Applied Data Science Capstone Project, IBM Data Science

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

Robert Ci
15th March, 2021

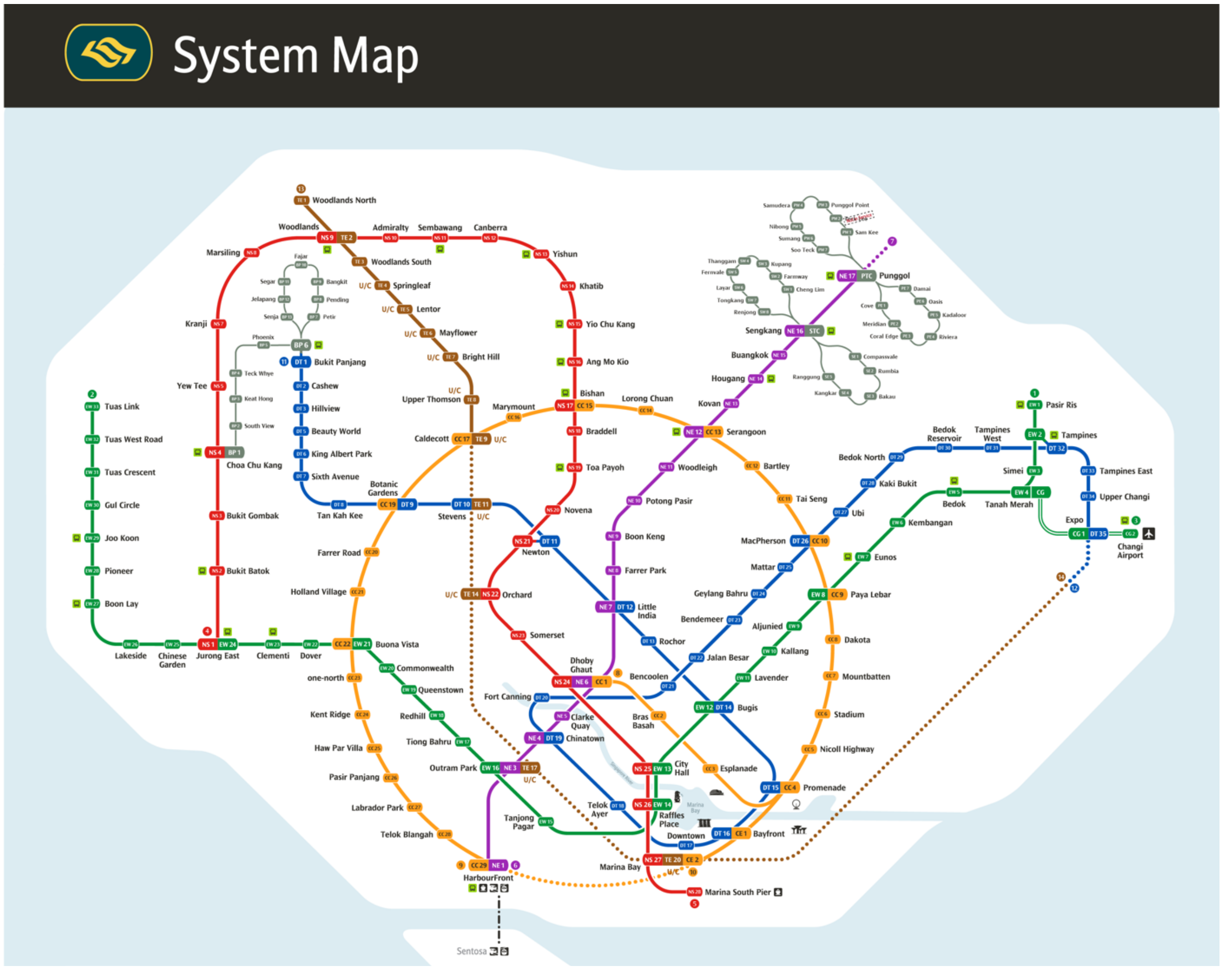


Table of Contents

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 six Mass Rapid Transit (“MRT”) lines and three Light Rail Transit (“LRT”) lines in operation with over 150 train stations scattered island wide.

Train stations in Singapore, big or small, are not places merely for commuter services. They are also connection points of various types of amenities nearby: coffee shops, convenience stores, ATMs, florists, eateries, clinics, markets, gyms, malls, and so many other.

On the basis of train stations’ 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 types of amenity are there around each train station within a given radius?
- Which train stations are similar to other train 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 train station (station names, geographic coordinates)
b. Additional information of each train station (station codes, line names)
c. Venues in each train station’s vicinity, and their types
d. Search radius around each train 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 station names, station codes, line names and their geographic 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 are thus 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).

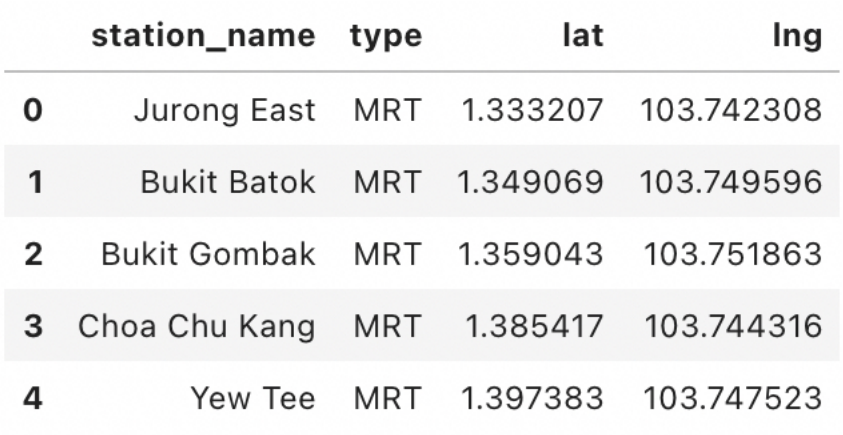
As for Field c, a dynamic dataset from Foursquare API is utilized. It is in .json format and contains the result of exploration, i.e. the venue names and types (categories), around an individual station.

