Clustering Neighborhoods in Brasilia

1. Business Understanding

1.1 Description of the problem

Provide support to different visitors to list and visualize Brazilian districts that fit their needs in term of culinary venues.

1.2 Discussion of the background

In the streets of Brazil's bustling capital city, visitors will find an overwhelming number dining options. There's seemingly no limit to the varieties of Brasilia's gastronomy.

I believe it's difficult for a travelers, especially restaurant-goers, to make a choice from among many options since there is also too much information on the web because everybody's got their own take of where to go and it's all so fragmented that you have to assemble it yourself especially if you're wanting non-touristy recommendations.

How could we leverage Foursquare location data and machine learning to help us make decision and find appropriate neighborhoods? This is the problem I would like to address in this paper taking Brasilia (federal capital of Brazil) as an example. In this paper, I am going to use Foursquare location data and clustering methods to group the districts to different group by their restaurant venues information.

2. Data Requirements

For this project we need following data:

2.1 Brasilia data that contains list districts

Description: Administrative Regions (RAs) of the 31 districts:
http://www.seduh.df.gov.br/wp-conteudo/uploads/joomla/a8a55e3a8c7454a52db1d0c0a552b557.pdf

Data source: *https://github.com/emersonslima/coursera/blob/master/codeplan_dados.xlsx*

2.2 Brasília data containing latitude and longitude

Description: Coordinates of the 31 districts

Datasource: https://github.com/emersonslima/coursera/blob/master/coordenadas_df.xlsx

2.3 Restaurants in each neighborhood

Data source: Foursquare APIs

Description : By using this API we will get all the venues in each neighborhood. I can filter these venues to get only restaurants.

3. Methodology

In this paper Cross Industry Standard Process for Data Mining or Cross Industry Standard Process Model for Data Mining (CRISP-DM), which is one of the most common methodologies has been applied. CRISP-DM was developed as an open standard by leading KDD appliers and a tool supplier [1].

The current CRISP-DM Process Model for Knowledge Discovery in Databases (KDD) provides an overview of the life cycle of a project as shown in figure 1.

Figure 1: Process diagram showing the relationship between the different phases of CRISP-DM [2]

This methodology contains some steps, as it is shown in the figure 1, including business understanding, data understanding, data preparation, modeling, evaluation, and deployment [3];

→ Business Understanding: Understand the project objectives and requirements from a business perspective, and then convert this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

→ Data Understanding: Start by collecting data, then get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses about hidden information.

→ Data Preparation: Includes all activities required to construct the final data set (data that will be fed into the modeling tool) from the initial raw data. Tasks include table, case, and attribute selection as well as transformation and cleaning of data for modeling tools.

