Analyzing & Predicting Massachusetts Crash Data

Mildred Orwenyo

October 9th, 2020

1. Introduction

1.1 Background

Every crash reported to a police department in the United States of America gets its details recorded by the police and added to the respective cities' database. These records have attributes that help provide data and statics regarding the kind and severity of accidents on the roads. This information can be used to monitor the statistics of crashes over time and can be key in determining where to implement transportation and infrastructure changes.

1.2 Problem

The volume of crashes in Massachusetts is not where the leaders would them to be; the volumes need to be reduced dramatically. Massachusetts Department of Transportation (DOT) has therefore been tasked with the mission of coming up with ways to reduce the number of crashes, damages, injuries and fatalities. I am involved in a study of available data to determine the next steps to take.

1.3 Interest

The State of Massachusetts wants to rank as number 1 in the **'The Safest US State to Drive'** list by 2027. This ranking will aid in attracting more tourists and business to the state. In order to complete this mission, data needs to be analyzed regarding the crashes that happened in 2019 so that major issues/areas of concern can be addressed and specific plans of action can be set in place in order to achieve this goal. Moreover, budget allocation for accomplishing this mission will depend on the scope of how much work is needed by county depending on what the analysis reveals.

2. Data Acquisition

2.1 Data Sources

First data source is on the link: MassDOT Crash Open Data Portal. The 2019 Crashes data is being used to provide the analysis needed by looking at crash severity and road conditions of where crashes tend to occur the most.
The second data source that shows Massachusetts population is from following link: 2019 Massachusetts Population by County. This data is used to provide a comparison between population vs crash volumes

2.2 Data Cleaning & Feature Selection

Data loaded from the first source has 139,109 rows and 116 features. This dataset has many features that I did not use in the crash analysis since they were unnecessary or redundant for this analysis and/or a significant number of them had missing values. Other than the missing values, in some of the features, the dataset was pretty clean therefore, there was no need to do more cleaning.

After reviewing all the features available, I selected only 16 features to use for this analysis. These 16 features did have some missing values in some of them but the number of records was high enough to where what data was available could provide sufficient analysis.

Table 1: Examples of the features selected vs those not selected:

Kept Features	**Dropped Features**	**Reasons for Dropping Features**
OBJECTID	CRASH_NUMB	Both are distinct and could be used to accomplish the same task e.g. determining count of crashes in particular categories
CITY_TOWN_NAME	CITY	Two similar features; providing the same information
CRASH_SEVERITY_DESCR	POLC_AGNCY_TYPE_DESCR	Dropped feature does not help in analyzing types of accidents that will help find out where focus in reduction should begin

3. Exploratory Data Analysis

3.1 Determination of Crash Statistics by County

Crash statistics by county was determined by grouping the total count of OBJECTIDs by each county. This simple data analysis shows us that MIDDLESEX is the county with the most Crashes. Efforts for changes may need to begin here

CNTY_NAME	CRASH_VOL
MIDDLESEX	31018
WORCESTER	19668
ESSEX	16394
BRISTOL	15293
NORFOLK	14073
HAMPDEN	13581
PLYMOUTH	10514
SUFFOLK	6187
BARNSTABLE	5207
HAMPSHIRE	2854
BERKSHIRE	2658
FRANKLIN	1256
NANTUCKET	222
DUKES	184

Figure 1: Crash Volumes by County

3.2 Comparing Population to Crashes

Sorting and Ranking

Before digging into this data, it was my assumption that the more the number of people in an area the more the number of crashes are likely to occur. This assumption was for the most part supported by the data. I sorted both the Crash volume data and the Population volume data by volume per county. I then ranked the respective features' volumes accordingly with the highest volumes ranking first (highest) for both features.

Counties with higher populations have higher volumes of crashes in comparison to those with lower population. Thus, Population rankings are comparable to the respective county's Crash rankings. This rankings are comparable in most of the counties except for one extreme outlier i.e. SUFFOLK county. There are far fewer accidents in SUFFOLK compared to the population size. The mean crash Ratio in Massachusetts as a whole was 56 in 2019 i.e. there was an average of one accident per 56 people. SUFFOLK County has the highest Crash Ratio value i.e. one accident per 130 people.

