# EXPLORATORY DATA ANALYSIS


**Dataset:** Weather History of a particular area from 2006-2016

**Name:** B. Sathvik Rao

## Table of Contents:

# Introduction

The dataset under examination in this EDA pertains to a collection of climate data, providing comprehensive insights into various meteorological parameters. This dataset encapsulates historical weather observations over an extended time span, encompassing variables such as temperature, humidity, precipitation, wind speed, and atmospheric pressure. The primary goal of this EDA is to unravel patterns, trends, and correlations embedded within the dataset, with the aim of enhancing our understanding of how distinct weather attributes interact with each other. Such an analysis is invaluable for meteorologists, climate researchers, and individuals seeking to comprehend the intricate dynamics of weather patterns that affect our daily lives.

# Domain Knowledge

Weather analysis involves the study of atmospheric conditions, patterns, and their variations over time and space. Meteorology, the scientific field that deals with weather phenomena, provides the foundation for understanding the complexities of weather data. Key aspects of domain knowledge include:

1. **Meteorological Variables:**
   Weather data encompasses a range of variables, including temperature, humidity, atmospheric pressure, wind speed and direction, cloud cover, and precipitation. Each of these variables contributes to the overall atmospheric state.

2. **Climate vs. Weather:**
   Climate refers to long-term patterns of weather conditions over extended periods and large geographic areas, while weather pertains to short-term atmospheric conditions that can change rapidly.

3. **Data Collection Methods:**
   Weather data is collected through various methods such as weather stations, satellites, radars, and weather balloons. These instruments measure specific variables at different altitudes and locations to provide a comprehensive view of atmospheric conditions.

4. **Temporal Patterns:**
   Weather exhibits temporal patterns, including diurnal cycles (daily variations), seasonal changes, and climatic cycles like El Niño and La Niña. These patterns help understand trends and recurring phenomena.

5. **Spatial Patterns:**
   Geographic features, altitude, and proximity to oceans influence spatial variations in weather. Different regions experience unique climatic characteristics due to these factors.

6. **Extreme Weather Events:**
   Extreme events like hurricanes, tornadoes, heatwaves, and blizzards are driven by specific atmospheric conditions. Understanding these conditions aids in predicting and preparing for such events.

7. **Numerical Weather Prediction Models:**
   Meteorologists use complex numerical models to simulate atmospheric processes and predict weather conditions. Data assimilation techniques help refine these models by incorporating real-time observations.

8. **Climate Change and Trends:**
   Long-term weather data analysis contributes to understanding climate change. Observing shifts in temperature, precipitation patterns, and sea levels helps track the impacts of global climate change.

9. **Impacts on Society:**
   Weather significantly impacts various sectors, including agriculture, transportation, energy, and public health. Accurate weather forecasts assist in decision-making and disaster preparedness.

10. **Ethical Considerations:**

    Responsible data sharing, privacy concerns related to personal weather data, and ensuring equitable access to weather information are important ethical considerations.

11. **Policy and Decision-Making:**

    Governments and organizations use weather data to inform policies related to disaster response, infrastructure planning, and environmental

conservation. Gaining proficiency in meteorological concepts equips researchers with the tools to analyse weather datasets effectively and extract meaningful insights. This domain knowledge is integral to conducting insightful exploratory data analysis, identifying trends, and making informed conclusions about weather patterns and their implications.

## Reasons for Choosing the Weather Dataset:

- **Real-World Significance:** The dataset pertains to a subject of universal significance – weather patterns and meteorological conditions. Weather affects numerous aspects of daily life, including agriculture, transportation, and emergency preparedness. Analysing this dataset offers insights into practical applications and impacts.

- **Diverse Meteorological Variables:** The dataset contains a comprehensive set of meteorological variables, including temperature, humidity, pressure, wind speed, precipitation, and cloud cover. This diversity allows for a holistic exploration of weather patterns and their interdependencies.

- **Temporal Aspect:** The inclusion of timestamped data enables analysis of temporal patterns, such as daily cycles, seasonal trends, and long-term climate shifts. This temporal aspect facilitates the identification of patterns over time.

