Applied Data Science Capstone Project, IBM Data Science

Battle of Train Stations: Singapore Train Station Vicinity Exploration and Clustering

Robert Ci
15th March, 2021

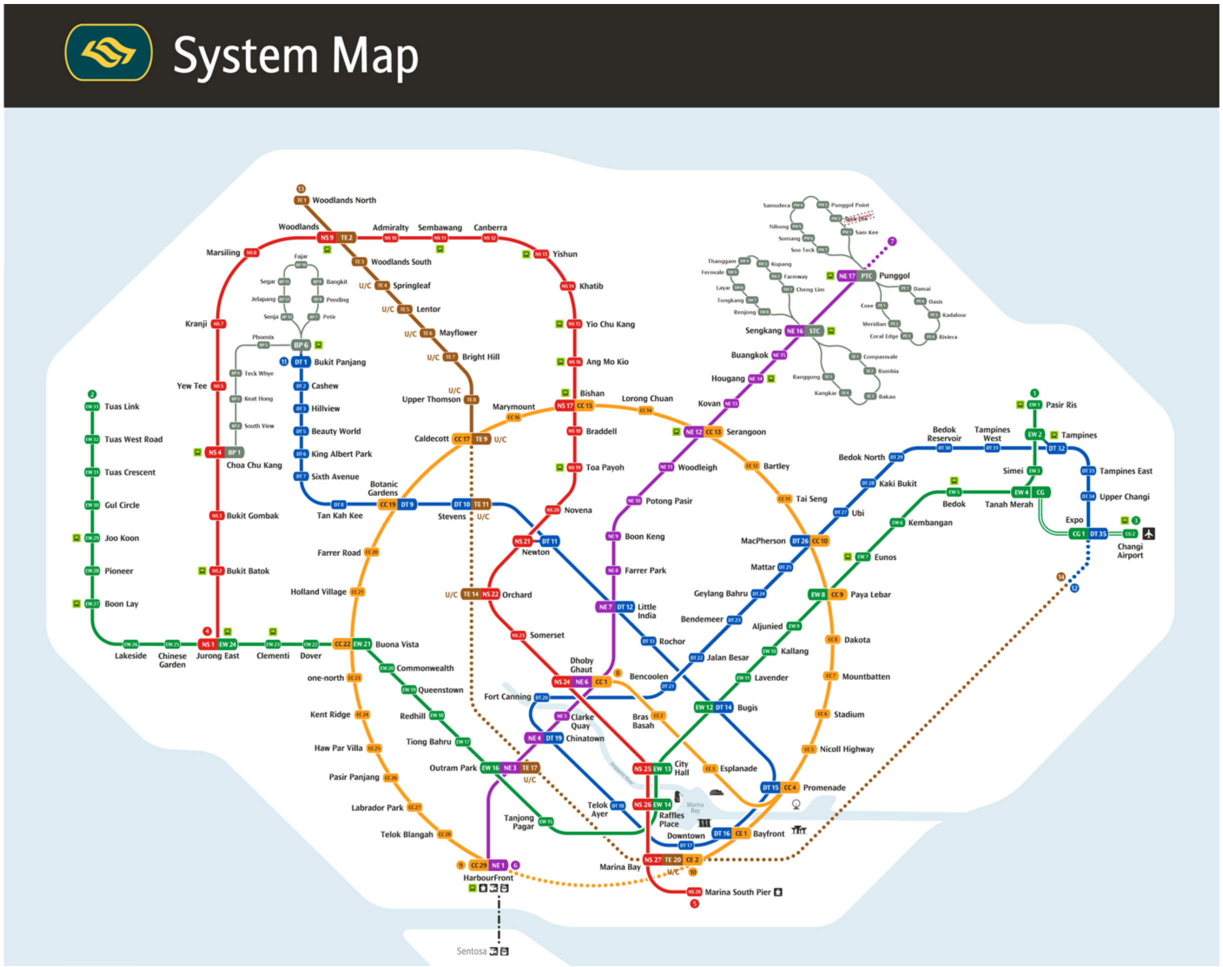


Table of Contents

1. Introduction

1.1. Background

Just like London Underground and New York City Subway, the railway system in Singapore forms an indispensable and inseparable part for everyone's life in this small island city-state. It's fast, safe, efficient and passenger-friendly. According to Land Transport Authority ("LTA"), currently there are 6 Mass Rapid Transit ("MRT") and 3 Light Rail Transit ("LRT") lines in operation with over 150 train stations scattered island wide.

