Analysis Summary

Shervan Shahparnia and Nathan Cohn - Group 1

### **Result 1: Correlation Between Weather Features and Severity Plots:**

### 

**Summary:** The feature correlation heatmap shows that **Severity** has **very weak correlations** with all other environmental features, ranging between **-0.05 and 0.04**. This indicates that accident severity is not strongly influenced by temperature, humidity, visibility, or precipitation on their own.  
 Notable relationships among weather features include:

* **Temperature(F) vs. Humidity(%):** -0.36  
   Moderate negative correlation -> warmer air reduces relative humidity.
* **Humidity(%) vs. Visibility(mi):** -0.40  
   Higher humidity tends to reduce visibility, likely due to fog or rain.
* **Visibility(mi) vs. Precipitation(in):** -0.12  
   Slight negative correlation -> precipitation somewhat reduces visibility.
* **Precipitation(in) vs. Severity:** 0.00  
   No meaningful correlation between precipitation and severity.

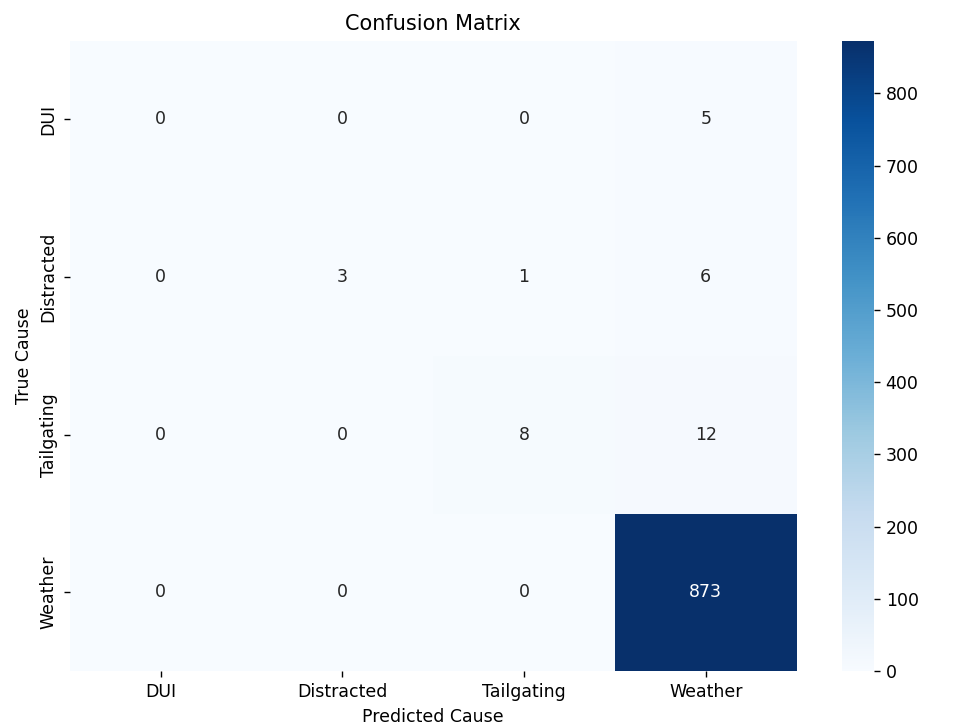
**Implications:** These results suggest that **no single weather variable serves as a strong predictor** of accident severity. Even though weather may influence driving conditions, it's likely that the **severity** of crashes depends more on external factors like **road design, driver behavior, and traffic volume**.

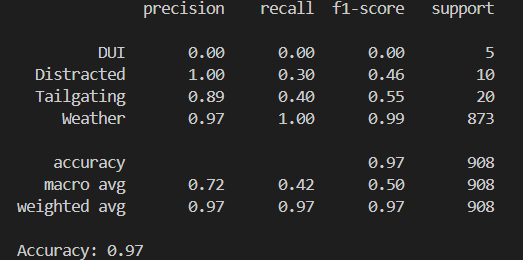
**Next Steps:** To improve our predictive modeling, we should explore variables such as:

* **Time of day** (e.g., rush hour, night driving)
* **Road conditions** (e.g., intersections, highways, curves)
* **Behavioral causes** (e.g., DUI, distracted driving)

Weather conditions may still contribute to crash frequency, but they alone don’t explain the level of severity. Future models may benefit from multivariable interactions or context-aware features.

### **Result 2: NLP-Based Classification of Accident Cause**

**Plots: **

****

**Summary:** We used basic keyword-based NLP techniques to label accident causes from the **Description** field (e.g., DUI, Weather, Distracted, Tailgating). These labels were then used to train a Random Forest Classifier, which achieved a surprisingly high accuracy of 97%.  
 However, the confusion matrix and classification report show that the model predominantly predicted one class: **Weather**. Nearly all misclassified instances were predicted as Weather, regardless of their true label.

* **Weather:** Precision = 0.97, Recall = 1.00, F1 = 0.99
* **All other classes** (DUI, Distracted, Tailgating): low recall (≤ 0.40) and in some cases, no predictions at all.

**Implications:** While the overall accuracy seems impressive, the imbalance in predictions reveals that our model is highly *biased toward the Weather label*. This is likely because the Weather class dominates the training set. The extremely low performance on other causes, especially **DUI (0.00 F1-score)**, shows that the current labeling method isn't capturing enough meaningful variation in the **Description** field.

