**Introduction**

The retail industry is one of the business fields that can easily and effectively be evolved with the aid of data science and analytics. Information about customer needs, historical sales figures, competitor and market analysis and other relevant data can assist in the development of a certain business, by allowing investors to take optimal decisions. One of the specific retails sectors that is of great interest to the author is the sports and fitness field. There is an extremely wide range of sports and activities that are being exercised daily by numerous people.

**Business problem/ Shareholder Interest**

The current social distancing measures worldwide have maybe urged a wider audience to actively engage with sports exercised by less than two people at a time. At the same time, it might be expected that after the lockdown measures are lifted, there will be a massive interest in participating in group sports again, or simply participate in athletic activities using shared space and equipment, such as gyms, pools and basketball courts. For these reasons, investing in supplying sports equipment appears like a reasonable consideration. This report aims to develop a statistical methodology of deciding in which neighborhood of New York it would be best for a group of shareholders to invest in a brand-new sports equipment retail store.

**Data**

In order to analyze the various factors that affect the outcome of a hypothetical investment, Foursquare location data will be utilized in order to compare between neighborhoods. More specifically, for reasons which will become clear in the methodology section, the following attributes of a neighborhood will suffice for our analysis:

* Total number of sports venues: This is the sum of all enlisted sports and fitness facilities in which people can exercise. The Foursquare database included a variety of venues, such as basketball, baseball and badminton courts
* Neighborhood Median Salary: The median salary for each NYC neighborhood
* Total population for each neighborhood: The population of each neighborhood

The total number of sports facilities can be found by utilising the Foursquare API, while the total population of each neighborhood an be found on the NYC OpenData website: <https://data.cityofnewyork.us/City-Government/New-York-City-Population-By-Neighborhood-Tabulatio/swpk-hqdp>. The median salary data were retrieved from <https://ny.curbed.com/2017/8/4/16099252/new-york-neighborhood-affordability>.

Further data that were used throughout the analysis included location data for the sports venues, such as latitude and longitude, as well as venue ID. All the relevant information is available by Foursquare.

