**DATA ANALYST INTERN PROJECT PHASE – 1**

**Name: Sonam Singh**

**Project: 2**

**Topic: Weather Analysis**

**Company: Nexus Info**

**Date: 20/08/2024**

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| SR. NO. | INDEX | PG. NO. |
| 1 | **Introduction** | **3-4** |
| 2 | **Data Exploration** | **5-7** |
| 3 | **Basic Analysis** | **8-12** |
| 4 | **Advanced Analysis and Visualizations** | **13-16** |
| 5 | **Correlation and Regression Analyses** | **17-21** |
| 6 | **Conclusion** | **22** |

**INTRODUCTION**

This project involves the analysis of a weather dataset, named ‘Weather Prediction’ by Ananth R., obtained from the Kaggle website. Kaggle is a popular online platform that hosts data science and machine learning competitions, provide datasets, and offers a collaborative environment for data scientists and analysts to share code, solutions, and insights. The weather analysis aims to uncover patterns and insights that could be useful for predicting weather conditions.

**OBJECTIVES OF THE PROJECT**

The main purpose of this project is –

1. To preprocess and analyze the weather dataset using Python to obtain the basic analysis, check for outliers and other inconsistencies in the data.
2. To perform advanced data analysis and create different visualizations on the weather dataset using Power BI.
3. To perform correlation analysis to identify relationships between different weather parameters.
4. And to implement regression analysis to predict one weather parameter based on others.

**DATA METHODOLOGY**

The following steps outline the methodology adopted to analyze the Weather dataset –

1. **Data Loading:** The Weather dataset was downloaded from Kaggle then the CSV file was loaded in Python. The dataset was converted into a Data Frame for easier manipulation and analysis.
2. **Data Exploration:** Basic descriptive statistics were calculated to understand the central tendencies, dispersion, and shape of the dataset’s distribution; and the dataset was checked for any missing values or inconsistencies that could affect the analysis.
3. **Data Analysis:** Various plots such as histograms, box plots, & scatter plots were generated to visualize the distributions and identify any relationships between the numerical features in the dataset.
4. **Data Visualization with Power Bi:** The processed data from Python was imported into Power Bi to create interactive dashboards. Interactive dashboards and visualizations were created to facilitate a deeper understanding of the data.
5. **Correlation and Regression Analysis:** The correlation and regression analyses were implemented using Python to identify the relationships between the variables.
6. **Documentation:** Each step of the analysis was documented, including the Python code, Power BI visualizations, and insights derived from the data. This documentation serves as a comprehensive guide to the project’s workflow and findings.

**DATA OVERVIEW**

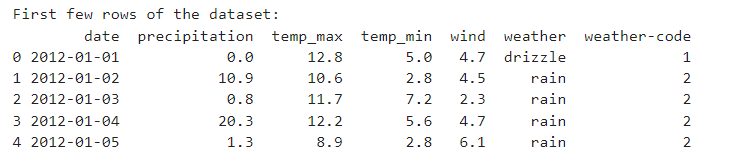
The ‘Weather Prediction’ dataset includes daily weather observations over a four-year period, from January 2012 to December 2015, totaling 1461 entries. The data contains following variables –

* Date: The date of the observation, formatted as dd-mm-yyyy.
* Precipitation: The amount of rainfall or other precipitation measured on that day.
* Maximum Temperature: The highest temperature recorded on that day.
* Minimum Temperature: The lowest temperature recorded on that day.
* Wind: The wind speed or other wind-related information.
* Weather: The type of weather observed, categorized into five distinct conditions – drizzle (encoded as 1), rain (2), sun (3), snow (4), and fog (5).

**DATA EXPLORATION**

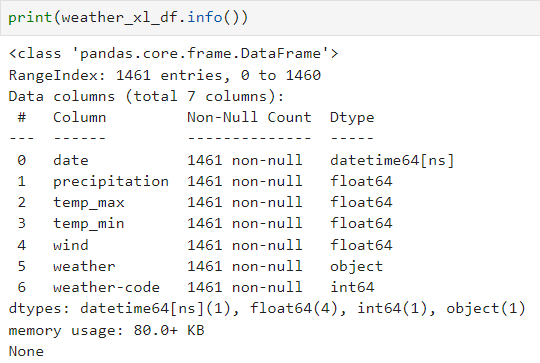
**LOADING THE DATASET**

The weather dataset was accessed from Kaggle website. The CSV file was loaded in Python. The method ‘.head()’ was used to view first few rows of the dataset to get a clear picture of the dataset. The output for the same is –



Figure

The dataframe was also created in Python for easier manipulation. The ‘.info()’ method in Python is used with pandas DataFrames to display a concise summary of the DataFrame. It provides essential information like data types of each column, about the DataFrame. The output for the same is –

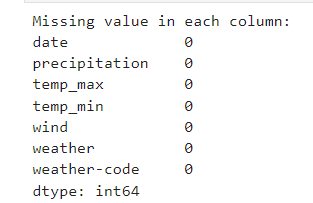


Figure

Thus, we can see that ‘weather’ column has object type, ‘weather-code’ contains integers as codes specified to weather conditions, and the rest are float variables.

