Team Members

Name	ID
Omar Abdelrahman AbdelLatif Hasaballa	**22010166**
Ebrahim Mohamed Abdel Ghani	**22010010**
Youssef Tarek Mahmoud Fathallah	**22010305**
Abdelrahman Osama Mohamed Ali	**22010143**
Akram Yasser El Saffy Mabrouk	**22010053**
Mazen Mohamed Nasr Mohamed	**22011940**

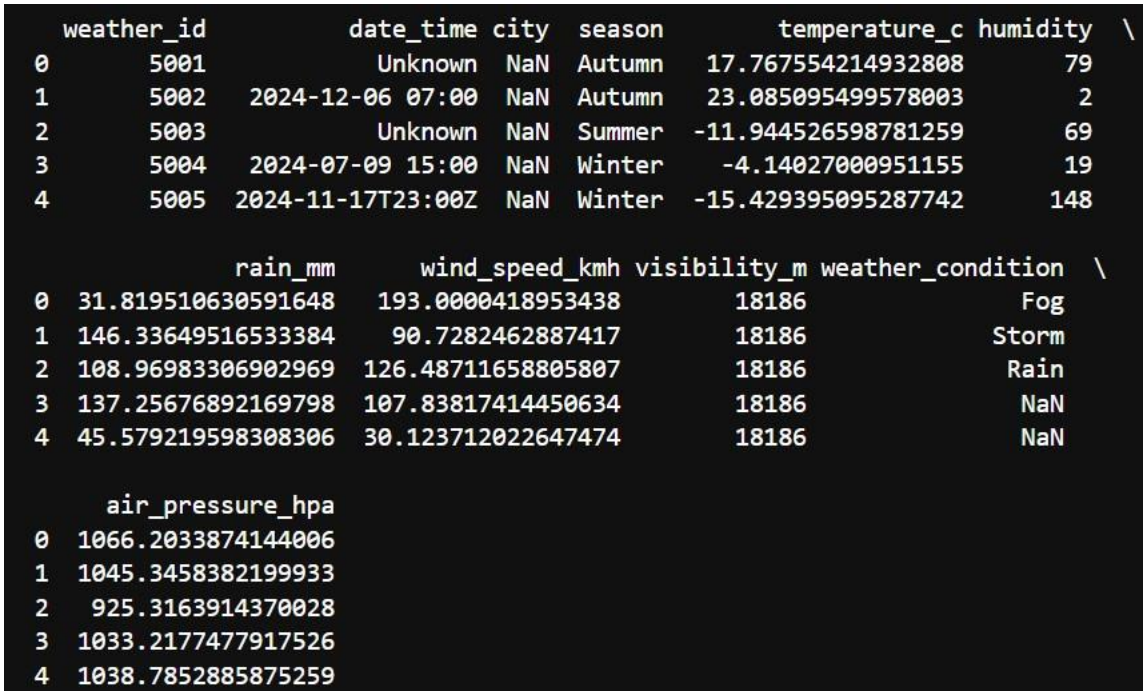
Introduction

This project focuses on studying the relationship between weather conditions and traffic in London. We collected weather and traffic data, cleaned and merged them, and then applied Factor Analysis and Monte Carlo Simulation to understand how weather affects congestion and accidents.

Data Cleaning Summary

2.1 Weather Data

- **Issues:** duplicate records, inconsistent date formats, outliers, and missing values.



	weather_id	date_time	city	season	temperature_c	humidity
0	5001	Unknown	NaN	Autumn	17.767554214932808	79
1	5002	2024-12-06 07:00	NaN	Autumn	23.085095499578003	2
2	5003	Unknown	NaN	Summer	-11.944526598781259	69
3	5004	2024-07-09 15:00	NaN	Winter	-4.14027000951155	19
4	5005	2024-11-17T23:00Z	NaN	Winter	-15.429395095287742	148

	rain_mm	wind_speed_kmh	visibility_m	weather_condition
0	31.819510630591648	193.0000418953438	18186	Fog
1	146.33649516533384	90.7282462887417	18186	Storm
2	108.96983306902969	126.48711658805807	18186	Rain
3	137.25676892169798	107.83817414450634	18186	NaN
4	45.579219598308306	30.123712022647474	18186	NaN

	air_pressure_hpa
0	1066.2033874144006
1	1045.3458382199933
2	925.3163914370028
3	1033.2177477917526
4	1038.7852885875259

- **Actions Taken:**
  - Removed duplicates and corrupted records.
  - Unified all date and time formats.
  - Corrected unrealistic values for temperature, humidity, and wind speed.
  - Filled missing numeric values with averages and inferred missing text values based on context.

