Data Mining: Project 1
Stop-Question-and-Frisk

Tasks: Data Cleaning, Preprocessing, Exploration, and Visualization
Assigned: August 17th,2017
Due: September 14th,2017
Points: 90

"The Stop-Question-and-Frisk program in New York City is a practice of the New York City Police Department by which police officers stop and question hundreds of thousands of pedestrians annually, and frisk them for weapons and other contraband." (Wikipedia)

An analysis by the New York Civil Liberties Union (NYCLU) revealed that innocent New Yorkers have been subjected to police stops and street interrogations more than 5 million times since 2002, and that black and Latino communities continue to be the overwhelming target of these tactics. Nearly 9 out of 10 stopped-and-frisked New Yorkers have been completely innocent. (https://www.nyclu.org/en/stop-and-frisk-data)

For this project we will look at stop-question-and-frisk data for the year 2016. The data and data description can be obtained from: Moodle

The modified dataset contains 12,404 objects and 107 attributes.

Follow the CRISP-DM framework

1. Business Understanding [10]

1.1 What is the purpose of the Stop-and-frisk program?

1.2 How do you define and measure effectiveness for such a program?

1.3 What data would be needed to judge its effectiveness?

2. Data Understanding [80]

2.1 Identify type for each attribute (nominal, ordinal, interval, ratio) in the data file. [5 points]

2.2 Verify data quality: [30 Points]

- Locate missing values, duplicate data, outliers, inconsistent data.
- Implement data cleaning approaches in R.
- Explain what you do for data cleaning.
- Provide the R codes (R markdown).

2.3 Give simple appropriate statistics (range, mode, mean, median, variance, counts, etc.) for the most 10 important attributes and describe what they mean or if you found something interesting. [15 points]

2.4 Visualize the most 10 important attributes appropriately (histogram, bar chart, etc.). Provide an interpretation for each chart. [15 points]

2.5 Explore relationships between attributes for at 10 relationships: Look at the attributes and then scatter plots, correlation, etc. as appropriate. Explain the results. [15 points]