Analysis of Montgomery County Crash

Reporting Drivers Data.

George Mason University
AIT580-005| Prof. Dr. Alejandro Álvarez

Praneeth Ravirala
George Mason University
Fairfax, Virginia.
praviral@gmu.edu

Abstract— The Proposed Study depicts the analysis for the causes of road crashes in Montgomery County based on the data collected from the Automatic Crash Reporting System (ACRS). The proposed study can also be used as reference model to analyse the road accidents, Identify the factors involving the road accidents, thereby taking preventive actions to reduce the frequency of accidents and fatalities in future. It also helps the Government in making the road Infrastructure Sustainable and Safer to the People by altering the Traffic rules and Road Engineering Methods.

Keywords— Data Cleaning, Data Transformation, Statistics, Visualisation, Relationship, Accidents, Factors,

I. INTRODUCTION

Accidents can occur in several ways i.e., by road, fire, air and rail accidents, one of the most common is the road crashes which lead to 42,939 fatalities in USA as per National Highway Traffic Safety Administration (NHTSA) record in the year 2021.So, it is Important to find out the significant causes of road crashes In order to prioritise the actions that to be taken Immediately to prevent the fatalities in future. For this study Montgomery county's Crash Reporting Drivers Dataset is used which contains different features such as Weather, Light, Traffic Control, Injury Severity, Speed Limit, Driver Substance Abuse etc which gives the Information about the Environmental, Driver and Road Infrastructural Conditions at which the road crash has occurred. our aim is to analyse and find out the potential features which Tend to increase the no of accidents. Our Study Involves the steps which Include Loading the Dataset, Data Cleaning, Data Transformation, Statistics and Visualisation, and Interpretation. The study includes utilisation of various programming languages such as python, R and Sql. Tools Include Jupyter Notebook, MySql workbench, RStudio and MS-Excel. Data Cleaning Phase Involves removal of null values and Inconsistencies, Data Transformation Involves Converting data into a suitable format for further processing, followed by applying statistical techniques such as grouping on different features and applying visualization techniques on each feature. The analysis also includes all its three types i.e. Univariate, Bi-Variate and Multivariate Analysis.

II. RELATED WORK

There are several previous works which have Included analysis on road accident data and finding the relationship between features but there are no significant references of study which were able to find out the environmental, driver and vehicle conditions during events which lead to the accidents. Our study helps to find out the potential factors from the above which are responsible for increasing road accidents and fatalities.

Mohamed Aljaban's Study explores the reasons for auto accidents in the United States, emphasizing the identification of contributory elements. The study comes to the conclusion that the main factors causing traffic accidents are a city's population density and peak travel times. The author deftly refutes popular belief by stating that factors such as weather, speed, and lighting shouldn't have a major influence on the frequency of accidents. This result offers a unique viewpoint, which encourages the researcher to use the Montgomery collision reporting drivers dataset to investigate and evaluate the real effects of weather, speed, and lighting on auto accidents. It encourages a deeper investigation into the subtleties of automobile design by setting the stage for a critical assessment of these aspects within the context of the researcher's own investigation[2].

The study proposed by Chen Chen et al. uses a sophisticated driving simulation model to examine how drivers' perceived risk during car following is affected by unfavorable weather. To estimate and classify the likelihood of accidents, the research takes into account 11 different weather variables, such as gloomy and wet circumstances. The study suggests that snow has a positive effect on driver behavior, perhaps affecting the overall probability of accidents. Multiple linear regression is used as the modeling technique. The researcher's goal to comprehend the connection between weather and injury severity is supported by this finding. The simulation-driven method helps to clarify the intricacies of road safety during inclement weather by providing insightful information on how different weather types can affect drivers' perceptions of risk[3].

