Guilherme Pichinelli de Carvalho

https://github.com/pichinelli

Launching Teahouse stores in Australia

FULL REPORTING of The Battle of Neighborhoods Capstone Project required for Applied Data Science Capstone course by IBM on Coursera platform

IBM Data Science Professional Certificate

Applied Data Science Capstone by IBM

Index

[Glossary 2](#_Toc9526268)

[What is this document for? 2](#_Toc9526269)

[Introduction 2](#_Toc9526270)

[Business Problem 3](#_Toc9526271)

[Data 3](#_Toc9526272)

[Methodology 4](#_Toc9526273)

[Data Wrangling, Data Preparation and Data Analysis 4](#_Toc9526274)

[Use of Machine Learning 8](#_Toc9526275)

[Results 9](#_Toc9526276)

[Discussion 11](#_Toc9526277)

[Conclusion 12](#_Toc9526278)

# Glossary

|  |  |  |
| --- | --- | --- |
| Term | Definition | Source |
| Teahouse | A teahouse (mainly in Asia) or tearoom (also tearoom) is an establishment which primarily serves tea and other light refreshments | <https://en.wikipedia.org/wiki/Teahouse> |

# What is this document for?

This document was created as a form of submission for the Full Report of Capstone Project required for Applied Data Science Capstone course provided by IBM on Coursera platform. “The Battle of Neighborhoods” is the course section name which refers to the learning context defined by IBM for the whole course, which focused on the use of geo data providers, such as Foursquare, to leverage location data for Data Science projects use.

The scope of this work is limited to:

* Introduction of the business problem and who would be interested in this project.
* Data used to solve the problem and the source of the data.
* Methodology section which represents the main component of the report, containing discussion and description of exploratory data analysis inferential statistical testing performed, and what machine learnings were used and why.
* Results section
* Discussion section with any observations noted and any recommendations based on the results.
* Conclusion section

# Introduction

This project intends to address the following hypothetical business scenario:

Industry: Retail

Category: Teahouse chain

Country: Australia

Business problem: A world leading Teahouse chain retailer is running an expansion business program having Australia as a new country and it has targeted the launching of 5 stores in the next business quarter. The company is in the process of determining the best locations considering the known historical aspects that drive the best business outcome. In a recent past, the company was using a traditional criterion to determine the places for new stores launching. As the Teahouse business has grown significantly in Australia and many competitors got presence in the country, the company is seeking for a data driven decision process to help them make the best decision where to launch the 5 new stores.

# Business Problem

As briefly described in the previous section, the Teahouse chain retailer is aiming to use data to drive the business decision where launching new stores in Australia. At this point of time, there is a baseline scenario to be considered in the project:

* **Target cities**: Sydney and Melbourne are preferred locations given the higher average ticket compared to other cities in Australia (driver got from external research), which is approximated +3% more profitable than other cities. The other cities are Adelaide, Brisbane, and Perth.
* **Customer profile**: Teahouse customer profile in more than 90% of total customer base are young people, in the age of 16-25 years, academic students. The company records excellent sales achievements in stores close to universities (up to 1500m distant) in other countries where it already operates.
* **Competitors:** Beyondthe competitorsplaying in the exact same field, some Coffee shops also competes with the same customer profiles.

Having the baseline scenario above, the project intends to use data from multiple sources and machine learning models to drive best decision making considering as starting point the below criterion:

1. Locations where customer profile population concentrates in presence, in a radius of 1,5 km of Universities locations;
2. Focus on the 5 target cities, taking in consideration the research insight of better results made in Sydney and Melbourne. Target cities are considered 40k radius from the central point from the respective City;
3. Teahouse business presence vs. customer profile population in the region, to be considered as a criterion;
4. Presence of competitor, having in consideration those in the same field and the indirect competitors as well.

As expected in a data science project, data analysis may guide for baseline changes and/or refinement. The approach used by the company in previous stores launching was not data driven and would not be a surprise in this project new insights and perspectives during its execution.

# Data

In this section it is described what are the core data sources identified for the project execution:

**Source: Foursquare API**

**Description:** API service

**Purpose:** Obtain location data geolocation of venues (competitors & universities). During the project, it will be required to find out the geolocation of universities as point of interest, then map the competitors in the respective region (radius of 1,5km).

