Capstone Project 1 Data Wrangling

By: Charlie Lee
Date: 1/28/2020

1. Data cleaning approach

Pandas and Numpy packages were imported. Main dataframe was generated by combining 3 different CSV files (2016, 2017, and 2018 respectively). These 3 different CSV files had the same number of columns and names. As each were loaded with the Pandas method, the data type object (dtype) for ‘Date’ column was set to Datetime so the month can be extracted easier later on.

Unnecessary columns such as: 'Arrest', 'Domestic', 'X Coordinate', 'Y Coordinate', 'Updated On', and 'Location' were dropped. Upon observing the dtypes for the columns, it was observed that ‘District’, ‘Ward’, and ‘Community Area’ were in floats rather than integers which was incorrect.

Before being able to convert the dtype, NaN objects had to be taken care of. There were: ‘District’ - (1), ‘Ward’ - (5), and ‘Community Area’ - (2) NaN values. NaN values found in ‘Location Description’, ‘Latitude’ and ‘Longitude’ were ignored since these won’t be used as part of data analysis.

After NaN values were taken care of, the dtypes were changed from float64 to int64 then exported to a CSV file.

A pivoted dataframe was generated by selecting ‘Year’ and ‘Month’ (extracted from ‘Date’ column) as the index, and ‘Primary Type’ as the column. This pivoted dataframe shows how many specific crimes occurred during a specific month of that year.

2. Missing values? NaN?

For each NaN value, the dataframe was searched to look for other data with the same conditions (‘District’, ‘Ward’, ‘Community Area’) excluding column which had NaN. The NaN values were replaced by the most frequent or only value.

Also, it was observed that three of the ‘Community Area’ had a value of 0 which was not correct.

3. Outliers?

There were no outliers in the data.