Assign_part2a_1701433C

Tan Wei Ping

20 December 2018

Part 2a

R Markdown - Qn_1

1a) Examine the first line of the CSV file.

```
# read the data
data_with_na <- read.csv("C:/Users/Wei Ping/Documents/SEM 2.2/DMTR/Assignment/dmtr_assign_part2a.csv")

# display first line of data
head(data_with_na,1)
```

```
##           id date_account_created timestamp_first_active
## 1 gdka1q5ktd            1/10/2010               2.01e+13
##   date_first_booking gender age signup_method signup_flow language
## 1          1/10/2010 FEMALE  29         basic           0       en
##   affiliate_channel affiliate_provider first_affiliate_tracked signup_app
## 1            direct             direct                untracked        Web
##   first_device_type first_browser country_destination
## 1       Mac Desktop        Chrome                  FR
```

Observations: Based on the output of the first line of the csv file, there are 16 attributes/variables. All the attribute and its values seems to have a variety of data types (numeric, integer, etc.), with different level of measurements (*nominal* - 'id' ,*ordinal*,*interval* - 'timestamp_first_active',*ratio* - 'age' ,*binary* - 'gender').

One of the attribute named as 'date_account_created' consists of data like 1/10/2010 which are known as *date* types that can be can be categorised as a factor variable. Another attribute named as 'timestamp_first_active' consists of data like 20100000000000, which is a numeric variable.

From these 2 attributes, it can be interpreted that this data set consists of transactional data for the users who made a signup for the unknown website (with clues from attributes like 'signup_app'/'first_browser'.)

1b) Examine the name, data type for each variable

```
#name of each variable
names(data_with_na)
```

```
##  [1] "id"                     "date_account_created"
##  [3] "timestamp_first_active" "date_first_booking"
##  [5] "gender"                 "age"
##  [7] "signup_method"          "signup_flow"
##  [9] "language"               "affiliate_channel"
## [11] "affiliate_provider"     "first_affiliate_tracked"
## [13] "signup_app"             "first_device_type"
## [15] "first_browser"          "country_destination"
```

```r
#data type of each variable
for (i in 1:length(data_with_na)){
  print(class(data_with_na[,i]))
}
```

```
## [1] "factor"
## [1] "factor"
## [1] "numeric"
## [1] "factor"
## [1] "factor"
## [1] "integer"
## [1] "factor"
## [1] "integer"
## [1] "factor"
## [1] "factor"
## [1] "factor"
## [1] "factor"
## [1] "factor"
## [1] "factor"
## [1] "factor"
## [1] "factor"
```

Observations: Among the 16 attributes/variables, there are 3 different data types - factor, numeric & integer. However most of the attribute/variable are 'factors'. This means that this dataset focusses a lot more on recording categorical than numeric variables.

1c) Examine the number of rows, columns of the dataset

```r
#number of rows
nrow(data_with_na)
```

```
## [1] 19813
```

```r
#number of columns
ncol(data_with_na)
```

```
## [1] 16
```

Observations: There are 19813 rows and 16 columns. This means that there are 16 variables/attributes and 19813 items.

2a) Data cleaning (to remove missing data)

```r
#replace blanks with NA
data_with_na <- read.csv("C:/Users/Wei Ping/Documents/SEM 2.2/DMTR/Assignment/dmtr_assign_part2a.csv",
                         header=T, na.strings=c("","NA"))

#print the results
head(data_with_na)
```

```
##           id date_account_created timestamp_first_active
## 1 gdka1q5ktd            1/10/2010               2.01e+13
```

```
## 2 qdubonn3uk            1/10/2010               2.01e+13
## 3 qsibmuz9sx            1/10/2010               2.01e+13
## 4 80f7dwscrn            1/11/2010               2.01e+13
## 5 jha93x042q            1/11/2010               2.01e+13
## 6 7i49vnuav6            1/11/2010               2.01e+13
##   date_first_booking    gender age signup_method signup_flow language
## 1          1/10/2010    FEMALE  29         basic           0       en
## 2          1/18/2010 -unknown-  NA         basic           0       en
## 3          1/11/2010      MALE  30         basic           0       en
## 4          1/11/2010 -unknown-  40         basic           0       en
## 5               <NA> -unknown-  NA         basic           0       en
## 6               <NA>    FEMALE  40         basic           0       en
##   affiliate_channel affiliate_provider first_affiliate_tracked signup_app
## 1            direct             direct               untracked        Web
## 2            direct             direct                    <NA>        Web
## 3            direct             direct                  linked        Web
## 4               seo             google               untracked        Web
## 5             other         craigslist               untracked        Web
## 6               seo             google               untracked        Web
##   first_device_type first_browser country_destination
## 1       Mac Desktop        Chrome                  FR
## 2     Other/Unknown     -unknown-                  US
## 3       Mac Desktop        Chrome                  US
## 4            iPhone     -unknown-                  US
## 5       Mac Desktop        Safari                 NDF
## 6       Mac Desktop       Firefox                 NDF
```

```
#check for any NA values
any(is.na(data_with_na))
```

