Weather Analysis

- Vaibhav Saini

Table Of Contents

Introduction

- Analyze the meteorological data of 1year from Weather.csv file to understand the weather pattern based on different factors.

- Conducted **data cleaning and preprocessing steps**, including handling missing values and outliers,encoding of categorical variables

.

- Explore and analyze weather data using advanced data analysis tool known as Tableau to uncover trends and patterns in the weather dataset using different type of visualizations such as bar, line, box, scatter plot etc.

- Conducted Correlation analysis on the data to get the relation between different factors of weather.

- Prepared a nearly accurate regression model to predict a weather parameter based on the other parameters.

Methodology

- **Data Understanding:**

  Printing out the information of data to get to know about the various weather parameter and their data types, number of null values in data ,memory usage. Also, the statistical information related to different factors such as, mean ,max value, standard deviation, etc.

- **Data Preparation:**

  # Replacing the null values in numerical columns with the mean of respective of column.

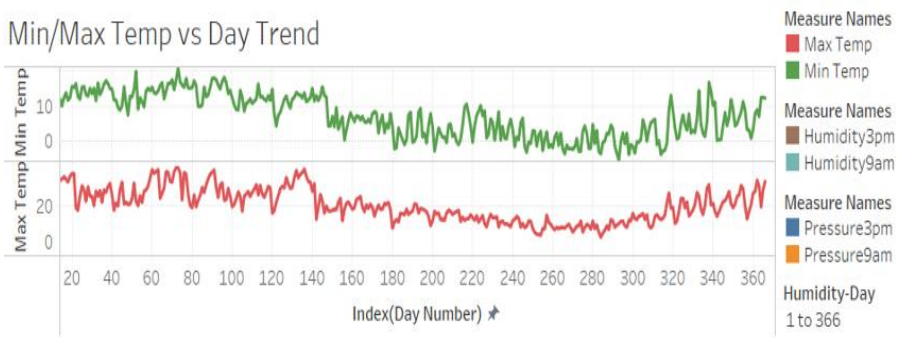
  # Replacing the null values in categorical columns here (RainToday,RainTommorow,Winddir) with mode value of respective column.

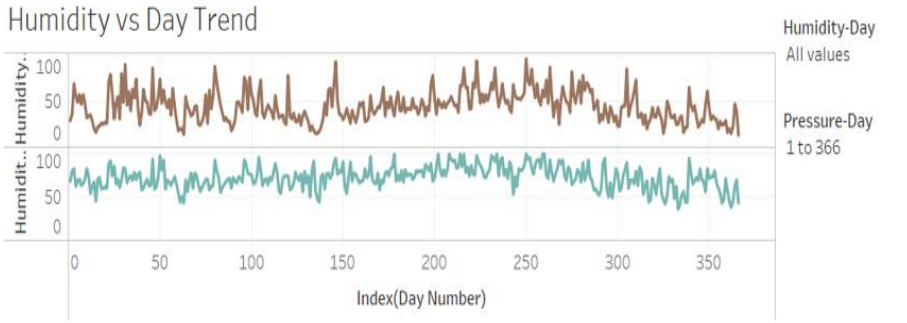
  # Use the boxplot to detect all the outliners and replace them with max limit of concern factor.

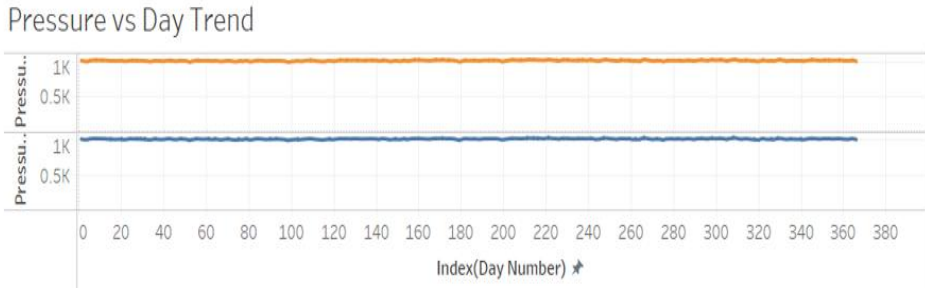
  # Use pairplot to get the idea how each factors are related with each other.

- **Data Vizualization Using Tableau:**

Line plot:
The plot is indicating how **Min/Max temperature** and **Pressure** is changing with each day.



