Big Data Project Technical Report

Table of Contents

Data Cleaning Summary

Traffic Data

Number of rows before cleaning: **5150**
Number of rows after cleaning: **4733**
Rows Removed (Duplicates/Unfixable): **417**

Cleaning Summary:

- Row Duplicates Removed: 150 occurrences fixed/handled.
- Unfixable Timestamp Errors: 267 occurrences fixed/handled.
- Standardized Congestion Categories: 88 occurrences fixed/handled.
- Imputed Missing City: 96 occurrences fixed/handled.
- Imputed Missing avg_speed_kmh by Area: 98 occurrences fixed/handled.
- Imputed Missing vehicle_count by Area: 96 occurrences fixed/handled.
- Corrected Negative Speeds: 25 occurrences fixed/handled.

Issue Type	Columns Affected	Probability	Approx. Count	Cleaning Method
Duplicates (Rows)	All	3%	**150** (Explicitly added)	Drop duplicates
Missing Values	Almost all	2%	**~100** per column	Imputation (Area Mean)
Outliers (Counts)	vehicle_count, accident_count	1%	**~50** per column	Clipping (Min/Max bounds)
Invalid Speeds	avg_speed_kmh	1%	**~50** (Negative/Extreme)	Absolute value / Clipping
Duplicate IDs	traffic_id	1%	**~50** (Reused IDs)	Deduplication
Inconsistent Categories	congestion_level	2%	**~100** ("Non_Standard"/"4")	Standardization mapping
Date Formats	date_time	5%	**~267**	Dropped

Weather Data

Number of rows before cleaning: **5150**
Number of rows after cleaning: **5000**
Rows Removed (Duplicates/Unfixable): **150**

Cleaning Summary:

- Row Duplicates Removed: 150 occurrences fixed/handled.
- Imputed Missing City/Condition: 218 occurrences fixed/handled.
- Imputed Missing Season: 95 occurrences fixed/handled.
- Imputed temperature_c by Season: 105 occurrences fixed/handled.
- Imputed humidity by Season: 96 occurrences fixed/handled.
- Imputed air_pressure_hpa by Season: 97 occurrences fixed/handled.
- Corrected Temp Outliers: 47 occurrences fixed/handled.
- Corrected Humidity Outliers: 62 occurrences fixed/handled.

Issue Type	Columns Affected	Probability	Approx. Count	Cleaning Method
Duplicates (Rows)	All	3%	**150** (Explicitly added)	Drop duplicates
Missing Values	Almost all	2%	**~100** per column	Imputation (Mode/Mean)
Outliers (Numeric)	temperature_c, humidity, rain_mm, wind_speed_kmh	1%	**~50** per column	Clipping (Min/Max bounds)
Duplicate IDs	weather_id	1%	**~50** (Reused IDs)	Deduplication
Mixed Data Types	visibility_m	2%	**~100** (Contains "LOW")	Coercion to numeric

Monte Carlo Simulation

1. The goal of Monte Carlo is to attempt to simulate many weather conditions often bad to estimate the traffic behavior (traffic jam and accident risk) in each scenario

2. We assumed 5 different weather scenarios:
    a. Heavy rain
    b. Temperature extremes (Low & High Temperatures)
    c. High humidity
    d. Low visibility
    e. Strong winds

3. We calculated the actual traffic_jam probability (when congestion_level = 'High') and the actual accident risk probability (when accident_count > 0) from the original dataset (merged_dataset.parquet)

4. We used **bootstrapping (non-parametric resampling) algorithm** to simulate 10,000 different datasets for each scenario then, we calculate traffic jam and accident risk probability for every dataset in every scenario

5. After this step we will have 2 lists of length 10,000 [traffic_jam_monte_carlo_probs, accident_risk_monte_carlo_probs] for each scenario then we calculate the following statistics for the two lists:

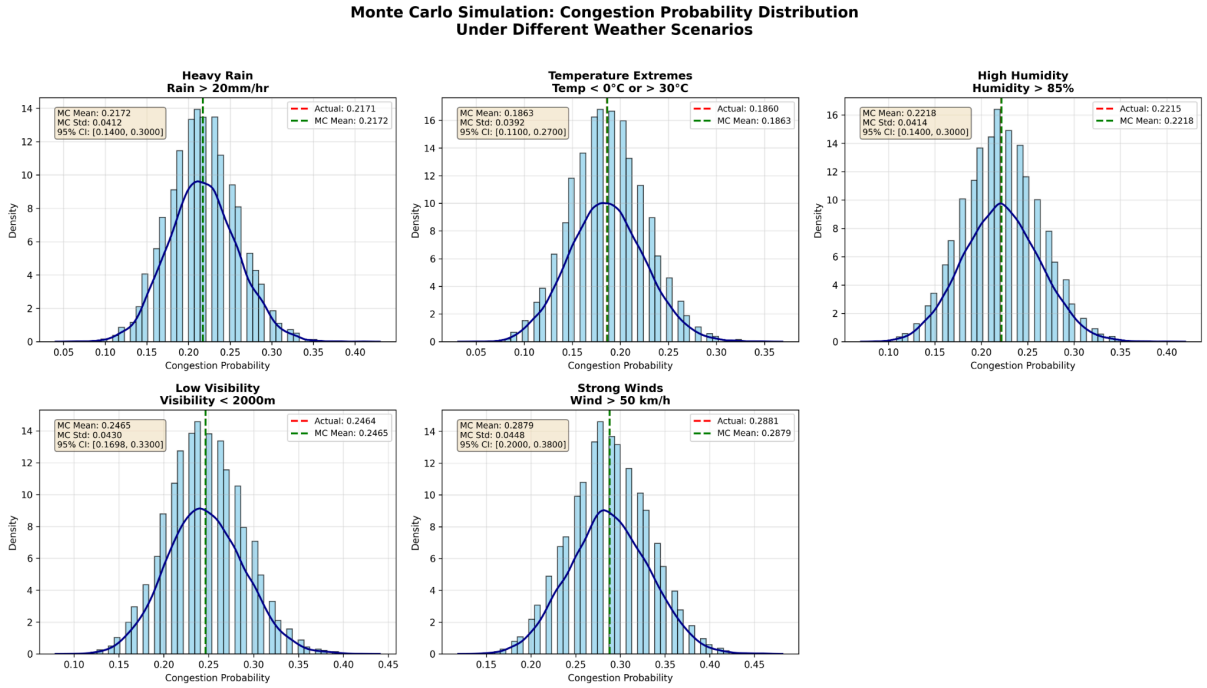
    - monte_carlo_mean
    - monte_carlo_std
    - monte_carlo_ci_95

    Note: As we increase n_simulation (number of datasets we create for each scenario) the monte_carlo_traffic_jam_mean be closer to actual traffic_jam probability from the original dataset and the CI be more tighter, the same for accident_risk

6. After finishing the algorithm and getting the results (actual probabilities and Monte Carlo results for the traffic jam and accident risk) we extracted a csv file that summarizes those results for the 5 scenarios but we excluded the full lists of the traffic_jam and accident risk values and store just the mean and std of them

Scenario	actual_traffic jam_prob	monte_carlo_ traffic_jam_mean	monte_carlo traffic_jam_std	monte_carlo traffic_jam ci_95_lower	monte_carlo traffic_jam ci_95_upper	actual_accident risk_prob	monte_carlo accident_risk_mean	monte_carlo accident_risk_std	monte_carlo accident_risk ci_95_lower	monte_carlo accident_risk ci_95_upper
Heavy Rain	0.217142857	0.217249	0.04119347	0.14	0.3	0.865714286	0.865966	0.033937396	0.8	0.93
Temperature Extremes	0.186046512	0.186295	0.039154859	0.11	0.27	0.813953488	0.814321	0.038860506	0.74	0.89
High Humidity	0.221508828	0.221834	0.041357665	0.14	0.3	0.911717496	0.911873	0.028290668	0.85	0.96
Low Visibility	0.246376812	0.246508	0.043034474	0.16975	0.33	0.886128364	0.886244	0.03164763	0.82	0.94
Strong Winds	0.288100209	0.287947	0.044802402	0.2	0.38	0.901878914	0.901589	0.029923988	0.84	0.96

7. Finally, we drew a plot about Distribution of congestion probabilities (monte_carlo_traffic_jam_probabilites) using matplotlib and seaborn libraries for each Scenario

