Coursera Capstone Project

Clustering and Comparing two neighborhoods in Mexico City to find the best place to relocate

Gillian Escudero

May 31, 2020

Part 1: Introduction and Business Problem

Before deciding on moving to a new area, families, and individuals, need to analyze a vast array of information to choose the location that is best suited to their needs. It is often the case that what prompts a move is a promotion, which forces families to investigate neighborhoods that are nearer the new office to avoid a long commute. In this scenario, finding a new place can be daunting, especially if we are not familiar with the area we are moving to. Some of the factors that could play a role in this decision are the proximity to bus stops, schools, parks, or other venues.

In Mexico City, a lot of people choose to live close to their work, because of the traffic. Two of the most popular municipalities to live in are Benito Juarez and Miguel Hidalgo. In this project we will explore, study, analyze, cluster, and compare the neighborhoods of these two boroughs. The goal is to provide valuable information for individuals looking to move to any of these two areas. We will compare the neighborhoods in each borough based on the quantity and category of the venues in the vicinity. By studying the neighborhoods, we will better understand what types of businesses thrive in each area, as well as finding how they are similar or how they are different.

Background information

Both Benito Juárez and Miguel Hidalgo ranked as the best quality of life in the country, in the study 'Transforming Mexico from the local', UNDP from 2015.

The HDI (Human Development Index) of Benito Juárez stands at 0.944 points. This is one of the best indexed, rivaling that of Switzerland which stands at 0.942.

In second place, Miguel Hidalgo, with an HDI of 0.917, is right on par with that of the United Kingdom, which stands at 0.918 points.

Benito Juarez Borough

Benito Juarez was created early in the 1940's. It occupies an area of 26.63 km2. It's 385,439 inhabitants share it with over 2 million visitors each day. It is characterized by the intense economic activity.

According to statistical data from the 2010 census, the delegation has a population of 385,439. The population density is 14,435 inhabitants per square kilometer, with an average of 2.7 occupants per dwelling. Men represent 45.77%, and women 54.23%. The median age is 36 years.

Miguel Hidalgo

Miguel Hidalgo was created on December 29, 1970 as one of the new 16 delegations that make up the Federal District. The area that encompasses its territory is a fusion of the pre-Hispanic settlements of Tacuba, Tacubaya and Chapultepec, along with residential colonies. It houses many of the most powerful men and women in the country in the political and business sectors and has a great historical tradition.

According to statistical data from the 2010 census, Miguel Hidalgo has a total of 372,890 inhabitants. Of this number, 172,668 were men and 200,222 were women. It has a density of 7,452 inhabitants per square kilometer.

Part 2: Data acquisition and preparation

To carry out this project two types of data are needed:

- A list of the neighborhoods in each borough, including geo location; and
- The venues data for each neighborhood.

The geolocation data for each neighborhood will be used to extract a list of venues in their vicinity. The venues data will be used to do a K-means clustering analysis. The objective is to cluster groups of neighborhoods based on their venue proximity and composition.

2.1 Neighborhood Data

The data for the neighborhoods for each borough can be extracted from Mexico City Data Portal (https://datos.cdmx.gob.mx/explore/dataset/coloniascdmx/table/). The city has an open database that includes the list of neighborhoods by borough along with the geolocation. The data can be consumed through an API or directly downloading the data in any of the file formats they offer (csv, json, or excel). For ease of processing, the direct download in csv will be used.

Since the full dataset includes more information than what is necessary for this analysis, the data will need to be filtered and adjusted to fit our needs. This includes renaming the columns, splitting the Geo location data from one column ("Geo Point") to separate columns for Latitude and Longitude, converting Latitude and Longitude data types from Object to float. Figure 1: shows the code to complete these steps.

Code

```
## Read csv file into a pandas dataframe
link='https://datos.cdmx.gob.mx/explore/dataset/coloniascdmx/download/?format=csv&timezone=Ame
rica/Guatemala&lang=es&use_labels_for_header=true&csv_separator=%2C'
data=pd.read_csv(link)
data.head()

## We can see that the latitude and longitude data is stored in a column labeled "Geo Point".
We will need to split this column in two to and store it in a temporary dataframe

Lat_Long=data['Geo Point'].str.split(',', expand=True)

## Now we will create a new dataframe using only the columns Colonia (Neighborhood), and
Alcaldia (Burough)

Mexico_city=data[['ALCALDIA','COLONIA']].copy()

## Next we will add the latitude and longitude from the temporary dataframe and rename the
columns.
Mexico_city['Latitude']=Lat_Long[0]
Mexico_city['Longitude']=Lat_Long[1]
Mexico_city.columns=['Borough','Neighborhood','Latitude','Longitude']

## We'll need to convert Latitude and Longitud to real numbers to avoid problems down the road
```

```
Mexico_city['Latitude'] = Mexico_city['Latitude'].astype(float)
Mexico_city['Longitude'] = Mexico_city['Longitude'].astype(float)

## Because we will be using only two boroughs from the whole dataset we will create dataframe
for each borough. MH for Miguel Hidalgo and BJ for Benito Juarez

MH_data = Mexico_city[Mexico_city['Borough'].isin(['MIGUEL HIDALGO'])].reset_index(drop=True)
BJ_data = Mexico_city[Mexico_city['Borough'].isin(['BENITO JUAREZ'])].reset_index(drop=True)
```

Figure 1: Code to extract, clean and prepare data.

