**Battle of World Capitals!!!**

**Capstone Project, Coursera-IBM Specialization**

Aníbal J Guerra S, [ajguerras@gmail.com](mailto:ajguerras@gmail.com)

Medellín, Colombia

In this project we are meant to put a lot of the knowledge we have gotten through the whole IBM data science specialization. My work is intended to exhibit the application of a methodology, the usage of the right tools, data treatment, tools for visualization and to provide a final analysis that answer the initial questions.

1. **Introduction**

I remember the days when playing the simulation game SIM City ® I could be the hero of my very own city as I designed and created a beautiful and bustling metropolis. Through a long chain of apparently simple decisions, like creating schools, libraries, hospitals, entertainment places, etc.; you could make your city a good and attractive place to live, so it would get larger and more intricate. That could be defined as success. On the opposite case, people in your town would start leaving and the whole system finally would fail. A perfect balance which sounds very hard to find.

I grew up in Latin America, in a country that we were told that we were “on the road to become a first world country”. But it always seemed like a never-ending process. After traveling a lot, and visiting very organized cities with a high standards for living, I started to ask myself: what factors really needed for a city to achieve the perfect balance that makes it attractive, providing enough quality of life to consider such city as a well-developed city.

1. **Problem Description**

Far beyond my youthful meditations, finding the answers to such questions is a very relevant problem. For public servers in the government (i.e. major, governors) having clarity about such factors would give them clarity about the decisions that should be made to keep the city on the road to development. Two questions seem obvious: where on that road is the city right now (what does the city “have”) and what should we “have” as a city in order to be considered as developed city in the future.

As I am not an expert in such extent, I find those answers hard to find. But for sure data science has the answers. I would use the Foursquare data related to all of the venues that the capitals of the world have right now, and see if through data science I get to stratify (or cluster) those cities to see if there are specific venues that are characteristic of the cities development in each class.

**Hypothesis**: It is possible to discriminate groups of cities according to the amount of certain types of venues in them.

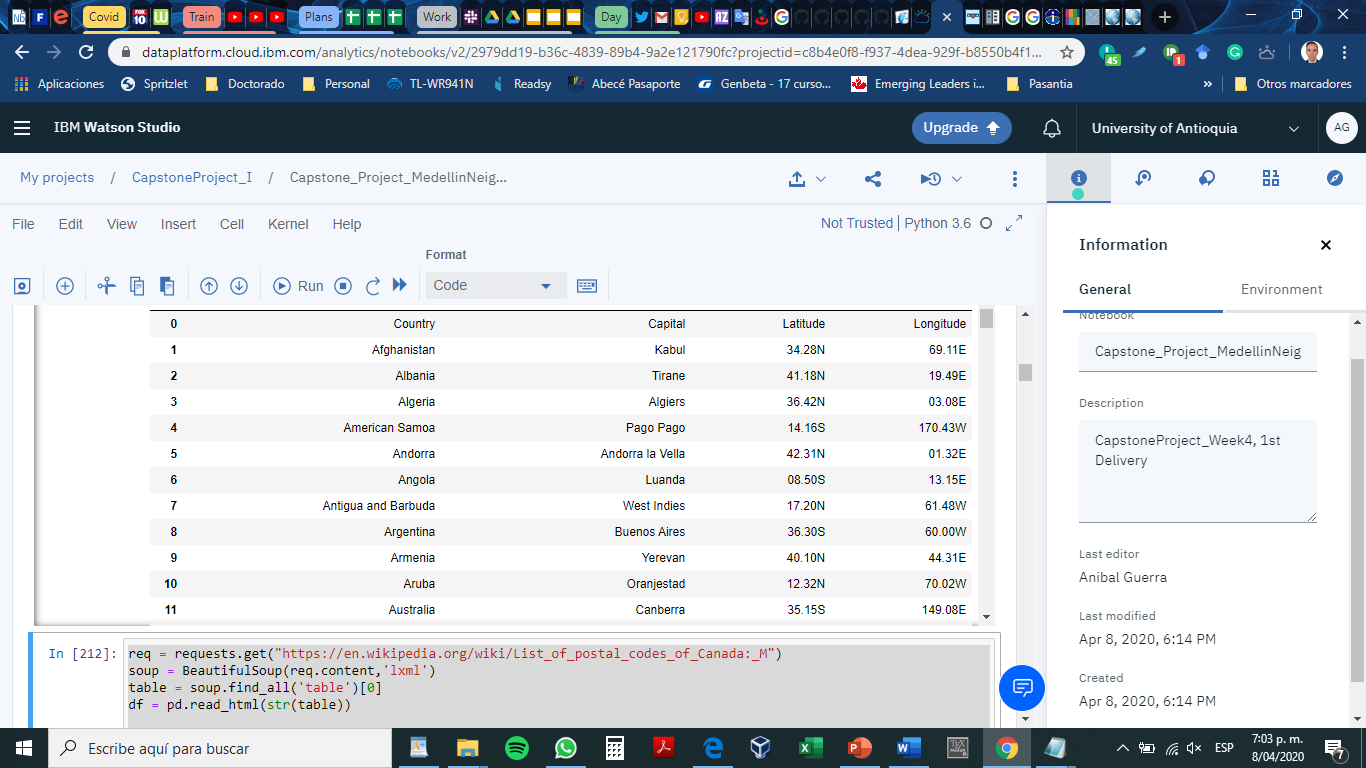
If so, I will have to figure out if such groups correspond to an already known classification.

1. **Data Description**
   1. Capitals of the world and along with their geo-localization data.

