Capstone Project
“Battle of the Neighbourhoods”

1. Introduction and Business Understanding

1.1. Description of the problem

The problem is Mexico is a big city, and in order for a venue to be successful, said venue has to have a good location and depending on that location the type of Restaurant, in this case, it must be, therefore an analysis of the types of venue on certain locations can help make an informed decision.

1.2. Background

Mexico City (CDMX) is the capital and largest city of Mexico, and if that wasn’t enough it is also the most-populous city in North America, with more than 9 million people. It is known for being a hub for business, finance, arts, culture and economics. Mexico City offers a variety of cuisines: restaurants specializing in the regional cuisines of Mexico's 31 states are available in the city, and the city also has several branches of internationally recognized restaurants. CDMX is also known for having some of the freshest fish and seafood in Mexico's interior. La Nueva Viga Market is the second largest seafood market in the world after the Tsukiji fish market in Japan.

With more than 45,000 Restaurants in CDMX, if you are planning on opening a restaurant, get ready for a competitive market. New business ventures need to know how to play its cards, therefore they need to carefully examine the data gathered on the city, its population and its preferences to help reduce the risk of failure**.**

2. Data Requirements

2.1. Districts

The districts of CDMX were obtained from a federal database.
Source:
https://datos.cdmx.gob.mx/explore/dataset/coloniascdmx/table/?location=10,19.36105,-99.14801&dataChart=eyJxdWVyaWVzIjpbeyJjb25maWciOnsiZGF0YXNldCI6ImNvbG9uaWFzY2RteCIsIm9wdGlvbnMiOnt9fSwiY2hhcnRzIjpbeyJhbGlnbk1vbnRoIjp0cnVlLCJ0eXBlIjoiY29sdW1uIiwiZnVuYyI6IkFWRyIsInlBeGlzIjoiZW50aWRhZCIsInNjaWVudGlmaWNEaXNwbGF5Ijp0cnVlLCJjb2xvciI6IiM2NmMyYTUifV0sInhBeGlzIjoibm9tYnJlIiwibWF4cG9pbnRzIjo1MCwic29ydCI6IiJ9XSwidGltZXNjYWxlIjoiIiwiZGlzcGxheUxlZ2VuZCI6dHJ1ZSwiYWxpZ25Nb250aCI6dHJ1ZX0%3D)

2.2. Venues

The venues were located using the Foursquare API and filtering these venues for only Restaurants.

3. Methodology

3.1 Data preparation

The data obtained from the federal database was downloaded locally and inserted, to create a data frame, using pandas, to contain all the information such as District, Geolocalization, Federal Entity, etc.

```
df = pd.read_csv('/resources/coloniascdmx.csv')
df.head()
```

	COLONIA	ENTIDAD	Geo Point	Geo Shape	CVE_ALC	ALCALDIA	CVE_COL	SECC_COM	SECC_PAR
0	LOMAS DE CHAPULTEPEC	9.0	19.4228411174,-99.2157935754	{"type": "Polygon", "coordinates": [[[-99.2201...	16	MIGUEL HIDALGO	16-042	4924, 4931, 4932, 4935, 4936, 4940, 4987	4923, 4937, 4938, 4939, 4942
1	LOMAS DE REFORMA (LOMAS DE CHAPULTEPEC)	9.0	19.4106158914,-99.2262487268	{"type": "Polygon", "coordinates": [[[-99.2296...	16	MIGUEL HIDALGO	16-044	4963	4964
2	DEL BOSQUE (POLANCO)	9.0	19.4342189235,-99.2094037513	{"type": "Polygon", "coordinates": [[[-99.2082...	16	MIGUEL HIDALGO	16-026	NaN	4918, 4919
3	PEDREGAL DE SANTA URSULA I	9.0	19.314862237,-99.1477954505	{"type": "Polygon", "coordinates": [[[-99.1458...	3	COYOACAN	03-135	433, 500, 431, 513, 501	424, 425, 426, 430, 499
4	AJUSCO I	9.0	19.324571116,-99.1561602234	{"type": "Polygon", "coordinates": [[[-99.1585...	3	COYOACAN	03-128	376, 377, 378, 379, 404, 493, 498	374

After the data frame was created, it was manipulated to only contain the pertinent information such as District.