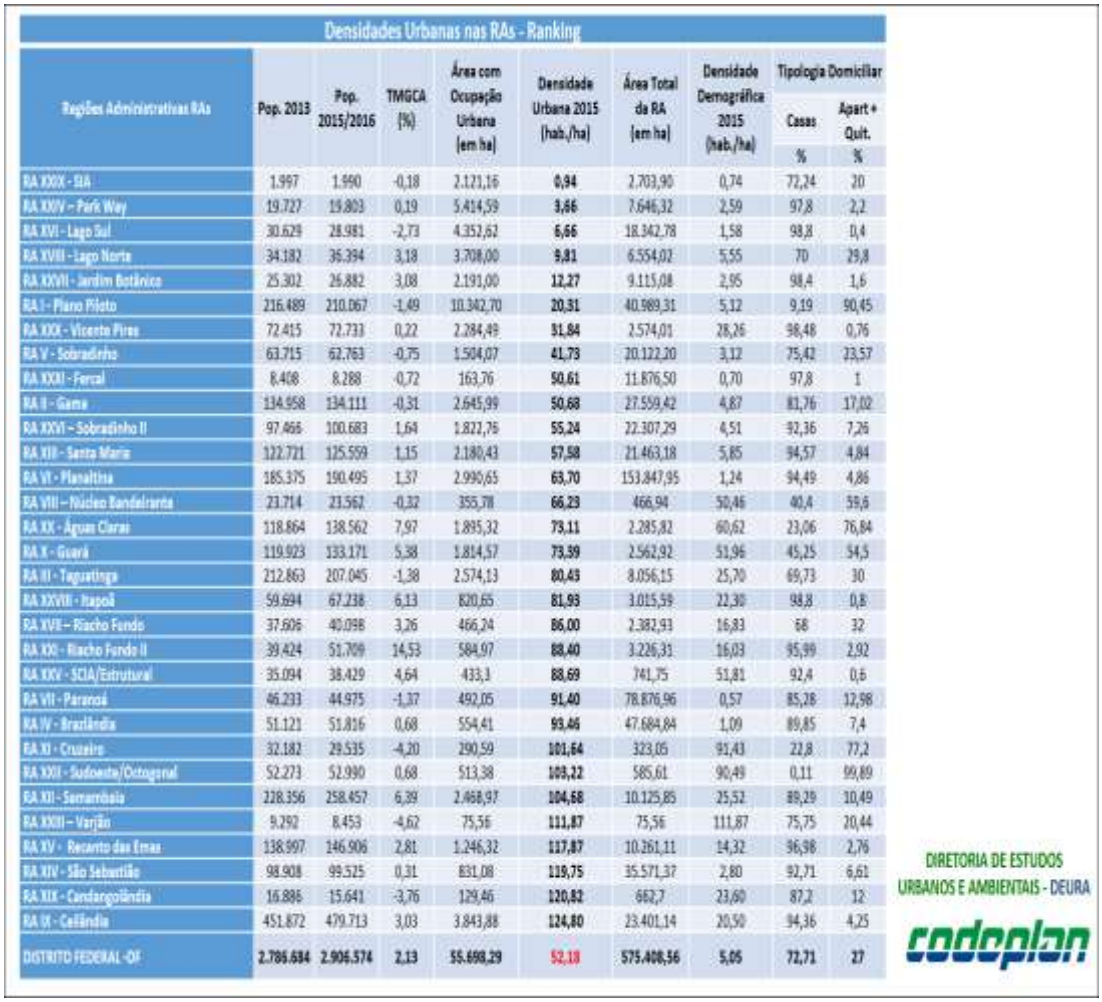
→ Modeling: Select and apply a variety of modelling techniques, and calibrate tool parameters to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.

→ Evaluation: Thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. Determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results is reached.

→ Deployment: Organize and present the results of data mining. Deployment can be as simple as generating a report or as complex as implementing a repeatable data mining process.

4. Data Preparation

To create a data-frame, I consider the occupied area of the RAs obtained at the Geoportal of the State Secretariat for Urban Development and Housing (SEDUH)1 as shown in figure 2.



