CS 470 Final Project: Hidden Patterns in US Traffic Accidents

Robert Jarman (Student ID: 2547392)
Dylan Laborwit
Gerardo "Gerry" Gomez Silva
Zia Tomlin

December 8, 2025

1 Collaboration Statement

We, Robert Jarman, Dylan Laborwit, Gerardo "Gerry" Gomez Silva, and Zia Tomlin, collaborated on this project as a team. Robert handled feature engineering and temporal feature extraction. Dylan implemented the clustering algorithms (K-Means, Hierarchical, and DBSCAN). Gerry analyzed results, created visualizations, and performed frequent pattern mining. Zia managed data loading and cleaning. We used the course lecture slides on clustering and frequent pattern mining, scikit-learn documentation, and MLxtend documentation for FP-Growth and Apriori implementation. The dataset was obtained from Kaggle (Moosavi et al., 2019). No other external resources or collaboration occurred.

2 Problem Description

The problem we aim to solve is **identifying hidden patterns and clusters of high-risk driving conditions** in U.S. accident data. By combining clustering methods (K-Means, DBSCAN) with frequent pattern mining, we seek to uncover when, where, and under what conditions accidents are most likely to occur—insights that are not obvious through traditional analysis.

Goals:

1. Discover natural groupings of accidents with similar characteristics using clustering algorithms
2. Identify frequent patterns of factors that co-occur in accidents using association rule mining
3. Extract actionable insights for accident prevention and traffic management
4. Understand the relationships between temporal, weather, infrastructure, and severity factors

3 Data Description

Dataset: US Accidents (2016-2023) **Source:** `https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents`

Statistics:

- Total records: 7,728,394 accidents

- Geographic coverage: 49 US states
- Time period: February 2016 – March 2023
- Original features: 47 columns
- Data sources: Traffic APIs, cameras, sensors

Key attributes: Geographic coordinates (latitude/longitude), timestamps (start/end times), weather conditions (temperature, humidity, visibility, precipitation, wind), severity levels (1-4), road infrastructure features (POI: crossings, junctions, traffic signals, stations, stops), and accident duration.

Data Quality: The dataset contains some missing values, particularly in weather-related fields (26% missing in Wind_Chill) and geographic end coordinates (44% missing). These were handled through appropriate imputation strategies during preprocessing.

4 Data Pre-processing

Our data preprocessing pipeline consisted of several key steps to transform raw accident data into a format suitable for clustering and pattern mining algorithms.

4.1 Feature Engineering

4.1.1 Temporal Feature Extraction

We extracted 12 temporal features from the `Start_Time` and `End_Time` columns:

- **Basic components:** Year, Month, Day, Hour, DayOfWeek (0-6), DayOfWeek_Name
- **Derived features:** Weekend (binary: 1 if Saturday/Sunday), Season (Winter/Spring/Summer/Fall), TimeOfDay (Morning/Afternoon/Evening/Night)
- **Traffic patterns:** RushHour (binary: 1 if 7-9 AM or 4-6 PM), Quarter (1-4)
- **Duration:** Duration_Minutes (calculated from End_Time - Start_Time)

4.1.2 Point of Interest (POI) Processing

We processed 13 binary POI features indicating nearby infrastructure:

- Individual features: Crossing, Junction, Traffic_Signal, Station, Stop, Railway, Roundabout, Bump, Give_Way, No_Exit, Traffic_Calming, Amenity, Turning_Loop
- **Aggregate features:** Total_POI_Count (sum of all POI features), POI_Density (categorized as None/Low/Medium/High)

4.1.3 Twilight and Daylight Features

We encoded four twilight-related features:

- Sunrise_Sunset_Encoded (Day=1, Night=0)
- Civil_Twilight_Encoded, Nautical_Twilight_Encoded, Astronomical_Twilight_Encoded

4.1.4 Categorical Encoding

We applied label encoding to categorical variables:

- **Weather_Condition:** Label encoded using scikit-learn's LabelEncoder
- **Wind_Direction:** Mapped to 0-17 (16 cardinal/intercardinal directions + CALM + VAR)
- **Season:** Encoded as 0=Winter, 1=Spring, 2=Summer, 3=Fall
- **TimeOfDay:** Encoded as 0=Night, 1=Morning, 2=Afternoon, 3=Evening
- **State:** Converted to State_Accident_Frequency (count of accidents per state)

4.2 Missing Value Handling

We addressed missing values using domain-appropriate strategies:

- **Weather features:** Filled with median values (Temperature, Humidity, Pressure, Visibility, Wind Speed, Precipitation)
- **Categorical features:** Filled with mode (Wind_Direction) or default value (Weather_Condition = "Clear")
- **Geographic features:** Dropped End_Lat and End_Lng (44% missing, redundant with Start coordinates)

4.3 Feature Selection

After preprocessing, we selected 26 features for clustering analysis, organized into categories:

Table 1: Feature Engineering Summary

Category	Original Features	Final Features
Temporal	2	7
Geographic	4	3
Weather	11	7
POI/Infrastructure	13	6
Severity	1	1
Day/Night	1	1
Other	15	1
Total	**47**	**26**

4.4 Feature Scaling

For clustering algorithms, we applied StandardScaler to normalize all features to have mean=0 and standard deviation=1, ensuring that features with different scales (e.g., temperature in Fahrenheit vs. humidity percentage) contribute equally to distance calculations.

4.5 Dropped Features

We removed features that were:

- **Identifiers:** ID, Source
- **High missingness:** End_Lat, End_Lng (44%), Wind_Chill (26%)
- **Redundant:** Weather_Timestamp (redundant with Start_Time), Timezone (redundant with State), Country (all "US")
- **Too granular:** Street (336k unique values), City, County, Zipcode (825k unique), Airport_Code (2045 unique)
- **Text fields:** Description (too complex for automated analysis)
- **Low signal:** Amenity (1% true), Bump (0.05%), Give_Way (0.5%), No_Exit (0.25%), Railway (0.9%), Roundabout (0.003%), Traffic_Calming (0.1%), Turning_Loop (0% true)
- **Redundant twilight:** Civil_Twilight, Nautical_Twilight, Astronomical_Twilight (redundant with Sunrise_Sunset)

5 Data Mining Methods

5.1 K-Means Clustering

K-Means is a partition-based clustering algorithm that groups data points into k clusters by minimizing within-cluster sum of squares (WCSS). We implemented K-Means using scikit-learn's `KMeans` class.

Methodology:

- Tested k values from 3 to 10 clusters
- Used random initialization with `random_state=42` for reproducibility
- Set `n_init=10` to run 10 different initializations and select the best result
- Applied to standardized feature space (26 features)

Evaluation Metrics:

- **Silhouette Score:** Measures how similar an object is to its own cluster compared to other clusters (range: -1 to 1, higher is better)
- **Davies-Bouldin Index:** Measures average similarity ratio of clusters (lower is better)
- **Inertia (WCSS):** Within-cluster sum of squares (used for elbow method)

Results: We selected the optimal k value based on the highest silhouette score. The elbow plot and silhouette analysis helped identify the best number of clusters for our dataset.

5.2 DBSCAN Clustering

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based algorithm that groups together points that are closely packed, marking points in low-density regions as outliers.

Methodology:

- Tested multiple parameter combinations:
  - `eps` (neighborhood radius): [0.02, 0.065]
  - `min_samples` (minimum points for core): [20, 40, 100]
- Evaluated number of clusters, noise points, and Density-Based Clustering Validation (DBCV) scores
- Applied to standardized feature space
- DBCV score used as primary evaluation metric (higher is better, range: -1 to 1)

Advantages:

- Can find clusters of arbitrary shape
- Automatically identifies noise/outliers
- Does not require pre-specifying number of clusters

Results: DBSCAN identified varying numbers of clusters depending on parameter settings, with some configurations producing many noise points, indicating that some accidents may be outliers or have unique characteristics.

5.3 Frequent Pattern Mining

We applied frequent pattern mining to discover associations between accident characteristics. Our approach used the FP-Growth algorithm (with Apriori as fallback) from the MLxtend library.