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

```
#number of values with NA
sum(is.na(data_with_na))
```

```
## [1] 19602
```

```
#omit data with na
data_without_na<-na.omit(data_with_na)
head(data_without_na)
```

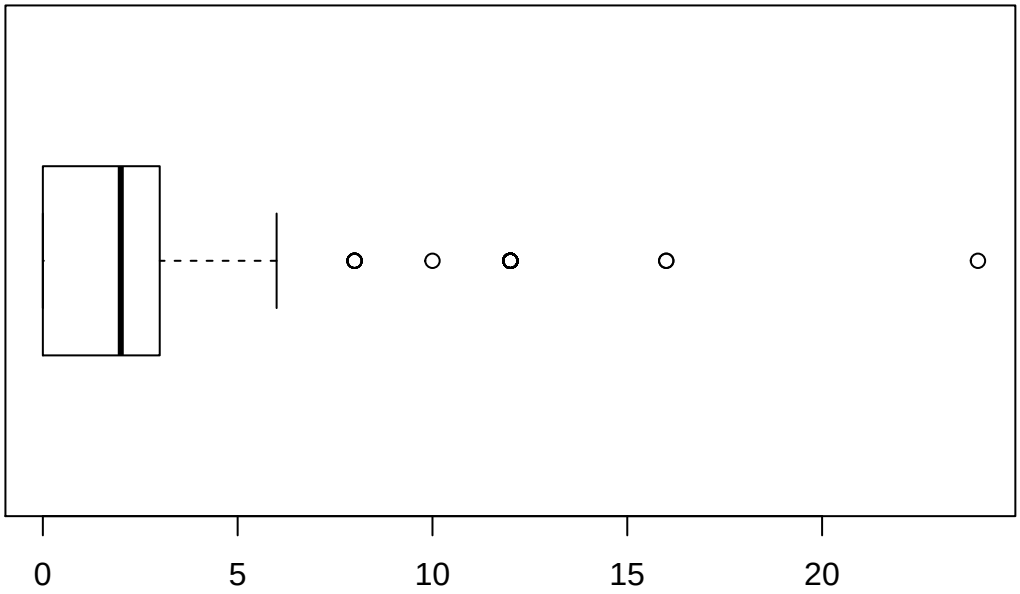
```
##           id date_account_created timestamp_first_active
## 1  gdka1q5ktd            1/10/2010               2.01e+13
## 3  qsibmuz9sx            1/10/2010               2.01e+13
## 4  80f7dwscrn            1/11/2010               2.01e+13
## 7  al8bcetz0g            1/12/2010               2.01e+13
## 9  hfrl5gle36            1/12/2010               2.01e+13
## 11 hql77nu2lk            1/13/2010               2.01e+13
##    date_first_booking    gender age signup_method signup_flow language
## 1           1/10/2010    FEMALE  29         basic           0       en
## 3           1/11/2010      MALE  30         basic           0       en
## 4           1/11/2010 -unknown-  40         basic           0       en
```

```
## 7           1/15/2010    FEMALE  26         basic           0       en
## 9           1/22/2010    FEMALE  32         basic           0       en
## 11          1/19/2010 -unknown-  37         basic           0       en
##    affiliate_channel affiliate_provider first_affiliate_tracked signup_app
## 1             direct             direct               untracked        Web
## 3             direct             direct                  linked        Web
## 4                seo             google               untracked        Web
## 7              other         craigslist               untracked        Web
## 9              other         craigslist               untracked        Web
## 11            direct             direct               untracked        Web
##    first_device_type first_browser country_destination
## 1        Mac Desktop        Chrome                  FR
## 3        Mac Desktop        Chrome                  US
## 4             iPhone     -unknown-                  US
## 7        Mac Desktop        Chrome                  FR
## 9     Desktop (Other)        Chrome                  US
## 11     Android Tablet     -unknown-                  US
```

Observations: There seems to be a lot of empty/blank cells in the dataset for attributes/variables like 'date_first_booking', 'age' and 'first_affiliate'. One way to clear away these empty/missing cells will be to replace them with 'NA' and later using codes like 'na.omit' to delete away these initial empty cells.