**CHECKING FOR MISSING VALUES**

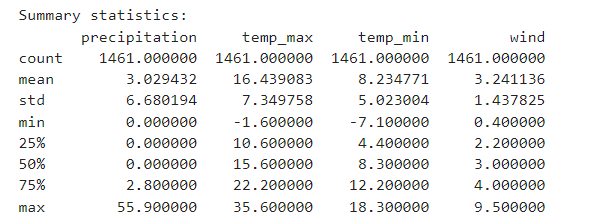
To ensure the integrity of the analysis, we first checked for missing values using the ‘isnull().sum()’ method in Python. This method provides a count of missing values in each feature. Fortunately, the weather dataset contained no missing values, allowing us to proceed with the analysis without the need for data imputation. The output for the same obtained from Python is –



Figure

**SUMMARY STATISTICS**

After confirming that the dataset contains no missing values, the next step was to generate summary statistics for each numerical feature in the dataset. This was done using the ‘describe()’ method in Python, which provides key statistics such as the mean, standard deviation, minimum, and maximum values, as well as, the 25th, 50th(median), and 75th percentiles. The summary statistics obtained from Python is given below –



Figure

The summary statistics provide insights into the overall distribution of the data. From the above figure, we can interpret that –

* **Precipitation:** The dataset shows an average precipitation of 3.03 units with a high variability (standard deviation: 6.68). Most days had no precipitation since median is 0, but there were occasional heavy rainfall events (maximum value: 55.9).
* **Maximum Temperature:** The average maximum temperature is 16.44⁰C with moderate variability of 7.40⁰C. Temperatures ranged from -1.6⁰C to 35.6⁰C, indicating a mixture of cold and warm days. The maximum recorded temperature of 35.6⁰C indicates that some days reached quite warm conditions.
* **Minimum Temperature:** The average minimum temperature is 8.23⁰C with a standard deviation of 5.02⁰C, suggesting some variability but not as much as the maximum temperature. Minimum temperatures ranged from -7.1⁰C to 18.3⁰C, showing occasional freezing days.
* **Wind:** The average wind speed is 3.24m/s with a moderate variability (since std: 1.44m/s). Wind speeds ranged from 0.4 to 9.5 units, indicating mostly mild to moderate winds with some stronger gusts.

**General Insights:**

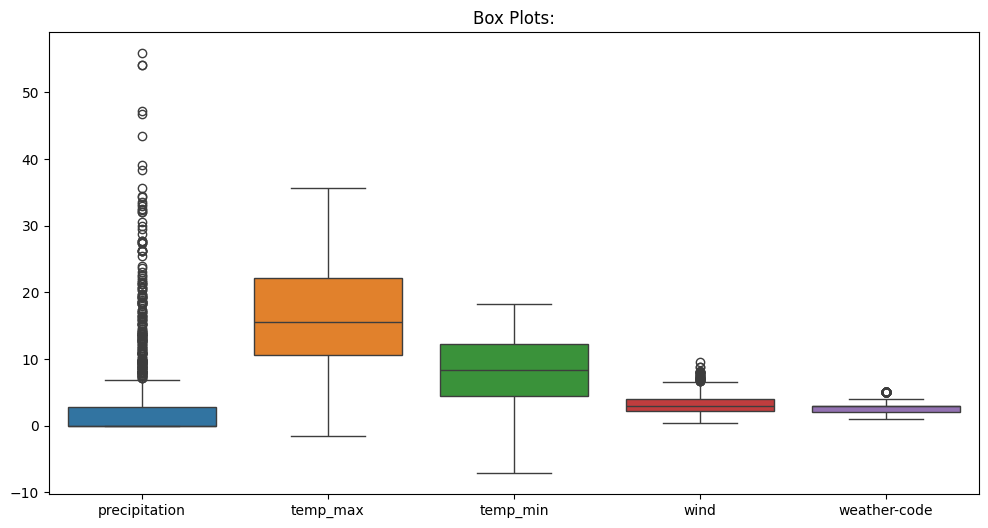
* The precipitation data shows a significant number of dry days, but when it rains, it can rain heavily.
* The temperature data indicates that there is a fairly even distribution of warmer and colder days, with occasional extreme cold and hot days.
* The wind data suggests that most days have a mild to moderate wind, with some instances of stronger winds.

**BASIC ANALYSIS**

This section involves summarizing the main characteristics of the dataset, often with visual methods, to understand the data’s underlying structure, detect anomalies and discover patterns. For this project, analysis was performed on the Weather dataset using Python to gain insights into the distribution and relationships of the various features.

**BOX PLOTS**

Box plots helps in identifying the outliers which may present in any of the features of the dataset.



