COMPREHENSIVE ANALYSIS OF TRAFFIC ACCIDENTS

- SrinivasaBharadwajChakilam(11614600)
- SrinivasaBharadwajChakilam@my.unt.edu Role: **Project Manager**
- Rakshitha Dabbara(11708197)
- RakshithaDabbara@my.unt.edu Role: **Data visualizer**
- Likitha Kolluru Balaji(11708196)
- LikithaKolluruBalaji@my.unt.edu Role: **Data Analyst**
- Chandrika Gogineni(11609940)
- ChandrikaGogineni@my.unt.edu Role: **Statistician**

Motivation

- The main motivation of this project is to improve the road safety and to reduce the impact of road accidents on lives in community and worldwide. This could be done by analyzing the various factors that influence accident severity such as weather conditions, time of day etc.
- This project also aims to provide data driven insights which inform effective policies, innovations in technologies, and public awareness campaigns.
- At last, the main objective is to harness the power of data analytics to create proactive measures which helps to prevent the accidents and to promote safer road actions which helps to contribute a safer and more resilient transportation for everyone.

Abstract

This project aims on comprehensive analysis of traffic accidents. Such as focusing on the relationship between speed, time of the day, and rate of accidents. By using statistical analysis and visualization techniques on the dataset. It also helps in quantifying the impact of speed on accidents , identifying optimal speed regulations, and investigating temporal patterns to highlight the high-risk hours such as rush hours.

By performing the regression analysis, time series modeling, and using clustering algorithms. The main objective of this study is to provide the actionable insights for urban planners, policy makers and public users to facilitate the evidence-based interventions in creating safer roads and efficent transportation system.

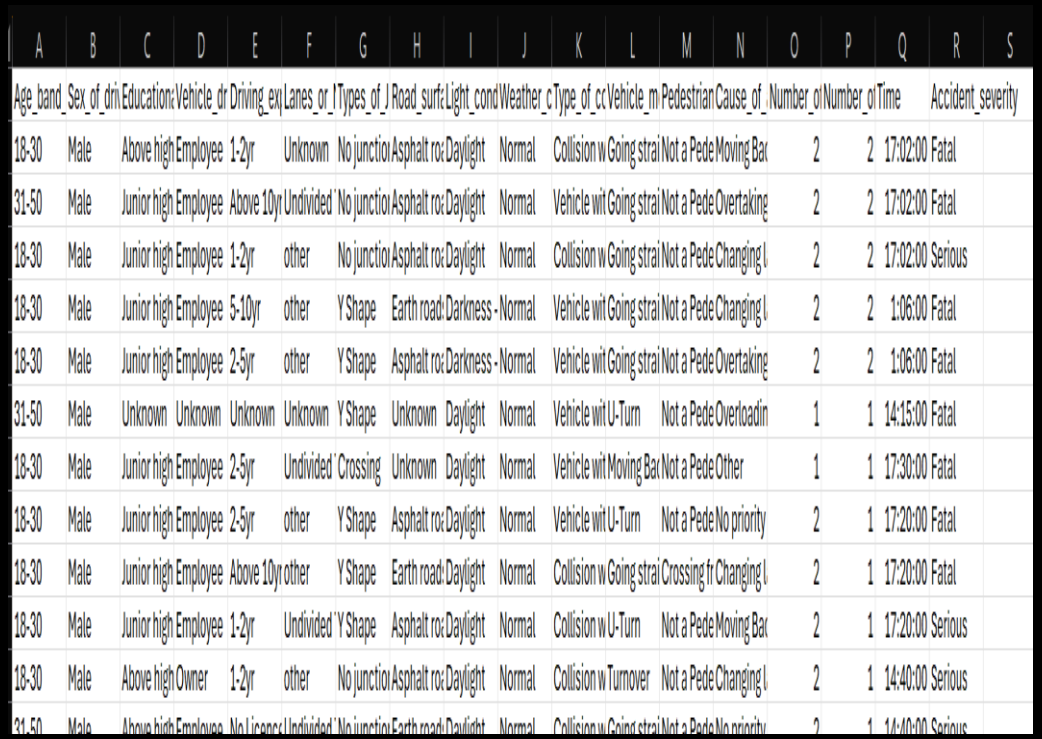
Data Set

- **Data Set**: road-traffic-accidents

Our project makes use of the 'cleaned_dataset.csv' file, which was derived from complete traffic accident records. This dataset contains essential information including driver demographics, accident severity, and environmental factors.

