**Coursera- Applied Data Science Capstone Week 4 and 5**

**Background**

In weeks 1 til 3 of the Applied data Science Capstone we learned how to use data in a real world scenario’s (New York and Toronto) using multiple data sources including Foursquare data obtained using the Foursquare API and scraping Wiki data. We used the Folium library and the K-Means Clustering algorithm to analyze and visualize the data/clusters.

For week 4 and 5 we will continue by defining a business case and apply the obtained knowledge and skills to obtain the answer to the defined business question.

In the defined scenario we are looking for the optimal location for the opening of a new Vegetarian/Vega restaurant based on the defined strategy which includes objectives, constraints and policy highlights. One of the main requirements of this hypothetical company “Company A” is that decisions are made “Data Driven” i.e. requiring support of Analytics. Consequently it is important that the conclusion of the analyses aligns with the business question, in other words the identification of a cluster may not be sufficient/detailed enough.

**Business Case**

Following a recent re-evaluation of their business strategy Company A has adopted an aggressive strategy with the opening of new locations. The following elements are part of this new strategy:

* Following ongoing changes in the customers preferences in relation to types of food/restaurant types i.e. more and more customers preferring healthy food (less meat, less calories, less fat) we want to locate each new restaurant to so called “American Restaurants”, preferably in the location of multiple of these restaurants.
* The Company adopted the concept of Data Driven Decision making i.e. it requires data analytics to support the decision making process related to the fact that the funds are limited.
* The focus is on the city of Toronto.

**Data**

1. General

* Toronto District and Venue data, especially data in relation to the location of American Restaurants.
* Latitude and Longitude data of each district

1. Data Sources

* Toronto Neighbourhood Data <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)>
* Geospatial Data (Latitude, Longitude), intended to be obtained from Geocoder package. As this technically did not work during the execution of the project the alternative data source (‘GeoSpatial\_Coordinates.csv’) was used.
* Venue data of the Toronto districts, via Foursquare API.

**Approach and Methodology**

1. **Python packages used**

* Pandas
* NumPy
* Requests
* Json
* Matplotlib
* Folium
* Sklearn

1. **Data exploration and cleaning activities**

**Toronto Neighbourhood Data**

Wikipedia dataset contains 3 columns: “PostCode”, “Borough” and “Neighbourhood”

* Transformation of the data into a DataFrame format.
* The “Borough” columns has various “Not assigned” (null values): excluded
* Creation unique PostCode/Borough combinations i.e. remove duplicates
* Identification of “Not assigned” (null values) in the “Neighbourhood” column and replacement with the value in the “Borough” column (1 row/value)

**Geospatial Data**

* Transformation of the data into a DataFrame format.
* No other cleaning activities necessary.

**Merging of Toronto Neighbourhood Data and Geospatial Data/Additional filtering/viz**

* Latitude and Longitude data is added to the Toronto Neighbourhood Data by merging with the Geospatial Data, using the Postal code field as key.
* Due to limitations of the Foursquare sandbox account we filter the resulting dataset using only “Borough” with “Toronto” in its name (i.e. West, East, Downtown and Central)
* To obtain additional insights the resulting dataframe(using the Latitude, Longitude, Borough and Neighbourhood columns) is displayed on a Folium map of Toronto

**Foursquare data**

* Using the foursquare API we obtain all venues within a range of 500 meter of the neighborhoods locations (latitude, longitude)
* Transformation of the data into a DataFrame format.
* Exploration of the size of the obtained data including obtaining the size of the resulting DataFrame, the number of unique Venue Categories overall and the number of Venue categories which involve Restaurants only **(I)**
* To prepare the data for effective use by the KMeans clustering algorithm we use one-hot-encoding function basically transforming the data to a frame where a separate, binary, column is created for each Venue Category, displaying either a 0 or 1 (resulting dataframe consists of 231 columns) and the relative frequency of each venue category in each neighborhood is created.
* From the resulting dataframe we obtain the Neighbourhood column and the relative frequency of American Restaurants as input for the clustering algorithm.