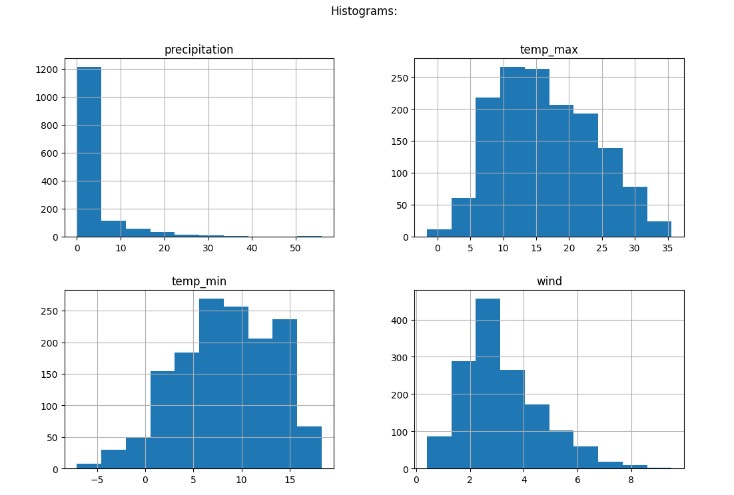
Figure

From the above figure, we interpret that –

* **Precipitation:** The box plot for precipitation shows a large number of outliers, indicating that while most precipitation values are low, there are several instances of extremely high precipitation. The IQR (Interquartile Range) is relatively small, and the median is close to the lower bound of the box, indicating that precipitation is typically low.
* **Maximum Temperature:** The box plot for temp\_max shows a more symmetric distribution. The temperature ranges between 0 and 35 degrees, with a median of around 15 degrees. There are no significant outliers, indicating a relatively consistent range of maximum temperature.
* **Minimum Temperature:** The variable temp\_min shows a similar distribution as temp\_max, but is generally lower, ranging from -5 to 15 degrees. The median is around 5 degrees. There are also no significant outliers here.
* **Wind:** The wind data shows a tighter distribution with fewer outliers. Most wind speed fall between 0 and 10 units, with a median around 2 units. The presence of a few outliers indicates some instances of higher wind speeds.
* **Weather-code:** The variable weather-code is an encoded variable with five categories of weather types. The box plot shows that these categories are distributed across different weather conditions, with some outliers indicating less common weather patterns.

**HISTOGRAMS**

The histograms provide a view of the frequency distribution of each feature in the Weather dataset.



Figure

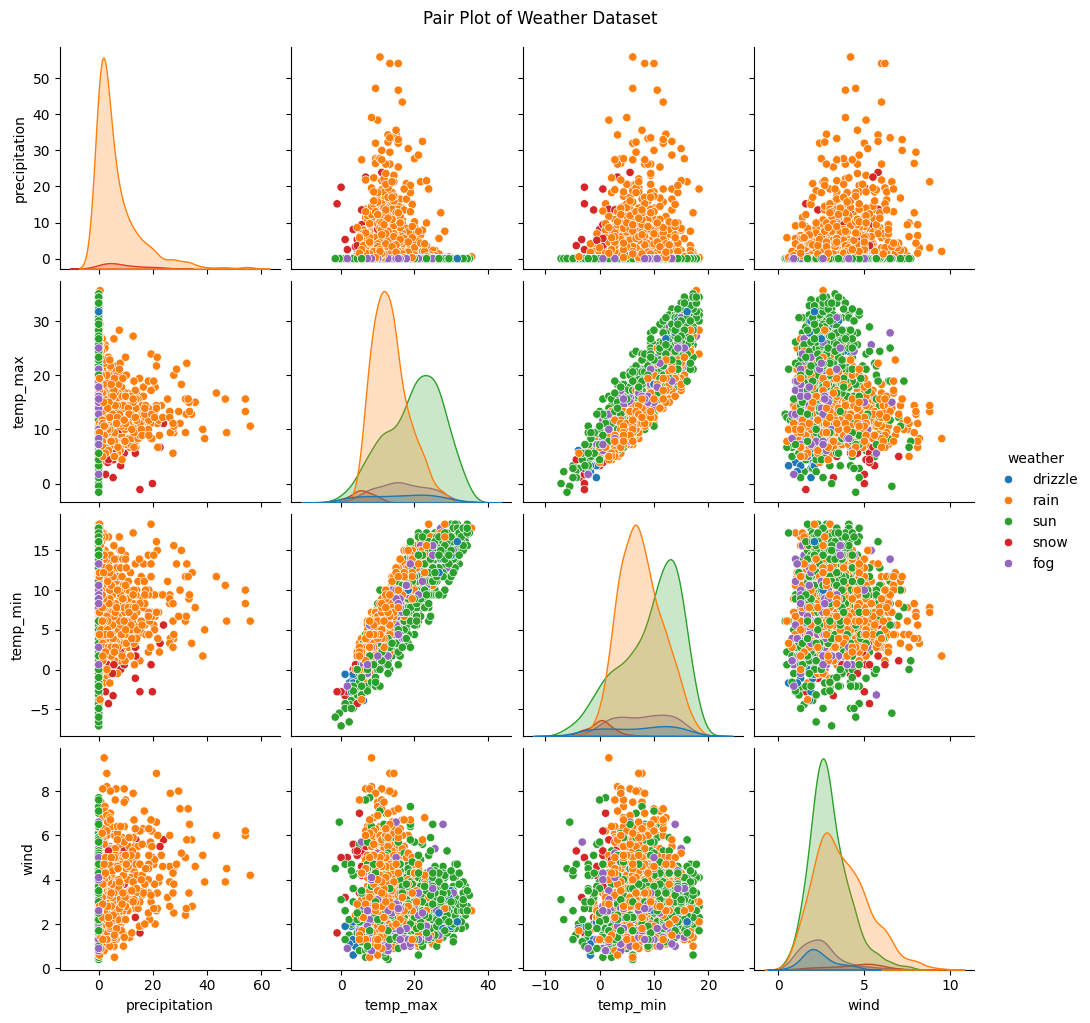
From the above figure, we can interpret that –

* **Precipitation:** The histogram for precipitation is highly right-skewed, with the majority of observations showing low precipitation. This skewness indicates that while precipitation is typically low, there are occasional instances of much higher precipitation.
* **Maximum Temperature:** The histogram for temp\_max shows a normal distribution, with most observations centered around 15 degrees. This suggests a typical bell-curve temperature distribution.
* **Minimum Temperature:** Similarly, the temp\_min histogram also shows a normal distribution, with most observations between 5 and 10 degrees.
* **Wind:** The histogram for wind shows a right-skewed distribution, indicating that while most wind speeds are low, there are occasional instances of higher wind speeds.