- **Outcome:** Clean weather dataset (4975 rows), saved as Parquet in Silver layer, ready for analysis.

	weather_id	date_time	city	season	temperature_c	humidity	\
0	5002.0	2024-06-12 07:00:00	London	Autumn	23.085095	2	
1	5003.0	NaT	London	Summer	-11.944527	69	
2	5004.0	2024-09-07 15:00:00	London	Winter	-4.140270	19	
3	5011.0	NaT	London	Spring	-12.398642	26	
4	5012.0	NaT	London	Autumn	27.750814	67	

	rain_mm	wind_speed_kmh	visibility_m	weather_condition	\
0	146.336495	90.728246	18186	Storm	
1	108.969833	126.487117	18186	Rain	
2	137.256769	107.838174	18186	None	
3	72.169772	108.460015	18186	Snow	
4	124.076806	48.493849	18186	None	

	air_pressure_hpa
0	1045.345838
1	925.316391
2	1033.217748
3	1032.701396
4	917.339112

2.2 Traffic Data

- **Issues:** negative speeds, extreme vehicle counts, inconsistent dates, missing values.

	traffic_id	date_time	city	area	vehicle_count
0	9001	01/08/2024 03AM	London	NaN	1210
1	9002	04/11/2024 05PM	London	Chelsea	15625
2	9003	24/01/2024 02PM	NaN	Islington	6846
3	9004	2099-13-40 25:61	NaN	Kensington	11205
4	9005	2099-13-40 25:61	London	Kensington	18182

	avg_speed_kmh	accident_count	congestion_level	road_condition
0	69.56417018849834	37	NaN	NaN
1	-19.71282632639115	10	Low	Dry
2	130.37489694749635	13	Low	NaN
3	131.72744387361186	54	Low	NaN
4	93.66666962353031	45	High	Damaged

	visibility_m
0	17123
1	17123
2	17123
3	17123
4	17123

- **Actions Taken:**
  - Converted negative speeds to zero.
  - Capped extreme vehicle counts to realistic limits.
  - Unified date formats and filled missing values for congestion_level, city, and area.
- **Outcome:** Clean traffic dataset (4968 rows), stored in Parquet format in Silver layer, ready for analysis.

	traffic_id	date_time	city	area	vehicle_count
0	9003.0	2024-01-24 14:00:00	London	Islington	6846
1	9007.0	2024-06-21 16:00:00	London	Islington	13146
2	9008.0	2024-06-30 04:00:00	London	Kensington	8233
3	9010.0	NaT	London	Southwark	13399
4	9011.0	NaT	London	Islington	11418

	avg_speed_kmh	accident_count	congestion_level	road_condition	visibility_m
0	130.374897	13	Low	None	17123
1	8.577776	0	None	Wet	17123
2	58.794062	16	Medium	None	17123
3	76.843779	16	Low	None	17123
4	64.399451	9	Low	Dry	17123

Merged Dataset

- Linked each traffic record to the corresponding weather record for the same time and location.
- Used **Inner Join** on date_time and city.
- Output: **merged_dataset.parquet**, ready for advanced analysis.
- **Features of the merged dataset:**
  - Complete analytical coverage: includes all weather and traffic variables.
  - Ready for Monte Carlo Simulation and Factor Analysis.
  - Parquet format ensures fast reading and efficient performance on large datasets.

