IBM Data Science Professional Certificate

Capstone Project-Nima Afshari

Sydney Business Locator

1. Introduction

1.1. Background

The main issue in starting a new business in food or retail or some other industry that needs to be in touch with customers is finding a suitable location. Traditionally, people used to make up their minds based on words of mouth of others no matter of how reliable they were. Moreover, since national and global businesses as well as franchises can't rely on that method to make this underlying decision, the demand for a modern method seems obvious. Data Science as the warrior for unknown and vague issues all around the science come to the field to propose a solution for making a change in the decision procedure.

1.2. Business Problem

The main goal of this project is to locate the best possible place for a particular type of business like restaurant (it can be easily generalized to other types). The target audience of this project could be various sized businesses with a broad scope of objectives planning to open a new center in one of the LGAs(Local Government Areas) of Sydney. For the purpose of generalization, top-level features will be analyzed for all LGAs of Australia. The result will be narrow down to NSW with some extra added features to make the decision more precise. The approach is based on a wide variety of features including population density, income of population, number of currently available businesses in that LGA at the related field which is provided through the foursquare web services and many others.

2. Data Acquisition and cleaning

2.1. Data Sources

The data is acquired from three legitimate sources for covering separate dimensions of problem. For the first part of analysis a comprehensive dataset of demography, income and population density is needed. Census data from Australian Bureau of Statistics is the most reliable resource in this regard. These data after collecting and cleaning will be used for clustering LGAs all around australia based on those desired features.

Another source of data is used to more accurately indicate how dense the business of interest is in each LGA. This part is collected through Foursquare API web services for more specific and limited places.

For taking into account prospect future customers, a wide variety of different datasets could be used based on what the biz is about to provide. For instance, if it's a restaurant that cooks a certain kind of food for a specific appetite like Chinese food, it should check where they have dwelled. To make it more generalized, I've used the population projection up to 2036 for all of LGAs of Sydney provided by an NSW Government Planning and Environment. That would be the most reliable prediction data available for how the population is changing according to the demography of LGAs.

Fusion of these three data sources will let an analysis that took into account all concentration, income, density of prospect customers, and time dimensions. The result could be personalized for each business case study.

2.2. Data Cleaning

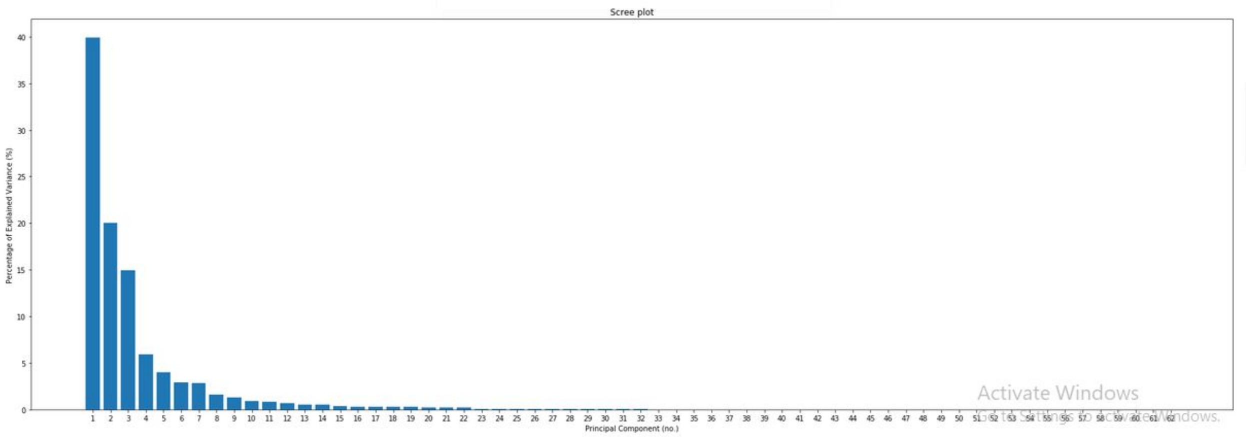
Census data from Australian Bureau of Statistics is a very large dataset and a selection between columns is needed for desired features. Each part of it is accessible through some .xls files including Population_Density, Economy_Industry, Income, Family_Community, and Land_Environment. Moreover, missing values that were found in the data is dropped due to small size of populations of LGAs.