Chris Jurewicz et.al study aims to determine the complex

relationship between vehicle impact speed and the likelihood of fatal and serious injuries in order to inform safer road design. The study aims to identify particular impact speeds linked to different levels of damage severity, offering important information for determining the ideal speed limits in traffic safety tactics. Interestingly, the study finds that impact speeds of 30, 50, and 70 kmph, respectively, are associated with a 10% death risk for head-on, side, and pedestrian collisions. In addition to providing a basis for applying these insights to the analysis of the Montgomery crash reporting drivers dataset and aiding in the development of practical road safety measures, this investigation proves relevant for the researcher's interest in comprehending the relationship between speed and injury severity[4].

Tanisha Mahajan et.al conduct an experimental evaluation using crash reporting drivers dataset from Montgomery County to investigate car crashes associated with driver drug abuse. They convert the temporal dimension into cyclical data and concentrate on the "hour of the day". The study emphasizes how crucial it is to use encoding techniques like hot encoding, ordinal encoding, and sin/cos when utilizing cyclical data as predictors in machine learning algorithms. This method is used by the authors to evaluate the effectiveness of machine learning models, such as decision trees, logistic regression, and k-nearest neighbors (knn). The research project benefits greatly from the study's insights into the efficacy of various encoding methods and machine learning algorithms, especially in the areas of data analysis and model construction, which provide direction for maximizing performance with the Montgomery crash reporting drivers dataset[5].

III. DATASET

I have Used "Montgomery County Crash Reporting Driver's Dataset" which includes Report No, County, Place, Muncipality and Road, Time and Location, Weather and Light, Driver's License Number, Driver Substance Abuse, Driver At Fault and finally speed limit, equipment problems. The above variables can be categorized to following 'NOIR' characteristics[1].

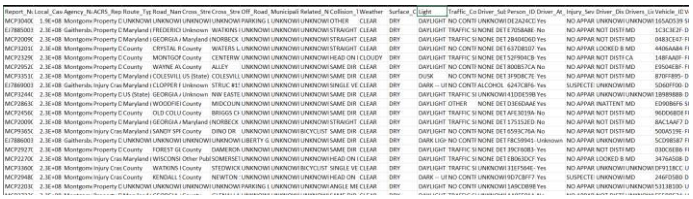
Nominal: Weather, Light, Driver_At_Fault.

Ordinal: Driver Substance Abuse, Injury Severity.

Ratio: Speed Limit.

Interval: Location.

Figure 1



Using the dataset we can able to analyze and it helps to answer the below research questions:

1.What are the features in your dataset that you want to focus to reach your goal?

2.What are the causes of vehicle accidents on Montgomery county?

3.Is there any relation between speed limit and injury severity?

IV. ANALYSIS AND INTERPRETATION

The Initial Analytical phase of the study is data cleaning in which the data is treated to remove null values and Inconsistencies for appropriate analysis and accurate results. Our study included the following steps:

Step1: Filling Null Values with 'UNKNOWN' in categorical features.

Figure 2

```
In [73]: #categorical variables
         fp_cat=fp.drop(columns=['Latitude','Longitude','Vehicle Year','Speed Limit','Location'],axis=1)

In [75]: columns_cat=['Route Type', 'Road Name','Cross-Street Type','Cross-Street Name','Off-Road Description','Municipality','Related
         for col in columns_cat:
             fp_cat[col].fillna('UNKNOWN',inplace=True)
             print(fp_cat[col].isna().sum())

         0
         0
         0
         0
         0
         0
         0
         0
         0
```

For Some categorical variables like 'Route Type' and 'Route Name' etc, It is advisable to fill the null values with either 'NA' or 'UNKNOWN' using fillna() function which replaces null values with 'UNKNOWN' value.

Step2: Imputing with Mode of the categorical features.

