GROUP 3 CAPSTONE PROJECT

DSF-PT08 PHASE V

GROUP MEMBERS

1. Mercy Juma mercy.juma@student.moringaschool.com
2. Shadrack Macharia shadrack.macharia@student.moringaschool.com
3. Hezekia Asaava hezekia.asaava@student.moringaschool.com
4. Mitchelle Joy Wayua mitchelle.wayua@student.moringaschool.com
5. Ndung'u Mburu ndung'u.mburu@student.moringaschool.com
6. AbdirahMan Abdi abdirahman.abdi@student.moringaschool.com

TITLE:PREDICTING TRAFFIC ACCIDENT SEVERITY

PRESENTATION OUTLINE

1. PROJECT BACKGROUND
2. OBJECTIVE
3. DATA UNDERSTANDING
4. EXPLORATORY DATA ANALYSIS (EDA)
5. MODELING & ANALYSIS
6. RESULTS & FINDINGS
7. CONCLUSION

BACKGROUND

Traffic accidents are a global concern, often due to the devastating consequences such as injuries, fatalities, and financial losses. Accidents severity is influenced by environmental, temporal and human factors. Understanding the role this variables play in determining accident severity can be critical for addressing public safety challenges.

The interplay between these factors creates a complex web of influences on accident severity. This project develops a predictive model that can estimate the severity of traffic incidents based on various parameters such as weather, nature of roads surface, time, day, among others .This initiative is not just about reducing the severity but also addresses the broader goal of integrating data science into urban planning and public safety decisions.

OBJECTIVES

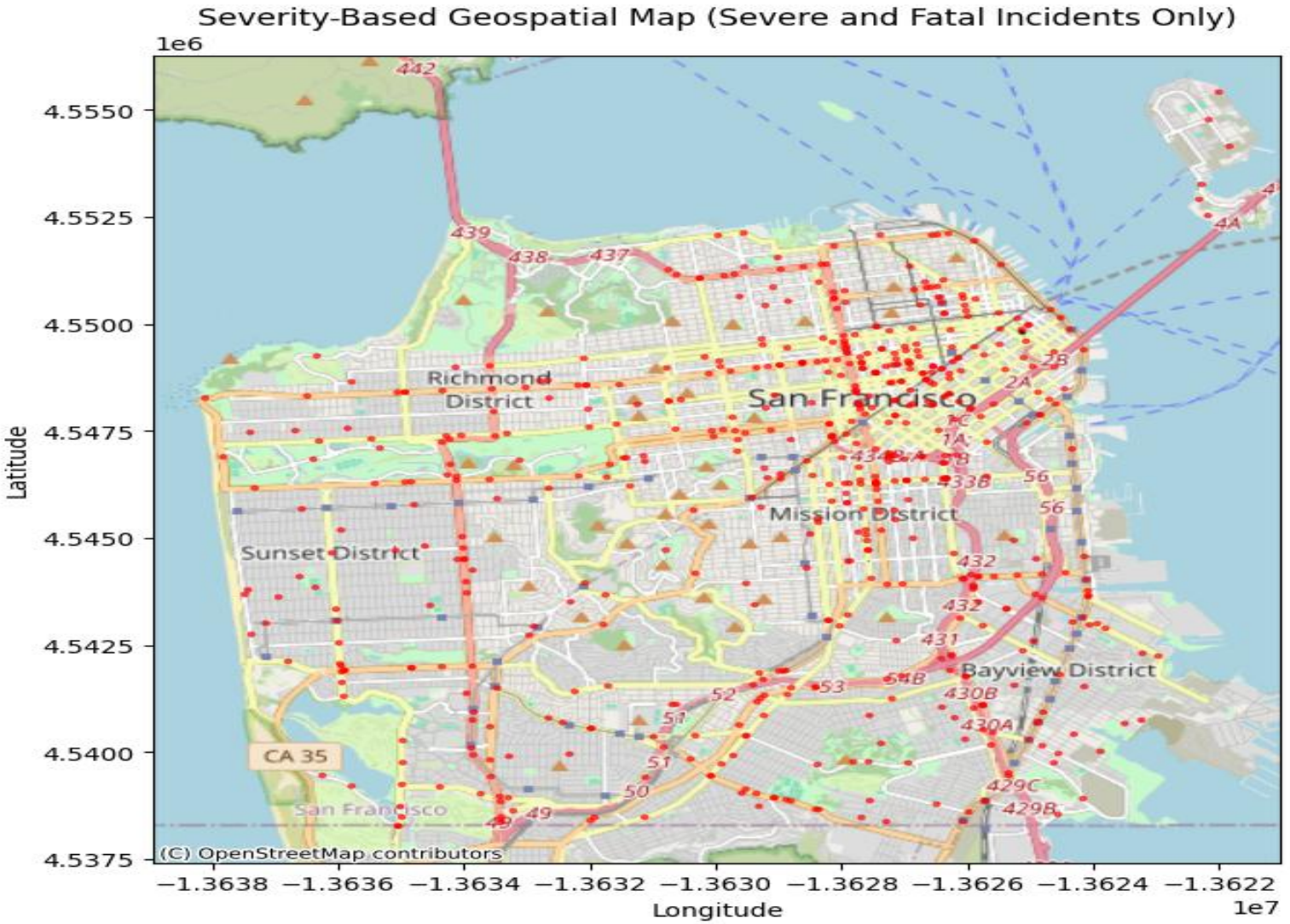
1. Identify high-risk neighborhoods and geographical hotspots for traffic incidents through geospatial mapping.
2. Build a predictive model: develop a machine learning model capable of predicting the severity of road traffic accidents based on relevant features such as casualty details, road type, location, weather conditions, and time-related factors.
3. Identify the most influential factors for predicting accident severity.

