Chicago Crime Data Analysis

Data Wrangling:

In this project, I will be dealing with the data set containing of all the crimes that are reported by the police in their directory from 2001-present.

This dataset is available in the Chicago city data repository. This dataset consisting of 7 million row of data by the columns of ID, Case Number, Date, Block, IUCR, Primary Type, Description, Location Description, Arrest, Domestic, Beat, District, Ward, Community Area, FBI Code, X Coordinate, Y Coordinate, Year, Updated On, Latitude, Longitude, Location

On this data first objective was to remove the unwanted data, columns containing the unwanted data or un relevant information like ID, Case Number, Block, IUCR, Beat, Updated On and location were removed. The second objective was to convert some of the object type columns to integers. The questions we are associated to different types of crimes and how they are located.

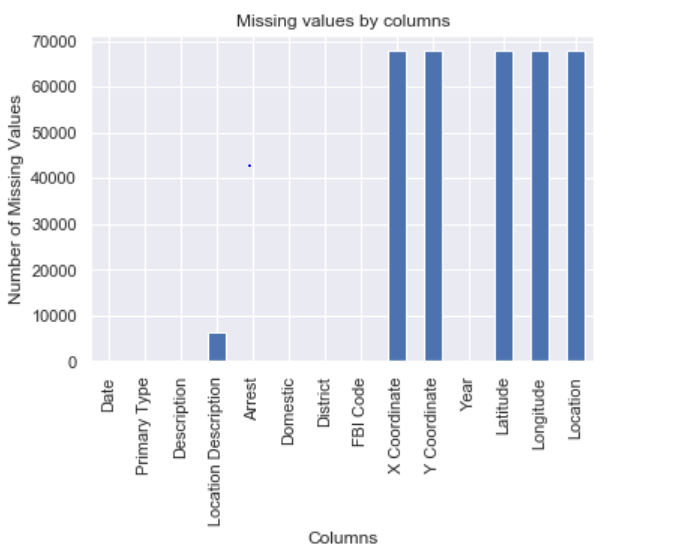
After preprocessing, the data frame consists of following information with our dropping null values

Dropped the following columns from the main Data Frame called crimes



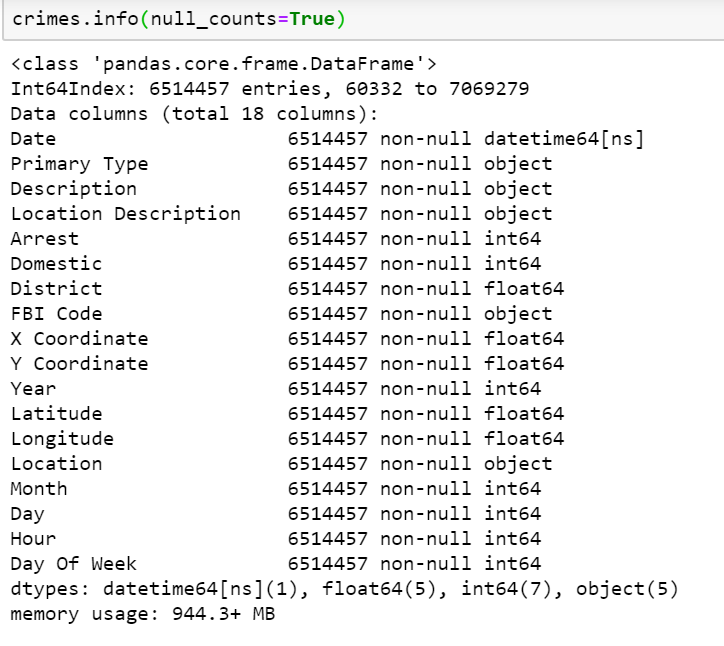
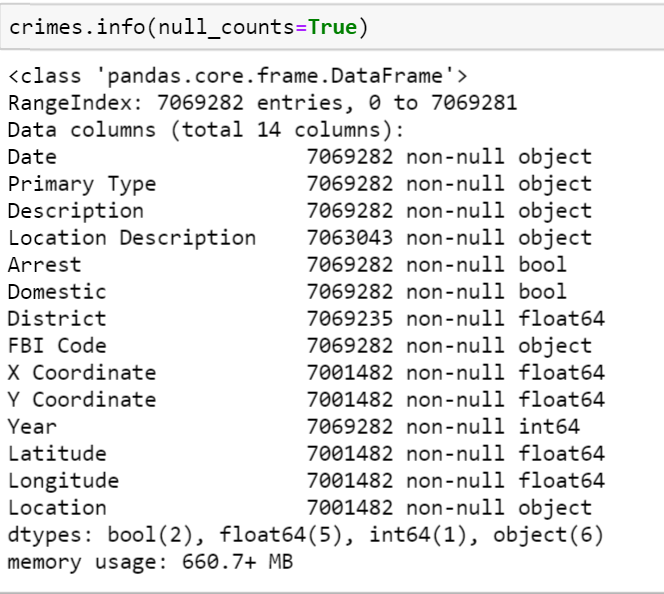
Difference between the columns containing null values and without null values (which is after dropping all null values)