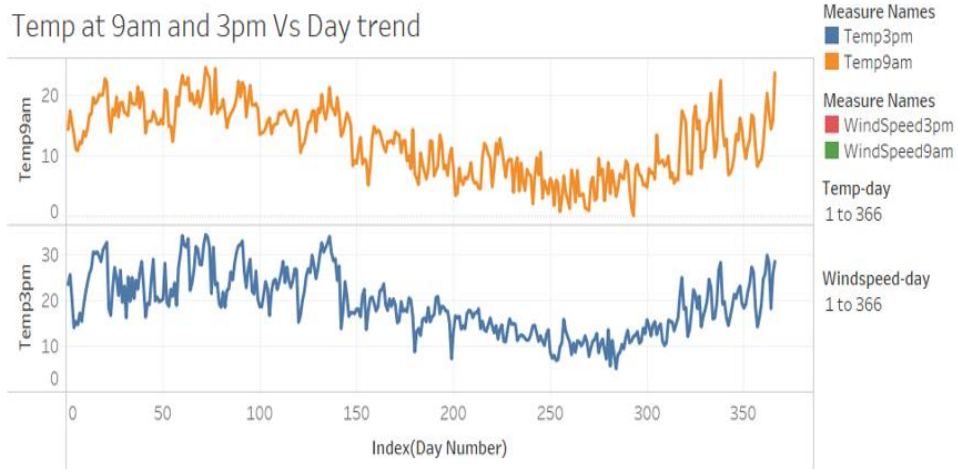


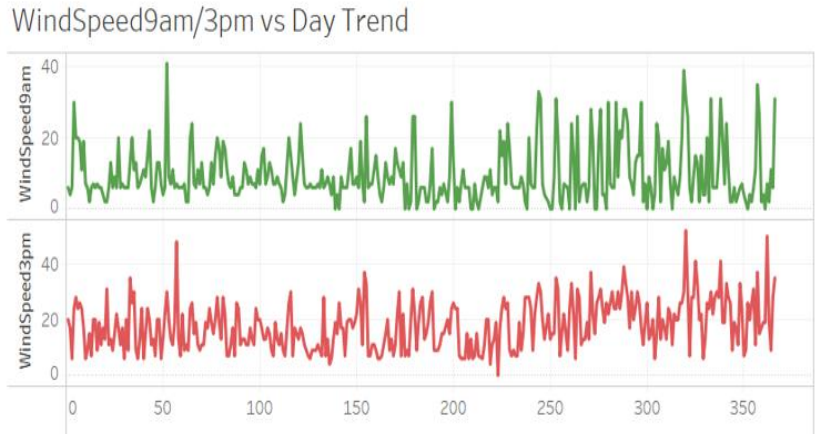


Insights:
*Range of MinTemp of a day throughout the year is [-5.30, 20.90], MaxTemp of a day is [7.60,35.80], Humidity at 3pm is [13,96], Humidity at 9am is [36,99], Pressure at 3pm is [996.80,1033.20], Pressure at 9am is [996.50,1035.70].

*The line plot depicting max/min temperatures reveals a notable peak in the range of days [0-150], followed by a decline in the subsequent range [150-293]. Afterward, temperatures begin to rise once more.

This **line plot** is indicating how **temperature and wind speed at 9am and 3pm** is changing with each day.





Insights:

* Range of Temp at 9 am throughtout the year is [0.10,24.70], Temp at 3 pm is [5.10-34.50], wind speed at 9 am is [0-41],wind speed at 3 pm is [0-52].

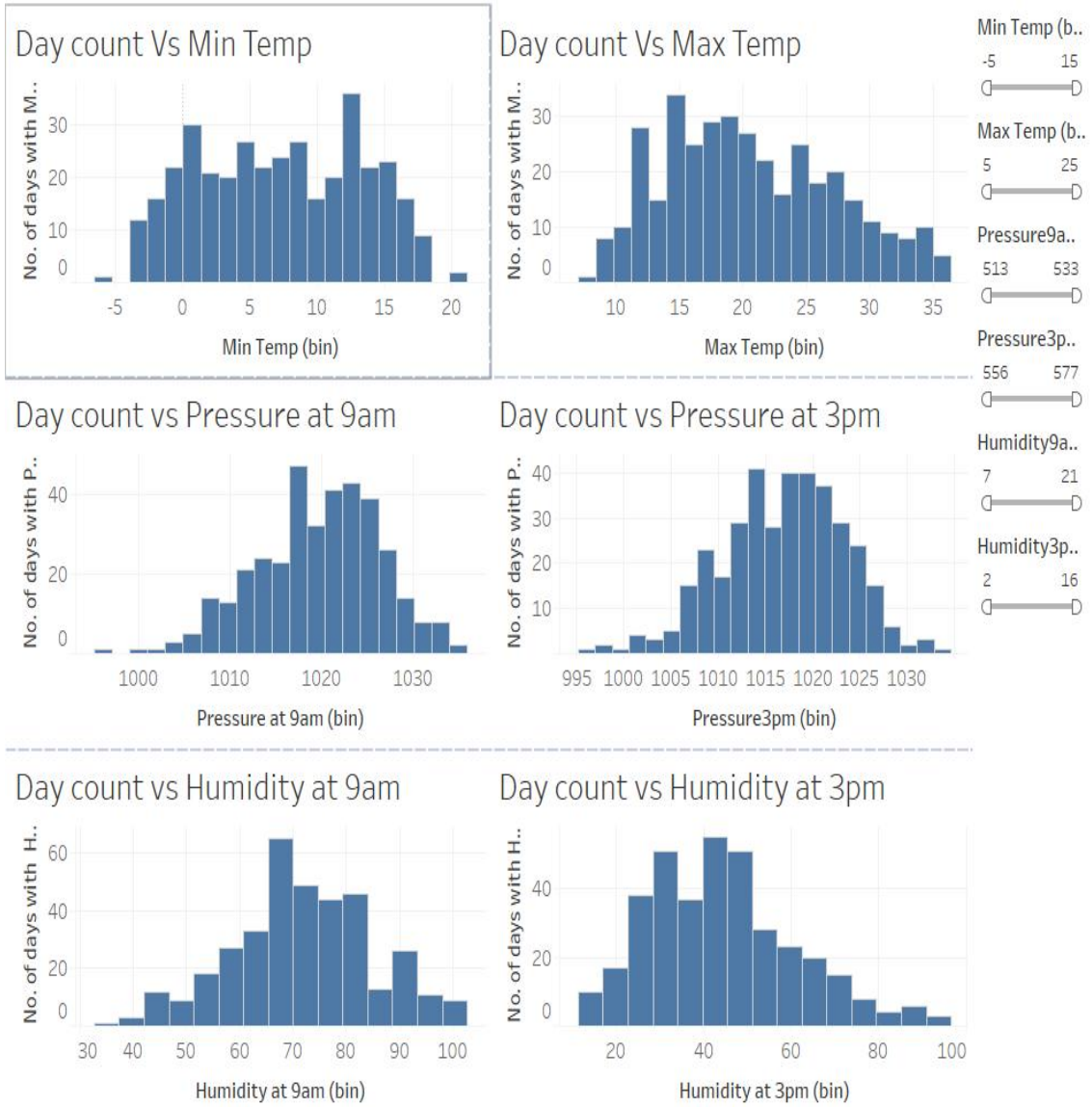
* Day number [0-148 and 318-366] has higher temp. at 9 am of the day as compared to rest of days.

