R Notebook for data incubator_minjie XU_2

Code

This is an R Markdown (http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

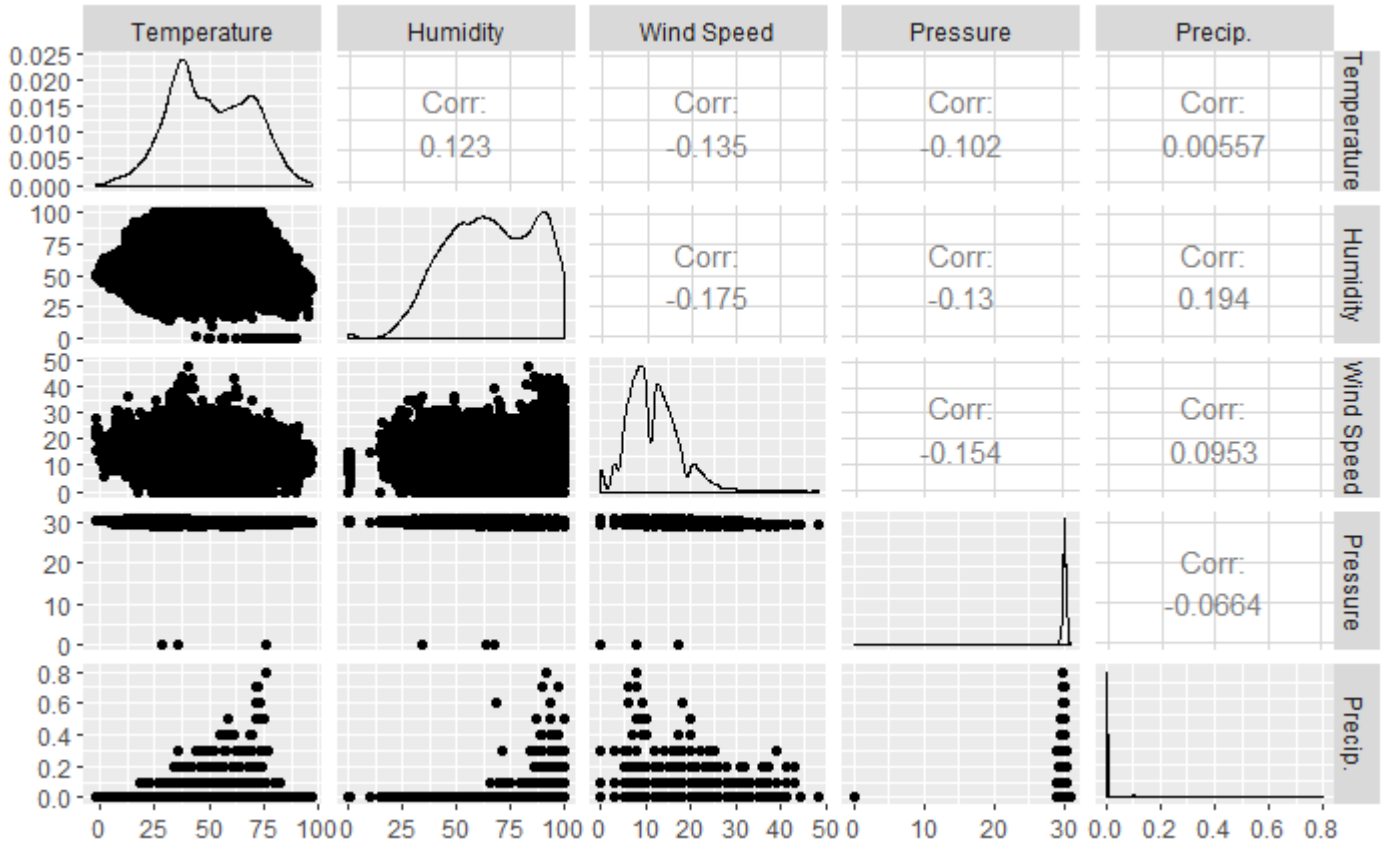
Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

Hide

```
weather<-as.data.frame(fread("C:\\Users\\Minjie\\Documents\\2018 Fall Master\\Data Mining\\project\\weather2.csv"))
weather[(weather$Date %in% c("2017-03-12","2018-03-11","2019-03-10") & weather$Time==2),"Time"]<-3
```

