RESTAURANT RECOMMENDER SYSTEM

SYDNEY

IBM Capstone Project

Table of Content

Introduction:

Problem Background:

Sydney is the state capital of New South Wales and the most populous city in Australia. After World War II , it experienced mass migration and become one of the most multicultural cities in the world. Sydney has a very diverse population with people migrating from different countries such as Britain, Greek, Lebanon, India , China, Korea etc. As people migrated, they carried their cuisine and culture with them to Sydney which resulted in giving this city a huge variety of cuisine to choose from.

Problem Description:

Food is an integral part of any culture, when people migrate they carry their food and culture with them. Living in such multicultural place restaurant recommendation seems quite helpful. In such scenarios, we need to find the right place, at reasonable cost, to serve us the best possible way. So there are few questions that must be addressed, such as :

1. How many types of foods are available in the restaurant ?
2. Which is the most nearest to me with good rating ?
3. How many "similar" restaurants are available near by me ?
4. Do the "similar" restaurants cost more ? if so, what speciality do that have ?

To address such question, XXYZ company's manager decides to allocate this project to me not just to find out solutions to the questions but also build a system that can help in recommending new places based on their rankings compared to the previously visited by me. Expectations from this recommender system is to get answer for the questions, and in such a way that it uncovers all the perspective of managing recommendations. It is sighted to show :

1. What types of restaurants are present in a particular area ?
2. where are the similar restaurant present based on a preference to particular food ?
3. How do different restaurants rank with respect to my preferences ?

Target Audience:

Target audiences for this project does not limit to a person who loves to try food from their own place but everyone. People could simply decide to look for a similar restaurant all the time because they are addicted to a specific category of food. People who rarely use restaurants would prefer to have the most rated restaurants nearby them and all this could be easily handed by our recommender system. So target for this project is basically everyone who is exploring different places or similar places.

Success Rate:

With restaurants evolving, new food categories emerge, hybrid food starts to become more popular, we need a system that could help us access vast number of food varieties. It is impossible for a person to ask each and everyone about their visit to a particular place and also not everyone remembers everything. On the other hand, Computers are good at remembering things, and with Machine learning to its peak, it high time technology will be our personal guide and help us personally based on our likes and dislikes. So people would care about this project as their personal assistance and success rate could certainly increase with time.

Data:

Data Requirement:

To find a solution to the questions and build a recommender model, we need data and lots of data. Data can answer question which are unimaginable and non answerable by humans because humans do not have the tendency to analyze such large dataset and produce analytics to find a solutions.
Let's consider the base scenario :
Suppose i want to find a restaurant, then logically, i need 3 things :

1)Its geographical coordinates(latitude and longitude) to find our where exactly it is located.

2)Population of the neighborhood where the restaurant is located.

3)Average income of neighborhood to know how much is the restaurant worth.

Lets take a closer look at each of these :
1)To access location of a restaurant, its Latitude and Longitude is to be known so that we can point at its coordinates and create a map displaying all the restaurants with its labels respectively.

2)Population of a neighborhood is very important factor in determining a restaurant's growth and amount of customers who turn up to eat. Logically, the more the population of a neighborhood, the more people will be interested to walk openly into a restaurant and less the population, less number of people frequently visit a restaurant. Also if more people visit, better the restaurant is rated because it is accessed by different people with different taste. Hence is is very important factor.

3)Income of a neighborhood is also very important factor as population was. Income is directly proportional to richness of a neighborhood. If people in a neighborhood earns more than an average income, then it is very much possible

that they will spend more however not always true with very less probability. So a restaurant assessment is proportional to income of a neighborhood.

