Urban Air Quality And Health Impact Analysis - EDA Project - Team 2

Abilasha Singh, Aswin Balaji Thippa Ramesh, Lixing Pan

2024-10-21

Contents

Dataset : Urban Air Quality and Health Impact Dataset

Variable | Definition

DateTime | Timestamp of the recorded data.

City | The U.S. city where the data was recorded (e.g., Phoenix, San Diego, New York City).

Temp_Max | Maximum temperature for the day (°F).

Temp_Min | Minimum temperature for the day (°F).

Temp_Avg | Average temperature for the day (°F).

Feels_Like_Max | Maximum "feels like" temperature for the day (°F).

Feels_Like_Min | Minimum "feels like" temperature for the day (°F).

Feels_Like_Avg | Average "feels like" temperature for the day (°F).

Dew_Point | Dew point temperature (°F).

Humidity | Relative humidity percentage.

Precipitation | Total precipitation for the day (inches).

Precip_Prob | Probability of precipitation (percentage).

Precip_Cover | Coverage of precipitation (percentage).

Precip_Type | Type of precipitation (e.g., rain, snow).

Snow | Amount of snowfall (inches).

Snow_Depth | Snow depth (inches).

Wind_Gust | Maximum wind gust speed (mph).

Wind_Speed | Average wind speed (mph).

Wind_Direction | Wind direction (degrees).

Pressure | Atmospheric pressure (hPa).

Cloud_Cover | Cloud cover percentage.

Visibility | Visibility distance (miles).

Solar_Radiation | Solar radiation (W/m^2).

Solar_Energy | Solar energy received (kWh).

UV_Index | UV index level.

Severe_Risk | Risk level of severe weather (e.g., low, moderate, high).

Sunrise | Sunrise time (HH:MM:SS).

Sunset | Sunset time (HH:MM:SS).

Moon_Phase | Phase of the moon (e.g., new moon, full moon).

Conditions | General weather conditions (e.g., clear, cloudy).

Description | Detailed description of the weather conditions.

Icon | Weather icon representation.

Stations | Weather stations reporting data.

Source | Data source information.

Temp_Range | Temperature range for the day (difference between max and min temperatures).

Heat_Index | Heat index value for the day.

Severity_Score | Score representing the severity of weather conditions.

Condition_Code | Code representing specific weather conditions.

Month | Month of the year.

Season | Season of the year (e.g., winter, spring).

Day_of_Week | Day of the week.

Is_Weekend | Indicator if the day is a weekend.

Health_Risk_Score | Score representing the potential health risk based on weather and air quality conditions.

```
library(dplyr)
library(ezids)
library(ggplot2)
library(tidyr)
library(reshape2)
library(gridExtra)
```

Data Loading

```
#loading the data
df=read.csv("air_quality_health_impact_data.csv")
head(df)
```

```
##     datetime datetimeEpoch tempmax tempmin temp feelslikemax feelslikemin
## 1 07-09-2024      1.73e+09    89.0    62.1 73.3         88.6         62.1
## 2 08-09-2024      1.73e+09    89.0    60.0 72.4         87.9         60.0
## 3 10-09-2024      1.73e+09    79.4    59.6 67.8         79.4         59.6
## 4 11-09-2024      1.73e+09    77.3    57.6 66.3         77.3         57.6
## 5 12-09-2024      1.73e+09    79.2    57.8 67.4         79.2         57.8
## 6 13-09-2024      1.73e+09    83.2    58.9 69.6         82.2         58.9
##   feelslike  dew humidity precip precipprob precipcover snow snowdepth windgust
## 1      73.3 59.8     66.3      0          0           0    0         0     16.1
## 2      72.3 57.6     62.5      0          0           0    0         0     13.9
## 3      67.8 57.2     70.7      0          0           0    0         0     17.4
## 4      66.3 56.8     73.1      0          4           0    0         0     23.0
## 5      67.4 55.6     68.3      0          5           0    0         0     17.9
## 6      69.5 54.2     60.5      0          0           0    0         0     16.1
##   windspeed winddir pressure cloudcover visibility solarradiation solarenergy
## 1       9.2     311     1012       12.0       10.0            268        23.4
## 2       8.1     310     1012       15.6        9.8            279        24.1
## 3       9.8     290     1012       18.8       12.4            275        23.8
## 4      13.4     274     1010       17.3       15.0            264        22.6
## 5      10.7     286     1007       14.2       15.0            262        22.6
## 6       8.9     288     1007        5.9       15.0            263        22.5
##   uvindex severerisk  sunrise sunriseEpoch   sunset sunsetEpoch moonphase
## 1       9         10 06:43:31     1.73e+09 19:26:34    1.73e+09      0.16
## 2       9         10 06:44:20     1.73e+09 19:25:03    1.73e+09      0.19
## 3       9         10 06:45:59     1.73e+09 19:22:01    1.73e+09      0.25
## 4       8         10 06:46:48     1.73e+09 19:20:29    1.73e+09      0.29
## 5       8         10 06:47:38     1.73e+09 19:18:57    1.73e+09      0.32
## 6       8         10 06:48:27     1.73e+09 19:17:25    1.73e+09      0.36
##   conditions                          description      icon source     City
## 1      Clear Clear conditions throughout the day. clear-day   comb San Jose
## 2      Clear Clear conditions throughout the day. clear-day   fcst San Jose
## 3      Clear Clear conditions throughout the day. clear-day   fcst San Jose
## 4      Clear Clear conditions throughout the day. clear-day   fcst San Jose
## 5      Clear Clear conditions throughout the day. clear-day   fcst San Jose
## 6      Clear Clear conditions throughout the day. clear-day   fcst San Jose
##   Temp_Range Heat_Index Severity_Score Month Season Day_of_Week Is_Weekend
## 1       26.9       75.8           3.41     9   Fall    Saturday       True
## 2       29.0       75.9           3.19     9   Fall      Sunday       True
## 3       19.8       73.5           3.54     9   Fall     Tuesday      False
## 4       19.7       72.9           3.90     9   Fall   Wednesday      False
## 5       21.4       74.3           3.39     9   Fall    Thursday      False
## 6       24.3       75.8           3.21     9   Fall      Friday      False
##   Health_Risk_Score
## 1              9.85
## 2              9.59
## 3              9.85
## 4             10.14
## 5              9.75
## 6              9.52
```

Data Description and Summmary

```r
#Shape of the dataset
paste("Row Count:",dim(df)[1],"Column Count:",dim(df)[2])
```

```
## [1] "Row Count: 27674 Column Count: 43"
```

```
#Structure of the dataset
str(df)
```

```
## 'data.frame':    27674 obs. of  43 variables:
##  $ datetime         : chr  "07-09-2024" "08-09-2024" "10-09-2024" "11-09-2024" ...
##  $ datetimeEpoch    : num  1.73e+09 1.73e+09 1.73e+09 1.73e+09 1.73e+09 ...
##  $ tempmax          : num  89 89 79.4 77.3 79.2 83.2 81.4 78.3 81.2 82.3 ...
##  $ tempmin          : num  62.1 60 59.6 57.6 57.8 58.9 59.4 59.8 59.3 60.9 ...
##  $ temp             : num  73.3 72.4 67.8 66.3 67.4 69.6 68.8 66.8 68.9 68.5 ...
##  $ feelslikemax     : num  88.6 87.9 79.4 77.3 79.2 82.2 81.3 78.3 79.6 80.2 ...
##  $ feelslikemin     : num  62.1 60 59.6 57.6 57.8 58.9 59.4 59.8 59.3 60.9 ...
##  $ feelslike        : num  73.3 72.3 67.8 66.3 67.4 69.5 68.8 66.8 68.6 68.4 ...
##  $ dew              : num  59.8 57.6 57.2 56.8 55.6 54.2 55.5 47.3 44.4 46.6 ...
##  $ humidity         : num  66.3 62.5 70.7 73.1 68.3 60.5 64.2 52.9 43.5 47.6 ...
##  $ precip           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ precipprob       : num  0 0 0 4 5 0 1 3.2 0 0 ...
##  $ precipcover      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ snow             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ snowdepth        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ windgust         : num  16.1 13.9 17.4 23 17.9 16.1 16.6 9.8 7.8 8.5 ...
##  $ windspeed        : num  9.2 8.1 9.8 13.4 10.7 8.9 9.8 8.9 8.1 10.3 ...
##  $ winddir          : num  311 310 290 274 286 ...
##  $ pressure         : num  1012 1012 1012 1010 1007 ...
##  $ cloudcover       : num  12 15.6 18.8 17.3 14.2 5.9 8.9 3.1 8.6 14.2 ...
##  $ visibility       : num  10 9.8 12.4 15 15 15 14.9 14.9 15 14.9 ...
##  $ solarradiation   : num  268 279 275 264 262 ...
##  $ solarenergy      : num  23.4 24.1 23.8 22.6 22.6 22.5 22.3 22 21.7 21.3 ...
##  $ uvindex          : num  9 9 9 8 8 8 8 8 8 8 ...
##  $ severerisk       : num  10 10 10 10 10 10 10 10 10 10 ...
##  $ sunrise          : chr  "06:43:31" "06:44:20" "06:45:59" "06:46:48" ...
##  $ sunriseEpoch     : num  1.73e+09 1.73e+09 1.73e+09 1.73e+09 1.73e+09 ...
##  $ sunset           : chr  "19:26:34" "19:25:03" "19:22:01" "19:20:29" ...
##  $ sunsetEpoch      : num  1.73e+09 1.73e+09 1.73e+09 1.73e+09 1.73e+09 ...
##  $ moonphase        : num  0.16 0.19 0.25 0.29 0.32 0.36 0.39 0.42 0.46 0.5 ...
##  $ conditions       : chr  "Clear" "Clear" "Clear" "Clear" ...
##  $ description      : chr  "Clear conditions throughout the day." "Clear conditions throughout the da
##  $ icon             : chr  "clear-day" "clear-day" "clear-day" "clear-day" ...
##  $ source           : chr  "comb" "fcst" "fcst" "fcst" ...
##  $ City             : chr  "San Jose" "San Jose" "San Jose" "San Jose" ...
##  $ Temp_Range       : num  26.9 29 19.8 19.7 21.4 24.3 22 18.5 21.9 21.4 ...
##  $ Heat_Index       : num  75.8 75.9 73.5 72.9 74.3 ...
##  $ Severity_Score   : num  3.41 3.19 3.54 3.9 3.39 3.21 3.26 2.58 2.38 2.45 ...
##  $ Month            : int  9 9 9 9 9 9 9 9 9 9 ...
##  $ Season           : chr  "Fall" "Fall" "Fall" "Fall" ...
##  $ Day_of_Week      : chr  "Saturday" "Sunday" "Tuesday" "Wednesday" ...
##  $ Is_Weekend       : chr  "True" "True" "False" "False" ...
##  $ Health_Risk_Score: num  9.85 9.59 9.85 10.14 9.75 ...
```

Appropriate Datatype Conversion

```
#Converting required columns as a factor variable
df$conditions <- as.factor(df$conditions)
df$description <- as.factor(df$description)
df$icon <- as.factor(df$icon)
df$source <- as.factor(df$source)
df$City <- as.factor(df$City)
df$Month <- as.factor(df$Month)
df$Season <- as.factor(df$Season)
df$Day_of_Week <- as.factor(df$Day_of_Week)
df$Is_Weekend <- as.factor(df$Is_Weekend)
```

```
#Checking for datatype of every column
sapply(df, class)
```

```
##          datetime     datetimeEpoch           tempmax           tempmin
##       "character"         "numeric"         "numeric"         "numeric"
##              temp      feelslikemax      feelslikemin         feelslike
##         "numeric"         "numeric"         "numeric"         "numeric"
##               dew          humidity            precip        precipprob
##         "numeric"         "numeric"         "numeric"         "numeric"
##       precipcover              snow         snowdepth          windgust
##         "numeric"         "integer"         "numeric"         "numeric"
##         windspeed           winddir          pressure        cloudcover
##         "numeric"         "numeric"         "numeric"         "numeric"
##        visibility    solarradiation       solarenergy           uvindex
##         "numeric"         "numeric"         "numeric"         "numeric"
##        severerisk           sunrise      sunriseEpoch            sunset
##         "numeric"       "character"         "numeric"       "character"
##       sunsetEpoch         moonphase        conditions       description
##         "numeric"         "numeric"          "factor"          "factor"
##              icon            source              City        Temp_Range
##          "factor"          "factor"          "factor"         "numeric"
##        Heat_Index    Severity_Score             Month            Season
##         "numeric"         "numeric"          "factor"          "factor"
##       Day_of_Week        Is_Weekend Health_Risk_Score
##          "factor"          "factor"         "numeric"
```

```
#summary of the dataset
summary(df)
```

```
##    datetime         datetimeEpoch         tempmax        tempmin
##  Length:27674       Min.   :1.73e+09   Min.   :70.2   Min.   :50.5
##  Class :character   1st Qu.:1.73e+09   1st Qu.:77.3   1st Qu.:58.4
##  Mode  :character   Median :1.73e+09   Median :81.4   Median :61.2
##                     Mean   :1.73e+09   Mean   :80.6   Mean   :61.4
##                     3rd Qu.:1.73e+09   3rd Qu.:83.2   3rd Qu.:64.6
##                     Max.   :1.73e+09   Max.   :90.5   Max.   :74.9
##
##       temp       feelslikemax   feelslikemin    feelslike         dew
##  Min.   :60.1   Min.   :68.8   Min.   :50.2   Min.   :59.3   Min.   :41.1
##  1st Qu.:66.5   1st Qu.:77.3   1st Qu.:58.0   1st Qu.:66.6   1st Qu.:47.5
##  Median :70.5   Median :81.0   Median :61.0   Median :70.4   Median :53.8
##  Mean   :69.9   Mean   :80.2   Mean   :61.4   Mean   :70.0   Mean   :53.1
```

```
##  3rd Qu.:73.4   3rd Qu.:83.1   3rd Qu.:64.7   3rd Qu.:73.3   3rd Qu.:58.5
##  Max.   :79.7   Max.   :90.3   Max.   :76.5   Max.   :80.1   Max.   :65.9
##
##     humidity        precip           precipprob     precipcover          snow
##  Min.   :38.5   Min.   :-0.02088   Min.   :-5.90   Min.   :-1.481   Min.   :0
##  1st Qu.:51.9   1st Qu.:-0.00288   1st Qu.:-0.29   1st Qu.:-0.282   1st Qu.:0
##  Median :57.5   Median : 0.00000   Median : 1.00   Median : 0.000   Median :0
##  Mean   :57.3   Mean   : 0.00091   Mean   : 2.35   Mean   : 0.050   Mean   :0
##  3rd Qu.:63.0   3rd Qu.: 0.00505   3rd Qu.: 3.28   3rd Qu.: 0.392   3rd Qu.:0
##  Max.   :76.6   Max.   : 0.02460   Max.   :20.81   Max.   : 2.228   Max.   :0
##
##    snowdepth    windgust       windspeed        winddir       pressure
##  Min.   :0   Min.   : 3.50   Min.   : 4.89   Min.   : 21   Min.   :1006
##  1st Qu.:0   1st Qu.: 9.21   1st Qu.: 8.25   1st Qu.:164   1st Qu.:1012
##  Median :0   Median :13.69   Median : 9.20   Median :201   Median :1016
##  Mean   :0   Mean   :13.11   Mean   : 9.33   Mean   :208   Mean   :1016
##  3rd Qu.:0   3rd Qu.:16.28   3rd Qu.:10.43   3rd Qu.:279   3rd Qu.:1021
##  Max.   :0   Max.   :23.42   Max.   :15.00   Max.   :330   Max.   :1031
##
##    cloudcover      visibility    solarradiation  solarenergy      uvindex
##  Min.   :-4.40   Min.   : 9.53   Min.   :214    Min.   :18.8   Min.   : 5.72
##  1st Qu.: 3.43   1st Qu.:11.93   1st Qu.:250    1st Qu.:21.6   1st Qu.: 6.89
##  Median :10.26   Median :14.90   Median :258    Median :22.2   Median : 7.98
##  Mean   :11.03   Mean   :13.63   Mean   :260    Mean   :22.5   Mean   : 7.68
##  3rd Qu.:16.63   3rd Qu.:15.05   3rd Qu.:268    3rd Qu.:23.3   3rd Qu.: 8.37
##  Max.   :29.63   Max.   :15.71   Max.   :313    Max.   :26.8   Max.   :10.00
##
##    severerisk      sunrise           sunriseEpoch         sunset
##  Min.   : 8.21   Length:27674       Min.   :1.73e+09   Length:27674
##  1st Qu.: 9.58   Class :character   1st Qu.:1.73e+09   Class :character
##  Median :10.00   Mode  :character   Median :1.73e+09   Mode  :character
##  Mean   :10.06                      Mean   :1.73e+09
##  3rd Qu.:10.63                      3rd Qu.:1.73e+09
##  Max.   :12.06                      Max.   :1.73e+09
##
##   sunsetEpoch         moonphase                conditions
##  Min.   :1.73e+09   Min.   :0.141   Clear           :22826
##  1st Qu.:1.73e+09   1st Qu.:0.231   Partially cloudy: 4848
##  Median :1.73e+09   Median :0.325
##  Mean   :1.73e+09   Mean   :0.346
##  3rd Qu.:1.73e+09   3rd Qu.:0.450
##  Max.   :1.73e+09   Max.   :0.647
##
##                               description                  icon
##  Becoming cloudy in the afternoon.   :  404   clear-day        :22826
##  Clear conditions throughout the day.:22422   partly-cloudy-day: 4848
##  Clearing in the afternoon.          :  404
##  Partly cloudy throughout the day.   : 4444
##
##
##
##   source                 City        Temp_Range     Heat_Index   Severity_Score
##  comb:  505   San Jose     :7777   Min.   : 8.1   Min.   :72.5   Min.   :1.85
##  fcst:27169   New York City:6060   1st Qu.:16.9   1st Qu.:76.1   1st Qu.:2.44
```

```
##               Philadelphia :3939   Median :19.8   Median :77.0   Median :2.72
##               Chicago      :3838   Mean   :19.3   Mean   :77.1   Mean   :2.85
##               Los Angeles  :3838   3rd Qu.:21.7   3rd Qu.:78.0   3rd Qu.:3.23
##               Dallas       : 909   Max.   :29.8   Max.   :81.6   Max.   :4.32
##               (Other)      :1313
##  Month      Season         Day_of_Week   Is_Weekend    Health_Risk_Score
##  9:27674   Fall:27674   Friday   :3737   False:18483   Min.   : 8.41
##                         Monday   :4343   True : 9191   1st Qu.: 9.06
##                         Saturday :3333                 Median : 9.28
##                         Sunday   :5858                 Mean   : 9.34
##                         Thursday :3535                 3rd Qu.: 9.59
##                         Tuesday  :2727                 Max.   :10.70
##                         Wednesday:4141
```

