IBM Capstone Project

Using data science to help identify the best location for a new Italian Restaurant in Sydney

Introduction

For this capstone project, we are going to look at the case study of an Italian restaurateur. Italian cuisine is one of the most popular cuisines in the world and for a prospective restaurateur looking to open a new Italian restaurant, it is important to identify the right location. Location is single handedly the most important factor for a restaurant that can affect its success

Business Problem

The objective of this capstone project is to explore the city of Sydney, Australia and to find the most appopriate location for opening a new italian restaurant. We will be using a variety of data science analytical and visualization methods such as clustering. We will aim to answer the following question: what is the best location to open an Italian restaurant in Sydney?

Target Audience

The target audience is any entrepreneur who is interested in opening a new italian restaurant in the city of Sydney. This approach can be replicated in any city in the world and modified for any other cuisine restaurant easily

Data

To solve this problem, we need the following Data:

- List of neighborhoods in Sydney, Australia
- Latitude and Longitude coordinates of these neighborhoods
- Venue data related to Italian restaurants. This will help us identify the neighborhoods suitable to open an Italian restaurant

Extracting the Data

- Scraping the wikipedia page to scrape data
- Getting latitude and longitude data of these neighborhoods via geocoder package
- Using foursquare API to get venue data related to these neighborhoods

Methodology

First we get the list of neighborhoods in Sydney, Australia. This is possible by parsing the wikipedia page using the Beautiful Soup library at the link

https://en.wikipedia.org/wiki/List_of_Sydney_suburbs

Using the appropriate CSS class ids and other cleaning techniques, we extract the neighborhood names and then run each neighborhood names via the geocoder package to get the individual lat/long co-ordinates and store them in a pandas dataframe

We then use the foursquare api to pull the list of top 100 venues within a 500 meter radius. We need to create a foursquare developer account to get access to this. We then get the names , categories, latitude and longitude of the venues. We identify the unique categories from this list. We analyze each neighborhood by grouping the rows per neighborhood and taking the mean on the frequency of occurrence for each venue category. This helps us for clustering

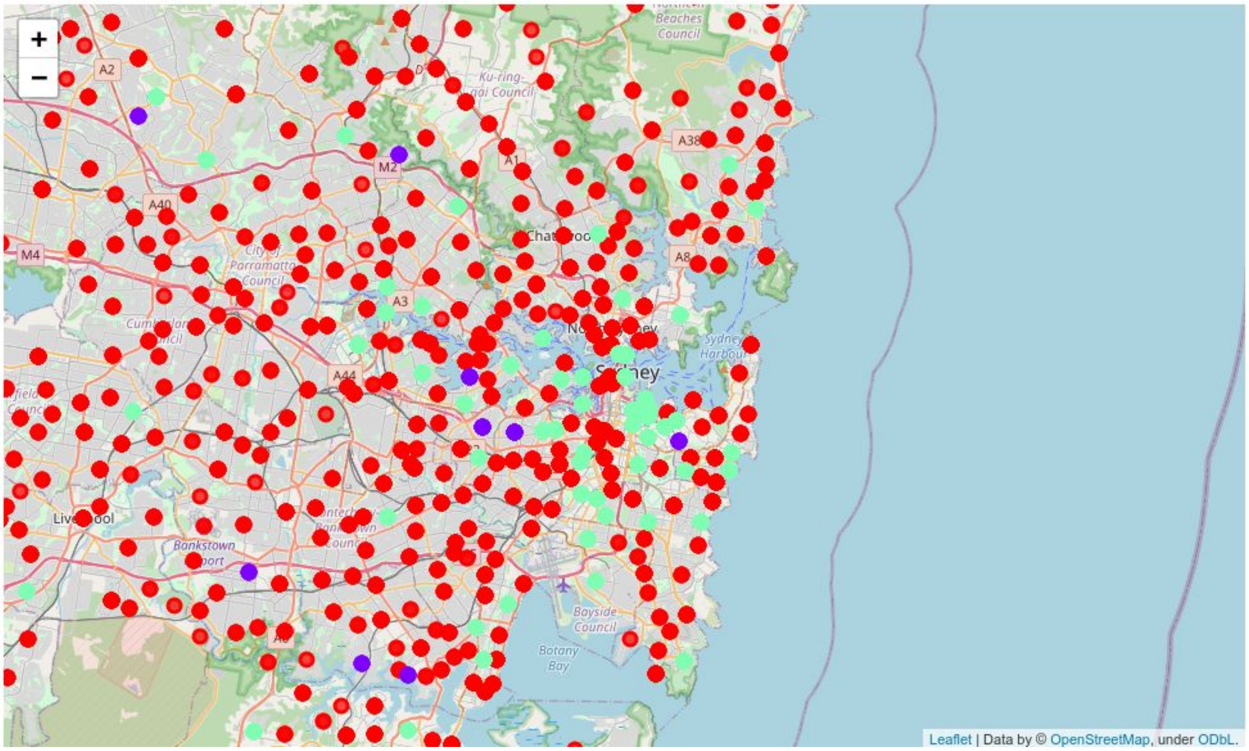
We primarily look at 'Italian restaurants', because it helps us understanding where out competition is.