**PAIR PLOT**

The pair plot shows the relationships between each pair of variables in the weather dataset, with scatter plots on the lower triangle, KDE plots on the diagonal, and scatter plots again on the upper triangle. From the pair plot of weather dataset features, fig. 8, we can see that –

* **Precipitation:** Generally, precipitation does not seem to have a strong linear relationship with other variables, as indicated by the scatter plots which show a wide spread. However, some clustering is visible in the low precipitation range, indicating certain weather conditions are associated with lower precipitation.
* **Maximum & Minimum Temperature:** There’s a clear positive correlation between maximum temperature and minimum temperature, which is evident from the linear clustering of points in their scatter plots. Higher maximum temperatures are generally associated with higher minimum temperatures.
* **Wind:** Wind does not show a strong correlation with other variables. Its scatter plots are more dispersed.

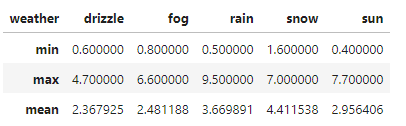


Figure

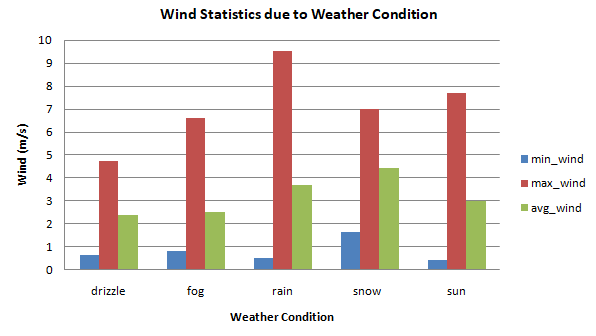
* **Weather Conditions:** The scatter plots show that different weather condition tend to cluster in different areas of the plots, indicating certain weather patterns are associated with specific ranges of temperature, precipitation and wind.

**WIND AND WEATHER**

To examine the relationship between wind speed and the weather condition, clustered bar graph was plotted in Python after distinguishing the dataset according to the weather types. The output is given below –



Figure

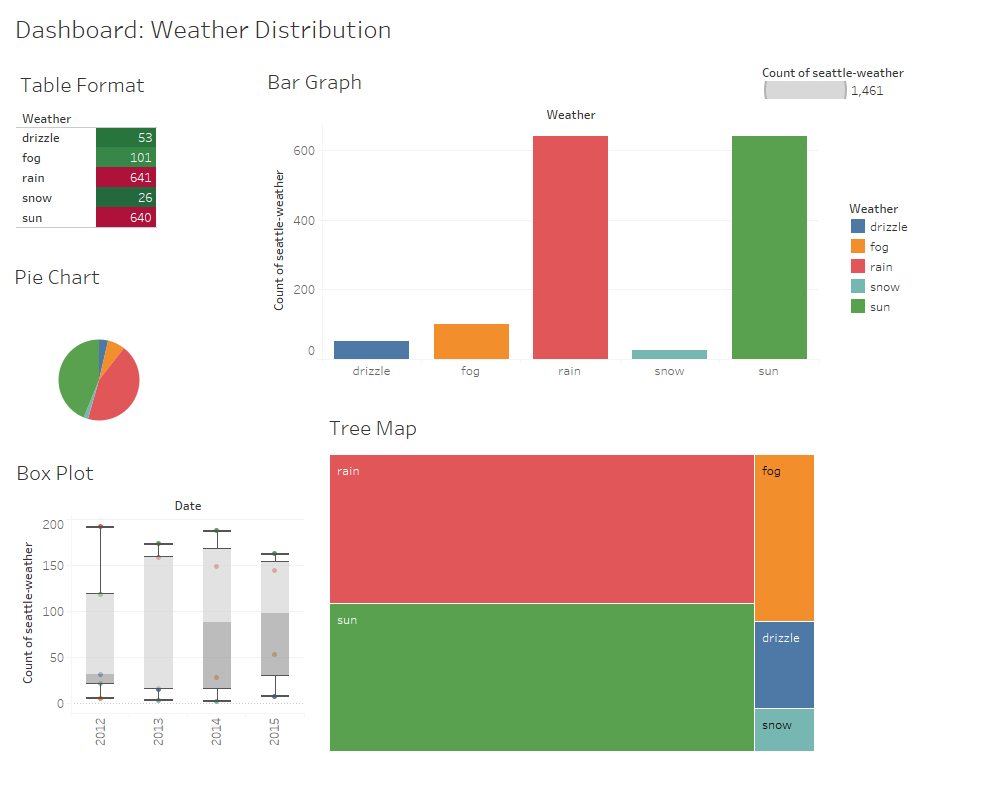


Figure

The above bar graph visually represents the minimum, maximum, and average wind speeds for each weather condition. ‘Drizzle’ has relatively low wind speeds with a max around 4.7m/s. ‘Fog’ shows slightly higher wind speeds than drizzle, with a maximum of 6.6m/s. ’Rain’ exhibits the highest maximum wind speed at 9.5m/s, indicating stronger winds during rain. ‘Snow’ also has significant wind speeds but lower than rain, with a maximum of 7m/s. ‘Sun’ has a moderate wind speed profile, with a maximum of 7.7m/s. The wind speed tends to vary significantly depending on the weather condition. Rain and snow tend to have higher average wind speeds while, drizzle and sun have lower wind speeds.

