US Traffic Accidents Analysis and Risk Pattern Identification

Shivangi Gupta
Shriya Tarun
Stiles Clements
Vedant Mehrotra

Abstract:

This study examines traffic accident patterns in the United States to predict severity and identify spatial–temporal high-risk zones. Using a 500K sample from the US Accidents Dataset, the analysis applied K-Means clustering to locate hotspots and a Random Forest model to predict severity levels (1–4). Key predictors included hour of day, month, season, weather conditions, and road features such as traffic signals, junctions, and crossings. Results show accident frequency peaks, as well as, hotspots concentrated in metropolitan areas and major intersections. Findings provide actionable insights for policy, infrastructure improvements, and optimized emergency response.

Problem Significance:

Road traffic accidents remain a significant public safety and economic concern in the United States. In 2022, motor vehicle crashes caused nearly 44,000 fatalities and generated over $470 billion in total costs, including medical expenses and the valuation of lives lost (Centers for Disease Control and Prevention [CDC], 2024). In 2023, there were approximately 40,901 fatalities, with the economic cost—covering medical treatments, property damage, lost productivity, and legal expenditures—estimated at around $340 billion (Insurance Institute for Highway Safety [IIHS], 2024; National Highway Traffic Safety Administration [NHTSA], 2024).

The ability to accurately predict accident severity and identify high-risk locations and time periods holds substantial value for public agencies, policymakers, and urban planners. Such predictive capabilities can facilitate evidence-based interventions, optimize emergency resource allocation, and guide strategic infrastructure development aimed at reducing accident frequency and severity.

This project is designed to achieve two primary objectives:

- Severity Prediction – Develop a machine learning model to predict the severity level (1–4) of a traffic accident using temporal, environmental, geographic, and infrastructure-related features.
- Risk Zone Identification – Apply spatial clustering methods to detect accident hotspots
- Temporal Pattern Analysis - Identify periods of elevated risk

Dataset Description:

The dataset, sourced from Kaggle, originally has 7 million accident observations with 46 features. It covers accidents across 49 U.S. states from Feb 2016 to Mar 2023. It includes temporal, geographic, environmental, and road infrastructure features. The target variable is *Severity* (Levels 1–4, from least to most severe).

Exploratory Analysis:

A structured exploratory assessment and feature engineering process was undertaken to prepare the dataset for analytical modeling. After filtering for valid timestamps, the working dataset comprised approximately 451,837 accident records, with a primary focus on temporal attributes derived from *Start_Time* and key road-environment variables such as *Amenity* and *Junction*. Temporal features included hour of day, month, a binary weekday/weekend classification, and seasonal groupings (Winter, Spring, Summer, Fall). Additionally, boolean infrastructure indicators were created to flag the presence of *Traffic_Signal*, *Junction*, *Crossing*, *Railway*, and *Stop* signs, enabling a quantifiable view of infrastructural influence on accident occurrence and severity.

Feature Engineering and Preprocessing:

1. Using the “Start Time” of accidents, converting it to datetime and creating new columns for “Hour” of the day, “Day” of the Week, “Month”
2. Using the columns “Sunrise Sunset”, “Civil Twilight”, “Nautical Twilight”, and “Astronomical Twilight” to create an ordinal feature to capture “Light Level”
3. Converting Boolean Road Features into 0-1 values

Interesting Findings:

The analysis identified distinct temporal, spatial, and environmental patterns influencing traffic accidents. Moderate-severity accidents (Severity 2) dominate the dataset, comprising approximately 77–80% of all recorded cases, while high-severity incidents (Severity 3 and 4) occur less frequently but pose greater operational and safety implications. Seasonal trends show that Fall registers the highest overall accident volume, though severity proportions remain consistent across the year. Weekday traffic patterns present sharp commuter peaks, whereas weekends display a more balanced hourly distribution.

Environmental factors were found to significantly impact accident severity, with adverse weather particularly rainfall and reduced visibility correlating with elevated risk. These findings offer a data-driven basis for designing proactive safety measures, optimizing emergency response allocation, and informing evidence-based transportation policy.

1. Upon creating the feature that determines the Light Level during the time of the accident, it was peculiar to see that most accidents occurred when the Light Level was the highest. This might potentially be due to the fact that more people drive in the day than at night.
2. Weekdays had the highest count of accidents, peaking on Friday, before drastically dropping in numbers over the weekend. This might be related to the fact that people prefer staying in on weekends.
3. Spikes in accidents during the day, especially weekdays, were from 6-9am and 3-6pm. This happens to be the peak period as it coincides with school and office timings.
4. The counties of Los Angeles and Miami-Dade have the highest number of accidents amongst all other counties.
5. Created calendars to identify which days of the year were more prone to having accidents. For example, in the county of LA, 14th Feb which happens to be Valentine’s Day had the highest number of accidents.

