Weather Conditions and Road Traffic Accidents in Milan (2020–2024)

Table of Contents

1. Introduction

Road traffic accidents are a complex urban phenomenon influenced by infrastructural, behavioral, and environmental factors. Among these, weather conditions play a critical role, as they directly affect road surface conditions, visibility, vehicle control, and driver behavior. Understanding how meteorological factors interact with road safety is essential for prevention strategies and policy-making.

The objective of this project is to investigate whether weather conditions have a measurable relationship with the frequency and severity of road traffic accidents in the city of Milan. In particular, the analysis focuses on temperature, rainfall, and snowfall, examining their association with the number of accidents, injured individuals, and fatalities over time.

The central research question guiding this study is whether changes in weather conditions correspond to systematic variations in road accident patterns. To address this question, the project follows a structured data management pipeline that includes data acquisition from heterogeneous sources, data cleaning, transformation, integration, and exploratory analysis.

2. Data Sources

The project relies on two independent data sources that differ in origin, structure, and granularity, but share a common temporal dimension. The use of multiple heterogeneous data sources is a deliberate choice, as the research question cannot be addressed using a single dataset. Weather data provide information on environmental conditions, while traffic accident data capture road safety outcomes. Only by combining these sources is it possible to study their relationship.

2.1 Web API – Weather Data

Weather data were collected using the Meteostat API, accessed through RapidAPI. This API provides historical meteorological observations at a daily resolution, allowing fine-grained access to weather conditions over time. The retrieved variables include daily average temperature, daily precipitation, and daily snowfall for the city of Milan.

Data were collected programmatically using Python scripts, ensuring reproducibility of the data acquisition process. API responses were returned in JSON format and subsequently parsed and stored in a structured tabular form suitable for further processing.

The selected time range spans from January 2020 to December 2024, which was chosen to ensure full temporal alignment with the available road traffic accident data. This temporal consistency is essential for later stages of data integration and analysis.

2.2 Open Data – Road Traffic Accidents

Road traffic accident data were obtained from the official open data portal of the Municipality of Milan (dati.comune.milano.it). This dataset provides monthly aggregated statistics on road accidents occurring within the city of Milan.

The accident data include information on the number of accidents, the number of injured individuals, and the number of fatalities, as well as the nature of each accident. The data are published by an official institutional source, ensuring reliability and consistency.

Although the accident data are provided at a monthly level, they are particularly suitable for this project because they allow direct alignment with weather data once meteorological observations are aggregated to the same temporal granularity.

2.3 Traffic Density (Area C Vehicle entries)

Although traffic density data are not directly integrated into the main dataset, they are used as an external explanatory variable in later analytical stages.

To complement the accident and weather datasets, traffic density data were obtained from the official open data portal of the Municipality of Milan (dati.comune.milano.it). The dataset provides monthly counts of vehicles entering the Area C zone, a restricted central traffic area, and serves as an indicator of urban mobility levels within the city.

The data were downloaded directly from the municipality's website and cover the year 2023, ensuring temporal consistency with the accident and weather observations used in the analysis. Vehicle entries are reported as absolute monthly counts and were extracted in their raw form without modification at this stage. These data were subsequently used in later analytical phases to examine the interaction between traffic volume, weather conditions, and accident frequency.

3. Data Exploration

Before integrating the datasets, an extensive exploratory phase was conducted to understand their structure, limitations, and compatibility.

3.1 Data Cleaning

A major challenge during the cleaning phase was the difference in temporal granularity between the two datasets. Weather data were available at a daily level, while accident data were provided monthly. To address this mismatch, daily weather observations were aggregated into monthly indicators.

For each month, average temperature was computed by averaging daily values, while total rainfall and total snowfall were obtained by summing daily precipitation and snowfall. Additionally, the number of rainy days and snowy days per month were calculated.

3.2 Data Quality

In the initial exploratory phase, weather conditions were classified into dry, rainy, and snowy categories. However, data exploration revealed that dry months were extremely rare in Milan during the selected time period, resulting in strong class imbalance. This made the initial classification statistically unreliable.

As a result, the methodology was revised to adopt an intensity-based approach, focusing on quantitative rainfall measures rather than qualitative labels.

4. Data Integration

The data integration phase represents a central step in this project, as the core research question cannot be answered using either dataset in isolation. Weather data alone do not provide information about road safety outcomes, and accident data alone do not explain the environmental conditions under which accidents occur. Therefore, integration is necessary to enable joint analysis of meteorological conditions and traffic accident patterns.

The primary purpose of data integration in this project is to align weather conditions and accident outcomes within the same temporal context, allowing us to observe how changes in weather correspond to variations in accident frequency and severity.

4.1 Schema Transformation

A major challenge in the integration process was the difference in temporal granularity between the two datasets. Weather observations were available at a daily level, while accident data were provided as monthly aggregates. Direct integration was therefore not possible without transformation.

To resolve this issue, daily weather data were aggregated into monthly indicators. This transformation was not only a technical requirement but also a methodological choice. Monthly aggregation reduces short-term noise in daily weather fluctuations and produces indicators that are more comparable to monthly accident statistics. As a result, both datasets were brought to the same temporal scale, ensuring meaningful comparison.

4.2 Correspondences Investigation

Before merging the datasets, a correspondence investigation was conducted to identify valid and reliable integration keys. Several potential dimensions were considered, including spatial and temporal attributes.

Since both datasets exclusively refer to the city of Milan, spatial matching was unnecessary. The only meaningful and consistent correspondence between the datasets was time, represented by year and month. These variables were therefore selected as the integration key and combined into a unified temporal identifier in the format YYYY-MM.

This decision ensures that each accident observation is associated with the correct weather conditions occurring during the same month.

4.3 Data Preparation

This step also reduces the risk of overfitting the analysis to short-term meteorological fluctuations.

Prior to integration, additional preparation steps were performed to improve robustness and interpretability. Raw daily weather fields with missing values were excluded from the integration process. Instead, derived monthly indicators such as total rainfall, number of rainy days, and snowfall measures were used.

