*Data Visualization and Communication (INFO 5709) – Final Project*

**WEATHER PREDICTION**

Saipradeep Bomma

Student ID: 11687754

Instructor : Gahangir Hossain

**1. INTRODUCTION**

Accurate and fast weather predictions are crucial in today's society, as our everyday activities are heavily influenced by the weather. Predicting various weather conditions, such as drizzle, rain, sun, snow, and fog, is the focus of my project. Planning outside activities, figuring out the best routes around, and preparing for any weather that may arise are all made much easier by understanding and forecasting these weather patterns.

My aim is to learn more about the patterns and factors that contribute to the occurrence of each type of weather by analyzing large amounts of data and employing sophisticated computer models to forecast the future. It's not just about knowing what the weather will be like; it's about helping people and communities make smart choices based on what the weather might do.

My goal is to develop a useful tool that goes beyond standard weather predictions. With the help of this application, individuals will be able to prepare and make wise decisions by being informed about various weather conditions in advance. Giving people the information they need to deal with various weather scenarios—whether they're preparing for a day out or something more serious—is the main goal of my initiative. It's similar to having a helpful handbook to help us deal with any weather conditions that arise.

**2. MOTIVATION**

In our daily lives, weather conditions play a pivotal role in influencing a myriad of activities, ranging from planning outdoor events to optimizing transportation strategies. Accurate weather predictions are essential for individuals, businesses, and communities to make informed decisions and enhance overall preparedness. The Weather Prediction Dataset serves as a valuable resource for developing predictive models that can offer insights into various weather phenomena, including precipitation, temperature, wind, and more. By leveraging this dataset, I aim to contribute to the advancement of weather forecasting, providing users with reliable tools for anticipating and preparing for diverse weather scenarios. This endeavor aligns with the broader goal of harnessing data-driven approaches to enhance our understanding of atmospheric patterns and empower individuals and organizations to navigate the dynamic challenges posed by weather conditions.

**3. DATASET DESCRIPTION**

The Seattle Weather Prediction dataset, obtained through Kaggle, offers a thorough understanding of the climate dynamics of the energetic metropolis of Seattle. This dataset, which includes important meteorological metrics including wind speed, precipitation, maximum and minimum temperatures, and weather descriptors, is essential to my comprehension of the complex patterns that influence Seattle's weather. The dataset's insights are priceless since weather forecasts have a significant impact on many elements of daily living, including outdoor activities and commute planning. I want to use exploratory data analysis (EDA) to determine how precipitation and temperature extremes are related, evaluate how wind affects temperature perception, and investigate how different weather types correspond with different precipitation amounts. By delving into this dataset, I endeavor to contribute to a more comprehensive understanding of Seattle's weather patterns and enhance my ability to make informed predictions.

The dataset consists of the following attributes:

1. Precipitation:

- Definition: The amount of rainfall or snowfall measured in millimeters.

- Role: Indicates the level of precipitation, helping assess the wetness of the weather.

2. Temp\_max:

- Definition: Maximum temperature recorded in degrees Celsius.

- Role: Reflects the highest temperature reached during a specific period, aiding in understanding temperature variations.

3. Temp\_min:

- Definition: Minimum temperature recorded in degrees Celsius.

- Role: Provides information about the lowest temperature experienced, contributing to the overall temperature profile.

4. Wind:

- Definition: Wind speed measured in kilometers per hour.

- Role: Indicates the speed of the wind, influencing factors like temperature perception and atmospheric conditions.

5. Weather:

- Definition: Descriptive category representing the prevailing weather conditions.

- Role: Qualitatively characterizes the weather, encompassing factors like drizzle, rain, sun, snow, and fog.

6. Date:

A date attribute represents a specific point in time, typically denoted by day, month, and year, facilitating chronological organization of temporal data.

# 4. TOOLS USED

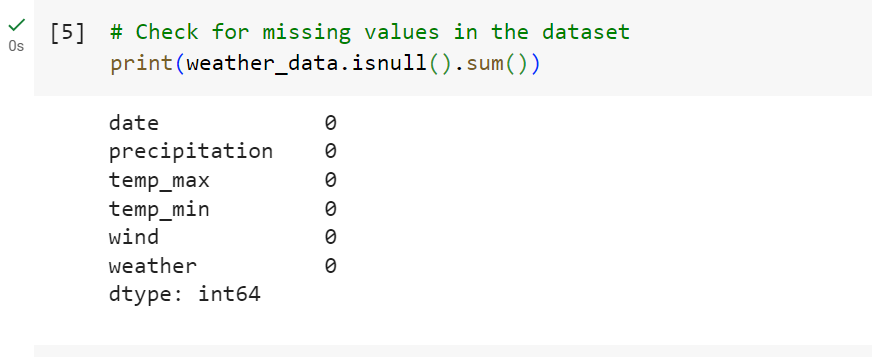
* Tools used in the project are Pandas, Numpy, and Seaborn and Tableau.
* Pandas: Python library for data science, built on NumPy, offering extensive data manipulation tools.
* NumPy: Python module for array operations, including matrix operations and linear algebra functions.
* Seaborn: Python data visualization package based on Matplotlib, known for creating advanced and visually appealing graphics.
* Tableau: Powerful business intelligence tool for convenient and fast data visualization, aiding in exploration and sharing of insights.

**5. DATA PREPROCESSING**

**5.1 Data Cleaning:**

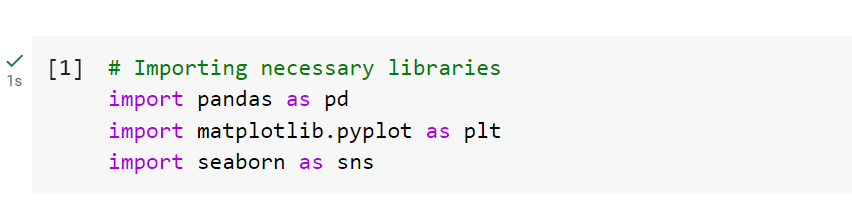
1. Replaced the missing values of attributes such as temp\_max, temp\_min with the most frequent value.