CNTY_NAME	POPULATION	POP_RANK	CRASH_VOL	CRASH_RANK	EQUAL	CRASH_RATIO	CRASH_POP_RATIO_RANK
MIDDLESEX	1614714	1	31018	1	True	52	5.0
WORCESTER	830839	2	19668	2	True	42	11.0
SUFFOLK	807252	3	6187	8	False	130	1.0
ESSEX	790638	4	16394	3	False	48	9.0
NORFOLK	705388	5	14073	5	True	50	7.0
BRISTOL	564022	6	15293	4	False	36	13.0
PLYMOUTH	518132	7	10514	7	True	49	8.0
HAMPDEN	470406	8	13581	6	False	34	14.0
BARNSTABLE	213413	9	5207	9	True	40	12.0
HAMPSHIRE	161355	10	2854	10	True	56	3.5
BERKSHIRE	126348	11	2658	11	True	47	10.0
FRANKLIN	70963	12	1256	12	True	56	3.5
DUKES	17352	13	184	14	False	94	2.0
NANTUCKET	11327	14	222	13	False	51	6.0

Figure 2: Crash & Population Sorted and Ranked by Volumes and by County

Correlation: Relationship between Crash Volumes and Population

There is a positive direct correlation between Population and Crash Volume. Therefore, population is a good predictor of crash volume

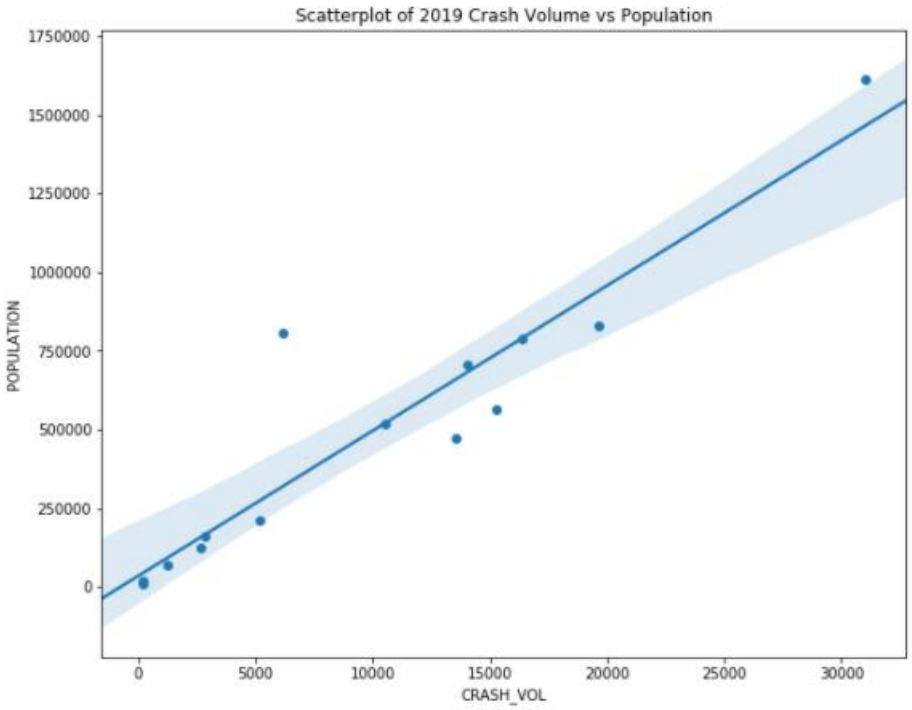

Figure 3: Relationship between Crash Volumes and Population Volumes

Relationship between Crash Volume Ranking and Population Ranking

There is a positive direct correlation between Crash Ranking and Population Ranking. Therefore, population ranking is a good predictor of crash ranking