This choice was motivated by data quality considerations: aggregated indicators are more stable, less sensitive to missing values, and better suited for explaining monthly accident outcomes.

In addition to weather data preparation, traffic density data from the Area C dataset were processed to ensure temporal compatibility with the integrated dataset. Raw vehicle entry records were aggregated to a monthly level by standardizing date-time fields and computing total monthly vehicle counts. This preprocessing step aligned the traffic data with the monthly granularity adopted throughout the project, enabling consistent comparison with accident and weather indicators in subsequent analyses.

4.4 Schemas Integration and Methodological Refinement

The integration was performed using a temporal join on the monthly identifier. Each accident record was matched with the corresponding monthly weather indicators.

At this stage, an important methodological refinement was introduced. In the initial exploratory phase, weather conditions had been categorized qualitatively into dry, rainy, and snowy months. However, data exploration revealed that dry months were extremely rare in Milan, resulting in severe class imbalance.

To address this limitation, the integrated dataset adopted a quantitative, intensity-based representation of weather conditions, particularly for rainfall. Instead of relying on categorical labels, rainfall was modeled using total monthly precipitation and the number of rainy days. Months were subsequently classified into low, medium, and high rainfall intensity levels.

This integration strategy serves two purposes. First, it preserves more information from the original weather data. Second, it enables more reliable statistical analysis by avoiding underrepresented categories. As a result, the integrated dataset provides a robust foundation for analyzing how weather intensity influences traffic accidents.

5. Database

5.1 Database Construction and Storage

All processed datasets were stored in a SQLite relational database to ensure reproducibility. Statistical summaries were computed using SQL queries, revealing higher accident and injury counts during months with higher rainfall intensity.

5.2 Statistical Analysis on the Integrated Dataset

Following data integration, schema alignment, and storage of the cleaned datasets in a relational database, a comprehensive statistical analysis was conducted to quantify the relationship between meteorological conditions and road traffic accidents in Milan. All analyses presented in this section are based on the final monthly integrated dataset described in the previous section.

5.2.1 Milan Traffic Accident Data Analysis (2020–2024)

The provided charts summarize traffic accidents in Milan over the last five years, focusing on numerical distribution, accident severity, and causes of death. All analysis rates are calculated based on the total number of accidents (SUM(Incidenti)).

5.2.1.1 Accident Trends and Pandemic Impact

The total number of accidents (SUM(Incidenti)) stood at approximately 5,300 in 2020.

Starting from 2021, there was a significant increase, with numbers rising to and stabilizing within the 7,200 – 7,600 range.

This sharp increase reflects the return to normal urban mobility levels following the lifting of pandemic-related restrictions.

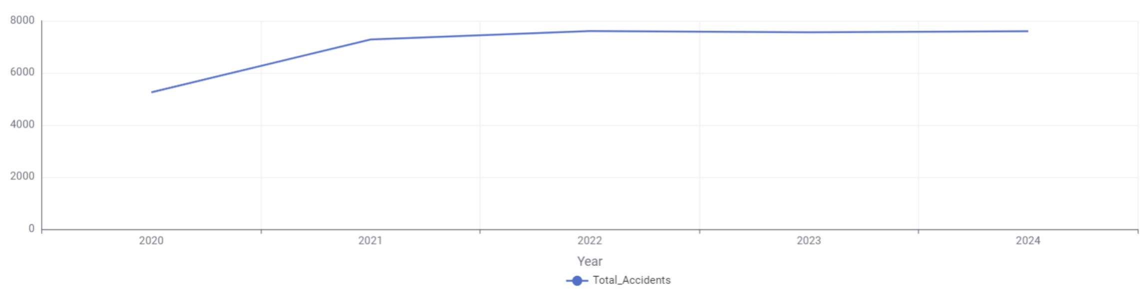

Figure 1: Annual Trend of Road Traffic Accidents in Milan (2020–2024)

5.2.1.2 Injury Intensity and Calculation Methodology

In this analysis, the "Injury Rate" is derived by dividing the total number of injuries by the total number of accidents:

$$\text{Injury Rate Percent} = \left(\frac{\sum \text{Feriti}}{\sum \text{Incidenti}} \right) \times 100$$

This rate is consistently between 123% and 126% throughout the analyzed period.

Since the value exceeds 100%, it technically proves that, on average, more than one person (approximately 1.25 people) is injured per accident.

This indicates that accidents in Milan are high-impact events that typically affect multiple individuals rather than being isolated to a single person.

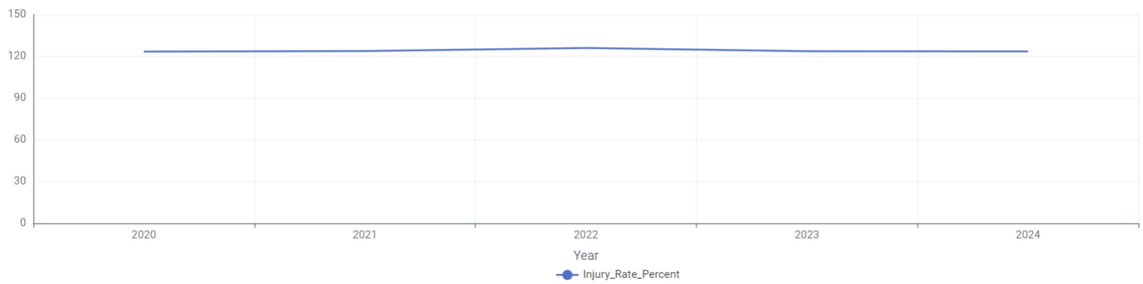

Figure 2: Annual Injury Rate Percentage (Injuries per Accident).

- ***5.2.1.3 Fatality Rate Analysis***

The fatality rate is calculated as the ratio of total deaths to the total number of accidents:

$$\text{Fatality Rate Percent} = \left(\frac{\sum \text{Morti}}{\sum \text{Incidenti}} \right) \times 100$$

While total accident numbers remained stable after 2021, the fatality rate showed more fluctuation.