After Replacing, Checking for Missing Values:



The weather dataset that has been loaded is examined for missing values using the code {print(weather\_data.isnull().sum())}. It gives a brief summary of the dataset's completeness by printing the total of the missing values for each attribute. This data is crucial for evaluating the quality of the data and aids in determining the best course of action for handling missing values during data preparation and in this dataset there are no missing values.

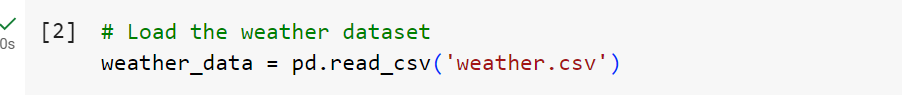
1. Standardized the date to a specific format.

**5.2 Importing necessary libraries:**

Here,

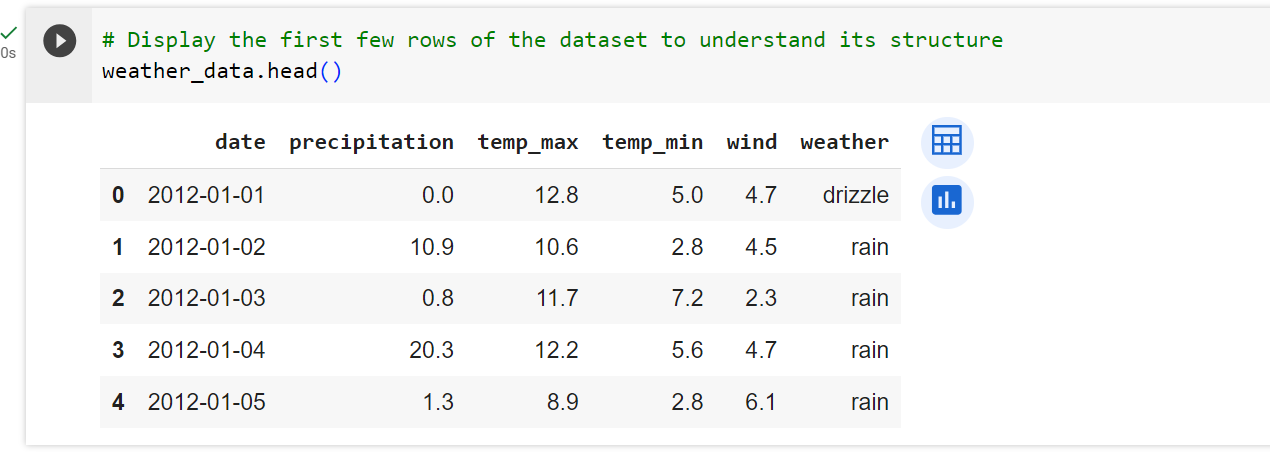
* Pandas (`pd`): Used for data manipulation and analysis.
* Matplotlib (`plt`): A widely-used plotting library for creating static, animated, and interactive visualizations.
* Seaborn (`sns`): A data visualization library based on Matplotlib, providing a high-level interface for drawing attractive and informative statistical graphics.

**5.3 Loading the weather dataset:**



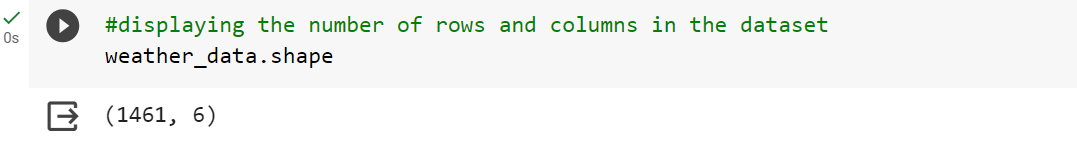
In essence, this method puts the weather dataset into a Pandas dataframe from the given CSV file, allowing for more Python weather data exploration and analysis.

**5.4 Displaying the first few rows of the dataset:**



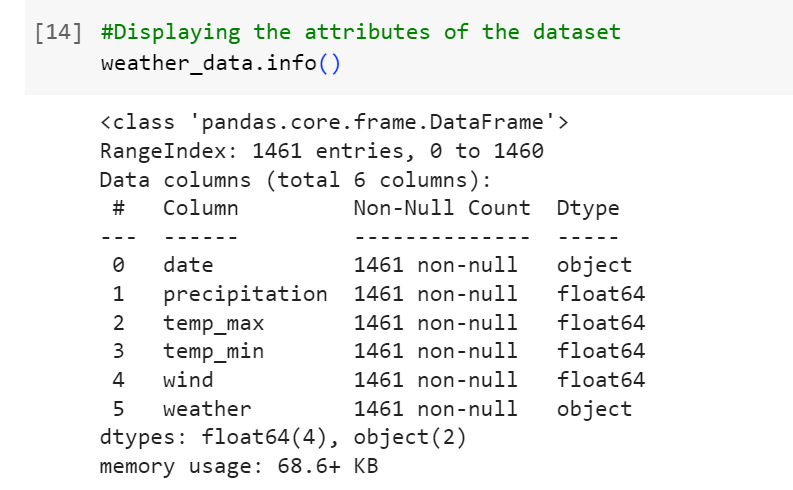
To provide a brief overview of the weather dataset's first rows and enable a rapid comprehension of its composition and organization, the `weather\_data.head()` command is utilized.

**5.5 Displaying the Shape:**



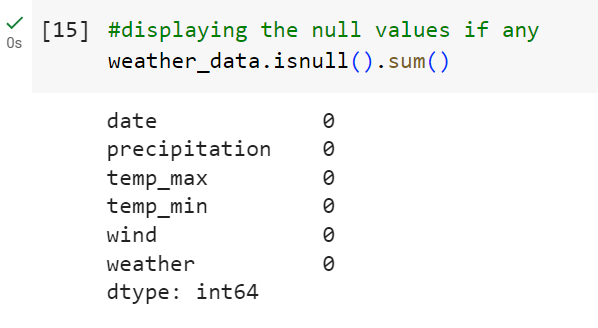
The tuple that represents the weather dataset's dimensions is returned by the code snippet {weather\_data.shape}. A brief summary of the size and organization of the dataset is given by the first value in the tuple, which is the number of rows, and the second value, which is the number of columns. Here, the output is (1461, 6), it means that the dataset contains 1461 rows and 6 columns.

