Study to Open a Restaurant Curitiba – PR, Brazil

Sidney Comandulli

March 21, 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

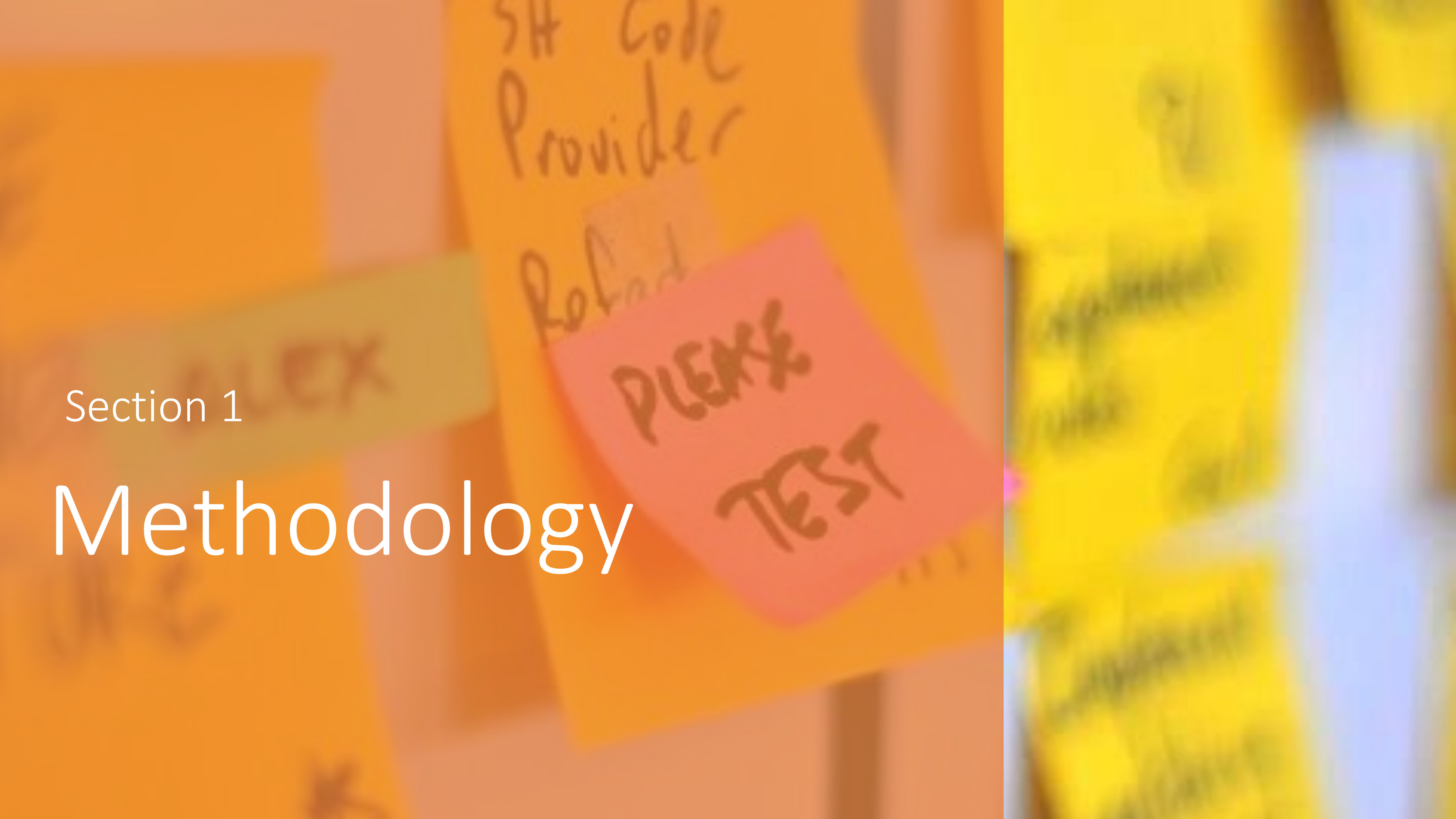
- This work intends to help the investor chose the best place to open a new restaurant in Curitiba
- For this, data will be collected from the web for launches (web scrapping), pre-treated (standardized) and trained in an unsupervised classification model using the K-Means algorithm.
- The results showed that through clustering it is possible to find not only the location, but also to guide the investor in the most appropriate type of restaurant to open.

Introduction

According to FourSquare API, there are more than 1800 restaurants in Curitiba and about 1,900,000 people (2022). That is why opening a new restaurant there can be an extremely challenging task. Choosing a restaurant type and a good spot, an entrepreneur usually carelessly relies on common sense and domain knowledge. Needless to say that too often an inconsiderate decision leads to a poor income and inevitable bankruptcy. According to several surveys, up to 40% of such start-ups fail in the very first year. Let's suppose, an investor has enough time and money, as well as a passion to open the best eating spot in Curitiba. What type of restaurant would it be? What would be the best place for it? Is there a better way to answer these questions rather than guessing?