Dropping of unwanted columns

```
#Checking for unique values in all the columns
unique_counts <- sapply(df, function(x) length(unique(x)))
unique_counts
```

```
##          datetime     datetimeEpoch           tempmax           tempmin
##                15               262               268               264
##              temp     feelslikemax     feelslikemin        feelslike
##               270               270               264               271
##               dew          humidity            precip       precipprob
##               272               269               234               243
##       precipcover              snow         snowdepth         windgust
##               234                 1                 1               261
##         windspeed           winddir          pressure        cloudcover
##               252               273               271               271
##        visibility    solarradiation       solarenergy           uvindex
##               242               274               260               238
##        severerisk           sunrise     sunriseEpoch            sunset
##               234                45               274                44
##       sunsetEpoch         moonphase        conditions       description
##               274               250                 2                 4
##              icon            source              City        Temp_Range
##                 2                 2                 9               269
##        Heat_Index    Severity_Score             Month            Season
##               274               268                 1                 1
##       Day_of_Week        Is_Weekend Health_Risk_Score
##                 7                 2             27674
```

```
#Dropping the unwanted columns
df<- subset(df, select = -c(Season,Month,snow,snowdepth))
head(df)
```

```
##     datetime datetimeEpoch tempmax tempmin temp feelslikemax feelslikemin
## 1 07-09-2024      1.73e+09    89.0    62.1 73.3         88.6         62.1
## 2 08-09-2024      1.73e+09    89.0    60.0 72.4         87.9         60.0
## 3 10-09-2024      1.73e+09    79.4    59.6 67.8         79.4         59.6
## 4 11-09-2024      1.73e+09    77.3    57.6 66.3         77.3         57.6
## 5 12-09-2024      1.73e+09    79.2    57.8 67.4         79.2         57.8
## 6 13-09-2024      1.73e+09    83.2    58.9 69.6         82.2         58.9
##   feelslike  dew humidity precip precipprob precipcover windgust windspeed
```

```
## 1      73.3 59.8     66.3      0          0           0     16.1       9.2
## 2      72.3 57.6     62.5      0          0           0     13.9       8.1
## 3      67.8 57.2     70.7      0          0           0     17.4       9.8
## 4      66.3 56.8     73.1      0          4           0     23.0      13.4
## 5      67.4 55.6     68.3      0          5           0     17.9      10.7
## 6      69.5 54.2     60.5      0          0           0     16.1       8.9
##   winddir pressure cloudcover visibility solarradiation solarenergy uvindex
## 1     311     1012       12.0       10.0            268        23.4       9
## 2     310     1012       15.6        9.8            279        24.1       9
## 3     290     1012       18.8       12.4            275        23.8       9
## 4     274     1010       17.3       15.0            264        22.6       8
## 5     286     1007       14.2       15.0            262        22.6       8
## 6     288     1007        5.9       15.0            263        22.5       8
##   severerisk  sunrise sunriseEpoch   sunset sunsetEpoch moonphase conditions
## 1         10 06:43:31     1.73e+09 19:26:34    1.73e+09      0.16      Clear
## 2         10 06:44:20     1.73e+09 19:25:03    1.73e+09      0.19      Clear
## 3         10 06:45:59     1.73e+09 19:22:01    1.73e+09      0.25      Clear
## 4         10 06:46:48     1.73e+09 19:20:29    1.73e+09      0.29      Clear
## 5         10 06:47:38     1.73e+09 19:18:57    1.73e+09      0.32      Clear
## 6         10 06:48:27     1.73e+09 19:17:25    1.73e+09      0.36      Clear
##                           description      icon source     City Temp_Range
## 1 Clear conditions throughout the day. clear-day   comb San Jose       26.9
## 2 Clear conditions throughout the day. clear-day   fcst San Jose       29.0
## 3 Clear conditions throughout the day. clear-day   fcst San Jose       19.8
## 4 Clear conditions throughout the day. clear-day   fcst San Jose       19.7
## 5 Clear conditions throughout the day. clear-day   fcst San Jose       21.4
## 6 Clear conditions throughout the day. clear-day   fcst San Jose       24.3
##   Heat_Index Severity_Score Day_of_Week Is_Weekend Health_Risk_Score
## 1       75.8           3.41    Saturday       True              9.85
## 2       75.9           3.19      Sunday       True              9.59
## 3       73.5           3.54     Tuesday      False              9.85
## 4       72.9           3.90   Wednesday      False             10.14
## 5       74.3           3.39    Thursday      False              9.75
## 6       75.8           3.21      Friday      False              9.52
```

```
dim(df)
```

```
## [1] 27674    39
```

Duplicates and missing values removal

```
#Checking for duplicate rows
duplicates <- df %>% duplicated() %>% sum()
print(paste("Number of duplicate rows:", duplicates))
```

```
## [1] "Number of duplicate rows: 0"
```

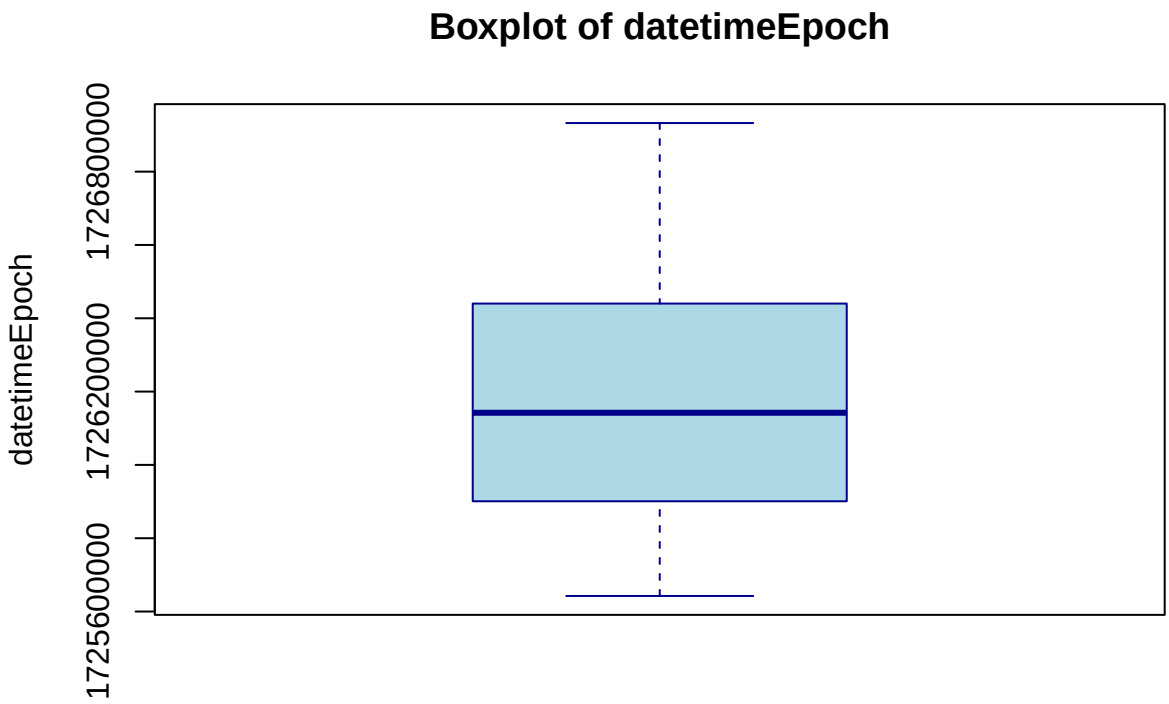
```
#Checking for missing values
missing_values <- colSums(is.na(df))
print(missing_values)
```

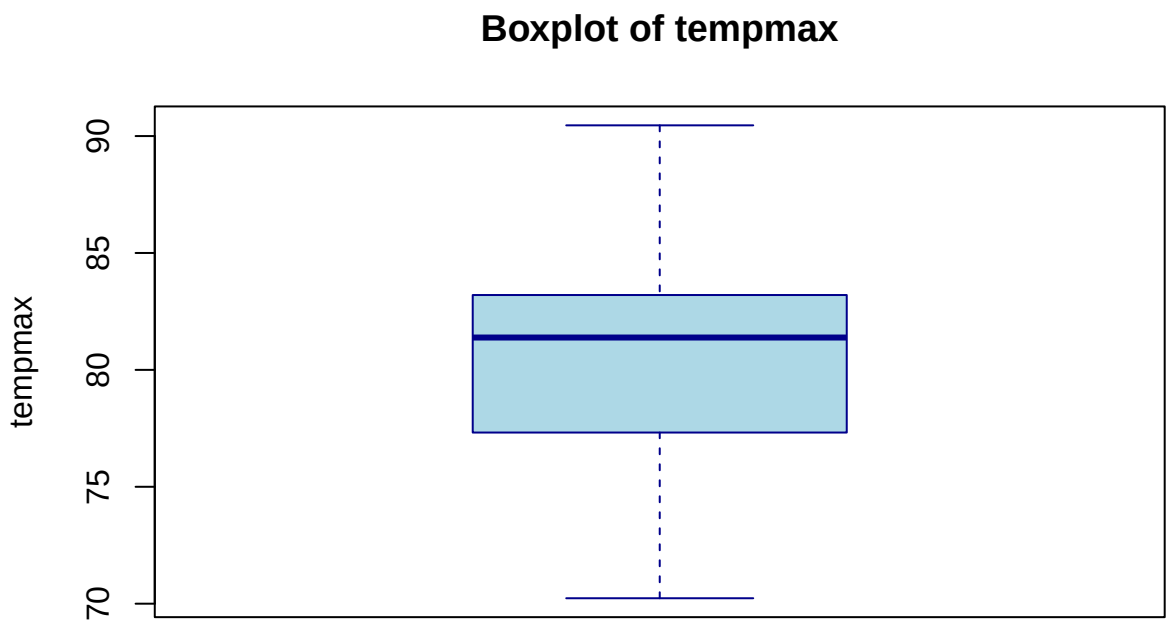
```
##          datetime     datetimeEpoch           tempmax           tempmin
##                 0                 0                 0                 0
##              temp      feelslikemax      feelslikemin         feelslike
##                 0                 0                 0                 0
##               dew          humidity            precip       precipprob
```