**ADVANCED ANALYSIS & VISUALIZATIONS**

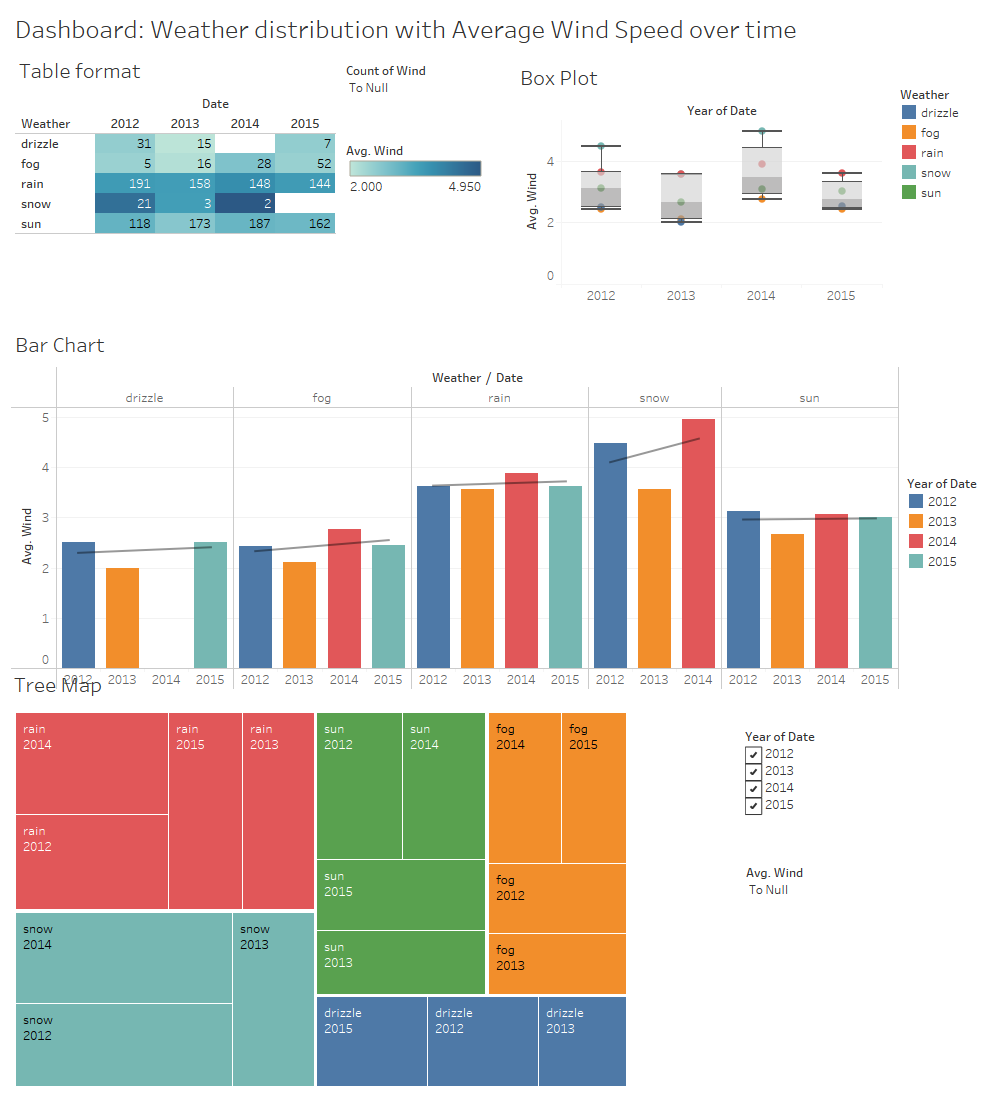
**WEATHER DISTRIBUTION**

****

Figure

The above dashboard helps in identifying the weather distribution in the dataset. The data indicates that ‘Rain’ and ‘Sun’ are the most frequent weather conditions, each occurring over 600 times, which can be seen in Pie chart also, while ‘Drizzle and ‘Snow’ are the least frequent. The box plot shows that weather conditions over the years shows variability, with ‘Rain’ being consistent across the years, and the same is seen in the Tree map visualization.