What if there is a way to cluster city neighborhoods, based on their near-by restaurant similarity? What if we can visualize these clusters on a map? What if we might find what type of restaurant is the most and least popular in each location? Equipped with that knowledge, we might be able to make a smart choice from a huge number of restaurant types and available places.

Let us allow machine learning to get the job done. Using reliable venue data, it can investigate the city neighborhoods, and show us unseen dependencies. Dependencies that we are not aware of.



Section 1

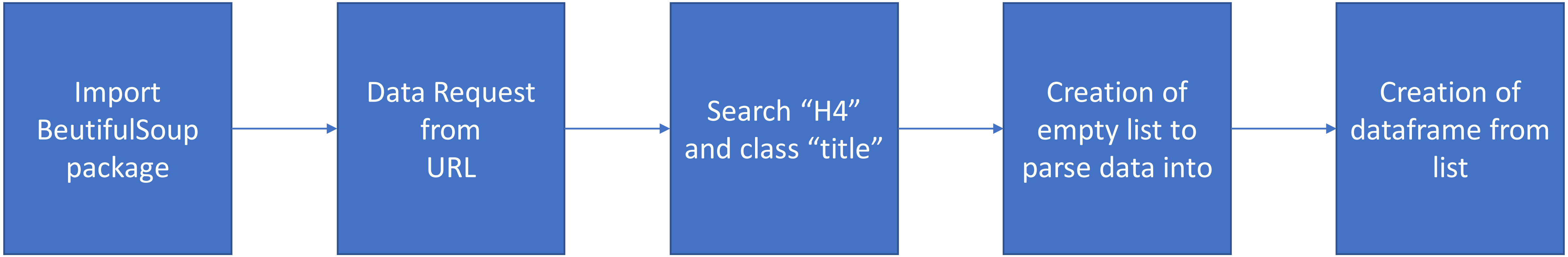
Methodology

Methodology

- **Data collection methodology**
  - Using a table on https://cepbrasil.org/parana/curitiba, collect information about Curitiba neighborhoods.
- **Data analysis**
  - Use the Geopy and Folium libraries to get the coordinates of all locations and map geospatial data on a Curitiba map.
  - Using Foursquare API, collect the top 100 restaurants and their categories for each location within a 500 meter radius.
- **Perform exploratory data analysis (EDA) using visualization and SQL**
- **Perform Clustering Data**
  - Group collected restaurants by location and by taking the mean of the frequency of occurrence of each type, preparing them for clustering.
  - Cluster restaurants by k-means algorithm and analyze the top 10 most common restaurants in each cluster.
- **Perform interactive visual analytics using Folium and Plotly Dash**
  - Visualize clusters on the map, thus showing the best locations for opening the chosen restaurant.

Data Collection

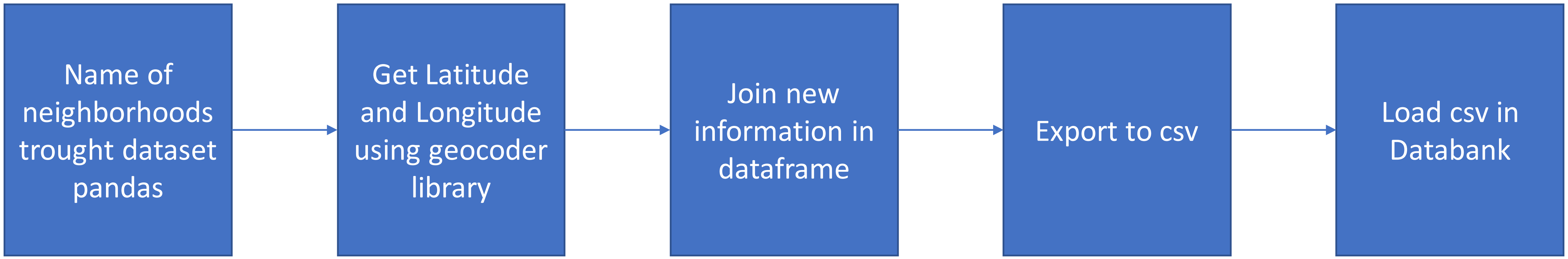
- Curitiba neighborhoods were web scraped with BeautifulSoup. Data was extracted from a webpage ([cepbrasil.org]) and parsed into a Pandas dataframe.