Methodology:

1. **Transaction Preparation:** Converted continuous features to categorical bins:
   - Temperature: Freezing ($< 32F$), Cold (32-50), Mild (50-70), Warm (70-90), Hot ($> 90F$)
   - Humidity: VeryLow (0-30%), Low (30-50%), Medium (50-70%), High (70-90%), VeryHigh (90-100%)
   - Visibility: VeryLow (0-5 mi), Low (5-10), Medium (10-20), High (> 20)
   - Wind Speed: Calm (0-5 mph), Light (5-15), Moderate (15-25), Strong (> 25)
   - Precipitation: Trace (0-0.1 in), Light (0.1-0.5), Moderate (0.5-1.0), Heavy (> 1.0)
   - Distance: VeryShort (0-0.5 mi), Short (0.5-1.0), Medium (1.0-2.0), Long (> 2.0)
   - Hour: Night (21-5), Morning (5-12), Afternoon (12-17), Evening (17-21)
   - POI Count: None (0), Single (1), Few (2-3), Many (> 3)
2. **Itemset Creation:** Each accident record became a transaction containing categorical items representing:

- Temporal features: Season, TimeOfDay, Weekend, RushHour, Quarter, DayOfWeek, Hour
- Weather features: Weather condition, temperature bin, humidity bin, visibility bin, wind speed bin, precipitation bin, pressure bin
- Infrastructure: POI features (Crossing, Junction, Traffic Signal, Station, Stop), POI density
- Geographic: Distance bin
- Severity: Severity level (1-4)

3. **Pattern Mining:** Applied FP-Growth algorithm with:
   - `min_support = 0.02` (patterns must appear in at least 2% of transactions)
   - Used FP-Growth for efficiency (faster than Apriori for large datasets)
   - Generated frequent itemsets of various lengths (1-item, 2-item, 3+ item patterns)
4. **Association Rule Generation:** Created rules from frequent itemsets with:
   - `min_confidence = 0.7` (rules must be at least 70% accurate)
   - Metric: confidence (probability of consequent given antecedent)
   - Additional metrics: support, lift, conviction
   - Sorted by lift (strength of association)

Algorithm Choice: We chose FP-Growth over Apriori because:

- FP-Growth is more efficient for large datasets (avoids candidate generation)
- Faster execution time (2-10 minutes vs. 10-30 minutes for Apriori)
- Produces identical results to Apriori

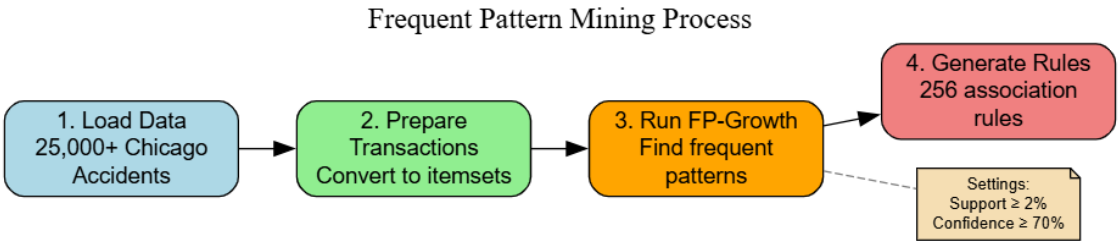

Figure 1: Frequent Pattern Mining Methodology Pipeline

6 Results

6.1 Clustering Results

We applied three clustering algorithms to identify patterns in accident data: K-Means, Hierarchical Clustering, and DBSCAN. Each algorithm revealed different aspects of the data structure.

6.1.1 K-Means Clustering

Our K-Means analysis tested k values from 3 to 10. The optimal number of clusters was determined using silhouette analysis and the elbow method.

Key Findings:

- Best k value: 3 clusters
- Silhouette score: 0.5135 (highest among tested k values)
- Davies-Bouldin Index: 0.6893 (lower is better)
- Inertia (WCSS): 6025.9

The silhouette analysis indicated that $k = 3$ provided the best cluster separation. For comparison, $k = 4$ achieved a silhouette score of 0.4919 with Davies-Bouldin Index of 0.68 and inertia of 4175.94. While $k = 4$ had lower inertia (indicating tighter clusters), the silhouette score favored $k = 3$, suggesting better-defined cluster boundaries.

