1.0 INTRODUCTION

In 2019, the global tourism industry was worth over $5 Trillion dollars according to the United Nations World Tourism Organization (UNTWO), with the United States contributing a whopping $580 Billion (>11%).

UNTWO defines a tourist as someone who travels atleast 80km from his or her home for atleast 24 hours for business, leisure and/or other reasons. From the regal streets of Grand Bazaar in Istanbul to Nakamise street of Sensoji Temple in Tokyo, over 1.5 billion people thronged to different travel destinations in 2019.

Tourism is a great economic contributor and its impact can be felt across the following industries:

1. Accommodation
2. Food and Beverage Services
3. Recreation and Entertainment
4. Transportation
5. Travel Services
6. Retail Trade (souvenirs and the like)

For the aforementioned reasons, developing countries are looking to standardize their current tourism sites in order to attract international, continental and local tourists.

2.0 BUSINESS PROBLEM

My client is a West African country with breathtaking rainforests and a vast variety of wildlife in its savannah alongside other historical sites. The objective of this project is to investigate the ancillary infrastructure surrounding the best tourist sites in the world with the view of strategically replicating such infrastructure to ensure the best experience for potential tourists in order to maximize the impact on the local economy.

This objective will be achieved by randomly selecting 10 popular tourist attractions across 5 continents, exploring existing outlets within 600 metres radius of the tourist site by using Foursquare location API and clustering similar outlets using K-Means Clustering algorithm.

The table below contains the biggest tourist spenders of 2018;

Country	Amount Spent ($)
China	277 billion
United States	144 billion
Germany	94 billion
United Kingdom	76 billion
France	48 billion
Australia	37 billion

Russia	35 billion
Canada	33 billion
South Korea	32 billion
Italy	30 billion

How does this West African country build its tourism infrastructure to attract these billions of dollars?

3.0 DATA

The data for this project was retrieved and process through multiple sources, however, the core data required are in two segments;

- Longitude and Latitude of the 10 tourist locations
- Location exploration using the above data points

The longitude and latitude were obtained manually because of the randomness of the locations while the location exploration will be obtained using Foursquare API.

3.1 Tourist Site Selection

The major criteria for selecting the tourist sites are as follows:

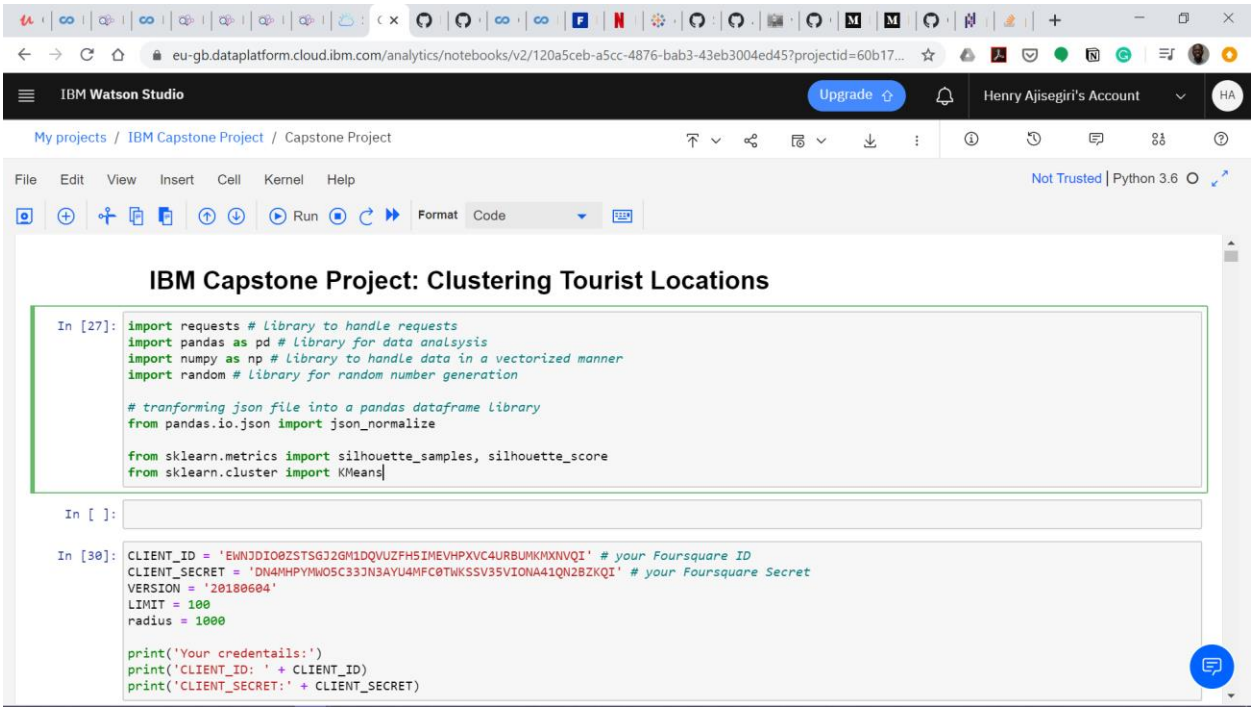
- The site must receive atleast 1m visitors yearly
- The availability of the site's data on foursquare's database