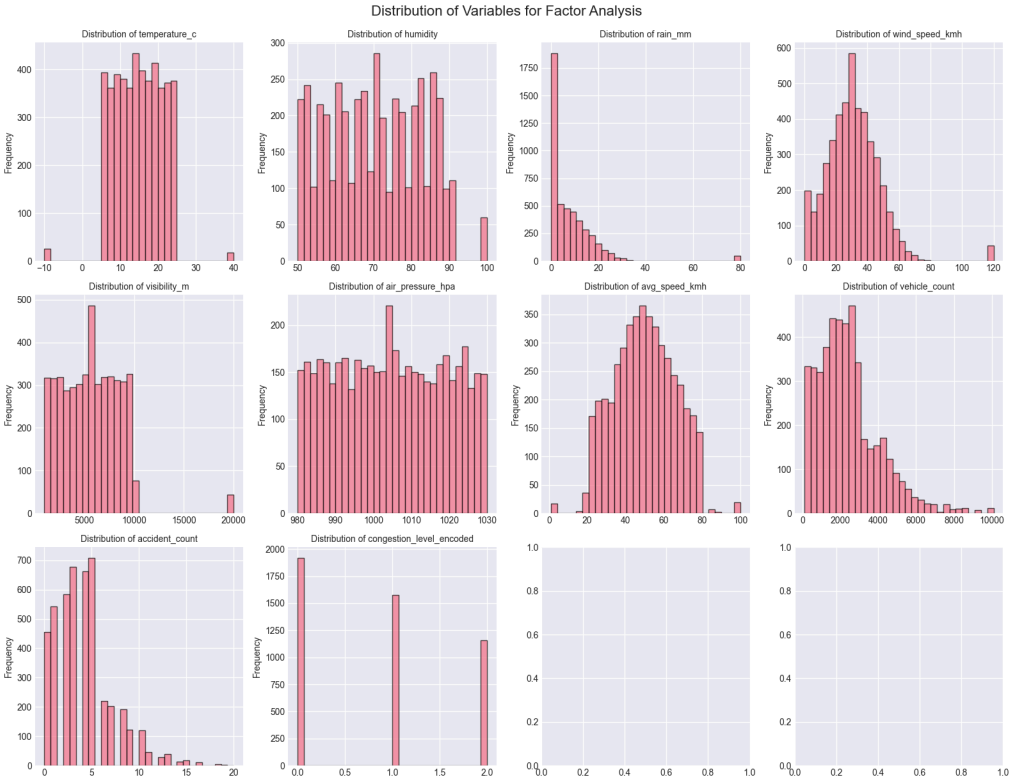


Factor Analysis Interpretation

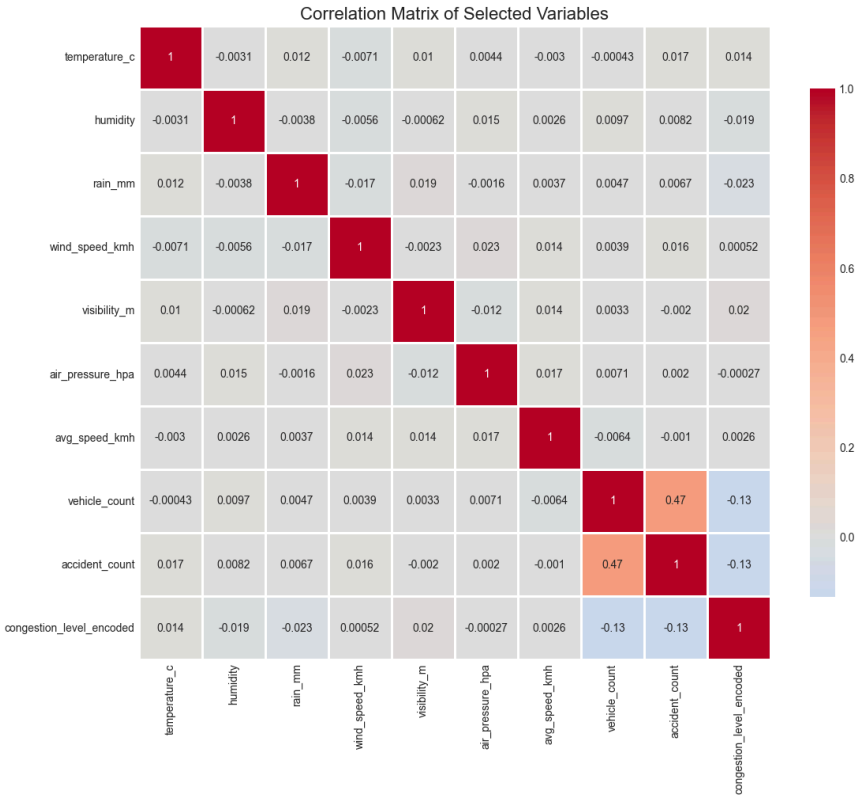
Our factor analysis process involved the following steps:

- **Correlation Analysis:** We created a correlation matrix to pinpoint the weather variables with the highest impact on traffic data.
- **Factor Extraction:** Using eigenvalues to determine significance, we extracted three primary factors: Weather Severity, Traffic Flow Stress, and Accident Risk.
- **Interpretation:** We generated a factor loadings table to interpret the composition of these factors.
- **Impact Assessment:** Finally, we tested the influence of each identified factor across each category of congestion level."

Below are our results:



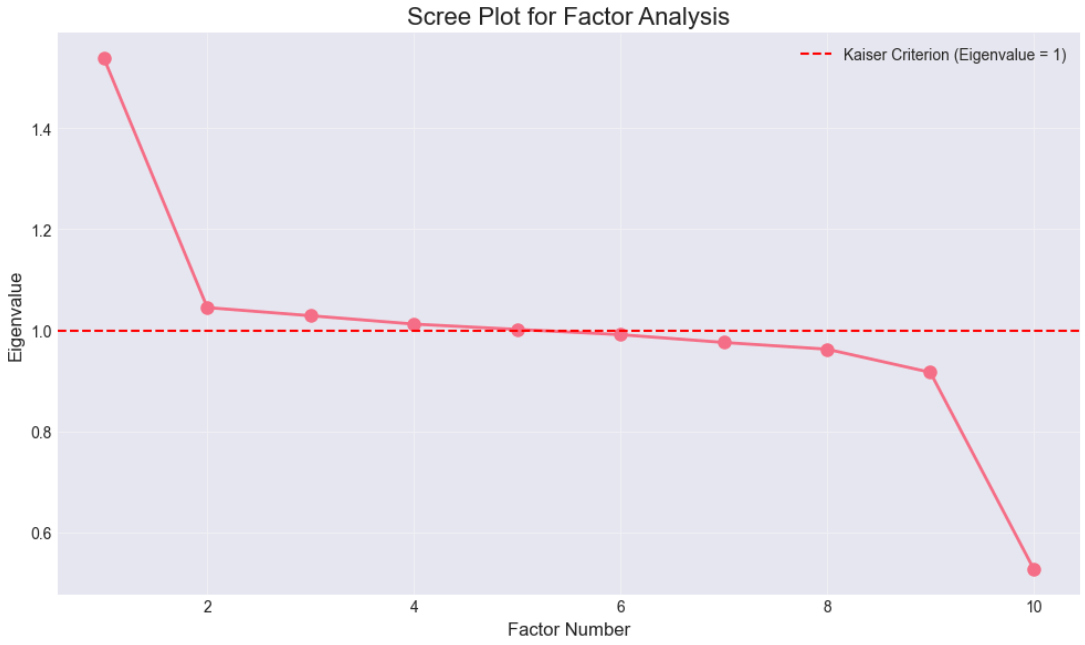
Variable Distributions



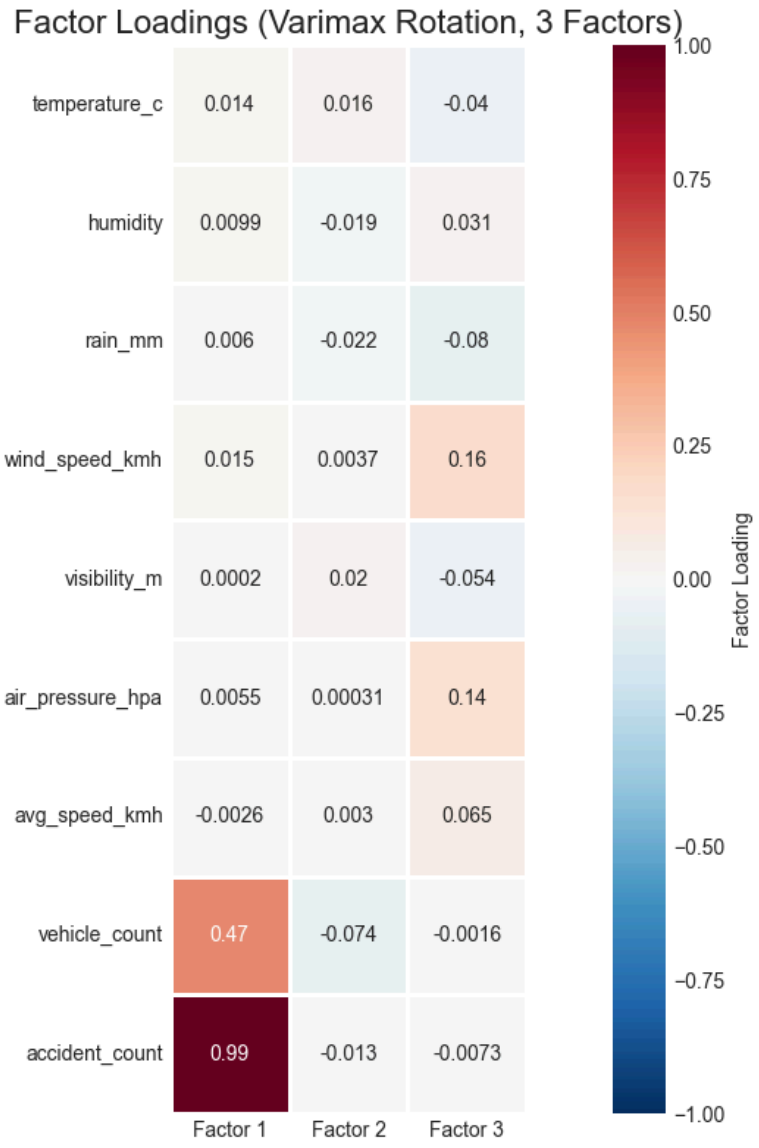
using heatmap to find and analyse the relationship of each variable

```
Strong Correlations (|r| > 0.3):
  vehicle_count ↔ accident_count: 0.503

2. Bartlett's Test of Sphericity:
  Chi-square: 1425.03
  p-value: 0.000000
  ✓ Data is suitable for factor analysis (p < 0.05)

3. Kaiser-Meyer-Olkin (KMO) Test:
  Overall KMO: 0.511
  ✗ KMO value (0.511) indicates data may not be suitable

  Individual KMO values:
                Variable   KMO
           temperature_c 0.481
                humidity 0.487
                 rain_mm 0.480
          wind_speed_kmh 0.483
            visibility_m 0.474
        air_pressure_hpa 0.629
           avg_speed_kmh 0.494
           vehicle_count 0.507
          accident_count 0.507
congestion_level_encoded 0.705
```