The lowest fatality rate was recorded in 2021 at 0.43%.

The rate peaked in 2023 at 0.57%, suggesting that accidents in that specific year resulted in more lethal outcomes compared to others.

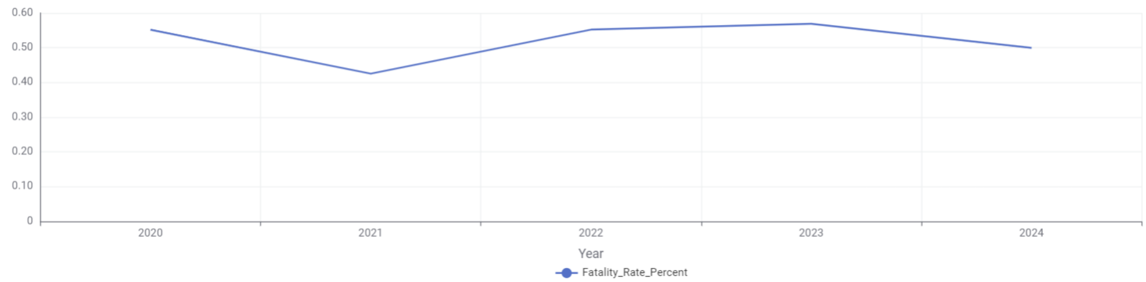

Figure 3: Annual Fatality Rate Percentage (Deaths per Accident).

5.2.1.4 Primary Causes of Fatalities

An examination of "Accident Types of Deaths" reveals the following:

"Pedestrian collision" is the leading cause of death, accounting for 67 fatalities.

This is followed by "Frontal-lateral collision", which accounts for 38 fatalities when combined.

Other significant causes include "Collision with obstacle" (28 deaths) and "Running off the road / Skidding" (17 deaths).

5.2.2 Correlation Analysis

5.2.2.1 Milan Weather VS Traffic Accidents

An initial correlation matrix (Heatmap) was generated to test for linear relationships between meteorological variables and accident counts.

Rainfall and Accidents: The Pearson correlation coefficient between total_rain and Incidenti (Accidents) was found to be 0.04. Although positive, this low value indicates that on a monthly scale, total rainfall volume alone does not have a strong linear relationship with the total number of accidents.

Rainy Days: The correlation for rain_days and accidents was slightly higher at 0.06. This suggests that the frequency of rainfall (how many days it rains) has a marginally greater impact on accident recurrence than the total volume of rain.

Injuries vs. Accidents: A correlation of 0.99 was observed between accidents and injured persons (Feriti), indicating a nearly perfect linear relationship; as accidents increase, injuries rise proportionally.

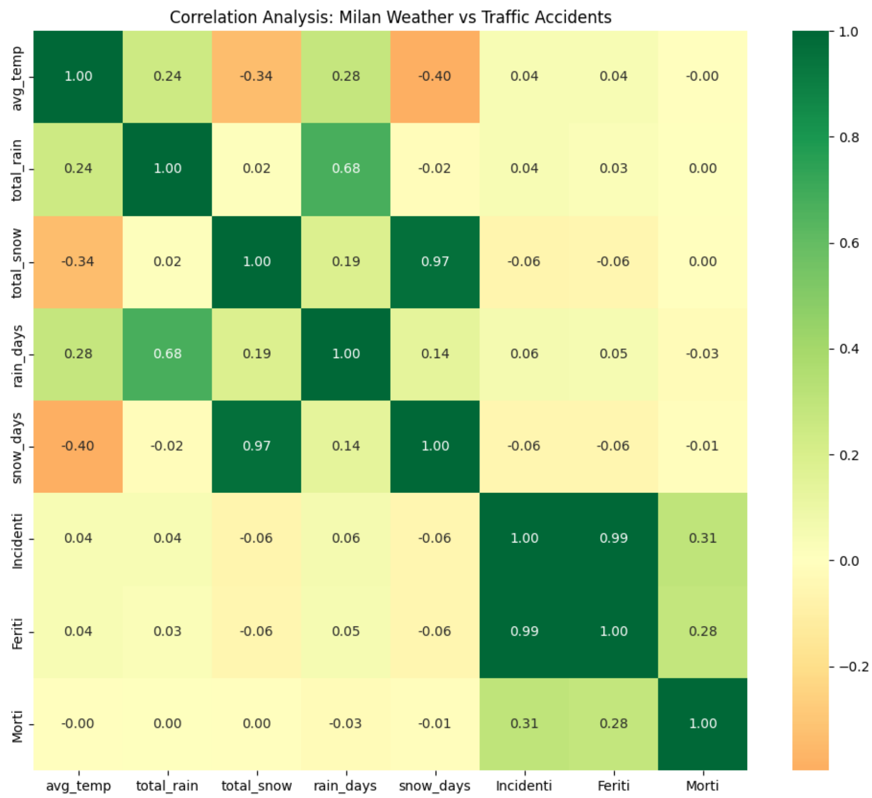

Figure 4: Correlation Matrix Heatmap: Weather Variables vs. Accident Statistics.

5.2.2.2 Correlation Between Rainfall and Traffic Accidents in Milan

This chart illustrates the relationship between total rainfall (mm) and the number of accidents in Milan, while also highlighting the intensity of fatalities within those incidents. Each data point represents a specific time period.