	weather_id	date_time	city	season	temperature_c	humidity	rain_mm	\
0	5003.0	NaT	London	Summer	-11.944527	69	108.969833	
1	5003.0	NaT	London	Summer	-11.944527	69	108.969833	
2	5003.0	NaT	London	Summer	-11.944527	69	108.969833	
3	5003.0	NaT	London	Summer	-11.944527	69	108.969833	
4	5003.0	NaT	London	Summer	-11.944527	69	108.969833	

	wind_speed_kmh	visibility_m_x	weather_condition	air_pressure_hpa	\
0	126.487117	18186	Rain	925.316391	
1	126.487117	18186	Rain	925.316391	
2	126.487117	18186	Rain	925.316391	
3	126.487117	18186	Rain	925.316391	
4	126.487117	18186	Rain	925.316391	

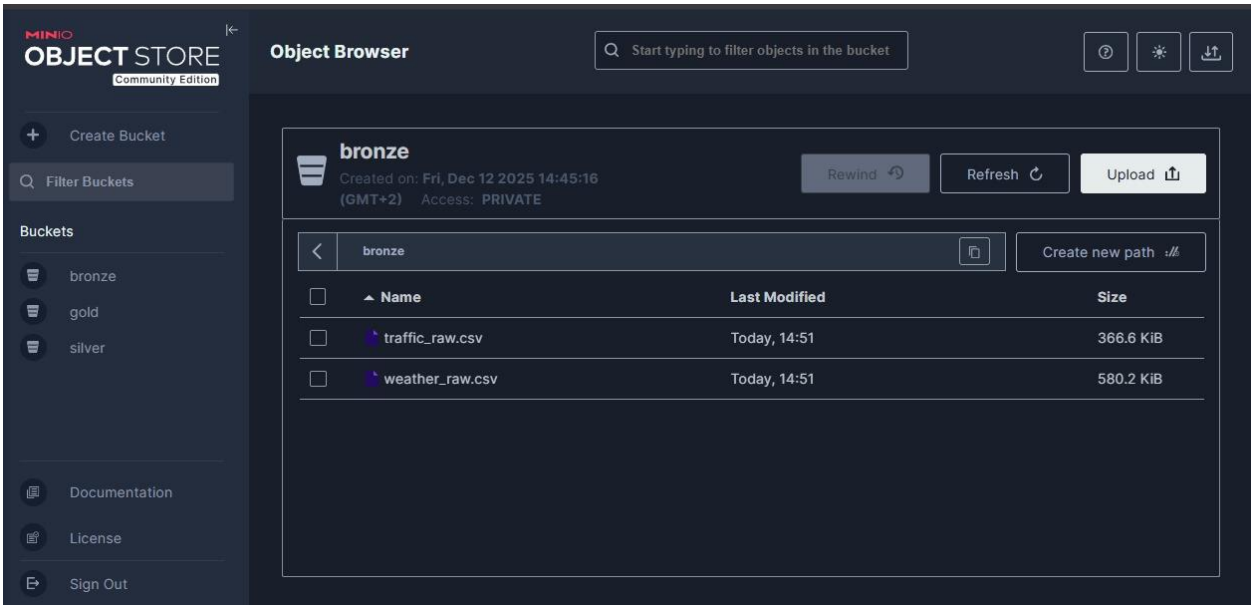
	traffic_id	area	vehicle_count	avg_speed_kmh	accident_count	\
0	9010.0	Southwark	13399	76.843779	16	
1	9011.0	Islington	11418	64.399451	9	
2	9012.0	Chelsea	6913	115.911074	29	
3	9018.0	Unknown	7019	120.454942	19	
4	9023.0	Kensington	12055	135.907217	19	

	congestion_level	road_condition	visibility_m_y
0	Low	None	17123
1	Low	Dry	17123
2	High	Dry	17123
3	Low	Damaged	17123
4	High	Snowy	17123