barlett's test and kaiser-meyer-oklin test to find weather the data suitable for factor analysis



pick best factors based on eigen values



loading table for factors to interpret them

Factor Interpretation and Naming

Factor 1:

accident_count: 0.992
vehicle_count: 0.474

Proposed Name: Traffic Stress Factor
Interpretation: Represents high traffic volume and congestion levels

Factor 2:

Proposed Name: Visibility & Speed Factor
Interpretation: Related to visibility conditions and average speed

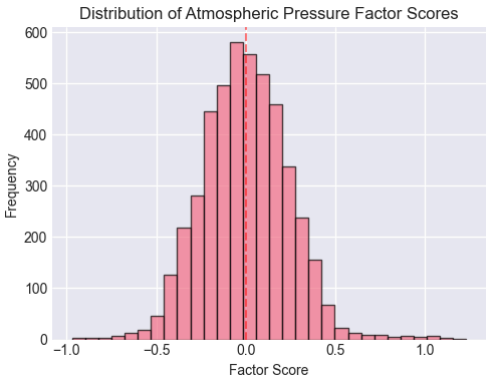
Factor 3:

wind_speed_kmh: 0.159
air_pressure_hpa: 0.141

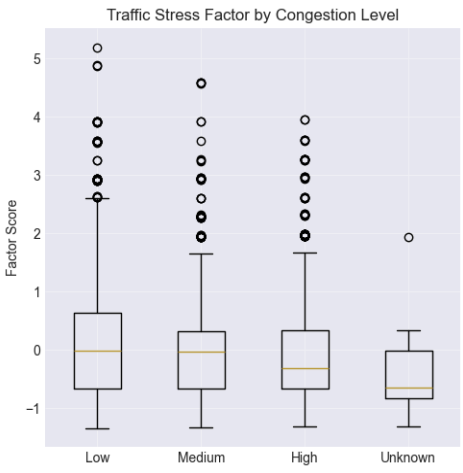
Proposed Name: Atmospheric Pressure Factor
Interpretation: Mainly captures atmospheric pressure variations

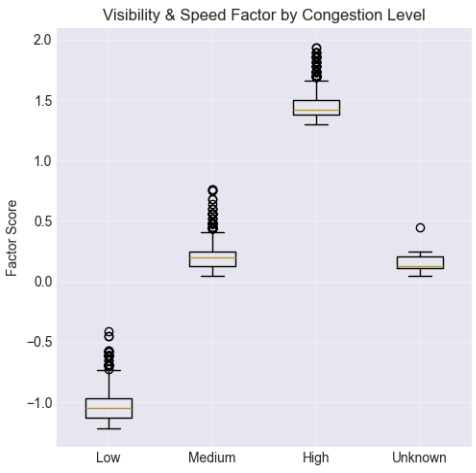
Factor Interpretation and Table

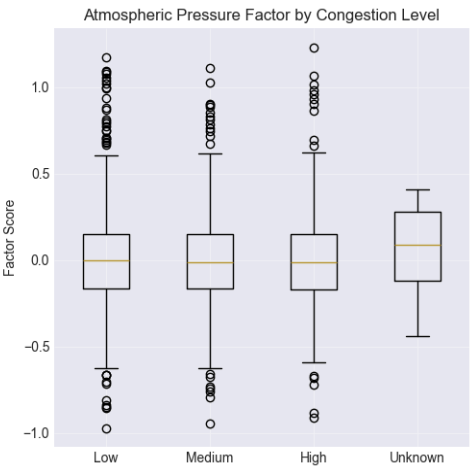
Factor	Proposed Name	Key Variables (Loadings)	Interpretation	Variance Explained
Factor 1	Traffic Stress Factor	accident_count (0.99) vehicle_count (0.47)	Represents high traffic volume and congestion levels.	1.225192
Factor 2	Visibility & Speed Factor	visibility_m (0.99)	Related to visibility conditions and average speed.	0.993476
Factor 3	Atmospheric Pressure Factor	wind_speed_kmh (0.16) air_pressure_hpa (0.14)	Mainly captures atmospheric pressure variations.	0.061410



Factors Distributions







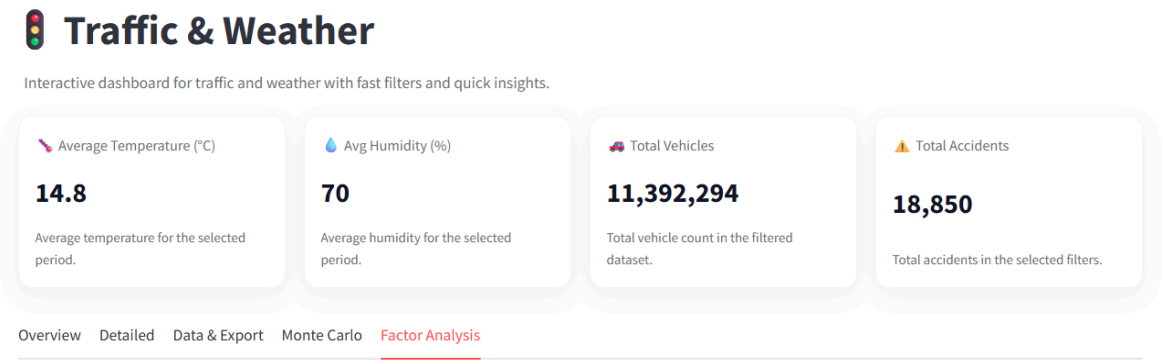
Analyzing Factor Relationships with Congestion Level

Data Visualization Dashboard :

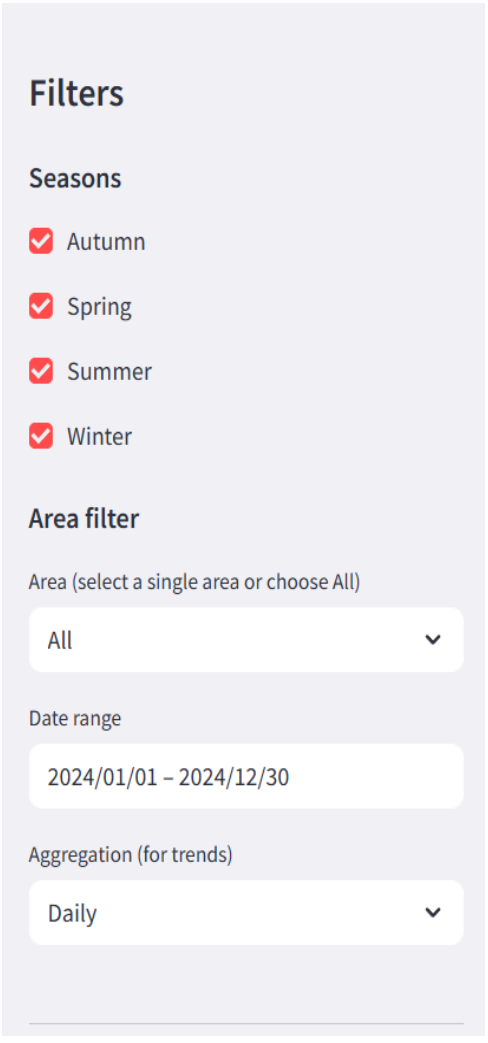
An interactive dashboard was built using **Streamlit** to provide quick and detailed insights into the complex relationship between traffic conditions and weather patterns.