DATA UNDERSTANDING

The project utilizes traffic incident data from San Francisco open data source, to develop a model to predict the likelihood of incidents leading to severe outcomes.

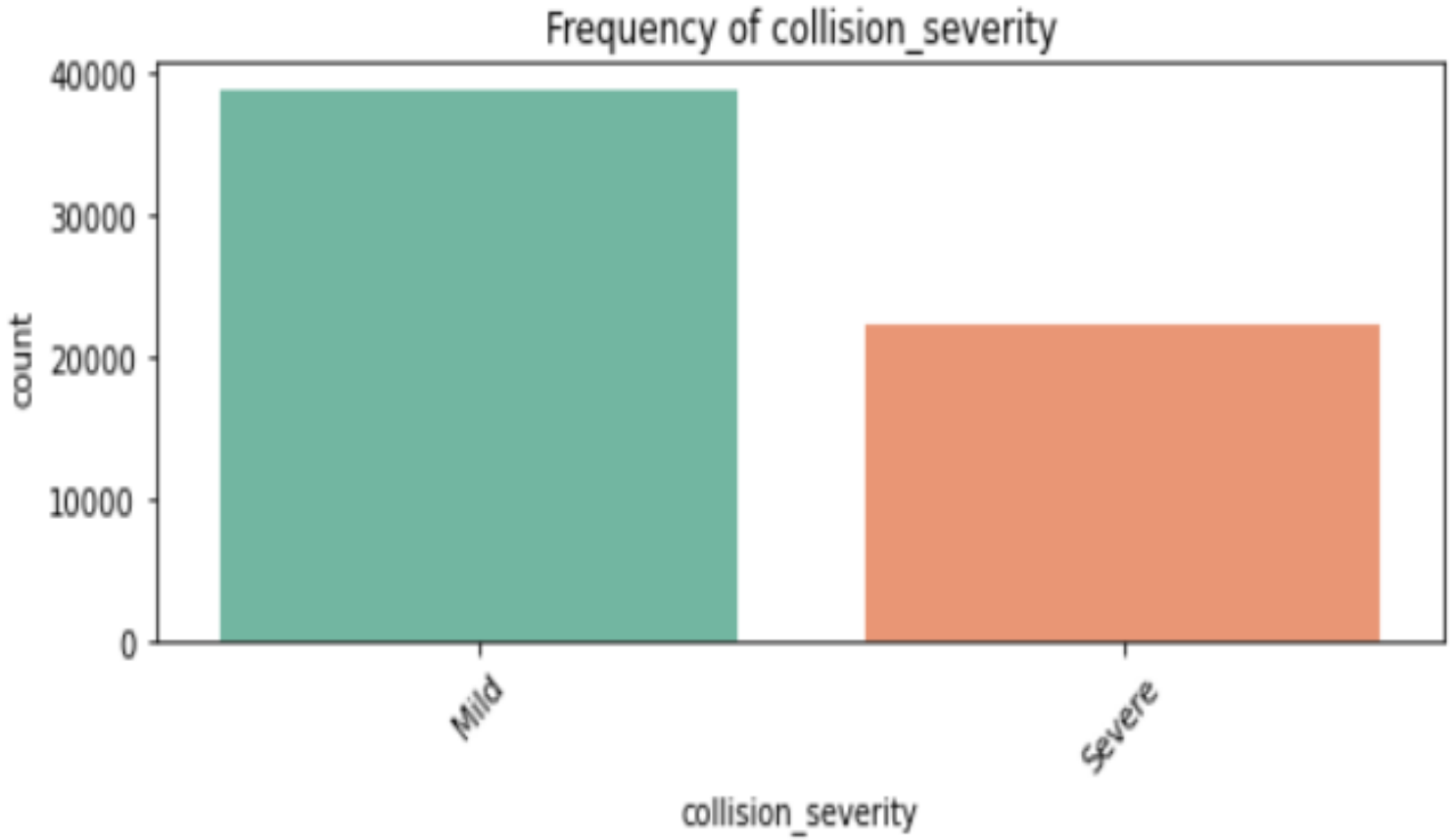
The data comprises of information on weather conditions(rainy, snow, fog etc.), time-related elements (time of the day, day of the week, month etc.), temporal factors such as direction of travel, existence of control devices, road surface among others.

