BEST DISTRICT OF A SMALL BUSINESS USING DATA SCIENCE IN THE CITY OF MADRID.

Jesús A Molero Cano

October 6, 2020

1 Introduction

Knowing where it is the best place to open a business in a city is a hard task. For this reason, by using DataScience tools, useful Data and the proper methodology we may find the answer for a specific city, in this case the city of Madrid.

The opening of a new business should have a significant study of the areas of market and opening beyond biases or common sense.

Having a strategic study of what areas are more suitable for starting a new business will help entrepreneurs in their search of the best location.

This is a strategic study in general terms that may help a further deeper study of the specific locations to open a new business.

The utility of this study case will result in a better economic recovery of the region after the events of the COVID-19 pandemic.

At the end of this report, the possible small businesses owners will see what Districts of Madrid are more suitable to open a new business if we consider Real State Prices (Rental), possible customers (Area population) and competence (Number of other small businesses in the area).

2 Data collection and sources

For this Data Science project we prepare a Data collection plan. In this case we download the information from websites of the local Public Administration, this data is updated and it comes from the end of 2019 to the summer of 2020. For this reason, this study report is updated.

First of all, we obtain data from `https://es.wikipedia.org/wiki/Anexo:DistritosdeMadrid`, from this webpage we obtain the different neighbourhoods of Madrid.

We obtain the second Source of information from `https://www.idealista.com/sala-de-prensa/informes-precio-vivienda/venta/madrid-comun`

idad/madrid-provincia/madrid/ where we will find the different prices of real state of the different neighbourhoods and districts in AUGUST 2020.

Finally, we obtain the third Source of information where we will find economic statistics of the different districts of Madrid and specifically its number of locals (in JULY 2020) from https://www.madrid.es/UnidadesDescentralizadas/UDCEstadistica/Nuevaweb/Econom%C3%ADa/Empresas%20y%20Locales/censo/D2110320.xlxs

All of this data together creates the Dataframe used for the purpose of this study, which will be used in a classification Algorithm.

3 Methodology

1) Defining purpose and goal. We define the goal and scope of the problem to solve: ***Strategic Study of the best districts to open any business***

2) Data Collection. All data is downloaded and it is presented in different Pandas Dataframes

3) Data preparation. Data is filtered and processed to create an unique Dataframe useful for the study purpose

4) Data Analysis. We create and develop the algorithm. In this case, we use a K-means algorithm to determine the most suitable neighbourhoods to open a small business. ***The K-means algorithm will see the similarity of the districts in three key areas: Real State Prices, Possible customers and competitors in the area***. As the comparison is defined we check for the clusters and centroid that presents a lower real state price for rental, more possible customers and less competitors in the area.

5) Study results Conclusions. ***The results will be shown in a map, that will represent the most suitable areas to open a small business and also a list of the best Districts to do so.***

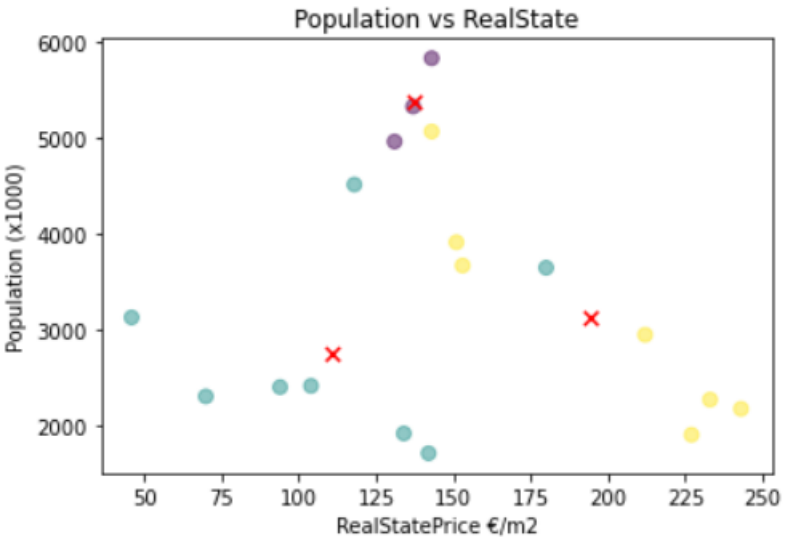
4 Results

The model process can be checked in GITHUB PAGE, there it can be followed the process and code used to determine the conclusions.

We use a K-Means algorithm to classify the different districts of Madrid in different types according with 3 main attributes: Real State prices, population and the number of other small business competitors in the area.

We classify the algorithm in 3 sets: Bad place (k=0), normal place (k=2), good place (k=1) to open a business comparing posible buyers, realstate prices and number of possible local businesses competitors.