Field d is determined by calculation in a later section of this project.

2.3. Data Preprocessing

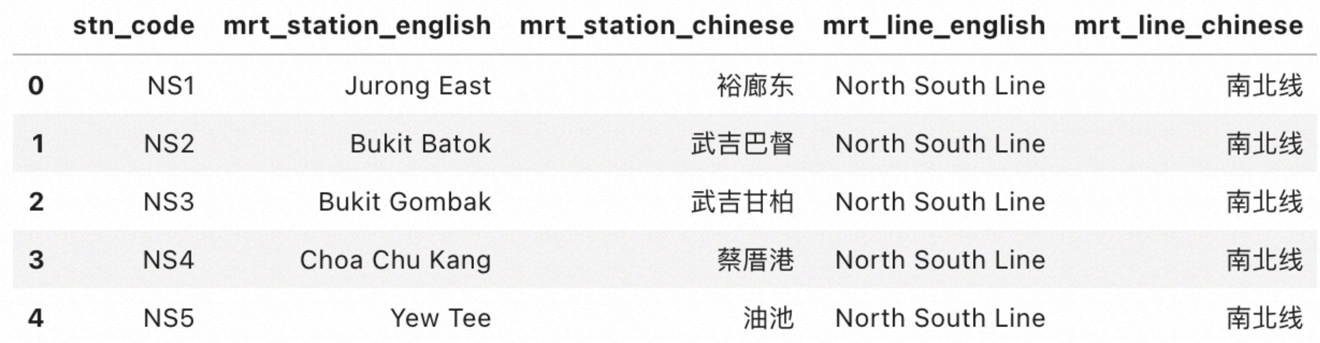
2.3.1. Loading Datasets

The aforementioned first two datasets are read and loaded as Pandas dataframes in Python (Pandas Development Team, 2021).



	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

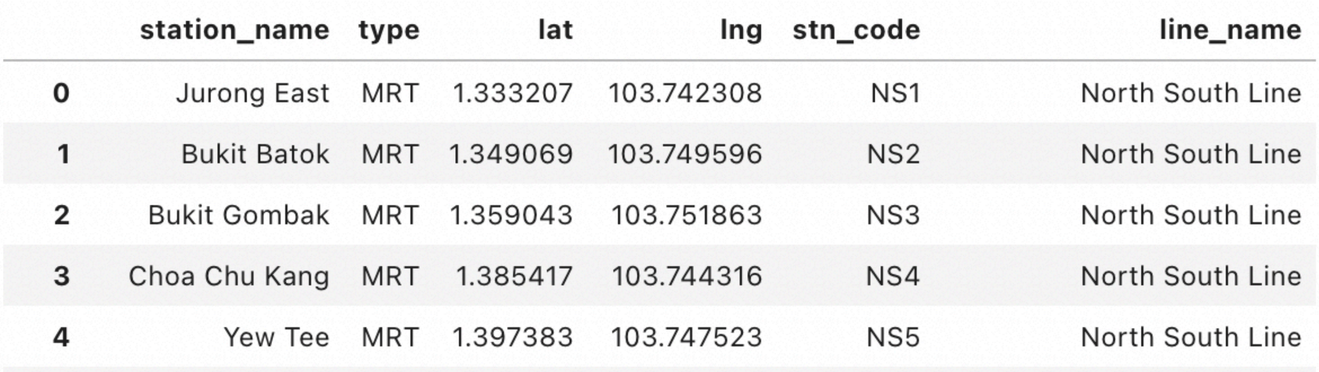
2.3.2. Merging Datasets

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

2.3.3. Cleaning the Dataset

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

Moreover, some train 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 entries and removed. This removal will not affect the results of this project.



	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 train stations in total.

A map of all train stations within the research scope is plotted.