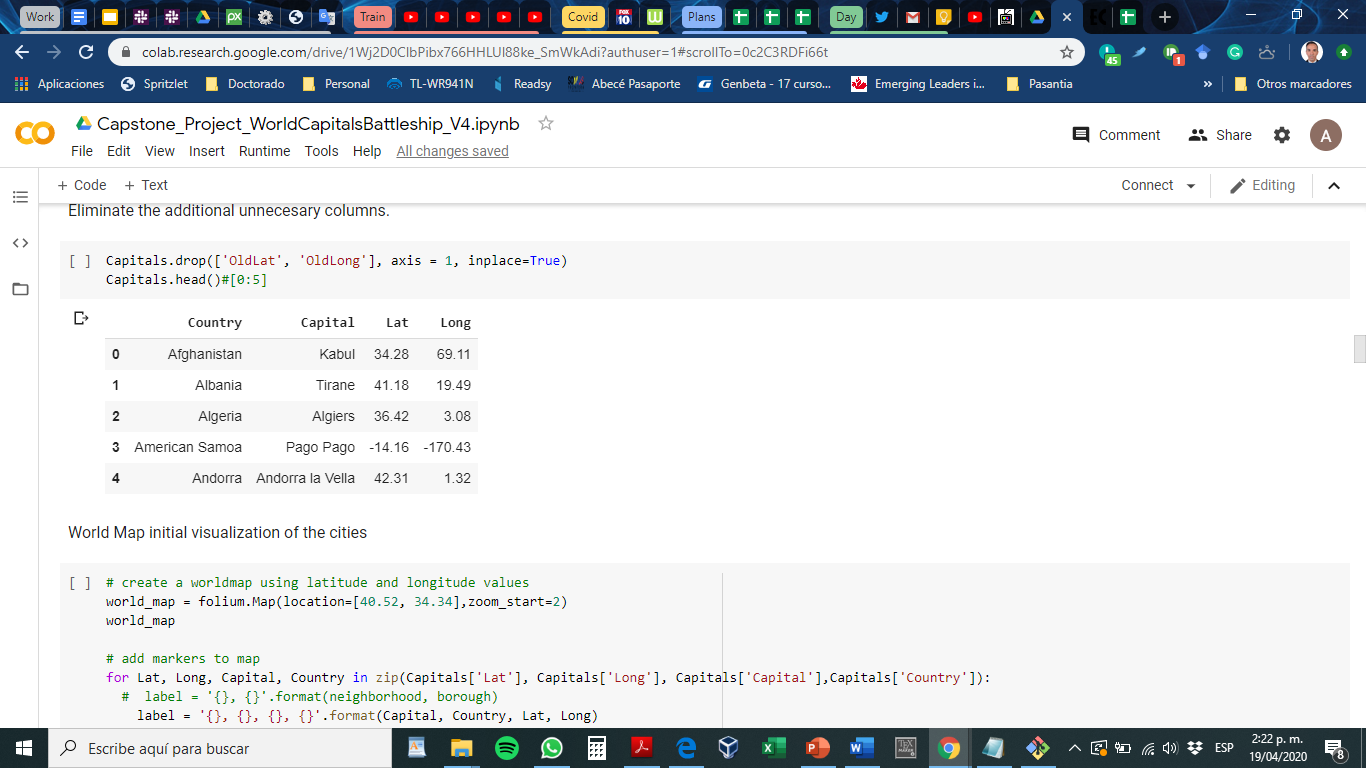
I used python’s BeautifulSoup library to get the initial data, the geo-localization of each capital from this website:

"https://lab.lmnixon.org/4th/worldcapitals.html").

The following figure is an example of the initial input. For every capital of the world, the respective geo-localization coordinates and the country they belong to.

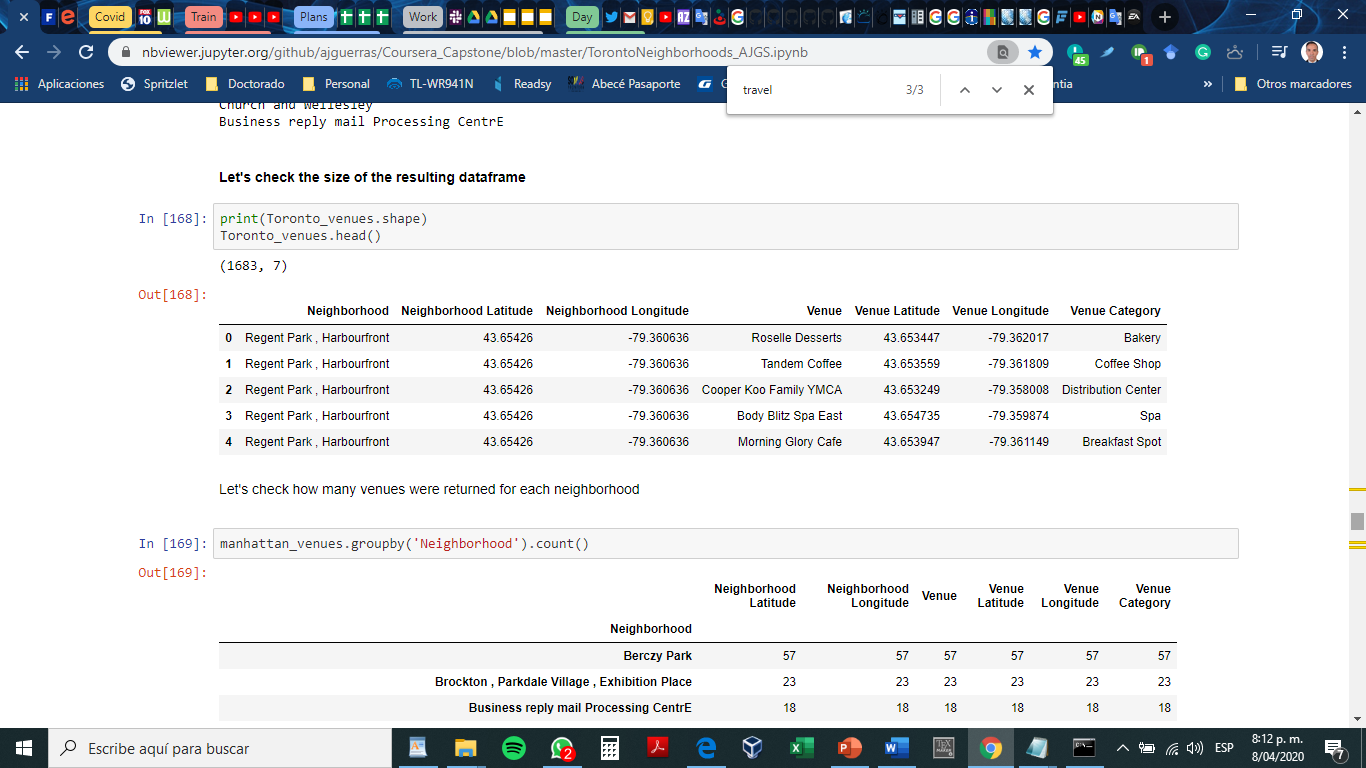


I had to make a couple of transformations to get such data in a suitable format for the Machine Learning process I was going to do after. Final data looked like this:



* 1. Venue data

From Foursquare database we can retrieve relevant information in the following format:



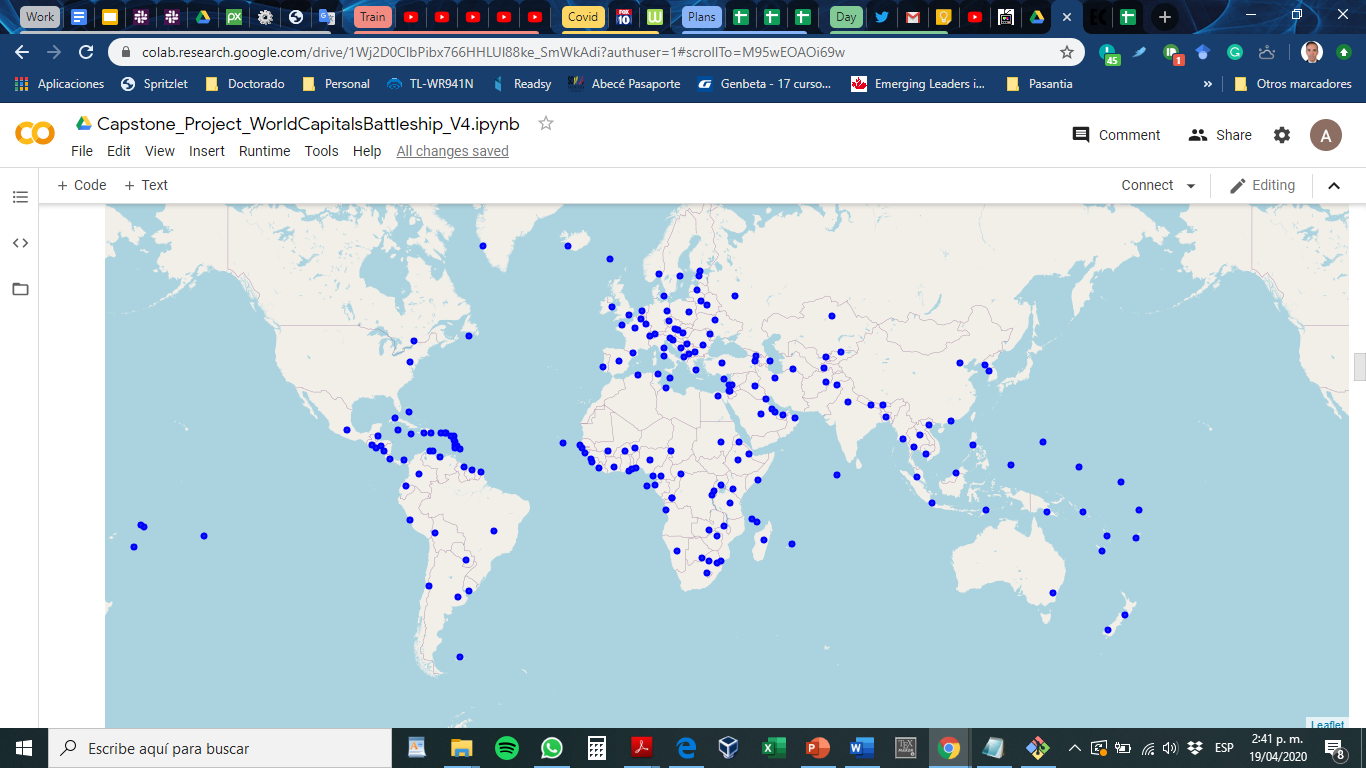
A maximum of 100 venues is returned by Foursquare for each request. Also, the maximum number of requests per day was 950, which was very little considering that I worked with more than 200 capitals cities of world.

1. **Methodology** 
   1. **Folium Maps**

All of the maps, and the respective clusters will be displayed using the Folium library. Folium builds on the data wrangling strengths of the Python ecosystem and the mapping strengths of the leaflet.js library. Manipulate your data in Python, then visualize it in on a Leaflet map via folium.

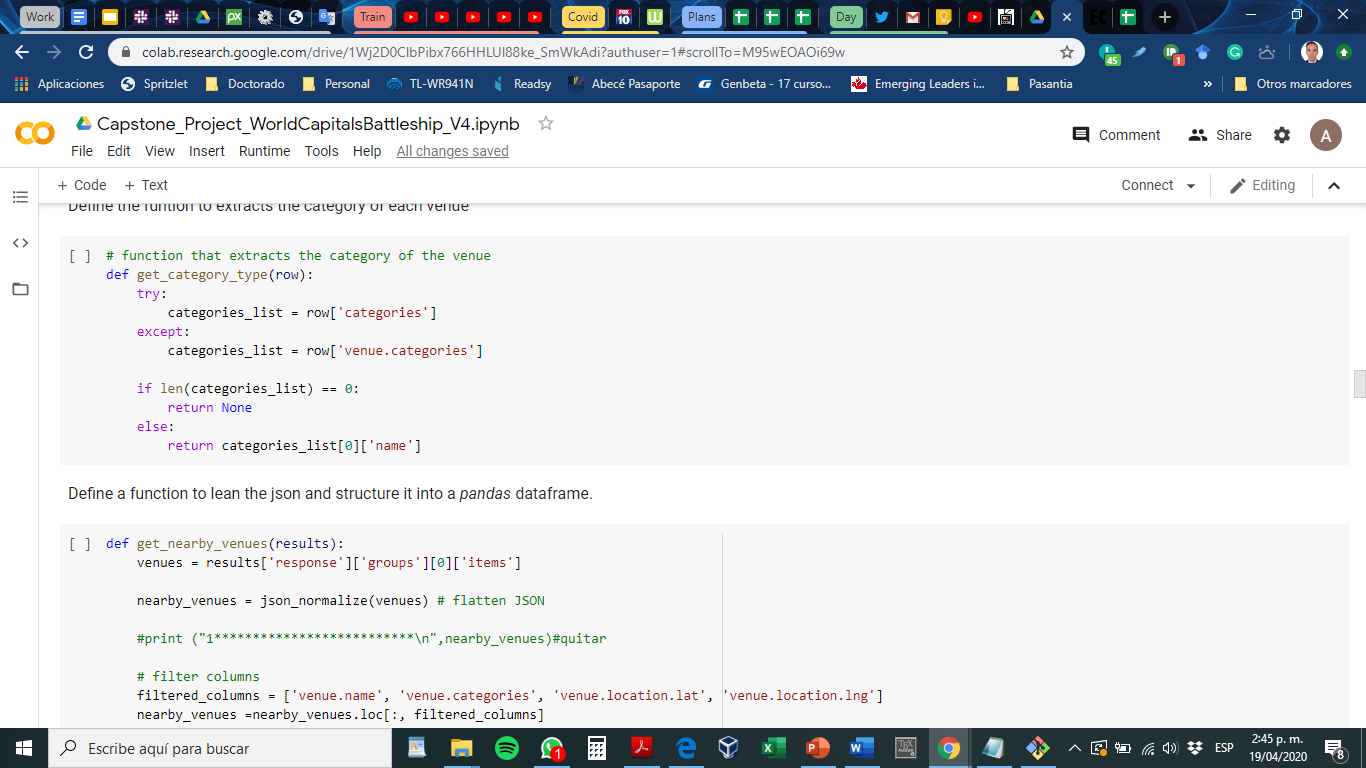
Folium makes it easy to visualize data that’s been manipulated in Python on an interactive leaflet map. It enables both the binding of data to a map for choropleth visualizations as well as passing rich vector/raster/HTML visualizations as markers on the map. For more information, visit: <https://python-visualization.github.io/folium/>

