Exploratory Analysis of the Weather of Puebla, Mexico

Feb 2023 - Jan 2024

Prepared by :

Aqib Nisar

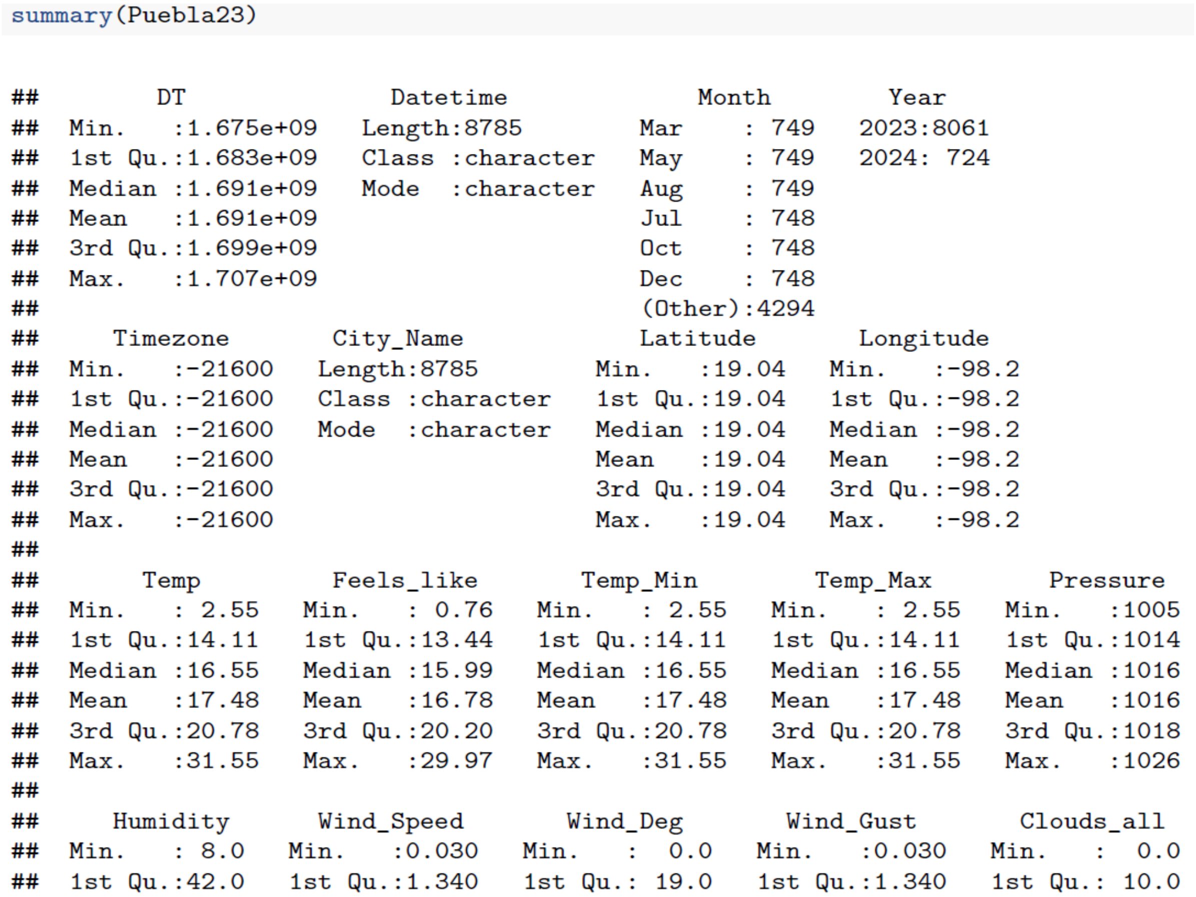
Table of CONTENTS

Introduction

Puebla is a city in east-central Mexico, southeast of Mexico City. It is the fourth largest city in Mexico, after Mexico City, Monterrey, and Guadalajara. It has a population of about 3.3 million with an elevation of 2135 meters above sea level. This exploratory analysis focuses on the weather of the city of Puebla extracted from the OpenWeather. This data was collected at hourly intervals from February 2023 to January 2024. The data was extracted using Python and the analysis is done using R programming language. The analysis gives a thorough picture of the climatic patterns observed throughout the year by looking at elements like temperature, humidity, atmospheric pressure, and wind.

Our data has a total of 8785 observations with 22 different variables, out of which 5 are converted to categorical variables with different levels to facilitate the analysis and 10 are numerical variables that will help us interpret the information.

