Data Analysis of "Motor Vehicle Crashes" for New York State

Overview: Motor Vehicle Crashes datasets consists of 3 datasets and provides information on motor vehicle crashes involved in traffic collision. The main dataset provides the case information, and the other three datasets: Vehicle Information (provides detailed report about vehicles involved) an Individual Information (provides detailed report about individual involved) This dataset shows each collision data recorded over a span of three years (2012-2014).

Goal: The major goals for this analysis are to find out when are the most dangerous times of the day to be driving, note which counties has the maximum number of vehicular crashes, plot the roads with maximum fatal accident over past three years, find the year which involved alcohol as a major contributing factor and find out factors such as road conditions, weather conditions and lighting conditions which have contributed to the accidents so that measure can be taken by the Traffic and Safety Department and Police Department to prevent it in the future.

Background: The dataset reports details of all traffic collisions occurring on county roadways within New York State and provided by the NYS Department of Motor Vehicles.
The dataset from https://data.ny.gov/ website is open source data and contains data from the year 2012 - 2014 in the form of csv split. Some of the columns are 'Road Surface Condition' which represents Condition of roadway surface, 'DOT Reference Marker Location' which Department of Transportation reference marker present at location of crash, 'Traffic Control Device' which represents reported traffic control device present where the crash occurred and so on.

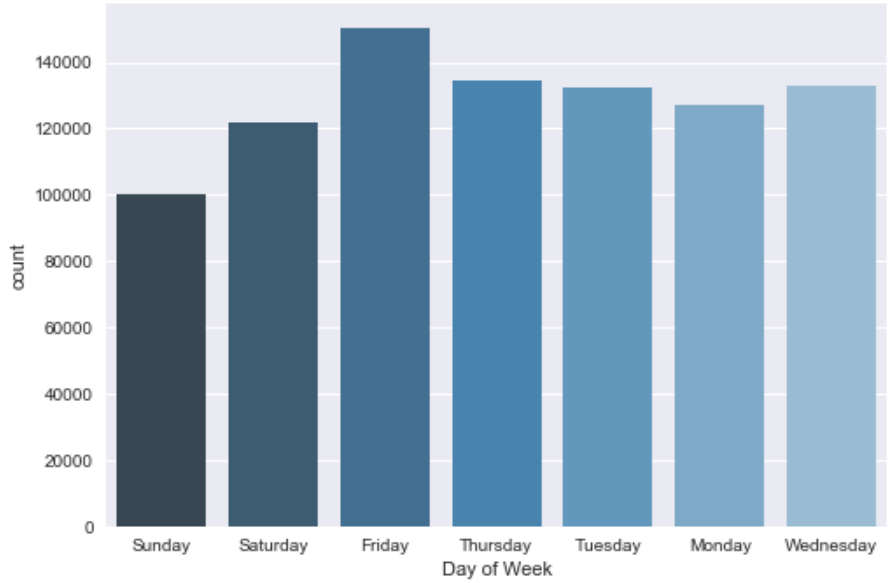
Analysis:

1.

```
NASSAU          99773
SUFFOLK         90425
QUEENS          61127
KINGS           60483
ERIE            49509
WESTCHESTER     42605
MONROE          42281
NEW YORK        38142
BRONX           32213
ONONDAGA        29367
Name: County Name, dtype: int64
```

It is observed that the above 10 counties have had maximum number of crashes.

2.



In the above graph, Y-axis is the count of crashes, X-axis is the Day of Week. The bar plot shows the day of the week with maximum and minimum number of crashes.

3.

6AM-12PM

	Day of Week	count
0	Friday	19892
2	Saturday	19574
5	Tuesday	17448
6	Wednesday	17392
4	Thursday	17344
1	Monday	17129
3	Sunday	15366