2b) Data cleaning (detect outliers) - **Numerical**

```
#deleting outlier for numeric variable - signup_flow
signup_flow1 <- data_without_na$signup_flow
boxplot(signup_flow1, horizontal=T)
```



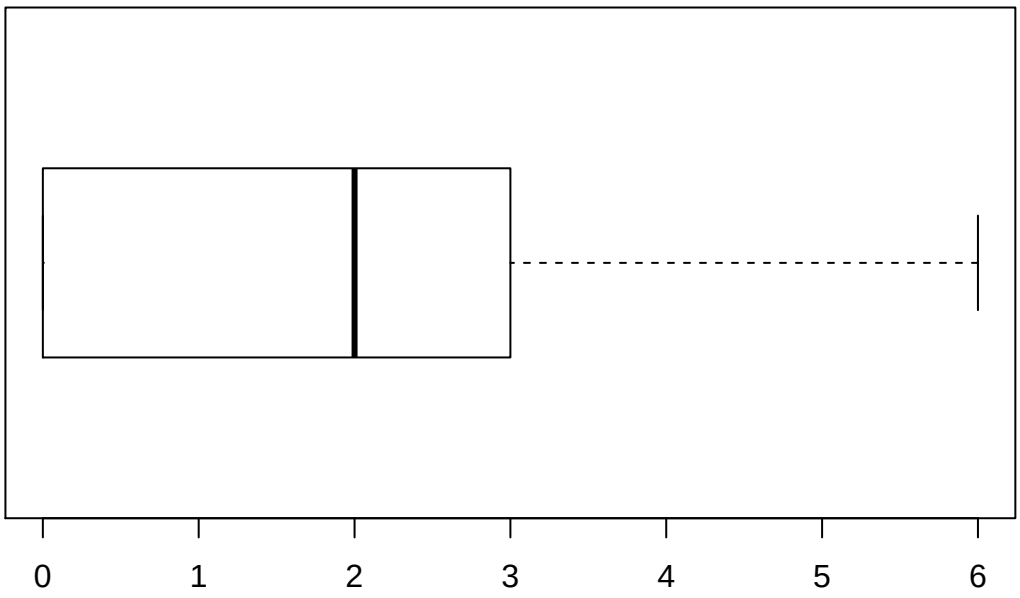
```r
summary(signup_flow1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   2.000   1.582   3.000  24.000
```

```r
#setting the benchmark to exclude outliers
bench_sf <- 3.00 + 1.5*IQR(signup_flow1)
bench_sf
```

```
## [1] 7.5
```

```r
#visualising the updated with boxplot without outliers
signup_flow1<-signup_flow1[signup_flow1<bench_sf]
boxplot(signup_flow1, horizontal = T)
```



