FINAL PROJECT

PROFESSOR: Dr. HARPREET SHARMA

COURSE NUMBER: 202315

NAME: CHRISTINA NIKITHA STANLEY

INDEX

1.EXPLORATORY DATA ANALYSIS

```
> str(climate)
'data.frame':   7804 obs. of  24 variables:
 $ Date                       : chr  "1/1/2009" "1/2/2009" "1/3/2009" "1/4/2009" ...
 $ Average.temperature..Â.F.: num  37.8 43.2 25.7 9.3 23.5 24.8 34.2 42.1 30.3 26.2 ...
 $ Average.humidity....      : int  35 32 60 67 30 42 60 41 46 38 ...
 $ Average.dewpoint..Â.F.    : num  12.7 14.7 12.7 0.1 -5.3 4.6 21.6 20 11.4 3.6 ...
 $ Average.barometer..in.    : num  29.7 29.5 29.7 30.4 29.9 29.8 29.7 29.8 30 30.4 ...
 $ Average.windspeed..mph.   : num  26.4 12.8 8.3 2.9 16.7 16 20.4 17.5 6.9 18.2 ...
 $ Average.gustspeed..mph.   : num  36.8 18 12.2 4.5 23.1 23.9 30 25.2 10.6 24.6 ...
 $ Average.direction..Â.deg.: int  274 240 290 47 265 276 276 265 292 258 ...
 $ Rainfall.for.month..in.   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Rainfall.for.year..in.    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Maximum.rain.per.minute   : logi  NA NA NA NA NA NA ...
 $ Maximum.temperature..Â.F.: num  40 52 41 19 30 29 39 51 41 31 ...
 $ Minimum.temperature..Â.F.: num  34 37 6 0 15 19 27 36 19 22 ...
 $ Maximum.humidity....      : int  4 4 8 7 5 5 8 5 8 4 ...
 $ Minimum.humidity....      : int  27 16 35 35 13 27 46 28 27 29 ...
 $ Maximum.pressure          : num  29.8 29.7 30.2 30.6 30.2 ...
 $ Minimum.pressure          : num  29.6 29.3 29.3 30.2 29.6 ...
 $ Maximum.windspeed..mph.   : num  41.4 35.7 25.3 12.7 38 29.9 38 35.7 24.2 31.1 ...
 $ Maximum.gust.speed..mph.  : num  59 51 38 20 53 48 54 49 36 46 ...
 $ Maximum.heat.index..Â.F.  : num  40 52 41 32 32 32 39 51 41 32 ...
 $ Date1                      : chr  "1/1/2009" "1/2/2009" "1/3/2009" "1/4/2009" ...
 $ Month                      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ diff_pressure              : num  0.166 0.401 0.972 0.339 0.665 0.242 0.258 0.256 0.871
 0.198 ...
 $ Region                     : chr  "South" "South" "West" "South" ...
```

Figure 1 Structure of the data set

The structure of the data set reveals the no.of observations and the no.of variables in the data set. It alsotells what type of data consists of this dataset. This dataset has 7804 observations and 24 variables. [2]

```
> summary(climate)
     Date           Average.temperature..Â.F. Average.humidity.... Average.dewpoint..Â.F. Average.barometer..in. Average.windspeed..mph.
 Length:7804        Min.   :-12.10            Min.   : 9.00        Min.   :-22.20         Min.   :28.20          Min.   : 0.000
 Class :character   1st Qu.: 33.70            1st Qu.:36.00        1st Qu.: 12.10         1st Qu.:29.70          1st Qu.: 2.700
 Mode  :character   Median : 45.10            Median :47.00        Median : 22.50         Median :29.90          Median : 4.600
                    Mean   : 44.67            Mean   :48.88        Mean   : 23.13         Mean   :29.88          Mean   : 5.759
                    3rd Qu.: 58.00            3rd Qu.:61.00        3rd Qu.: 35.40         3rd Qu.:30.00          3rd Qu.: 8.000
                    Max.   : 76.30            Max.   :94.00        Max.   : 55.10         Max.   :31.00          Max.   :26.400
 Average.gustspeed..mph. Average.direction..Â.deg. Rainfall.for.month..in. Rainfall.for.year..in. Maximum.rain.per.minute Maximum.temperature..Â.F.
 Min.   :  0.00          Min.   :  0               Min.   :0.0000          Min.   : 0.000         Mode:logical            Min.   :-6.10
 1st Qu.:  4.50          1st Qu.:116               1st Qu.:0.0500          1st Qu.: 0.980         NA's:7804               1st Qu.:43.90
 Median :  7.10          Median :253               Median :0.2200          Median : 5.080                                 Median :57.25
 Mean   : 10.01          Mean   :216               Mean   :0.4511          Mean   : 5.486                                 Mean   :57.56
 3rd Qu.: 12.10          3rd Qu.:282               3rd Qu.:0.6700          3rd Qu.: 9.050                                 3rd Qu.:73.20
 Max.   :240.40          Max.   :360               Max.   :4.4800          Max.   :16.410                                 Max.   :92.70
 Minimum.temperature..Â.F. Maximum.humidity.... Minimum.humidity.... Maximum.pressure Minimum.pressure Maximum.windspeed..mph. Maximum.gust.speed..mph.
 Min.   :-27.70            Min.   :  1.00       Min.   : 0.00        Min.   :29.34    Min.   :13.27    Min.   :  0.00          Min.   :  0.00
 1st Qu.: 23.00            1st Qu.: 63.00       1st Qu.:15.00        1st Qu.:29.87    1st Qu.:29.56    1st Qu.: 13.80          1st Qu.: 19.60
 Median : 32.80            Median : 81.00       Median :22.00        Median :30.02    Median :29.71    Median : 18.40          Median : 27.60
 Mean   : 31.23            Mean   : 73.67       Mean   :26.02        Mean   :30.05    Mean   :29.70    Mean   : 19.84          Mean   : 33.97
 3rd Qu.: 41.80            3rd Qu.: 89.00       3rd Qu.:32.00        3rd Qu.:30.20    3rd Qu.:29.87    3rd Qu.: 24.20          3rd Qu.: 34.50
 Max.   : 65.70            Max.   :100.00       Max.   :90.00        Max.   :31.20    Max.   :30.86    Max.   :181.70          Max.   :255.30
 Maximum.heat.index..Â.F.    Date1              Month        diff_pressure        Region
 Min.   :-6.10            Length:7804        Min.   : 1.000   Min.   : 0.0000   Length:7804
 1st Qu.:43.90            Class :character   1st Qu.: 3.000   1st Qu.: 0.2200   Class :character
 Median :57.20            Mode  :character   Median : 6.000   Median : 0.2930   Mode  :character
 Mean   :58.09                               Mean   : 6.396   Mean   : 0.3438
 3rd Qu.:77.30                               3rd Qu.: 9.000   3rd Qu.: 0.3950
 Max.   :88.40                               Max.   :12.000   Max.   :16.6020
```

Figure 2 Summary of the data set

The summary of the data set provides details of the dataset like minimum, maximum values, quartilevalues, mean and median values of every variable in the dataset. [2]

```
> colSums(is.na(climate))
                    Date Average.temperature..Â.F.      Average.humidity....
                       0                         0                         0
   Average.dewpoint..Â.F.     Average.barometer..in.    Average.windspeed..mph.
                       0                         0                         0
  Average.gustspeed..mph. Average.direction..Â.deg.    Rainfall.for.month..in.
                       0                         0                         0
   Rainfall.for.year..in.    Maximum.rain.per.minute Maximum.temperature..Â.F.
                       0                      7804                         0
Minimum.temperature..Â.F.        Maximum.humidity....      Minimum.humidity....
                       0                         0                         0
         Maximum.pressure          Minimum.pressure    Maximum.windspeed..mph.
                       0                         0                         0
 Maximum.gust.speed..mph.   Maximum.heat.index..Â.F.                     Date1
                       0                         0                         0
                    Month              diff_pressure                    Region
                       0                         0                         0
```

