**FUNDATMENTALS OF DATA SCIENCE ASSIGNEMNT**

**ANALYSIS OF FREQUENCY AND CAUSE OF ACCIDENTS ON & NEAR INDIAN STATE HIGHWAYS**

**Group Number: 3**

**Group Members**

Diksha Mothdiya (23030141019)

Swapnil Kothari (23030141034)

Rhea Shree S (23030141050)

Tushar Bansal (23030141074)

Yash Sharma (23030141080)

Topics covered under the assignment:

* Exploratory Data Analysis
* Problem Statement
* Data Preprocessing
* Summary of basic statistics
* challenges or limitations
* conclusion

1. **EXPLORATORY DATA ANALYSIS**

Exploratory Data Analysis(EDA) is an important step of data analysis and involves examining the dataset to identify patterns, outliers, and anomalies, and generating conclusions based on our understanding of the dataset. It encompasses producing summary statistics for numerical data within the dataset and creating various graphical representations to better understand the data.

Our Datasets are primary data downloaded from Kaggle. We have used Python Libraries, mainly Pandas, for data cleaning purposes. We will be using Microsoft PowerBI for the data analysis process. The datasets we have explored are: Road\_Accidents\_data (Month and Hourly wise) and India\_Injury\_Road Accident\_Fatality\_2017. We have combined the common columns of these datasets such as Year and Time values in the main dataset that contains over 120000 records and 34 Fields. The main dataset contains information about the Road Allignment, Driver Information, the effects and cause of accidents as well as the information about the severity of the casuality.

1. **PROBLEM STATEMENT**

**CHALLENGE:** Lower the number and seriousness of accidents that occur on and near the Indian state highways and roads.

Highway & near highway crashes continue to be a serious problem since they result in a high death toll, serious injuries, and financial losses. Effective preventative measures are hampered by the state's incomplete awareness of the circumstances causing these tragedies.

It is challenging to analyse trends and spot patterns in the fragmented and fragmentary data that is currently available, such as crash and police report data.

1. **DATA PREPROCSESING**

Data gathered from different sources have many fields and records which need filtering and changes to make the data normalized and more effective for the analysts to draw insights and visuals out of it. Steps followed for formatting the dataset are as below:

* **Data profiling:**

The process of examining the datasets gathered from various sources and sorting down the files or the records and fields which are relevant and effective for our analysis.

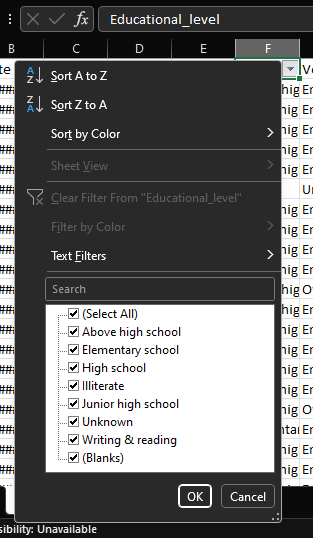
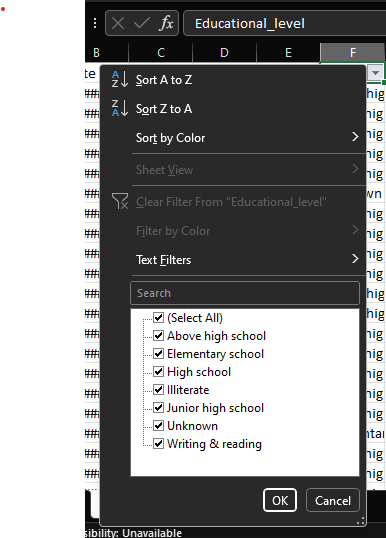
As of our datasets from various sources, there were two data files in .csv formats (normalized with reference to the main Road.csv file) which had data which can be easily derived from the main file Road.csv. So, to reduce the redundant data from processing and reducing the analysation time we decided to finalize and use only one input file that is Road.csv.

