Motor Vehicle Collision Analysis: AIT 664 Team - NITRO

Akhila Kudupudi
Arika Bhattarai
Suparna Mannava

GROUP MEMBERS

1. Akhila Kudupudi
2. Suparna Mannava
3. Arika Bhattarai

S.No	Name	Responsibility
1.	Akhila	Data Cleaning and Feature Engineering
2.	Suparna	Data Analysis and Documentation
3.	Arika	Machine Learning Modelling

INTRODUCTION

Traffic collisions are a significant public health and safety concern globally, causing millions of injuries and thousands of deaths every year. In densely populated urban areas, the complexity of road networks, combined with high traffic volumes, increases the likelihood of accidents. With the availability of large datasets on vehicle collisions, it is now possible to analyze vast amounts of data to uncover patterns that can inform policy decisions, infrastructure improvements, and targeted safety interventions. By leveraging data analytics, machine learning, and statistical models, researchers can gain deeper insights into the contributing factors of vehicle accidents, such as driver behavior, vehicle characteristics, and external conditions like weather and road infrastructure.

This research aims to explore comprehensive vehicle collision datasets to identify the most significant factors contributing to road accidents and their severity. In particular, the focus will be on analyzing how variables such as driver distraction, vehicle type, weather conditions, and road infrastructure interact to influence crash outcomes. For example, driver inattention or distraction has been identified as a leading cause of vehicle accidents, yet the extent to which this factor interacts with environmental conditions, such as rain or poor lighting, remains underexplored. Similarly, while larger vehicles such as SUVs may offer greater protection to drivers, their

involvement in collisions can pose a higher risk to other road users, including pedestrians and smaller vehicles.

Furthermore, the research will explore whether predictive models can be developed to anticipate high-risk areas or times for collisions based on historical data, potentially informing traffic management and safety campaigns. Ultimately, this study aims to contribute to a deeper understanding of vehicle collisions and to provide actionable insights that can lead to safer roads and more informed traffic policies.

RESEARCH QUESTIONS

1. What are the monthly and yearly trends in crash occurrences? Are there specific months or years with significantly higher crash rates?
2. How does the time of day influence the frequency and severity of crashes? Are certain times (morning, afternoon, evening) more prone to accidents?
3. What role does the driver's sex play in crash outcomes? Is there a noticeable difference in crash involvement between male and female drivers?
4. How does the number of vehicle occupants relate to crash severity? Does having more occupants increase the likelihood of injuries?
5. Which vehicle types are most frequently involved in collisions?

SOURCE:

https://catalog.data.gov/dataset/motor-vehicle-collisions-vehicles

Size: The dataset contains **10,48,576 rows** and **25 columns**.

ATTRIBUTES:

The dataset contains the following attributes (features):

1. UNIQUE_ID - Integer
2. COLLISION_ID - Integer
3. CRASH_DATE - Date
4. CRASH_TIME - Time
5. VEHICLE_ID - String

6. STATE_REGISTRATION - String
7. VEHICLE_TYPE - String
8. VEHICLE_MAKE - String
9. VEHICLE_MODEL - String
10. VEHICLE_YEAR - Integer
11. TRAVEL_DIRECTION - String
12. VEHICLE_OCCUPANTS - Integer
13. DRIVER_SEX - String
14. DRIVER_LICENSE_STATUS - String
15. DRIVER_LICENSE_JURISDICTION - String
16. PRE_CRASH - String
17. POINT_OF_IMPACT - String
18. VEHICLE_DAMAGE - String
19. VEHICLE_DAMAGE_1 - String
20. VEHICLE_DAMAGE_2 - String
21. VEHICLE_DAMAGE_3 - String
22. PUBLIC_PROPERTY_DAMAGE - String
23. PUBLIC_PROPERTY_DAMAGE_TYPE - String
24. CONTRIBUTING_FACTOR_1 - String
25. CONTRIBUTING_FACTOR_2 – String

DATA PREPROCESSING/CLEANING:

Data Loading:

The original dataset was loaded into a Pandas DataFrame. The dataset contained columns such as UNIQUE_ID, COLLISION_ID, CRASH_DATE, VEHICLE_ID, VEHICLE_TYPE, and others. Due to the large dataset size, a decision was made to filter the data from 2020 onwards for analysis.