The selected tourist sites are in the table below;

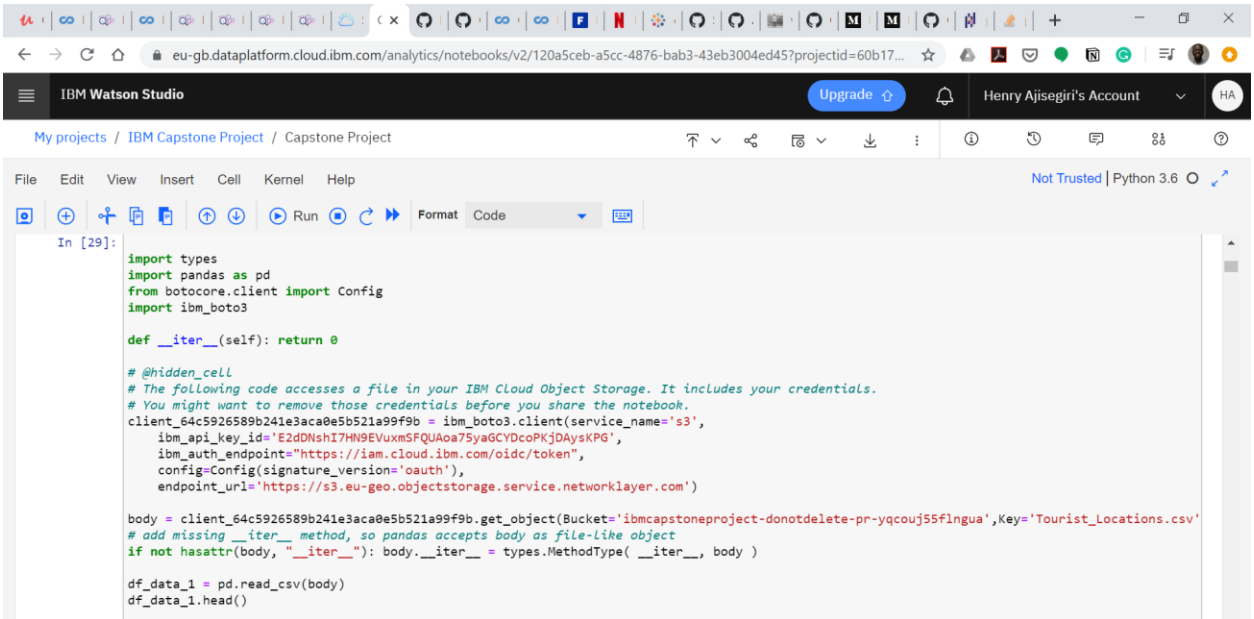
Tourist Site	Country	Country Code
The Forbidden City	China	CN
The Grand Palace	Thailand	TH
The Grand Bazaar	Turkey	TR
Sacre-Coeur Basilica	France	FR
St. Peter's Basilica	Vatican/Italy	VT/IT
Taj Mahal	India	IN
The Acropolis	Greece	GR
Eiffel Tower	France	FR
Sensoji Temple	Japan	JP
Burj Khalifa	UAE	AE