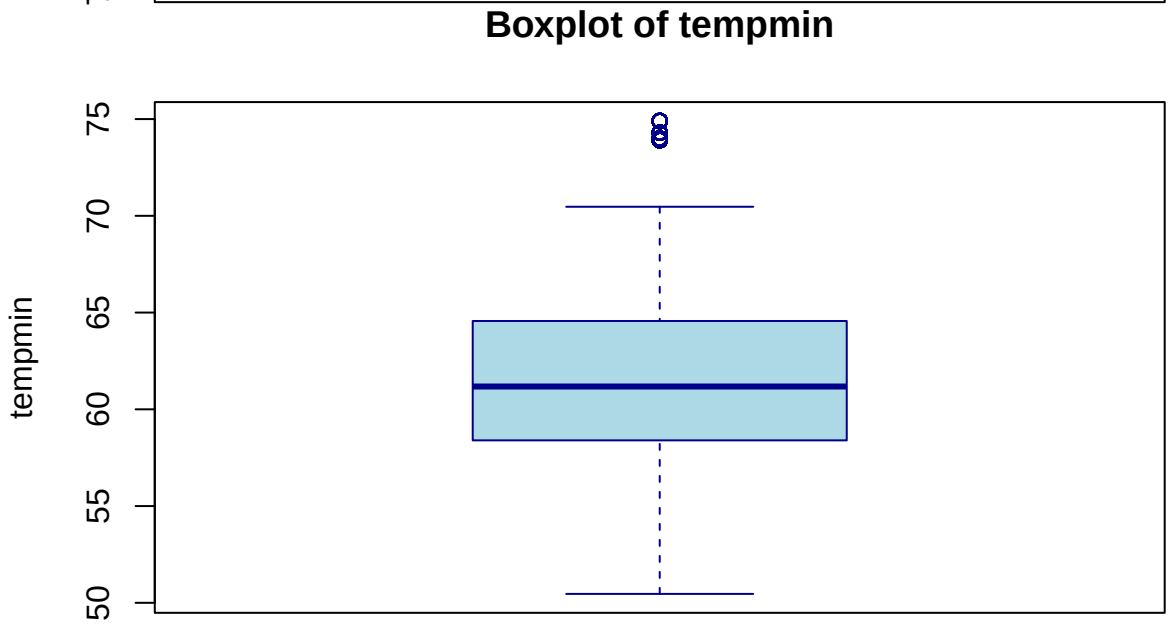
```
##                0                0                0                0
##       precipcover        windgust        windspeed          winddir
##                0                0                0                0
##          pressure       cloudcover       visibility    solarradiation
##                0                0                0                0
##       solarenergy          uvindex       severerisk          sunrise
##                0                0                0                0
##      sunriseEpoch           sunset      sunsetEpoch        moonphase
##                0                0                0                0
##        conditions      description             icon           source
##                0                0                0                0
##              City       Temp_Range       Heat_Index   Severity_Score
##                0                0                0                0
##       Day_of_Week       Is_Weekend Health_Risk_Score
##                0                0                0
```

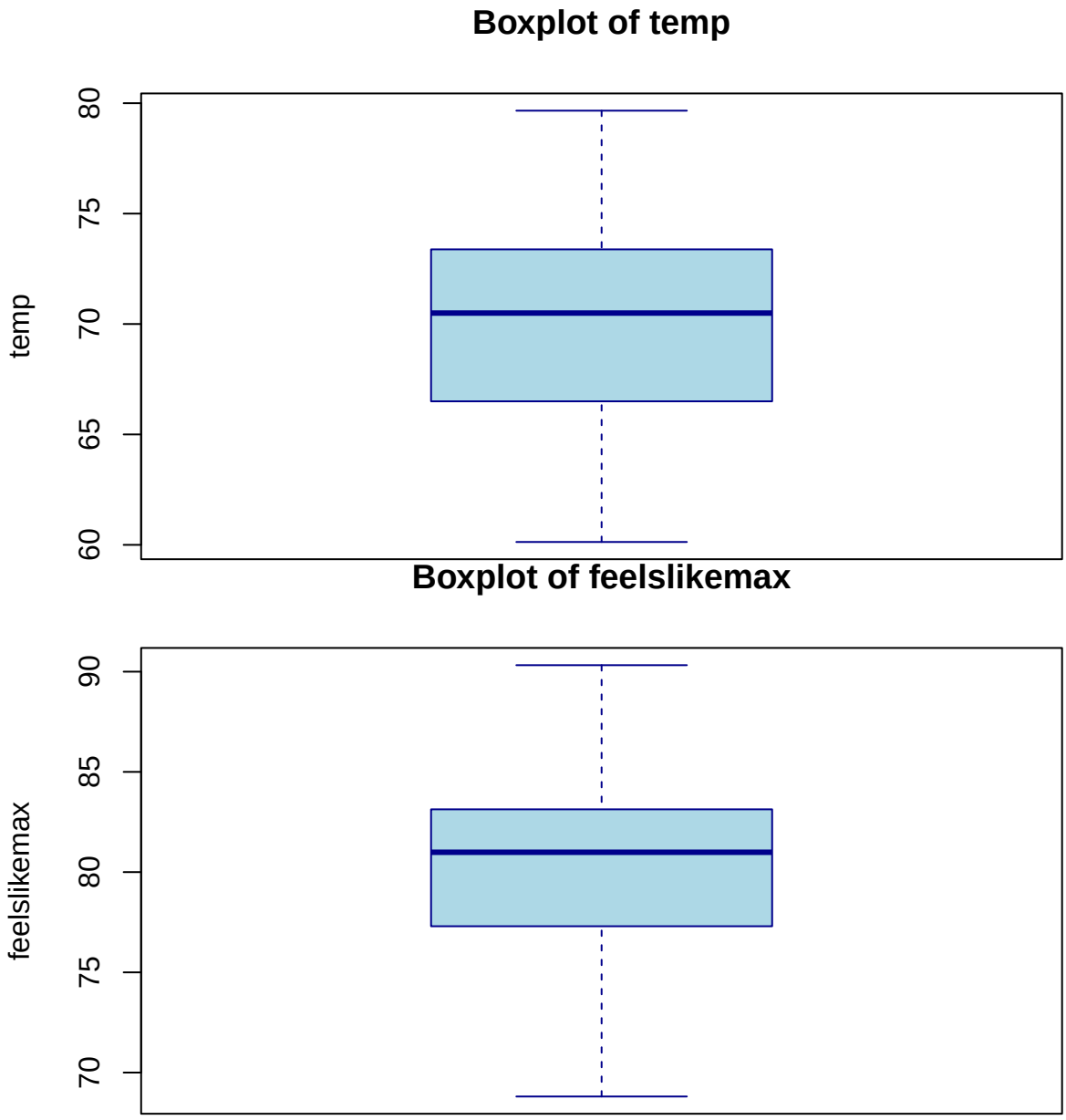
Outliers Removal

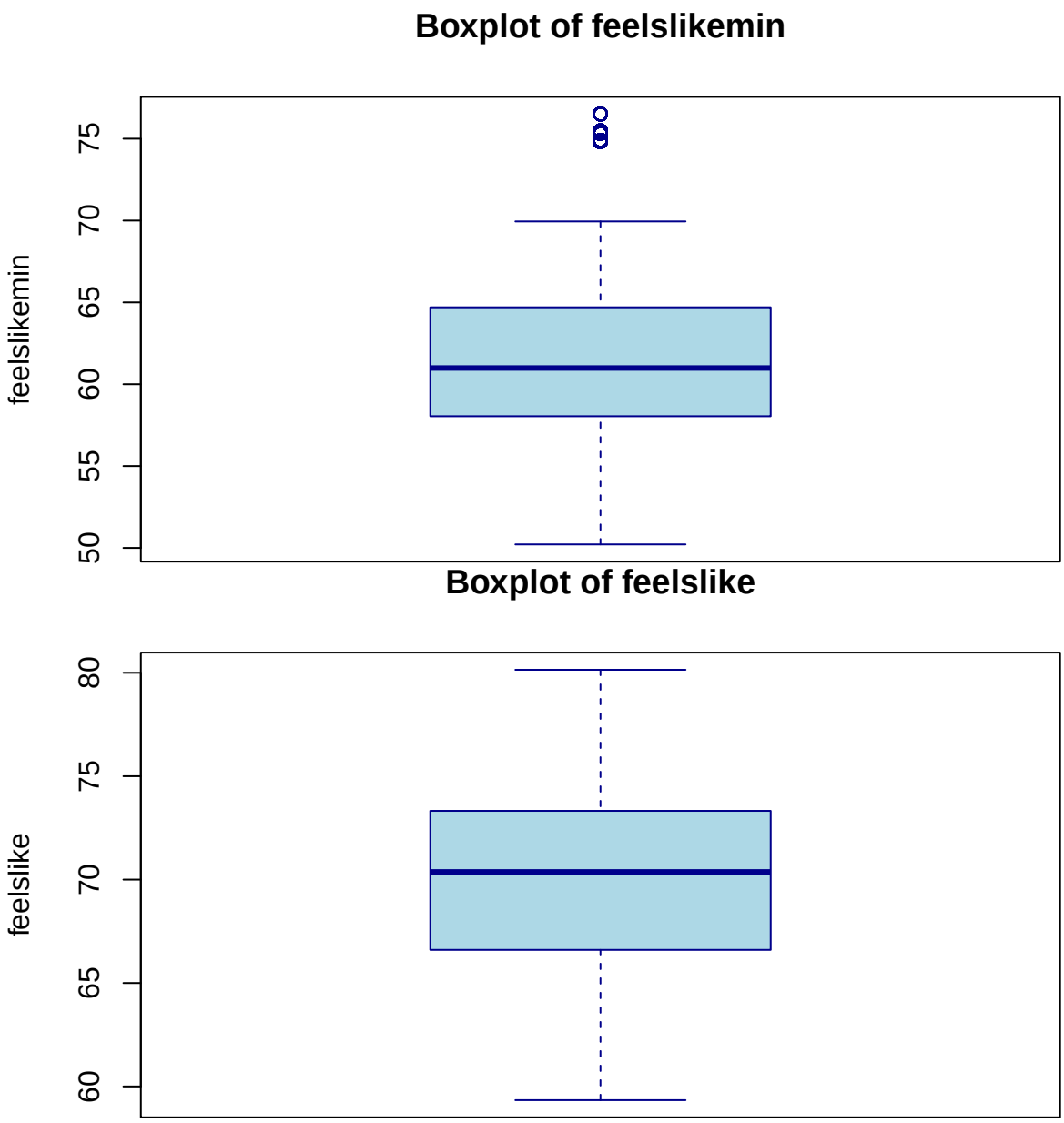
```r
#Visualizing the outliers
numeric_cols <- df %>% select_if(is.numeric)
num_cols <- ncol(numeric_cols)

# Loop through numeric columns and create box plots
for (col in colnames(numeric_cols)) {
  boxplot(numeric_cols[[col]],
          main = paste("Boxplot of", col),
          ylab = col,
          col = "lightblue",
          border="darkblue")
}
```

