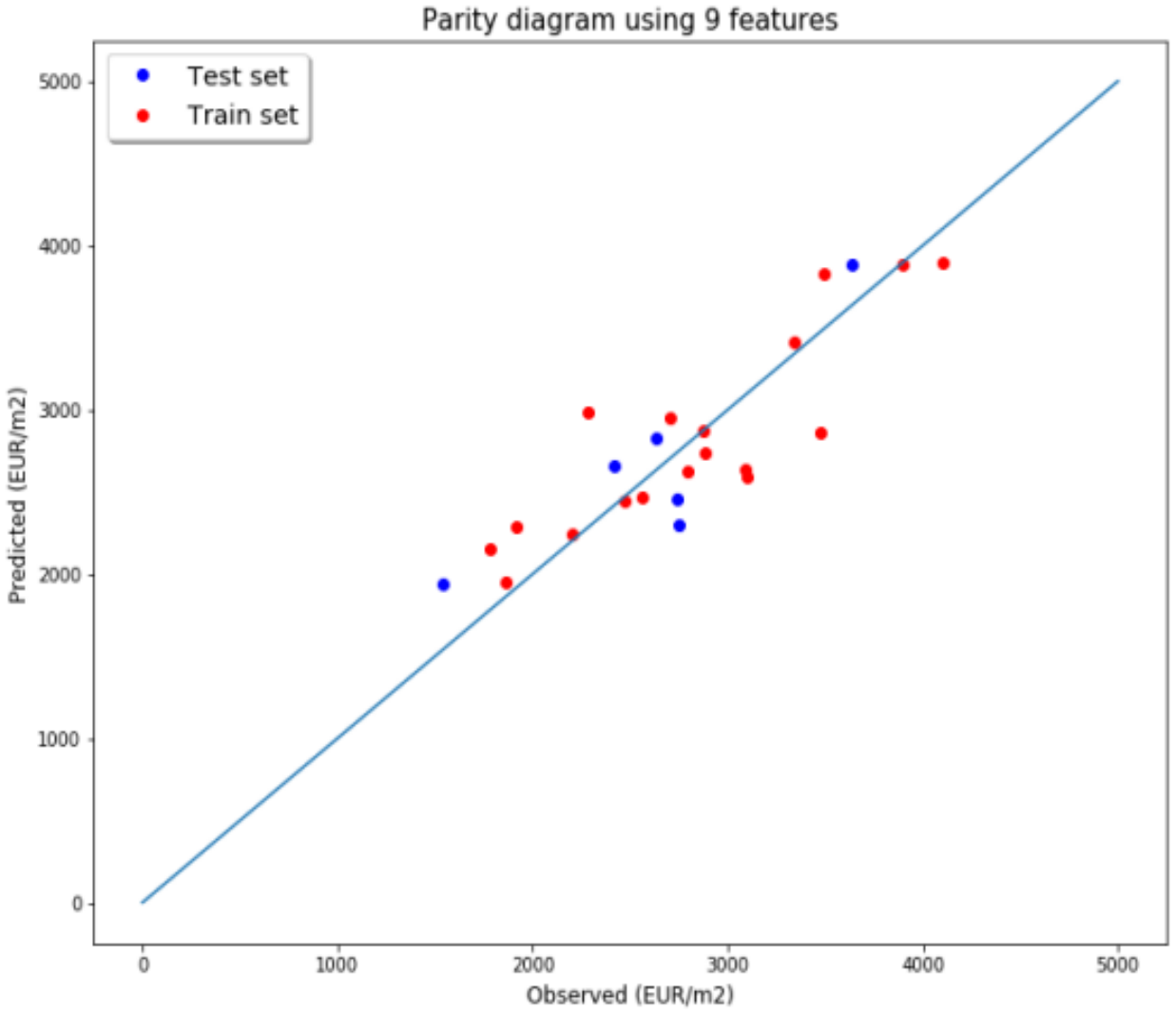
Multiple Linear Regression - I



Feature	Absolute correlation factor
Food and drink	0.767761
Hotels and accomodation	0.644956
Arts, entertainment and nightlife	0.605600
Shopping	0.534702
Population	0.465174
Public buildings	0.457006
Area	0.356242
Outdoors	0.204509
Athletics and sports	0.194633

- Max $R^2 = 0.732$
- 9 features used

Multiple Linear Regression - II



Variable	Coefficient
Intercept	2559.86
Food and drink	-4.0805
Hotels and accomodation	44.4691
Arts, entertainment and nightlife	34.5041
Shopping	28.4188
Population	-0.0141
Public buildings	110.1465
Area	-23.3758
Outdoors	148.1406
Athletics and sports	84.5804

- Reasonably good fit
- $R^2 = 0.732$
- Residuals randomly distributed

Conclusions

Conclusions

- Foursquare data ✚ Area and population data ⟹ Predict property prices.
- Foursquare data was useful in predicting property prices.
  - Most impactful venues: *'Food and drink'*, *'Arts, entertainment and nightlife'* and *'Hotels and accomodation'*.
- Property prices successfully predicted using multiple linear regression.
  - R^2 = 0.732; Residuals randomly distributed.
- Obtained correlation can be an useful tool to predict the property prices evolution in Lisbon's boroughs.
  - Continuously monitoring of the evolution of each of the features required.

References

References

1. C. D. Lalaine, “Investment Analysis of Portuguese Real Estate Market,” *GlobalPropertyGuide*, 06-Feb-2019. [Online]. Available: https://www.globalpropertyguide.com/Europe/Portugal/Price-History. [Accessed: 21-May-2019].
2. Wikipédia, “Lista de freguesias de Lisboa,” 2019. [Online]. Available: https://pt.wikipedia.org/wiki/Lista_de_freguesias_de_Lisboa. [Accessed: 10-Apr-2019].
3. N. Carregueiro, “Mapa: Só numa freguesia de Lisboa e Porto os preços das casas não sobem mais de 10%,” *Negócios*, 30-Oct-2018.
4. “Foursquare API,” *Foursquare*, 2019. [Online]. Available: https://developer.foursquare.com/. [Accessed: 22-May-2019].

Thank you. ☺