The provided data from NSW Government Planning and Environment is available as an .xlsx file for download from their site. The downloaded file includes different sheets and the desirable one is addressed using sheet_name argument in the read_excel function. The sheet also has extra cells with irrelevant information that has cropped using header, usecols, and skipfooter. For missing data that were exists due to importing, elimination used because there were no valuable data in it. The result is a neat pandas dataframe of population projection.

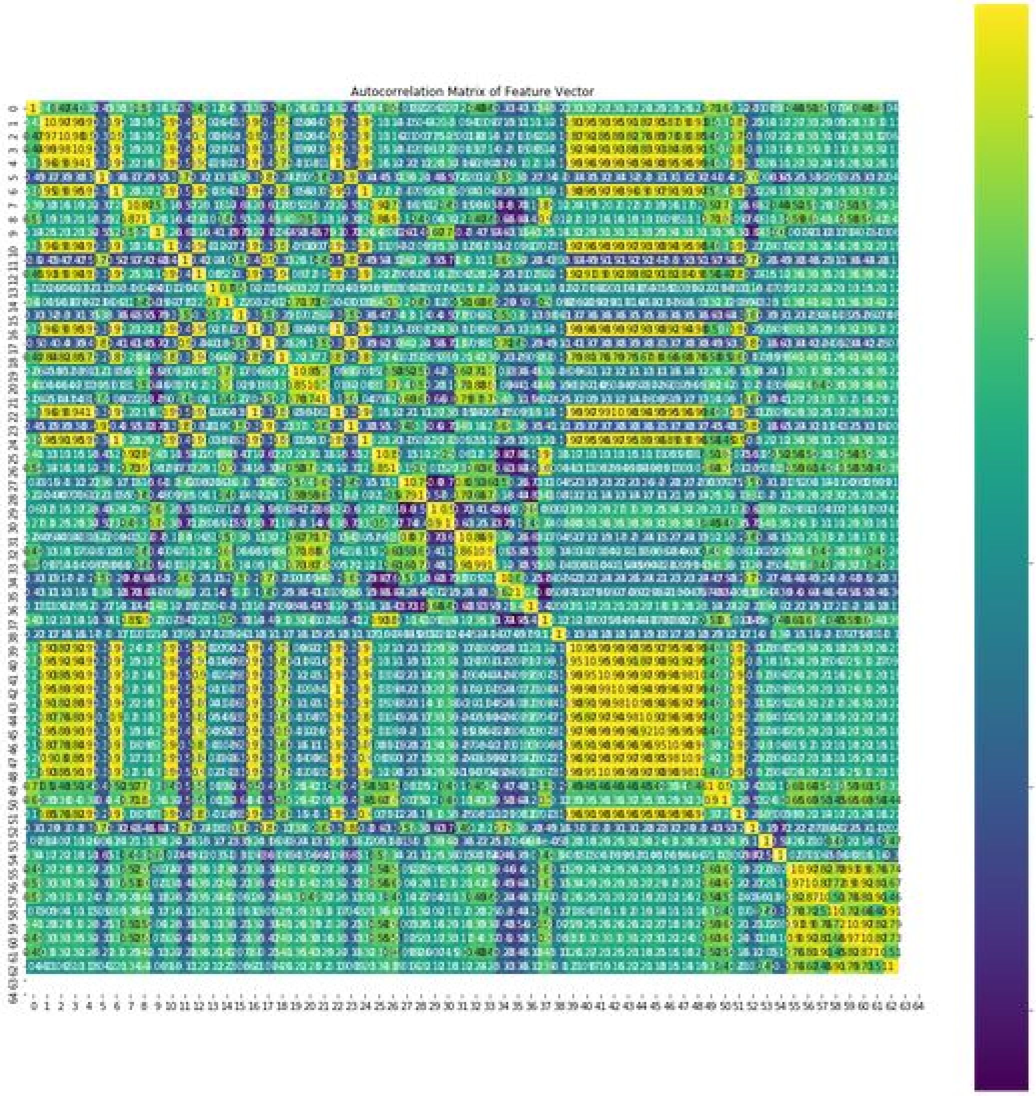
The RESTful API of foursquare location data provides the data as a JSON file. For further analysis it should be converted to pandas dataframe. It’s accomplished using each LGA and corresponding latitude and longitude of each one that is provided after calling the API.

