Data Wrangling Capstone Project

First, I separated the dataset of Incident Level Complaint Data – 2006 through most recent full year into 8 different years of dataset of CVS files from the year of 2010 to 2017. This is due to the fact that dataset is too large for my computer to handle. From 2010 – 2017, there are more than 4 million rows of data.

Second, I started to clean my data for Incident Level Complaint Data – current year through most recent full quarter (2018) first. I started by filtering out the columns that are not useful for my project which are (columns 1, 2, 6:13, 15, 17:20, 22, 23, 27, 31:34). After that, I changed the column name for each of the columns in order to make it easier to read and also make all the blank data into UNKNOWN. For the Complaint Date column, I started format to (%m/%d/%Y) and filtered everything before January 1st of 2018. Finally, I changed all data that doesn’t make sense in Suspect Age Group, Victim Age Group, Suspect Sex, and Victim Sex into UNKNOWN.

Lastly, since all the dataset from the year of 2010 – 2017 are all from the same CVS file. I used the same steps throughout the datasets of year 2010 – 2017. First, I simply deleted the useless columns which are (columns 1, 4:8, 10:12, 15, 17:23, 27:29, 31:32). After that, I rearranged the order of the columns to match the 2018 dataset. I also formatted the Complaint Date to (%m/%d/%Y) to match the 2018 dataset. Then, I changed the column names for each of the column for the purpose to making it easier to read. Finally, I changed the blank data and all the data that does not make sense into UNKNOWN.

You can find all the R scripts that I used on the each dataset on the same GitHub file.