Column Selection and Dropping Irrelevant Columns:

Certain columns were dropped due to high percentages of missing values, such as VEHICLE_MODEL, VEHICLE_YEAR, PUBLIC_PROPERTY_DAMAGE_TYPE, VEHICLE_DAMAGE_1, VEHICLE_DAMAGE_2, and VEHICLE_DAMAGE_3.

Dropping these columns helped reduce the complexity of the analysis and eliminated non-informative attributes.

Handling Missing Values:

Missing values in critical columns like VEHICLE_MAKE, DRIVER_LICENSE_STATUS, DRIVER_LICENSE_JURISDICTION, PRE_CRASH, POINT_OF_IMPACT, VEHICLE_DAMAGE, PUBLIC_PROPERTY_DAMAGE, CONTRIBUTING_FACTOR_1, and CONTRIBUTING_FACTOR_2 were filled with a placeholder value, 'Unknown', to indicate that the information was not available.

This approach allowed the analysis to proceed without discarding potentially valuable records.

Removing Duplicates:

Duplicate records were identified and dropped to ensure data integrity and prevent skewing of the analysis.

Transforming Data:

The TRAVEL_DIRECTION column was transformed into a numerical representation (TRAVEL_DIRECTION_CODE), mapping directions such as 'North', 'South', 'East', and 'West' to specific integer values (1, 2, 3, and 4). Any other direction or missing values were set to 0.

Validating Data:

Rows where the number of VEHICLE_OCCUPANTS was less than or equal to zero were filtered out, as such records would not be useful for analysis.

Aggregating Data:

An aggregation was performed to sum the number of occupants for each vehicle make, which provided insights into the distribution of vehicle occupants by different makes.

Filtering Data by Date:

Data was filtered to include only records from the year 2020 onwards, given the large dataset size and to focus on more recent trends in the analysis.

Further Cleaning by Dropping Rows with Missing Values in Critical Columns:

Columns deemed essential for the analysis (STATE_REGISTRATION, VEHICLE_TYPE, TRAVEL_DIRECTION, and DRIVER_SEX) were checked, and rows with missing values in these columns were dropped.

The dataset was also cleaned of rows with missing values in the VEHICLE_OCCUPANTS column.

Saving the Cleaned Data:

The cleaned and preprocessed DataFrame was saved to a new CSV file (filtered_motor_vehicle_collisions_2020_onwards.csv), ensuring the refined data could be used for further analysis and modeling.

These data cleaning and preprocessing steps were carried out using Jupyter Notebook, which facilitated the interactive exploration and transformation of the dataset.

FEATURES AND FEATURE SELECTION:

FEATURE SELECTION:

- Original Features:
  The original dataset contained the following features:
    - UNIQUE_ID
    - COLLISION_ID
    - CRASH_DATE
    - CRASH_TIME
    - VEHICLE_ID
    - STATE_REGISTRATION

  - VEHICLE_TYPE
  - VEHICLE_MAKE
  - VEHICLE_MODEL
  - VEHICLE_YEAR
  - TRAVEL_DIRECTION
  - VEHICLE_OCCUPANTS
  - DRIVER_SEX
  - DRIVER_LICENSE_STATUS
  - DRIVER_LICENSE_JURISDICTION
  - PRE_CRASH
  - POINT_OF_IMPACT
  - VEHICLE_DAMAGE
  - VEHICLE_DAMAGE_1
  - VEHICLE_DAMAGE_2
  - VEHICLE_DAMAGE_3
  - PUBLIC_PROPERTY_DAMAGE
  - PUBLIC_PROPERTY_DAMAGE_TYPE
  - CONTRIBUTING_FACTOR_1
  - CONTRIBUTING_FACTOR_2

- Selected Features After Cleaning:

After analyzing the relevance and quality of the data, the following features were retained:

  - UNIQUE_ID
  - COLLISION_ID
  - CRASH_DATE
  - CRASH_TIME
  - VEHICLE_ID

  - STATE_REGISTRATION
  - VEHICLE_TYPE
  - VEHICLE_MAKE
  - TRAVEL_DIRECTION
  - VEHICLE_OCCUPANTS
  - DRIVER_SEX
  - DRIVER_LICENSE_STATUS
  - DRIVER_LICENSE_JURISDICTION
  - PRE_CRASH
  - POINT_OF_IMPACT
  - VEHICLE_DAMAGE
  - PUBLIC_PROPERTY_DAMAGE
  - CONTRIBUTING_FACTOR_1
  - CONTRIBUTING_FACTOR_2
  - TRAVEL_DIRECTION_CODE (a newly derived feature)