Data Lake Layers

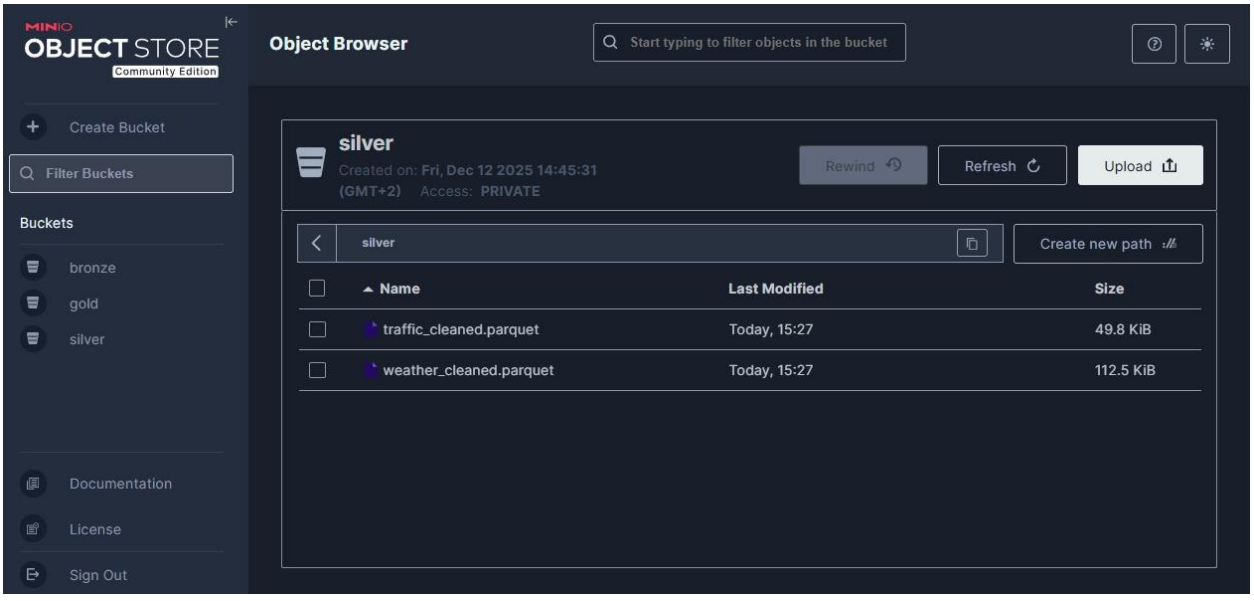
Bronze Layer

- **Description:** Raw, unprocessed data collected directly from sources.
- **Weather Data:** Raw weather readings with duplicates, missing values, inconsistent timestamps, and outliers.
- **Traffic Data:** Raw traffic counts, speeds, congestion, and accidents, often containing errors or missing fields.
- **Purpose:** Keep original data intact as a reference and backup before cleaning.