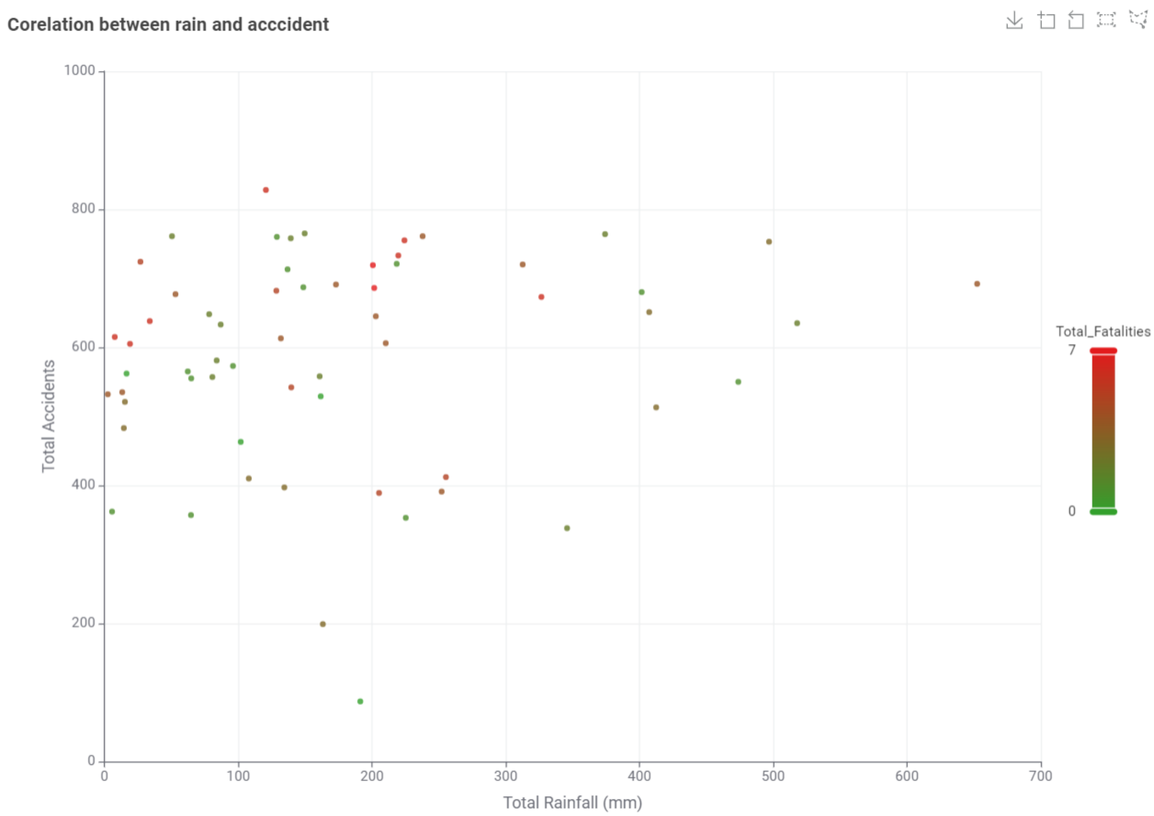

Figure 5: Scatter Plot: Correlation Between Total Monthly Rainfall and Total Accidents (Colored by Fatalities).

1. Correlation Between Precipitation and Accident Volume

Wide Statistical Dispersion: A primary observation is the wide spread of data points across the graph.

Non-Exclusive Impact: This indicates that high accident volumes (frequently reaching the 600–800 range) occur during periods of low rainfall or dry conditions as well as during rainy periods. Accidents are not exclusively triggered by extreme weather.

2. Rainfall Distribution and Density

Data Clustering: The vast majority of observations are clustered within the 0 to 250 mm rainfall range.

Stabilization at Extremes: When rainfall exceeds 400 mm, there is no exponential "explosion" in the number of accidents; instead, the figures remain within similar ranges seen at lower precipitation levels.

3. Fatality Intensity (Color Gradient Analysis)

The color of each point represents the total fatalities (Total_Fatalities), ranging from Green (0 deaths) to Red (7 deaths):

High-Risk Zones: Notably, the red and orange points—representing the highest fatality counts—are concentrated during "moderate" rainfall periods (between 100–250 mm).

Volume Threshold: These fatal incidents most frequently occur when the total number of accidents exceeds 600.

5.2.2.3 Correlation Between Weather Conditions and Traffic Accidents

This analysis examines the impact of different weather conditions on traffic safety using two primary datasets: Distribution of Days and Total Number of Accidents by Weather Condition.

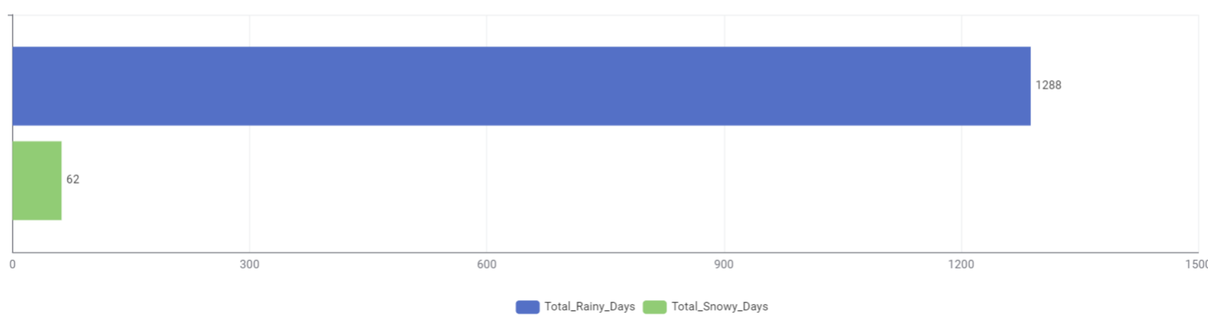

Figure 6: Frequency Distribution of Rainy Days vs. Snowy Days (2020-2024).