It has 19 columns and 12316 rows of data with column name as:

1. Age_band_of_driver
2. Sex_of_driver
3. Educational_level
4. Vehicle_driver_relation
5. Driving_experience
6. Lanes_or_Medians
7. Types_of_Junction
8. Road_surface_type
9. Light_conditions
10. Weather_conditions
11. Type_of_collision
12. Vehicle_movement
13. Pedestrian_movement
14. Cause_of_accident
15. Number_of_vehicles_involved
16. Number_of_casualties
17. Time
18. Accident_severity
19. Hour



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
Age_band_	Sex_of_dri	Education	Vehicle_dr	Driving_ex	Lanes_or_	Types_of_J	Road_surfa	Light_cond	Weather_c	Type_of_cc	Vehicle_m	Pedestrian	Cause_of_	Number_of	Number_of	Time	Accident_severity	
18-30	Male	Above high	Employee	1-2yr	Unknown	No junctio	Asphalt roa	Daylight	Normal	Collision w	Going strai	Not a Pede	Moving Bac	2	2	17:02:00	Fatal	
31-50	Male	Junior high	Employee	Above 10yr	Undivided	No junctio	Asphalt roa	Daylight	Normal	Vehicle wit	Going strai	Not a Pede	Overtaking	2	2	17:02:00	Fatal	
18-30	Male	Junior high	Employee	1-2yr	other	No junctio	Asphalt roa	Daylight	Normal	Collision w	Going strai	Not a Pede	Changing l	2	2	17:02:00	Serious	
18-30	Male	Junior high	Employee	5-10yr	other	Y Shape	Earth road	Darkness -	Normal	Vehicle wit	Going strai	Not a Pede	Changing l	2	2	1:06:00	Fatal	
18-30	Male	Junior high	Employee	2-5yr	other	Y Shape	Asphalt roa	Darkness -	Normal	Vehicle wit	Going strai	Not a Pede	Overtaking	2	2	1:06:00	Fatal	
31-50	Male	Unknown	Unknown	Unknown	Unknown	Y Shape	Unknown	Daylight	Normal	Vehicle wit	U-Turn	Not a Pede	Overloadin	1	1	14:15:00	Fatal	
18-30	Male	Junior high	Employee	2-5yr	Undivided	Crossing	Unknown	Daylight	Normal	Vehicle wit	Moving Bac	Not a Pede	Other	1	1	17:30:00	Fatal	
18-30	Male	Junior high	Employee	2-5yr	other	Y Shape	Asphalt roa	Daylight	Normal	Vehicle wit	U-Turn	Not a Pede	No priority	2	1	17:20:00	Fatal	
18-30	Male	Junior high	Employee	Above 10yr	other	Y Shape	Earth road	Daylight	Normal	Collision w	Going strai	Crossing fr	Changing l	2	1	17:20:00	Fatal	
18-30	Male	Junior high	Employee	1-2yr	Undivided	Y Shape	Asphalt roa	Daylight	Normal	Collision w	U-Turn	Not a Pede	Moving Bac	2	1	17:20:00	Serious	
18-30	Male	Above high	Owner	1-2yr	other	No junctio	Asphalt roa	Daylight	Normal	Collision w	Turnover	Not a Pede	Changing l	2	1	14:40:00	Serious	
31-50	Male	Above high	Employee	No Licens	Undivided	No junctio	Earth road	Daylight	Normal	Collision w	Going strai	Not a Pede	No priority	2	1	14:40:00	Serious	

Technologies used

- Python: Primarily used for scripting analyses.
- Pandas: Used for data manipulation and analysis, especially for data tables and time series.
- NumPy: It is mainly used for numerical operations with pandas library.
- Matplotlib: It is a plotting library for creating static, interactive visulaizations.
- Seaborn: A high-level visualization library based on Matplotlib.
- SciPy: Used for scientific and technical computing, including optimization, linear algebra, integration, and statistics.
- Statsmodels: Used for statistical modeling and hypothesis testing.
- Scikit-learn: Used for machine learning, model building, evaluation, and validation.
- Jupyter Notebook: Interactive computing environment for Python code execution and display.