**5.6 Displaying the Attributes:**



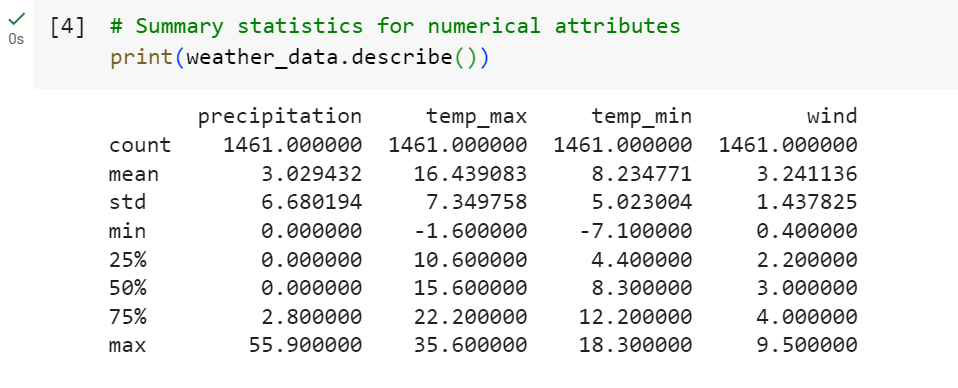
The `info()` method is utilized to present a brief overview of the dataset's characteristics. It offers a brief summary of the composition and comprehensiveness of the meteorological dataset by supplying details on data kinds, non-null counts, and memory utilization.

**5.7 Displaying the Null Values**



The `isnull()` Pandas DataFrame method is used in this bit of code to find missing values in the weather dataset. The `sum()` function that follows adds up all of the null values for every attribute, giving a brief overview of how comprehensive the data is. Understanding any data gaps and selecting suitable handling techniques during data analysis depend heavily on this information. Here there are 0 null values in this dataset.

**5.8 Describing the dataset**

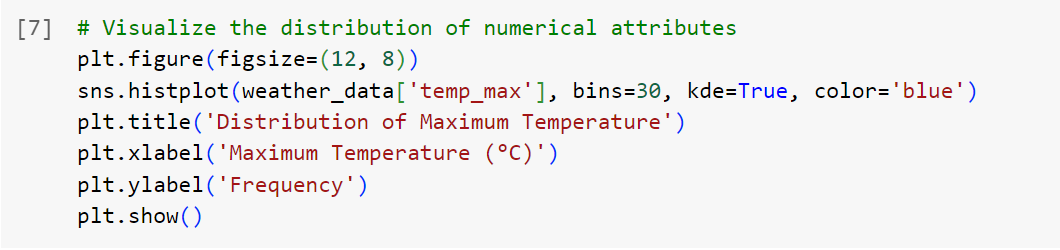


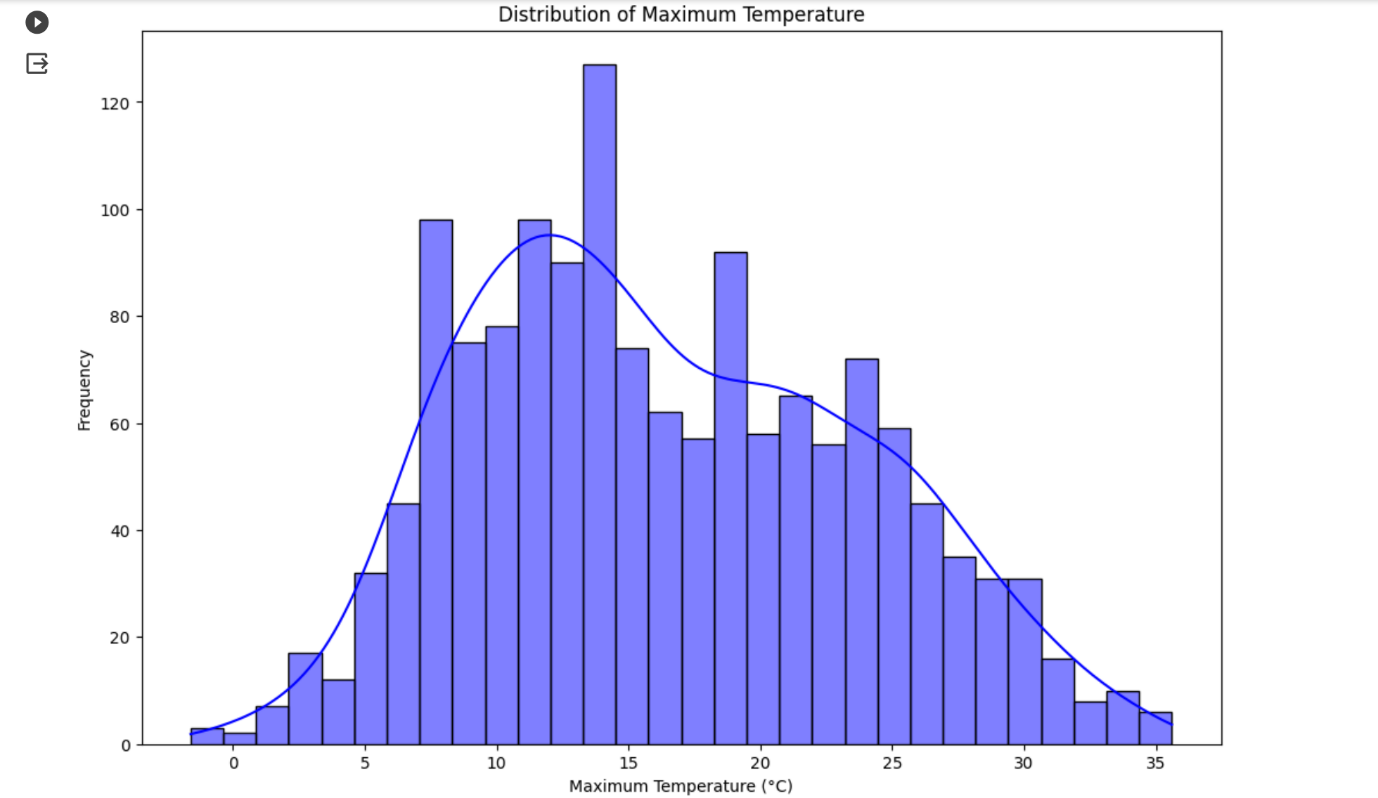
A brief synopsis of the numerical attributes in the weather dataset is given by the `describe()` function. Important statistical data are included, including the mean, standard deviation, minimum, 25th percentile (Q1), median (50th percentile or Q2), 75th percentile (Q3), and maximum quantities. To facilitate preliminary data exploration and comprehension, this summary provides a brief synopsis of the numerical variables' distribution, spread, and central tendency.

