Project report for the course of IMSE – 586: Big data analytics and visualization

Project title:

Analysis and comparison of traffic crash data of Southeast Michigan for the year of 2015 and 2020.

Summitted to

Fred Feng

Assistant Professor

Industrial and Manufacturing Systems Engineering

University of Michigan-Dearborn

Submitted by

Apurba Saha

Student of IMSE-586 &

MS in Information Systems and Technology

University of Michigan, Dearborn

Abstract:

Reducing traffic crashes has emerged as a burning challenge of mankind over the past few decades for the entire world. The developed countries like United States spend massive amount of money for compensating the losses associated with traffic crashes; therefore, accident data analysis is treated as the supreme research area. Although Southeast Michigan is a very important locality of the Mid-west where traffic accident is a major concern, the research related to traffic crash analysis for this region is very limited. This study tried to analyze the accident data collected from the depository of Southeast Michigan Council of Governments by deploying several statistical tools. The data sets were analyzed by handful libraries of Python like Pandas, Numpy and so on as well as with the help of several plots. In spite of having a difference in total number of crashes between the year of 2015 and 2020, the two separate data sets having same attributes showed very identical behavioral patterns. Both the data sets showed that Detroit City is more prone to accident in the region of Southeast Michigan. The accident rate is apex at 3-6 PM and the nadir point lies in between 3-5 AM which collaborates with the natural human behavior of movements. In addition, winter months have higher number of crashes compared to the other months. The weather and lighting do not affect the accidents significantly because data sets showed that majority of the accidents happened in clear weather condition and in daylight. Although the study does not suggest any strong recommendation, avoiding rush hour and careful driving habit can minimize accidents.

Table of content

Content	Page
1. Introduction	4
2. Research questions	4-5
3. Literature review	5
4. Methods 4a. Data source, 4b. Data description, 4c. Tools to be used, 4d. Data cleaning, 4e. Data implementation and Algorithm.	6-12
5. Result (Data analysis and visualization) 5.a Number of accidents, 5.b Accident by crash severity, 5.c Accidents based on area, 5.d Accidents based on months, 5.e Accidents based on Weekdays, 5.f Accidents based on time of day, 5g. Accidents based on lighting condition, 5h. Accidents based on road condition, 5i. Factors contributions to crashes	12-19
6. Discussion	19-20
7. Conclusions	21
8. References	22

1. Introduction:

Road safety is considered as a major health concern all over the world because related data shows that more than three thousand people around the globe die in a day due to road traffic injury. Moreover, road crashes have a huge impact on economy which is as high as US $ 520 billion per year. In spite of having modern research to minimize the traffic accidents, the increase of number vehicles on road augmented the number and rate of traffic crashes every year. Although the importance of motorized vehicle is undoubtedly accepted by everybody, the overall damage (human and property) caused by vehicles is inevitably a great concern for modern civilization. The study and analysis of traffic crashes not only provide a wider picture of reasons for accidents but also suggest the better way of traffic management and control.

As automotive transport has become a part of our day-to-day life in the USA particularly in Michigan, we cannot imagine a single day without of it. However, we need a lot of work to build an entirely secure and sustainable transportation system. In order to build a sustainable transportation system our government and private agencies need tons of fruitful research-based information. This project work is a very little effort towards this huge journey where the traffic crash data of Southeast region of Michigan will be analyzed scientifically with the help of modern data analytics tools for finding out some fruitful outcomes.

2. Research questions:

lot of accidents are happening every day because of numerous numbers of reasons directly or indirectly related to the road condition, drivers' performance, weather condition, pedestrians' interaction and many other things. In spite of having various reasons for traffic crashes, one or two causes can be identified as the pivotal causes. The area of jurisdiction of this study is the traffic crash data of Southeast Michigan, the Wayne County and other neighboring counties. The reasons for majority of the accidents will be the main category of questions. Here are some research questions that will be searched in this study:

- What are the time and day when most accidents occurred for both 2015 and 2020?
- What are the factors (age, drugs, alcohols) responsible for most crashes for both 2015 and 2020?
- Which areas are more prone to accidents for both 2015 and 2020?
- Which weather condition is more probable for accident for both 2015 and 2020?

- What are the differences and similarities in accidents of year 2016 and 2020?
- What are the recommendations to minimize the accidents based on 2020 data?

3. Literature review:

Reducing traffic accidents is an important public safety challenge, therefore, accident analysis and prediction has been a topic of much research over the past few decades but using small-scale datasets with limited coverage, being dependent on extensive set of data, and being not applicable for real-time purposes are the important shortcomings of the existing studies [Sobhan Moosavi, Mohammad Hossein et el. 2019, 27th ACM]. Furthermore, the traffic crash data analysis mainly uses data mining and machine learning techniques, focusing on identifying factors that affect the severity of an accident. There are a variety of reasons that contribute to accidents in which some of them are internal to the driver but many are external [Shristi Sonal and Saumya Suman, 2018, ICETIETR]