**Dataset:** <https://developer.foursquare.com/>

**Source: Docs Education.gov**

**Description:** Educational data published by official entity in Australia related to Universities

**Purpose:** Obtain important data about Universities, mainly number of students to drive business decision. This dataset will be determinant to identify the ratio of students per quantity of competitors and then map places where the Teahouse business have a greater potential for return of investment.

**Dataset:** <https://docs.education.gov.au/system/files/doc/other/2018_first_half_year_student_summary_tables.xls>

**Source: Google Geocoder API**

**Description:** API Service

**Purpose:** Obtain central points of Target cities. For this purpose, other services could be used, but intentionally Google Geocoder API was chosen as an experimentation for the service. Also Foursquare API lacks this capability.

**Dataset:** <https://cloud.google.com/>

# Methodology

With the Business Problem determined, the project started looking after the main dataset for use which should be one ideally published by an official entity, recent, containing features such us University name, City, Number of students, at least. That would be the starting point and then use Foursquare API to help analysing competitors and preference locations to open new stores.

After a good amount of search, it was found a consistent and free dataset, published by an official entity, with recent data, containing the expected main features. The only feature missing was the City of the respective University facility, which would help identifying points of interest following the Target Cities idea given as Business drive.

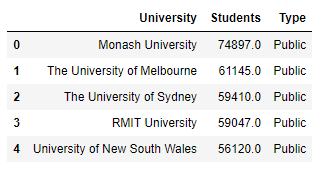
## Data Wrangling, Data Preparation and Data Analysis

The main dataset was found at **Docs Education.gov** (more information in Data section). It is a MS Excel file with Information about Universities in Australia. This dataset is preprocessed and the data is grouped by Type of Institution (Public and Prvate) and by State which in Australia are New South Whales (NSW), Victoria (VIC), Queensland (QLD), Western Australia (WA), South Australia (SA), Tasmania (TAS), Australian Capital Territory (ACT) and a category called Multi-State. The features available are:

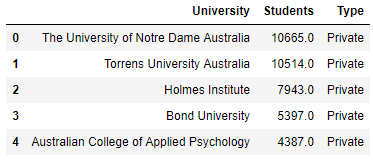
* Type of University (group)
* State of University (group)
* Institution
* 2017 Commencing Students – Number of students (per University)
* 2018 Commencing Students – Number of students (per University)
* 2018 Commencing Students – % of total (per University)
* 2018 Commencing Students – % change from 2017
* 2017 All Students – Number of students (per University)
* 2018 All Students – Number of students (per University)
* 2018 All Students – % of total (per University)
* 2018 All Students – % change from 2017

The data of interest was stored in the sheet named "4", it was found easier get the data in 2 pieces and then concatenate them in one unique Datafrrame. First getting the Public universities data, then Private universities data.

That was the first Public Universities Dataframe:



Then, in the sequence a Private Universities Dataframe:



By taking this approach was easier to concatenate and made one unique Dataframe.

After concatenating, a new feature was created and the Dataframe was prepared to receive more data from Foursquare API. Details below:

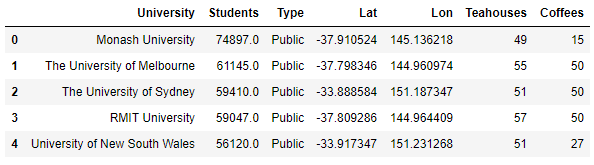
* New feature **Percentage** to contain % of students compared with the total
* Column **Lat** for Latitude of University geo location
* Column **Lon** for Longitude of University geo location
* Column **Teahouses** to indicate the number of Teahouses and Tearooms located close to the University
* Column **Coffees** to indicate the number of Coffees located close to the University

With the basis of Dataframe established, using Geolocator Geocoder and Foursquare API, critical Features were fulfilled to be used as the basis for this project.

Some instability was perceived using Geolocator service. Based on testing the timeout parameter was set to 3 in Nominatim function. With this less timeout errors were received, but Geolocator was still showing undesirable instability.

Based on this experience it was decided to try Google Geocode API instead of Geolocator Geocoder. The result was much better.