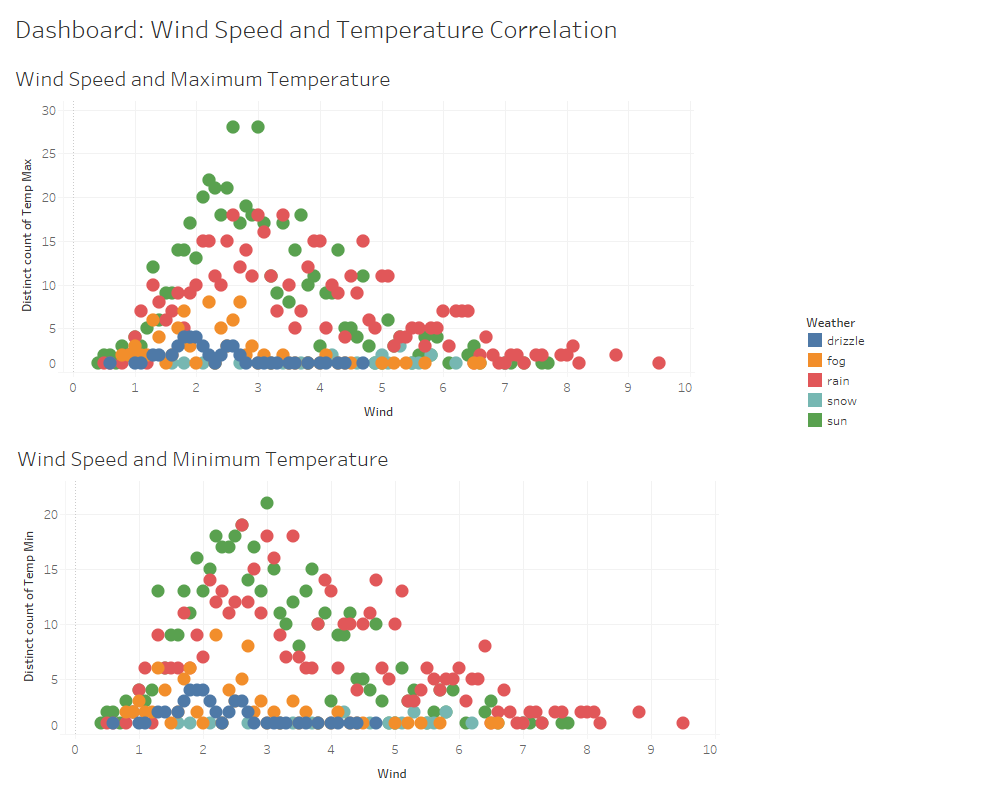
**WEATHER DISTRIBUTION WITH AVERAGE WIND SPEED OVER TIME**



Figure

In fig.11, the box plot shows that ‘Rain’ & ‘Sun’ not only dominate in frequency but also show consistent wind speeds over the years. ‘Snow’ in 2014 shows a peak in wind speed. The bar chart shows that over the time, the average wind speed has gradually increased, particularly during snowy & rainy weather and the tree map highlights that 2014 had the most significant wind events. The temporal trends shows that there is an observable increase in average wind speeds over the years, particularly in adverse weather conditions like ‘snow’ and ‘rain’.

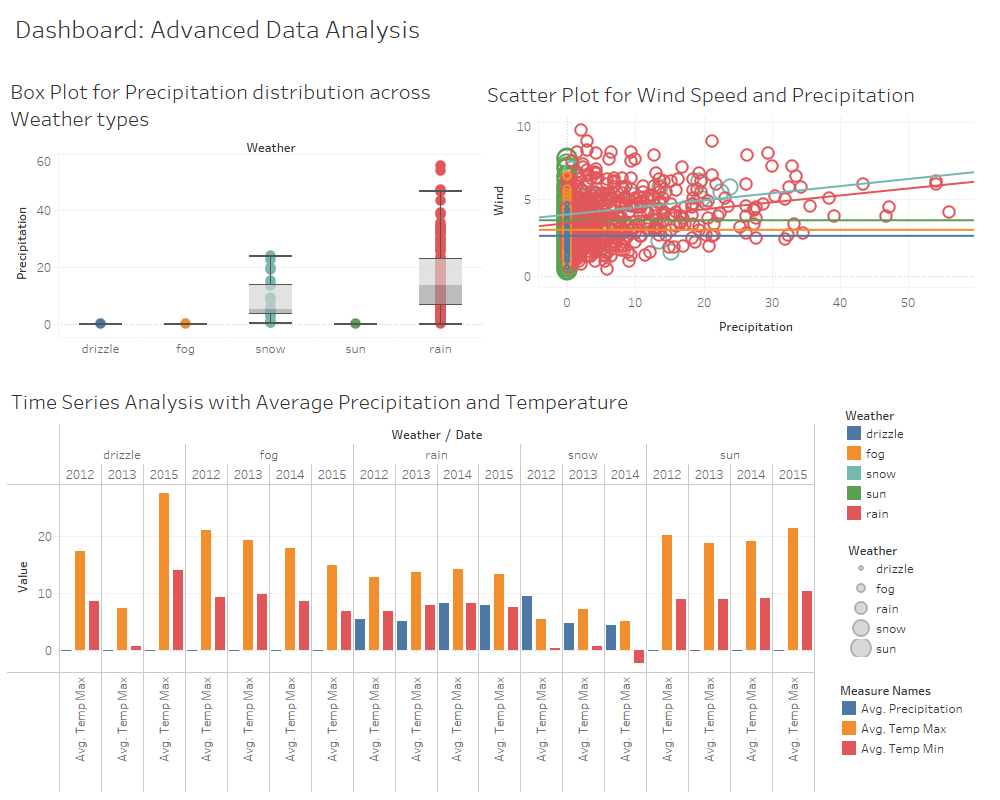
**WIND SPEED AND TEMPERATURE CORRELATION**



Figure

* The scatter plot for wind speed and max. temperature reveals that maximum temperatures are generally higher when wind speeds are lower. ‘Sun’ and ‘Rain’ dominate the higher temperature ranges, while ‘Snow’ and ‘Fog’ appear at low temperatures, often corresponding with lower wind speeds.
* Similar to the max. temperature correlation, the scatter plot for wind speed and min. temperature also shows an inverse relationship with wind speed showing ‘Sun’ and ‘Rain’ more prevalent at higher temperatures.

