# Analysis and Visualizations Plan -

# Shervan Shahparnia and Nathan Cohn - Group 1

* For the analysis methods

| **Method Name** | **Input to the Algorithm** | **Output of the Algorithm** | **Goal / Hypothesis** | **Why This Method? / Relevance to Dataset** |
| --- | --- | --- | --- | --- |
| Correlation Matrix | Continuous variables: temperature, humidity, visibility, wind speed, severity | Correlation coefficients matrix | Investigate linear relationships (e.g., whether low visibility correlates with higher severity) | Identifies multicollinearity and relationships between predictors, informing feature engineering |
| Random Forest Classifier | All relevant features: time, weather, road layout, POIs | Predicted accident likelihood or severity class, feature importances | Predict accident likelihood or severity; assess which features most influence outcomes | Handles non-linear relationships, works well with tabular data, robust to missing values and imbalances |
| Gradient Boosted Trees (XGBoost) | Same inputs as Random Forest | Same outputs, typically higher accuracy | Improve predictive performance with boosting; better feature weighting and control of overfitting | Highly effective in structured datasets like this; often outperforms Random Forests on tabular data |
| Geospatial Clustering (DBSCAN) | Latitude, longitude, severity | Clusters of high-density accident zones (hotspots) | Detect natural clusters of accidents (e.g., busy intersections) without assuming fixed number of clusters | Unlike k-means, DBSCAN handles noise and irregular density patterns, better for spatial anomaly detection |
| Time Series Decomposition | Daily or weekly accident counts over time | Seasonal, trend, residual components | Determine whether accidents follow trends or seasonal cycles (e.g., holidays, weather) | Helps understand temporal cycles and anomalies (e.g., COVID-era traffic changes) |
| PCA (for Feature Reduction) | Numerical features: weather, environment, traffic density | Reduced-dimension principal components | Reduce dimensionality, eliminate redundancy, and visualize high-dimensional data | Helps improve model efficiency and visualize complex feature interactions |

* For the visualization plan

| **Visualization Type** | **X-Axis** | **Y-Axis / Y2-Axis (if applicable)** | **Purpose / Insight Aimed** |
| --- | --- | --- | --- |
| Line Plot - Temporal Trends | Hour of day / Day of week / Month | Number of accidents | Identify daily, weekly, and seasonal accident patterns (e.g., rush hours, holidays, weather shifts) |
| Heatmap - Weather vs. Severity | Weather condition categories | Average severity | Reveal which weather types (e.g., fog, rain) correlate with more severe accidents |
| Interactive Map - Clusters | Latitude & Longitude (on map view) | Heat overlay = accident density | Visualize spatial hotspots, especially near intersections or complex road networks |
| Scatter Plot - Visibility vs. Severity | Visibility (miles) | Severity | Visualize whether reduced visibility leads to more severe accidents |
| Feature Importance Plot - Model | Feature name (from Random Forest/XGBoost) | Importance score | Show which features contribute most to predicting severity or accident risk |
| Radar Chart - State Safety Profiles | Safety factors (e.g., avg severity, weather events, visibility) | State (1 radar per state) | Compare multiple safety metrics across regions to identify the most and least safe states |
| Gantt Chart - Accident Duration Timeline | Start time | Duration bars | Explore accident duration across a timeline — useful for traffic flow analysis and resource planning |