The final step is to create two separate dataframes, one for Benito Juarez (`BJ_data`) and one for Miguel Hidalgo (`MH_data`), by filtering the "Borough" column in the Mexico City dataframe. Figures 2 and 3 show some basic information of these two dataframes, including the number of records for each borough.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 88 entries, 0 to 87
Data columns (total 4 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Borough       88 non-null     object
 1   Neighborhood  88 non-null     object
 2   Latitude      88 non-null     float64
 3   Longitude     88 non-null     float64
dtypes: float64(2), object(2)
memory usage: 2.1+ KB
```

Figure 2: Benito Juarez dataframe information.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 64 entries, 0 to 63
Data columns (total 4 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Borough       64 non-null     object
 1   Neighborhood  64 non-null     object
 2   Latitude      64 non-null     float64
 3   Longitude     64 non-null     float64
dtypes: float64(2), object(2)
memory usage: 1.6+ KB
```

Figure 3: Miguel Hidalgo dataframe information.

Having the neighborhood data with the coordinates allows us to draw a map of each borough using Folium Python package. Figure 4 shows the map generated of Benito Juarez alongside the borough boundaries taken from Mexico City Data Portal. Figure 5 shows the map of the data extracted for Miguel Hidalgo.

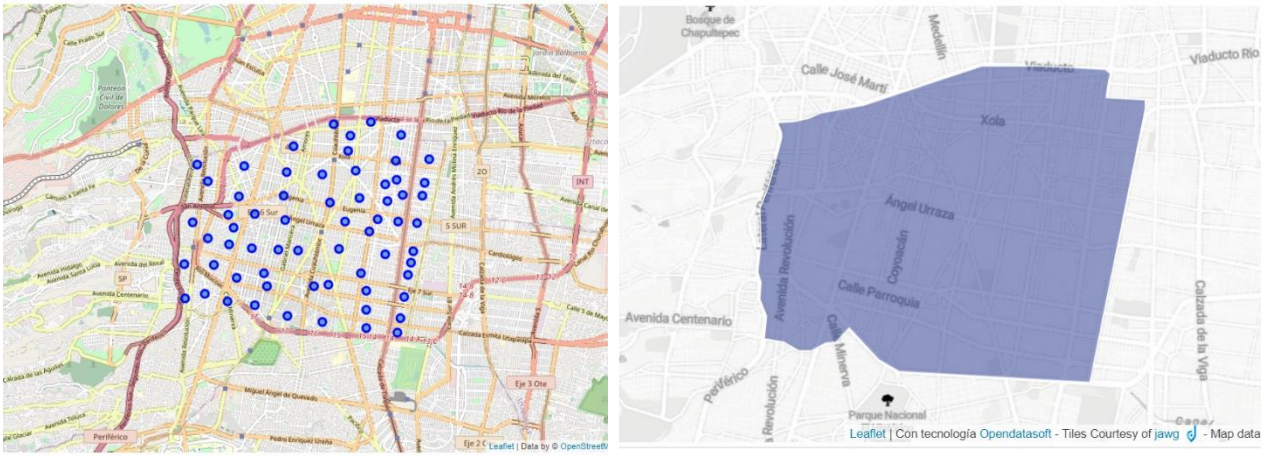


Figure 4: Map of Benito Juarez neighborhoods location and boundaries.