Figure3

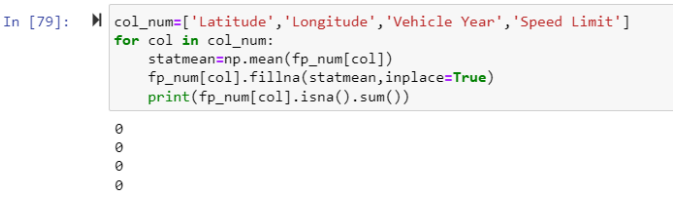
```
In [76]: columns_cat1=['Collision Type','Weather','Surface Condition','Light','Traffic Control','Driver Substance Abuse','Non-Motorist
         for col in columns_cat1:
             statmode=statistics.mode(fp_cat[col])
             fp_cat[col].fillna(statmode,inplace=True)
             print(fp_cat[col].isna().sum())

         0
         0
         0
         0
         0
```

For Some categorical variables like 'Collision_Type' and 'Weather', It is recommended to Impute the null values with its Mode using fillna() function and statistics.mode() function.

Step3: Imputing with mean of the numerical variables.

Figure4



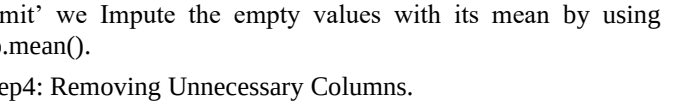
```
In [79]: col_num=['Latitude','Longitude','Vehicle Year','Speed Limit']
         for col in col_num:
             statmean=np.mean(fp_num[col])
             fp_num[col].fillna(statmean,inplace=True)
             print(fp_num[col].isna().sum())

         0
         0
         0
         0
```

For Numerical variables like 'latitude', 'Longitude', 'Speed Limit' we Impute the empty values with its mean by using np.mean().

Step4: Removing Unnecessary Columns.

Figure5



```
In [77]: fp_cat.drop('Non-Motorist Substance Abuse',axis=1,inplace=True)
         fp_cat.drop('Circumstance',axis=1,inplace=True)
```

Variables 'Circumstance' and 'Non-Motorist Substance Abuse' contain 90% values which are empty which is not suitable for further analysis and is recommended for dropping those two variables using drop() function.

The Second Phase Includes Data Transformation in which the 'Crash Date/Time' variable is converted into datetime format using datetime() function.

Figure6

```
In [83]:  # Convert the 'DateTime' column to datetime format
          fp['Crash Date/Time'] = pd.to_datetime(fp['Crash Date/Time'], errors='coerce')
          fp['Crash Date/Time'].fillna('2023-01-01 00:00:00',inplace=True)
          print(fp['Crash Date/Time'].head(10))
          print(fp['Crash Date/Time'].isna().sum())

0   2019-05-31 15:00:00
1   2023-07-21 17:59:00
2   2023-07-20 15:10:00
3   2023-07-23 12:10:00
4   2023-07-24 06:10:00
5   2023-07-11 07:40:00
6   2023-07-12 20:28:00
7   2023-07-05 23:25:00
8   2023-07-21 07:14:00
9   2023-07-19 19:00:00
Name: Crash Date/Time, dtype: datetime64[ns]
0
```

The result contains the suitable date time format 'YYYY-MM-DD HH:MM:SS' which is suitable for statistics and visualisation.

The next phase is Applying Statistical Techniques such as Grouping, Min, Max and Mode using Sql language with MySql Workbench. Our study does the following steps:

Step1: Statistics based on Light, Weather, Injury Severity, Traffic Control.

Figure7

```
select Light,Weather,Injury_Severity,Traffic_Control,
avg(Speed_Limit) as average_speed_limit,
min(Speed_Limit) as min_speed_limit,
max(Speed_Limit) as max_speed_limit
from cleaned_data
group by Light,Weather,Injury_Severity,Traffic_Control;
```

