R Notebook

Outliers

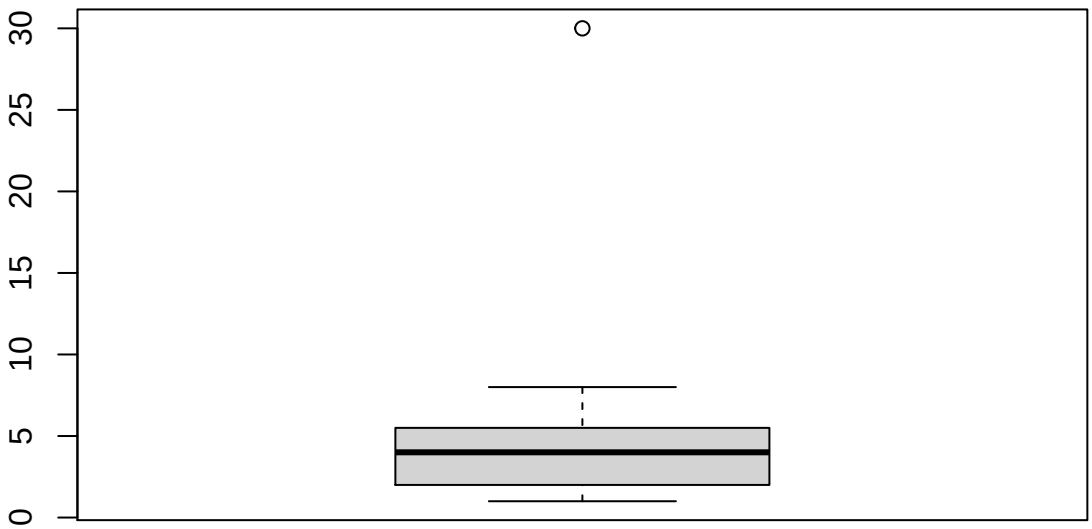
1. Screening for Outliers

In the process of producing, collecting, processing and analyzing data, outliers can come from many sources and hide in many dimensions. An outlier is an observation that is numerically distant from the rest of the data. When reviewing a boxplot, an outlier is defined as a data point that is located outside the fences ("whiskers") of the boxplot.

```
## Example

# Let's create the vector A

A <- c(3, 2, 5, 6, 4, 8, 1, 2, 30, 2, 4)

# then print it out
A
```

```
##  [1]  3  2  5  6  4  8  1  2 30  2  4
```

```
# We then plot a boxplot to help us visualise any existing outliers

boxplot(A)
```




```
# Then use the function boxplot.stats which lists the outliers in the vectors

boxplot.stats(A)$out
```

```
## [1] 30
```

Outliers should be investigated carefully. Often they contain valuable information about the process under investigation or the data gathering and recording process. Before considering the possible elimination of these points from the data, one should try to understand why they appeared and whether it is likely similar values will continue to appear. Of course, outliers are often bad data points

2. Obvious Inconsistencies

An obvious inconsistency occurs when a record contains a value or combination of values that cannot correspond to a real-world situation. For example, a person's age cannot be negative, a man cannot be pregnant and an under-aged person cannot possess a drivers license.

```
## Example
# Say from our vector x above, values above 20 are obvious inconsistencies then we using logical indice

non_greater_than_20 <- A > 20

# printing out non_greater_than_20
non_greater_than_20
```

```
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
```