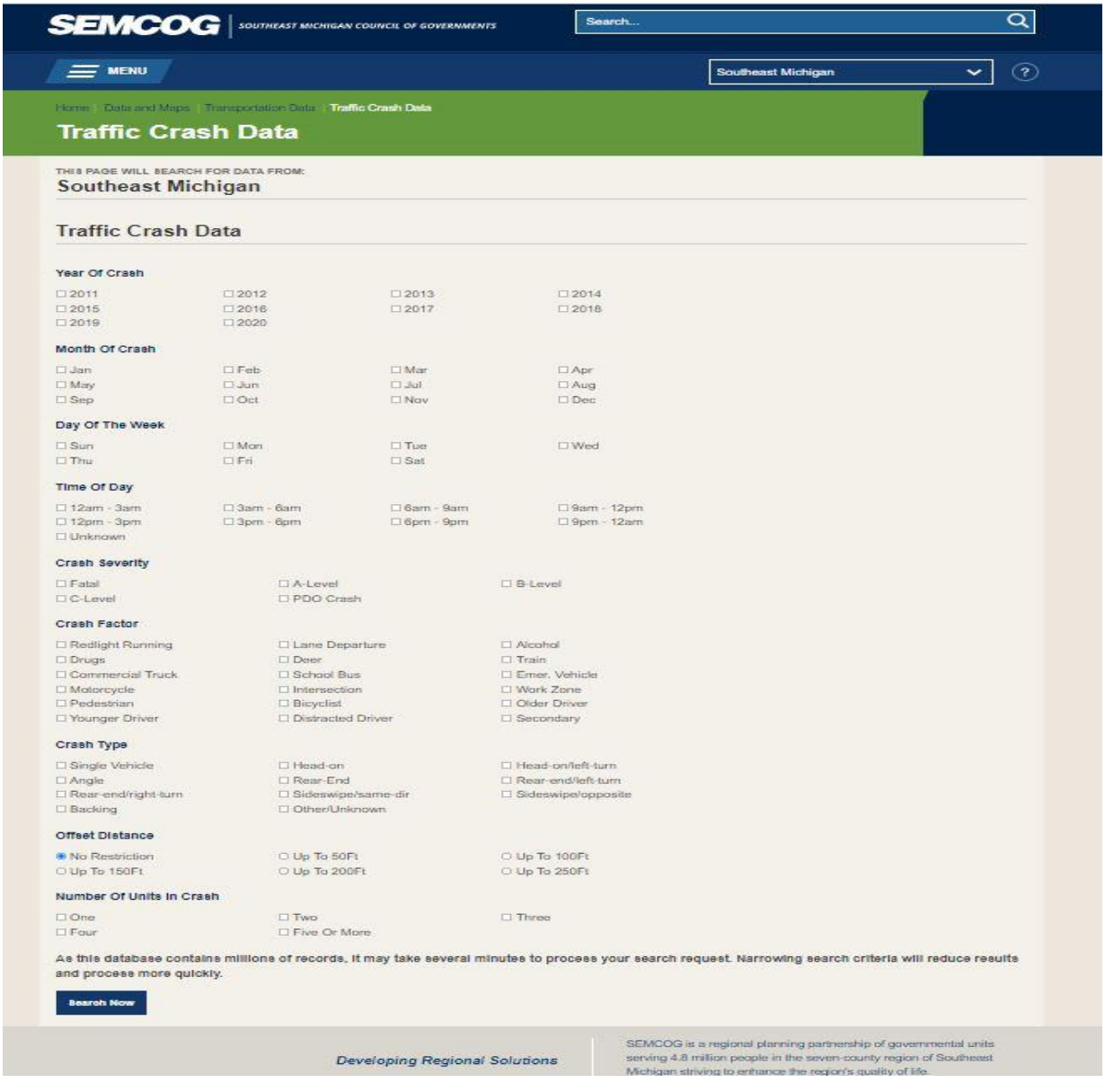
The article titled "Data Analysis of Road Traffic Accidents to Minimize the rate of Accidents" by [Prashant Krishnan, Vivek Chandra et el, 2018, 3rd IEEE] utilized the data sets having nearly one million of tuples consist of numerous features like location, type of accident, lighting condition, speed, severity, alcohol consumption, time and date of accident etc. for analyzing the road accidents. Python open libraries like scikit-learn, Matplotlib, Seaborn, Party were used primarily for statistical analysis and visualization. According to their findings, the tendency of purchasing cars among the general people had increased the rate of traffic crashes throughout the world. They uncovered some patterns and trends to draw helpful conclusions which can be used as recommendations for law enforcing and government agencies. The paper "Data analysis using python" by [Kiranbala Nongthombam and Deepika Sharma, 2021, IJERT] studied the analysis of any suitable data using Python where the very basic processes like cleaning, transforming, modeling were illustrated. They also discussed the pros and cons of using python as a tool and the major steps involved in Python usages.

4. Methods:

This section will discuss the data source and description, data pre-processing, tools to be used as well as data implementation and algorithm. In order to answer the research questions, this section may be considered as the backbone of the entire study.

4a. Data source:

In the case of data analysis, the crucial task is to get right and trustable data as well as understanding the content and structure of the data. The data was taken form https://semcog.org/ for this project. This website is also known as Southeast Michigan Council of Governments which is a public agency.



4b. Data description:

https://semcog.org/ is the website of Southeast Michigan Council of Governments (SEMCOG) that supports local planning through its technical, data, and intergovernmental resources. The works of SEMCOG improves the quality of the region's water, makes the transportation system safer and more efficient, revitalizes communities, and spurs economic development. The traffic crash data sets can be generated as CSV format from this website by setting up some parameters. Two separated data were generated based on the year of 2015 and 2020. Each dataset has the same columns of parameters and they are year, month, day, time, crash severity, crash factors, crash types and number of units. These datasets are believed to be satisfactory for a Python based analysis and comparison. For the case of 2015, the number of records is 108026 and which is 85727 for the year of 2020. The attributes for both the datasets are same and they are displayed below:

Sl no	Attribute	Description
1.	CRSH_ID	This is a unique identifier of the accident record.
2.	ROADNAME	The road name of place of accident.
3.	MILE	The mile location from Detroit downtown of place of accident.
4.	PR	This is a unique identifier of the accident record may be used by law enforcing agency.
5.	MONTH	Month of the accident.
6.	DATE	Date of the month of the accident.
7.	YEAR	Year of the accident.
8.	DAY	Day of the accident.
9.	TIME	Time of the accident.
10.	SEVERITY	Severity of the accident. Fatal, A-Level, B-Level, C-Level and PDO Crash were used to measure the severity of the crash. Fatal means death, A-Level means high injury, B-Level means medium injury, A-Level means low injury and PDO means property damage only.
11.	TYPE	The type of accident. Data was categorized by eleven types and they are Single Vehicle, Head-on, Head-on/left-turn, Angle, Rear-End, Rear-end/left-turn, Rear-end/right-turn, Sideswipe/same-dir, Sideswipe/opposite, Backing, Other/Unknown.

12.	WEATHER	The weather condition at the time of accident. Data was categorized by ten types and they are Clear, Cloudy, Rain, Sleet / Hail, Snow, Unknown, Sever Crosswinds, FOG, Uncoded, Blowing Snow.
13.	LIGHTING	The lighting condition when the accident occurred. A total of seven lighting conditions such as Daylight, Lights, Dark, Dusk, Dawn, Unknown, Other were used in this case.
14.	ROADCONDITION	The road condition at the time of accident. Data was categorized by ten types and they are Dry, Icy, Slushy, Wet, Snowy, Other, Debris, Muddy, Uncoded and Unknown.
15.	OFFSETDISTANCE	Offset distance between the centerlines of the intersecting legs (minor road) at the intersection.
16.	UNITS	The number of units involved in the accident.
17.	FACTOR	The crash factor was defined by one or more combination of seventeen parameters. The parameters are, Redlight Running (R), Drugs (DR), Commercial Truck (C), Motorcycle (M), Pedestrian (P), Younger Driver(Y), Lane Departure(L), Deer (DE), School Bus(S), Intersection(I), Bicyclist(B), Distracted Driver(D), Alcohol(A), Train(T), Emer. Vehicle(E), Work Zone(W), Older Driver(O).