Data Collection:

1) Collecting geographical coordinates is not difficult but after googling for few days, it was not available on open source data websites such as wikipedia, census report websites etc. Hence I created my own dataset, where I selected all the suburbs in Sydney city and googled their latitude and longitude. Here is the link of the website from where I got the suburbs data : https://www.cityofsydney.nsw.gov.au/learn/research-and-statistics/the-city-at-a-glance

	Borough	Neighborhoods	Latitude	Longitude	Population	City	AverageIncome
0	South	Alexandria	-33.9080	151.1903	8262	Sydney	2421
1	West	Annandale	-33.8814	151.1707	9451	Sydney	2337
2	North	Barangaroo	-33.8638	151.2022	189	Sydney	3661
3	West	Beaconsfield	-33.9110	151.1999	987	Sydney	2330
4	West	Camperdown	-33.8862	151.1791	10341	Sydney	1796

2) Population by neighborhood is again easy to find out given that its readily available. This link https://www.cityofsydney.nsw.gov.au/learn/research-and-statistics/the-city-at-a-glance has population data for all the suburbs in Sydney. The data frame for Sydney neighborhood population looks like:

:

	Borough	Neighborhoods	Population
0	South	Alexandria	8262
1	West	Annandale	9451
2	North	Barangaroo	189
3	West	Beaconsfield	987
4	West	Camperdown	10341

3) Income by neighborhood is again easy to find out given that its readily available. This link https://www.cityofsydney.nsw.gov.au/learn/research-and-statistics/the-city-at-a-glance has income data for all the suburbs in Sydney.This is the average income per week for the households in each of the suburb. The data frame for Sydney neighborhood population looks like:

	Borough	Neighborhoods	AverageIncome
0	South	Alexandria	2421
1	West	Annandale	2337
2	North	Barangaroo	3661
3	West	Beaconsfield	2330
4	West	Camperdown	1796

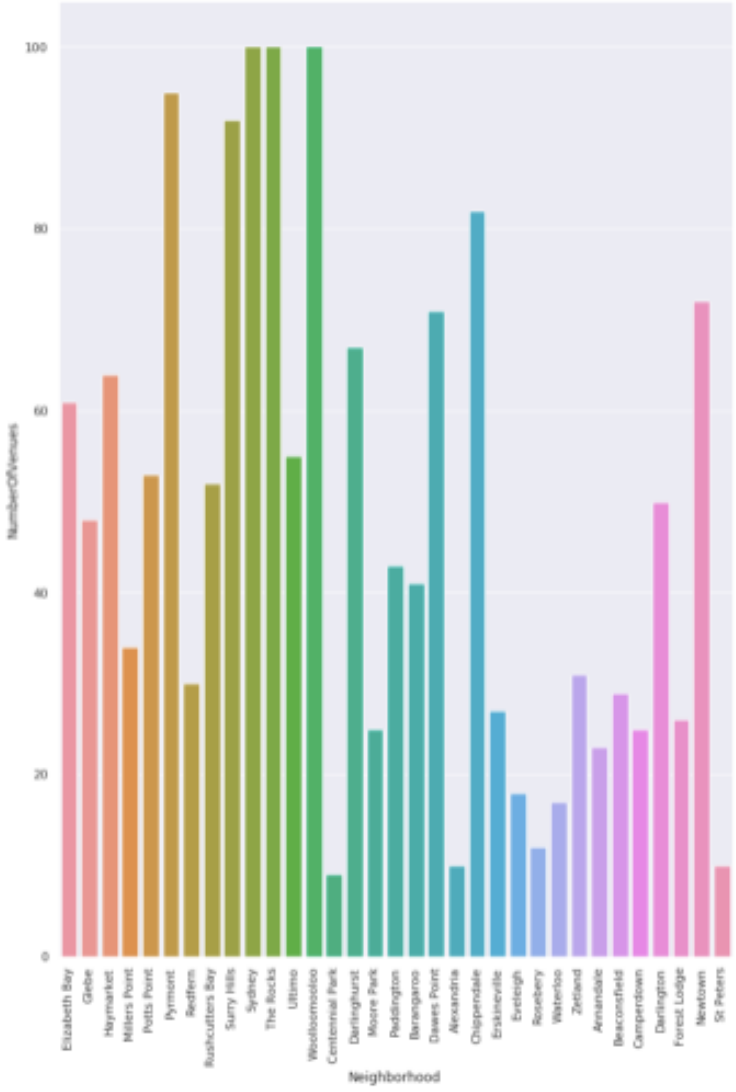
4) FourSquare API : Use of foursquare is focused to fetch nearest venue locations so that we can use them to form a cluster. Foursquare api leverages the power of finding nearest venues in a radius(in my case : 500 mts) and also corresponding coordinates, venue location and names. After calling, the following data frame is created:

	Neighborhood	Borough	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Alexandria	South	-33.908	151.1903	Pino's Vino e Cucina	-33.905480	151.190950	Italian Restaurant
1	Alexandria	South	-33.908	151.1903	La Cachette	-33.904849	151.189727	Café
2	Alexandria	South	-33.908	151.1903	Blackbird & Co	-33.906612	151.187861	Café
3	Alexandria	South	-33.908	151.1903	The Copper Mill	-33.906068	151.188426	Café
4	Alexandria	South	-33.908	151.1903	The Grounds of Alexandria	-33.910774	151.194406	Café

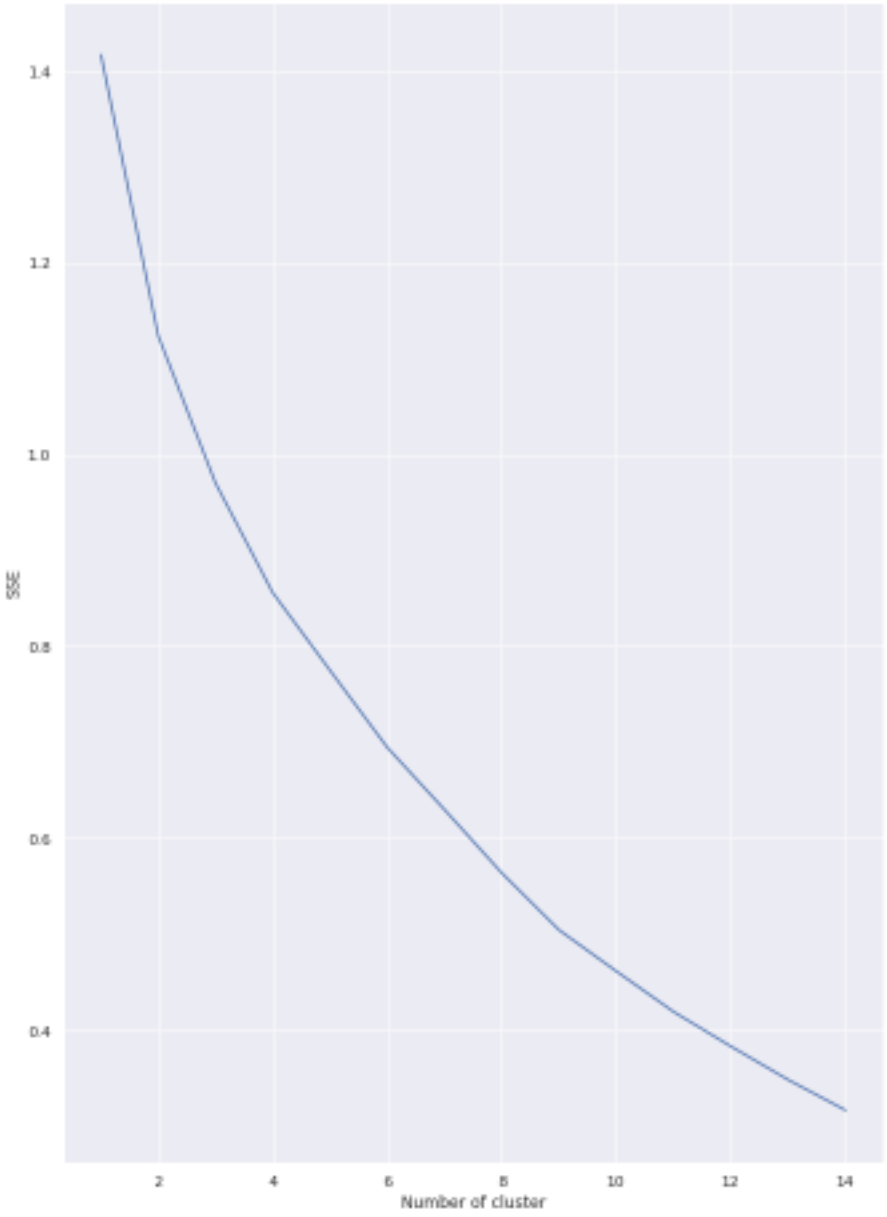
Methodology:

Exploratory Analysis:

Scrapping the data from different sources and then combining it to form a singleton dataset is a difficult task. To do so, we need to explore the current state of dataset and then list up all the features needed to be fetched.
Exploring the dataset is important because it gives you initial insights and may help you to get partial idea of the answers that you are looking to find out from the data.
While exploring the dataset, I found out that Surry hills, Sydney & Woolloomooloo has most number of venues while Centennial Park & St. Peters has the least.