Data Cleaning Steps

- We have meticulously analyzed the information, addressing problems such as missing values, normalizing categorical types, and removing unnecessary entries such as dropping duplicates and null values.

```
# Basic Data Cleaning
# Remove rows where any of the following columns are missing
df.drop_duplicates(inplace=True)
df = df.dropna(subset=['Number_of_vehicles_involved', 'Accident_severity', 'Weather_conditions'])
```

Statistical Tests

- Statistical Tests: In this project, we have implemented the statistical tests such as T-test, chi-square test, ANOVA and regression analysis.

Statistical Tests

- **T-Test:** We have performed t-test to compare the number of automobiles involved in two accident severity categories ('Fatal' and 'Serious'), yielding a statistically significant difference.
- T-statistic = 9.738
- P-value = 2.50×10−222.50×10−22.
- The findings indicate that the observed difference in the number of cars involved in 'Fatal' and 'Serious' accidents is unlikely to be attributable to random chance but also have a statistical meaning.

```
# Statistical Analysis
# T-Test: Compare 'Number_of_vehicles_involved' between two 'Accident_severity' groups
# For simplicity, let's compare two severity levels: 'Fatal' and 'Serious' (adjust based on your actual data categories)
group1 = df[df['Accident_severity'] == 'Fatal']['Number_of_vehicles_involved']
group2 = df[df['Accident_severity'] == 'Serious']['Number_of_vehicles_involved']

# Ensure there are enough data points for each group
if len(group1) > 1 and len(group2) > 1:
    t_stat, p_val = stats.ttest_ind(group1, group2, nan_policy='omit')
    print(f"T-Test: T-statistic = {t_stat}, P-value = {p_val}")
else:
    print("Not enough data points for a T-Test comparison.")
```

```
T-Test: T-statistic = 9.738058576065248, P-value = 2.503689789558747e-22
```

Statistical Tests

- **Chi-squared test:** The Chi-squared test determines if there is a significant relationship between category variables.
- In this we performed Chi-square test to determine the correlation between meteorological(weather) conditions and accident severity.
- Chi2 Stat value – 41.66
- P-value= 0.0004
- The low p-value rejects the null hypothesis, indicating a significant correlation between meteorological conditions and accident severity.

```
# Chi-Square Test: Relationship between 'Weather_conditions' and 'Accident_severity'
contingency_table = pd.crosstab(df['Weather_conditions'], df['Accident_severity'])
chi2_stat, p_val, dof, ex = stats.chi2_contingency(contingency_table)
print(f"Chi-Square Test: Chi2 Stat = {chi2_stat}, P-value = {p_val}")
```

```
Chi-Square Test: Chi2 Stat = 41.66198252891352, P-value = 0.0004430700682616468
```

Statistical Tests

- **ANOVA:** The ANOVA (Analysis of Variance) is a statistical approach for comparing the means of many groups and determining if at least one group's mean varies substantially from the others.
- In this we performed ANOVA test was used to investigate the variation in accident severity at different times of day.
- F-statistic: 2.892
- P-value: 2.21 x 10^(-162).
- This little p-value, which is significantly below the usually accepted alpha threshold of 0.05, demonstrates that there is a statistically significant variation in accident severity across different times of day.

```python
from scipy.stats import f_oneway

if 'Time' in df.columns and 'Accident_severity' in df.columns:
    # Converting 'Accident_severity' into numerical values for analysis
    severity_mapping = {'Fatal': 2, 'Serious': 1}
    df['Accident_severity_numerical'] = df['Accident_severity'].map(severity_mapping)

    # Droping rows where 'Time' or 'Accident_severity_numerical' is NaN after the mapping
    df = df.dropna(subset=['Time', 'Accident_severity_numerical'])

    # Grouping data by 'Time' and collect 'Accident_severity_numerical' values
    time_groups = df.groupby('Time')['Accident_severity_numerical'].apply(list)

    # Performing ANOVA only if there are at least two groups to compare
    if len(time_groups) >= 2:
        anova_result = f_oneway(*time_groups)

        print(f"ANOVA Result: F-statistic = {anova_result.statistic}, P-value = {anova_result.pvalue}")

        # Interpretation
        if anova_result.pvalue < 0.05:
            print("There is a statistically significant difference in accident severity across different times of the day.")
        else:
            print("There is no statistically significant difference in accident severity across different times of the day.")
    else:
        print("Not enough groups for ANOVA.")
else:
    print("The dataset does not contain the required 'Time' and/or 'Accident_severity' columns.")
```

```
ANOVA Result: F-statistic = 2.8916371110401466, P-value = 2.2111019820814385e-162
There is a statistically significant difference in accident severity across different times of the day.
```