Outcomes:

The Random Forest Classifier achieved the best balance of accuracy and recall, outperforming logistic regression, particularly for high-severity but less frequent cases. Key predictive factors included hour of day, season, weather conditions, and infrastructure elements such as traffic signals, junctions, and crossings. These findings directly support interventions like commuter-hour enforcement, weather-responsive traffic control, and intersection redesign.

Spatial analysis via K-Means Clustering revealed consistent hotspots in urban centers and along major corridors. It identified structured patterns, providing a comprehensive view of risk concentration. The density hotspots corroborated the earlier findings relating to the high accident counts in Los Angeles and Miami-Dade counties.

Practical Applications:

By aggregating accident data into a calendar format (Fig 5), it becomes easy to spot periods - such as specific months, weeks, or weekdays - when severe incidents peak. For instance, intense clusters of red indicate higher severity, guiding traffic authorities to deploy more patrols or implement preventive measures during those high-risk intervals. Medical aid teams can use these insights to strategically position ambulances and staff, anticipating surge periods for severe accidents. Similarly, fire safety resources can be optimized for quick response whenever the calendar reveals patterns of elevated risk. Ultimately, this visualization transforms raw data into a practical dashboard for proactive resource allocation, enabling agencies to match real-world deployment with predictable accident trends and ensuring a safer environment throughout the year.

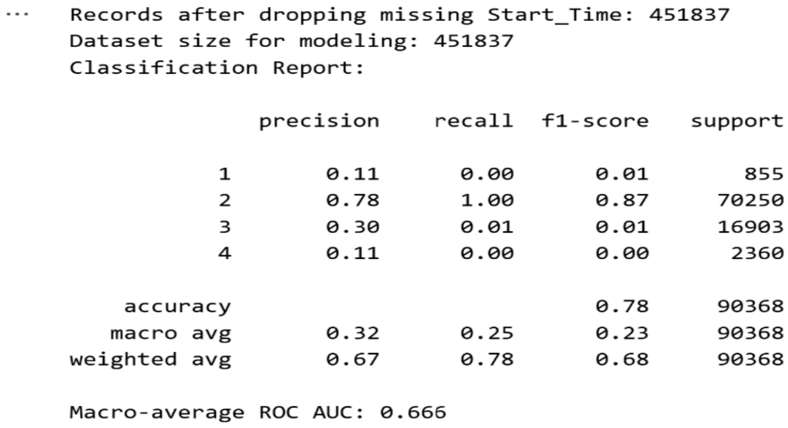
Conclusion:

This analysis demonstrates that combining predictive modeling with spatial clustering offers a robust framework for anticipating accident severity and identifying high-risk zones. Time of day, seasonal trends, weather conditions, and traffic control infrastructure emerged as the most influential predictors enabling targeted commuter-hour enforcement, weather-responsive traffic management, and intersection redesign.

Spatial analysis revealed consistent hotspots in urban centers and along major corridors. K-Means identified structured patterns, providing a comprehensive view of risk concentration. These insights support data-driven infrastructure upgrades, targeted policing, and optimized emergency response placement - shifting safety management from reactive measures to proactive, long-term planning.

Link to charts for spatial analysis:
https://spoofygoofy.xyz/

Random Forest Analysis



```
...   Records after dropping missing Start_Time: 451837
      Dataset size for modeling: 451837
      Classification Report:

                   precision    recall  f1-score   support

                 1       0.11      0.00      0.01       855
                 2       0.78      1.00      0.87     70250
                 3       0.30      0.01      0.01     16903
                 4       0.11      0.00      0.00      2360

          accuracy                           0.78     90368
         macro avg       0.32      0.25      0.23     90368
      weighted avg       0.67      0.78      0.68     90368

      Macro-average ROC AUC: 0.666
```