Another problem with not including other 2 .csv files is the record count in 2 secondary files is not matching with the primary dataset. It can result into random data generation which might hamper the results generated and the result might be biased. So, we have finalised to use only one dataset.

* **Data cleansing:**

In this step we have modified the data to make it of specific format, which is correct for processing. For example, there are some fields in the dataset which have data which is not correct when the processing is in picture. The field **Educational\_level** of the driver who is involved in the accident is a field.

The records with blanks in this field is a problem when we are processing this data. As there are records with value for the same field as Unknown, we have written a program in Python to replace the **blank values** into **Unknown**.

**Before Python Program Run After Python Program Run**

Similarly, as shown in above screenshots we have replaced the blanks to value Unknown for below-mentioned fields so that the data processing would be easier and effective.

* Educational\_level
* Vehicle\_driver\_relation
* Type\_of\_vehicle
* Owner\_of\_vehicle
* Types\_of\_Junction
* Type\_of\_collision
* Vehicle\_movement
* Work\_of\_casuality
* **Data reduction:**

Removal of data from the input datasets which do not bring any necessary facts or insights through them is a part of this step. Identifying such fields and analysing the effects after we remove them is an important factor considered in this step.

Our selected dataset had certain fields as well as records which were of no use to analyse the problem statement nor of use to draw any insights from them.

Data fields which have Blanks and Unknown value in the field and replacing the value is not a correct approach as it will manipulate the data and it will turn out as randomised data which can not be used to draw insights. Below-mentioned are some of the fields for which we have removed the records if the field value is either Blanks or value Unknown.

|  |  |  |
| --- | --- | --- |
| **Field Name** | **Record Count** | **Values in the record** |
| Sex\_of\_driver | 178 | Unknown |
| Driving\_experience | 829 + 33 | Unknown + Blanks |
| Area\_accident\_occured | 239 | Blanks |

Few other below listed fields are those who has NA value and replacing or removing them will not make any difference in the analysis of the dataset. So instead of changing the field values for the records which have NA values, we removed the records completely.

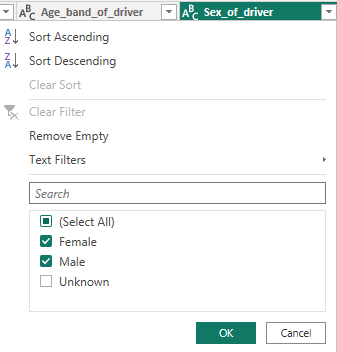
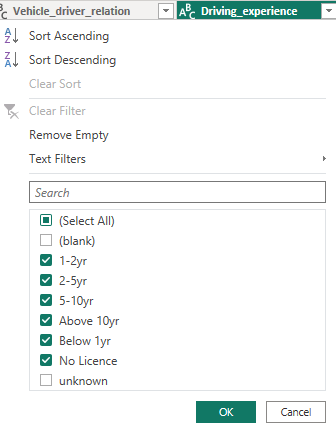
|  |  |  |
| --- | --- | --- |
| **Field Name** | **Record Count** | **Values in the record** |
| Casualty\_class | 4443 | NA |
| Sex\_of\_casualty | 4443 | NA |
| Age\_band\_of\_casualty | 4443 | NA |

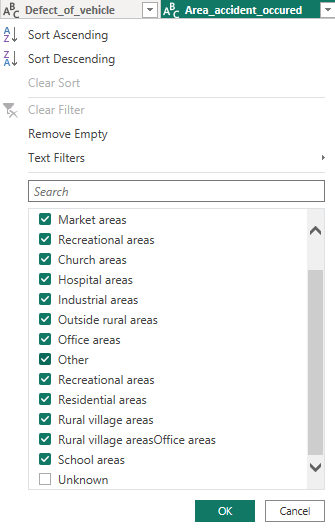
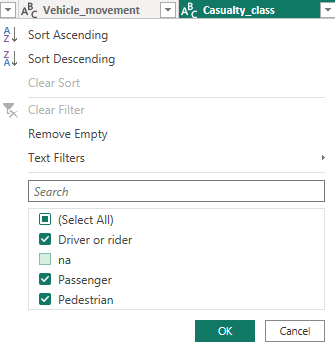
Also, there are few fields which are non-relevant to our analysis and through which we can not draw any insights. So, we have removed those fields. Below-mentioned field is removed from the input dataset before processing.

