DATA ANALYSIS ON FAST FOOD RESTAURANT BUSINESS

CAPSTONE PROJECT - IBM DATA SCIENCE PROFESSIONAL CERTIFICATE SPECIALIZATION

BY

MUNMUN MANNA

INDEX

INTRODUCTION

This project focuses on fast food restaurant chain business in New York city. It looks for answers to the following questions –

To maximize the profit: -

- what kind of customers base should the business focus on?

For example - should the business target high income group, young aged group or high density area?

- What should be the price tier for the restaurant?

Should it be uniform across places or differentiated based on some criterion?

DATA SECTION

4 stages

1) The zip codes are collected for New York county in the New York state from the data source. The data contains all the zip codes of USA along with latitude and longitude geographical attributes.

2) The demographics are collected for the New York county zip codes. Information like median income, population, median age per zip codes are gathered here.

3) Using the Foursquare API, a list of fast food restaurants are found for every zipcode. The restaurants are uniquely identified by its Id. Since, regular calls to Foursquare are limited, these information are stored in a csv file named 'export_dataframe1.csv‘

4) Using the Foursquare API, all the details are collected for each fast food restaurant using unique restaurant Id. Information like No. of likes, Ratings and Price_Tier are received at this stage. Since, regular calls to Foursquare are limited, these information are stored in a csv file named 'export_dataframe2.csv'

DATA SECTION CONTD…

S.No.	Details	Source
1	Collection of zip codes for New York county in the New York state from the data source.	https://www.unitedstateszipcodes.org/ny/
2	Collection of demographic information per zip code.	https://zipwho.com/
3	Collection of list of fast food restaurants using **Foursquare API**.	https://api.foursquare.com/v2/venues/search? &client_id={}&client_secret={}&v={}&ll={},{} &radius={}&limit={}&categoryId={} CategoryId='4bf58dd8d48988d16e941735' is used for fast food restaurant as stated in the following link: https://developer.foursquare.com/docs/resources/categories
4	Collection of details of each fast food restaurant using **Foursquare API**.	https://api.foursquare.com/v2/venues/{}? &client_id={}&client_secret={}&v={}

Example

Considering the zip code - 10001. We found 10 restaurants in its surroundings using the Foursquare API as described in stage 3.

	zip	Latitude	Longitude	Name	Id	Venue Category
0	10001	40.748724	-74.003422	SUBWAY	4ecbe682f9f4a82587658269	Sandwich Place
1	10001	40.747569	-73.997088	McDonald's	4b1331eef964a5205c9523e3	Fast Food Restaurant
2	10001	40.752250	-74.005727	Enfes	59c143aa6cf01a3c2418005a	Turkish Restaurant
3	10001	40.748357	-74.004128	Subway	5a64e13cea1e445448e69273	Sandwich Place
4	10001	40.746259	-73.997676	Subway	4d6e842538363704f4d8a9d0	Sandwich Place
5	10001	40.751434	-73.993487	Subway	4aef3eb9f964a520e3d621e3	Sandwich Place
6	10001	40.749272	-73.995269	Taco Bell / Pizza Hut	4f18c538e4b09594f839d493	Fast Food Restaurant
7	10001	40.753343	-73.995942	Subway	4ddedbc3cc3f89c0826fa9c8	Sandwich Place
8	10001	40.754320	-73.998480	Subway Restaurant	4b64bfb7f964a52016cd2ae3	Sandwich Place
9	10001	40.745323	-73.997815	Boston Market	3fd66200f964a52065e91ee3	American Restaurant

There are 6 subways, 1 McDonald's, 1 Enfes, 1 Taco Bell/Pizza Hut, 1 Boston Market. We will focus on each Restaurant and receive details about it. For example - Let us focus on Restaurant with Id='4ecbe682f9f4a82587658269', **Foursquare API** provides the following (next slide) information