**6. EXPLORATORY DATA ANALYSIS(EDA)**

Exploratory Data Analysis (EDA) is an essential initial step in meteorological analysis that identifies patterns and insights in a dataset. From January 1, 2012, to December 31, 2015, our three-year weather dataset includes daily records of critical parameters like date, precipitation, temperature max and minimum, wind speed, and weather conditions. We want to use EDA to find patterns, variances, and connections in the data so that we can fully comprehend the weather dynamics over this long period of time.

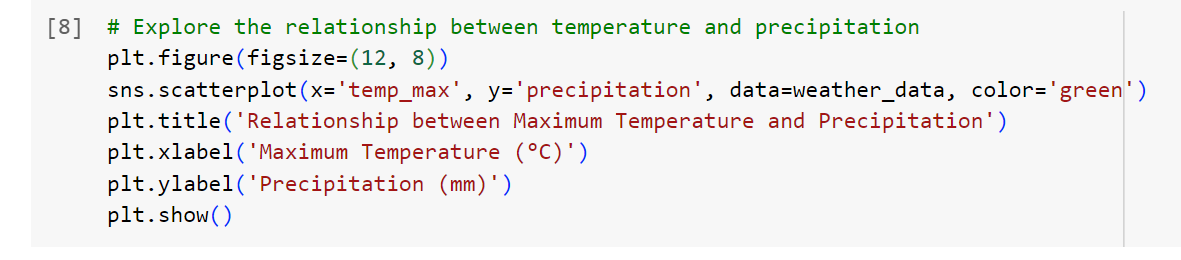
**6.1 Visualizing Temperature Distribution Over Time:**

