Farah Javed

Exercise 6.1

Data On Road Traffic Accidents In Germany 2024

Data dowloaded via link: Verkehrsunfälle mit Personenschaden in Deutschland (Unfallatlas)

Data Summary

Data source:

This is an external data source. The *Unfallatlas* dataset on opengeodata.nrw is published under an open data license **Datenlizenz Deutschland – Namensnennung – Version 2.0** and is hosted by the Landesbetrieb Information und Technik Nordrhein-Westfalen (IT.NRW). The opengeodata.nrw portal is part of the geoinformation services of the state government of North Rhine-Westphalia and is operated by IT.NRW as a platform for providing open geospatial data.

Data Collection:

- **Primary source:**
  Based on *road traffic accident reports* filed by police departments.
- **Geo-coordinate validation :**
  Accident locations are only included if their coordinates can be reliably matched to the official road network (from the state's surveying/mapping authorities).
    - Accidents with invalid or unmatchable coordinates are excluded.
    - Matching success rate is usually >90%, but was ~84% for NRW in 2022.

- **Inclusion criteria:**
    - Only accidents with personal injury are included (no property-only accidents).
    - Accidents must have valid geo-location data and pass plausibility checks.
- **Spatial aggregation (for map display):**
    - At low zoom levels: Only major roads (e.g. Autobahnen, Bundesstraßen); segments ~5 km in length.
    - At high zoom levels: All road types; segments shortened to ~250 m.
- **Attributes included in dataset:**
    - Number of deaths, seriously injured, and slightly injured
    - Location type (urban / rural)
    - Road category (e.g. Autobahn, Bundesstraße, etc.)

Data Contents:

This dataset contains georeferenced data on road traffic accidents involving personal injury across Germany in 2024, including accident location, severity, road category, and type of vehicle involved.

Data Limitations:

1. **Only includes accidents with personal injury**
    - Accidents involving only property damage are excluded, which may limit analyses of total traffic incident rates or hazardous locations with frequent non-injury collisions.
2. **Missing or excluded data due to geo-matching issues**
    - Accident reports are excluded if their coordinates cannot be reliably matched to official road geometries.
    - This leads to incomplete spatial coverage — especially in rural areas or where GPS inaccuracies occur.

    - For example, in some years up to ~15% of accidents were excluded in NRW.

3. **Lack of contextual accident information**
    - No details are provided on accident causes, weather conditions, driver behavior, or legal responsibility — limiting causal or behavioral analysis.
4. **No personal or demographic data**
    - While necessary for privacy, the lack of data on age, gender, or other demographics prevents analysis of risk factors by population subgroup.
5. **Annual update cycle**
    - The data is updated only once per year, which limits its usefulness for real-time analysis or responsive traffic safety planning.
6. **Excludes non-public road accidents**
    - Accidents on private property or non-public roads (e.g., farm paths, industrial areas) are typically not included.

Ethical Considerations:

- The dataset is anonymized and contains no personally identifiable information, which is appropriate and aligns with GDPR and German data protection standards.
- Excluding accidents without reliable geo-coordinates may skew spatial analyses or lead to underrepresentation of certain areas (e.g., rural regions with poor GPS data quality).
- Users may assume areas with no accidents in the data are safe, when in fact the data may be incomplete or filtered out.

Why I chose this data set:

- I will be able to showcase the skills that current job market in junior data analyst positions require like, tableau, python, data cleaning, interactive dashboard etc.

- This data set is real-world German data, giving me the opportunity to develop my skills further using relevant data to the market.
- This portfolio project has relevance to many different sectors and companies in Germany like Mobility and transportation, smart city NRW and traffic safety programs, other consulting and analytics firms that handle data projects for the public sector or mobility clients.

Data cleaning:

- No missing values found.
- No duplicates found.
- No mixed data type columns found.
- No outliers found.
- Changed column names from German to English for clarity.

Data Profile

Variables	Time variant/-invariant	Structured/Unstructured	Qualitative/Quantitative	Qualitative: Nominal/Ordinal Quantitative: Discrete/Continuous
OID_	Time invariant	Structured	Qualitative	Ordinal
accident_id	Time invariant	Structured	Qualitative	Ordinal
state	Time invariant	Structured	Qualitative	Ordinal
administrative_region	Time invariant	Structured	Qualitative	Nominal
administrative_district	Time invariant	Structured	Qualitative	Nominal
municipality	Time invariant	Structured	Qualitative	Nominal
year_of_accident	Time invariant	Structured	Qualitative	Ordinal
month_of_accident	Time invariant	Structured	Qualitative	Ordinal
hour_of_accident	Time variant	Structured	Quantitative	continuous

day_of_week	Time invariant	Structured	Qualitative	Ordinal
accident_category	Time invariant	Structured	Qualitative	Ordinal
collision_type	Time invariant	Structured	Qualitative	Ordinal
accident_behavior_type	Time invariant	Structured	Qualitative	Ordinal
light_conditions	Time invariant	Structured	Qualitative	Ordinal
accident_with_bicycle	Time invariant	Structured	Qualitative	Ordinal
accident_with_pessenger_car	Time invariant	Structured	Qualitative	Ordinal
accident_with_pedestrian	Time invariant	Structured	Qualitative	Ordinal
accident_with_motorcycle	Time invariant	Structured	Qualitative	Ordinal
involving_heavy_goods_vehicle	Time invariant	Structured	Qualitative	Ordinal
other_vehicle_involved	Time invariant	Structured	Qualitative	Ordinal
road_surface_conditions	Time invariant	Structured	Qualitative	Ordinal
UTM_X	Time invariant	Structured	Quantitative	ratio
UTM_Y	Time invariant	Structured	Quantitative	ratio
Longitude_WGS84	Time invariant	Structured	Quantitative	interval
Latitude_WGS84	Time invariant	Structured	Quantitative	interval
location_check_level	Time invariant	Structured	Qualitative	Ordinal

Derived columns:

Variables	**Time variant/-invariant**	**Structured/Unstructured**	**Qualitative/Quantitative**	**Qualitative: Nominal/Ordinal Quantitative: Discrete/Continuous**
State_name	Time invariant	Structured	Qualitative	ordinal

population	Time variant	Structured	Quantitative	continuous
Accidents_per_100k	Time variant	Structured	Structured	continuous

Clarifying Questions:

1. Which months, days of the week, and hours have the highest accident frequency?
2. Are there noticeable rush-hour or seasonal peaks?
3. Which states have highest accident rate/month?
4. How do accidents differ between urban and rural areas?
5. What type of vehicles get into accidents the most?
6. What is the percentage of serious accidents (loss of life) in each state?
7. Do light conditions have any impact on accident occurrence?
8. Do weather conditions influence accident frequency?
9. Where do most accidents occur? Type of road.
10. What kind of accidents are most common, involving a bicycle, pedestrian, heavy load vehicles, passenger car?
11. Can we identify geographical clusters of high-severity accidents?
12. Can we find types of accidents that tend to co-occur in similar conditions (e.g., late-night, poor lighting, weekend)?