Analyzing & Predicting Massachusetts Crash Data

1. Introduction

1.1 Background

Crash records have attributes that help provide data and statics regarding the kind and severity of accidents on the roads.
This information provides statistics of crashes over time and will be used in determining where to implement transportation and infrastructure changes.

1.2 Problem & 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.
The 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.
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 plan of action can be set in place in order to achieve this goal
Budget allocation for this mission will depend on the scope of how much work is recommended through the analysis done.

2. Data Acquisition

2.1 Data Sources

1st data set is from the following 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 2nd data set showing 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

- 1st data set has 139,109 rows and 116 features
- Cleaned data contains 16 features only

3. Methodology & Results: Study of Counties and Towns

3.1 Crash Statistics by County

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

- MIDDLESEX is the county with the most Crashes.

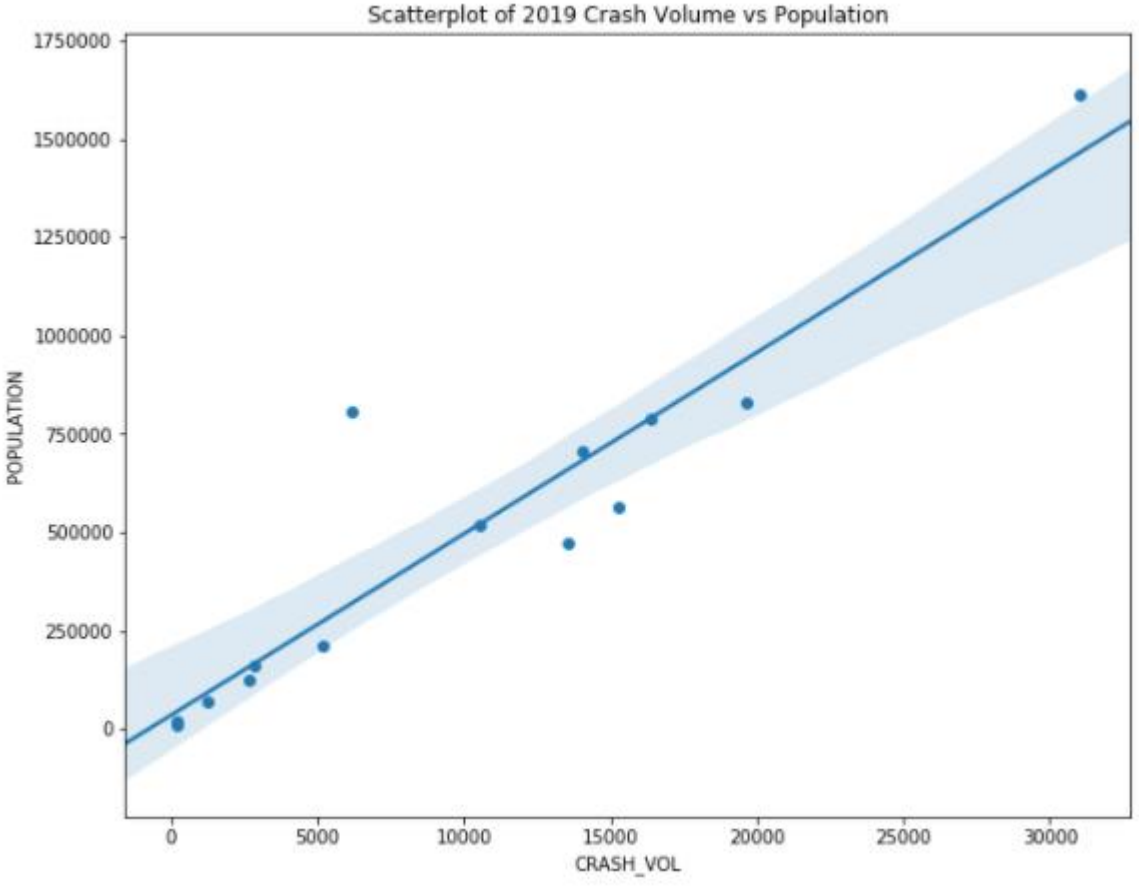
3.2 Comparing Population to Crashes

Sorting and Ranking