When we look at the summary of our weather data, we notice that some variables, like temperature, show very little difference from their minimum and maximum temperature variables.



```
summary(Puebla23)
```

```
##        DT                Datetime              Month          Year     
##  Min.   :1.675e+09   Length:8785        Mar      : 749   2023:8061  
##  1st Qu.:1.683e+09   Class :character   May      : 749   2024: 724  
##  Median :1.691e+09   Mode  :character   Aug      : 749              
##  Mean   :1.691e+09                      Jul      : 748              
##  3rd Qu.:1.699e+09                      Oct      : 748              
##  Max.   :1.707e+09                      Dec      : 748              
##                                         (Other):4294              
##     Timezone        City_Name             Latitude       Longitude   
##  Min.   :-21600   Length:8785        Min.   :19.04   Min.   :-98.2  
##  1st Qu.:-21600   Class :character   1st Qu.:19.04   1st Qu.:-98.2  
##  Median :-21600   Mode  :character   Median :19.04   Median :-98.2  
##  Mean   :-21600                      Mean   :19.04   Mean   :-98.2  
##  3rd Qu.:-21600                      3rd Qu.:19.04   3rd Qu.:-98.2  
##  Max.   :-21600                      Max.   :19.04   Max.   :-98.2  
##                                                                     
##       Temp          Feels_like        Temp_Min         Temp_Max         Pressure   
##  Min.   : 2.55   Min.   : 0.76   Min.   : 2.55   Min.   : 2.55   Min.   :1005  
##  1st Qu.:14.11   1st Qu.:13.44   1st Qu.:14.11   1st Qu.:14.11   1st Qu.:1014  
##  Median :16.55   Median :15.99   Median :16.55   Median :16.55   Median :1016  
##  Mean   :17.48   Mean   :16.78   Mean   :17.48   Mean   :17.48   Mean   :1016  
##  3rd Qu.:20.78   3rd Qu.:20.20   3rd Qu.:20.78   3rd Qu.:20.78   3rd Qu.:1018  
##  Max.   :31.55   Max.   :29.97   Max.   :31.55   Max.   :31.55   Max.   :1026  
##                                                                                
##     Humidity        Wind_Speed        Wind_Deg        Wind_Gust        Clouds_all  
##  Min.   : 8.0   Min.   :0.030   Min.   :  0.0   Min.   :0.030   Min.   :  0.0  
##  1st Qu.:42.0   1st Qu.:1.340   1st Qu.: 19.0   1st Qu.:1.340   1st Qu.: 10.0  
##  Median :57.0   Median :2.200   Median :118.0   Median :2.200   Median : 57.0
##  Mean   :57.1   Mean   :2.472   Mean   :118.8   Mean   :2.472   Mean   : 52.8
##  3rd Qu.:74.0   3rd Qu.:3.260   3rd Qu.:191.0   3rd Qu.:3.260   3rd Qu.: 93.0
##  Max.   :98.0   Max.   :9.160   Max.   :360.0   Max.   :9.160   Max.   :100.0
##
##    Weather_Id    Weaather_main        Weather_Description  Weather_icon
##  Min.   :500.0   Clear :2262   clear sky        :2261     04n    :1961
##  1st Qu.:800.0   Clouds:5238   overcast clouds  :2145     04d    :1662
##  Median :802.0   Rain  :1285   broken clouds    :1478     01d    :1281
##  Mean   :757.9                 light rain       :1072     01n    : 981
##  3rd Qu.:803.0                 scattered clouds: 953      10n    : 673
##  Max.   :804.0                 few clouds       : 662     10d    : 612
##                                (Other)          : 214     (Other):1615
```

By using the sum() function we can see that the variables ‘Temp’, ‘Temp_min’, and ‘Temp_max’ have only 3 out of 8785 observations that are different from each other. Similarly, variables ‘Wind speed’ and ‘Wind Gust’ have zero difference in their values. Therefore, ignoring the ‘Temp_min’, ‘Temp_max’, and ‘Wind Gust’ variables in our analysis is justifiable.

```
sum(Puebla23$Temp != Puebla23$Temp_Min)
```

```
## [1] 3
```

```
sum(Puebla23$Temp != Puebla23$Temp_Max)
```

```
## [1] 2
```

```
sum(Puebla23$Temp_Max != Puebla23$Temp_Min)
```

```
## [1] 3
```

```
sum(Puebla23$Wind_Speed != Puebla23$Wind_Gust)
```

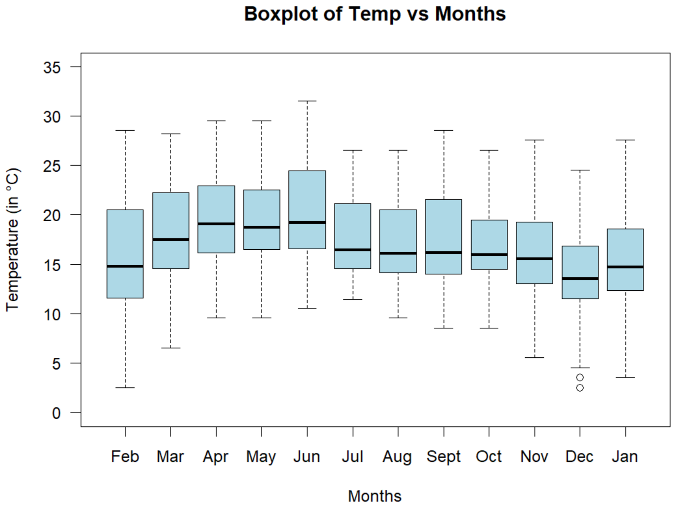
```
## [1] 0
```

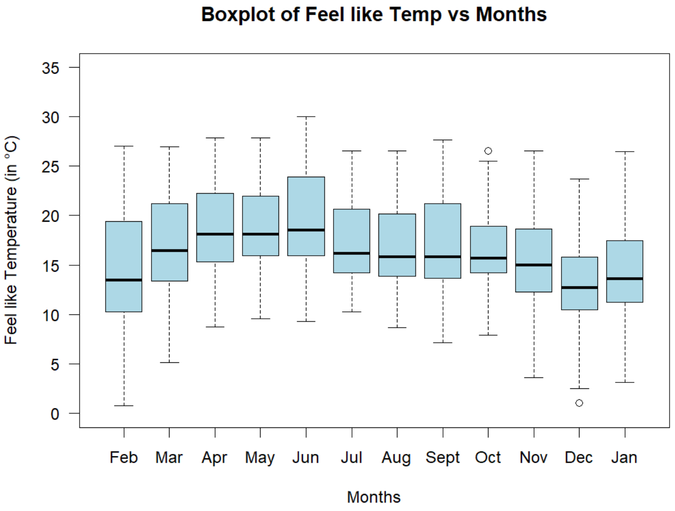
Description of Features

Features of dataset	Variable Type	Range	Description
DT	-	1675238400 - 1706677200	Time of data calculation, Unix, UTC
DateTime	-	(2023-02-01 03:00:00) - (2024-01-31 00:00:00)	Date and time of the recorded weather data (Hourly)
Month	Categorical	Feb 2023 – Jan 2024	The month of the recorded weather data
Year	Categorical	2023 - 2024	The year of the recorded weather data.
Timezone	-	N/A	Indicates the timezone of the city (in seconds)
City_Name	-	N/A	Name of the city (Puebla)
Latitude	-	N/A	Geographic coordinates (19.03793)
Longitude	-	N/A	Geographic coordinates (-98.20346)
Temp	Numerical	2.55 – 31.55 (in °C)	Temperature recorded in the city
Feels_like	Numerical	0.76 – 29.97 (in °C)	How the temperature feels like to a human.
Temp_Min	Numerical	2.55 – 31.55 (in °C)	Minimum temperature recorded in a period
Temp_Max	Numerical	2.55 – 31.55 (in °C)	Maximum temperature recorded in a period
Pressure	Numerical	1005 – 1026 (in hPa)	Atmospheric pressure recorded in the city
Humidity	Numerical	8.0 - 98.0 (%)	Indicates the level of moisture in the air
Wind_Speed	Numerical	0.030 - 9.160 (m/s)	Speed of the wind at the given location and time
Wind_Deg	Numerical	0.0 - 360.0 (in degrees)	Direction of the wind
Wind_Gust	Numerical	0.030 - 9.160 (m/s)	The maximum wind speed recorded during the specified period.
Clouds_all	Numerical	0.0 – 100.0 (%)	The percentage of the sky covered by clouds.
Weather_Id	-	500.0 – 804.0	A numerical code that represents the specific weather conditions.
Weather_main	Categorical	N/A	A general categorization of the weather conditions
Weather_Description	Categorical	N/A	A more detailed description of the weather conditions
Weather_icon	Categorical	N/A	A code or symbol representing the weather conditions