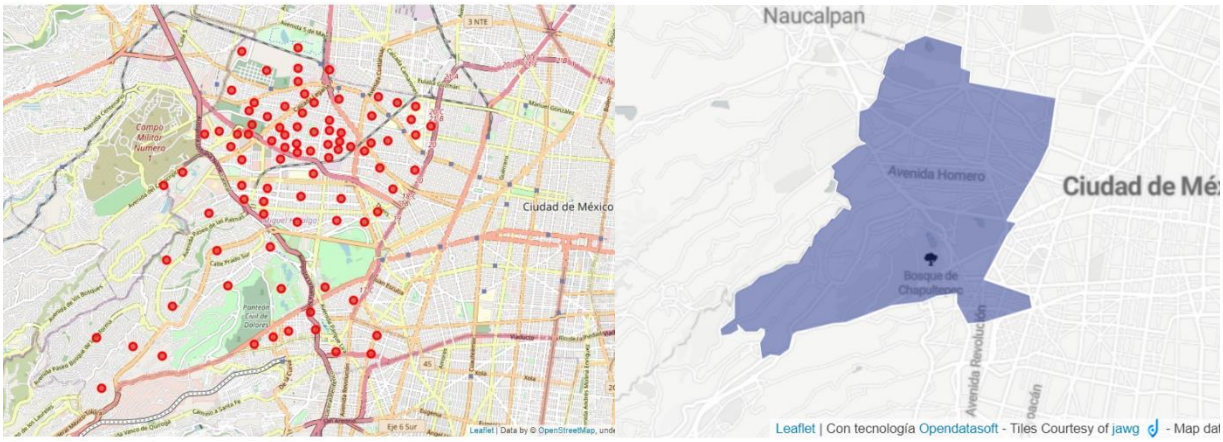


Figure 5: Map of Miguel Hidalgo neighborhoods location and boundaries.

2.2 Venues Data

We will extract the venues data from Foursquare. The Places API offers real-time access to Foursquare's global database of rich venue data. The venue data is obtained by passing the required parameters for each neighborhood to the Places API. We will create a dataframe for each borough to contain the extracted venue data.

Figure 6 shows the code used to create a function that takes the neighborhood names, latitudes and longitudes as inputs, creates the get request for Places API, and it returns a dataframe with the list of venues.

Code

```
def getNearbyVenues(names, latitudes, longitudes, radius=1000):

    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        print(name)

        # create the API request URL
        url =
'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={
}&v={}&ll={},{}&radius={}&limit={}'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            LIMIT)

        # make the GET request
        results =requests.get(url).json()["response"]['groups'][0]['items']

        # return only relevant information for each nearby venue
        venues_list.append([(
```

```
            name,
            lat,
            lng,
            v['venue']['name'],
            v['venue']['location']['lat'],
            v['venue']['location']['lng'],
            v['venue']['categories'][0]['name']) for v in results])

    nearby_venues = pd.DataFrame([item for venue_list in venues_list for
item in venue_list])
    nearby_venues.columns = ['Neighborhood',
                  'Neighborhood Latitude',
                  'Neighborhood Longitude',
                  'Venue',
                  'Venue Latitude',
                  'Venue Longitude',
                  'Venue Category']

    return(nearby_venues)
```

Figure 6: Function to Get nearby venues for each neighborhood.

After using this function for each Borough dataframe (`BJ_data` and `MH_data`), we get the following results:

- Benito Juarez: 6,400 venues with 225 venue categories.
- Miguel Hidalgo: 7,817 venues with 273 venues categories.

Part 3: Methodology

3.1 Exploratory Data Analysis

Now that we have collected the data we need; it is time to do some exploratory analysis of each borough.

Most Common Venue Categories

We will start our analysis by discovering what are the categories that have more venues in each Borough. We will use a bar plot with the number of occurrences for each category.

As shown on Figures 7 and 8, both boroughs have a terribly similar distribution of venues categories. Not so unexpectedly, the most popular venue categories in both boroughs are Mexican Restaurants and Taco places, followed closely by Coffee shops, Bakeries, and Cafes.

From this we can interpret that both Boroughs have a widespread array of food choices. Restaurants to dine out are common in each.