4c. Tools to be used:

Python: Python is an easy to understand and commonly used programming language. it is an object-oriented language used for numerous purposes particularly for data analysis and visualization.

Benefits of Python:

- Data Analysis
- Development of Website
- Development of Application

Why Python?

- Python is available on different platforms (Windows, Mac, Linux, Raspberry Pi, etc.).
- Python is more straightforward as the English language, i.e., simple coding.

 Python has a syntax that allows developers to write programs with fewer lines than some other programming languages.
 Python runs on an interpreter system, meaning that code can be executed as soon as it is written. This means that prototyping can be rapid.
 Python can be treated procedurally, an object-orientated way, or a practical way.
In this project, python has used for data cleaning, data manipulation and data visualization purposes. Jupyter Notebook has used for python. All analysis is done in Python 3.0.

4d. Data cleaning:

Although the data sets have thousands of records, they are almost ready to use for analysis. There is may be some anomalies in the dataset like:
 Null records
 Date format
 Day is missing
 Duplicate records
 Mismatched column

To Address all these anomalies in data, data cleaning is the most important and mandatory step. At this state, data sets will be manually checked for any preprocessing needed to accomplish the required analysis. We can see that there are two unnecessary columns in both the data sets which can be dropped because they will not be required for the analysis. The columns ['PR', 'Unnamed: 17'] can be dropped from both the data sets by the following python commands.

```
crash15.drop(['PR', 'Unnamed: 17'], axis=1, inplace=True)
crash20.drop(['PR', 'Unnamed: 17'], axis=1, inplace=True)
```

The data sets are now completely ready to use.

4e. Data implementation and Algorithm:

Once the preprocessing of data is done, the data analysis comes further. For analyzing the data, I need some tool that can make the task simplified. As I had a robust idea of Python data analytics and visualization, I preferred Python. This programming language not only simplifies a lot of work but also has numerous useful built-in packages for analysis purposes. Anaconda is a free open-source distribution of Python for large scale data processing, analytics and scientific computing.

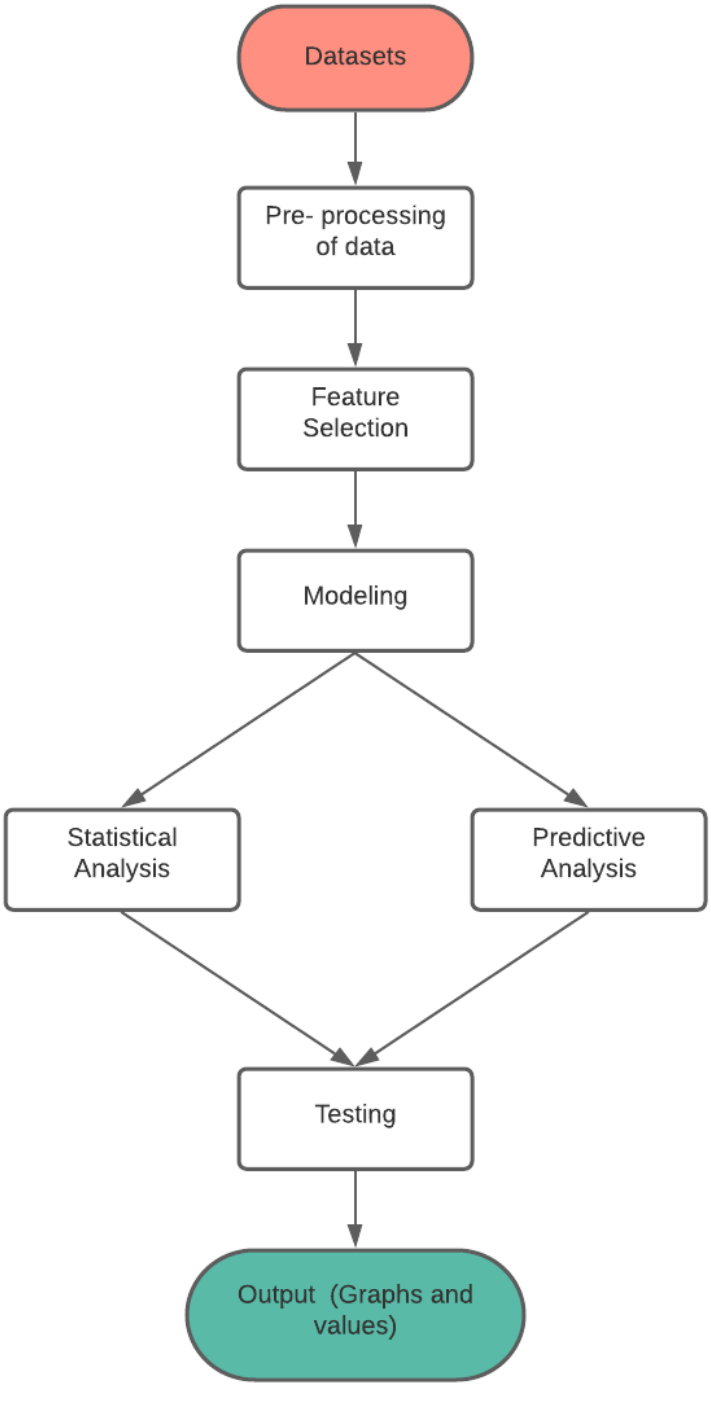
In this analysis, I used Anaconda because it has the most useful packages for Mathematics and it contains all the popular python libraries that can be used for data science. The most important being scikit-learn, numpy, pandas, scipy etc. In addition, it also comes with the jupyter notebook and Ipython distribution. So, it saved work and time from importing numerous libraries separately.

Jupyter notebook was utilized for this analysis where the produced documents contains both Python code and rich text elements such as figures, numbers as well as executable documents which can be run to perform data analysis. The Jupyter Notebook App is a server-client application that allows editing and running notebook documents via a web browser. The Jupyter Notebook App can be executed on a local desktop requiring no internet access or can be installed on a remote server and accessed through the internet. In addition to displaying/editing/running notebook documents, the Jupyter Notebook App has a "Dashboard" (Notebook Dashboard), a "control panel" showing local files and allowing to open notebook documents or shutting down their kernels.

The Python packages pandas and numpy were used widely for data manipulation and analysis purposes. Pandas provides data structures and operations for manipulating numerical tables and time series. On the other hand, numpy stands for "Numeric Python" which is an open source module that provides fast computation on arrays and metrics. The packages which played a major role in the analysis are pandas and numpy. Pandas is used for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It provides high- performance, easy to use structures and data analysis tools. NumPy is the fundamental package for scientific computing with Python that has the following characteristics.