[7]:

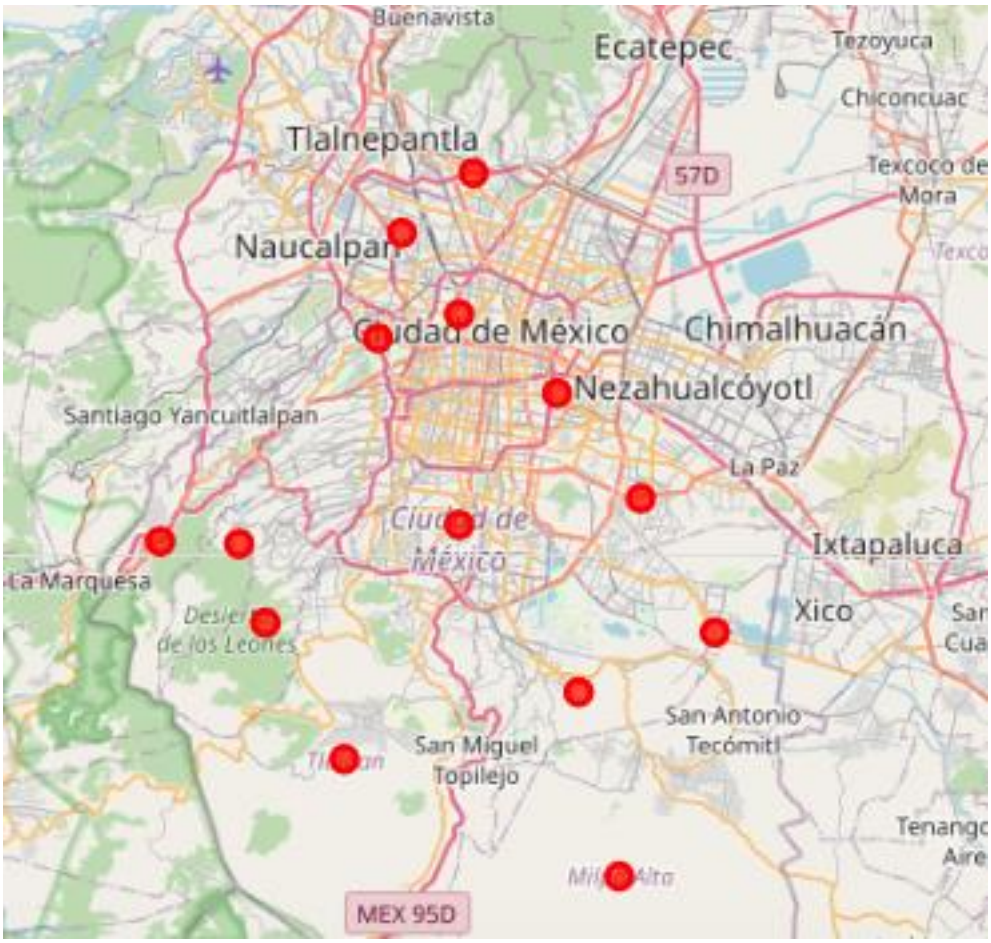
	ALCALDIA
0	MIGUEL HIDALGO
3	COYOACAN
6	VENUSTIANO CARRANZA
24	GUSTAVO A. MADERO
38	TLALPAN
46	XOCHIMILCO
50	MILPA ALTA
62	IZTACALCO
64	AZCAPOTZALCO
68	ALVARO OBREGON
71	CUAUHTEMOC
73	TLAHUAC
96	CUAJIMALPA DE MORELOS
97	IZTAPALAPA
113	BENITO JUAREZ
120	LA MAGDALENA CONTRERAS

Next, the coordinates for the districts were obtained using geocoder from geopy as seen below.

[8]:

	ALCALDIA	Latitude	Longitude
0	MIGUEL HIDALGO	19.429614	-99.198638
3	COYOACAN	19.328040	-99.151063
6	VENUSTIANO CARRANZA	16.308984	-92.637935
24	GUSTAVO A. MADERO	19.518545	-99.143640
38	TLALPAN	19.200877	-99.217012

The python library **Folium** was used to visualize the city with its districts as shown below.



