R_data_cleaning_Example

Demetrious Lloyd

2024-01-25

```r
#Shut off the warnings
options(warn = -1)
#install and load libraries
library(data.table)
library(stats)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

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

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

```r
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
library(broom)
library(ggplot2)
library(BSDA)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'BSDA'
```

```
## The following object is masked from 'package:datasets':
##
##     Orange
```

```r
library(naniar)

#open csv as DataFrame with the correct filepath
df<- read.csv("C:/Users/raven/Desktop/[03] Portfolio Projects/Sofia Air Quality/2017-07_bme280sof.csv", header=TRUE)
str(df)
```

```
## 'data.frame':    701548 obs. of  9 variables:
##  $ X          : int  1 5 7 9 10 11 13 22 25 30 ...
##  $ sensor_id  : int  2266 2292 3096 3428 3472 1952 1846 3512 2228 3438 ...
##  $ location   : int  1140 1154 1558 1727 1750 976 923 1770 1120 1732 ...
##  $ lat        : num  42.7 42.7 42.7 42.6 42.7 ...
##  $ lon        : num  23.3 23.3 23.4 23.4 23.3 ...
##  $ timestamp  : chr  "2017-07-01T00:00:07" "2017-07-01T00:00:08" "2017-07-01T00:00:10" "2017-07-01T00:00:12" ...
##  $ pressure   : num  95270 94356 95156 94680 94328 ...
##  $ temperature: num  23.5 23.1 26.5 28.3 26.3 ...
##  $ humidity   : num  62.5 59.5 44.4 38.3 46.4 ...
```

```r
#Convert  the timestamp column from character to time

df$timestamp <- as.POSIXct(df$timestamp)

#observe structure of DataFrame
str(df)
```

```
## 'data.frame':    701548 obs. of  9 variables:
##  $ X          : int  1 5 7 9 10 11 13 22 25 30 ...
##  $ sensor_id  : int  2266 2292 3096 3428 3472 1952 1846 3512 2228 3438 ...
##  $ location   : int  1140 1154 1558 1727 1750 976 923 1770 1120 1732 ...
##  $ lat        : num  42.7 42.7 42.7 42.6 42.7 ...
##  $ lon        : num  23.3 23.3 23.4 23.4 23.3 ...
##  $ timestamp  : POSIXct, format: "2017-07-01" "2017-07-01" ...
##  $ pressure   : num  95270 94356 95156 94680 94328 ...
##  $ temperature: num  23.5 23.1 26.5 28.3 26.3 ...
##  $ humidity   : num  62.5 59.5 44.4 38.3 46.4 ...
```

```r
#remove ID column
df_2017_07_Air <-df[, 2:9]

#Double Check
str(df_2017_07_Air)
```

```
## 'data.frame':    701548 obs. of  8 variables:
##  $ sensor_id  : int  2266 2292 3096 3428 3472 1952 1846 3512 2228 3438 ...
##  $ location   : int  1140 1154 1558 1727 1750 976 923 1770 1120 1732 ...
##  $ lat        : num  42.7 42.7 42.7 42.6 42.7 ...
##  $ lon        : num  23.3 23.3 23.4 23.4 23.3 ...
##  $ timestamp  : POSIXct, format: "2017-07-01" "2017-07-01" ...
##  $ pressure   : num  95270 94356 95156 94680 94328 ...
```

```
##  $ temperature: num  23.5 23.1 26.5 28.3 26.3 ...
##  $ humidity   : num  62.5 59.5 44.4 38.3 46.4 ...
```

```
head(df_2017_07_Air[1])
```

```
##   sensor_id
## 1      2266
## 2      2292
## 3      3096
## 4      3428
## 5      3472
## 6      1952
```