4.0 METHODOLOGY

The entire analysis was performed using Python Language and Machine Learning algorithm (K Means) in a Jupyter IDE hosted on IBM Watson Studio. The libraries used are shown in the image below;



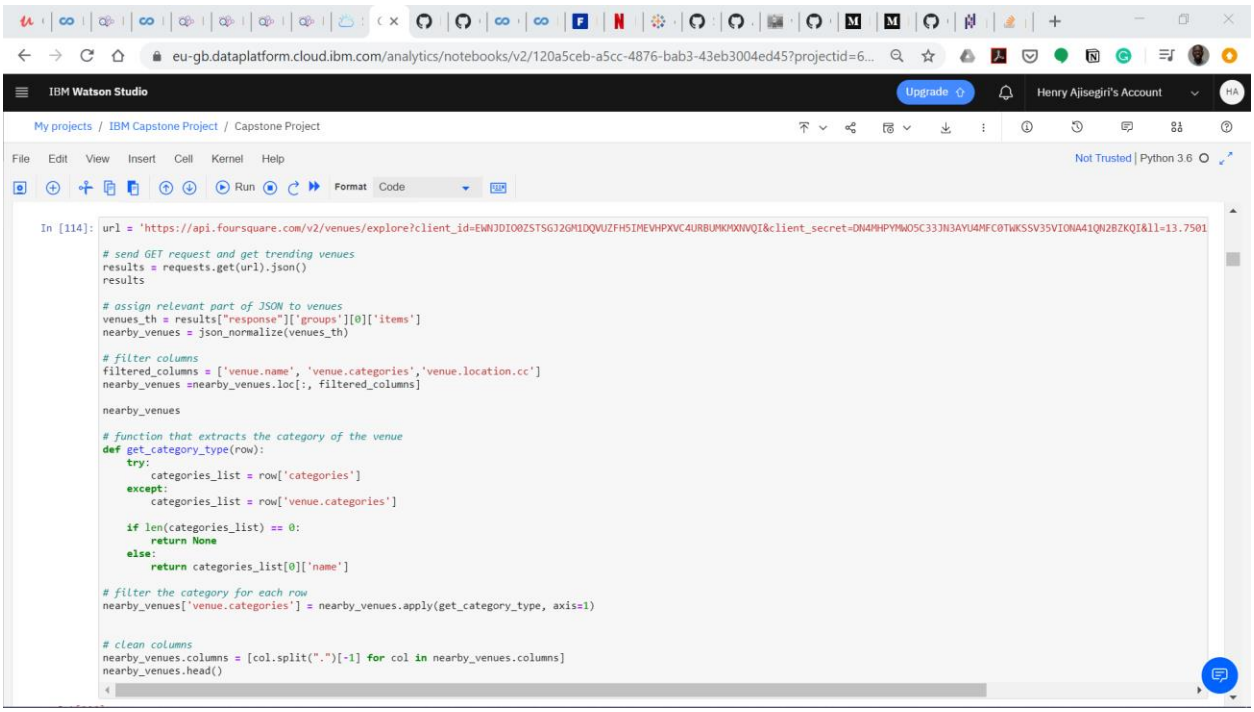
The manually created csv file was loaded directly into Watson studio as shown below;



5.0 ANALYSIS

5.1 Foursquare

Foursquare's API was used to extract the details of the most popular spots around the tourist sites earlier mentioned. The specified radius coverage is 1000m and the number of spots extracted was limited to 100 per site. The image below shows the Foursquare API call used to extract the required information and the data wrangling process to make the data presentable; this code was run 10 times for the 10 different locations selected.

