# I. Objective

The purpose of this document is to describe the background of problem & data acquisition methods used for IBM data science capstone project.

# II. Introduction/Business Problem

Singapore is an island nation with most number of visitors at any given time. Considering a well-known tourist destination, Singapore already has a lot websites in place to recommend places to visit or stay. However, all these available websites focus mostly on usual tourist attractions which may not be of interest for backpackers. The purpose of this project is to,

- Choose one of the recently opened MRT lines in Singapore, Downtown line (DT) and analyze the most common venues located along the MRT line so that visitors can choose a place to stay near any DT MRT station and explore all the venues.
- The project also intends to cluster the MRT stations based on the venues so that the visitors can choose to spend more time on the stations with venues of their interest.
- The project not only provides the top most common venues but also provides a table of 10 most common venues

For the project, MRT station names are extracted from the Wikipedia page on Singapore MRT lines and common venues in the neighborhood of MRT stations in downtown line.

# III. Data Acquisition

This project will use the following data source: <br>
MRT station names and associated details are retrieved from the open data source in Wikipedia page on all the MRT line in Singapore. <br>
Website: https://en.wikipedia.org/wiki/List_of_Singapore_MRT_stations <br>

The original data source contains stations for all MRT lines with names in multiple languages, alpha numeric codes, interchanges and other details like opening date etc. For the purpose of this project, only the alpha numeric code and respective station names in English has been chosen. Then the data has been modified to choose locations only in downtown line. <br> <br>

For the MRT stations of interest, location data has been retrieved using geopy package. The top venues recommendations are retrieved from Foursquare API. FourSquare API to explore neighborhoods in selected towns in Singapore. The Foursquare API was also used to explore the most common venues along the downtown MRT line. <br>

# IV. Methodology

Data is extracted from the Singapore MRT stations Wikipedia page.
Below were the steps done in ***data pre-processing*** to retrieve the required information from the Wikipedia page:

- Remove the redundant columns in the data frame extracted from Wikipedia page. Include only the MRT station names & the corresponding alpha numeric code
- Filter the data to extract only the MRT stations in downtown line which if this project's interest
- Singapore MRT has multiple interchanges & many MRT stations in downtown line are repeated in the data frame since those stations fall under multiple MRT lines.
    - Cleanup the table based on alpha numeric code to ensure there were no duplicates in the table

After finalizing the data table ***use geopy package*** to pull the latitude & longitude details of all the MRT stations.

Once the data is ready for analysis, use the below methods for retrieving the required results:

- Use ***Foursquare API & explore option*** to retrieve the recommended venues nearby the MRT stations in downtown line
    - Initially analyze the venues near Cashew MRT station to understand the data
    - Expand the analysis to all the MRT stations in the downtown line
- Analyze the data to summarize the number of restaurants from different cuisines along the downtown line
- ***Use matplotlib bar chart and visualize*** the top 20 common venue categories along downtown line and the frequency of occurrence.
- Analyze each neighborhood using ***onehot encoding***
- Use ***KMeans clustering*** to study the similarity of locations in downtown line & cluster them
- Study the clusters & understand the results.