2.3. Dimensionality Reduction

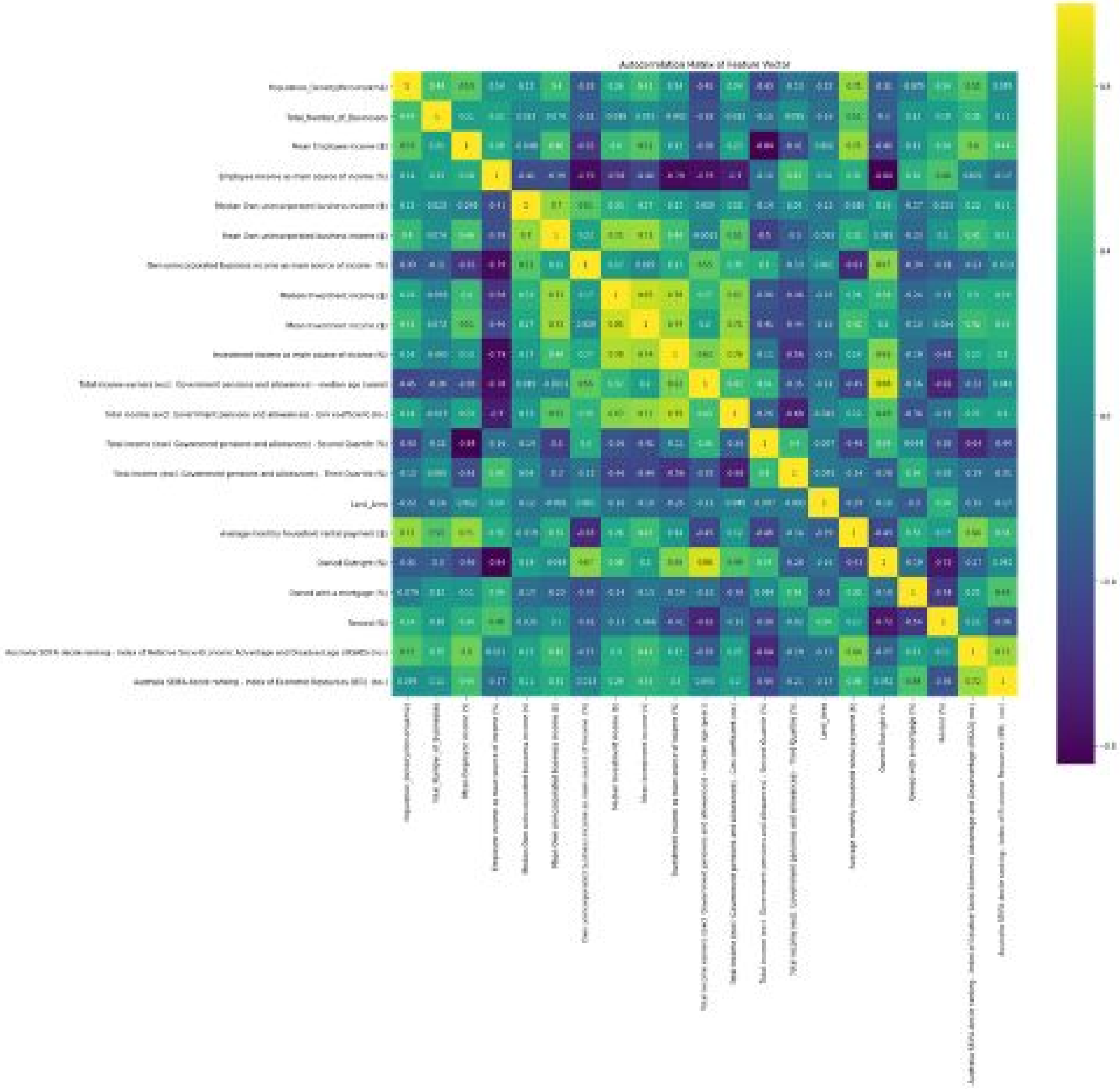
Feature selection from raw data was totally based on intuition and there is an obvious need for eliminating redundant features. For this purpose, different methodologies can be used. At first, I try to reduce the dimension of dataframe using PCA. Here is the Scree Plot of percentages of each PCA Component.