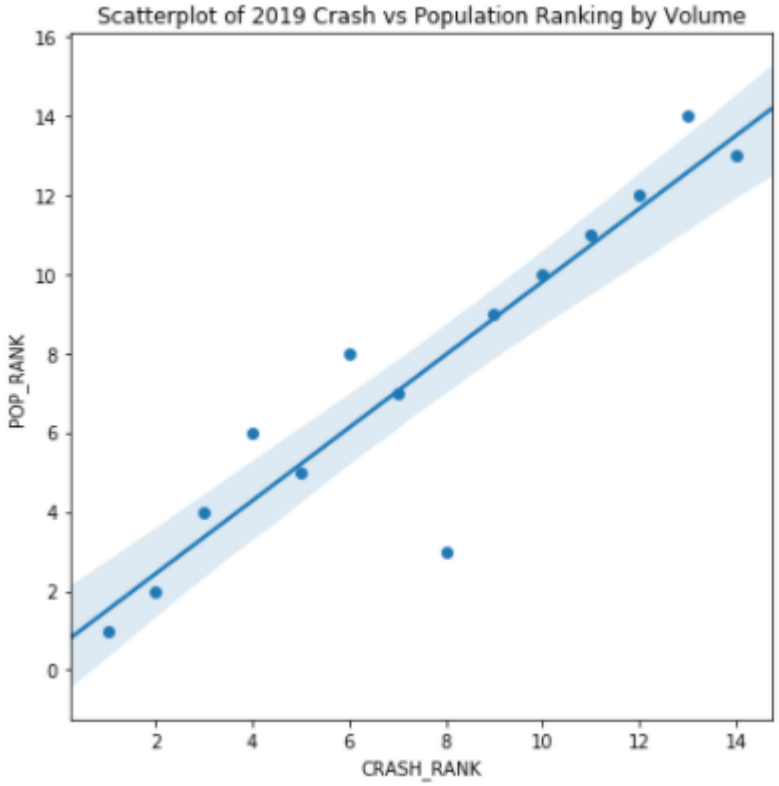

Figure 4: Relationship between Crash Ranking and Population Ranking

Figure 4 has one point that is an outlier/anomaly. This point has a quite high ranking in Population (a low number == high ranking) compared to ranking quite low in Crashes (high number == low ranking). It is for SUFFOLK county (coordinates: 8, 3). This needs to be looked into further to see what SUFFOLK county is doing that the other counties are not doing so that they can implement changes that will help mirror SUFFOLK County's Crash vs Population Ranking

Variances: Counties whose Population Ranking are not equal (not comparable) to the Crash Ranking

Even though there is a general positive direct correlation between Crash Ranking and Population Ranking, there are a few counties that somewhat fall outside the regressing line. Counties with negative RANK_DIFF are doing great; they have far fewer Crashes when compared to Population volume. Counties with positive RANK_DIFF need to improve; they have a larger volume of Crashes when compared to Population volume

CNTY_NAME	POPULATION	POP_RANK	CRASH_VOL	CRASH_RANK	EQUAL	CRASH_RATIO	CRASH_POP_RATIO_RANK	RANK_DIFF
SUFFOLK	807252	3	6187	8	False	130	1.0	-5
ESSEX	790638	4	16394	3	False	48	9.0	1
BRISTOL	564022	6	15293	4	False	36	13.0	2
HAMPDEN	470406	8	13581	6	False	34	14.0	2
DUKES	17352	13	184	14	False	94	2.0	-1
NANTUCKET	11327	14	222	13	False	51	6.0	1

Figure 5: Counties whose Population Ranking are not equal (not comparable) to the Crash Ranking

3.3 Comparing Fatalities per County & City: Volume Analysis

Even though it has been determined that there is a positive direct correlation between Crash Volume and Population Volume, further analysis reveals that this might not be the case when comparing Crash Volume to Fatalities Volume. Even though MIDDLESEX is the county with the most Crashes, the county with the most fatalities during the 2019 period is BRISTOL. Boston City is in a County that had the most fatalities but the least crashes

CNTY_NAME	NUMB_FATAL_INJR
BRISTOL	51
WORCESTER	47
MIDDLESEX	43
HAMPDEN	42
ESSEX	35
NORFOLK	33
PLYMOUTH	32
SUFFOLK	24
BERKSHIRE	13
BARNSTABLE	7
HAMPSHIRE	6
FRANKLIN	5