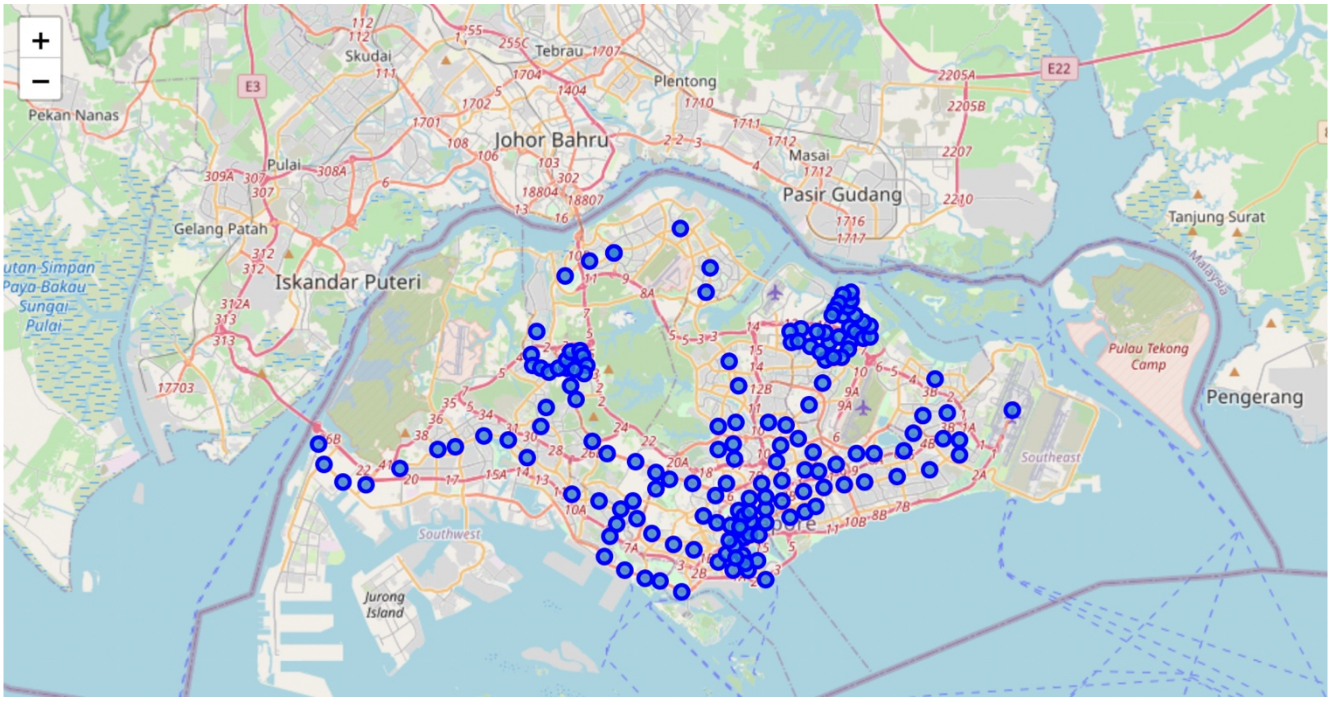


Figure 4. Map of Train Stations to be Studied

3. Methodology

3.1. Exploration

3.1.1. Finding the Optimal Radius

Foursquare API provides search results of recommended venues around a specific geographic place of interest. To find all recommended venues around a place within a given search radius, the ‘explore’ request is applied (Foursquare, 2020).

Before utilizing the Foursquare API for exploration, a search radius mentioned in the above Field d has to be properly calculated as it is a required parameter in making API calls.

In geometric representations, in order to avoid any overlapping of exploration areas around one station and around its closest neighbor station, the optimal radius for exploration around a station shall not be longer than ½ of the minimum distance between two closest train stations in Singapore.

The exhaustion approach is applied for such calculation:

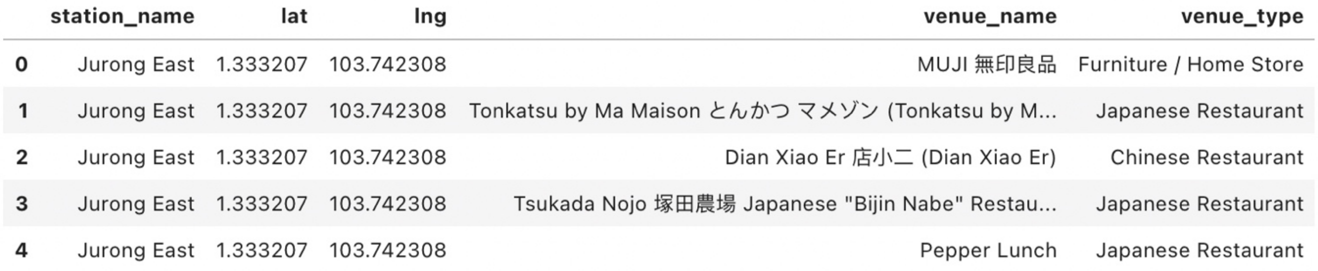
1) The distance between any two train stations in Singapore is calculated and stored;
2) The minimum of all the distance values is found out to represent the distance of the two closest stations;
3) The optimal radius value is the minimum distance divided by 2 and rounded down to an integer.

Above approach is executed utilizing GeoPy library in Python (GeoPy Contributors, n.d.). After the calculation, the optimal radius value is 92, in meters (GeoPy Contributors, n.d.).

3.1.2. Exploration using Foursquare API

API calls are made through programmed processes and all the returned exploration results around each train station are stored in a new dataframe.

The ‘limit’ parameter of each API call is set to 200, i.e., an exploration result around one train station shall contain at most 200 recommended venues.



	station_name	lat	lng	venue_name	venue_type
0	Jurong East	1.333207	103.742308	MUJI 無印良品	Furniture / Home Store
1	Jurong East	1.333207	103.742308	Tonkatsu by Ma Maison とんかつ マメゾン (Tonkatsu by M...	Japanese Restaurant
2	Jurong East	1.333207	103.742308	Dian Xiao Er 店小二 (Dian Xiao Er)	Chinese Restaurant
3	Jurong East	1.333207	103.742308	Tsukada Nojo 塚田農場 Japanese "Bijin Nabe" Restau...	Japanese Restaurant
4	Jurong East	1.333207	103.742308	Pepper Lunch	Japanese Restaurant