Statistical Tests

- **Pearson Analysis:** The Pearson Correlation coefficient measures linear correlation between two sets of data.
- In this we have performed the test between time of day and the number of cars involved in accidents.
- Pearson Analysis statistic: 0.025
- P-value: 0.005
- As the p-value is less than 0.05, we can conclude that there is a statistically significant variation in accident severity across the different times of the day.

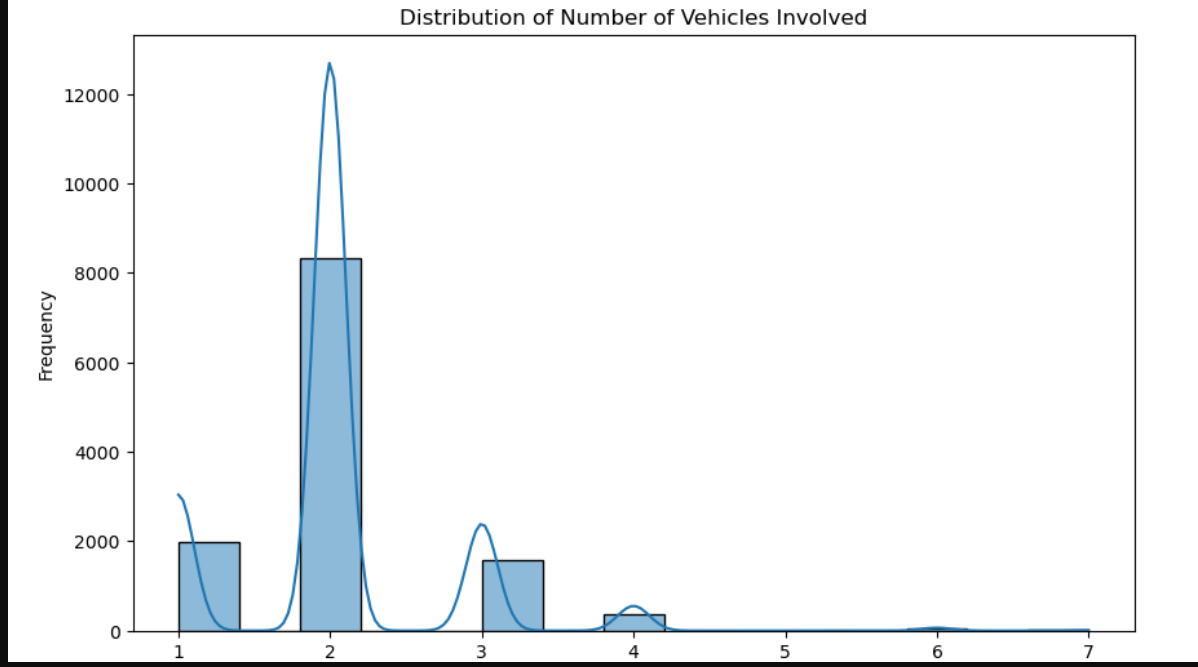
```python
import pandas as pd
from scipy.stats import pearsonr
# Function to convert time to numerical value (minutes past midnight)
def convert_time_to_numerical(time_str):
    h, m, s = map(int, time_str.split(':'))
    return h * 60 + m + s / 60.0

# Apply the conversion to the 'Time' column
df['Time_numerical'] = df['Time'].apply(convert_time_to_numerical)

# Compute the Pearson correlation coefficient between 'Time_numerical' and 'Number_of_vehicles_involved'
corr_coefficient, p_value = pearsonr(df['Time_numerical'], df['Number_of_vehicles_involved'])

print(f"Pearson Correlation Coefficient: {corr_coefficient}, P-value: {p_value}")

# Interpretation
if p_value < 0.05:
    print("There is a statistically significant correlation between the time of day and the number of vehicles involved in accide
else:
    print("There is no statistically significant correlation between the time of day and the number of vehicles involved in accic
```

```
Pearson Correlation Coefficient: 0.02528669956453332, P-value: 0.005303282333252398
There is a statistically significant correlation between the time of day and the number of vehicles involved in accidents.
```

