

```
# Outliers

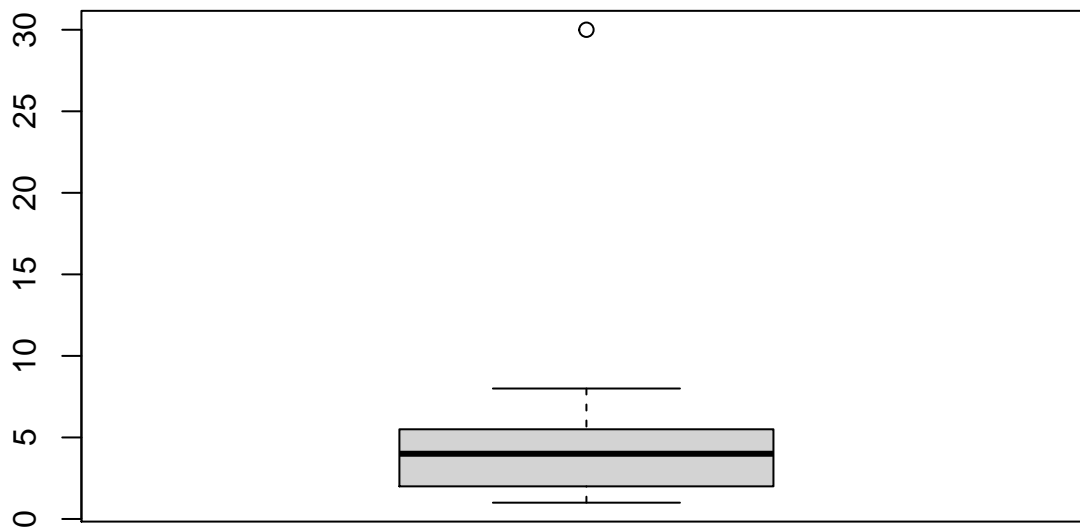
## 1. Screening for Outliers

## 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
```

2. Obvious Inconsistencies

```
## Example
# ---
# Say from our vector x above, values above 20 are obvious inconsistencies
```

```
# then we using logical indices to check for
# ---
#
non_greater_than_20 <- A > 20
```

```
# printing out non_greater_than_20
non_greater_than_20
```

```
## [1] 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
# ---
# Dataset url = http://bit.ly/BusNairobiWesternTransport
# ---
# OUR CODE GOES BELOW
#
# Importing our database
# ---
#
# install.packages("data.table") # install package data.table to work with data tables
library(data.table) # load package
#install.packages("tidyverse") # install packages to work with data frame - extends into visualization
library(tidyverse)
```

```
## -- 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
```

```
## -- 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()
```

```
bus_dataset <- fread('https://raw.githubusercontent.com/cimplival/datasets/master/buses-western-Nairobi')
```

```
# 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 5781 ...
## $ seat_number : chr "15A" "14A" "8B" "19A" ...
## $ payment_method : chr "Mpesa" "Mpesa" "Mpesa" "Mpesa" ...
## $ payment_receipt: chr "UZUEHCBUS0" "TIHLBUSGTE" "EQX8Q5G190" "SGP18CLOME" ...
## $ 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>
```

```
dim(bus_dataset)
```

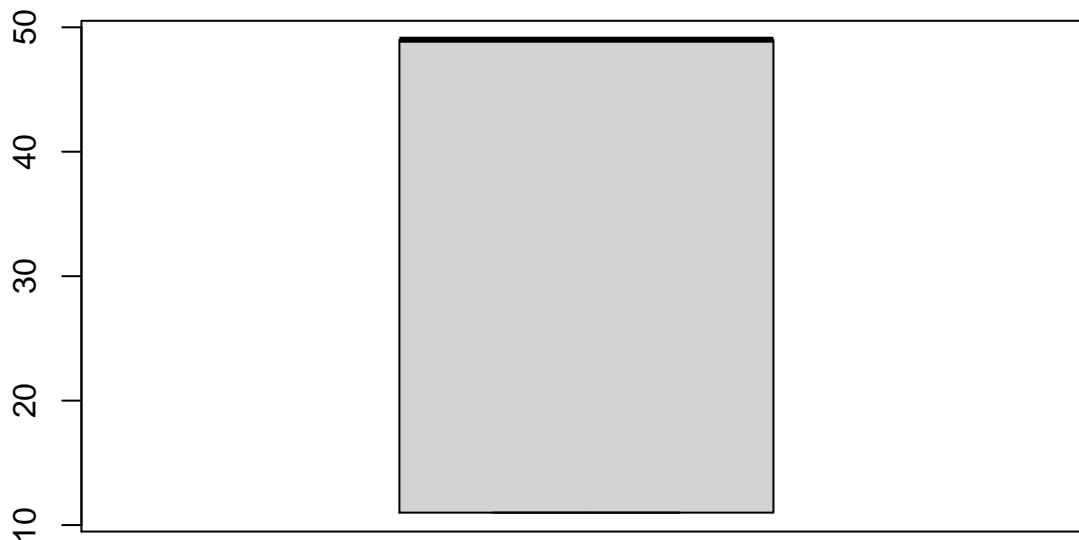
```
## [1] 51645 10
```

```
class(bus_dataset)
```

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

```
# Identifying the numeric class in the data and evaluating if there are any outliers
# ---
# OUR CODE GOES BELOW
#
```

```
boxplot(bus_dataset$max_capacity)
```



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

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