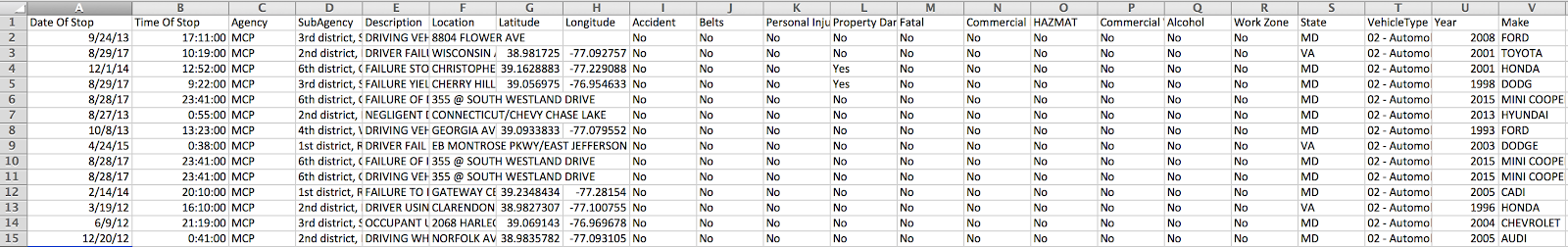
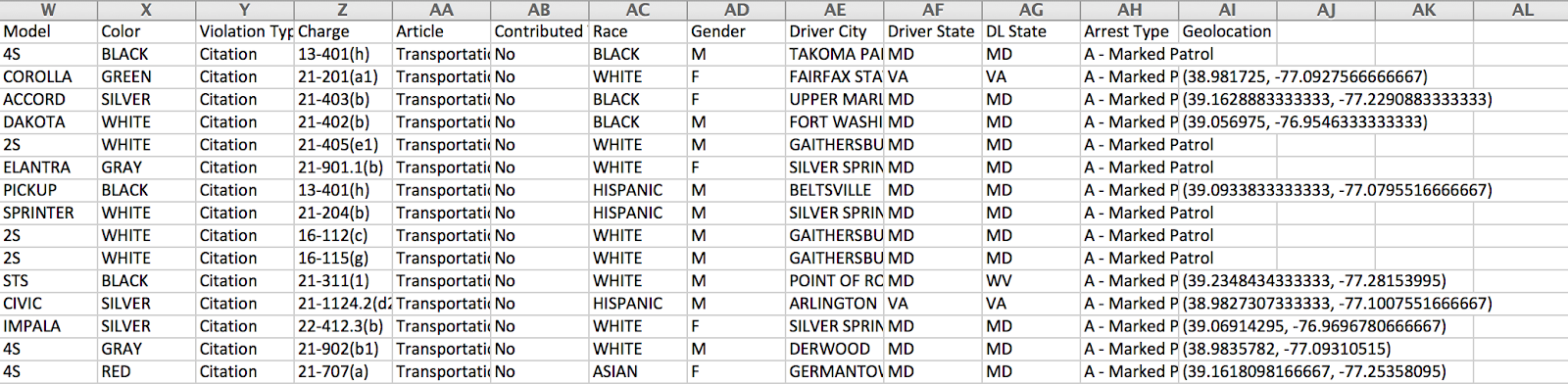
The Data is taken from the following source

:<https://catalog.data.gov/dataset/traffic-violations-56dda/resource/450018e7-f6c0-43fd-b5c9-a83de293b206> is updated till August 9th, 2018

The cvs file has 1373693 rows and 35 columns taken for entire 50 states across USA