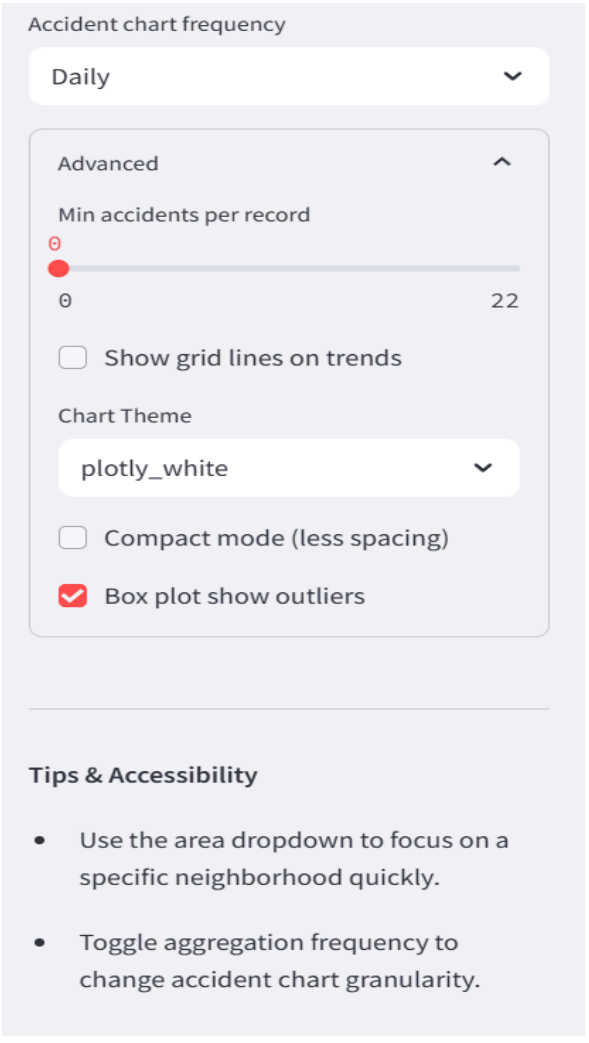
Key Features

- **Key Performance Indicators (KPIs):** Instant summary display of average temperature, humidity, total vehicle count, and total accidents.



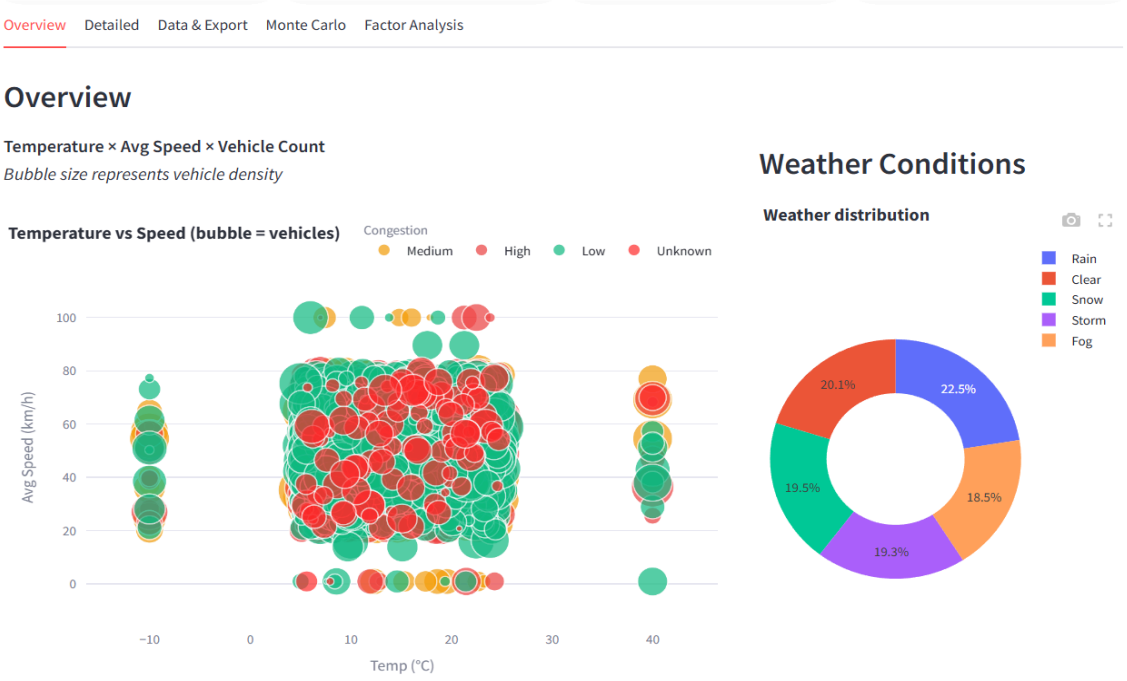
- **Filtering Controls:** Data filtering based on season, area, date range, and minimum accident threshold.



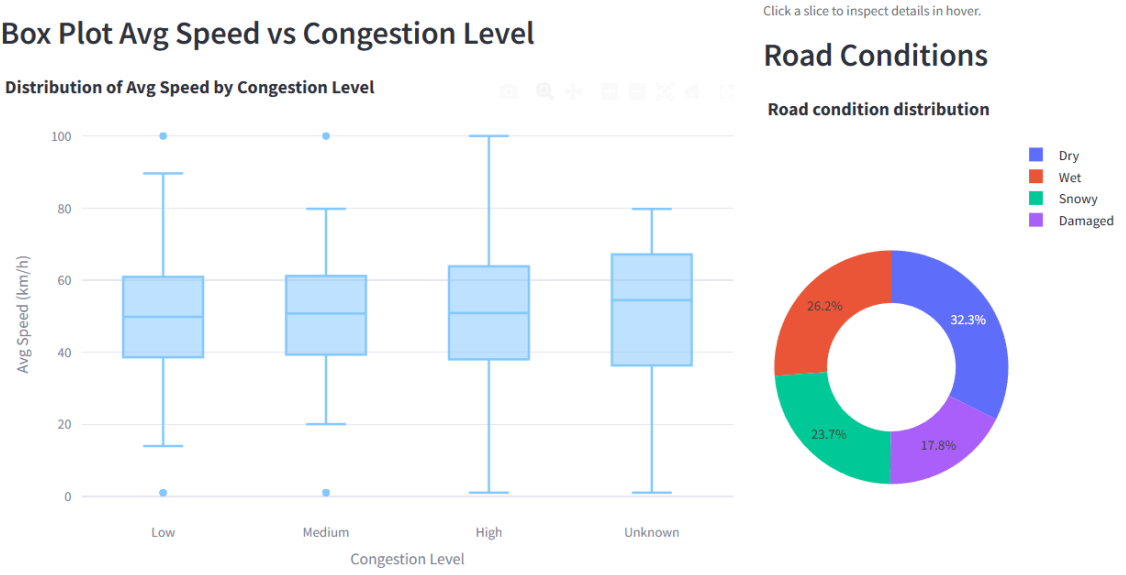


- **Tabs for Comprehensive Analysis:** The dashboard is organized into several tabs:
    - **Overview:**

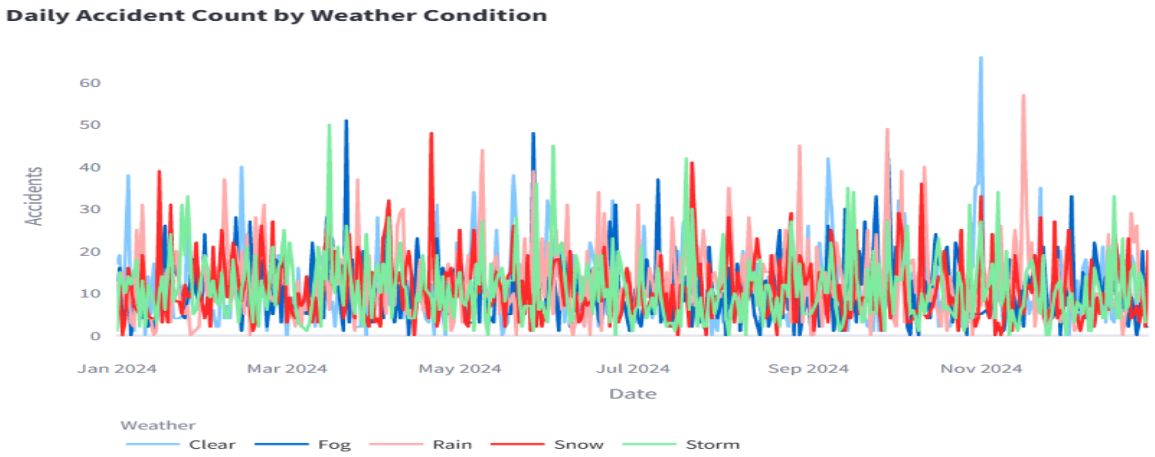
        - A **Scatter Plot** correlating Temperature with Average Speed (bubble size represents vehicle count) .
        - **Weather Condition Pie Chart:** Shows the proportional **distribution of different weather conditions** (e.g., Clear, Cloudy, Rain) in the filtered dataset.