12PM-6PM

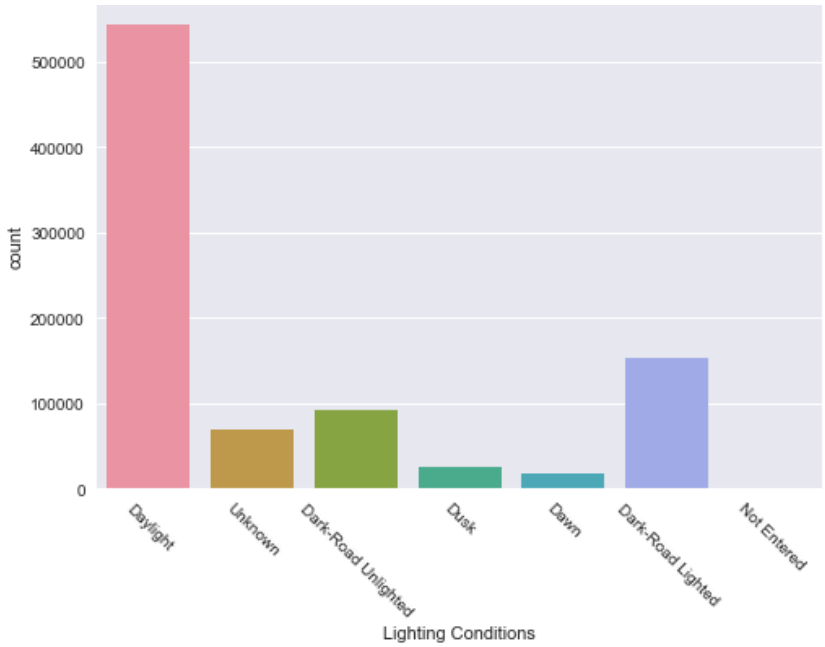
	Day of Week	count
0	Friday	60041
5	Tuesday	52900
6	Wednesday	52781
4	Thursday	52698
1	Monday	50435
2	Saturday	43686
3	Sunday	35824

6PM-12AM

	Day of Week	count
0	Friday	36244
2	Saturday	32318
4	Thursday	31130
6	Wednesday	29270
5	Tuesday	28763
1	Monday	26986
3	Sunday	26652

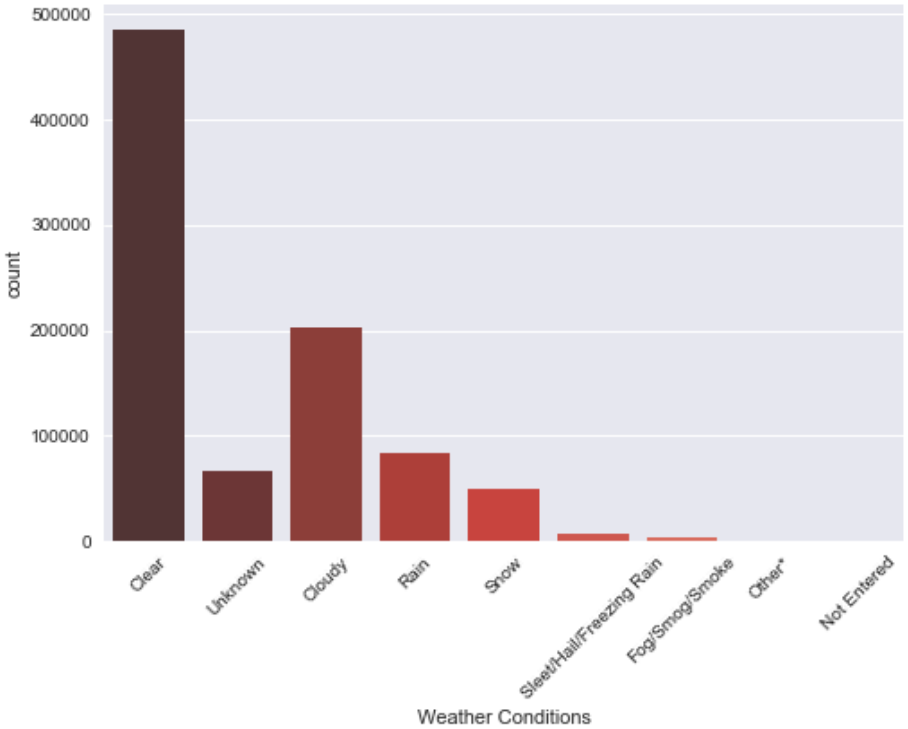
The above table is distributed as morning hours(6Am-12Pm), afternoon hours(12Pm-6Pm) and evening hours(6Pm-12Am) and the crashes count based on the Day of Week.

4.