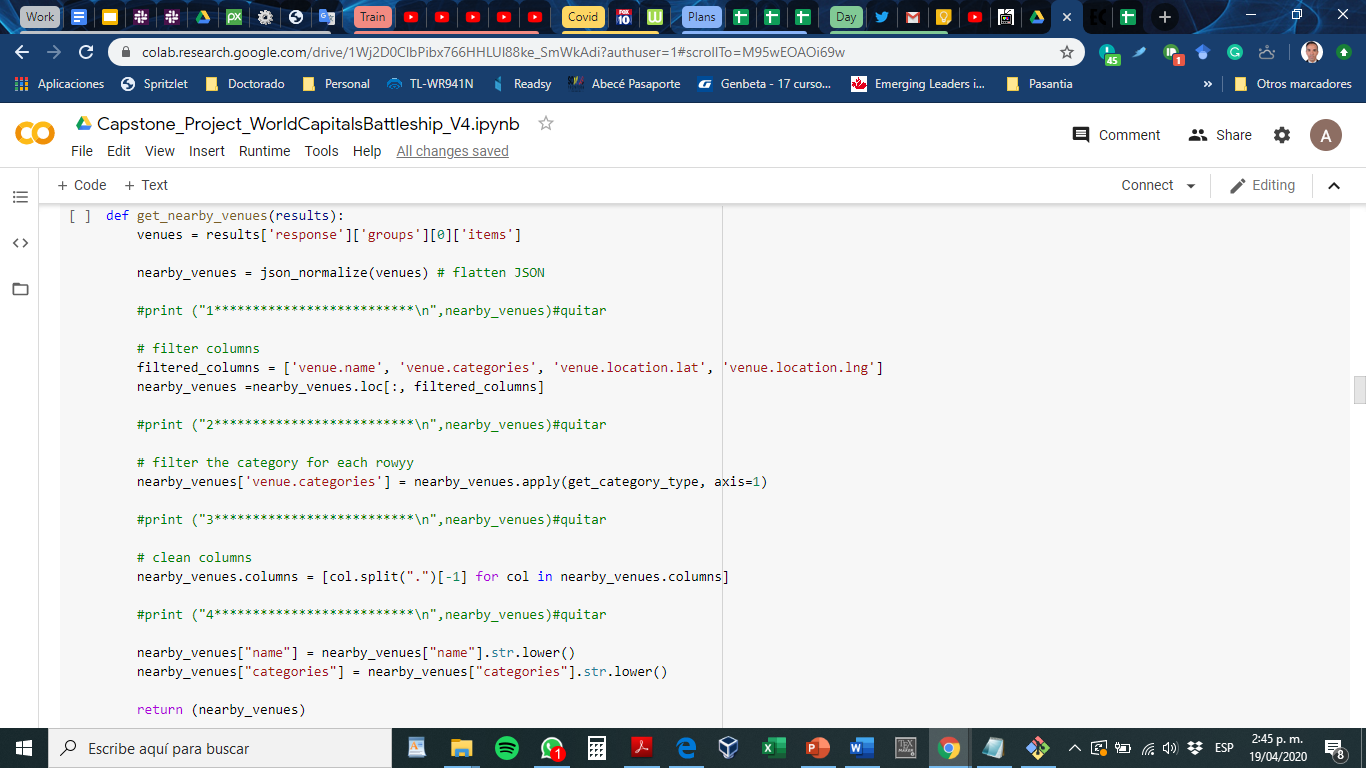
This is the first view of the world capitals using folium:

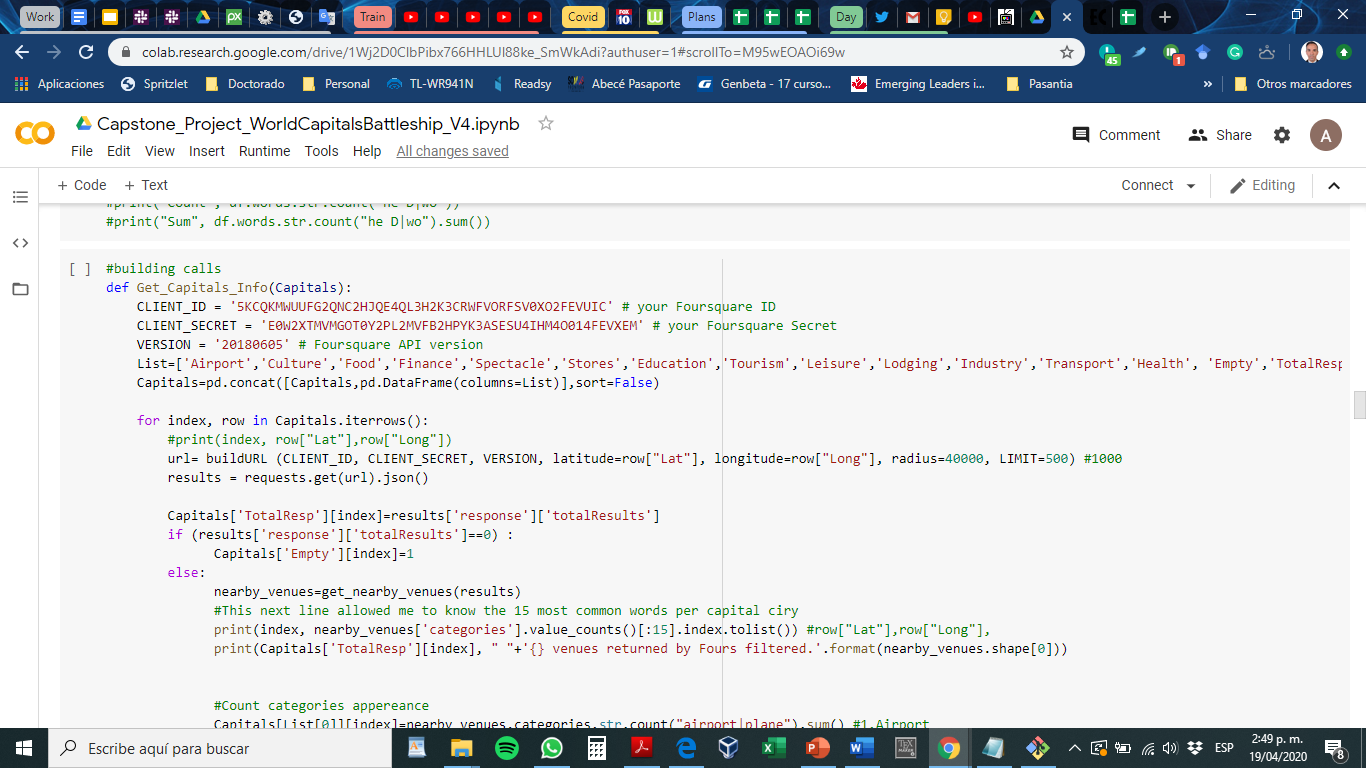


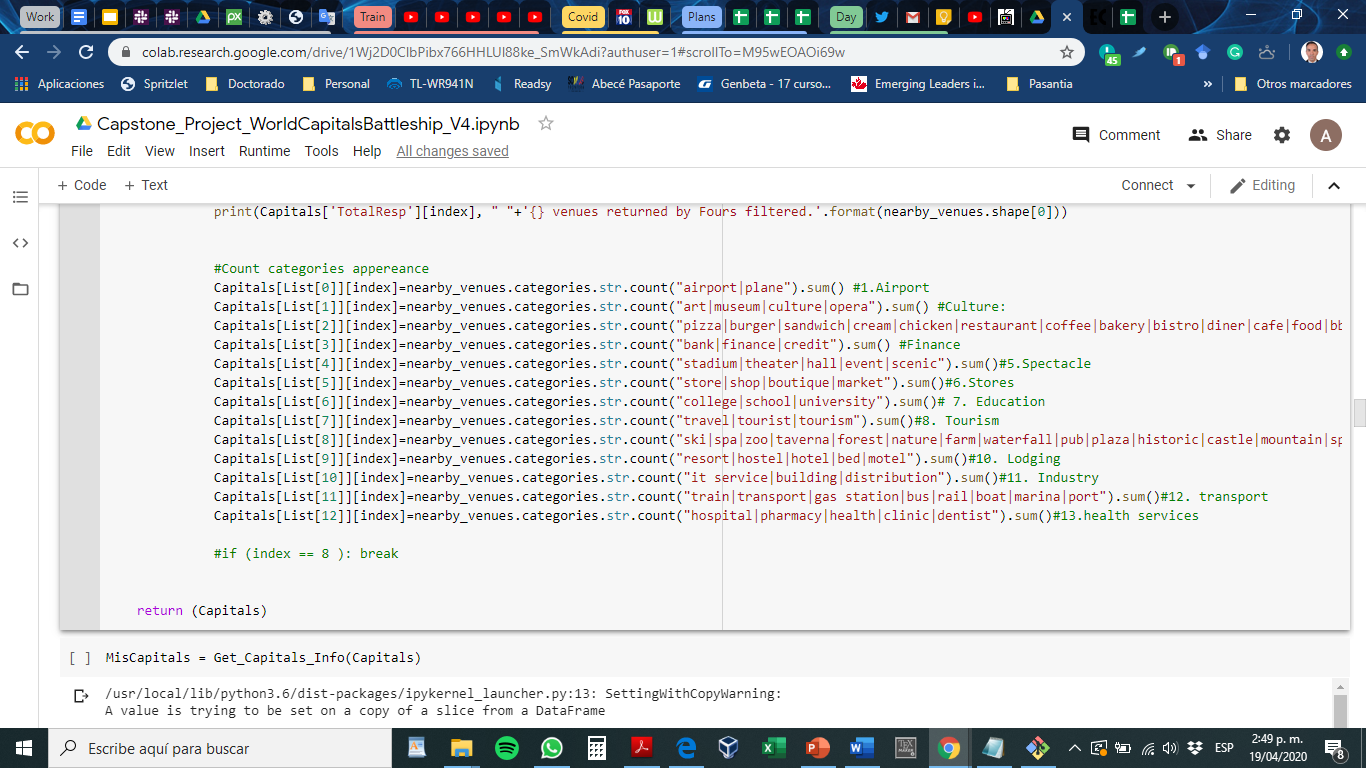
* 1. **Data preprocessing**

I needed to extract and structure to the information retrieved from the Foursquare database. I based my work on useful functions provided in the 3rd week of the IBM Capstone project.