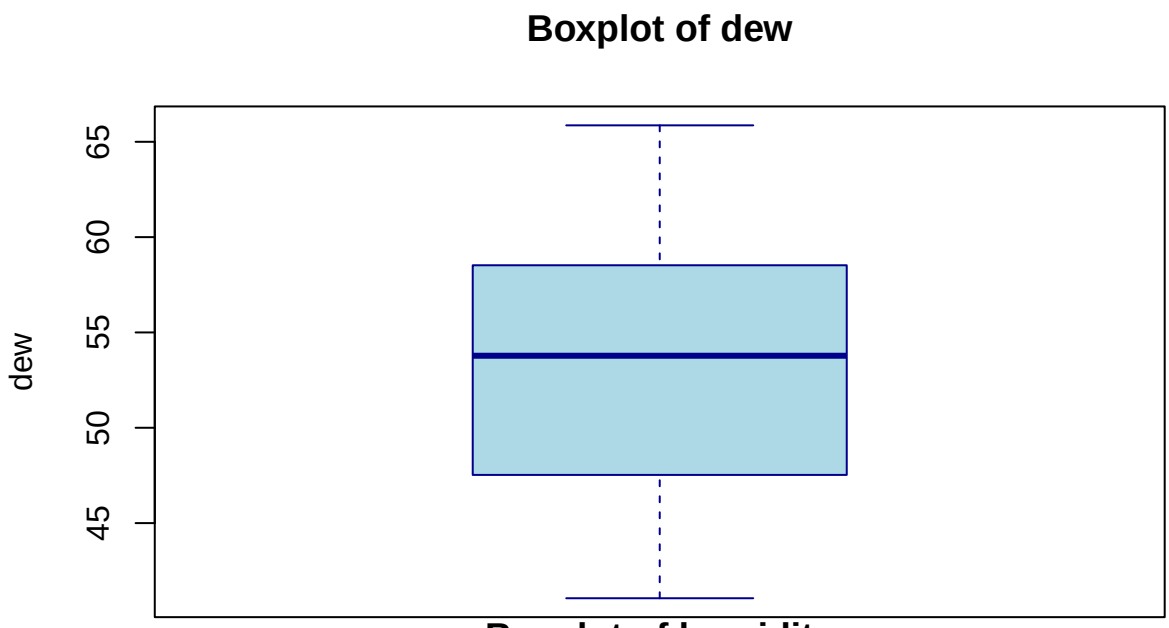


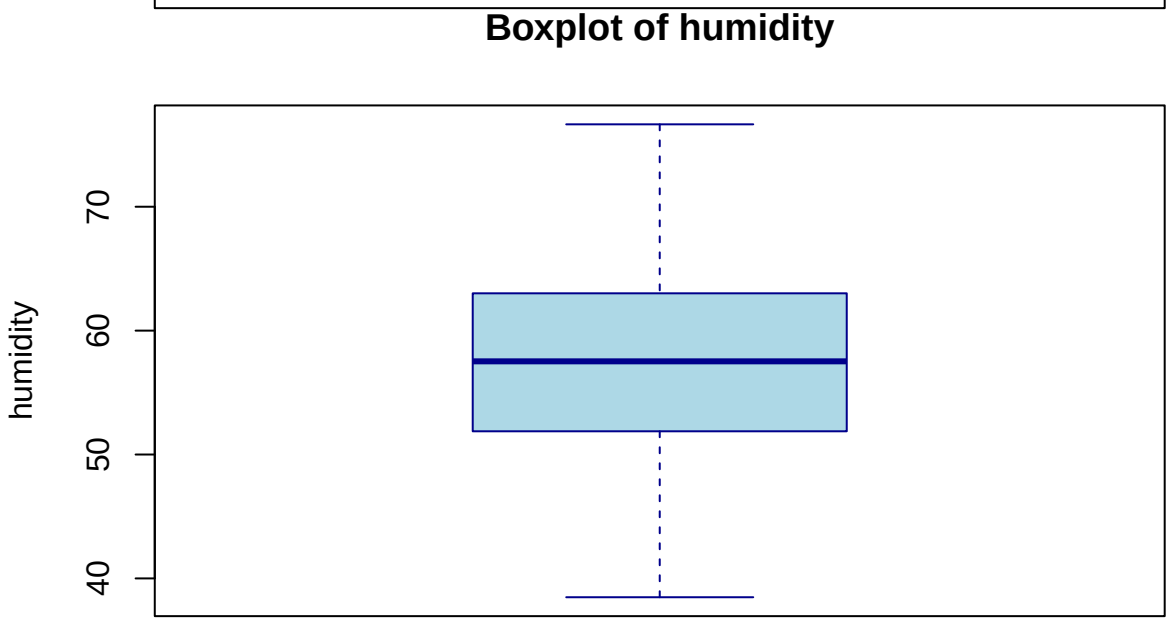

Boxplot of tempmax
tempmax
70
75
80
85
90

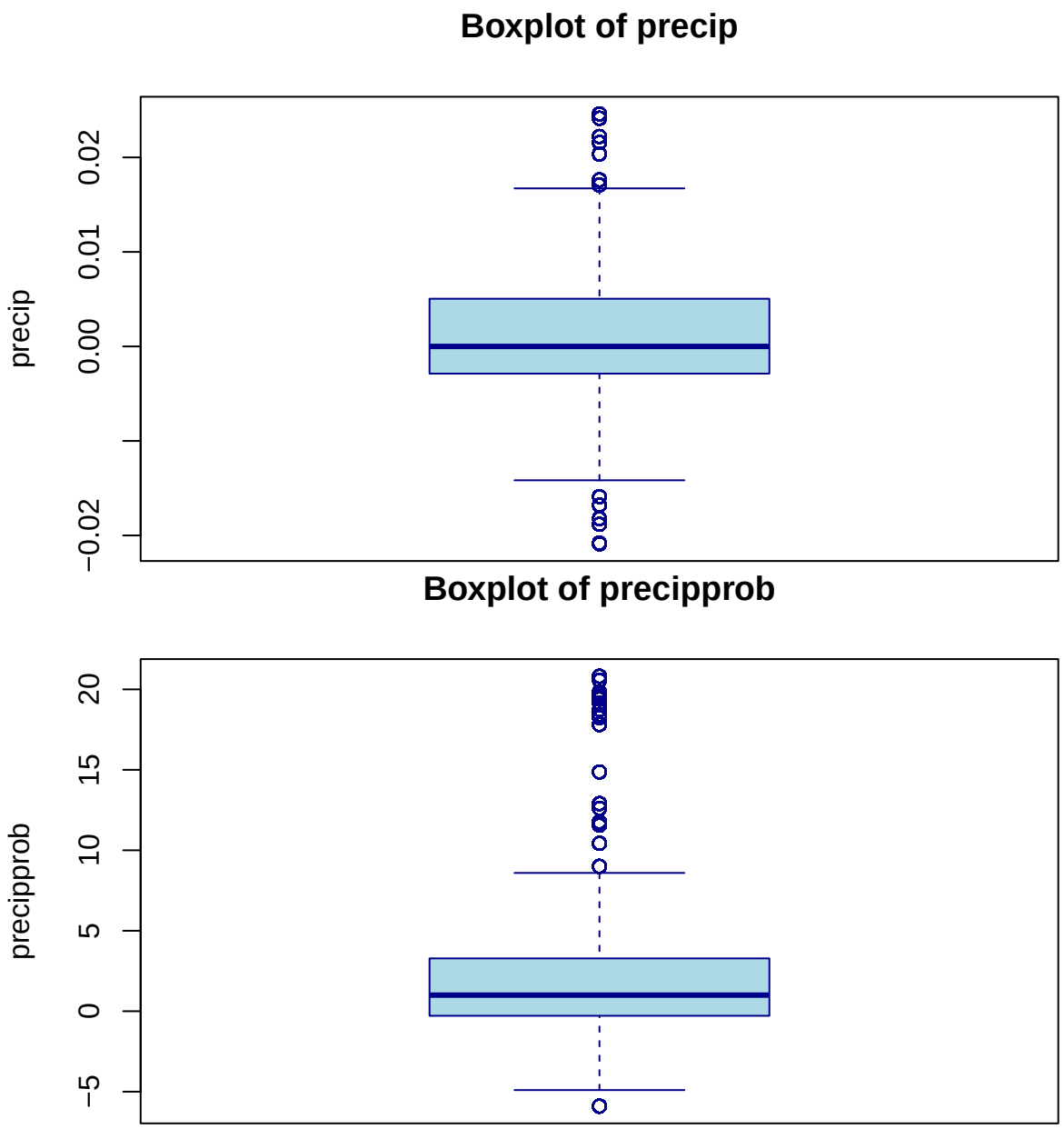

Boxplot of tempmin
tempmin
50
55
60
65
70
75

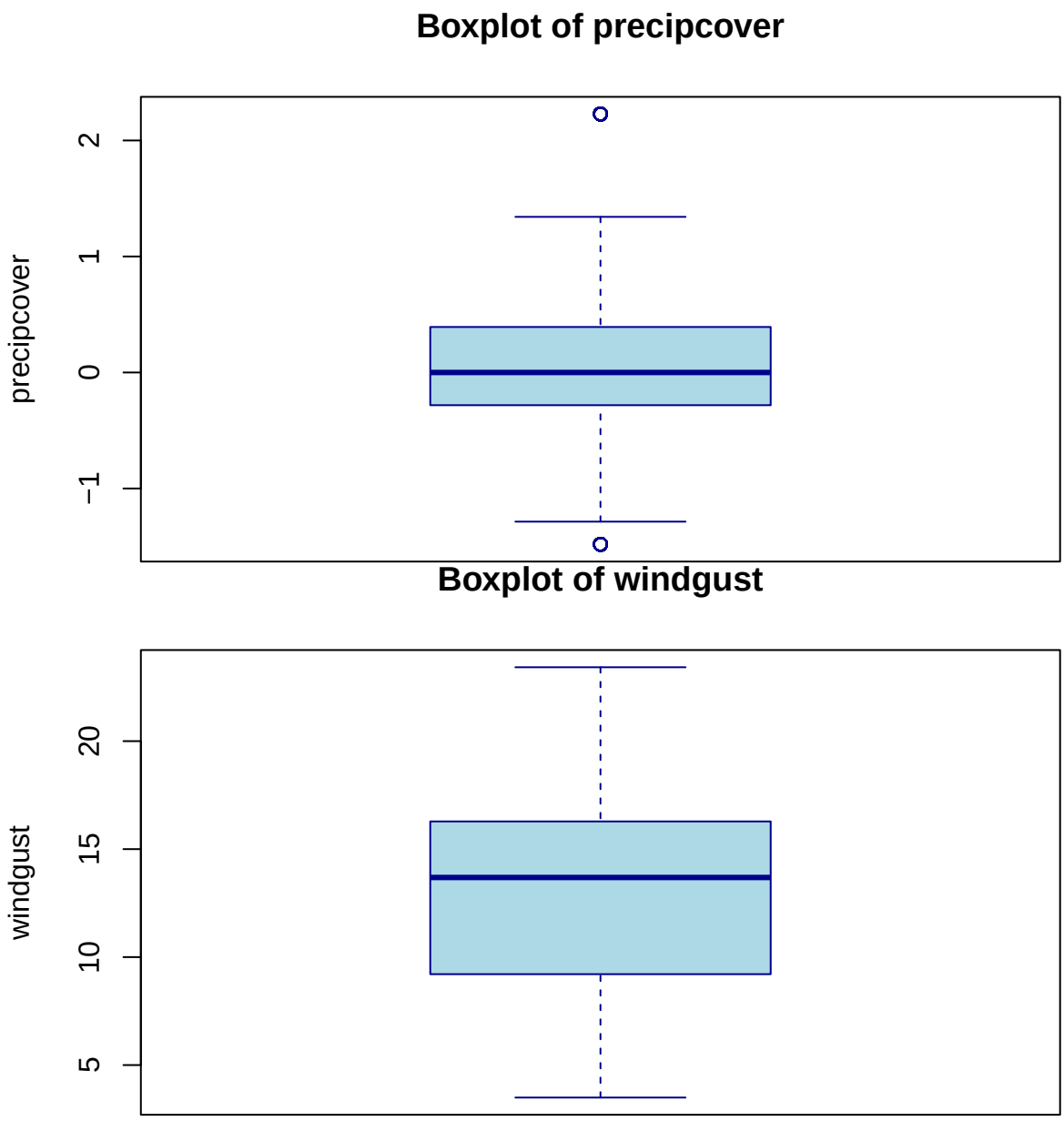

Boxplot of temp
temp
60
65
70
75
80
Boxplot of feelslikemax
feelslikemax
70
75
80
85
90

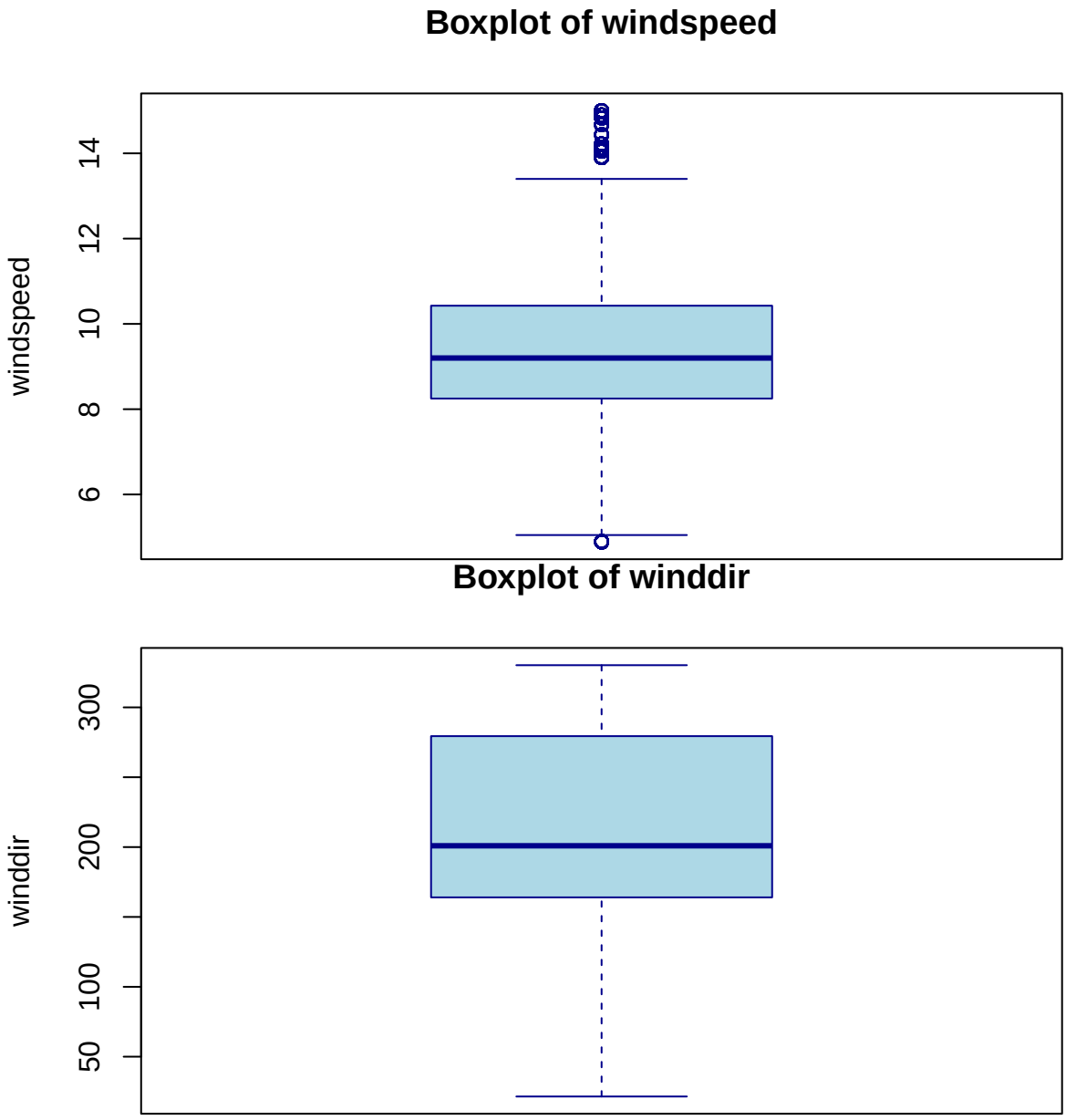

Boxplot of feelslikemin
feelslikemin
50
55
60
65
70
75
Boxplot of feelslike
feelslike
60
65
70
75
80