CNTY_NAME	CITY_TOWN_NAME	NUMB_FATAL_INJR
SUFFOLK	BOSTON	20
HAMPDEN	SPRINGFIELD	9
WORCESTER	WORCESTER	8
BRISTOL	ATTLEBORO	8
BRISTOL	TAUNTON	8
HAMPDEN	CHICOPEE	8
ESSEX	METHUEN	6
MIDDLESEX	MARLBOROUGH	5
HAMPDEN	WEST SPRINGFIELD	5
HAMPDEN	HOLYOKE	5

Figure 6: Fatalities by County and by City

3.4 Comparing Crash Severity per County

The data reveals that there are two counties that had no reported Crash Fatalities in 2019; DUKES and NANTUCKET. This is great news that should be shared with all counties in Massachusetts and be looked at as ideal counties as far as crash fatalities are concerned.

CNTY_NAME CRASH_SEVERITY_DESCR	BARNSTABLE	BERKSHIRE	BRISTOL	DUKES	ESSEX	FRANKLIN	HAMPDEN	HAMPSHIRE	MIDDLESEX	NANTUCKET	NORFOLK	PLYMOUTH	SUFFOLK	WORCESTER
Fatal injury	7	12	46	0	34	5	38	6	40	0	31	32	21	45
Non-fatal injury	1,290	546	3,890	56	3,579	263	3,689	599	6,417	22	3,448	3,064	1,623	4,078
Not Reported	117	76	699	10	653	51	625	87	1,481	51	417	260	327	785
Property damage only (none injured)	3,731	1,980	10,335	117	11,810	923	8,967	2,103	22,198	133	9,940	6,998	4,092	14,195
Unknown	62	44	323	1	318	14	262	59	882	16	237	160	124	565
Total	5,207	2,658	15,293	184	16,394	1,256	13,581	2,854	31,018	222	14,073	10,514	6,187	19,668

Figure 7: Crash Severity by County

Visualization: Severity Description

The data reveals that most crashes result in **Property damage** i.e. 70.1%, followed by **Non-fatal Injury** at 23.4%. These account for 93.5% of the effects of crashes. Even though the goal is to reduce all crashes, a closer look at the majority severity descriptions will help figure out what to do to reduce them and in turn reduce overall crash volume.