**Methodology**

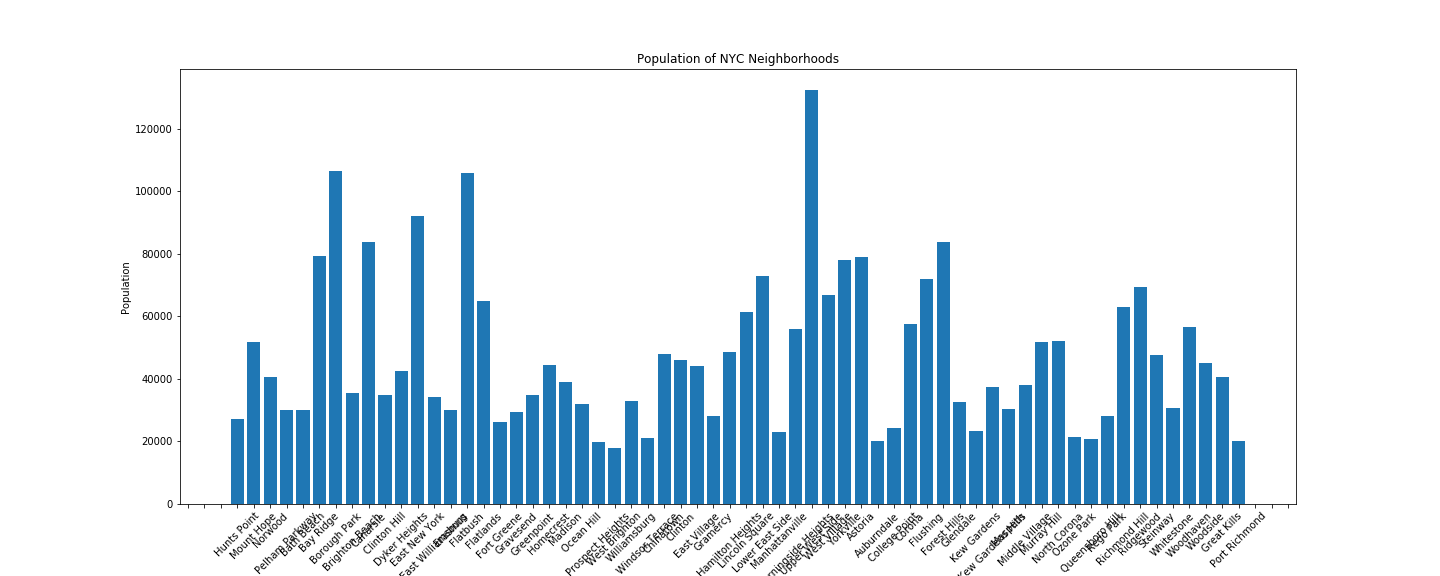
**Overview**

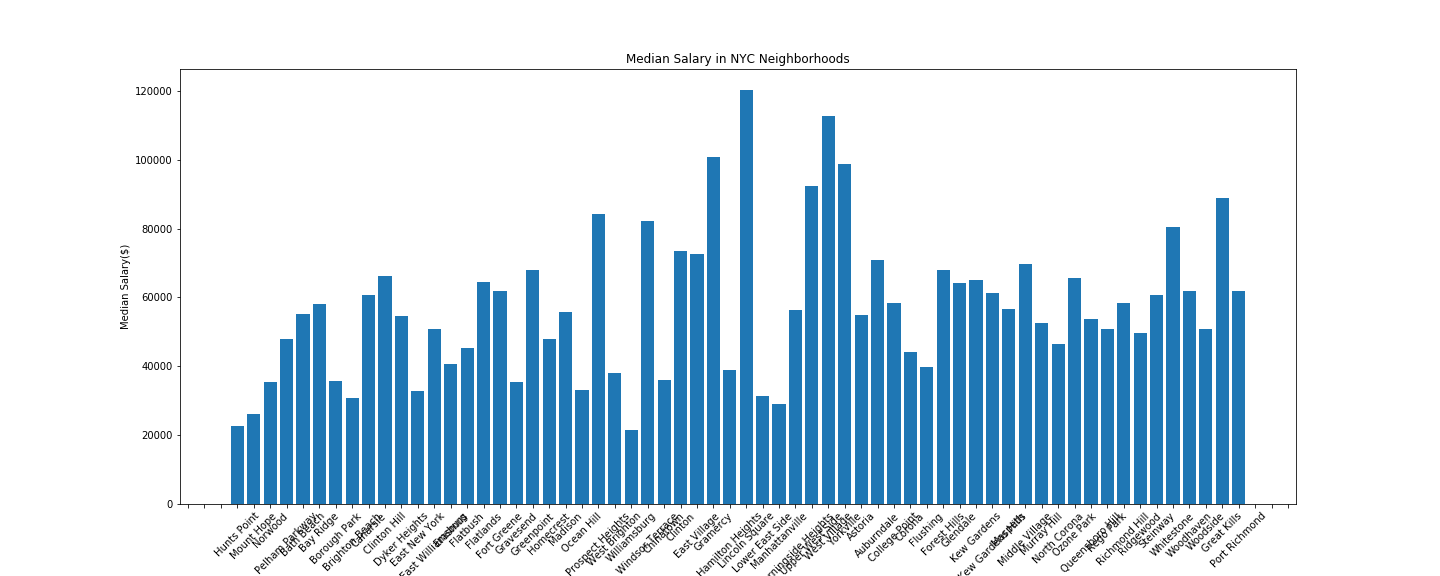
To get a sense of which neighbourhood(s) would be characterised as advantageous to invest in sports retail, various factors of each neighbourhood must be taken under account. The total number of sports venues and facilities existing in the neighbourhood is the first one we considered, as the demand for sports equipment is likely to be positively related to the number of places at which sports can be exercised. A neighbourhood’s median salary is also a socioeconomic indicator that might affect someone’s expenditure on hobbies such as exercising and is therefore also included in the analysis. Finally, the population of each neighbourhood defines the most direct exposure of the customer target audience to the products provided by the hypothetical retail store.