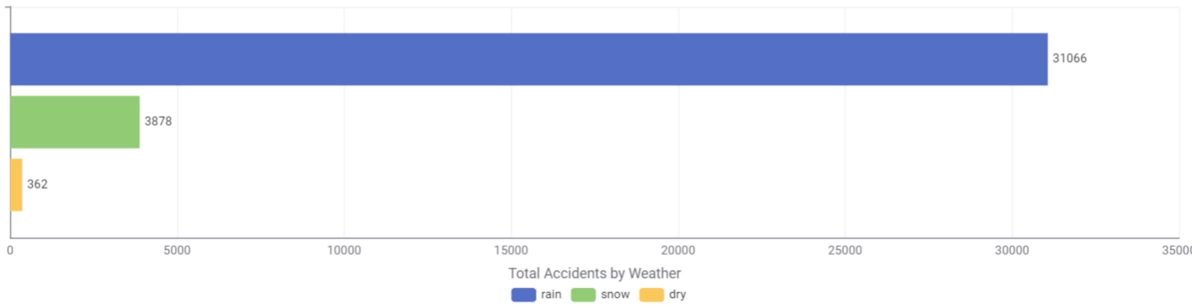

Figure 7: Total Number of Accidents Categorized by Weather Condition.1. Data Overview and Distribution

The first chart illustrates the frequency of specific weather conditions during the observed period. The data shows that there were 1,288 rainy days, whereas only 62 snowy days were recorded. This suggests that the region is either geographically prone to rain or that the data spans a period where rain is the predominant form of precipitation.

2. Accident Volume vs. Weather Type

The second chart displays the total number of accidents categorized by weather:

Rainy Weather: 31,066 accidents

Snowy Weather: 3,878 accidents

Dry Weather: 362 accidents

At first glance, rainy weather appears to be the most dangerous due to the high volume of accidents. However, this is largely a result of the much higher frequency of rainy days (1,288) compared to other conditions.

3. Statistical Insight: Risk Per Day

To understand the actual risk level, we must calculate the average number of accidents per day for each weather type:

Accidents per Rainy Day: Approximately 24.1

Accidents per Snowy Day: Approximately 62.5

So the most significant finding from this data is that while rainy weather results in a higher total number of accidents, snowy weather increases the individual risk of an accident by approximately 2.5 times. The high accident rate during the relatively few snowy days indicates that snow is a much more hazardous factor for road safety, likely due to reduced traction and visibility. This highlights the need for specialized traffic management and driver caution during winter conditions.

5.2.3 Categorization and Trends

5.2.3.1 Rainfall Intensity Categorization and Accident Distribution

This section examines accidents occurring on rainy days in Milan, categorized into three primary groups based on the statistical distribution of total rainfall. The methodology used for grouping and the subsequent findings are detailed below:

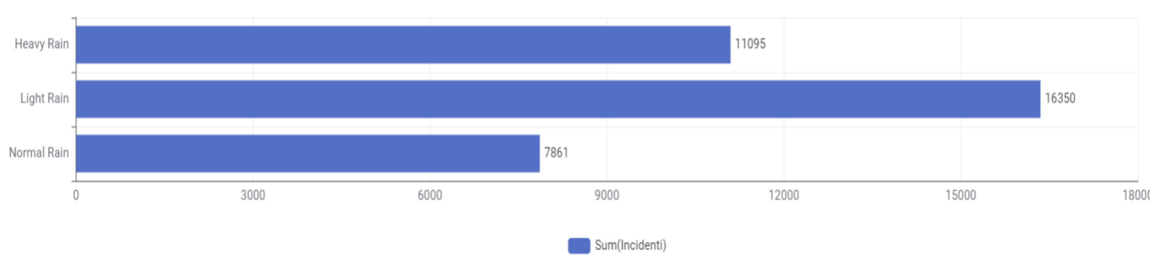

Figure 8: Distribution of Accidents by Rainfall Intensity Category (Light, Normal, Heavy)

Methodology and Grouping Logic

To classify the intensity of rainfall objectively, the following statistical threshold values were determined based on total rain by month:

Light Rain: total rain <136.8 mm

Normal Rain: 136.8 < total rain< 212.25 mm

Heavy Rain: total rain > 212.25 mm

2. Accident Distribution by Category

Based on this statistical grouping, the distribution of accidents according to rainfall intensity is as follows:

- Light Rain :This category leads the list with a total of 16,350 accidents. This can be explained by the fact that rainfall in Milan frequently remains within this statistical range, and drivers may be more likely to neglect safety measures (such as reducing speed or maintaining following distance) during "low" rainfall.
- Heavy Rain: This group, representing the highest precipitation levels, recorded 11,095 accidents. The lower accident count compared to light rain, despite more challenging weather, may result from reduced traffic volume during extreme downpours or a higher "risk perception" leading to more cautious driving.
- Normal Rain :A total of 7,861 accidents occurred within this range.

3. General Evaluation and Risk Analysis

The data proves that accident numbers differ not only by the presence of rain but also by its statistical intensity:

Frequency Factor: Despite a total of 1,288 rainy days recorded over the 2020–2024 period, the clustering of a significant portion of accidents in the "Light Rain" category indicates that this is the most common condition, thus driving the total accident volume.

Accident Severity and Fatalities: The previously established injury rate of 123%–126% and the high fatality risk for pedestrians (67 deaths) become critically important during these "Light Rain" periods, where driver vigilance may be lower.

Statistically, traffic accidents in Milan are most concentrated in the "Light Rain" category. This reveals that road safety initiatives should focus not only on extreme weather events but also on frequent, light-rain days where accident frequency remains high.

5.2.3.2 Categorization of Fatality Rates Based on The Cause of Accidents

This chart illustrates that the majority of fatal accidents in Milan result from pedestrian collisions or striking fixed objects, rather than vehicle-to-vehicle impacts. In the following sections of our study, we will analyze the weather conditions (rainy or sunny) associated with these incidents to uncover the underlying causes behind these fatalities.