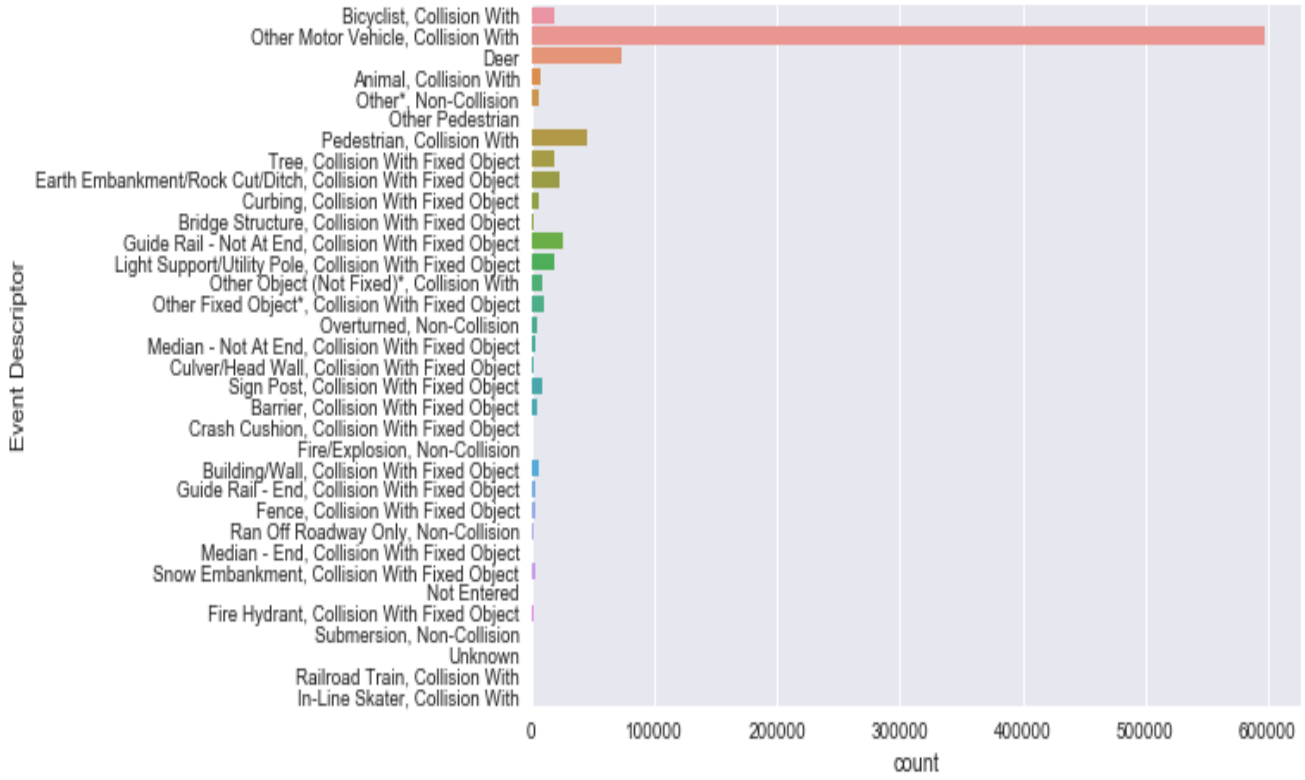
In the above graph, Y-axis is the count and X-axis is the Lighting Conditions which have contributed to the crashes.

5.