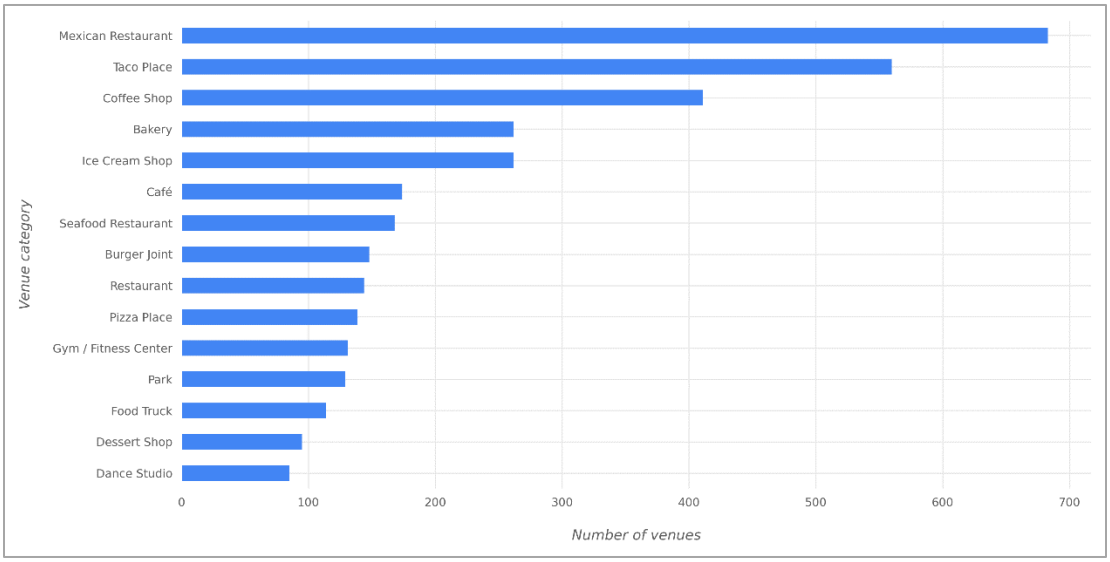


Figure 7: Benito Juarez venues by category.

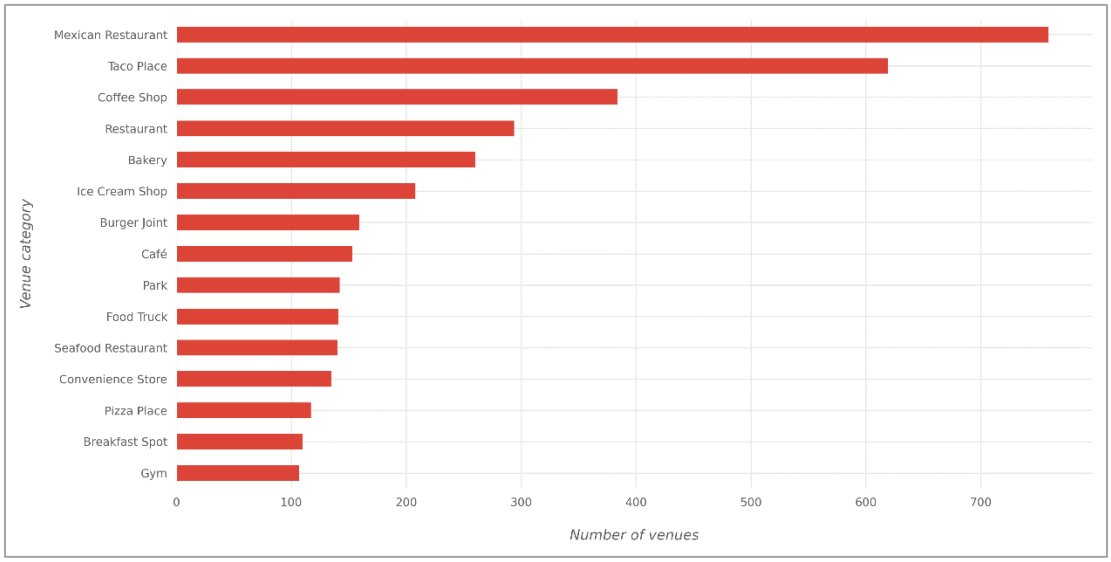


Figure 8: Miguel Hidalgo venues by category.

Most Widespread Venue Categories

With this analysis we will know what venue categories exist in most neighborhoods. The categories with a quantity of neighborhoods that is close to the total number of neighborhoods in each borough, will be those that are most widespread.

As shown on Figure 9, in Benito Juarez, the most widespread categories, that is those venue categories with more than 60 neighborhoods, are Coffee Shop, Taco Place, Ice Cream Shop, Seafood and Mexican Restaurant. This means that in Benito Juarez almost all neighborhoods have at least one of these venues within a radius of 1 kilometer.

Figure 9: Benito Juarez number of neighborhoods by category.

Figure 10 shows the results for Miguel Hidalgo. In this borough the most widespread categories, that is those venue categories with more than 80 neighborhoods, are Mexican Restaurants, Bakeries, Taco Place, Restaurants and Coffee Shops. This means that in Miguel Hidalgo almost all neighborhoods have at least one of these venues within a radius of 1 kilometer.