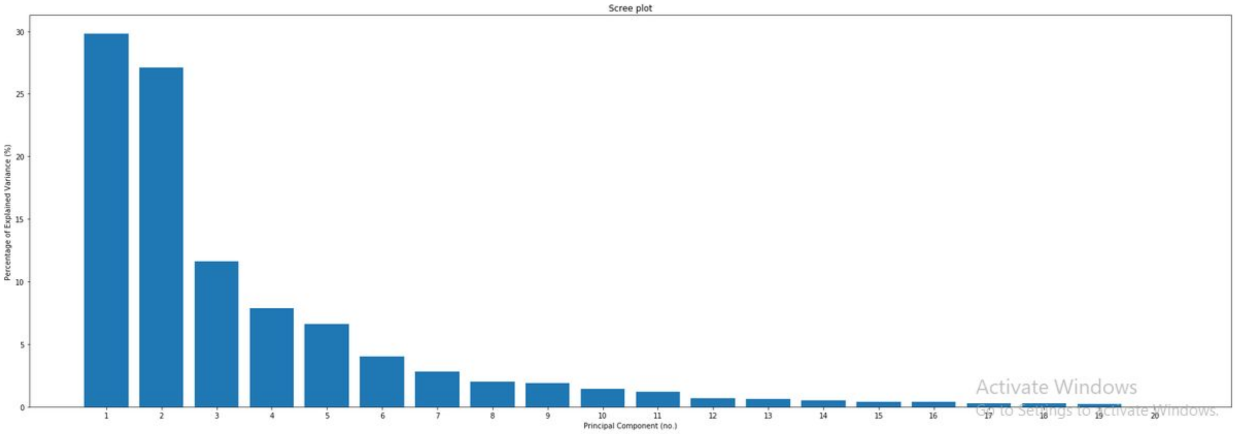


As it can be seen from above graph almost half the features could safely be removed because of redundancy. However, for the sake of meaningfulness of features, I decided not to use projection of dimensions in this project. An alternative method could be removing of highly correlated features. In this way, just highly correlated features would remain in the dataframe. The following heatmap shows autocorrelation matrix of feature vector.



Totally purple and yellow cells are highly correlated ones and should be selectively removed from the dataframe. After removal of those features the it can be seen from the heatmap and PCA that the redundancy has been minimized.


Autocorrelation Matrix of Feature Vector
Population_Density(persons/km2)
Total_Number_of_Businesses
Mean Employee income ($)
Employee income as main source of income (%)
Median Own unincorporated business income ($)
Mean Own unincorporated business income ($)
Own unincorporated business income as main source of income (%)
Median investment income ($)
Mean investment income ($)
Investment income as main source of income (%)
Total income earners (excl. Government pensions and allowances) - median age (years)
Total income (excl. Government pensions and allowances) - Gini coefficient (no.)
Total income (excl. Government pensions and allowances) - Second Quartile (%)
Total income (excl. Government pensions and allowances) - Third Quartile (%)
Land_Area
Average monthly household rental payment ($)
Owned outright (%)
Owned with a mortgage (%)
Rented (%)
Australia SEIFA decile ranking - Index of Relative Socio-economic Advantage and Disadvantage (IRSAD) (no.)
Australia SEIFA decile ranking - Index of Economic Resources (IER) (no.)
0.8
0.4
0.0
-0.4
-0.8


Scree plot
Percentage of Explained Variance (%)
0
5
10
15
20
25
30
Principal Component (no.)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