*Day number [0-140 and 318-366] has higher temp. at 3 pm of the day as compared to rest of days.

*Generally most mornings of the year are more windy then evenings.

Histogram:

The provided histograms illustrate the distribution of Min and Max temperatures, as well as Pressure and Humidity levels at 9 am and 3 pm throughout the year. By examining the frequency of respective values, it aids in assessing the overall weather conditions for the entire year.



Insights:

* Most **common min temp. value** throughout the year is 12.
  Most **common max temp. value** throughout the year is 14.
  Most **common pressure at 9 am** of the day throughout the year is 1017.
  Most **common pressure at 3 pm** of the day throughout the year is 1013.
  Most **common humidity at 9 am** of the day throughout the year is 65.
  Most **common humidity at 3 pm** of the day throughout the year is 40.

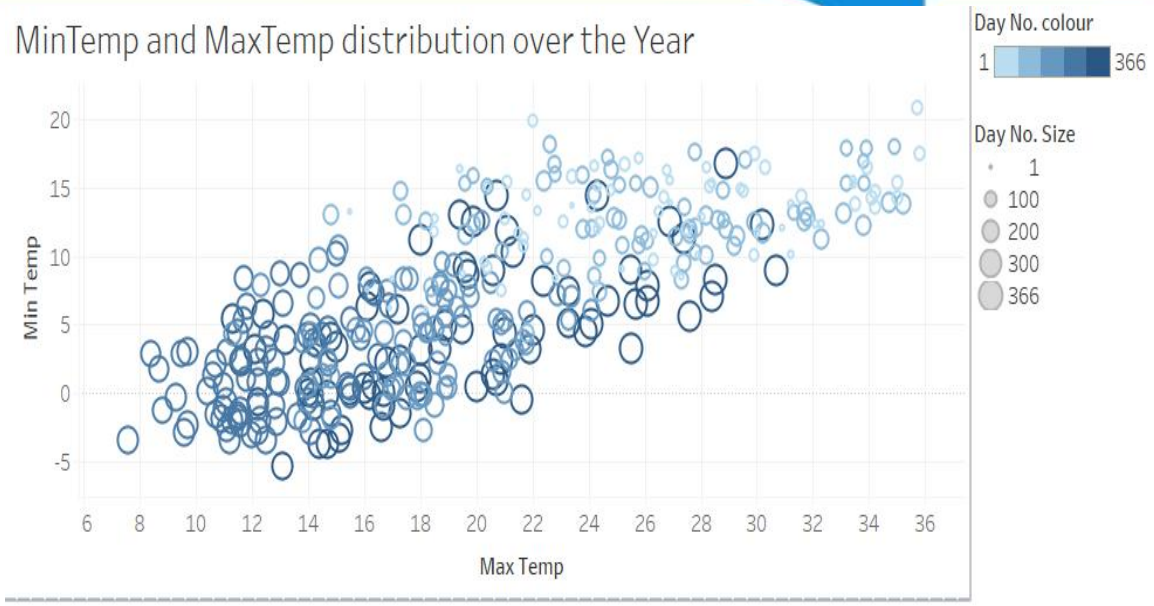
Scatter Plot:

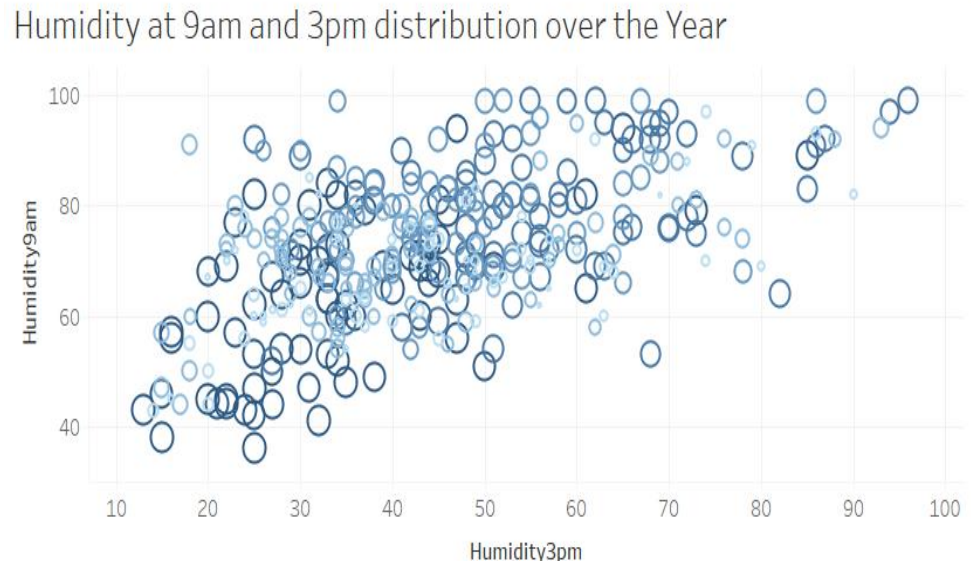
Given plot shows the distribution of **min and max temperature** distribution throughout the year and other is the plot showing distribution of **humidity at 9 am and 3 pm** of the day.

In the given plot , circles with bigger size and dark blue colour are representing the latter day in the year then ones with small and light blue colour circles.

Insights:

* It can be depicted that start of the year is having higher temperature as compared to the later time frame of the year.
*It is difficult to depict any pattern for humidity as it is randomly distributed throughout the year but it can be said that most of days is having the median value of humidity.