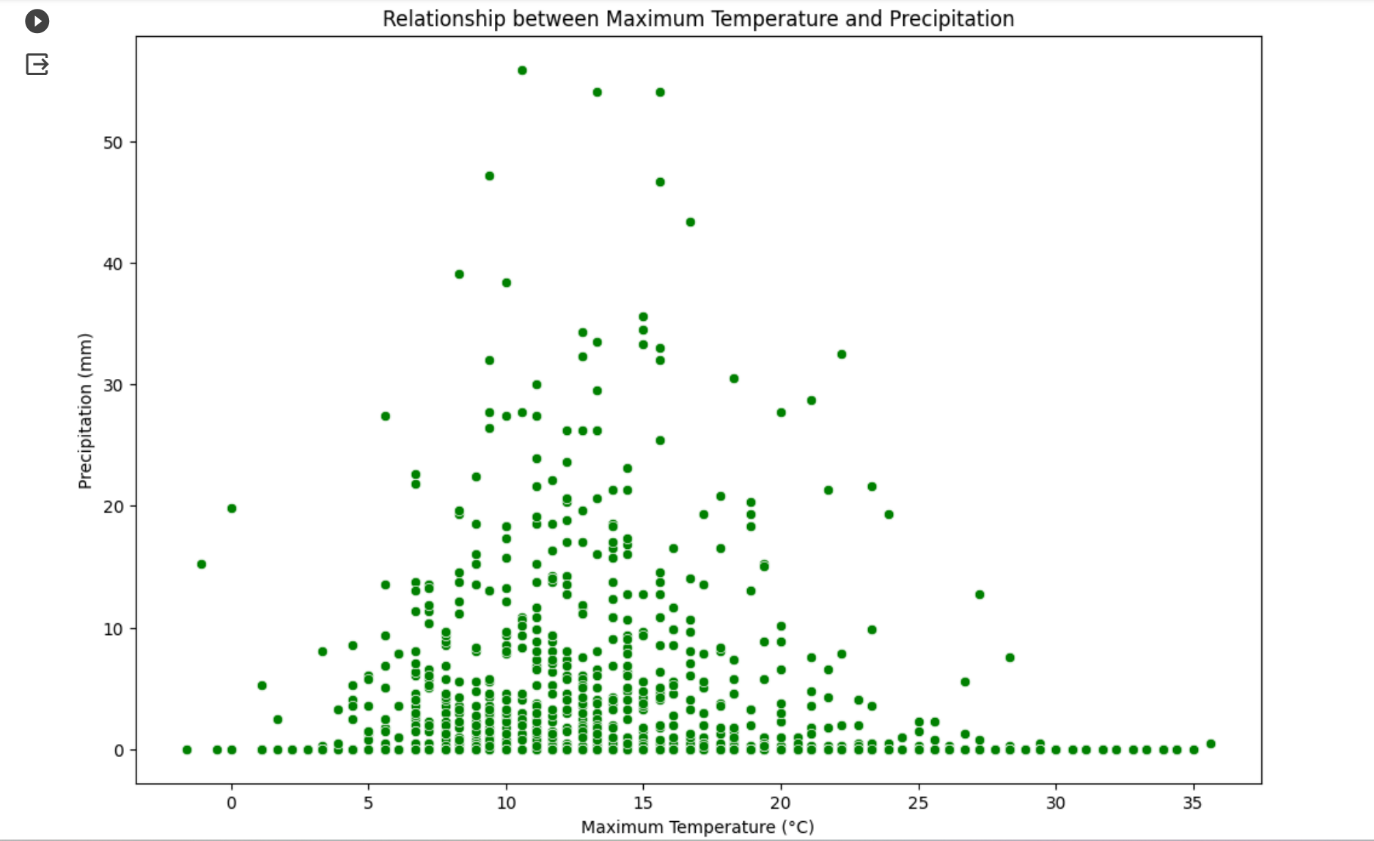




This is a histogram that shows the distribution of the greatest temperature ({temp\_max}) in the weather dataset using Matplotlib and Seaborn. The number of intervals in the histogram is specified by the `bins=30} parameter, and a kernel density estimate is superimposed by `kde=True}. The generated figure helps to identify temperature patterns in the dataset by offering insights into the frequency distribution of maximum temperatures.

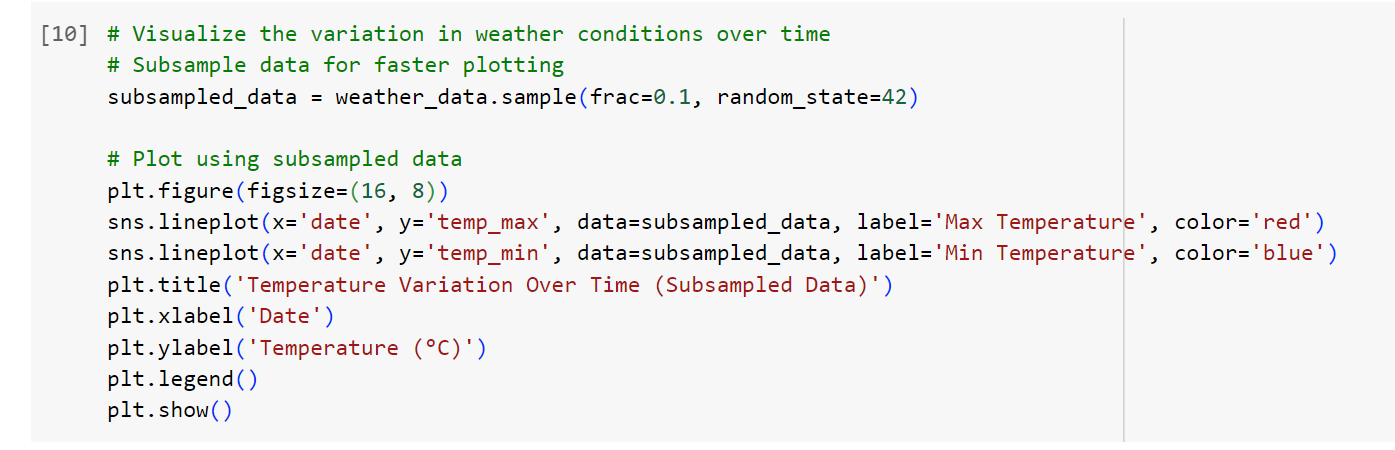
**6.2 Exploring the Relationship Between Attributes:**