EXPLORATORY DATA ANALYSIS: GEOSPATIAL MAP



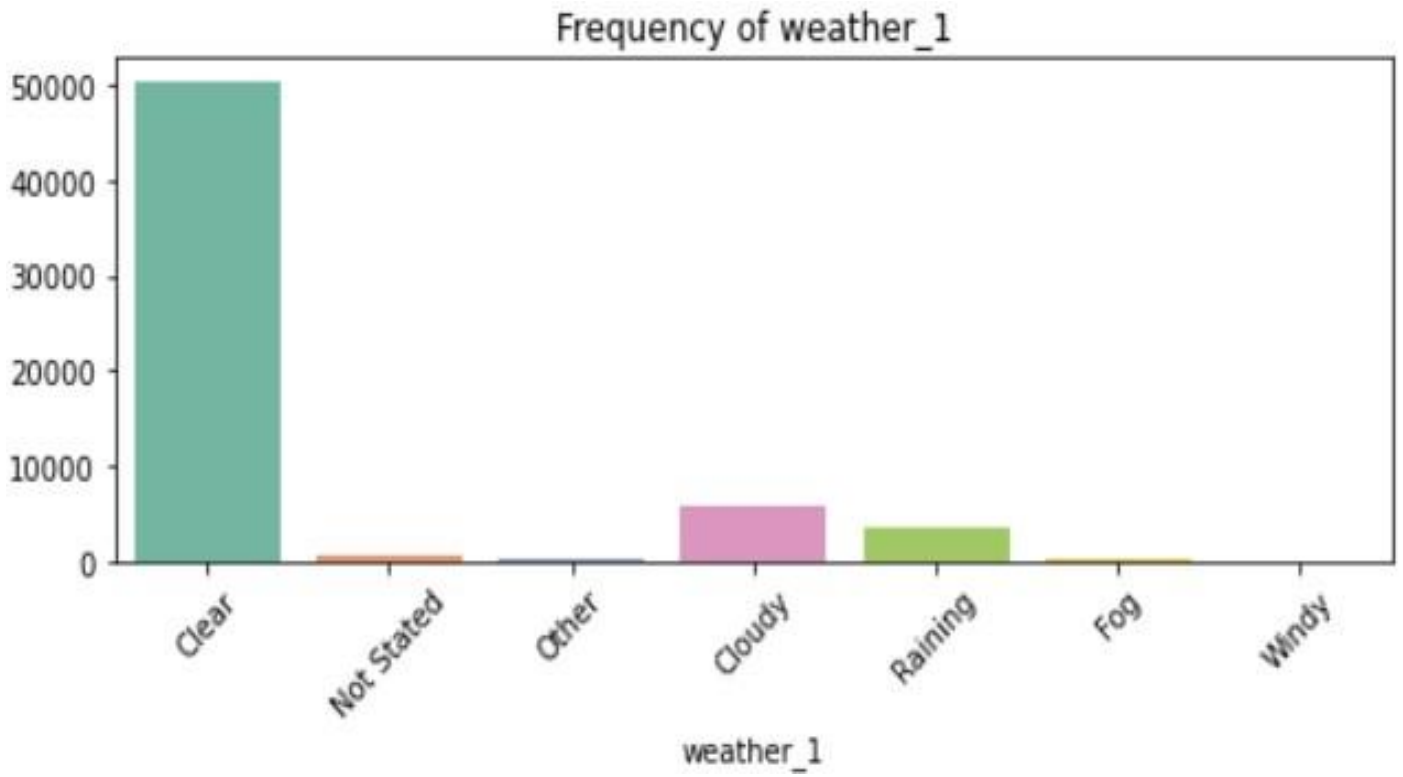
- *The map displays the geographical distribution of traffic incidents in SanFrancisco, categorized by severity.*
- *The red points highlight the areas where fatal accidents occurred.*
- *It is evident that the red points are clustered at intersections showing that these are blackspots.*

FREQUENCY OF COLLISION SEVERITY



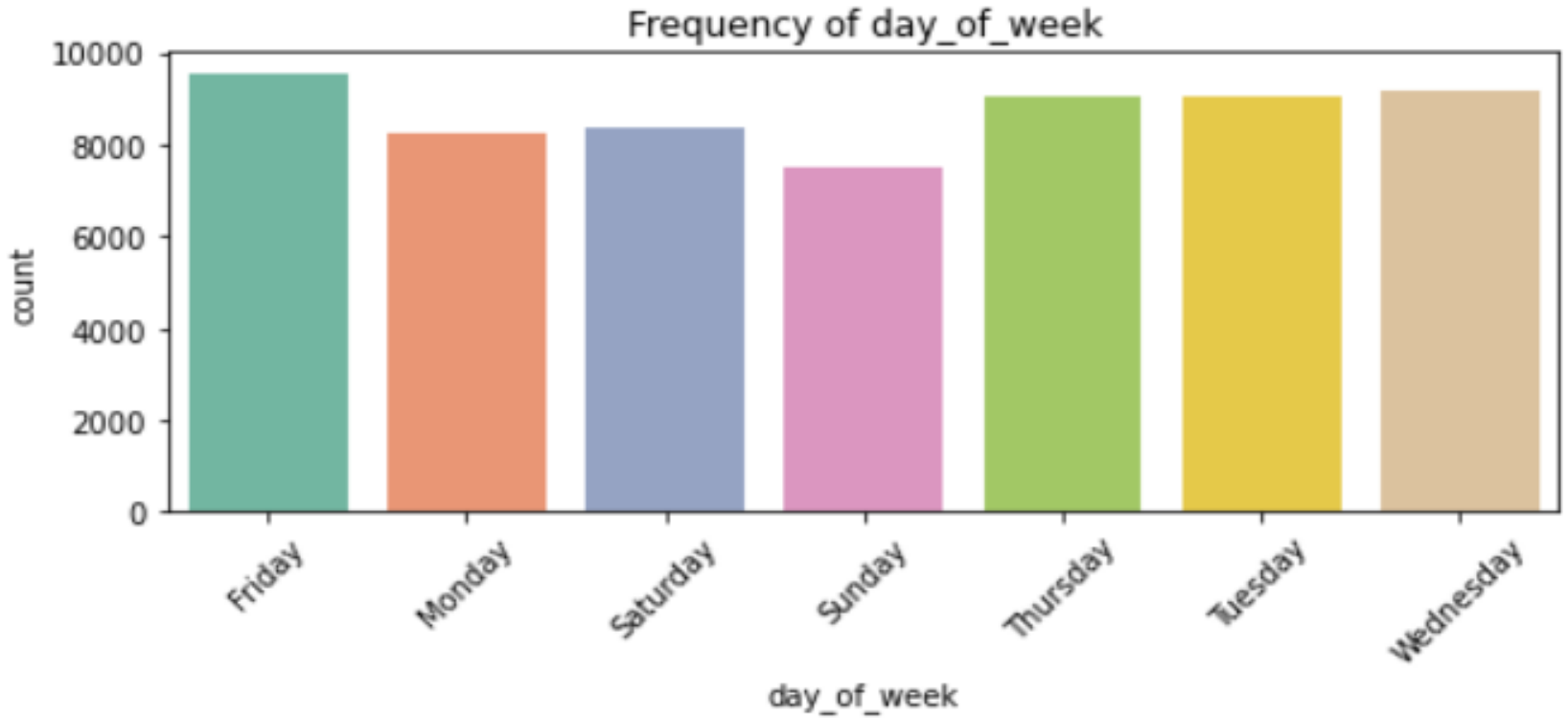
- *Most accidents are classified as mild.*
- *Takeout- There is class imbalance and hence most of the collision severities turned out to be mild than severe.*