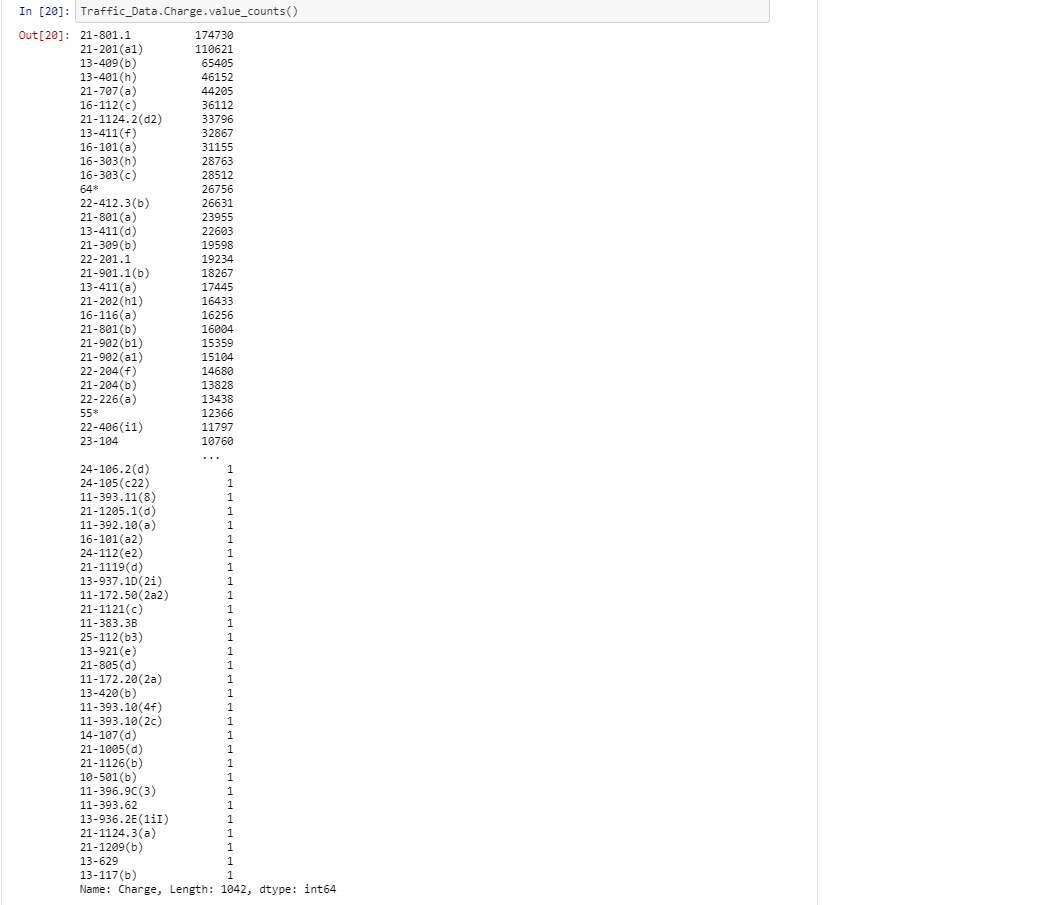
The columns have the following information in them