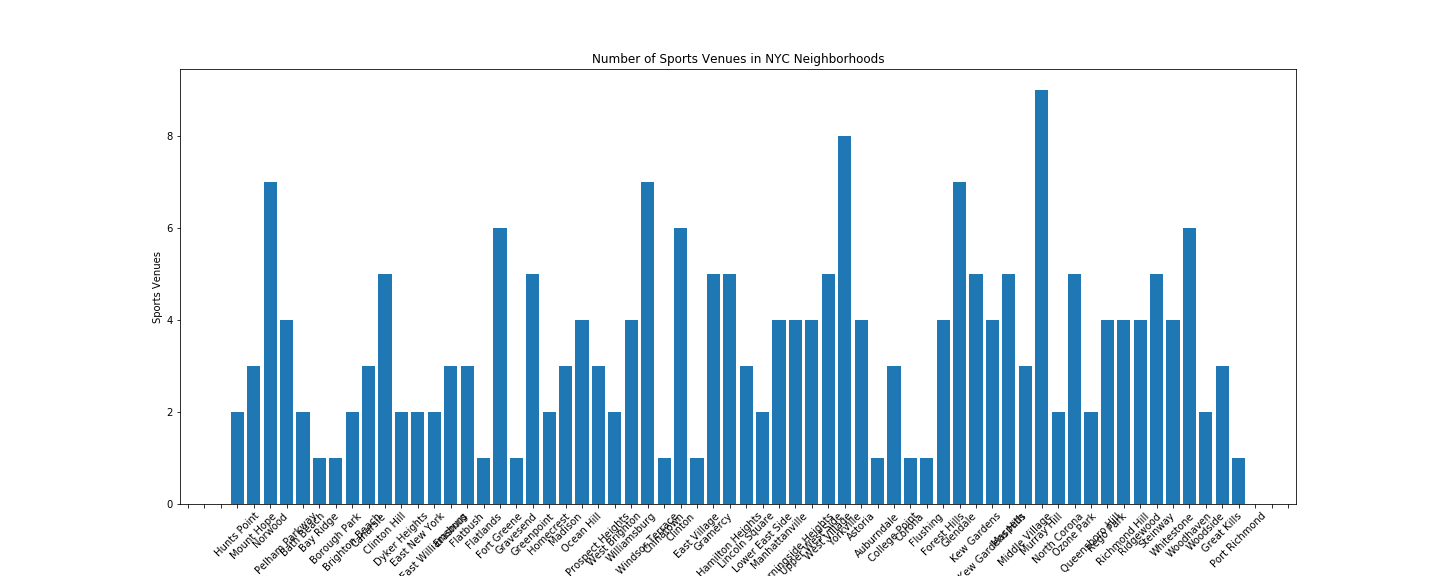
The above characteristics were retrieved for various NYC neighbourhoods from various sources, outlined in the Data section. Using the Foursquare developer services, key information about NYC neighbourhoods’ sports facilities and location data were identified and combined with other characteristics.