- a powerful N-dimensional array object
- sophisticated (broadcasting) functions
- tools for integrating C/C++ and Fortran code
- useful linear algebra, Fourier transform, and random number capabilities

An algorithm was developed for answering all the research questions which includes the methods, steps and set of Python or other codes for accomplishing the research work. In this case, as there are two separate data sets of traffic accident for the year of 2015 and 2020, all the research questions were for these two years. Below is the algorithm for this study:

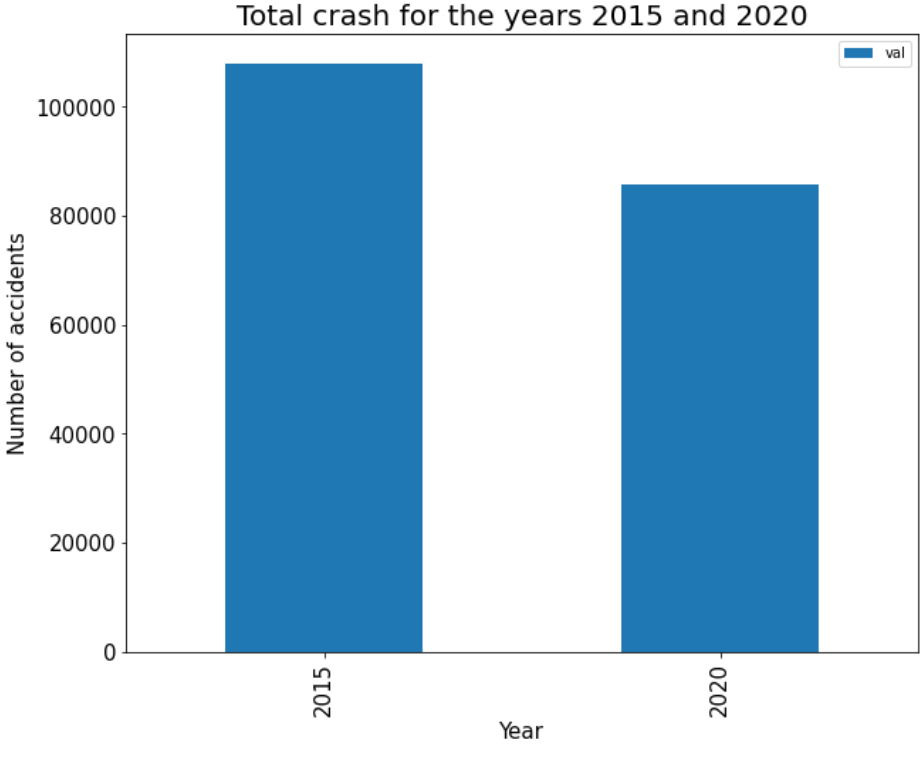


Python open libraries like scikit-learn, Matplotlib, Seaborn, Party was imported and utilized. The essential coding to find out the maximum, minimum and average values or frequencies for required parameters was be calculated when necessary for answering the questions. A number of plots from matplotlib library was created to visualize the facts. In addition, the data was also be inspected for fitting into any model discussed in the lectures.

5. Result (Data analysis and visualization):

5a. Number of accidents:

The number of accidents was 108026 and 85727 for the year of 2015 and 2020 respectively. So, we can see that the total number of crashes reduced in 2020 compared to the year of 2015.



5b. Accident by crash severity:

The table below shows the number of accidents in terms of severity where it can be seen that onl y 354 crashes concurred death out of 108026 for the year of 2015 (0.3%). Most of the crashes ha ppened classified as property damage only.

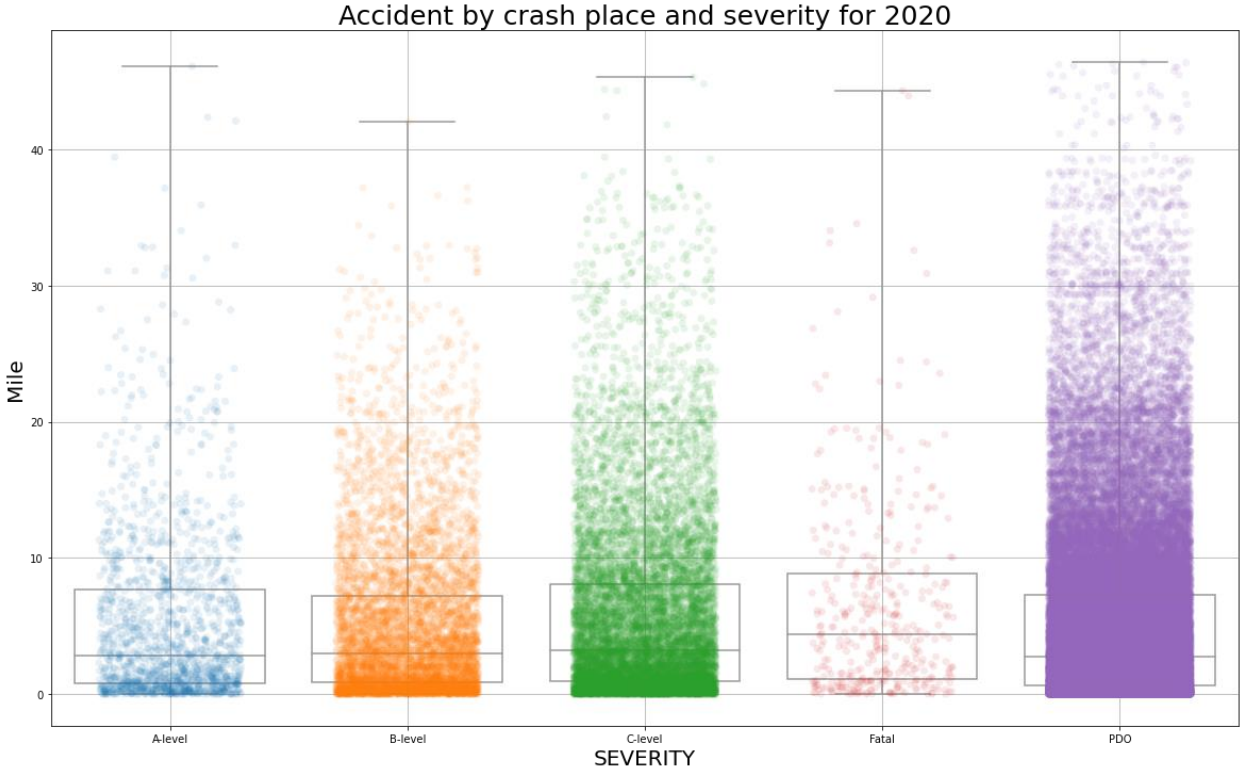
PDO	83773
C-level	16530
B-level	5933
A-level	1436
Fatal	**354**

5c. Accidents based on area:

These data sets have two locational parameters named ‘ROADNAME’ and ‘MILE’ where the ‘ROADNAME’ column does not indicate any analysis significance. The ‘MILE’ column represents the crash location distance in miles from the downtown Detroit. So, based on this ‘MILE’ column, we can analyze the location of crash. Moreover, these datasets have five types of crash severity; Fatal, A-level, B-level, C-level and PDO, where Fatal is the crash associated with death of any person. A-level meaning serious or grievous bodily injury. B-level is medium injury and C-level is mild or little injury. However, PDO means property damage only which has no bodily injury.

