IBM – Coursera

Data Science Professional Specialization

Capstone Project

Correlation between Airbnb rooms in the Lisbon neighborhood and its surrounding venues

Alexandre da Cunha Pires

2020

This notebook is for the capstone project for the 9-courses specialization in Data Science of IBM. Platform Coursera. Data Science Specialization This project will try to solve a problem or question by applying data science methods on the location data gotten from FourSquare API.

1. Introduction

Whoever has come to Lisbon will be enchanted by her beauty and atmosphere, with vibrant life, never-ending to-do list and by the expertly-planned transport system which can take you everywhere in the city. Lisbon is famous worldwide as warm place, the best gastronomy and "tourist must go to" city in Europe. No wonder the real estate market is becoming one of the priciest. It's where penthouses at the Principe Real, just steps from the Cais do Sodré, is sold for more than 2 million. But that is just an outlier example. A quick search can show us the real estate price can vary by a large margin from neighbourhoods to neighbourhoods. For example, a 2-bedrooms condo in Parque das Nações, can cost 700 thousands on average; while in Odivelas, just 20 minutes aways, it's only $150 thousands and Moscavide is only 5 minutes cost less than 400 thousands.

So what aspects of a neighbourhood that can affect the price of real estates to such extend? One hypothesis is that the surrounding venues can be a decision factor. Surely anyone, who has attempted to find an accommodation for rent or buy, has seen advertisements such as: This condo is located near the subway station, malls, supermarkets, dinners, etc. And it's likely that the price will be higher than others with locations not as "convenient".

The Airbnb arrived and changed the tourism market. In the past use to have only hotels and small amount of people renting their own house to tourist to have some extra money. Today, with a site that collect many different options of rooms, studio and houses for a competitive price and some services quite similar that you can find in hotels many new investor having invest a lot of money in this sector of market. As the tourism increase every year and the preference of the consumer (tourist) incline to Airbnb rooms, we can see many buildings been reformed to offer new rooms scattered in many neighborhoods. But are there neighborhood more attractive than others? and are there potential neighborhood to explorer? Can the venues surrounding an accommodation affect market value? And what kind of venues can affect the most?

Target Audience to this report:

- Potential Buyers – investor that look for gold mine to increase your revenue;
- House Sellers – knowing the real value of the building can be used to sell to the right buyers for the highest price;
- Small Business – if someone want to open a new restaurant, café, bakery and tourism shop, it would be a good information where are new regions with accommodations for tourists;
- For everyone that have interest in see what can be done with data science and to prove myself that I can do it.

II. Description of the data

Lisbon city neighborhood were chosen as the target of this project for the following reasons:

- Is the city in Europe that having been increasing interest for tourism, technological business investments and a home for many foreigners looking for live changing;
- Is the best dataset that it was found and suit well for this project;
- Beside Lisbon to be one of the best city in world to live, it is where I want to find my home.

1. Data collection process:

The average price will be scrapped from the Kaggle website, It was chosen Airbnb Lisbon 2017 as dataset to work on. For each neighbourhood, call Geocoder Python to get its coordinate. For each neighbourhood's coordinate, call FourSquare API to get the surrounding venues. Count the occurrences of each venue type and attach that information to each neighbourhood. Each row represents a neighbourhood. Each column will be the count of one type of venue in that neighbourhood.

2. Using data to solve the question:

First, it will be done different graphics to understand which feature has more influence in the price and it will transpose the data in the map to have better visualization how the Airbnb rooms are spread in Lisbon.

Second, correlation between price and surrounding venues will be checked. Next step, if correlated, machine learning techniques will be used to analyse the data. The output will be a list of venues types that effect the most on price, along with their weight on the result.

Questions that it will be use to explorer the data:

If the surrounding venues can affect the price of Airbnb rooms? what kind of surrounding venues, and to what extent, can affect the price? if we can use the surrounding venue to estimate the value of an accommodation over the average price of one area? And to what degree of confidence? The result can be useful for home buyers, who can roughly estimate the value of a target house over the average. Or to planners, who can decide which venues to place around their product, so that the price is maximized.