Train stations in Singapore, big or small, are not simply places for commuter services. They are also connection points of various types of amenities around: coffee shops, convenience stores, ATMs, florists, eateries, clinics, malls, and so many other. Based on their versatility, it becomes an interesting topic for city adventurers and residents in the neighborhood to take a deeper look on them and explore the surroundings of each station for new re-discoveries.

1.2. Objective

This project will develop, analyze and try to answer below questions regarding the main topic:

- What kinds of amenities are there around each train station within a given radius?
- Which stations are similar to other stations, in terms of amenity types?

2. Data

2.1. Data Scope

To fulfill above objective, we would require data of the following fields:

a. Basic information of each station (station name, geographic coordinates)
b. Additional information of each station (station codes, line names)
c. Venues in each station's vicinity, and their types
d. Search radius around each station

2.2. Data Acquisition

Unfortunately, neither the LTA nor the Urban Redevelopment Authority ("URA") has provided a correct, clean and concise all-in-one dataset of train station names, station codes, line information and their corresponding coordinates in WGS 84 (also known as "EPSG:4326") format (EPSG, 1984).

In order to reasonably simplify the process and save time, two separate public datasets in .csv format from Kaggle and Land Transport DataMall will thus be used to cover field a and b. They are:

- 'mrt_lrt_data.csv' (Lee, 2019), and

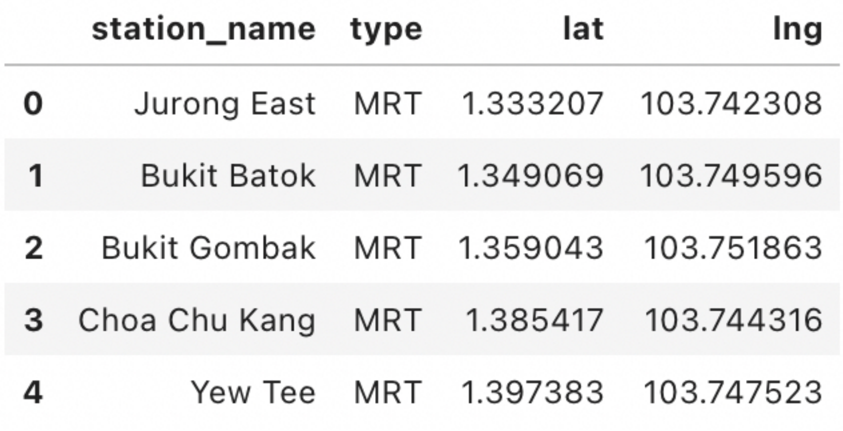
- ‘Train Station Codes and Chinese Names.csv’ (Land Transport Authority, 2018).

As for field c, a dynamic dataset from Foursquare API will be utilized. It is in .json format and contains the result of exploration, i.e. the venue names and types (categories), around an individual station.

Field d will be determined by calculation. The optimal radius for exploration around a station shall not be longer than the ½ of the minimum distance between two stations, in order to avoid any overlapped exploration areas of one station and the other. The calculation will be demonstrated in a later section of this project.