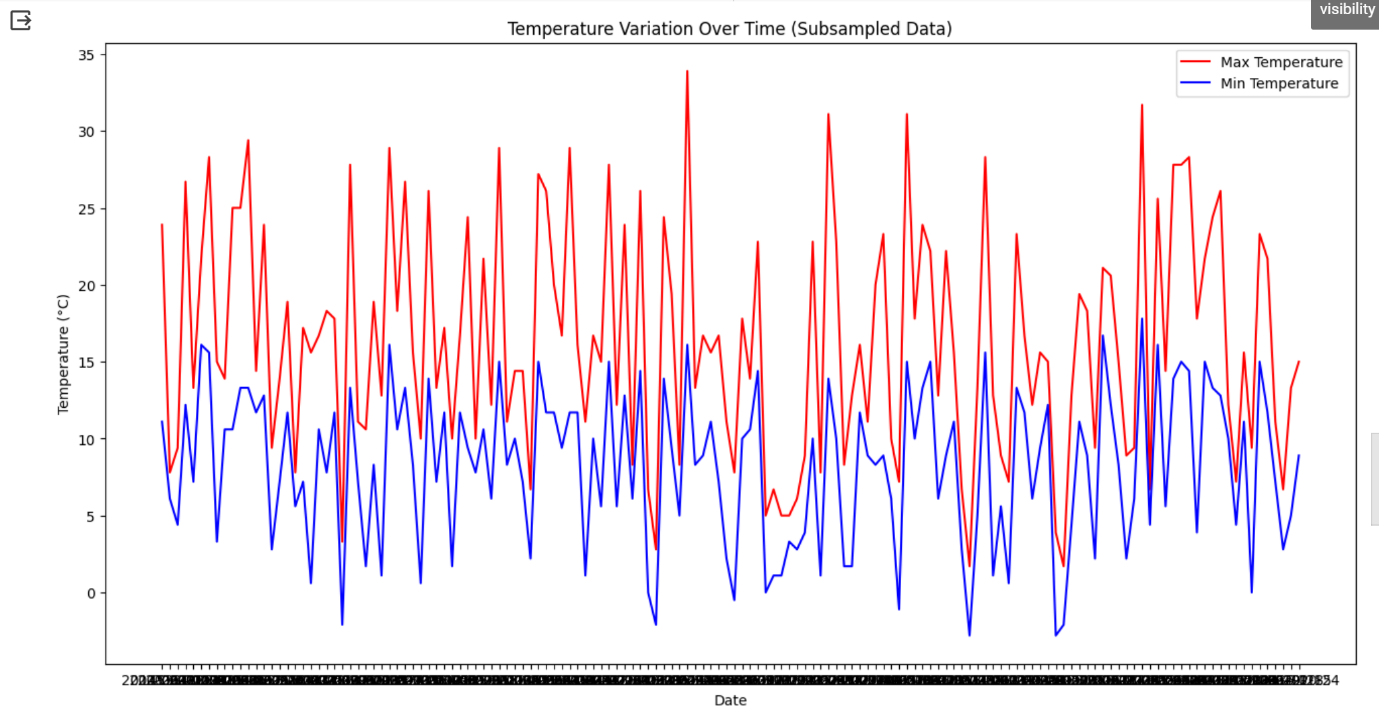
****

****

This is a scatter plot examining the association between maximum temperature ({temp\_max}) and precipitation in the weather dataset using Seaborn and Matplotlib. The graphic helps identify any trends or connections in the data by illuminating how changes in maximum temperature relate to various precipitation levels. The data points are shown by the green dots, where the y-axis shows the amount of precipitation in millimeters and the x-axis shows the maximum temperature in degrees Celsius.

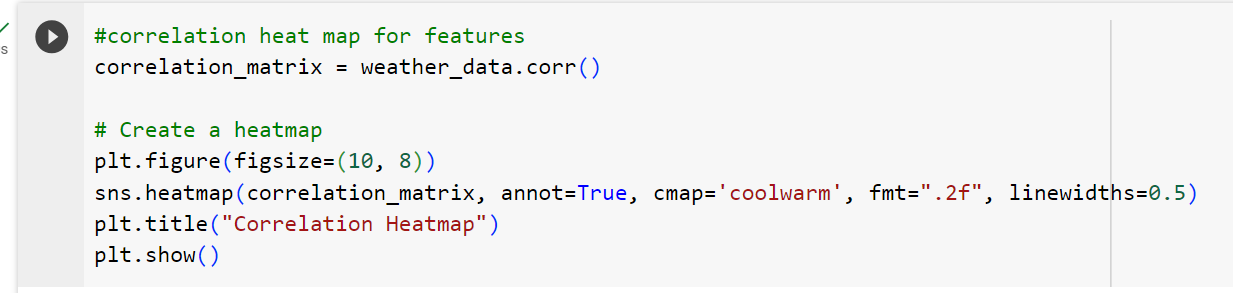
**6.3 Exploring Temperature Trends Over Time:**

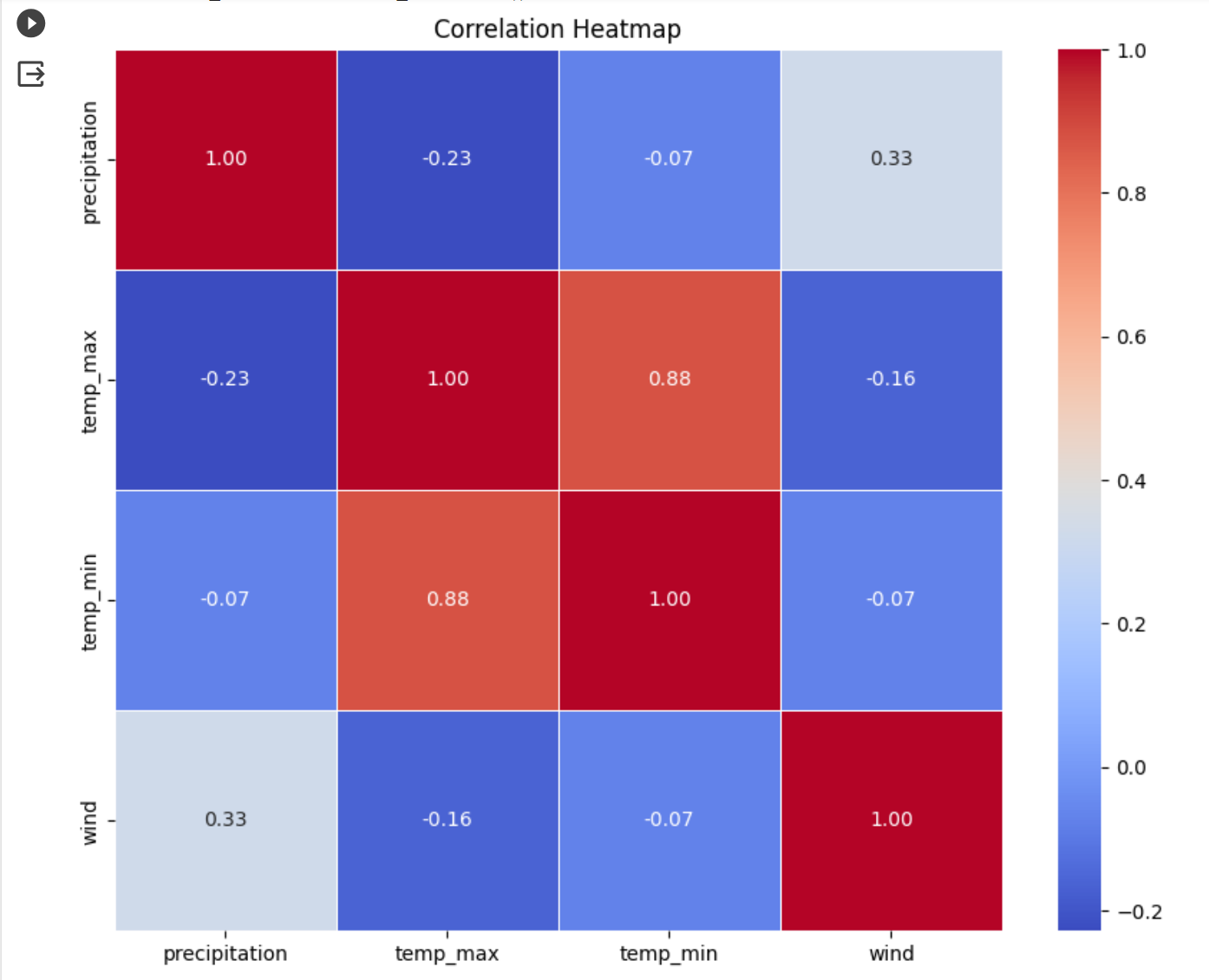
****

****

To speed up plotting in this visual analysis of meteorological data, a subset of the dataset is used. The resultant line plot provides a concise but informative depiction of temperature dynamics from the larger dataset by showing variations in both maximum and minimum temperatures across time. The trends for the highest and lowest temperatures are shown by the red and blue lines, respectively.

