**A**

# MINI PROJECT REPORT ON

## Weather Analysis Using Python

**FOR**

Term Work Examination

***Bachelor of Computer Application in Artificial Intelligence and Machine Learning (BCA - AIML)***

**Year 2024-2025**

**Ajeenkya DY Patil University, Pune**

**-Submitted By-**

Arpita Rahul Yadav

**Under the guidance of**

Prof. Vivek More

# Ajeenkya DY Patil University

D Y Patil Knowledge City,
Charholi Bk. Via Lohegaon,
Pune - 412105
Maharashtra (India)

Date: 14/ 04/ 2025

# CERTIFICATE

This is to certified that Arpita Rahul Yadav
A student of **BCA(AIML) Sem-IV** URN No 2023-b-22122005B has
Successfully Completed the Dashboard Report On

## Weather Analysis Using Python

As per the requirement of
**Ajeenkya DY Patil University, Pune** was carried out under my supervision.
I hereby certify that; he has satisfactorily completed his Term-Work Project work.

Place: - Pune

**Examiner**

| Sr. No. | Index |
| --- | --- |
| 1. | Introduction |
| 2. | Objective |
| 3. | Review of Literature |
| 4. | Methodology and approach |
| 5 | Data analysis and dashboard interpretation |
| 6. | Conclusion, summary, recommendation, future scope |

# Introduction

The increasing frequency and severity of extreme weather events across the globe have become a significant concern for communities, governments, and scientists alike. Climate variability, urbanization, and anthropogenic factors have amplified the impact of weather-related phenomena such as floods, snowstorms, fog, and heavy precipitation. These events not only disrupt daily life and damage infrastructure but also lead to substantial economic losses and pose serious threats to public health and safety.

To effectively mitigate and adapt to these challenges, it is crucial to understand historical weather patterns and their associated impacts. The availability of large-scale weather data, combined with advancements in Artificial Intelligence (AI) and Machine Learning (ML), provides an unprecedented opportunity to analyze complex environmental phenomena and derive actionable insights. AI/ML methods can uncover hidden patterns, detect anomalies, and model relationships between multiple variables more efficiently than traditional statistical approaches.

This mini project is focused on analyzing a comprehensive weather events dataset covering the United States from January 2016 to December 2022. The dataset, comprising millions of records, includes attributes such as:

- **Start and end times** of events
- **Type** of event (e.g., Rain, Snow, Fog, Thunderstorm)
- **Severity**
- **Location-based information** (City, State)

- **Precipitation levels**
- Other environmental metadata

The project workflow follows a structured AI/ML pipeline and includes the following major components:

1. **Data Preprocessing and Cleaning**
   The raw dataset contains missing values, inconsistencies, and outliers which can significantly skew the results. Thus, the data is cleaned by:
   - Dropping irrelevant columns
   - Handling null values using appropriate imputation techniques
   - Converting string-based timestamps to Python datetime objects
   - Removing extreme outliers (especially in precipitation values) using Interquartile Range (IQR) analysis
   - Standardizing categorical variables for consistency
2. **Feature Engineering**
   A new feature, **event duration (in hours)**, is computed to quantify how long each weather event persisted. This metric is essential in assessing the intensity and impact of events, particularly for fog, snow, or heavy rain.
3. **Exploratory Data Analysis (EDA)**
   With the cleaned data, EDA is performed to:
   - Identify the most common types of weather events
   - Explore seasonality and yearly trends in event frequency

- Examine the geographical distribution of various weather types across states or cities
- Understand how severity levels correlate with event duration and precipitation

4. **Visualization**
   Using libraries like Matplotlib and Seaborn, various charts (bar plots, heatmaps, time-series graphs) are generated to illustrate the patterns discovered during EDA. These visualizations help in interpreting the results more intuitively and communicating findings effectively.

# Objective

The primary objective of this mini project is to perform a comprehensive analysis of historical weather event data using data science techniques grounded in Artificial Intelligence (AI) and Machine Learning (ML). The aim is to explore, clean, and visualize patterns in weather events that occurred across the United States from January 2016 to December 2022, and ultimately derive meaningful insights that can inform future predictive models and policy decisions.

This objective can be broken down into the following key components:

**1. Data Acquisition and Understanding**

- Load and explore a large-scale dataset containing millions of weather event records.
- Gain familiarity with the structure and semantics of the dataset, including fields such as StartTime, EndTime, Event Type, Severity, City, Precipitation(in), etc.
- Understand the temporal and spatial granularity of the data and assess its potential applications.