**Exploratory Data Analysis**

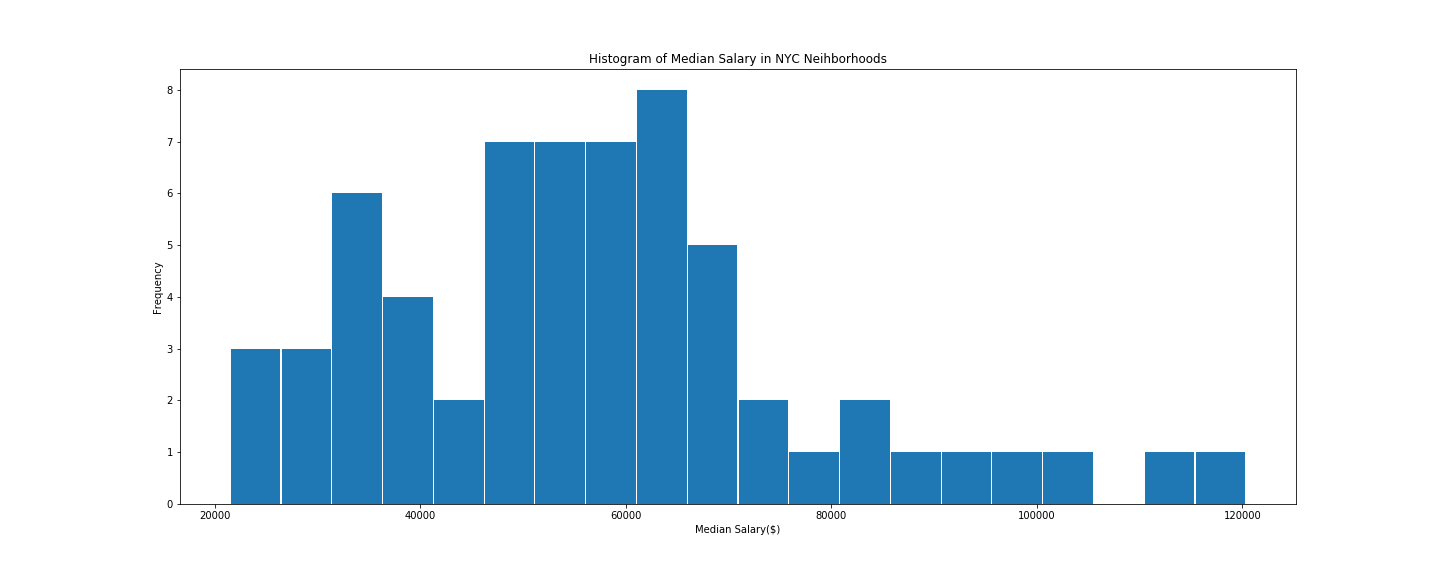
It is expected that neighborhoods with high median salary, huge population and lots of sports venues and facilities will provide a better investment location, so it is of interest to identify such neighborhoods. For this reason, in the exploratory data analysis, barplots of Population, Median Salary and Number of Sports Venues by neighborhood were plotted respectively, and are shown below:







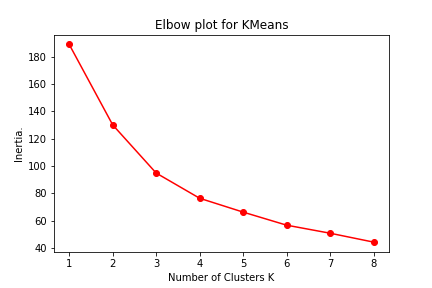
In addition, in order to get an overview of the distribution of median salaries across neighborhoods, a histogram of median salaries was plotted:



As we can see, neighborhoods such as West Village, with a population of 66880, a median salary of $112689 and 5 different sports venues constitute good candidates for investing in sports retail. The distribution of median salaries is skewed to the right, indicating the fact that quite a few neighborhoods are much wealthier than most of the others, and hence focus should be prioritized to those.

**K Means Clustering**

In order to quantify the extend to which the above attributes distinguish neighborhoods into different investment opportunity groups, the data matrix including the attributes was standardized and used to run K means with random start. To select the optimal value for K, the elbow method was used; the total inertia of a clustering configuration (sum of squared distances of each data point to its closest cluster centre) was plotted against K, and the optimal number of clusters was found to be 3, as it is at this point where the inertia substantially reduces:



Therefore, K Means was fit to the design matrix using K=3. The NYC neighborhoods were clustered in terms of the attributes described above, and the outcome was interpreted as defining distinct classes closely related to the opportunity for investment.

**Results**

After the algorithm was run, the resulting clustered neighborhoods were visualized on top of a NYC map using Folium. It was decided to compare the clusters using the average sum of the standardized attributes across neighborhoods, for each cluster. For example, Norwood, a neighborhood allocated to cluster 0, has standardized population, number of sports venues and median salary equal to -0.2862, 1.8621 and -1.0581, respectively. The sum of the above is 0.5178, and the average of such sums for cluster 0 is 1.5091. Similarly, the average sums for clusters 1 and 2 are 0.9096 and -1.5006. A higher average sum of attributes indicates a higher average number for the combination of population, median salary and number of sports venues, and therefore a better investment opportunity. Hence, neighborhoods in cluster0, with the highest average sum of attributes, are the most advantageous to invest in.