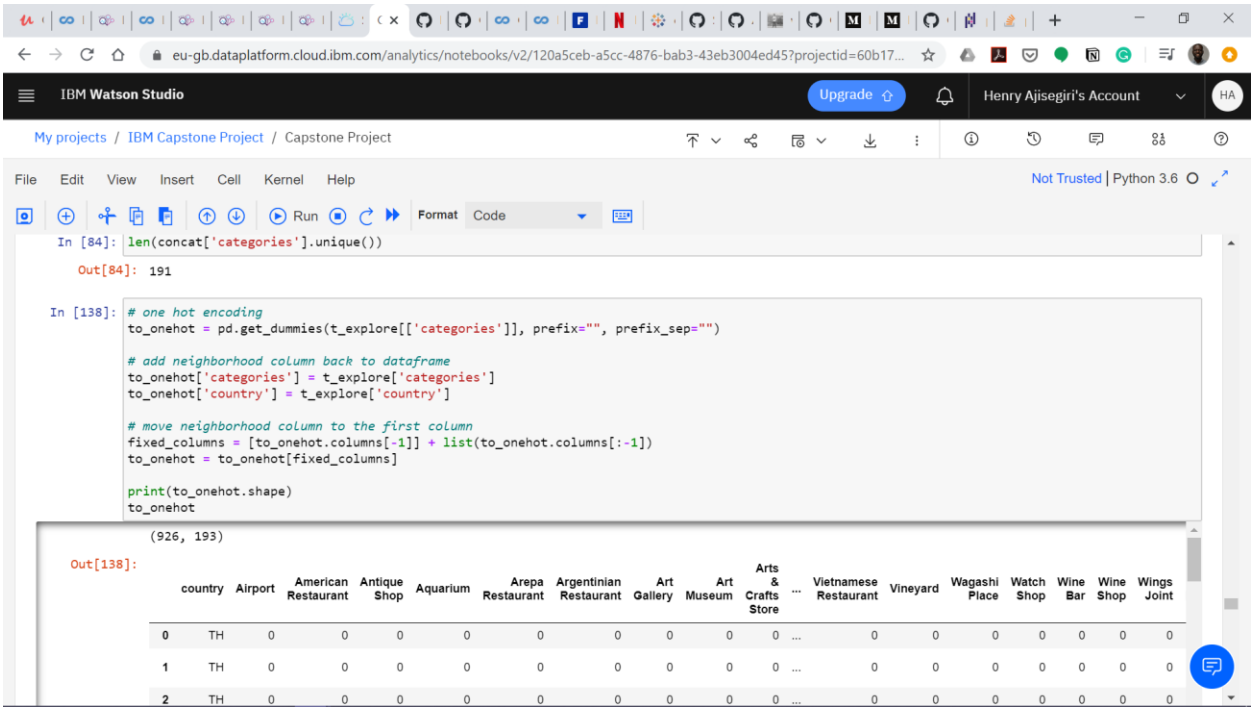


Key steps

- The required authentication details were filled into the API url
- Necessary information was read from the json file into a dataframe
- The dataframe was cleaned up for better presentation and easier operations down the line