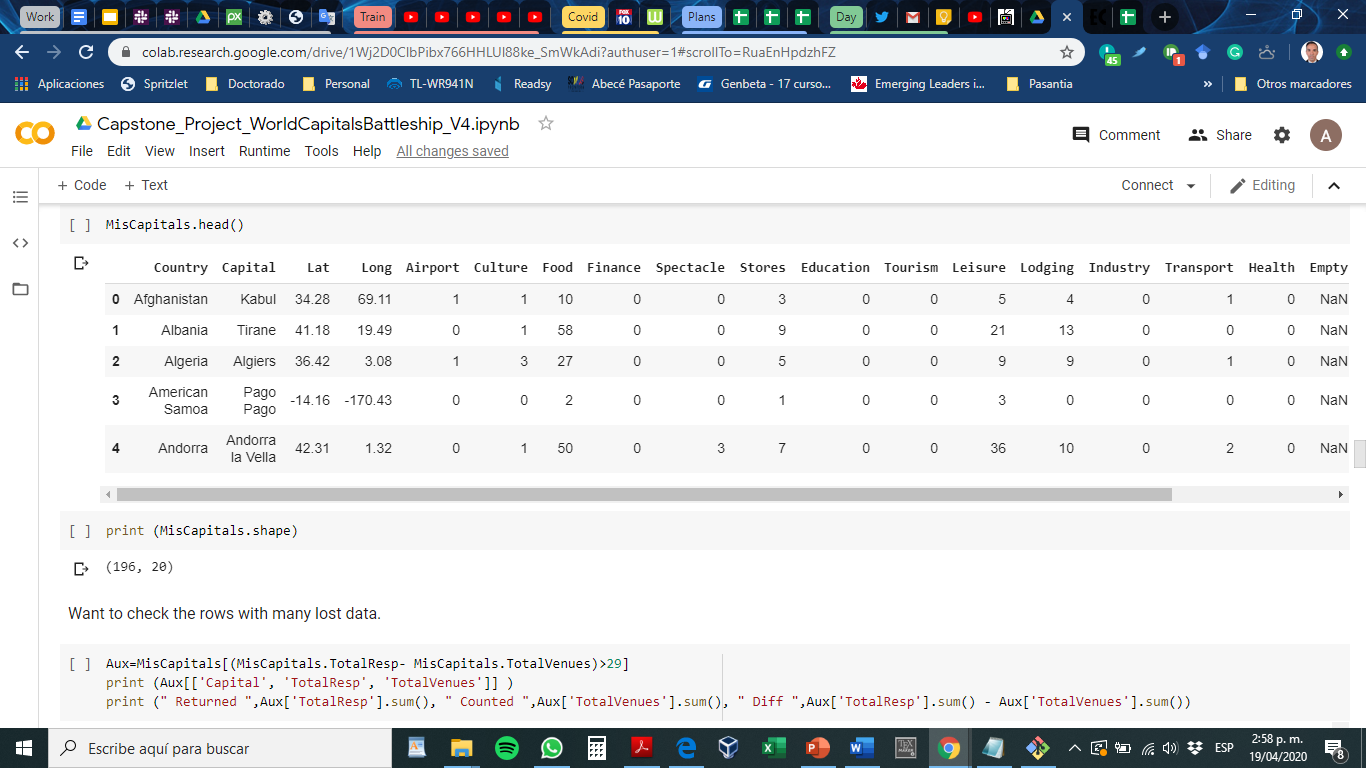


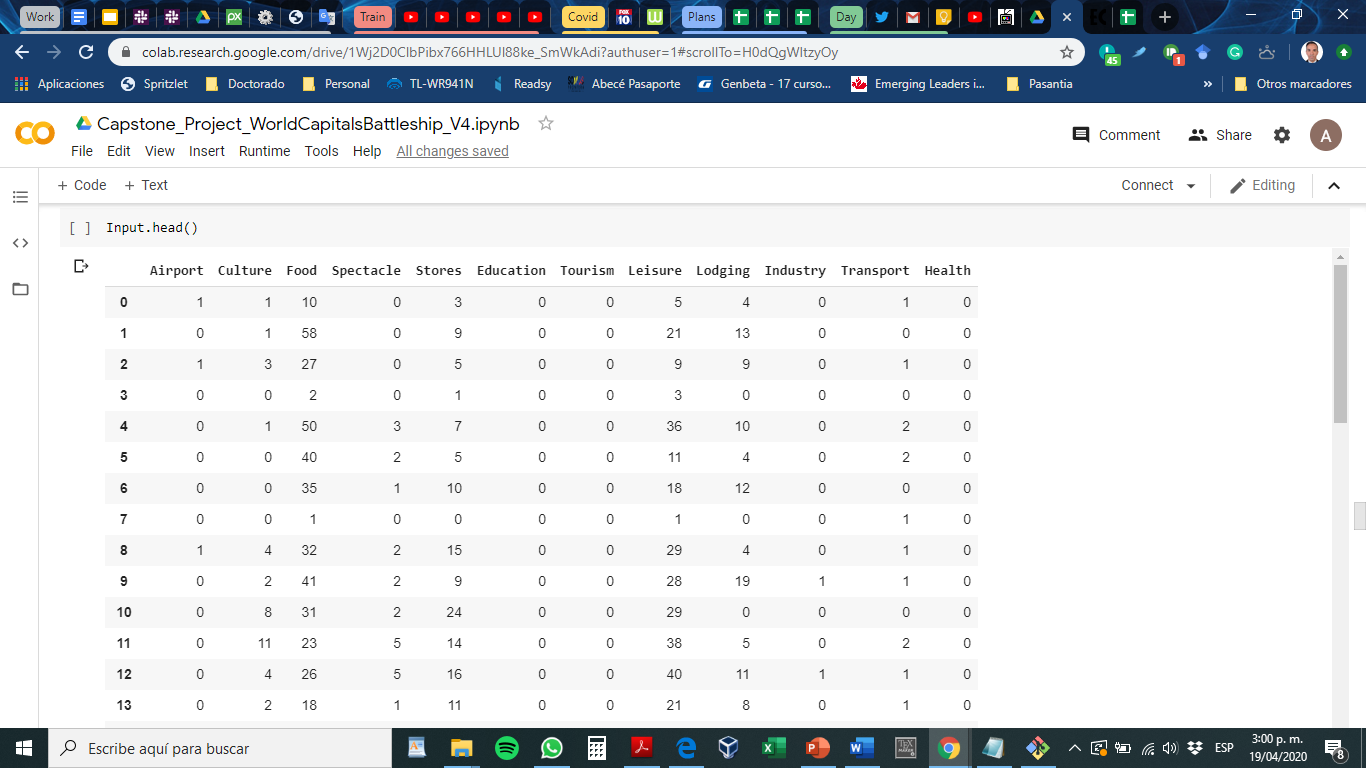


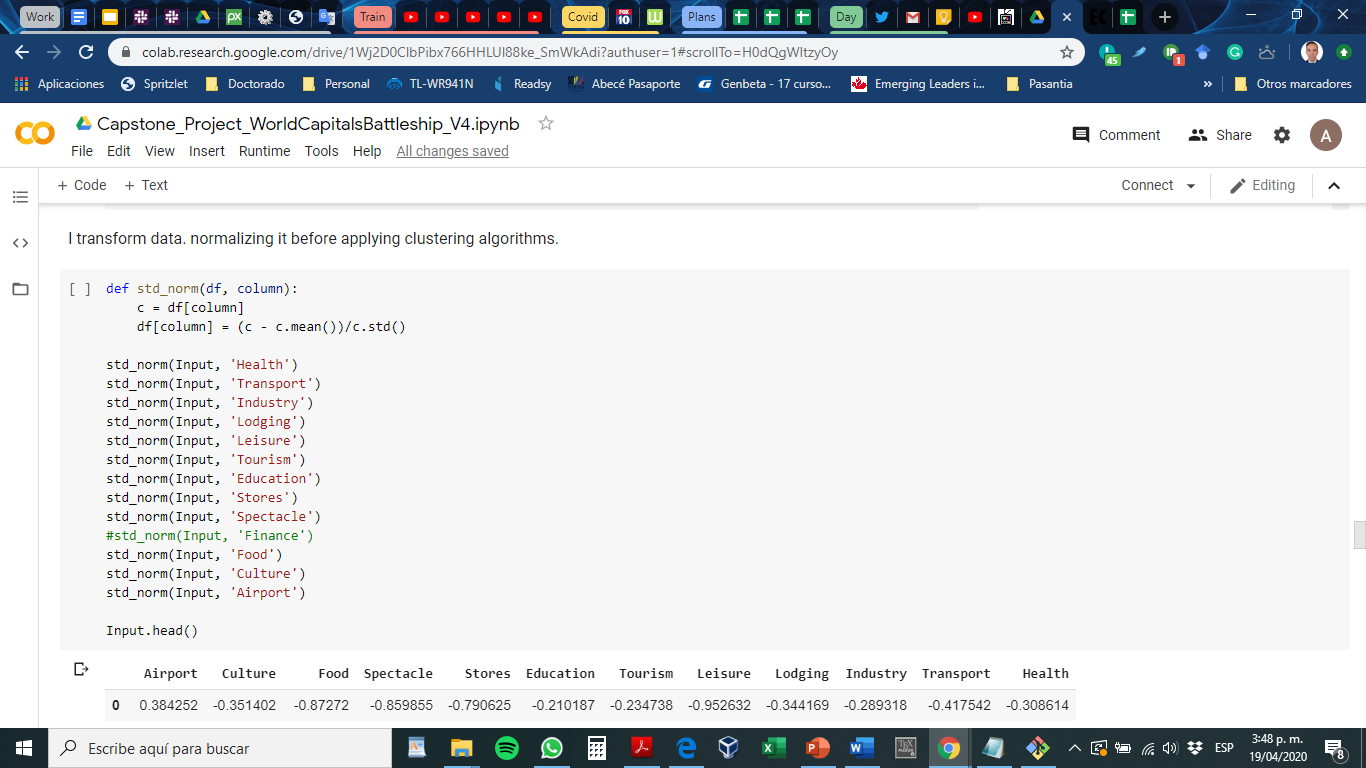


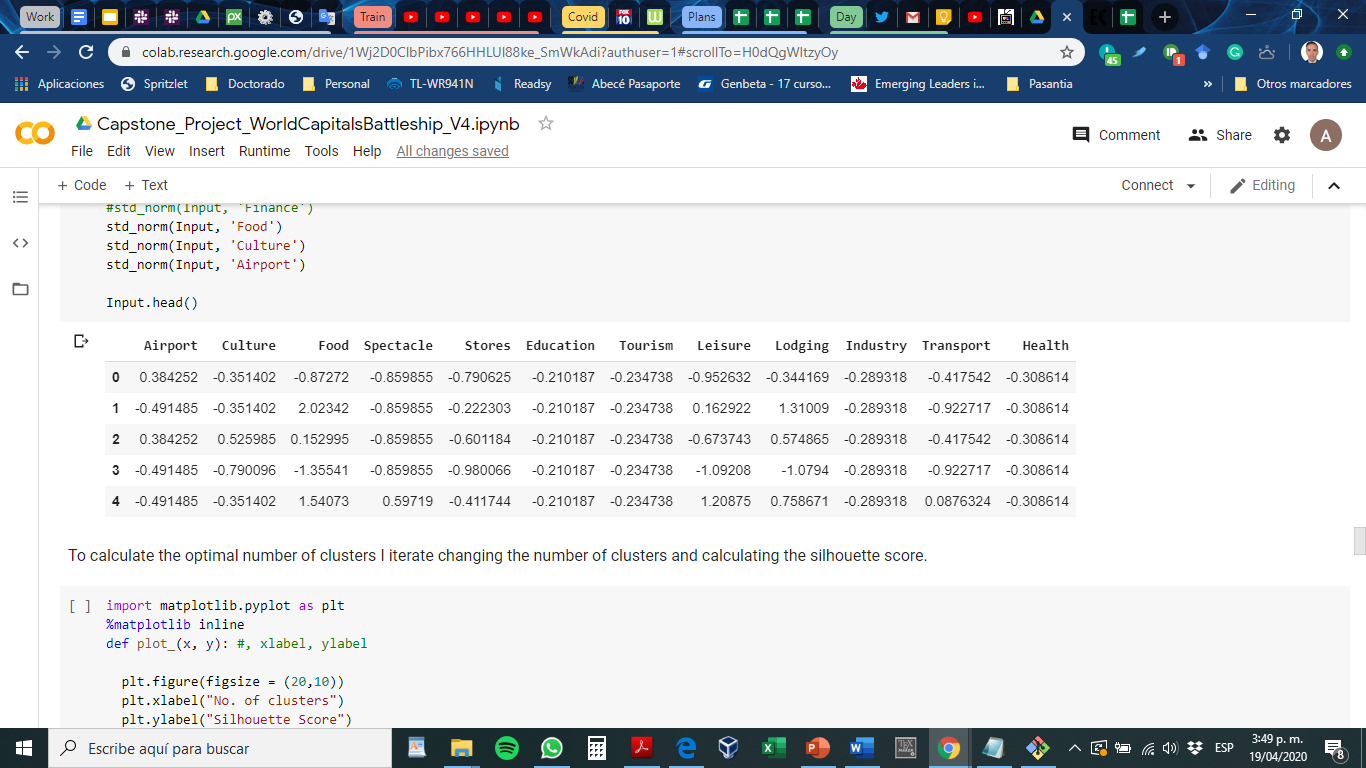


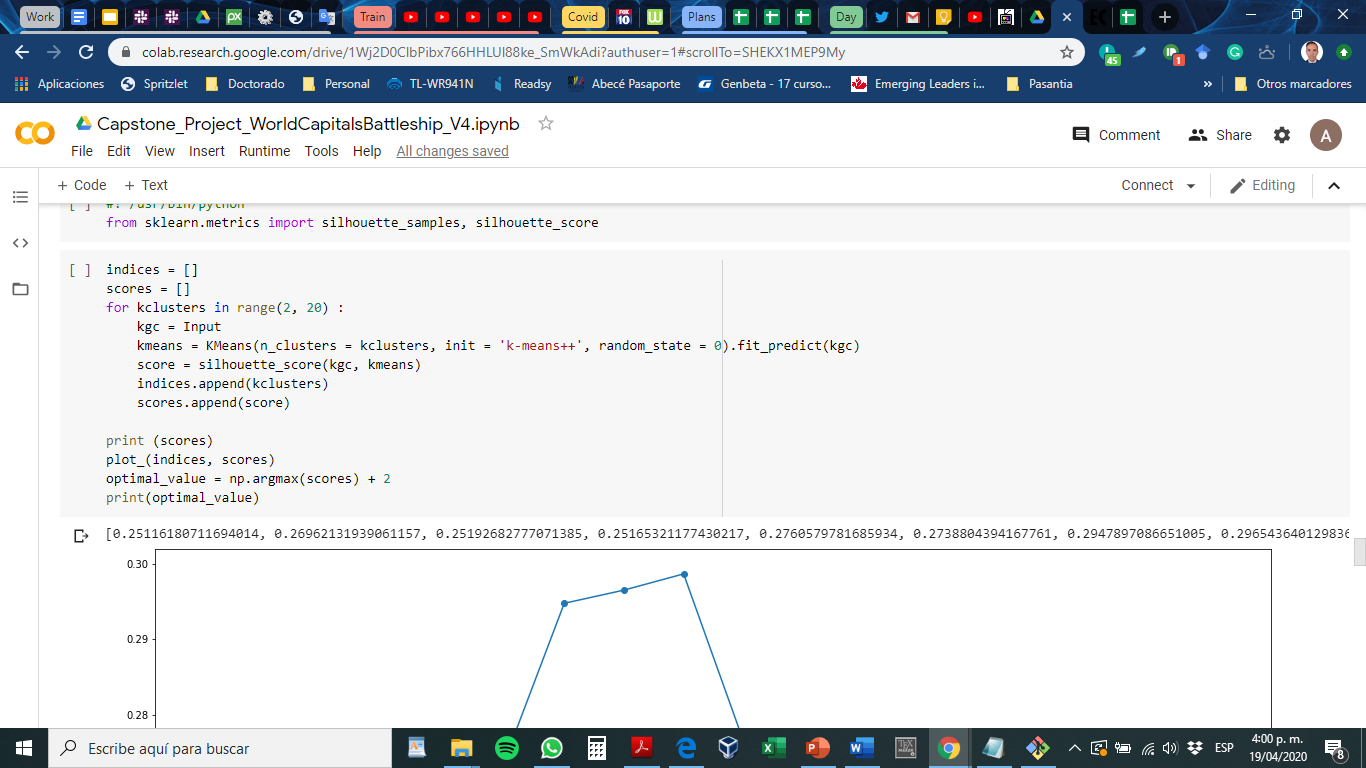


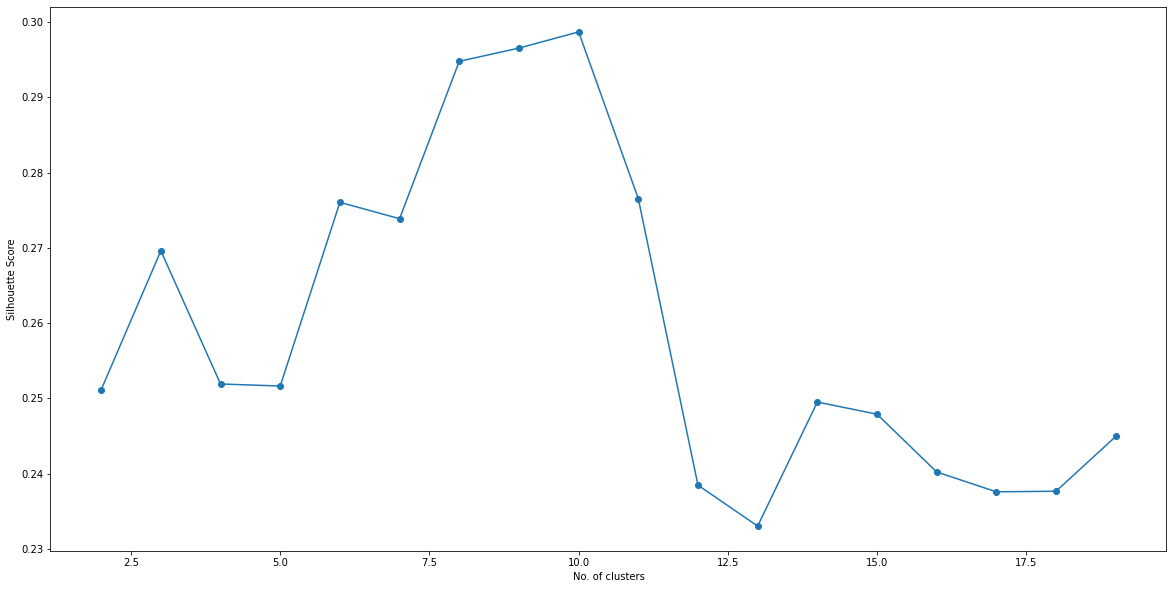




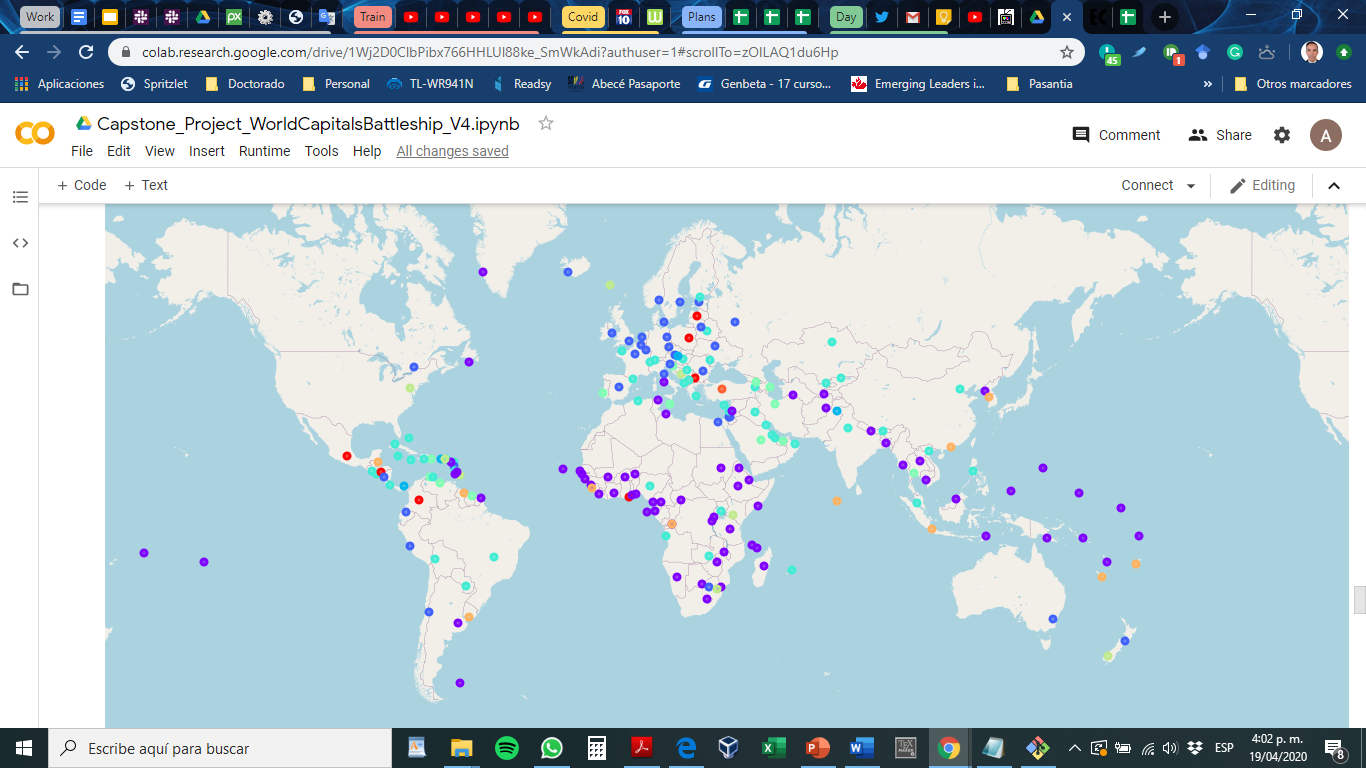






****

**Machine Learning – Clustering**



The decision model would be based on the amount of venues from certain categories per city. So, prior to counting the venues, I plan to apply some kind of filter that create classes inside the venue’s dataset. For example: Food Services (Coffee, Restaurants, Breakfast places), health services (Hospital, Pharmacies), lodging services (hostels, hotels), etc.

Optimal number of clusters.

Silhouette Score is a measure of how similar an object is to its own cluster (cohesion) compared to

other clusters (separation). The silhouette ranges from -1 to +1, where a high value indicates that the

object is well matched to its own cluster and poorly matched to neighboring clusters. Based on the

Silhouette Score of various clusters below 20, the optimal cluster size is determined.

**4.6 K-means clustering**

The venue data is then trained using K-means Clustering Algorithm to get the desired clusters to base

the analysis on. K-means was chosen as the variables (Venue Categories) are huge, and in such situations

K-means will be computationally faster than other clustering algorithms.

7

**Most common venues per cluster:**

**5 Results**

The neighbourhoods are divided into n clusters where n is the number of clusters found using the optimal

approach. The clustered neighbourhoods are visualized using different colours so as to make them

distinguishable.

Code