- Features Dropped:

Certain features were dropped due to high percentages of missing values or lack of relevance to the analysis, including:

  - VEHICLE_MODEL
  - VEHICLE_YEAR
  - PUBLIC_PROPERTY_DAMAGE_TYPE
  - VEHICLE_DAMAGE_1
  - VEHICLE_DAMAGE_2
  - VEHICLE_DAMAGE_**3**

EXPLORATORY DATA ANALYSIS:

1. Descriptive Statistics (Value Counts):

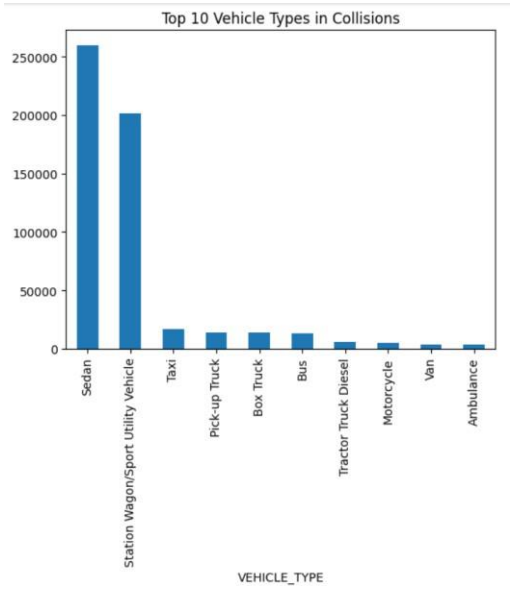
```
VEHICLE_TYPE
Sedan                                259738
Station Wagon/Sport Utility Vehicle  201161
Taxi                                  17238
Pick-up Truck                         14153
Box Truck                             13776
                                      ...
SAVANA                                    1
Kubota bac                                1
Ice cream                                 1
Postoffice                                1
MDX                                       1
Name: count, Length: 1099, dtype: int64
DRIVER_SEX
M    418302
F    142994
U      1270
Name: count, dtype: int64
```

This table provides the frequency count of the unique values in two categorical columns named VEHICLE_TYPE and DRIVER_SEX.

By far the most frequent collision vehicle types are Sedans (259,738) followed by Station Wagons/SUVs (201,161).

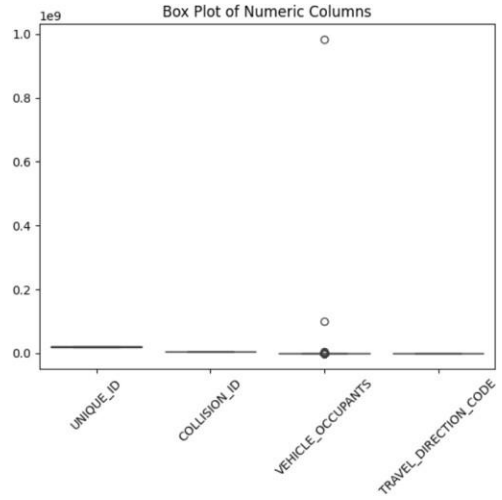
For DRIVER_SEX, Males (M) are involved in more collisions - 418,302 compared to Females (F) - 142,994. Unknown values - U - are comparatively few.

2. Bar Chart: Top 10 Vehicle Type in Collisions:



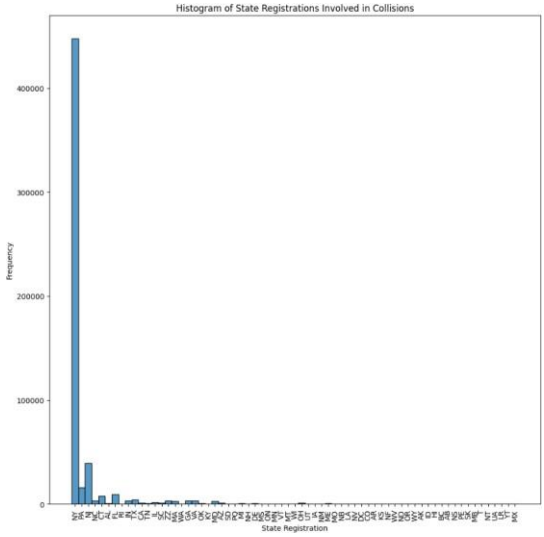
This is a bar chart showing the top 10 vehicle types involved in collisions.The most frequent types are sedans and station wagons/SUVs, by far outnumbering other types such as Taxis and Pick-up Trucks. This shows that Sedans and SUVs are the most common types of vehicles owned.

