Analysis of Valencian Municipalities for a new Bar

For IBM Data Science Capstone Project, 7th April 2021

This document will contain an introduction to the problem, a breakdown of the data required and links to relevant data.

GITHUB LINK TO JUPYTER NOTEBOOK (CODE ONLY):

https://github.com/andrew-wwg/Coursera_Capstone/blob/94faf2fa9eab9a784a827e2c6df79055b428ff45/Final%20ass%20-%20Capstone%20Project.ipynb

1. The problem

The results of this project will be relevant for people moving to the city in search of accomodation, business opportunities and places to visit. The project is specifically intended to help a prospective bar owner make an informed decision on the location of their new bar.

Valencia is a growing city in Spain. It has a unique value proposition for Spaniards and tourists alike. Beautiful beaches, a comparatively young demographic, popular traditional cuisine and a myriad of museums and historical sites it has the potential to become a major European city. Valencia does suffer from lack of investment from new projects however as Barcelona draws many tourists and business owners away. That being so, Valencia's weaker economy is an opportunity for businesses to start up with lower costs and get on at the first floor of a promising future of economic growth.

A bar owner is interested in opening a new bar in the city of Valencia, Spain. They want to open their bar in a region of the city where the average inhabitant's salary is high, but where there are not currently many bars and therefore limited competition.

2. Methodology

Data will be extracted from the Spanish Tax website which includes information about median salaries of inhabitants across postcodes in Spain. This dataset will be filtered so that we can focus on the postcode in the city of Valencia. This will be done by inspecting the HTML, saving as a text file, opening as an XLS and converting to a CSV file. Alternatively, BeautifulSoup may be used to scrape the data from the tax website. After data analysis, the top 5 discticts of Valencia will be investigated further using FourSquare data.

Using an open source dataset of coordinates for Valencian postcodes, a map will be generated to clearly mark the top 5 districts.

FourSquare Data

Foursquare location data will be used to explore the coordinates (longitude and latitude) of bars in these 5 municipalities. Using a map visualization, a suitable location for a future bar, far from competition can be determined.

In order the get salary data about different postcodes in Valencia it was necessary to extract information from the government website by inspecting the HTML as beautiful soup would not read the website in its current state. I therefore put the HTML into a textfile and opened it in a spreadsheet programme - from there I could convert it to a CSV file and filter it that way.

The second task was to combine this postcode salary data with coordinate data. Only once the data was merged did I realise that many of the places had been given the same generic coordinates. It was therefore important to get accurate up-to-date information using the Nominatim geolocator for each place and update the information in the dataframe.

Once all the data had been collected and cleaned the next step was to perform a Kmeans clustering analysis to determine the best area to open a bar based on location and salary data. After this analysis it was concluded that Pla de Remei would be an ideal location.

Next, using Foursquare, a map was generated to show the locations of bars within a 500m radius of Pla de Remei. This map allowed me to identify a few roads where there was an absence of bars and therefore an opportunity to open a new bar with demand in the borough.

These were the libraries used in the investigation:

In [1]:
```
import requests
import pandas as pd
pd.set_option("display.max_rows", 20, "display.max_columns", 20)
import numpy as np # library to handle data in a vectorized manner
import matplotlib.cm as cm
import matplotlib.colors as colors
import random # library for random number generation
import json # library to handle JSON files
import urllib
#The following packages are useful but I did not use here
#!pip install beautifulsoup4
#from bs4 import BeautifulSoup
#!pip install lxml

!pip install geopy
from geopy.geocoders import Nominatim # module to convert an address into latitude and longitude values

# libraries for displaying images
from IPython.display import Image
from IPython.core.display import HTML

# tranforming json file into a pandas dataframe library
from pandas.io.json import json_normalize

# import k-means from clustering stage
from sklearn.cluster import KMeans

! pip install folium==0.5.0
import folium # plotting library

print('Folium installed')
print('Libraries imported.')
```

I also adapted a section of code which was auto-generated by Back4app in order to fetch filtered data specifically from their Spain geodata API (which is where the data was located online):

In [4]:
```
#adding a condition to filter out the majority of unneccessary data for elsewhere in Spain
where = urllib.parse.quote_plus("""
{
    "Postal_Code": {
        "$gt": 45999
    }
}
""")
url = 'https://parseapi.back4app.com/classes/Spainpostalcode_Spain_Postal_Code?limit=50&order=Postal_Code&keys=Place_Name,
headers = {
    'X-Parse-Application-Id': 'L1J6TLuzAJOD0PTSbaxxAL6MumtHfwkyr2Fg41Xq', # This is your app's application id
    'X-Parse-REST-API-Key': '0W3mihPdwBy2bud7Tj7fSS4CLR223AX1Qii5zZYd' # This is your app's REST API key
}
rawdata = json.loads(requests.get(url, headers=headers).content.decode('utf-8')) # Here you have the data that you need
print(rawdata)
```