Complete notebook in: capstone-ibm-project/Capstone_Project Curitiba Final Version.ipynb at main · sdncmndll/capstone-ibm-project (github.com)

Data Collection - Geocoder

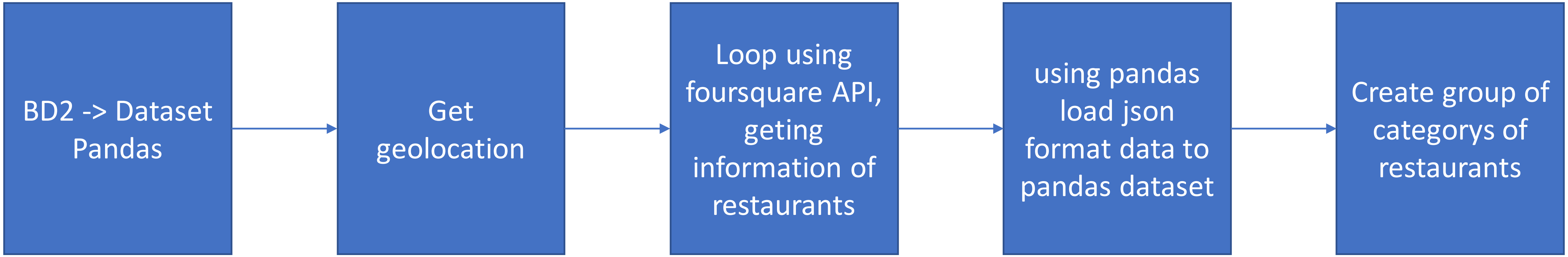
- After capturing the neighborhoods of Curitiba, we took the respective geolocations
- Data is persisted in a DB2 IBM database for use in the application.



Complete notebook in: [capstone-ibm-project/Capstone_Project Curitiba Final Version.ipynb at main · sdncmndll/capstone-ibm-project (github.com)]

Data Collection - Foursquare

- Using the data from the database with neighborhoods and geolocations, we will perform a search on the foursquare API



Complete notebook in: [capstone-ibm-project/Capstone_Project_Curitiba_Final_Version.ipynb at main · sdncmndll/capstone-ibm-project (github.com)]

EDA with Data Visualization

- The datas collect are from Curitiba city ?
- What kind of category of restaurant appear with more frequency ?
- What is the best parameter value to use in k in K-Means ?
- How was the result of clustering looking in the map of Curitiba city ?

Complete notebook in: [capstone-ibm-project/Capstone_Project_Curitiba_Final_Version.ipynb at main · sdncmndll/capstone-ibm-project (github.com)]

EDA with SQL

- In this study we used the database to store geolocation data for each neighborhood.
- The objective was to use the information that does not change in a database in the cloud

Complete notebook in: [capstone-ibm-project/Capstone_Project_Curitiba_Final_Version.ipynb at main · sdncmndll/capstone-ibm-project (github.com)]

Build an Interactive Map with Folium

- Using Folium and lat/long data all neighborhood were marked in a map. A circle and a marker were added to the map.

- On a second map was possible to include different colors for markers considering the clustering

Complete notebook in: [capstone-ibm-project/Capstone_Project_Curitiba_Final_Version.ipynb at main · sdncmndll/capstone-ibm-project (github.com)]

Clustering

- Using seaborn and the K-Means algorithm, we pass information about the occurrences of types of restaurants.
- With an initial K parameter, we evaluate the optimal k parameter through the Silhouette Score
- We show the clusters on the map and in lists that allow evaluating the occurrences of each type of restaurant

Complete notebook in: capstone-ibm-project/Capstone_Project_Curitiba_Final_Version.ipynb at main · sdncmndll/capstone-ibm-project (github.com)

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Clustering analysis results

Complete notebook in: [capstone-ibm-project/Capstone_Project_Curitiba_Final_Version.ipynb at main · sdncmndll/capstone-ibm-project (github.com)]

Section 2

Insights drawn from EDA

Table of neighborhoods with geolocation

```
#query statement to retrieve all rows in CWB_DATA_DB table
selectQuery = "select * from cwb_data_db"

#retrieve the query results into a pandas dataframe
cwb_data_db = pd.read_sql(selectQuery, pconn)
cwb_data_db.head(10)
```

	NEIGHBOURHOOD	LATITUDE	LONGITUDE
0	Abranches ...	-25.37028	-49.27007
1	Água Verde ...	-25.44746	-49.28556
2	Ahú ...	-25.40486	-49.26329
3	Alto Boqueirão ...	-25.52542	-49.24917
4	Alto da Glória ...	-25.41970	-49.26181
5	Alto da Rua XV ...	-25.42645	-49.25011
6	Área Rural de Curitiba ...	-25.43998	-49.27654
7	Atuba ...	-25.43333	-49.23333
8	Augusta ...	-25.45520	-49.37563
9	Bacacheri ...	-25.39847	-49.23038

By scrapping the CEP site, combining data with the geopy library, it was possible to record the data in a DB2 database in the cloud

Complete notebook in: [capstone-ibm-project/Capstone_Project Curitiba Final Version.ipynb at main · sdncmndll/capstone-ibm-project (github.com)]