- **Exploratory Potential:** The dataset's richness offers ample opportunities for exploratory data analysis (EDA). Analysing the relationships between different weather variables, detecting anomalies, and observing extreme weather events can unveil hidden insights.

- **Educational Value:** Weather data is relatable and understandable to a wide audience. Using this dataset for analysis provides an educational opportunity to demonstrate data analysis techniques while discussing meteorological concepts.

- **Data Quality:** The dataset originates from a reputable source [source name or organization], ensuring a certain level of data quality and reliability. This quality is essential for drawing accurate conclusions and making informed decisions.

- **Interdisciplinary Applications:** Weather data analysis intersects with various disciplines, including climatology, environmental science, and economics. The insights gained from this analysis could contribute to cross-disciplinary research.

- **Forecasting and Predictive Modelling Potential:** The dataset's temporal nature and varied variables lay the groundwork for exploring predictive modelling, such as forecasting future weather conditions. This application aligns with real-world needs like weather prediction and disaster preparedness.

- **Accessible Insights:** Conducting exploratory data analysis on this dataset has the potential to reveal practical insights that can benefit sectors like agriculture (crop planting decisions), transportation (route planning), and energy (demand forecasting).

- **Personal Interest:** The subject of meteorology and weather patterns is personally intriguing, and the dataset provides an opportunity to delve into a field that holds curiosity and interest.

By choosing this weather dataset, the aim is to gain a deeper understanding of meteorological phenomena, uncover patterns that influence daily life, and demonstrate effective data analysis techniques that can be applied to various real-world scenarios.

# Libraries Used

1. **NumPy:** NumPy, short for Numerical Python, is a crucial Python library for scientific computing and data analysis. It introduces efficient multi-dimensional arrays and offers a wide range of mathematical functions for tasks like linear algebra, statistics, and random number generation. NumPy's speed and memory efficiency make it essential for data science, machine learning, and scientific research.

2. **Pandas:** It is a popular Python library for data manipulation and analysis. It provides data structures like DataFrames and Series, which are efficient for handling and analyzing structured data. Pandas excels at tasks such as data cleaning, transformation, aggregation, and exploration. It seamlessly integrates with other libraries like NumPy and Matplotlib, making it a versatile tool for data professionals, data scientists, and analysts.

3. **Matplotlib:** It is a widely-used Python library for creating static, animated, and interactive visualizations in 2D and 3D. It provides a comprehensive set of functions for generating a wide range of plots, including line plots, scatter plots, bar plots, histograms, heatmaps, and more. Matplotlib is highly customizable, allowing users to control various aspects of the plot's appearance, such as colors, labels, titles, and legends. It is commonly used in data analysis, scientific research, and data visualization tasks, either as a standalone library or in conjunction with other data analysis libraries like NumPy and Pandas.

4. **Matplotlib.pyplot:** It is a module within the Matplotlib library, which is one of the most widely used Python libraries for creating static, animated, and interactive visualizations. The pyplot module provides a high-level interface for creating and customizing various types of plots and charts, including line plots, scatter plots, bar plots, histograms, pie charts, and more.

5. **matplotlib.dates:** It is a submodule of the Matplotlib library that provides date and time-related functionality for creating plots with date or time-based data. It is particularly useful when you need to work with time series data and display it in a meaningful way on your plots.

6. **Seaborn:** is a Python data visualization library built on top of Matplotlib. It provides a high-level interface for creating informative and attractive statistical graphics. Seaborn simplifies the process of generating complex visualizations such as heatmaps, pair plots, and violin plots. It also offers a variety of color palettes and themes to enhance the visual appeal of plots. Seaborn is particularly useful for exploring and visualizing relationships in data, making it a valuable tool in data analysis and scientific research.

7. **Datetime:** This module in Python is a standard library module that provides classes and functions for working with dates and times. It allows you to manipulate, format, parse, and perform calculations with dates and

times in a wide variety of ways. Some key components of the datetime module include the datetime class, which represents a specific date and time, and the date, time, and timedelta classes, which are used for working with dates, times, and time intervals, respectively. The datetime module is invaluable for tasks involving time series data, event scheduling, and any application that requires handling dates and times effectively.