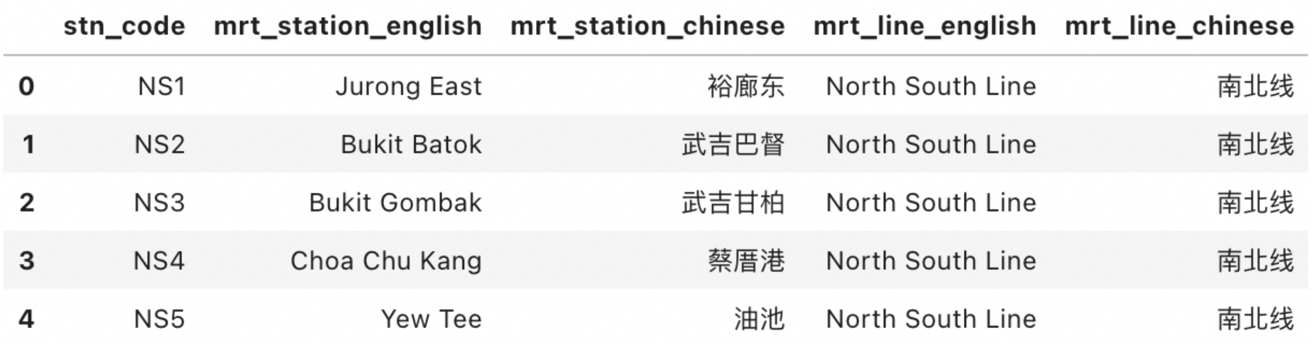
2.3. Data Preprocessing

The aforementioned first two datasets are read and loaded as Pandas dataframes in Python.



	station_name	type	lat	lng
0	Jurong East	MRT	1.333207	103.742308
1	Bukit Batok	MRT	1.349069	103.749596
2	Bukit Gombak	MRT	1.359043	103.751863
3	Choa Chu Kang	MRT	1.385417	103.744316
4	Yew Tee	MRT	1.397383	103.747523

Figure 1. Dataframe of Basic Information of Each Station



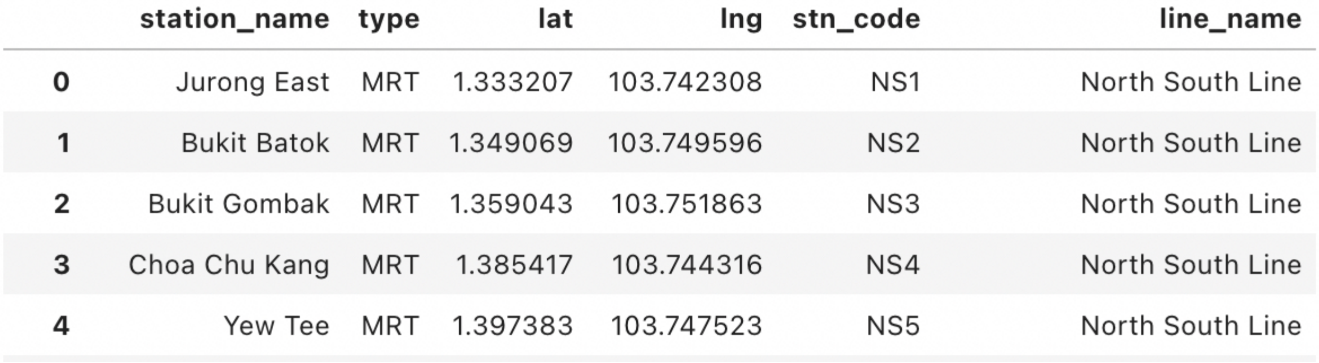
	stn_code	mrt_station_english	mrt_station_chinese	mrt_line_english	mrt_line_chinese
0	NS1	Jurong East	裕廊东	North South Line	南北线
1	NS2	Bukit Batok	武吉巴督	North South Line	南北线
2	NS3	Bukit Gombak	武吉甘柏	North South Line	南北线
3	NS4	Choa Chu Kang	蔡厝港	North South Line	南北线
4	NS5	Yew Tee	油池	North South Line	南北线

Figure 2. Dataframe of Additional Information of Each Station

The two dataframes are merged to further find out the stations without geographic coordinates. One station (Ten Mile Junction LRT Station) is found and removed from the merged dataframe, as geographic coordinates are necessary features.

Since this project will be carried out all in English, two features in Chinese are regarded as redundant information and removed.

Moreover, some stations are interchange stations and each has more than one station code and more than one line name in nature, as each station code represents a station's place in one particular line in sequential order and an interchange station can belong to more than one line. Such being the case, occurrence of each station name other than the first is regarded as duplicated entry and removed. This removal will not affect the results.



	station_name	type	lat	lng	stn_code	line_name
0	Jurong East	MRT	1.333207	103.742308	NS1	North South Line
1	Bukit Batok	MRT	1.349069	103.749596	NS2	North South Line
2	Bukit Gombak	MRT	1.359043	103.751863	NS3	North South Line
3	Choa Chu Kang	MRT	1.385417	103.744316	NS4	North South Line
4	Yew Tee	MRT	1.397383	103.747523	NS5	North South Line

Figure 3. Dataframe after Data Preprocessing

The dataframe after data preprocessing is ready for the processes in later sections. It contains information of 5 features for 157 trains stations in total.