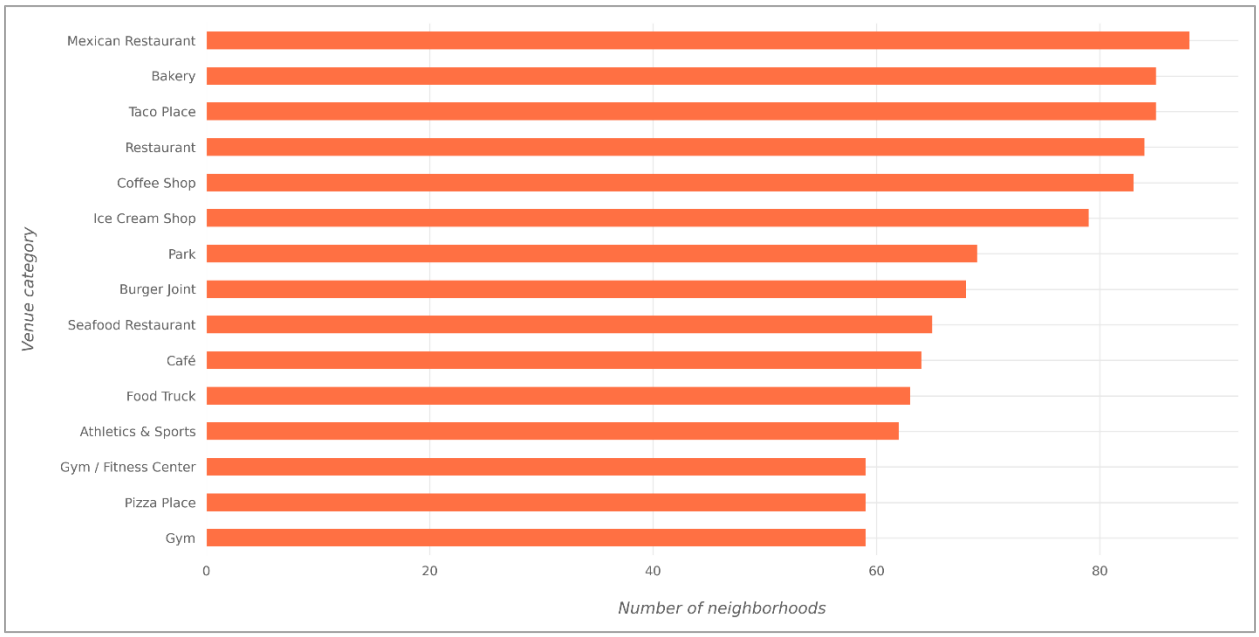


Figure 10: Miguel Hidalgo number of neighborhoods by category.

3.2 Clustering of Neighborhoods

To cluster the neighborhoods in each borough based on the venue categories, k-means cluster was employed. This machine learning algorithm is an unsupervised clustering technique that searches for a pre-determined number of clusters within an unlabeled multidimensional dataset.

The model is based on two key assumptions: The cluster center is the arithmetic mean of all the points belonging to the cluster; and each point is closer to its own cluster center than any other cluster center in the dataset.

One-hot encoding

Before using the K-means clustering algorithm, we first need to prepare our data. For this purpose, one-hot encoding will be applied on the "Venue Category" feature and the result of the encoding will be used for the clustering. Figure 11 shows the code to perform one-hot encoding on the Miguel Hidalgo data. The same process is applied to Benito Juarez venue data.

Code

```
# one hot encoding
MH_onehot = pd.get_dummies(MH_venues[['Venue Category']], prefix="",
prefix_sep="")

# add neighborhood column back to dataframe
MH_onehot['Neighborhood'] = MH_venues['Neighborhood']

# move neighborhood column to the first column
fixed_columns = [MH_onehot.columns[-1]] + list(MH_onehot.columns[:-1])
MH_onehot = MH_onehot[fixed_columns]
```

Figure 11: Miguel Hidalgo venue data one-hot enconding.

The next step is aggregating the values for each neighborhood so that each neighborhood is represented in a single row. This will be done by grouping rows by neighborhood and taking the mean of the frequency of occurrence of each category.

After producing the aggregated data for each borough separately, it will be combined in a single dataframe.

Combining the two dataframes

First any venue category that is in one borough, but not the other is detected and stored in a set to replicated in the combined dataframe. Then, to distinguish the neighborhoods in Benito Juarez and Miguel Hidalgo, a text string is added to the end of each neighborhood name before merging the dataframes. Figure 12 shows the code to combine the two dataframes.

Code

```
clmns_BJ_only = set(BJ_grouped.columns).difference(set(MH_grouped.columns))
clmns_MH_only = set(MH_grouped.columns).difference(set(BJ_grouped.columns))
```

```
BJ_grouped_ = BJ_grouped.copy()
BJ_grouped_['Neighborhood'] = BJ_grouped_['Neighborhood'].apply(lambda x: x
+ '_Benito Juarez')
MH_grouped_ = MH_grouped.copy()
MH_grouped_['Neighborhood'] = MH_grouped_['Neighborhood'].apply(lambda x: x
+ '_Miguel Hidalgo')

for c in clmns_BJ_only:
    MH_grouped_[c] = 0
for c in clmns_MH_only:
    BJ_grouped_[c] = 0

all_clmns_sorted = ['Neighborhood'] +
sorted(list(BJ_grouped_.drop('Neighborhood', axis=1).columns),
key=str.lower)
BJ_grouped_ = BJ_grouped_[all_clmns_sorted]
MH_grouped_ = MH_grouped_[all_clmns_sorted]

BJ_MH_grouped = pd.concat([BJ_grouped_,
MH_grouped_]).reset_index(drop=True)
```

Figure 12: Combining the venues data from both boroughs

The most common categories for each neighborhood

Due to the variety of venues, only the top 10 common venues are selected for each neighborhood as the features to train the K-means clustering algorithm. This dataframe is created by retrieving the 10 categories with the largest values for each neighborhood. Figure 13 shows the resulting dataframe.