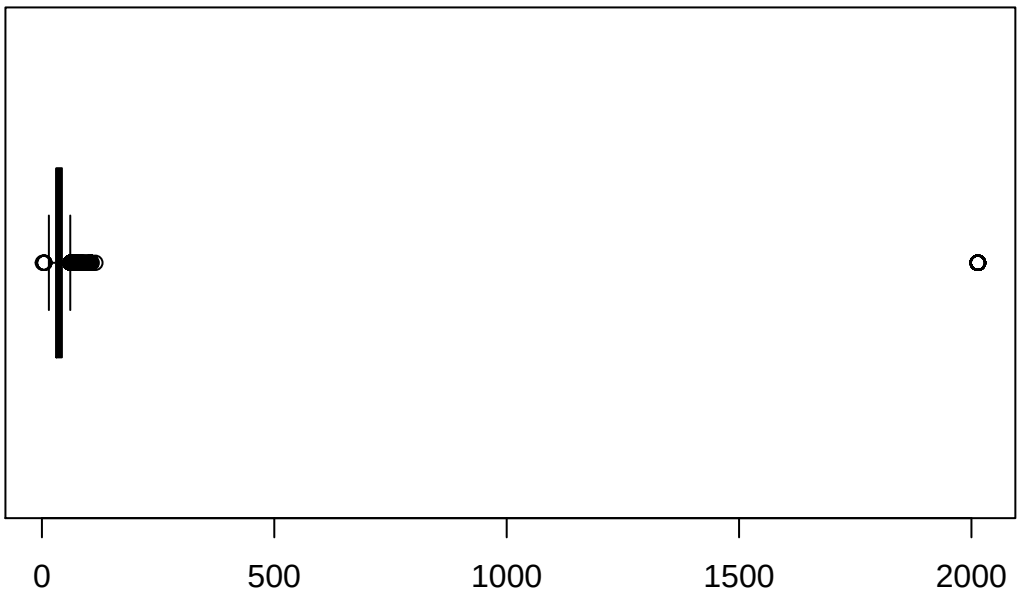

```r
#update the actual dataset the new values
data_without_na<-data_without_na[data_without_na$signup_flow <bench_sf, ]
head(data_without_na)
```

```
##           id date_account_created timestamp_first_active
## 1  gdka1q5ktd            1/10/2010               2.01e+13
## 3  qsibmuz9sx            1/10/2010               2.01e+13
## 4  80f7dwscrn            1/11/2010               2.01e+13
```

```
## 7  al8bcetz0g            1/12/2010              2.01e+13
## 9  hfrl5gle36            1/12/2010              2.01e+13
## 11 hql77nu2lk            1/13/2010              2.01e+13
##    date_first_booking    gender age signup_method signup_flow language
## 1           1/10/2010    FEMALE  29         basic           0       en
## 3           1/11/2010      MALE  30         basic           0       en
## 4           1/11/2010 -unknown-  40         basic           0       en
## 7           1/15/2010    FEMALE  26         basic           0       en
## 9           1/22/2010    FEMALE  32         basic           0       en
## 11          1/19/2010 -unknown-  37         basic           0       en
##    affiliate_channel affiliate_provider first_affiliate_tracked signup_app
## 1             direct             direct               untracked        Web
## 3             direct             direct                  linked        Web
## 4                seo             google               untracked        Web
## 7              other         craigslist               untracked        Web
## 9              other         craigslist               untracked        Web
## 11            direct             direct               untracked        Web
##    first_device_type first_browser country_destination
## 1        Mac Desktop        Chrome                  FR
## 3        Mac Desktop        Chrome                  US
## 4             iPhone     -unknown-                  US
## 7        Mac Desktop        Chrome                  FR
## 9     Desktop (Other)        Chrome                  US
## 11     Android Tablet     -unknown-                  US
```

Observations: In order to effectively visualise what are the outlier values, the attribute - signup_flow has its data visualised into a boxplot diagram. The formula 3rd_Quartile + 1.5*IQR(data) has also been used to set the benchmark for which data (known as outliers) must be excluded. Based on the benchmark value, for values that are more than 7.5, it must be excluded from the overall dataset.

```
#deleting outlier for numeric variable - age
age1 <- data_without_na$age
boxplot(age1, horizontal=T)
```




```
summary(age1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0    31.0    35.0    56.7    43.0  2014.0
```