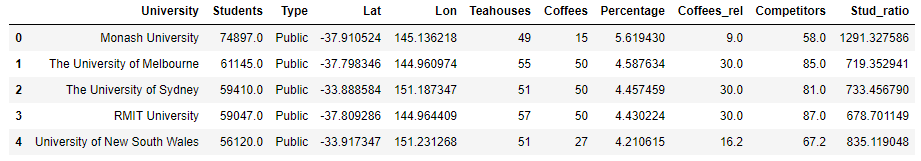
Here it is the result of the Dataframe enriched using Google API.



Analysing the resulted Dataframe, some Feature Engineering was found as valuable. Details below:

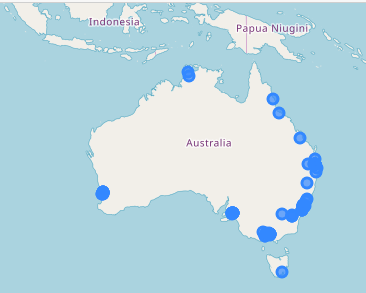
* New feature **Percentage** to contain % of students compared with the total;
* New feature **Coffees\_rel**, based on business assumptions a Coffee shop represents a partial competition. The ratio assumed based on business assumptions is 0.6, so, for each 1 Coffee shop, it has been considered 0.6 competitor in nature
* New feature **Competitors** is the sum of Teahouse and Coffees\_rel columns to show a relative total number of competitors in the defined radius taking the respective University location as centroid.
* New feature **Stud\_ratio** is the ratio of Students per number of Competitors. This feature gives a drive of how many target customers exists for each competitor.

This is how the Dataframe became:



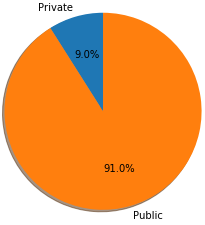
With the Dataframe built, some initial analysis about the data could be performed.

The first analysis was done using Folium map, to visualise the locations (better visualization in the Jupyter Notebook).



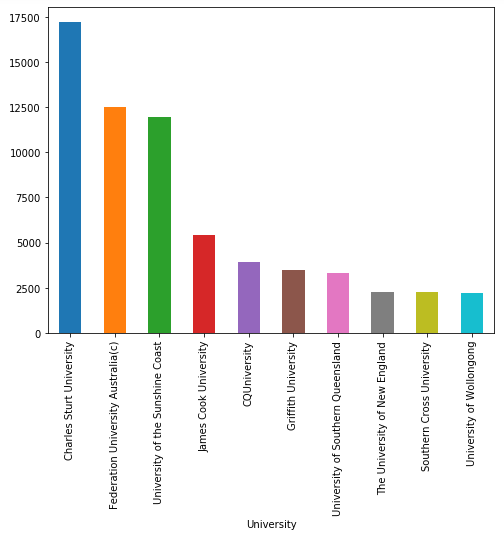
Analysis: Folium map shows a concentration of Universities in places like Sydney, Brisbane and Melbourne.

A comparison of number of students in Public vs. Private Universities:



Analysis: immense majority of Students coming from Public Universities.

The **Stud\_ratio** feature seems a main driver for decision. Below is a chart of Top 10 Universities by Stud\_ratio feature:



**Challenge with Australia Postcode structure, suburbs and cities datasets**

As a surprise in Australia the official Postcode dataset is not freely available. The country postal office agency, Australian Post, commercialize this information as an API service. Most of the alternative datasets found did not appear very 100% accurate. That was not the main challenge.

**Suburbs and Cities in Australia**

Different to other countries, in the Australia Postcode structure, the Postcodes are distributed across the states with some exceptions (details here: <https://en.wikipedia.org/wiki/Postcodes_in_Australia> ). The difficult part is, when you consume Google, Foursquare or any other API, the attribute 'city' comes with the name of the Suburb and not with the City name. That was a not predicted challenge in the beginning of the project. As stated in the Business Problem, there is a business drive of Target Cities, which should be considered given the initial assumption of preference locations. Due this challenge the project came up with a different solution approach which consists of:

* Identify the central latitude/longitude coordinates of Target cities
* For each Target City, identify in a radius of 40km all Universities, flagging them

An auxiliary Dataframe containing the central points of Target Cities was created to help mapping them accordingly. That’s how it was created:

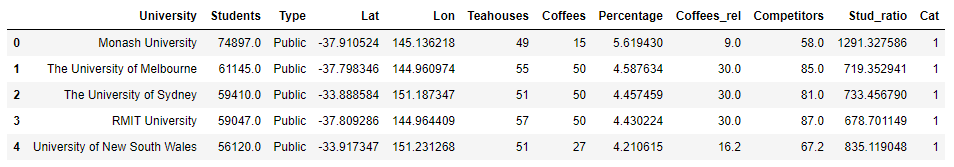


With this table, instead of adding to each University a respective City, given the mechanism of considering the radius of central points in the city, it was found better create Categories, using the following rule:

* **Category 0:** Any other location in Australia different to Target Cities
* **Category 1**: Sydney or Melbourne. Target Cities with theorical higher return of investment (refer to Business Problem section).
* **Category 2:** Adelaide, Brisbane or Perth.