3. Box Plot Numeric Columns:



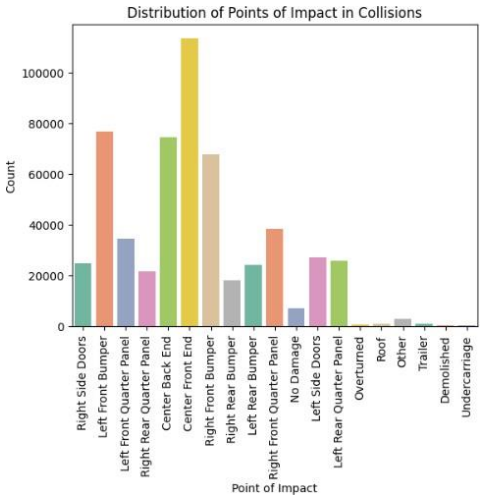
A box plot showing outliers for numeric columns in the data. For example, VEHICLE_OCCUPANTS Outliers are fairly apparent in the VEHICLE_OCCUPANTS column as there are a few highly dissimilar points from the trend of the rest. Other numeric columns like UNIQUE_ID, COLLISION_ID, etc., have very dense values and no significant outliers.

4. Univariate Analysis - Histogram: State Registrations:



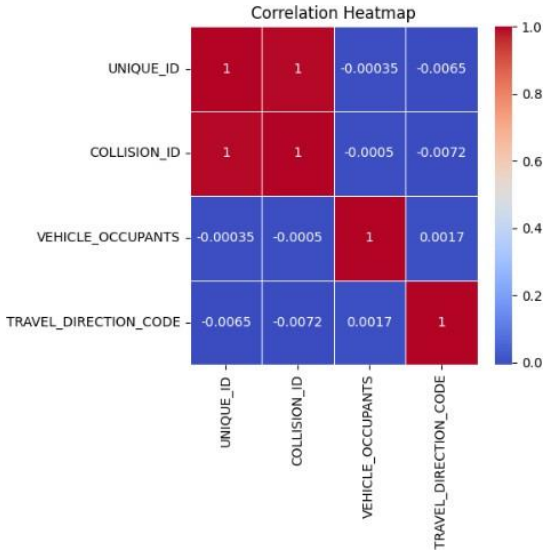
This histogram represents the frequency of vehicle registrations by state involved in collisions. The count of registrations is dominated by a single state, probably the focus of this dataset, for example, NY. Other states have much lower frequencies.

5. Bar Chart-Distribution of Points of Impact:



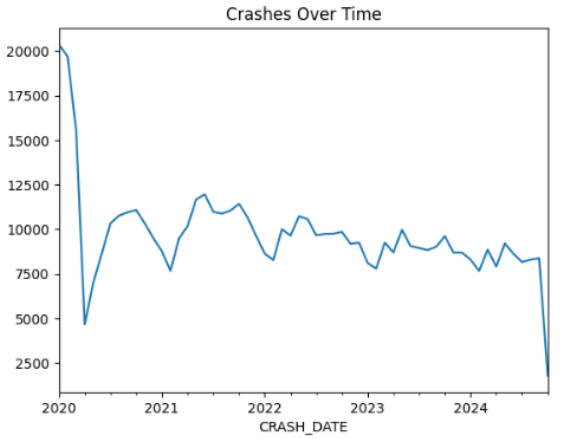
This is a bar chart showing the distribution of collision points of impact-in other words, points such as Center Front, Right Side Doors, among others. Center Front End is the most common point of impact, followed by areas like Left Front Bumper and Right Front Bumper. These follow the trends of frontal collision being the most frequent likely due to head-on or rear-end collisions.