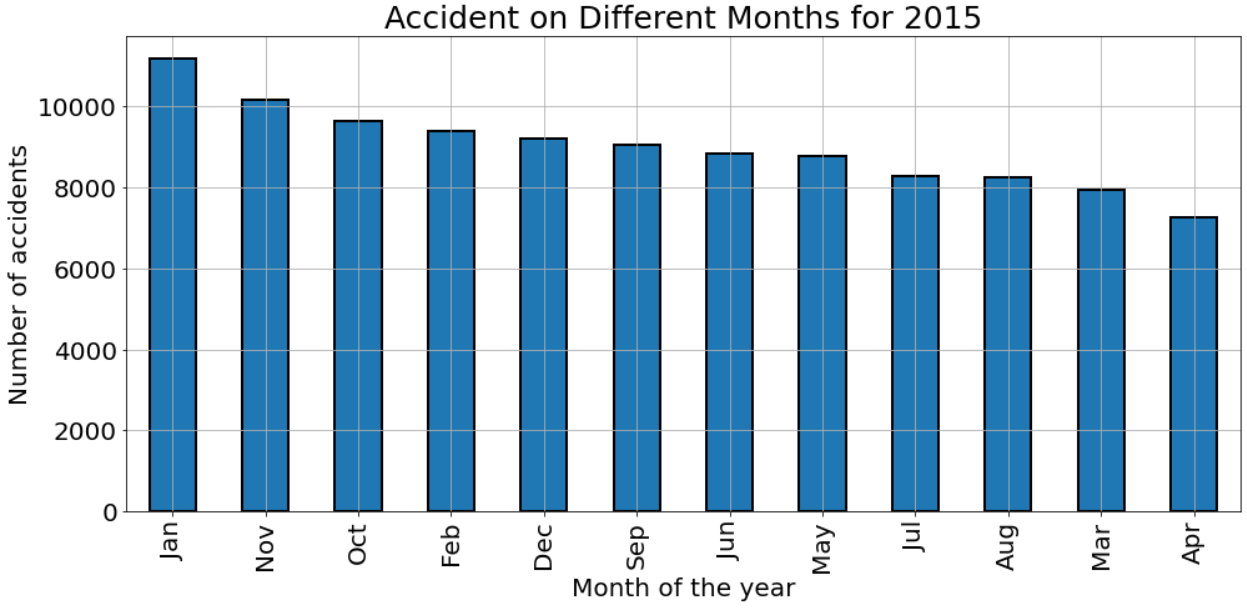
Form the following seaborn strip plots, it can be seen that majority of the accidents take place in Detroit downtown to ten miles. The number of crashes decrease with the increase of distance from downtown. So, most of the accidents occur near downtown for both the year of 2015 and 2020. In addition, both the year shows similar patterns.


Accident by crash place and severity for 2015
Mile
0
10
20
30
40
A-level
B-level
C-level
Fatal
PDO
SEVERITY


Accident by crash place and severity for 2020
Mile
0
10
20
30
40
A-level
B-level
C-level
Fatal
PDO
SEVERITY

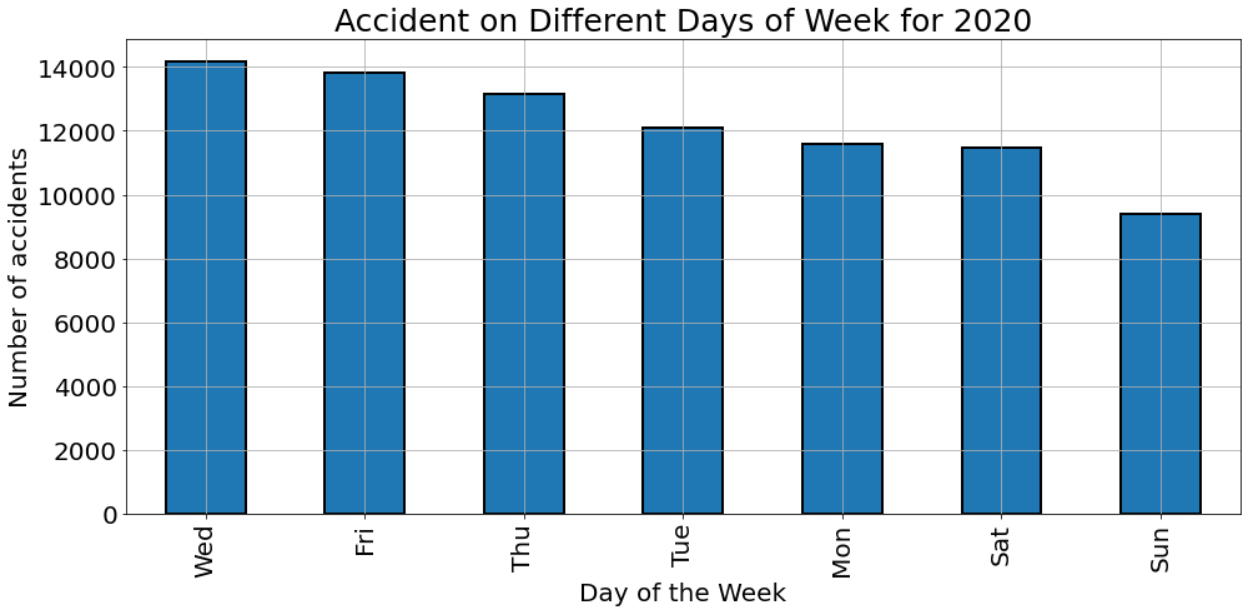
5d. Accidents based on months:

In the region of South East Michigan, the highest number of accidents occurred in January for 2015 and February for 2020. However, for the year of 2020, both January and February have nearly equal values. April has the lowest number of accidents for both the years. So, we may assume that Winter has the highest accident rates and Spring has the lowest for both 2015 and 2020 in the context of Michigan.



