Weather

Stavrova Valeria

19 11 2019

# Loading data

weather <-readRDS("C:/Users/Adm/Downloads/weather.rds")

# Explorying data

Let’s have a look at the structure of the data.

str(weather)

## 'data.frame': 286 obs. of 35 variables:  
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ year : int 2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 ...  
## $ month : int 12 12 12 12 12 12 12 12 12 12 ...  
## $ measure: chr "Max.TemperatureF" "Mean.TemperatureF" "Min.TemperatureF" "Max.Dew.PointF" ...  
## $ X1 : chr "64" "52" "39" "46" ...  
## $ X2 : chr "42" "38" "33" "40" ...  
## $ X3 : chr "51" "44" "37" "49" ...  
## $ X4 : chr "43" "37" "30" "24" ...  
## $ X5 : chr "42" "34" "26" "37" ...  
## $ X6 : chr "45" "42" "38" "45" ...  
## $ X7 : chr "38" "30" "21" "36" ...  
## $ X8 : chr "29" "24" "18" "28" ...  
## $ X9 : chr "49" "39" "29" "49" ...  
## $ X10 : chr "48" "43" "38" "45" ...  
## $ X11 : chr "39" "36" "32" "37" ...  
## $ X12 : chr "39" "35" "31" "28" ...  
## $ X13 : chr "42" "37" "32" "28" ...  
## $ X14 : chr "45" "39" "33" "29" ...  
## $ X15 : chr "42" "37" "32" "33" ...  
## $ X16 : chr "44" "40" "35" "42" ...  
## $ X17 : chr "49" "45" "41" "46" ...  
## $ X18 : chr "44" "40" "36" "34" ...  
## $ X19 : chr "37" "33" "29" "25" ...  
## $ X20 : chr "36" "32" "27" "30" ...  
## $ X21 : chr "36" "33" "30" "30" ...  
## $ X22 : chr "44" "39" "33" "39" ...  
## $ X23 : chr "47" "45" "42" "45" ...  
## $ X24 : chr "46" "44" "41" "46" ...  
## $ X25 : chr "59" "52" "44" "58" ...  
## $ X26 : chr "50" "44" "37" "31" ...  
## $ X27 : chr "52" "45" "38" "34" ...  
## $ X28 : chr "52" "46" "40" "42" ...  
## $ X29 : chr "41" "36" "30" "26" ...  
## $ X30 : chr "30" "26" "22" "10" ...  
## $ X31 : chr "30" "25" "20" "8" ...

summary(weather)

## X year month measure   
## Min. : 1.00 Min. :2014 Min. : 1.000 Length:286   
## 1st Qu.: 72.25 1st Qu.:2015 1st Qu.: 4.000 Class :character   
## Median :143.50 Median :2015 Median : 7.000 Mode :character   
## Mean :143.50 Mean :2015 Mean : 6.923   
## 3rd Qu.:214.75 3rd Qu.:2015 3rd Qu.:10.000   
## Max. :286.00 Max. :2015 Max. :12.000   
## X1 X2 X3 X4   
## Length:286 Length:286 Length:286 Length:286   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## X5 X6 X7 X8   
## Length:286 Length:286 Length:286 Length:286   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## X9 X10 X11 X12   
## Length:286 Length:286 Length:286 Length:286   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## X13 X14 X15 X16   
## Length:286 Length:286 Length:286 Length:286   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## X17 X18 X19 X20   
## Length:286 Length:286 Length:286 Length:286   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## X21 X22 X23 X24   
## Length:286 Length:286 Length:286 Length:286   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## X25 X26 X27 X28   
## Length:286 Length:286 Length:286 Length:286   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## X29 X30 X31   
## Length:286 Length:286 Length:286   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##

# Tidying data

We see that values are stored in columns, not rows. We need to fix that.

library(tidyr)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

newweather <-weather[,-1] %>% gather(key = day, value = val, X1 : X31) %>%   
 pivot\_wider(names\_from = measure, values\_from = val)

Also we need to sort out the date of observation. I decided to keep days, months and years in different columns, because it may be convenient to select certain years/months/days as in the examples below.

library(data.table)

##   
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

draftday<-data.table(newweather$day)   
day<- separate(draftday, V1, c("X", "day"),sep = 1)[,2]  
finalweather<-cbind(day, newweather[,-3])[,c(2,3,1,4:25)]

We can also see that the columns that should be numeric are actually character. Precipitation column has some T values, but it may not be so, it’s not a logical variable. We need to fix that too.

finalweather1 <- finalweather %>% mutate\_at(vars(-Events), as.numeric) %>% filter(PrecipitationIn != "T")

## Warning: РІ СЂРµР·СѓР»СЊС‚Р°С‚Рµ РїСЂРµРѕР±СЂР°Р·РѕРІР°РЅРёСЏ СЃРѕР·РґР°РЅС‹ NA

Examples.