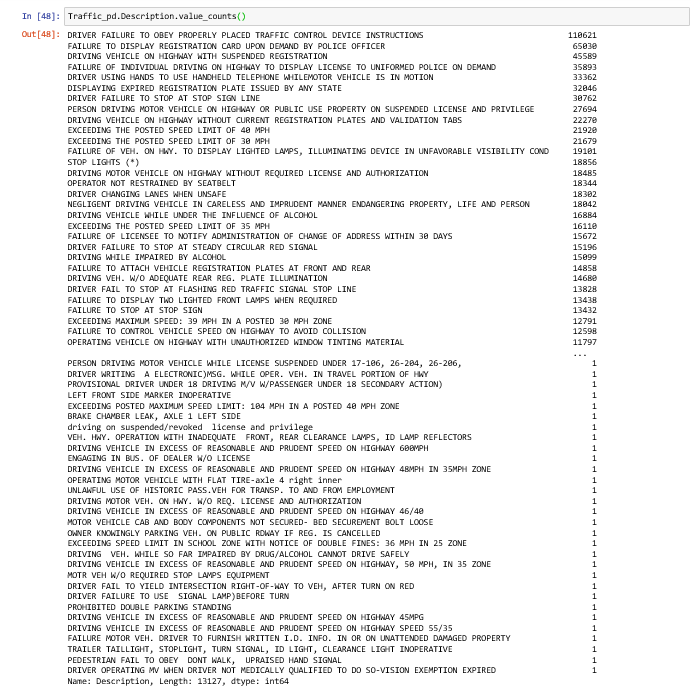




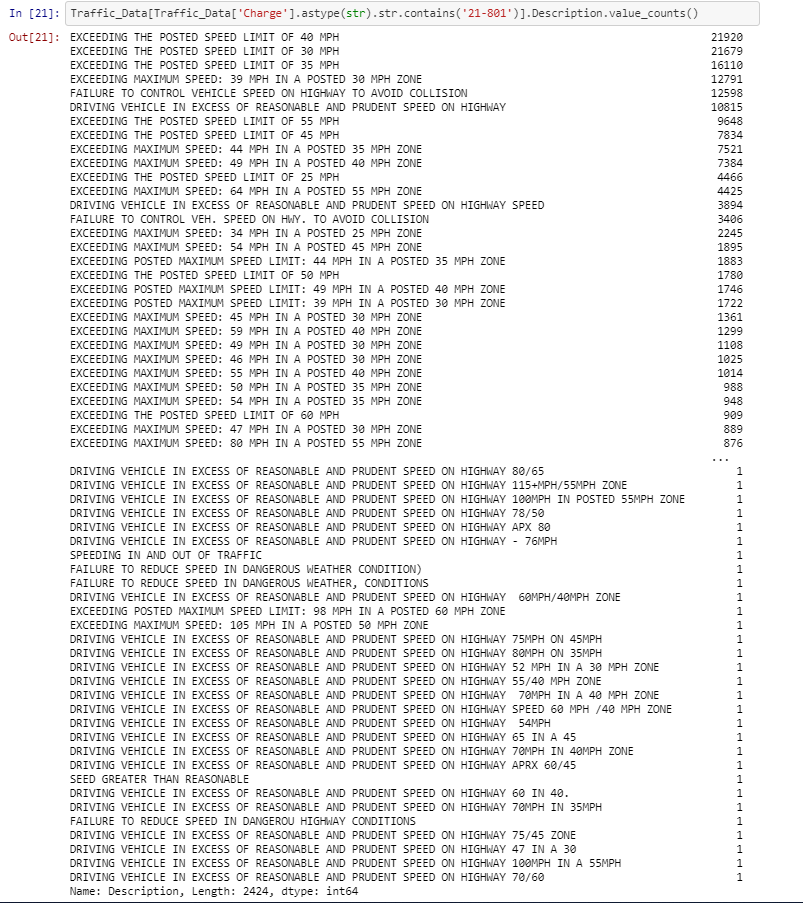
Our focus was on the columns Description, Charge, Race, Gender,Date