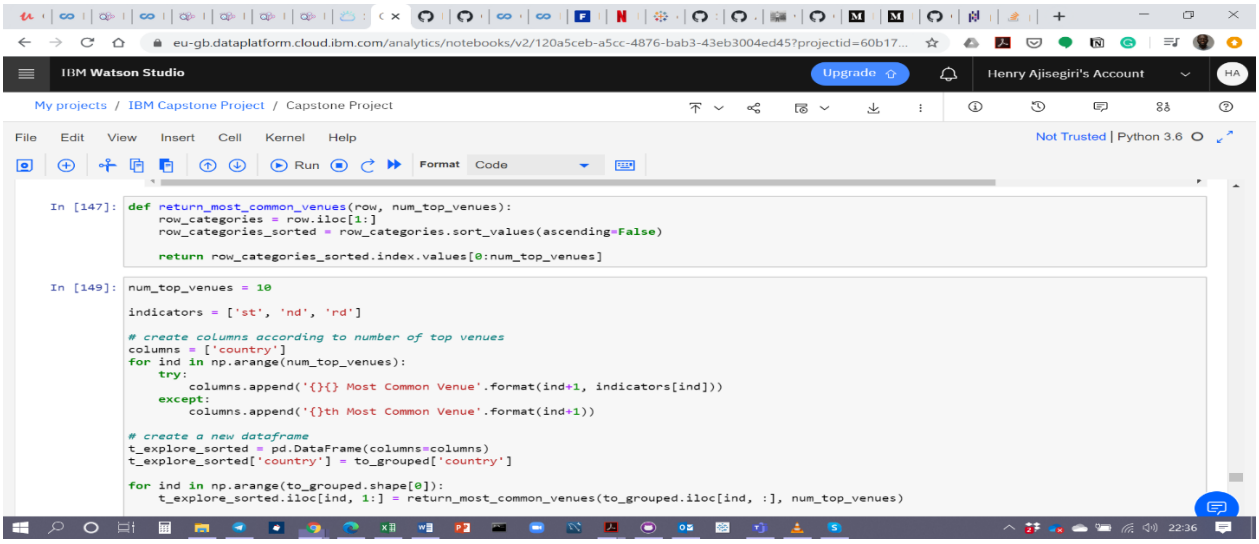
5.2 One Hot Encoding

One hot encoding is the process through which categorical variables are converted into a form (usually numerical) that can be provided to Machine Learning algorithms in order to perform efficiently during fitting and prediction. As stated earlier, K Means Clustering algorithm was used for this project and all unique items in the 'categories' column was one hot encoded.



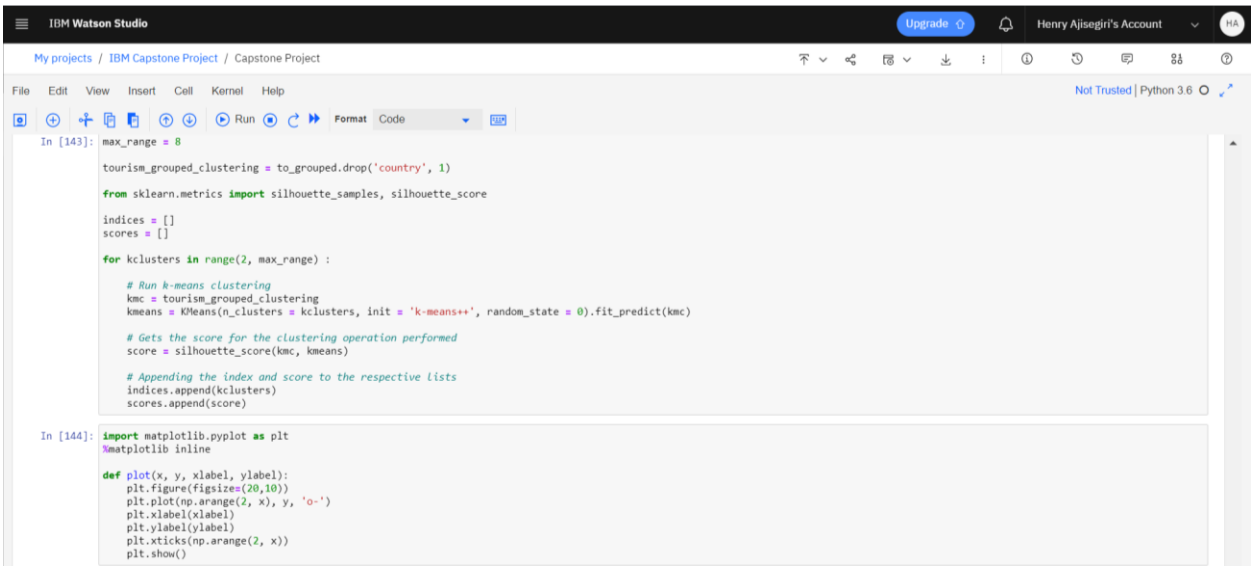
5.3 Ten (10) Most Common Venues

From over 900 locations and 191 unique categories, the 10 most popular venues were selected into a new dataframe for the purpose of training the K Means Clustering Algorithm.

