CAPSTONE PROJECT REPORT- THE BATTLE OF NEIGHBORHOODS

SUMMARY

- Introduction: What project addresses and why. The problem and who it will help
- Data: description and source of data.
- Methodology: How I am addressing the problem.
- Results: discuss the analysis results.
- Observations noted and any recommendations you can make based on the results.
- Conclusion

INTRODUCTION

A tech entrepreneur is looking to start an IT Training business in Nigeria and would like to make an informed decision on where to start. At this time, they would like a location from one of these two states: Oyo state and Rivers. Although, they are aware that a lot of people when deciding to branch out an international organization to Nigeria or move back to Nigeria, they often choose Lagos, because it is considered the economic Capital of Nigeria, they would rather one of the other 2 states mentioned. Both states have a good mix of the kind of environment they are wanting, and they belief there is a lot of untapped potential.

They are looking to start this education-business because, there is a big gap between the skills that are needed in the Nigerian market and the world, versus the training the education system is providing to students. The field of IT is one of many fields with such gap. The client would like to work towards bridging this gap while also taking advantage of this business opportunity. They are wanting a location with a vibrant population of young people.

This project would help anyone who is looking for how to choose the best location to start an IT technical training school.

DATA DESCRIPTION AND SOURCE

It was very challenging gathering data on Nigeria. Nigeria is a third world country that is only beginning data exploration and as a result, not much has been gathered. After spending some time researching for data and not coming up with much, I recruited the help of a friend in Nigeria- John Ajayi. He was able to gather some data. The data for this project is based on data of Nigeria in the 2016 Census - https://data.world/ocha-nigeria/a7c3de5e-ff27-4746-99cd-05f2ad9b1066

The data is a compilation of the population of different age groups in the two states. Since the client is looking to start an IT training business, we know that certain age groups will be more likely to enroll in such training than others. The data consist of the population of various age groups in the different regions of each state. I would have preferred the data to be of cities, instead of regions as it would have been narrowed down a little more. But it was quite difficult to get data with such specifications. I will be analyzing the size of the preferred age groups in each region, and then comparing whichever regions have a higher population in both states.

Meaning, based on our analysis, I will select one region from each state and then compare both selected regions. From the selected regions of both states, I will be comparing the number of secondary schools (High school) each has. The list of schools will be gotten using Foursquare's venue feature as using Foursquare is part of the requirement of this project.

OYO DATA

State regions	Age 4 to 5 Pop	Age 6 to 13 Pop	Age 14 to 17 Pop	Age 18 to 22 Pop	Age 23+ Pop	Total Pop	Latitude	Longitude
Oyo North	77,552	558,224	358,344	335,667	1,329,787	2,659,574	8.525903	3.616589
Oyo South	172,518	265,887	345,668	624,567	1,000,654	2,567,254	7.374447	3.271742
Oyo Central	200,584	156,716	338,547	613,453	1,257,954	2,409,294	7.968394	3.571628

RIVER DATA

State regions	Age 4 to 5 Pop	Age 6 to 13 Pop	Age 14 to 17 Pop	Age 18 to 22 Pop	Age 23+ Population	Tota Pop	Latitude	Longitude
Rivers East	215,760	225,884	335,486	785,774	1,228,554	2,791,458	4.889828	7.107192
Rivers South East	205,901	265,524	300,258	424,125	602,655	1,798,463	4.645197	7.441428
Rivers West	356,254	385,263	412,549	658,416	679,009	2,491,491	4.831936	6.592625

As the project progressed, I realized that foursquare does not have data on my locations of choice. So I switched to use locations we had recently used during the course. I am using New York state and Ontario Province. New york's population data that I used was gotten from https://www.newyork-demographics.com/cities_by_population and Ontario's population data was gotten from https://www.citypopulation.de/en/canada/cities/ontario/. I was also able to get data on the list of High schools in this location using Foursquare.

METHODOLOGY

In this project K-Means Clustering and Groupby were used to analyze the data. Based on the table of data gathered, a city with the highest population was selected from each state. Then the data of schools in those cities were gathered using Foursquare. A cluster of schools was then created by postal code.