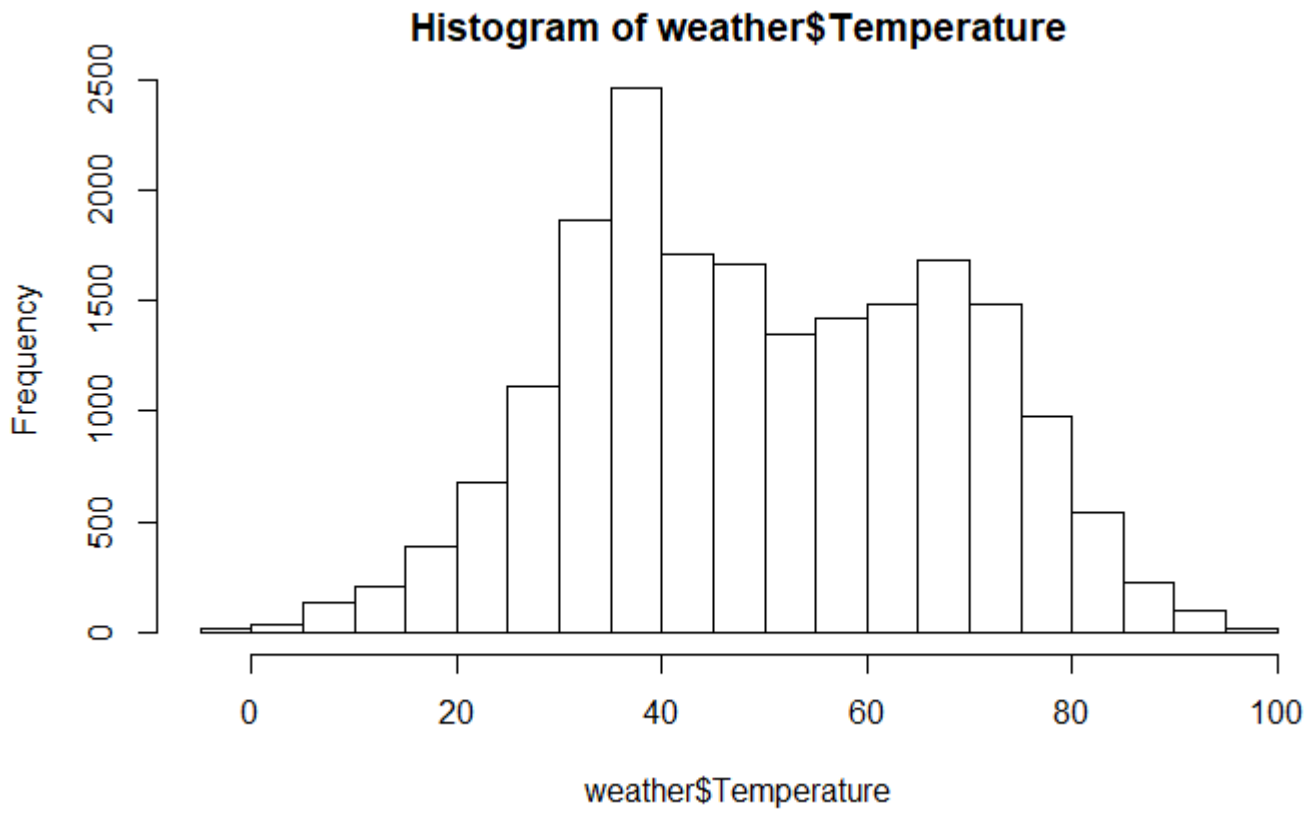
Hide

```
library(GGally)
ggpairs(weather[, c("Temperature","Humidity","Wind Speed","Pressure","Precip.")])
```