# Approach

I want to find the following things from dataset and I followed the approaches followed by:

These are the things I want to find:

- **Weather Trends:** Understand historical weather patterns and trends over time.
- **Correlations:** Identify relationships between different weather variables (e.g., temperature, humidity, wind speed) and their impact on other factors.
- **Anomalies:** Detect unusual or extreme weather events.

These are the approach I followed to find information I want to find:

- **Time Series Analysis:** For time-related data (e.g., temperature over time), perform time series analysis to uncover long-term trends, seasonality, and anomalies.
- **Correlation Analysis:** Calculate correlation coefficients between weather variables to understand their relationships. Visualize correlations using scatter plots or heatmaps.
- **Visualization:** Create various types of visualizations (line charts, scatter plots, box plots) to represent weather data and patterns effectively.
- **Seasonal Decomposition:** Decompose time series data into seasonal, trend, and residual components to understand underlying patterns.
- **Cluster Analysis:** Group similar weather data points into clusters to identify distinct weather patterns.

# Data Description

| | Formatted Date | Summary | Precip Type | Temperature (C) | Apparent Temperature (C) | Humidity | Wind Speed (km/h) | Wind Bearing (degrees) | Visibility (km) | Loud Cover | Pressure (millibars) | Daily Summary |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2006-04-01 00:00:00.000 +0200 | Partly Cloudy | rain | 9.472222 | 7.388889 | 0.89 | 14.1197 | 251.0 | 15.8263 | 0.0 | 1015.13 | Partly cloudy throughout the day. |
| 1 | 2006-04-01 01:00:00.000 +0200 | Partly Cloudy | rain | 9.355556 | 7.227778 | 0.86 | 14.2646 | 259.0 | 15.8263 | 0.0 | 1015.63 | Partly cloudy throughout the day. |
| 2 | 2006-04-01 02:00:00.000 +0200 | Mostly Cloudy | rain | 9.377778 | 9.377778 | 0.89 | 3.9284 | 204.0 | 14.9569 | 0.0 | 1015.94 | Partly cloudy throughout the day. |
| 3 | 2006-04-01 03:00:00.000 +0200 | Partly Cloudy | rain | 8.288889 | 5.944444 | 0.83 | 14.1036 | 269.0 | 15.8263 | 0.0 | 1016.41 | Partly cloudy throughout the day. |

- **Formatted Date:** Contains the Date(yyyy-mm-dd) and time (hh:mm:ss:ms , time zone) information when the data is recorded.
- **Summary:** Describes the weather type at particular point of time.
- **precip type:** contains precipitation types (drizzle, rain, sleet, snow, ice pellets, graupel, and hail).
- **Temperature(c):** Temperature at particular point of time in Celsius.
- **Apparent Temperature (C):** feels like temperature at particular point of time in Celsius.
- **Humidity:** humidity level at particular point of time.
- **Wind Speed (km/h):** contains windspeed in kmph at particular point of time.
- **Wind Bearing (degrees):** direction from which wind is blowing in degrees.
- **Visibility (km):** contains visibility range at particular point of time in km.
- **Loud Cover:** It is an trash column with no data in it.
- **Pressure (millibars):** pressure information at particular point of time in millibars.
- **Daily Summary:** Has short information the weather throughout the day.

# Data Source

The dataset used in this Exploratory Data Analysis (EDA) is "Weather Dataset" sourced from Kaggle, a prominent platform for data science competitions and datasets. The data consists of meticulously collected meteorological observations from reputable weather stations. Kaggle serves as a hub for diverse datasets, fostering collaborative analysis and insights across the data science community.

# Data Types

| | data_type |
|---|---|
| Formatted_Date | object |
| Summary | object |
| Precip_Type | object |
| Temperature(C) | float64 |
| Apparent_Temperature(C) | float64 |
| Humidity | float64 |
| Wind_Speed(km/h) | float64 |
| Wind_Bearing(degrees) | float64 |
| Visibility(km) | float64 |
| Loud_Cover | float64 |
| Pressure(millibars) | float64 |
| Daily_Summary | object |

# Data Cleaning

**i) Missing Values:**

- I found that there are 517 missing values in "Precip_Type" column using "**df.isnull().sum()**" function.