Figure 3 Checking for missing values

We check for any missing values using the is.na() function and add up for all the variables using the colSums() function. From the above figure we can see that Maximum rain per minute variable has 7804 missing values. [5]

```
> climate <- subset(climate, select = -c(Maximum.rain.per.minute) )
> colnames(climate)
 [1] "Date"                        "Average.temperature..Â.F."
 [3] "Average.humidity...."        "Average.dewpoint..Â.F."
 [5] "Average.barometer..in."      "Average.windspeed..mph."
 [7] "Average.gustspeed..mph."     "Average.direction..Â.deg."
 [9] "Rainfall.for.month..in."     "Rainfall.for.year..in."
[11] "Maximum.temperature..Â.F."   "Minimum.temperature..Â.F."
[13] "Maximum.humidity...."        "Minimum.humidity...."
[15] "Maximum.pressure"            "Minimum.pressure"
[17] "Maximum.windspeed..mph."     "Maximum.gust.speed..mph."
[19] "Maximum.heat.index..Â.F."    "Date1"
[21] "Month"                       "diff_pressure"
[23] "Region"
```

Figure 4 Removing an unnecessary column

Now, the data set is being subset as we do not require the maximum rain per minute variable since it contains only empty values.

```
> #renaming month
> climate$Month[climate$Month == 1] = "January"
> climate$Month[climate$Month == 2] = "February"
> climate$Month[climate$Month == 3] = "March"
> climate$Month[climate$Month == 4] = "April"
> climate$Month[climate$Month == 5] = "May"
> climate$Month[climate$Month == 6] = "June"
> climate$Month[climate$Month == 7] = "July"
> climate$Month[climate$Month == 8] = "August"
> climate$Month[climate$Month == 9] = "September"
> climate$Month[climate$Month == 10] = "October"
> climate$Month[climate$Month == 11] = "November"
> climate$Month[climate$Month == 12] = "December"

> unique(climate$Month)
 [1] "January"   "February"  "March"     "April"     "May"       "June"     
 [7] "July"      "August"    "September" "October"   "November"  "December" 
```

