The best and the worst Bucharest district to open a Pizzeria
Liviu Leordeanu
June 19, 2020

1. Introduction
1.1 Background
Data science can help businesses, this is common knowledge. Today the information is more and more accessible. It is easy to learn information about the number of restaurants opened in a district, the type of restaurants and the population density in a specific district
1.2 Problem
Data that might contribute to determine witch district is the best district to open a new pizza place
2. Data acquisition and cleaning
2.1 Data sources
I will use the following data sources:
A. infos about other pizzerias in Bucharest from Foursquare API
B. an cvs files that contains all the Postal codes in every district downloaded from: https://data.gov.ro/dataset/coduri-postale-romania/resource/29a5c89e-0f23-4a42-aa05-765ef04177b6
C. https://en.wikipedia.org/wiki/Sectors_of_Bucharest

2.2 Data cleaning
A. infos about other pizzerias in Bucharest from Foursquare API
First the data looked like this:

[73]:

	id	name	categories	referralId	hasPerk	location.address	location.lat	location.lng	location.labeledLatLngs	location.distanc
0	52de583a498e0f25c138ceaf	Pizzeria Al Ritrovo	[{'id': '4bf58dd8d48988d1ca941735', 'name': 'P...	v-1592517826	False	Piața Unirii nr. 1	44.428371	26.104164	[{'label': 'display', 'lat': 44.428371429443360...	83
1	4d80c04855fea1cd4ae911c5	Pizzeria Classic	[{'id': '4bf58dd8d48988d1ca941735', 'name': 'P...	v-1592517826	False	Bd. Unirii nr. 31, bl. A1	44.426887	26.105763	[{'label': 'display', 'lat': 44.426886878403860...	105

After:
- i kept only columns that include venue name, and anything that is associated with location,
- filtered the category for each row,
- cleaned column names by keeping only last term,
- droped the columns 'categories', 'lat', 'lng', 'labeledLatLngs', 'distance', 'cc', 'city', 'state', 'country', 'formattedAddress','crossStreet', ‘neighborhood' and ‘id'
- inserted the mising postal codes,
- droped the “adress" column,
- renamed columns ‘postalCode' into ‘Codpostal’,
- The API file looked like this:

B. https://en.wikipedia.org/wiki/Sectors_of_Bucharest
Initially the data look like this:

[40]:

	Tip artera	Denumire artera	Numar	Codpostal	Sector	Oficiu distribuire	SIRUTA SECTOR	NIV	SIRSUP
0	Stradă	Mincu Ion, arh.	nr. 21-T	11357	1	Bucureşti 2	179141	3	179132
1	Stradă	Mincu Ion, arh.	nr. 14-T	11359	1	Bucureşti 2	179141	3	179132
2	Stradă	Porumbaru Emanoil	nr. 1-25	11421	1	Bucureşti 2	179141	3	179132
3	Stradă	Porumbaru Emanoil	nr. 27-45	11422	1	Bucureşti 2	179141	3	179132
4	Stradă	Porumbaru Emanoil	nr. 47-69	11423	1	Bucureşti 2	179141	3	179132
...	...	...	...	...	...	...	...	...	...
12395	Intrare	Stavru Tudor	NaN	14074	1	Bucureşti 18	179141	3	179132
12396	Intrare	Talianu Ion	NaN	14075	1	Bucureşti 18	179141	3	179132
12397	Intrare	Atanasiu Niky	NaN	14076	1	Bucureşti 18	179141	3	179132
12398	Piaţă	Strasbourg	NaN	11818	1	Bucureşti 63	179141	3	179132
12399	Piaţă	Nicolau Irina	NaN	10226	1	Bucureşti 15	179141	3	179132

I began the cleaning of the data:
- dropping the unnecessary 'Tip artera', 'Denumire artera', 'Numar','Oficiu distribuire', 'SIRUTA SECTOR', 'NIV' and 'SIRSUP' columns
- turning df2.Codpostal and df1.Codpostal into the same format - string in order to merge the 2 data sets. After the merger the data frame 3 looked like this:

[ ]:

	name	Codpostal	Sector
0	Pizzeria Al Ritrovo	30119	3
1	Pizzeria Classic	30821	3
2	Pizzeria 3 Monelli	10784	1
3	Pizzeria Bellini	20082	2
4	Pizzeria Mamma Mia Crangasi - Giulesti	60286	6
5	Pizzeria da Michele	20616	2
6	Pizzeria David Obor	21151	2
7	Pizzeria Volare	14453	1
8	Pizzeria Firenze	21151	2
9	Pizzeria Due Amici	40342	4
10	Pizzeria Don Corleone	24102	2
11	Pizzeria Gili	31623	3
12	Pizzeria Athos 2	41322	4
13	Pizzeria Unirii	31281	3
14	Pizzeria Florina	41322	4

It had the restaurants from the FOURSQUARE API and POSTAL CODE and the District from the cvs.
After some more cleaning:
- removing the NaN
- grouping by district and counting the restaurants in every district
- renaming the columns 'name' into 'Number of pizzeria' and 'Sector' into 'District'

The 3rd dataframe was like this:

[46]:

	District	Number of pizzeria
0	1	5
1	2	7
2	3	6
3	4	4
4	5	1
5	6	2

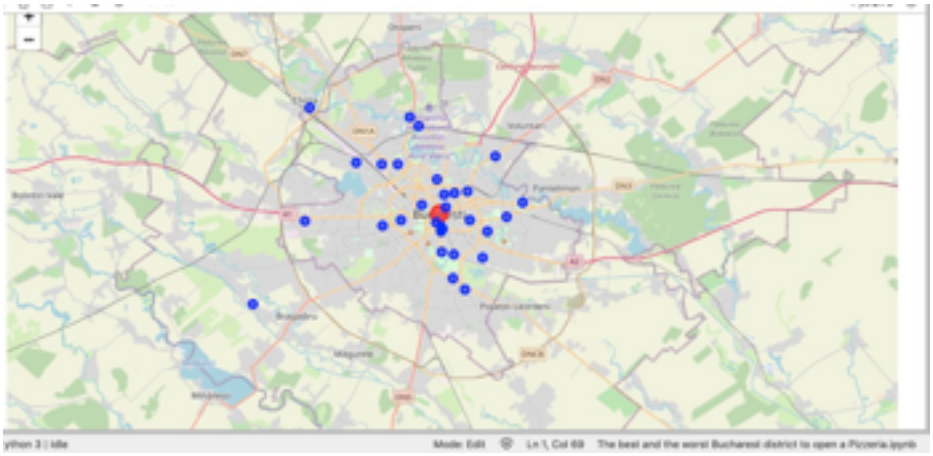
C. https://en.wikipedia.org/wiki/Sectors_of_Bucharest
After finding the Population density table in the page the dat frame look like this:

[30]:

	Rank	District	Population density
0	1\n	Sector 3	11,336\n
1	2\n	Sector 2	10,793\n
2	3\n	Sector 6	9,678\n
3	4\n	Sector 5	9,053\n
4	5\n	Sector 4	8,466\n
5	6\n	Sector 1	3,340\n

The cleaning was:
- replacing /n with empy space
- deleting ',' from the Population density numbers
- dropping the unnecessary Rank column