# V. Results / Discussion

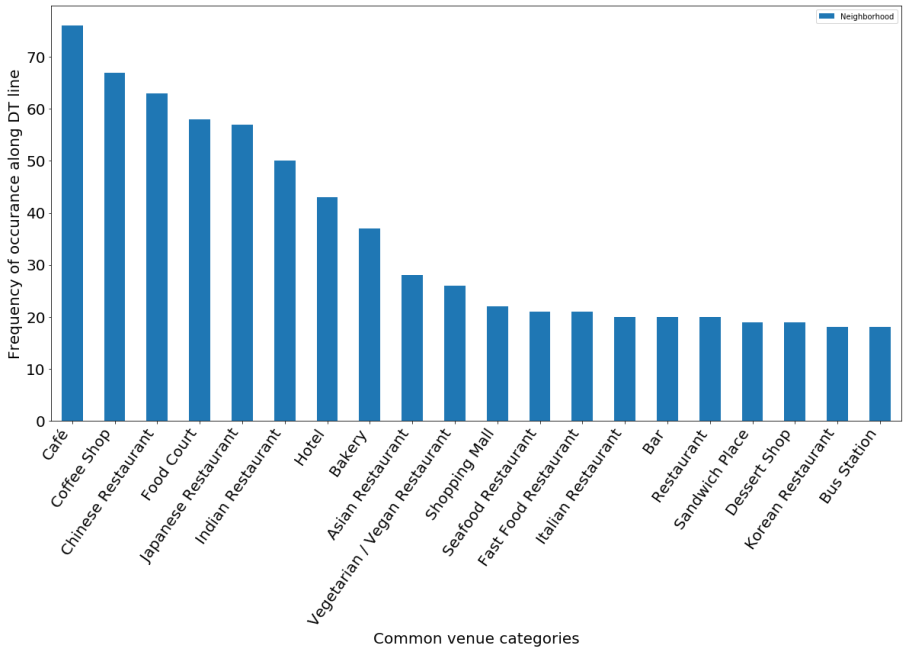
This section discusses about the results and observations from the above capstone project.

- Venues near Cashew MRT station:
    - Initial analysis of venues around Cashew MRT station showed most frequently occurring categories are bus station, cafes, seafood & Italian restaurants. Following this, the analysis was extended to all MRT stations.

- Analysis for all MRT stations in neighborhood:
    - Below are results of number of restaurants from different cuisines along the downtown line
        - `The number of Indian restaurants in downtown line are 50`
        - `The number of Seafood restaurants in downtown line are 21`
        - `The number of Chinese restaurants in downtown line are 63`
        - `The number of Italian restaurants in downtown line are 20`
    - Based on onehot encoding, each MRT station was analyzed for most common venues

| | Neighborhood | 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 | Bayfront | Hotel | Garden | Boutique | Scenic Lookout | Lounge | Bar | Casino | Roof Deck | Waterfront | Japanese Restaurant |
| 1 | Beauty World | Café | Korean Restaurant | Chinese Restaurant | Food Court | Shopping Mall | Asian Restaurant | Thai Restaurant | Bus Line | Noodle House | Massage Studio |
| 2 | Bedok North | Coffee Shop | Asian Restaurant | Fast Food Restaurant | Bakery | Other Great Outdoors | Steakhouse | Supermarket | Bus Station | Park | Food Court |
| 3 | Bedok Reservoir | Reservoir | Food Court | Scenic Lookout | Bus Stop | Yoga Studio | Event Space | Furniture / Home Store | Frozen Yogurt Shop | Fried Chicken Joint | French Restaurant |
| 4 | Bencoolen | Café | Hotel | Japanese Restaurant | Coffee Shop | Chinese Restaurant | Theater | Sports Bar | Restaurant | Bookstore | Gaming Cafe |

- Visualization of top 20 common venues along DT line




- KMeans Clustering
  - The KMeans clustering was done to segregate the downtown line into 7 clusters based on the similarity. The results using KMeans clustering showed most stations fell under cluster 0 due to the availability of restaurants with multiple cuisines & cafes as a similarity between them.
  - Visitors who would like to explore multiple international cuisines like American, Japanese, Korean, Italian etc. during their visit should choose to stay near any of the MRT stations in cluster 0
  - Visitors like backpackers who prefer food joints with economic rates can choose cluster 1 which has food courts as the most common venue. Food courts in Singapore always come with cheaper food options
  - Visitors who prefer vegan restaurants or Indian cuisine are better to stay at 'Little India' MRT station which is in cluster 4.
  - Visitors who would like to explore nature & trails are advised to choose cluster 3 which is near Bedok reservoir

The clustering has shown the similarities between neighborhoods along downtown line & will be helpful for visitors who would like to stay along downtown line which has access to Singapore Island wide.