finalweather1 %>% select(c(1:5)) %>% subset(year==2014)

## year month day Max.TemperatureF Mean.TemperatureF  
## 1 2014 12 1 64 52  
## 11 2014 12 2 42 38  
## 23 2014 12 3 51 44  
## 35 2014 12 4 43 37  
## 45 2014 12 5 42 34  
## 55 2014 12 6 45 42  
## 66 2014 12 7 38 30  
## 77 2014 12 8 29 24  
## 86 2014 12 9 49 39  
## 97 2014 12 10 48 43  
## 109 2014 12 11 39 36  
## 142 2014 12 14 45 39  
## 153 2014 12 15 42 37  
## 172 2014 12 17 49 45  
## 182 2014 12 18 44 40  
## 193 2014 12 19 37 33  
## 221 2014 12 22 44 39  
## 228 2014 12 23 47 45  
## 240 2014 12 24 46 44  
## 249 2014 12 25 59 52  
## 260 2014 12 26 50 44  
## 269 2014 12 27 52 45  
## 281 2014 12 28 52 46  
## 292 2014 12 29 41 36  
## 301 2014 12 30 30 26  
## 311 2014 12 31 30 25

finalweather1 %>% select(year, month, day, Events) %>% subset(month==12) %>% subset(day==1)

## year month day Events  
## 1 2014 12 1 Rain  
## 10 2015 12 1 Rain

Then we deal with the missing values.

finalweather2<-na.omit(finalweather1)

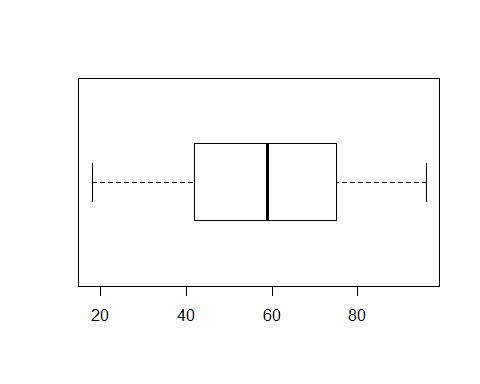
# Let’s have a look at the result we’ve got

summary(finalweather2)

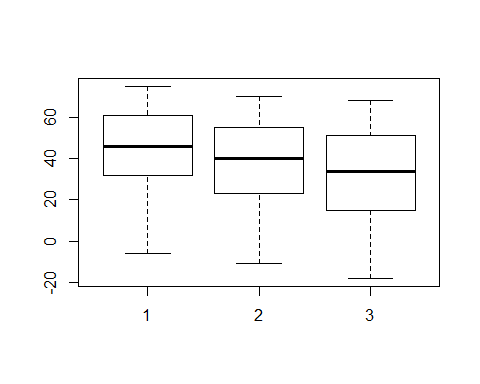
## year month day Max.TemperatureF  
## Min. :2014 Min. : 1.000 Min. : 1.0 Min. :18.00   
## 1st Qu.:2015 1st Qu.: 3.000 1st Qu.: 8.0 1st Qu.:42.00   
## Median :2015 Median : 6.000 Median :15.0 Median :59.00   
## Mean :2015 Mean : 6.463 Mean :15.7 Mean :58.24   
## 3rd Qu.:2015 3rd Qu.: 9.500 3rd Qu.:23.5 3rd Qu.:75.00   
## Max. :2015 Max. :12.000 Max. :31.0 Max. :96.00   
## Mean.TemperatureF Min.TemperatureF Max.Dew.PointF MeanDew.PointF   
## Min. : 8.0 Min. :-3.00 Min. :-6.00 Min. :-11.00   
## 1st Qu.:36.0 1st Qu.:30.00 1st Qu.:32.00 1st Qu.: 23.00   
## Median :52.0 Median :44.00 Median :46.00 Median : 40.00   
## Mean :50.7 Mean :42.59 Mean :44.79 Mean : 38.14   
## 3rd Qu.:67.5 3rd Qu.:59.50 3rd Qu.:61.00 3rd Qu.: 55.00   
## Max. :84.0 Max. :74.00 Max. :75.00 Max. : 70.00   
## Min.DewpointF Max.Humidity Mean.Humidity Min.Humidity   
## Min. :-18.00 Min. : 39.00 Min. :28.00 Min. :16.00   
## 1st Qu.: 14.50 1st Qu.: 73.00 1st Qu.:55.50 1st Qu.:34.00   
## Median : 33.00 Median : 86.00 Median :66.00 Median :45.00   
## Mean : 31.35 Mean : 85.86 Mean :65.68 Mean :47.87   
## 3rd Qu.: 51.00 3rd Qu.: 93.00 3rd Qu.:76.00 3rd Qu.:59.00   
## Max. : 68.00 Max. :1000.00 Max. :98.00 Max. :96.00   
## Max.Sea.Level.PressureIn Mean.Sea.Level.PressureIn Min.Sea.Level.PressureIn  
## Min. :29.58 Min. :29.49 Min. :29.16   
## 1st Qu.:30.00 1st Qu.:29.88 1st Qu.:29.75   
## Median :30.14 Median :30.04 Median :29.94   
## Mean :30.17 Mean :30.05 Mean :29.93   
## 3rd Qu.:30.32 3rd Qu.:30.20 3rd Qu.:30.09   
## Max. :30.88 Max. :30.77 Max. :30.64   
## Max.VisibilityMiles Mean.VisibilityMiles Min.VisibilityMiles Max.Wind.SpeedMPH  
## Min. : 2.000 Min. :-1.000 Min. : 0.000 Min. :10.00   
## 1st Qu.:10.000 1st Qu.: 8.000 1st Qu.: 2.000 1st Qu.:16.00   
## Median :10.000 Median :10.000 Median :10.000 Median :20.00   
## Mean : 9.891 Mean : 8.727 Mean : 6.595 Mean :20.85   
## 3rd Qu.:10.000 3rd Qu.:10.000 3rd Qu.:10.000 3rd Qu.:24.00   
## Max. :10.000 Max. :10.000 Max. :10.000 Max. :38.00   
## Mean.Wind.SpeedMPH Max.Gust.SpeedMPH PrecipitationIn CloudCover   
## Min. : 4.00 Min. : 0.00 Min. :0.0000 Min. :0.000   
## 1st Qu.: 8.00 1st Qu.:21.00 1st Qu.:0.0000 1st Qu.:3.000   
## Median :10.00 Median :25.00 Median :0.0000 Median :5.000   
## Mean :10.83 Mean :27.17 Mean :0.1195 Mean :4.598   
## 3rd Qu.:13.00 3rd Qu.:32.00 3rd Qu.:0.0700 3rd Qu.:7.000   
## Max. :22.00 Max. :94.00 Max. :2.9000 Max. :8.000   
## Events WindDirDegrees   
## Length:311 Min. : 1.0   
## Class :character 1st Qu.:113.5   
## Mode :character Median :223.0   
## Mean :201.4   
## 3rd Qu.:278.0   
## Max. :360.0