ACCIDENT FREQUENCY BY WEATHER



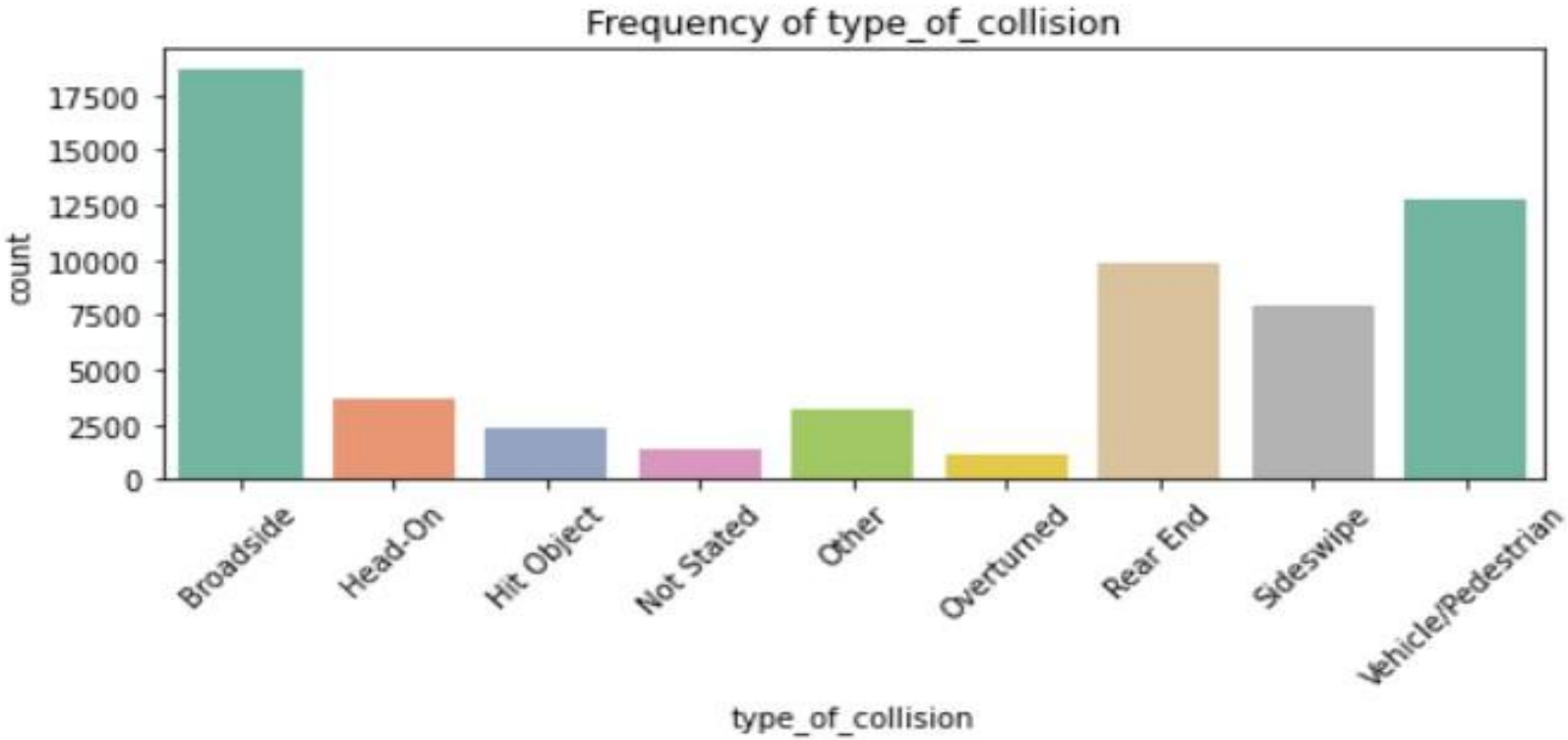
- *Most accidents occurred under clear weather.*
- *Takeout- Collisions often occur in clear weather hence weather conditions are not the only factors contributing to accidents.*