Figure 5 Characterizing the month variable

The month variable contains numeric month values and hence it is being characterized accordingly.

```
> climate %>%
+   describe()
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Date*	1	7804	1951.50	1126.48	1951.50	1951.50	1446.28	1.00	3902.00	3901.00	0.00	-1.20	12.75
Average.temperature..Â.F.	2	7804	44.67	15.33	45.10	45.42	18.24	-12.10	76.30	88.40	-0.39	-0.40	0.17
Average.humidity....	3	7804	48.88	17.44	47.00	48.32	17.79	9.00	94.00	85.00	0.27	-0.55	0.20
Average.dewpoint..Â.F.	4	7804	23.13	14.63	22.50	23.27	17.05	-22.20	55.10	77.30	-0.06	-0.78	0.17
Average.barometer..in.	5	7804	29.88	0.25	29.90	29.87	0.30	28.20	31.00	2.80	0.27	1.16	0.00
Average.windspeed..mph.	6	7804	5.76	4.02	4.60	5.26	3.41	0.00	26.40	26.40	1.13	1.04	0.05
Average.gustspeed..mph.	7	7804	10.01	14.12	7.10	8.16	4.74	0.00	240.40	240.40	9.10	107.16	0.16
Average.direction..Â.deg.	8	7804	216.04	97.67	253.00	224.71	45.96	0.00	360.00	360.00	-0.80	-0.73	1.11
Rainfall.for.month..in.	9	7804	0.45	0.60	0.22	0.33	0.30	0.00	4.48	4.48	2.47	8.07	0.01
Rainfall.for.year..in.	10	7804	5.49	4.53	5.08	5.14	6.02	0.00	16.41	16.41	0.43	-0.93	0.05
Maximum.temperature..Â.F.	11	7804	57.56	17.75	57.25	58.06	21.57	-6.10	92.70	98.80	-0.21	-0.77	0.20
Minimum.temperature..Â.F.	12	7804	31.23	14.12	32.80	32.20	13.79	-27.70	65.70	93.40	-0.67	0.46	0.16
Maximum.humidity....	13	7804	73.67	20.38	81.00	76.80	14.83	1.00	100.00	99.00	-1.34	1.39	0.23
Minimum.humidity....	14	7804	26.02	15.62	22.00	23.65	11.86	0.00	90.00	90.00	1.47	2.24	0.18
Maximum.pressure	15	7804	30.05	0.26	30.02	30.03	0.25	29.34	31.20	1.87	0.55	0.57	0.00
Minimum.pressure	16	7804	29.70	0.45	29.71	29.71	0.23	13.27	30.86	17.59	-17.01	508.46	0.01
Maximum.windspeed..mph.	17	7804	19.84	12.23	18.40	18.95	6.82	0.00	181.70	181.70	7.19	79.00	0.14
Maximum.gust.speed..mph.	18	7804	33.97	38.63	27.60	27.92	10.23	0.00	255.30	255.30	5.10	26.32	0.44
Maximum.heat.index..Â.F.	19	7804	58.09	17.95	57.20	59.00	25.95	-6.10	88.40	94.50	-0.26	-0.95	0.20
Date1*	20	7804	1951.50	1126.48	1951.50	1951.50	1446.28	1.00	3902.00	3901.00	0.00	-1.20	12.75
Month*	21	7804	6.49	3.40	7.00	6.50	4.45	1.00	12.00	11.00	-0.02	-1.18	0.04
diff_pressure	22	7804	0.34	0.41	0.29	0.31	0.13	0.00	16.60	16.60	22.74	739.39	0.00
Region*	23	7804	2.82	1.46	3.00	2.77	1.48	1.00	5.00	4.00	0.20	-1.32	0.02

Figure 6 Description of the dataset

The describe function from the package "psych" is used to give a description of the entire dataset. This is helpful to analyze the data which one single function. [1][2]

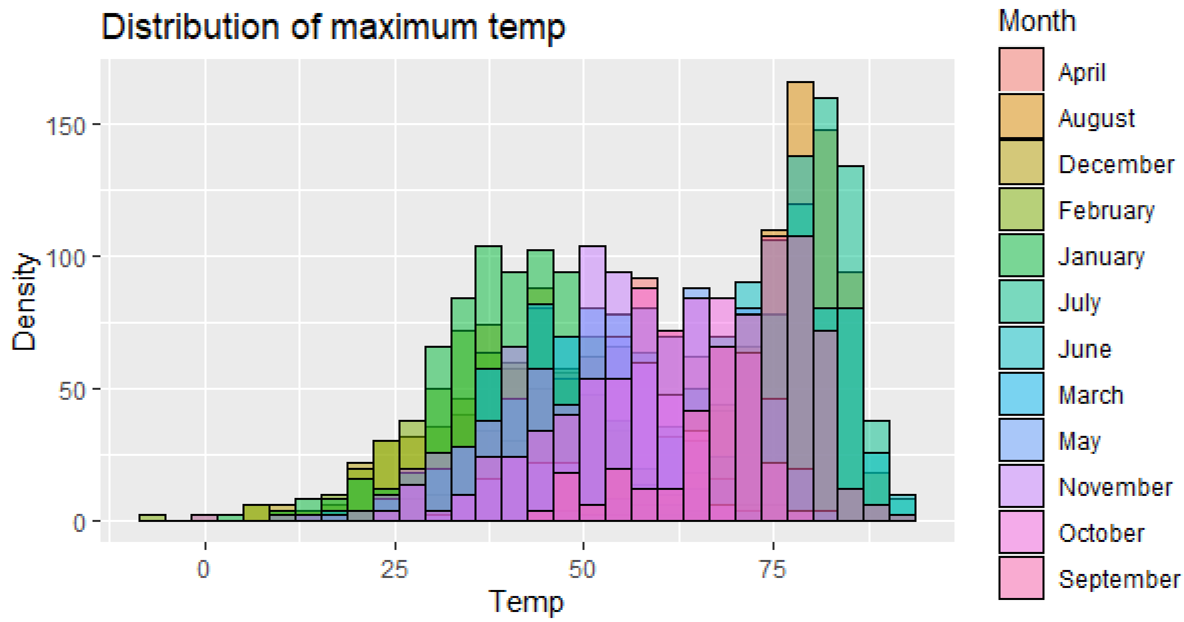

Figure 7 Distribution of maximum temperature by month

The above histogram is distribution of the maximum temperature which is differentiated by months. The highest temperature is about 80°F especially in the month of August. [4]

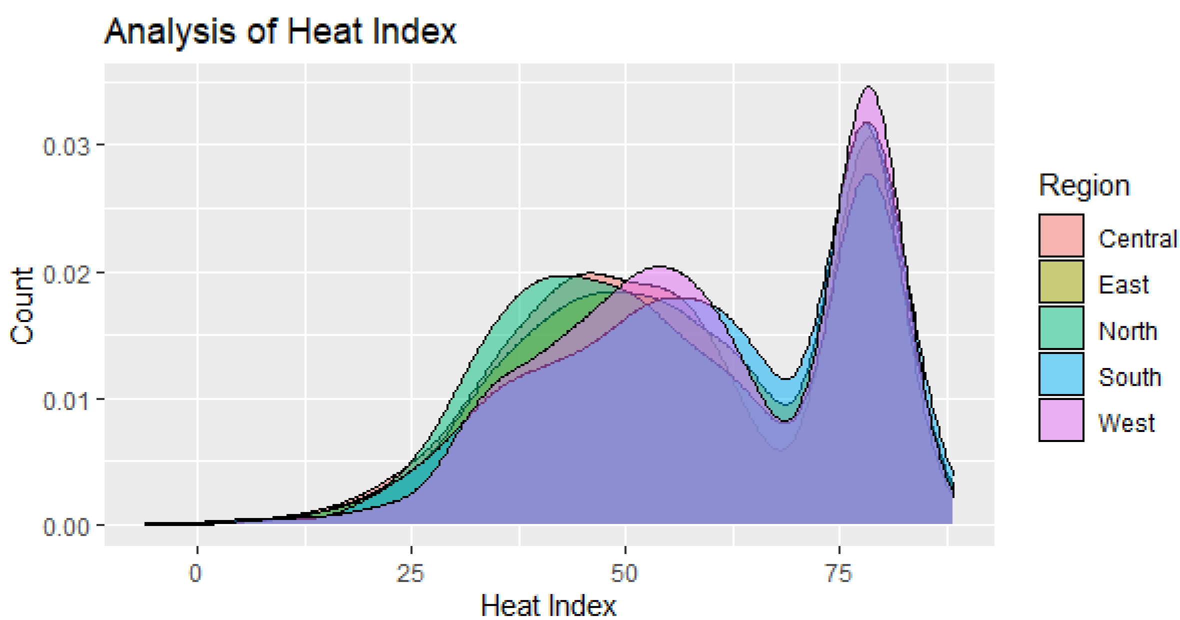

Figure 8 Analysis of heat index

The above plot is a density plot on the heat index which is compared by regions. The highest heat index is recorded in the western region. [4]

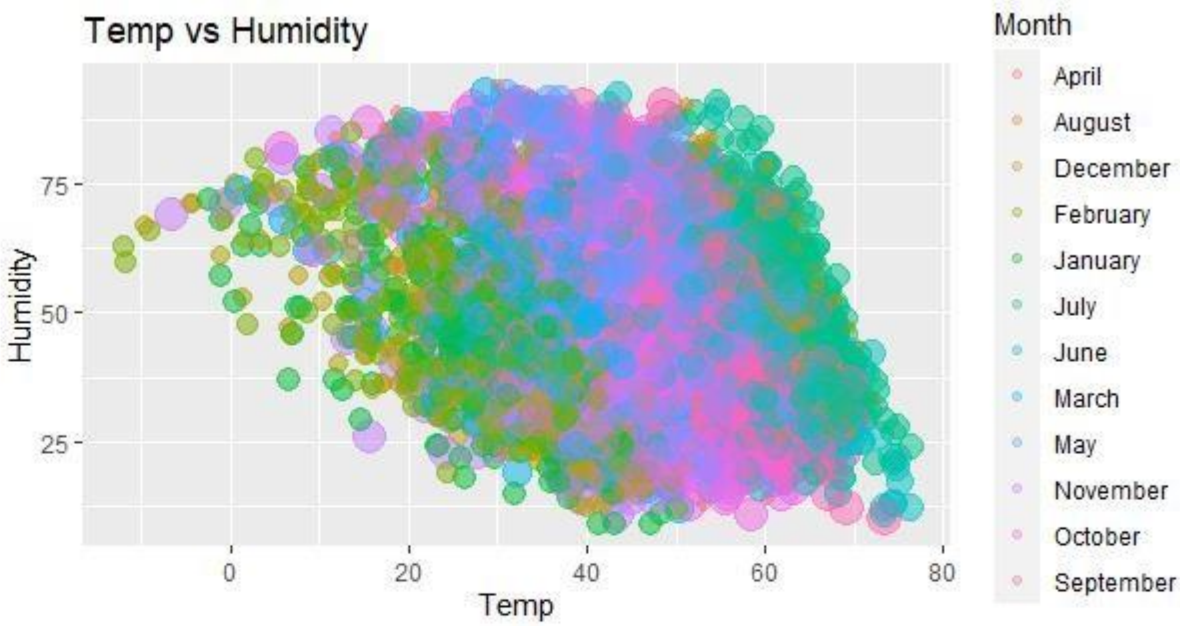

Figure 9 Analysis of Temperature vs Humidity

The scatter plot is an analysis of maximum temperature vs maximum humidity. It is differentiated by months and the size is differentiated by region. The highest and the most frequent temperature and humidity is recorded in the months of April and May.

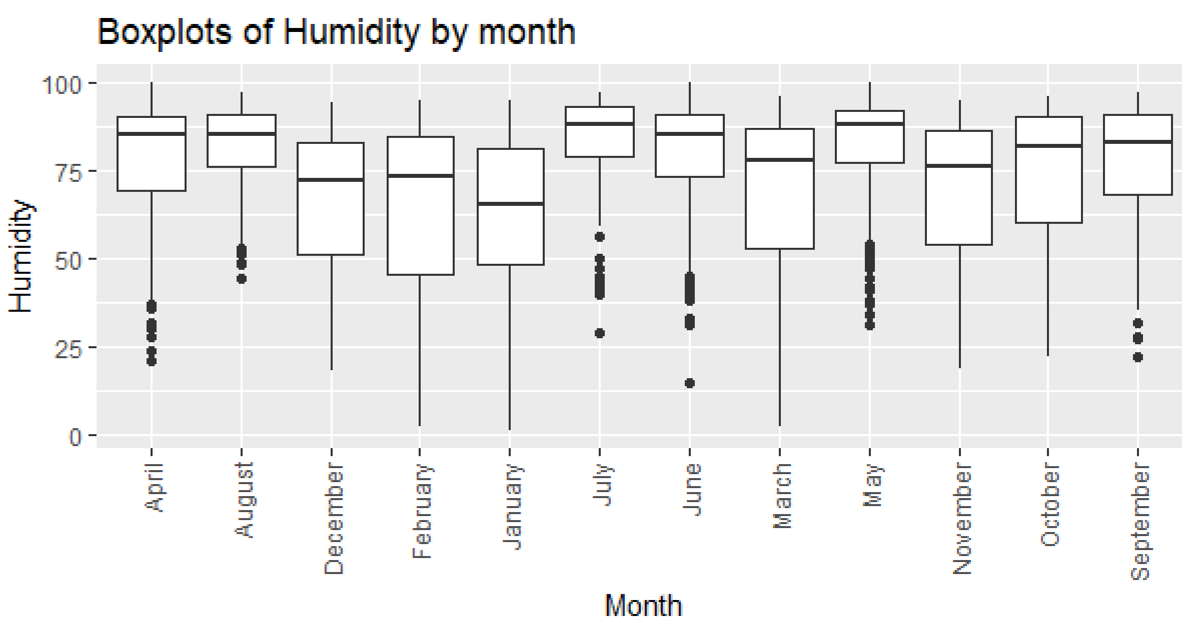

Figure 10 Analysis of Humidity by month

The above is an analysis of humidity based on each month of the year. The least humidity is recorded in the month of January and the highest is recorded in the months April, May and June.

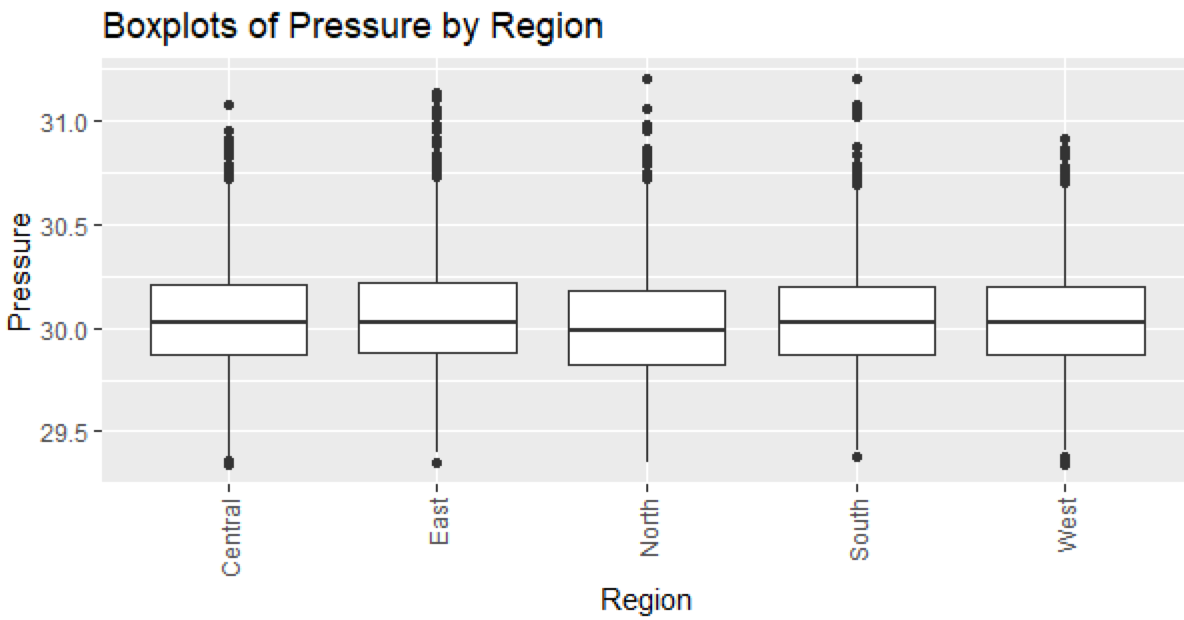

Figure 11 Analysis on pressure by region

The above is an analysis of pressure based on the regions. The least pressure is recorded in the western region and the highest is recorded in the northern region.

Table 1 Cross table on maximum and minimum pressure by region

.id	label	variable	Region				
			Central	East	North	South	West
Maximum.pressure	Maximum.pressure	Min / Max	29.3 / 31.1	29.4 / 31.1	29.4 / 31.2	29.4 / 31.2	29.3 / 30.9
		Med [IQR]	30.0 [29.9;30.2]	30.0 [29.9;30.2]	30.0 [29.8;30.2]	30.0 [29.9;30.2]	30.0 [29.9;30.2]
		Mean (std)	30.0 (0.3)	30.1 (0.3)	30.0 (0.3)	30.0 (0.3)	30.0 (0.3)
		N (NA)	1964 (0)	1700 (0)	1441 (0)	1189 (0)	1510 (0)
Minimum.pressure	Minimum.pressure	Min / Max	23.8 / 30.6	23.6 / 30.7	13.3 / 30.9	27.5 / 30.9	13.3 / 30.6
		Med [IQR]	29.7 [29.6;29.9]	29.7 [29.6;29.9]	29.7 [29.5;29.9]	29.7 [29.6;29.9]	29.7 [29.6;29.9]
		Mean (std)	29.7 (0.3)	29.7 (0.4)	29.7 (0.7)	29.7 (0.3)	29.7 (0.6)
		N (NA)	1964 (0)	1700 (0)	1441 (0)	1189 (0)	1510 (0)

This is a cross table gives us brief detail of the maximum and minimum pressure in each region. It mentions about the maximum/minimum values, the median, the mean and the no.of missing values. The maximum pressure is recorded in the northern and southern region and the lowest pressure is recorded in the central and western region. [3]

2.QUESTIONS

1. Are there discrepancies between the sample mean temperature and the fixed temperature value that are statistically significant?
2. Are the mean average temperatures equal in the eastern and western region?
3. Are the mean maximum temperatures in the months of April and August the same?

T-TESTS

1. One sample t-test

   Hypothesis

   H_0: The sample mean temperature and actual mean temperature are equal
   H_1: The sample mean temperature and actual mean temperature are not equal