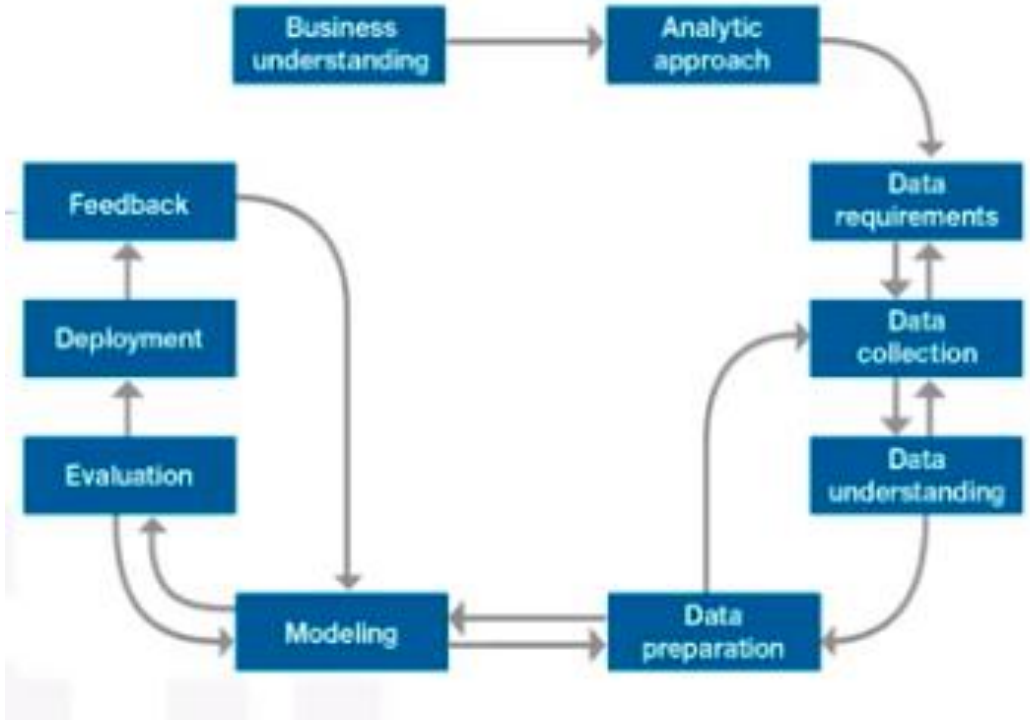
EXAMPLE CONTD…

	zip	Latitude	Longitude	Name	Id	Venue Category	Likes	Price_Tier	Ratings
0	10001	40.748724	-74.003422	SUBWAY	4ecbe682f9f4a82587658269	Sandwich Place	4	1	6.1

We can clearly see, there are 4 likes, Price_Tier is 1 (i.e. cheap) and ratings is 6.1

Hence, it is apparent that **Foursquare API** has played a vital role in data collection.

METHODOLOGY SECTION



Picture taken from IBM slide on Data Science Methodology

1. Business Understanding

To find the best customer base for a startup fast food restaurant chain business in New York.

It is important to do an analysis and start a company wisely by selecting **right factors and resources** so that the company starts making profit from day one and sustain for a long time.

2. Analytic Approach

This project will focus on areas of New York county and compare their demographics such as Annual Income, Population, Age with their fast food restaurant details such as ratings, count of likes, price tier.
It is needed to study under what ideal factors, the fast food restaurants are popular. For example - Should fast food be lowly or highly priced? Are fast food more popular among young, middle-aged or old people?
This project attempts to find answers to some of these questions.

3. Data Requirements

a) Firstly, the information that is needed are zipcodes of new york.

b) Secondly, demographics of zip codes of New York are in demand.

c) Thirdly, list of all fast food restaurants in New York will be collected and used.

d) Finally, details of all the fast food restaurants listed in the third step will be stored and saved.

For the 3rd and last step, **Foursquare API** plays a vital role in Data Collection process.

4. Data Collection

The data section (above – slide 4) clearly states which sources have been used to retrieve data listed in the data requirement section.