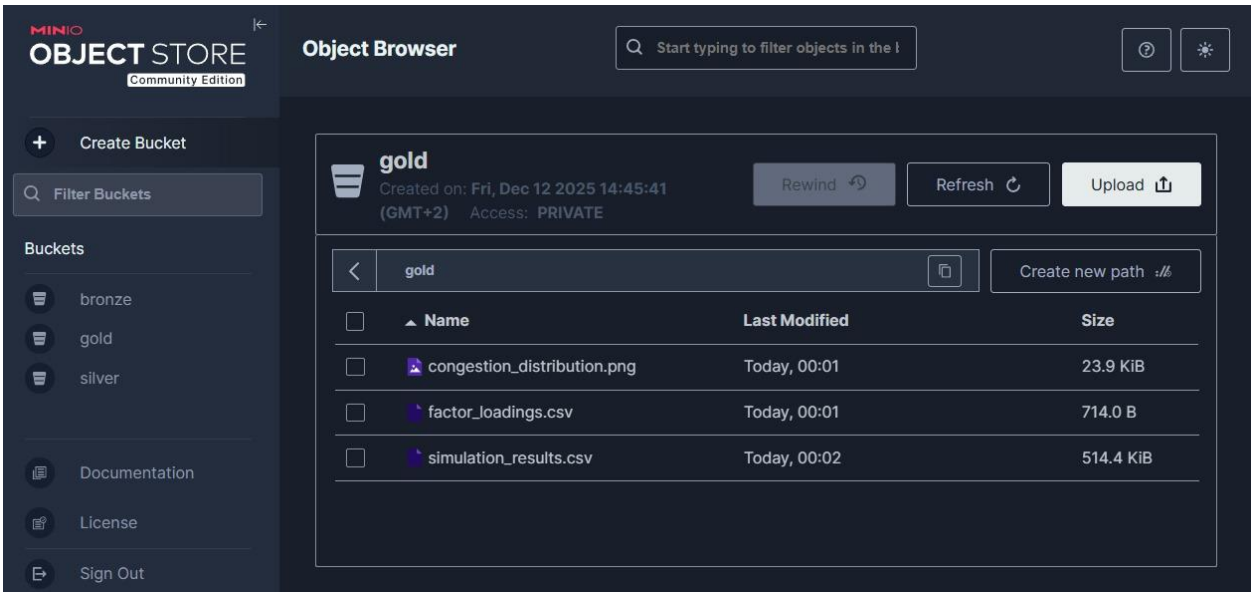
Silver Layer

- **Description:** Cleaned and preprocessed data ready for analysis.
- **Actions Taken:**
  - Removed duplicates and corrupted records.
  - Fixed outliers and missing values.
  - Unified date and time formats for merging.
- **Outcome:**
  - Weather and Traffic datasets cleaned and stored in **Parquet** format.
  - Merged dataset created by joining weather and traffic data on date_time and city.
- **Purpose:** Provide reliable, high-quality data for Monte Carlo simulations and Factor Analysis.



Gold Layer

- **Description:** Analysis-ready data and derived results.
- **Components:**
    - Monte Carlo simulation results (probabilities of congestion and accidents).
    - Factor Analysis outputs (factor loadings, heatmaps, interpretations).
- **Purpose:**
    - Gold layer contains insights and outputs that can be directly used for decision-making.
    - Acts as the final product of data processing and analysis, ready to feed dashboards or reports.



Factor Analysis

- Applied Factor Analysis to understand hidden relationships between weather and traffic.
- **Main factors extracted:**
    1. **Traffic Volume Factor:** vehicle_count and avg_speed_kmh; inversely related, reflects daily traffic flow, independent of weather.
    2. **Pressure & Severity Factor:** air pressure, rain, and humidity; strongly affects accident count and congestion. Most important factor.
    3. **Wind & Mobility Stress Factor:** wind_speed_kmh; independent factor that requires monitoring.
- **Conclusion:** Focus on Factor 2 (weather fluctuations) for predicting congestion and accidents, while monitoring wind as an independent factor.

	Factor_1	Factor_2	Factor_3
temperature_c	-0.000102351	-0.434030102	-2.489031613
humidity	8.98E-05	0.444306442	-0.13062354
rain_mm	6.80E-05	-0.845158689	-4.331076841
wind_speed_kmh	-0.00123359	-2.555183111	30.93055233
visibility_m_x	0	1.66E-24	-5.72E-23
air_pressure_hpa	-0.002235755	-57.24305171	-0.020107605
vehicle_count	-4323.336586	2.11E-06	-8.62E-09
avg_speed_kmh	-2.311723278	5.94E-05	-0.001007848
accident_count	0.431100642	-0.0001433	-0.000120939

Monte Carlo Simulation

- Conducted **10,000 simulations** to estimate the probability of high congestion and accidents under various weather scenarios.
- **Results:**
  - Worst multi-risk scenario: 65% congestion, 32% accidents.
  - Heavy rain + high humidity: 45% congestion, 22% accidents.
  - Heavy rain + extreme temperature: 45% congestion, 22% accidents.
  - Best case (Baseline): 5% congestion, 2% accidents.
- **Insight:** Results confirm Factor 2 (air pressure + rain + humidity) is the main driver of traffic risk, aligning with Factor Analysis findings.