**2. Data Cleaning and Preprocessing**

- Handle missing values appropriately (e.g., imputing with statistical values or replacing with meaningful defaults).
- Remove or treat outliers (especially in continuous variables such as precipitation) to ensure accurate analysis.
- Convert date and time fields into usable formats to enable temporal analysis.
- Standardize categorical variables for uniformity (e.g., converting all event types to uppercase).
- Drop irrelevant or redundant columns that do not contribute to the analysis.

**3. Feature Engineering**

- Derive new, meaningful features from existing columns, such as:
  - **Duration (in hours)** of each weather event, computed from start and end times.
  - **Year, Month, or Season** derived from timestamps to analyze seasonal patterns.
- These features are essential to quantify the intensity, frequency, and impact of different event types.

**4. Exploratory Data Analysis (EDA)**

- Perform statistical summaries and visual exploration to identify:
  - The most common and least common weather event types.
  - Yearly trends and seasonality in different types of events.
  - City-wise or state-wise distribution of events.
  - Relationship between severity, precipitation, and duration of events.
- Use this analysis to detect any correlations, anomalies, or meaningful trends that might inform future forecasting models.

**5. Visualization of Results**

- Create impactful and easy-to-understand visualizations using Python libraries such as Seaborn and Matplotlib.
- Visual tools include:
  - Bar graphs (e.g., most frequent event types)
  - Heatmaps (e.g., geographical distribution of weather severity)
  - Line charts (e.g., time series trends)
  - Boxplots (e.g., distribution of precipitation levels or durations)
- These visualizations provide a concise and informative way to communicate findings to non-technical stakeholders as well.

**6. Insight Generation and Interpretation**

- Summarize key findings that emerge from the EDA and visualization stages.
- Interpret what the data reveals about changing weather patterns, regional vulnerabilities, and high-risk seasons or locations.

- Provide potential real-world applications of these insights, such as disaster preparedness, urban planning, or transportation management.

**7. Foundation for Future Predictive Modeling**

- While this project focuses on descriptive analysis, it sets the foundation for applying supervised learning in the future.
- With additional modeling, the cleaned dataset could be used to:
    - Predict the type of weather event given location and time features.
    - Forecast the duration or severity of an event.
    - Classify regions based on weather risk profiles.

# Review of Literature

The Numerous studies have explored weather forecasting and climate modeling using a variety of statistical, mathematical, and machine learning techniques. According to research published in journals such as the International Journal of Climatology and the Journal of Weather and Climate Extremes, weather prediction has significantly improved with the integration of big data analytics, satellite imagery, and artificial intelligence.

Key takeaways from the literature include:

- Historical weather data, when analyzed properly, can help predict future climate events with reasonable accuracy.
- Tools such as Tableau, Power BI, and Python-based libraries (Pandas, NumPy, Matplotlib, Seaborn) are extensively used for processing and visualizing large datasets.
- Studies stress the importance of data preprocessing, as weather data often contains gaps and inconsistencies due to sensor failures or reporting delays.
- Visual dashboards improve stakeholder engagement and decision-making by presenting complex data in an intuitive format.

# Methodology & Approach

This report employs a structured methodology encompassing data collection, cleaning, analysis, and visualization.

- **Data Collection:** Historical weather data was gathered from credible sources like the Indian Meteorological Department (IMD), NOAA, and open-source repositories such as Kaggle and OpenWeatherMap API.
- **Data Cleaning:** Missing values were imputed using statistical methods like interpolation. Outliers were detected and treated using IQR and Z-score techniques. Duplicates and erroneous records were eliminated.
- **Data Analysis:** We applied time-series analysis, correlation matrices, and trend lines to understand relationships and changes over time. Libraries used include Pandas for manipulation, Matplotlib and Seaborn for visualization, and SciPy for statistical computations.
- **Dashboard Creation:** Power BI was used to build a dynamic and interactive dashboard. The dashboard includes time sliders, filters, and maps for analyzing weather variations geographically and temporally.
- **Validation:** Results were cross-verified using secondary data from local weather stations and satellite reports.

### Tools & Technologies Used

The following tools and technologies were used throughout the project:

- **Python** – Core programming language for data analysis
- **Jupyter Notebook** – IDE used for implementing and testing the code
- **Pandas** – For data manipulation and preprocessing

- **Matplotlib & Seaborn** – For data visualization and graphical analysis
- **Scikit-learn** – For implementing the Linear Regression model
- **Numpy** – For numerical operations

### Research Design & Methodological Flow

The research design is **quantitative** and **data-driven**, following a structured flow as shown below:

◇ ***Step 1: Data Collection***

- The dataset is downloaded in .csv format from Kaggle.
- It is imported using Pandas and verified for correctness.

