1. DATA ACQUISITION AND CLEANING

1.1. DATA SOURCES

All the metropolitan restaurants data were brought in through Foursquare API and it is set into a Data frame. To analyse the restaurants data I need council areas of Victoria which I found them here.

1.2. DATA CLEANING

There was a lot to do with cleaning the data.!

As I've chosen to analyse the restaurants of Melbourne Metropolitan region, there is no way, I could collect the data from Foursquare API in a single request. So, I grabbed the data by sending request for every suburb and merged them into a data frame. Depending on the radius settings, when extracting data from Foursquare there is a chance that a venue might appear multiple times. In order to overcome this issue, I decided to drop all the duplicates based on the restaurant ID.

When I saw the insights of my data frame, I found many properties in one of the features, '**categories'**. I made each property into a new feature and dropped some of the unnecessary features.

Next, I tried visualizing top '5' cuisines using pie chart. Then I learnt, more than 15% of the restaurants did not mention the cuisines they serve. So, I tried to identify the cuisines based on the restaurant names, for example the name of a restaurant includes Chinese but their cuisine is given as 'Restaurant', then I modified it to 'Chinese Cuisine'.

Similarly, I observed some of the 'Bars & Restaurants' were mentioned as 'Australian Cuisines' where there is no 'Australian Cuisine'. Based on their names I modified them.

After fixing all these problems, I checked my council area data as it was in geojson format, I changed it into a data frame and later on merged it with restaurants data.

1.3. FEATURE SELECTION

After data wrangling, there were 1,157 samples and 15 features in the data. While examining each feature, some redundancy was found. There are two features **'lat'** and **'lng'** contained the same information as **'labeledLatLngs'** which makes the data frame a little clumsy. So, I dropped the feature, **'labeledLatLngs'.**

From the merged data frame, I've chosen features, **'categories', 'lat', 'lng', 'postalCode', 'Councilarea'** for the future analysis.