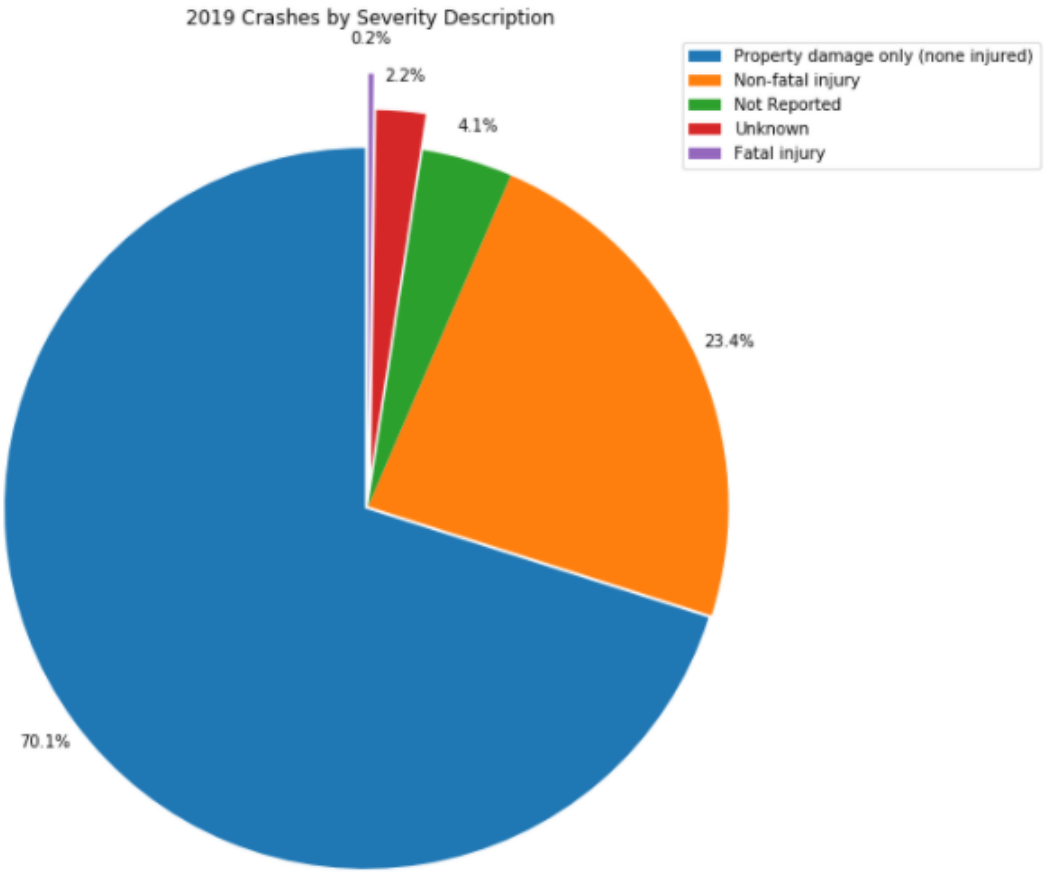

Figure 8: Percentage Proportions of Severity Descriptions

Visualization: Severity Description Volumes by County

The graph below shows that MIDDLESEX County has the highest volume of "Property Damage only" incidences and all "Injuries". The second County with such high volumes is WORCESTER and the third is ESSEX. These are where the most of crash volume reduction will need to start.

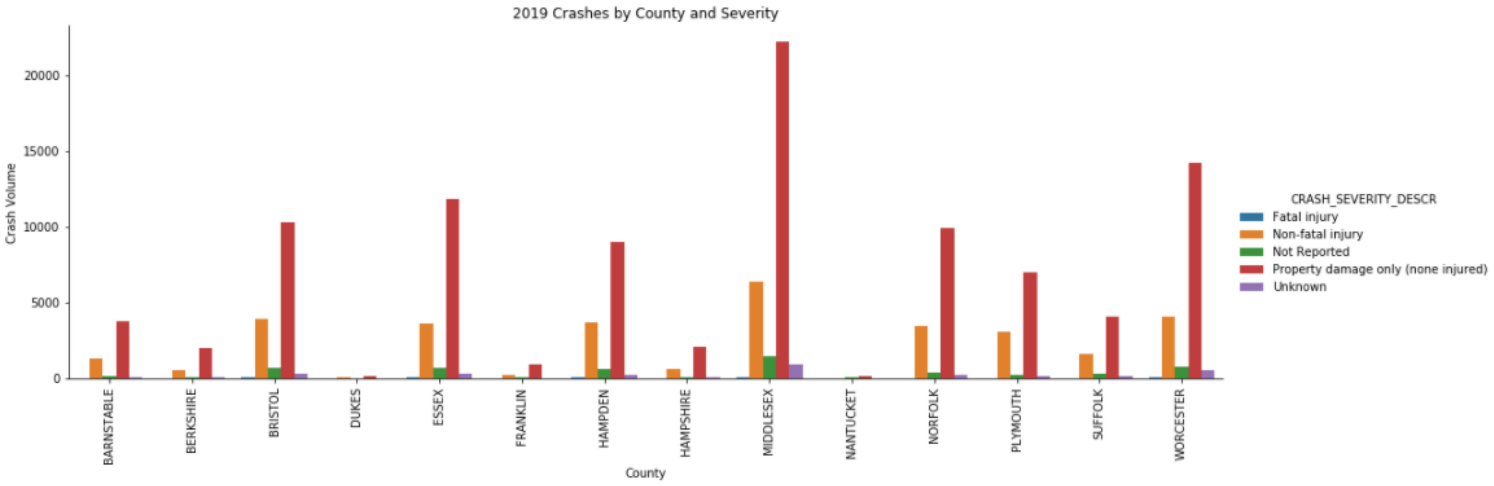

Figure 9: Severity Descriptions Volumes by County

3.5 Crash Lighting Statistics/Proportions

It does make sense that most crashes occur during daylight since this is when most people are out.