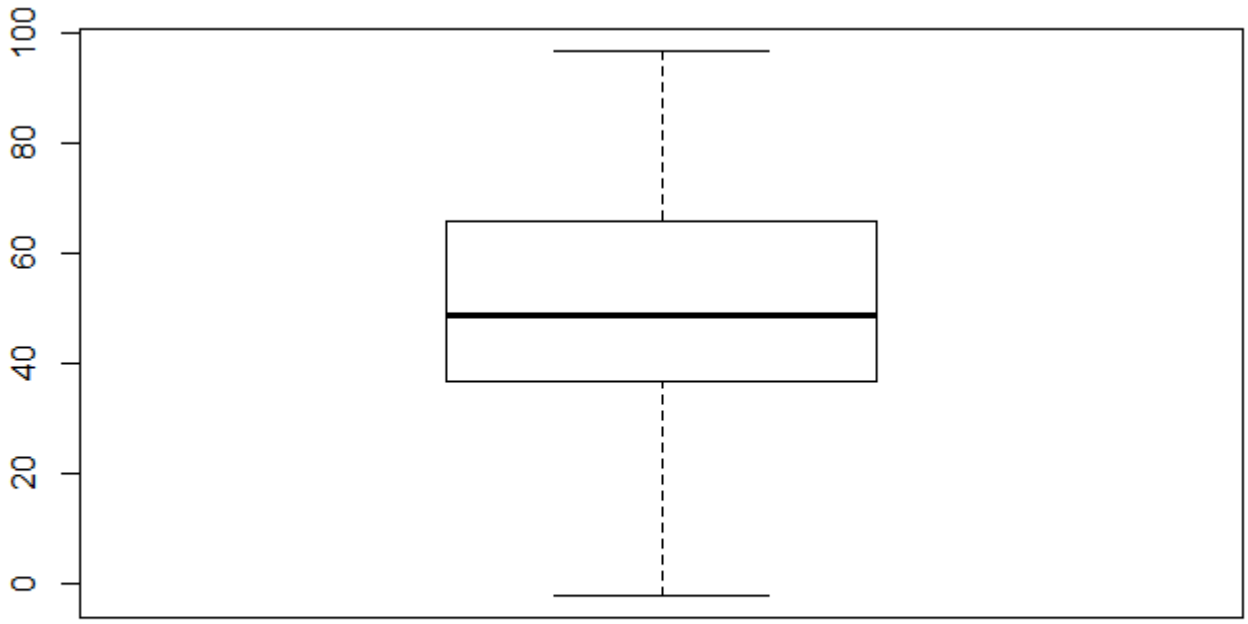
Hide

```
hist(weather$Temperature)
```



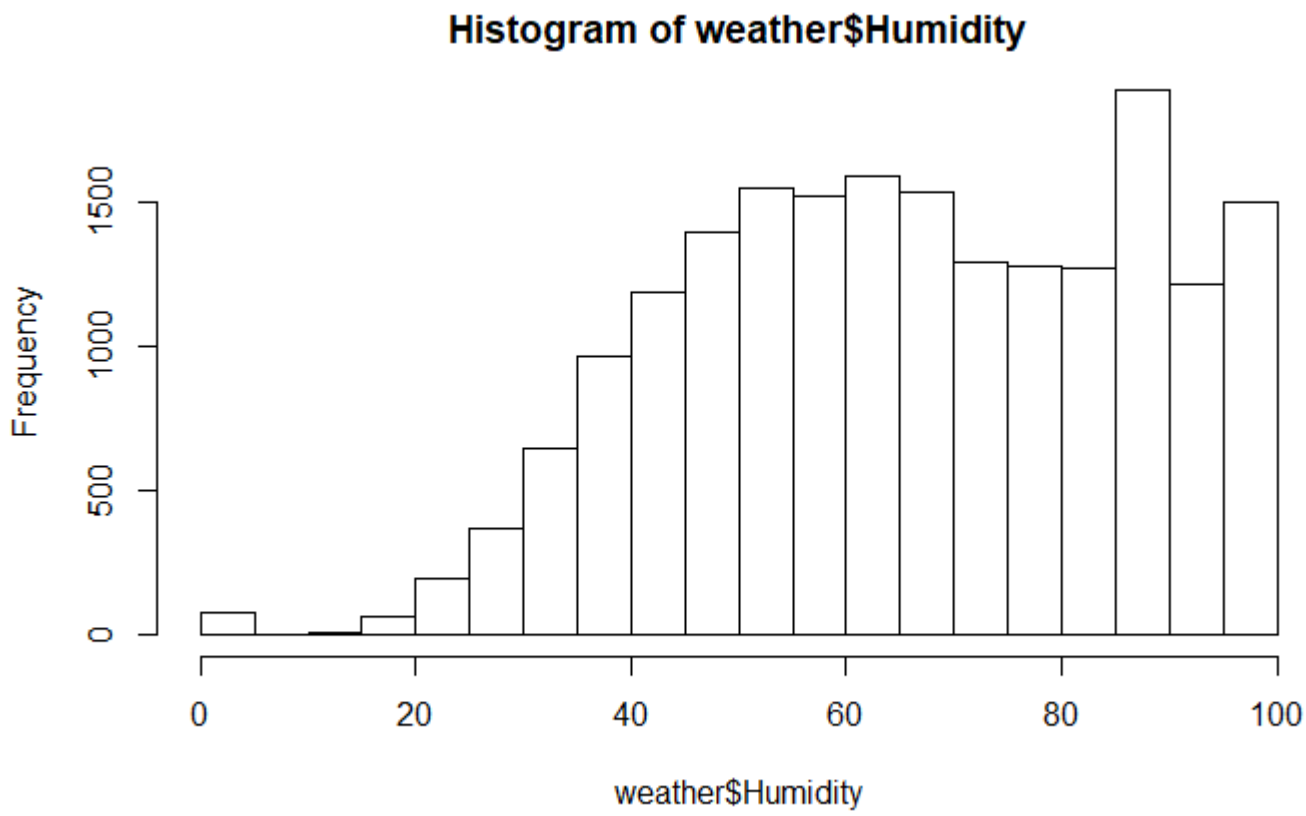
Hide

```
boxplot(weather$Temperature)
```