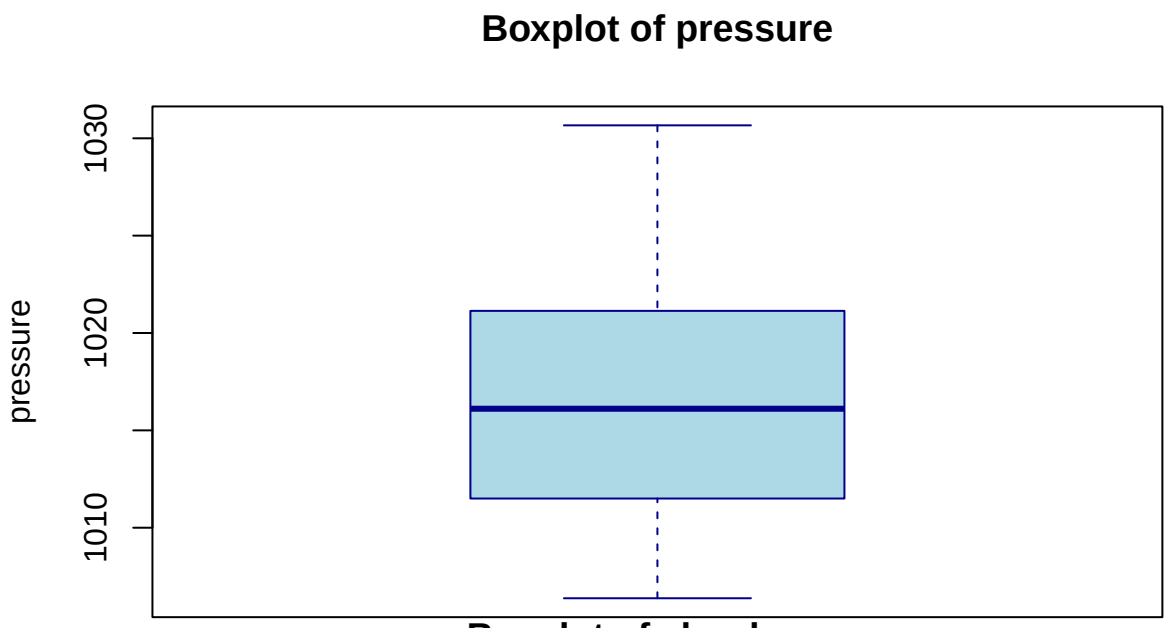

Boxplot of dew
dew
45
50
55
60
65

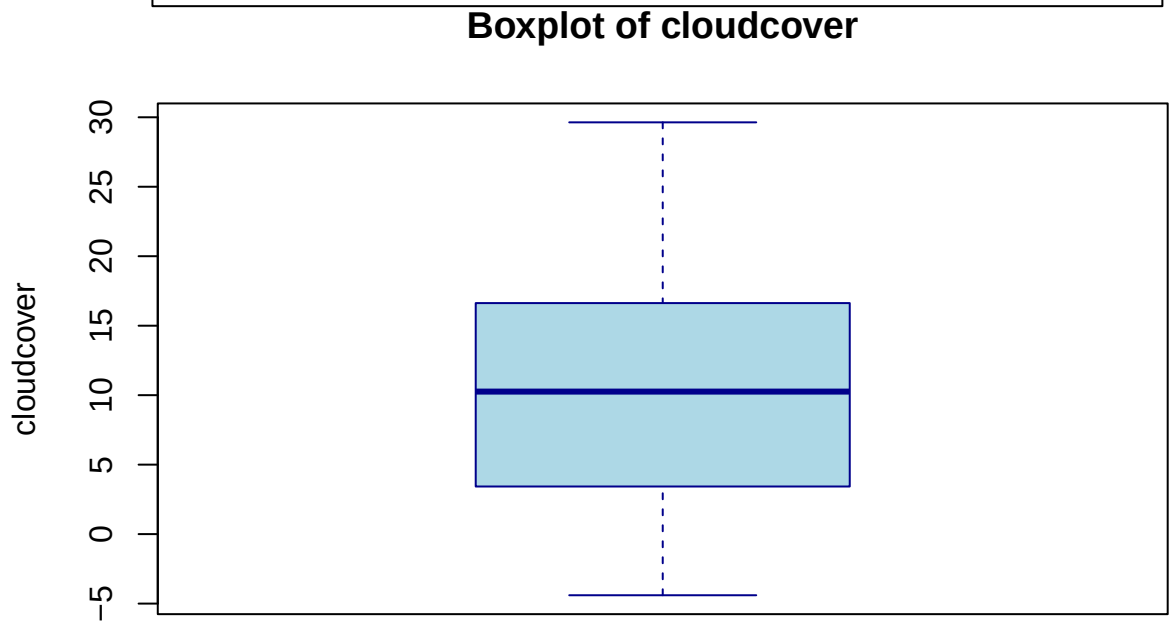

Boxplot of humidity
humidity
40
50
60
70

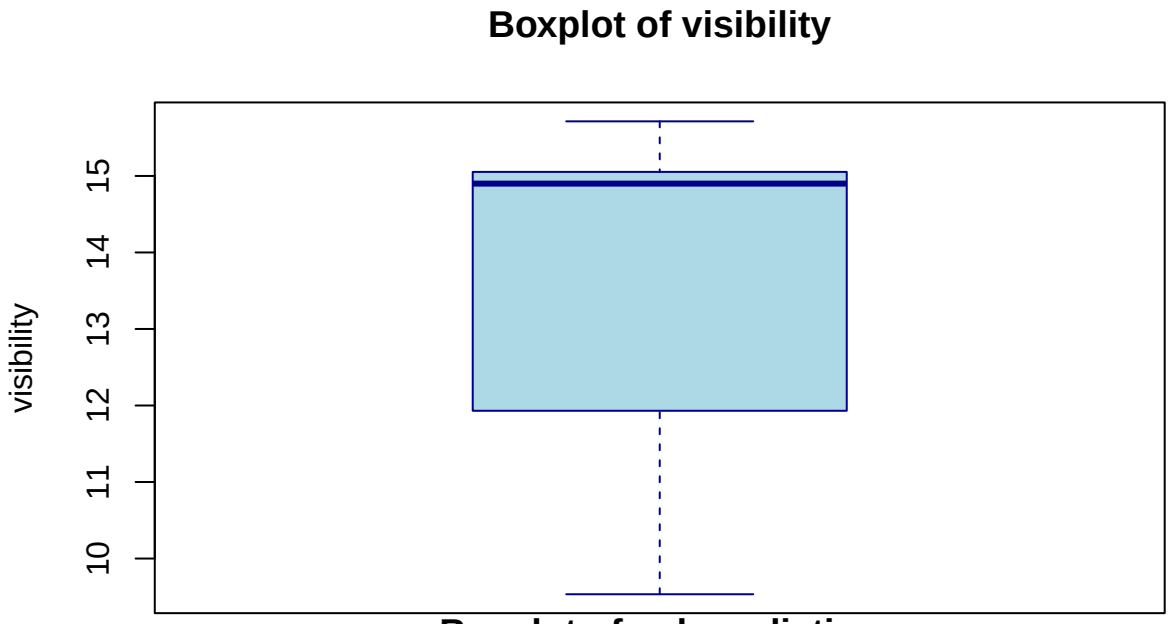

Boxplot of precip
precip
0.02
0.01
0.00
−0.02
Boxplot of precipprob
precipprob
20
15
10
5
0
−5

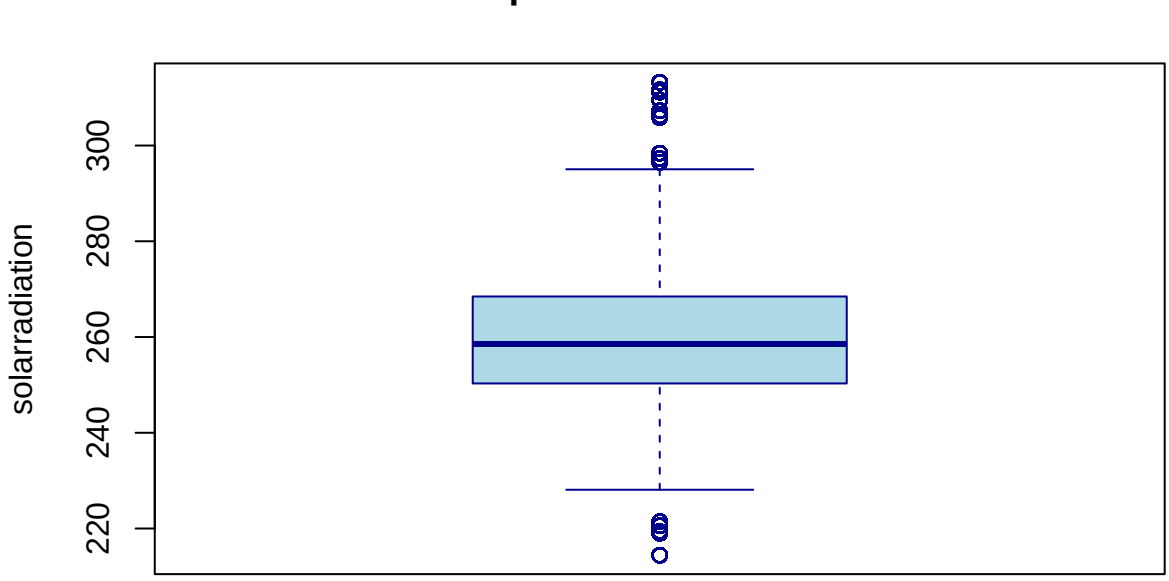

Boxplot of precipcover
precipcover
2
1
0
−1
Boxplot of windgust
windgust
20
15
10
5

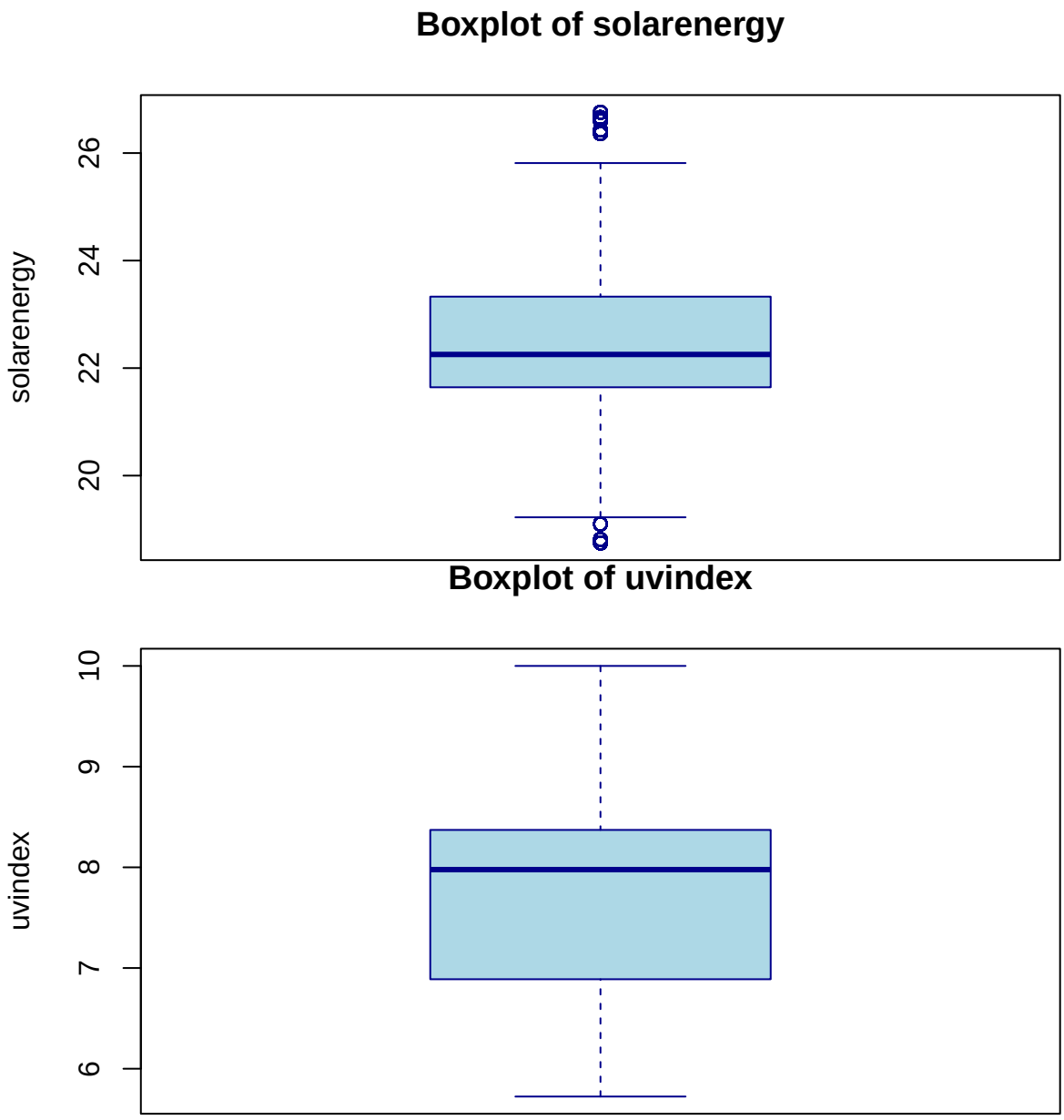

Boxplot of windspeed
windspeed
6
8
10
12
14
Boxplot of winddir
winddir
50
100
200
300