simulation	congestion_prob	accident_prob	heavy_rain	temp_extreme	high_humidity	low_visibility	strong_winds
1	0.25	0.12	FALSE	FALSE	FALSE	FALSE	TRUE
2	0.25	0.12	FALSE	FALSE	FALSE	TRUE	FALSE
3	0.45	0.22	TRUE	FALSE	TRUE	FALSE	FALSE
4	0.05	0.02	FALSE	FALSE	FALSE	FALSE	FALSE
5	0.25	0.12	FALSE	TRUE	FALSE	FALSE	FALSE
6	0.45	0.22	TRUE	TRUE	FALSE	FALSE	FALSE
7	0.25	0.12	FALSE	TRUE	FALSE	FALSE	FALSE
8	0.05	0.02	FALSE	FALSE	FALSE	FALSE	FALSE
9	0.25	0.12	TRUE	FALSE	FALSE	FALSE	FALSE
10	0.05	0.02	FALSE	FALSE	FALSE	FALSE	FALSE
11	0.05	0.02	FALSE	FALSE	FALSE	FALSE	FALSE
12	0.05	0.02	FALSE	FALSE	FALSE	FALSE	FALSE
13	0.45	0.22	TRUE	FALSE	FALSE	TRUE	FALSE
14	0.25	0.12	TRUE	FALSE	FALSE	FALSE	FALSE
15	0.25	0.12	TRUE	FALSE	FALSE	FALSE	FALSE
16	0.05	0.02	FALSE	FALSE	FALSE	FALSE	FALSE

Dashboard Analysis & Interpretation

The interactive dashboard provides a clear visual summary of the project results and can be divided into three main sections: dataset statistics, Monte Carlo simulation results, and factor analysis insights.

1. Dataset Statistics Overview

The dataset statistics section displays an initial preview and summary of the merged weather and traffic data. The sample data shows that the records mainly correspond to **London** during the **summer season**.

Summary statistics indicate:

- A large number of records, reflecting a rich dataset suitable for statistical analysis.

- An average temperature of approximately **19°C**, with average humidity around **50%**, which is reasonable for summer conditions in London.
- Higher maximum values for variables such as temperature and rainfall, representing extreme or rare weather conditions.

Some unusual values (such as very low temperatures during summer or missing values in certain columns) appear in the raw preview. These values represent edge cases or residual data issues and highlight the importance of the earlier data cleaning phase applied before analysis.

Overall, this section confirms the scale and diversity of the dataset used in the project.

📊 Traffic & Weather Analysis Dashboard

Dataset Statistics

	weather_id	date_time	city	season	temperature_c	humidity	rain_mm	wind_speed_kmh	visibility_m_x	weather_condition
0	5003	NaT	London	Summer	-11.9445	69	108.9698	126.4871	18186	Rain
1	5003	NaT	London	Summer	-11.9445	69	108.9698	126.4871	18186	Rain
2	5003	NaT	London	Summer	-11.9445	69	108.9698	126.4871	18186	Rain
3	5003	NaT	London	Summer	-11.9445	69	108.9698	126.4871	18186	Rain
4	5003	NaT	London	Summer	-11.9445	69	108.9698	126.4871	18186	Rain
5	5003	NaT	London	Summer	-11.9445	69	108.9698	126.4871	18186	Rain
6	5003	NaT	London	Summer	-11.9445	69	108.9698	126.4871	18186	Rain
7	5003	NaT	London	Summer	-11.9445	69	108.9698	126.4871	18186	Rain
8	5003	NaT	London	Summer	-11.9445	69	108.9698	126.4871	18186	Rain
9	5003	NaT	London	Summer	-11.9445	69	108.9698	126.4871	18186	Rain

Summary Statistics

	weather_id	temperature_c	humidity	rain_mm	wind_speed_kmh	visibility_m_x
count	548350	561313	561313	561313	561313	561313
mean	7492.1161	19.3147	49.6641	79.2008	75.4097	18186
std	1447.6057	23.4705	29.0578	43.0753	43.4056	0
min	5003	-19.901	0	0.1488	0.2486	18186
25%	6251	-0.5862	25	42.6871	35.9828	18186
50%	7452	18.8201	48	80.7158	76.19	18186
75%	8768	40.355	75	117.0091	113.2349	18186
max	9995	59.9555	100	149.9807	149.9535	18186

Monte Carlo Simulation Results

2. Monte Carlo Simulation Results

The Monte Carlo simulation section visualizes the probabilistic outcomes of traffic congestion and accidents under different weather scenarios.