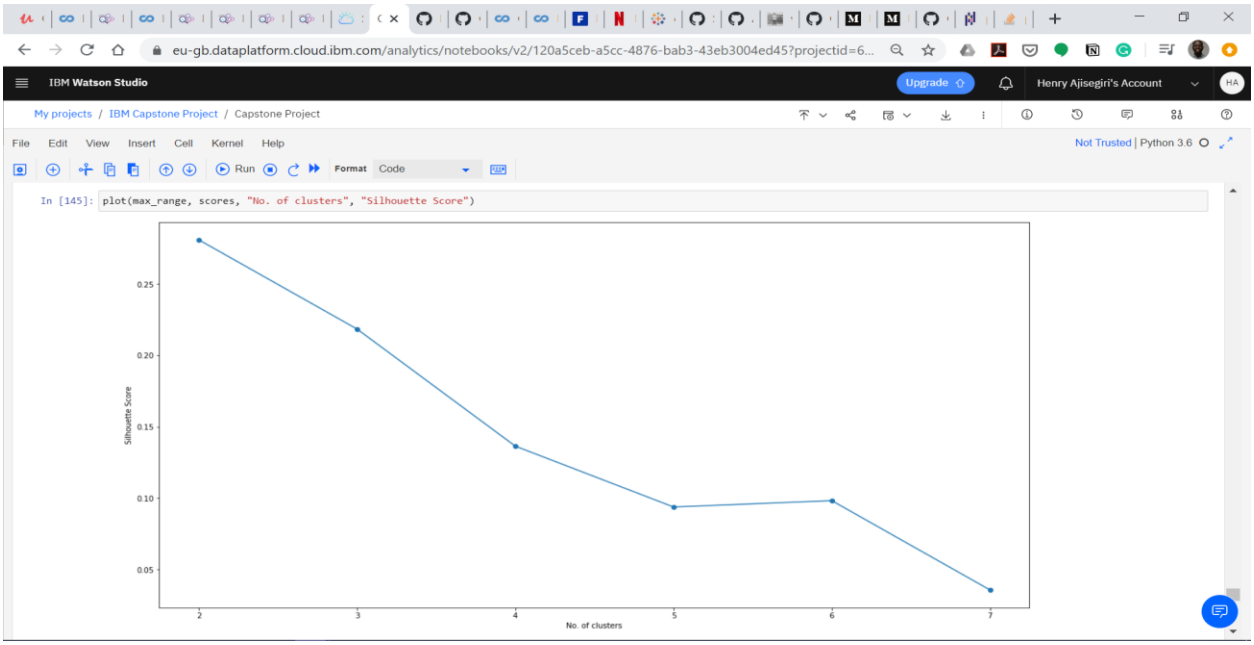


5.4 Silhouette Score

This 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 neighbouring clusters. Based on the Silhouette Score of various clusters below 20, the optimal number of clusters was determined.



With the codes above, the graph below was plotted;