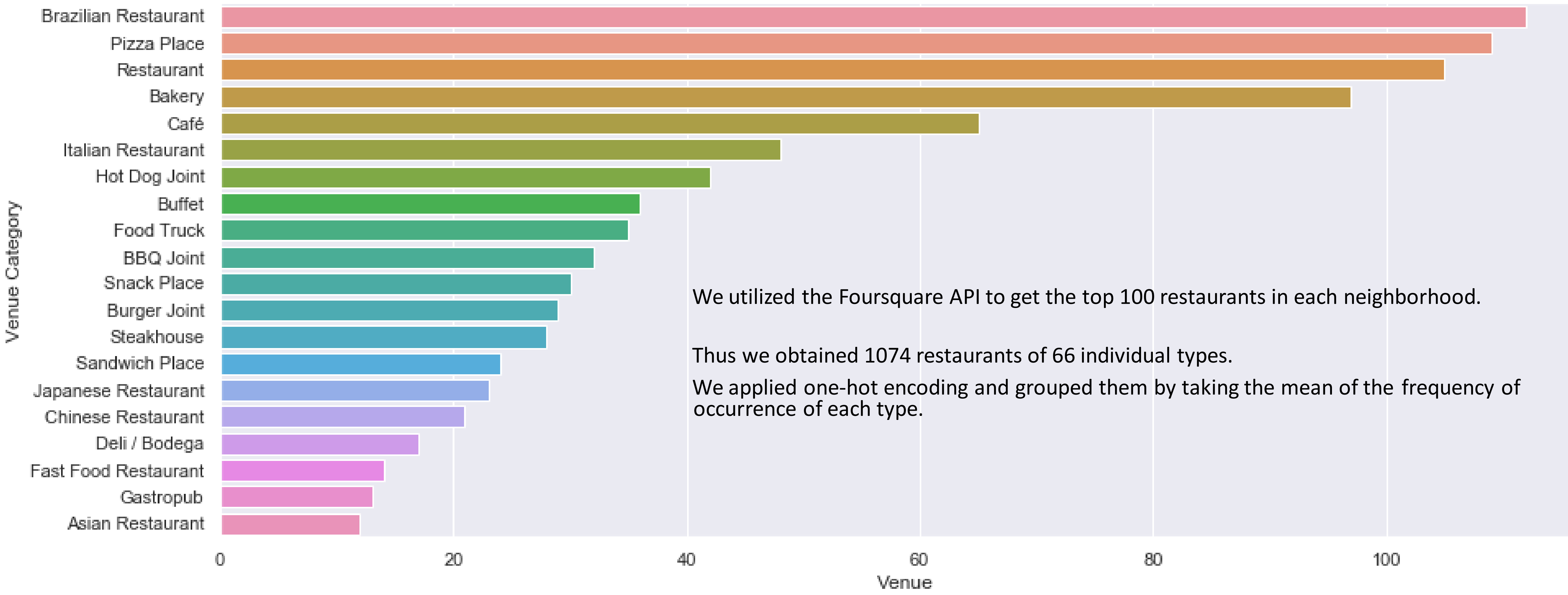
Map of neighborhoods

We can be seen that the neighborhoods were all correctly plotted on the map.

Curitiba is a city located in the south of Brazil

Complete notebook in: [capstone-ibm-project/Capstone_Project_Curitiba_Final_Version.ipynb at main · sdncmndll/capstone-ibm-project (github.com)]

Exploring Curitiba Restaurants



We utilized the Foursquare API to get the top 100 restaurants in each neighborhood.

Thus we obtained 1074 restaurants of 66 individual types.

We applied one-hot encoding and grouped them by taking the mean of the frequency of occurrence of each type.

Complete notebook in: [capstone-ibm-project/Capstone_Project Curitiba Final Version.ipynb at main · sdncmndll/capstone-ibm-project (github.com)]

Preparing Matrix to Clustering

Next, let's group rows by neighborhood and by taking the mean of the frequency of occurrence of each category preparing the dataframe for clustering.

In [135…

```
cwb_grouped = cwb_onehot.groupby('Neighborhood').mean().reset_index()
cwb_grouped
```