III. Methodology

Initially it will be done basic analyses with the dataset to understand more the data that are been used for this project

The assumption that price of Airbnb rooms is dependent on the surrounding venue. Thus, regression techniques will be used to analyse the dataset. The regression will be the occurrences of venue types. And the dependent variable will be standardized average prices.

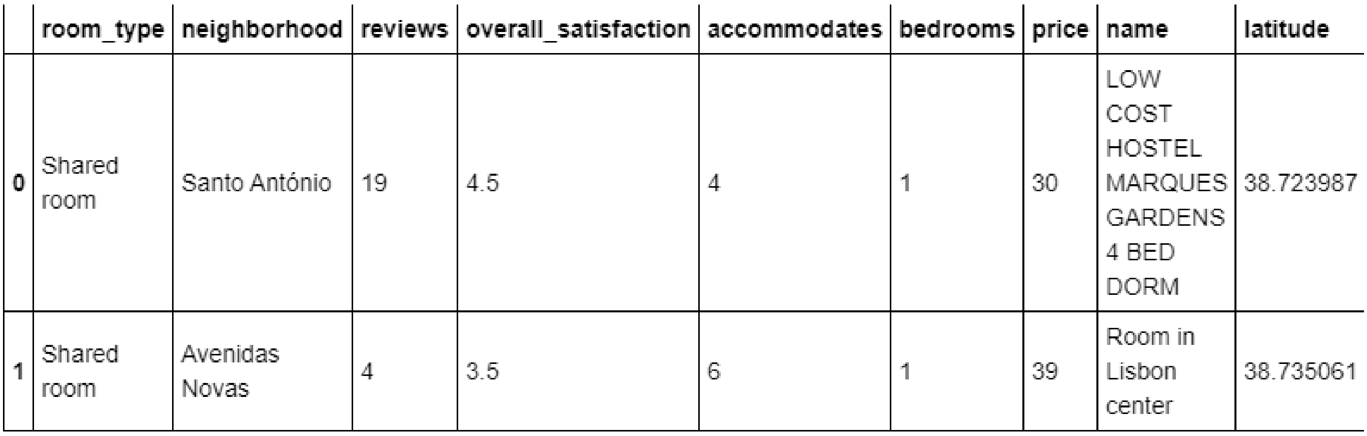
It will be used K-ment to identify clusters in the neighborhoods by the venues.

Python data science tools will be used to help analyse the data. Completed code can be found here:

https://github.com/ACPIRES-EU/Coursera_Capstone2/blob/master/Capstone_Price_Building.ipynb

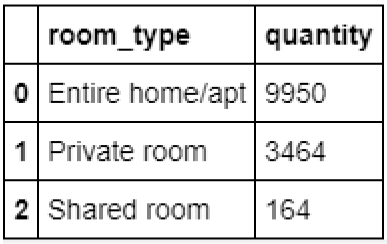
Analyse of the dataset of Airbnb

Dataset come with many data that it was known how would be the parameters of each column. It was made some groupby, sum, mean to understand the impact of each variable like review, overall_satisfacion and accommodations.

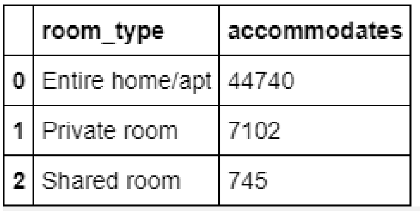


	room_type	neighborhood	reviews	overall_satisfaction	accommodates	bedrooms	price	name	latitude
0	Shared room	Santo António	19	4.5	4	1	30	LOW COST HOSTEL MARQUES GARDENS 4 BED DORM	38.723987
1	Shared room	Avenidas Novas	4	3.5	6	1	39	Room in Lisbon center	38.735061

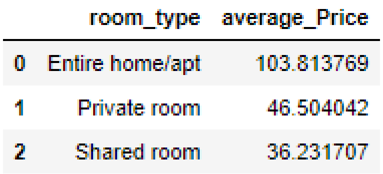
The pictures below we can see that the room type show us 3 types of rooms. The first one is "Entire home/apt" and have the majority of rooms. In "Private room" the tourist only rent one room, so it was created the second picture showing how many rooms there are in one apartment renting a single room and the value double. The last picture we can see the avg of price for type of room.