	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
0	ACACIAS_Benito Juarez	Coffee Shop	Mexican Restaurant	Bakery	Ice Cream Shop	Boutique	Cosmetics Shop	Steakhouse	Shopping Mall	Seafood Restaurant	Burger Joint
1	ACTIPAN_Benito Juarez	Ice Cream Shop	Café	Coffee Shop	Cosmetics Shop	Gym / Fitness Center	Argentinian Restaurant	Taco Place	Supermarket	Diner	Dessert Shop
2	ALAMOS I_Benito Juarez	Mexican Restaurant	Taco Place	Bakery	Burger Joint	Seafood Restaurant	Coffee Shop	Dessert Shop	Café	Ice Cream Shop	Restaurant
3	ALAMOS II_Benito Juarez	Taco Place	Mexican Restaurant	Burger Joint	Café	Coffee Shop	Ice Cream Shop	Bakery	Breakfast Spot	Dance Studio	Bar
4	ALBERT_Benito Juarez	Mexican Restaurant	Taco Place	Breakfast Spot	Food Truck	Coffee Shop	Bakery	Gym	Soccer Field	Pool	Flea Market

Figure 13: Most Common categories in each neighborhood

K-Means clustering

The dataframe show in Figure 13 is used to apply the clustering algorithm of Scikit-learn library as shown on Figure 14. The clustering algorithm assigns a cluster label from 0 to 4 to each neighborhood, these labels denote the cluster assigned to each record.

Code

```
# the number of clusters
kclusters = 5

BJ_MH_grouped_clustering = BJ_MH_grouped.drop('Neighborhood', 1)
# run k-means clustering
kmeans = KMeans(n_clusters=kclusters,
random_state=0).fit(BJ_MH_grouped_clustering)

# add clustering labels
BJ_MH_neighborhoods_categories_sorted.insert(0, 'Cluster Labels',
kmeans.labels_)

BJ_MH_merged =
BJ_MH_neighborhoods_categories_sorted.set_index('Neighborhood')
```

Figure 14: Cluster algorithm using 5 clusters

By merging the resulting dataframe with the original data sets from Benito Juarez and Miguel Hidalgo, we will add the latitude and longitude to be able to visualize the clusters in a folium map. The code to perform this process is shown in Figure 15.

Code

```
BJ_data_ = BJ_data.copy()
BJ_data_['Neighborhood'] = BJ_data_['Neighborhood'].apply(lambda x: x +
'_Benito Juarez')
MH_data_ = MH_data.copy()
MH_data_['Neighborhood'] = MH_data_['Neighborhood'].apply(lambda x: x +
'_Miguel Hidalgo')

BJ_MH_data_=pd.DataFrame()
BJ_MH_data_=BJ_data_.append(MH_data_)
BJ_MH_merged_=BJ_MH_neighborhoods_categories_sorted.copy()

# merge data to add latitude/longitude for each neighborhood
BJ_MH_merged_ = BJ_MH_merged_.join(BJ_MH_data_.set_index('Neighborhood'),
on='Neighborhood')
```

Figure 15: Merging neighborhoods with cluster labels with latitude and longitude data

Part 4: Results

The output of the clustering operation is 5 clusters with labels 0, 1, 2, 3, and 4. Each cluster is comprised of a group of neighborhoods that are similar based on the most common venue categories in each neighborhood. The clustering algorithm was run on the 64 neighborhoods in Benito Juarez and the 88 neighborhoods of Miguel Hidalgo.

Table 1 shows the number of neighborhoods in each cluster.

Table 1: Cluster distribution		
Cluster Label	Number of Neighborhoods	Color
0	24	Pale green
1	34	Orange
2	34	Red
3	14	Purple
4	46	Sky blue

Because both Boroughs are close together in Mexico City, we can plot the clusters in the same map. Figure 16 shows a map of Mexico City with the neighborhoods clusters for each borough.