Oyo North region has the highest number of population of age 18+ for Oyo state, while Rivers East region has the highest population for the same age range for Rivers state. But unfortunately, I could not proceed with analyzing these locations as Foursquare could not provide data on them and using Foursquare is part of the requirement of this project. So, for the purpose of this project I changed my location to New York State and Ontario province. Analyzing this to show what could be done if we could get foursquare data on Nigeria.

According to this website, https://www.newyork-demographics.com/cities_by_population, New York City is the city in New York state with the highest population.

Rank	City	Population
1	New York	8,336,817
2	Hempstead	766,980
3	Brookhaven	480,763
4	Islip	329,610
5	Oyster Bay	298,391

And according to this website, https://www.citypopulation.de/en/canada/cities/ontario/ Toronto is the city in Ontario Province with the highest population.

Major Cities

	Name	Population Estimate (E) 2019-07-01
1	Toronto	2,965,713
2	Ottawa	1,028,514
3	Mississauga	769,050
4	Brampton	696,975
5	Hamilton	574,263

By using groupby and K-means, I was able to figure out which postal codes in the two selected cities has the highest number of high schools in its cluster. Whichever postal code has more schools in it cluster, will be chosen. And this will automatically determine which of the two cities will be selected for the location of the IT technical schools.

The idea behind the population and high school cluster is that, the city with a higher population provides the possibility of a greater population reach by this training school, and the more high schools within a particular cluster, the more reach to people graduating high school and looking to go to college, a technical school or start a career. Anyone looking to do any of these 3 things mentioned will be a good target market. This may also be a good opportunity that they did not know they needed.

RESULT

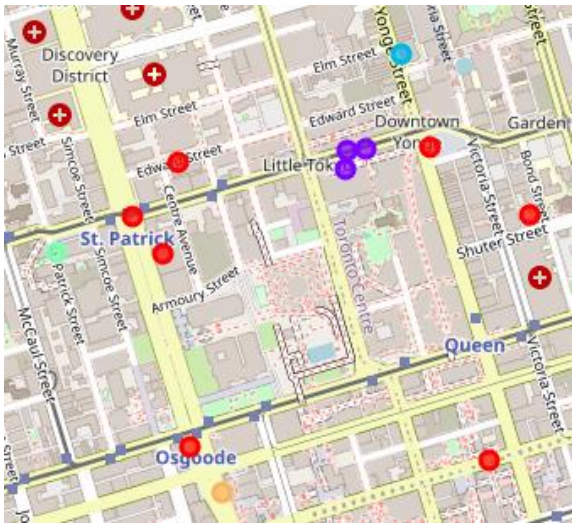
Groupby and K-means of Toronto

From the analysis done we can see that the zip code M5G 2C5 has more high schools than the other zip codes. In the cluster map, the purple dots are clusters of schools in this zip code.

```
Torontodataframe_filtered.groupby('postalCode').count()
```

]:

postalCode	name	categories	city	lat	lng
L3R 0M1	1	1	1	1	1
M2J 1Y3	1	1	1	1	1
M4P 1A9	1	1	1	1	1
M5B 2B9	1	1	1	1	1
M5B 2K1	1	1	1	1	1
M5G 1G6	1	1	1	1	1
M5G 2C3	1	1	1	1	1
M5G 2C5	2	2	2	2	2
M5H 3E5	1	1	1	1	1
M5M 3G5	1	1	1	1	1
M5T 2X7	1	1	1	1	1
N1H 7V2	1	1	1	1	1



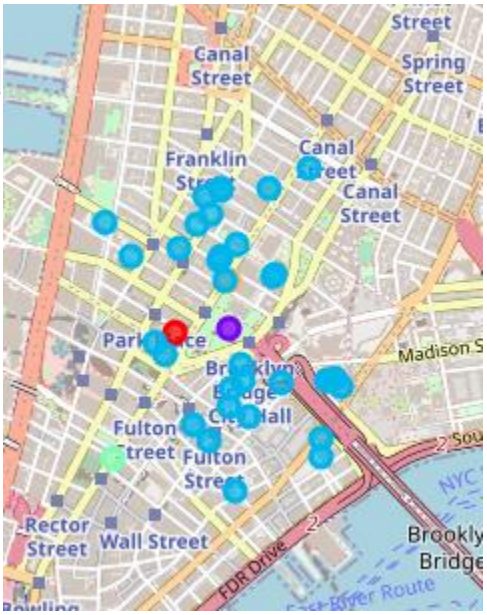
Groupby and K-means of New York City

From the analysis done we can see that the zip code 10038 has more high schools than the other zip codes. And following close to it is 10007.

```
NYdataframe_filtered.groupby('postalCode').count()
```