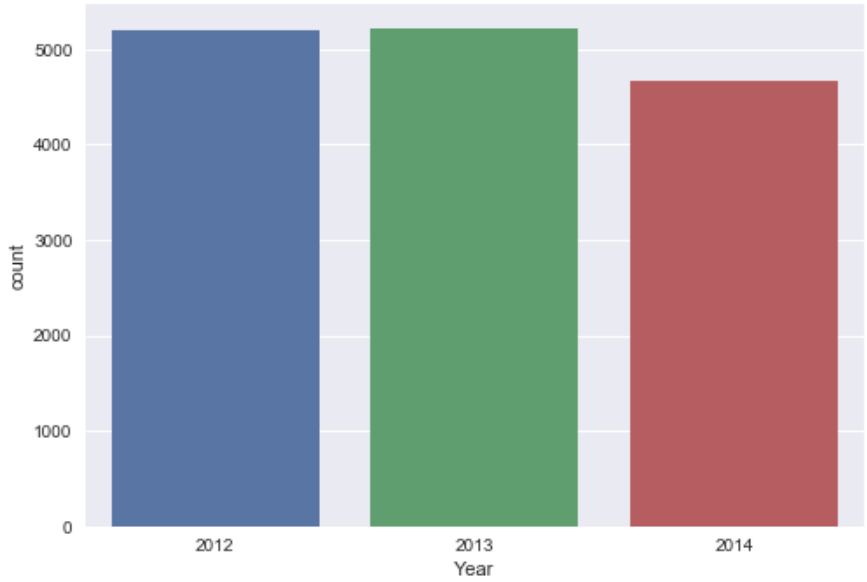
In the above graph, Y-axis is the count and X-axis is the Weather Conditions which have contributed to the crashes.

6.



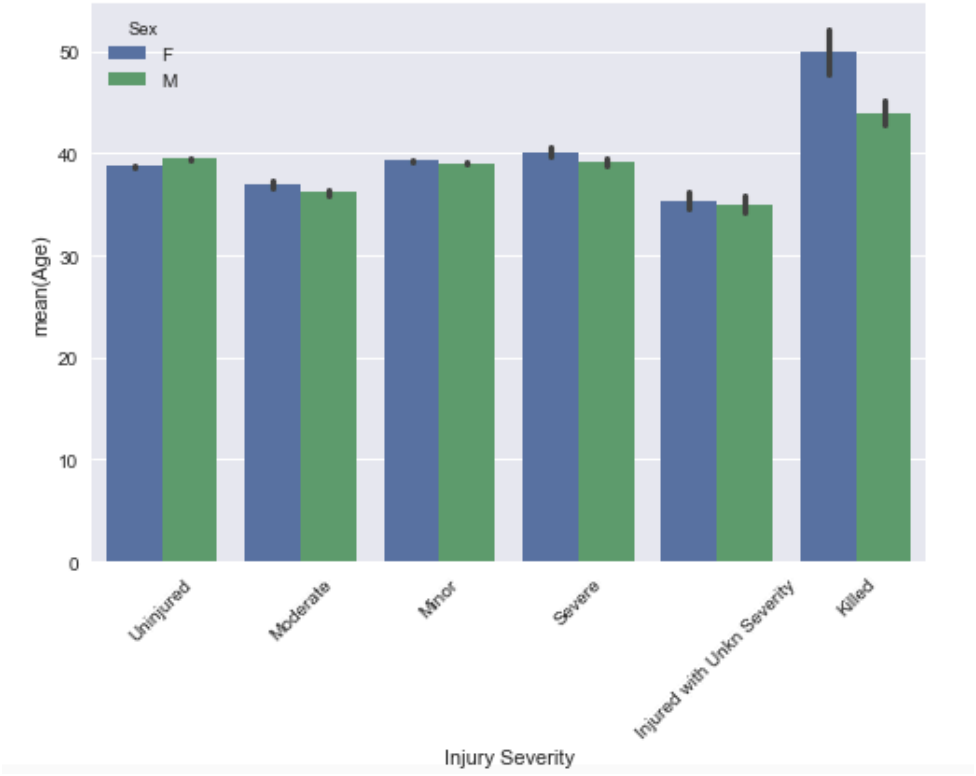
In the above graph, X-axis is the count of crashes, Y-axis is the Event Description responsible for vehicle crashes.

7.



In the above graph, X-axis is the Year, Y-axis is number of crashes. The above bar chart shows alcohol involvement contributing to motor vehicle crashes.

7.



In the above graph, X-axis represents the injury severity, Y-axis represents the Age(mean). The legends are the Male and Female ratio.

8.