```
#Check for duplicate values
sum(duplicated(df_2017_07_Air))
```

```
## [1] 4026
```

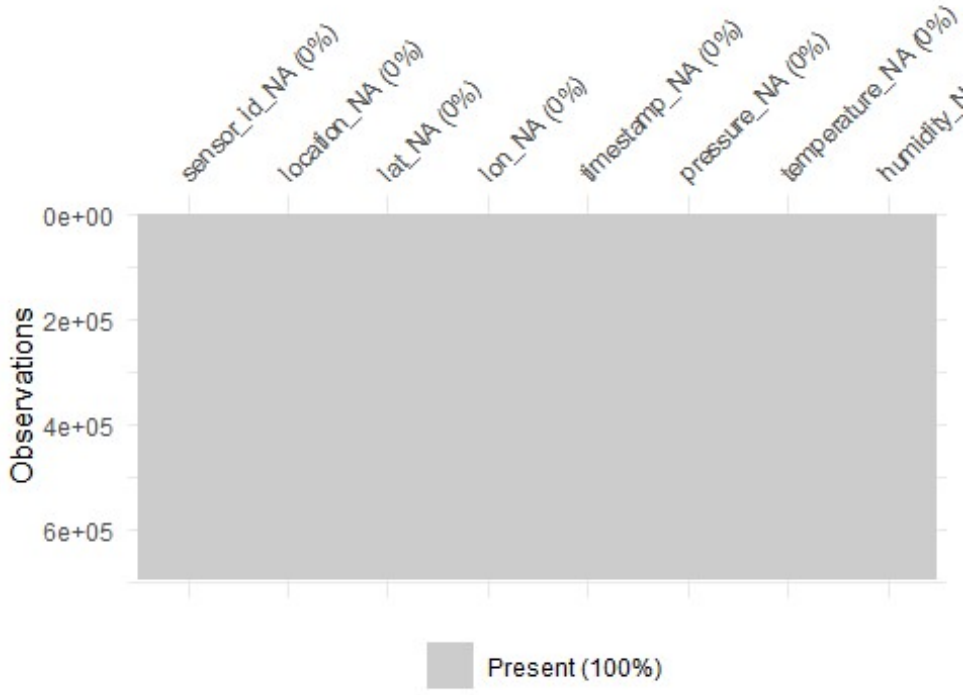
```
#4026 duplicates in dataframe
# remove them with unique and check
df_2017_07_Air <- unique(df_2017_07_Air)
sum(duplicated(df_2017_07_Air))
```

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

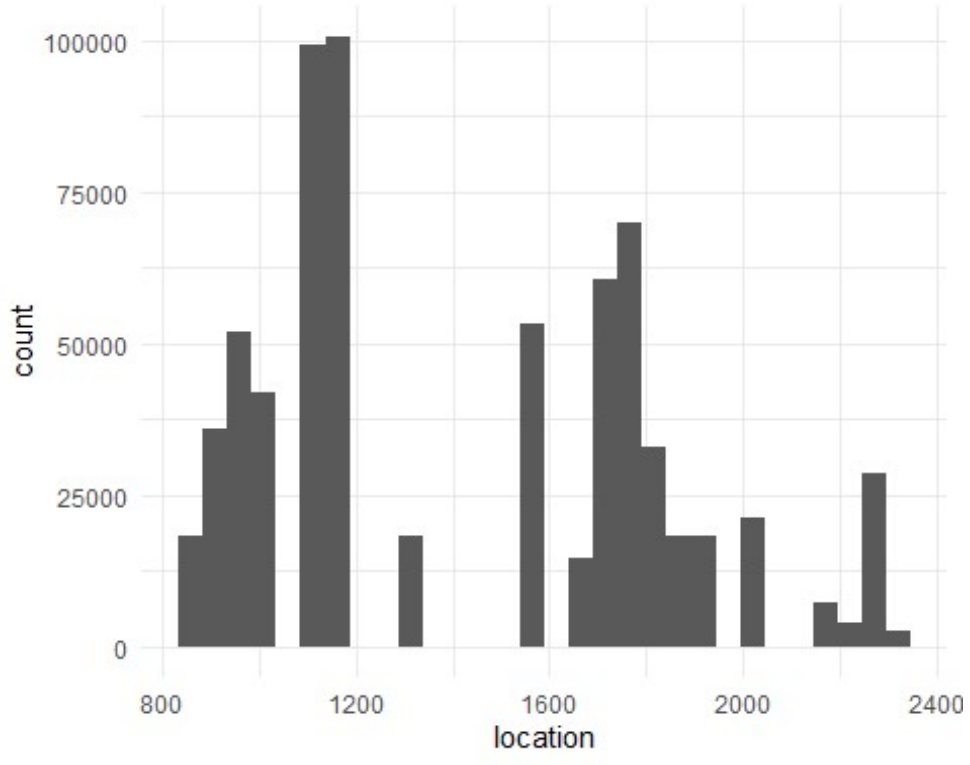
```
#Count missing values in each column
colSums(is.na(df_2017_07_Air))
```