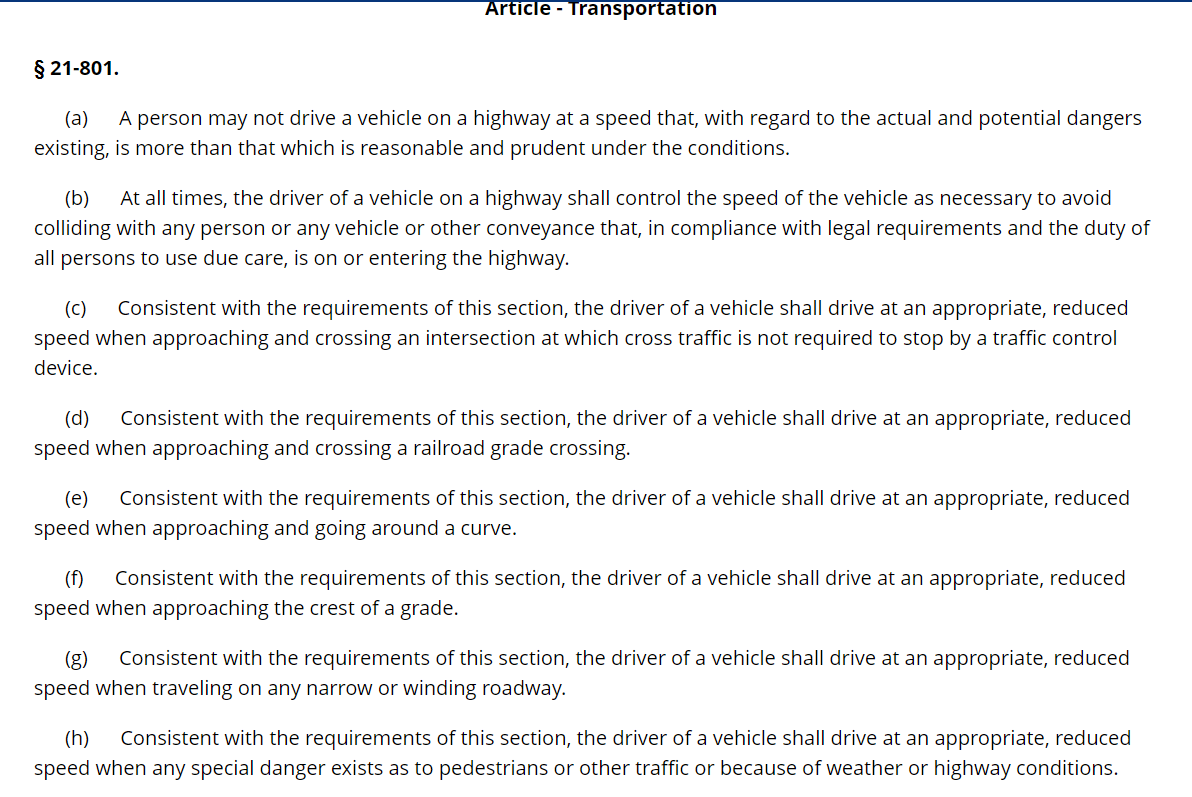
If we look at the Description and Charge in our dataframe



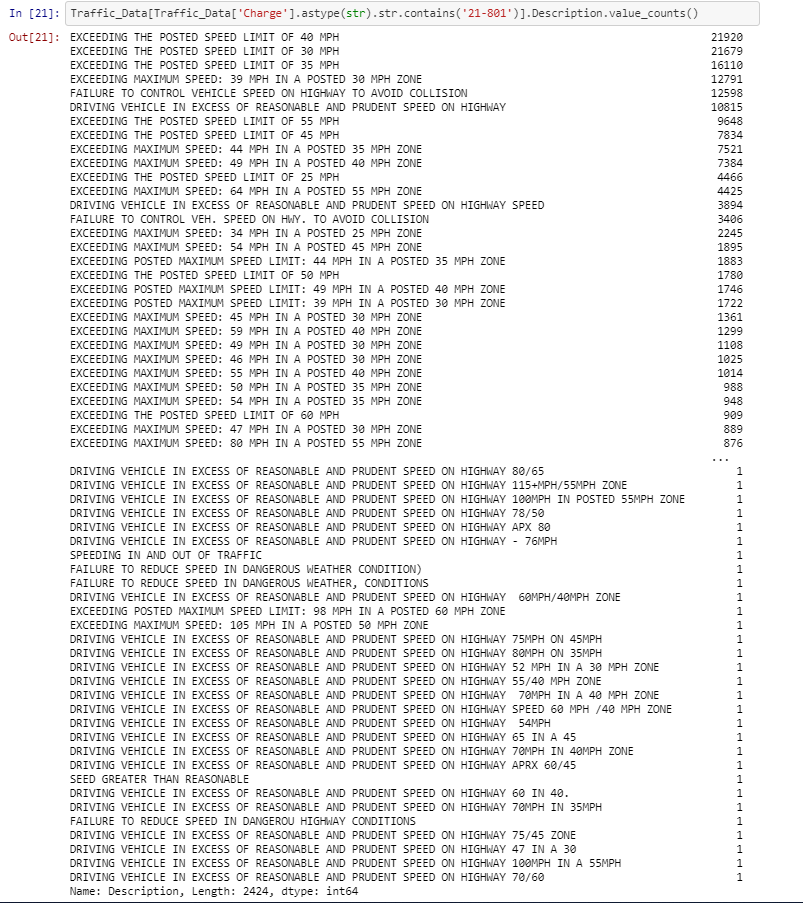
I

In order to justify the top 20 Charges and Description ,the search was based on Charge.