```
        One Sample t-test

data:  climate$Average.temperature...F.
t = 49.979, df = 7803, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 36
95 percent confidence interval:
 44.33065 45.01081
sample estimates:
mean of x
 44.67073
```

Figure 12 One sample t-test for Average temperature

The test's p-value reveals that the difference from 0.05 is incredibly minor. Therefore, we are forced to adopt the alternative hypothesis rather than the null hypothesis. This indicates that there is a considerable disparity in the sample mean temperature and actual mean temperature. The sample estimated mean 44.67073°F. Therefore, there is a 95% confidence level that the mean temperature lies between 44.33065°F and 45.01081°F.

2. Two sample t-test

   Hypothesis

   H_0: The sample mean rainfall in a year is equal in the eastern and western regions
   H_1: The sample mean rainfall in a year is not equal in the eastern and western regions

```
        Welch Two Sample t-test

data:  east$Rainfall.for.year..in. and west$Rainfall.for.year..in.
t = 4.706, df = 3207.7, p-value = 2.633e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.4519137 1.0974325
sample estimates:
mean of x mean of y
 6.246229  5.471556
```

Figure 13 Two sample t-test for rainfall

The test's p-value reveals that the difference from 0.05 is incredibly minor. Therefore, we are forced to adopt the alternative hypothesis rather than the null hypothesis. This indicates that there is a considerable disparity in the mean rainfall in the eastern and western regions. The mean rainfall in the eastern region is 6.24 in, while the mean rainfall in the western region is 5.47 in. The gap in their rainfall can range from 0.45 to 1.1 in. Therefore, there is a 95% chance that there is more rainfall in the eastern region than western region.

3. Two sample t-test

Hypothesis

H_0: The sample mean humidity is equal in the months of April and August
H_1: The sample mean humidity is not equal in the months of April and August

```
        Welch Two Sample t-test

data:  April$Maximum.humidity.... and August$Maximum.humidity....
t = -4.9923, df = 1142.1, p-value = 6.894e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.311913 -2.314603
sample estimates:
mean of x mean of y
 78.13396  81.94721
```

Figure 14 Two sample t-test for humidity

The test's p-value reveals that the difference from 0.05 is incredibly minor. Therefore, we are forced to adopt the alternative hypothesis rather than the null hypothesis. This indicates that there is a considerable disparity in the mean humidity in the months of April and August. The mean humidity in the month of April is estimated as 78.1%, while the mean humidity in the month of August is estimated as 82%. Therefore, there is a 95% chance that the humidity in the month of April is less than the month of August

3.CORRELATION

```
> correlation_table_climate
```

	Average.temperature...F.	Average.humidity....	Average.dewpoint...F.	Average.windspeed..mph.	Average.gustspeed..mph.
Average.temperature...F.	1.0000000000	-0.2581030	0.7648304	-0.1671616	0.0007368343
Average.humidity....	-0.2581030386	1.0000000	0.4045572	-0.5161409	-0.1977588109
Average.dewpoint...F.	0.7648304167	0.4045572	1.0000000	-0.4553552	-0.1126577443
Average.windspeed..mph.	-0.1671615527	-0.5161409	-0.4553552	1.0000000	0.3936660412
Average.gustspeed..mph.	0.0007368343	-0.1977588	-0.1126577	0.3936660	1.0000000000

Figure 15 Correlation table

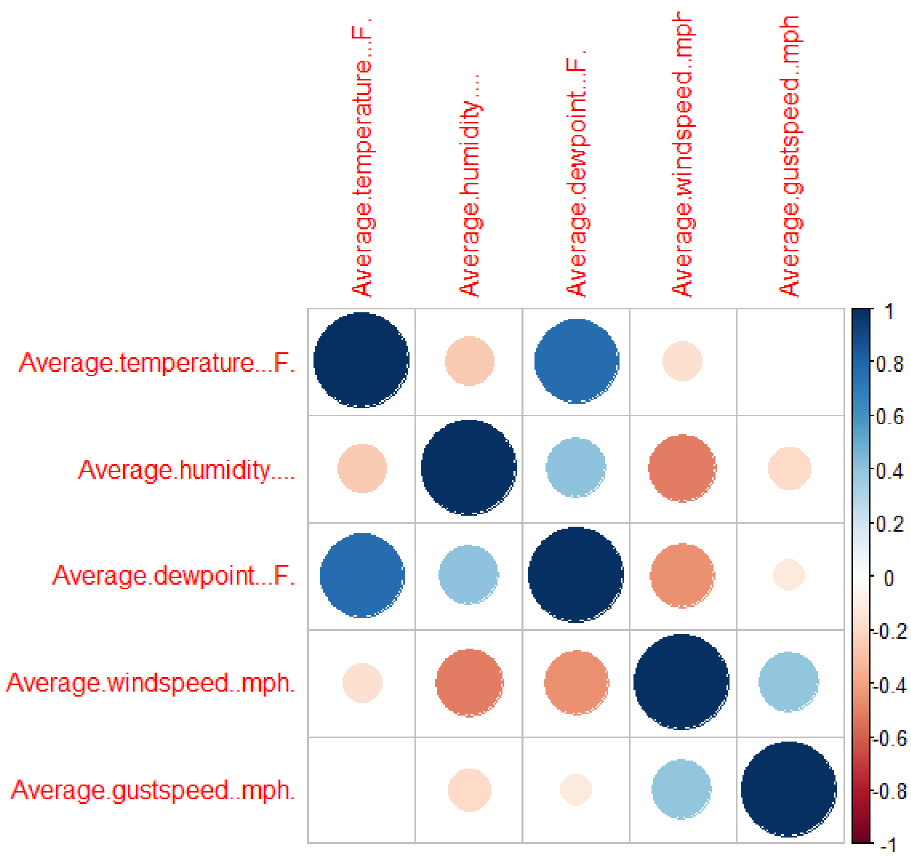

Figure 16 Correlation plot

From the above correlation table and plot we can infer that the average dewpoint and average temperature has the highest positive correlation. (0.7648304)

On the contrary, there is a negative correlation between average humidity and windspeed (-0.5161409) which indicates that the as the humidity rises the speed of the wind decreases. [6][7]

It is advised not to utilise more than five variables in a correlation table for reporting purposes. It will be challenging to find the correlation between all the variables, let alone describe the relationship between the dependent and independent variables, if there are more than five variables.

4.REGRESSION

Temperature vs Humidity