```
##   sensor_id    location         lat         lon   timestamp    pressure
##           0           0           0           0           0           0
## temperature    humidity
##           0           0
```

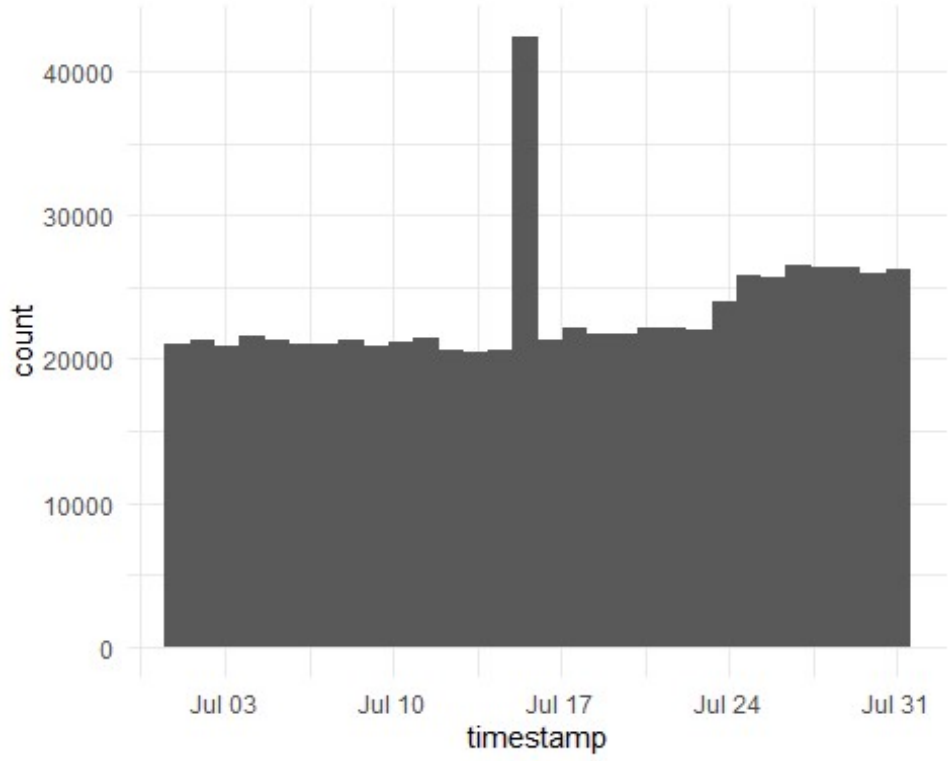
```
#Visualize missing values with a shadow matrix
vis_miss(as_shadow(df_2017_07_Air), warn_large_data = FALSE)
```



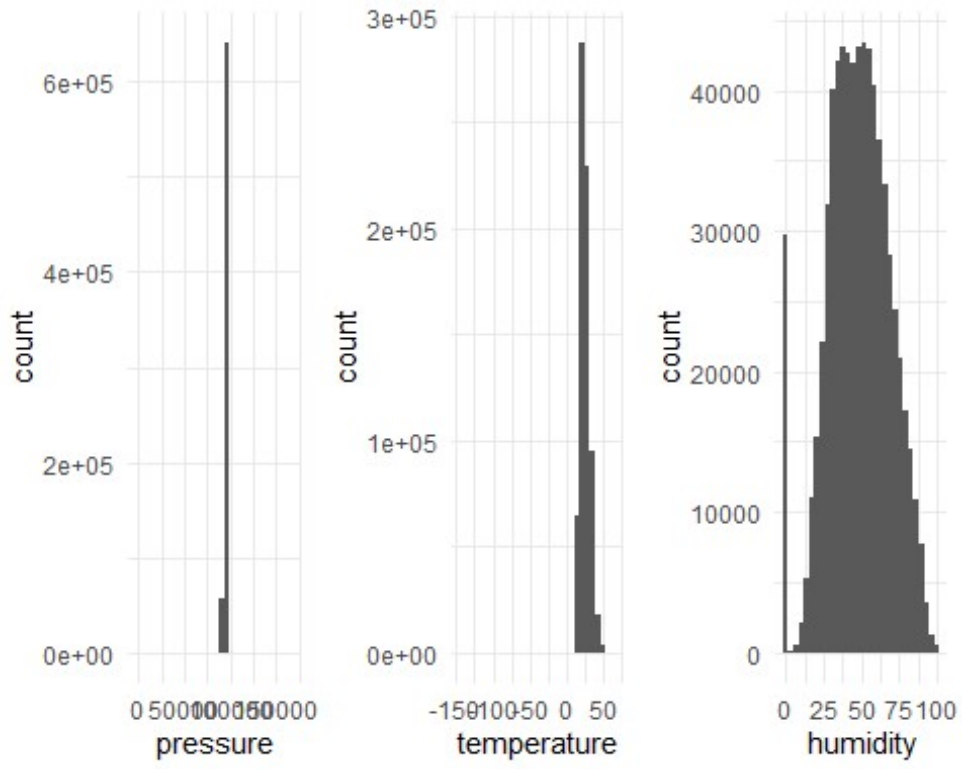