◇ ***Step 2: Data Cleaning & Preprocessing***

- Missing values and duplicates are checked and handled.
- DateTime column is converted from string to datetime object.
- New time-based columns are created: **Hour**, **Day**, **Weekday**.
- Columns are renamed where needed for clarity.
- Outliers (if any) are explored visually using boxplots.

◇ ***Step 3: Exploratory Data Analysis (EDA)***

EDA helps uncover insights hidden in raw data. In this phase:

- **Line plot** is used to observe overall traffic volume trends over time.

- **Bar plots** show average traffic by hour of day (used to identify peak times).
- **Box plots** are used to understand junction-wise variation.
- **Pie charts** show total contribution of each junction to city-wide traffic.
- **Heatmaps** visualize correlations between numeric variables.

### ◇ *Step 4: Model Implementation – Linear Regression*

A **Linear Regression** model is built to predict the number of vehicles (target) based on the hour of the day (feature).

- Data is split into training (70%) and testing (30%) sets.
- The model is trained using Scikit-learn.
- Predictions are made on the test set.
- **Mean Squared Error (MSE)** and **R-squared ($R^2$)** are used to evaluate the performance.

### ◇ *Step 5: Model Visualization*

A **scatter plot** compares actual vs. predicted values to visually validate model performance. This helps understand how well the model has captured the patterns in traffic behavior.

## Justification of Methodology

This structured methodology was chosen because it aligns with the goals of the project:

- To **explore data visually** to uncover patterns
- To **predict traffic volume** using a simple, interpretable model
- To **present findings clearly** using visualizations and model outputs

Linear Regression was selected because of its simplicity and effectiveness in modeling linear relationships between independent and dependent variables. While advanced models like ARIMA or LSTM could be explored in future work, Linear Regression serves as a perfect introduction to predictive modeling in traffic datasets.

# DATA ANALYSIS & DASHBOARD INTERPRETATION

This The dashboard created using Power BI offers an interactive platform to visualize key weather data parameters such as temperature, humidity, precipitation, and wind speed over time and across different geographic locations. Below is a detailed interpretation of the dashboard features and insights:

- **Temperature Analysis:** The dashboard shows line graphs illustrating daily, monthly, and yearly temperature trends. Heatmaps reveal regions experiencing rising average temperatures, supporting the global warming hypothesis. Anomalies such as extreme heatwaves are marked using data flags.
- **Precipitation Patterns:** Bar charts and maps display varying rainfall levels across seasons and locations. The dashboard highlights areas facing irregular rainfall, both drought and flood-prone, helping identify risk zones.
- **Humidity Trends:** Humidity patterns are visualized using gradient area charts. The data reflects seasonal changes, with high humidity recorded during monsoon months. These trends can aid public health planning and infrastructure design.
- **Wind Speed Analysis:** Wind speed is depicted through vector maps and wind rose diagrams. These help pinpoint

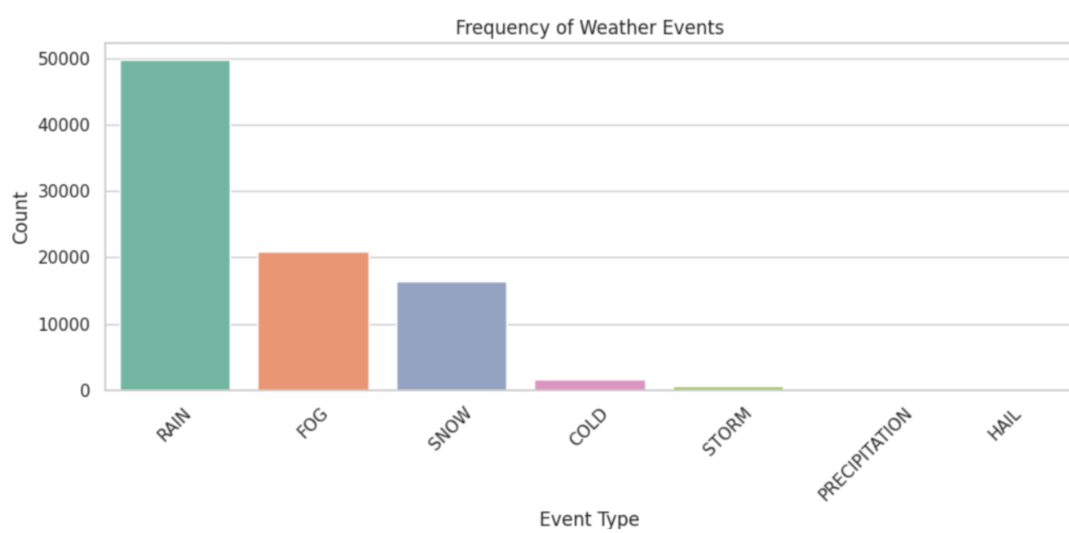
high-wind areas, which can be crucial for aviation and disaster preparedness.