Light	Weather	Injury_Severity	Traffic_Control	average_speed_limit	min_speed_limit	max_speed_limit
CLEAR	UNKNOWN	0AA7B06B-3574-4C91-8A89-5F853C359911	DRY	0.0000	0	0
CLEAR	UNKNOWN	5630CBCF-E8D4-42D6-A9B1-F90DF4B697EA	DRY	0.0000	0	0
DARK -- UNKNOWN LIGHTING	BLOWING SNOW	NO APPARENT INJURY	TRAFFIC SIGNAL	50.0000	50	50
DARK -- UNKNOWN LIGHTING	BLOWING SNOW	SUSPECTED MINOR INJURY	TRAFFIC SIGNAL	50.0000	50	50
DARK -- UNKNOWN LIGHTING	CLEAR	FATAL INJURY	NO CONTROLS	40.0000	40	40
DARK -- UNKNOWN LIGHTING	CLEAR	FATAL INJURY	TRAFFIC SIGNAL	40.0000	40	40
DARK -- UNKNOWN LIGHTING	CLEAR	NO APPARENT INJURY	FLASHING TRAFFIC SIGNAL	35.4545	30	45
DARK -- UNKNOWN LIGHTING	CLEAR	NO APPARENT INJURY	NO CONTROLS	28.3687	0	65
DARK -- UNKNOWN LIGHTING	CLEAR	NO APPARENT INJURY	OTHER	32.9167	5	55

We Grouped SpeedLimit based on the above four variables and find out summary statistics such as avg ,min, max.

Step2: No of Accidents based on weather and Injury Severity.

Figure8

```
Select Weather,Injury_Severity,count(*) as No_Of_Accidents
from  cleaned_data Group by Weather, Injury_Severity
order by No_Of_Accidents Desc;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Weather	Injury_Severity	No_Of_Accidents
CLEAR	NO APPARENT INJURY	99550
RAINING	NO APPARENT INJURY	15456
CLOUDY	NO APPARENT INJURY	13065
CLEAR	POSSIBLE INJURY	12087
CLEAR	SUSPECTED MINOR INJURY	8284
RAINING	POSSIBLE INJURY	2042
CLOUDY	POSSIBLE INJURY	1853
RAINING	SUSPECTED MINOR INJURY	1447

Arrangement of No of accidents in decreasing order with respect to weather and Injury Severity depicts on clear weather, No Apparent Injury the no of accidents are most ie 99550, followed by Raining Weather has more accidents.

Step3: No of Accidents based on Light and Injury Severity.

Figure9

```
Select Light,Injury_Severity,count(*) as No_Of_Accidents
from  cleaned_data Group by Light, Injury_Severity
order by No_Of_Accidents Desc;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Light	Injury_Severity	No_Of_Accidents
DAYLIGHT	NO APPARENT INJURY	90057
DARK LIGHTS ON	NO APPARENT INJURY	29866
DAYLIGHT	POSSIBLE INJURY	11553
DAYLIGHT	SUSPECTED MINOR INJURY	7515
DARK NO LIGHTS	NO APPARENT INJURY	3672
DARK LIGHTS ON	POSSIBLE INJURY	3449
DUSK	NO APPARENT INJURY	3002

Grouped No of Accidents based on Light and Injury Severity and arranging in descending order depicts no of accidents are more in Daylight and No apparent Injury Conditions ie 90057 followed by darklight has more accidents.

Step4: No of Accidents based on Speed Limit and Injury Severity.

Figure10

```
Select Speed_Limit,Injury_Severity,count(*) as No_Of_Accidents
from  cleaned_data Group by Speed_Limit, Injury_Severity
order by No_Of_Accidents Desc;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Speed_Limit	Injury_Severity	No_Of_Accidents
35	NO APPARENT INJURY	37932
40	NO APPARENT INJURY	24388
25	NO APPARENT INJURY	19205
30	NO APPARENT INJURY	17744
45	NO APPARENT INJURY	8945
15	NO APPARENT INJURY	5300
35	POSSIBLE INJURY	5155

We can depict from the above graph after grouping no of accidents based on speed limit and injury Severity that no of accidents are more in case of 35 speed and no apparent injury ie 37932, 40 speed has second most no of accidents.

Step5: No of Accidents based on Driver_Substance_Abuse

Figure11