```
#Step 3. detect Outliers
colnames(df_2017_07_Air)

## [1] "sensor_id"   "location"    "lat"         "lon"         "timestamp"
## [6] "pressure"    "temperature" "humidity"

str(df_2017_07_Air)

## 'data.frame':    697522 obs. of  8 variables:
##  $ sensor_id  : int  2266 2292 3096 3428 3472 1952 1846 3512 2228 3438 ...
##  $ location   : int  1140 1154 1558 1727 1750 976 923 1770 1120 1732 ...
##  $ lat        : num  42.7 42.7 42.7 42.6 42.7 ...
##  $ lon        : num  23.3 23.3 23.4 23.4 23.3 ...
##  $ timestamp  : POSIXct, format: "2017-07-01" "2017-07-01" ...
##  $ pressure   : num  95270 94356 95156 94680 94328 ...
##  $ temperature: num  23.5 23.1 26.5 28.3 26.3 ...
##  $ humidity   : num  62.5 59.5 44.4 38.3 46.4 ...

#Visualize location histogram
ggplot(df_2017_07_Air, aes(x = location)) + geom_histogram() +theme_minimal()

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



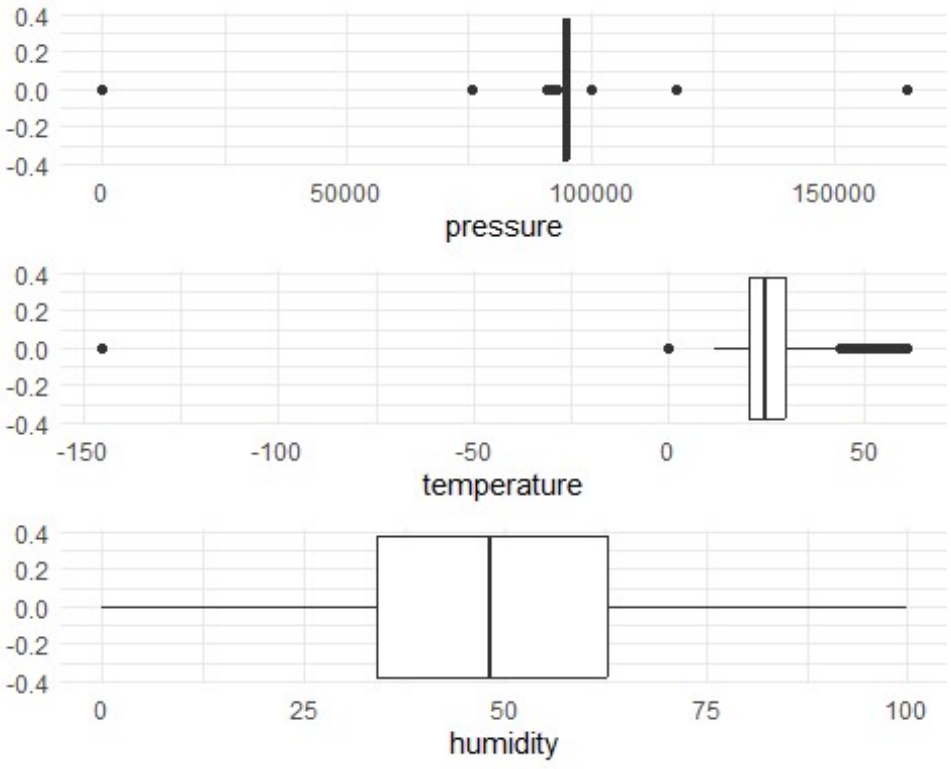

```
#Visualize timestamp histogram
ggplot(df_2017_07_Air, aes(x = timestamp)) + geom_histogram()
+theme_minimal()

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



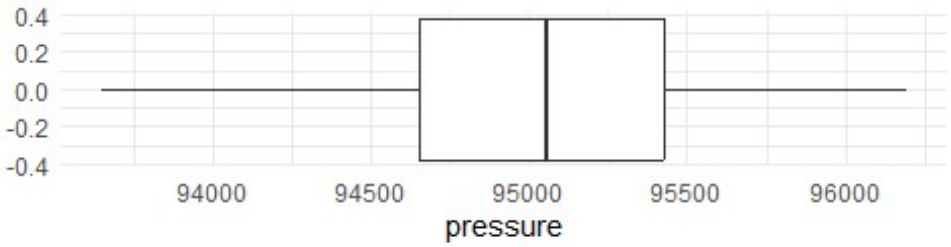