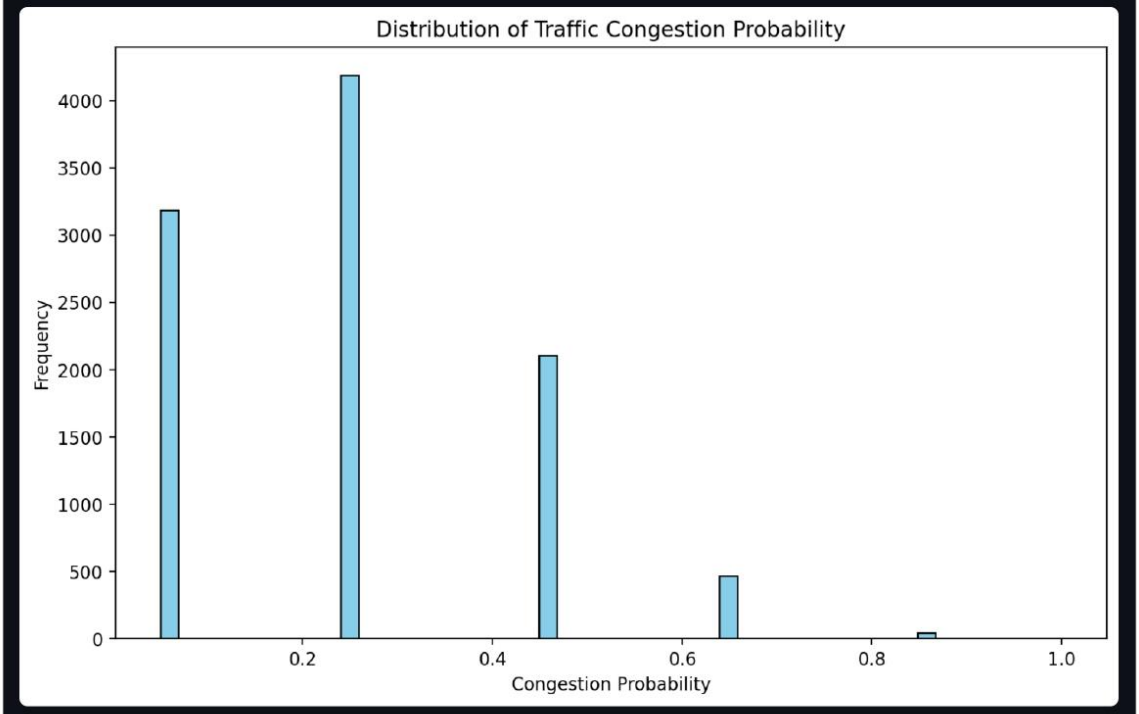
Key observations:

- **Congestion probability** values range approximately from **0.05 to 0.45** in the displayed samples.
- **Accident probability** values show a similar range, with most values concentrated at lower probabilities.
- The histogram of congestion probability shows that the most frequent risk levels are around **0.25–0.30**, indicating moderate congestion risk in many scenarios.
- The histogram of accident probability shows that most outcomes fall between **0.10 and 0.20**, suggesting that high accident risk is less frequent but still possible under certain conditions.

These visualizations help demonstrate how traffic risk increases gradually as weather conditions become more severe, rather than changing suddenly.

	simulation	congestion_prob	accident_prob	heavy_rain
0	1	0.25	0.12	☐
1	2	0.25	0.12	☐
2	3	0.45	0.22	☑
3	4	0.05	0.02	☐
4	5	0.25	0.12	☐
5	6	0.45	0.22	☑
6	7	0.25	0.12	☐
7	8	0.05	0.02	☐
8	9	0.25	0.12	☑
9	10	0.05	0.02	☐