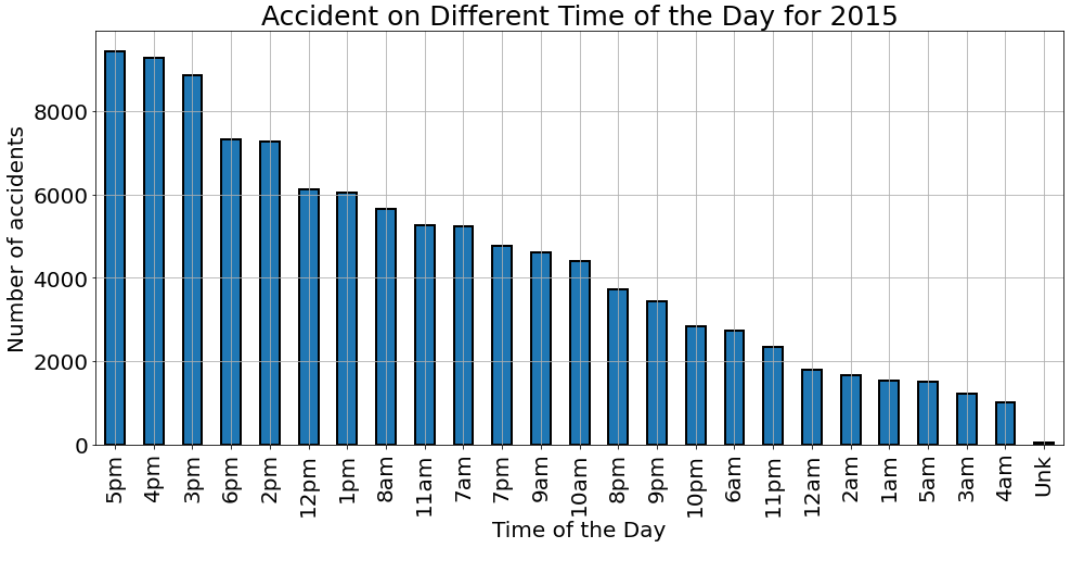
5e. Accidents based on weekdays:

Majority of the crashes occur in Friday and Wednesday for both the years. On the other hand, weekend has the fewest numbers. This result is quite reasonable because most of the people stay home during weekends and number of cars on street is low during this time.



5f. Accidents based on time of day:

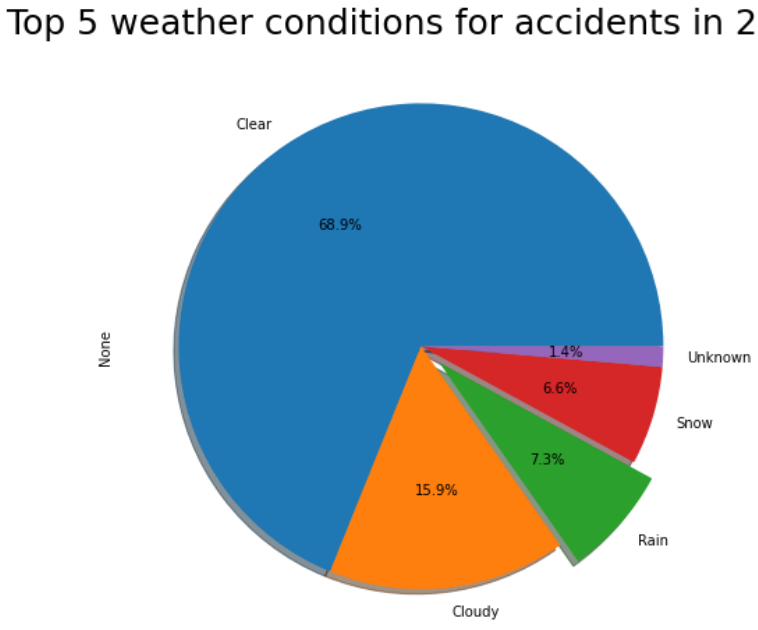
Afternoon time, particularly 3pm to 6pm when most of the people rush to home after work, is more probable to accidents compared with the late night time (3-5pm). This supports the natural philosophy of people's movement.



5f. Accidents based on weather condition:

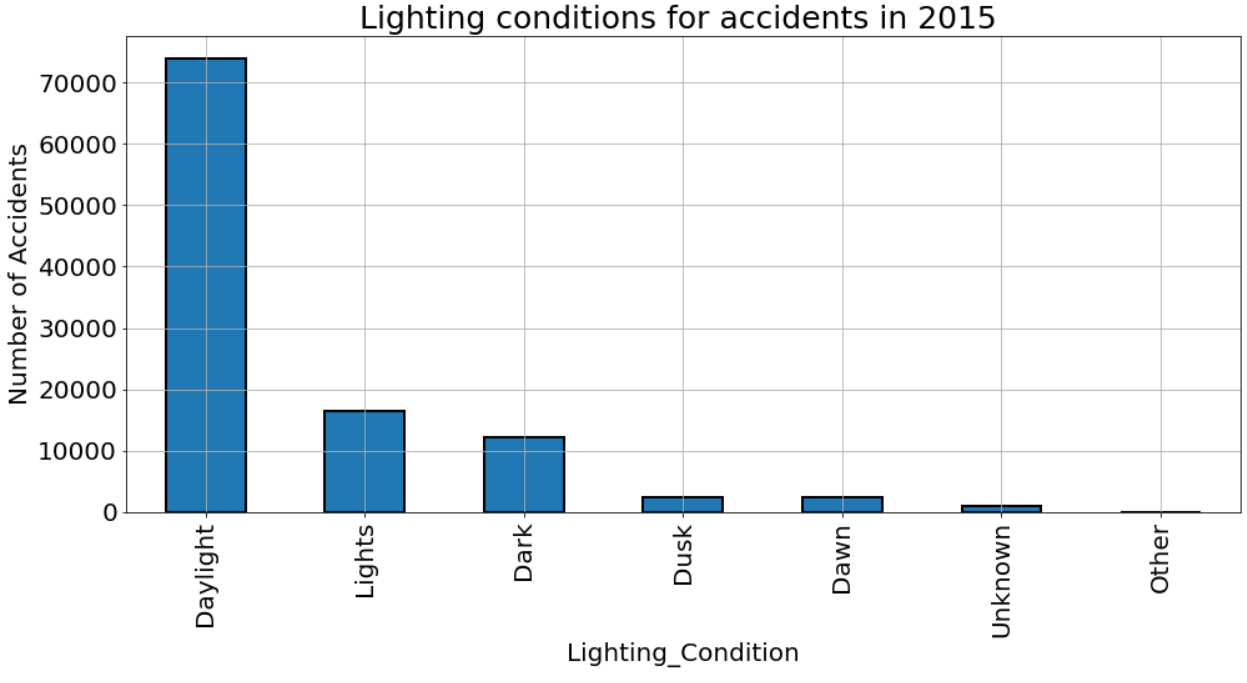
From the data sets of traffic crash of Southeast Michigan, it can be inferred that bad weather condition has little effects on accidents. It is observed that almost 63-68% accidents happened in clear weather condition for both the years of 2015 and 2020. Cloudy and rainy weather condition

