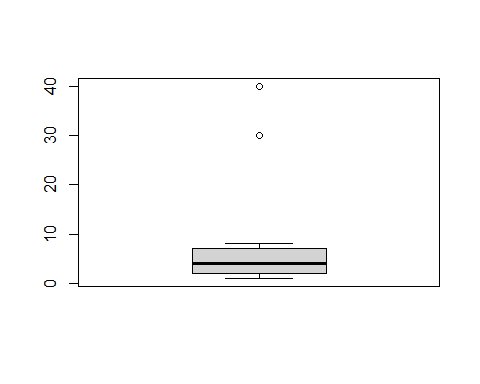
# Outliers

## 1. Screening for Outliers

## Example   
# ---  
# Let's create the vector A  
# ---  
#   
  
A <- c(3, 2, 5, 6, 4, 8, 1, 2, 30, 2, 40)  
  
# then print it out   
A

## [1] 3 2 5 6 4 8 1 2 30 2 40

# 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 40

## 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 FALSE TRUE FALSE TRUE

# 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 package data.table to work with data tables  
library(data.table) # load package  
# 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('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>

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  
# ---  
  
#  
boxplot(bus\_dataset$max\_capacity)

