The Battle of Neighborhoods

Introduction

Colombia is a diverse country, full of an impressive geography, with different cultures and rich in water, each region is a new culture, demarcated by customs that make them so unique.

Business problem

The objective is to be able to find the cities and the most optimal points in them to open a new business that seeks to reach a high purchasing power but that in the same way offers services that can be used by the whole society in general, seeking to stand out for its high quality, affordable prices and innovation.

Description of the data

Data on the geographic location of all of Colombia, the postal codes, the names of the neighborhoods were required, which served as a starting point for the construction of the base on which the analysis and the final model would finally be carried out.

1. Colombia postal codes
2. Dane name and codes georeferencing
3. Layers of Dane's apples
4. Dane multidimensional poverty survey information
5. Colombia Census 2018 Dane
6. Dane household income and expenditure survey
7. ArcGIS API
8. Foursquare API data

Methodology

We will create our model with the help of Python, so we start by importing all the required packages.

```
# procesamiento
import pandas as pd
import numpy as np
import time
import datetime
from tqdm.notebook import tqdm
from pandas.io.json import json_normalize
from sklearn.cluster import KMeans
import json
import xml
from sklearn import preprocessing
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
```

```
#geolocalización
import arcgis
import requests
import geocoder
import folium
from geopy.geocoders import Nominatim
from bs4 import BeautifulSoup
from arcgis.geocoding import geocode
from arcgis.gis import GIS
gis = GIS()

from arcgis.geometry import Geometry, Polyline, Point, union
from arcgis.geometry.filters import intersects, contains, overlaps, crosses,
from arcgis.geometry.filters import envelope_intersects, index_intersects
from arcgis.geoenrichment import Country, enrich
from arcgis.features import FeatureLayer,FeatureSet, FeatureCollection
from arcgis.geoenrichment import *
from arcgis.geocoding import geocode,Geocoder, get_geocoders,reverse_geocode
import arcgis.features.use_proximity as use_proximity
from time import time
from geopy.distance import geodesic, great_circle
```

The Dane layers were exported to be able to obtain the segment of customers to be impacted, obtain the visualization of the blocks on the map and later the model was built, the map was drawn with the clusters corresponding to the location and then we compared and we discuss the findings.

Data collection

In the data collection stage, we begin with the collection of the necessary data for the cities of Medellin. We need data that has the specific zip codes, neighborhoods and districts of each of the cities.

Foursquare API data

We will need data on different places in different neighborhoods in that specific district. To obtain that information, we will use the location information from "Foursquare." Foursquare is a provider of location data with information on all kinds of places and events within an area of interest. This information includes place names, locations, menus, and even photos. As such, the foursquare location platform will be used as the sole source of data, as all required information indicated can be obtained via the API.

After finding the list of neighborhoods, we connect to the Foursquare API to collect information about the places within each neighborhood. For each neighborhood, we have chosen the radius to be 500 meters.

The data retrieved from Foursquare contained information for places within a specified distance of the longitude and latitude of the zip codes. The information obtained by location is as follows:

Neighborhood: Neighborhood name

Neighborhood Latitude: Neighborhood Latitude

Neighborhood length: neighborhood length

Place: Place name

Place latitude: Place latitude

Length of place: length of place

Venue Category: Venue Category

Based on all the information collected for Medellin, we have enough data to build our model. We group neighborhoods based on similar place categories. We then present our observations and findings. With this data, our stakeholders can make the necessary decision.

Data preprocessing

The blocks that have internet service and that are located in Medellin are selected.

```
coordenadas=MZN[MZN['nombre_municipio']=='MEDELLIN']
coordenadas=coordenadas[coordenadas['INTERNET']==1]
coordenadas=coordenadas[['LONGITUD','LATITUD']].drop_duplicates(subs
coordenadas=coordenadas.reset_index()
coordenadas=coordenadas.iloc[:,1:]
coordenadas
cor=get_loc(coordenadas)
cor
```

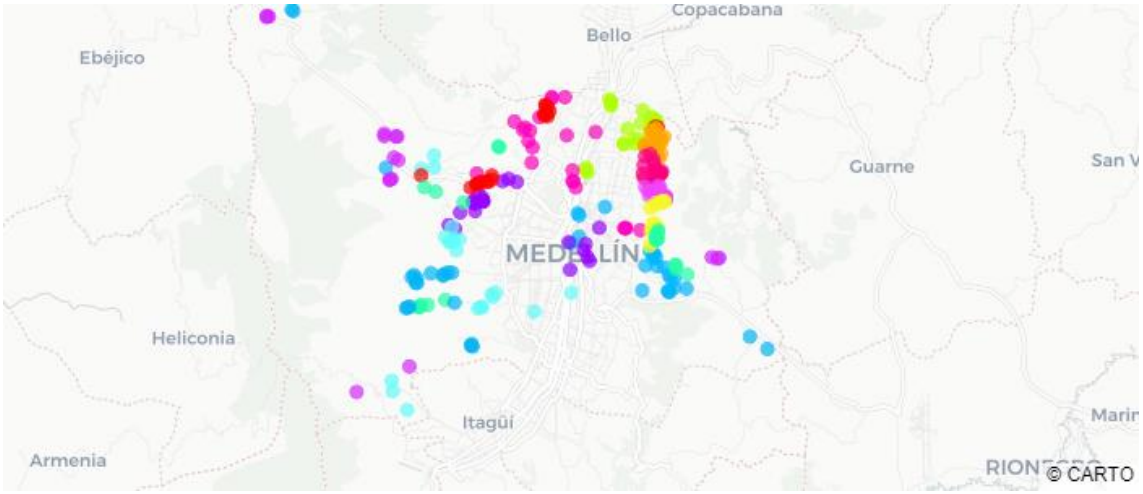
K means

We will use the K Means Clustering machine learning algorithm to group similar neighborhoods. There are many different cluster sizes we can select from, we will go with a cluster number of 6 to keep it as optimized as possible.

```
df=pd.merge(MZN,cor,on=['LONGITUD','LATITUD'])
kmeans = KMeans(n_clusters=6, random_state=0).fit(df[['CATEGORIA','i
       'INTERNET', 'PERSONAS','PRIMAR', 'SECUND', 'SUPERI',
       'POSTGR', 'NINGUN_ED','Estrato']])
df['cluster']=kmeans.labels_
df
```

Display of grouped neighborhoods

Our data is processed, the missing data is collected and compiled. The model is built. All that remains is to see the neighborhoods grouped together on the map.



conclusion

The purpose of this project was to explore the neighborhoods of Medellin and see how attractive it is to tourists and potential migrants. We scan based on their zip codes and then extrapolate the commonplaces present in each of the neighborhoods and finally conclude with the grouping of similar neighborhoods.