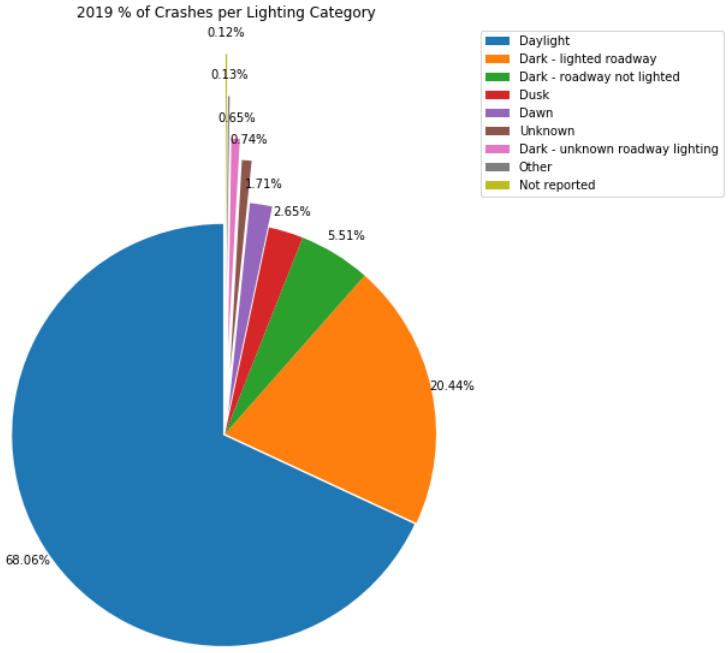

Figure 10: Crashes per Lighting Category

4. Discussion & Recommendations

- Efforts for changes need to begin in the counties with the highest volume of crashes; there is a greater opportunity for improvement in such counties (Section 3.1)
- The variance between the county with the highest vs lowest Crash Ratio is quite high. It shows that there is room for improvement for counties with lower Crash Ratios (Section 3.2)
- Population is a pretty good predictor of crash volume
  - In most cases, county population compared to crash volume relationship is consistent and we would expect it to continue to be so as population increases year after year. The goal is to have an overall decrease in crashes as time goes by (Section 3.2)
  - The committee tasked with this project's mission needs to look at what SUFFOLK county is doing for it to have such a low volume of crashes compared to its population (Section 3.2)
- A closer look at the high volume of fatalities in BRISTOL is needed to see what can be done to reduce this number (Section 3.3)
- Focus on reducing "Property Damage" incidences and all "Injuries" should start in MIDDLESEX County. This county has the most volume of crashes and the most volume of this severity description. (Section 3.4)
  - Counties with fewer crashes per person should share with MIDDLESEX some ideas of why their numbers are lower.
- Massachusetts DOT should ensure that all police officers are encouraged and trained to record/populate **severity** data and specific **lighting conditions** for every accident that they enter into their systems so that the 'Not Reported" and "Unknown" severity and lighting may become bucketed & addressed appropriately. (Section 3.4)
- Massachusetts DOT should determine if the 31.94% crashes that happen when it is not daylight happen in areas where lighting needs to be improved upon and why some dark roadways have no lighting. Such areas need to have lighting installed (Section 3.5)

5. Conclusion

In this study, I analyzed the relation between crash volumes vs population volumes, crash severities and crash lighting descriptions. Overall, I determined that population volume is a good predictor of crash volume as seen on the regression line on Figure 3 and when I calculated the correlation value (0.93). This can help the state predict what the future volume crashes are going to look like as the population increases if no action is taken to reduce these numbers. If proper action and follow up is done regarding these findings, there is great opportunity and probability of reducing crash volumes in the whole state of Massachusetts since a few counties already have low crash volumes when compared to their respective population volumes.