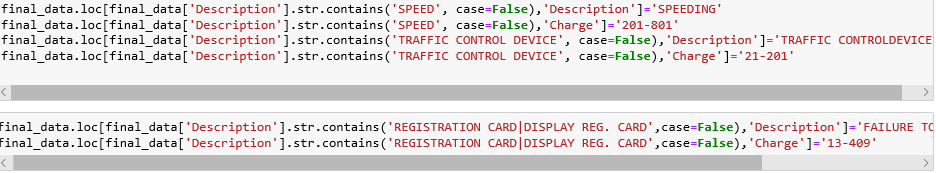
When we search the first and top valued  charge 21-801 gives 

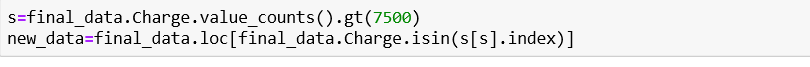


So Based on Charge-21-801,Speed is the keyword and all the counts are measured under it .

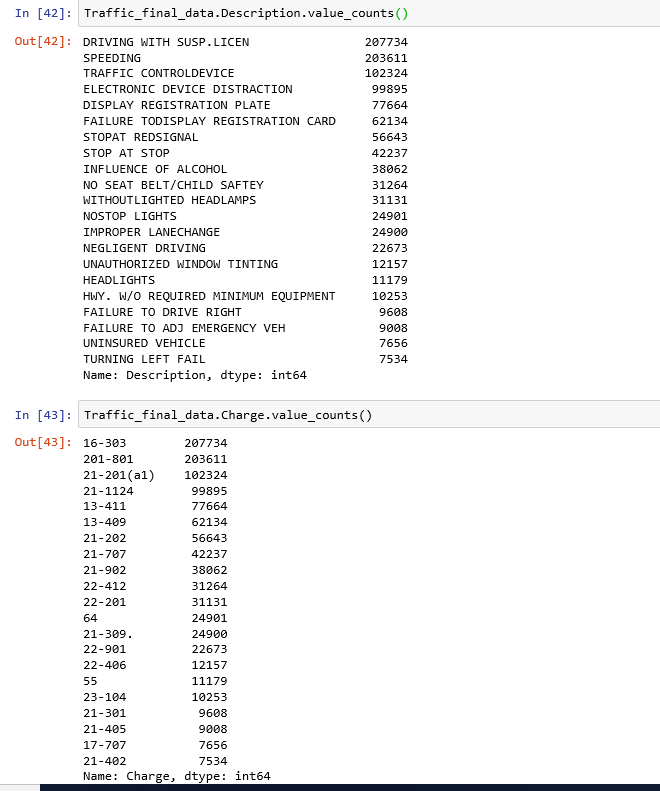
If we look under  Description column we have 2424 types being accounted here. 

The idea is to combine these Description with unique names ,when we combine all these Descriptions into one category of violation ie ‘SPEED’, the following  31 charges are combined together 21-801., 21-801(b), 21-801(a), 21-1116(a), 21-801(g), 21-301(b), 21-804(a), 21-405(e2), 21-802.1, 21-1125(1),21-801(h), 21-801(e), 21-804(b), 77, 21-303(e),21-804(c), 21-1205(a), 21-803.1(e), 21-1409, 21-806(b),21-801(c), 21-1125(3), 21-805.1(a), 21-801(d),21-1205.1(a), 21-1125(2), 22-221(j),'21-1125(4), 21-801(f), 21-604(e), 21-1116(b), 11-392.11

And around 220741 Decsriptions are combined together under one category “Violation due to Speeding” or SPEEDING using the code below.

This way top 20 Traffic violations are accounted and the remaining charges are dropped.

The cleaned data frame thus has



The cleaned csv file is

