Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Pizza Shop in Bangalore, Karnataka

By: Varun Behera

October 2019

INTRODUCTION

Today's Youngsters in India are most fascinating to living the lifestyle of the people of the west and also their food style is changing the same direction. Nowadays **pizza** an Italian originating food item is wide spreading in various parts of India. Not only young people but people of every age group are also preferring to eat pizza as snacks in the IT city of India.

As the taste for pizza is increasing day by day and more percentage of population in this city are leaning their taste in having such food in every hour of the day. Home delivery is also available in many outlets but the northern and southern regions of Bangalore take the whole charge. Thus, there is scarcity of pizza outlets in other regions of the places of Bangalore.

As a result, there are many pizza outlets in the city of Bangalore and many more are being built. Opening pizza shops allows food investors to earn consistent income. Of course, as with any business decision, opening a new pizza outlet requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the pizza shop is one of the most important decisions that will determine whether the pizza shop will be a success or a failure.

BUSINESS PROBLEM

The objective of this capstone project is to analyze and select the best locations in the **city of Bangalore, Karnataka** to open a new **pizza shop**. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of Bangalore, Karnataka, if a **Pizza franchise** is looking to open a new pizza outlet, where would you recommend that they open it?

TARGET AUDIENCE OF THIS PROJECT

The project is particularly useful to the **pizza franchise or pizza brands** who want to open their new outlet in different parts of the city to cover the major sections of the city. By opening new outlets, they can also minimize the hustle and late deliveries, thus increasing efficiency and customer satisfaction.

Also, this project targeted to the **new entrepreneurs** who want to try their luck in the food word as this will give an ample amount of information of where to open an outlet in order to drag people towards themselves and complete with the big sharks with new offers and customer-oriented techniques.

DATA

To solve the problem, we will need the following data:

• List of **neighborhoods** in **Bangalore**. This defines the scope of this project which is confined to the city of Bangalore, the capital city of the **Karnataka** state of India.

• **Latitude** and **longitude** coordinates of those **neighborhoods**. This is required in order to plot the map and also to get the venue data.

• Venue data, particularly data related to **Pizza places**. We will use this data to perform **clustering** on the neighborhoods.

SOURCES OF DATA AND METHODS TO EXTRACT THEM

- The data is stored in a **Wikipedia page** having the link as below: https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Bangalore
- This contains a list of neighborhoods in Bangalore with a total of 8 **boroughs** and 88 **neighborhoods**. The **boroughs** are central, northern, southern, eastern, western and many more.
- We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python **requests** and **beautifulsoup** packages.
- Then we will get the **geographical coordinates** of the neighborhoods using Python **Geocoder package** which will give us the latitude and longitude coordinates of the neighborhoods.
- After that, we will use **Foursquare API** to get the venue data for those neighborhoods. **Foursquare** has one of the largest databases of 105+ million places and is used by over 125,000 developers.
- **Foursquare** API will provide many categories of the venue data, we are particularly interested in the **Pizza places** category in order to help us to solve the business problem put forward.
- This is a project that will make use of many **data science skills** such as **web scraping** (Wikipedia), **working with API** (Foursquare), **data cleaning**, **data wrangling**, **machine learning** (K-means clustering) and **map visualization** (Folium).

METHODOLOGY

- The data is stored in a **Wikipedia page** having the link as below: https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Bangalore
- This contains a list of neighborhoods in Bangalore with a total of 8 **boroughs** and 88 **neighborhoods**. The **boroughs** are central, northern, southern, eastern, western and many more.
- We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python **requests** and **beautifulsoup** packages.
- Then we will get the **geographical coordinates** of the neighborhoods using Python **Geocoder package** which will give us the latitude and longitude coordinates of the neighborhoods.
- After gathering the neighborhood and coordinates data, we will populate the data into a pandas DataFrame and then visualize the neighborhoods in a map using Folium package.
- This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Bangalore.
- Next, we will use Foursquare API to get the top 100 venues that are within a radius of 1000 meters.
- We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key.
- We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude.
- With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues.
- Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category.
- Then we are also preparing the data for use in clustering. As we are analyzing the “Pizza Places” data, we will filter the “Pizza Places” as venue category for all neighborhoods.
- Then, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm determines k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.
- We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for “Pizza Places”.
- The results will allow us to identify which neighborhoods have higher concentration of pizza outlets while which neighborhoods have fewer number of pizza outlets.