Number of Vehicles vs Frequency of accidents

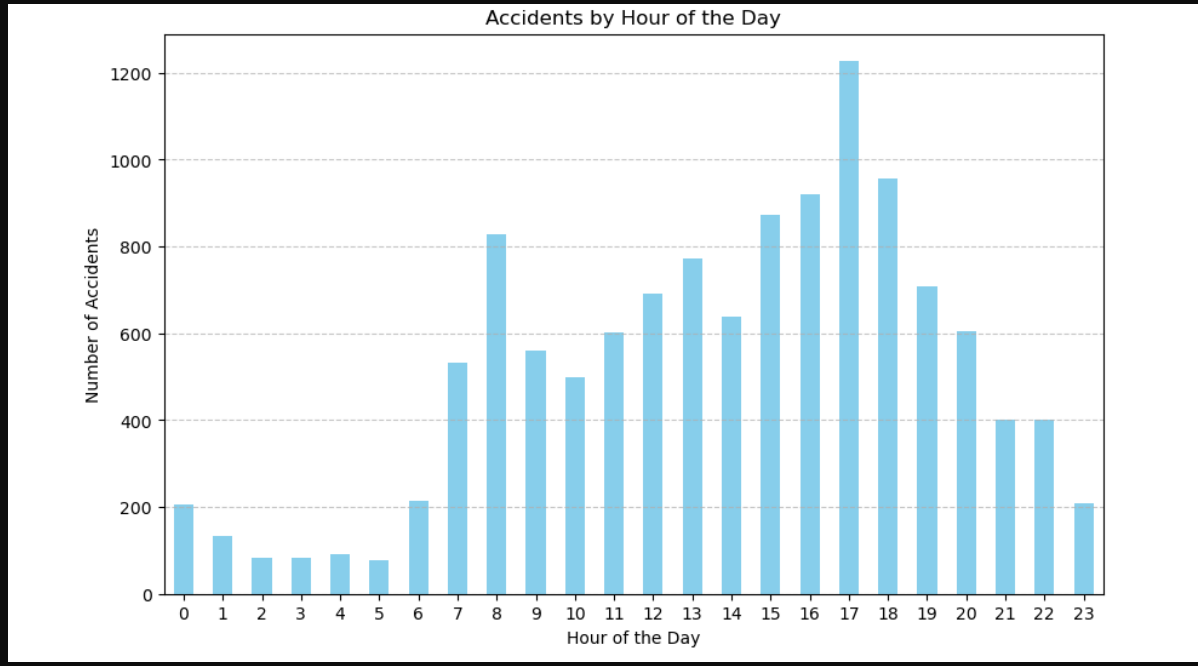
- From the graph we can conclude that the accident occurs more often when 2 vehicles are involved in accident.

```
# Distribution of 'Number_of_vehicles_involved'
plt.figure(figsize=(10, 6))
sns.histplot(df['Number_of_vehicles_involved'], kde=True)
plt.title('Distribution of Number of Vehicles Involved')
plt.xlabel('Number of Vehicles Involved')
plt.ylabel('Frequency')
plt.show()
```