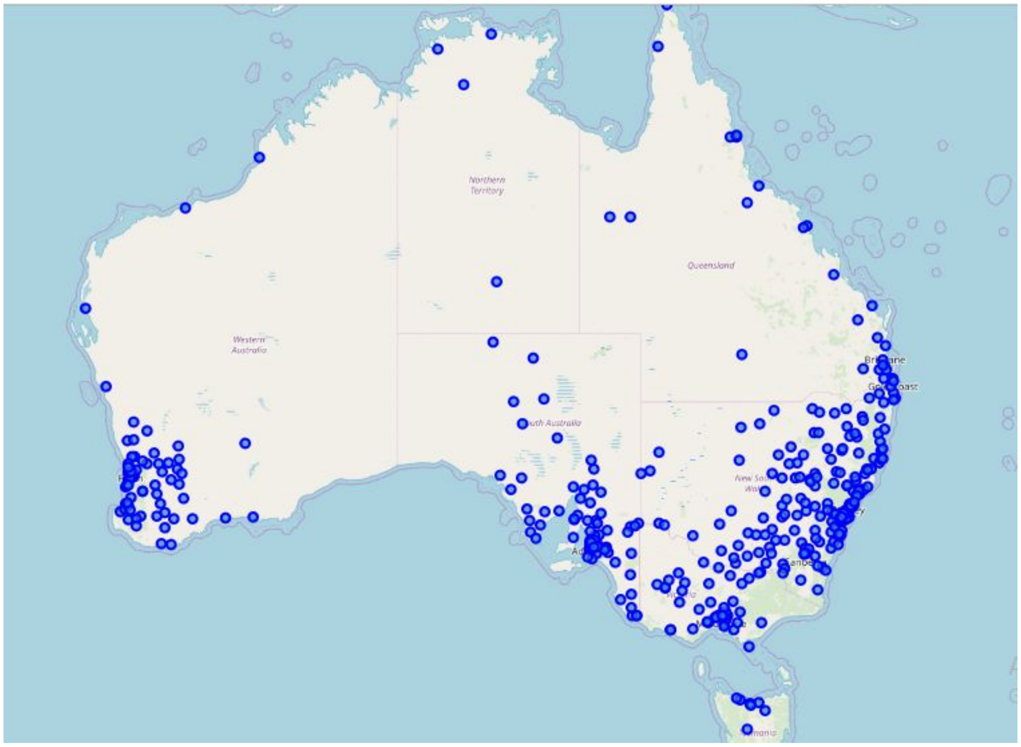
3. Methodology

The goal of this project is to find the best district for inaugurating a new business. Hence, narrowing down the result at each step is needed to make more accurate decisions. For this purpose, after clustering LGAs all over Australia, I limit the result on **Sydney Metropolitan LGAs** and augment the DataFrame with population projection from NSW Government Planning and Environment. First a K-Means clustering is operated on the DataFrame to separate the LGAs and rank them based on their capabilities. Next, from the cluster of high-ranked LGAs an example LGA is opted for further analysis and fine tuning the place of business using extra data. This extra data is about currently available businesses and is collected from Foursquare API. The result would be more precise locations for our purpose.