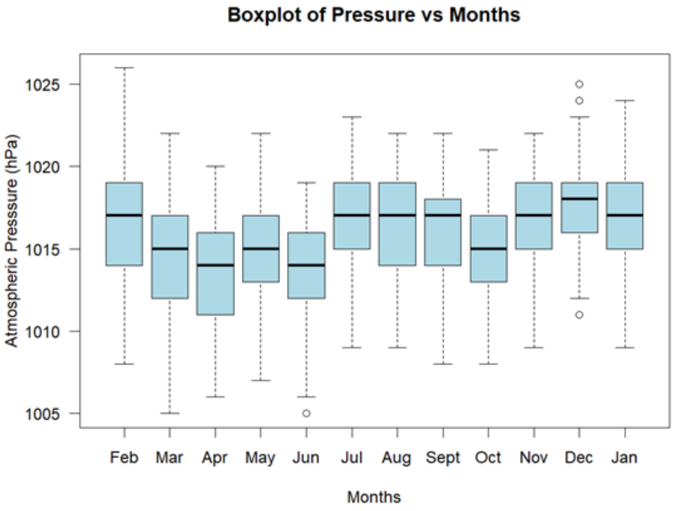
Visual Patterns

Since we are analysing monthly data, we will use boxplots to observe the patterns of our dataset. The temperature recorded is in degrees Celsius for both ‘Temp’ and ‘Feels Like’ variables. By looking at the boxplots of the two variables, we can see that both of them show similar trends with only minimal differences. According to these plots, the average temperature of Puebla stays between 15 to 20 °C with the maximum temperature reaching around 31°C during the month of June and the minimum in February around 2.5 °C. There are only a couple of outliers in our data which suggests that the temperatures recorded were well controlled and reliable. It also suggests that the temperature remained normal throughout the year with no extreme weather events.

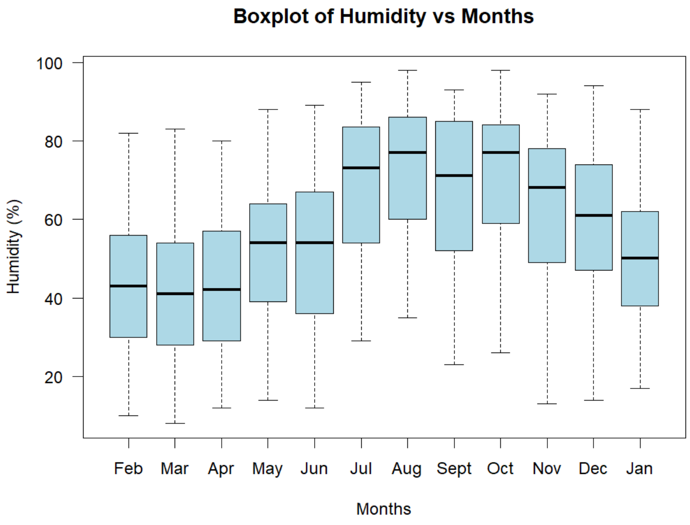




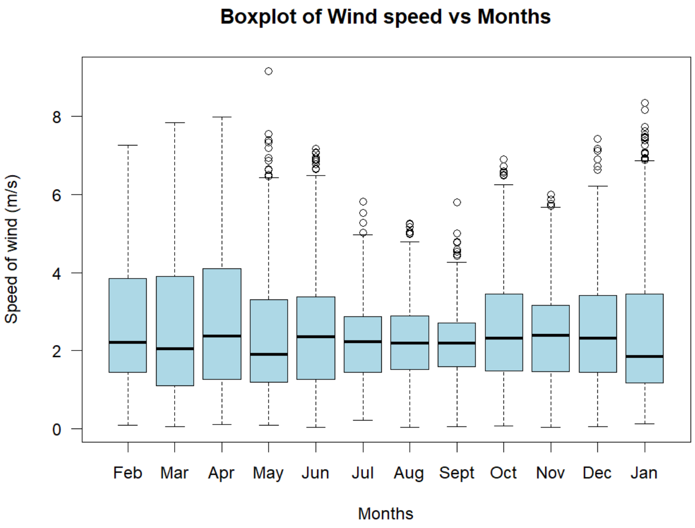
The boxplot of atmospheric pressure tells us that the pressure varies differently every month. However, most of it stays under 1020 hPa throughout the year. The month of February shows the highest variation in atmospheric pressure variable with the highest reaching 1026 hPa while December shows the least variation with only a few outliers. March recorded the lowest reading of 1002 hPa.