```
Call:
lm(formula = climate$Average.humidity.... ~ climate$Average.temperature...F.,
    data = climate)

Residuals:
    Min      1Q  Median      3Q     Max 
-40.838 -12.594  -0.669  11.687  45.038 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      61.99594    0.58771   105.5   <2e-16 ***
climate$Average.temperature...F. -0.29366    0.01244   -23.6   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.85 on 7802 degrees of freedom
Multiple R-squared:  0.06662,	Adjusted R-squared:  0.0665 
F-statistic: 556.8 on 1 and 7802 DF,  p-value: < 2.2e-16
```

Figure 17 Summary on regression between temperature and humidity

The regression is calculated between average humidity and temperature, the intercept value is 61.99594 with a standard error of 0.588 and the t value is 105.5 which is at the higher end also p value is almost equal to 0 which implies the predictor is more significant. Median is -0.669, the minimum and maximum value is -40.838 and 45.038 respectively with a residual standard error of 16.85. The R^2 value is 6.66% which indicates there is not much variation of the data, but it is significant. This issue can be solved by adding more variables which in turn will increase the R^2 value which will also improve the goodness of fit. [11][12]

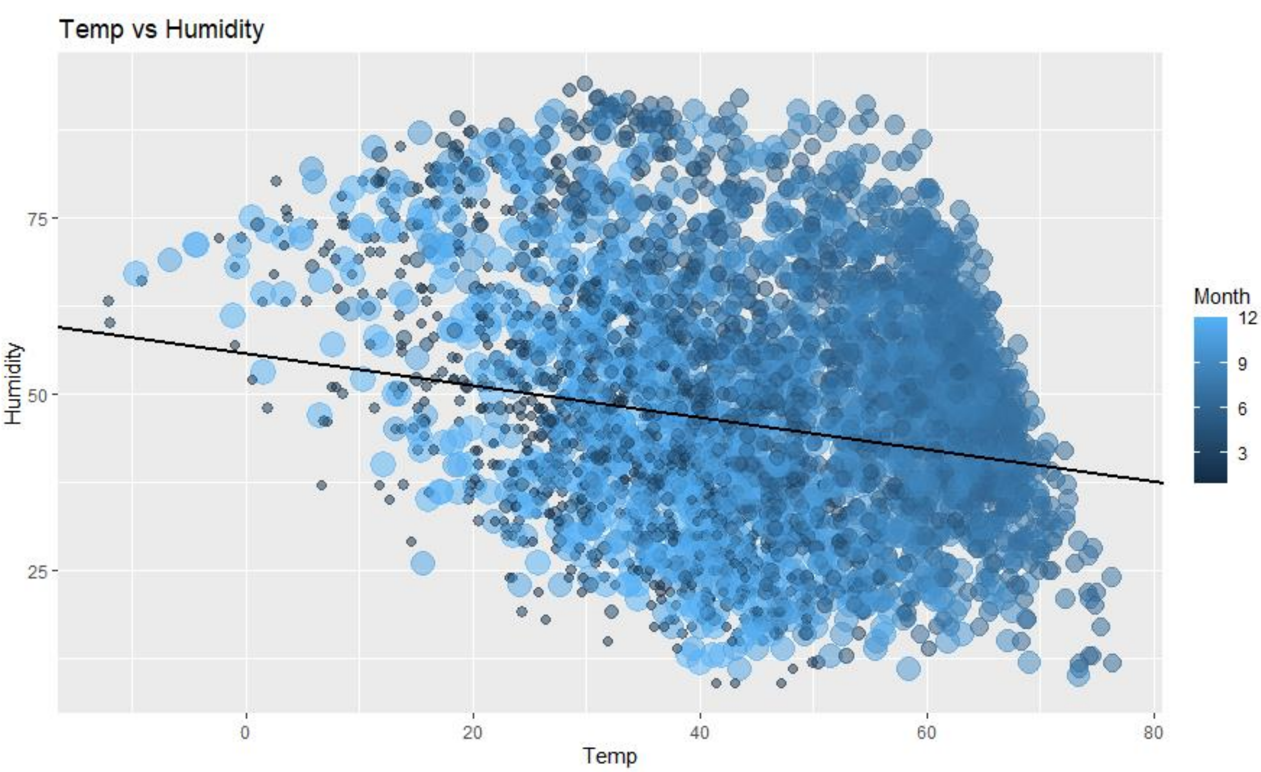

Figure 18 Temperature vs Humidity

The above plot represents a negative linear relationship between temperature and humidity. Although the best fit is weak, it is at least sufficient to infer that as the temperature increases, the humidity decreases. Hence, humidity is dependent on temperature. [11][12]

Temperature vs Rainfall

```
Call:
lm(formula = climate$Rainfall.for.year..in. ~ climate$Maximum.temperature...F.,
    data = climate)

