Road Accidents Analysis – Short Report

1. Project Overview

The analysis examines road accident records from Kenya (2016–2017) with the aim of identifying temporal, spatial, and behavioral patterns, assessing high-risk areas, and deriving actionable insights for road-safety interventions.

Key focus areas:

- Temporal trends (hourly, daily, monthly)
- Geographic hotspots by county and specific locations
- Causes and categories of accidents
- Victim demographics (age, gender)

---

2. Data Quality and Key Pitfalls

The dataset, while providing a useful overview of accident occurrences, presents **several major limitations** that must be addressed for robust analytics:

2.1 Missing Values

- Significant gaps in critical fields such as **victim age**, **cause code**, and **county identifiers**.
- Multiple records have missing or ambiguous accident details, limiting the reliability of demographic and cause analysis.

Remedy:

- Enforce mandatory fields during data collection.
- Implement structured digital reporting forms with validation rules.

2.2 Inconsistent Text and Categorical Entries

- County names, gender labels, and road descriptions are inconsistent (e.g., HOMABAY vs HOMA BAY, J for unknown gender, 2M & 3F for mixed victims).
- Cause codes are inconsistently applied, with over 200 instances labeled “cause not traced.”

Remedy:

- Standardize categorical values using controlled vocabularies.
- Automate mapping of raw input values to predefined categories during data ingestion.
- Introduce drop-downs or selection menus in digital reporting tools to minimize free-text errors.

2.3 Duplicates and Redundant Records

- The dataset contains exact and near-duplicate entries, particularly for high-traffic locations.
- Multiple records appear for the same incident, inflating counts and skewing analysis.

Remedy:

- Implement unique incident identifiers (e.g., combination of date, time, location).
- Apply automated deduplication algorithms during preprocessing.

2.4 Sparse Temporal Coverage

- 2017 entries are underrepresented, producing skewed monthly and yearly trend analyses.
- Missing timestamps and incomplete time data limit accurate hourly trend assessments.

Remedy:

- Establish systematic data collection schedules.
- Validate time and date entries at the point of capture.

2.5 Ambiguous and Unstructured Narratives

- Accident narratives are free-text and often incomplete or inconsistent.
- Lack of standardized terminology limits automated NLP analysis for cause extraction.

Remedy:

- Introduce structured reporting templates with predefined categories for causes, circumstances, and outcomes.
- Provide training to personnel responsible for filling accident reports.

---

3. Key Analysis Findings (Data-Driven Insights)

- **High-risk counties:** Nairobi, Kiambu, Nakuru.
- **Peak accident times:** Evening hours (17:00–20:00) and mid-year months (April–July).
- **Dominant accident causes:** Driver-related factors including speeding, overtaking, and losing control.
- **Victim demographics:** Predominantly male (~84%), with ages concentrated in early- to mid-30s.

Note: While these findings are indicative, the reliability is constrained by the underlying data quality issues outlined above.

---

4. Recommendations for Data Improvement

1. **Digital, standardized reporting system** to replace manual records.
2. **Mandatory fields** for all critical variables (date, time, location, victim demographics, cause code).
3. **Controlled vocabularies and drop-down options** to ensure consistent categorical entries.
4. **Automated validation and de-duplication** during data ingestion.
5. **Continuous training** for data collectors on accurate and complete reporting.
6. **Integration with GPS and traffic sensors** for precise location and temporal accuracy.

By implementing these measures, future datasets will be more reliable, enabling accurate modeling, predictive analytics, and evidence-based road safety interventions.

---

5. Conclusion

The current analysis highlights both important trends and the **limitations imposed by poor data quality**. Correcting these pitfalls is crucial for meaningful insights, predictive modeling, and policy-making. While existing patterns indicate high-risk times, locations, and driver behaviors, any operational decisions based on this dataset must account for its **missing, inconsistent, and incomplete records**.