Figure 1: Random Forest Classification Report

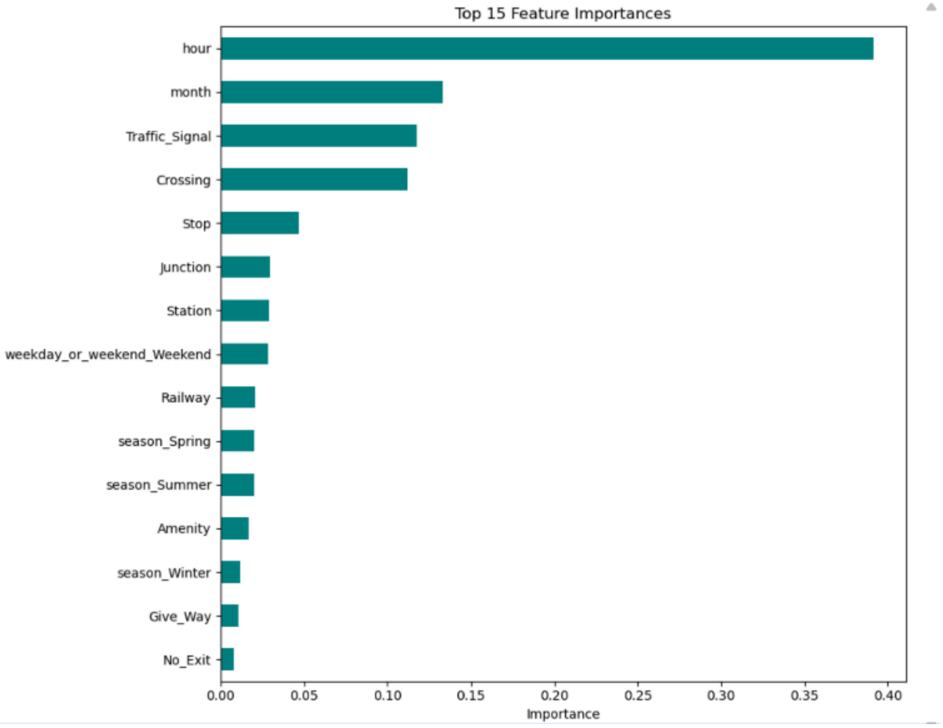


Figure 2: Random Forest Most Important Features

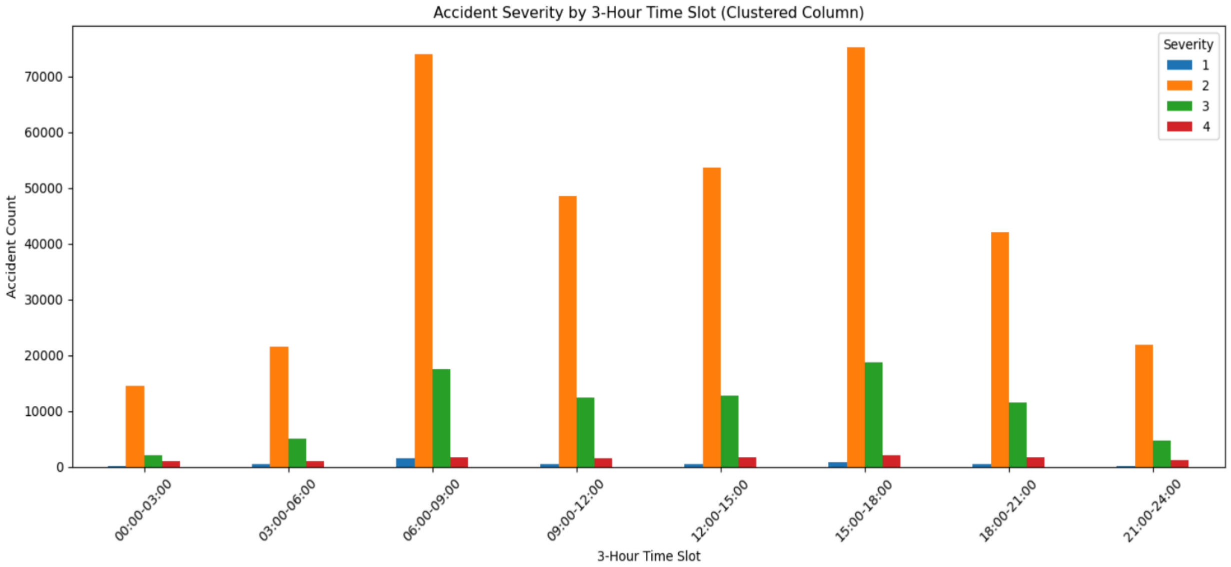


Figure 3: Accident Severity by 3-hour timeslot

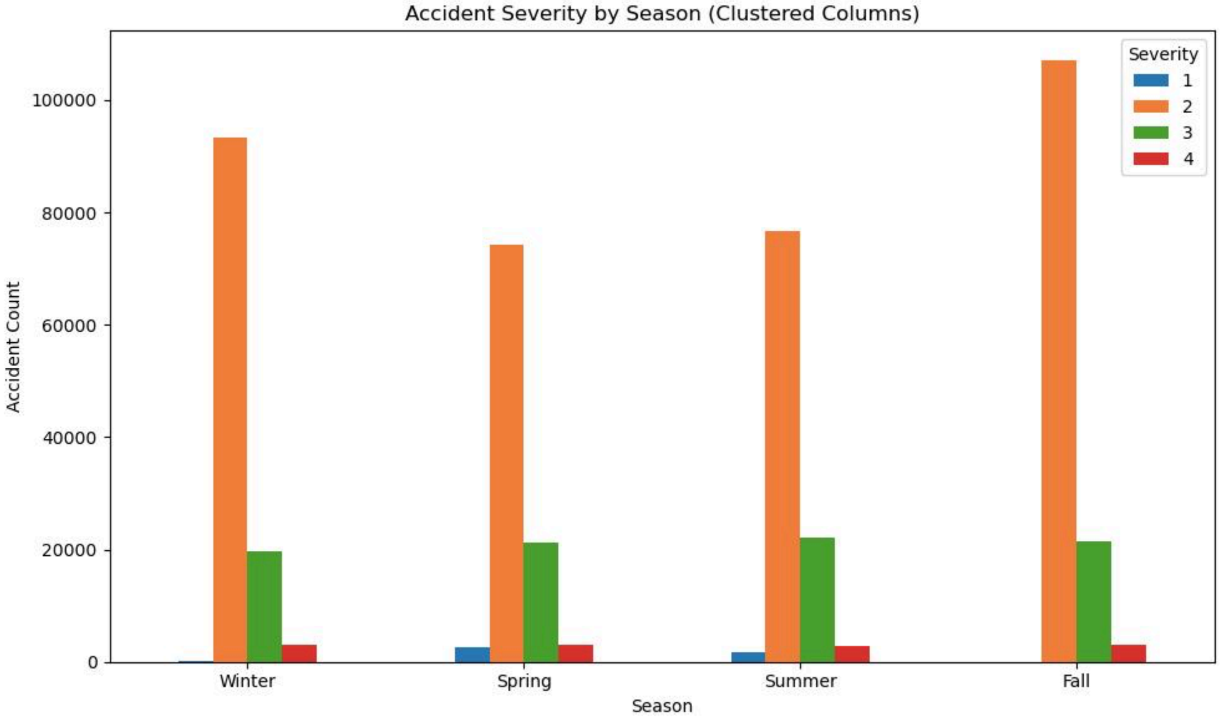


Figure 4: Accident Severity by Season

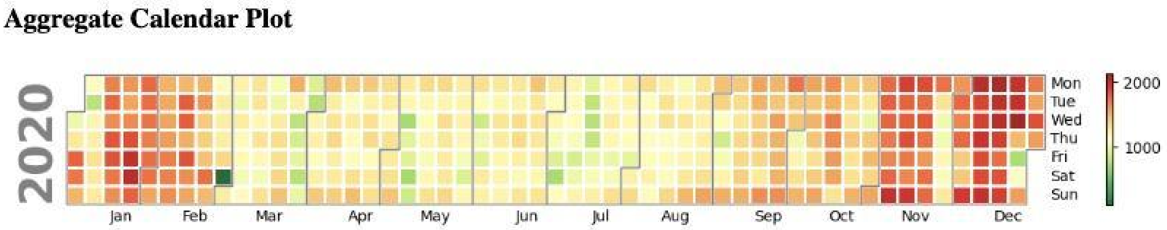


Figure 5: Accident severity across entire year

References:

1) Centers for Disease Control and Prevention. (2024). Transportation safety: About. https://www.cdc.gov/transportation-safety/about/index.html
2) Insurance Institute for Highway Safety. (2024). Fatality facts: Yearly snapshot. https://www.iihs.org/research-areas/fatality-statistics/detail/yearly-snapshot
3) Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, 2019.
4) Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.
5) National Highway Traffic Safety Administration. (2024). The economic and societal impact of motor vehicle crashes, 2019 (Report No. DOT HS 813 403). https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813403.pdf
6) Repairer Driven News. (2024, December 11). Report estimates fatal crashes cost U.S. $417 billion annually. https://www.repairerdrivennews.com/2024/12/11/report-estimates-fatal-crashes-cost-u-s-417-billion-annually