ACCIDENT FREQUENCY BY DAY OF THE WEEK



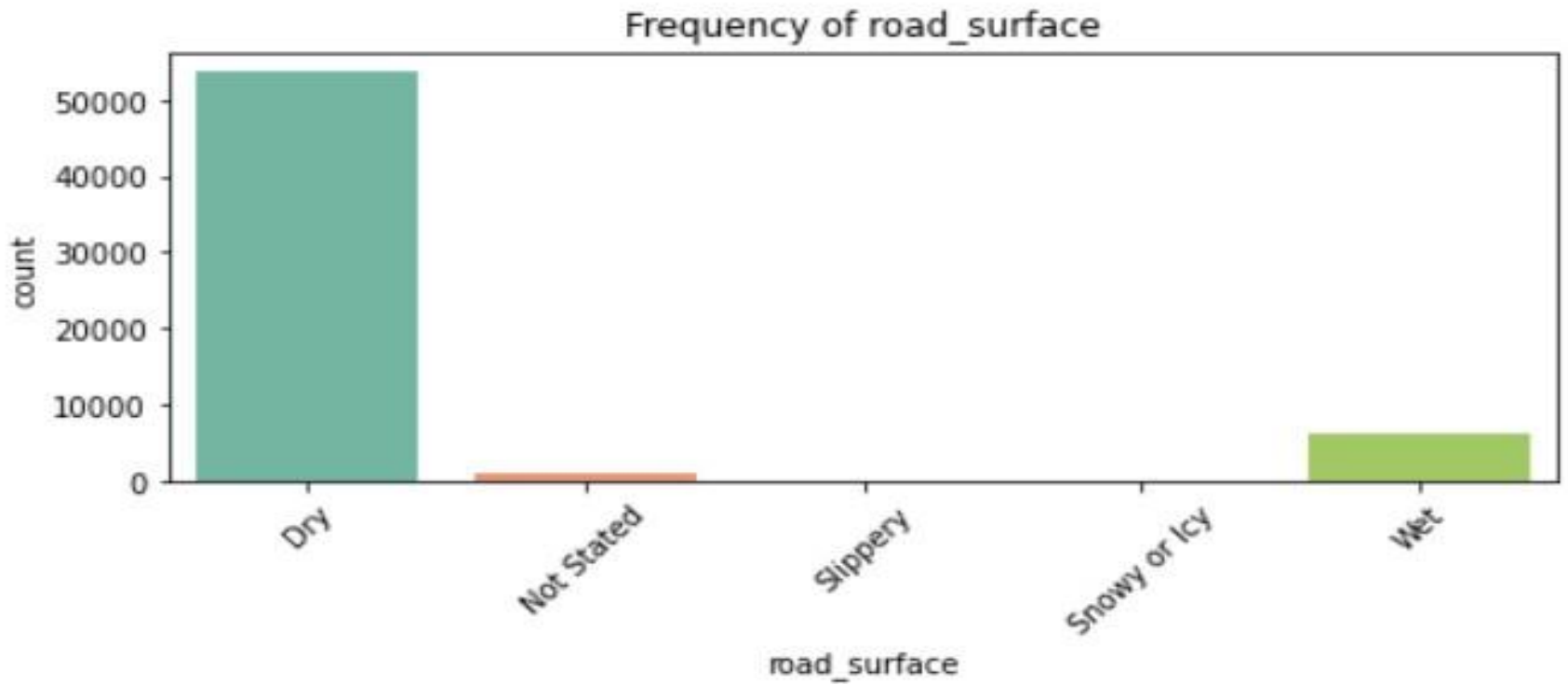
- *Most accident occur on Fridays , followed by Tuesday , Wednesday and Thursday.*
- *Takeout-The pattern of traffic accidents suggests that Fridays are particularly hazardous, with the highest number of incidents, likely due to increased activity and travel before the weekend.*