Densidades Urbanas nas RAs - Ranking

Regiões Administrativas RAs	Pop. 2013	Pop. 2015/2016	TMGCA (%)	Área com Ocupação Urbana (em ha)	Densidade Urbana 2015 (hab./ha)	Área Total da RA (em ha)	Densidade Demográfica 2015 (hab./ha)	Tipologia Domiciliar	
								Casas %	Apart + Quit. %
RA XXIX - SIA	1.997	1.990	-0,18	2.121,16	0,94	2.703,90	0,74	72,24	20
RA XXIV – Park Way	19.727	19.803	0,19	5.414,59	3,66	7.646,32	2,59	97,8	2,2
RA XVI - Lago Sul	30.629	28.981	-2,73	4.352,62	6,66	18.342,78	1,58	98,8	0,4
RA XVIII - Lago Norte	34.182	36.394	3,18	3.708,00	9,81	6.554,02	5,55	70	29,8
RA XXVII - Jardim Botânico	25.302	26.882	3,08	2.191,00	12,27	9.115,08	2,95	98,4	1,6
RA I - Plano Piloto	216.489	210.067	-1,49	10.342,70	20,31	40.989,31	5,12	9,19	90,45
RA XXX - Vicente Pires	72.415	72.733	0,22	2.284,49	31,84	2.574,01	28,26	98,48	0,76
RA V - Sobradinho	63.715	62.763	-0,75	1.504,07	41,73	20.122,20	3,12	75,42	23,57
RA XXXI - Fercal	8.408	8.288	-0,72	163,76	50,61	11.876,50	0,70	97,8	1
RA II - Gama	134.958	134.111	-0,31	2.645,99	50,68	27.559,42	4,87	81,76	17,02
RA XXVI – Sobradinho II	97.466	100.683	1,64	1.822,76	55,24	22.307,29	4,51	92,36	7,26
RA XIII - Santa Maria	122.721	125.559	1,15	2.180,43	57,58	21.463,18	5,85	94,57	4,84
RA VI - Planaltina	185.375	190.495	1,37	2.990,65	63,70	153.847,95	1,24	94,49	4,86
RA VIII – Núcleo Bandeirante	23.714	23.562	-0,32	355,78	66,23	466,94	50,46	40,4	59,6
RA XX - Águas Claras	118.864	138.562	7,97	1.895,32	73,11	2.285,82	60,62	23,06	76,84
RA X - Guará	119.923	133.171	5,38	1.814,57	73,39	2.562,92	51,96	45,25	54,5
RA XII - Taguatinga	212.863	207.045	-1,38	2.574,13	80,43	8.056,15	25,70	69,73	30
RA XXVII - Itapoã	59.694	67.238	6,13	820,65	81,93	3.015,59	22,30	98,8	0,8
RA XVII – Riacho Fundo	37.606	40.098	3,26	466,24	86,00	2.382,93	16,83	68	32
RA XXI - Riacho Fundo II	39.424	51.709	14,53	584,97	88,40	3.226,31	16,03	95,99	2,92
RA XXV - SCIA/Estrutural	35.094	38.429	4,64	433,1	88,69	741,75	51,81	92,4	0,6
RA VII - Paranoá	46.233	44.975	-1,37	492,05	91,40	78.876,96	0,57	85,28	12,98
RA IV - Brazlândia	51.121	51.816	0,68	554,41	93,46	47.684,84	1,09	89,85	7,4
RA XI - Cruzeiro	32.182	29.535	-4,20	290,59	101,64	323,05	91,43	22,8	77,2
RA XXII - Sudoeste/Octogonal	52.273	52.930	0,68	513,38	103,22	585,61	90,49	0,11	99,89
RA XII - Samambaia	228.356	258.457	6,39	2.468,97	104,68	10.125,85	25,52	89,29	10,49
RA XXIII – Varjão	9.292	8.453	-4,62	75,56	111,87	75,56	111,87	75,75	20,44
RA XV - Recanto das Emas	138.997	146.906	2,81	1.246,32	117,87	10.261,11	14,32	96,98	2,76
RA XIV - São Sebastião	98.908	99.525	0,31	831,08	119,75	35.571,37	2,80	92,71	6,61
RA XIX - Candangolândia	16.886	15.641	-3,76	129,46	120,82	662,7	23,60	87,2	12
RA IX - Ceilândia	451.872	479.713	3,03	3.843,88	124,80	23.401,14	20,50	94,36	4,25
DISTRITO FEDERAL -DF	2.786.684	2.906.574	2,13	55.698,29	52,18	575.408,56	5,05	72,71	27

DIRETORIA DE ESTUDOS URBANOS E AMBIENTAIS - DEURA

codeplan

Figure 2: Urban densities in Administrative Regions

[1] http://www.seduh.df.gov.br/wp-conteudo/uploads/joomla/a8a55e3a8c7454a52db1d0c0a552b557.pdf

From the information in figure 2, the following a “**codeplan_dados.xlsx**” data-frame was created:

```
In [333]: df = pd.read_excel("codeplan_dados.xlsx")
          df
```