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

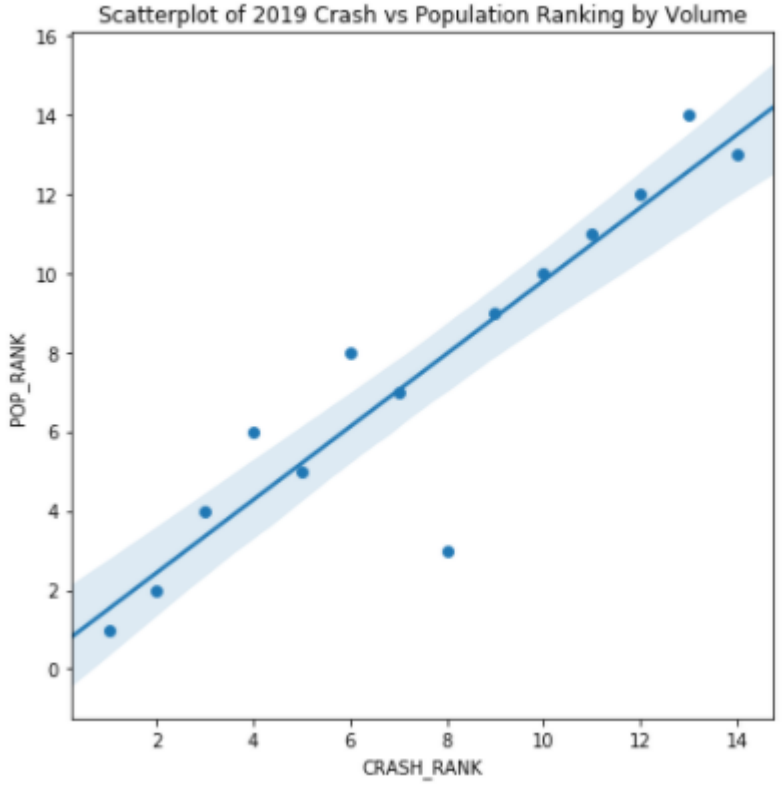
- The mean crash Ratio in Massachusetts as a whole was 56 in 2019 i.e. there was an average of 1 accident per 56 people
- SUFFOLK County has the highest Crash Ratio value i.e. 1 accident per 130 people.
- HAMPDEN County has the lowest Crash Ratio value i.e. 1 accident per 34 people

Correlation: Relationship between Crash Volumes and Population



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

Descriptive Statistics & Cluster Analysis: Relationship between Crash Volume Ranking and Population Ranking



- This scatterplot 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 ratio

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

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

- 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

3.3 Comparing Fatalities per County & City: Volume Analysis

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

- Even though MIDDLESEX is the county with the most Crashes, the county with the most fatalities during the 2019 period is BRISTOL.

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

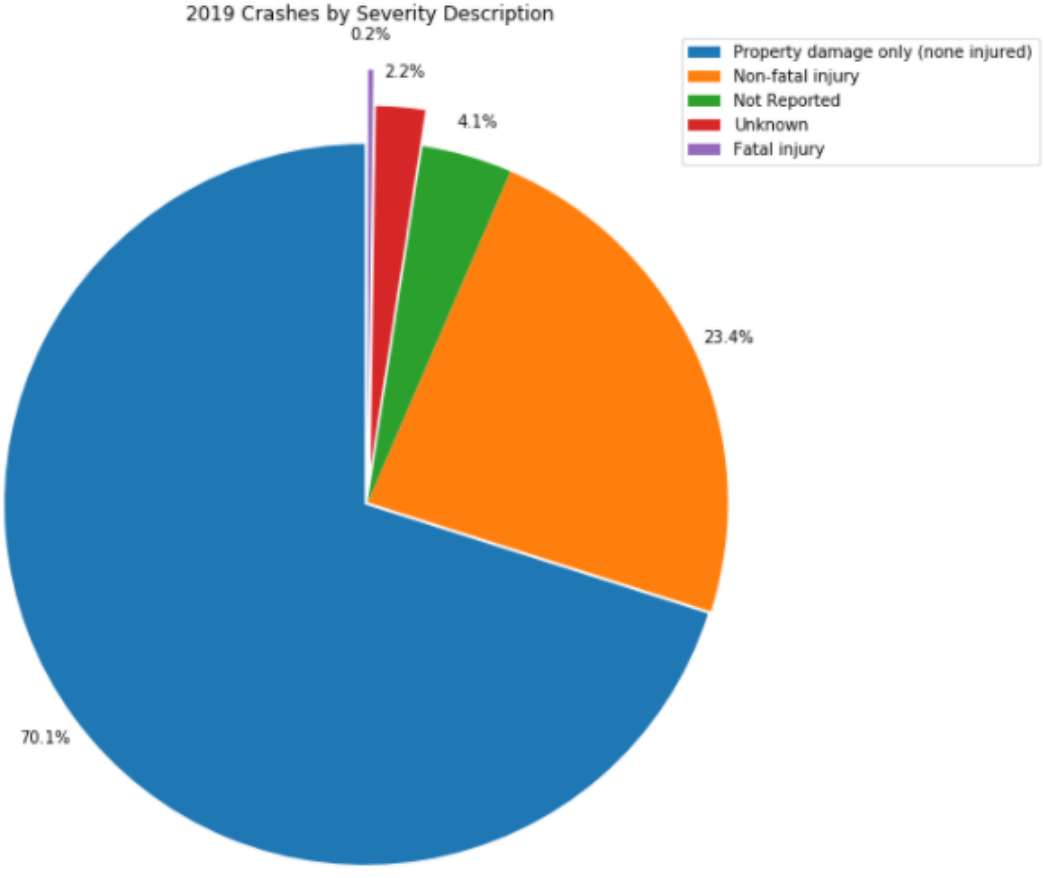
- Boston City in a County that had the most fatalities but the least crashes
- The circumstances surrounding this need to be looked into because it is an anomaly