Out[135…

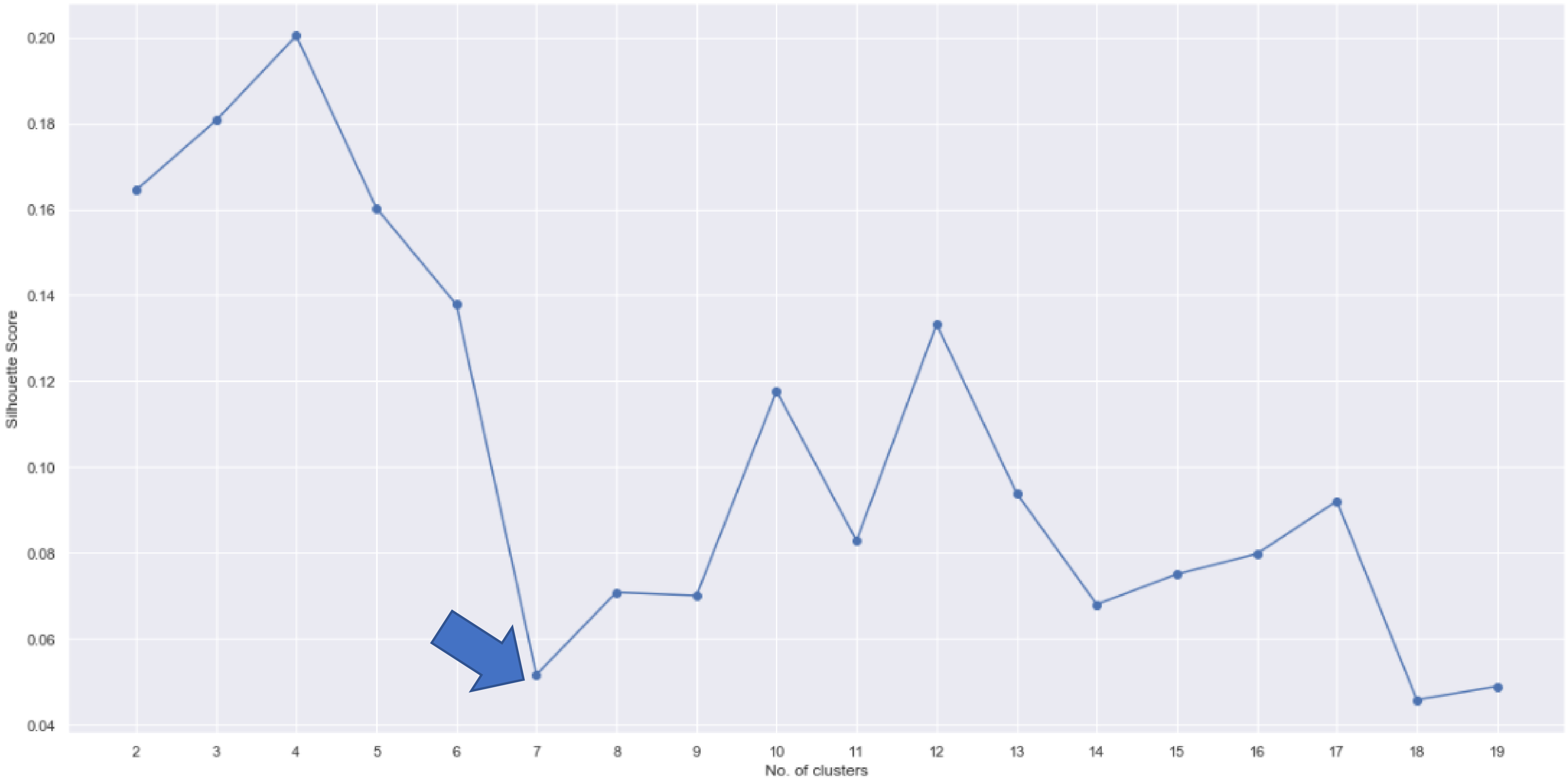
	Neighborhood	Acai House	Afghan Restaurant	American Restaurant	Argentinian Restaurant	Asian Restaurant	BBQ Joint	Bagel Shop	Bakery	Bistro	Brazilian Restaurant	Breakfast Spot	Buffet	Burger Joint	Cafeteria	Café
0	Abranches	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.750000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1	Ahú	0.000000	0.000000	0.000000	0.000000	0.000000	0.052632	0.000000	0.105263	0.000000	0.052632	0.052632	0.052632	0.000000	0.000000	0.052632
2	Alto Boqueirão	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.062500	0.000000	0.062500	0.062500	0.000000	0.062500	0.000000	0.000000
3	Alto da Glória	0.000000	0.000000	0.000000	0.000000	0.027027	0.027027	0.000000	0.108108	0.027027	0.108108	0.000000	0.081081	0.000000	0.000000	0.108108
4	Alto da Rua XV	0.000000	0.000000	0.000000	0.000000	0.000000	0.076923	0.000000	0.076923	0.019231	0.038462	0.000000	0.038462	0.019231	0.000000	0.038462
5	Atuba	0.000000	0.000000	0.000000	0.000000	0.000000	0.083333	0.000000	0.083333	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.083333
6	Augusta	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.333333	0.000000	0.000000	0.000000	0.000000	0.000000	0.333333	0.000000
7	Bacacheri	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.142857	0.000000	0.000000	0.000000
8	Bairro Alto	0.000000	0.000000	0.000000	0.000000	0.000000	0.090909	0.000000	0.272727	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
9	Barreirinha	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.125000	0.000000	0.125000	0.000000	0.000000	0.000000	0.000000	0.000000
10	Batel	0.029412	0.000000	0.000000	0.000000	0.029412	0.000000	0.000000	0.029412	0.000000	0.029412	0.029412	0.088235	0.029412	0.000000	0.147059
11	Bigorrilho	0.000000	0.000000	0.020833	0.000000	0.020833	0.062500	0.000000	0.062500	0.000000	0.104167	0.000000	0.000000	0.020833	0.000000	0.083333
12	Boa Vista	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.200000	0.000000	0.000000	0.000000	0.000000	0.200000	0.000000	0.000000
13	Bom Retiro	0.000000	0.000000	0.000000	0.000000	0.045455	0.090909	0.000000	0.000000	0.000000	0.090909	0.000000	0.090909	0.045455	0.000000	0.045455