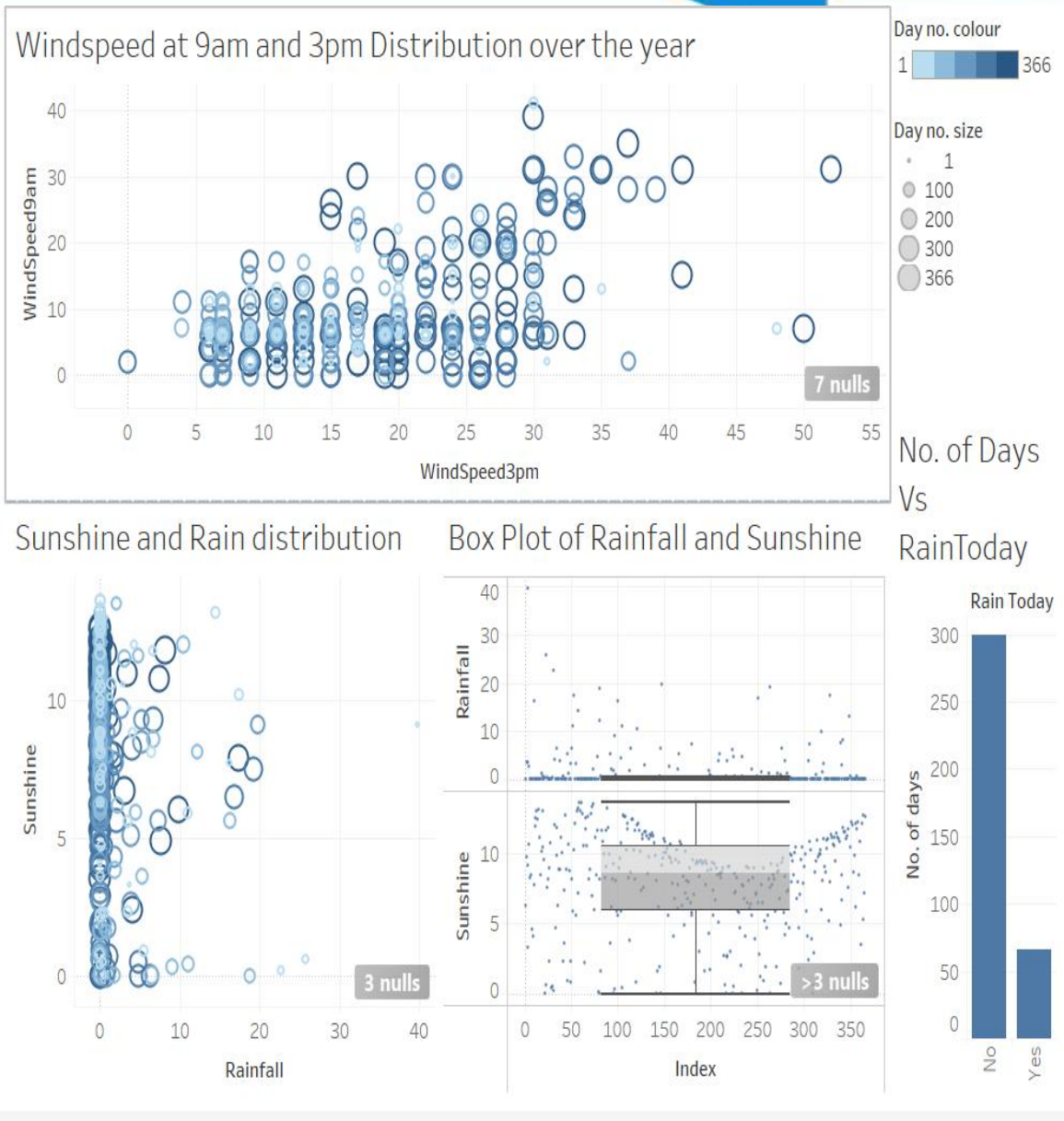


This is the another **scatter plot between wind speed at 9 am and 3 pm** of the day and another one is between **sunshine and rainfall** indexs representing their distribution throughout the year to analyze the approx weather
situation of the year.

In the given plot , circles with bigger size and dark blue colour are representing the latter day in the year then ones with small and light blue colour circles.

There is one **histogram plot of Rain Today**, showing how many days of the year it will be raining.

There is **box plot of Rainfall and Sunshine** giving the statistical measures insights such as median, max, value of the rainfall and sunshine index . Also it is also showing outliers if any.



Insights:

* Again here also, it is difficult to depict any wind pattern from the plot as it is randomly distributed throughtout the year, but it is surely showing that the mornings are more windy than evevings of the day.

* Most of the days of the year are sunny , and with only 66 days as rainy days.

* Range of sunlight index is [0-13.6] with median index as 8.6.

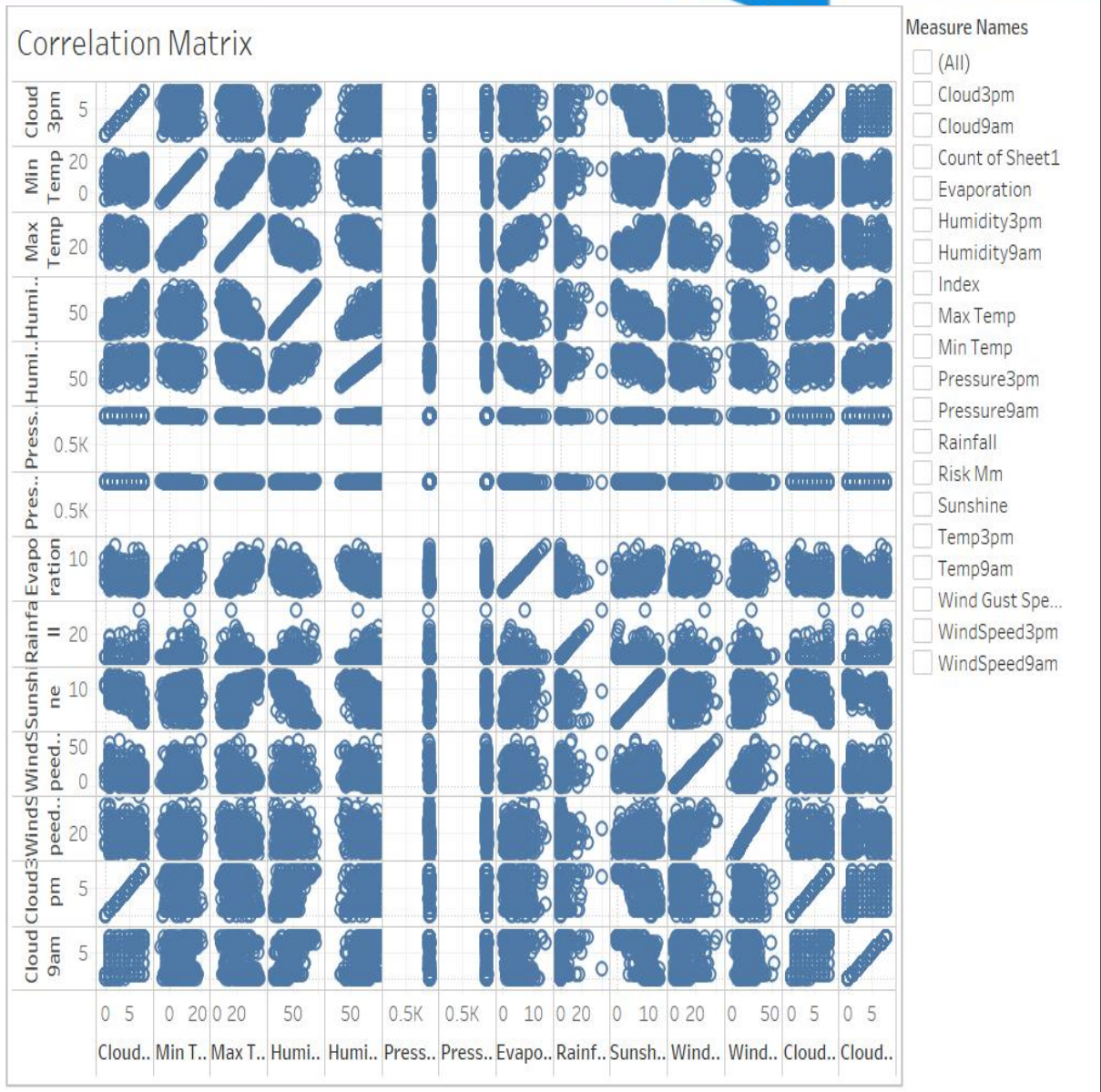
* Range of sunlight index is [0-0.4] with median index as 0.