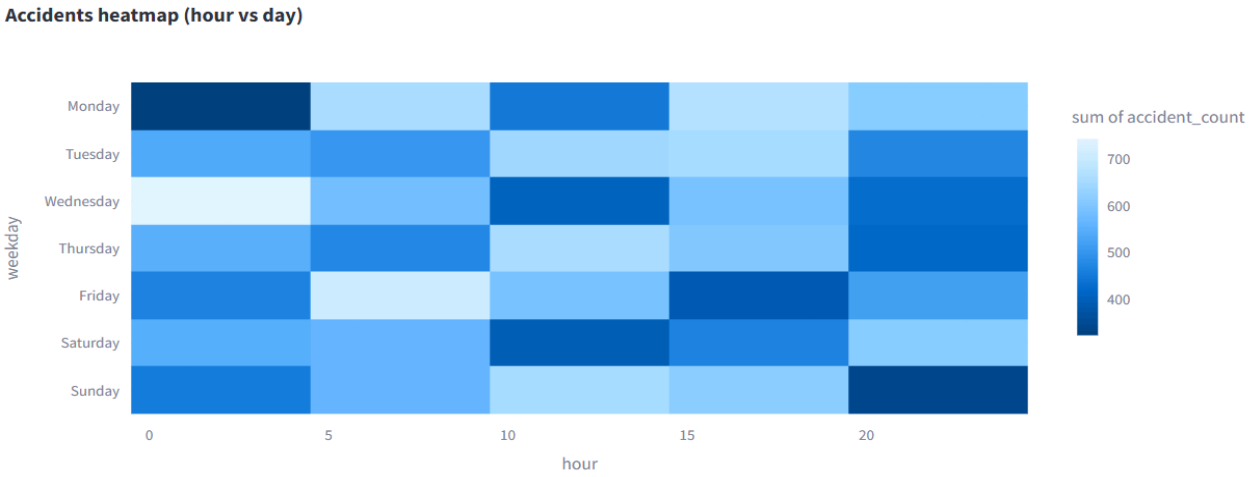
        - Pie charts showing the distribution of **Weather Conditions** and **Road Conditions** (Dry, Wet, Snowy, Damaged).
        - A **Box Plot** visualizing the distribution of average speed across different congestion levels.



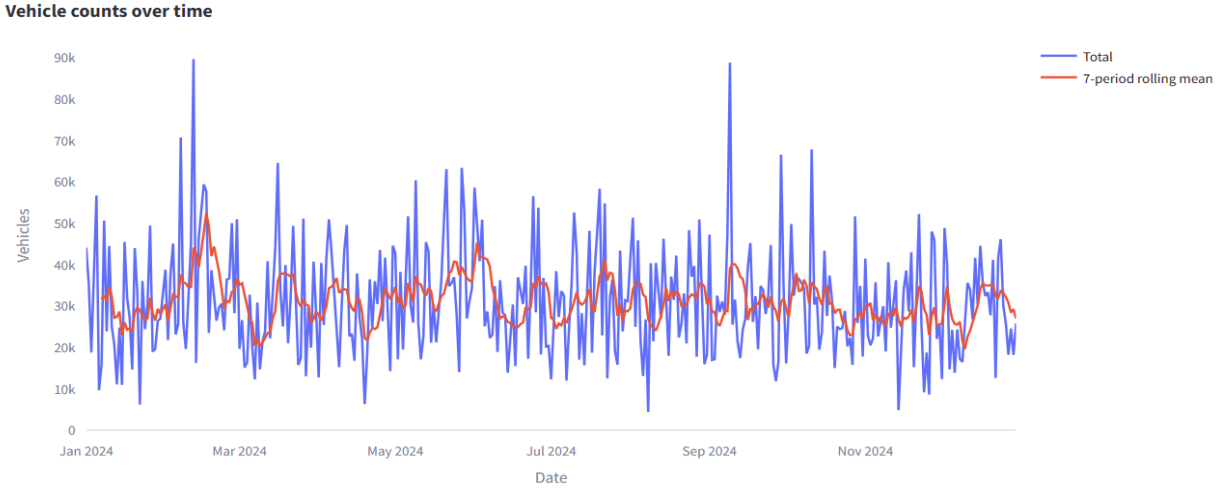
- **Accident Time Series Line Chart:** Tracks the **Accident Count over time** (Daily, Weekly, or Monthly, adjustable in the sidebar), segmented by the **Weather**



- **Detailed:**
  - A **Heatmap** visualizing accidents by hour and day of the week, highlighting peak risk times.

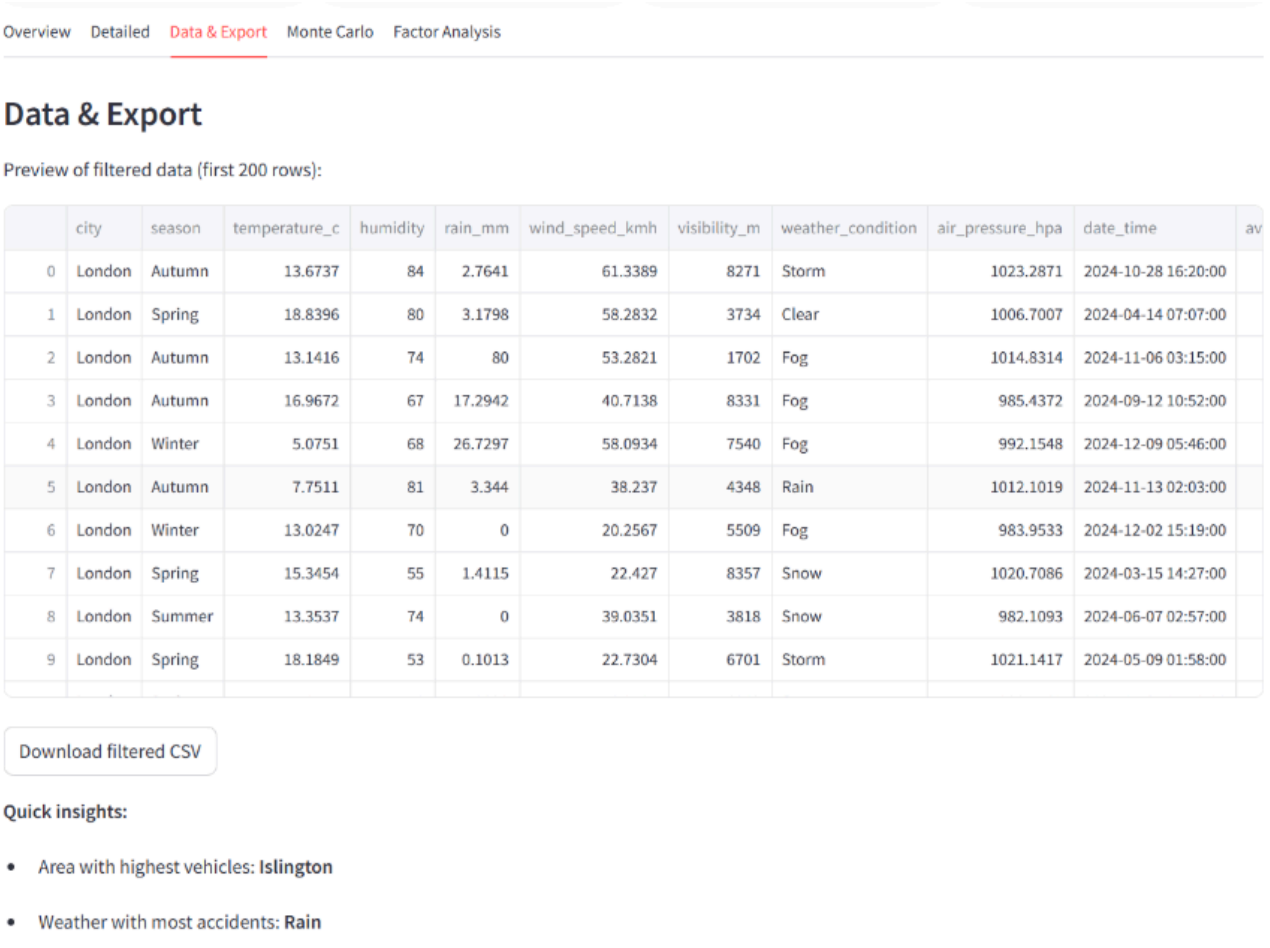


  - A **Line Chart** showing vehicle count trend over time with a rolling mean.



- A **Stacked Bar Chart** visualizing congestion distribution by area.