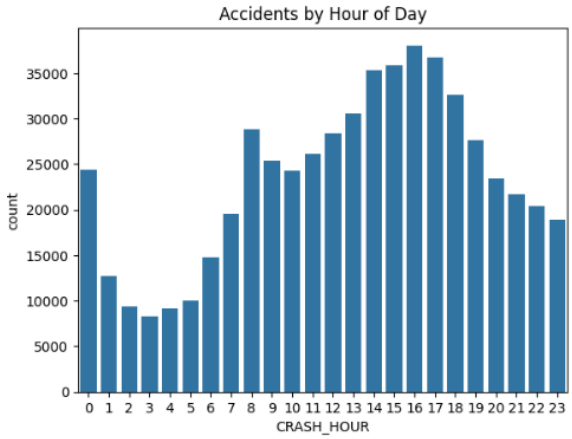
6. Bivariate Analysis - Correlation Heatmap:



The following represents the correlation among numeric columns like UNIQUE_ID, VEHICLE_OCCUPANTS, and TRAVEL_DIRECTION_CODE. For most columns, the correlations are near zero, suggesting weak or no linear relationships. Not much strong correlation of other variables with VEHICLE_OCCUPANTS.

7. Time Series Analysis:





Crashes Over Time: The line chart below represents the number of crashes over time - date. At constant values in years, crashes decrease in the last period since the dataset might be incomplete or missing records.

Accidents by Hour of Day: The following bar chart visualizes the distribution of crashes due to the hour of day. Crashes peak in the afternoon hours around 3 PM, there is arguably higher traffic volume due to either the commute or school pickup.

MACHINE LEARNING/ DATA MINING MODELS

1. What are the monthly and yearly trends in crash occurrences? Are there specific months or years with significantly higher crash rates?

Performed ARIMA model for the forecasting results.

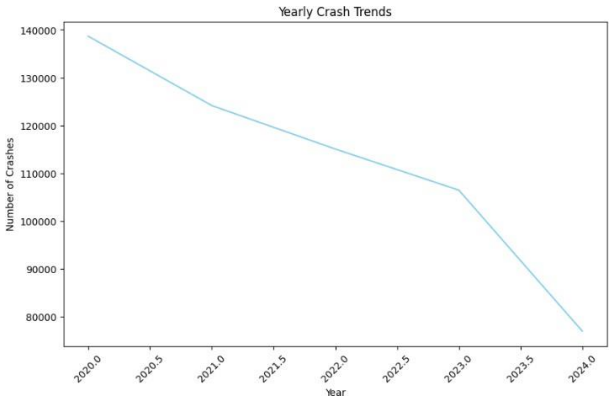


Fig: 1

Yearly Crash Trends

- **Description**: This line chart shows the total number of crashes per year.
- **Observation**:
    - Crashes have been consistently decreasing from 2020 to 2024.
    - The decline may indicate improvements in road safety or reduced travel during certain periods (e.g., pandemic restrictions).

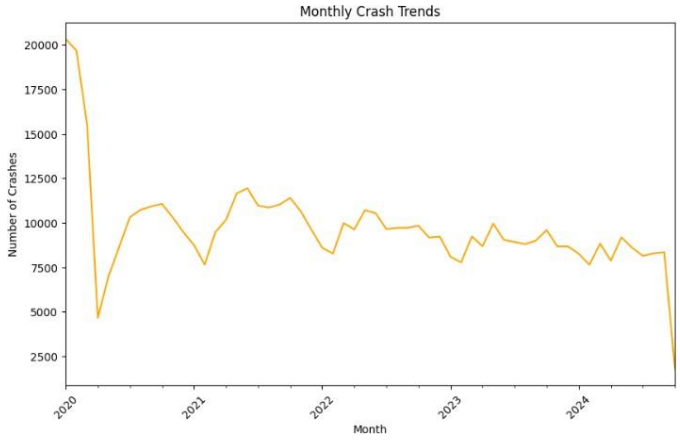


Fig: 2

Monthly Crash Trends

- **Description**: This line chart shows the total number of crashes per month from 2020 onward.
- **Observation**:
    - A sharp drop in crashes at the beginning of 2020, likely due to COVID-19 restrictions.
    - Fluctuations in crash numbers show periodic peaks, potentially influenced by seasonal factors like weather or holidays.
    - The overall trend aligns with the yearly decline.