**Fitness\_of\_casuality >** Non-relevant field from the dataset, removed under data reduction step.

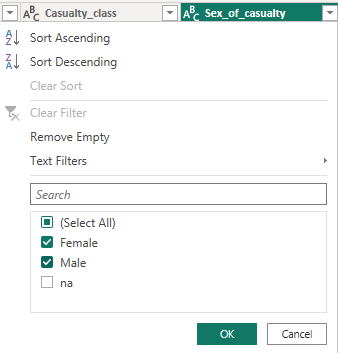
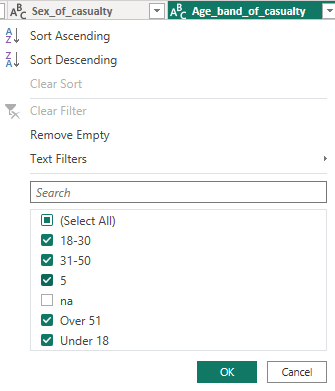
**Screenshots taken while doing data reduction using Power BI:**

Fields which had **Blanks** and value **Unknown** were filtered out before loading the data for further analysis part

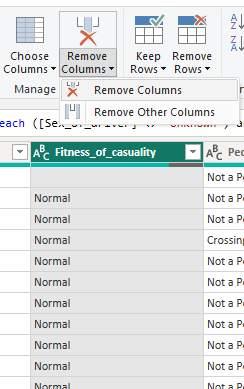
 

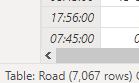
Fields which had value **NA**

**Removing the field column which is not relevant to the analysis from the data**



**Total record count after data reduction step**



* **Data transformation:**

Manipulating the records in the dataset to make it better for analysis and easier for visualization while retaining the original facts and figures and generating non-biased data is done in this step. Changes made in this step are as follows:

For below-mentioned records the Blank cells were replaced with value Unknown.

* Educational\_level
* Vehicle\_driver\_relation
* Type\_of\_vehicle
* Owner\_of\_vehicle
* Types\_of\_Junction
* Type\_of\_collision
* Vehicle\_movement
* Work\_of\_casuality

Instead of completely removing records due to absence of data might result into over-data reduction. To avoid this, we took average/mean values for some of the fields. They are as listed below:

* Service\_year\_of\_vehicle
* Age\_band\_of\_driver

For some fields, to replace the blank values we have used imputation with the mode that is most frequent value in the occurrence. Below are the fields with this type of changes:

* Road\_surface\_type
* Lanes\_or\_Medians
* Road\_allignment
* **Data validation:**

After the data transformation the fields with changes were reviewed to evaluate the replaced values and records which were removed compared to original dataset.