3.2 Exploratory Data Analysis

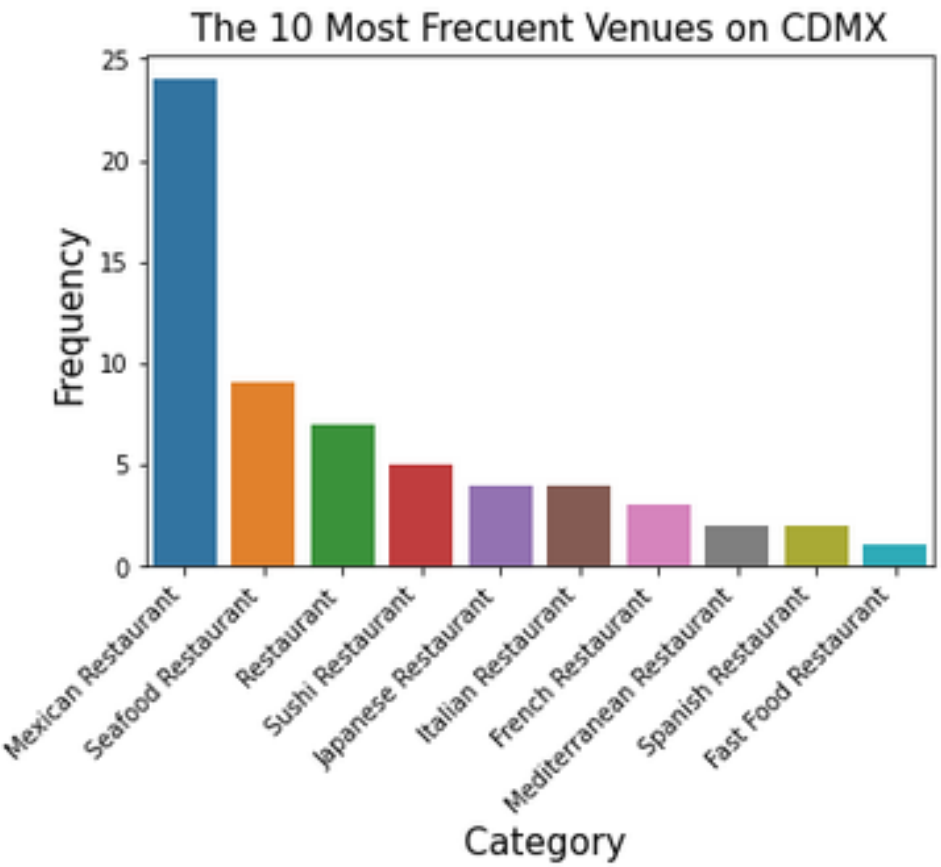
With the Foursquare API, we were able to observe some properties and insights of the data. For example, in the first district “Miguel Hidalgo”, we were able to observe that Mexican Restaurants are on top of the list.

```
Mexican Restaurant    9
Jewelry Store         5
French Restaurant     3
Bookstore             3
Italian Restaurant    3
Boutique              3
Ice Cream Shop        3
Hotel                 3
Park                  3
Cocktail Bar          2
Name: categories, dtype: int64
```

And that there are 17 different types of Restaurants in CDMX, if not more. I say if not more because there is a possibility that the many more “Restaurants” are labelled into another category

```
There are 17 uniques categories.
```

What can I say? We love our food, and the data shows that.



Now, let’s analyse each district so we can know more about the top venues in each one of them. So, we created a data frame with a one hot encoding for the venues category as shown below.

	City	African Restaurant	American Restaurant	Argentinian Restaurant	Brazilian Restaurant	Eastern European Restaurant	Fast Food Restaurant	French Restaurant	Italian Restaurant	Japanese Restaurant	Mediterranean Restaurant	R
1	MIGUEL HIDALGO	0	0	0	0	0	0	0	0	0	1	
2	MIGUEL HIDALGO	0	0	0	0	0	0	0	0	1	0	
3	MIGUEL HIDALGO	0	0	0	0	0	0	0	0	0	0	
4	MIGUEL HIDALGO	0	0	0	0	0	0	0	0	0	0	
5	MIGUEL HIDALGO	0	0	0	0	0	0	0	0	0	0	

We then, used the pandas groupby on the City Column and calculated the mean of the frequency of occurrence in each venue category as shown below.