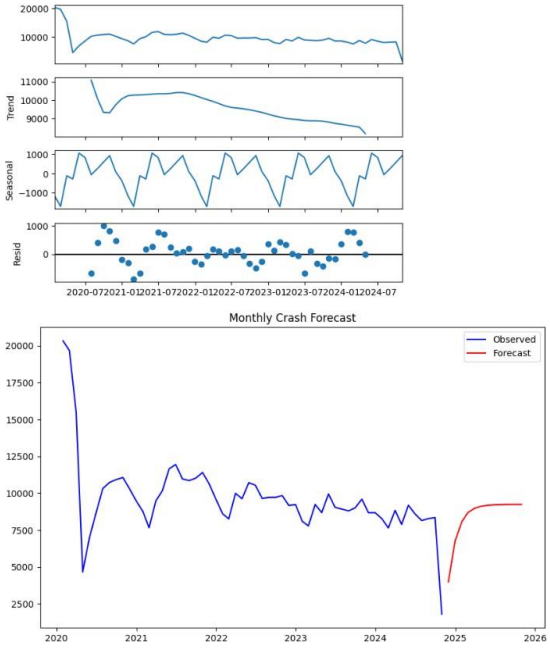


Fig: 3

Crash Forecast and Seasonal Decomposition

- **Decomposition**:
    - **Trend**: A clear downward trend in crashes over time.

- **Seasonal**: Regular monthly fluctuations show recurring patterns, likely linked to specific times of the year.
- **Residual**: Unexplained variations that are not captured by the trend or seasonality.

- **Forecast**:
  - The observed crash numbers (blue line) show historical trends.
  - The forecast (red line) predicts a stabilization or slight increase in crashes after 2024, suggesting potential changes in factors affecting crashes.

2. How does the time of day influence the frequency and severity of crashes? Are certain times (morning, afternoon, evening) more prone to accidents?

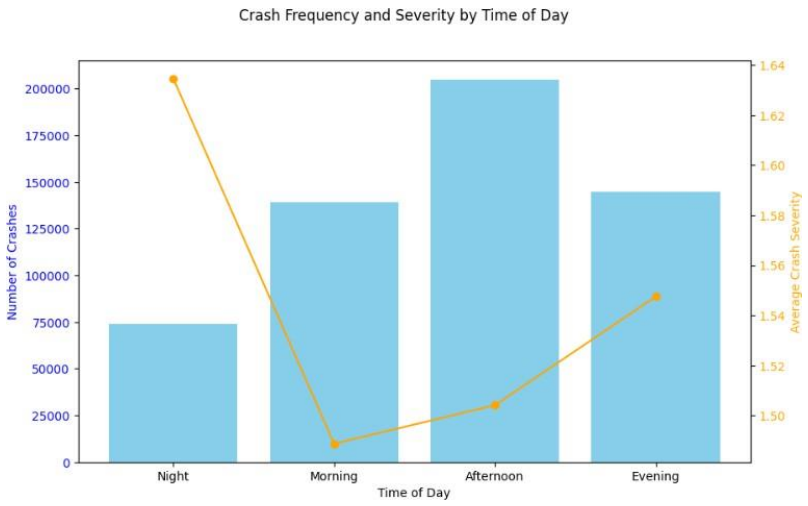


Fig: 4

Blue Bars (Left Y-Axis): Represent the number of crashes for each time of day:

- **Night**: Lowest number of crashes (~75,000).
- **Morning**: Moderate number of crashes (~125,000).
- **Afternoon**: Highest number of crashes (~200,000).
- **Evening**: Second-highest number (~150,000).

Orange Line (Right Y-Axis): Shows the average crash severity for each time of day:

- **Night**: Highest average severity (~1.62).
- **Morning**: Lowest average severity (~1.5).
- **Afternoon**: Gradual increase in severity (~1.52).
- **Evening**: Higher severity (~1.56).

Insights:

1. **Crash Frequency**:
    - **Afternoon** has the highest frequency of crashes, likely due to increased traffic during peak hours.
    - **Night** has the lowest frequency, possibly due to fewer vehicles on the road.
2. **Crash Severity**:
    - Crashes at **Night** are the most severe, likely due to poor visibility and higher speeds.
    - **Morning** crashes tend to be less severe, possibly due to slower speeds during morning commutes.
3. **Combined Insight**:
    - While afternoons see the highest crash frequency, nights have the highest crash severity, indicating the need for targeted interventions like improved street lighting and night patrols.