Out[333]:

	sequencial	Regiões Administrativas RAs	População	TMGCA (%)	Área com Ocupação Urbana (em ha)	Densidade Urbana (hab./ha)	Área Total da RA (em ha)	Densidade Demográfica (hab./ha)	Casas %	Apart + Quit. %
0	1	SIA	1990	-0.16	2121.18	0.94	2703.90	0.74	72.24	20.00
1	2	Park Way	19803	2.19	5414.89	3.66	7646.32	2.59	97.8	2.20
2	3	Lago Sul	28981	-2.73	4352.62	6.66	18342.78	1.58	99.6	0.40
3	4	Lago Norte	36394	3.16	3708.00	9.81	6554.02	5.55	79	20.80
4	5	Jardim Botânico	26882	3.88	2191.00	12.27	9115.08	2.95	98.4	1.60
5	6	Plano Piloto	210067	-1.48	10342.70	20.31	48869.31	4.12	9.16	90.46
6	7	Vicente Pires	72733	0.22	2284.49	31.84	2674.01	28.28	99.48	0.78
7	8	Sobradinho	62763	-0.75	1504.07	41.73	28122.28	2.12	75.42	23.57
8	9	Fercal	8288	-0.72	163.76	50.61	11876.58	0.70	97.8	1.00
9	10	Gama	134111	-0.31	2645.99	50.68	27589.42	4.87	81.79	17.02
10	11	Sobradinho II	100683	1.84	1822.78	55.24	22307.29	4.51	92.38	7.26
11	12	Santa Maria	125559	1.18	2180.43	57.58	21463.18	5.85	94.57	4.64
12	13	Planaltina	190495	1.37	2980.85	63.70	153847.98	1.24	94.49	4.88
13	14	Núcleo Bandeirante	23562	-0.32	355.79	66.23	466.94	50.46	40.48	9.60
14	15	Águas Claras	136662	7.97	1895.32	73.11	2206.62	60.62	23.88	75.84
15	16	Guará	133171	5.38	1814.57	73.39	2562.92	51.96	45.25	54.50
16	17	Taguatinga	207548	-1.38	2874.13	83.43	8086.15	25.70	69.73	30.00
17	18	Itapoã	67238	6.13	820.88	81.93	3019.89	22.30	98.3	0.80
18	19	Riacho Fundo	40098	3.26	466.24	88.90	2382.93	16.83	88	12.00
19	20	Riacho Fundo II	51709	14.53	594.97	88.40	3226.31	16.03	95.99	2.92
20	21	SCIA/Estrutural	38429	4.84	433.38	8.88	741.75	81.81	82.4	0.60
21	22	Paranoá	44975	-1.37	492.06	91.40	78876.96	0.57	89.28	12.98
22	23	Brazlândia	51816	0.68	554.41	93.46	47864.84	1.09	89.85	7.40
23	24	Cruzeiro	29535	-4.20	290.59	101.64	323.06	91.40	22.8	77.20
24	25	Sudoeste/Octogonal	53880	0.68	513.36	103.22	585.61	90.49	0.11	88.83
25	26	Samambaia	258457	5.39	2456.97	104.66	10125.66	25.52	89.29	10.49
26	27	Varjão	8453	-4.62	75.66	111.87	75.66	111.87	78.75	20.44
27	28	Recanto das Emas	146906	2.61	1246.32	117.87	10281.11	14.32	96.98	2.76
28	29	São Sebastião	99525	0.31	831.08	118.75	38871.37	2.60	92,718.81	NaN
29	30	Candangolândia	15641	-3.78	129.46	120.82	662.72	3.66	97.2	12.09
30	31	Ceilândia	479713	3.83	3843.66	124.80	23401.14	20.50	94.38	4.25

After little manipulation, the data frame is obtained as below:

```
df.head()
```

	sequencial	Regiao	Pouplacao	Densidade	Area
0	1	SIA	1990	0.94	2703.90
1	2	Park Way	19803	3.66	7646.32
2	3	Lago Sul	28981	6.66	18342.78
3	4	Lago Norte	36394	9.81	6554.02
4	5	Jardim Botânico	26882	12.27	9115.08

```
# print the number of rows of dataframe
df.shape
```