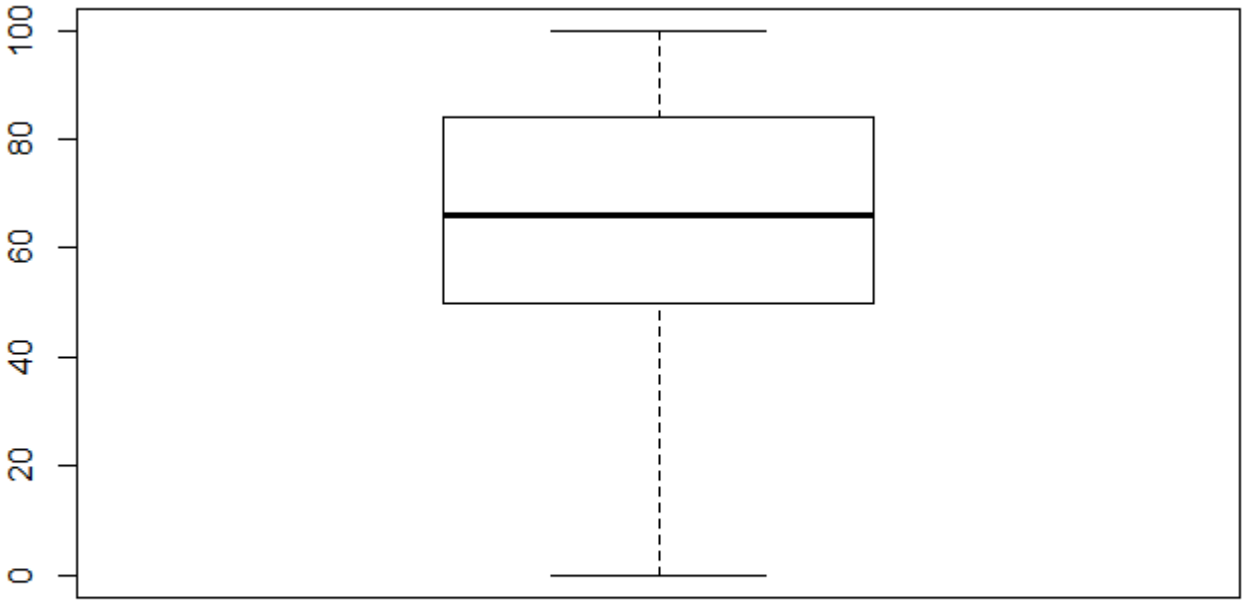
Hide

```
hist(weather$Humidity)
```



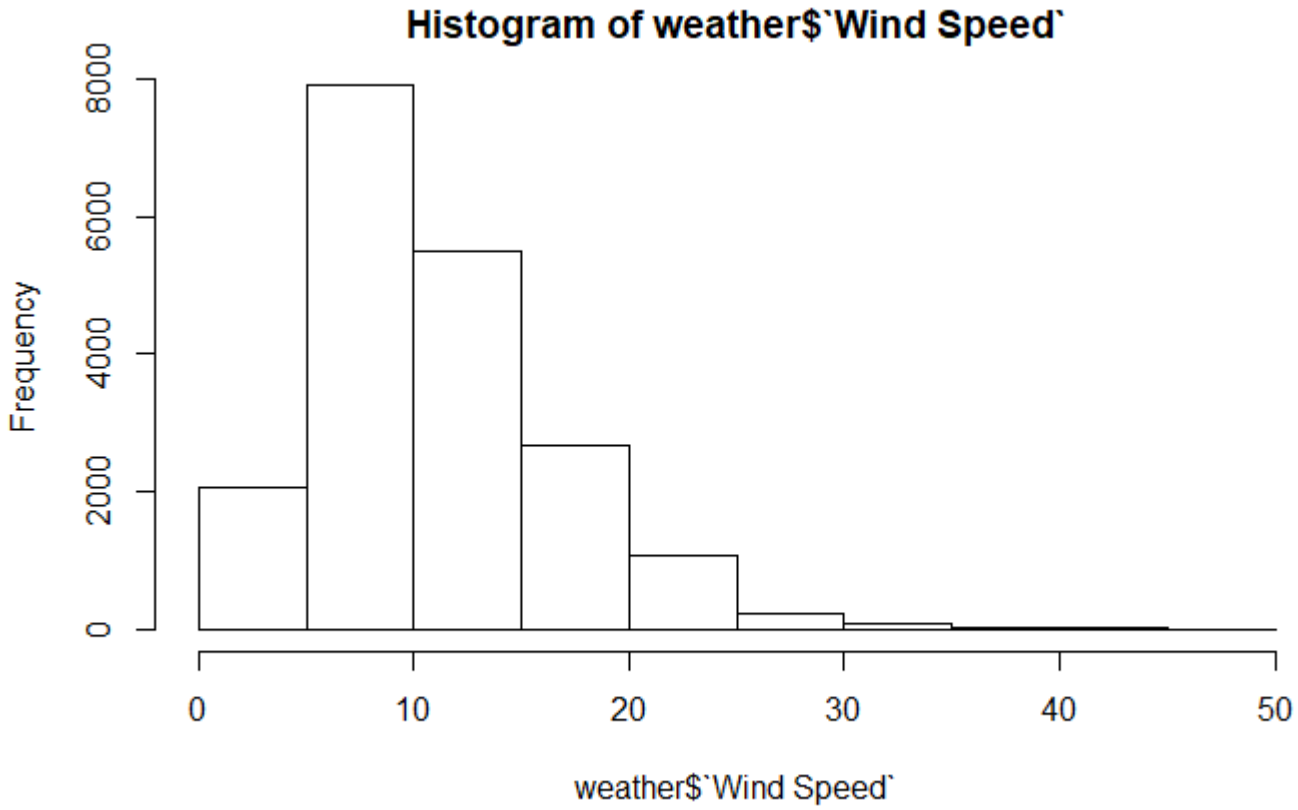
Hide

```
boxplot(weather$Humidity)
```