Residuals:
   Min     1Q Median     3Q    Max
-5.859 -3.939 -1.213  3.306 13.421

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                       2.48682    0.17050   14.59   <2e-16 ***
climate$Maximum.temperature...F.  0.05211    0.00283   18.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.439 on 7802 degrees of freedom
Multiple R-squared:  0.04163,   Adjusted R-squared:  0.04151
F-statistic: 338.9 on 1 and 7802 DF,  p-value: < 2.2e-16
```

Figure 19 Summary on regression between temperature and rainfall

The regression is calculated between maximum temperature and rainfall, the intercept value is 2.48682 with a standard error of 0.17 and the t value is 14.5 which is at the higher end also p value is almost equal to 0 which implies the predictor is more significant. Median is -1.213, the minimum and maximum value is -5.8 and 13.4 respectively with a residual standard error of 4.439. The R^2 value is 4.16% which indicates there is not much variation of the data, but it is significant. This issue can be solved by adding more variables which in turn will increase the R^2 value which will also improve the goodness of fit. [11][12]

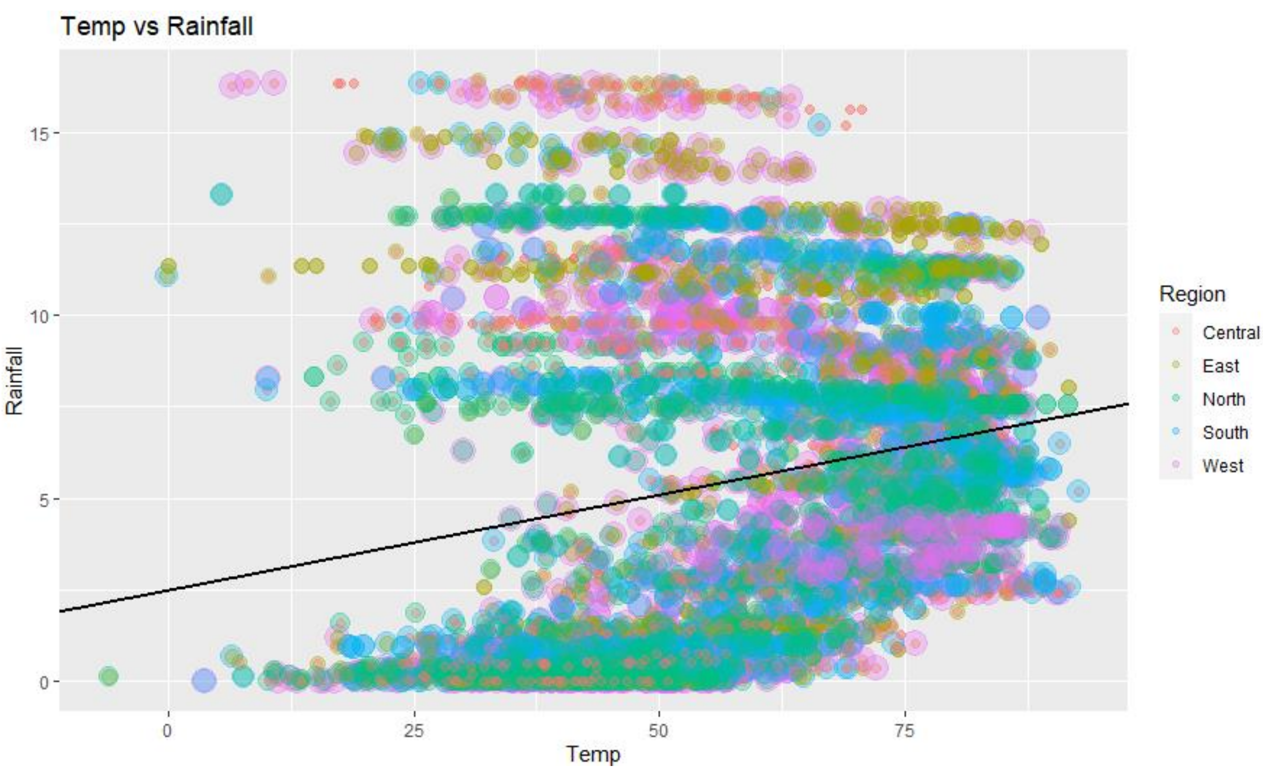

Figure 20 Temperature vs Rainfall

The above plot represents a positive linear relationship between temperature and rainfall. Although the best fit is weak, it is at least sufficient to infer that as the temperature increases, the rainfall also increases. Hence, rainfall is dependent on temperature. [11][12]

Pressure vs Heat Index

```
Call:
lm(formula = climate$Maximum.heat.index...F. ~ climate$Maximum.pressure,
    data = climate)

Residuals:
    Min      1Q  Median      3Q     Max
-48.361 -11.853   0.533  13.334  34.949

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1127.2274    20.3273   55.45   <2e-16 ***
climate$Maximum.pressure  -35.5841     0.6765  -52.60   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.43 on 7802 degrees of freedom
Multiple R-squared:  0.2618,    Adjusted R-squared:  0.2617
F-statistic:  2767 on 1 and 7802 DF,  p-value: < 2.2e-16
```

Figure 21 Summary on regression between pressure and heat index

The regression is calculated between maximum pressure and heat index, the intercept value is 1127.2274 with a standard error of 20.32 and the t value is 55.45 which is at the higher end also p value is almost equal to 0 which implies the predictor is more significant. Median is 0.533, the minimum and maximum value is -48.361 and 34.949 respectively with a residual standard error of 15.43. The R^2 value is 26.18% which indicates there is not much variation of the data, but it is significant. This issue can be solved by adding more variables which in turn will increase the R^2 value which will also improve the goodness of fit. [11][12]

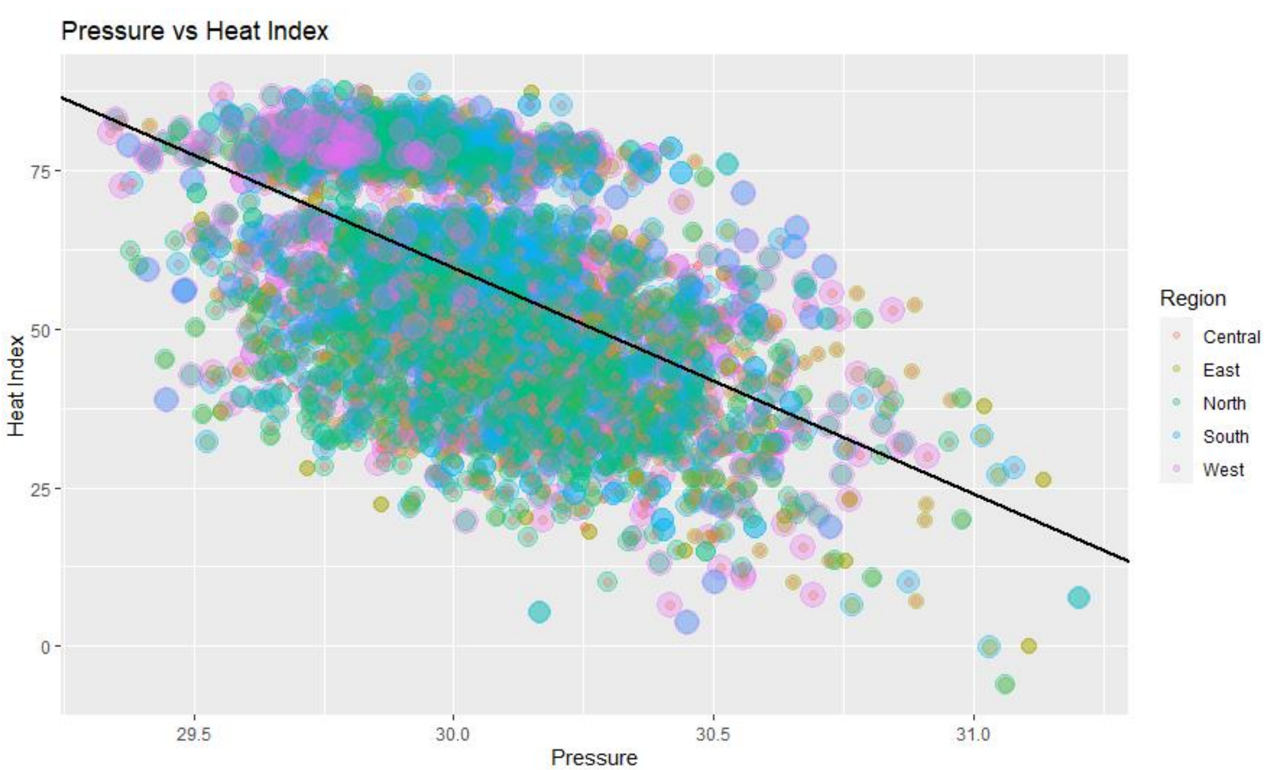

Figure 22 Pressure vs Heat Index

The above plot represents a negative linear relationship between pressure and heat index. Although the best fit is weak, it is at least sufficient to infer that as the pressure increases, the heat decreases. Hence, heat index is dependent on pressure. [11][12]

Windspeed vs temperature

```
Call:
lm(formula = climate$Average.temperature...F. ~ climate$Average.windspeed..mph. +
    Reg, data = climate)

Residuals:
    Min      1Q  Median      3Q     Max
-61.044 -10.124   1.569  12.287  32.917