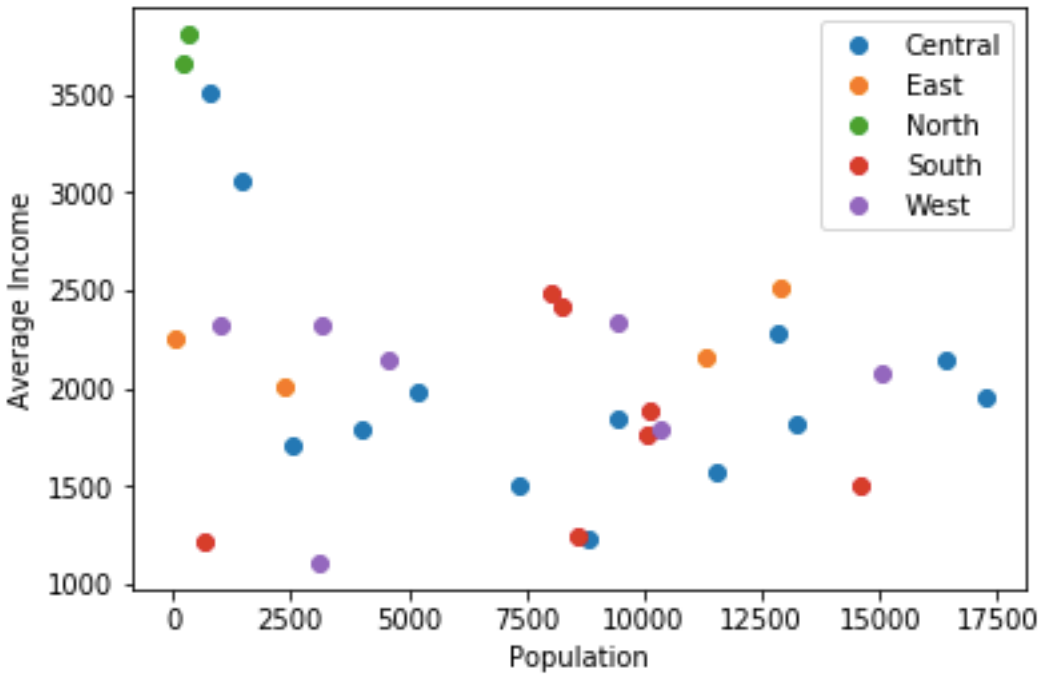
Also while producing graph for number of cluster, I produced a graph to explore all the values for n_clusters and then finding the best by exploring the elbow graph.



Inferential Analysis:

Most important factors while building the recommender system were population and income. They are the most import factor because they have a nonlinear relationship according to our dataset.

It needed to make some inferential analysis to understand this nonlinear relationship. As the amount of population increases, it does not necessarily mean that average income of a neighborhood will also increase. It is true to most of the case but also many cases differ to follow this trend. Similarly, a neighborhood with less number of people may not necessarily have less average income. It is possible to have less number of people and more income and vice versa. This can be inferred from the following graph:



Results:

The result of the recommender system is that it produces a list of top restaurants and the most common venue item that the user can enjoy. During the runtime of the model, a simulation was done by taking 'The Rocks' as the neighborhood and then processed through our model so that it could recommend neighborhoods with similar characters as that of 'The Rocks'.

