**Final Capstone Project report part2**

**PRACHI SINGLA**

**DATE: 16-04-2019**

1. **Data Acquisition and Cleaning**
   1. **Data sources**

I have extracted data of neighbourhoods of Canada from here <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>. Then after reading this data into pandas data frame, I have inserted geographical coordinates of the neighbourhoods, using the Geocoder package, here is a link to a csv file that has the geographical coordinates of each postal code: [http://cocl.us/Geospatial\_data](https://cocl.us/Geospatial_data).

Then I used foursquare API to get all venues in neighbourhoods of Toronto i.e. https://api.foursquare.com/v2/venues/explore?

After extraction of all required data, by using useful analytical techniques, I will analyse and give insights which will be helpful to many in different prospective.

* 1. **Data cleaning**

Data (neighbourhood data) downloaded or scraped from source as mentioned above and transform all neighbourhood information into pandas dataframe after importing all important libraries needed for project.

It’s time to explore & clean data.

**Firstly**, explore dataset and extract shape of dataset along with all columns names and their number which comes out to be as follows:

|  |  |
| --- | --- |
| Dataset shape | (288,3) |
| Postal codes | 103 |
| Boroughs | 11 |
| Neighbourhoods | 209 |
| Rows with Not assigned Borough | 77 |
| Rows with Not assigned Neighbourhood | 78 |
| Rows with Not assigned Neighbourhood and Borough | 77 |

But the problem is this that it includes not assigned values, which is not helpful at all and we need to clean data (either removed or replaced these not assigned values).

So **Secondly**, Eliminate null values (not assigned values) from Boroughs & replacing null values (not assigned values) with Borough names in neighbourhood column. Then get new dataset shape with all column names with nos. & types which comes out to be as follows:

|  |  |
| --- | --- |
| Dataset shape | (211,3) |
| Postal codes | 103 |
| Boroughs | 11 |
| Neighbourhoods | 209 |

Now data has been explored & cleaned accordingly for further processing.

Data frame before cleaning



Data frame after cleaning