	room_type	quantity
0	Entire home/apt	9950
1	Private room	3464
2	Shared room	164

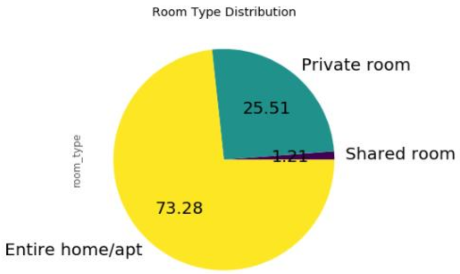


	room_type	accommodates
0	Entire home/apt	44740
1	Private room	7102
2	Shared room	745

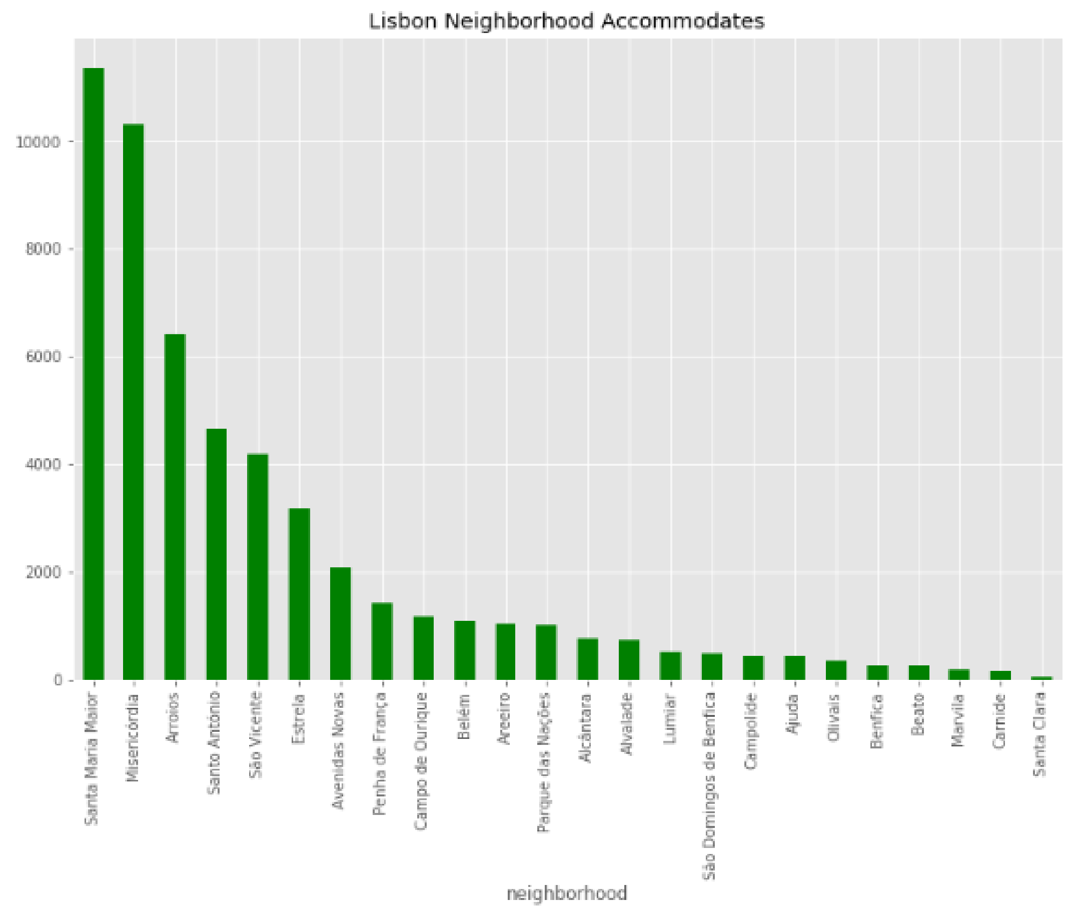


	room_type	average_Price
0	Entire home/apt	103.813769
1	Private room	46.504042
2	Shared room	36.231707