The following image shows the result:

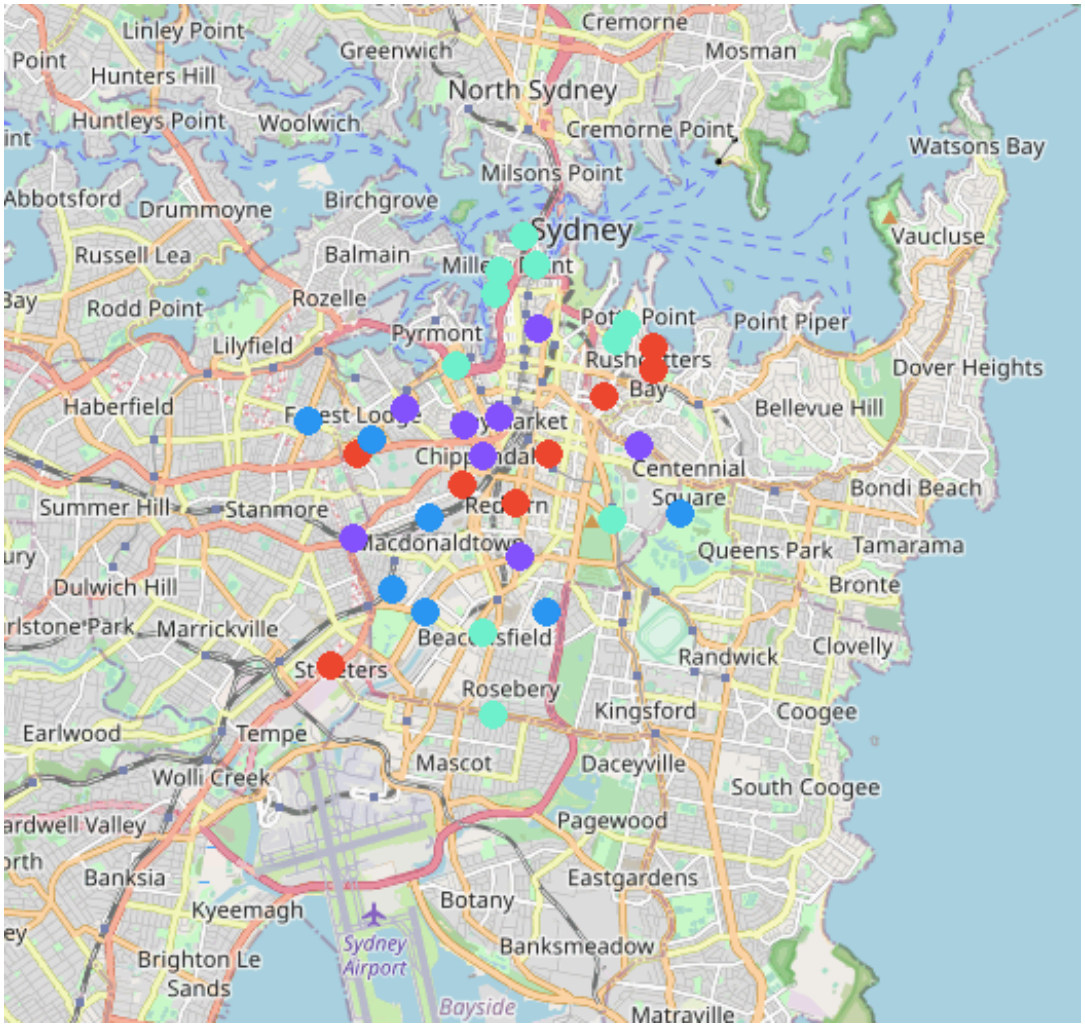
Out[103]:

	Neighborhoods	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	Ranking
0	Barangaroo	Venue Category_Café	Venue Category_Hotel	Venue Category_Bar	[0.4421434166390603]
1	Beaconsfield	Venue Category_Café	Venue Category_Coffee Shop	Venue Category_Furniture / Home Store	[0.3428723260123892]
2	Dawes Point	Venue Category_Café	Venue Category_Australian Restaurant	Venue Category_Pub	[0.46034662647808944]

Discussion:

Since there was a nonlinear relationship between income and population, it can be concluded that we must always perform inferential approach to find relationship among different set of features. Also during clustering, similar neighborhoods must be dumped into the right cluster.

The following graph shows the clusters:



Another observation that we can make is that choosing number of clustering could produce very diverse results. Some may be over fitted or some may be under fitted. Hence analysis of number of clusters must be done. Ref elbow_graph in the Methodology section.

Conclusion:

The recommender system is a system that considers factors such as population, income and makes use of Foursquare API to determine nearby venues. It is a powerful data driven model whose efficiency may decrease with more data but accuracy will increase. It will help users to finish their hunger by providing the best recommendation to fulfill all their needs.