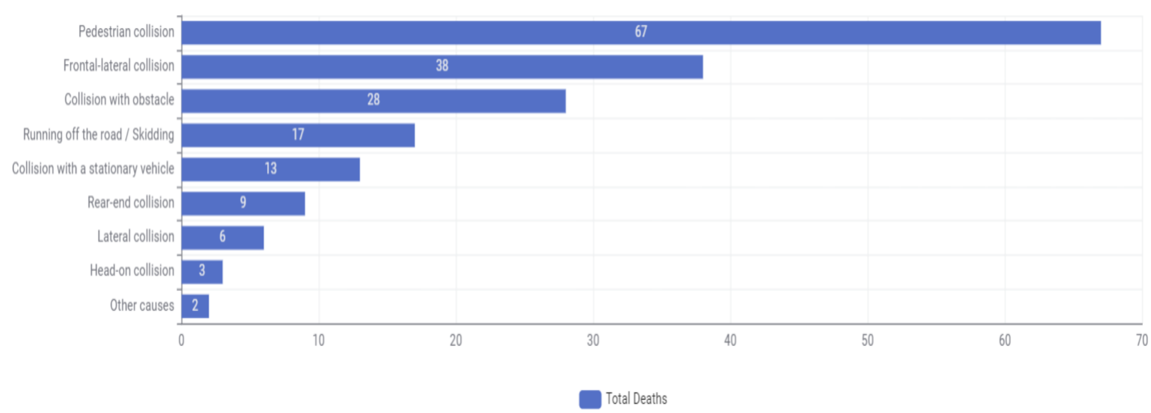

Figure 9: Total Fatalities by Accident Type (Cause of Death).

5.2.3.3 Seasonal Trends and Outliers

Seasonal Peak: The analysis identified a distinct seasonal pattern, with a sharp peak in road accidents during Autumn (averaging ~78 accidents per month). In contrast, winter months showed the lowest average (<60 accidents).

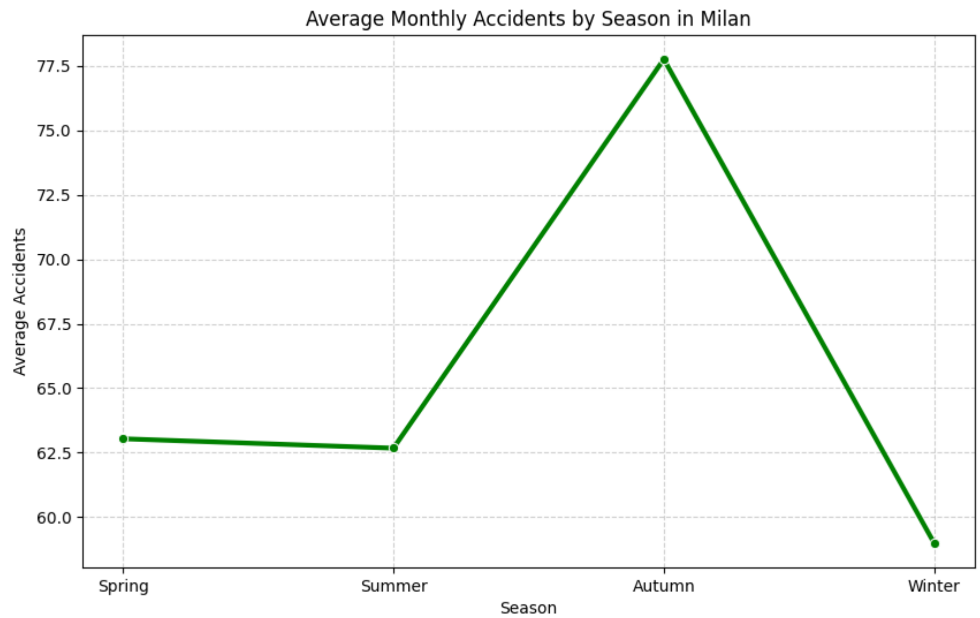

Figure 10: Average Monthly Accidents by Season (Seasonal Trend).

- Outlier Detection: Boxplot analysis highlighted specific critical months. The most extreme outlier was October 2021, which recorded 311 accidents and 120mm of rainfall. The alignment of accident outliers with periods of heavy rainfall provides visual evidence supporting the hypothesis.

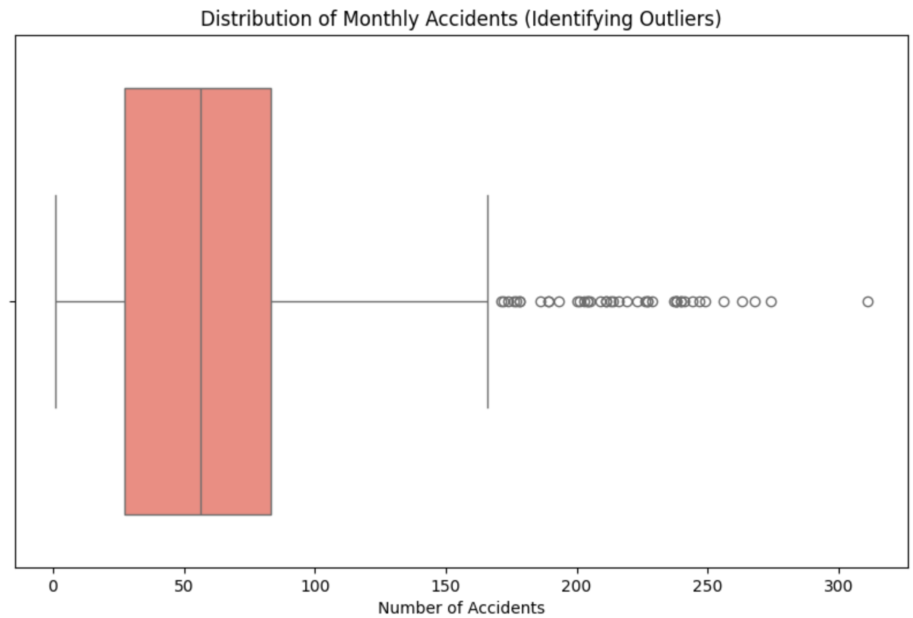

Figure 11: Boxplot of Monthly Accidents Identifying Outliers and Extreme Events.

5.2.4 Time Series and Regression Modeling

- A time series analysis from 2020 to 2024 revealed significant fluctuations, including a notable spike in accidents in late 2024 that coincided with peak rainfall. To quantify this, a Linear Regression model was applied.