* Their are no outliers for sunlight data while there are many ouliners for rainfall data as per the collected data.

Correlation matrix:

This is the representation of correlation matrix showing how one weather factor is dependent on other , it also shows the distribution of values of the concern weather parameter throughout the year.

These patterns will help in predicting the value of one factor , provided other weather parameters known.



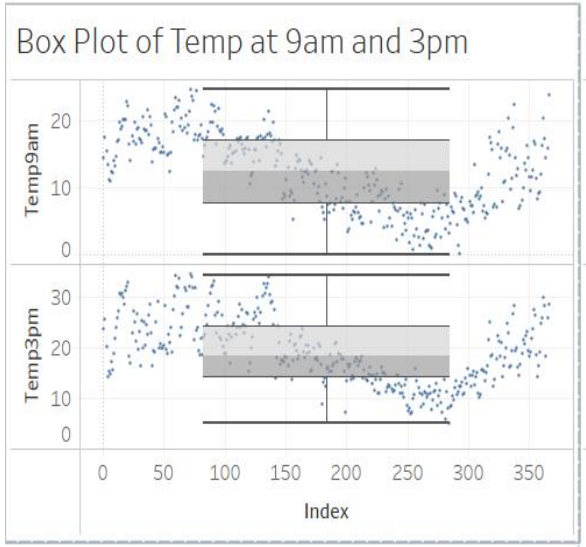
Box plot:

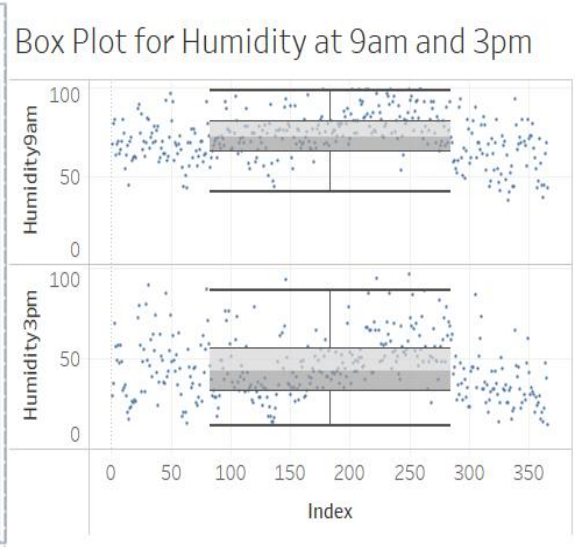
Given plot is showing the box plots of temperature, humidity, pressure and wind speed at 9 am and 3 pm of the day throughout the year.

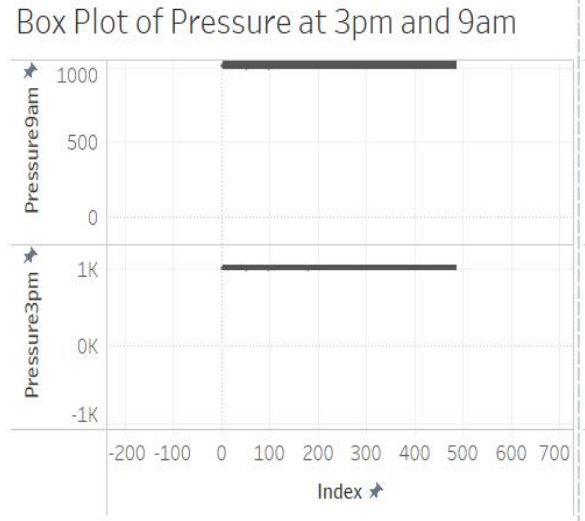
Givingn the statistical insight about them such as, median,max, min ,outliers,IQR(interquartile range) of the respective plot.

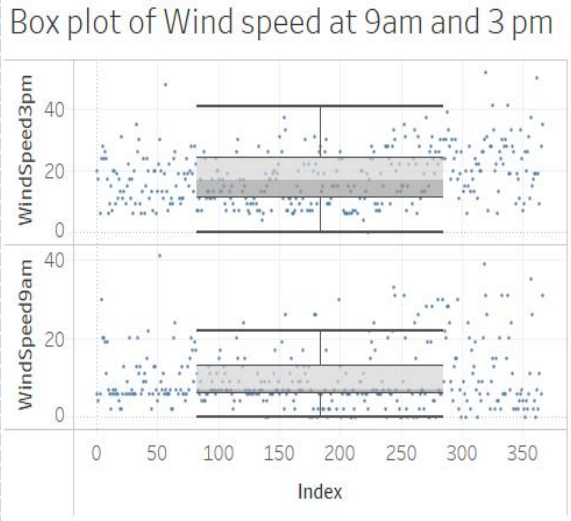
Insights:

Temp at 9 am	Temp at 3 pm
Range: [0.10,24.70] **IQR:** [7.60,17] **Median:** 12.55 **Outliers:** none	**Range:** [5.1,34.5] **IQR:** [14.1,24] **Median:** 18.55 **Outliers:** none









Humidity at 9 am

Range: [41,99]
IQR: [64,81]
Median: 72
Outliers: present

Humidity at 3 pm

Range: [13,88]
IQR: [32,55]
Median: 43
Outliers: present

Pressure at 9 am

Range: [1002.1,1035.7]
IQR: [1015.3,1024.5]
Median: 1020.15
Outliers: present

Pressure at 3 pm

Range: [1001.3,1033.2]
IQR: [1012.8,1021.5]
Median: 1017.4
Outliers: present

Wind Speed at 9 am

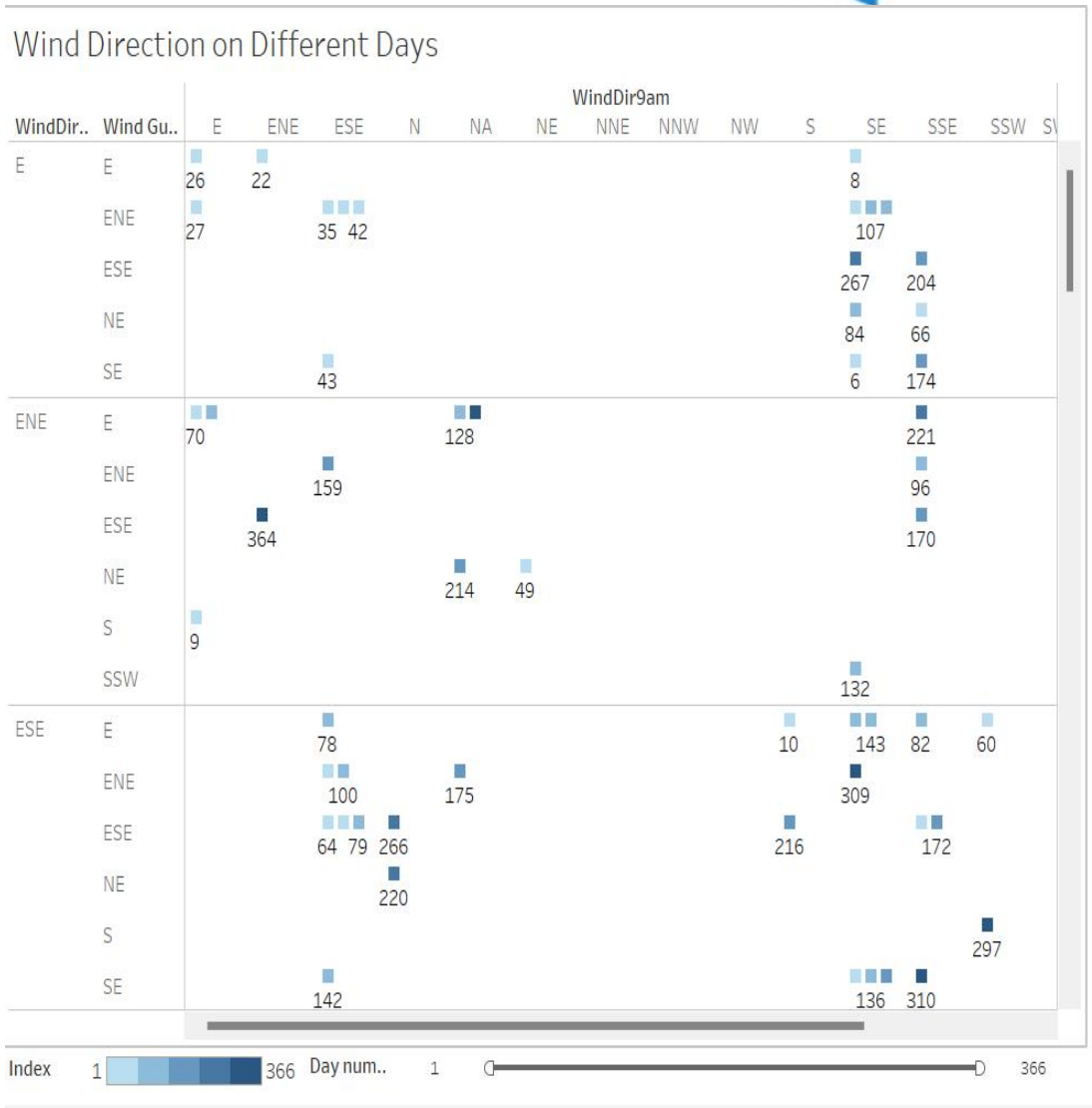
Range: [0,22]
IQR: [6,13]
Median: 7
Outliers: present

Wind Speed at 3 pm

Range: [0,41]
IQR: [11,24]
Median: 17
Outliers: present

Given plot is showing **direction of Wind at 9 am and 3 pm** of each day along with the **direction of gust wind**.

The number below each square is representing day number ,also the darker colour showing the latter day of the year then lighter colour ones.