Figure 5. Dataframe of Exploration Results

3.2. Analysis

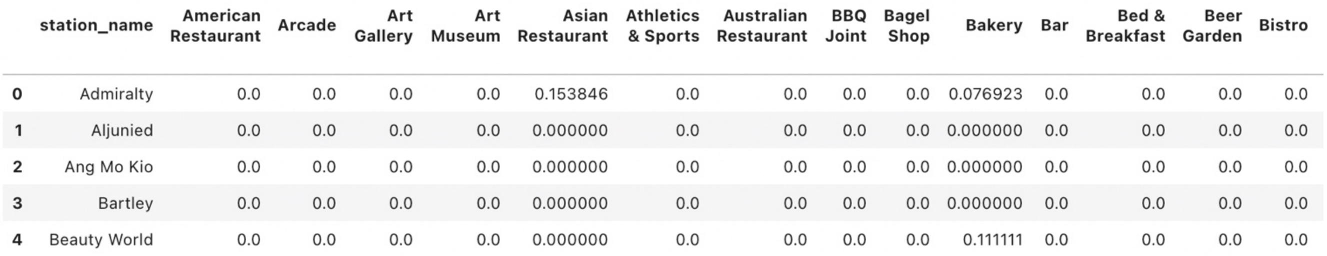
3.2.1. One-Hot Encoding

For the processes in later sections, the ‘venue_type’ attribute in the dataframe of exploration results is converted to a new dataframe of binary numerical values through one-hot encoding (Wikipedia, 2021).

As some attributes ('Boat or Ferry', 'Building', 'Bus Line', 'Bus Station', 'Bus Stop', 'Light Rail Station', 'Metro Station', 'Train Station') in the one-hot encoded dataframe are related to non-amenity types of venues, these attributes are thus not within the study scope and removed from the dataframe.

The ‘station_name’ and ‘venue_name’ attributes are also added to the one-hot encoded dataframe to make it more human-readable.

The rows, each representing a single venue, are grouped by ‘station_names’ on the mean of the occurrence of each venue type in each train station's vicinity. This mean value (frequency) is regarded as a normalized activeness score for a certain type of venue around that train station.

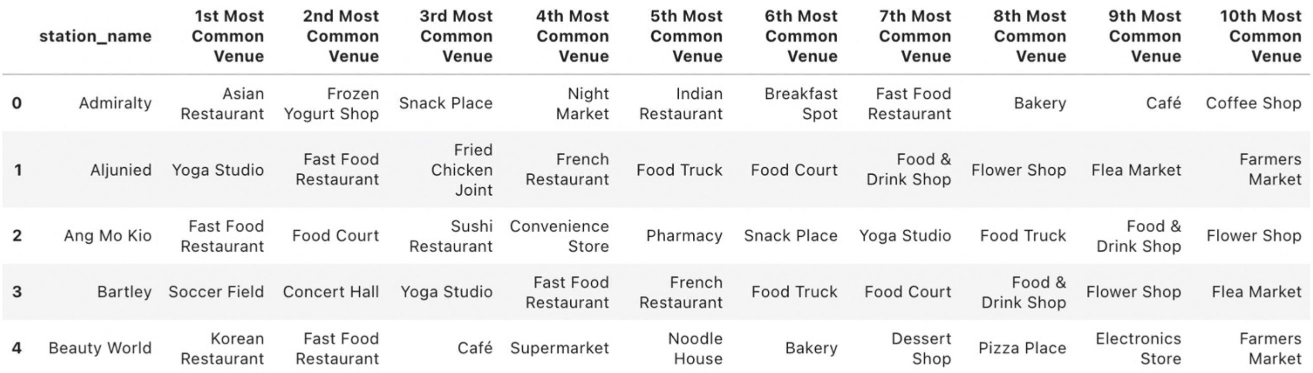


	station_name	American Restaurant	Arcade	Art Gallery	Art Museum	Asian Restaurant	Athletics & Sports	Australian Restaurant	BBQ Joint	Bagel Shop	Bakery	Bar	Bed & Breakfast	Beer Garden	Bistro
0	Admiralty	0.0	0.0	0.0	0.0	0.153846	0.0	0.0	0.0	0.0	0.076923	0.0	0.0	0.0	0.0
1	Aljunied	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
2	Ang Mo Kio	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
3	Bartley	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
4	Beauty World	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.111111	0.0	0.0	0.0	0.0

Figure 6. Grouped Dataframe after One-Hot Encoding

3.2.2. Getting Each Station's Top Venue Types

A new dataframe is created through programmed processes to record the sorted result of each train station’s top 10 venue types by the aforementioned activeness score. Till this stage, the most common amenity types around each train station are known.



	station_name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Admiralty	Asian Restaurant	Frozen Yogurt Shop	Snack Place	Night Market	Indian Restaurant	Breakfast Spot	Fast Food Restaurant	Bakery	Café	Coffee Shop
1	Aljunied	Yoga Studio	Fast Food Restaurant	Fried Chicken Joint	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Flower Shop	Flea Market	Farmers Market
2	Ang Mo Kio	Fast Food Restaurant	Food Court	Sushi Restaurant	Convenience Store	Pharmacy	Snack Place	Yoga Studio	Food Truck	Food & Drink Shop	Flower Shop
3	Bartley	Soccer Field	Concert Hall	Yoga Studio	Fast Food Restaurant	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Flower Shop	Flea Market
4	Beauty World	Korean Restaurant	Fast Food Restaurant	Café	Supermarket	Noodle House	Bakery	Dessert Shop	Pizza Place	Electronics Store	Farmers Market

Figure 7. Dataframe of Top 10 Amenity Types around Each Train Station

3.3. Clustering

To find the similarities among train stations and group them properly on the basis of the most common amenity types around, the unsupervised K-means algorithm is applied.