3.4 Comparing Crash Severity per County

CNTY_NAME CRASH_SEVERITY_DESCR	BARNSTABLE	BERKSHIRE	BRISTOL	DUKES	ESSEX	FRANKLIN	HAMPDEN	HAMPSHIRE	MIDDLESEX	NANTUCKET	NORFOLK	PLYMOUTH	SUFFOLK	WORCESTER	T
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	

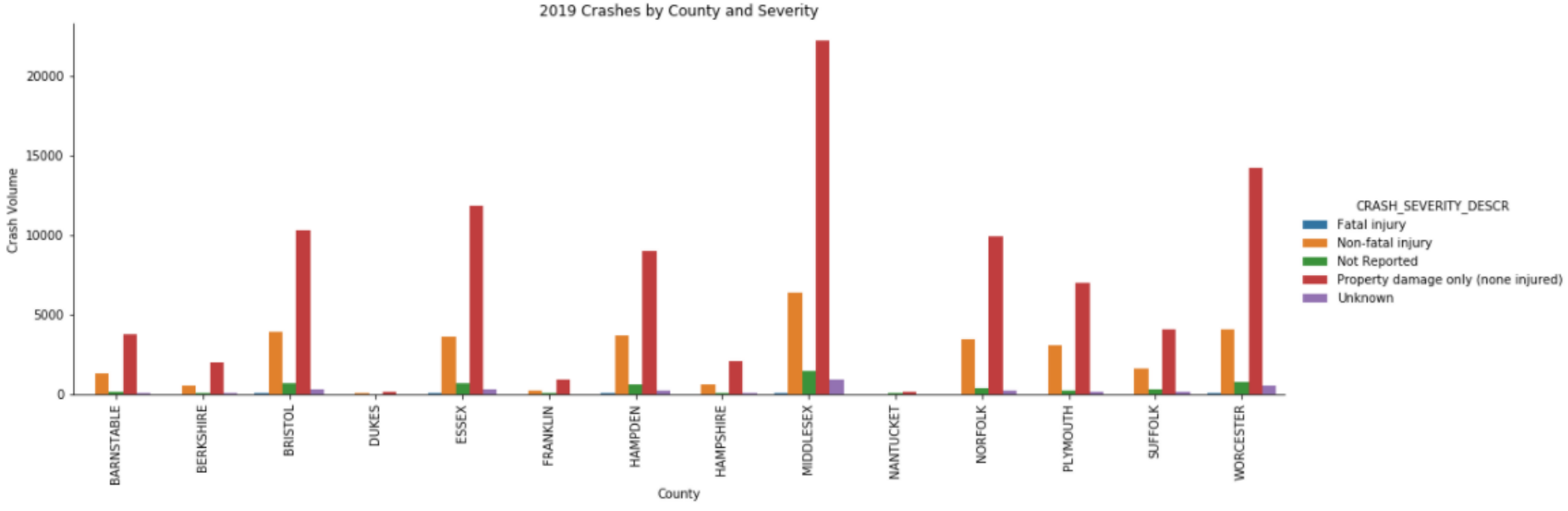
- There are 2 counties that had no reported crash related fatalities in 2019: DUKES and NANTUCKET. This is great news!