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     48.04306    0.31025 154.853  < 2e-16 ***
climate$Average.windspeed..mph. -0.62969    0.04255 -14.799  < 2e-16 ***
Reg                              1.66717    0.47621   3.501 0.000466 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.1 on 7801 degrees of freedom
Multiple R-squared:  0.02947,   Adjusted R-squared:  0.02922
F-statistic: 118.4 on 2 and 7801 DF,  p-value: < 2.2e-16
```

Figure 23 Summary on regression between windspeed and temperature

The regression is calculated between maximum average temperature and windspeed in the southern region, the intercept value is 48.04306 with a standard error of 0.31 and the t value is 154.853 which is at the higher end also p value is almost equal to 0 which implies the predictor is more significant. Median is 1.569, the minimum and maximum value is -61.044 and 32.917 respectively with a residual standard error of 15.1. The R^2 value is 2.94% which indicates there is not much variation of the data, but it is significant. This issue can be solved by adding more variables which in turn will increase the R^2 value which will also improve the goodness of fit. [11][12]

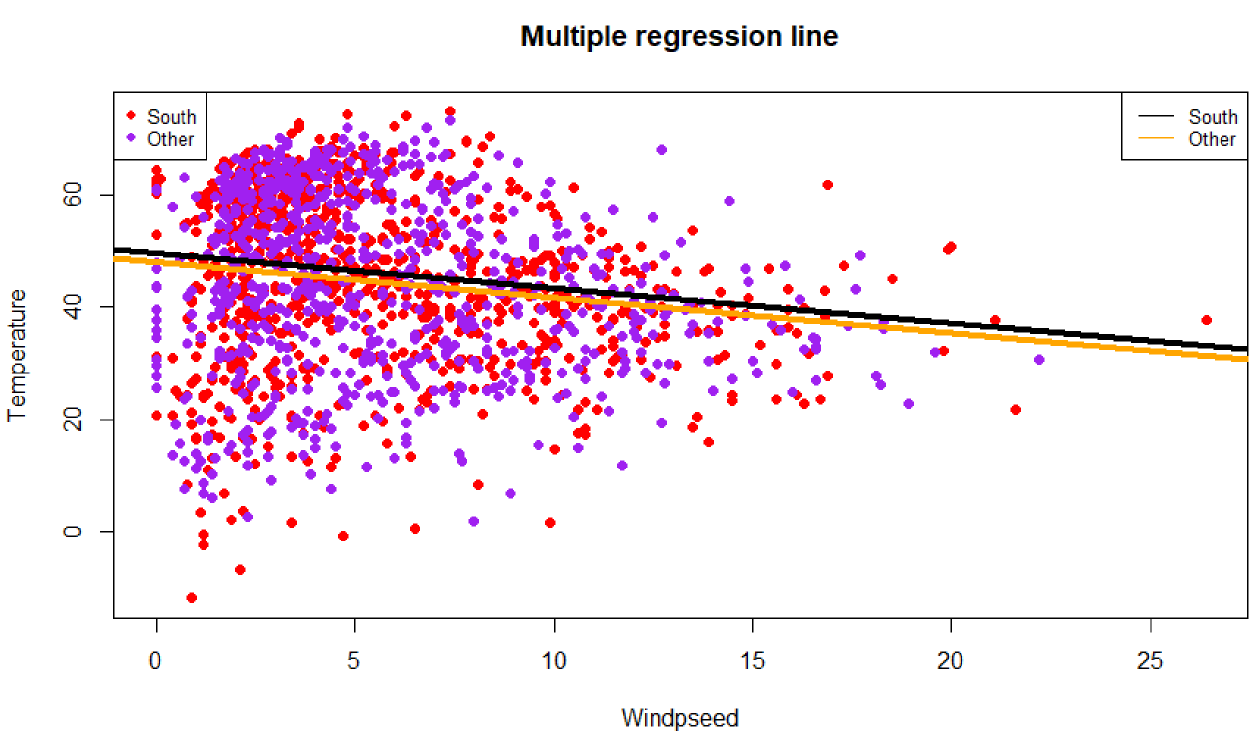

Figure 24 Multiple linear regression

The above plot represents a negative linear relationship between windspeed and temperature in southern region and all other regions. Although the best fit is weak, it is at least sufficient to infer that as the windspeed increases, the temperature decreases in all regions. Hence, temperature is dependent on windspeed. [10]

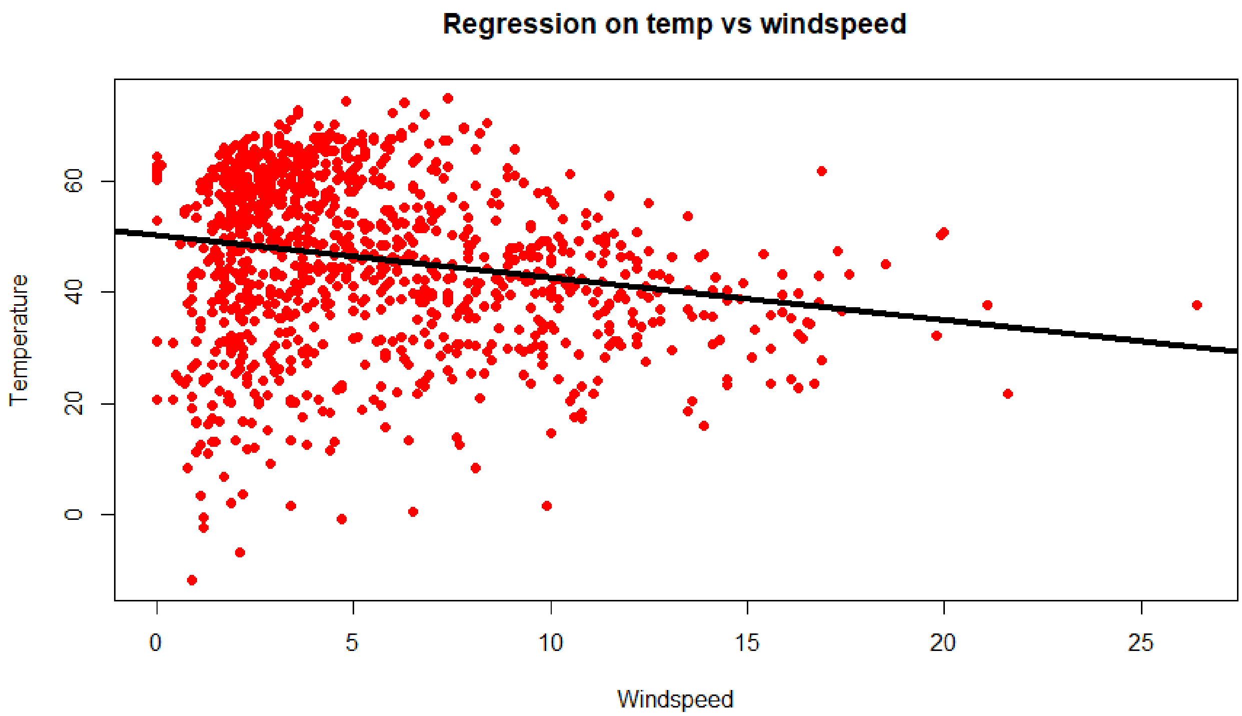

Figure 25 Windspeed vs Temperature in southern region

The above plot represents a negative linear relationship between windspeed and temperature only in the southern region. Although the best fit is weak, it is at least sufficient to infer that as the windspeed increases, the temperature decreases in the southern part. Hence, temperature is dependent on windspeed. [10]

5.SUMMARY

This dataset has 7804 observations and 24 variables.

The highest temperature is about 80°F especially in the month of August.

The highest heat index isrecorded in the western region.

The maximum pressure is recorded in the northern and southern region and the lowest pressure is recorded in the central and western region.

There is a considerable disparity in the sample mean temperature and actual mean temperature.

There is a considerable disparity in the mean rainfall in the eastern and western regions. The mean rainfall in the eastern region is 6.24 in, while the mean rainfall in the western region is 5.47 in. The gap in their rainfall can range from 0.45 to 1.1 in.

There is a considerable disparity in the mean humidity in the months of April and August. The mean humidity in the month of April is estimated as 78.1%, while the mean humidity in the month of August is estimated as 82%.

The average dewpoint and average temperature has the highest positive correlation.

As the temperature increases, the humidity decreases.

As the temperature increases, the rainfall also increases.

As the pressure increases, the heat decreases.

As the windspeed increases, the temperature decreases in the southern part.

6.BIBILIOGRAPHY

1. L., V., & K. (2021). *we are learning about… psych::describe*. RPub. https://www.rpubs.com/katherinewong/tuteweek10describe
2. McBride, L. (2020). *Time_Series_Multivariate_Weather*. Kaggle. https://www.kaggle.com/code/lunamcbride24/time-series-multivariate-weather/data?select=climate_data.csv
3. Chaltiel, D. (2022). *Introduction to Crosstable*. CRAN. https://cran.r-project.org/web/packages/crosstable/vignettes/crosstable.html
4. Kabacoff, R. I. (2017). *Histograms and Density Plots*. Quick-R. https://www.statmethods.net/graphs/density.html
5. (2022). *3 Ways to Find Columns with NA's in R [Examples]*. codingprof. https://www.codingprof.com/3-ways-to-find-columns-with-nas-in-r-examples/
6. Glen, S. (n.d.). *Residual Plot*. Statistics How to. https://www.statisticshowto.com/residual-plot/
7. Alammar, K. (2017). *If there is no correlation, is there a need to run linear regression?* Research Gate.https://www.researchgate.net/post/if_there_is_no_correlation_is_there_a_need_to_run_linear_regression
8. (n.d.). *Pearson Product-Moment Correlation*. Leard Statistics. https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php
9. Calvello, M. (2020). *Correlation vs. Regression Made Easy: Which to Use + Why*. G2. https://www.g2.com/articles/correlation-vs-regression
10. Blokhin, A. (2022). *Linear vs. Multiple Regression: What's the Difference?* Investopedia. https://www.investopedia.com/ask/answers/060315/what-difference-between-linear-regression-and-multiple-regression.asp#:~:text=A%20multiple%20regression%20formula%20has,the%20slope%20of%20the%20relationship.
11. Frost, J. (2018). *How To Interpret R-squared in Regression Analysis*. Statistics By Jim. https://statisticsbyjim.com/regression/interpret-r-squared-regression/
12. Fernando, J. (2021). *R-Squared Formula, Regression, and Interpretations*. Investopedia. https://www.investopedia.com/terms/r/r-squared.asp

7.APPENDDIX

