Delhi Accident Data Analysis Report

1. Dataset Description

1.1 Source

Internal Delhi road accident records dataset spanning multiple years with comprehensive accident details.

1.2 Columns

- **YEAR** – Year of accident occurrence
- **DISTRICT** – District in Delhi where the accident occurred
- **VEHICLE AT FAULT** – The type of vehicle at fault in the accident
- **VICTIM** – Type of victim (pedestrian, cyclist, vehicle occupant, etc.)
- **TYPE OF ACCIDENT** – Categorization of the accident (simple, fatal, non-injury, etc.)
- **# INJURED** – Number of persons injured in the accident
- **# KILLED** – Number of persons killed in the accident
- Additional fields (e.g., _c7, _c8, _c9) related to dataset specifics

1.3 Data Quality

- No significant missing or null values
- Column names standardized for clarity
- Numeric columns such as injured and killed counts are properly inferred
- Dataset is geographically distributed over various districts in Delhi

2. Operations Performed

2.1 Data Cleaning & Exploration

- Checked for null or missing values, found none
- Analyzed unique values in categorical columns like DISTRICT, VEHICLE AT FAULT, VICTIM
- Summarized numeric columns for mean, min, max, standard deviation
- Handled column names with special characters using proper PySpark syntax

2.2 Descriptive Analytics

- Counted total accident records per district
- Visualized accident types distribution using pie charts
- Calculated total injured and killed sums across the dataset
- Analyzed yearly trends of accidents using bar plots and scatter plots
- Generated district-wise accident frequency analysis

2.3 Relationship Analysis

- Correlation between vehicle type and accident severity
- Distribution of injured and killed by district over years
- Patterns in accident type relative to victim classification
- Min/max injured and killed counts by district for risk assessment

2.4 Advanced PySpark Operations

- Used groupBy operations for district-wise aggregations
- Applied filtering conditions for specific accident criteria
- Performed column transformations and data type handling
- Exported filtered datasets (e.g., accidents with injured > 1) to CSV format

3. Key Insights

3.1 Accident Distribution

- Certain districts show higher accident frequencies requiring targeted interventions
- Pedestrians and vehicle occupants constitute major victim groups
- Fatal accidents represent a significant subset needing immediate attention
- Geographic clustering of accidents suggests infrastructure-related factors

3.2 Injury and Fatality Numbers

- Injured counts vary widely, from zero in some incidents to multiple in others
- Fatalities, while lower in absolute numbers, reflect serious safety concerns
- Average injured and killed per accident were computed to guide risk analysis
- District-wise casualty patterns reveal safety hotspots

3.3 Temporal Trends

- Accident counts show variation year to year, indicating cyclical patterns
- Certain years witnessed spikes in accidents related to specific vehicle types or districts
- Seasonal and temporal analysis opportunities exist for deeper insights

3.4 Vehicle and Victim Patterns

- Specific vehicle types are associated with higher injury/fatality rates
- Victim type distribution helps identify vulnerable road user groups
- Accident type classification aids in understanding severity patterns

4. Technical Implementation

4.1 PySpark Configuration

- Configured SparkSession for local execution mode
- Handled file path issues with proper URI formatting
- Implemented fallback mechanisms for data export operations

4.2 Data Processing Challenges

- Resolved column name issues with special characters using backticks
- Managed Hadoop connectivity issues with local file system operations
- Implemented robust error handling for data export operations

4.3 Visualization and Analysis

- Created pie charts for accident type distribution
- Generated scatter plots for injured vs killed analysis
- Produced bar charts for district-wise accident counts
- Exported filtered datasets for further analysis

5. Recommendations

5.1 Safety Measures

- Implement focused road safety interventions in high-accident districts
- Launch enhanced pedestrian safety campaigns in vulnerable areas
- Review and regulate vehicle types linked to severe accidents
- Establish accident monitoring systems for real-time response

5.2 Urban Planning

- Use accident data to prioritize traffic signal placements and road design improvements
- Develop infrastructure improvements in accident-prone districts
- Enhance emergency response capabilities based on accident hotspots
- Create pedestrian-friendly zones in high-risk areas

5.3 Data-Driven Policy Making

- Develop evidence-based traffic safety policies using accident patterns
- Allocate resources efficiently based on district-wise accident analysis
- Create targeted awareness campaigns for high-risk vehicle types and victim groups
- Implement dynamic traffic management based on accident trends

5.4 Future Work

- Build predictive models for accident risk forecasting using machine learning
- Employ clustering techniques to group districts by accident profiles
- Integrate with external data sources (weather, traffic volume, demographics) for richer analysis
- Develop real-time accident monitoring and alert systems
- Create interactive dashboards for stakeholder decision-making

6. Conclusion

The Delhi Accident Data analysis reveals critical insights into road safety patterns across different districts and time periods. The dataset demonstrates varying severity levels with significant numbers of injured and killed across years and vehicle types. Geographic distribution highlights specific districts requiring immediate attention for safety interventions.

Key findings include district-wise accident concentration, vehicle type risk patterns, and victim vulnerability profiles. The analysis framework using PySpark provides a robust foundation for ongoing safety monitoring and policy development.

This comprehensive analysis serves as a strong basis for urban planning, traffic safety measures, and targeted interventions by local authorities. The systematic approach to data processing and analysis ensures reliable insights for evidence-based decision making in road safety management.

The dataset proves invaluable for accident trend analysis, risk factor identification, and public safety enhancement strategies. Future iterations should focus on predictive modeling and real-time monitoring capabilities to further enhance road safety outcomes in Delhi.