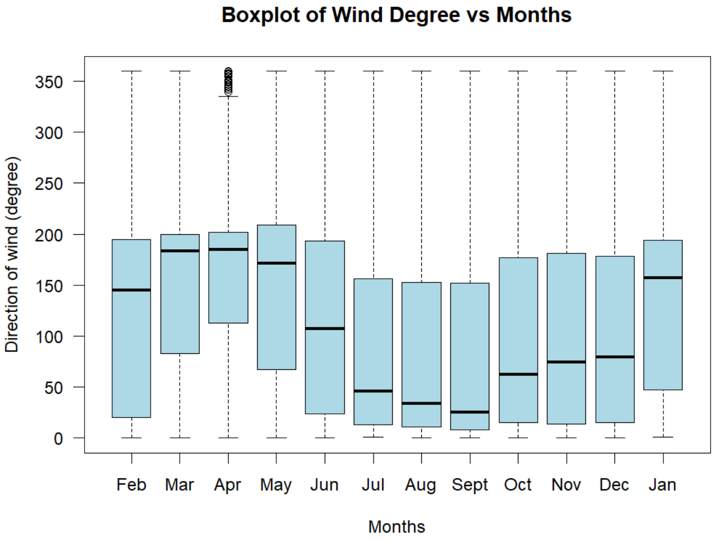


Also, the presence of a few outliers suggests that there were no major weather events for the year and the pressure remained normal.

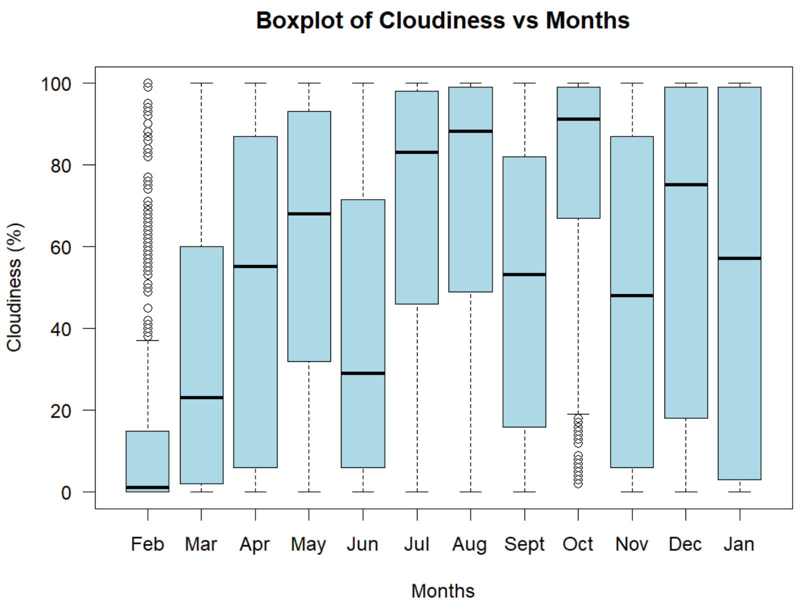


According to the box plot, the percentage of humidity shows an increase during the months from July to December. This can be due to the change in seasons, as the dry season of Puebla ends in May or June and the rainy season starts from July with the peak in the month of August. From the period of Jul to Dec, the average humidity remained within 60 to 80 %, and from Feb to Apr, the average didn't even cross the 45% mark with the minimum reaching only 8% in March for the year 2023. There are no outliers in this boxplot which once again tells us the reliability of our recorded data.