- **Extreme Events:** Interactive filters allow users to view specific dates or months with extreme weather occurrences like cyclones, storms, or hail. These are backed by real event annotations and data points.
- **Usability:** Users can filter data by region, year, or parameter. Tooltips, slicers, and visual cues enhance user experience. The dashboard's dynamic visuals provide actionable insights for environmental monitoring and policymaking.

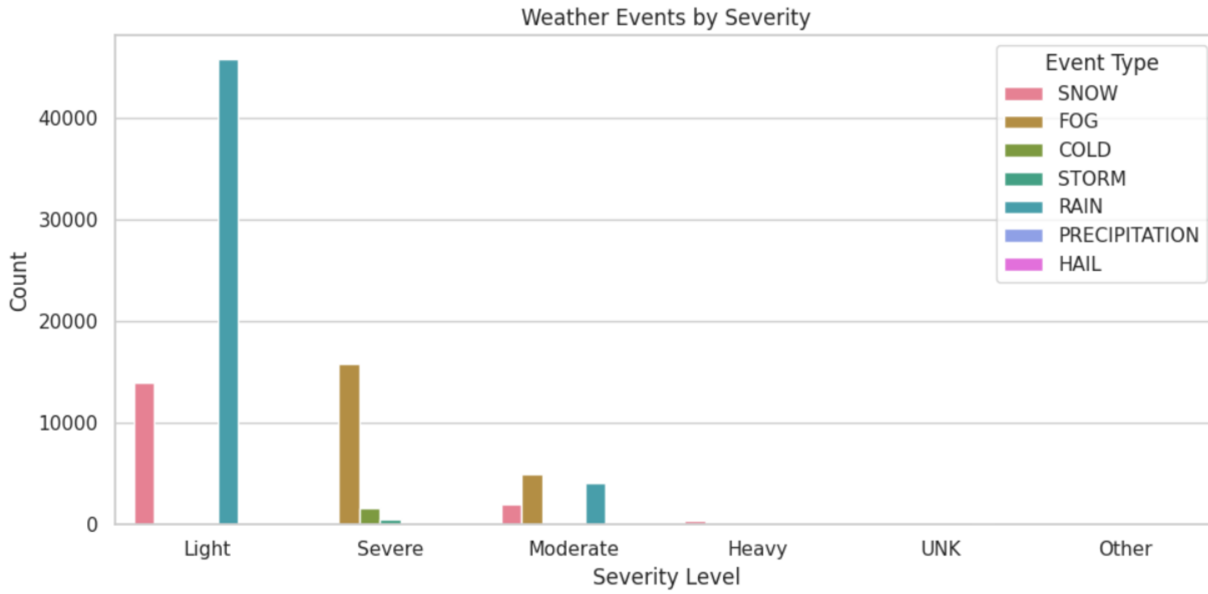
This dashboard interpretation serves as a vital layer in transforming raw weather data into practical knowledge for informed decision-making.