**ADVANCED DATA ANALYSIS**



Figure

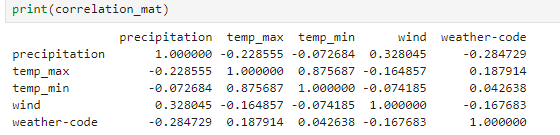
* Box plot for Precipitation distribution across Weather types shows that ‘Rain’ has the highest and most variable precipitation levels, with some outliers. ‘Snow’ also has a significant spread but is less frequent. ‘Fog’ & ‘’Sun’ have much lower and more consistent precipitation levels.
* The scatter plot for Wind speed and precipitation suggests a weak positive correlation between wind speed and precipitation particularly for ‘Rain’ conditions. However, most of the data points are clustered at lower wind speeds indicating higher wind speeds are less common.
* The time series chart shows seasonal trends in average precipitation and temperature across different weather types over several years. ‘Rain’ consistently contributes the most to precipitation, while ‘Sun’ generally shows higher temperatures.

**CORRELATION & REGRESSION ANALYSIS**

**CORRELATION ANALYSIS**

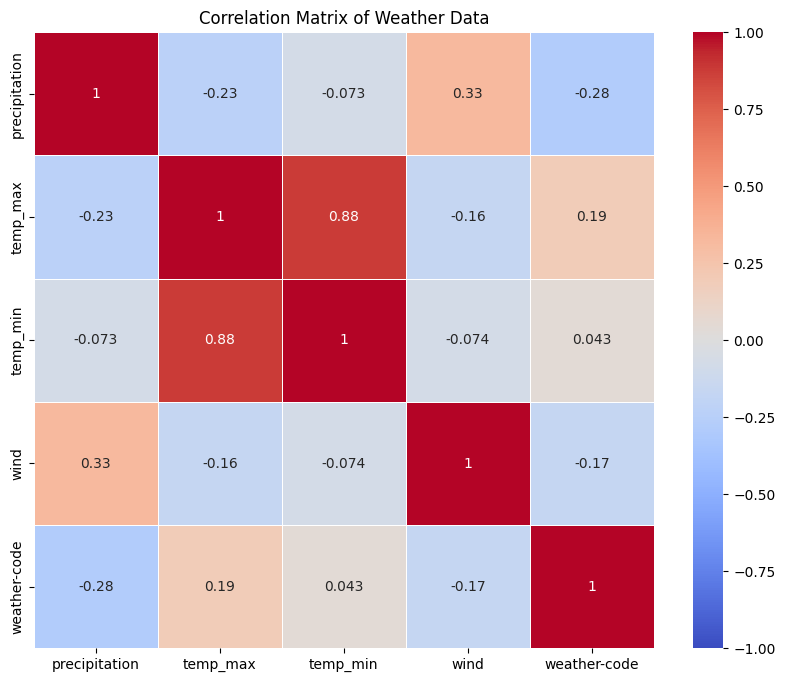
Correlation analysis is a statistical technique used to evaluate the strength and direction of the linear relationship between two quantitative variables. In this weather dataset, correlation analysis helps to identify how various factors such as temperature, precipitation, and wind are interrelated. Understanding these relationships is crucial for making informed predictions and understanding the underlying patterns in weather data.

To perform the correlation analysis, the Pearson correlation coefficient was calculated for each pair of variables using Python’s ‘.corr()’ method. The correlation matrix was then visualized using a heatmap to easily identify strong correlations between variables.



Figure

The correlation heatmap below visually represents the relationships between the variables in the dataset. Strong positive correlations (coefficients close to 1) are depicted in maroon, while moderate positive correlations (coefficients between 0.75 & 0.50) are shown in peach and weak positive correlations (coefficients below 0.25) are in light pink. Very weak or no correlation (coefficients close to 0) appear in light blue. Negative correlations are represented in the shades of blue, with coefficients between -0.25 and -0.50 appearing in sky blue, while stronger negative correlations (coefficients greater than -0.75) in dark blue. This visualization helps in identifying potential relationships and dependencies among the temperatures, precipitation, wind and weather type.



Figure

From the above results (fig. 11 & 12), we can interpret that –

* Precipitation is moderately positively correlated with wind (0.33), indicating that higher precipitation is often associated with stronger winds. While, there is a weak negative correlation between precipitation and maximum temperature (-0.23), which aligns with the negative coefficient observed in the regression analysis.
* There is a strong positive correlation between maximum temperature and minimum temperature (0.88), indicating a strong positive relationship between the temperatures. Whereas, maximum temperature show a weak positive correlation with the weather-code (0.19), suggesting that more favorable weather conditions slightly increase the maximum temperature.
* Wind shows a weak negative correlation with maximum temperature (-0.16) which suggests that higher wind speeds slightly reduce the maximum temperature, although the regression shows a positive impact, likely due to interaction with other variables. Also, wind speed is negatively correlated with weather-code (-0.17), indicating that certain weather conditions associated with lower wind speeds are present when weather codes indicate good weather.
* Weather-code shows a weak positive correlation of 0.19 with maximum temperature, confirming that better weather conditions are mildly associated with higher temperatures.