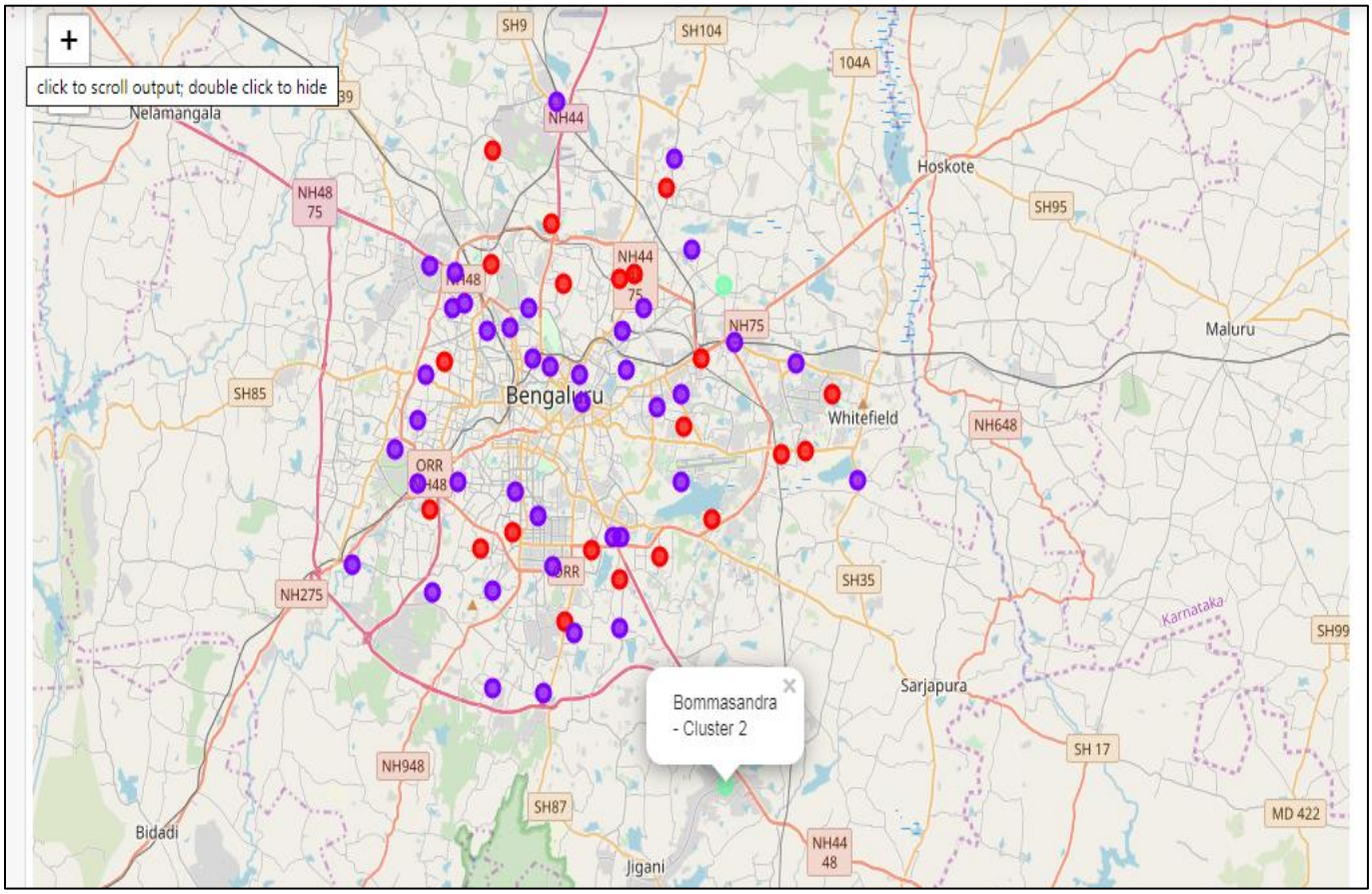
- Based on the occurrence of pizza outlets in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new pizza outlet.

RESULT

The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Pizza Places”:

- **Cluster 0**: Neighborhoods with high concentration of Pizza Places
- **Cluster 1**: Neighborhoods with low number to no existence of Pizza Places
- **Cluster 2**: Neighborhoods with moderate number of Pizza Places

The results of the clustering are visualized in the map below with cluster 0 in red color, cluster 1 in purple color, and cluster 2 in light blue color.



OBSERVATIONS

- Most of the pizza shops are concentrated in the Northern and Eastern area of Bangalore city, with the highest number in cluster 0 and moderate number in cluster 2.
- On the other hand, cluster 1 is having the lowest no of pizza places in the neighborhoods.
- This represents a great opportunity and high potential areas in opening new pizza palace as there is very little to no competition from existing pizza places.
- Meanwhile, pizza palace in cluster 0 are likely suffering from intense competition due to oversupply and high concentration of pizza shops.
- From another perspective, this also shows that the oversupply of pizza shops mostly happened in the northern, western and south-eastern areas of the city, with the central and southern area still have very few pizza shops.
- Therefore, this project recommends pizza shops investors to refer on these findings to open new outlets in neighborhoods in cluster 1 with little to no competition which is the central and southern parts of Bangalore.
- Pizza shops with unique selling propositions to stand out from the competition can also open new pizza shops in neighborhoods in cluster 2 with moderate competition.
- Lastly, pizza investors are advised to avoid neighborhoods in cluster 1 which already have high concentration of pizza shops and suffering from intense competition.

LIMITATIONS AND SUGGESTIONS FOR FUTURE RESEARCH

In this project, we only consider one factor i.e. frequency of occurrence of Pizza Places, there are other factors such as population and income of residents that could influence the location decision of a new Pizza Places.

But such data are not available to the neighborhood level required by this project. Future research could develop a methodology to estimate such data to be used in the clustering algorithm to determine the preferred locations to open a new Pizza Places.

In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results.

CONCLUSION

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. pizza brands and investors regarding the best locations to open a new Pizza Place.

To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in cluster 1 are the most preferred locations to open a new Pizza outlet. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new Pizza outlet.

REFERENCES

- Data of neighborhoods in Bangalore:
    - https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Bangalore
- Foursquare developers: https://developer.foursquare.com/docs
- https://www.webstaurantstore.com/article/42/how-to-start-a-pizzeria.html