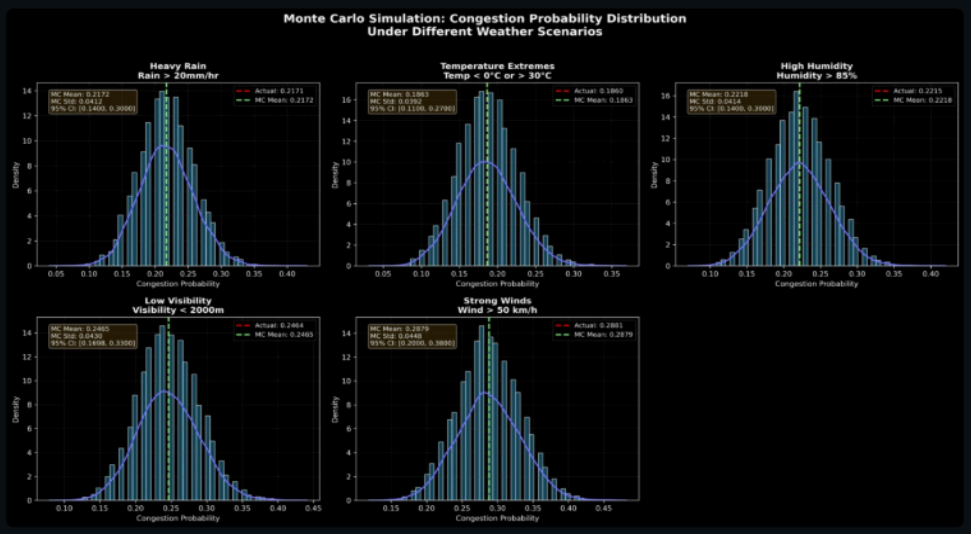
- **Data & Export:** Allows users to preview the filtered data and download it as a CSV file.



	city	season	temperature_c	humidity	rain_mm	wind_speed_kmh	visibility_m	weather_condition	air_pressure_hpa	date_time	av
0	London	Autumn	13.6737	84	2.7641	61.3389	8271	Storm	1023.2871	2024-10-28 16:20:00	
1	London	Spring	18.8396	80	3.1798	58.2832	3734	Clear	1006.7007	2024-04-14 07:07:00	
2	London	Autumn	13.1416	74	80	53.2821	1702	Fog	1014.8314	2024-11-06 03:15:00	
3	London	Autumn	16.9672	67	17.2942	40.7138	8331	Fog	985.4372	2024-09-12 10:52:00	
4	London	Winter	5.0751	68	26.7297	58.0934	7540	Fog	992.1548	2024-12-09 05:46:00	
5	London	Autumn	7.7511	81	3.344	38.237	4348	Rain	1012.1019	2024-11-13 02:03:00	
6	London	Winter	13.0247	70	0	20.2567	5509	Fog	983.9533	2024-12-02 15:19:00	
7	London	Spring	15.3454	55	1.4115	22.427	8357	Snow	1020.7086	2024-03-15 14:27:00	
8	London	Summer	13.3537	74	0	39.0351	3818	Snow	982.1093	2024-06-07 02:57:00	
9	London	Spring	18.1849	53	0.1013	22.7304	6701	Storm	1021.1417	2024-05-09 01:58:00	

- **Monte Carlo:** Dedicated tab to display the final results and visualizations from the simulation.

Monte Carlo Simulation



This is the distribution of congestion level probability extracted from 10,000 samples under the different weather scenarios we have.

Simulation Results (CSV)

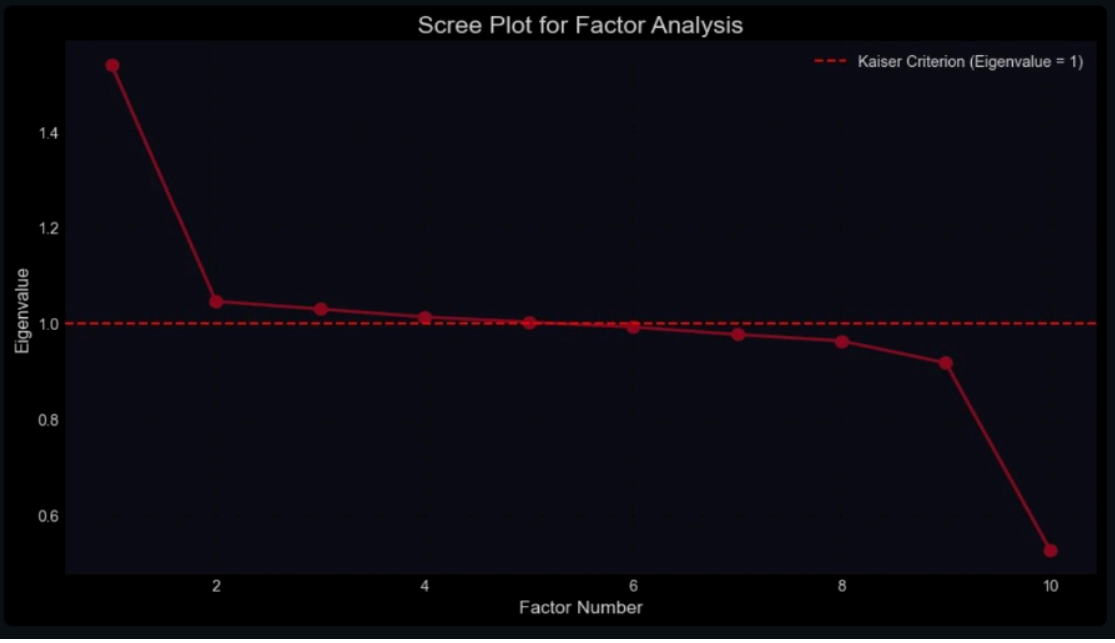
	Scenario	actual_traffic_jam_prob	monte_carlo_traffic_jam_mean	monte_carlo_traffic_jam_std	monte_carlo_traffic_jam_ci_95_lower	mont
0	Heavy Rain	0.2171	0.2172	0.0412	0.14	
1	Temperature Extremes	0.186	0.1863	0.0392	0.11	
2	High Humidity	0.2215	0.2218	0.0414	0.14	
3	Low Visibility	0.2464	0.2465	0.043	0.1698	
4	Strong Winds	0.2881	0.2879	0.0448	0.2	

Each row contains the actual probability of high congestion and accident risk calculated from the original dataset and the predicted probability extracted from the Monte Carlo simulation along with standard deviation and 95% confidence interval

Download Monte Carlo Results (CSV)

- **Factor Analysis:** Dedicated tab to display the final results and visualizations from factor analysis processes.