A new Feature was created then, called **Cat**.

With this additional feature, this how the Dataframe was looking:



## Use of Machine Learning

Using **Feature Engineering**, especially with the new feature **Stud\_ratio** & visualizing a pre-result in the **Folium** map, the initial orientation of preference of Target Cities started to be challenged. It was observed that Universities with higher **Stud\_ratio** are frequently outside the Target Cities. In order to visualize this better, an unsupervisioned clustering Machine Learning such as **K-Means** would help challenge/confirm if the business drive of Target Cities is a consistent assumption for business decision.

The new questions that came with the analysis were:

* "**Why take Target Cities as a hard business drive?"**
* **"What if the best places to launch the stores are not necessarily in the Target Cities?"**

PS: In a real business scenario, there are other factors to take in consideration in similar projects such as Facilities, Logistics, Workforce, and others. Here, the project purposes to focus on the dimension of customer profile population and competition.

**K-Means Clustering**

As explained before, K-Means Clustering was chosen given the capacity in performing unsupervisioned clustering. The expectation is to see if the algorithm shows reasonable categorisation from business perspective.

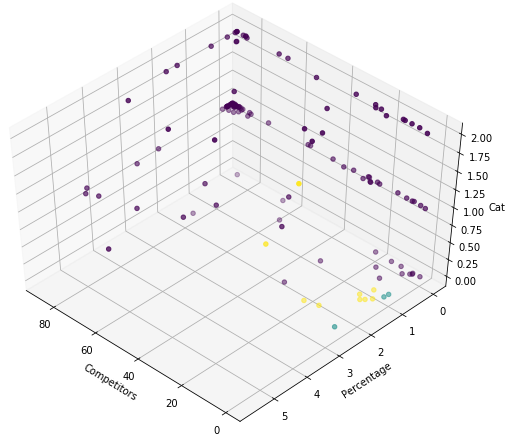
**The best K value for this model¶**

I'm not reporting the historical of all attempts using different values of K used in my analysis. I tried from 1 to 10. In order to judge what is the best K value for the model I had to run multiple rounds with different values of K and analyses the results. In this learning process it was realised best results coming from K=3. More than 3 clusters were giving a non-clear sense of clustering, making it confuse. Less than 3 clusters were not giving enough meaning to understand the options and drive a business decision. So, basically, K=3 made sense from business perspective, gave a better view to evaluate the options. The cost of different K values was calculated, but that did not help much in determining a reasonable clustering in this scenario (details in Jupyter Notebook).

# Results

The results are better visualised using charts, as below.

**3D Axis Plotting**

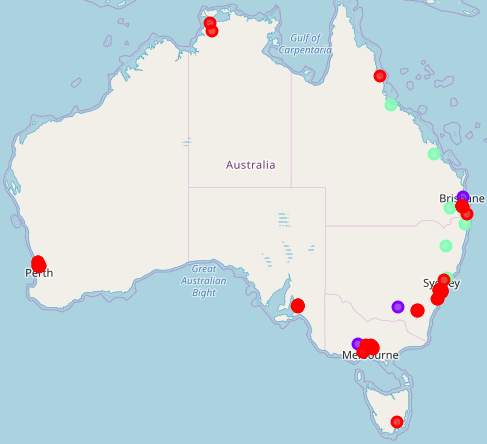


There are good observations about this chart.

The business orientation is to find the points with low competition and high percentage and then locate which categories they belong to in the chart.

* Most of the green and yellow points are grouped, making sense with the orientation for best results.
* Two yellow points with higher competition and low percentage has been perceived as noises.

