1- Introduction

In this report we will analyze crashes happening across the state of Maryland mainly from January 2015 through September 2021. The data is explored from the standpoints of time, fatality, location, weather, vehicle makes and models, license classes, and gender. We explore the time of accidents hourly, daily, weekly, monthly. Also, we provide insight about the number of fatal crashes happening each day by exploring the time of accidents. To explore fatality, we consider objects involved in an accident and briefly touch on the location of the occurrence of it. To further analyze the location, we will explore the mileage of an accident in a particular road to see if we can draw a conclusion. We also provide information about the road divisions which lead to the most number of accidents. At the end of this report, we look at weather, road, and surface conditions simultaneously to see if we can find any relation among these factors and the number of accidents. Due to exploring many factors, some of our observations including vehicle makes and models, license classes, and gender have moved to our code with appropriate explanation. Based on our analysis and our observations we will provide some insights and hypothesize some reasons behind the phenomenon that we have observed. In the end, we will make some suggestions based on exploratory data analysis.

2- Datasets

We are working on three crash datasets (accidents, persons, vehicles) gathered by the Police department of Maryland state from January 2015 through September 2021. The data has been published by the department of information and technology of the state of Maryland, and it is available on their public website. One of these datasets includes details about vehicles involved in accidents. This dataset has 1.39M rows and 49 columns. Another one includes details about people involved in accidents with 1.66M rows and 48 columns. The last one is data about each individual accident that has 743K rows where each one is described by 56 columns.

3- Method

To start our analysis, we came up with questions about different datasets and set answering them as our goal. Then, we identified columns that are associated with our questions and worked on those columns. Of course, identified columns should be cleaned and prepared for our exploratory data analysis. In addition, some analysis demands combining different datasets like accidents and vehicles. This part was the most challenging part because we deal with Big data and handling sizable datasets and merging them takes a lot of computational power. As a result, we applied two remedies. First, we picked the columns of interest and used them merge them to create new data to explore them. Second, we ran our final Jupyter Notebook locally because Google Colab has computational limitations and the data with this volume cannot be handled on that.

4- Main features of crashes

As we stated in the introduction, data is explored from different main standpoints: time, fatality, location, weather, the vehicle makes and models, license classes and gender. In other words, each standpoint is a feature characterizing a single crash. We first go over the time feature and state our results.

4-1- Time

When time is considered, each of these time frames elaborate on different aspects of crashes. For example, daily distribution provides us with the information of rush hours. Monthly distribution will help us to figure out which month of the year is the most crowded month. If a quarter has the highest number of crashes, it probably is the time of the year when people use their cars more. Finally yearly distribution helps to have an overall view of what is going on throughout a year. For example, compared to 2019 the number of crashes has decreased, and we believe that is the effect of the pandemic. While exploring the data time-wise we also looked at fatality features as well. For example we figure out that 1.9 fatal accidents happen daily on average across the state of Maryland. Also, the count of fatal injuries is high at night. In the next subsection, we will go through the fatality feature in detail.

4-2- Fatality

In this subsection, we will explore the fatality by considering Involving objects and Speed limit.

4-2-1 Involving objects

As we studied our datasets, it was observed that the most fatal injuries happen with "Single Vehicles". Since a single vehicle is involved, this suggests that there should be a fixed object with which the vehicle must have collided to turn into a fatal accident. When one looks deeper, she can observe that most of the fatal crashes involving single vehicles, collisions happened with "Tree Shrubbery". This is followed by crashes with Guardrails, Embankments and Curbs. Also, the second highest number of fatal crashes involves no fixed objects. These crashes take place in Right Turn and Left Turn Lanes.

4-2-2 Speed limit

Generally, at the speed limit of 25 we have the most number of crashes and the least number of crashes happen at the speed limit of 75. So the belief that if we drive fast the probability of an accident increases is wrong. By looking at the speed limit of fatal accidents, one can see the most fatal accidents happen at the speed limit 55. So, driving fast would not necessarily lead to accidents and fatalities.

4-3- Location

We analyze crashes from two different aspects based on their locations. One in the amount of mileage that has been passed or left in a particular road.

4-3-1 Mileage in a given road

We have done analysis on the location of a particular accident on a given road. Based on our analysis one can observe that the most number of crashes take place between the first mile of any given road/driveway. The first mile of any given road can be looked at as the entrance of that road where cars want to join a new road from other roads. Interestingly, the second highest number of crashes take place just one mile before the start of any given road. This part of the road is the place where cars try to exit the road. Considering these two observations, one may conclude that the speed change at the beginning or at the end of the road or changing the driving patterns increase the probability of getting involved in an accident. This could be a major reason for most accidents happening in these initial and end mile ranges of a given road. The aforementioned conclusion is drawn because as we move forward in distance, the number of accidents seems to keep decreasing.

4-3-2 Road division

According to the data, the most number of crashes happen in roads when there is no division between the two ways, i.e., in two-ways roads. Among these crashes, the ones with barriers have the most "Same Direction Rear End" collisions followed by "Same Movement Angle" and "Same Direction Sideswipe" collisions.

4-4 Weather, road condition, and surface condition

As opposed to a common belief which assumes accidents happen more in inclement weather, there is no evidence that can support this statement. Based on our analysis, the most accidents happen in clear weather where there are no defects with roads and surface condition is dry.

5- Conclusion

To summarize the complete analysis, maximum fatal injuries happen in the same direction rear end right turn at speed of 55mph, the state government allocating medical forces nearby these roads could be a way to lessen the chances of Fatal/Incapacitating injuries. It was surprising to see that there are more fatal injuries in Clear Weather in comparison to Severe Winds, however weather wasn't a major contributing factor to the analysis. Also, we realized that being on a road with a high-speed limit does not increase the probability of having a fatal crash because fatal crashes happen most at the speed limit of 25 or 55 mph. From the analysis so far, we are certain that maximum number of accidents happened during the rush hours of 7am-9am and 3pm-5pm, with Fatal accidents been the most during the afternoon. With the pandemic hitting worldwide and people working remotely, there was 20% decrease in the crashes during pandemic.

Limitation: We have a lot of values for junction codes, weather condition, road condition, surface condition wherever the values are Not Applicable - the count of accidents is 3rd largest. There are Unknown and Others values as well, so we have not considered for our projections. Hence, we are not able to address those collision types.

6- References

1. https://opendata.maryland.gov/Public-Safety/Maryland-Statewide-Vehicle-Crashes/65du-s3qu
2. https://opendata.maryland.gov/Public-Safety/Maryland-Statewide-Vehicle-Crashes-Person-Details-/py4c-dicf
3. https://opendata.maryland.gov/Public-Safety/Maryland-Statewide-Vehicle-Crashes-Vehicle-Details/mhft-5t5y
4. https://www.maryland-demographics.com/counties_by_population
5. https://stackoverflow.com/questions/17388213/find-the-similarity-metric-between-two-strings
6. https://stackoverflow.com/questions/7331462/check-if-a-string-is-a-possible-abbrevation-for-a-name
7. https://www.canva.com/design/DAEyMM2h0G4/share/preview?token=TU8kG8gUAQwvipMjQ_Acmw&role=EDITOR&utm_content=DAEyMM2h0G4&utm_campaign=designshare&utm_medium=link&utm_source=sharebutton