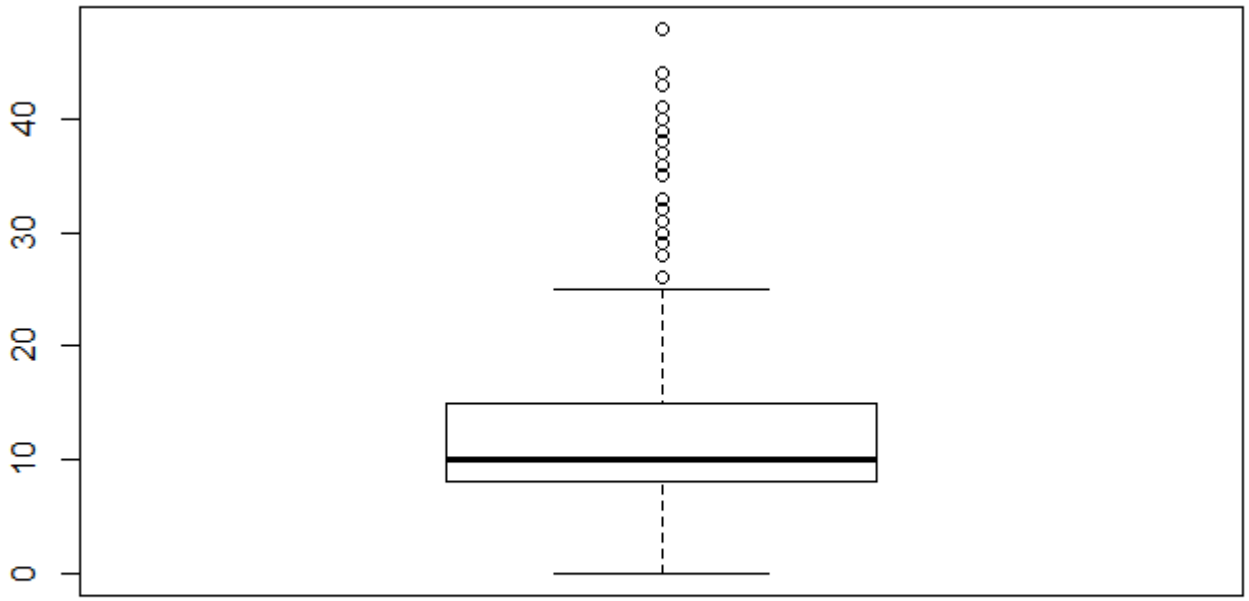
Hide

```
hist(weather$`Wind Speed`)
```



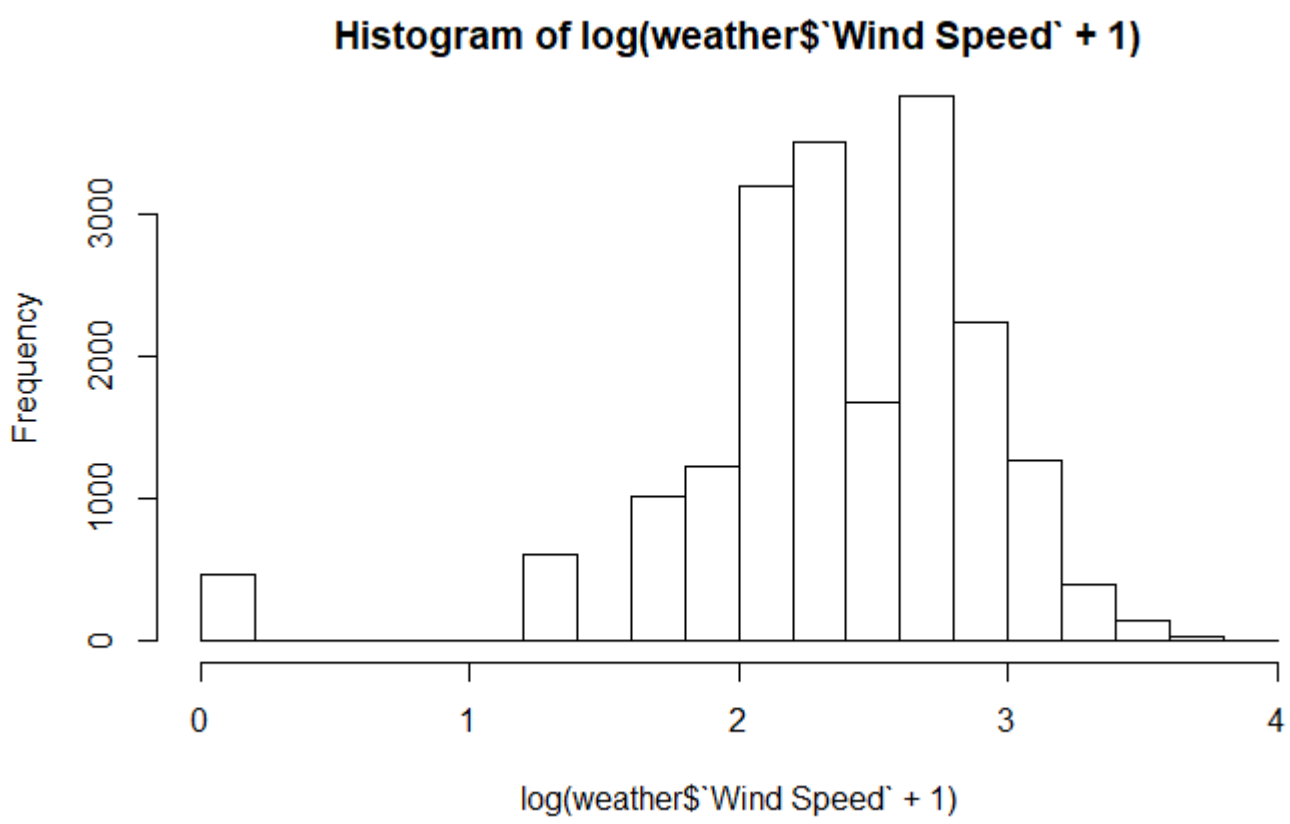
Hide

```
boxplot(weather$`Wind Speed`)# right skewed
```