(31, 5)

4.1 Getting Coordinates

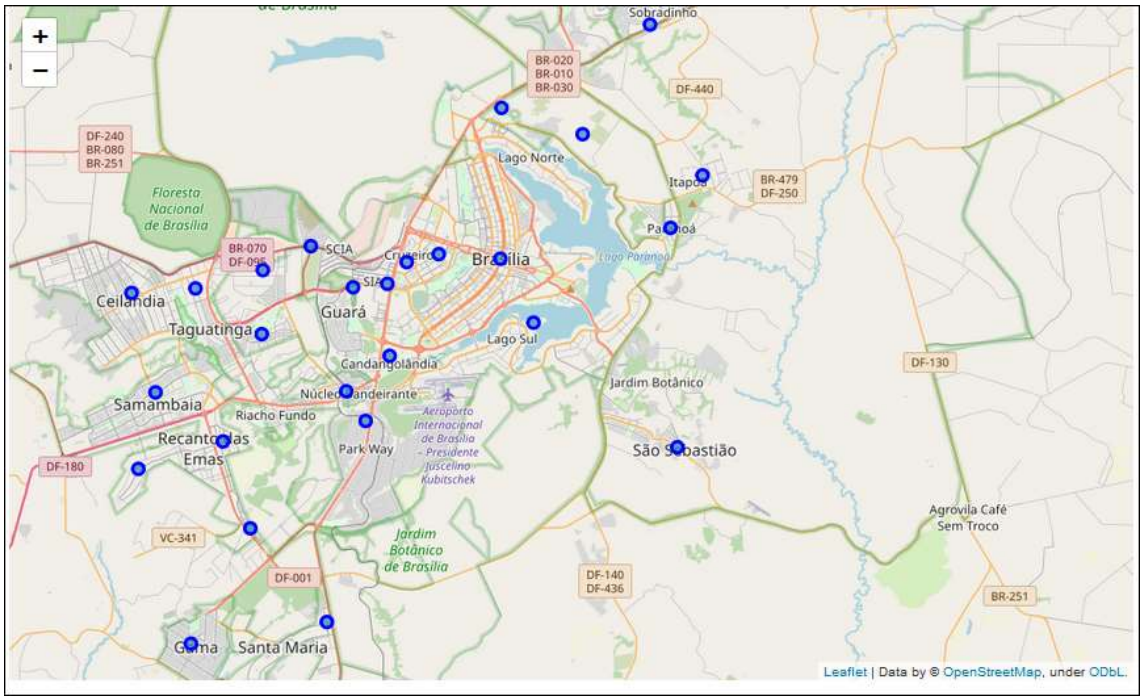
Next objective is to get the coordinates of these 31 districts using a "**coordenadas_df.xlsx**" dataframe:

```
In [216]: #Get data lat/long data from csv
          lat_long_coord_df = pd.read_excel("coordenadas_df.xlsx")
          lat_long_coord_df
```

Out[216]:

	Regiao	Latitude	Longitude
0	SIA	-15.8079	-47.9504
1	Park Way	-15.8872	-47.9635
2	Lago Sul	-15.8307	-47.8625
3	Lago Norte	-15.7213	-47.8329
4	Jardim Botânico	-25.4431	-49.2450
5	Plano Piloto	-15.7935	-47.8825
6	Vicente Pires	-15.8003	-48.0253
7	Sobradinho	-15.6580	-47.7925
8	Fercal	-15.5240	-47.7917
9	Gama	-16.0161	-48.0683
10	Sobradinho II	-15.6267	-47.8083
11	Santa Maria	-16.0036	-47.9872
12	Planaltina	-15.6216	-47.6522
13	Núcleo Bandeirante	-15.8699	-47.9753
14	Águas Claras	-15.8372	-48.0258
15	Guará	-15.8102	-47.9713
16	Taguatinga	-15.8107	-48.0658
17	Itapoã	-15.7455	-47.7609
18	Riacho Fundo	-15.8992	-48.0493
19	Riacho Fundo II	-15.9493	-48.0329
20	SCIA/Estrutural	-15.7863	-47.9968
21	Paranoá	-15.7757	-47.7799
22	Brazlândia	-15.6701	-48.2005
23	Cruzeiro	-15.7955	-47.9391
24	Sudoeste/Octogonal	-15.7912	-47.9196
25	Samambaia	-15.8706	-48.0902
26	Varjão	-15.7064	-47.8821
27	Recanto das Emas	-15.9151	-48.0999
28	São Sebastião	-15.9028	-47.7760
29	Candangolândia	-15.8496	-47.9490
30	Ceilândia	-15.8134	-48.1044