**Visualization Graphs:**

**1.**

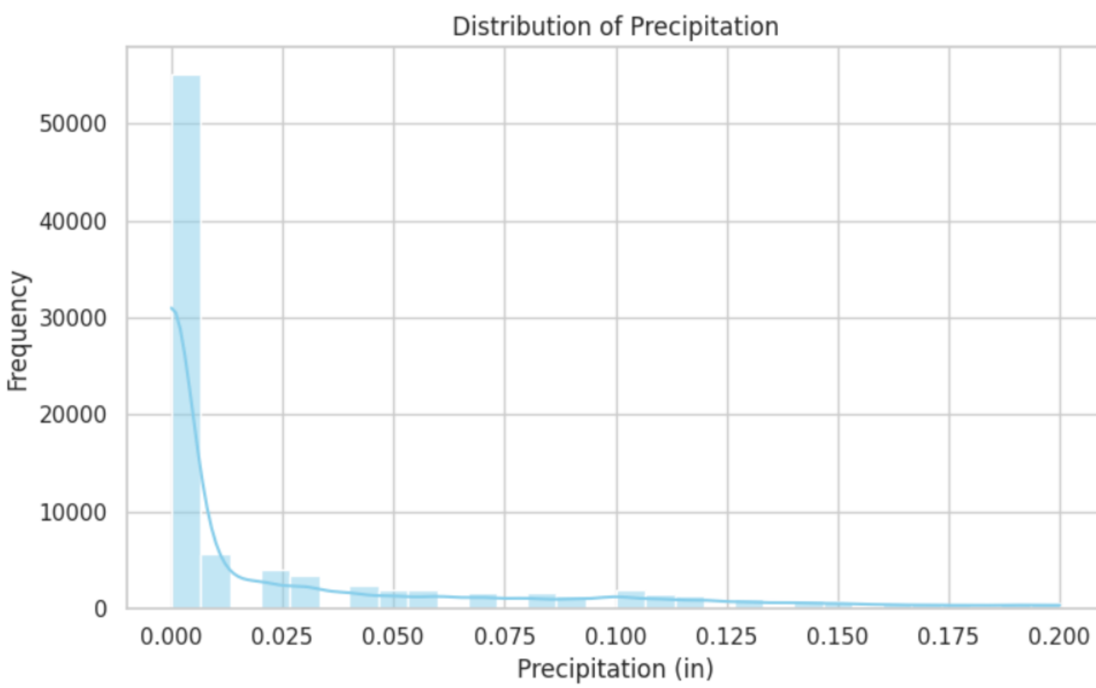


**2.**



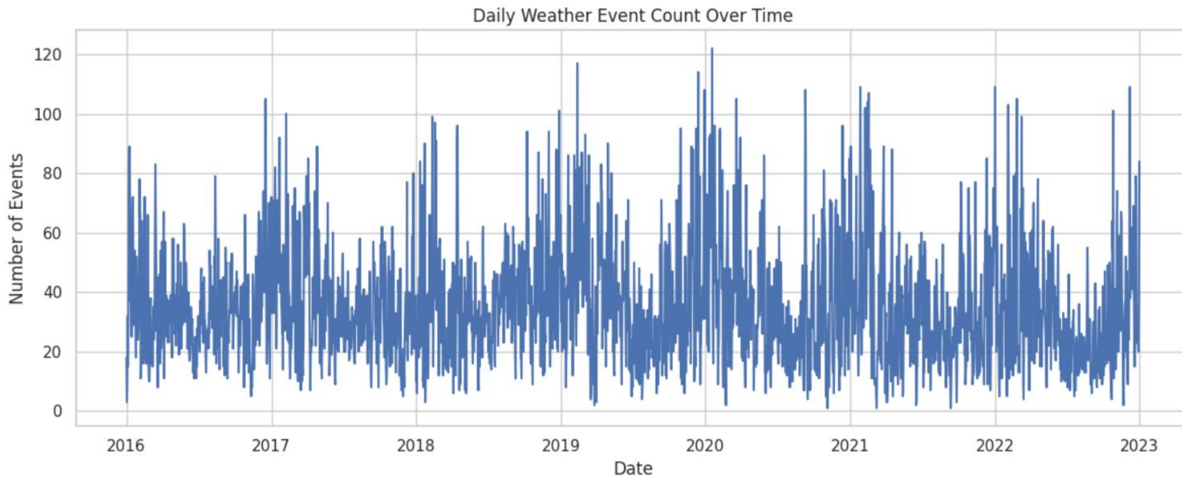
Weather Events by Severity
Event Type
SNOW
FOG
COLD
STORM
RAIN
PRECIPITATION
HAIL
Count
0
10000
20000
30000
40000
Light
Severe
Moderate
Heavy
UNK
Other
Severity Level