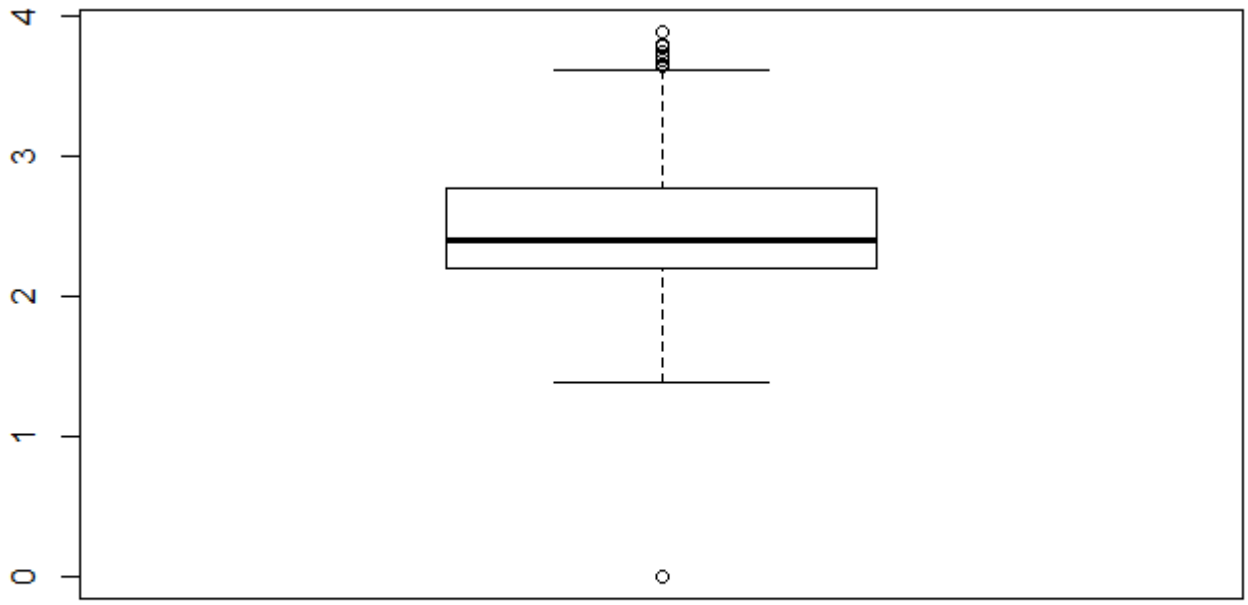
Hide

```
hist(log(weather$`Wind Speed`+1))
```



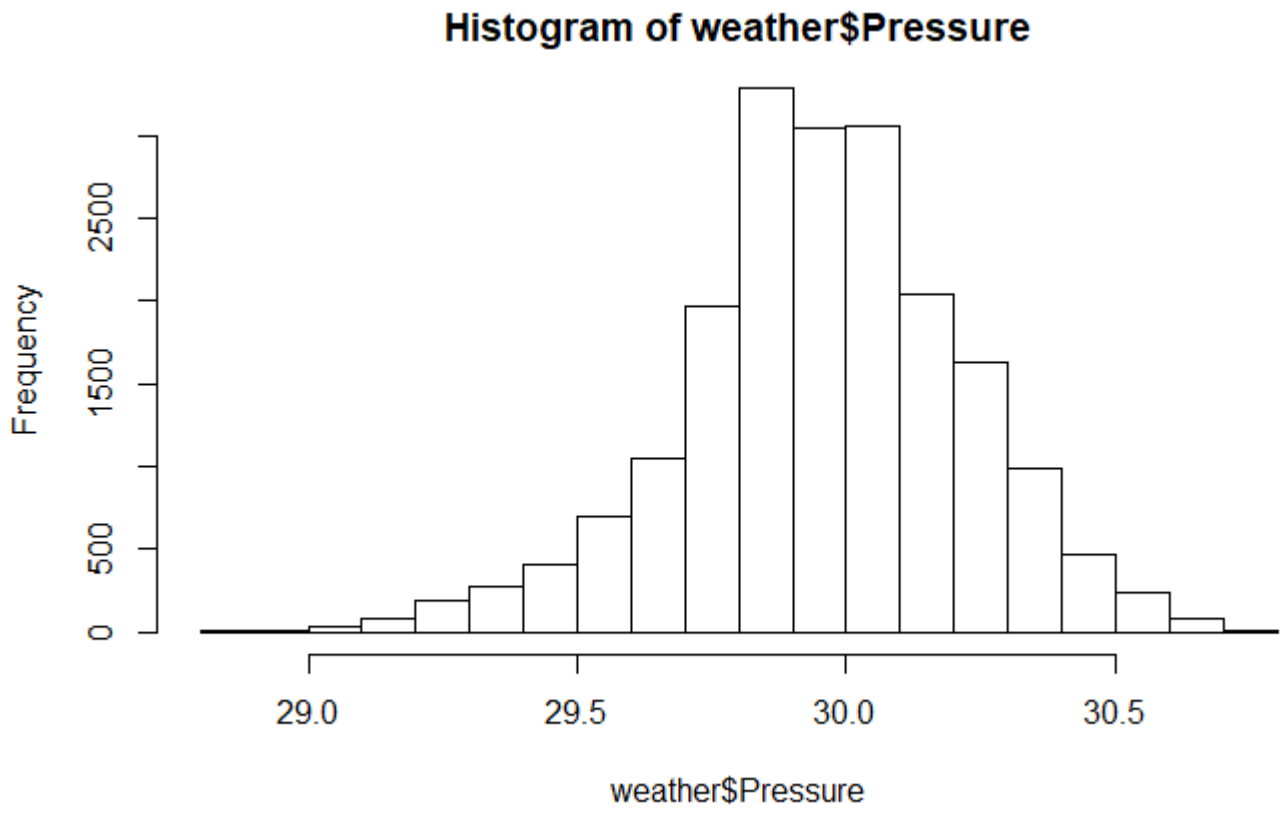
Hide

```
boxplot(log(weather$`Wind Speed`+1))
```