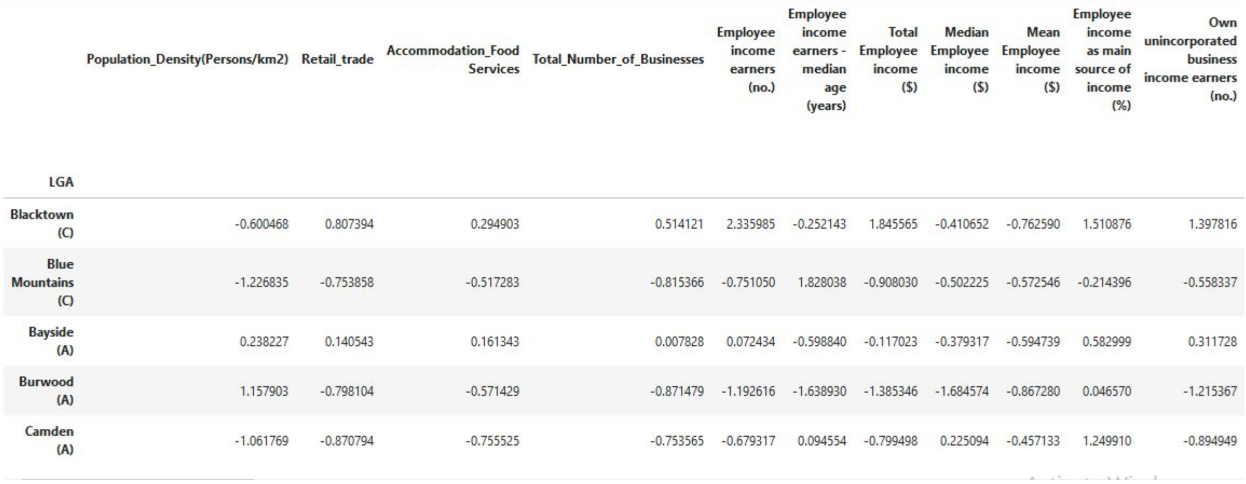
The advantage of this step by step narrowing down model is more useful for those companies that plan to make franchises or branches. This lets them to openly add or reduce features, change algorithms on each step to make the model more accurate based on their experience during launching some of their new branches.

4. Results

Figure below shows all the LGAs from the dataframe over the map before clustering.

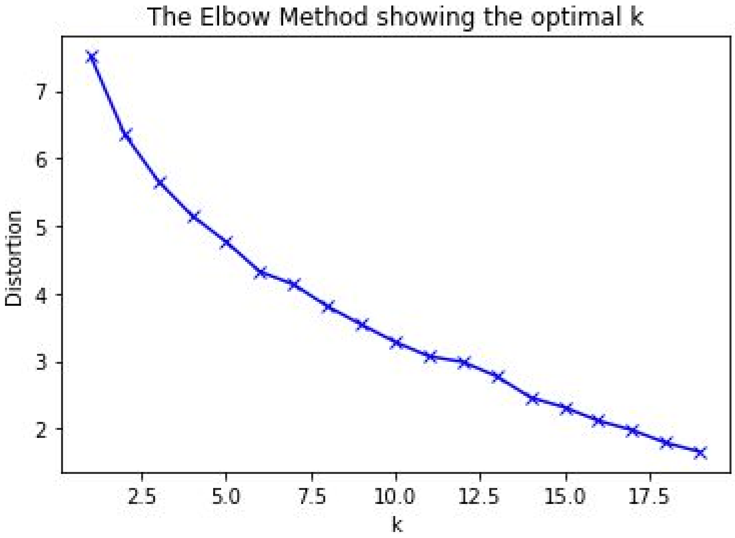


By narrowing down the dataset to Sydney Metropolitan Area and adding the features of population projection to the dataset the size of it become (33, 21) . The head of resulting DataFrame is as below.