]:

postalCode	name	categories	city	lat	lng	state
10007	10	10	10	10	10	10
10013	8	8	8	8	8	8
10038	15	15	15	15	15	15
10271	1	1	1	1	1	1
10279	1	1	1	1	1	1
10452	1	1	1	1	1	1
11201	1	1	1	1	1	1



DISCUSSION ON OBSERVATIONS AND RECOMMENDATION

The more High schools in a location means the more people who will soon be in the market for professional education; college, technical school etc. Based on the combination of highest population and k-mean, I observe that zip code 10038 in New York has more cluster of High schools among the selected cities. Also, New York city's population is way more that Toronto's population. From this analysis, New York city will be the best location to start this IT Training business. Specifically, the neighborhoods with the zip code 10038. 10038 has the highest cluster of high schools among the zip codes in New York city and Toronto. And, although not all the neighborhoods with this zip code are in close proximity to each other, a large percent of them are.

Another observation made was this; the cycle process from data collection and data analysis that the course talked about, is very important as it teaches one to be very flexible with the process. It helped me throughout the process of this project as I had to go back and forth and make several adjustments till, I got what was needed. It also reflected in how I had to adjust the locations used in the project.

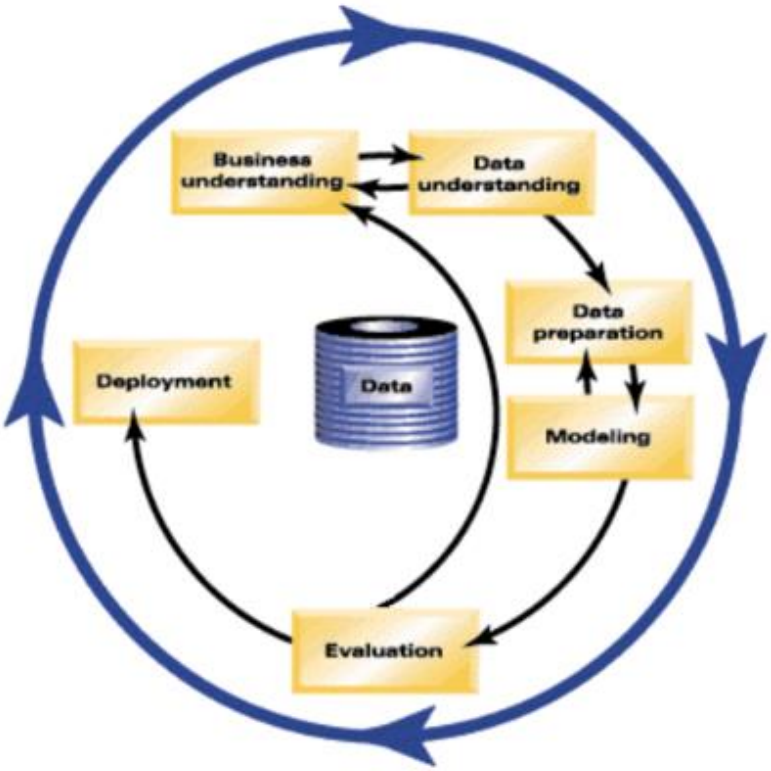


Fig.1 CRISP-DM model, IBM Knowledge Center, CRISP-DM Help Overview

The image is a screenshot from the Data Science Methodology course.

CONCLUSION

From everything gathered, I would highly recommend starting an IT training business in New York city. The business will have a better potential at succeeding in this location. If I was able to gather data from foursquare for the locations I had originally chosen for the project, this is the same process I would have followed to arrive at a location recommendation. This shows that there is a huge opportunity for data collection in Nigeria.