The unsupervised K-means algorithm is picked as each train station, in terms of amenity types, are not grouped by pre-defined categorical labels. The purpose of this algorithm is to form clusters to group all datapoints, with the number of clusters represented by the variable K. The algorithm works iteratively to assign each datapoint to one of the K clusters tessellated based on feature similarities, until convergence is reached.

3.3.1. Determining K's Value

Elbow method is applied to determine the best value of K. K's range is set between 3 and 9 to avoid underfitting or overfitting.

In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use (Wikipedia, 2020).

Above method is executed utilizing SciKit-Learn library in Python (SciKit-Learn Developers, n.d.). After the calculation, the optimal K value is 6.

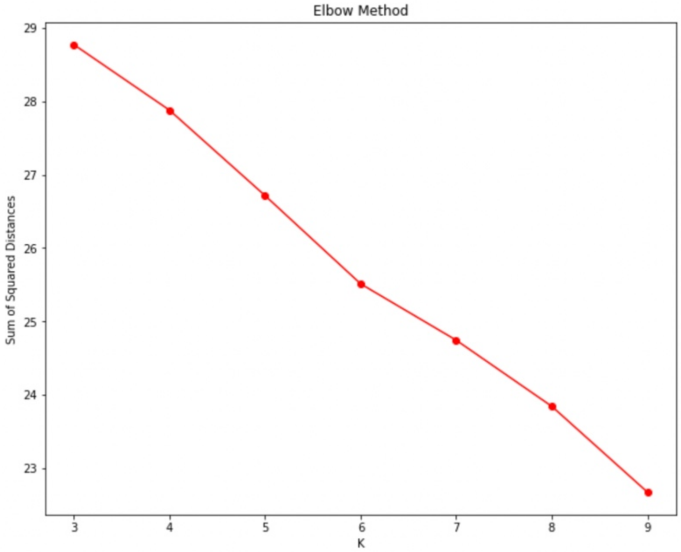
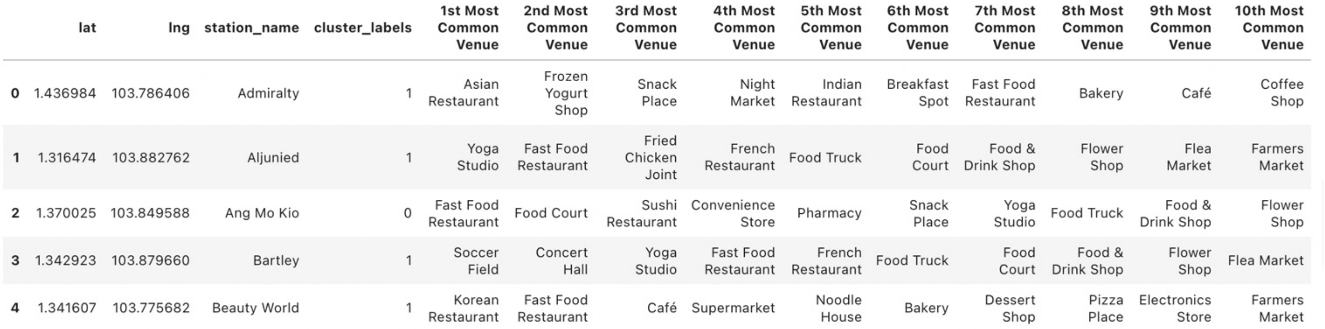


Figure 8. Elbow Method for K Determination

3.3.2. K-Means Clustering

K-means clustering (k=6) is executed utilizing SciKit-Learn library in Python (ibid.) taking the one-hot encoded dataframe as the input.

The cluster labels of train stations, together with the corresponding geographic coordinates of each train station, are added as attributes to the aforementioned dataframe of top 10 amenity types around each train station by merging, for cluster visualization purposes.



	lat	lng	station_name	cluster_labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	1.436984	103.786406	Admiralty	1	Asian Restaurant	Frozen Yogurt Shop	Snack Place	Night Market	Indian Restaurant	Breakfast Spot	Fast Food Restaurant	Bakery	Café	Coffee Shop
1	1.316474	103.882762	Aljunied	1	Yoga Studio	Fast Food Restaurant	Fried Chicken Joint	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Flower Shop	Flea Market	Farmers Market
2	1.370025	103.849588	Ang Mo Kio	0	Fast Food Restaurant	Food Court	Sushi Restaurant	Convenience Store	Pharmacy	Snack Place	Yoga Studio	Food Truck	Food & Drink Shop	Flower Shop
3	1.342923	103.879660	Bartley	1	Soccer Field	Concert Hall	Yoga Studio	Fast Food Restaurant	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Flower Shop	Flea Market
4	1.341607	103.775682	Beauty World	1	Korean Restaurant	Fast Food Restaurant	Café	Supermarket	Noodle House	Bakery	Dessert Shop	Pizza Place	Electronics Store	Farmers Market