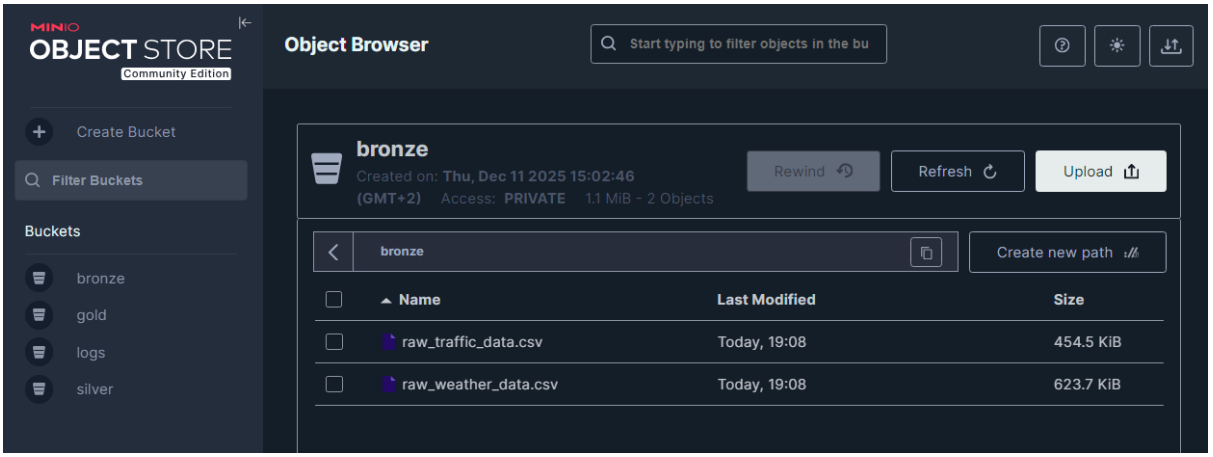
Factor Analysis



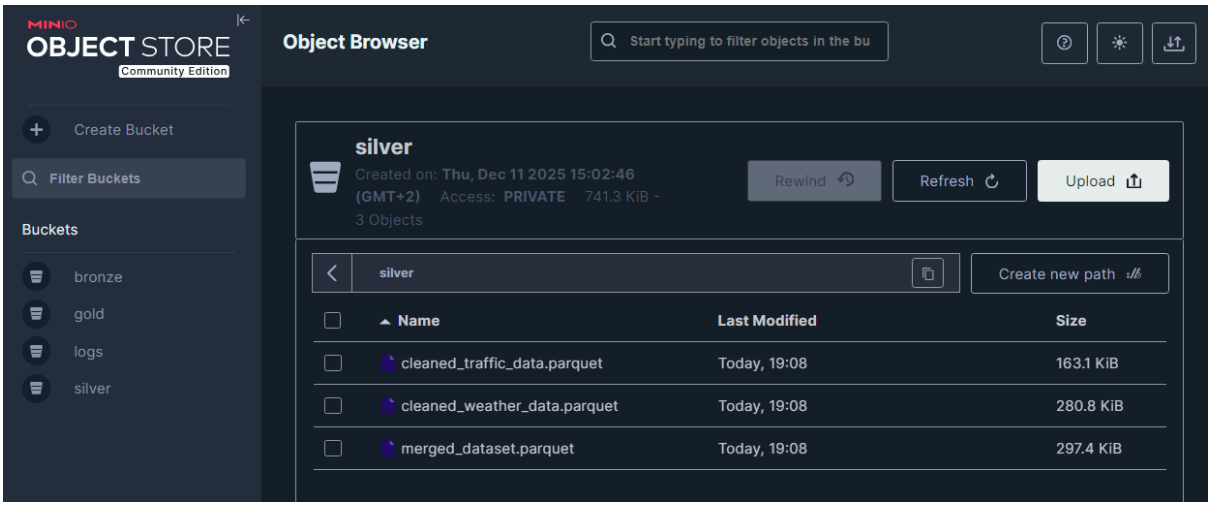
MinIO and HDFS Data Storage

MinIO

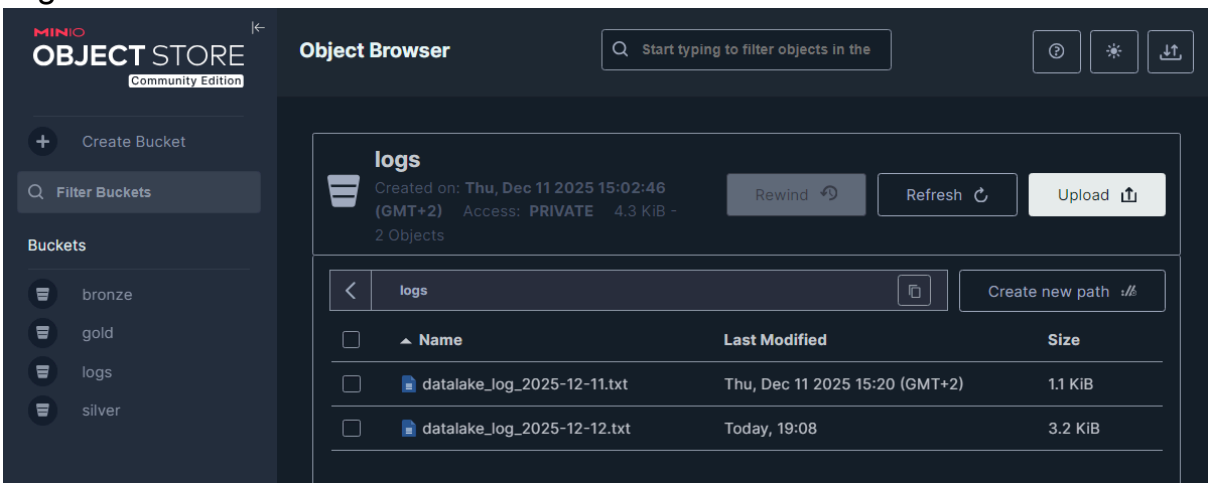
Bronze Bucket



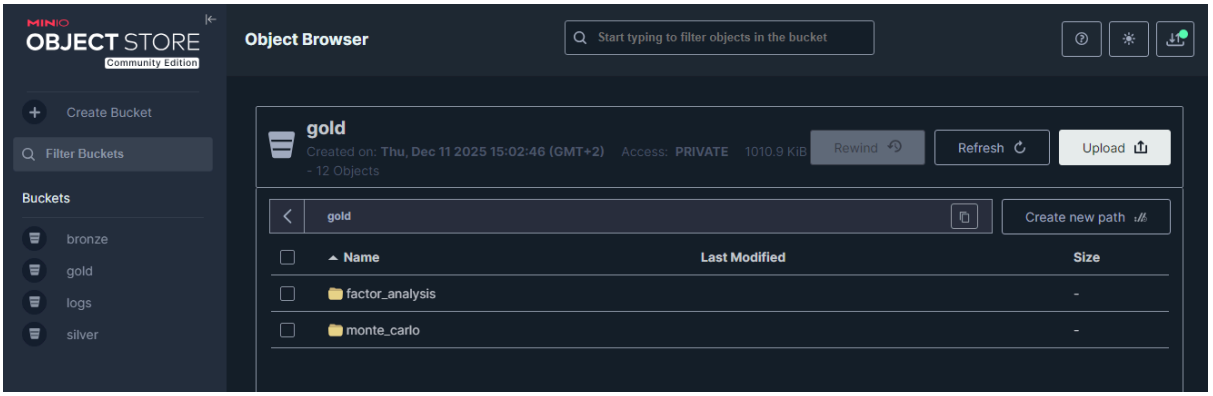
Silver Bucket



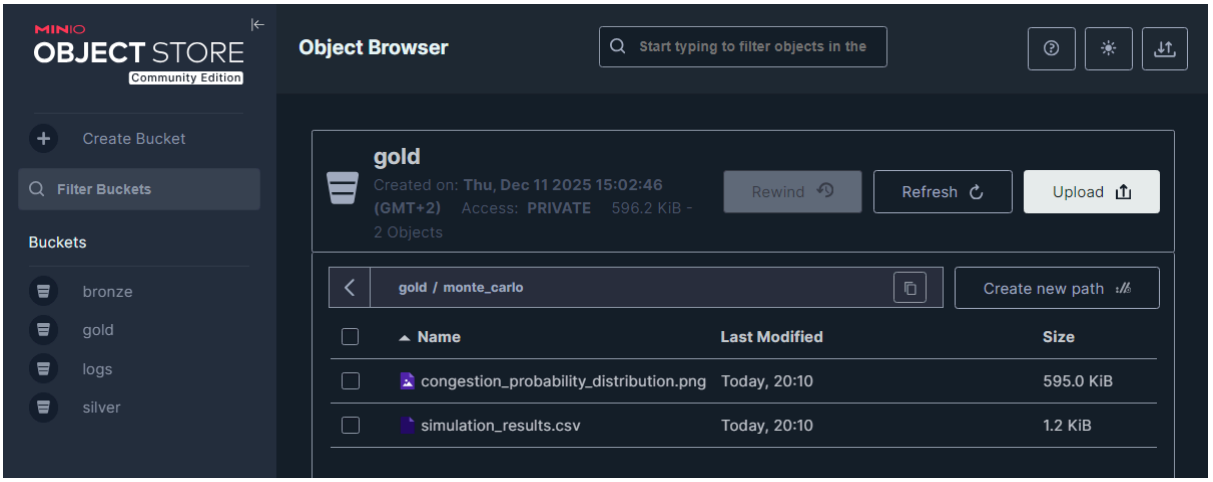
Logs Bucket



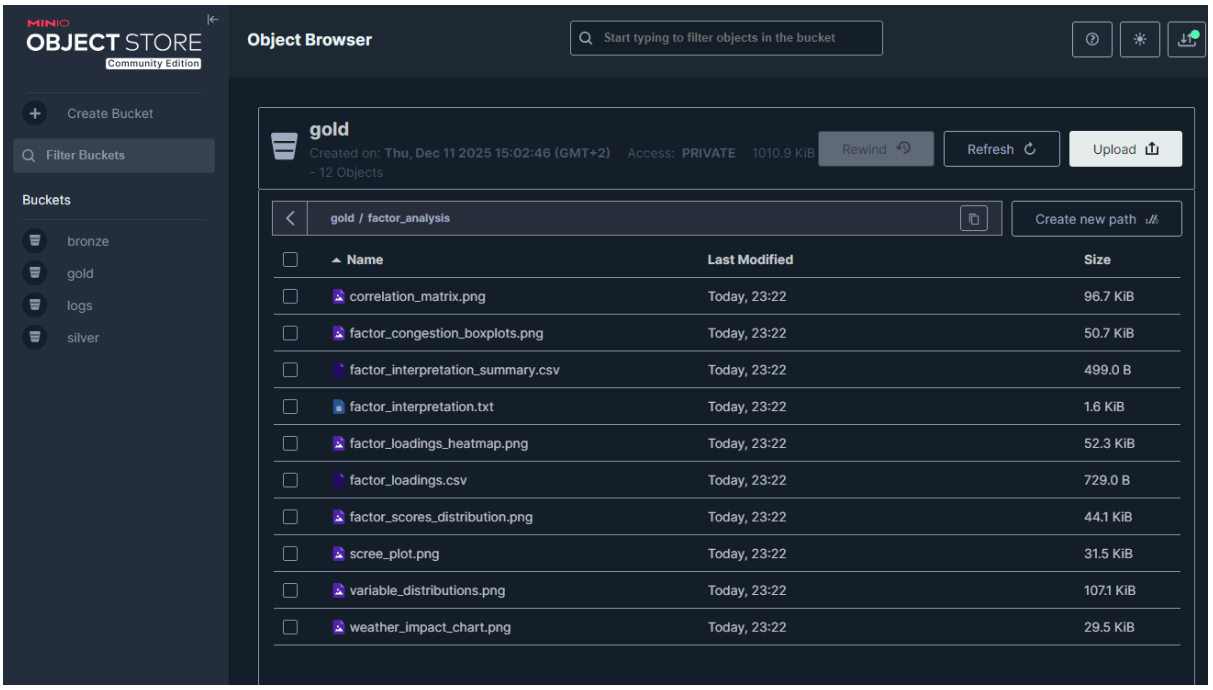
Gold Bucket



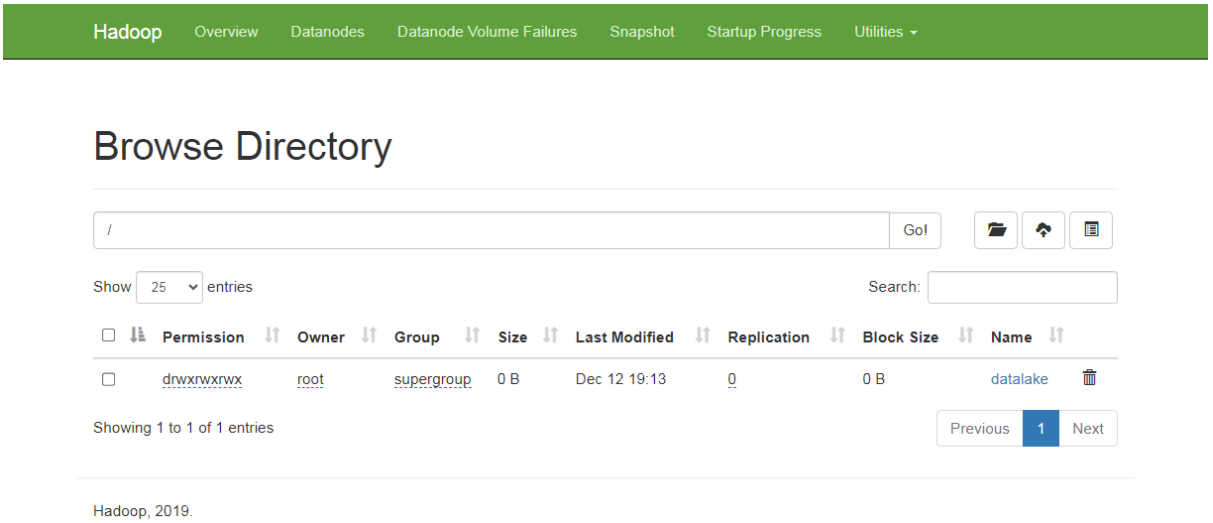
Gold Bucket/monte_carlo



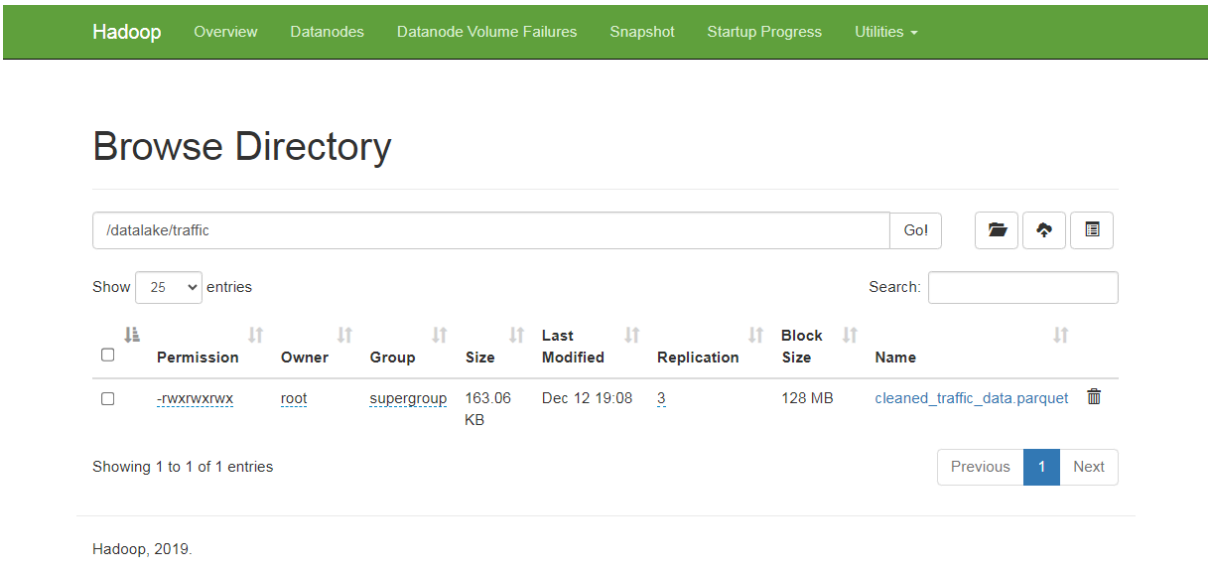
Gold Bucket/factor_analysis



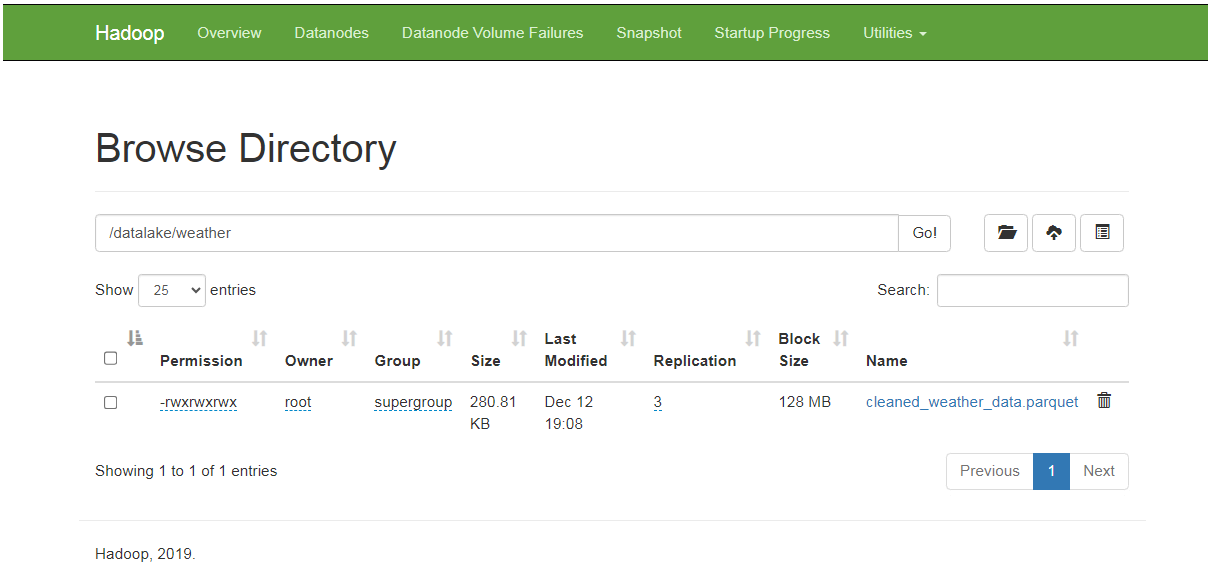
HDFS



Traffic



Weather



Final Insights & Recommendations

Monte Carlo

High congestion levels appear under these weather scenarios by these probabilities:

1. ~22% under heavy rain
2. ~19% under extreme temperatures
3. ~22% under high humidity
4. ~25% under low visibility
5. ~29% under Strong Winds

Factor Analysis

1. RAIN DOMINATES: 74% stronger impact than wind speed
   - 1mm rain increase → 13.2% higher congestion likelihood
2. VISIBILITY MATTERS: Clear days = 7.6% less congestion
   - Better visibility enables normal traffic flow
3. OTHER FACTORS MINIMAL: Temperature, humidity, pressure
   have negligible (<5%) direct impact on congestion

Dashboard Insights