Complete notebook in: [capstone-ibm-project/Capstone_Project Curitiba Final Version.ipynb at main · sdncmndll/capstone-ibm-project (github.com)]

Choosing the optimum K Parameter

In [142…

```
plot(max_range, scores, "No. of clusters", "Silhouette Score")
```



From the graph the optimal number is found to be 7

Complete notebook in: [capstone-ibm-project/Capstone_Project Curitiba Final Version.ipynb at main · sdncmndll/capstone-ibm-project (github.com)]

List of clusters discovered by K-Means

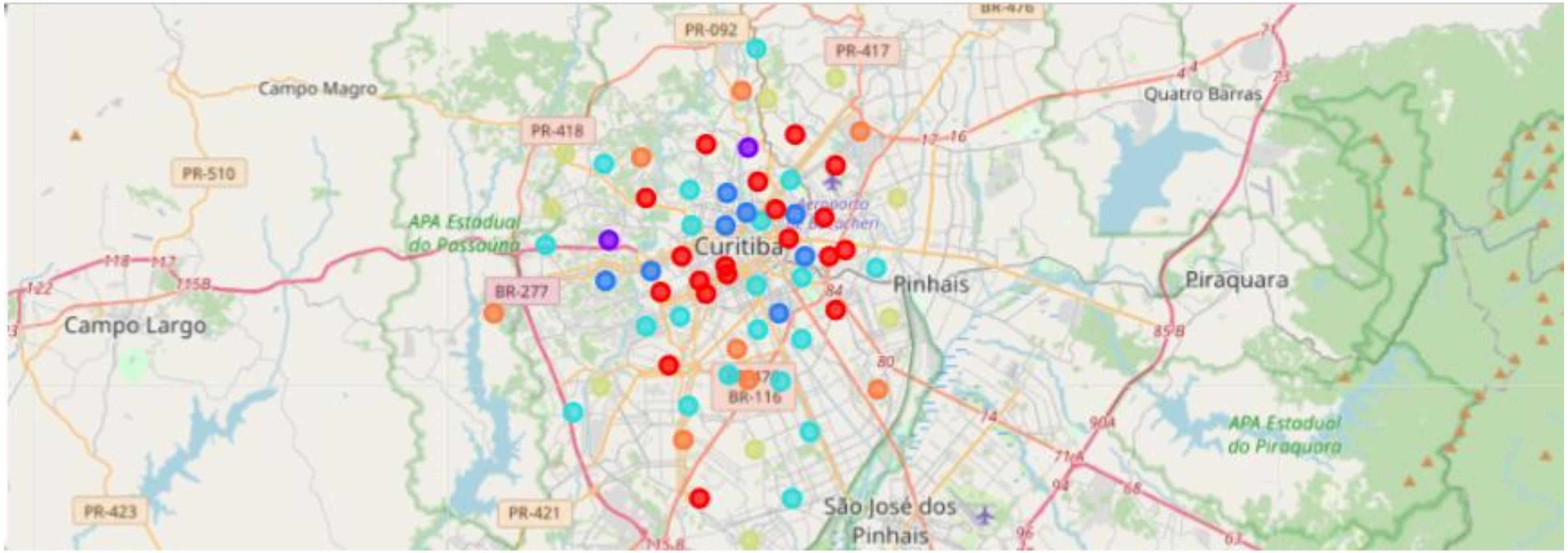
```
Now the cluster dataframe has 69 data rows.
```