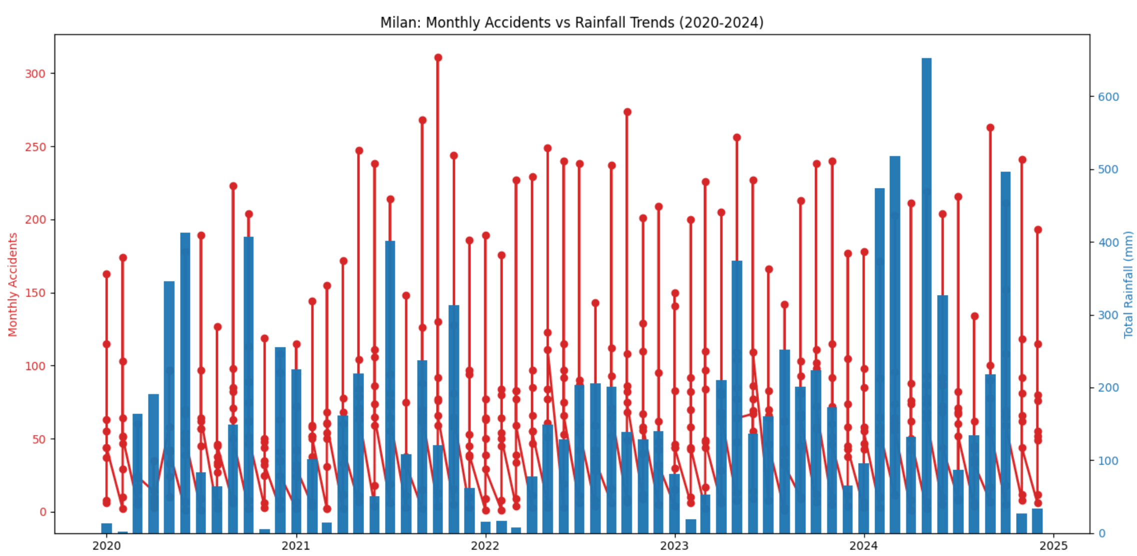

Figure 12: Time Series Analysis: Monthly Accidents (Bar) vs. Total Rainfall (Line) (2020–2024)

- Model Parameters: The model yielded an Intercept of 62.66 and a Slope of 0.0167.
- Interpretation: The baseline is approximately 63 accidents per month under dry conditions. The positive slope indicates that for every 100mm of additional rainfall, the monthly accident count increases by approximately 1.6 accidents.

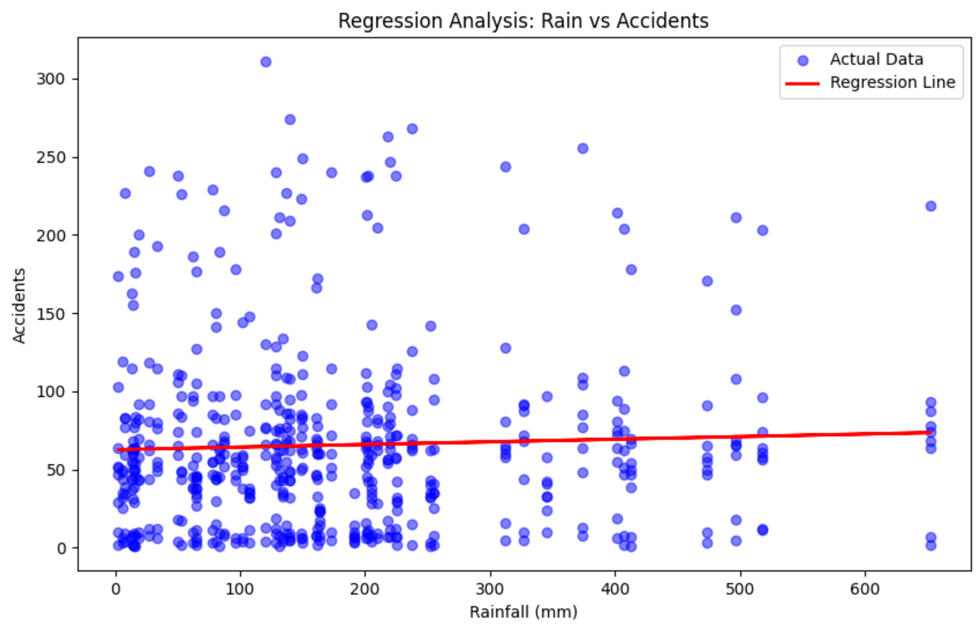

Figure 13: Linear Regression Model: Monthly Rainfall vs. Number of Accidents.

- Non-Linearity: Attempts to fit non-linear (polynomial) models showed that due to high data variance, a linear trend remained the most robust description, though the relationship is likely influenced by saturation effects (where extreme rain reduces traffic flow).

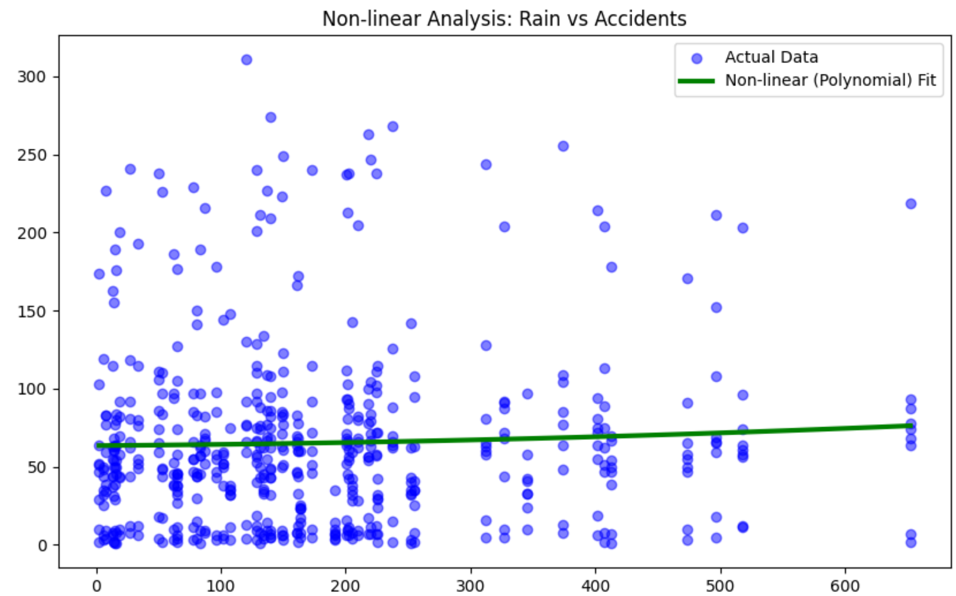

Figure 14: Non-linear (Polynomial) Regression Analysis: Rainfall vs. Accidents.

