Methods of Advanced Data Engineering

Project Report
Ajharali Shaikh
23072460

Analysing climate conditions and their involvement on road accidents.

Data Sources:

There are two primary datasets that were utilized in this study. The first dataset contains road accident data in Berlin, and the second dataset contains weather data for different cities all over Germany.

Dataset 1: Road Accident Dataset

This dataset provides detailed information on accident types, vehicle types, month-to-month road conditions, and other relevant factors. This dataset was chosen because it provides all the necessary values that are needed to accurately analyze the problem statement.

URL: https://www.statistik-berlin-brandenburg.de/opendata/AfSBBB_BE_LOR_Strasse_Strassenverkehrsunfaelle_2019_Datensatz.csv

Dataset 2: Weather Dataset

The second dataset comprises weather data for all of Germany, which includes precision, air pressure, and other meteorological variables.

URL: https://opendata.dwd.de/climate_environment/CDC/regional_averages_DE/monthly/precipitation/

Data Pipeline

The data pipeline process involves loading the two datasets from their respective URLs into a Python script, performing data cleaning and transformation, and merging both the road accident and weather datasets based on month. After that, save the final dataset in SQLite.

There are two primary datasets that were utilized in this study. The first dataset contains road accident data in Berlin, and the second dataset contains weather data for different cities all over Germany.

Data ingestion: fetching data from URLs

Data Cleaning:

Data Source 1:

1. All the unnecessary columns were removed from the dataset.

2. Column names were changed to English for easy understanding.

3. Month and road condition were changed from numerical to string values for better readability.

Data Source 2:

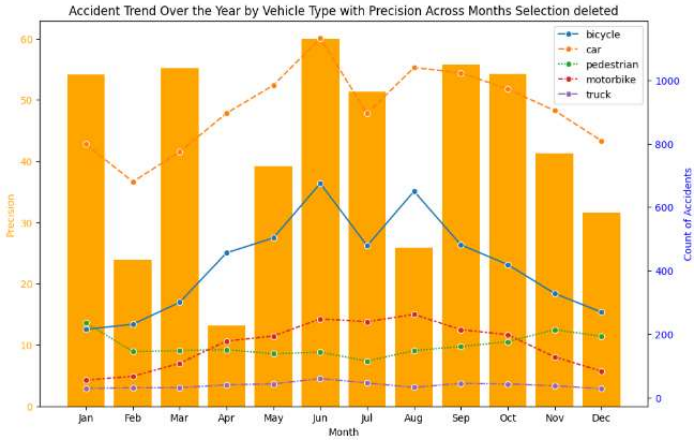
1. Data converted from a TXT file to a CSV.

2. Read the data only for Berlin in 2019.

3. Combine the data from precision, air pressure and sunshine duration in one data frame, referring to the month column.

Results:

Output Data:



Upon examining the graph above, it's clear that there is no direct correlation between the precision values from the weather dataset and the total number of road accidents involving different types of vehicles. This lack of correlation developed from the limited precision values available, as we only have 12 values for each month, making them insufficient for robust analysis. Consequently, I've opted to utilize road accidents as a reference point for analyzing the impact of weather conditions on the types of vehicles involved in road accidents.

Limitations:

Single-City Dataset: The limitation of having the accident dataset for a single city implies that the findings and patterns observed may not be universally applicable to other locations with different traffic dynamics, infrastructure, and socio-economic factors. Road safety issues can vary significantly between cities, so generalizing findings to a broader context may not be accurate.

Limited Time Frame: The dataset covering only one year may restrict the ability to identify long-term trends or seasonal variations that could influence road accidents. A more extended timeframe would provide a more comprehensive understanding of patterns and fluctuations, allowing for better-informed decision-making and the planning of safety measures.

Monthly Weather Data: The availability of weather data only monthly rather than daily or hourly restricts the granularity of the analysis. Accidents may be influenced by specific weather conditions at different times of the day or week, and having more detailed weather data would enhance the accuracy of identifying correlations between road accidents and weather.