Out[155…

	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Abranches	-25.37028	-49.27007	6	Bakery	Food Truck	Empanada Restaurant	Comfort Food Restaurant	Deli / Bodega	Diner	Doner Restaurant	Dumpling Restaurant	Empada House	Wings Joint
1	Água Verde	-25.44746	-49.28556	0	Café	Restaurant	Buffet	Japanese Restaurant	Chinese Restaurant	Food Truck	Bakery	Italian Restaurant	Brazilian Restaurant	Middle Eastern Restaurant
2	Ahú	-25.40486	-49.26329	0	Steakhouse	Pizza Place	Bakery	Restaurant	Food Truck	Brazilian Restaurant	Japanese Restaurant	Café	Italian Restaurant	Buffet
3	Alto Boqueirão	-25.52542	-49.24917	3	Hot Dog Joint	Diner	Pizza Place	Snack Place	Deli / Bodega	Breakfast Spot	Doner Restaurant	Burger Joint	Bakery	Japanese Restaurant
4	Alto da Glória	-25.41970	-49.26181	3	Café	Brazilian Restaurant	Bakery	Pizza Place	Buffet	Mediterranean Restaurant	Chinese Restaurant	Sushi Restaurant	Portuguese Restaurant	Hot Dog Joint
5	Alto da Rua XV	-25.42645	-49.25011	0	Restaurant	Pizza Place	Italian Restaurant	BBQ Joint	Bakery	Japanese Restaurant	Brazilian Restaurant	Café	Sandwich Place	Fried Chicken Joint
6	Área Rural de Curitiba	-25.43998	-49.27654	0	Restaurant	Italian Restaurant	BBQ Joint	Café	Bakery	Pizza Place	Japanese Restaurant	Brazilian Restaurant	Comfort Food Restaurant	Middle Eastern Restaurant
7	Atuba	-25.43333	-49.23333	0	Pizza Place	Food Truck	Fast Food Restaurant	Café	BBQ Joint	Mediterranean Restaurant	Bakery	Hot Dog Joint	Seafood Restaurant	Peruvian Restaurant
8	Augusta	-25.45520	-49.37563	6	Cafeteria	Restaurant	Bakery	Wings Joint	Dumpling Restaurant	Comfort Food Restaurant	Deli / Bodega	Diner	Doner Restaurant	Empada House

Complete notebook in: [capstone-ibm-project/Capstone_Project Curitiba Final Version.ipynb at main · sdncmndll/capstone-ibm-project (github.com)]

Map of clusters discovered by K-Means



MAP LEGEND
Cluster 1 - red dots
Cluster 2 - purple dots
Cluster 3 - blue dots
Cluster 4 - light blue dots
Cluster 5 - cian dots
Cluster 6 - green dots
Cluster 7 - beige dots
Cluster 8 - orange dots

The clustered restaurants using the k-means algorithm based on their types similarity. The k-means is an unsupervised machine learning algorithm for clustering unlabeled data

Complete notebook in: [capstone-ibm-project/Capstone_Project Curitiba Final Version.ipynb at main · sdncmndll/capstone-ibm-project (github.com)]

Insight Cluster 1

Top neighborhood of the [CLUSTER 1] and the restaurant styles of this neighborhood.

- This may indicate what people in this cluster prefer to consume

In [196…

```
cluster_1.describe(include='all')[1:4]
```

Out[196…

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
unique	19	11	8	14	12	13	15	11	15	13	15
top	Jardim Social	Pizza Place	Italian Restaurant	Italian Restaurant	Buffet	Bakery	Brazilian Restaurant	Bakery	Diner	Comfort Food Restaurant	Middle Eastern Restaurant
freq	1	3	4	2	3	3	3	4	2	3	3

Complete notebook in: [capstone-ibm-project/Capstone_Project_Curitiba_Final_Version.ipynb at main · sdncmndll/capstone-ibm-project (github.com)]

Insight Cluster 2

Top neighborhood of the [CLUSTER 2] and the restaurant styles of this neighborhood.

- This may indicate what people in this cluster prefer to consume

In [194…

```
cluster_2.describe(include='all')[1:4]
```

Out[194…

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
unique	3	1	3	3	3	3	2	2	2	2	3
top	Santo Inácio	Snack Place	Hot Dog Joint	Brazilian Restaurant	Comfort Food Restaurant	Comfort Food Restaurant	Diner	Doner Restaurant	Dumpling Restaurant	Wings Joint	Dumpling Restaurant
freq	1	3	1	1	1	1	2	2	2	2	1

Complete notebook in: capstone-ibm-project/Capstone_Project_Curitiba_Final_Version.ipynb at main · sdncmndll/capstone-ibm-project (github.com)

Insight Cluster 3

Top neighborhood of the [CLUSTER 3] and the restaurant styles of this neighborhood.

- This may indicate what people in this cluster prefer to consume

In [177…

```
cluster_3.describe(include='all')[1:4]
```

Out[177…

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
unique	8	3	5	7	7	8	8	8	7	8	8
top	São Francisco	Restaurant	Brazilian Restaurant	Café	Pizza Place	Buffet	Buffet	Comfort Food Restaurant	Deli / Bodega	Middle Eastern Restaurant	Comfort Food Restaurant
freq	1	4	3	2	2	1	1	1	2	1	1

Complete notebook in: [capstone-ibm-project/Capstone_Project_Curitiba_Final_Version.ipynb at main · sdncmndll/capstone-ibm-project (github.com)]

Insight Cluster 4

Top neighborhood of the [CLUSTER 4] and the restaurant styles of this neighborhood.