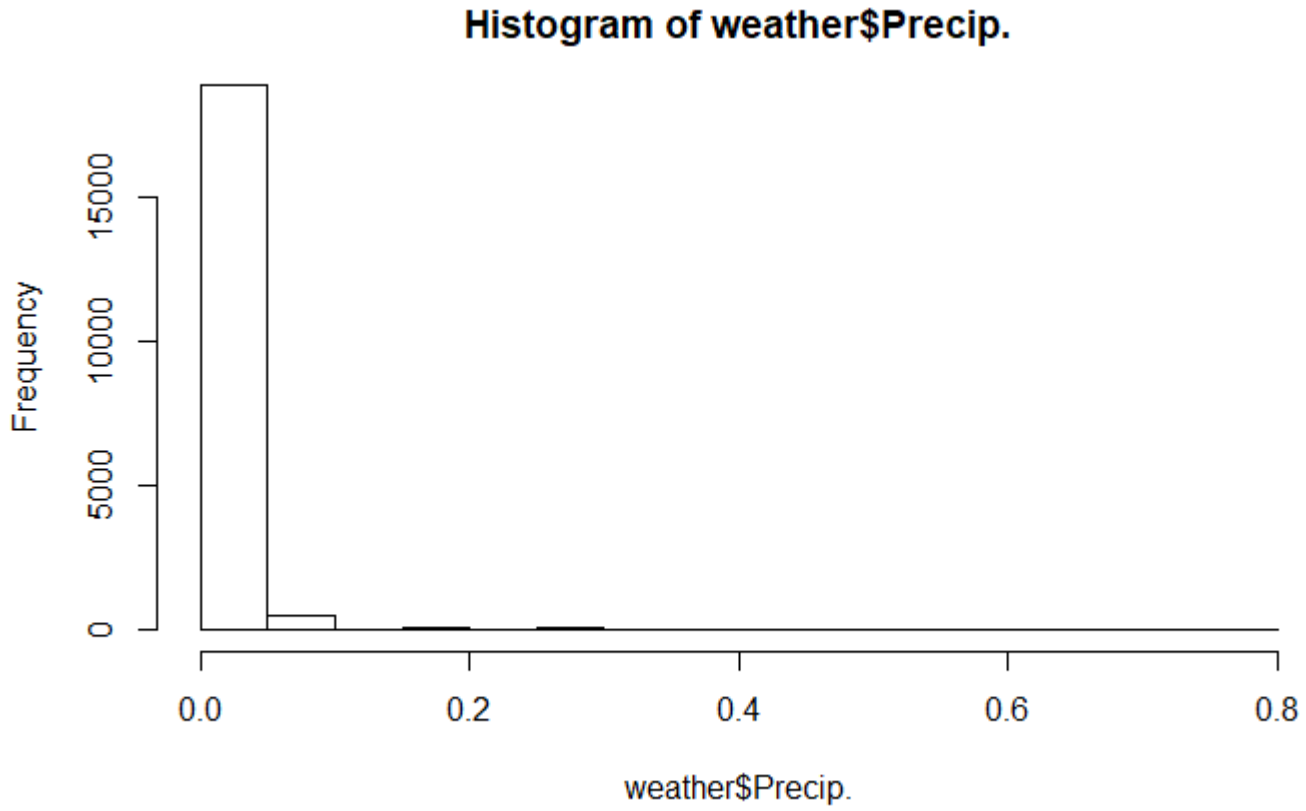
Hide

```
weather$`Wind Speed`<- log(weather$`Wind Speed`+1)
weather$Pressure[weather$Pressure==0]<-NA # remove outliers
hist(weather$Pressure)
```