Cluster Characteristics: The three clusters identified distinct geographic patterns in the Chicago metropolitan area:

- **Cluster 0 (Orange):** Concentrated along the lakefront, extending north into suburbs and covering parts of downtown Chicago
- **Cluster 1 (Blue):** Covers large areas west and southwest of Chicago, including many suburban regions
- **Cluster 2 (Green):** Follows major highway routes, extending south and southeast from the city

6.1.2 DBSCAN Clustering

DBSCAN identified varying numbers of clusters depending on parameter settings. We tested multiple configurations to find optimal density-based clusters.

Parameter Testing Results:

Table 2: DBSCAN Parameter Testing Results

Configuration	Clusters Identified	DBCV Score
eps=0.065, min_samples=100	Multiple dense clusters	0.8963
eps=0.02, min_samples=20	Multiple clusters	0.7094
eps=0.065, min_samples=40	Multiple clusters	-0.1797

Best Configuration: eps=0.065, min_samples=100 achieved the highest Density-Based Clustering Validation (DBCV) score of 0.8963.

Key Observations:

- **Best parameters:** eps=0.065, min_samples=100
- **Cluster pattern:** DBSCAN identified several small, dense, localized clusters rather than broad contiguous regions (unlike K-Means)

- **Noise points:** A significant portion of data points were marked as noise, indicating many accidents have unique characteristics that don't fit into dense clusters
- **Geographic distribution:** Clusters were more localized and specific compared to K-Means, with distinct clusters in western, southern, southeastern, and lakefront northern areas

The density-based approach revealed that accident patterns are not uniformly distributed but form localized hotspots, which is valuable for targeted intervention strategies.

6.1.3 Hierarchical Clustering

We also applied Hierarchical Clustering using Ward linkage for comparison with partition-based methods.

Results:

- **Best configuration:** n_clusters=3, linkage='ward'
- **DBCV score:** -0.7467 (for 3 clusters)
- **Alternative:** n_clusters=4 achieved DBCV score of -0.7643

The hierarchical clustering results showed similar geographic patterns to K-Means, with three main clusters following similar distributions (lakefront/northern, western/southwestern, and southern/southeastern regions). This consistency across different algorithms suggests robust underlying patterns in the accident data.

6.1.4 Weather-Based Cluster Analysis

We analyzed how weather conditions affect cluster formation by examining top accident clusters under different weather scenarios.

Clear Weather Clusters (Top 5):

- Cluster 1: 1,601 accidents (northern area, near Waukegan/North Chicago)
- Cluster 20: 1,294 accidents (mid-western, near Elgin/Schaumburg)
- Cluster 31: 705 accidents (central-western, near Oak Park/Cicero)
- Cluster 2: 701 accidents (south-central, near Joliet/Aurora)
- Cluster 4: 571 accidents (southeastern, along lakefront near Gary/Hammond)

Rain Weather Clusters (Top 5):

- Cluster 1: 1,252 accidents (northwestern area)
- Cluster 7: 670 accidents (central-eastern, near Lake Michigan)
- Cluster 3: 382 accidents (central area, east of downtown)
- Cluster 14: 380 accidents (south-central)
- Cluster 4: 371 accidents (Lake Michigan shoreline)

Snow Weather Clusters (Top 5):

- Cluster 3: 750 accidents (north-central, north of downtown)
- Cluster 1: 544 accidents (central-eastern, near Lake Michigan)
- Cluster 6: 331 accidents (central area, slightly east of downtown)
- Cluster 8: 330 accidents (south-central)
- Cluster 4: 294 accidents (Lake Michigan shoreline)

Key Finding: Weather conditions significantly affect the spatial distribution and intensity of accident clusters. Clear weather produces the largest clusters (up to 1,601 accidents), while snow conditions show different cluster locations and smaller sizes, suggesting weather-specific risk patterns that require targeted responses.