Figure 9. Dataframe for Cluster Visualization

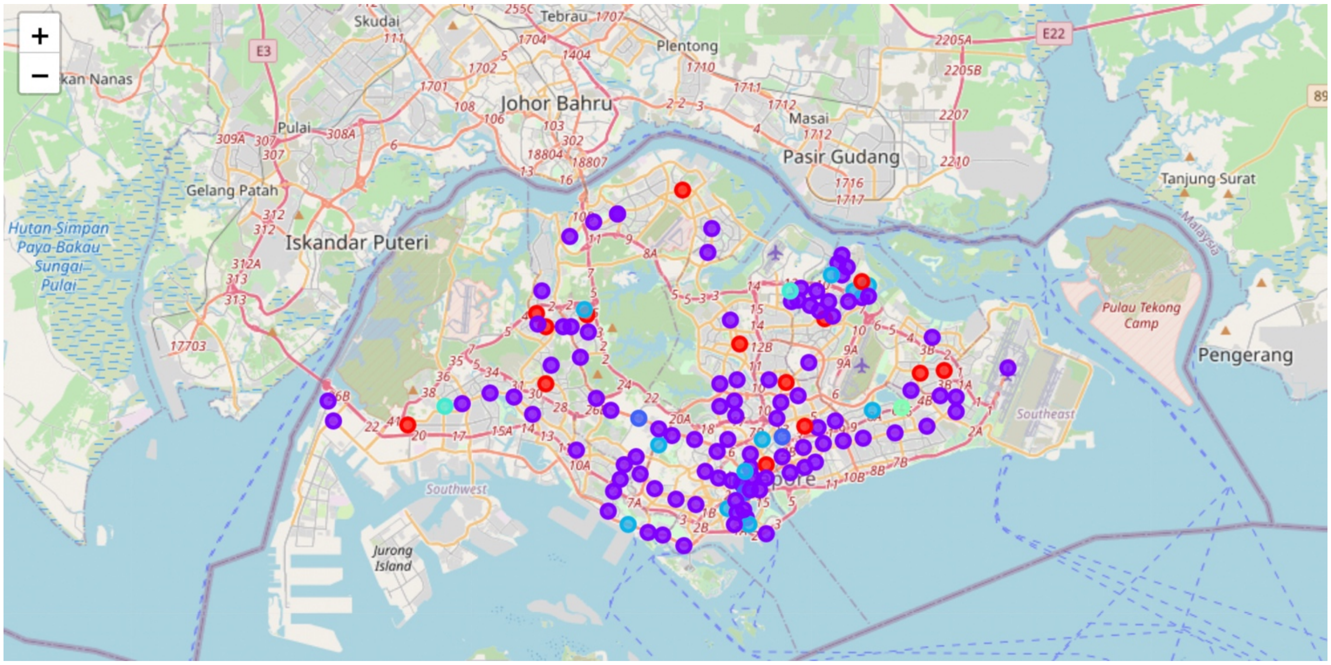


Figure 10. Map of Train Stations Clustered

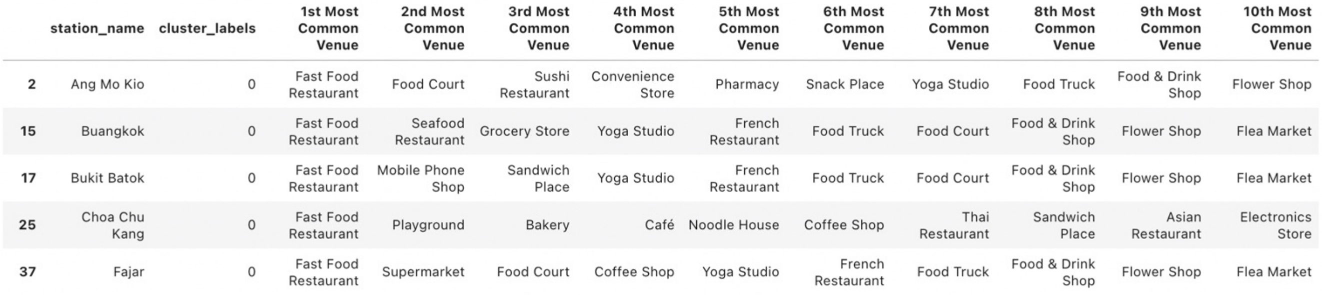
4. Results

Below is a summary of the aforementioned K-means clustering.

	Marker Color	Number of Member Datapoints
Cluster 0	red	14
Cluster 1	violet	125

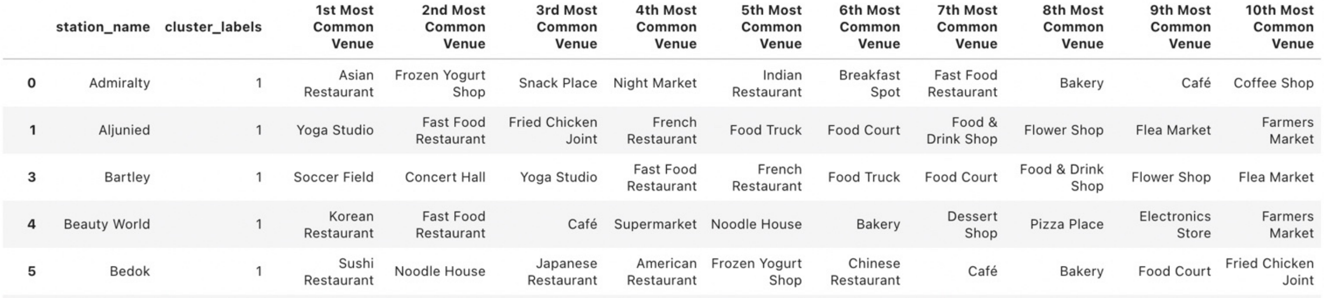
	Marker Color	Number of Member Datapoints
Cluster 2	indigo	4
Cluster 3	blue	11
Cluster 4	cyan	2
Cluster 5	green	1
Total		157