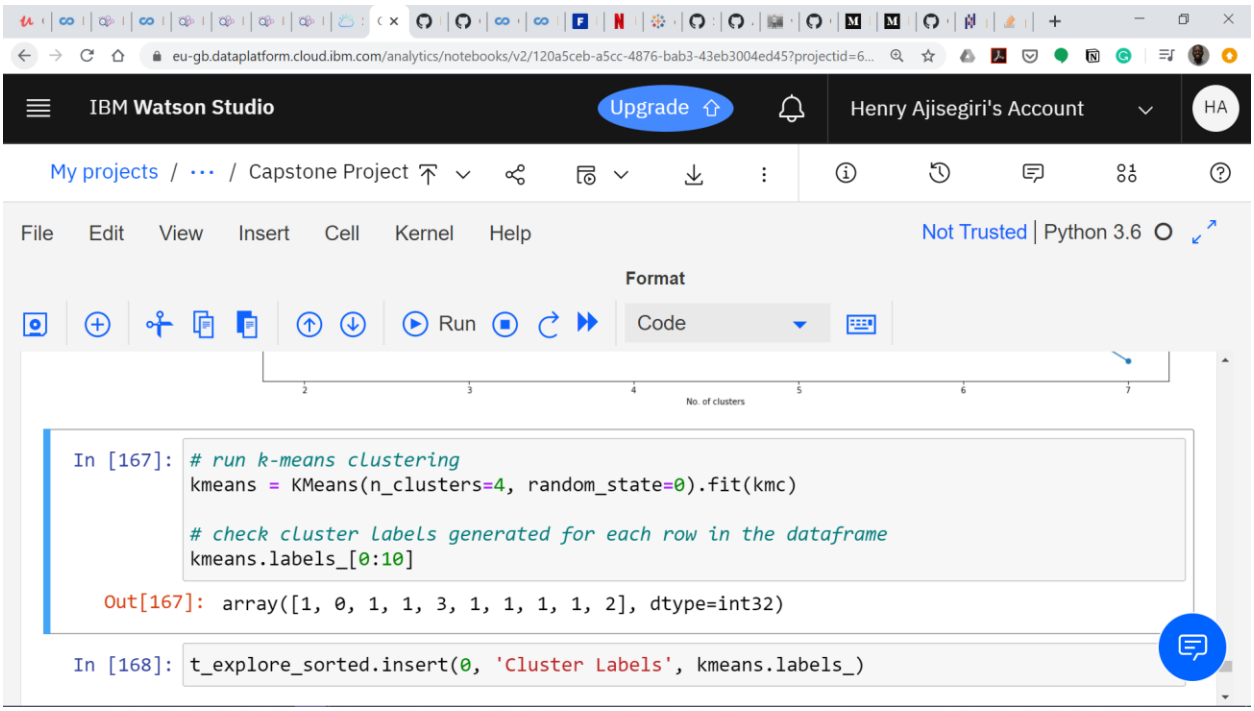


From the Silhouette Score, the optimal number of clusters is two (2). However, since the business problem is not strongly based on high precision and accuracy, the K Mean Clustering algorithm was performed with four (4) clusters.

5.5 K Means Clustering

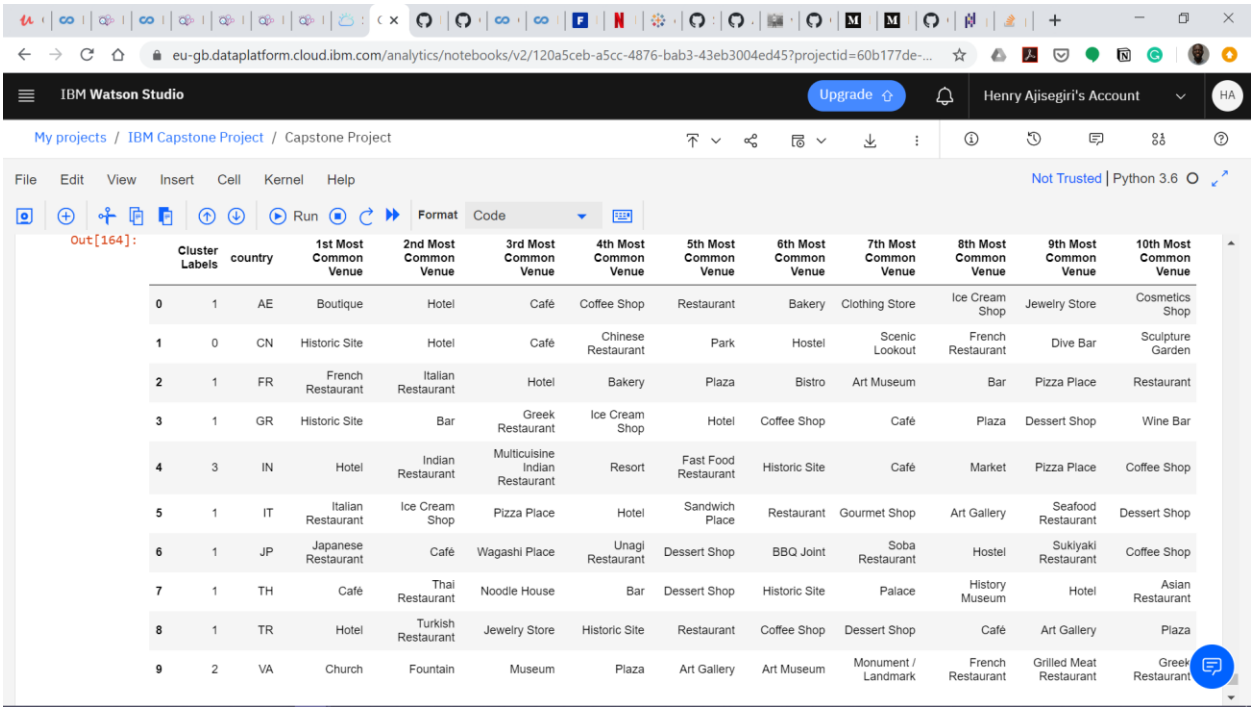
The data obtained from the tourist sites' exploration was trained using the K Means Clustering algorithm in order to group the variation of ancillary tourist attractions into clusters which will serve as insight(s) to help the client take the best approach in creating an enabling environment for tourism to thrive in the country.