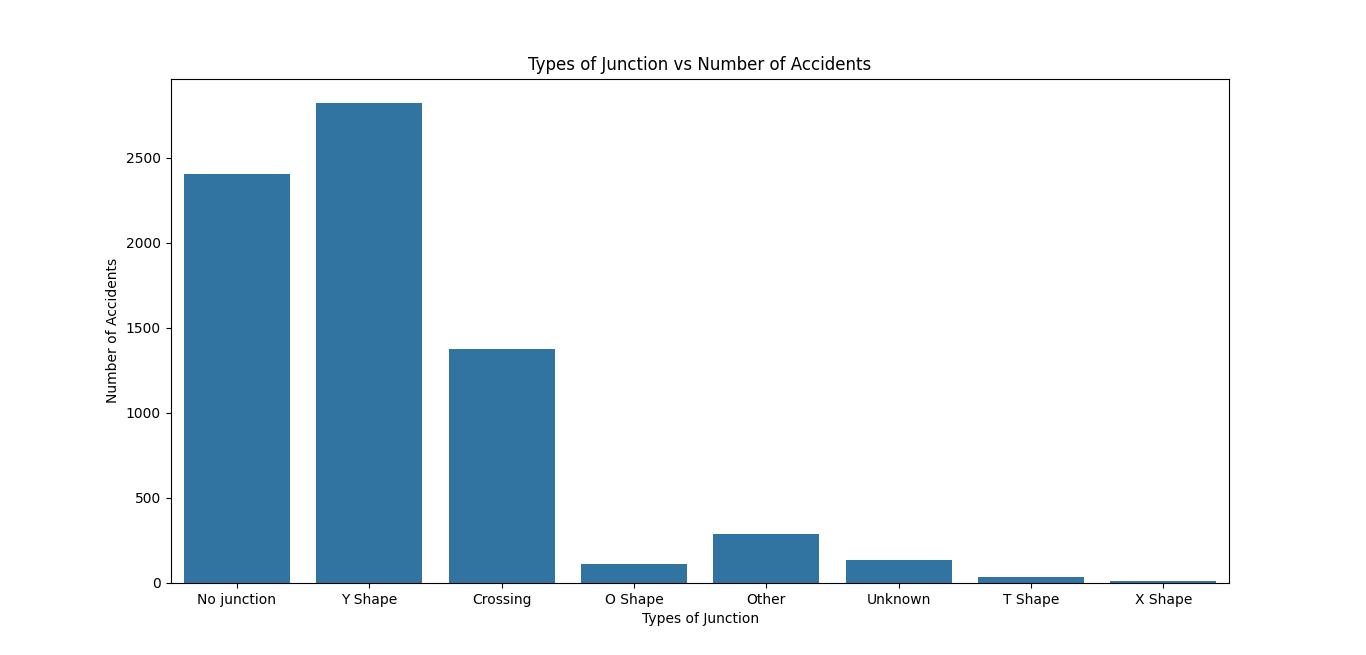
1. **SUMMARY STATISTICS**

After data preprocessing using python program and data analysis under excel spreadsheet application. Below tables are descriptive statistics of two fields from the data.

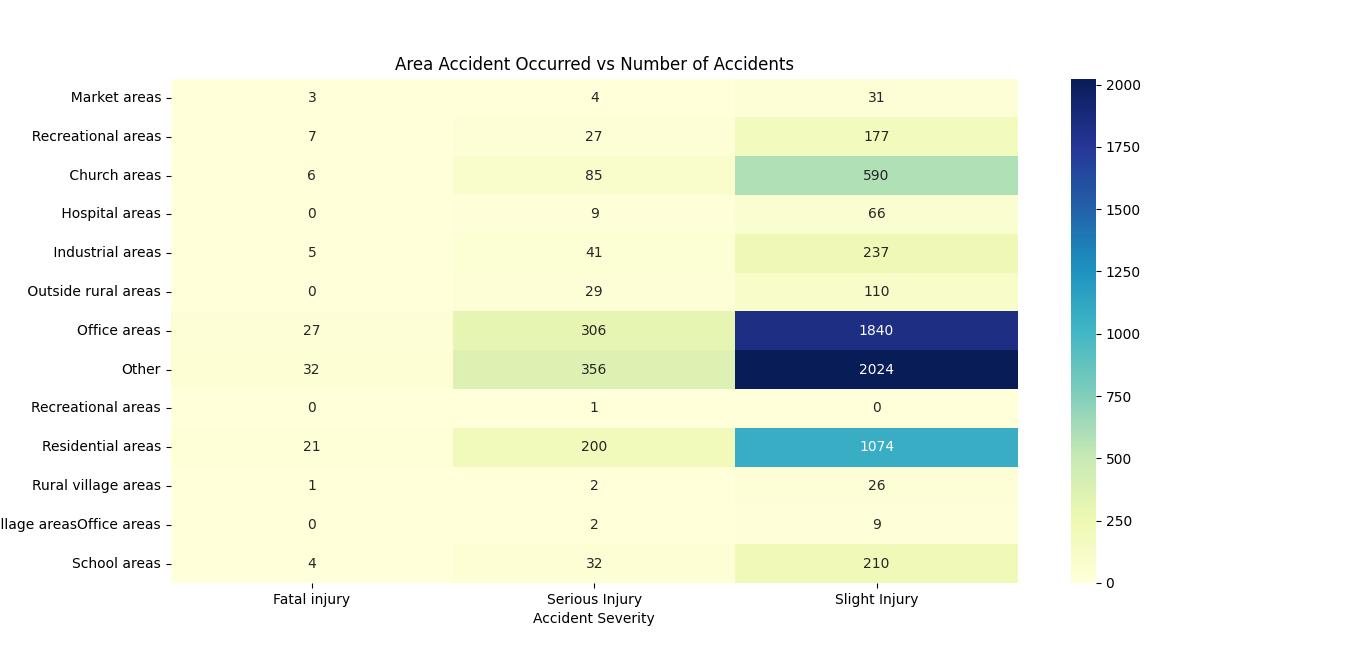
|  |  |
| --- | --- |
| ***Service\_year\_of\_vehicle*** | |
| Mean | 5.778211815 |
| Standard Error | 0.02747549 |
| Median | 5.779745686 |
| Mode | 5.779745686 |
| Standard Deviation | 2.417843082 |
| Sample Variance | 5.84596517 |
| Kurtosis | 0.640467591 |
| Skewness | 0.39519617 |
| Range | 10.5 |
| Minimum | 0.5 |
| Maximum | 11 |
| Sum | 44746.4723 |
| Count | 7744 |