We then perform the clustering method using k-means clustering. K-means clustering identifies the k number of centroids and allocates every data point to the nearest clustering, keeping the centroids as small as possible. We have clustered the neighborhoods of Sydney into 3 clusters based on the occurrence of 'Italian restaurants'

Based on the resultant output, we can analyze and recommend the ideal locations for the restaurant.

Results

The resultant clusters



The results from the k-means clustering show that we can categorize the Sydney neighborhoods into 3 clusters based on how many italian restaurants are there in each neighorhood

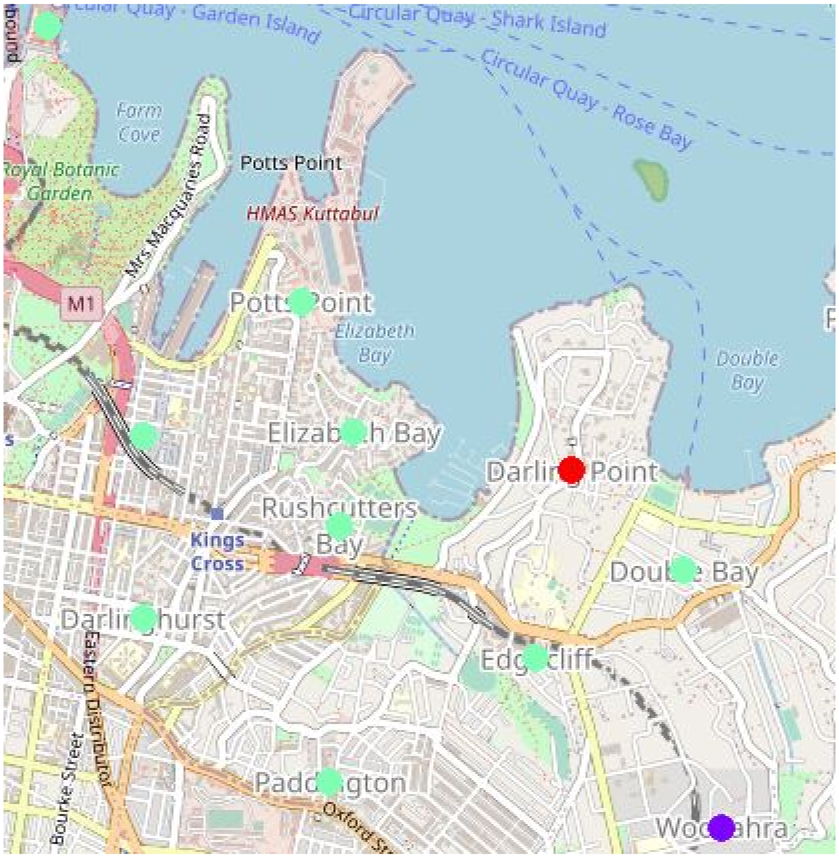
Red cluster / Cluster 0 : those with zero or very few Italian restaurants

Purple Cluster/ Cluster 1: neighborhoods with few italian restaurants

Green Cluster/ Cluster 2 : neighborhoods with high level of Italian restaurants

Recommendations

There are two approaches for setting up new restaurants, which depends on the business model of the restaurant. There are those who prefer opening up in areas with established market and compete with the competitor on basis of price/quality. For these type we would recommend areas around Sydney CBD which have high density of green dots such as Rushcutter’s bay and between Darlinghurst and Double Bay



The other approach is to find areas that are less served with Italian restaurants and compete for the additional market. In that case, the best strategy would be to find any purple dot cluster in the map and set up a restaurant there.

Limitations and Future Research

In this project, we are only considering the existence of Italian restaurants as the sole factor in determining where to open a new restaurant. The other factors that can be looked at are zoning classification (business vs residential), ethnic demographics of the customers, and income levels. It would be out of the scope of this capstone project as that would be a very big and intensive project.

Conclusion

In this project we have provided data driven recommendations to the stake holder by identifying a business problem, acquiring data and then preparing it and using machine learning models and visualization techniques using k-means clustering.

References

- List of neighborhoods in Sydney
  https://en.wikipedia.org/wiki/List_of_Sydney_suburbs
- Foursquare api documentation for places
  https://developer.foursquare.com/docs/places-api

- The code for this project can be found here

  https://github.com/ajayfsm/Coursera_Capstone/blob/master/sydneysoup.ipynb