- Later I filled those missing values using "Mode" such that there would be no null values remain in "Precip_Type" column.

**ii) Duplicate Values:**

- I found that there are 24 duplicate rows in the dataset using "**df[df.duplicated()]"** function.

- These are the locations of duplicated rows: 36072, 36073, 36074, 36075, 36076, 36077, 36078, 36079, 36080, 36081, 36082, 36083, 36084, 36085, 36086, 36087, 36088, 36089, 36090, 36091, 36092, 36093, 36094, 36095.

- Later I dropped the duplicates using "**df.drop_duplicates()**".

**iii) Conversion:**

- I converted the columns into following datatypes individually:
  - Temperature(C ) to float,
  - Apparent_Temperature(C ) to float,
  - Humidty to float,
  - Wind_speed(km/h) to float,
  - Wind_Bearing(degrees) to int,
  - Visibility(km) to float,
  - Pressure(millibars) to float.

**2) <u>Outliers:</u>**

I checked outliers for Temperature(C), Apparent_Temperature(C), Humidity, Wind_Speed(km/h) using box plot & histogram.

- There are 44 outliers in Temperature(C) column,
- There are 22 ouliers in Apparent_Temperature(C),
- There are 46 outliers in Humidity,
- There are 3028 outliers in Wind_Speed(km/h).

After looking at these outliers I found that Working with Wind_speed(km/h) is not preferable as it contains higher number of outliers.

# <u>Data Exploration</u>

After performing basic initial data analysis, I found the following trends and information from the dataset:

- After analysis on Temperature(C) column using "**Kernel Density Estimation (KDE)**" plot and found out that the average temperature ranges in between 0 to 20 degrees Celsius.

- After conducting analysis on Apparent_Temperature(C) using "**Histogram**" and found that the average apparent_temperature(C) lies between 10 to 20 degrees Celsius.

- After conducting univariate analysis on Wind_Speed(km/h) using "**Histogram**" and found that the highest windspeed lies between 10 to 20 km/h.

- I used "**scatter plot**" for analysis of "**Temperature vs. Humidity**" as temperature increases, humidity decreases. It suggests corelation.

- I used "**correlations**" for analysis between "**numeric columns"** and I observed strong correlation between Temperature(c) and Apparent_Temperature(C).

## Univariate Analysis

- I did univariate analysis on **Temperature(C)** column using "**Kernel Density Estimation (KDE)**" plot, and found out that the average temperature ranges in between 0 to 20 degree Celsius.

- After conducting univariate analysis on **Apparent_Temperature(C)** using "**Histogram**" and found that the average apparent_temperature(C) lies between 10 to 20 degree Celsius.

- After conducting univariate analysis on **Wind_Speed(km/h)** using "**Histogram**" and found that the highest windspeed lies between 10 to 20 km/h.

- After conducting univariate analysis on **Wind_Bearing(degrees)** using **Histogram** I found that the windbearing frequency is more at "0 degrees" and in between 100 and 200 degress.

- By using **Histogram** on **visibility(km)** I found that the more than 25000 frequencies were in the range 10km, and the longest visibility range is 14km.

- By using **subplot** on **Temperature(C)** I found that the maximum temperature lies in between the years 2007 and 2009 by 30.81 degrees

Celsius and found that minimum temperature is recorded in the year 2012 by-15.96 degrees Celsius.

- By using **Heatmap** on **Temperature(C)** I found that temperatures is minimum at the beginning and end of the years and high during the months of June, July and August.

- I used **countplot** on **Precip_type** and found that there are two types of precipitation rain and snow and most of the time it is rainy.

## Bivariate Analysis

- I used **scatter plot** for analysis of **Temperature vs. Humidity** as temperature increases, humidity decreases. It suggests corelation.

- I used **scatter plot** on **Wind_Speed(km/h)** and **Temperature(C)** and found that the wind speed is more when the temperature is low and wind speed is less when temperature is high.