- This may indicate what people in this cluster prefer to consume

In [190…

```
cluster_4.describe(include='all')[1:4]
```

Out[190…

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
unique	21	9	11	12	13	16	15	15	16	12	15
top	Capão da Imbuia	Brazilian Restaurant	Brazilian Restaurant	Pizza Place	Restaurant	Bakery	Deli / Bodega	Diner	Doner Restaurant	Doner Restaurant	Dumpling Restaurant
freq	1	8	4	6	4	3	3	3	3	4	3

Complete notebook in: [capstone-ibm-project/Capstone_Project_Curitiba_Final_Version.ipynb at main · sdncmndll/capstone-ibm-project (github.com)]

Insight Cluster 5

Top neighborhood of the [CLUSTER 5] and the restaurant styles of this neighborhood.

- This may indicate what people in this cluster prefer to consume

In [179…

```
cluster_5.describe(include='all')[1:4]
```

Out[179…

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
unique	1	1	1	1	1	1	1	1	1	1	1
top	Ganchinho	Comfort Food Restaurant	Wings Joint	Chinese Restaurant	Food Truck	Food Court	Food	Fondue Restaurant	Fish & Chips Shop	Fast Food Restaurant	Empanada Restaurant
freq	1	1	1	1	1	1	1	1	1	1	1

Complete notebook in: [capstone-ibm-project/Capstone_Project Curitiba Final Version.ipynb at main · sdncmndll/capstone-ibm-project (github.com)]

Insight Cluster 6

Top neighborhood of the [CLUSTER 6] and the restaurant styles of this neighborhood.

- This may indicate what people in this cluster prefer to consume

In [184…

```
cluster_6.describe(include='all')[1:4]
```

Out[184…

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
unique	9	3	5	8	7	9	5	7	6	6	5
top	Cajuru	Pizza Place	Pizza Place	Brazilian Restaurant	Food Truck	Burger Joint	Comfort Food Restaurant	Comfort Food Restaurant	Sandwich Place	Dumpling Restaurant	Doner Restaurant
freq	1	6	3	2	2	1	2	2	2	2	2

Complete notebook in: [capstone-ibm-project/Capstone_Project_Curitiba_Final_Version.ipynb at main · sdncmndll/capstone-ibm-project (github.com)]

Insight Cluster 7

Top neighborhood of the [CLUSTER 7] and the restaurant styles of this neighborhood.

- This may indicate what people in this cluster prefer to consume

In [187…

```
cluster_7.describe(include='all')[1:4]
```

Out[187…

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
unique	8	3	7	8	6	6	6	6	4	5	4
top	Uberaba	Bakery	Restaurant	Snack Place	Empada House	Comfort Food Restaurant	Deli / Bodega	Diner	Doner Restaurant	Dumpling Restaurant	Wings Joint
freq	1	6	2	1	3	3	3	3	3	3	4

Complete notebook in: [capstone-ibm-project/Capstone_Project_Curitiba_Final_Version.ipynb at main · sdncmndll/capstone-ibm-project (github.com)]

Discusion

Analyzing the most popular restaurants in each cluster, the stakeholder should prefer the *least* popular types as a safe choice. There is no sense in opening the 20th Japanese restaurant in the same street. Of course, there might be more than 10 types in a location. And one might object, that following this logic, the stakeholder must prefer the last type in a full list, and not the 10th one. But bear in mind that descending on the popularity list we might face an absence of demand for this type of food, and open a restaurant that is not needed in this particular location. Presence of interested customers is a must for a successful business. That is why in our recommendations we offer to stop on 10th and 9th positions.

Recommendations, based on description of each cluster:

Based on each analyzed cluster, you can know what types of existing restaurants are and their frequency of occurrences. An important recommendation is to observe the list generated for the TOP neighborhood. In this list you can observe the consumption trend of the cluster. If you set to invest in a particular Cluster (region) always consider what is missing in the neighborhood compared to the TOP neighborhood.

Complete notebook in: [capstone-ibm-project/Capstone_Project_Curitiba Final Version.ipynb at main · sdncmndll/capstone-ibm-project (github.com)]

Conclusion

In this report we worked out a methodology to determine what the most promising type of restaurant is and where it should be opened.
We collected information about Curitiba boroughs from "CEP Brasil", and using geospatial libraries mapped them. Using Foursquare API, we collected the top 100 restaurants and their types for each location within a radius 500 meters from its central point. Then we grouped collected restaurants by location and by taking the mean of the frequency of occurrence of each type, preparing them for clustering. Finally we clustered restaurants by the k-means algorithm and analize the top 10 most common restaurants in each cluster, making useful observations. Eventually we visualized clusters on the map, thus showing the best locations for opening the chosen type of restaurant.
This type of analysis can be applied to any city of your choice that has available geospatial information.
This type of analysis can be applied to any type of venue (shopping, clubs, etc.) that is available in Foursquare database.

Complete notebook in: [capstone-ibm-project/Capstone_Project Curitiba Final Version.ipynb at main · sdncmndll/capstone-ibm-project (github.com)]