ACCIDENT FREQUENCY BY TYPE OF COLLISION



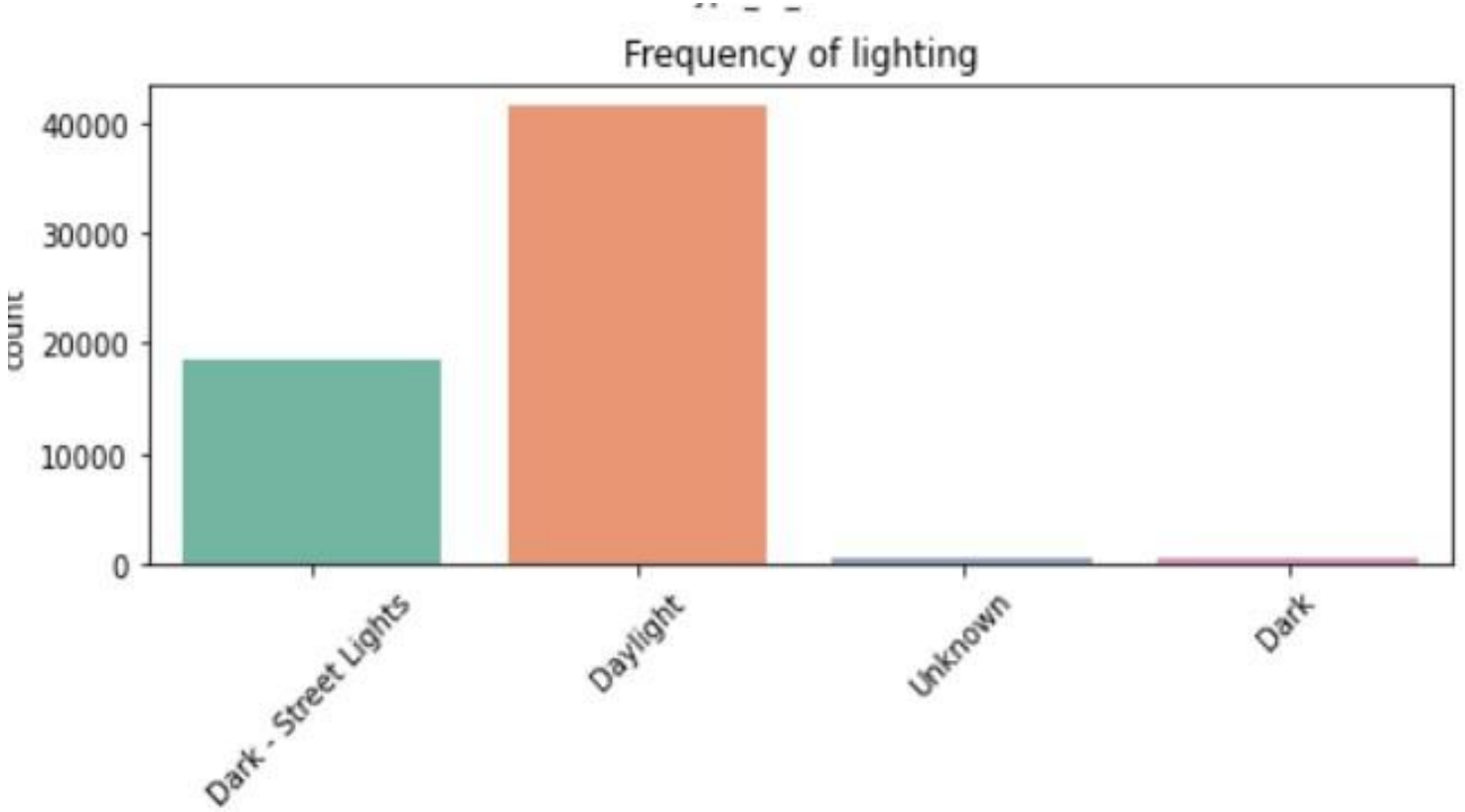
- *Broadside ,vehicle/pedestrian and rearend have the most frequent occurrences while overturned has the least occurrences.*
- *Takeout-Broadside collisions, often occurring at intersections, highlight the importance of improving traffic signal management and promoting driver awareness to mitigate intersection-related risks.*

ACCIDENT FREQUENCY BY ROAD SURFACE



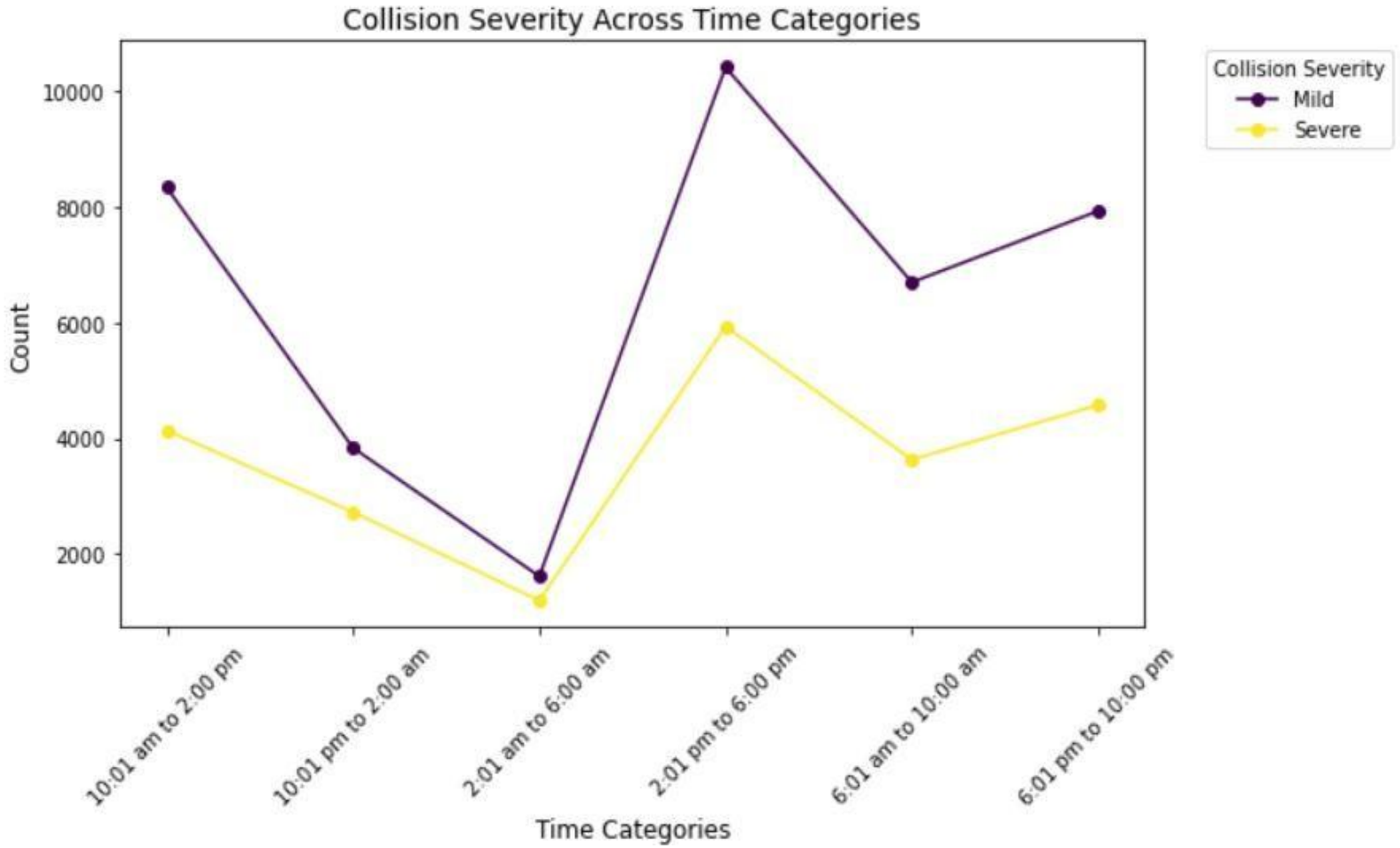
- *Most accidents occurred under dry road surface.*
- *Takeout-Road surface don't necessarily contribute to accidents , however other factors like driver fatigue sobriety and vehicle condition may have greater implication.*