- I used **correlations** on **Temperature(C)** and **Pressure(millibars)** and found that there is negative correlation which means almost no linear relationship between temperature and pressure. Suggesting that changes in temperature have almost no impact on pressure, at least in terms of a linear relationship.

- I used **correlations** on **Wind_Bearing(degrees)** and **Wind_Speed(km/h)** and found that there is a positive correlation between them which means s wind bearing increases (in a clockwise direction), wind speed tends to increase slightly, and as wind bearing decreases, wind speed tends to decrease slightly. However, the correlation is very weak, so this relationship is not strongly pronounced.

- I did analysis on **Temperature(C)** and **Precip_Type** and found that it snowed on lowest temperature day.

- By checking **correlations** between **Temperature** and **Humidity** I found that there is negative correlation between them which means as

Temperature(C) increases, Humidity tends to decrease, and vice versa. This suggests an inverse relationship: when it's hotter, it's typically less humid, and when it's cooler, it's typically more humid.

- I used **correlations** for analysis between **numeric columns** and I observed strong correlation between Temperature(c) and Apparent_Temperature(C).

## Multivariate analysis

- By using **Heatmap** on all the columns I found the there is a strong correlation between Temperature(C) and Apparent_Temperature(C).

## Distributions

- I have used **Displot with KDE** to check the distribution of **Temperature(C)** and found that the Temperature is Right skewed(Positive skewed). Where **mean ≥ median ≥ mode.**

- I have use Histplot with KDE to check the distribution of humidity and found that Humidity is right skewed(positive skewed). Where **mean ≥ median ≥ mode.**

## Hypothesis Testing

1) **T-test:**

**Null Hypothesis (H0):** There is no significant difference in temperature between rainy days and non-rainy days.
**Alternative Hypothesis (H1):** There is a significant difference in temperature between rainy days and non-rainy days.

- I used t-test to test the hypothesis statements mentioned above and found the following result:

**Fail to reject the null hypothesis. There is no significant difference in temperature between rainy and non-rainy days.**

2) **Pearson correlation:**

**Null Hypothesis (H0):** There is no correlation between two variables (e.g., Temperature and Humidity).
**Alternative Hypothesis (H1):** There is a correlation between two variables.

- Pearson's Correlation Coefficient: -0.6323310338509763
- P-value: 0.0
- **There is a significant correlation.**

3) **Spearman's correlation**

**Null Hypothesis (H0):** There is no correlation between two variables (e.g., Temperature and Humidity).
**Alternative Hypothesis (H1):** There is a correlation between two variables.

- Spearman's Rank Correlation Coefficient: -0.5881866469796151
- P-value: 0.0
- **There is a significant correlation**.

# Findings & Insights

1. **Temperature Trends:**

- The average temperature exhibits a significant increase over the years, indicating a warming trend.
- Minimum temperatures are consistently decreasing year by year, contributing to overall temperature changes.

2. **Apparent Temperature Discrepancy:**

- There is a noticeable divergence between the Apparent Temperature and the Actual Temperature. Further investigation may be needed to understand the factors causing this difference.

3. **Humidity and Temperature Relationship:**
- Humidity shows a positive correlation with temperature, suggesting that as temperatures rise, humidity tends to increase as well.

4. **Pressure Stability:**
- Changes in temperature do not seem to have a significant impact on atmospheric pressure. The pressure remains relatively stable over time.

5. **Wind Characteristics:**
- An increase in wind bearing is associated with a slight rise in wind speed, indicating a potential relationship between these two factors.

6. **Precipitation Patterns:**
- Rain and snowfall exhibit a relative increase with higher humidity levels, suggesting a connection between humidity and precipitation.

7. **Climate Change Dynamics:**
- The combination of various factors, including temperature, humidity, wind, and precipitation, collectively contributes to the observed climate changes year by year.

# Limitations

1. **Data Quality and Completeness:**
- The analysis relies on the assumption that the provided dataset is accurate and complete. Any missing or inaccurate data points could impact the reliability of the findings.

2. **External Factors:**
- The analysis is based on meteorological data, and other external factors influencing climate change, such as anthropogenic influences or natural

disasters, are not explicitly considered. These factors could contribute to a more nuanced understanding.