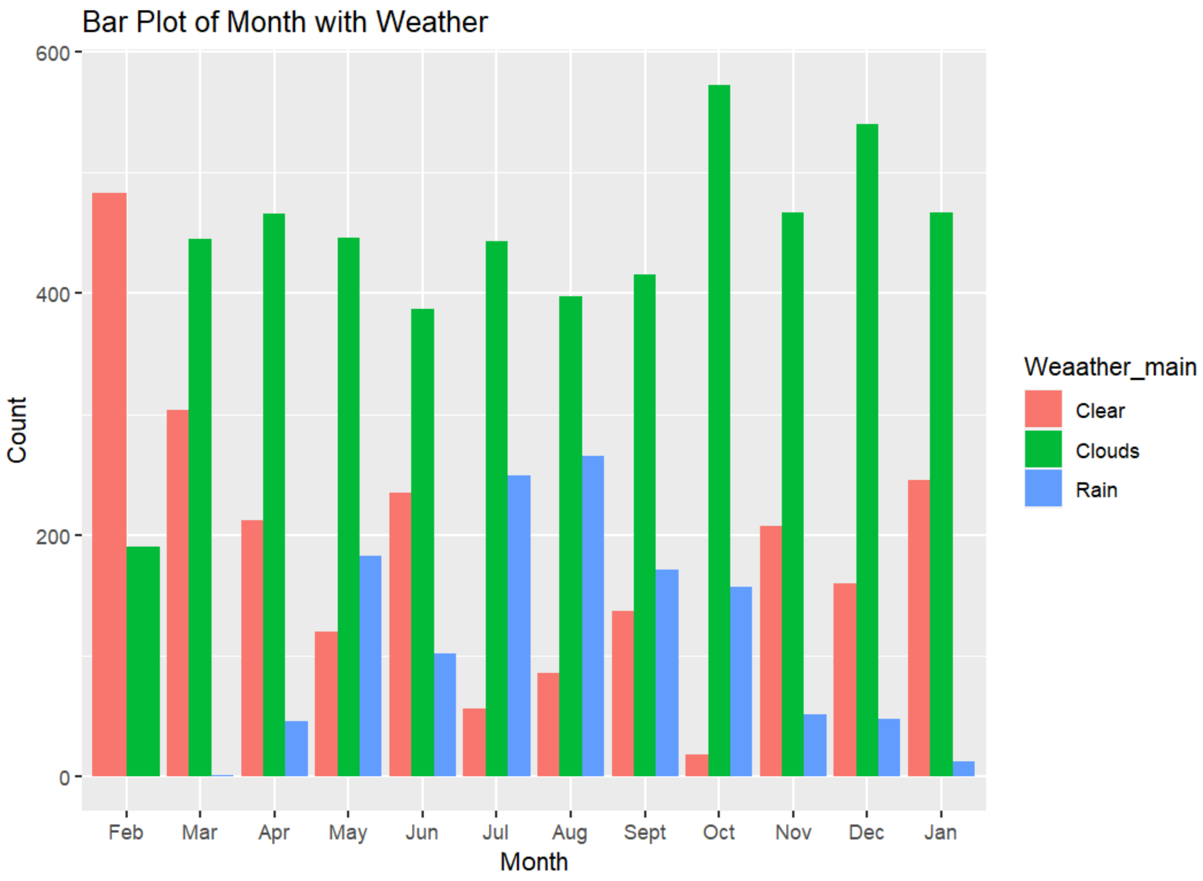




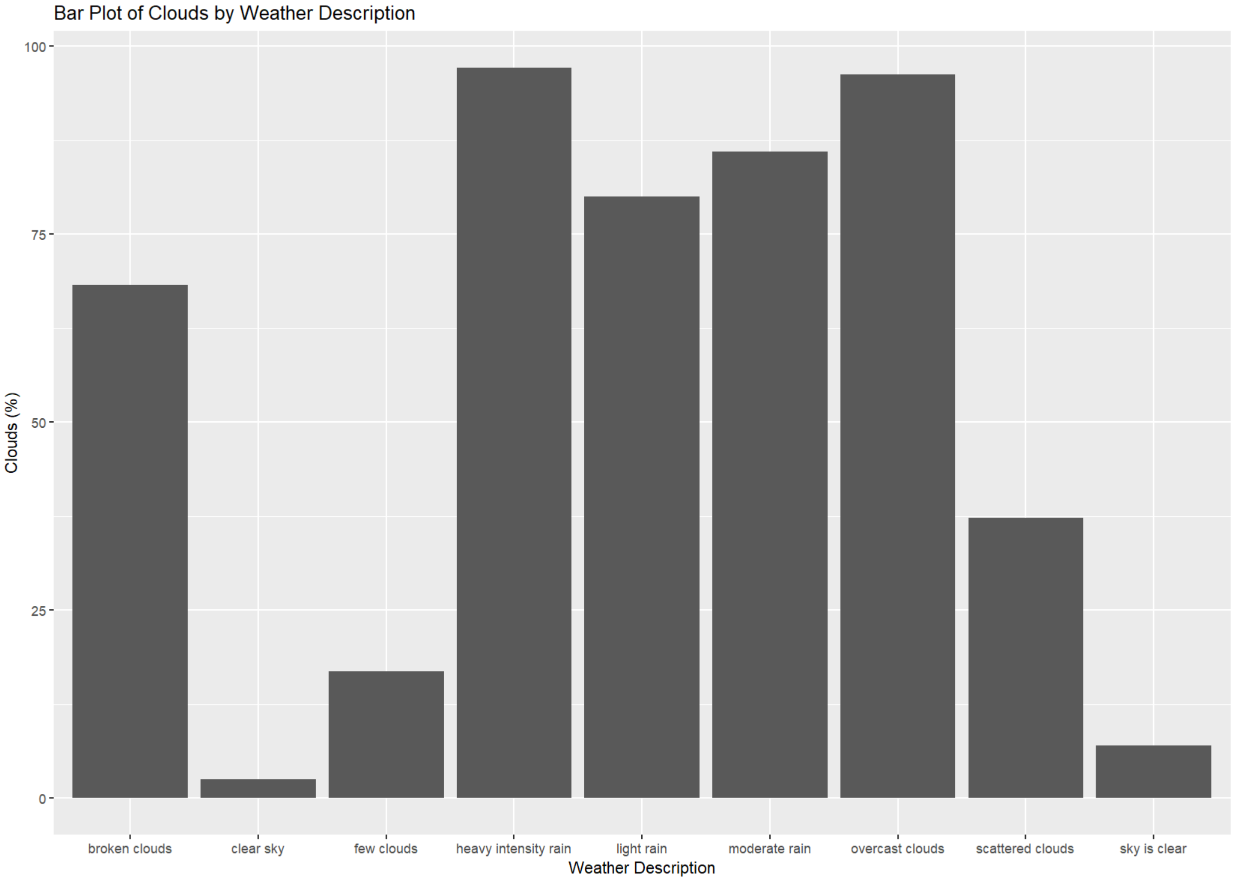
Looking at the wind speed plot, we observe that the average wind speed remained only around 2 m/s throughout the year and didn't show much dramatic changes. The months of Jun, Jul, and Aug show the least amount of variation in wind speeds while as Apr shows the highest, with the maximum reaching 9.16 m/s. The direction of the wind also mostly stayed under the 200-degree mark but the variation in direction remained the same throughout the year. The average direction of wind is different for each month.



The boxplot percentage of clouds shows huge variation from month to month, unlike other boxplots of our dataset. The months of Jul, Aug, and Oct show the highest percentage of clouds which of course coincides with the rainy seasons. February shows the least amount of cloudiness with the most outliers. The average percentage of cloudiness kept changing every month, with all of them reaching 0% and 100% at some point during the month.



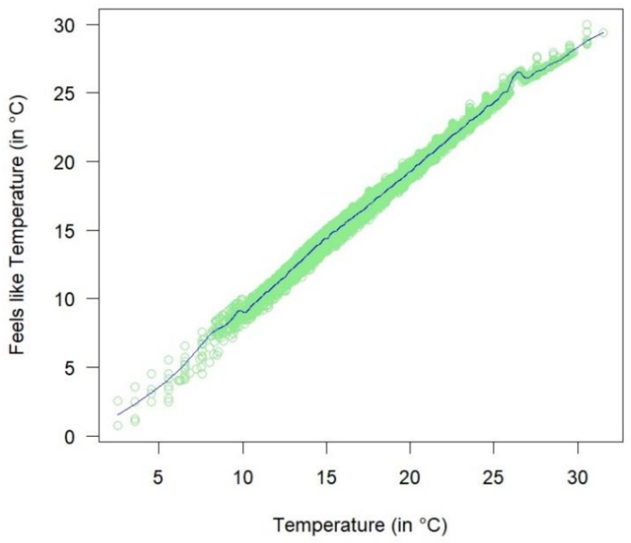
The Barplot above gives us information about the weather conditions throughout the year. According to this plot, the month of Feb 2023 was dry and clear and didn't see any rainfall with March being the second driest which recorded only a couple of hours of rain for the whole month. The actual rainy season started in July with the most amount of rain in August (around 260 hours) for the year 2023. The rainy season ended in the month of October, one month earlier than the normal season which ends in November. October had the cloudiest weather of about 570 hours recorded with only 10 to 20 hours of clear weather. For the whole period of Feb 2023 to Jan 2024, all the months recorded the highest counts of cloudy weather as compared to other weather conditions with the exception of February which had clear weather for the most number of counts.