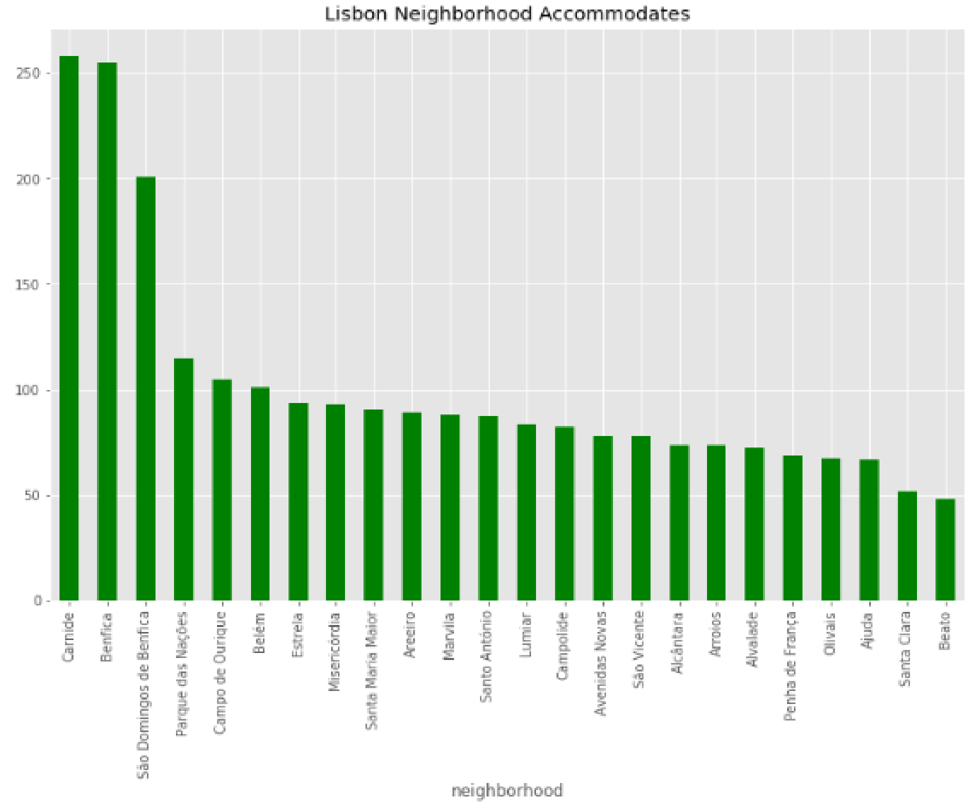
The weight of impact of the average of room price come from Entire home with 73.28% and Private room with 25.51%.