```
Select Driver_Substance_Abuse,count(*) as No_Of_Accidents
from cleaned_data group by Driver_Substance_Abuse
order by No_of_Accidents desc;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Driver_Substance_Abuse	No_Of_Accidents
NONE DETECTED	143203
UNKNOWN	11115
ALCOHOL PRESENT	3830
ALCOHOL CONTRIBUTED	1335
ILLEGAL DRUG PRESENT	241
MEDICATION PRESENT	113

Grouping No of Accidents based on Driver_Substance_Abuse illustrates at no driver substance abuse no of accidents are more ie 143203, no of accidents with alcohol present is 3830.

Step6: No of Accidents based on Driver_At_Fault

Figure12

```
Select Driver_At_Fault,count(*) as No_Of_Accidents
from cleaned_data group by Driver_At_Fault
order by No_of_Accidents desc;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Driver_At_Fault	No_Of_Accidents
Yes	86068
No	69793
Unknown	4311
NO CONTROLS	4
ALCOHOL CONTRIBUTED	1
7B20DCE4-ADC1-4222-B6BB-9338BA58CD0E	1
NONE DETECTED	1

The above grouping result of no of accidents based on Driver_At_Fault shows that no of accidents are more when driver at fault.

Step7: Summary Statistics of Speed Limit

Figure13

```
Select avg(Speed_Limit) as average_speed,
min(Speed_Limit) as minimum_speed,
max(Speed_Limit) as maximum_speed
from cleaned_data;
```

Result Grid | Filter Rows: | Export:

average_speed	minimum_speed	maximum_speed
32.5671	0	75

The Summary Statistics of Speed Limit Include Average, Minimum and Maximum.

Step8: No of Accidents based on Equipment Problems.

Figure14

```
Select Equipment_Problems,count(*) as No_Of_Accidents
from cleaned_data group by Equipment_Problems
order by No_of_Accidents desc;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Conte

Equipment_Problems	No_Of_Accidents
NO MISUSE	146756
UNKNOWN	13089
AIR BAG FAILED	159
OTHER	105
BELT(S) MISUSED	32
BELTS/ANCHORS BROKE	11

The result shows that even after no equipment problem the no of accidents are more ie 146756, When Air bag failed the accidents are 159.

The final phase is Visualising the statistics to analyse the data and interpret the results. It helps us to answer the research questions.

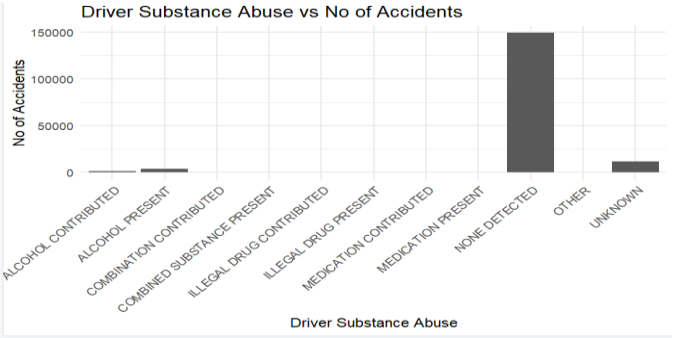
In Visualisation we first do the Univariate Analysis, below are the respective plots:

Plot1: Using ggplot no of accidents are plotted based on driver substance abuse using bar plot