A map of all train stations as research subjects are plotted.

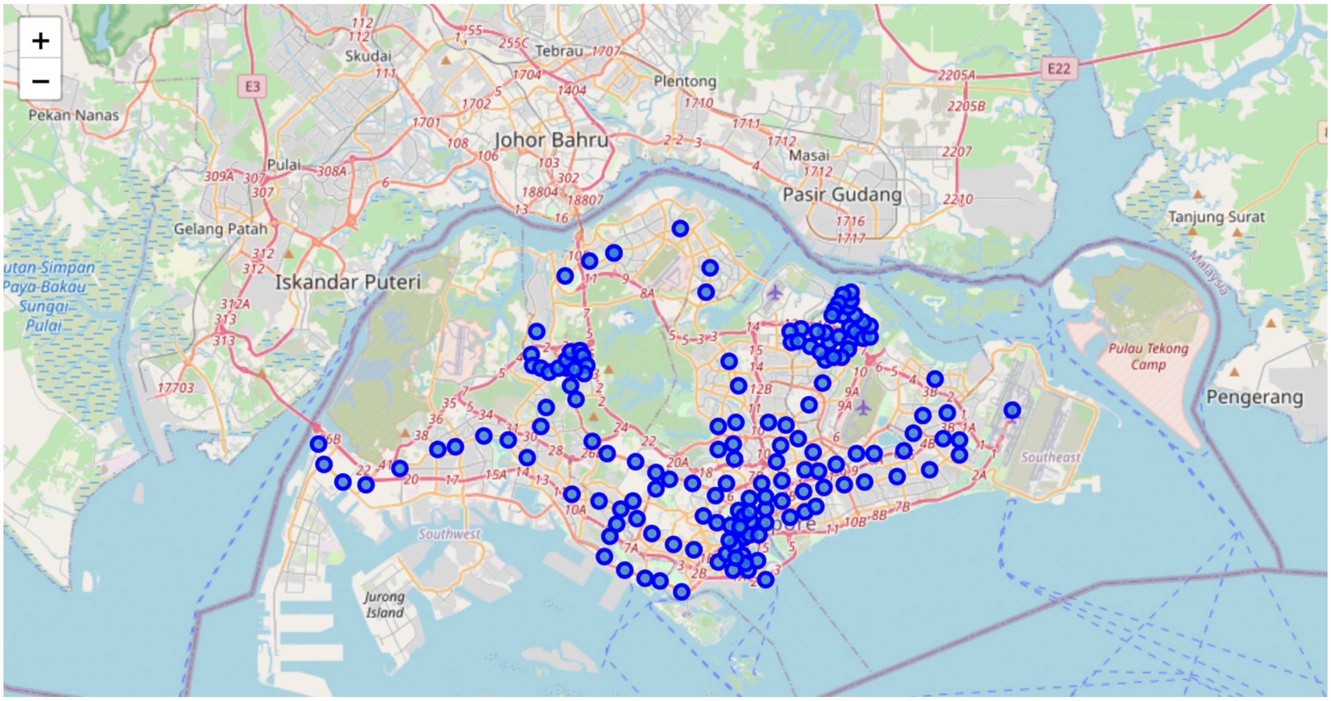


Figure 4. Map of Train Stations to be Researched

Bibliography

EPSG. (1984). *EPSG:4326*. Retrieved from https://epsg.io/4326

Land Transport Authority. (2018, March 19). *Static Datasets.* Retrieved from Land Transport DataMall: https://datamall.lta.gov.sg/content/datamall/en/static-data.html

Lee, Y. (2019, August 3). *Singapore Train Station Coordinates.* Retrieved from Kaggle: https://www.kaggle.com/yxlee245/singapore-train-station-coordinates