|  |  |
| --- | --- |
| ***Age\_band\_of\_driver*** | |
| Mean | 33.48640293 |
| Standard Error | 0.117633047 |
| Median | 33.65796805 |
| Mode | 24 |
| Standard Deviation | 10.35170815 |
| Sample Variance | 107.1578616 |
| Kurtosis | -0.975722553 |
| Skewness | 0.257218606 |
| Range | 35 |
| Minimum | 17 |
| Maximum | 52 |
| Sum | 259318.7043 |
| Count | 7744 |

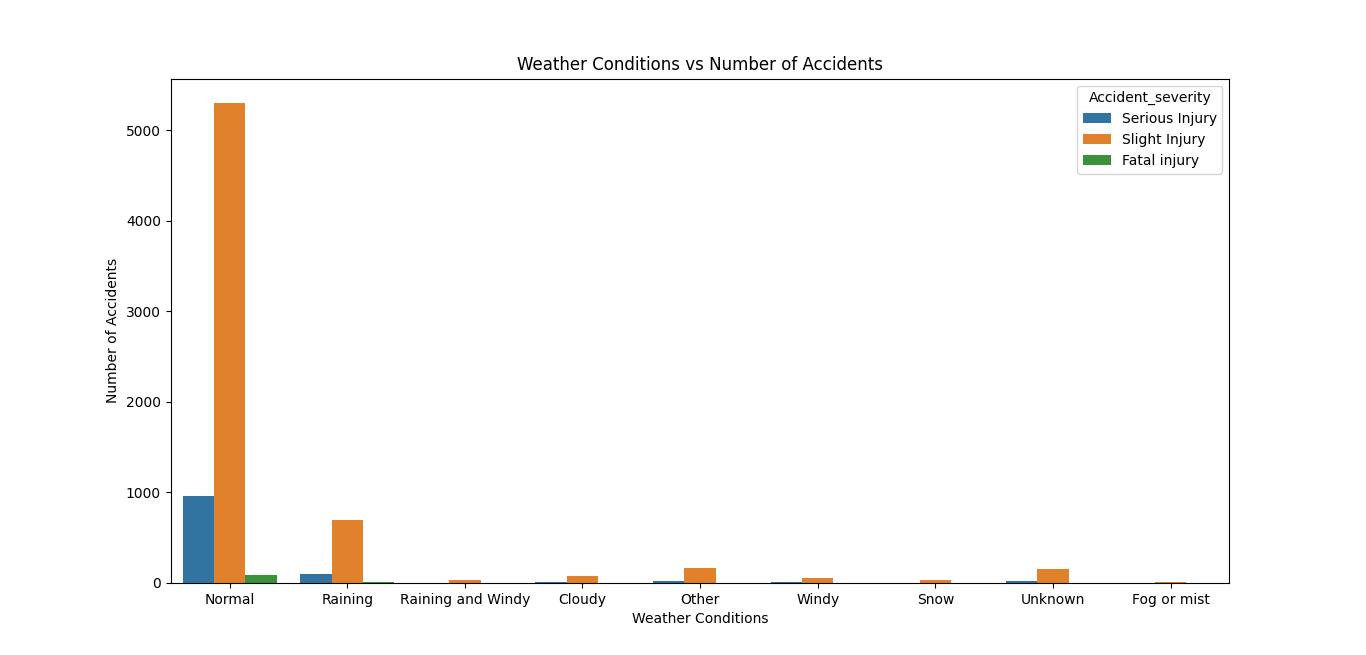
Below shown are some graphs from different fields of the datasets to show inter-relation.