Hours vs Frequency of accidents

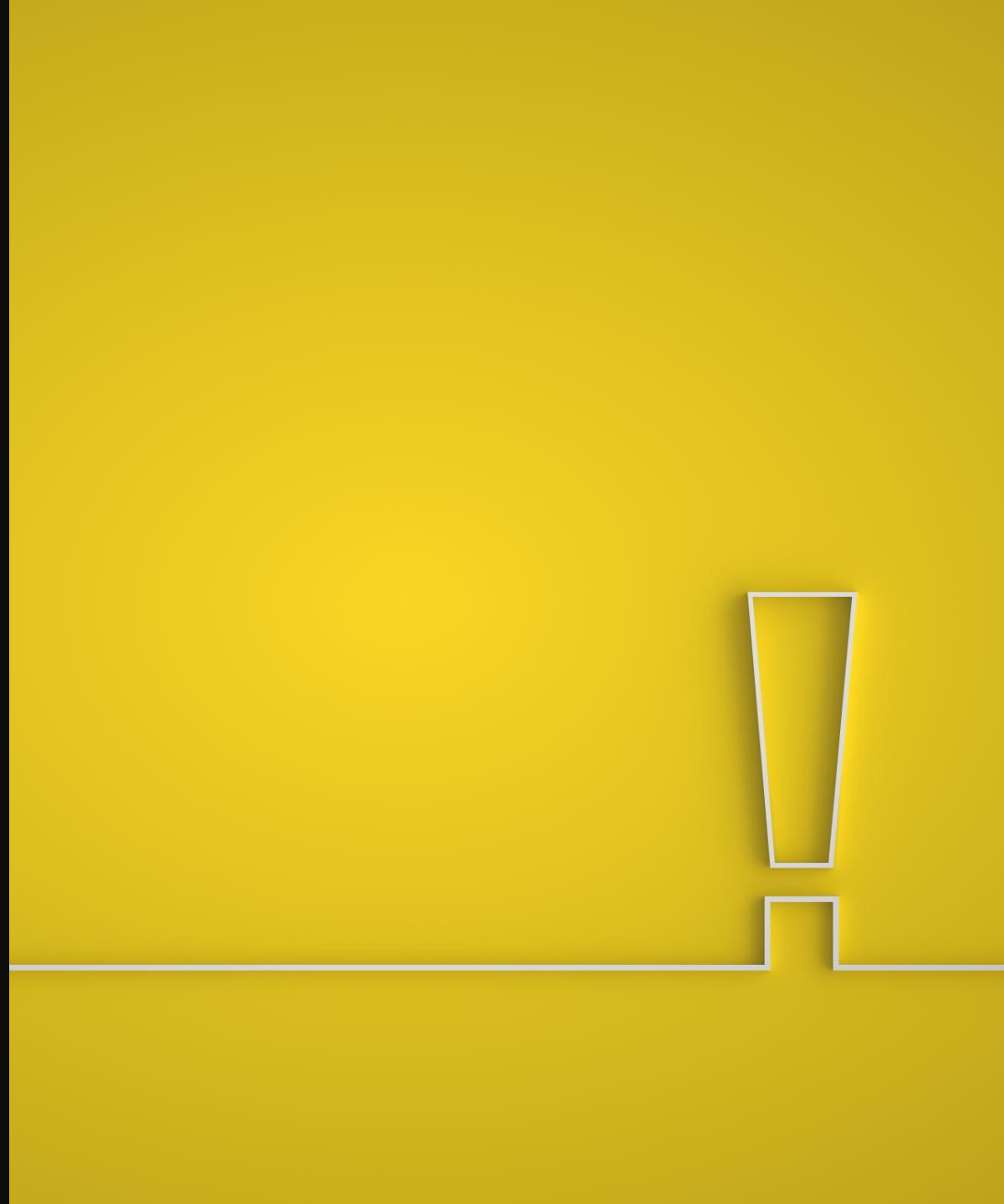
- It can be inferred that peak hours in the day are 8am,3-7pm which usually are busy hours where people tend to go to work etc,



Key Points

- Used features to run stats on and to predict the association and p-value.
- Features like ‘Accident severity’ and ‘number of vehicles involved’ are used to group the entire dataset into two basic groups called ‘Fatal’ and ‘Serious’ which are later used to find mean number of accidents per day.

- Performed ANOVA on features ‘Weather condition’ and ‘Number of casualities’ to find significant differences in number of casaulities across different weather conditions.
- Performed ANOVA on time and accidents severity to find if time has any effect on accidents.



Resources and Related Projects

Projects:

1)Road traffic Accidents-identification of major causes of the accident by analyzing it using different machine learning classification algorithms algorithms.

Road Traffic Accidents (kaggle.com)

2) Road traffic accidents:

This article discusses road accident data sources, analytical methodology, algorithms, operational obstacles, risk variables, road safety measures efficiency, and future methodological approaches. It also covers operational issues, risk considerations, and the evaluation of future solutions..

Road traffic accidents: An overview of data sources, analysis techniques and contributing factors - ScienceDirect

Resources and Related Projects

Projects:

3)Road traffic accidents : In recent years, there has been increased concern over traffic accidents. The National Highway Traffic Safety Administration is initiating the "Speeding Wrecks Lives" effort to discourage speeding, following a 14-year high in speed-related fatalities in 2021.