I used python folium library to visualize geographic details of Brasilia and its 31 regions and I created a map of Brasilia with boroughs superimposed on top. I used latitude and longitude values to get the visual as below:



4.2 Exploratory Data Analysis

I will use exploratory data analysis(EDA) to uncover hidden properties of data and provide useful insights to the reader

4.3 Using Foursquare Location Data

let's make use of Foursquare API and get the top 100 venues that are In the region of Aguas Claras.

I noticed that 18 unique venue categories were returned by Foursquare and Sandwich Place Restaurants in the top of the list.

```
print ('{} unique categories in Aguas Claras'.format(nearby_venues['categories'].value_counts().shape[0]))
```

```
18 unique categories in Aguas Claras
```

```
print (nearby_venues['categories'].value_counts()[0:10])
```

```
Sandwich Place            2
Bakery                    2
Pizza Place               2
Comfort Food Restaurant   1
Steakhouse                1
Pharmacy                  1
Ramen Restaurant          1
Health & Beauty Service   1
Pet Store                 1
Salon / Barbershop        1
Name: categories, dtype: int64
```

Analyzing each neighborhood to know about the top 5 venues of each one:

- Create a data-frame with pandas one hot encoding for the venue categories.

```
# add neighborhood column back to dataframe
ac_onehot['Neighborhood'] = ac_venues_only_restaurant['Neighborhood']

ac_onehot.head()
```

	Brazilian Restaurant	Central Brazilian Restaurant	Chinese Restaurant	Comfort Food Restaurant	Fast Food Restaurant	Goiano Restaurant	Italian Restaurant	Japanese Restaurant	Mineiro Restaurant	Northeastern Brazilian Restaurant	Ramen Restaurant	Restaurant	Seafoc Restaura
1	1	0	0	0	0	0	0	0	0	0	0	0	
2	1	0	0	0	0	0	0	0	0	0	0	0	
3	0	0	1	0	0	0	0	0	0	0	0	0	
4	1	0	0	0	0	0	0	0	0	0	0	0	
5	1	0	0	0	0	0	0	0	0	0	0	0	

- Use pandas groupby on neighborhood column and calculate the mean of the frequency of occurrence of each venue category.

```
ac_grouped = ac_onehot.groupby('Neighborhood').mean().reset_index()
ac_grouped
```