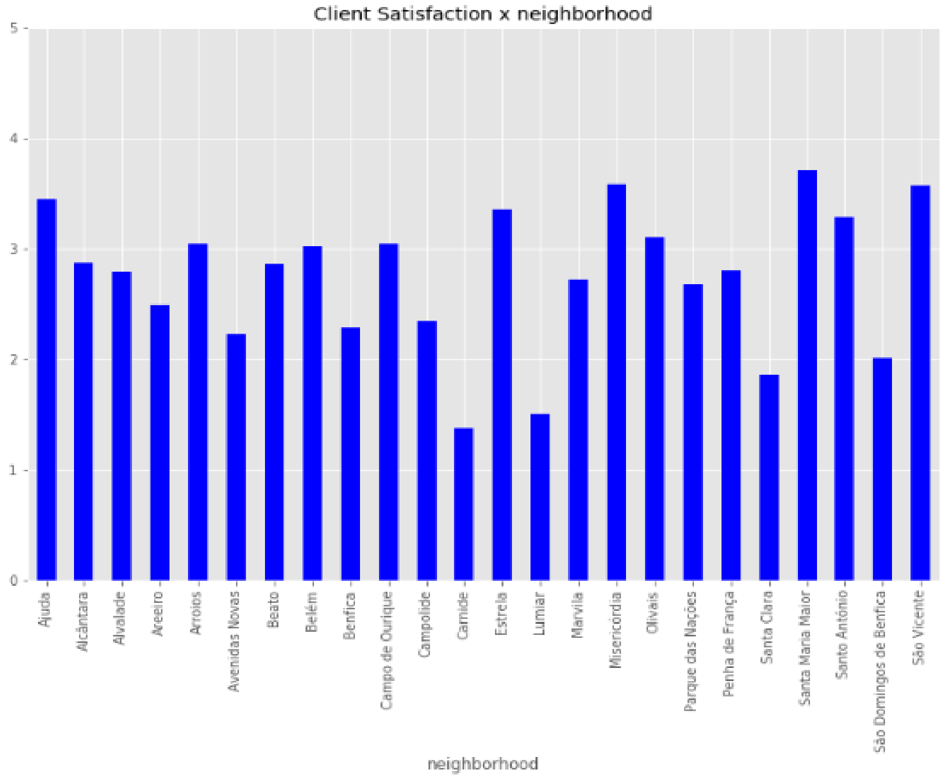


In the matter how the Airbnb is spread in the city, it is clear that 80% of all rooms are in only in 5 of 24 neighborhoods in Lisbon. This can be a limitation by transport, venues, strategic place, building availability or potential growth.



The average price for neighborhood it is very stable with an exception in 3 neighborhood of 24. It is interesting when it analysed overall satisfaction of the clients per neighborhood, price it is not assurance to have good quality and high satisfaction.

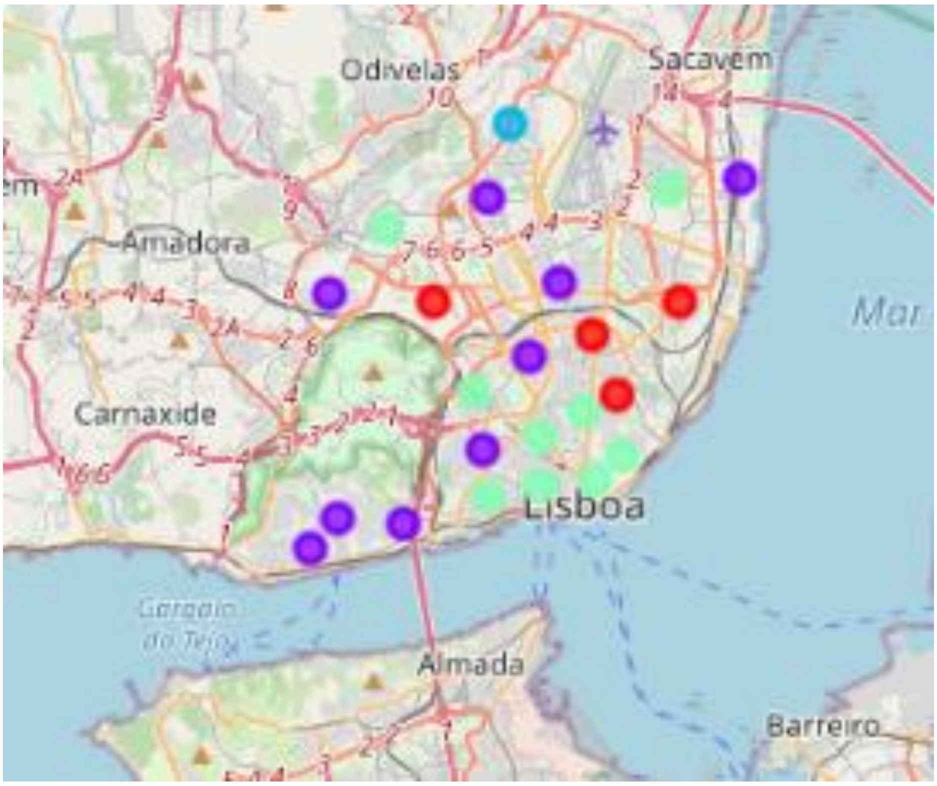




Analyse of Clusters using K-ment

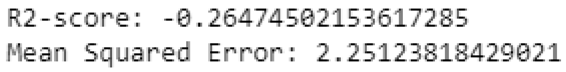
Applying K-meat with venues of the neighborhood, it was created 4 clusters that define quite precise with the client satisfaction and tourism interest.