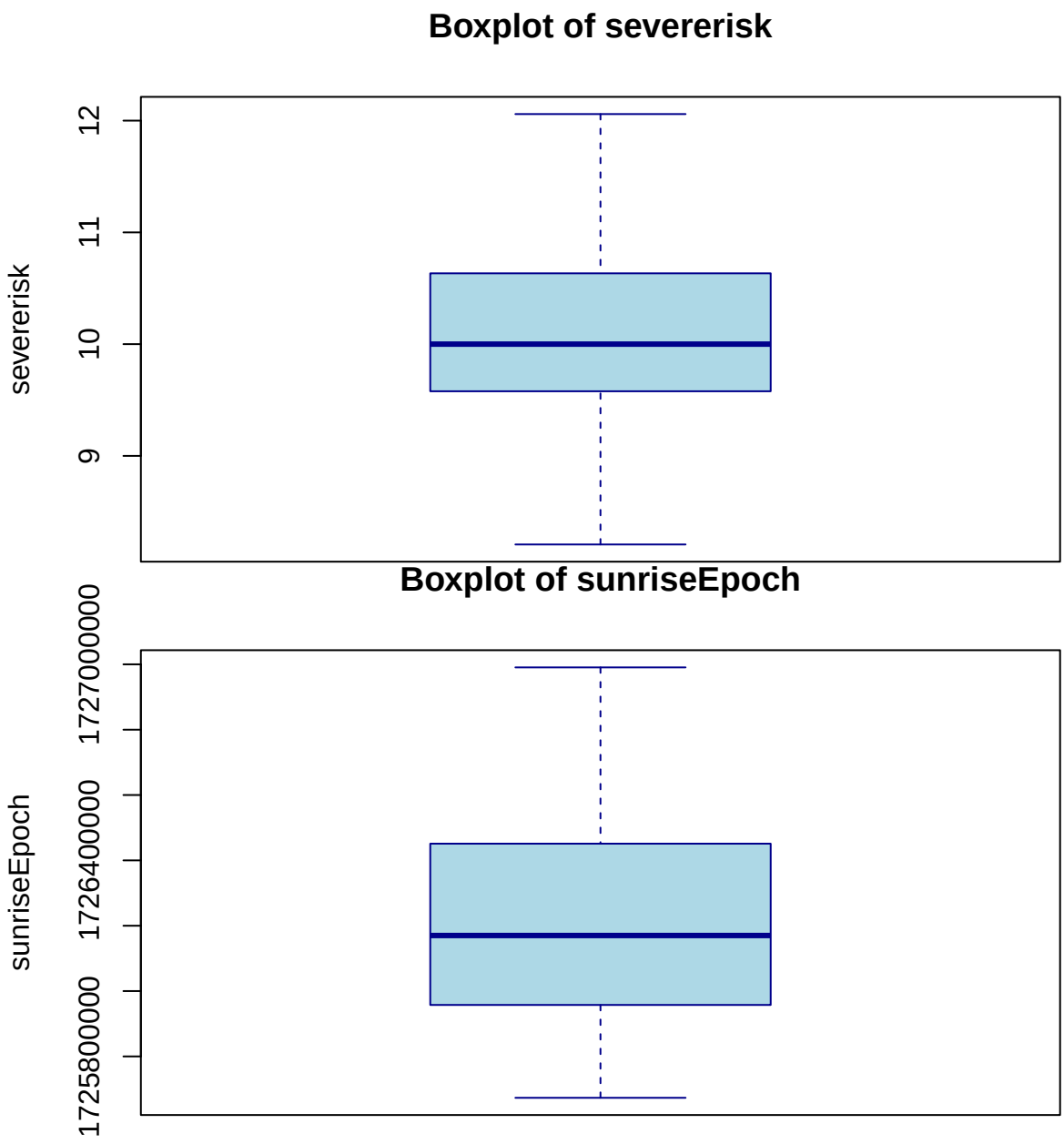

Boxplot of pressure
pressure
1030
1020
1010

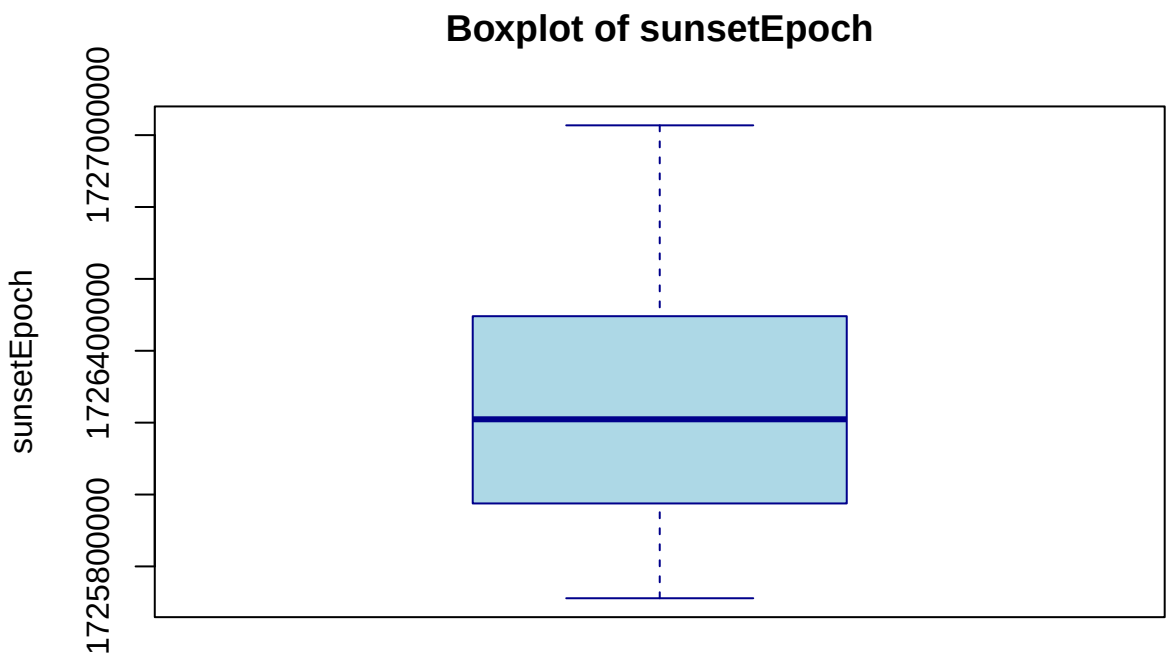

Boxplot of cloudcover
cloudcover
30
25
20
15
10
5
0
−5

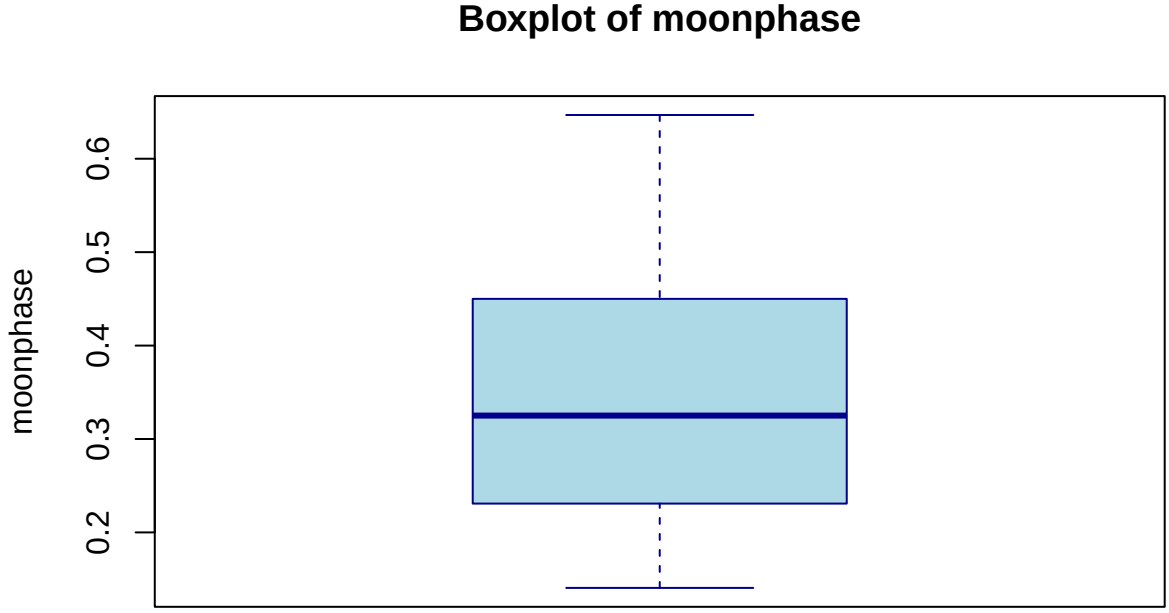

Boxplot of visibility
visibility
10
11
12
13
14
15

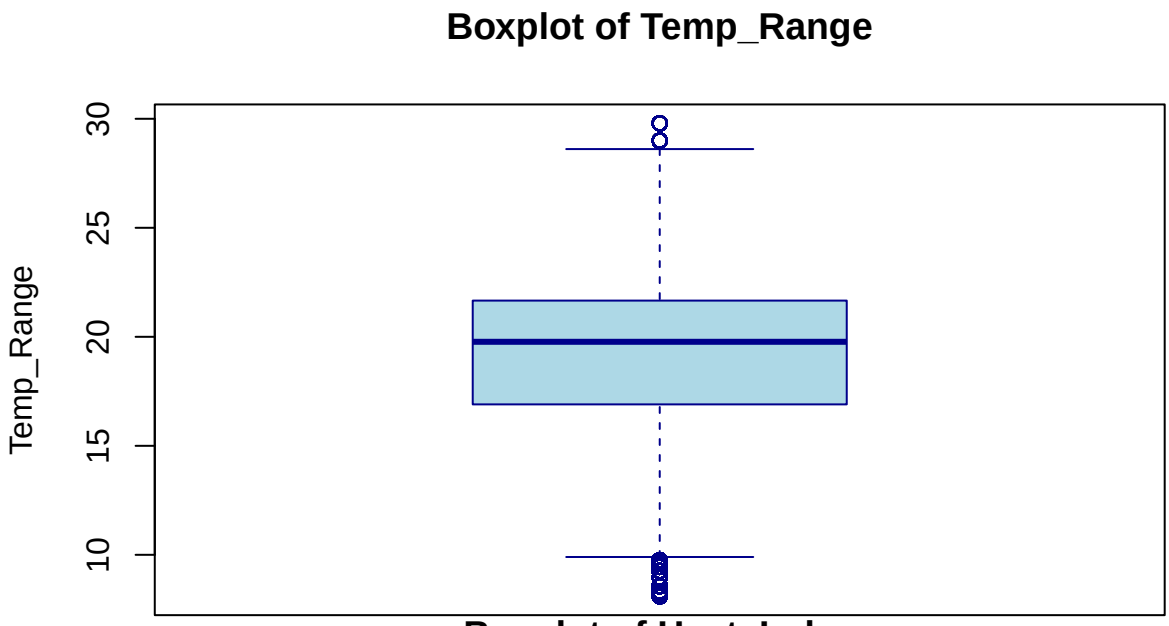

Boxplot of solarradiation
solarradiation
220
240
260
280
300

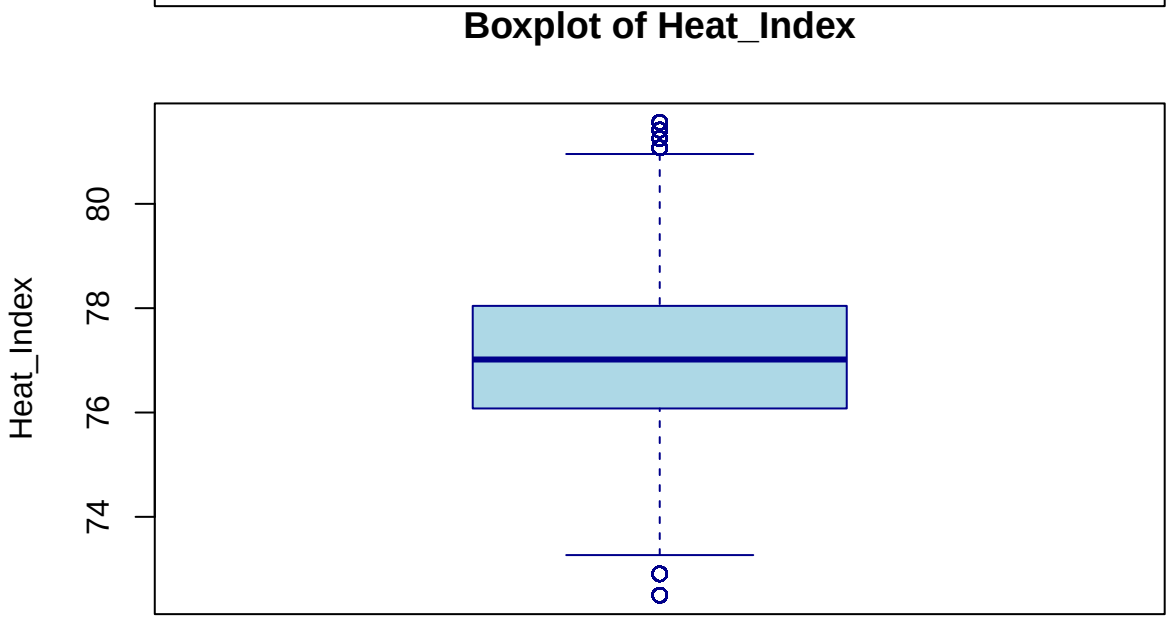

Boxplot of solarenergy
solarenergy
20
22
24
26
Boxplot of uvindex
uvindex
6
7
8
9
10

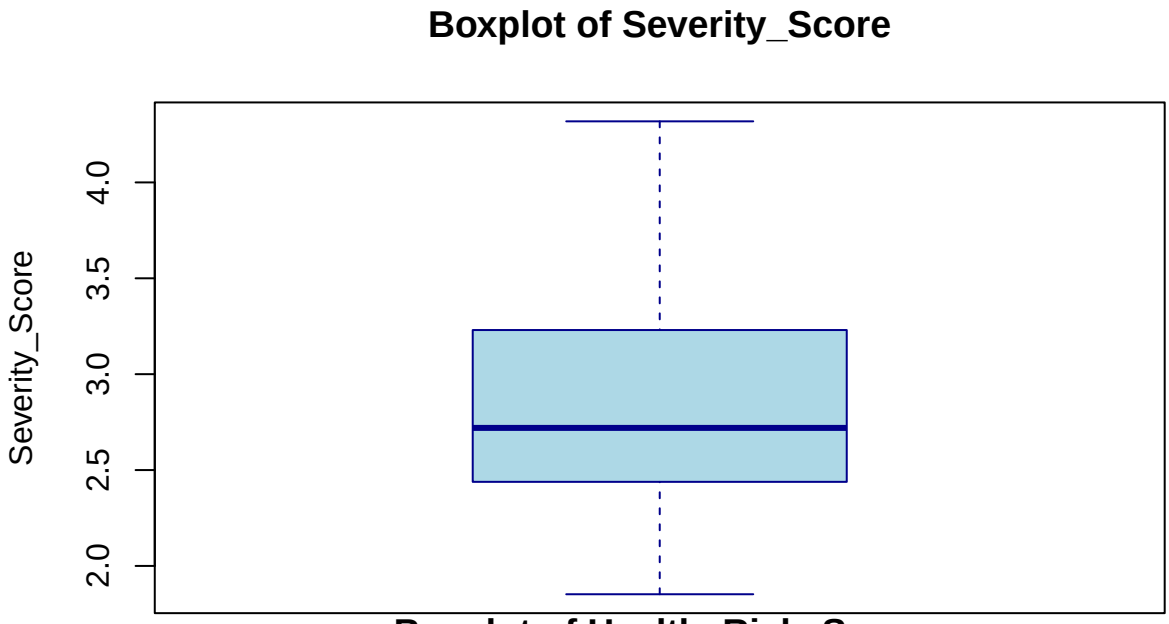

Boxplot of severerisk
severerisk
12
11
10
9
Boxplot of sunriseEpoch
sunriseEpoch
1727000000
1726400000
1725800000

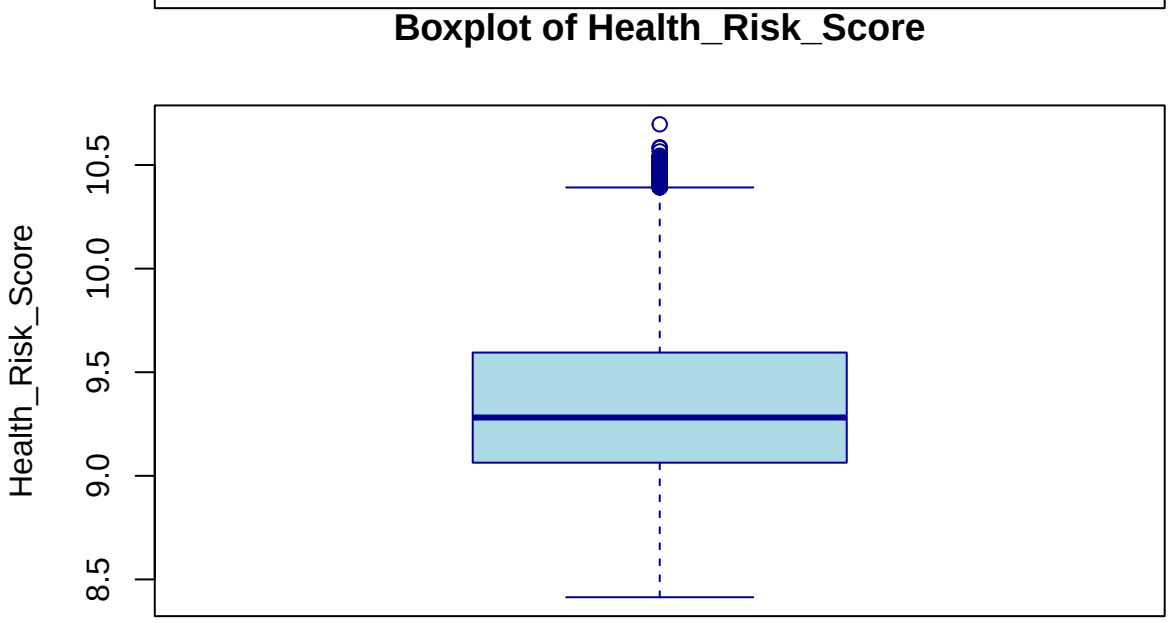

Boxplot of sunsetEpoch
sunsetEpoch
1725800000
1726400000
1727000000


Boxplot of moonphase
moonphase
0.2
0.3
0.4
0.5
0.6


Boxplot of Temp_Range
Temp_Range
10
15
20
25
30


Boxplot of Heat_Index
Heat_Index
74
76
78
80






```r
#Defining function to remove outliers using IQR approach
remove_outliers <- function(data) {
  for (col in names(data)) {
    if (is.numeric(data[[col]])) {
      Q1 <- quantile(data[[col]], 0.25)
      Q3 <- quantile(data[[col]], 0.75)
      IQR <- Q3 - Q1
      lower_bound <- Q1 - 1.5 * IQR
      upper_bound <- Q3 + 1.5 * IQR

      data <- data[data[[col]] >= lower_bound & data[[col]] <= upper_bound, ]
    }
  }
  return(data)
}
```

```r
#remove outliers by applying the function
df <- remove_outliers(df)
head(df)
```