3. What role does the driver's sex play in crash outcomes? Is there a noticeable difference in crash involvement between male and female drivers?

```
Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00      9339
           1       0.39      0.00      0.00     34826
           2       0.60      1.00      0.75     67152
           3       0.00      0.00      0.00       943

    accuracy                           0.60    112260
   macro avg       0.25      0.25      0.19    112260
weighted avg       0.48      0.60      0.45    112260

              Feature  Importance
2  Time_Range_Encoded    0.443139
1   VEHICLE_OCCUPANTS    0.394175
0  DRIVER_SEX_Encoded    0.162685
```

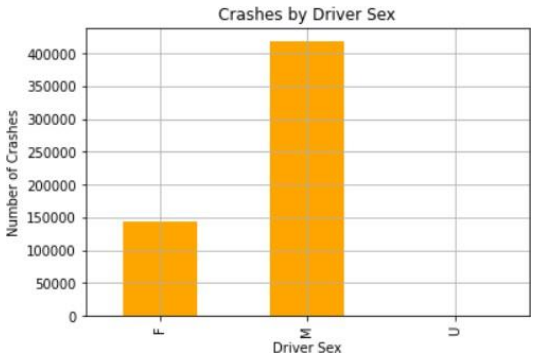


Fig: 5

Performed Random Forest Classifier for the results. Male drivers account for the majority of crashes (over 400,000), while female drivers are involved in significantly fewer (~150,000). Crashes with unspecified driver sex are negligible. This suggests male drivers are more frequently involved, possibly due to higher exposure or risk-taking behaviors.

4. Which vehicle types are most frequently involved in collisions?

Fig: 6

The heatmap illustrates the distribution of crashes based on **vehicle type** and **number of occupants**:

1. **Most Common Vehicle Types**: Sedans and SUVs dominate crashes, especially with **1 occupant**.
2. **Single-Occupant Crashes**: Crashes involving only **1 occupant** are the most frequent across all vehicle types.
3. **Higher Occupant Counts**: Vehicles like **buses**, **SUVs**, and **vans** are more associated with crashes involving higher occupant counts (4-6+).
4. **Motorcycles and Pickup Trucks**: These mostly involve crashes with **1-2 occupants**, as expected due to their design.

CONCLUSION

In conclusion, this study provides valuable insights into the critical trends and factors influencing motor vehicle collisions. The analysis highlights seasonal variations in crash frequency and severity, the significant role of driver demographics, and vehicle types in collision outcomes. Targeted interventions, such as improved night visibility, enhanced traffic management strategies, and public awareness campaigns, are crucial for mitigating these risks.

By leveraging data analytics and machine learning, this study provides actionable insights to improve road safety and traffic management. Collaborative efforts among policymakers, engineers, and the public can create safer roads for all. These findings serve as a foundation for evidence-based decisions to reduce collision rates and enhance transportation safety.

FUTURE WORK

To further enhance the impact of this research, future work will focus on:

1. **Advanced Predictive Modeling**: Incorporating deep learning techniques and additional external factors, such as traffic flow data and road conditions, to improve model accuracy.
2. **Broader Data Integration**: Expanding the analysis to include datasets from multiple states or countries for more generalizable insights.

3. **Real-Time Analysis**: Developing real-time monitoring systems using sensor data or live traffic feeds to predict and prevent crashes.
4. **Policy Recommendations**: Collaborating with traffic authorities to translate findings into actionable strategies, such as redesigning high-risk intersections or improving driver education programs.
5. **Interactive Dashboards**: Building interactive visualization tools to make data-driven insights accessible to policymakers and stakeholders for decision-making.

By addressing these areas, this research can significantly contribute to reducing motor vehicle collisions and enhancing road safety on a broader scale.

REFERENCES

[1] U.S. Government Open Data. (n.d.). Motor Vehicle Collisions - Vehicles Dataset. Retrieved from https://catalog.data.gov/dataset/motor-vehicle-collisions-vehicles

[2] McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly Media.

[3] Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.

[4]Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. Retrieved from https://otexts.com/fpp3/arima.html

[5] Wickham, H., & Grolemund, G. (2016). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media.

[6] National Highway Traffic Safety Administration. (2022). Traffic Safety Facts Annual Report. Retrieved from https://crashstats.nhtsa.dot.gov

[7] World Health Organization. (2018). Global Status Report on Road Safety 2018. Retrieved from https://www.who.int/publications/i/item/9789241565684

[8] Seaborn Library Documentation. (n.d.). Retrieved from https://seaborn.pydata.org

Matplotlib Documentation. (n.d.). Retrieved from https://matplotlib.org/stable/index.html

[9] Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2015). Time Series Analysis: Forecasting and Control. Wiley.