- Cluster 0 (red): Tourism area (Because has Park, Gym, Pub, Restaurant, shopping);
- Cluster 1 (purple): Business area (Because has Hotels, many coffee shops and gift shop);
- Cluster 2 (blue): Living Area with less population and far from the business and tourism area (Because has less options and interaction with Gym, for example);
- Cluster 3 (light green): Living Area very close to the business area and tourism area



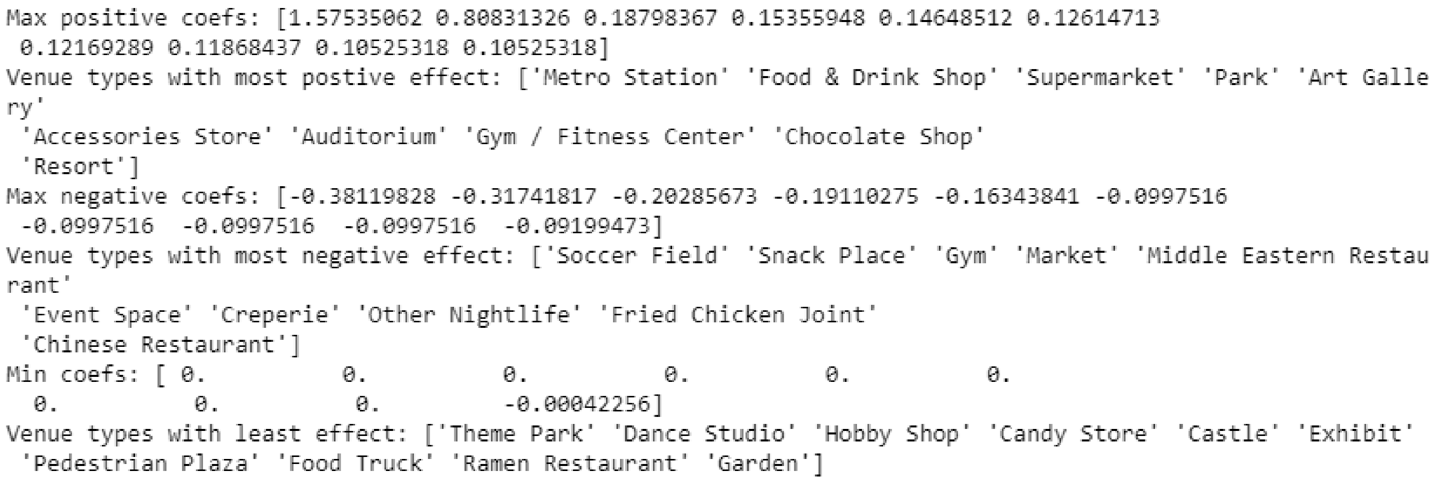
Analyse of Linear Regression and Principal Component Regression

The Linear Regression model will contain a list of coefficients corresponding to venue types. R2 score (or Coefficient of determination) and Mean Squared Error (MSE) will be used to see how well the model fit the data. The result doesn't seem very promising. R2 score is small, which means the model may not be suitable for the data. The MSE is very high, so it is definitely a bad result.



```
R2-score: -0.26474502153617285
Mean Squared Error: 2.25123818429021
```

It isn’t a totally lost the model results. It is capable to extract some good analyse from the list of venues that have positive and negative effects in the average price.