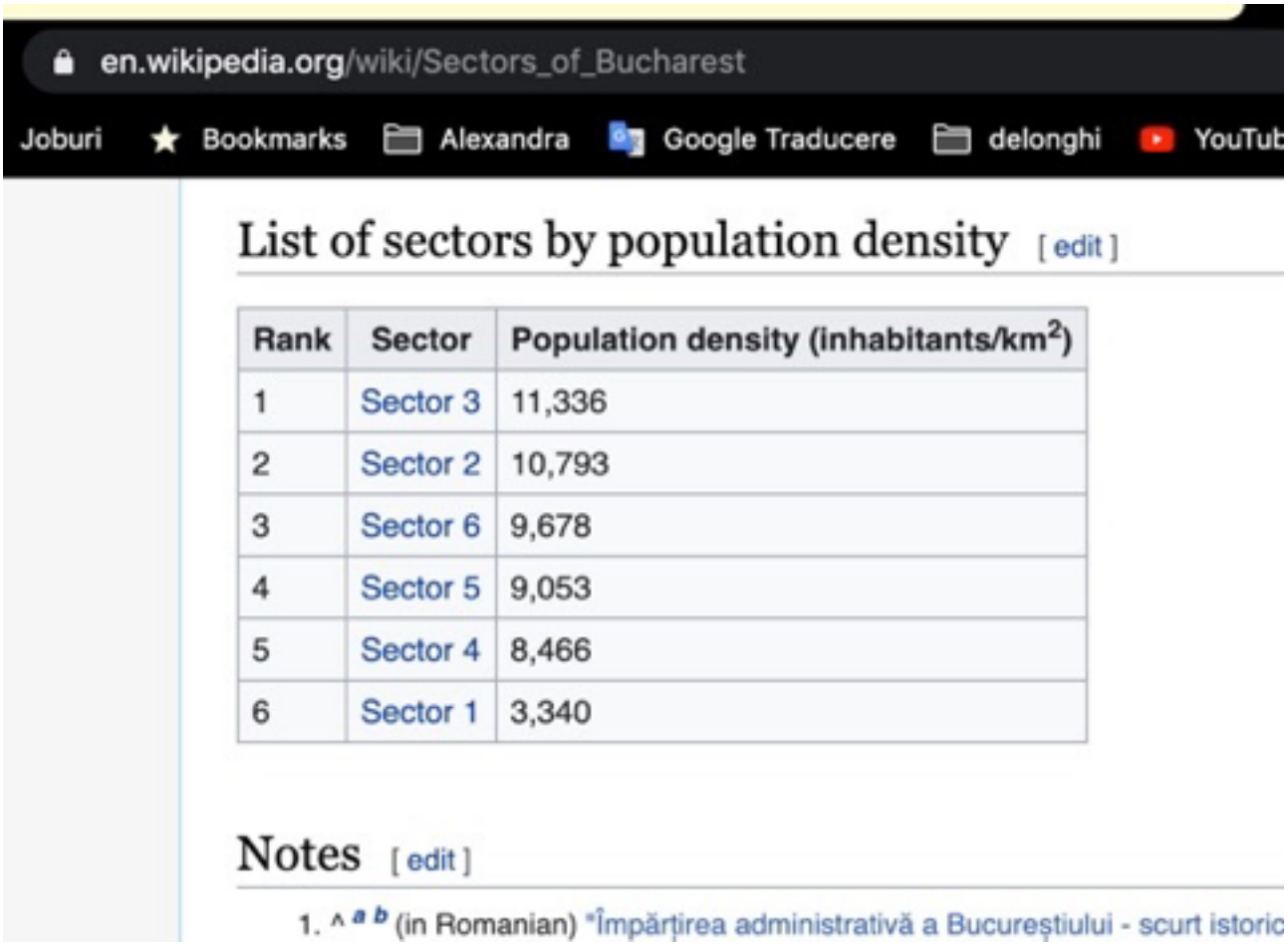
3. Exploratory Data Analysis
3.1 Visualise the Italian restaurants that are in the 1500 radius of Bucharest on the map



I used a folium map to place a red dot for Bucharest, and blue circles for every pizzeria in Bucharest.
For the location of the Pizzerias i used the column Longitude and Latitude from the Foursquare API

3.2 Scraping a wiki page for data

On this web page https://en.wikipedia.org/wiki/Sectors_of_Bucharest, I found some relevant information for my project:



This needed to be scraped so i used the BeautifulSoup library.
After importing the library i parsed the HTML from our URL into the BeautifulSoup parse tree format. After looking at the html code i found that the fourth table is relevant for my project, the one with the population density.
Then I looped through the rows and i constructed a dataframe containing the rank, the district and the population density.

3.3 Data normalisation

The ideal district is one where the number of pizzerias is the lowest and where population density is at its heightist.

In order to manipulate the population density and the number of pizzerias in every district i normalise the data. So the minimum population density became 0 and the maximum became 1. Same for the number of pizzerias in every district.

Then I subtracted the normalised number of pizzerias from the normalised population density in order to establish the top of the best Pizzeria districts.

5. Conclusions

After doing all the calculation looks like the worst district to open a pizzeria in Bucharest is DISTRICT 1 because it has the lowest density of population and the second number of pizzerias per district. So there are to manny pizza places for a district with not so manny people.

The best district to open a pizzeria in Bucharest is DISTRICT 5. It has the lowest number of pizzerias and the 4th populations density per district. So there are not enough pizzerias for a district that has a high population density.