1. **Selection of machine learning type/supervised or unsupervised learning**

Evaluating the objective/strategy stated in the business case, the availability of data and the fact that we do not have labelled data available we opt for a clustering (unsupervised) algorithm, using 4 clusters. In general the clustering algorithm determines (in this case) 4 centroids and allocates each data point to one of the (4) clusters by i) minimizing the distance of each datapoint vs the cluster centroid and ii) maximizing the distance between the 4 centroids of the different clusters iii) using multiple tries thereby moving the cluster centroids to obtain the optimal solution.

The input data for the clustering algorithm is a dataset containing the Neighbourhood column and the relative frequency of American restaurants in each Neighbourhood.

1. **Next steps**

Next we add(merge) the clustering labels and the “American Restaurant” column (reflecting the relative frequency of occurring in each neighbourhood) to the dataframe created in the step indicated above with **(I)** and explore the resulting dataframe (size, datatypes) for the creation of the folium map, visualizing neighbourhoods and cluster label.

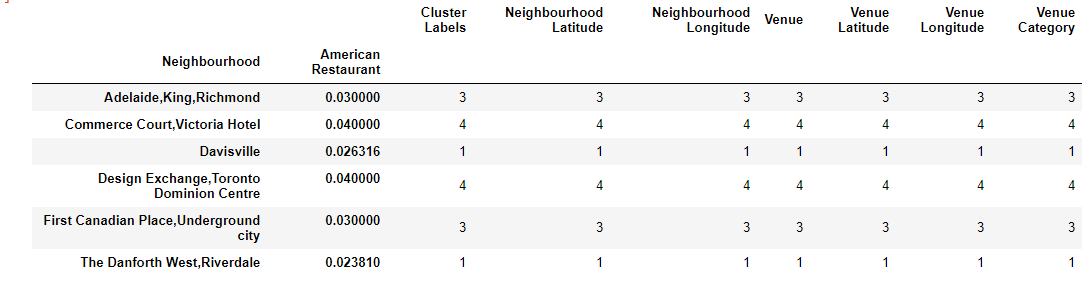
1. **Further analyses of the clusters to answer the business question and conclusion**

To be able to answer the business question we explore each cluster, identifying

* Cluster 0 contains 4 American Restaurants
* Cluster 1 contains 0 American Restaurants
* Cluster 2 contains 16 American Restaurants
* Cluster 3 contains 6 American Restaurants

As Cluster 2 is superior in terms of number of American Restaurants versus the other clusters we will further explore this cluster only to be able to answer the business question (note: the identification of the superior cluster is not sufficient to answer the business question as there are still 6 neighbourhoods in the cluster while we are looking for best location to open 1 restaurant. In other words the clustering data needs to be broken down using the number of American Restaurants (count) and relative frequency of the American Restaurants in each Neighbourhood.

Below screenshot shows the resulting DataFrame:



**Conclusion**

To answer the question where to locate a new Vegetarian/Vega restaurants we used 3 different data sources to identify, using a clustering algorithm, to find the optimal location I.e. a neighborhood with the most American restaurants

Evaluating the different clusters we concluded that there was 1 superior cluster containing 16 of the overall 26 American restaurants in Toronto. As the superior cluster contained 6 different neighbourhoods we however did not have the answer for the optimal location of 1 Vegetarian/Vega restaurant yet. To be able to obtain the answer to the business question we executing 1 further steps which resulted in above dataframe. As both the number of American Restaurants and the relative frequency are the same for the 2 neighbourhoods “Commerce Court, Victoria Hotel” and “Design Exchange, Toronto Dominion Centre” and we do not have additional feature available to further analyze/evaluate we can only conclude that 1 of these neighbourhoods is the optimal solution based on the criteria identified in the Business Case.

**Further thoughts/discussion**

To be able to have a suitable and usable case(limited by for instance process time) for the Capstone Project the business case was simplified and does not reflect a real world scenario. The evaluation only on number of American Restaurants is not realistic in practice but only used as ficticious scenario to apply the skills and knowledge obtained in this course. Also the limitation of the Foursquare sandbox account in terms of maximum number of rows and features to be extracted required simplifications. In real word scenario’s there will be more criteria to be evaluated/features to be extracted for Data Driven Decision making to be effective. Examples of additional features are real estate prices, number of residents per neighbourhood, additional demographics like average age, average income , restaurant turnovers and additional location features.