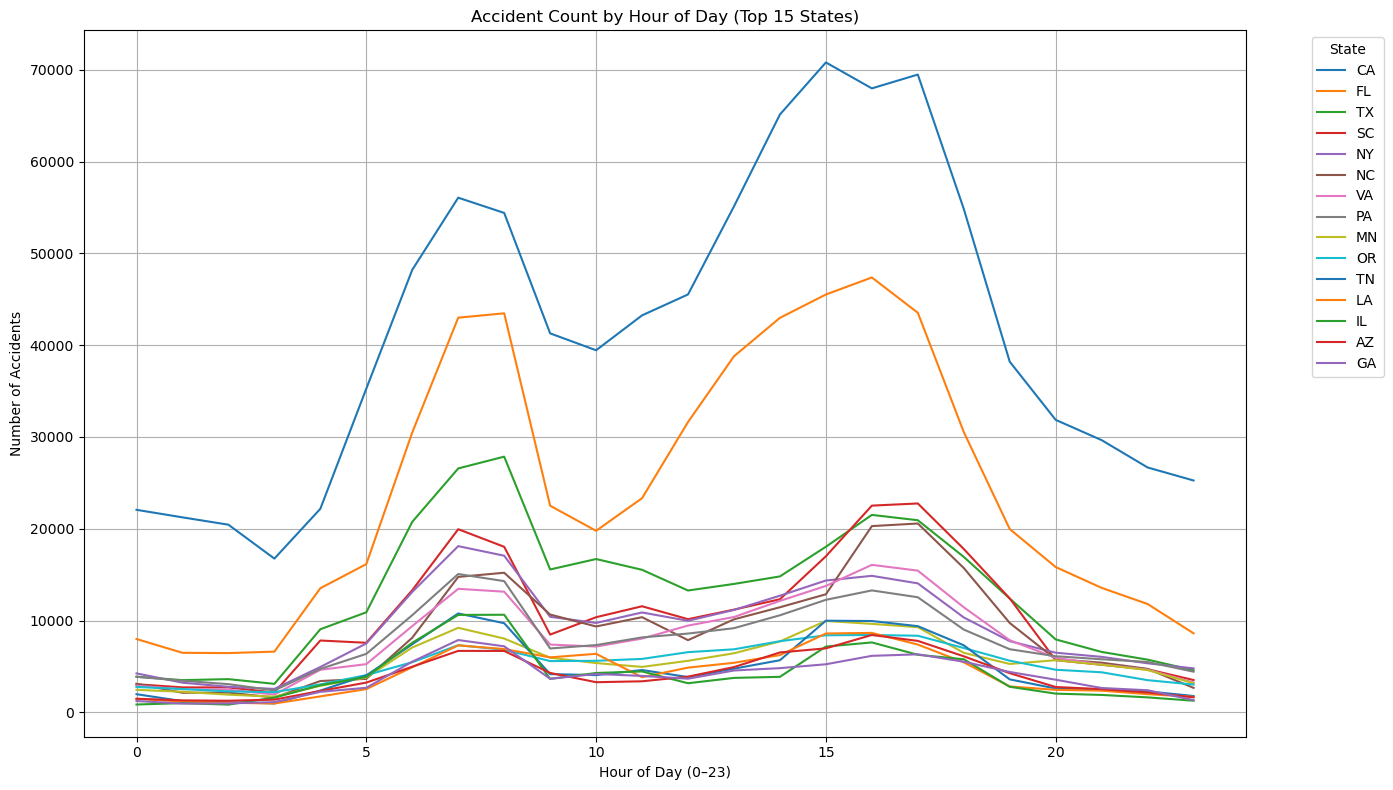
This suggests that:

* Our current keyword labeling method lacks precision and overlooks many relevant descriptions
* The model relies heavily on how often each label appears rather than truly learning meaningful patterns

**Next Steps:**

* Refine the labeling logic by incorporating NLP pipelines (e.g., named entity recognition, phrase matching, or transformer models like BERT).
* Balance the dataset by upsampling underrepresented causes or using SMOTE.
* Introduce multi-label classification or allow overlapping causes for richer context.
* Reevaluate the current preprocessing to ensure **Description** fields are fully captured and cleaned.

### **Result 3: Hourly Accident Trends Across Top 15 States**

**Plot: **

**Summary:** This line graph shows how accident counts vary by hour across the top 15 states with the highest total accident volume. All states demonstrate a clear **bimodal pattern**, with spikes during the **morning (7–9 AM)** and **evening (4–6 PM)** commonly associated with **commuting hours**.  
 California (CA), Florida (FL), and Texas (TX) have the highest volumes overall. California stands out with **pronounced spikes** during both rush hours, while other states like Arizona and Georgia have more evenly distributed patterns throughout the day. Although these states have the most drivers on the road, it’s also likely that the data could be a bit skewed because they also have the most entries within the dataset itself (which also lines up with our previous findings in our EDA where we showed the entries by state).

**Implications:** These patterns reinforce the idea that traffic volume and commuting behavior are major contributors to accident frequency. The consistency of rush hour peaks across states suggests that time-of-day is a stronger factor in predicting accident occurrence than weather alone (as seen in Result 1).  
 States with flatter curves may have more consistent traffic throughout the day or different urban planning layouts (e.g., less centralized commuting).

**Next Steps:**

* Incorporate 'Hour of Day' into future predictive models to better anticipate high-risk periods.
* Consider regional differences in traffic laws, road conditions, or city design that might affect accident timing.  
  Explore whether certain causes (like DUI or distracted driving) spike during specific hours, especially outside of typical commute times.