Recommendation for Business Opportunities in Toronto

1. Data Acquisition and Preprocessing

In the week 3 of the capstone project, we already used Foursquare data, also the neighborhood data and the geospatial data, to explore and cluster the neighborhoods in Toronto, based on the different venue categories.

In the scope of this project we will reuse the same data, but we will make minor changes in preprocessing and apply a different approach after preprocessing.

1.1. Loading borough information and geospatial data for Toronto

List of postal codes of Canada:

The postal codes, boroughs and related neighborhoods of Toronto will be loaded from the following Wikipedia link:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

This data will be transferred into the pandas data frame "toronto_data", which will have the three columns "Postal Code", "Borough" and "Neighborhood".

Geospatial data:

The geospatial data for the postal codes of Toronto will be loaded from the following link, as also requested in the week 3 of the capstone project:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

This data will be transferred to the pandas data frame "df_geo", which will have the three columns "Postal Code", "Latitude" and "Longitude" for all the postal codes of Toronto.

Merging data frames:

After this step, the geospatial data will be merged with "toronto_data" imported in the first step, using Postal Code as a common attribute. The merged data frame will contain latitude and longitude data for all the neighborhoods of Toronto.

1.2. Loading Foursquare data for venues

The "explore" endpoint of Foursquare will be used, to capture all the venues including id, name, location, category id and category information for every neighborhood (analogue to the week 3 assignment in capstone project). This data will be saved in the pandas data frame "toronto_venues".

In the first step of preprocessing, this data frame will be grouped by the three attributes "Neighborhood", "Category ID" and "Venue Category". The rest of the attributes will be counted. The attribute "Venue ID" will be renamed as "Frequency", which will represent

the number of the venues in every category and every neighborhood. This grouped data frame will be named as “venues_by_category”.

1.3. Preparation of input venues

For the selected neighborhood “Willowsdale East”, an input data frame will be created. For this purpose, the data records from the data frame “venues_by_category” will be taken over to a new data frame “inputNeigh”, where the neighborhood is “Willowsdale East” in the data record.

In the second step, the columns „Neighborhood“, “Neighborhood Latitude“, “Neighborhood Longitude“, “Venue“, “Venue Latitude“, “Venue Longitude“ will be dropped from inputNeigh. At the end, only the three columns “Category ID”, “Venue Category” and “Frequency” will remain in the data frame.

This input data frame represents all the venue categories and their frequencies (number venues in this category) for the selected neighborhood “Willowsdale East”.

1.4. Preparation of venue subset

A venue subset will be created by taking over all records from the data frame “venues_by_category” to a new data frame, considering the condition that the category is included also in inputNeigh. This new data frame will be named as “venueSubset” and will include only the categories involved in inputNeigh.

In the second step, the columns “Neighborhood Latitude“, “Neighborhood Longitude“, “Venue“, “Venue Latitude“, “Venue Longitude“ will be dropped from venueSubset. As result, only the four columns “Neighborhood”, “Category ID”, “Venue Category” and “Frequency” will remain in the data frame.

In the third step, the data frame “venueSubset” will be grouped by “Nieghborhood”, so that it can be used easily in the for-loop in the next step, to calculate the Pearson coefficient for the neighborhoods. This grouped data frame will be named as “venueSubsetGroup”.