Visualization: Severity Description



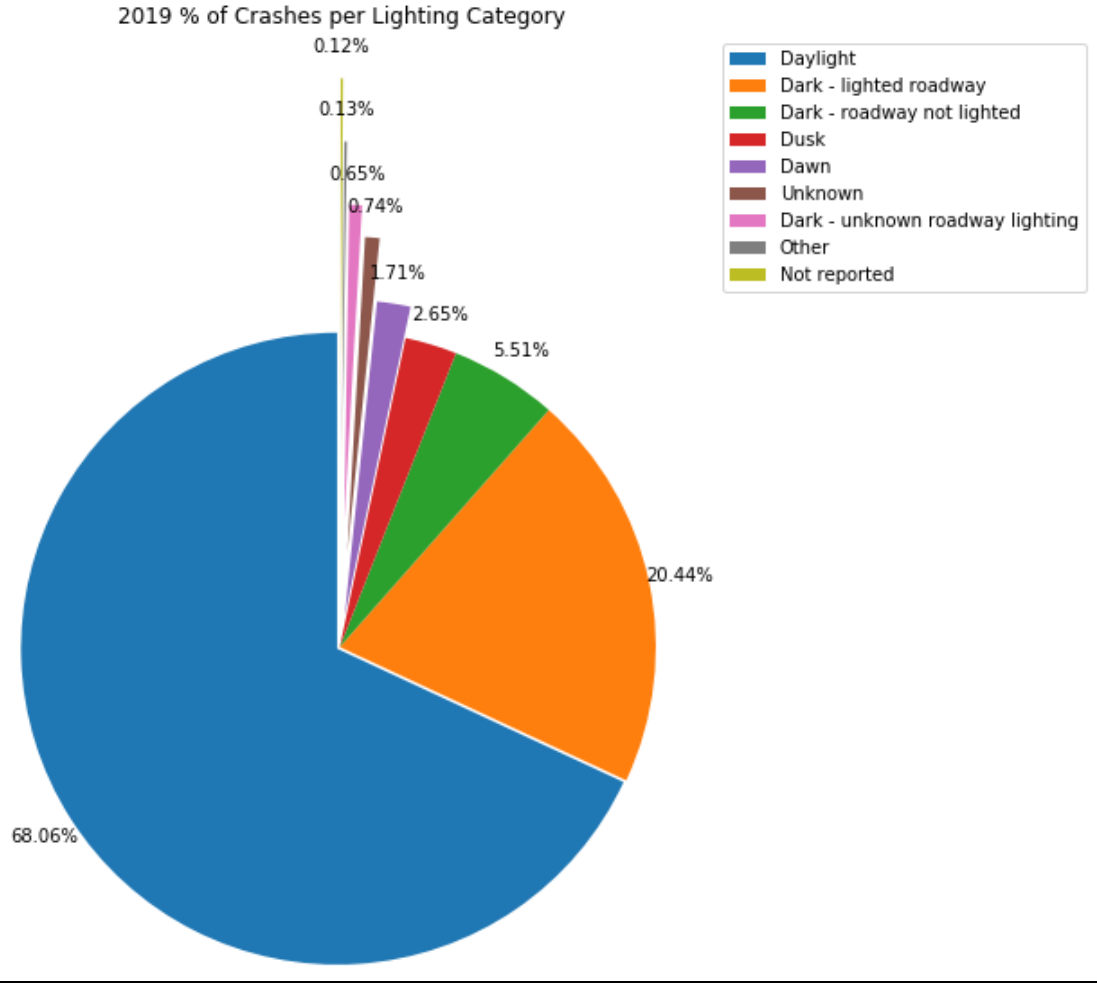
- Most crashes result in ***Property damage*** i.e. 70.1%, followed by ***Nonfatal injury*** at 23.4%. These 2 descriptions account for 93.5% of the effects of crashes; the goal is to reduce all crashes thus reducing all the consequences

Visualization: Severity Description by Count



- MIDDLESEX County has the highest volume of "Property Damage" incidences and all "Injuries"

3.5 Crash Lighting Statistics



- Most crashes happen during daylight. Specific circumstances contributing to these crashes need to be determined.

4. Discussion, Recommendations & Conclusion

- 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)

- *If proper action and follow up is done regarding these findings, there is a great opportunity and probability of reducing crash volumes in the whose state of Massachusetts since a few counties already have low crash volumes when compared to their respective population volumes*