Therefore, we see that there is a strong relationship between maximum and minimum temperatures in the dataset. However, the results also highlights some weaker relationships, which are still important but less impactful in predicting maximum temperature compared to other factors.

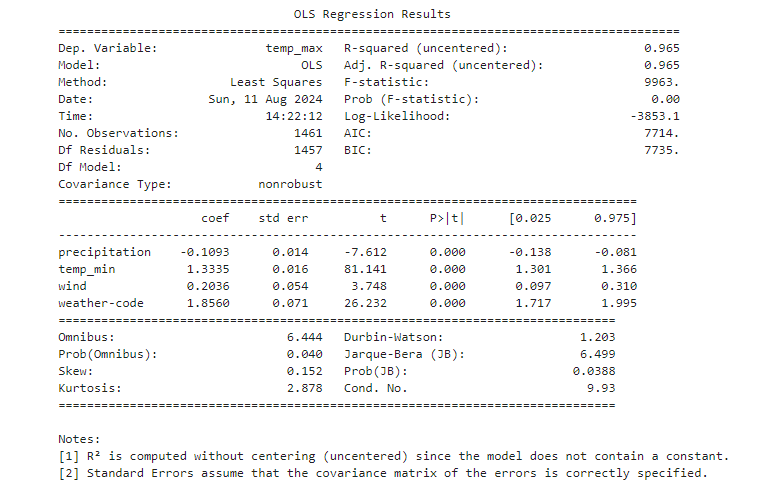
**REGRESSION ANALYSIS**

Regression analysis is a powerful statistical method used to model and analyze the relationships between a dependent variable and one or more independent variables. It helps in predicting the value of the dependent variable based on the known values of the independent variables.

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called Linear models.

In this weather dataset, regression analysis is used to predict weather-related outcomes such as precipitation or temperature based on other variables.

The regression model was developed using the ‘statsmodels’ library in Python. In this model, maximum temperature was chosen as the dependent variable, while precipitation, minimum temperature, wind and weather-code, were selected as independent variables. The Ordinary Least Squares (OLS) method was used to estimate the coefficients of the regression model.

The OLS Regression result obtained from Python is –

Figure

From the above results, we can say that –

**Model Summary:**

* R-square is 0.965, this indicates that 96.5% of the variance in temp\_max is explained by the predictors in the model.
* The Adjusted R-square (0.965) also reflects a high explanatory power, adjusted for the number of predictors.
* F-statistic is 9963, which is very high suggesting that the model as a whole is statistically significant.
* Prob(F-statistic):0.00, confirms that overall regression model is significant (p-value<0.05).

**Coefficients:**

* **Precipitation:** The regression coefficient for precipitation is -0.1093. This negative coefficient suggests that an increase in precipitation tends to decrease the maximum temperature (temp\_max). The effect is significant with a p-value of 0.000.
* **Minimum Temperature:** The regression coefficient for temp\_min is 1.3335. Thus minimum temperature is positively related to the maximum temperature, meaning that a 1-unit increase in temp\_min is associated with a 1.3335 unit increase in temp\_max. This relationship is highly significant (p-value: 0.000).
* **Wind:** The regression coefficient for wind speed is 0.2036. Wind speed has a positive and significant effect on maximum temperature, with a 1-unit increase in wind speed associated with a 0.2036 unit increase in temp\_max.
* **Weather-code:** The coefficient for weather type is 1.8560. The weather-code (likely categorical, indicating different weather conditions) has a strong positive effect on maximum temperature and the relationship is significant (p-value: 0.000).

**Significance:** All the predictors (precipitation, minimum temperature, wind and weather-code) have p-values less than 0.05, indicating that they all are statistically significant in predicting the maximum temperature.

Thus, regression analysis shows that all the predictors are significant in explaining variations in maximum temperature, with minimum temperature having the largest positive impact which can also be seen in correlation results.

**CONCLUSION**

In this project, we performed a comprehensive analysis on the Weather dataset using Python and advanced analysis was then visualized using Tableau. Through various analyses like correlation, regression and visualizations, we gained valuable insights into the distribution and relationships between various weather conditions, precipitation, temperatures and wind speed.

The analysis revealed that ‘Rain’ and ‘Sun’ are the most frequent weather conditions, dominating the dataset while ‘Snow’ and ‘Drizzle’ occur much less frequently. The consistent distribution across years suggests a predictable pattern.

Higher precipitation is closely associated with ‘Rain’, while other weather types have lower and more variable precipitation levels. Temperature fluctuations also align with different weather types, with snow corresponding to lower temperatures and sun to higher temperatures. Wind speed also exhibits varying degrees of correlation with temperatures and precipitation across different weather types.

Despite the insights gained, there are some limitations to this analysis. The dataset may not capture all the complexities & nuances of weather patterns, especially in extreme or rare events. Additionally, the analysis is constrained by the specific timeframe of the data, which may not fully represent long term climate trends. Furthermore, the data used in the analysis might be subject to recording errors or biases that could affect the accuracy of the results.

Overall, the Weather dataset provided efficient findings that offer a clear understanding of how weather variables interact, which can be crucial for forecasting and making data-driven decisions in weather-dependent industries. The analysis highlights the importance of continuous monitoring and detailed examination of weather data to improve predictive models and operational efficiency.