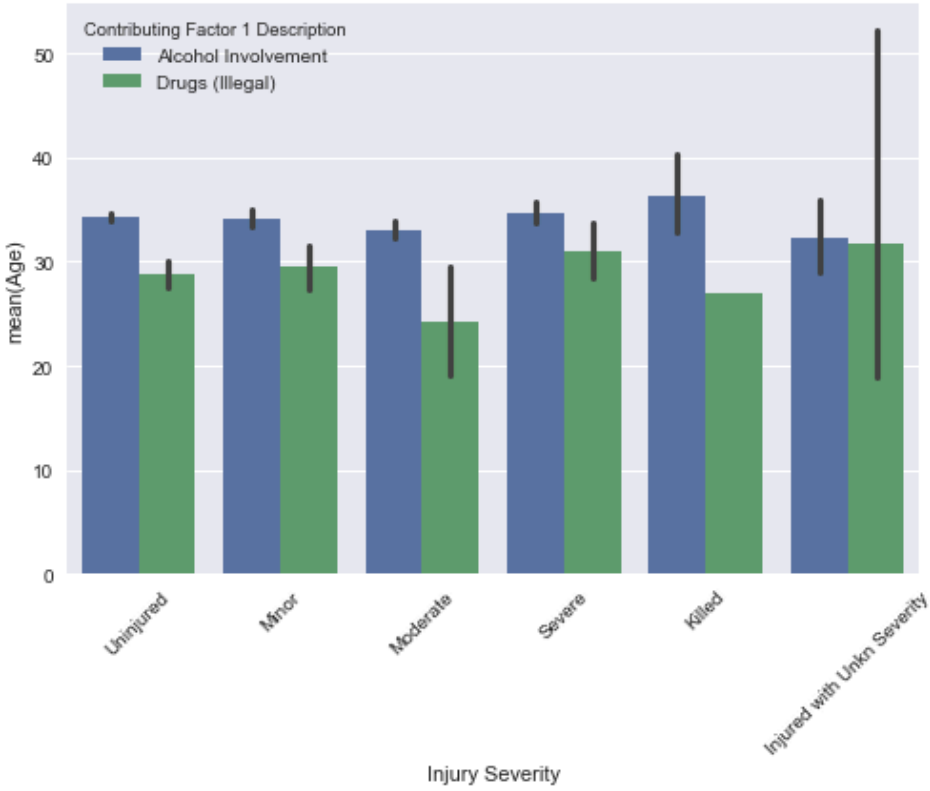
Top 10 Counties where Snow and rain affected the number of Crashes

County Name	Snow Count
ERIE	5827
MONROE	4931
SUFFOLK	3965
ONONDAGA	3734
NASSAU	2894
ONEIDA	2246
WESTCHESTER	2158
ORANGE	2087
ALBANY	1897
NIAGARA	1573

County Name	Rain Count
NASSAU	16535
SUFFOLK	14665
ERIE	8661
QUEENS	8312
KINGS	8183
MONROE	7800
WESTCHESTER	7405
ONONDAGA	6152
NEW YORK	5181
ORANGE	4341

The above tables depicts, top 10 counties based on count where snow and rainy weather affected the Crashes.

9.



In the above graph, X-axis represents the injury severity, Y-axis represents the Age(mean). The legends are the contributing factors for the vehicle crash.

Results and Conclusions:

1. The maximum number of crashes happen during weekdays with Friday being the most and Sunday being the least.
2. The average number of crashes in morning hours is 17K, afternoon hours is 50K and evening hours is 30K.
3. The maximum accidents take place during daylight, dark-road lighted and dark-road unlighted.
4. Top 4 weather conditions affecting crashes are clear, cloudy, rain and snow.
5. It is seen that maximum crashes occur when collision with other motor vehicle and also collision with deer.
6. Erie, Nassau, Westchester, Suffolk, Onondaga and Monroe are the counties which had maximum number of crashes during snow and rains.
7. Individuals between age group of 25-30 were involved in crashes involving illegal drugs. Individuals between age group of 30-35 were involved in crashes involving Alcohol.

Scalability Challenges: Motor Vehicle Crash Case Information dataset is best utilized in conjunction with the 2 other datasets posted: Crash Individual and Crash Vehicle. On an average, each of these 3 datasets individually consist of 1.5M rows and 20 columns. On merging these datasets, the number of rows is 3.3M and columns is 51. However, Case Information dataset does not have any ink to the other two datasets. In order to solve this issue, the analysis was performed on 2 separate datasets- 1. Case Information 2. Vehicle and Individual Information Combined.

Implementation: Pandas dataframe is used to read and parse data and Seaborn for visualization. Programming language used for the project is Python. The entire project was implemented on local system.