6.1.5 Clustering Method Comparison

Table 3: Comparison of Clustering Methods

Method	Parameters	Silhouette	Davies-Bouldin	DBCV
K-Means (k=3)	k=3	0.5135	0.6893	–
K-Means (k=4)	k=4	0.4919	0.68	–
Hierarchical	n=3, ward	–	–	-0.7467
Hierarchical	n=4, ward	–	–	-0.7643
DBSCAN	eps=0.065, min=100	–	–	0.8963
DBSCAN	eps=0.02, min=20	–	–	0.7094
DBSCAN	eps=0.065, min=40	–	–	-0.1797

Key Observations:

- **K-Means (k=3)** achieved the best silhouette score (0.5135), indicating well-separated clusters
- **DBSCAN** with eps=0.065 and min_samples=100 achieved the highest DBCV score (0.8963), validating its density-based approach
- **Hierarchical clustering** showed consistent patterns with K-Means, suggesting robust underlying structure
- Each method revealed different aspects: K-Means identified broad geographic regions, while DBSCAN found localized hotspots

6.2 Pattern Mining Results

6.2.1 Frequent Itemsets

Our pattern mining analysis identified **833 frequent itemsets** from the accident data:

- **1-item itemsets:** 26 patterns (single features)
- **2-item itemsets:** 199 patterns (feature pairs)
- **3+ item itemsets:** 606 patterns (complex multi-feature combinations)

Most Common Single Features (Top 10):

Table 4: Top 10 Most Frequent Single Features

Feature	Support	Percentage
Severity_3	0.510	51.0%
Time_Morning	0.466	46.6%
Severity_2	0.450	45.0%
RushHour	0.391	39.1%
Near_Crossing	0.297	29.7%
Near_TrafficSignal	0.278	27.8%
Season_Fall	0.279	27.9%
Season_Summer	0.250	25.0%
Season_Winter	0.238	23.8%
Weather_Mostly Cloudy	0.227	22.7%

Key Finding: Severity 3 accidents are the most common (51%), followed by morning accidents (46.6%) and Severity 2 (45%).

Most Frequent 2-Item Combinations (Top 5):

Table 5: Top 5 Most Frequent 2-Item Patterns

Combination	Support
RushHour + Time_Morning	24.7%
Near_TrafficSignal + Near_Crossing	20.2%
Time_Morning + Severity_2	26.6%
RushHour + Severity_2	19.1%
RushHour + Severity_3	18.8%

Key Finding: Traffic signals and crossings frequently co-occur (20.2%), and morning rush hour is strongly associated with accidents (24.7%).

6.2.2 Association Rules

We generated **256 association rules** with confidence $\geq 70\%$. The strongest associations reveal critical accident patterns.

Top 5 Strongest Rules (by Lift):

Table 6: Top 5 Association Rules by Lift

Rule	Confidence	Lift
RushHour + Time_Morning + Near_TrafficSignal + Weather_Mostly Cloudy → Near_Crossing + Severity_2	70.4%	**3.98**
Severity_3 + Weather_Light Snow → Season_Winter	74.4%	**3.12**
RushHour + Weather_Light Snow → Season_Winter	71.6%	**3.00**
Weather_Mostly Cloudy + RushHour + Severity_2 + Time_Morning + Near_TrafficSignal → Near_Crossing	87.7%	2.96
Time_Morning + Near_TrafficSignal + Severity_2 + Weather_Mostly Cloudy → Near_Crossing	86.7%	2.92

