Analyzing the Impact of Weather on Traffic Collision Dynamics in Los Angeles

The study examines whether daily weather variables—such as temperature, humidity, visibility, and wind speed—are correlated with the traffic accidents. By understanding these potential connections, the study aims to provide valuable insights for city planners, traffic managers, and public safety officials to enhance road safety and reduce the risk of weather-related collisions

1 Question

The main question is: What is the correlation between weather conditions and traffic collisions in Los Angeles from January to June 2020, and how do weather factors impact traffic collisions?

2 Data Sources

2.1 Traffic Collision Data

- **Metadata URL**: https://data.lacity.org/Public-Safety/Traffic-Collision-Data-from-2010-to-Present/d5tf-ez2w/about$_data$
- **Data URL**: https://data.lacity.org/resource/d5tf-ez2w.csv
- **Data Type**: CSV
- **Description**: This file contains the data provided by the Los Angeles Police Department for January to June 2020 and dataset owner is LAPD OpenData
- **License**: CC0 1.0 Universal

  Canonical URL: https://creativecommons.org/publicdomain/zero/1.0/legalcode.

2.2 Weather Data

- **Metadata URL**: `https://www.visualcrossing.com/resources/documentation/weather-data/weather-data-documentation/`
- **Data URL**: `https://weather.visualcrossing.com/VisualCrossingWebServices/rest/services/retrievebulkdataset?&key=NGNV4J2JDYHQS2AT24WRV28NH&taskId=aae51ed3104c7803e597cd073839ce9b&zip=false`
- **Data Type**: CSV

- **Description**: This file contains temperature, windspeed, humidity, and visibilty of the Los Angeles from January to June 2020.
- **License**: Visual Crossing Corporation. Visual Crossing Weather. 2021, `https://www.visualcrossing.com/`.

2.3 Compliance with Data Source Licenses

The traffic flow data is licensed under the Creative Commons Attribution CC0 1.0 Universal license, which allows for sharing and adaptation of the data, provided appropriate credit is given. The weather data from Visual Crossing is provided under their usage terms, which typically involve acknowledging the source and not using the data for commercial purposes without permission. The data is used strictly only for educational and research purpose.

3 Data Pipeline

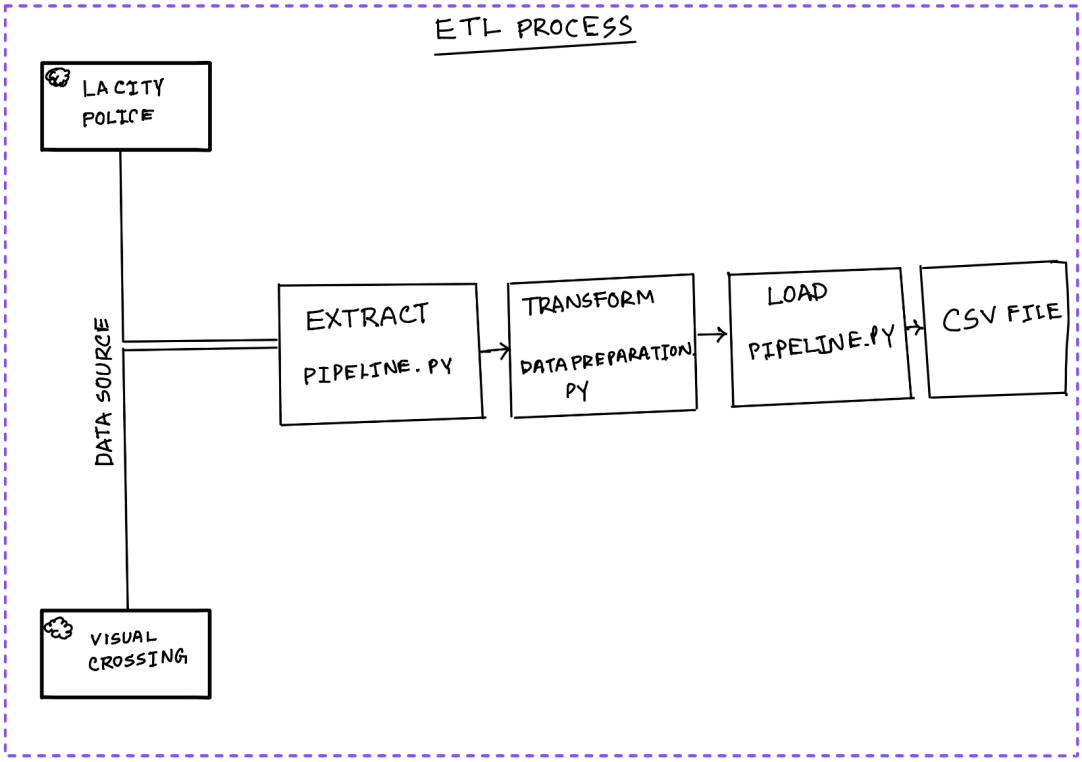

Figure 1: Data Pipeline

3.1 Data Pipeline

The pipeline can be run using the bash script `pipeline.sh` present in the project directory. It executes `pipeline.py`, which in turn executes `dataPreparation.py` for the data transformation steps.

Technology Used: Python, Pandas, Requests library, logging module, and CSV file format.

3.1.1 Extract

- **get_url_data()**: Fetches the data from the given URL and loads it into a DataFrame. Technically, this method sends a GET request to the provided URL, and reads the response into a DataFrame. Also, handles any errors.

3.1.2 Data Preparation

The functions in `process.py` handles dataset-specific cleaning:

- **collect_traffic_data()**: Fetches traffic data using pagination to handle API row limits. Combines batches into a single DataFrame. To get the dataset there are some limitation we were iónly able to download 1000 data so to bypass the API limit we had to use different function and parameters. So get_url_data Retrieves data directly from a provided URL and converts it into a DataFrame and dataframe_to_CSV function saves the given DataFrame to a CSV file.
- **weatherDataProcess()**: Process weather data by dropping unwanted columns for our project visit. We import this function from the dataDatapreparation.py file. The pre-processing os data is done there. The columns considered for further analysis are *datetime*, *temp*, *humidity*, *windspeed*, and *visibility*.
- **dataframe_to_CSV()**: This help to save the file in CSV format.

4 Results and Limitations

4.1 Preferred Data Format for Pipeline Output

The data pipeline outputs two structured CSV datasets: one for traffic collision data with columns date_occ and flow, and another for weather data with columns datetime, temp, humidity, windspeed, and visibility. Both datasets are filtered for January to June 2020 and aligned temporally to facilitate correlation analysis. Data quality is ensured through cleaning, including handling missing values, and dropping irrelevant columns. The final datasets are complete, relevant, consistent, and well-suited for further analysis.

4.2 Assessment of Data Quality and Potential Challenges

- **Traffic Collision Data**: The dataset is pretty decent quality I was confused initially how to utilize it for longer period to find the correlation traffic collision is an unexpected event so there could be many other reason to consider.Considerded only the date occurred as the I felt other columns were not that necessary for the analysis I am doing. Since I did not considered so many columns the dataset worked pretty fine for me.
- **Weather Data**: The traffic collision data might not cover all areas within Los Angeles equally, and the weather data could lack granularity or exhibit temporal gaps, potentially limiting the depth and accuracy of the insights derived from the analysis. I didn't consider taking the longer range because which could give a more concrete result but for short version this data was not good enough as I was only providing averaged longer periods.