	Neighborhood	Brazilian Restaurant	Central Brazilian Restaurant	Chinese Restaurant	Comfort Food Restaurant	Fast Food Restaurant	Goiano Restaurant	Italian Restaurant	Japanese Restaurant	Mineiro Restaurant	Northeastern Brazilian Restaurant	Ramen Restaurant
0	Ceilândia	0.400000	0.000000	0.000000	0.0	0.200000	0.000000	0.000000	0.0	0.000000	0.000000	0.0
1	Cruzeiro	0.666667	0.000000	0.000000	0.0	0.000000	0.000000	0.333333	0.0	0.000000	0.000000	0.0
2	Gama	0.200000	0.000000	0.000000	0.0	0.600000	0.000000	0.000000	0.0	0.000000	0.000000	0.0
3	Guará	1.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0
4	Jardim Botânico	0.000000	0.000000	0.500000	0.0	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0
5	Núcleo Bandeirante	0.500000	0.166667	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.000000	0.166667	0.0
6	Paranoá	1.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0
7	Planaltina	0.666667	0.000000	0.000000	0.0	0.166667	0.000000	0.000000	0.0	0.000000	0.166667	0.0
8	Plano Piloto	0.285714	0.000000	0.000000	0.0	0.285714	0.000000	0.000000	0.0	0.000000	0.000000	0.0
9	Riacho Fundo	0.000000	0.000000	0.000000	0.0	0.500000	0.000000	0.000000	0.0	0.000000	0.000000	0.0
10	SIA	0.800000	0.000000	0.200000	0.0	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0
11	Samambaia	0.000000	0.000000	0.333333	0.0	0.000000	0.000000	0.000000	0.0	0.333333	0.000000	0.0
12	Sudoeste/Octogonal	0.000000	0.000000	0.000000	0.0	0.333333	0.000000	0.000000	0.0	0.000000	0.000000	0.0
13	São Sebastião	1.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0
14	Taguatinga	0.166667	0.000000	0.000000	0.0	0.166667	0.166667	0.166667	0.0	0.166667	0.000000	0.0
15	Vicão	1.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.0
16	Vicente Pires	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	1.0	0.000000	0.000000	0.0
17	Águas Claras	0.000000	0.000000	0.000000	0.5	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.5

- Output each neighborhood along with the top 5 most common venues

```
In [296]: num_top_venues = 5

for hood in ac_grouped['Neighborhood']:
    print("----"+hood+"----")
    temp = ac_grouped[ac_grouped['Neighborhood'] == hood].T.reset_index()
    temp.columns = ['venue','freq']
    temp = temp.iloc[1:]
    temp['freq'] = temp['freq'].astype(float)
    temp = temp.round({'freq': 2})
    print(temp.sort_values('freq', ascending=False).reset_index(drop=True).head(num_top_venues))
    print('\n')
```

```
4          Fast Food Restaurant   0.0


----Vicente Pires ----
                      venue  freq
0          Japanese Restaurant   1.0
1          Brazilian Restaurant   0.0
2  Central Brazilian Restaurant   0.0
3           Chinese Restaurant   0.0
4       Comfort Food Restaurant   0.0


----Águas Claras ----
                      venue  freq
0       Comfort Food Restaurant   0.5
1            Ramen Restaurant   0.5
2          Brazilian Restaurant   0.0
3  Central Brazilian Restaurant   0.0
4           Chinese Restaurant   0.0
```