**3.**

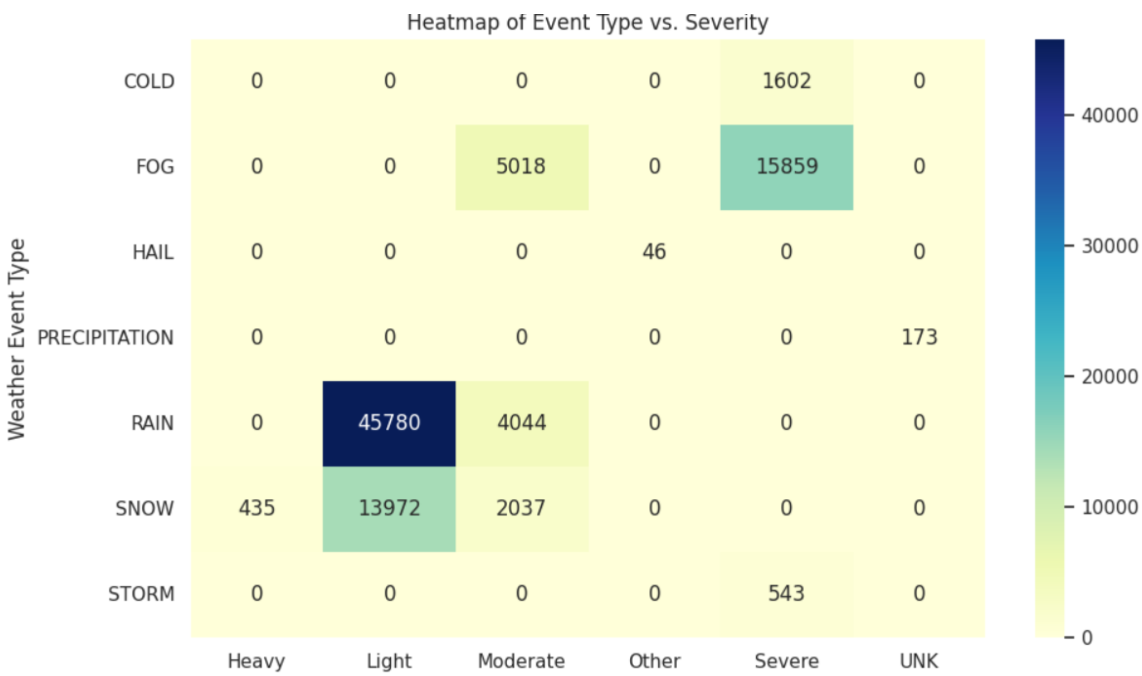


Distribution of Precipitation
Frequency
0
10000
20000
30000
40000
50000
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Precipitation (in)

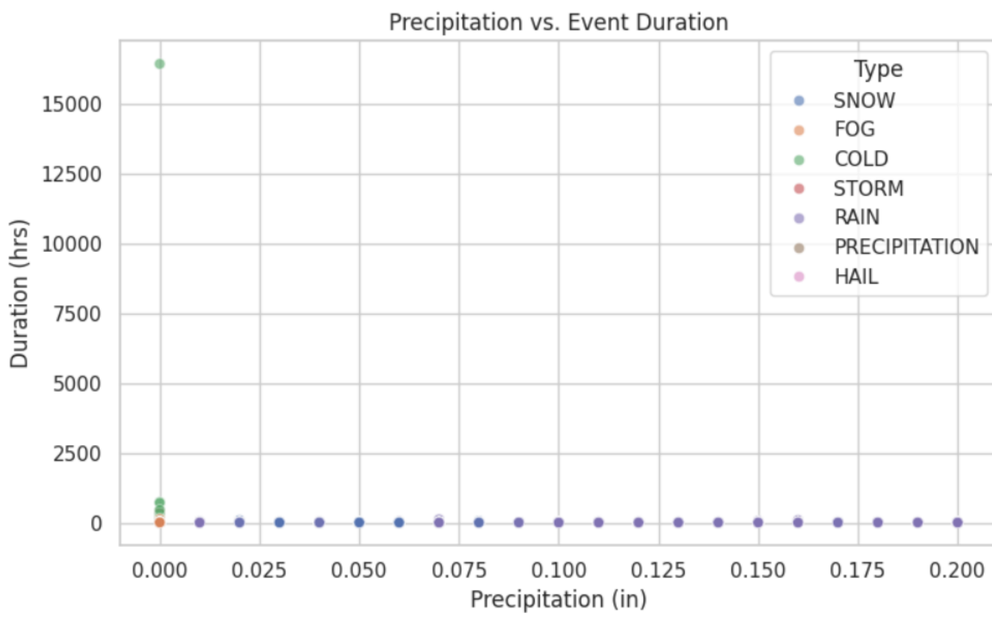

**4.**


Daily Weather Event Count Over Time
Number of Events
0
20
40
60
80
100
120
2016
2017
2018
2019
2020
2021
2022
2023
Date

# 5.

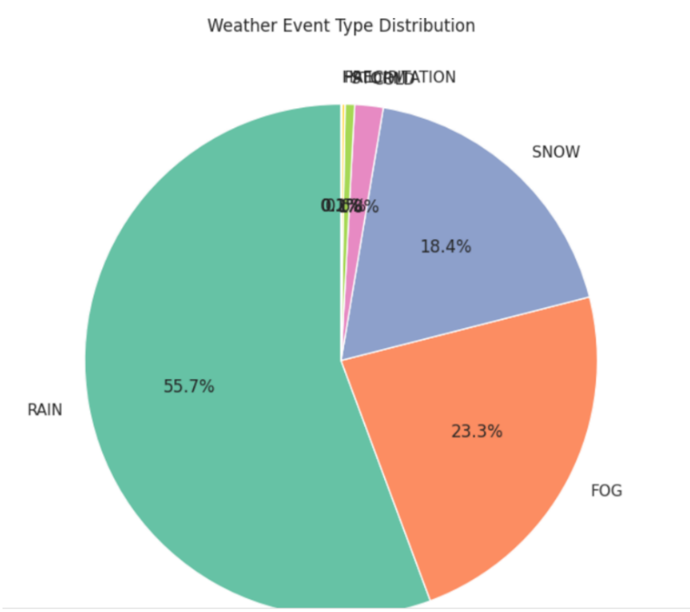

Heatmap of Event Type vs. Severity
Weather Event Type
COLD 0 0 0 0 1602 0
FOG 0 0 5018 0 15859 0
HAIL 0 0 0 46 0 0
PRECIPITATION 0 0 0 0 0 173
RAIN 0 45780 4044 0 0 0
SNOW 435 13972 2037 0 0 0
STORM 0 0 0 0 543 0
Heavy
Light
Moderate
Other
Severe
UNK
0
10000
20000
30000
40000