1. The **Stacked Bar Chart** for **Vehicles by Congestion Level per Area** indicates that areas like **Islington** and **Kensington** exhibit the highest total vehicle counts and substantial volumes of "High" congestion incidents, making them priority zones for traffic management optimization.

2. The **Accidents Heatmap (hour vs day)** clearly shows that the highest accident risks are concentrated during **weekday peak commuting hours** (e.g., 7 AM to 9 AM and 4 PM to 6 PM), requiring targeted deployment of resources during these times.

3. The **Box Plot of Avg Speed vs Congestion Level** confirms a steep drop in the median average speed as congestion moves from "Low" to "High," but also shows a wide variance in speed during "Medium" congestion, indicating high uncertainty and volatility in traffic flow at this level.

Individual Contributions

Name	ID	Contribution
عبدالله عاطف خميس عبد الونيس	22010140	Data Generation Data Cleaning
الحسين ياسر إبراهيم السيد	22010056	Monte Carlo Simulation
أحمد عماد عبد الفتاح عبد الغني	22010027	Monte Carlo Simulation
فارس أحمد ابوالفتوح عبدالفضيل	22010182	Factor Analysis
عمر حافظ مأمون محمد حسن	22011562	Data Visualization Dashboard
سيف أيمن أبو اليزيد حسين	22010117	Data Visualization Dashboard
طارق مصطفى محمد عز الدين زين	22010124	MinIO & HDFS Report
أحمد محمد محمود محمود محمدي	22010038	MinIO & HDFS Report