I will use prescriptive analytics to help a traveler decide a location to go for a restaurant. I will use clustering (KMeans). I cluster the 31 regions based on the venue categories and use K-Means clustering. So, expectative would be based on the similarities of venue categories, these regions will be clustered:

```
# Cluster Neighborhoods
kclusters = 5

ac_grouped_clustering = ac_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(ac_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

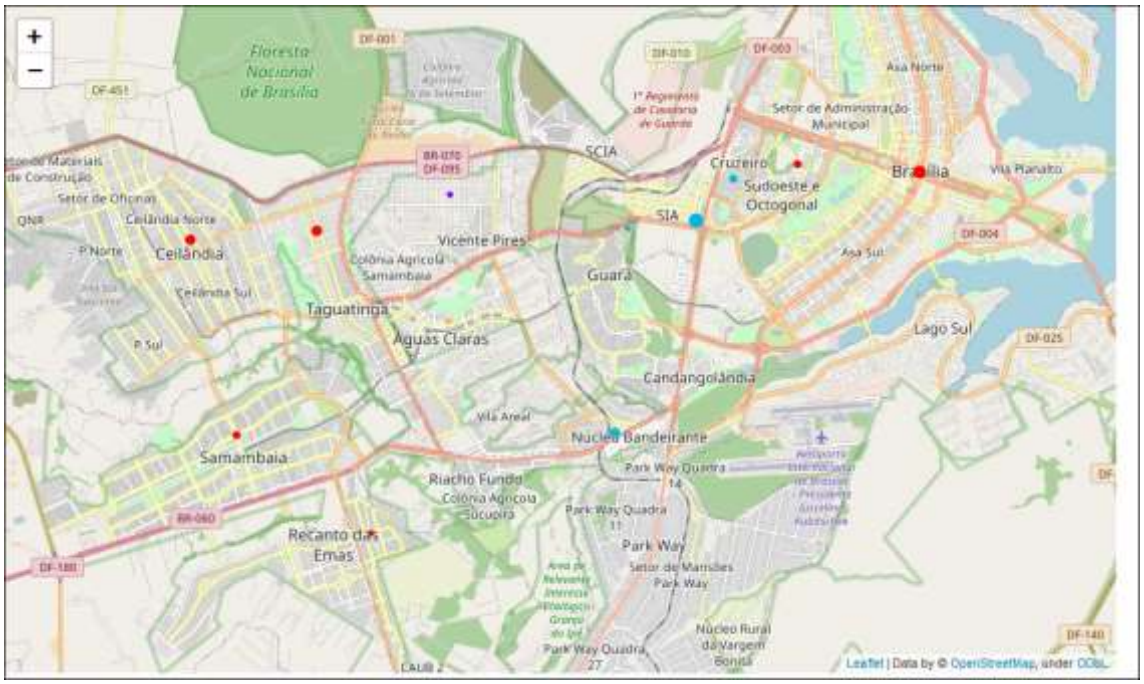
```
array([0, 2, 0, 2, 3, 2, 2, 2, 0, 0])
```

```
# create a new dataframe that includes the cluster as well as the top 10 venues for each neighborhood.
# add clustering labels

neighborhoods_venues_sorted.insert(0, 'Cluster Label', kmeans.labels_)

ac_merged = df_lat_long
```

I can represent these clusters in a leaflet map using Folium library



5. Results

I got a glimpse of the Restaurants in Brasilia and were able to find out some interesting insights which might be useful to travelers as well as people with business interests. Let's summarize our findings:

- Brazilian and Fast Food restaurants top the charts of most common venues in the 31 regions.
- Paranoá, São Sebastião, Varjão and Vicente Pires has the least number of restaurants.
- Plano Piloto has maximum number of restaurants.

The clustering is completely based on the most common venues obtained from Foursquare data. However, i ignored other factors like distance of the venues from closest stations, range of prices of restaurants and so on, since we don’t have such data and it would be difficult to farm it for a small exploratory study. Hence, our analysis only helps travelers to get an overview of Restaurants distribution by categories in the 31 regions districts of Brasilia.

6. Conclusion

Like seen in the example above, data was used to cluster neighborhoods in Brasilia based on the most common food venues (Restaurants) in its 32 regions. The results can help a traveler to decide about the region that fit the most his needs. I used Foursquare API to explore the major districts of Brasilia and saw the results of segmentation of regions using Folium leaflet map. Similarly, data can also be used to solve other problems, which most people face in metropolitan cities.

References

[1] Li, T., & Ruan, D. An extended process model of knowledge discovery in database. Journal of Enterprise Information Management. Available: https://www.deepdyve.com/lp/emerald-publishing/an-extended-process-model-of-knowledge-discovery-in-database-1jF0qjDCBE

[2] Kenneth Jensens, Process diagram showing the relationship between the different phases of CRISP-DM (2012), Available:

https://es.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining#/media/Archivo:CRISP-DM_Process_Diagram.png

[3] Yuan, Wang & Yihua, Zhang. (2009). Research on Classification and Subdivision Model of Telecom Rural Channel Based on Clustering Analysis. International Conference on Information Management, Innovation Management and Industrial Engineering. 3. 531-534. 10.1109/ICIII.2009.438. Available: https://ieeexplore.ieee.org/document/5369755

[4] Data Mining Concepts. Data Mining Process. Available: https://docs.oracle.com/cd/B19306_01/datamine.102/b14339/5dmtasks.htm