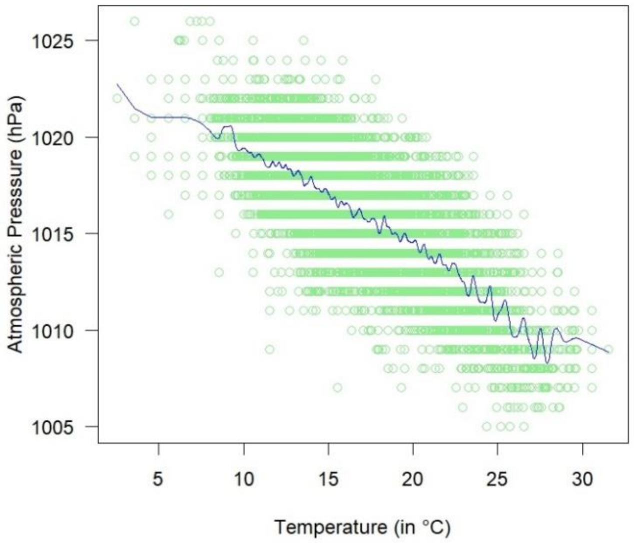


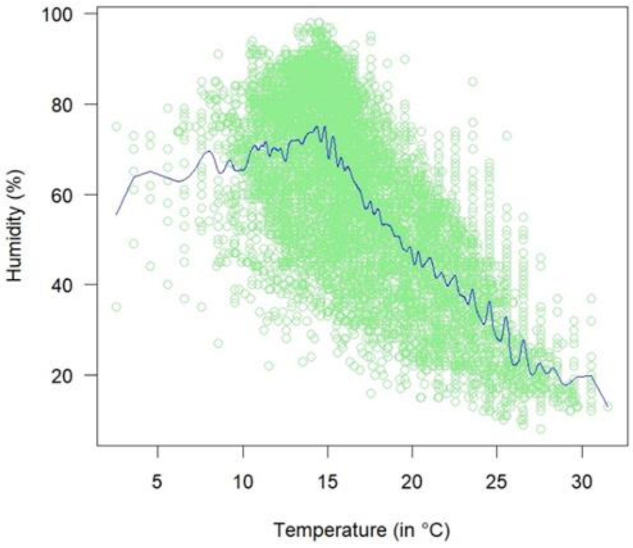
The above Barplot tells us about the percentage of clouds and their corresponding weather conditions. According to this graph, there is a high chance of rainfall when the percentage of clouds crosses the 75% mark however it can also result in an "Overcast clouds" condition. It also says that if the clouds are below 10% we can classify that as a clear weather condition. Moreover, the weather can be called "cloudy" when the percentage of clouds exceeds the 25% mark. The graph also suggests that high-intensity rainfall will occur when the cloud percentage crosses the 90% mark.

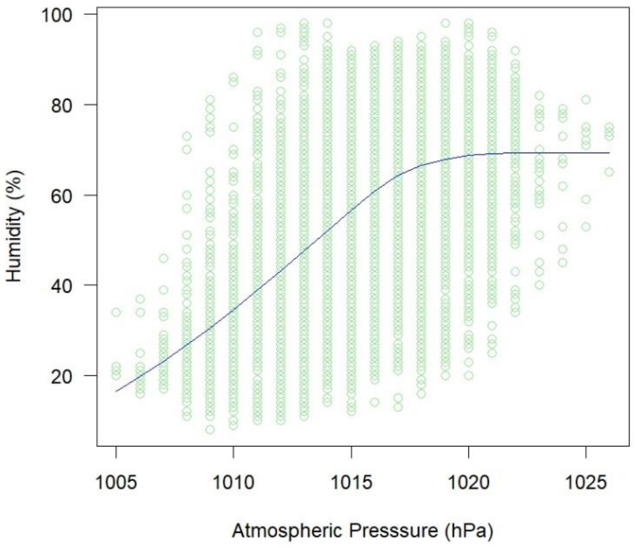
After plotting graphs between different numerical variables, which show have strong correlation by using the cor() function, we can take a look at their scatterplots below and find some insights. Firstly, as expected the graph between temperature and feels like temperature shows a strong linear trajectory, which means both the variables increase at the same rate. The second graph shows the relation between temperature and pressure. According to this graph, atmospheric pressure decreases gradually as the temperature starts to increase. The graph between temperature and humidity shows some interesting details. The graph shows that from

0 to 15 °C, humidity remains at 60 to 70 %, however after the mark of 15 °C the humidity starts decreasing sharply and continues to go down until it reaches around 10% at 31 °C. This tells us that warmer temperature is associated with lower relative humidity for the city of Puebla. The last scatterplot is between pressure and humidity, it shows a gradual increase in both variables, however after reaching atmospheric pressure of around 1020 hPa, the humidity flatlines and stays at about 70% and doesn't show any change with the increase in pressure.