```
#Visualize physical properties
#histograms
Phist <- ggplot(df_2017_07_Air, aes(x = pressure)) + geom_histogram()
+theme_minimal()
Thist <-ggplot(df_2017_07_Air, aes(x = temperature)) + geom_histogram()
+theme_minimal()
Hhist <- ggplot(df_2017_07_Air, aes(x = humidity)) + geom_histogram()
+theme_minimal()
grid.arrange(Phist, Thist, Hhist, ncol = 3)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

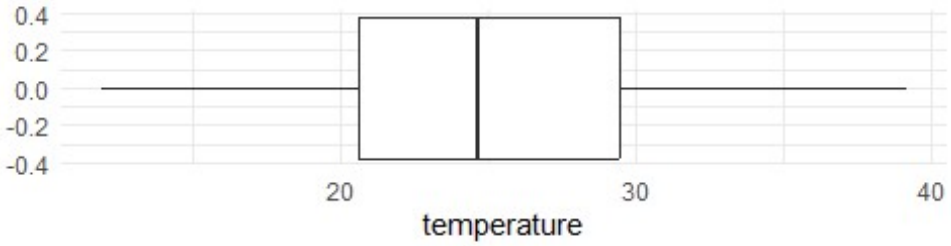


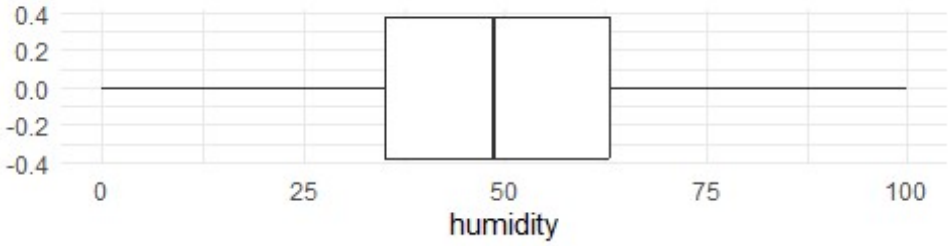

```r
#boxplots
Pbox <- ggplot(df_2017_07_Air, aes(x = pressure)) + geom_boxplot()
+theme_minimal()
Tbox <- ggplot(df_2017_07_Air, aes(x = temperature)) + geom_boxplot()
+theme_minimal()
Hbox <- ggplot(df_2017_07_Air, aes(x = humidity)) + geom_boxplot()
+theme_minimal()
grid.arrange(Pbox, Tbox, Hbox, nrow = 3, ncol = 1)
```




```r
#Treatment Plan
#Remove pressure and temperature outliers using the 1.5 iqr rule
Plower <- quantile(df_2017_07_Air$pressure, 0.25) -
IQR(df_2017_07_Air$pressure)
Pupper <- quantile(df_2017_07_Air$pressure, 0.75) +
IQR(df_2017_07_Air$pressure)
df_2017_07_Air <- filter(df_2017_07_Air, df_2017_07_Air$pressure > Plower)
df_2017_07_Air <- filter(df_2017_07_Air, df_2017_07_Air$pressure < Pupper)


Tlower <- quantile(df_2017_07_Air$temperature, 0.25) -