Inspecting the features, we see that all the features that have a large count of missing values are features that relate to the geographical location of the crime scene. This is No Surprise as the Chicago Crime Dataset is based on firsthand accounts of people involved in or around the crime. It is not necessary that such firsthand reports need to contain the specific locations of the crime. We have 3,45,286 missing values in the whole dataset that are present in Location Description, Community, X Co-ordinate, Y Co-ordinate, Latitude, Longitude and Location.

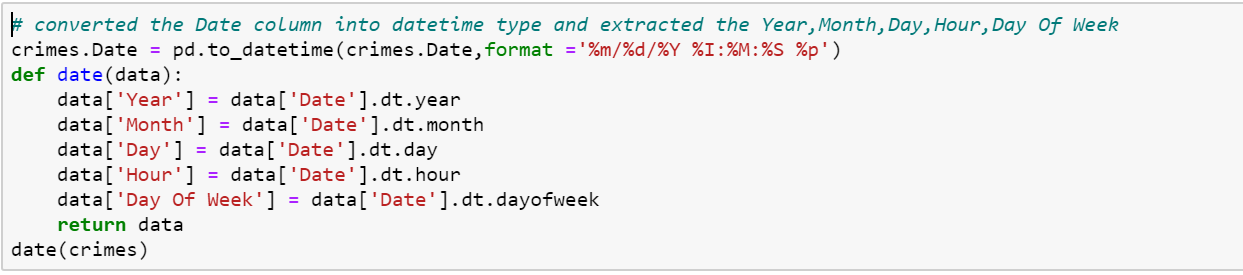


Since, these features are not direct numeric values, we can't use summary statistical functions to fill in the missing values. Hence I thought to remove all values containing null using dropna().92.5 percent of data is retained after dropping null values.

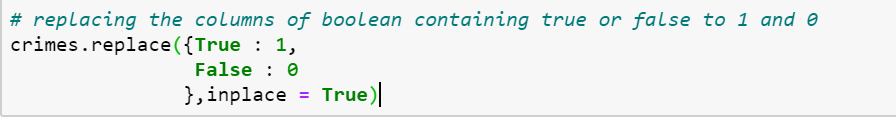
Below figure shows the comparison of no-null values counts before and after dropping null values



I have created 3 additional columns Day, Hour, Day Of Week by converting the Date column to the datetime type and then slicing the required column attributed for the Date. Below figure gives the code that I used for slicing. This will be helpful for me to identify the crime pattern in different sections of day



The columns [‘Arrest’, ‘Domestic’] contains the values in terms Boolean values, I have converted to their respective 1’s and 0’s in the column places using the following code



All the changes done to the Data Frame is saved in the form of CSV file and PICKLE files