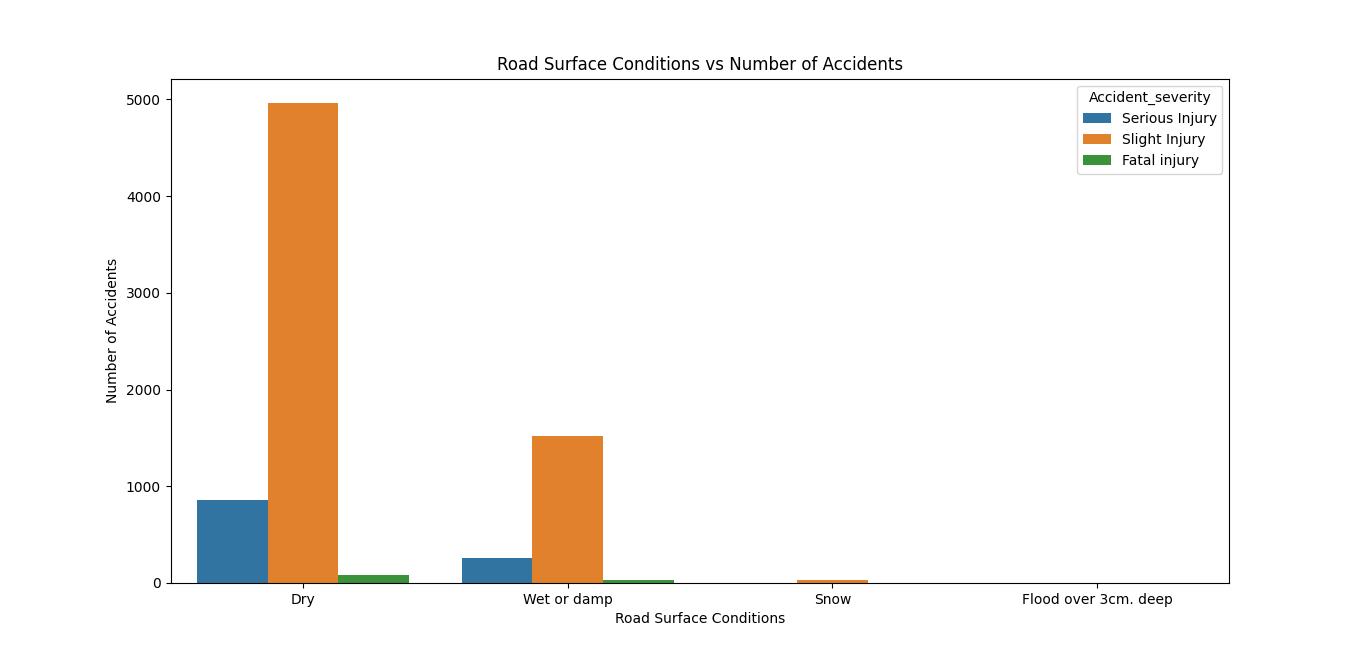
**Graph 1**



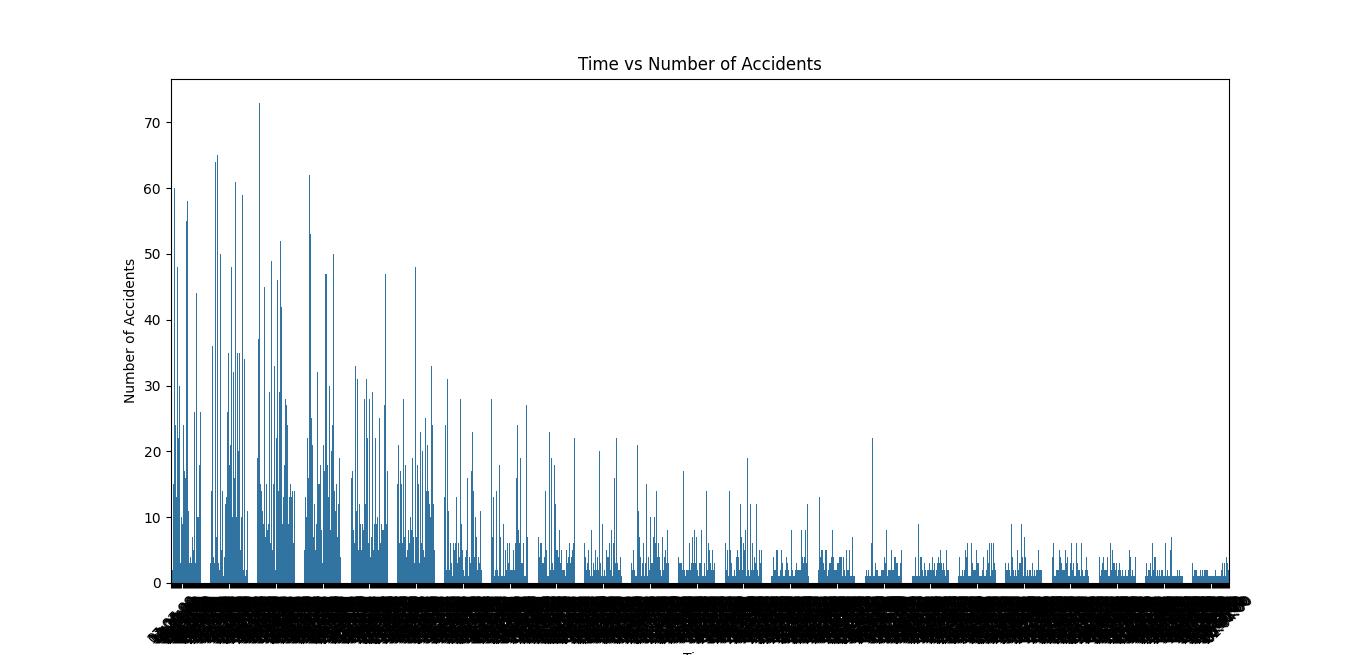
**Graph 2**



**Graph 3**



Graph 4



**Graph 5**

1. **CHALLENGES/LIMITATIONS**

* Incomplete or missing data in certain variables may affect the accuracy of analyses.
* The assumption of linear relationships in correlation analysis may oversimplify complex interactions.
* External factors not included in the dataset (e.g., driver behavior, law enforcement presence) may influence accident patterns.
* Data quality issues, such as inaccuracies in reporting, could impact the reliability of the findings.
* The dataset's temporal and spatial resolution may limit the granularity of the analysis.

1. **CONCLUSION**

Under this assignment we completed below tasks:

* Understood the data processing methods and implemented them on the acquired dataset.
* Developed a predictive model using statistical and machine learning techniques to identify high-risk areas and predict potential accident locations.
* Integrated the model with a data visualization platform that displays real-time insights and historical trends.

1. **PROGRAM FILES**

**   **