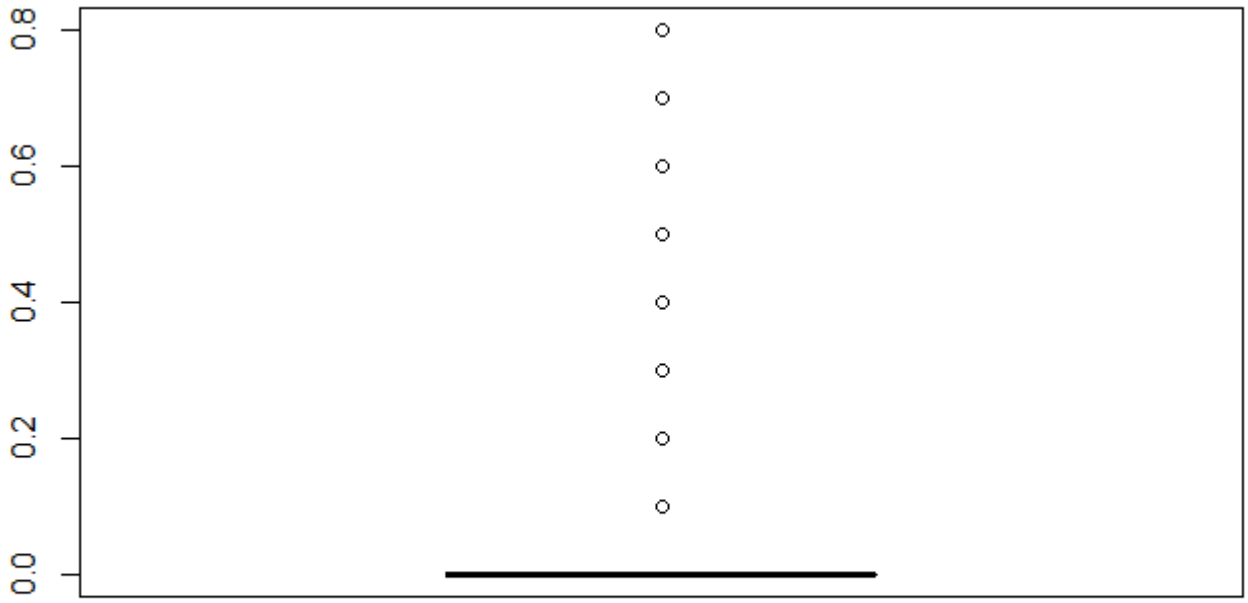
Hide

```
hist(weather$Precip.)
```



Hide

```
boxplot(weather$Precip.)
```



Hide

```
table(weather$Precip.) # majority is 0
```

```
    0   0.1   0.2   0.3   0.4   0.5   0.6   0.7   0.8
18957   499    55    23     6     4     3     2     1
```

Hide

```
## word cloud for weather condition
weather$Condition<-gsub("T-Storm", "TStorm",weather$Condition)
#install.packages(c("tm", "SnowballC", "wordcloud", "RColorBrewer", "RCurl", "XML"))
library("tm")
```

```
package <U+393C><U+3E31>tm<U+393C><U+3E32> was built under R version 3.5.3Loading required packa
ge: NLP
package <U+393C><U+3E31>NLP<U+393C><U+3E32> was built under R version 3.5.2
Attaching package: <U+393C><U+3E31>NLP<U+393C><U+3E32>

The following object is masked from <U+393C><U+3E31>package:ggplot2<U+393C><U+3E32>:

    annotate
```

Hide

```
library("SnowballC")
```

```
package <U+393C><U+3E31>SnowballC<U+393C><U+3E32> was built under R version 3.5.2
```

Hide

```
library("wordcloud")
```

```
package <U+393C><U+3E31>wordcloud<U+393C><U+3E32> was built under R version 3.5.3Loading require
d package: RColorBrewer
package <U+393C><U+3E31>RColorBrewer<U+393C><U+3E32> was built under R version 3.5.2
```

Hide

```
library("RColorBrewer")
# Load the data as a corpus
docs <- Corpus(VectorSource(weather$Condition))
#remove special characters from the text.
toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
docs <- tm_map(docs, toSpace, "/")
```

```
transformation drops documents
```

Hide

```
docs <- tm_map(docs, toSpace, "-")
```

```
transformation drops documents
```

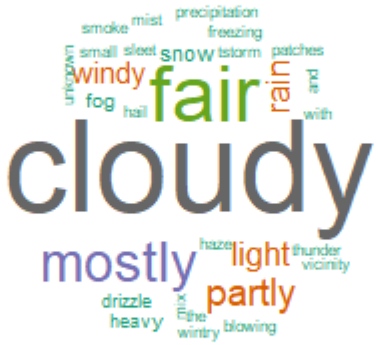
Hide

```
# Convert the text to lower case
docs <- tm_map(docs, content_transformer(tolower))
```

```
transformation drops documents
```

Hide

```
# Build a term-document matrix
dtm <- TermDocumentMatrix(docs)
m <- as.matrix(dtm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
# Build a term-document matrix
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
          max.words=Inf, random.order=FALSE, rot.per=0.15,
          colors=brewer.pal(8, "Dark2"))
```



Hide

```
conditionlist<-as.data.frame(table(weather$Condition))
index1<-c(3,4,7,8,29,30,31,32,33) # weather condition category 1
index2<-c(6,9,10,11,28,34,39) #weather condition category 2
index3<-c(5,18,19,20,21,22,23,24,25,26,27,35,36,37,38,47) #weather condition category 3
index4<-c(1,2,12,13,14,15,16,17,40,41,42,43,44,45,46,48,49) #weather condition category 4
conditionlist$category<-NA
conditionlist$category[index1]<-1
conditionlist$category[index2]<-2
conditionlist$category[index3]<-3
conditionlist$category[index4]<-4
for (name in conditionlist$Var1)
        weather$Condition[weather$Condition==name]<-conditionlist$category[conditionlist$Var1==n
ame]
weather$newtime<-as.POSIXct(paste(weather$Date, weather$Time), format="%Y-%m-%d %H")
```

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.