The regression analysis produced an R-squared value of 0.0017. The extremely low R-squared value indicates that rainfall volume alone is not a strong linear predictor for accident counts, reinforcing the complexity of the phenomenon and the presence of other dominant factors like traffic density.

5.2.5 Traffic Density As a Confounding Factor

Combined Impact of Traffic Density and Weather on Accident Volume (2023)

This analysis compares the number of vehicles entering the "Area C" zone (Traffic Density) with the monthly accident counts across the city, based on Milan municipality data. While Area C represents a restricted central district, its vehicle volume serves as a significant indicator of Milan's overall urban mobility.

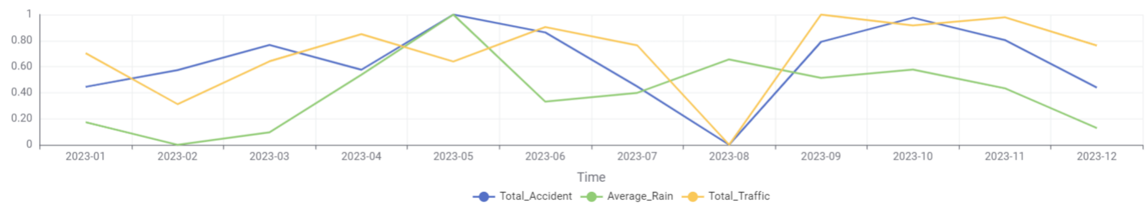

Figure 15: Normalized Trends: Comparison of Total Accidents, Average Rainfall, and Area C Traffic Density (2023).

1. Traffic Density: The Primary Determinant

Visual Correlation: The Total Accident (blue) and Total Traffic (yellow/orange) trends exhibit near-perfect parallelism throughout the year.

The August Trend: In August, both traffic volume and accident counts drop to their lowest points simultaneously, suggesting that accident frequency is primarily driven by vehicle mobility.

Statistical Generalization: The data indicates that traffic density remains the most consistent predictor of accident frequency citywide.

2. The Role of Rain: Secondary Trigger

Inconsistencies: The Average Rain (green) data does not always align with accident trends.

Comparative Analysis: While rain and accidents both peak in May, the period between September and October shows a sharp increase in accidents despite relatively low rainfall.

The Trigger Effect: This confirms that while rainfall influences safety, traffic volume is the foundational factor. Rain acts as a "trigger" that increases risk levels within existing high-traffic periods.

3. Holistic View with Statistical Grouping

By integrating the rainfall intensity categories established earlier we reach the following conclusions:

The majority of accidents occur during Light Rain because traffic density remains high and driver vigilance typically decreases.

The 2023 data demonstrates that regardless of weather, higher vehicle counts on the road lead to a proportional increase in the number of accidents.

6. Conclusion and Future Developments

This study investigated the relationship between weather conditions and road traffic accidents in Milan over the period 2020–2024, with a particular focus on rainfall, snowfall, and accident severity. By integrating heterogeneous datasets from meteorological and traffic accident sources into a unified monthly framework, the analysis was able to assess whether variations in weather conditions correspond to systematic changes in accident patterns.

The results indicate that weather conditions do influence road traffic accidents, but not in a simple linear or deterministic way. Rainfall does not act as the sole driver of accident occurrence; rather, it functions as a contextual risk factor that interacts with underlying traffic density. While high accident volumes are observed across both dry and rainy periods, statistical analyses show that accident frequency tends to increase during months characterized by higher rainfall intensity, especially when traffic volumes remain elevated. This confirms that weather conditions contribute to accident risk, but primarily as a catalyst rather than an independent cause.

A key methodological insight of this project is that rainfall intensity provides more explanatory power than total rainfall volume alone. Intensity-based indicators, such as the number of rainy days and categorized rainfall levels, capture variations in road conditions and driver behavior more effectively than aggregate precipitation measures. Moderate rainfall periods, in particular, are associated with higher fatality risk, possibly due to reduced driver vigilance compared to extreme weather events, where traffic volume and speed tend to decrease.

Snowy conditions present a different dynamic. Although the total number of accidents during snowy periods is relatively low, the average number of accidents per snowy day is significantly higher than during rainy days. This suggests that snow increases per-day accident risk due to reduced traction and visibility, while simultaneously discouraging overall mobility. As a result, snow leads to fewer total accidents but higher individual risk when driving occurs.

Despite these findings, the analysis is subject to several limitations. First, the use of monthly aggregation masks short-term dynamics that may occur at daily or hourly levels, particularly during extreme weather events. Second, the absence of spatial granularity prevents differentiation between high-risk road segments and safer areas within the city. Finally, traffic density was only approximated using proxy indicators, limiting the precision of multivariate interaction effects.

Future developments should focus on integrating daily or hourly accident data, incorporating spatial dimensions such as road type or district-level analysis, and applying multivariate statistical models that explicitly account for traffic volume, weather conditions, and temporal effects simultaneously. Such extensions would enable a more comprehensive understanding of how environmental and mobility factors jointly shape road safety outcomes in urban environments.

To ensure transparency and reproducibility of the data management pipeline, the full source code used for data collection, preprocessing, integration, and analysis is available in a public GitHub repository:

https://github.com/MParpenchi/DM_Project/tree/main