Table 1. Summary of Clustering



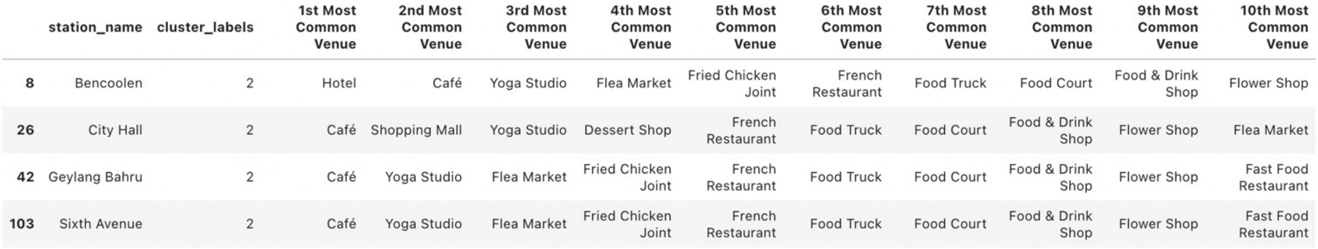
	station_name	cluster_labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	Ang Mo Kio	0	Fast Food Restaurant	Food Court	Sushi Restaurant	Convenience Store	Pharmacy	Snack Place	Yoga Studio	Food Truck	Food & Drink Shop	Flower Shop
15	Buangkok	0	Fast Food Restaurant	Seafood Restaurant	Grocery Store	Yoga Studio	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Flower Shop	Flea Market
17	Bukit Batok	0	Fast Food Restaurant	Mobile Phone Shop	Sandwich Place	Yoga Studio	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Flower Shop	Flea Market
25	Choa Chu Kang	0	Fast Food Restaurant	Playground	Bakery	Café	Noodle House	Coffee Shop	Thai Restaurant	Sandwich Place	Asian Restaurant	Electronics Store
37	Fajar	0	Fast Food Restaurant	Supermarket	Food Court	Coffee Shop	Yoga Studio	French Restaurant	Food Truck	Food & Drink Shop	Flower Shop	Flea Market

Figure 11. Cluster 0 (Excerpt of First 5 Members)



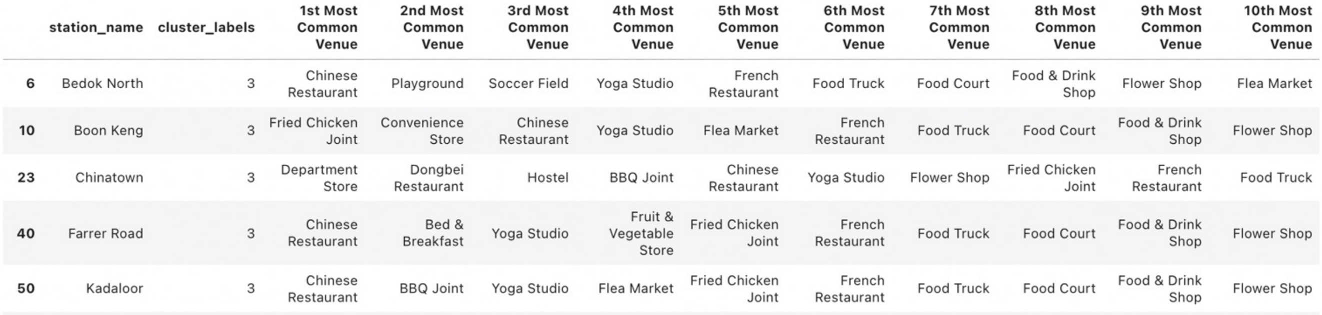
	station_name	cluster_labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Admiralty	1	Asian Restaurant	Frozen Yogurt Shop	Snack Place	Night Market	Indian Restaurant	Breakfast Spot	Fast Food Restaurant	Bakery	Café	Coffee Shop
1	Aljunied	1	Yoga Studio	Fast Food Restaurant	Fried Chicken Joint	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Flower Shop	Flea Market	Farmers Market
3	Bartley	1	Soccer Field	Concert Hall	Yoga Studio	Fast Food Restaurant	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Flower Shop	Flea Market
4	Beauty World	1	Korean Restaurant	Fast Food Restaurant	Café	Supermarket	Noodle House	Bakery	Dessert Shop	Pizza Place	Electronics Store	Farmers Market
5	Bedok	1	Sushi Restaurant	Noodle House	Japanese Restaurant	American Restaurant	Frozen Yogurt Shop	Chinese Restaurant	Café	Bakery	Food Court	Fried Chicken Joint

Figure 12. Cluster 1 (Excerpt of First 5 Members)



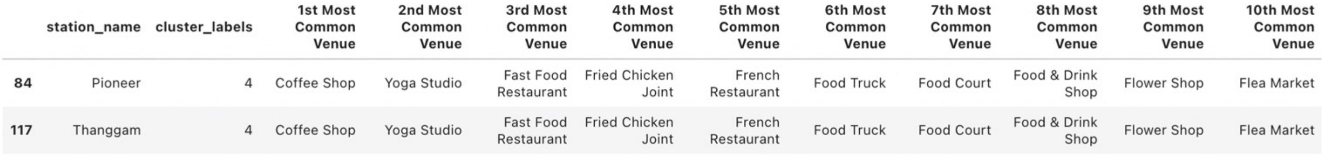
	station_name	cluster_labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
8	Bencoolen	2	Hotel	Café	Yoga Studio	Flea Market	Fried Chicken Joint	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Flower Shop
26	City Hall	2	Café	Shopping Mall	Yoga Studio	Dessert Shop	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Flower Shop	Flea Market
42	Geylang Bahru	2	Café	Yoga Studio	Flea Market	Fried Chicken Joint	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Flower Shop	Fast Food Restaurant
103	Sixth Avenue	2	Café	Yoga Studio	Flea Market	Fried Chicken Joint	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Flower Shop	Fast Food Restaurant

