2. Data Acquisition and Preparation

2.1 Data Acquisition: Since the area of study is Toronto and neighbourhoods, to get the neighbourhood details, use Wikipedia as the source. Downloaded the data in html format using the link Toronto postal Codes. But this data doesn't include geographical coordinates. To get the geographical coordinates corresponding to each neighborhood, used the facility in website (Toronto neighborhood Geospatial_data). This is a csv format file. This study requires the details of schools and Tennis facilities in each neighborhood to identify the most required locations for new Tennis facility. I am using Foursquare data (Foursquare) for venues to get current details of schools and Tennis facilities in neighborhoods in North York borough (Toronto).

2.2 Data cleaning: Since the source of data for this project is from different sources and in different format, to build a clean data frame for further analysis, different data cleaning activities have to be done. The data scraped from Wikipedia for data related to Toronto neighborhood, contains neighborhood with not assigned boroughs.

Since our study is about the neighborhoods in North York only, neighborhoods without borough assigned were dropped to confirm the selected neighborhoods belong to North York itself. Since the geographical coordinates are towards postal codes, to avoid duplications in Geographical coordinates and postal codes in data, combined neighborhoods with same postal codes and made postal codes unique.

The venues list obtained from foursquare for schools include some other offices like school board and those ones which are not educational institutions were dropped. In Tennis venues list obtained from Foursquare, include shops which purchase Tennis goods also were removed from data.

2.3 Data preparation/ Feature Selection: Since the problem is to find out the low Tennis facilitated schools in neighboring areas, after validating various parameters, considered schools and Tennis facilities in 500M from each neighborhood center. Distance from neighborhood center is important feature in this study. But distance parameter in venues data of Tennis facilities and schools, obtained from Foursquare, is continuous. For future analysis based on distance, the continuous variable distance is binned to put it under 4 categories based on distance range. Another categorization did was based on type of school. Various school types categorized into two groups - 'High school and below' and 'Above High School' which include all colleges, universities and other institutions. After completing data

cleaning and feature selection, conducted various exploratory and descriptive statistics and visualization to learn about the overall quality of data and to get a preliminary idea about the relationship between selected features.