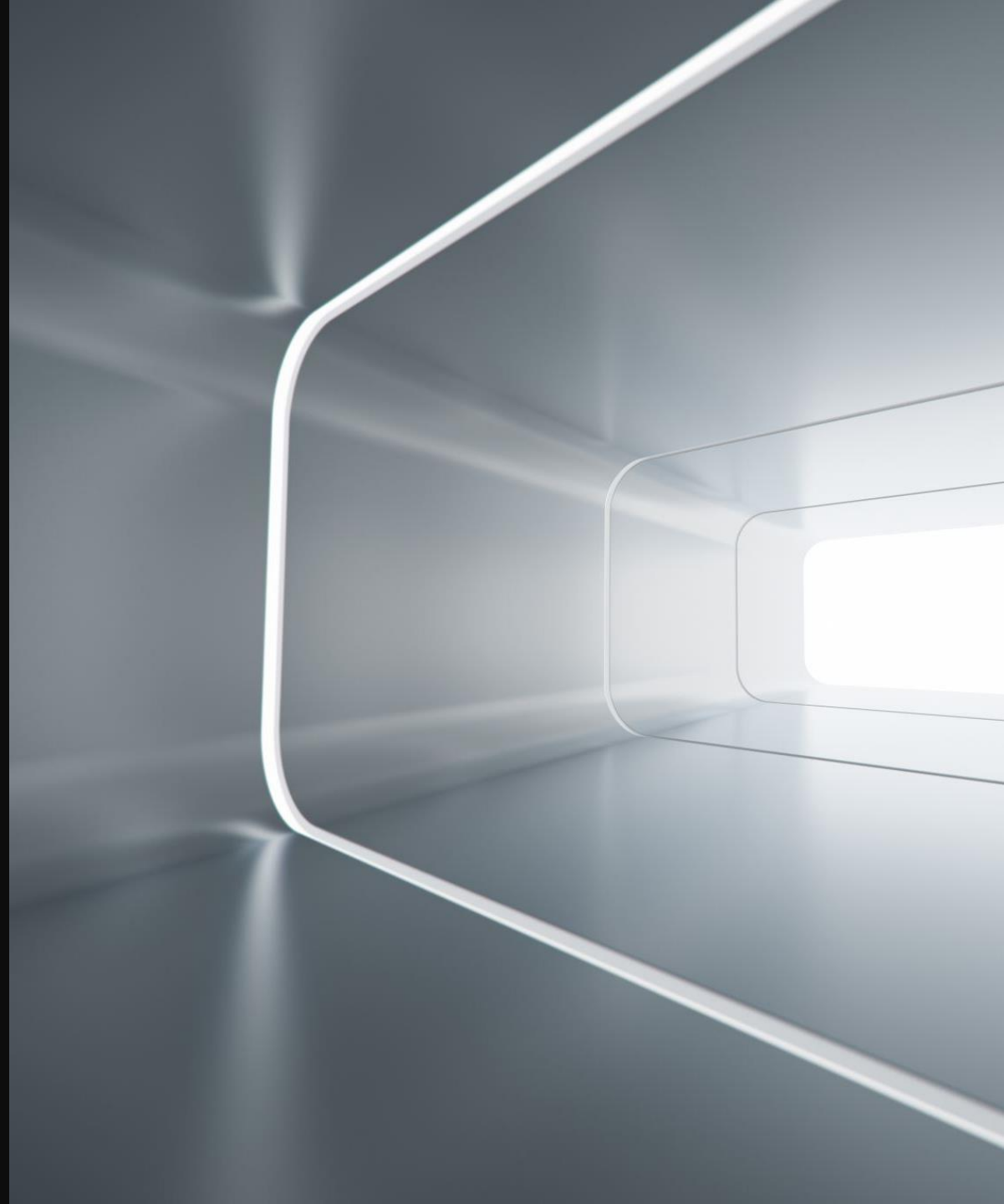
https://www.nhtsa.gov/press-releases/speed-campaign-speedingfatalities-14-year-high

Prediction

- The project aims to offer insights and predictions to inform policymakers, transportation authorities, and road safety organizations about accident severity factors, guiding the development of strategies to mitigate accidents.

Future Scope

- We can use machine learning models to find the hotspots and help us find routes with minimal traffic.




Thank You!