**6 Discussion**

After analyzing the various clusters produced by the Machine learning algorithm, cluster no.14, is a

prime fit to solving the problem of finding a cluster with common venue as a train station mentioned

before.

Figure 4: Cluster having Train Station as most common venue

The five places namely Bijpur, Garshyamnagar, Halisahar, Hind Motor and Kodalia fall in the

outskirts of the city of Kolkata, hence the demographic of the population in these areas fall under

the lower middle class of the society.

According to most organizations, like the World Bank and the Organization for the Economic Cooperation

and Development (OECD), people living on less than US $2 a day are con- sidered poor. For those

in the middle classes, the earnings typically lie in the range of US $10 to $100 per day, as expressed in

the 2015 purchasing power parities.

India is expected to see a dramatic growth in the middle class, from 5 to 10 percent of the population in

2005 to 90 percent in 2039, by which time a billion people will be added to this group. In 2005, the mean

per capita household expenditure was just US $3.20 per day, and very few households exceeded incomes

of US $5 per day. Yet, by 2015, half the population had crossed this threshold. By 2025, half the Indian

population is expected to surpass US $10 per day.

#! /usr/bin/python

Venues Categories:

1.- Airport: has the word “airport”, plane

2.- Food services: Restaurant, coffee, bakery, bistro, diner, cafe

3.- Culture: Art, Museum

4.- Bank: Bank

5.- Spectacles, and big events: Stadium, Theater, hall

6.- Shops: Store, shop, boutique, market

7.- education services: College

8.- Tourism: travel