Scatterplot of Temp vs Feels like

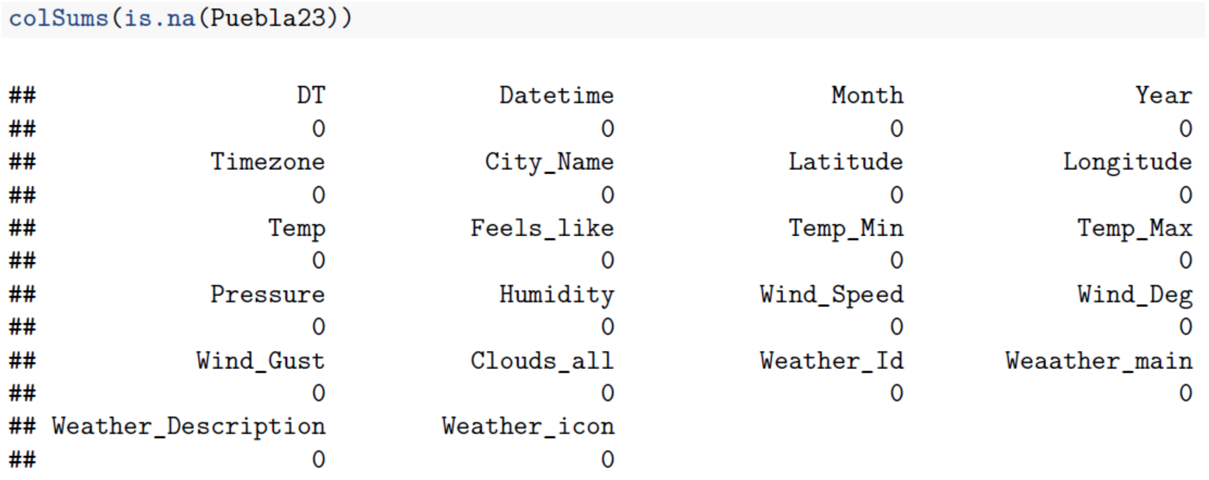

Scatterplot of Temp vs Pressure


Scatterplot of Temp vs Humidity


Scatterplot of Pressure vs Humidity

Missing Values

By using the colSums() function we can conclude that there are no missing values in our dataset and don’t have to fix any missing data.



```
colSums(is.na(Puebla23))
```

```
##                  DT            Datetime               Month                Year
##                   0                   0                   0                   0
##            Timezone           City_Name            Latitude           Longitude
##                   0                   0                   0                   0
##                Temp          Feels_like            Temp_Min            Temp_Max
##                   0                   0                   0                   0
##            Pressure            Humidity          Wind_Speed            Wind_Deg
##                   0                   0                   0                   0
##           Wind_Gust          Clouds_all          Weather_Id       Weaather_main
##                   0                   0                   0                   0
## Weather_Description        Weather_icon
##                   0                   0
```

Outliers

Previously, as we saw there were many outliers in our monthly boxplots, now to calculate these outliers we will use the formula LL = Q1 – 1.5*IQR & UL = Q3 + 1.5*IQR (where Q1 and Q3 are the values in 1st and 3rd quartile, and IQR is the interquartile range i.e. IQR = Q3 – Q1). Anything below the Lower Limit (LL) and above the Upper Limit (UL) are the outliers.

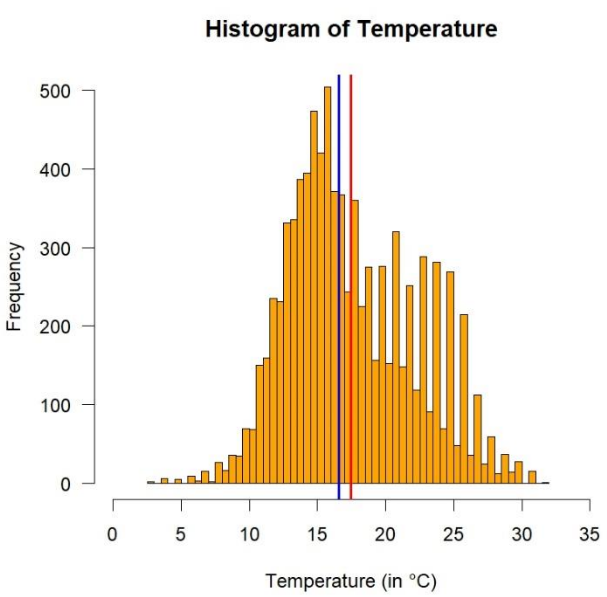
To get the upper and lower limit of a specific month we use the summary() function for that month and then calculate the IQR. After that, we used the Sum() function and put our conditions in its arguments to get the number of outliers per month.

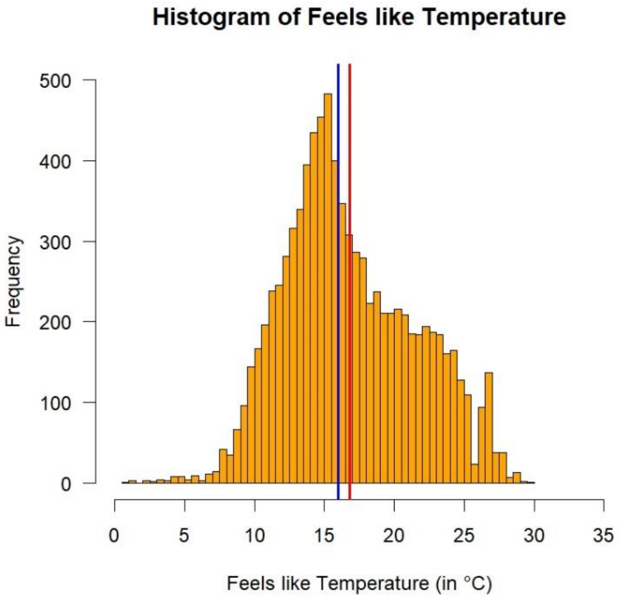
The following table gives us the number of outliers in each variable per month:

Months	**Outliers in Temp**	**Outliers in Feels like Temp**	**Outliers in Pressure**	**Outliers in Humidity**	**Outliers in Wind speed**	**Outliers in Wind degree**	**Outliers in Clouds**
Feb	-	-	-	-	-	-	105
Mar	-	-	-	-	-	-	-
Apr	-	-	-	-	-	26	-
May	-	-	-	-	13	-	-
Jun	-	-	1	-	12	-	-
Jul	-	-	-	-	4	-	-
Aug	-	-	-	-	6	-	-
Sept	-	-	-	-	8	-	-
Oct	-	1	-	-	7	-	45
Nov	-	-	-	-	5	-	-
Dec	3	1	3	-	6	-	-
Jan	-	-	-	-	19	-	-