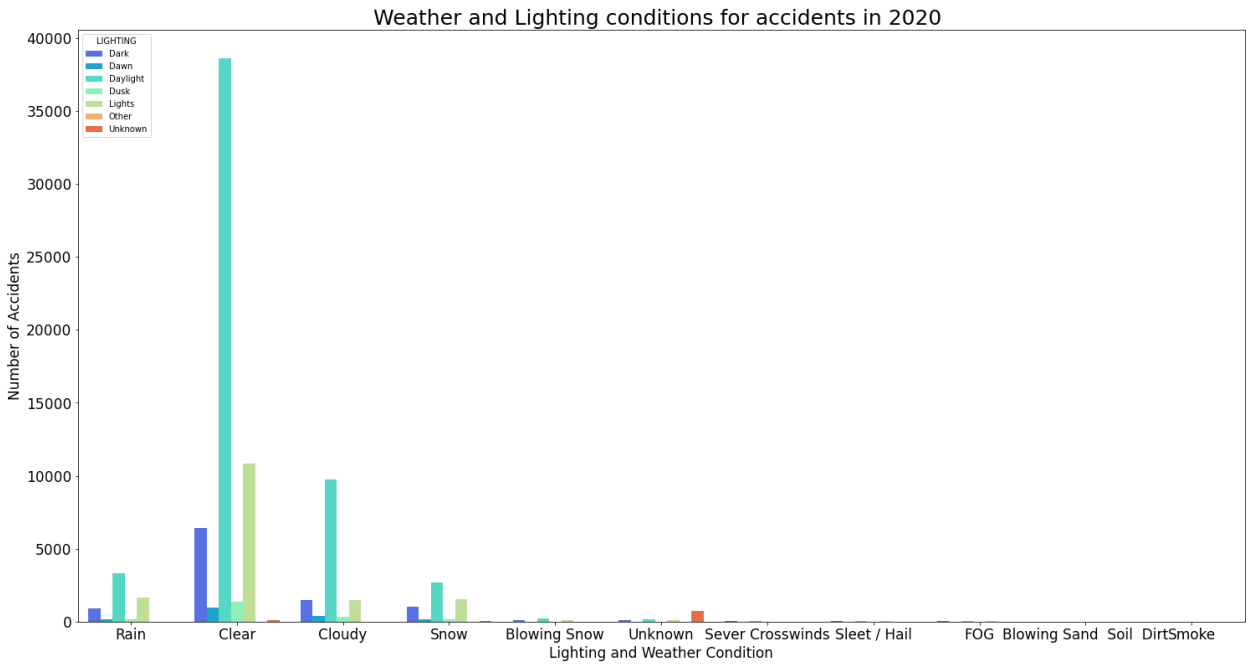
is also responsible but not to a great extent.



5g. Accidents based on lighting condition:

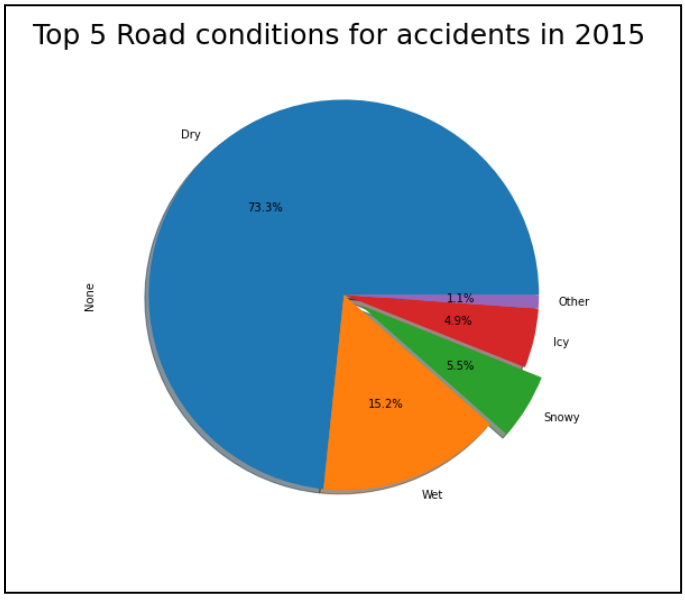
Similar to weather condition, lighting also has a little impact on accidents. It is observed that more than seventy percent crashes happened on daylight for both the years with clear weather condition.

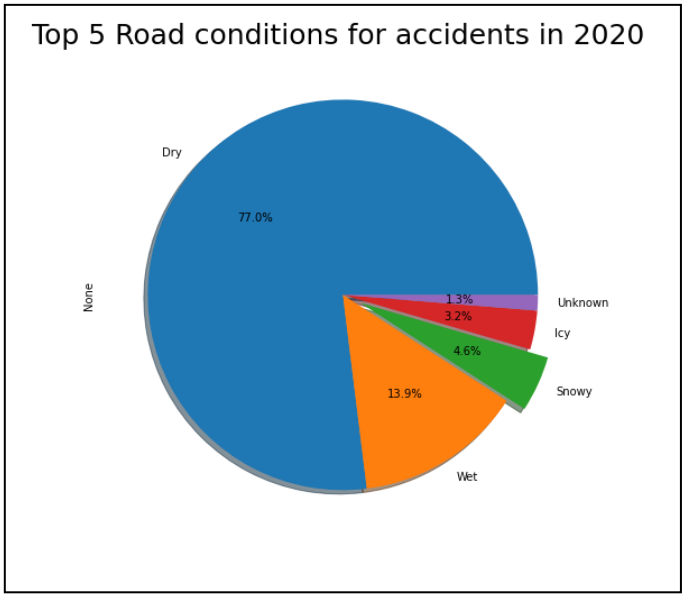




5h. Accidents based on road condition:

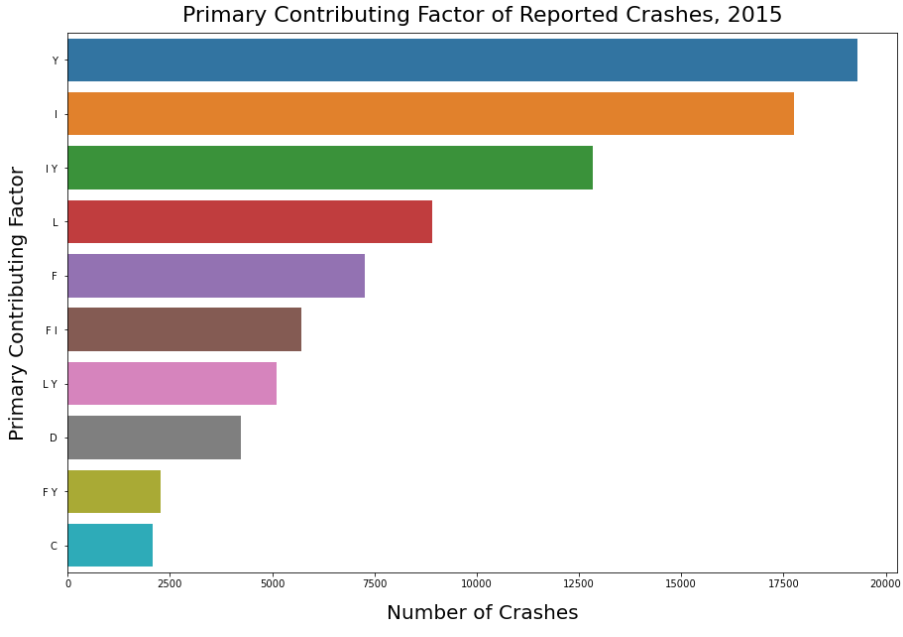
Road condition has negligible effect on traffic accidents because data showed that around 75% crashes occurred in dry roads. Icy and snowy road conditions have very little percentage of accidents.





5i. Factors contributions to crashes:

The top two identified factors reasonable for traffic crashes are younger drivers and intersection points. However, which represents a small portion of this
massive number.



6. Discussions:

Although road accidents are unavoidable but the risk of these crashes can be reduced by simple precautions and thinking ahead (Pablo, 2019). This study tried to find out the answers of some research questions at the same time mentioning some recommendations. The first research question was the time, day and month of maximum and minimum crashes. The study showed that this question followed a similar pattern for both the years where most of the accidents happened on 3-5 PM at Friday and Wednesday in the month of January. Reversely, the least number crashes occurred on 3-5 AM at Saturday and Sunday in April for both the years. So, we can assume that accidents are much higher at office closing time and in winter season. The second research question was to find out the main factors that are responsible for accidents. After analyzing the data sets of Southeast Michigan, we found that the prime contributing factors for accidents are young driver and intersections.

Thirdly, which areas are more probable for accident was the next question. There are two locational parameters named ‘ROAD’ and ‘MILE’ in the data sets where road does not imply much significance. But mile is the measure of distance of accident place from downtown Detroit and that

can be utilized. The tow seaborn strip plots for the year of 2015 and 2020 represent that majority of the accidents occurred in the areas of downtown to ten miles and the number of crashes with the increase of mile. Thus, we can assume that Detroit city is more prone to accident. The fourth question was about the weather and lighting condition responsible for accidents. After careful analysis, we discovered that most of the accident happened at clear weather condition with daylight. So, bad weather condition does not represent accidents.

The fifth question was the similarities and dissimilarities between the behavior of two data sets for the year of 2015 and 2020. From the data analysis and visualization, we observed that the total number of accidents was less for 2020 compared to 2015. But the behavior of data sets in the context of several parameters was almost similar and identical. For this reason, we may assert that although the number of accidents is different in these years, they behaved very identically. The final research point was to make some recommendations that can reduce the number and severity of the accidents. At this point, the writer is unable to make some strong recommendation in this regard but can assert some suggestions that may reduce the accident numbers and severity.

It will be wise to avoid unnecessary driving in the rush hour (3-6 PM) and be extra cautious while driving during this time. In addition to this, drivers should be more careful when crossing the intersection. Nevertheless, people should be careful and sincere during driving will be a must suggestion.

The data sets lack some useful locational parameters to display the accident location on any type of maps which is the main limitation of this study. The data sets have two locational parameters named as ‘ROAD’ and ‘MILE’ which can not be used technically to make a map.

7. Conclusions:

With the augmentation of urban population and automotive vehicles, Traffic accident has become the most crucial public safety issue. Scientists and engineers are relentlessly perusing for sustainable developments that can minimize the number of traffic crashes. A lot of researches are going on for the improvement of automotive communication where traffic crash data analysis plays the most important role. This study based on the data of Southeast Michigan, analyzed and compared the data of two different years (2015 and 2020). The writer tried to find out several parameters responsible for traffic crashes. The data suggested that both the year have very similar patterns where most of the accidents occurred on late afternoon time when the offices end. In addition, Detroit area are more prone to accident in Southeast Michigan region. Although there was no specific strong recommendation, the writer suggested that avoiding unnecessary driving in the rush hour (3-6 PM) may help reducing the crashes. This research can be utilized as a tool for further research and study because the writer found no relevant article based on the data of Southeast Michigan.

8. References:

- Nikam, Swapnil Kisan, "ANALYSIS OF US ACCIDENTS AND SOLUTIONS" Electronic Theses, Projects, and Dissertations. 979, 2020.

- Prashant Krishnan V, Vivek Chandra Sheel, M. V. S. Viswanadh and et el, “Data Analysis of Road Traffic Accidents to Minimize the rate of Accidents”. 3rd IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions, 2018.

- Kiranbala Nongthombam and Deepika Sharma, “Data analysis using python”, IJERT, 2021.

- Najjar, A., Kaneko, S., & Miyanaga, Y., “Combining Satellite Imagery and Open Data to Map Road Safety”, Thirty-First AAAI Conference on Artificial Intelligence, 2017.

- Anderson D, Anderson J., “Electric and hybrid cars: A history. McFarland, Jefferson”, 2004.

- Moosavi, S., Samavatian, M. H., Nandi, A.,Parthasarathy, S. & Ramnath R., “Short and Long-term Pattern Discovery Over Large-Scale Geo-Spatiotemporal Data”, 2019.

- Wolfgang K, Rybicki J, Mauve M, “On the nature of inter-vehicle communication. Communication in Distributed Systems” ITG-GI Conference. VDE, 2007.

- Xu W., Hu S, Cui Y, Ma Y, “Study on Evaluation of Harmonious Development Between Transportation and Economy”, JBJU (Social Sciences Edition) ; 2:004, 2005.

- Mohamed A Abdel-Aty, A Essam Radwan – “Modeling traffic accident occurrence and involvement”, 2000.