5. Data Understanding

- In the zipcode file, it is found that all the counties of New York are covered as the data was available as per a specific state. So it is filtered so that the project can focus on areas of New York county only.
- Also due to limitation in data availability, the project focuses on those zipcodes for which demographic information are found. This information contains median income and median age and not the means. Thus, we can rely more on the data because considering the averages, it is possible that the figures are exploited by the outliers.
- Using each zipcode, list of restaurants is generated using **Foursquare API**. Each restaurant is represented uniquely by an Id.
- Using these Ids, **Foursquare API** generates details of each and every restaurant. Since with free account with **Foursquare**, limited regular and premium calls can be made per day, nearly a week was spent in data collection process. Therefore, time consumption is a heavy cost for this process.

6. Data Preparation

- At many stages, filtrations of data were done to highlight the relevant factors.
  For example, zipcodes of the new york state were narrowed to zipcodes of the new york county.
- These were further filtered to zipcodes for which demographic information exist in the data collection stage. This was done using dataframes merging technique.
- Also, the dropping tool was used a few times to jump to the relevant information.
  For example, demographic information also contained extra information like education, cost of living index, male to female ratio, etc. for each zip code. These columns were dropped to focus on income, age and population.
- For certain restaurants, **Foursquare** misses information regarding ratings or price tier. While price tier varies from 1 (cheap) to 4 (costly), ratings range from 1(poor) to 10(very good). As part of data cleaning, the records with missing information were dropped so that the results are apparent.

7. Modelling

- Machine learning method K Means Clustering is done to cluster the restaurants in clusters based on the data collected. No. of clusters was selected to be 5 and value of K was taken higher so that a state of convergence is reached.
- Simple scatter plots will show only a type of relationship (like linear or polynomial) between 2 parameters but a scatter plot among the clusters will provide a clear picture which cluster or group is driving the crowd to build a trend and hence will help in taking executive decisions.

8. Evaluation

All the questions are answered with details supported by the data science method. Please see the Results Section (Slide # 19) below for details.

9. Deployment

Since the data are taken from real life sources, the results are true to the picture. Thus the suggestions made in the Discussion Section (Slide #20 & 21) below should be followed to build the best environment for a profitable business venture.

10. Feedback

These results will be discussed with my client who wants to startup the business and is looking for suggestions. All these recommendations will be made. Looking forward for his feedback. This project can be extended to - analysis on type of fast food, analysis on popularity of fast food among teens, etc.

RESULTS SECTION

All results are shown in form of graphs in the following link:

https://github.com/munmunmanna/Coursera_Capstone/blob/master/Capstone_Project_Coding%23c.ipynb

DISCUSSION SECTION

Following observations are seen through the analysis:

1. High ratings cluster is witnessed in the 60000 to 70000 income bracket.
2. Highest ratings cluster is seen in medium density regions between 40000 to 60000.
   Density higher than 60000 has seen relatively lower to average ratings.
   Density between 50000 to 60000 population has seen ratings of 8 and above.
3. Fast food does not seem to be popular among high aged (age>40) people except for people in the age group of 46.
   People aged between 34 to 36 years have provided the highest ratings followed by people in the age group of 38 to 39.
   Middle and higher aged people may be health conscious or might be having other responsibilities like family expenses or medical expenses.
4. Fast food restaurants which serve food in the price tier of 1 (i.e. cheaper) has achieved the highest ratings ranging from 4 to 9.5.
   So price is an important choice criteria for fast food restaurants.

DISCUSSION SECTION CONTD…

5. Fast food is liked by three distinct category of people. One, Higher income and lower age which provide ratings between 4 to 8.
   Two, Higher income and middle aged which provide ratings between 4 to 6. And three lower income and younger age group which provide ratings between 4 to 6.
   There is also a small cluster of lower age and higher income which have provided lower ratings.
6. Low price tier restaurants are dense with any income and any age groups.
7. Higher ratings are received when there are lower age group within low price tier restaurants.
8. Low dense area with higher income group like fast food.
9. Low dense area with lower age group like fast food.

CONCLUSION

Fast food restaurant business environment is robust and changes dynamically to pull the crowd. To set up business in the USA's financial capital area, New York,

- one should open restaurants in those areas where there is **low density population but people are young and earn higher income**.
- Price is an important factor in business. For fast food, crowd is attracted to qualitative food at **lowest price (Tier 1)**.

THANK YOU!!