Guilherme Pichinelli de Carvalho

https://github.com/pichinelli

The Battle of Neighborhoods Capstone Project (Week 1)

Submission of first part (Week 1) 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](#_Toc8998421)

[What this document is for 2](#_Toc8998422)

[Introduction 2](#_Toc8998423)

[Business Problem 3](#_Toc8998424)

[Data 4](#_Toc8998425)

# Glossary

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

# What this document is for

This document was created as a form of submission for the first part 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:

* Clearly define a problem or an idea of choice, where it would need to leverage the Foursquare location data to solve or execute it.
* As any data science problems always target an audience, it means that the project should help a group of stakeholders in solving a business problem. So, it is mandate to explicitly describe the target audience and why they would care about the described business problem
* The submission must cover:
  + A description of the problem and a discussion of the background;
  + A description of the data and how it will be used to solve the problem.

# 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 a major 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 to launch 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 (data 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. As previously stated, the common 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/>