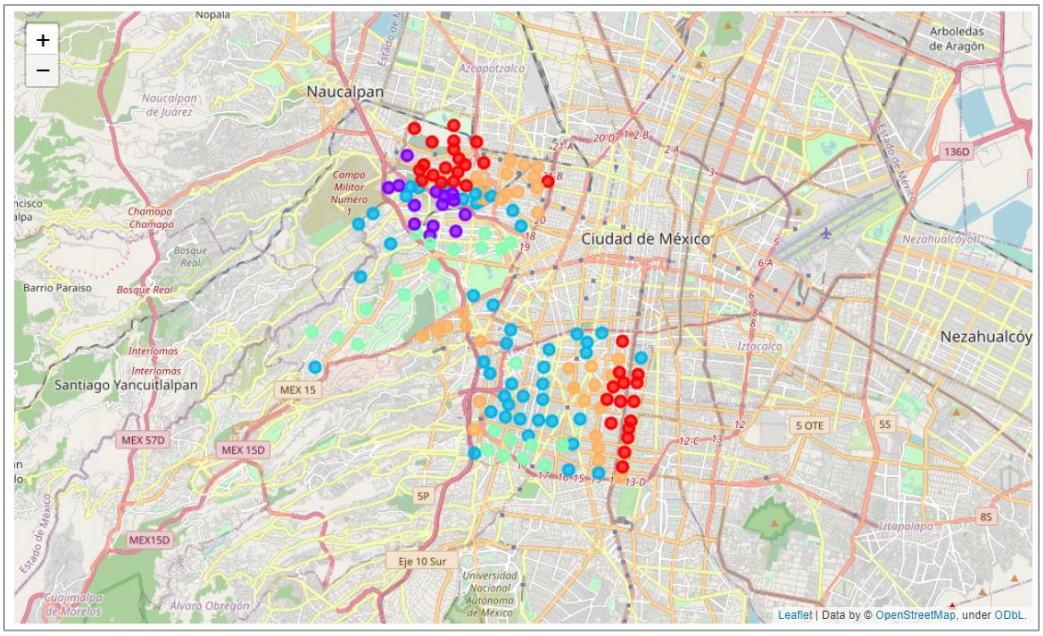
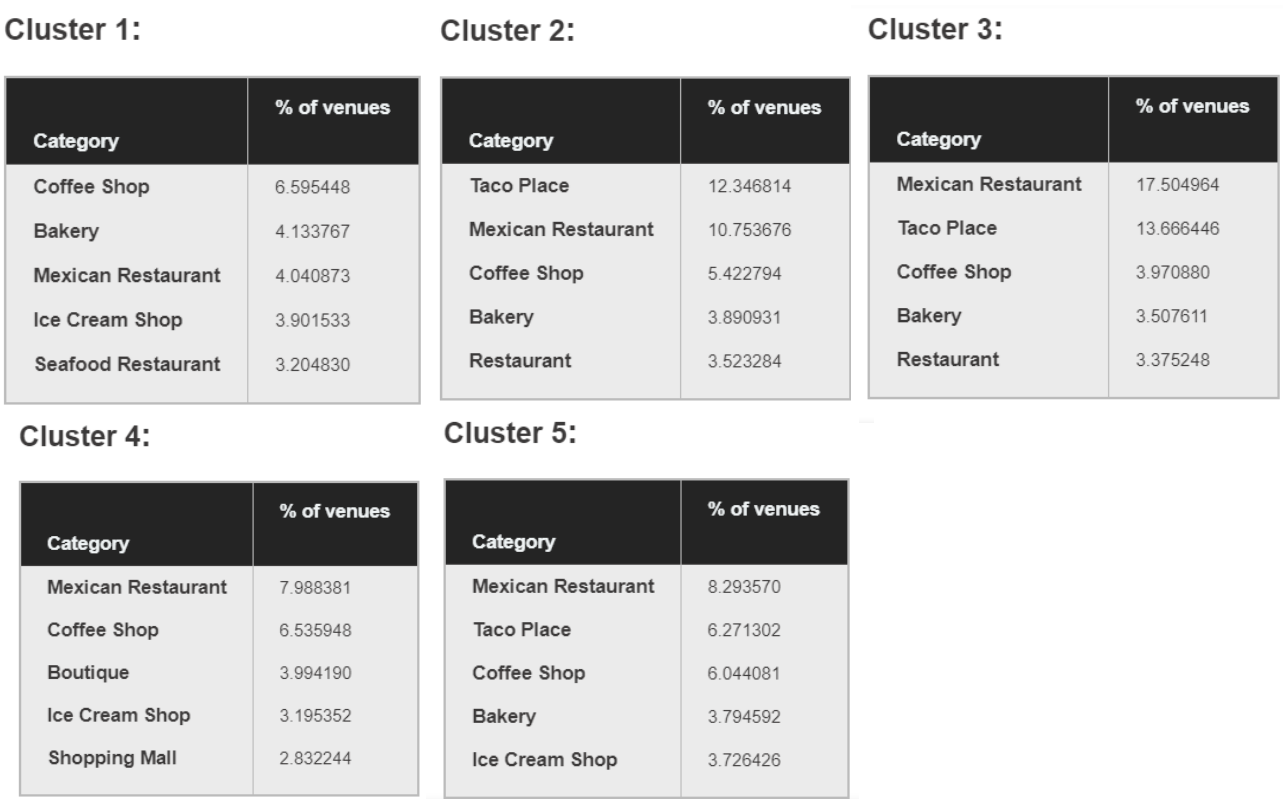


Figure 16: Map of Mexico City with Benito Juarez and Miguel Hidalgo neighborhood clusters

Part 5: Discussion

Figure 17 shows the distribution of the top 5 venue categories in each cluster. For each cluster, the percentage of the venue category is shown.