Figure 13. Cluster 2



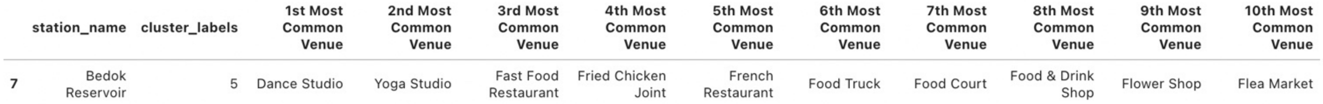
	station_name	cluster_labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
6	Bedok North	3	Chinese Restaurant	Playground	Soccer Field	Yoga Studio	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Flower Shop	Flea Market
10	Boon Keng	3	Fried Chicken Joint	Convenience Store	Chinese Restaurant	Yoga Studio	Flea Market	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Flower Shop
23	Chinatown	3	Department Store	Dongbei Restaurant	Hostel	BBQ Joint	Chinese Restaurant	Yoga Studio	Flower Shop	Fried Chicken Joint	French Restaurant	Food Truck
40	Farrer Road	3	Chinese Restaurant	Bed & Breakfast	Yoga Studio	Fruit & Vegetable Store	Fried Chicken Joint	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Flower Shop
50	Kadaloor	3	Chinese Restaurant	BBQ Joint	Yoga Studio	Flea Market	Fried Chicken Joint	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Flower Shop

Figure 14. Cluster 3 (Excerpt of First 5 Members)



	station_name	cluster_labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
84	Pioneer	4	Coffee Shop	Yoga Studio	Fast Food Restaurant	Fried Chicken Joint	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Flower Shop	Flea Market
117	Thanggam	4	Coffee Shop	Yoga Studio	Fast Food Restaurant	Fried Chicken Joint	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Flower Shop	Flea Market

Figure 15. Cluster 4



	station_name	cluster_labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
7	Bedok Reservoir	5	Dance Studio	Yoga Studio	Fast Food Restaurant	Fried Chicken Joint	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Flower Shop	Flea Market

Figure 16. Cluster 5

5. Discussion

5.1. Discussion about Output

From a descriptive point of view, below observations on the clusters can be made:

- The volume of one cluster can be very different from that of another cluster. The largest cluster (Cluster 1) has 125 member datapoints, whilst the smallest cluster (Cluster 5) only has 1 member datapoint.
- The majority of the amenities explored in the vicinities of train stations is in general related to cuisine. 'Food Court' and 'Food Truck' are the most often appeared amenity types.
- Out of non-cuisine related amenities, 'Yoga Studio' appears in high frequency.
- Cluster 1 has the greatest number of unique amenity types that do not appear in any other clusters, such as 'Frozen Yogurt Shop', 'Art Museum', 'Jazz Club', etc.
- Cluster 4's two member datapoints are sharing completely same top 10 most common venue types.
- Cluster 2's member datapoint are sharing at least 8 out of top 10 most common venue types.

5.2. Discussion about Model Defect

By its own design, the unsupervised K-means clustering algorithm randomly chooses initial mean points from the dataset, and repeats partitioning and centroid-finding steps until convergence is reached. This characteristic would affect the clustering results as different outputs may be produced each time the model is initiated. The output accuracy and the best value of K are also subject to changes.

6. Conclusion

The objective of the project is to explore the amenities around train stations in Singapore and group them based on their similarities regarding amenity types. In this study, we have loaded two public datasets, introduced multiple Python libraries to clean and wrangle the data, used Foursquare API for vicinity exploration, and clustered the datapoints.

From the output of the K-means clustering model, it can be told on the whole that the amenities around each train station in Singapore are homogeneous but also kaleidoscopic in terms of the most common venue types. It means a passenger can easily access various kinds of facilities from a station within walking distance, and also implies the public transport infrastructure in Singapore is well-designed and service-oriented.

Going further from this study, a potential storekeeper or restaurant owner can make use of the information in the project such as venue types and activeness scores to consider the feasibility of setting up a new place of business around certain train station or to evaluate the competitiveness of an existing asset in operation.

References

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

Foursquare. (2020). *Venue Recommendations*. Retrieved from Places API Reference: https://developer.foursquare.com/docs/api-reference/venues/explore/

GeoPy Contributors. (n.d.). *Calculating Distance*, Revision c1c5abf7. Retrieved from https://geopy.readthedocs.io/en/stable/#module-geopy.distance

GeoPy Contributors. (n.d.). *Units Conversion*, Revision c1c5abf7. Retrieved from https://geopy.readthedocs.io/en/stable/#module-geopy.units

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

Pandas Development Team. (2021, March 2). *pandas.read_csv*, 1.2.3. Retrieved from Pandas Documentation: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_csv.html#pandas.read_csv

SciKit-Learn Developers. (n.d.). *sklearn.cluster.KMeans*, 0.24.1. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html#sklearn.cluster.KMeans

Wikipedia. (2020, December 11). *Elbow method (clustering)*. Retrieved from https://en.wikipedia.org/wiki/Elbow_method_(clustering)

Wikipedia. (2021, January 25). *One-hot*. Retrieved from https://en.wikipedia.org/wiki/One-hot