Figure15

```
ggplot(df, aes(x = df$Driver_Substance_Abuse)) +
  geom_bar(stat = "count") +
  labs(title = "Driver Substance Abuse vs No of Accidents ",
       x = "Driver Substance Abuse",
       y = "No of Accidents") +
  theme_minimal()+theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Figure16



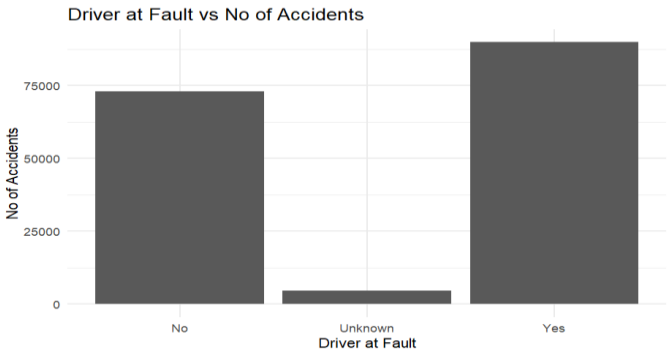
The above plot shows that no of accidents are less if alcohol is present

Plot 2: No of accidents are plotted based on driver at fault.

Figure17

```
ggplot(df, aes(x = df$Driver_At_Fault)) +
  geom_bar(stat = "count") +
  labs(title = "Driver at Fault vs No of Accidents ",
       x = "Driver at Fault",
       y = "No of Accidents") +
  theme_minimal()
```

Figure18



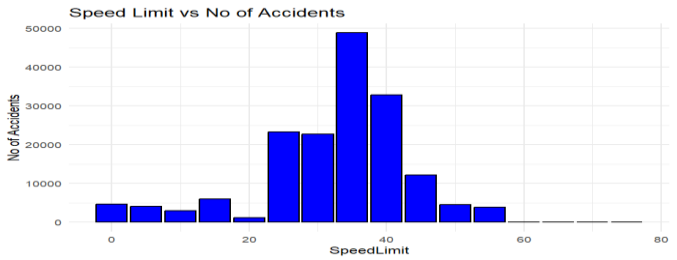
The above plot depicts that Driver fault is present for most of the accidents.

Plot3: Plotting Speedlimit vs no of accidents

Figure19

```
ggplot(df, aes(x = df$Speed_Limit)) +
  geom_histogram(stat = "count",fill='blue',color='black')+
  labs(title = "Speed Limit vs No of Accidents ",
       x = "SpeedLimit",
       y = "No of Accidents") +
  theme_minimal()
```

Figure20



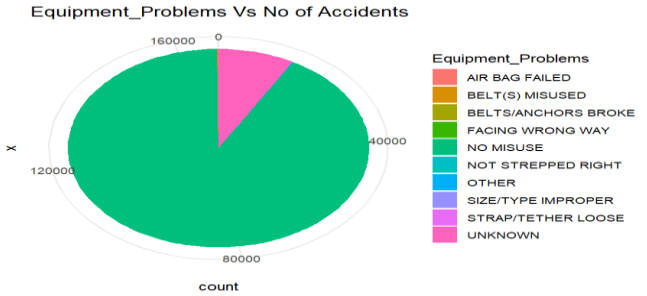
The above bar chart depicts for speed limit of 35 has highest no of accidents.

Plot4: Plotting Equipment Problems Vs No of Accidents

Figure21

```
ggplot(result, aes(x = "", y = count, fill = Equipment_Problems)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  theme_minimal() +
  labs(title = "Equipment_Problems Vs No of Accidents")
```

Figure22



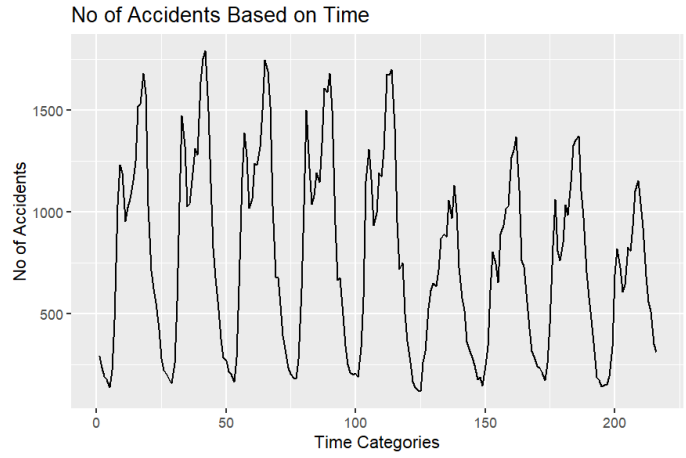
Below pie chart depicts no of accidents are more even if there is no problem with the equipment, followed by strap/tether loose.

Plot5: Time Series Plot

Figure23