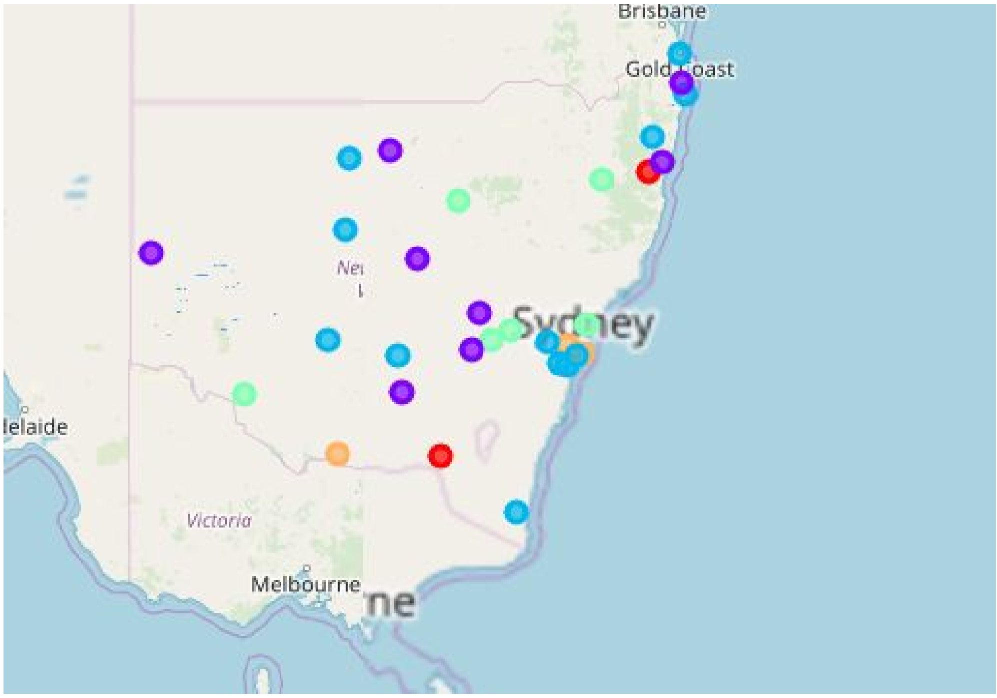


	Population_Density(Persons/km2)	Retail_trade	Accommodation_Food Services	Total_Number_of_Businesses	Employee income earners (no.)	Employee income earners - median age (years)	Total Employee income ($)	Median Employee income ($)	Mean Employee income ($)	Employee income as main source of income (%)	Own unincorporated business income earners (no.)
LGA											
Blacktown (C)	-0.600468	0.807394	0.294903	0.514121	2.335985	-0.252143	1.845565	-0.410652	-0.762590	1.510876	1.397816
Blue Mountains (C)	-1.226835	-0.753858	-0.517283	-0.815366	-0.751050	1.828038	-0.908030	-0.502225	-0.572546	-0.214396	-0.558337
Bayside (A)	0.238227	0.140543	0.161343	0.007828	0.072434	-0.598840	-0.117023	-0.379317	-0.594739	0.582999	0.311728
Burwood (A)	1.157903	-0.798104	-0.571429	-0.871479	-1.192616	-1.638930	-1.385346	-1.684574	-0.867280	0.046570	-1.215367
Camden (A)	-1.061769	-0.870794	-0.755525	-0.753565	-0.679317	0.094554	-0.799498	0.225094	-0.457133	1.249910	-0.894949

For selecting number of clusters in K-Means algorithm elbow method is used. It is worth mentioning that the result of elbow method show how separable the whole dataset is. As shown in below figure the dataset is not perfectly separable and this can be overcome by doing more feature engineering. For the purpose of this project we assume this is enough for making a decision about which LGA should be used for next step.



Result of clustering on the map is shown in the next figure.



Retaining meaningfulness of the features was for this step to check that which cluster is more suitable for our purpose. This search is much simpler than previous blind search. For example, as per restaurant, someone can check best LGA based on number of available food and retail store, the income of region, the population projection of region, etc. based on these inspection, I selected “Penrith (C) (NSW)” as the case for further analysis.

Making calls to the Foursquare API for available restaurant in that region and doing one-hot encoding will lead to a new dataframe of interest. The final decision should be taken based on this even through another clustering or by selecting an important feature like type of restaurant.

4. Conclusion

In this project a top-down approach to make a decision using data science has been proposed. At first, it is essential to select reliable and comprehensive set of features to feed to the other steps. almost all the algorithms need eliminating of redundancy between features. Thus, an essential part of the project would be feature engineering. Next, clustering the dataset will make the options easier to be chosen. by narrowing down the response space based on previous stage, other more precise features could be added to the DataFrame to make it richer. These new features are opted from Foursquare API and are related to the venue of interest like restaurant.

Based on the result of the analysis, it's possible to define a new metric for the best place to found the new business. Moreover, some other features like how people of that LGA tend to eat out, or something like their income or number of family members, ethnicity could make our analysis more precise. By taking data from the prosperity level of businesses into account from premium calls of Foursquare API, a better outcome can be reached.

Based on this data examining, someone can check how one LGA is growing and what is the status of current businesses available.