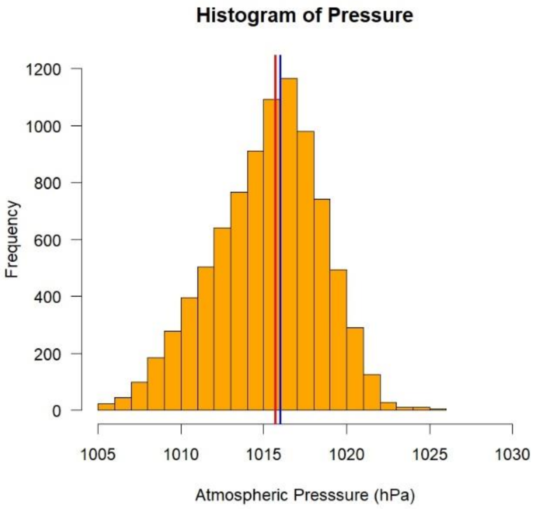
Distribution

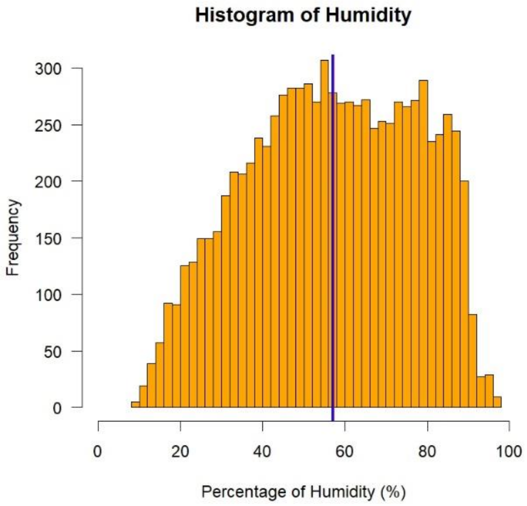
Using the hist() function we can see different histograms of the continuous variables of our dataset.



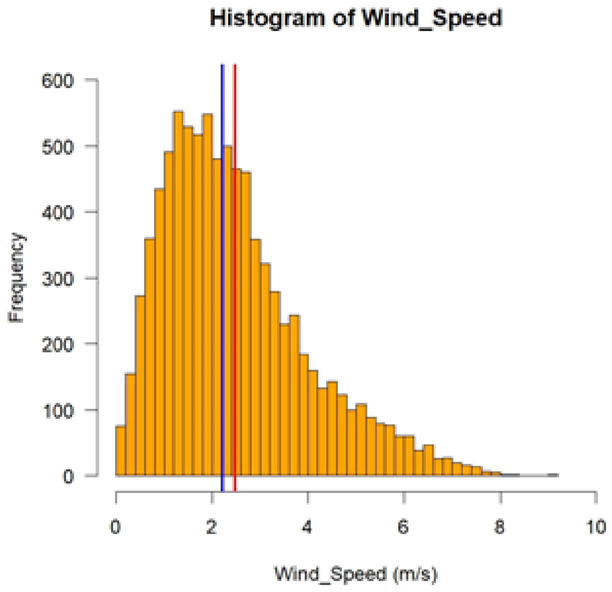


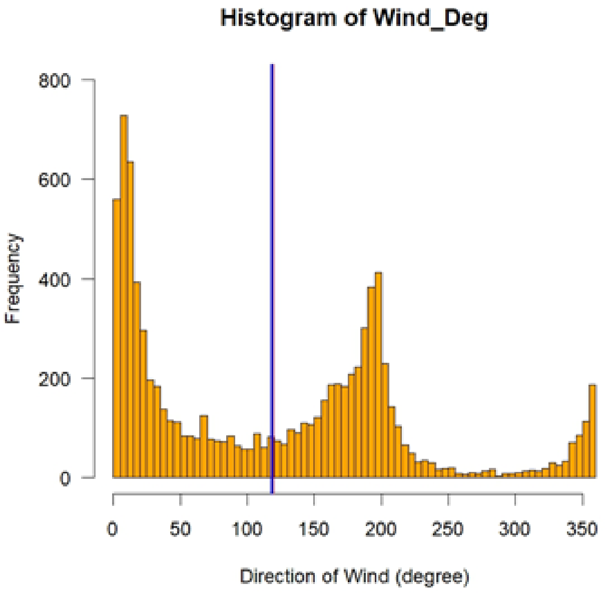
Firstly, we can see the similarity between the histograms of temperature and feels like temperature variables, both of which are symmetrical or Unimodal in shape. Also the mean and median are almost the same for both variables. Both have the peak frequency at around 15 °C which then gradually comes down.





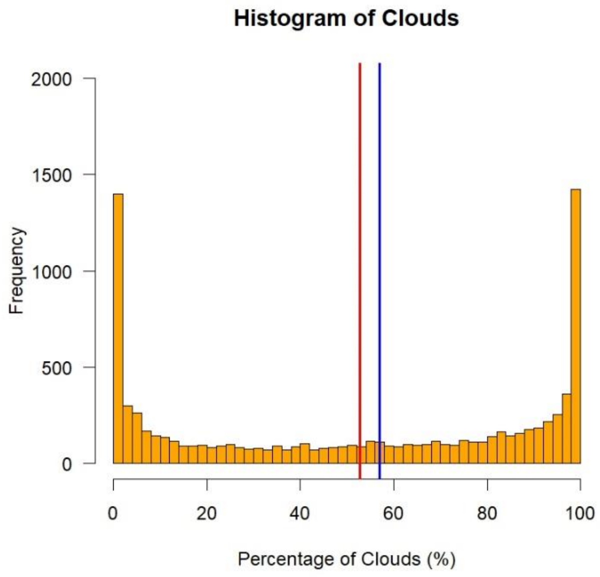
Similarly, the histogram of pressure also shows symmetry with increasing at first, reaching the peaks, and then going down. Looking at the humidity, its shape resembles more of a bimodal as there seem to be two peaks in it, the first one at around 55 and the second one at 80.





The wind speed histogram clearly shows right skewness at the peaks at the beginning which then go down slowly as the value of the speed increases. Meanwhile, the shape of the histogram of wind direction is multimodal as there are more than 2 peaks in it.

Finally, the last histogram of clouds shows a bimodal shape due to the 2 peaks at the beginning and the end of the graph with the rest of the graph having constant heights of bins.



Conclusion

This analysis gave us insights into Puebla’s climate for the year 2023, which also helped us understand how different elements of weather interact and relate to each other.

However, the too much uniformity between the ‘Temperature’, ‘Temp_min’, and ‘Temp_max’

Variables as well as ‘Wind speed’ wind_gust’, raises questions about the reliability of our data or whether certain weather conditions are causing these consistent patterns. Further investigating is needed to check these details so that we can have more accurate data to work with and to better understand the weather patterns in Puebla, Mexico.