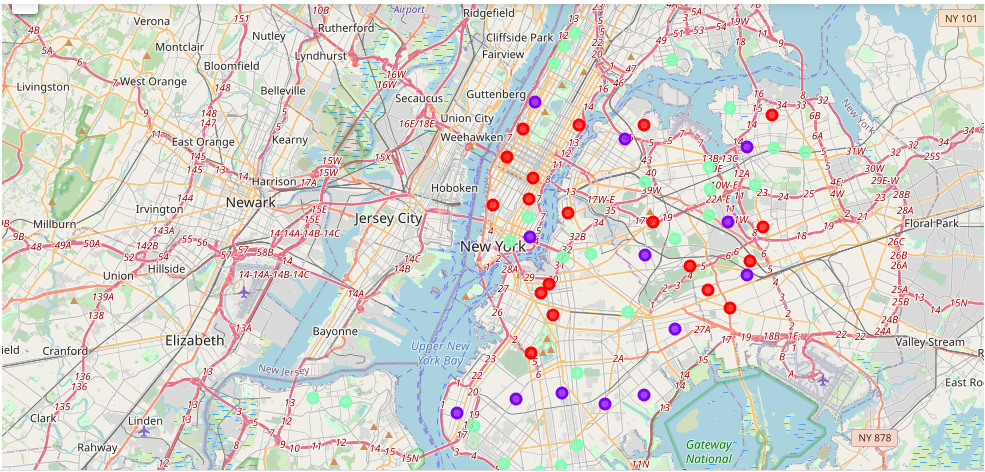


Figure 1: Clustered neighbourhoods in NYC

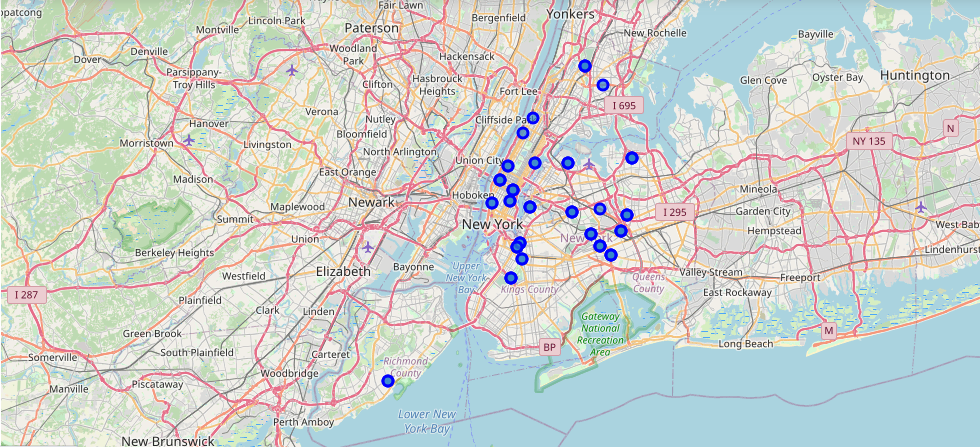


Figure 2: Neighbourhoods for optimal investment

**Discussion**

The analysis adopted throughout this report serves as a first basis approximation on deciding where and why to invest in sports retail in NYC. Based on the basic characteristics of a neighborhood that were utilized in the study, the following neighborhoods can be classified as “best”:



If historical data on the outcome of investment in sports retail (profitable/not profitable or similar) were available, classification techniques could be used to classify new hypothetical investments based on neighborhood attributes, so a suggestion for future analysis would be the collection of such data. In addition, retail performance metrics such as revenues and profits could be incorporated in the study by regression models that predict the level of such metrics based on neighborhood characteristics. However, with the limited amount of data available for this report, some fairly good results were obtained.

**Conclusion**

To conclude, our analysis shows that a basic consideration of location and demographic data of neighborhoods in NYC such as median salary, population and number of sports venues can lead to a successful segmentation of neighborhoods into classes of various interests for a potential investor in sports retail. A new branch of a retail store will have higher chances of flourishing in an area with high exposure to a wealthy audience, with plenty of places to exercise and put the sports equipment into use.

Further analysis, as aforementioned, could include further data collection in the form of surveys that could subsequently be used for more machine learning models and methods, such as regression models and classification trees.