9.- Entertainment, Activities, leisure: Trail, Gym, club, music, pool, park, playground, plaza, aquarium, bar, church, park, Mall

10 .- lodging services : hostel, hotel,

11.- Technological services: IT services

12.- Transport facilities: train , transportation, gas station, Bus, rail

13.- Health services: hospital, pharmacy

Intenta discriminar por cluster la categoria o variable dominante

**6 Discussion**

After analyzing the various clusters produced by the Machine learning algorithm, cluster no.14, is a

prime fit to solving the problem of finding a cluster with common venue as a train station mentioned

before.

Figure 4: Cluster having Train Station as most common venue

The five places namely Bijpur, Garshyamnagar, Halisahar, Hind Motor and Kodalia fall in the

outskirts of the city of Kolkata, hence the demographic of the population in these areas fall under

the lower middle class of the society.

According to most organizations, like the World Bank and the Organization for the Economic Cooperation

and Development (OECD), people living on less than US $2 a day are con- sidered poor. For those

in the middle classes, the earnings typically lie in the range of US $10 to $100 per day, as expressed in

the 2015 purchasing power parities.

India is expected to see a dramatic growth in the middle class, from 5 to 10 percent of the population in

2005 to 90 percent in 2039, by which time a billion people will be added to this group. In 2005, the mean

per capita household expenditure was just US $3.20 per day, and very few households exceeded incomes

of US $5 per day. Yet, by 2015, half the population had crossed this threshold. By 2025, half the Indian

population is expected to surpass US $10 per day.

Limpiar codigo del notebook

Reporte pdf

Presetnacion ppt