IQR(df_2017_07_Air$temperature)
Tupper <- quantile(df_2017_07_Air$temperature, 0.75) +
IQR(df_2017_07_Air$temperature)
df_2017_07_Air <- filter(df_2017_07_Air, df_2017_07_Air$temperature > Tlower)
df_2017_07_Air <-filter(df_2017_07_Air, df_2017_07_Air$temperature < Tupper)


#Visualize boxplots, no more negative and zero values on P and T
Pbox <- ggplot(df_2017_07_Air, aes(x = pressure)) + geom_boxplot()
+theme_minimal()
Tbox <- ggplot(df_2017_07_Air, aes(x = temperature)) + geom_boxplot()
+theme_minimal()
Hbox <- ggplot(df_2017_07_Air, aes(x = humidity)) + geom_boxplot()
+theme_minimal()
grid.arrange(Pbox, Tbox, Hbox, nrow = 3, ncol = 1)
```








```r
#convert humidity zero values to NA values and impute with sample humidity
mean
df_2017_07_Air$humidity <- ifelse(df_2017_07_Air$humidity == 0,
mean(df_2017_07_Air$humidity), df_2017_07_Air$humidity)

#check for missing values
colSums(is.na(df_2017_07_Air))
```

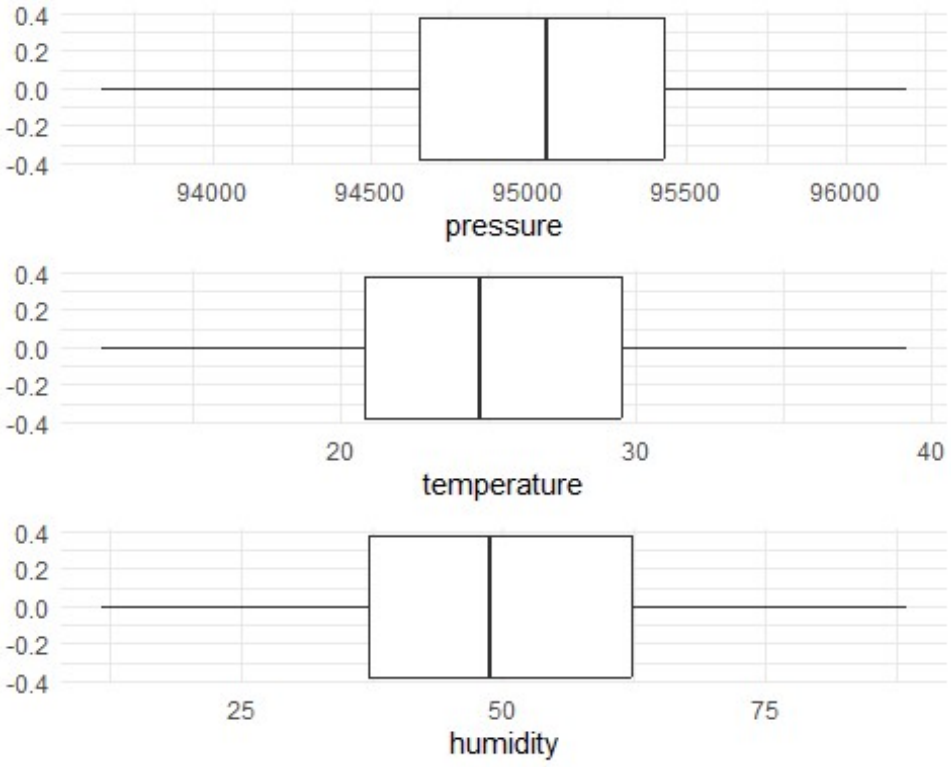
```
##   sensor_id    location         lat         lon   timestamp    pressure
##           0           0           0           0           0           0
## temperature    humidity
##           0           0
```

```r
#check for 0 values
sum(df_2017_07_Air$humidity == 0)
```