**6.4 Correlation Analysis of Weather Features:**

****

****

The pairwise correlation coefficients between various meteorological features are computed and captured in the `correlation\_matrix`. The intensity and direction of these relationships are graphically represented in the heatmap, which was made with Seaborn ({sns.heatmap}). Cooler hues (blues) denote negative correlations, whereas warmer hues (reds) suggest positive correlations. Precise correlation values are provided by annotations, which shed light on how the various weather parameters are related to one another.

### **7. HYPOTHESES:**

1. **Does the Type of Weather Influence the Occurrence of Extreme Temperatures?**

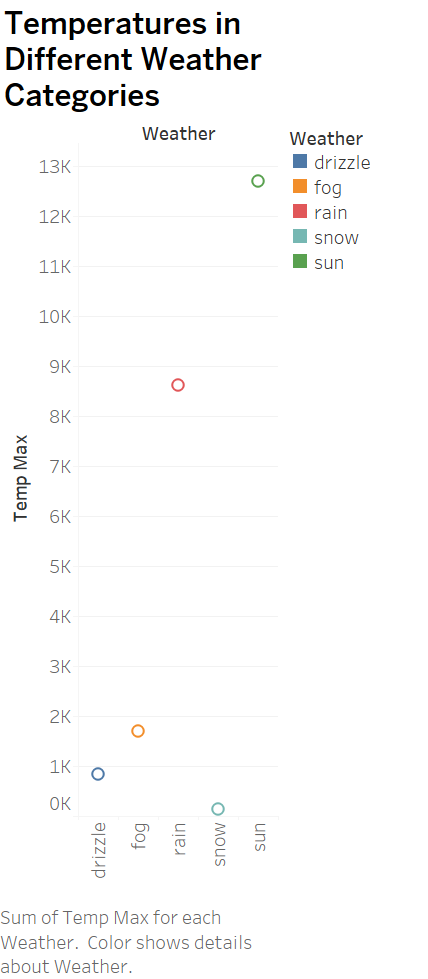
* Hypothesis: Extreme temperatures, either hot or cold, are more likely to occur during specific weather conditions, such as heavy rainfall or snowfall..
* Justification: The hypothesis posits that extreme temperatures, either hot or cold, may be more prevalent during specific weather conditions. This exploration aims to identify correlations between weather categories and the occurrence of intense temperature events, providing insights into the influence of weather patterns on extreme temperatures.

1. **How does temperature vary over the years?**
   * Hypothesis: There is a discernible trend in the variation of temperature (both maximum and minimum) over the years.
   * Justification: The hypothesis posits that an observable trend exists in the fluctuation of temperatures over the years. By examining historical temperature data, we anticipate patterns in both maximum and minimum temperatures, revealing potential long-term climatic shifts or consistent seasonal variations.
2. **Do Specific Weather Categories Exhibit Higher Precipitation Levels?**

* Hypothesis: Certain weather categories are associated with elevated levels of precipitation.
* Justification: Different weather conditions may contribute to varying levels of precipitation. We'll categorize weather conditions and examine their correlation with precipitation.

### **8. HYPOTHETICAL VISUALIZATIONS**

**1. Interactive Scatter Plot - Weather Influence on Extreme Temperatures:**

* *Elementary Perceptual Tasks:* Position (for scatter points), Color (for weather categories), Size (for temperature intensity).
* *Visualization Format:* Scatter plot with interactive elements (hover effects).
* *Spatial Design:* Two-dimensional space with weather categories on the x-axis and temperature on the y-axis.
* *Color Rationale:* Different colors represent different weather categories, facilitating the identification of specific conditions. The size of points indicates the intensity of extreme temperatures.

**Explanation**:

This scatter plot visualizes the relationship between weather conditions and the occurrence of extreme temperatures. Each point on the plot represents a data point, with weather categories on the x-axis, temperatures on the y-axis, and different colors indicating specific weather conditions. The size of the points reflects the intensity of extreme temperatures. The scatter plot allows for an immediate assessment of whether extreme temperatures are more prevalent during certain weather conditions. Hover effects enhance the interactivity, providing users with detailed information about specific data points, fostering a deeper understanding of the relationship between weather and extreme temperatures.

**Result**:

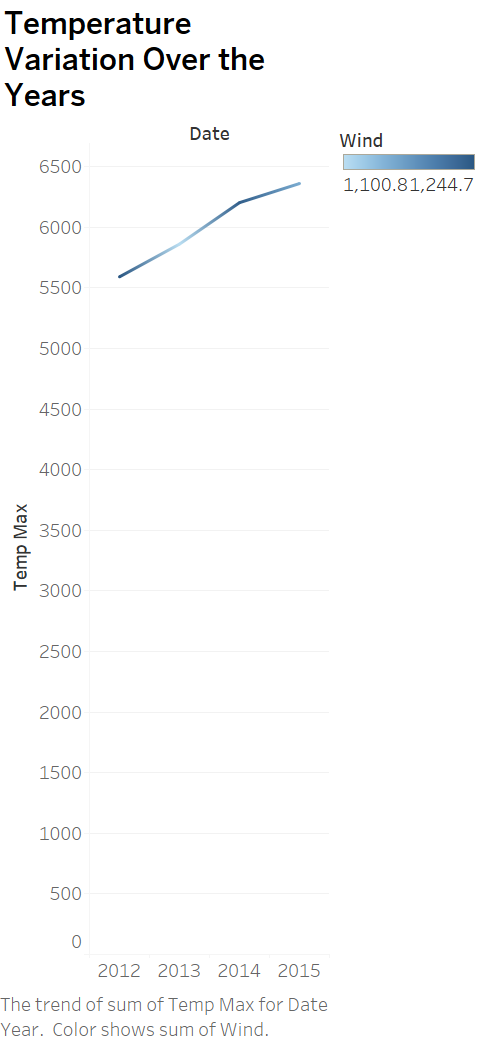
Yes, the data analysis affirms the hypothesis, as observed trends indicate that during sunny weather conditions, temperatures tend to be higher. Conversely, lower temperatures coincide with rainy, foggy, drizzly, and snowy weather respectively. This alignment underscores the influence of specific weather patterns on the occurrence of extreme temperatures, validating the initial hypothesis.