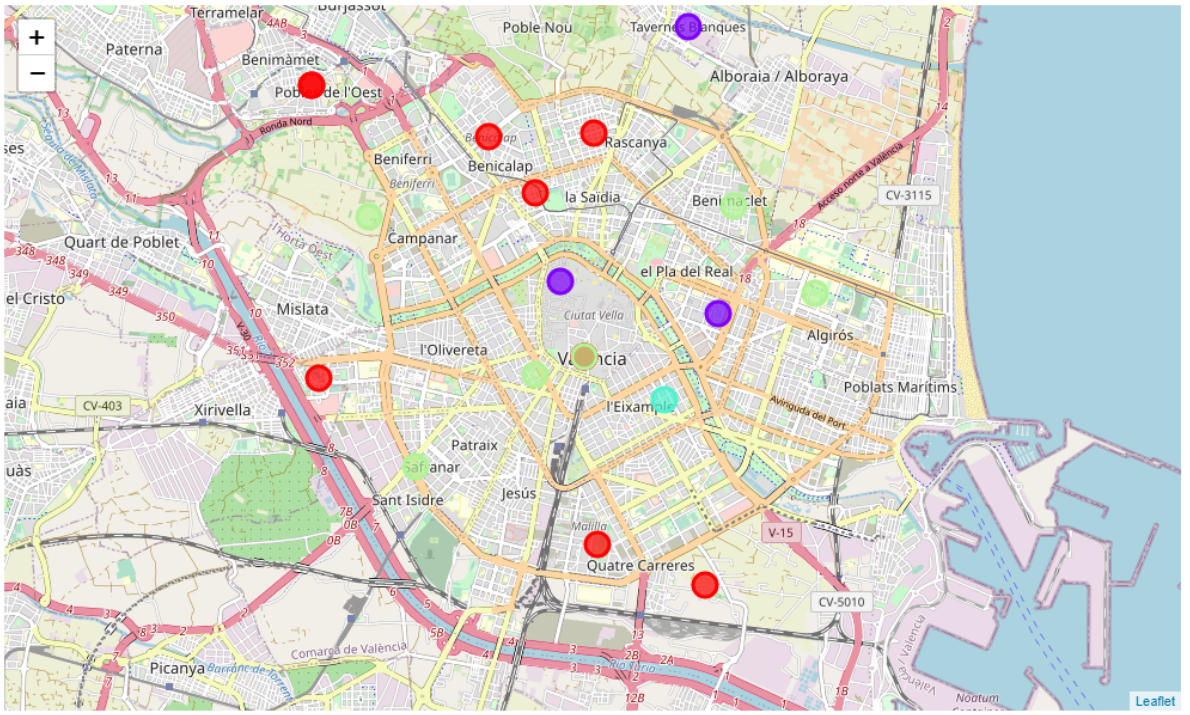
Sources of data:

1. Information on median salaries can be extracted from the Spanish Government's Taxation website: https://www.agenciatributaria.es/AEAT/Contenidos_Comunes/La_Agencia_Tributaria/Estadisticas/Publicaciones/sites/irpfCodPostal/2018/jrubikf241580c2986609e03ee3216d79d3f457701c254e.html
2. Information about the municipalties such as their longitude and latitude can be found using this freely available dataset: https://www.back4app.com/database/back4app/spain-zip-code-list

3. Results

Below is the Folium map showing kmeans clustering of postcodes in the city of Valencia where green and orange dots represent favorable zones based on salary (and therefore disposable income):

Out[94]:



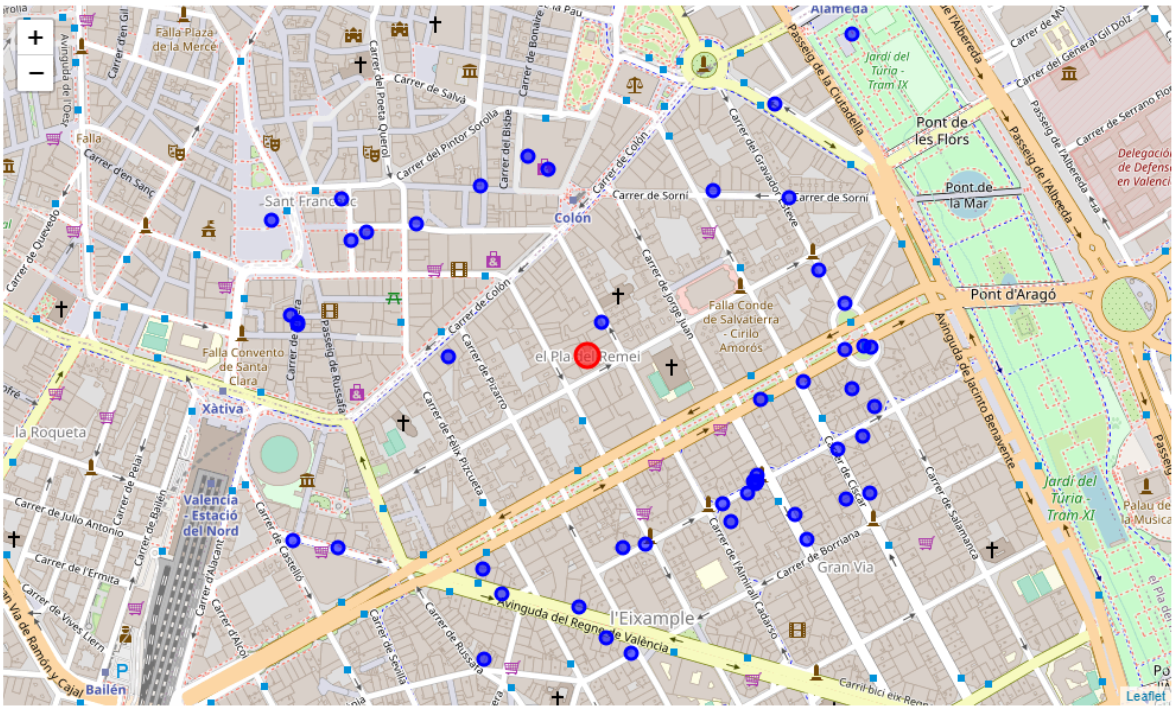
Here is the dataframe from which this map was generated. It shows that Pla de Remei, El Pilar, Extamurs and La Crus de Grao are the ideal locations as their salaries are significantly better than the other 25 barrios (boroughs). It also shows that postcodes 46004 and 46023 have the same coordinates. Through further investigation this was confirmed the be the case as they share a central governing building.

Out[95]:

	PostCode	Neighbourhood	NumberOfTaxpayers	MedianSalary	PlaceName	Latitude	Longitude	Cluster
3	46004	Pla de Remei	4917	72577	Valencia	39.469800	-0.377400	5
0	46001	El Pilar	5704	36408	Valencia	38.389457	-0.509706	4
7	46008	Extramurs	15422	34061	Valencia	39.467878	-0.384593	4
22	46023	La Crus Del Grao	15513	38053	Valencia	39.469800	-0.377400	4
20	46021	Algirós	17531	33191	Valencia	39.476849	-0.344421	4
...	...	...	...	...	...	...	...	...

The diagram below shows the location of the 50 ‘bar’s closest to the central coordinates of the borough selected for further investigation. Further analysis shows that the majority are in fact restaurants or betting shops.

Out[119]:



4. Discussion

The above diagrams seem to suggest 4 key points for discussion:

1. The richest neighborhoods of Valencia are located centrally close to the old historical centre and away from the coast or the outer limits of the city. Perhaps this is due to higher rents and therefore may not really mean higher disposable incomes. It could also be the case that the population are of a significantly older generation and less likely to be the target market of the proposed bar. Further analysis on postcode demographics would need to be conducted to determine this.

2. That there are really 4 neighborhoods with significantly higher median salaries (>34K) and therefore more disposable income (assuming similar living costs). Furthermore, there is one postcode in particular that has a much greater median salary at approximately twice the figure of the next highest. This seems almost anomalous as the other 28 neighborhoods all seem to fit a regular gradual curve of median salary distributions between 20K-36K. This would need to be verified possibly though a second round of quantitative analysis with another dataset or qualitatively by visiting the neighborhood.

3. The richest neighborhood also has a population significantly lower than other neighborhoods of comparable median salary. This may impact the business owner's decision to move as demand will be lower and a reliable customer base is essential to the success of this business like many others.

4. That the central point of this richest neighborhood has a very noticeable lack of bars. This does seems a little unusual and should be verified in person.

Conclusion

The results would seem to recommend that either the postcode 46004 or 46023 would be most sensible for opening a new bar. This is because the median salaries are good there, but also as these are neighboring postcodes the total population in this small neighborhood of Valencia is significant and not as low as could be initially inferred from the tax data. Having said that, there are 2 other neighborhoods where median salaries are above 33k and who belong to the same Kmeans cluster that might also be suitable location candidates.

In terms of competition, the folium map presenting Foursquare data on nearby business is tentatively promising. It appears to show that there is a significant area of the neighborhood which is not catered by a bar, and those nearby locations that have been identified as bars are not bars in the primary sense (they are restaurants or betting shops). This is promising, but should be investigated further either quantitatively through another round of analysis using a different database or qualitatively by visiting the neighborhood in person and making an assessment that way.

GITHUB LINK TO JUPYTER NOTEBOOK (CODE ONLY):
https://github.com/andrew-wwg/Coursera_Capstone/blob/94faf2fa9eab9a784a827e2c6df79055b428ff45/Final%20ass%20-%20Capstone%20Project.ipynb