By this way we obtain in the classification those districts with low real state prices, high population and low number of business competitors. This is concluded comparing the results of the K-means study in accordance with the following pictures.


Population vs RealState
Population (x1000)
RealStatePrice €/m2
6000
5000
4000
3000
2000
50
75
100
125
150
175
200
225
250


Prices vs Businesses
RealStatePrice €/m2
Number of Businesses
6000
5000
4000
3000
2000

5 Code

In this section we add the most significant code used in the development of this program.

In [ ]:
```
# Importing all necessary libraries
from pandas import DataFrame
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans

# Training and creating K-means
kdf = DataFrame(df,columns=['Population','RealStatePrice','Number of Businesses'])
kmeans = KMeans(n_clusters=3, random_state=0).fit(kdf)

# We store the clusters results, this will help in a map representation
df['Kmeans'] = kmeans.labels_
centroids = kmeans.cluster_centers_

print(centroids)
```

In [ ]:
```
plt.scatter(df['Population'], df['RealStatePrice'], c= kmeans.labels_.astype(float), s=50, alpha=0.5)
plt.scatter(centroids[:, 0], centroids[:, 1], c='red', s=50, marker='x')
plt.ylabel('Population (x1000)')
plt.xlabel('RealStatePrice €/m2')
plt.title('Population vs RealState')
plt.show()
```

In [ ]:
```
plt.scatter(df['Number of Businesses'], df['RealStatePrice'], c= kmeans.labels_.astype(float), s=50, alpha=0.5)
plt.ylabel('RealStatePrice €/m2')
plt.xlabel('Number of Businesses')
plt.title('Prices vs Businesses')
plt.show()
```

In [ ]:
```
import folium

MAD = folium.Map(location=[40.4167, -3.70325])
```

In [ ]:
```
# Here we will define some colors to the df and its after representation
def regioncolors(df):
    if df['Kmeans'] == 0:
        return 'red'
    elif df['Kmeans'] == 1:
        return 'green'
    else:
        return 'yellow'

df['color'] = df.apply(regioncolors, axis=1)
df.head()
```

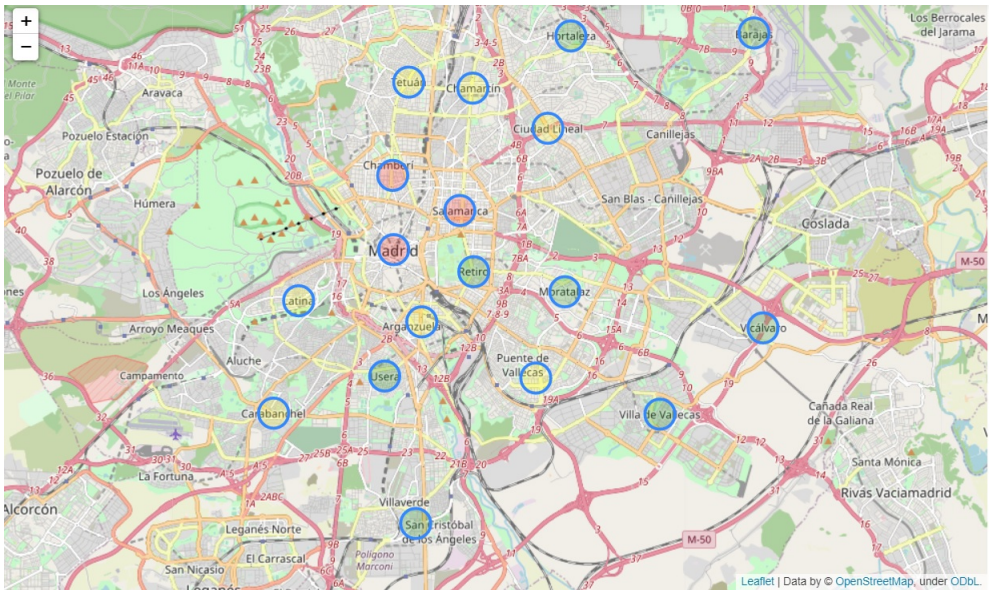
In [ ]:
```
locations = df[['latitude', 'longitude']]
locationlist = locations.values.tolist()
for i in range(0, len(df)):
    folium.CircleMarker(locationlist[i],radius=15, popup=df['Districts'][i], fill_color=df['color'][i]).add_to(MAD)

MAD
```

6 Map Result

In this section we can observe the results of the study. We can observe the most suitable districts highlighted with green color.

7 Conclusion

In general terms the most suitable districts to open a new small business whereas in the form of a restaurant, a shop or a clinic are: Hortaleza, Retiro, Moratalaz, Barajas, Vicálvaro and Cristobal de los Reyes.



If we added more data to the model such as average salaries of the districts we may find a similar and more accurate result that would define more specifically the bes districts for a new business in the city of Madrid.