CHALLENGE

```
# Question: Use the given bus dataset below, determine whether there are any obvious inconsistencies

url = "http://bit.ly/BusNairobiWesternTransport"

# Importing our database
library(data.table) # load package
```

```
## Warning: package 'data.table' was built under R version 4.0.4
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.0.4
```

```
## Warning: package 'tibble' was built under R version 4.0.4
```

```
## Warning: package 'tidyr' was built under R version 4.0.4
```

```
## Warning: package 'readr' was built under R version 4.0.4
```

```
## Warning: package 'purrr' was built under R version 4.0.4
```

```
## Warning: package 'dplyr' was built under R version 4.0.4
```

```
## Warning: package 'stringr' was built under R version 4.0.4
```

```
## Warning: package 'forcats' was built under R version 4.0.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::between()   masks data.table::between()
## x dplyr::filter()    masks stats::filter()
## x dplyr::first()     masks data.table::first()
## x dplyr::lag()       masks stats::lag()
## x dplyr::last()      masks data.table::last()
## x purrr::transpose() masks data.table::transpose()
```

```r
bus_dataset <- fread('http://bit.ly/BusNairobiWesternTransport')

# Previewing the dataset
# View(bus_dataset)
str(bus_dataset)
```

```
## Classes 'data.table' and 'data.frame':   51645 obs. of  10 variables:
##  $ ride_id        : int  1442 5437 5710 5777 5778 5777 5777 5778 5778 5781 ...
##  $ seat_number    : chr  "15A" "14A" "8B" "19A" ...
##  $ payment_method : chr  "Mpesa" "Mpesa" "Mpesa" "Mpesa" ...
##  $ payment_receipt: chr  "UZUEHCBUSO" "TIHLBUSGTE" "EQX8Q5G19O" "SGP18CL0ME" ...
##  $ travel_date    : IDate, format: "0017-10-17" "0019-11-17" ...
##  $ travel_time    : chr  "7:15" "7:12" "7:05" "7:10" ...
##  $ travel_from    : chr  "Migori" "Migori" "Keroka" "Homa Bay" ...
##  $ travel_to      : chr  "Nairobi" "Nairobi" "Nairobi" "Nairobi" ...
##  $ car_type       : chr  "Bus" "Bus" "Bus" "Bus" ...
##  $ max_capacity   : int  49 49 49 49 49 49 49 49 49 49 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

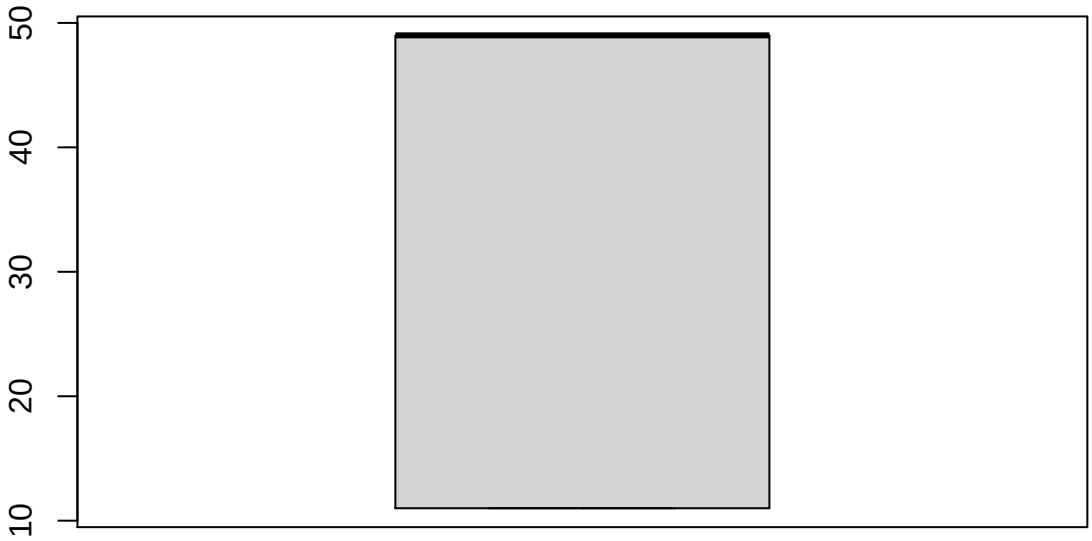
```r
dim(bus_dataset)
```

```
## [1] 51645    10
```

```r
class(bus_dataset)
```

```
## [1] "data.table" "data.frame"
```

```r
# Identifying the numeric class in the data and evaluating if there are any outliers

boxplot(bus_dataset$max_capacity)
```




```
boxplot.stats(bus_dataset$max_capacity)$out
```

```
## integer(0)
```

- The numerical column (max_capacity) contains no outliers