Data Wrangling

NYC Crime Statistics from 2015 to 2019

The data came from:
https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243/data.

The data was extracted from the website above through a csv file. I was able to filter 5-year data and convert it to a CSV file.
It was a challenge to find the most efficient method to import the data and libraries that are needed to create a single coherent dataframe. But as soon as I successfully made the dataframe, I was able to play around it. My next step is to make a time series of the frequency of crimes committed in each date in each borough after data cleaning.

a.) What kind of cleaning steps did you perform? How did you deal with missing values, if any?

First, I deleted unnecessary columns. There were 35 columns originally. It was cut down to 26 columns. I left the coordinates (last four columns on the right) of the reported crimes so if I have enough time I can do data visualization exploration on it.

Next, all the columns pertaining to dates were converted to datetime format. Though I have the time column, I will merge it later on.

Lastly, I made sure that the columns with unique values don't have any duplicates. Fortunately, it was all unique.

b.) How did you deal with missing values, if any?

I am starting to work on the time series for the frequency of boroughs. My initial approach is to fill all the NaN values in the boroughs column and replace it with “UNKNOWN”. I need to figure out ways to replace NaT values on date columns so the columns will be able to read the “NaT” values.

My data is fraught with duplicate frequency of dates but the frequency is very random. Some dates have 10 crimes committed while some only have 2, 23, etc. To solve this, I filter the DATES and BOROUGHS by boroughs and graph the frequency of crimes in each date using df.values_count().

c.) Were there outliers, and how did you handle them?

As I plot the whole 5-year data, I noticed that there is way fewer data in years 2015 to 2018. If I plot all the 5-year data altogether, the years before 2019 would look like outliers. My mentor and I have agreed to use only the crime data in 2019. I filtered the dateframe so all the values are within the 2019.