Congestion Probability Distribution



Accident Probability Distribution

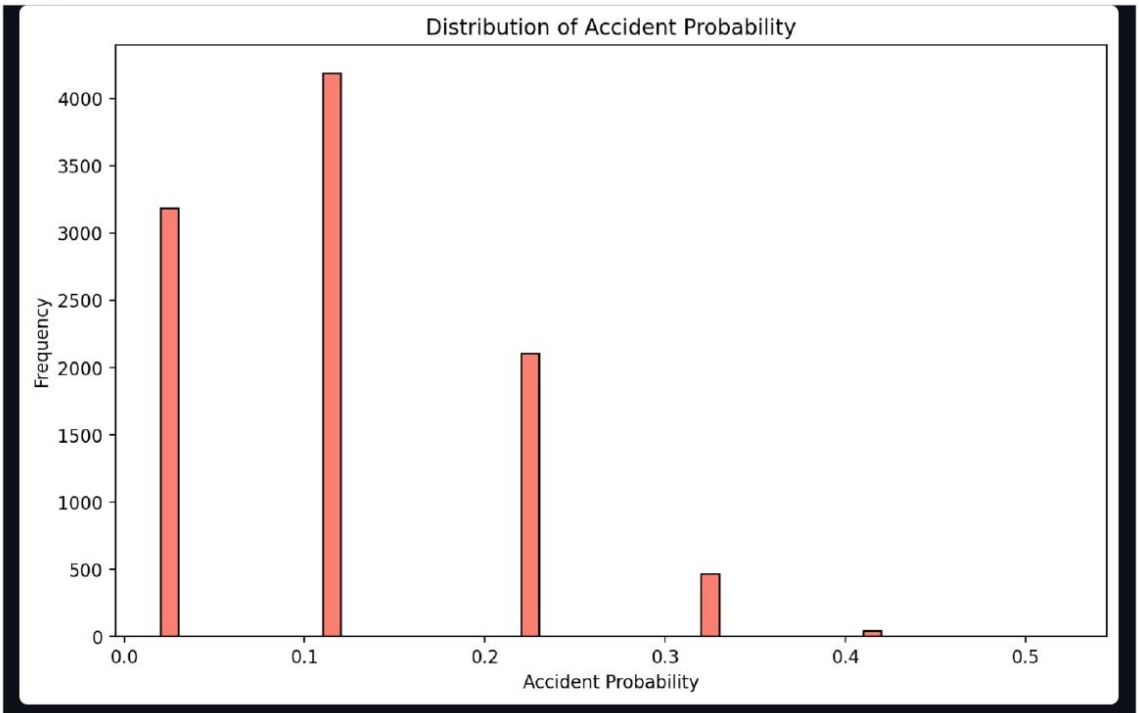
3. Factor Analysis Insights

The factor analysis section explains how different weather and traffic variables are grouped into a smaller number of underlying factors.

Main insights from the dashboard:

- **Factor 1** is strongly influenced by vehicle count, average speed, and accident count. This factor mainly represents traffic intensity and its direct outcomes.
- **Factor 2** is influenced by humidity, rainfall, and average speed, indicating the impact of weather conditions on traffic performance.
- **Factor 3** is dominated by wind speed, temperature, and rainfall, representing general weather conditions that may independently affect mobility.

The factor loadings table and heatmap make it easier to visually identify which variables have the strongest influence on each factor.



Factor Analysis Insights

	Factor_1	Factor_2
temperature_c	-0.0001	
humidity	0.00009	
rain_mm	0.00007	
wind_speed_kmh	-0.0012	
visibility_m_x	0	
air_pressure_hpa	-0.0022	
vehicle_count	-4323.3366	
avg_speed_kmh	-2.3117	
accident_count	0.4311	

Factor Loadings Heatmap

Heatmap file not found. Please generate it first.

Top Variables per Factor

Factor_1 likely influenced by: vehicle_count, avg_speed_kmh, accident_count

Factor_2 likely influenced by: air_pressure_hpa, wind_speed_kmh, rain_mm

Factor_3 likely influenced by: wind_speed_kmh, rain_mm, temperature_c

Overall Dashboard Conclusion

The dashboard successfully transforms complex analytical results into intuitive visual insights.
It clearly demonstrates the relationship between weather conditions (such as rain, humidity, and wind) and traffic behavior (congestion and accidents).

By combining cleaned data, Monte Carlo simulation results, and factor analysis outputs, the dashboard acts as the final presentation layer of the project and supports data-driven understanding of traffic risks under different weather scenarios.

Recommendations

1. Focus on monitoring fluctuating weather to predict congestion and accidents in advance.
2. Monitor high wind speed as an independent warning factor.
3. Use the probabilities from simulations to activate traffic management plans during high-risk weather.
4. Continuously update and clean datasets to maintain prediction accuracy.

GitHub Repository:
https://github.com/OMARg404/bigData