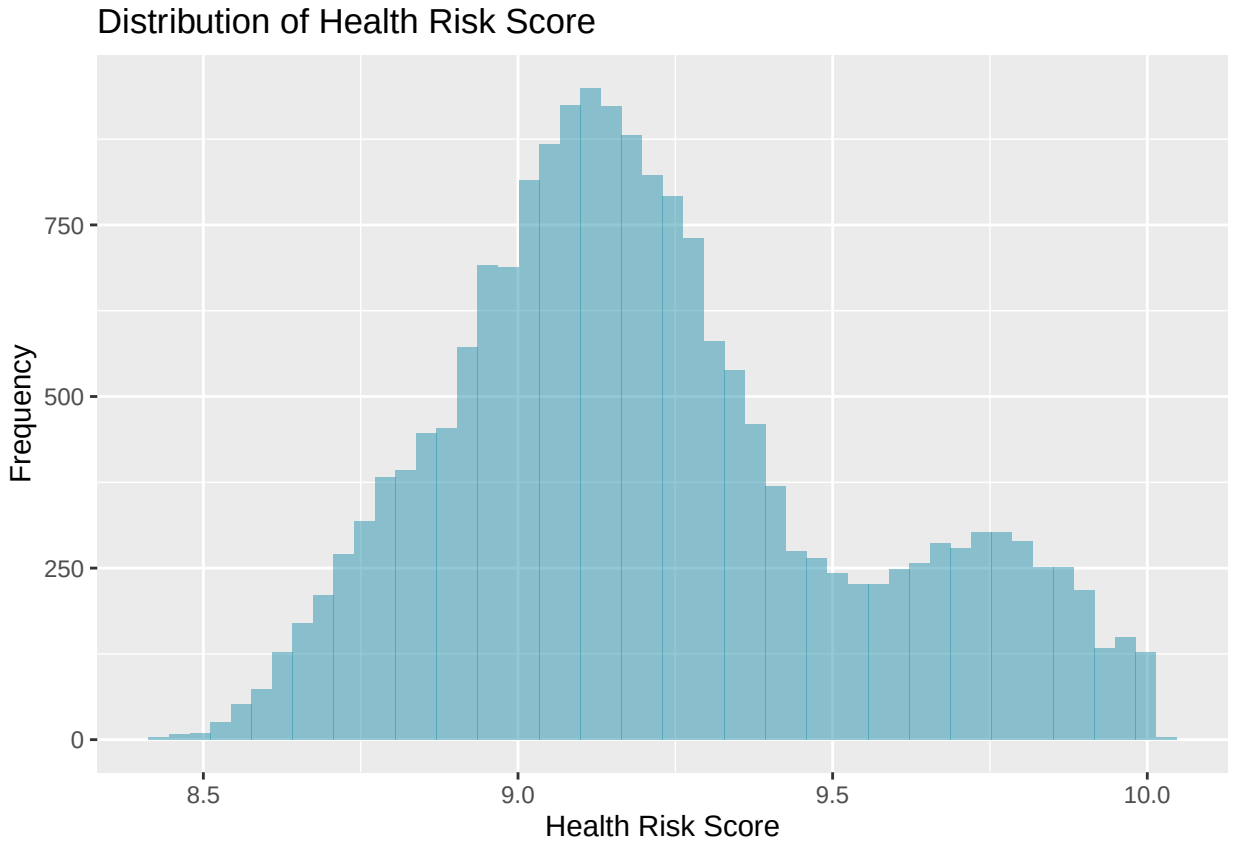
```
##     datetime datetimeEpoch tempmax tempmin temp feelslikemax feelslikemin
## 1 07-09-2024      1.73e+09    89.0    62.1 73.3         88.6         62.1
## 3 10-09-2024      1.73e+09    79.4    59.6 67.8         79.4         59.6
## 5 12-09-2024      1.73e+09    79.2    57.8 67.4         79.2         57.8
## 6 13-09-2024      1.73e+09    83.2    58.9 69.6         82.2         58.9
## 7 14-09-2024      1.73e+09    81.4    59.4 68.8         81.3         59.4
## 8 15-09-2024      1.73e+09    78.3    59.8 66.8         78.3         59.8
##   feelslike  dew humidity precip precipprob precipcover windgust windspeed
## 1      73.3 59.8     66.3      0        0.0           0     16.1       9.2
## 3      67.8 57.2     70.7      0        0.0           0     17.4       9.8
## 5      67.4 55.6     68.3      0        5.0           0     17.9      10.7
## 6      69.5 54.2     60.5      0        0.0           0     16.1       8.9
## 7      68.8 55.5     64.2      0        1.0           0     16.6       9.8
## 8      66.8 47.3     52.9      0        3.2           0      9.8       8.9
##   winddir pressure cloudcover visibility solarradiation solarenergy uvindex
## 1     311     1012       12.0       10.0            268        23.4       9
## 3     290     1012       18.8       12.4            275        23.8       9
## 5     286     1007       14.2       15.0            262        22.6       8
## 6     288     1007        5.9       15.0            263        22.5       8
## 7     263     1010        8.9       14.9            259        22.3       8
## 8     256     1012        3.1       14.9            255        22.0       8
##   severerisk  sunrise sunriseEpoch   sunset sunsetEpoch moonphase conditions
## 1         10 06:43:31     1.73e+09 19:26:34    1.73e+09      0.16      Clear
## 3         10 06:45:59     1.73e+09 19:22:01    1.73e+09      0.25      Clear
## 5         10 06:47:38     1.73e+09 19:18:57    1.73e+09      0.32      Clear
## 6         10 06:48:27     1.73e+09 19:17:25    1.73e+09      0.36      Clear
## 7         10 06:49:17     1.73e+09 19:15:53    1.73e+09      0.39      Clear
## 8         10 06:50:06     1.73e+09 19:14:21    1.73e+09      0.42      Clear
##                           description      icon source     City Temp_Range
## 1 Clear conditions throughout the day. clear-day   comb San Jose       26.9
## 3 Clear conditions throughout the day. clear-day   fcst San Jose       19.8
## 5 Clear conditions throughout the day. clear-day   fcst San Jose       21.4
## 6 Clear conditions throughout the day. clear-day   fcst San Jose       24.3
## 7 Clear conditions throughout the day. clear-day   fcst San Jose       22.0
## 8 Clear conditions throughout the day. clear-day   fcst San Jose       18.5
##   Heat_Index Severity_Score Day_of_Week Is_Weekend Health_Risk_Score
## 1       75.8           3.41    Saturday       True              9.85
## 3       73.5           3.54     Tuesday      False              9.85
## 5       74.3           3.39    Thursday      False              9.75
## 6       75.8           3.21      Friday      False              9.52
## 7       75.2           3.26    Saturday       True              9.61
## 8       77.6           2.58      Sunday       True              9.12
```

```r
#Checking shape of the data after removing outliers
paste("Row Count:",dim(df)[1],"Column Count:",dim(df)[2])
```

```
## [1] "Row Count: 18885 Column Count: 39"
```

Univariate Analysis

```
# Distribution of Health Risk Score
ggplot(df, aes(x = Health_Risk_Score)) +
  geom_histogram(bins = 50, fill = "#2993ae", alpha = 0.5) +
  labs(title = "Distribution of Health Risk Score", x = "Health Risk Score", y = "Frequency")
```




```
paste("The majority of the population has health risk scores around 9.0, with a small subset showing el
```

```
## [1] "The majority of the population has health risk scores around 9.0, with a small subset showing el
```

Bivariate and Multivariate Analysis

```
paste("Is there a statistically significant difference in the Health Risk Score on weekends compared to
```

```
## [1] "Is there a statistically significant difference in the Health Risk Score on weekends compared to
```

```
# Separate Health Risk Scores by weekends and weekdays
weekend_scores <- df$Health_Risk_Score[df$Is_Weekend == "True"]
weekday_scores <- df$Health_Risk_Score[df$Is_Weekend == "False"]

# Display hypothesis statements
cat("Hypothesis Statements:\n")
```

```
## Hypothesis Statements:
```

```
cat("Null Hypothesis (H0): There is no significant difference in the Health Risk Score between weekends
```

```
## Null Hypothesis (H0): There is no significant difference in the Health Risk Score between weekends ar
```

```r
cat("Alternative Hypothesis (H1): There is a significant difference in the Health Risk Score between wee
```

```
## Alternative Hypothesis (H1): There is a significant difference in the Health Risk Score between weeke
```

```r
t_test_result <- t.test(weekend_scores, weekday_scores, alternative = "two.sided")

# Display the t-test results
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  weekend_scores and weekday_scores
## t = 30, df = 5925, p-value <2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.175 0.200
## sample estimates:
## mean of x mean of y
##      9.36      9.17
```

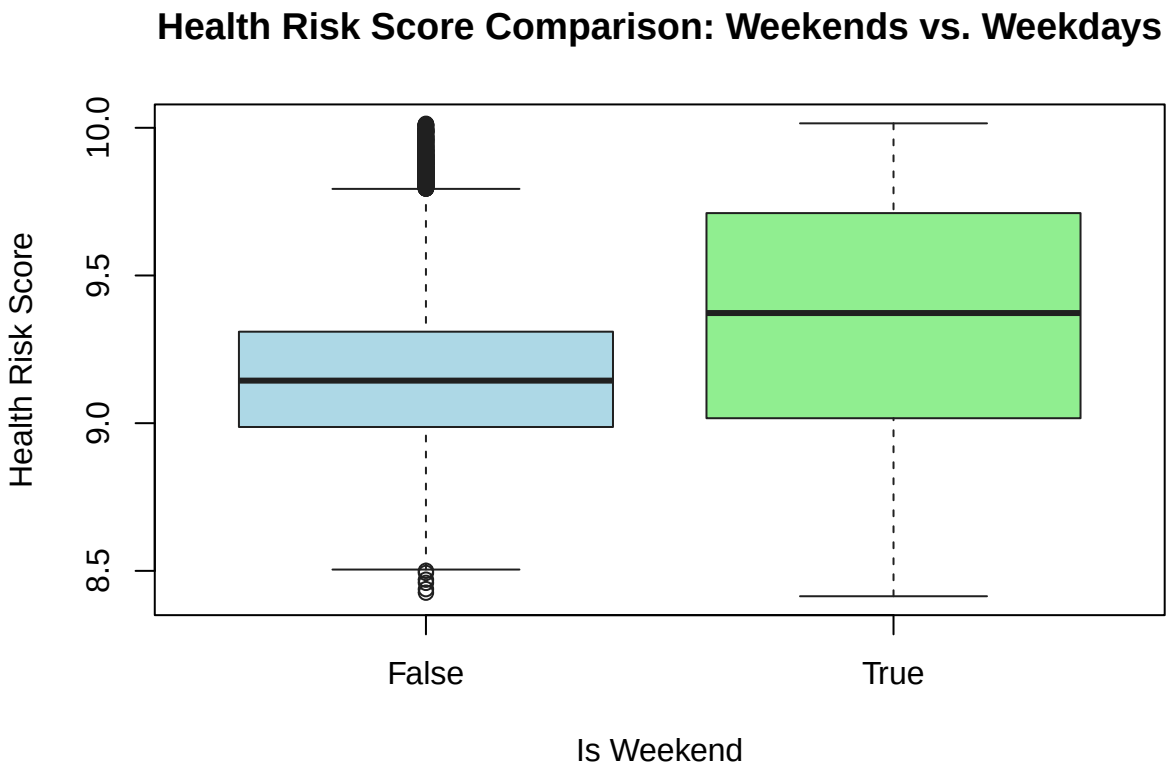
```r
# Interpretation
cat("\nInterpretation:\n")
```

```
##
## Interpretation:
```

```r
if(t_test_result$p.value < 0.05) {
  cat("Reject the null hypothesis: There is a statistically significant difference in the Health Risk Sc
} else {
  cat("Fail to reject the null hypothesis: There is no statistically significant difference in the Healt
}
```

```
## Reject the null hypothesis: There is a statistically significant difference in the Health Risk Score
```

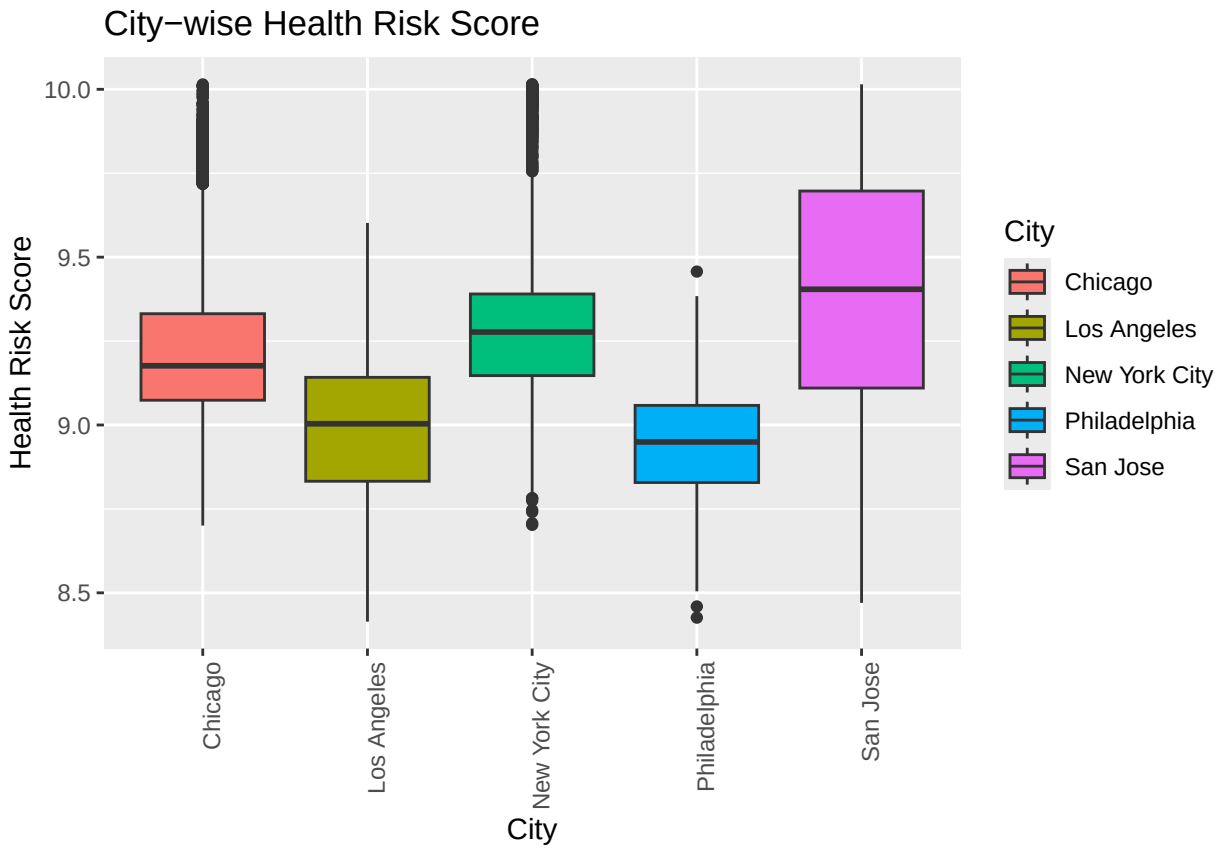
```r
boxplot(Health_Risk_Score ~ Is_Weekend, data = df,
        main = "Health Risk Score Comparison: Weekends vs. Weekdays",
        xlab = "Is Weekend",
        ylab = "Health Risk Score",
        col = c("lightblue", "lightgreen"),
        border = "#202020")
```