```
ggplot(result1, aes(x = data1$category, y = count)) +
  geom_line() +
  labs(title = "No of Accidents Based on Time",
       x = "Time Categories", y = "No of Accidents")
```

Figure24



Each time category has been coded to a numeric value so that visualisation looks good.

The next step is to bivariate analysis to find out the relationship between two variables.

Relationship Between Injury Severity and AverageSpeed

Figure25

	Injury_Severity	AverageSpeed
	<chr>	<dbl>
1	FATAL INJURY	38.8
2	NO APPARENT INJURY	31.9
3	POSSIBLE INJURY	35.4
4	SUSPECTED MINOR INJURY	35.4
5	SUSPECTED SERIOUS INJURY	36.4

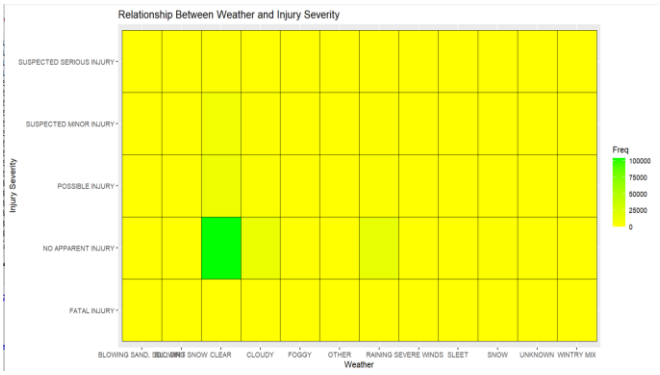
If we group Injury Severity Based on Average Speed we can illustrate Fatal Injury has highest average speed.

Plot 6: Relationship Between Weather and Injury Severity

Figure26

```
#Bivariate
ggplot(df_heatmap, aes(x = Var1, y = Var2, fill = Freq)) +
  geom_tile(color = "black") +
  scale_fill_gradient(low = "yellow", high = "green") +
  labs(title = "Relationship Between Weather and Injury Severity",
       x = "Weather",
       y = "Injury Severity")
```

Figure27



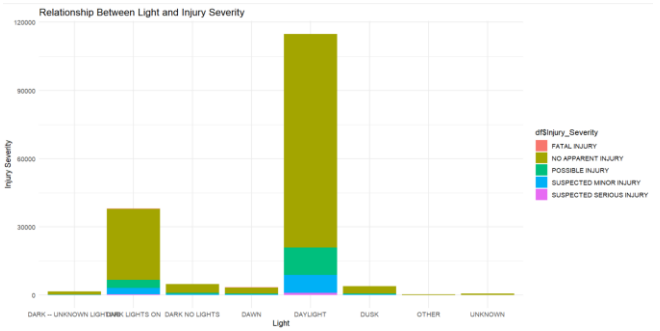
Using heatmap we find out the no of accidents are more even there is no apparent injury and clear weather, raining weather has second most no of accidents.

Plot7: Relationship between Light and Injury Severity

Figure28