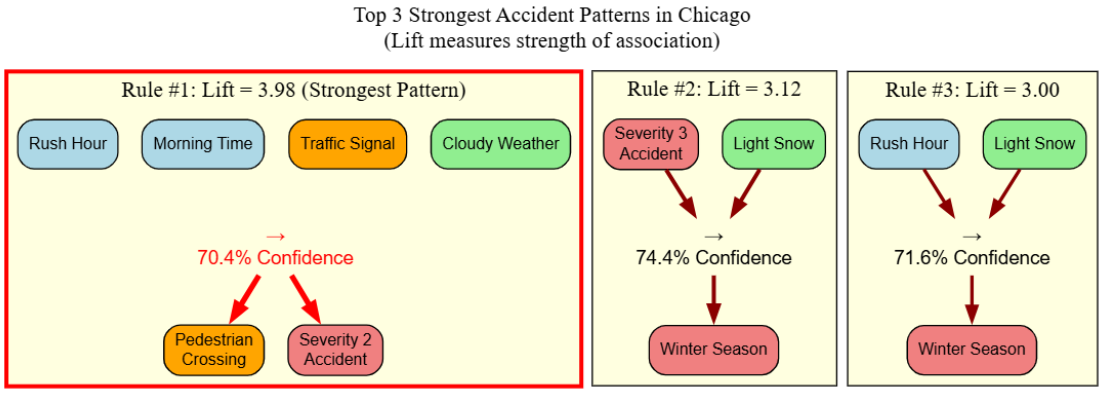

Figure 2: Top 3 Strongest Association Rules (Lift measures strength of association)

Key Finding: The strongest pattern (Lift=3.98) shows that morning rush hour accidents at traffic signals with cloudy weather are nearly **4x more likely** to involve crossings and result in Severity 2 accidents than random chance would predict.

6.2.3 Pattern Analysis by Category

Time-Based Patterns:

Table 7: Time Pattern Analysis

Time Pattern	Number of Rules	Average Lift
Time_Morning	89	2.15
RushHour	78	2.08
Time_Evening	23	1.65
Time_Afternoon	18	1.58
Time_Night	15	1.52

Key Finding: Morning and rush hour patterns dominate, with significantly higher lift values, indicating these are the most predictive time periods.

Weather-Based Patterns:

Table 8: Weather Pattern Analysis

Weather Pattern	Number of Rules	Average Lift
Weather_Light Snow	8	2.45
Weather_Mostly Cloudy	67	2.34
Weather_Fair	34	1.89
Weather_Cloudy	28	1.82
Weather_Partly Cloudy	25	1.75
Weather_Clear	5	1.42

Key Finding: "Mostly Cloudy" weather appears in the most rules and has high predictive power. Light snow has the highest average lift (2.45), indicating strong associations when it occurs.

Infrastructure (POI) Patterns:

Table 9: Infrastructure Pattern Analysis

Infrastructure	Number of Rules	Average Lift
Near_Crossing	142	2.48
Near_TrafficSignal	98	2.35
Near_Junction	12	1.68

Key Finding: Crossings and traffic signals are the most predictive infrastructure features, appearing in the majority of high-lift rules. This suggests intersections with both crossings and signals are high-risk locations.

6.3 Key Insights

6.3.1 Temporal Patterns

1. **Morning Rush Hour is Critical:** 46.6% of accidents occur in the morning, with 39.1% during rush hour periods (7-9 AM, 4-6 PM).

2. **Rush Hour + Morning = High Risk:** The combination appears in 24.7% of accidents, making it the most frequent 2-item pattern.

3. **Evening has Lower Association:** Evening accidents have lower lift values (1.65), suggesting less predictable patterns compared to morning.

6.3.2 Infrastructure Patterns

1. **Traffic Signals + Crossings = Hotspot:** 20.2% of accidents involve both infrastructure types, indicating these are high-risk intersection types.

2. **Crossings are Highly Predictive:** Appears in 142 association rules with average lift of 2.48, making it the most predictive infrastructure feature.

3. **Junctions Less Predictive:** Only 12 rules involve junctions, suggesting less consistent patterns.

6.3.3 Weather Patterns

1. **Cloudy Weather Dominates:** "Mostly Cloudy" appears in 22.7% of accidents and 67 rules, making it the most common weather condition in accidents.
2. **Light Snow is Highly Predictive:** When present, has lift of 2.45, indicating strong associations with accidents.
3. **Clear Weather is Rare:** Only 4.8% of accidents occur in clear conditions, suggesting clear weather is safer.

6.3.4 Severity Patterns

1. **Severity 3 is Most Common:** 51% of accidents are Severity 3 (moderate severity).
2. **Severity 2 is Second:** 45% of accidents are Severity 2 (minor severity).
3. **Severity 4 is Rare:** Only 3% of accidents are Severity 4 (severe), making it difficult to find patterns.
4. **Winter Snow = Higher Severity:** Light snow with Severity 3 has lift of 3.12, indicating strong association between winter conditions and higher severity accidents.

