Capstone Project 1 Update

19th March 2018

Data Wrangling

The dataset being used for this project is from the City of Phoenix Open data portal. The dataset in question is the Crime Data for Phoenix. The data provided is from November 1st, 2015 forward through 7 days prior to the newest posting date. The data used for the capstone project is from 11/1/15 to 2/20/18 (the csv file was downloaded on 2/27/18). This dataset has 148468 rows, 7 columns and 2 data types: object(6) and float64(1).

Reviewing missing data showed 4 columns have missing data:

1. Date_occurred_on = 336
2. Date_occurred_to = 43719
3. Zip code = 3
4. Presmise_type = 805

To clean the missing values for the date columns they will be filled in with the date/time from the populated date column of that incident. A check was run to display the rows that were NaN for date_occurred_on and date_occurred to and it returned 0 rows where that condition is true. Since we had missing values of 336 and 43719 we cannot omit those rows. The 43719 missing values are due to the victims knowing the exact date/time of the incident therefore the date_occurred_to field is left blank. The 3 zip code rows will be removed since 3 is a very insignificant number that will not affect the project when removed. The premise_type missing values will be classified as unknown, there is already an unknown variable entry. The 805 could not be removed since it is a large number of rows that could affect the project.

Overall the data is pretty clean aside from the large number of missing values. The clean data set is called *crime_clean* and has 148465 rows and 7 columns.