- Introduction where you discuss the business problem and who would be interested in this project.

  (Hypothetical situation) My friend Mac owns some shops for rent in Hong Kong. Recently, one of tenants planned to move. After he did some researches, he has a plan to invite one of the below 2 chains to move into his shop (at To Kwa Wan, Kowloon, Hong Kong), and he is sure both of them are interested to expend their business:

  1. McDonxxx's ("the M" for short): a American fast food restaurant chain serves over 69 million customers daily in over 100 countries across approximately 40,000 outlets. And there are around 200 outlets in Hong Kong
  2. Starbxxxs ("the S" for short): a coffeehouse chain found in Seattle, operating around 28,000 coffee shops worldwide, including around 150 coffee shops in Hong Kong.

  But, my friend Mac is not sure if his place fits the M's shop or the S's shop, and he do not know if there are any data can answer the question. So, he came to me for help. We did some discussions and finally we think:

  We can use data analytics to analysis the nearby venues of the two chains in Hong Kong. And see if the nearby venues of my friend's place are simular to the M's or the S's. So we can try to use the mechine learning Classification method to classify Mac's shop into Class M or Class S according to the all nearby venues data of the existing known M's and S's shops.

  The steps of the data analysis will be:

  1. Collecting and understanding the data
  2. Perparing the data
  3. Modeling and evalution
  4. Deployment

- Data where you describe the data that will be used to solve the problem and the source of the data.

1. Things before introducing the data

Collecting data and understand the data

In the Introduction, I think I can use "Classification method" to analysis the nearby venues of the M & S current shops, build a model and then predict which class should Mac's place be classified. So, this should be simular to the Week 3 project: both use nearby venues to analysis. But there are few different things:

1. As we will provide the classification results to the model and train the result

(supervide learning), our problem is a classification problem and Week 3 project is a clustering problem. So we will build the model, train the result using the data of M & S shops, and predict the result of the targeted place.

2. Instead of analysis the venues near each neighborhoods, this project will analyse the venues near M, S and our targeted shops.

About the ipynb files I uploaded

The notebook files use Python 3 kernel to run. Some notebooks need "folium", "BeautifulSoup" and "geopy" libraries. These libraries may not be included into the default Jupyter notebook, and therefore, pre-instsalling them may be needed. Also, if you view the file in Github, I found that the Folium maps in my notebooks will not shown here. If you want to see the maps (although the maps are only for me to analysis the data), you may need to download all notebook files and run them all in your Jupyter notebook in the sequence of:

mtr.ipynb --> M_S_location.ipynb --> Combined.ipynb --> Classification-part1.ipynb

Each notebook will generate one or two excel files that will be used in the next notebook.

2. Location data of M and S shops

At first, I want to generate locations from their company websites, and they should list out all the shops in Hong Kong. However, both the websites do not provide lists of their location and we can imported. And I cannot found any lists in the internet.

Then, I plan to use Foursquare Search API and think that "Search for venue" API can be used to get the shops locations. After I do a test search, there are some limitation:

1. If there are less than 50 results that are exact matched per API response, there may be some results that are not exact match. it means that the API may return some results that are not related to the M or S shops.
2. Although maximum "radius" can be set largh enough (100,000 meters), the maximum results to return for my Foursquare account is only 50 results per search. As mentioned in the Introduction, there are around 200 M shops and 150 S Shops in Hong Kong. We need to run the search more than at least 4 times and using different latitudes and longitudes to get more results.

The first issue can be solved by adding a name check step before we add the data into our lists. For the second issue, thank to MTR, one of the most mature public transport network in the world, I can use the MTR station location data to search M & S shops nearby. Although All M & S shops in the list cannot be guaranteed, I still can get most of the results in the urban area.

Belows are the related data files (in PANDAS dataframe format and exported as Excel files) and the related Jupyter nootebook file:

1. MTR station location data *(mtr.xlsx -- the exported file of the dataframe of mtr*

stations locations, mtr.ipynb -- the notebook file to generate dataframe from the a website): There is a table in the website [https://www.exploremetro.com/blog/hong-kong-mtr-station-names-in-cantonese-jyutping/]. And the table has English names, Chinese names and the Cantonese Pinyins of the stations. We can get all the stations names, change them to addresses (just simply add ", Hong Kong" at the end to become an address), and use GeoPy library to convert the addresses into geographical coordinates. Then we create a dataframe with columns: Name, Latitude & Longitude and export it to an excel file for future use.

2. M shops and S shops location data *(m_shops.xlsx -- the exported file of the dataframe of M shops locations, s_shops.xlsx -- the exported file of the dataframe of S shops locations, M_S_location.ipynb -- the file to generate the dataframes of M shops and S shops locations)*: use each MTR station location data, to generate M shops and S shops data around 1000 meters of each station using Foursquare API, then remove the dupicate shops with the same ID. The dataframes include columns: the name, latitude data and longitude data, ...
3. The list combined M shops and S shops *(full_list.xlsx -- the exported file of the dataframe of M shops and S shops locations, Combined.ipynb -- the file to generate the dataframe)*: When I try to combine the above 2 dataframes into a dataframe, I found that there are many places has both M shop and S shop. So if a M shop and a S shop are within 50 meter, I will add the point in the middle of the 2 points into the full list, and these 2 points will be not added into the list. As a result, the full list will have 3 types of locations data: type 0 -- location data of M shops, type 1 -- location data of S shops and type 2 -- location data of the place that has both 2 shops.

3. Nearby venues data of M and S shops

Once I got the location data of M & S shops, the following step is almost the same as the Week 3 project: try to get the venues data near each M or S shop by sending request to Foursquare explore venues API, and then convert json file of the response into PANDAS dataframe. M shops and S shops venues data in PANDAS dataframe format and exported as Excel files: *(venues.xlsx -- the exported file of the dataframe of the nearby venues of all shops. Classification-part1.ipynb -- to generate the dataframe)*: the dataframe includes the infomation columns of the venues (such as name, location, category of the venues), the column of the shop location, the type column (to indicate that the place is type 0: M shop, type 1: S shop, or type 2: the place has both 2 shops).

4. Location data and Nearby venues data of our targeted place (Mac's shop at

To Kwa Wan, Kowloon, Hong Kong)

As I have the full address of the shop, the location data is easy to get by using Python GeoPy library. And the nearby venues data can be easily generate from Foursquare API. The location data and the venues data will be generated next week when I build model.

- Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, and what machine learnings were used and why.

- Results section where you discuss the results.

- Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.

- Conclusion section where you conclude the report.