```
ggplot(df, aes(x = df$Light, fill = df$Injury_Severity)) +
  geom_bar(position = "stack") +
  labs(title = "Relationship Between Light and Injury Severity",
       x = "Light",
       y = "Injury Severity") +
  theme_minimal()
```

Figure29



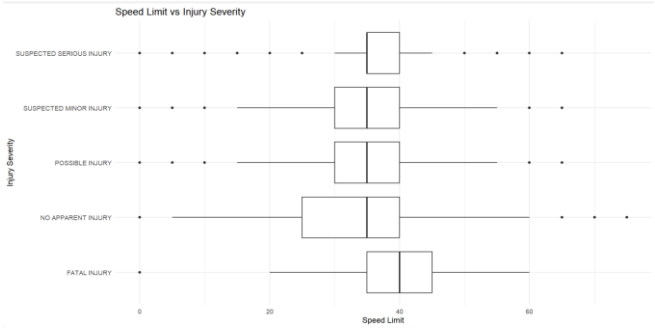
The above Stacked Bar graph depicts No of accidents are more in daylight.

Plot8: Relationship between Speed Limit and Injury Severity

Figure30

```
ggplot(df, aes(x = df$Speed_Limit, y = df$Injury_Severity)) +
  geom_boxplot() +
  labs(title = "Speed Limit vs Injury Severity",_
       x = "Speed Limit",
       y = "Injury Severity") +
  theme_minimal()
```

Figure31



The below boxplot depicts Fatal Injury has high speed limit ie 60.

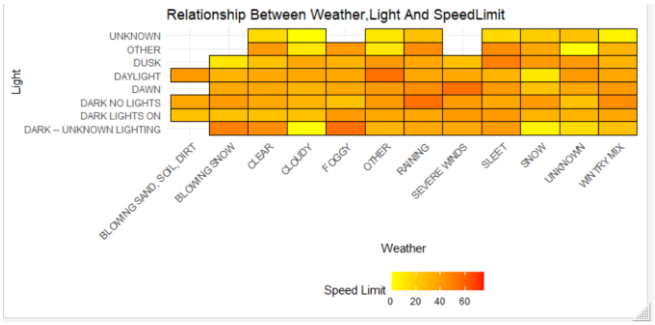
The final step of visualisation is to perform multivariate analysis to find out the relationship between three or more variables.

Plot9: Multivariate Analysis of weather,light and speed limit

Figure32

```
ggplot(df, aes(x = Weather, y = Light, fill = Speed_Limit)) +
  geom_tile(color = "black") +
  labs(
    title = "Relationship Between Weather,Light And SpeedLimit",
    x = "Weather",
    y = "Light",
    fill = "Speed Limit"
  ) +
  scale_fill_gradient(low = "yellow", high = "red") +
  theme_minimal()+theme(axis.text.x = element_text(angle = 45, hjust = 1),legend.position = "bottom")
```

Figure33



The above hashmap depicts in dark light & foggy weather the speed limit is more.

V. SUMMARY

Below are the answers to the research questions:

1.What are the features in your dataset that you want to focus to reach your goal?

The analysis focuses on numerous features, such as light conditions, speed limit, weather, and time, injury severity in order to achieve the goal of reducing collisions and severity in Montgomery County.

2. What are the causes of vehicle accidents on Montgomery county?

The main causes of vehicle accidents are Driver's Fault, Light and Weather.

3. Is there any relation between speed limit and injury severity?

Fatal Injury severity has high average speed limit 38 and Suspected Serious Injury has average speed limit 36.4

VI. REFERENCES

[1].Crash Reporting - Drivers Data | Open Data Portal. (2023, October 20).

[2]. Aljaban, Mohamed, "Analysis of Car Accidents Causes in the USA" (2021).

[3]. Chen, C., Zhao, X., Liu, H., Ren, G., & Liu, X. (2019). Influence of adverse weather on drivers' perceived risk during car following based on driving simulations. Journal of Modern Transportation, 27(4), 282–292.

[4].Jurewicz, C., Sobhani, A., Woolley, J., Dutschke, J., & Corben, B. (2016). Exploration of Vehicle Impact Speed – Injury Severity Relationships for application in Safer Road Design. Transportation Research Procedia, 14, 4247–4256.

[5].Mahajan, T. (2021). An experimental assessment of treatments for cyclical data. ScholarWorks