**2. Line Chart - Temperature Variation Over the Years:**

* *Elementary Perceptual Tasks:* Position (for line points), Color (for temperature categories).
* *Visualization Format:* Line chart.
* *Spatial Design:* Two-dimensional space with time (years) on the x-axis and temp\_max on the y-axis.
* *Color Rationale:* Different colors represent maximum and minimum temperatures, allowing for a clear visual representation of temperature trends over the years.

**Explanation**:

This visualization explores the hypothesis of how temperature varies over the years. Using a line chart, you can plot the trend of both maximum and minimum temperatures across different years. Each line represents a temperature category, with colors distinguishing between maximum and minimum temperatures. This allows for a comprehensive understanding of the temperature dynamics and trends observed over the specified time period.

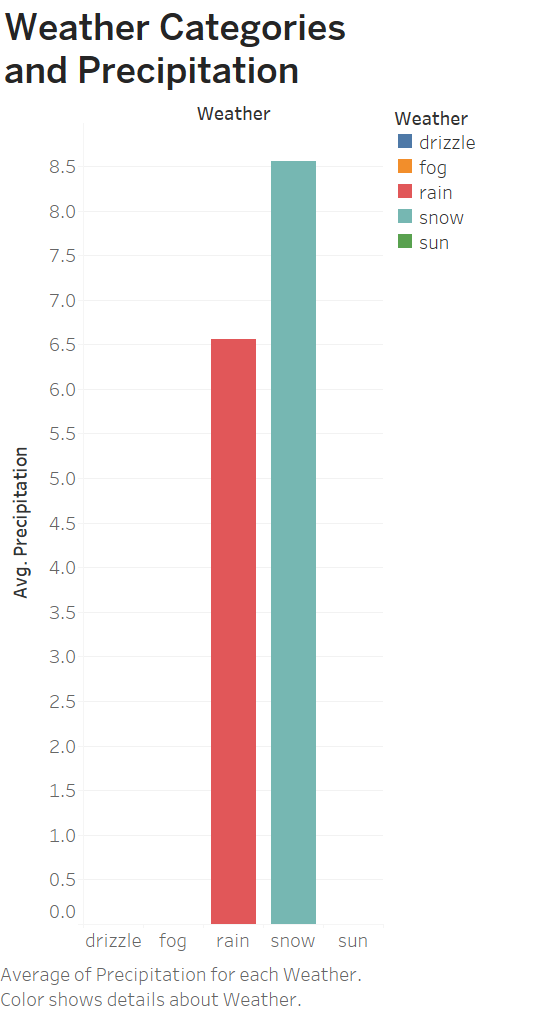


**Result**:

The second hypothesis posits that there is a discernible upward trend in temperature, both maximum and minimum, from the years 2012 to 2015. This implies a gradual increase in temperature over this time period. By analyzing historical temperature data, patterns and trends are anticipated, indicating a potential shift towards warmer climatic conditions or consistent seasonal variations during this specific timeframe.

**3. Bar Chart - Weather Categories and Precipitation:**

* Elementary Perceptual Tasks: Length (for bar heights), Color (for weather categories).
* Visualization Format: Grouped bar chart.
* Spatial Design: Two-dimensional space with weather categories on the x-axis and average precipitation on the y-axis.
* Color Rationale: Distinct colors for each weather category facilitate easy differentiation, emphasizing the variations in precipitation levels.

****

**Explanation**:

The bar chart addresses the question of whether specific weather categories exhibit higher precipitation levels. In this design, elementary perceptual tasks leverage length to convey variations in precipitation, with bar heights representing average precipitation levels. The use of distinct colors for each weather category enhances the visualization, aiding easy differentiation. The spatial design employs a two-dimensional layout, with weather categories on the x-axis and average precipitation on the y-axis, facilitating a straightforward comparison between different weather conditions. It offers a clear overview of how weather conditions relate to precipitation, allowing for easy visual comparison between categories.

**Result**:

Yes, **Weather Categories such as rain and snow exhibit higher precipitation levels.**

**9. CONCLUTION:**

**In the culmination of this project, the exploration and analysis of the Weather Prediction Dataset have yielded valuable insights into the complex interplay of meteorological variables. Through data visualizations and hypotheses testing, I have identified correlations between weather attributes, validated assumptions, and uncovered patterns influencing temperature and precipitation. Leveraging Python and visualization libraries like Matplotlib and Seaborn, I developed interactive and informative visualizations to enhance understanding. This project not only contributes to the domain of weather prediction but also underscores the significance of data-driven methodologies in deciphering and preparing for diverse weather scenarios. As I navigate the dynamic landscape of atmospheric conditions, the findings from this project underscore the potential of data analytics to empower informed decision-making and enhance my readiness to face the challenges posed by the ever-changing weather.**

**10. REFERENCES:**

1. J. D. Hunter, "Matplotlib: A 2D Graphics Environment," Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007, <https://doi.org/10.1109/MCSE.2007.55>

2. M. Waskom et al., "seaborn: statistical data visualization," 2022. [Online]. Available: <https://seaborn.pydata.org/>

3. A. Ravi, "Weather Prediction Dataset," Kaggle, 2019. [Online]. Available: <https://www.kaggle.com/datasets/ananthr1/weather-prediction>

4. Python Software Foundation, "Python Language Reference, version 3.10," 2023. [Online]. Available: <https://www.python.org/doc/3/>