Cluster 1:

Category	% of venues
Coffee Shop	6.595448
Bakery	4.133767
Mexican Restaurant	4.040873
Ice Cream Shop	3.901533
Seafood Restaurant	3.204830

Cluster 2:

Category	% of venues
Taco Place	12.346814
Mexican Restaurant	10.753676
Coffee Shop	5.422794
Bakery	3.890931
Restaurant	3.523284

Cluster 3:

Category	% of venues
Mexican Restaurant	17.504964
Taco Place	13.666446
Coffee Shop	3.970880
Bakery	3.507611
Restaurant	3.375248

Cluster 4:

Category	% of venues
Mexican Restaurant	7.988381
Coffee Shop	6.535948
Boutique	3.994190
Ice Cream Shop	3.195352
Shopping Mall	2.832244

Cluster 5:

Category	% of venues
Mexican Restaurant	8.293570
Taco Place	6.271302
Coffee Shop	6.044081
Bakery	3.794592
Ice Cream Shop	3.726426

Figure 17: Most common venue-categories in each of the 5 clusters

From this figure we can distinguish some of the differences in each cluster:

- In the first cluster, Coffee shops are the most common venues.
- In the second cluster, the most common venues are Taco Places, with 12.34% of the venue's categories.
- Mexican Restaurants are the most common venue categories in Clusters 3, 4 and 5.
- The fourth cluster is the only one with Shopping Malls and Boutiques in the top 5 venue categories.

Figure 18 shows the number of neighborhoods from each borough in each of the resulting clusters. From this bar chart, we can examine how the neighborhoods from each borough are distributed I each cluster.

Considering Miguel Hidalgo had ~20 more neighborhoods than Benito Juarez, it is natural to see more neighborhoods from this borough in each cluster. This is the case for most clusters, except for Cluster 5, where there are more neighborhoods from Benito Juarez.

Cluster 4 does not have any neighborhoods from Benito Juarez. Let's remember this is the only cluster where there were shopping malls and boutiques in the top 5 venue category.

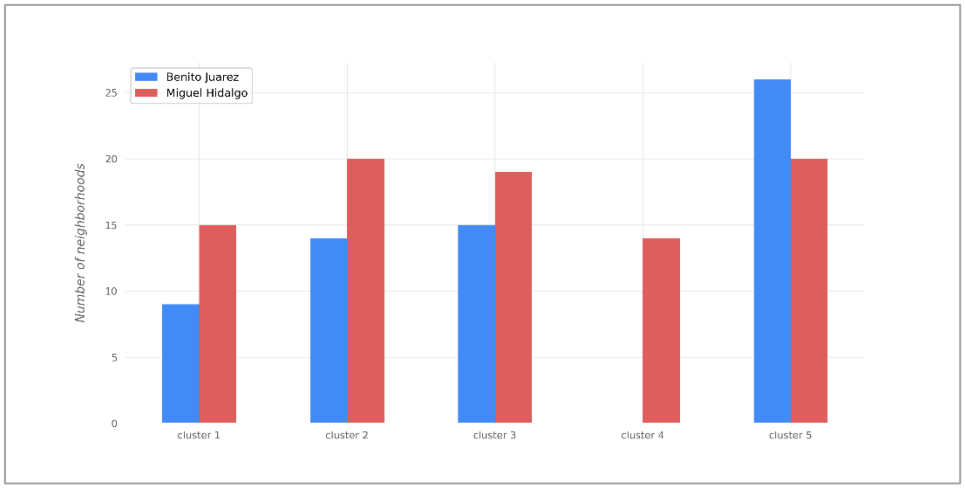


Figure 18: Number of neighborhoods from Benito Juarez and Miguel Hidalgo in each cluster

Part 6: Conclusions

The goal of this project was to help people decide the best neighborhoods to relocate based on the proximity of venues in Mexico City. By using public data, the neighborhoods of the two of the hottest boroughs from Mexico City were analyzed. The results show that both boroughs, Benito Juarez and Miguel Hidalgo, are remarkably similar. There are clusters of neighborhoods that share common characteristics in the two boroughs, as well as some that do not.

This analysis could be expanded by adding crime rate data, which is also available in Mexico City's Data portal. Another feature to add would be a seismic activity feature, considering Mexico is prone to earthquakes.