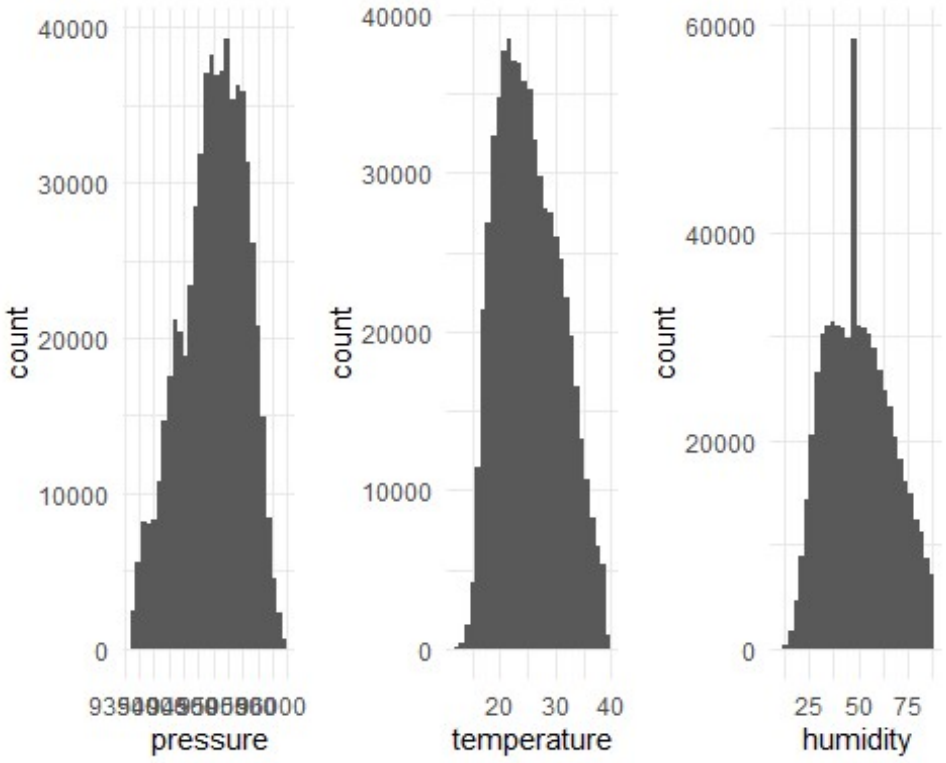
```
## [1] 0
```

```r
#Apply the IQR Rule
Hlower <- quantile(df_2017_07_Air$humidity, 0.25) -
IQR(df_2017_07_Air$humidity)
Hupper <- quantile(df_2017_07_Air$humidity, 0.75) +
IQR(df_2017_07_Air$humidity)
df_2017_07_Air <- filter(df_2017_07_Air, df_2017_07_Air$humidity > Hlower)
df_2017_07_Air <- filter(df_2017_07_Air, df_2017_07_Air$humidity < Hupper)

#Final visualization of boxplots, no more negative and zero values on P and T
```

```
Pbox <- ggplot(df_2017_07_Air, aes(x = pressure)) + geom_boxplot()
+theme_minimal()
Tbox <- ggplot(df_2017_07_Air, aes(x = temperature)) + geom_boxplot()
+theme_minimal()
Hbox <- ggplot(df_2017_07_Air, aes(x = humidity)) + geom_boxplot()
+theme_minimal()
grid.arrange(Pbox, Tbox, Hbox, nrow = 3, ncol = 1)
```




```
#Final visualization of histograms, sharp peak at humidity mean due to
imputation
Phist <- ggplot(df_2017_07_Air, aes(x = pressure)) + geom_histogram()
+theme_minimal()
Thist <-ggplot(df_2017_07_Air, aes(x = temperature)) + geom_histogram()
+theme_minimal()
Hhist <- ggplot(df_2017_07_Air, aes(x = humidity)) + geom_histogram()
+theme_minimal()
grid.arrange(Phist, Thist, Hhist, ncol = 3)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


40000
30000
20000
10000
0
count
pressure
20
30
40
temperature
60000
40000
20000
25
50
75
humidity