ACCIDENT FREQUENCY BY LIGHTING



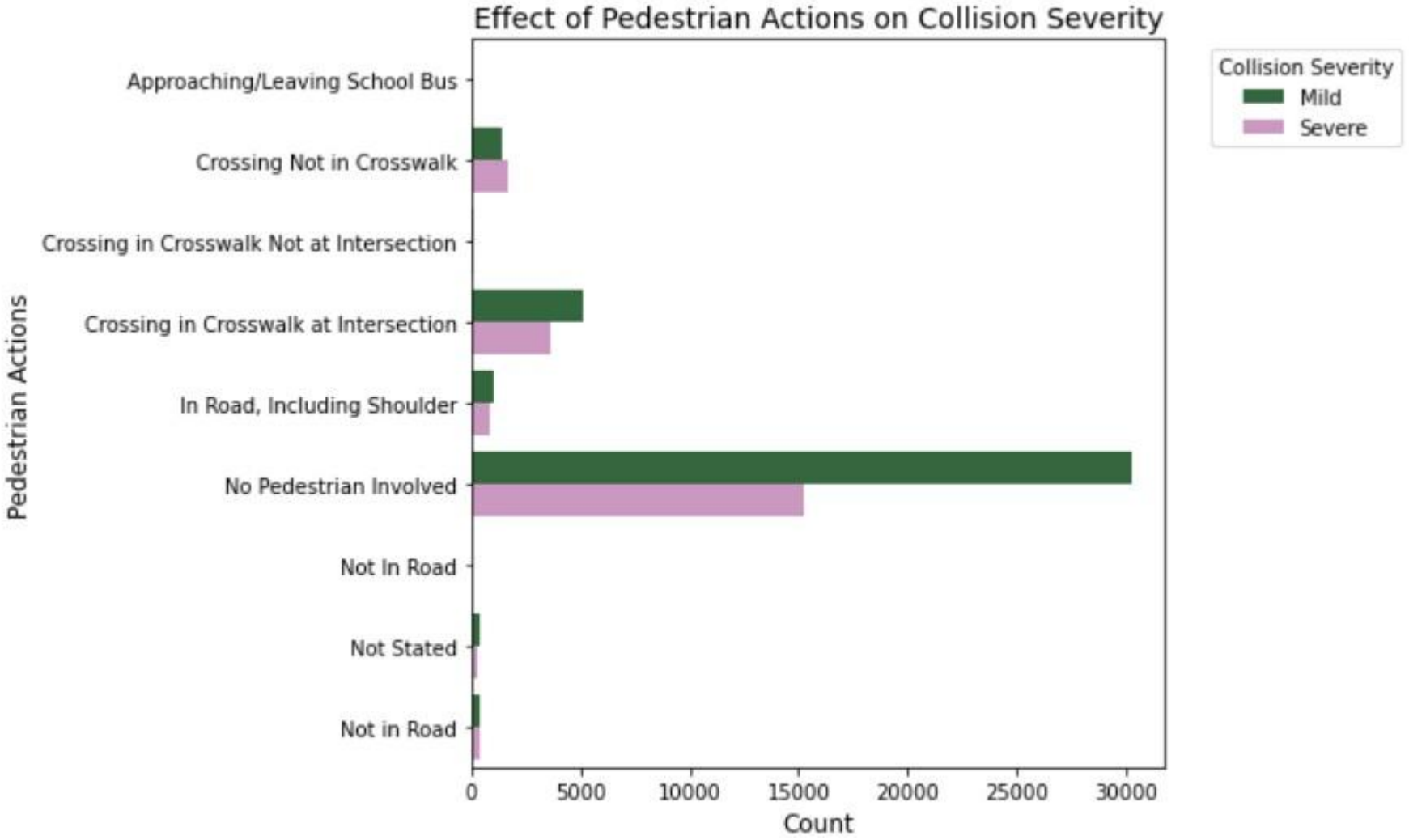
- *Collision are most frequent during daylight condition followed by street lighted condition at night.*
- *Takeout-Since most accidents have occurred during well lit conditions , hence proper lighting has minimal impact on occurrence of accidents*