```
#loading the data set
climate <-read.csv("C:/Users/nikit/Downloads/climate_data.csv")

#structure and summary
str(climate)
summary(climate)

#checking for missing values
colSums(is.na(climate))

#removing a column
climate <- subset(climate, select = -c(Maximum.rain.per.minute) )
colnames(climate)

#renaming month
climate$Month[climate$Month == 1] = "January"
climate$Month[climate$Month == 2] = "February"
climate$Month[climate$Month == 3] = "March"
climate$Month[climate$Month == 4] = "April"
climate$Month[climate$Month == 5] = "May"
climate$Month[climate$Month == 6] = "June"
climate$Month[climate$Month == 7] = "July"
climate$Month[climate$Month == 8] = "August"
climate$Month[climate$Month == 9] = "September"
climate$Month[climate$Month == 10] = "October"
climate$Month[climate$Month == 11] = "November"
climate$Month[climate$Month == 12] = "December"
unique(climate$Month)

#describe()
climate %>%
  describe()
```

```
#histogram of max temp
ggplot(climate, aes(x=Minimum.temperature...F., fill=Month)) +
  geom_histogram(colour = "black",  alpha = 0.5, position = "identity") + ggtitle("Distribution of maximum temp")+
  xlab("Temp")  + ylab("Density")


#density
ggplot(climate, aes(x= Maximum.heat.index...F., fill=Region)) +
  geom_density(alpha=0.5) +
  xlab("Heat Index")+
  ylab("Count") +
  ggtitle("Analysis of Heat Index") +
  scale_fill_discrete(name = "Region")


#scatter plot of temp vs humidity
ggplot(data=climate,aes(x=Average.temperature...F.,y=Average.humidity....,
                        color = Month, size=factor(Month)))+
  geom_point(alpha=0.3)+
  xlab("Temp")+
  ylab("Humidity") +
  labs(color="Month") +
  guides(size=FALSE)+
  ggtitle("Temp vs Humidity")



#boxplot of humidity by month
ggplot(climate, aes(x = Month, y = Maximum.humidity....)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  geom_boxplot() + xlab("Month") +
  ylab("Humidity") +
  ggtitle("Boxplots of Humidity by month")
```

```
#boxplot of maximum pressure
ggplot(climate, aes(x = Region, y = Maximum.pressure)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  geom_boxplot() + xlab("Region") +
  ylab("Pressure") +
  ggtitle("Boxplots of Pressure by Region")

#jitter plot of temp vs humidity
ggplot(climate, aes(Maximum.temperature..?.F., Maximum.humidity....,
                    color=Month )) + geom_jitter() +
  ggtitle("Scatterplot of temperature vs humidity")

#crosstable
library(gmodels)
crosstable(climate, c(Maximum.pressure, Minimum.pressure), by=Region) %>%
  as_flextable(keep_id=TRUE)
mean(climate$Average.temperature...F.)

east <- subset(climate, subset=(Region=="East"))
west <- subset(climate, subset=(Region=="West"))

April <- subset(climate, subset=(Month==4))
August <- subset(climate, subset=(Month==8))

#t test
t.test(climate$Average.temperature...F., mu=36, alternative = "two.sided")

#Two sample t-test
t.test(east$Rainfall.for.year..in., west$Rainfall.for.year..in.,var.equal = F)

#Two sample t-test
t.test(April$Maximum.humidity...., August$Maximum.humidity....,var.equal = F)
#filtering the necessary attributes
```

```
new_climate <- climate[c("Average.temperature...F.", "Average.humidity....", "Average.dewpoint...F.", "Average.windspeed..mph.", "Average.gustspeed..mph.")]

head(new_climate)



#correlation table

correlation_table_climate <- cor(new_climate)

correlation_table_climate


#correlation plot

corrplot(correlation_table_climate)


#regression plot of temp vs humidity


lr1 = lm(climate$Average.temperature...F. ~ climate$Average.humidity...., data=climate)

lr1

summary(lr1)


ggplot(data=climate,aes(x=Average.temperature...F.,y=Average.humidity....,

                        color = Month, size=factor(Month)))+

  geom_point(alpha=0.3)+

  geom_abline(slope=lr1$coefficients[2],

                                   intercept=lr1$coefficients[1],

                                   color="black",

                                   size=1)+

  xlab("Temp")+

  ylab("Humidity") +

  labs(color="Month") +

  guides(size=FALSE)+

  ggtitle("Temp vs Humidity")



#regression plot of temp vs rainfall
```

```
lr2 = lm(climate$Rainfall.for.year..in. ~ climate$Maximum.temperature...F., data=climate)
lr2
summary(lr2)

ggplot(data=climate,aes(x=climate$Maximum.temperature...F.,y=climate$Rainfall.for.year..in.,
                        color = Region, size=factor(Region)))+
  geom_point(alpha=0.3)+
  geom_abline(slope=lr2$coefficients[2],
              intercept=lr2$coefficients[1],
              color="black",
              size=1)+
  xlab("Temp")+
  ylab("Rainfall") +
  labs(color="Region") +
  guides(size=FALSE)+
  ggtitle("Temp vs Rainfall")

#regression plot of pressure vs heat index

lr3 = lm(climate$Maximum.heat.index...F. ~ climate$Maximum.pressure, data=climate)
lr3
summary(lr3)

ggplot(data=climate,aes(x=climate$Maximum.pressure,y=climate$Maximum.heat.index...F.,
                        color = Region, size=factor(Region)))+
  geom_point(alpha=0.3)+
  geom_abline(slope=lr3$coefficients[2],
              intercept=lr3$coefficients[1],
              color="black",
              size=1)+
  xlab("Pressure")+
  ylab("Heat Index") +
```

```
  labs(color="Region")  +
  guides(size=FALSE)+
  ggtitle("Pressure vs Heat Index")

#ifelse to convert categorical to dummy variable: Male 1 Female 0
climate$Reg <- ifelse(climate$Region=='South', 1,0)
climate$Reg
climate$Reg <- as.numeric(climate$Reg)

#regression plot of windspeed vs temperature in southern region

lr4 = lm(climate$Average.temperature...F. ~ climate$Average.windspeed..mph.+Reg, data=climate)
lr4
summary(lr4)

plot(climate$Average.windspeed..mph.[climate$Reg==1],
climate$Average.temperature...F.[climate$Reg==1], xlab="Windpseed",
     ylab="Temperature", main = "Multiple regression line", pch=19, col="red")
legend("topleft", legend=c("South", "Other"),
       col=c("red",  "purple"), pch=19,cex=0.8)
legend("topright", legend=c("South", "Other"),
       col=c("Black",  "orange"), lty = 1,cex=0.8)
points(climate$Average.windspeed..mph.[Salaries$Male==0],
climate$Average.temperature...F.[Salaries$Male==0], col="Purple", pch=19)

#coefficients of model
lr4$coefficients

#other regions
abline(a=lr4$coefficients[1], b=lr4$coefficients[2], col="Orange", lwd = 4)

#south
abline(a=lr4$coefficients[1]+lr4$coefficients[3], b=lr4$coefficients[2], col="Black", lwd = 4)
```

```
#creating subset

south = subset(climate,subset = (climate$Region=="South"))


#head of subsets

head(south, 10)


#summary

summary(south)


#regression analysis plot

plot(x=south$Average.windspeed..mph., y=south$Average.temperature...F., data=south, col="red",
ylab="Temperature", xlab="Windspeed",

    main="Regression on temp vs windspeed", pch=19)


#regression analysis

southline = lm(south$Average.temperature...F. ~ south$Average.windspeed..mph., data=south)

summary(southline)

abline(southline, lwd=4)
```