	City	African Restaurant	American Restaurant	Argentinian Restaurant	Brazilian Restaurant	Eastern European Restaurant	Fast Food Restaurant	French Restaurant	Italian Restaurant	Japanese Restaurant	Mediterranean Restaurant
0	AZCAPOTZALCO	0.0	0.0	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.111111	0.000000
1	COYOACAN	0.0	0.0	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000
2	CUAUHTEMOC	0.0	0.0	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.200000
3	IZTACALCO	0.0	0.1	0.000000	0.000000	0.000000	0.1	0.000000	0.000000	0.000000	0.000000
4	IZTAPALAPA	0.0	0.0	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000
5	MIGUEL HIDALGO	0.0	0.0	0.032258	0.032258	0.032258	0.0	0.096774	0.096774	0.096774	0.032258
6	TLAHUAC	0.0	0.0	0.000000	0.000000	0.000000	0.0	0.000000	0.166667	0.000000	0.000000
7	XOCHIMILCO	1.0	0.0	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000

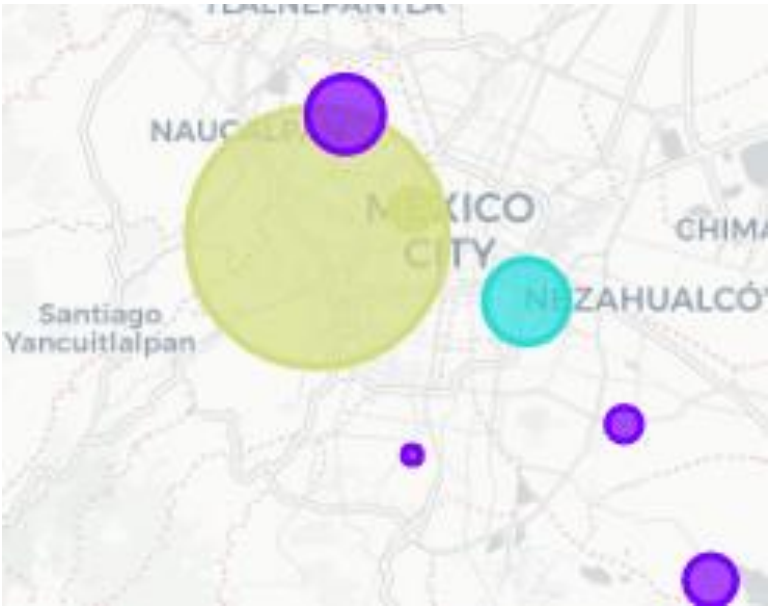
We then printed each district with the 5 top venues.

```
----AZCAPOTZALCO----
                      venue  freq
0         Mexican Restaurant  0.67
1         Seafood Restaurant  0.22
2        Japanese Restaurant  0.11
3  Mediterranean Restaurant  0.00
4           Sushi Restaurant  0.00


----COYOACAN----
                      venue  freq
0         Seafood Restaurant   0.5
1         Mexican Restaurant   0.5
2         African Restaurant   0.0
3  Mediterranean Restaurant   0.0
4           Sushi Restaurant   0.0


----CUAUHTEMOC----
                      venue  freq
0         Mexican Restaurant   0.4
1  Mediterranean Restaurant   0.2
2           Sushi Restaurant   0.2
3         Seafood Restaurant   0.2
4         African Restaurant   0.0
```

We then used prescriptive analytics to help decide where to build its new Restaurant, in this case we used kmeans to accomplish that. We then tried to cluster these districts based on the venues category as shown below.

	District	Latitude	Longitude	Cluster Labels
0	MIGUEL HIDALGO	19.429614	-99.198638	3.0
3	COYOACAN	19.328040	-99.151063	1.0
46	XOCHIMILCO	19.236978	-99.082300	0.0
62	IZTACALCO	19.398975	-99.095312	2.0
64	AZCAPOTZALCO	19.485815	-99.184206	1.0

And finally, we represented these clusters on a map using the Folium library as shown below.



4. Results and Discussion

We can have a glimpse at the Restaurant category in CDMX and some interesting insight that may help stakeholders take a better-informed decision about the location of a new Restaurant and its type of food it serves. We can say that indeed Mexicans love their food so much it is on top of the chart in many districts. I acknowledge that many factors have not been taken into account, such as range of prices, Michelin Stars, etc., and that is because of the difficulty it involves mining such data.

5. Conclusion

Many problems can be solved using data, in a globalized world data can be found almost anywhere. As shown here, data mined from federal databases as well as public ones, can be used to solve interesting problems. Here, we used the data to cluster districts in Mexico City based on the most common venue. This result can help stakeholder plan more carefully a new venue and predict where said venue is going to be more successful. With the help of the python libraries, the internet community and sufficient data, many great things can be accomplished. And last but not least, I also known that this work may not be either perfect or accurate, however I think it´s a step in the right direction.