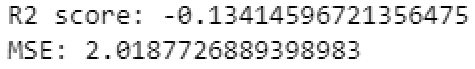


```
Max positive coefs: [1.57535062 0.80831326 0.18798367 0.15355948 0.14648512 0.12614713
 0.12169289 0.11868437 0.10525318 0.10525318]
Venue types with most postive effect: ['Metro Station' 'Food & Drink Shop' 'Supermarket' 'Park' 'Art Galle
ry'
 'Accessories Store' 'Auditorium' 'Gym / Fitness Center' 'Chocolate Shop'
 'Resort']
Max negative coefs: [-0.38119828 -0.31741817 -0.20285673 -0.19110275 -0.16343841 -0.0997516
 -0.0997516  -0.0997516  -0.0997516  -0.09199473]
Venue types with most negative effect: ['Soccer Field' 'Snack Place' 'Gym' 'Market' 'Middle Eastern Restau
rant'
 'Event Space' 'Creperie' 'Other Nightlife' 'Fried Chicken Joint'
 'Chinese Restaurant']
Min coefs: [ 0.          0.          0.          0.          0.          0.
  0.          0.          0.         -0.00042256]
Venue types with least effect: ['Theme Park' 'Dance Studio' 'Hobby Shop' 'Candy Store' 'Castle' 'Exhibit'
 'Pedestrian Plaza' 'Food Truck' 'Ramen Restaurant' 'Garden']
```

PCR can be explained simply as the combination of Principal Component Analysis (PCA) with Linear Regression. (Wikipedia, n.d.) PCR employs the power of PCA, which can convert a set of values of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. As the result, the number of features is reduced while keeping most of the characteristic of the dataset.

Then PCR use Linear Regression on the converted set to return a coefficient list, just like in normal Regression techniques.

Again, R2 score and MSE are used to see how well the model fit the dataset.



```
R2 score: -0.13414596721356475
MSE: 2.0187726889398983
```

The result is not promising as it shows worst numbers over the simple Linear Regression.

IV. Results

There are many neighborhood close to tourism and business area with good transport like metro, bus, taxi, uber that it hasn’t been explorer, but as Lisbon is an old city, many neighborhoods have old buildings that need high investments to be suitable to use as Airbnb apartment. The venues shows that these neighborhood has enough attraction to assure client satisfaction in this matter.

About neighborhood far from tourism centre is totally advisable not invest there. These neighborhood have the lower overall satisfaction from the client because have less venues, take more to go to tourism area or business area.

The predict model doesn’t have a good score and even after applying a more sophisticate method, the model is still not suitable for the dataset. Thus, it can’t be used to precisely predict a neighborhood average price of Airbnb rooms.

Explanations for the poor model can be:

- The price takes some time to find a balance between supply and demand, as Airbnb still are expanding and the demand still high. So, there are price fluctuation;
- The machine learning techniques are chosen or applied poorly;
- It can exist a mix of interest from customer that create differences value between the venues increasing the complexity math to find a suitable model.

Even though, it was found the positive effect venues that can help to take a decision in new investments in new rooms.

V. Discussions

We have in nowadays vast information in the internet. But to find data about specific subject is very difficult and if you find it, it has a high probability that it won't be for free. There are some sites that has many different datasets if you want to practice your skills.

Another challenge is, choosing the suitable technique to construct the model is also a worthwhile process. As this report shows that, by applying a different method, you can extract different information or find better results.

About find a better place to invest is interesting to merge the data set with another data that contain the average noise in the neighborhood and crime statistic to check if can impact price and client satisfaction.

VI. Conclusion

It was possible to learning many things with the analyse in this project. Such as, transport, restaurant, coffee shop and market are important to settle any Airbnb investment, as well close to tourism (priority) and business area. Another interest fact is that only 4 neighborhood having been explorer massively and exist other neighborhoods that have high potential to new investments.

It's unfortunately that the analysis couldn't produce a precise model or showing any strong coefficient correlation for any venue type. But we can still get some meaningful and logical insights from the result.

Doing this project helps practicing every topic in the specialization, and thus, equipping learners with Data Science methodology and tools using Python libraries. Also doing a real project certainly helps one learns so much more outside the curriculum, as well as realizes what more to research into after completing the program. And as this report shows, there are surely a lot of things to dig into.

IT was an insightful experience and thank you for all participants from the course.