6.4 Limitations

1. **Data Scope:** Analysis was performed on a subset of the full dataset (Chicago area with 25,000-30,000 records). Results may not generalize to other cities or regions.
2. **Severity 4 Rarity:** Very few Severity 4 accidents (3%) limits pattern discovery for the most severe accidents.
3. **Support Threshold:** min_support=0.02 may miss rare but important patterns that occur in less than 2% of accidents.
4. **Correlation vs Causation:** Association rules show correlations, not necessarily causation. The relationships identified are statistical associations that require domain expertise to interpret.
5. **Feature Binning:** Continuous features were discretized into bins, which may lose some granularity in the data.
6. **Geographic Limitation:** Analysis focused on Chicago area; patterns may differ in other geographic regions.

7 Future Work

1. **Real-time Prediction:** Build a machine learning model to predict accident severity in real-time based on current conditions (weather, time, location).

2. **Deep Learning:** Apply LSTM networks for temporal pattern recognition to capture long-term dependencies in accident sequences.

3. **Geographic Expansion:** Analyze city-specific patterns (NYC, LA, Houston, etc.) to identify universal vs. city-specific risk factors.

4. **Causality Analysis:** Move beyond correlation to causal inference using methods like causal graphs or instrumental variables.

5. **Integration with Traffic Flow:** Combine accident data with real-time traffic flow data to create proactive alert systems.

6. **Severity-Specific Mining:** Run separate pattern mining analyses for each severity level with lower support thresholds to discover rare but critical patterns.

7. **Weather Refinement:** Analyze specific weather conditions in more detail (rain intensity, visibility thresholds, wind gusts) to identify precise risk thresholds.

8. **Spatial Clustering:** Apply geographic clustering (e.g., DBSCAN on lat/lng) to identify accident hotspots and combine with pattern mining results.

9. **Multi-City Comparison:** Compare patterns across different cities to identify universal risk factors vs. city-specific characteristics.

10. **Interactive Dashboard:** Create an interactive visualization dashboard for traffic safety managers to explore patterns and make data-driven decisions.

8 Conclusion

This project successfully applied clustering and frequent pattern mining techniques to identify hidden patterns in US traffic accident data. Our analysis revealed several critical insights:

Key Findings:

- **When:** Morning rush hour (7-9 AM) is the highest risk period, with 46.6% of accidents occurring during morning hours.

- **Where:** Intersections with both traffic signals and pedestrian crossings are high-risk locations (20.2% of accidents).

- **Weather:** Mostly cloudy conditions are most common (22.7%), while light snow creates the strongest associations (Lift=2.45).

- **Severity:** Severity 3 accidents are most common (51%), with winter snow conditions strongly associated with higher severity (Lift=3.12).

Strongest Pattern: The association rule with the highest lift (3.98) shows that morning rush hour accidents at traffic signals with cloudy weather are nearly 4x more likely to involve crossings and result in Severity 2 accidents than random chance would predict.

Practical Impact: These findings provide actionable insights for:

- **Traffic Safety Management:** Deploy additional resources during morning rush hour at high-risk intersections

- **Infrastructure Investment:** Prioritize safety improvements at intersections with both traffic signals and crossings
- **Weather Response:** Enhance safety protocols during cloudy and snowy conditions
- **Policy Development:** Focus traffic management interventions on identified high-risk time windows and locations

The combination of clustering and frequent pattern mining proved effective in uncovering non-obvious relationships that traditional statistical analysis might miss. These insights can inform evidence-based traffic safety strategies and contribute to reducing accident rates.

References

Moosavi, S., Samavatian, M. H., Parthasarathy, S., & Ramnath, R. (2019). A Countrywide Traffic Accident Dataset. *arXiv preprint arXiv:1906.05409*.

Scikit-learn: Machine Learning in Python. Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. Raschka, S. (2018). *Journal of Open Source Software*, 3(24), 638.