**Map chart analysis**



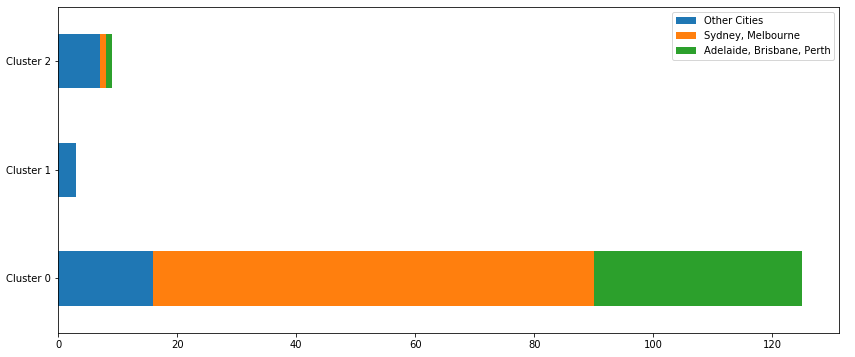
In these comments the Map + Individual clusters analysis (in below sections) are incorporated:

* **RED DOTS** - Cluster 0: these are 125 Universities with high competition and/or low number of students. Which suggests as not best locations to open new stores
* **PURPLE DOTS** - Cluster 1: these are 3 Universities with the low competition and/or high number of students. Which suggests as best locations to open new stores
* **GREEN DOTS** - Cluster 2: these are 9 Universities with reasonable low-mid competition and/or mid-high number of students. Which suggests some further analysis to pick-up the best 2 options out of 9.

**Individual Cluster analysis**

For better visualisation, a Stacked Bar chart reveals that Target Cities business drive is not consistent with current Australia Teahouse business scenario.

* Almost all Universities located in Target Cities, belong to **Cluster 0**, which suggests to not open a new Teahouse store given the high competition: it is a **No Go**.
* The Universities in **Cluster 1** are **Must Go**, given the favorable scenario in terms customer profile population vs. competition.
* **Cluster 2** shows good and favorable locations for new stores opening, it requires attention due some noises, even though it guides well for a **Should Go** classification.



* Almost all Sydney-Melbourne located Universities (orange color) belongs to Cluster 0.
* Better options in Other Cities (blue color), can be seen in Cluster 1 and Cluster 2
* Few Universities in Target Cities clustered in Cluster 2.

# Discussion

This project started with the objective of enabling the best decision process by a Teahouse chain, in the context of launching its first 5 stores in the Australia. Instead of using the known and traditional way of taking this decision, the company intended to use a data-driven approach. As a starting point some business drivers were given as domain inputs for the analysis. It was expected some new insights and gaining of new perspectives. The concept of Target Cities was a given driver.

During the project run was observed an inconsistency by taking business driver in the Australia Teahouse scenario. The first indication appeared during the Feature Engineering phase, by creating the **Stud\_ratio** feature, which brought a different perspective.

Which was not expected was the Postcode structure challenge in Australia, which further made the use of map charts very important to get the right direction to challenge the Target Cities concept.

With that on hand, the project gained an additional objective which was using the data to convey that Target Cities should not be an untouched condition.

The use of K-Means along with the results visualization helped making a strong argument to drive business in launching new stores focusing on locations in the Clusters 1 and 2, and not in Cluster 0.

This project was characterized by:

* Higher effort in Data wrangling, cleansing, feature engineering and programming challenges;
* Use of free-open datasets;
* Use of different APIs such Foursquare and Google;
* Reuse of free-available functions in Stackoverflow such as Harvesine and Progress Bar;
* Use of advanced Python Libraries such as Matplotlib, Folium, SciKit-Learn;
* Use of Data manipulation and Visualisation to demonstrate conclusions.

# Conclusion

From business perspective:

* Business drivers are important because they give a domain sauce to the analysis, but data can reveal different perspectives and change a given business driver. It is important to be open for this.
* **Cluster 0** reveals as **Must Go** if customer profile population vs. competition is the key analysis to take the decision
* **Cluster 1** reveals opportunities and should be treated as **Should Go**. There are some opportunities to make the model better or use another Machine Learning model to minimize or eliminate the noises.
* **Cluster 2** is a **No Go** for the first 5 stores launching.

From technical/project perspective:

* **Feature Engineering** is a very important step in the project and itself may reveal critical directions to the project.
* Geolocator is very unstable and the use of Google Geocoder API made the difference in terms of data quality