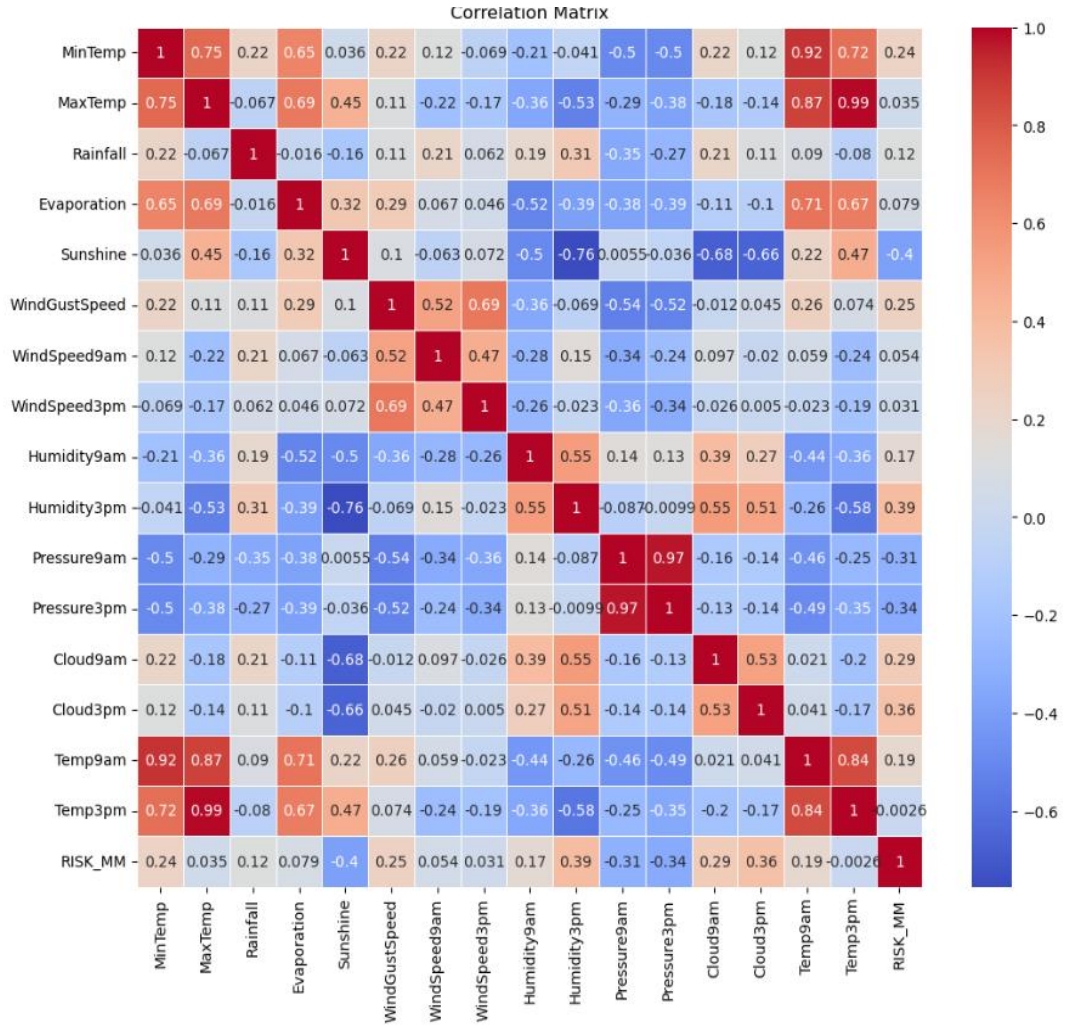
After performing data viualization and getting some important data insights we move towards next step that is correlation analysis and regression analysis-

- **Correlation Analysis:**

For this we visualize our correlation matrix of all the concern weather parameters using heatmap as shown in the plot given,

Here, all the numerical weather parameters are used with each cell containing a number representing the degree of dependence of parameter on x and y axis of the concern cell.The darker the colour context of the cell more the parameters are dependent on each other.

Positive number shows both parameters changes in same direction while negative implies they are changing in opposite directions.



- **Regression Analysis:**

For this we have to create a regression model with a decent accuracy such that it can predict one weather parameters on the basis of other parameters.

Since, we have our data well prepared upto now so now, for regression model of numerical parameter, we just have to select a target variables and dependent variables or which are known to us. After that we need two sets of data , test data and training data, so we split our known data into two parts:

Training set: Used to train our model to predict our weather parameters based in other weather parameters.

Test set: Used to test our model for accuracy of prediction.

*The only difference we will do while predicting categorical data such as ‘RainToday’ is that we need to encode our data present in categorical coloumn such that string data gets converted into some numerical encoding.

Conclusion

This weather analysis has provided valuable insights into various factors influencing weather patterns. Through exploratory data analysis, we uncovered trends in temperature fluctuations, rainfall, wind speed, and atmospheric pressure. The correlation matrix highlighted dependencies between different weather factors. Our predictive modeling accurately forecasted rain occurrences, demonstrating the effectiveness of machine learning in weather prediction.

THANK YOU