# Visualising

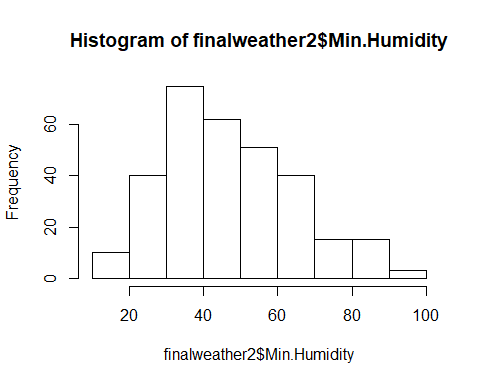
boxplot(finalweather2$Max.TemperatureF, horizontal = T)



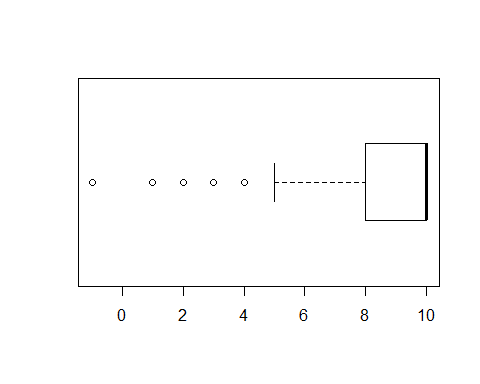
boxplot(finalweather2$Max.Dew.PointF,finalweather1$MeanDew.PointF,finalweather1$Min.DewpointF)



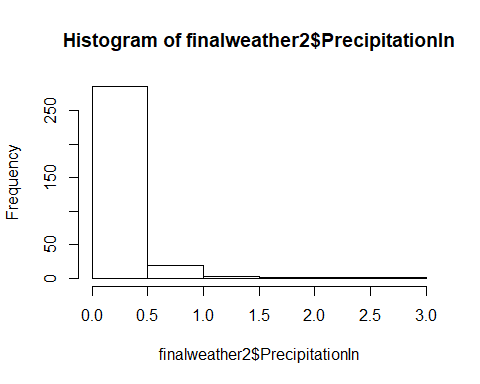
hist(finalweather2$Min.Humidity)



boxplot(finalweather2$Mean.VisibilityMiles, horizontal = T)



hist(finalweather2$PrecipitationIn)



*The End*