2. Data acquisition and cleaning

2.1 Data sources

The data required for this analysis where extracted from the following link: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. Beautiful soup library was also one of the tools that helped data manipulation. After the page was downloaded as an html file, the first table (containing the related data) was accessed (through an iteration) and converted to a pandas data frame. First five (5) rows displayed below.

[12]:

	PostalCode	Borough	Neighborhood
0	M1A\n	Not assigned\n	Not assigned\n
1	M2A\n	Not assigned\n	Not assigned\n
2	M3A\n	North York\n	Parkwoods
3	M4A\n	North York\n	Victoria Village
4	M5A\n	Downtown Toronto\n	Regent Park, Harbourfront

After having the data frame, a data cleaning process was necessary. The "\n" suffix was removed and any missing values were also excluded. The result was a (103,3) shaped data frame with 11 boroughs and 103 neighborhoods.
First five (5) rows of the data frame:

[14]:

	PostalCode	Borough	Neighborhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
5	M6A	North York	Lawrence Manor, Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

The third stage was the location data addition. This was achieved by using a geolocator function and creating two (2) new columns in the data frame in which the latitudes and longitudes were stored. Only one (1) missing value found and was excluded. First five rows of the updated data frame:

	PostalCode	Borough	Neighbourhood	latitude	longitude
2	M3A	North York	Parkwoods	43.7545	-79.3300
3	M4A	North York	Victoria Village	43.7276	-79.3148
4	M5A	Downtown Toronto	Regent Park, Harbourfront	43.6555	-79.3626
5	M6A	North York	Lawrence Manor, Lawrence Heights	43.7223	-79.4504

Most player stats, position, age, and draft position data can be found in two Kaggle datasets here and here. These two datasets, however, lack data for certain years. For example, the player stats dataset ends in 2017, and the player draft dataset starts in 1978 and ends in 2015. To complement these two datasets, I scraped basketball-reference.com for player season stats of 2018 and player draft positions of 1965-1977 and 2016-2017 (players drafted in 2018 has yet to play in NBA). 2.2 Data cleaning Data downloaded or scraped from multiple sources were combined into one table. There were a lot of missing values from earlier seasons, because of lack of record keeping. I decided to only use data from 1980 season and after, because of later seasons have fewer missing values and basketball was a lot different in the early years from today's game. There are several problems with the datasets. First, players were identified by their names. However, there were different players with the same names, which cause their data to mix with each other's. Though it was possible to separate some of them based on the years, teams, and positions they played, I decided that it was not worth the large effort to do so, because such players only accounted for ~1% of the data. Therefore, players with duplicate names were removed. Second, multiple entries existed for players who changed teams mid-season. This cause their seasonal data to represent multiple samples with incomplete data. I wrote script to extract total season stats for these players, and discarded partial season rows. Third, there were two short seasons in recent NBA history, during which less than the normal 82 games were played. This has caused stats in those seasons to be artificially smaller than other seasons. To correct that, I normalized

cumulative features such as points, rebounds, etc. as if 82 games were played. After fixing these problems, I checked for outliers in the data. I found there were some extreme outliers, mostly caused by some types of small sample size problem. For example, some players had only played a few games or a few minutes the entire season, and had performed extremely well or poor in those minutes. Therefore, seasons during which less than 20 games or 100 minutes were played were dropped from the dataset. Similarly, there were players who only took one 3-point shot, but made it, therefore had 100% shot accuracy. I changed the shot accuracies for players who shot less than 10 shots to missing values. There were 4 features which had missing values. Games started were imputed from minutes played because starters usually play more minutes. Missing 3-point accuracies were imputed with a very small value (0.05) because if a player rarely shoots 3s, it is probably because he is not very good at it. Missing free throw accuracies were imputed using the mean of all players. Missing draft positions, meaning undrafted, were imputed using position 61 (the position after the last position in the draft, 60th).