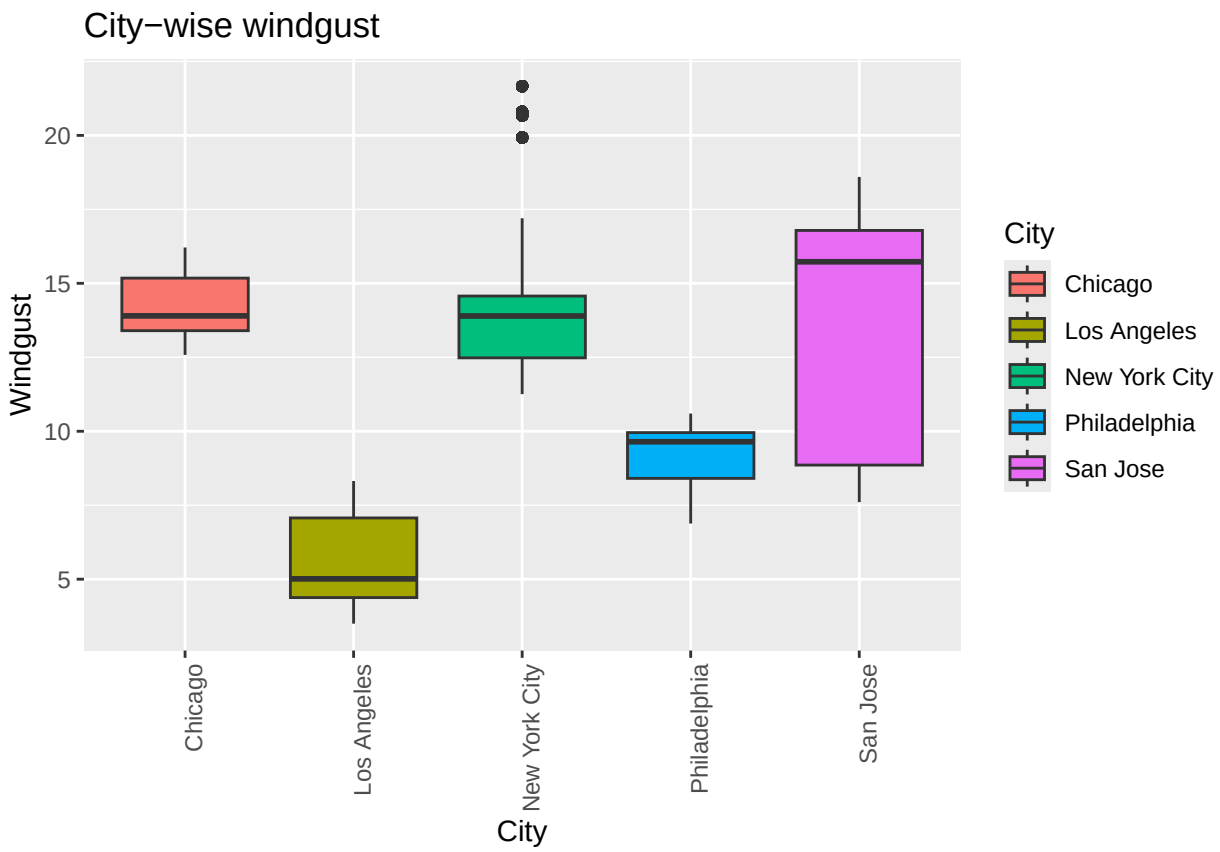

```
paste("The weak negative correlation of -0.24 suggests that higher temperatures are slightly associated
```

```
## [1] "The weak negative correlation of -0.24 suggests that higher temperatures are slightly associated
```

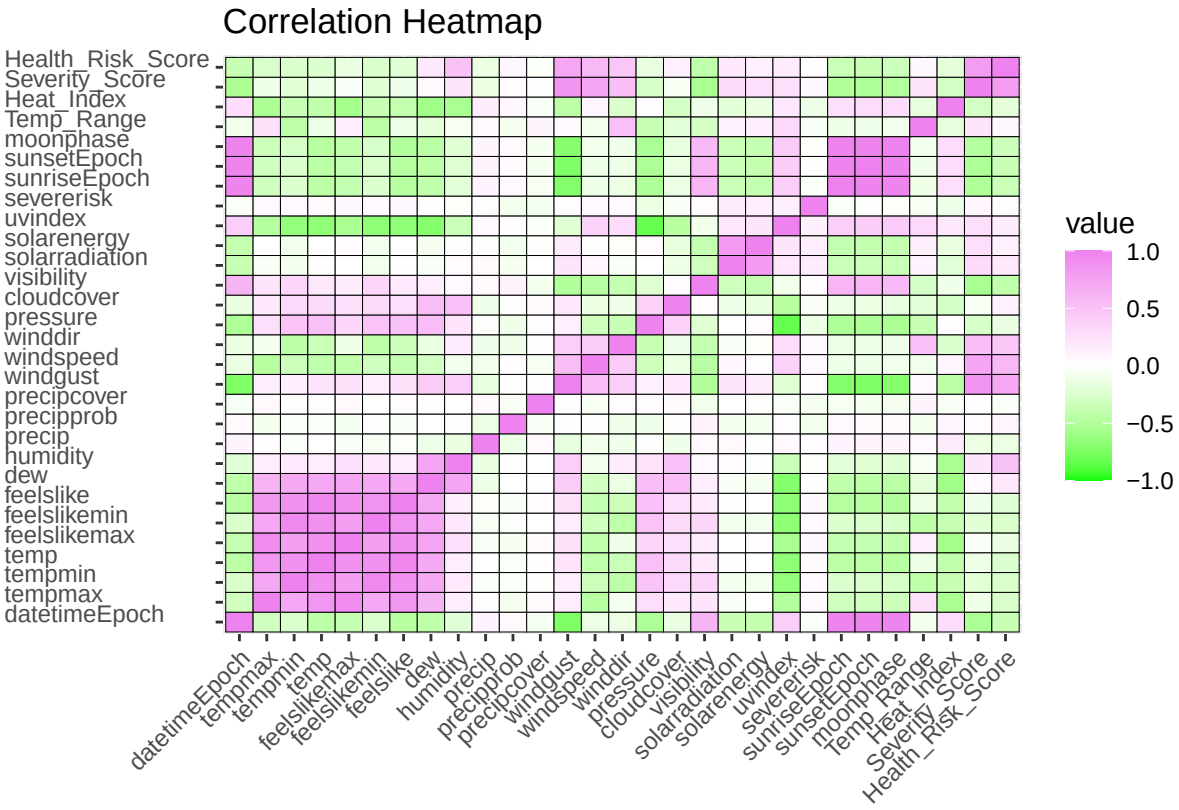
```
# Boxplot of Health Risk Score by City
ggplot(df, aes(x = City, y = Health_Risk_Score, fill = City)) +
  geom_boxplot() +
  labs(title = "City-wise Health Risk Score", x = "City", y = "Health Risk Score") +
  theme(axis.text.x = element_text(angle =90,hjust=1))
```




```r
# Boxplot of Windgust by City
ggplot(df, aes(x = City, y = windgust, fill = City)) +
  geom_boxplot() +
  labs(title = "City-wise windgust", x = "City", y = "Windgust") +
  theme(axis.text.x = element_text(angle =90,hjust=1))
```




```r
#Corrrelation between every columns
constant_columns <- sapply(df, function(x) length(unique(x)) == 1)
df <- df[, !constant_columns]

#correlation matrix
cor_matrix <- cor(df[, sapply(df, is.numeric)])

# Melt the correlation matrix for plotting
melted_cor_matrix <- melt(cor_matrix)

# Correlation heatmap
ggplot(melted_cor_matrix, aes(Var1, Var2, fill = value)) +
  geom_tile(color = "black") +
  scale_fill_gradient2(low = "green", high = "violet", mid = "white", midpoint = 0, limit = c(-1, 1)) +
  labs(title = "Correlation Heatmap", x = "", y = "") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1),axis.text.y = element_text(vjust=0,h
```




```
paste("Correlation between windgust and Health risk :",cor(df$windgust, df$Health_Risk_Score))
```

```
## [1] "Correlation between windgust and Health risk : 0.719211489784877"
```

```
plot1<-ggplot(df, aes(x = windgust, y = Health_Risk_Score)) +
  geom_point(color = "#2993ae") +
  geom_smooth(method = "lm", color = "black") + # Adds a regression line
  labs(title = "Wind Gusts vs Health Risk Score",
       x = "Wind Gust",
       y = "Health Risk Score") +
  theme_minimal()

# Calculate correlation
paste("Correlation between severity score and Health risk :",cor(df$Severity_Score, df$Health_Risk_Score
```

```
## [1] "Correlation between severity score and Health risk : 0.79812942759995"
```

```
plot2<-ggplot(df, aes(x = Severity_Score, y = Health_Risk_Score)) +
  geom_point(color = "#2993ae") +
  geom_smooth(method = "lm", color = "black") + # Adds a regression line
  labs(title = "Severity Score vs Health Risk Score",
       x = "Severity Score",
       y = "Health Risk Score") +
  theme_minimal()

paste("Correlation between wind speed and Health risk :",cor(df$windspeed, df$Health_Risk_Score))
```

```
## [1] "Correlation between wind speed and Health risk : 0.586553375240766"
```

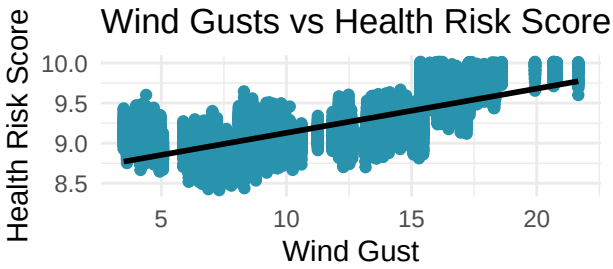
```r
plot3<-ggplot(df, aes(x = windspeed, y = Health_Risk_Score)) +
  geom_point(color = "#2993ae") +
  geom_smooth(method = "lm", color = "black") + # Adds a regression line
  labs(title = "Wind speed vs Health Risk Score",
       x = "Wind Speed",
       y = "Health Risk Score") +
  theme_minimal()

paste("Correlation between humidity and Health risk :",cor(df$humidity, df$Health_Risk_Score))
```

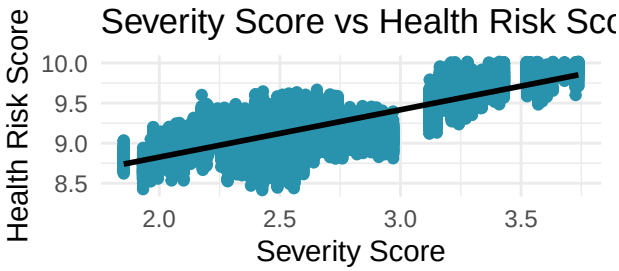
```
## [1] "Correlation between humidity and Health risk : 0.505210350812946"
```

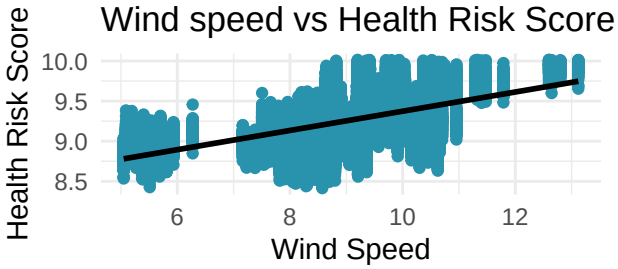
```r
plot4<-ggplot(df, aes(x = humidity, y = Health_Risk_Score)) +
  geom_point(color = "#2993ae") +
  geom_smooth(method = "lm", color = "black") + # Adds a regression line
  labs(title = "humidity vs Health Risk Score",
       x = "humidity",
       y = "Health Risk Score") +
  theme_minimal()

paste("Correlation between humidity and windgust :",cor(df$humidity, df$Health_Risk_Score))
```

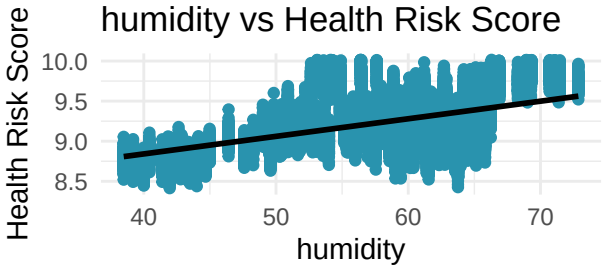
```
## [1] "Correlation between humidity and windgust : 0.505210350812946"
```

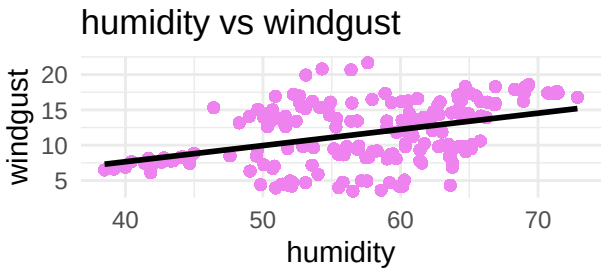
```r
plot5<-ggplot(df, aes(x = humidity, y = windgust)) +
  geom_point(color = "violet") +
  geom_smooth(method = "lm", color = "black") + # Adds a regression line
  labs(title = "humidity vs windgust",
       x = "humidity",
       y = "windgust") +
  theme_minimal()

grid.arrange(plot1, plot2,plot3,plot4,plot5, ncol =2 )
```


Wind Gusts vs Health Risk Score
Health Risk Score
10.0
9.5
9.0
8.5
5
10
15
20
Wind Gust


Severity Score vs Health Risk Sc
Health Risk Score
10.0
9.5
9.0
8.5
2.0
2.5
3.0
3.5
Severity Score


Wind speed vs Health Risk Score
Health Risk Score
10.0
9.5
9.0
8.5
6
8
10
12
Wind Speed


humidity vs Health Risk Score
Health Risk Score
10.0
9.5
9.0
8.5
40
50
60
70
humidity


humidity vs windgust
windgust
20
15
10
5
40
50
60
70
humidity