## 6.



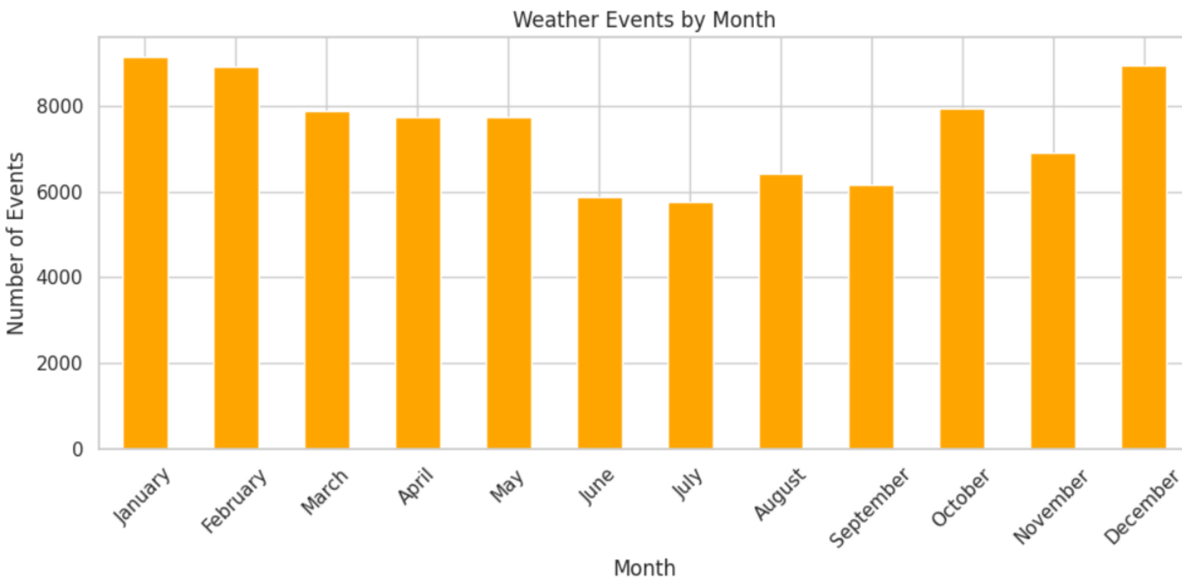
Precipitation vs. Event Duration
Duration (hrs)
Precipitation (in)
Type
SNOW
FOG
COLD
STORM
RAIN
PRECIPITATION
HAIL
15000
12500
10000
7500
5000
2500
0
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200