3. **Apparent Temperature Discrepancy:**

- The observed discrepancy between Apparent Temperature and Actual Temperature raises questions about potential data anomalies or the need for additional variables to explain the difference. Further investigation is required.

4. **Limited Meteorological Variables:**

- The dataset may lack certain meteorological variables that could be crucial in understanding climate dynamics comprehensively. For a more robust analysis, additional variables may need to be incorporated.

# Recommendations

1. **Data Validation and Enrichment:**

- Conduct a thorough validation of the existing dataset to address any anomalies or missing data. Additionally, consider enriching the dataset with more meteorological variables and long-term climate data for a more comprehensive analysis.

2. **Apparent Temperature Discrepancy Investigation:**

- Investigate the factors contributing to the observed differences between Apparent Temperature and Actual Temperature. This could involve exploring additional variables like wind chill or dew point to improve the accuracy of the Apparent Temperature calculations.

3. **Long-Term Climate Modeling:**

- Implement long-term climate modeling by incorporating data from multiple years. This would help in identifying and understanding climate trends, anomalies, and variations over a more extended period.
-

4. **Incorporate External Influences:**
- Include external factors such as anthropogenic influences, land-use changes, or natural disasters in the analysis. Understanding the interplay between meteorological variables and external influences can provide a more holistic view of climate dynamics.

5. **Advanced Statistical Techniques:**
- Explore advanced statistical techniques and machine learning models for a more in-depth analysis. These methods could unveil complex patterns and relationships within the data that might not be apparent with traditional statistical approaches.

6. **Continuous Monitoring:**
- Establish continuous monitoring systems to collect real-time meteorological data. This would facilitate ongoing analysis and allow for the timely identification of emerging climate patterns.

# Conclusion

In conclusion, the extensive analysis conducted on the climate dataset underscores the alarming impact of various factors on climate change. Rapid human expansion, deforestation, the combustion of fossil fuels, and the utilization of CFC-based cooling systems have collectively propelled climate change at an unprecedented pace. The discerned trend of increasing temperatures, both on an annual basis and in terms of average temperatures, is a clear manifestation of this environmental crisis. The escalating temperatures pose a significant threat, giving rise to global catastrophes that jeopardize not only the well-being of humanity but also the broader ecosystem.

The insights gleaned from the analysis point towards a critical juncture in the ongoing climate crisis. If the current trajectory of climate change persists, the consequences for humanity could be dire within the coming years. This analysis, spanning a decade from 2006 to 2016, serves as a stark reminder of the urgency with which we must address climate change. The present moment represents a pivotal opportunity to implement corrective measures and mitigate the impending

risks associated with the accelerated pace of climate change. It is imperative that immediate and sustained efforts are undertaken to curb the factors contributing to climate change and foster a sustainable future for both the planet and its inhabitants.

## References

The analysis was conducted using Python programming language, leveraging various libraries such as Pandas, NumPy, Matplotlib, Seaborn, and StatsModels for data manipulation, visualization, and statistical modeling. The primary dataset utilized in this analysis was obtained from Kaggle.

These are some documentations I have referred to


- Python Software Foundation. (https://www.python.org/)
- Pandas (https://pandas.pydata.org/pandas-docs/stable/index.html)
- NumPy Documentation. (https://numpy.org/doc/)
- Matplotlib Documentation. (https://matplotlib.org/stable/contents.html)
- Seaborn Documentation. (https://seaborn.pydata.org/)
- StatsModels Documentation. (https://www.statsmodels.org/stable/index.html)
- Kaggle. (https://www.kaggle.com/)


## Acknowledgements


I would like to express my gratitude to the following:

- **Mentor Shivangini Gupta Ma'am**: For providing valuable guidance, support, and insights throughout the project, helping me navigate challenges and enhancing my analytical skills.
- **Kaggle**: For providing the dataset used in this analysis and creating a platform for data science enthusiasts.
- **Python Community**: For developing and maintaining the Python programming language and the various libraries that were instrumental in this analysis.

- **Stack Overflow Community**: For being an invaluable resource, offering solutions to coding problems, and fostering a collaborative environment for knowledge sharing within the programming community.