K Means was selected because of the size of the variables as K Means is computationally faster than other clustering algorithms.



6.0 RESULTS & DISCUSSION

The countries were divided into 4 clusters but the clusters cannot be visualized with a map because the locations selected are vastly dispersed across the world. However, the result can be viewed on a table.



	Cluster Labels	country	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	1	AE	Boutique	Hotel	Café	Coffee Shop	Restaurant	Bakery	Clothing Store	Ice Cream Shop	Jewelry Store	Cosmetics Shop
1	0	CN	Historic Site	Hotel	Café	Chinese Restaurant	Park	Hostel	Scenic Lookout	French Restaurant	Dive Bar	Sculpture Garden
2	1	FR	French Restaurant	Italian Restaurant	Hotel	Bakery	Plaza	Bistro	Art Museum	Bar	Pizza Place	Restaurant
3	1	GR	Historic Site	Bar	Greek Restaurant	Ice Cream Shop	Hotel	Coffee Shop	Café	Plaza	Dessert Shop	Wine Bar
4	3	IN	Hotel	Indian Restaurant	Multicuisine Indian Restaurant	Resort	Fast Food Restaurant	Historic Site	Café	Market	Pizza Place	Coffee Shop
5	1	IT	Italian Restaurant	Ice Cream Shop	Pizza Place	Hotel	Sandwich Place	Restaurant	Gourmet Shop	Art Gallery	Seafood Restaurant	Dessert Shop
6	1	JP	Japanese Restaurant	Café	Wagashi Place	Unagi Restaurant	Dessert Shop	BBQ Joint	Soba Restaurant	Hostel	Sukiyaki Restaurant	Coffee Shop
7	1	TH	Café	Thai Restaurant	Noodle House	Bar	Dessert Shop	Historic Site	Palace	History Museum	Hotel	Asian Restaurant
8	1	TR	Hotel	Turkish Restaurant	Jewelry Store	Historic Site	Restaurant	Coffee Shop	Dessert Shop	Café	Art Gallery	Plaza
9	2	VA	Church	Fountain	Museum	Plaza	Art Gallery	Art Museum	Monument / Landmark	French Restaurant	Grilled Meat Restaurant	Greek Restaurant

From the results, most (UAE, France, Greece, Italy, Japan, Thailand and Turkey) of the countries have similar infrastructure supporting their tourism industry. China, India and the Vatican are all in different clusters.

It is clear that the client should study the tourist sites in cluster 1 as they are among the top visited tourist attractions in the world and the size of the cluster also shows that a lot of countries have shaped their infrastructure in similar fashion. It is also important to study the uniqueness of China, India and the Vatican.

From the data exploration phase, it was observed that Asian restaurants are well distributed in all tourist attractions with Italian restaurant coming a distant second. This may be a response to the huge number of Asians visiting several sites around the world or the fact that the world loves Asian food, I choose the former.

7.0 CONCLUSION

Tourism in many developing and least developed countries is the most viable and sustainable economic development option, and in some countries, the main source of foreign exchange earnings. For the client, tourism is a great option for diversification and enabling the support structure similar to cluster 1 countries will lead to more jobs via small to medium businesses, improved GDP and increased foreign exchange earnings.