COLLISION SEVERITY ACROSS TIME CATEGORIES



- *Time category 2:01pm to 6:00pm shows the highest collision counts for multiple severities while low collision counts occur during late night and early morning hours.*
- *This indicates a high concentration of traffic activity or risk factors during this time period, possibly due to peak hours or increased road usage.*
- *Takeout- Traffic safety interventions might be more effective if targeted at the afternoon to early evening and morning commute periods.*

EFFECT OF PEDESTRIAN ACTION ON COLLISION SEVERITY



- *The category where no pedestrian is involved dominates the data. It highlights the dominance of vehicle-to-vehicle collision.*
- *Takeout-This analysis highlights the importance of focusing safety interventions on high-risk pedestrian behaviors and locations, such as crosswalks at intersections.*

MODELING AND ANALYSIS

The project is a classification problem and uses Logistic regression technique for a data analysis with comparative analysis with Random Forest Classifier, XGBoost ,Support Vector classifier and the K-Nearest Neighbor classifier.

From the analysis, Support Vector Classifier, Logistic Regression, and XGBoost are the best performing models for Mild collisions, with high recall and F1-scores, the overall accuracy for both models remained unchanged at 0.66.

MODEL RESULTS

Model	Accuracy	Precision (Mild)	Recall (Mild)	F1-Score (Mild)	Precision (Severe)	Recall (Severe)	F1-Score (Severe)
Random Forest	0.64	0.68	0.82	0.74	0.49	0.31	0.38
Logistic Regression	0.66	0.68	0.89	0.77	0.56	0.24	0.34
Support Vector Classifier	0.66	0.67	0.91	0.77	0.55	0.20	0.30
k-Nearest Neighbors	0.61	0.67	0.76	0.72	0.44	0.33	0.38
XGBoost	0.66	0.69	0.87	0.77	0.55	0.29	0.38

- *This table summarizes the Accuracy, Precision, Recall, and F1-Score for various models used in predicting collision severity. These metrics were chosen to evaluate the models effectively, especially in cases of class imbalance.*

RESULTS AFTER FEATURE IMPORTANCE

- Since both Random Forest and XGBoost are tree-based models, they inherently provide a measure of feature importance based on how much each feature contributes to reducing uncertainty.
- After reviewing the top features from both models, we selected the following features for further modeling: 'distance', 'number_injured', 'day_of_week', 'party1_dir_of_travel', 'number_killed', 'type_of_collision', and 'party2_dir_of_travel'.
- Overall, both Logistic Regression and XGBoost perform better with Mild collision predictions but still miss a significant number of Severe collisions. Improvements in identifying Severe cases could involve tuning the models further or using specialized techniques for class imbalance, such as oversampling Severe cases or using weighted loss functions.

Model	Accuracy	Precision (Mild)	Recall (Mild)	F1-Score (Mild)	Precision (Severe)	Recall (Severe)	F1-Score (Severe)
Logistic Regression	0.66	0.66	0.95	0.78	0.60	0.13	0.22
XGBoost	0.65	0.67	0.91	0.77	0.53	0.17	0.26

RESULTS AFTER HYPERPARAMETER TUNING

Model	Accuracy	Precision (Mild)	Recall (Mild)	F1-Score (Mild)	Precision (Severe)	Recall (Severe)	F1-Score (Severe)
Logistic Regression	0.66	0.66	0.95	0.78	0.60	0.13	0.22
XGBoost	0.66	0.66	0.95	0.78	0.60	0.13	0.22

- We utilized Grid search CV for Hyperparameter tuning picked the best parameters getting the results as shown above.
- Both Logistic Regression and XGBoost perform similarly, with good performance on Mild collisions but significant difficulty with Severe collisions.
- The models could be improved for Severe collision prediction with further tuning, resampling techniques, or alternative modeling approaches.

CONCLUSION

Our project successfully utilized machine learning to analyze traffic accident data and predict severity based on environmental, temporal, and road-related factors.

- **Key Findings:** Most accidents occurred under clear weather, on Fridays, during broadside collisions, on dry road surfaces, and predominantly between 2 PM and 6 PM.
- **Model Performance:** Logistic Regression and XGBoost models showed consistent accuracy (66%), with better recall for mild cases than severe ones. Hypertuning enhanced performance slightly, but severe incident predictions remain a challenge.
- **Feature Insights:** Variables like distance traveled, number of injuries, and type of collision were identified as crucial factors influencing severity.

RECOMMENDATIONS

1. **Improve Severe Incident Prediction:** Explore ensemble methods or deep learning techniques to enhance recall for severe cases. Incorporate additional data points, such as driver behavior or road infrastructure details, to enrich the model.
2. **Application to Public Safety:** Use predictions to inform proactive measures like targeted safety campaigns during high-risk periods (e.g., Fridays, peak hours) and locations prone to broadside collisions. Equip emergency responders with severity predictions for optimized resource allocation.
3. **Expand Scope:** Validate the model across different cities to assess scalability and adaptability. Integrate these insights into real-time traffic management systems for immediate safety interventions.

CAUTION!

Remember to Stay alert, stay safe, and make every journey count. Your choices behind the wheel can save or take lives!