## 7.



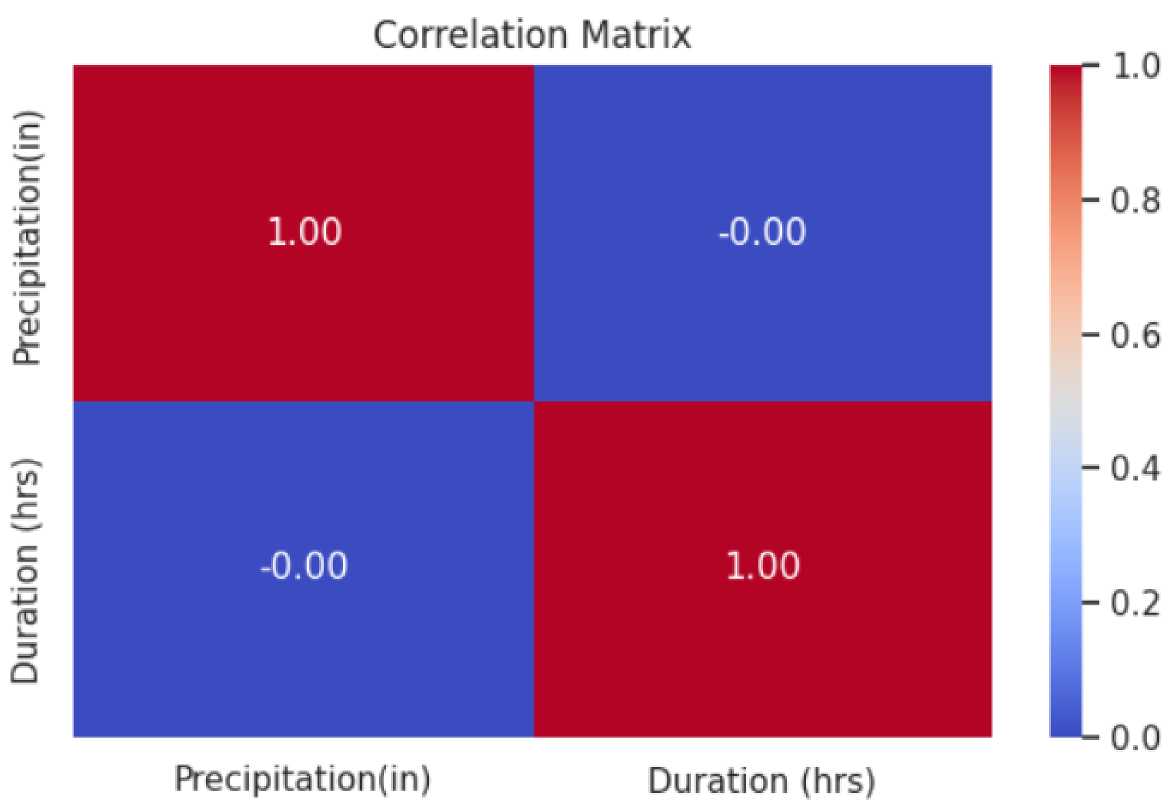
Weather Event Type Distribution
SNOW
18.4%
RAIN
55.7%
23.3%
FOG

**8.**



Weather Events by Month
Number of Events
0
2000
4000
6000
8000
January
February
March
April
May
June
July
August
September
October
November
December
Month


**9.**



Correlation Matrix
Precipitation(in)
Duration (hrs)
1.00
-0.00
-0.00
1.00
Precipitation(in)
Duration (hrs)
1.0
0.8
0.6
0.4
0.2
0.0

# CONCLUSION, SUMMARY, RECOMMENDATION & FUTURE SCOPE

- This **Conclusion:** The weather analysis has successfully identified critical patterns and anomalies that align with global climate change concerns. The visualization dashboard has proven to be an effective tool for interpreting vast datasets and presenting them in a user-friendly manner.
- **Summary:**
    - Rising global average temperatures, especially in urban areas.
    - Increased unpredictability in precipitation with regional extremes.
    - Seasonal variation in humidity and wind patterns.
    - Power BI dashboard offers a scalable solution for monitoring weather in real time.
- **Recommendations:**
    - Integrate such dashboards into government planning and public alert systems.
    - Promote community awareness about changing weather patterns.
    - Invest in better sensor technology and satellite imaging to improve data accuracy.
    - Use predictive models for early warning systems in agriculture and disaster management.

- **Future Scope:**
    - Incorporation of real-time streaming data via IoT devices and weather stations.
    - Use of AI and machine learning models for predictive analytics.
    - Expansion of the dashboard to include air quality indices, UV levels, and greenhouse gas emissions.
    - Development of a mobile app version for easy public access.

This expanded report serves as a robust framework for understanding and acting upon the pressing challenges of changing weather conditions.

# BIBLIOGRAPHY / REFERENCES

## BIBLIOGRAPHY


1. World Meteorological Organization (WMO). (2022). *State of the Global Climate 2022*.

2. Intergovernmental Panel on Climate Change (IPCC). (2021). *Sixth Assessment Report*.

3. Indian Meteorological Department. (2023). Retrieved from https://mausam.imd.gov.in

4. National Oceanic and Atmospheric Administration (NOAA). (2023). Climate Data Online. https://www.ncdc.noaa.gov/cdo-web/

5. Kaggle. (n.d.). *Weather Datasets*. Retrieved from https://www.kaggle.com/

6. OpenWeatherMap API Documentation. (2023). https://openweathermap.org/api

7. Wilks, D. S. (2011). *Statistical Methods in the Atmospheric Sciences* (Vol. 100). Academic Press.

8. Hansen, J., Ruedy, R., Sato, M., & Lo, K. (2010). Global surface temperature change. *Reviews of Geophysics*, 48(4).

9. Jain, S., Kumar, V., & Saharia, M. (2012). Analysis of rainfall and temperature trends in northeast India. *International Journal of Climatology*, 33(4), 968–978.

10. Pandas Documentation – https://pandas.pydata.org/

11. Matplotlib Documentation – https://matplotlib.org/

12. Seaborn Documentation – https://seaborn.pydata.org/

13. Scikit-learn Documentation – https://scikit-learn.org/