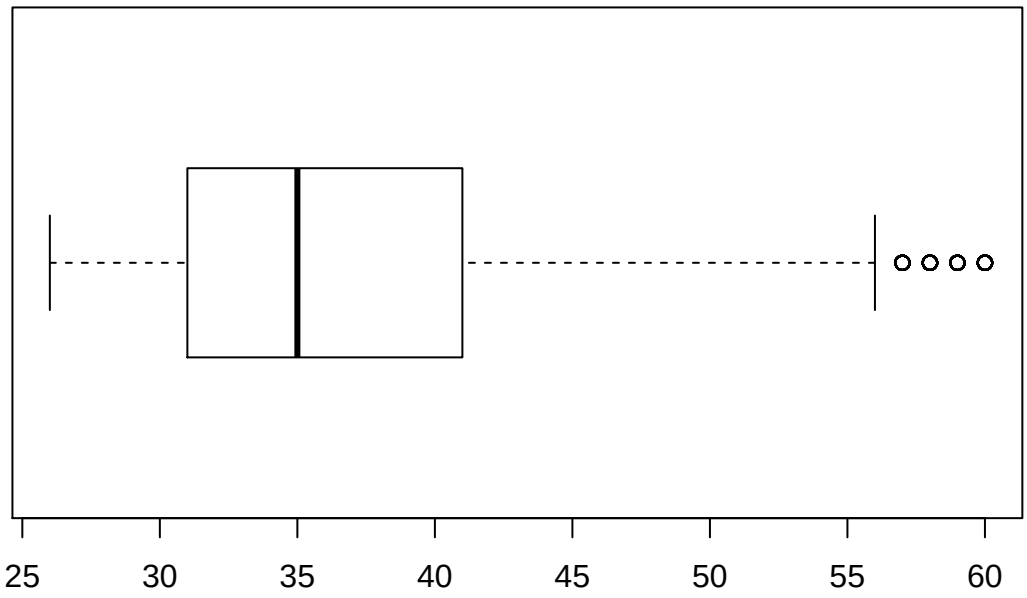
```
#setting the benchmark to exclude outliers
bench_ag <- 43.00 + 1.5*IQR(age1)
bench_ag
```

```
## [1] 61
```

```
bench_ag1 <- 43.00 - 1.5*IQR(age1)
bench_ag1
```

```
## [1] 25
```

```
#visualising the updated with boxplot without outliers
age1<-age1[age1 < bench_ag & age1 > bench_ag1]
boxplot(age1, horizontal = T)
```




```
#update the actual dataset the new values
data_without_na<-data_without_na[data_without_na$age<bench_ag & data_without_na$age>bench_ag1,]
head(data_without_na)
```

```
##           id date_account_created timestamp_first_active
## 1  gdka1q5ktd            1/10/2010               2.01e+13
## 3  qsibmuz9sx            1/10/2010               2.01e+13
## 4  80f7dwscrn            1/11/2010               2.01e+13
## 7  al8bcetz0g            1/12/2010               2.01e+13
## 9  hfrl5gle36            1/12/2010               2.01e+13
## 11 hql77nu2lk            1/13/2010               2.01e+13
##    date_first_booking    gender age signup_method signup_flow language
## 1           1/10/2010    FEMALE  29         basic           0       en
## 3           1/11/2010      MALE  30         basic           0       en
## 4           1/11/2010 -unknown-  40         basic           0       en
## 7           1/15/2010    FEMALE  26         basic           0       en
## 9           1/22/2010    FEMALE  32         basic           0       en
## 11          1/19/2010 -unknown-  37         basic           0       en
##    affiliate_channel affiliate_provider first_affiliate_tracked signup_app
## 1             direct             direct               untracked        Web
## 3             direct             direct                  linked        Web
## 4                seo             google               untracked        Web
## 7              other         craigslist               untracked        Web
## 9              other         craigslist               untracked        Web
## 11            direct             direct               untracked        Web
##    first_device_type first_browser country_destination
```

```
## 1        Mac Desktop        Chrome                  FR
## 3        Mac Desktop        Chrome                  US
## 4             iPhone     -unknown-                  US
## 7        Mac Desktop        Chrome                  FR
## 9    Desktop (Other)        Chrome                  US
## 11    Android Tablet     -unknown-                  US
```

Observations: In order to effectively visualise what are the outlier values, the attribute - age has its data visualised into a boxplot diagram. The formula 3rd_Quartile + 1.5xIQR(data) & 3rd_Quartile - 1.5xIQR(data) has also been used to set the benchmark for which data (known as outliers) must be excluded. Based on the benchmark value, for values that are more than 61 and less than 25, it must be excluded from the overall dataset.

Data cleaning (detect outliers) - Categorical

```
#delete values with the word '-unknown-' from gender
data_without_na<-data_without_na[!grepl("-unknown-", data_without_na$gender),]

#delete values with the word '-unknown-' from first_affiliate_tracked
data_without_na<-data_without_na[!grepl("-unknown-", data_without_na$first_affiliate_tracked),]

#delete values with the word '-unknown-' from first_browser
data_without_na<-data_without_na[!grepl("-unknown-", data_without_na$first_browser),]
head(data_without_na)
```

```
##           id date_account_created timestamp_first_active
## 1  gdka1q5ktd            1/10/2010               2.01e+13
## 3  qsibmuz9sx            1/10/2010               2.01e+13
## 7  al8bcetz0g            1/12/2010               2.01e+13
## 9  hfrl5gle36            1/12/2010               2.01e+13
## 18 7my0vrljxc            1/15/2010               2.01e+13
## 23 k15j7mbny0            1/19/2010               2.01e+13
##    date_first_booking gender age signup_method signup_flow language
## 1           1/10/2010 FEMALE  29         basic           0       en
## 3           1/11/2010   MALE  30         basic           0       en
## 7           1/15/2010 FEMALE  26         basic           0       en
## 9           1/22/2010 FEMALE  32         basic           0       en
## 18          1/15/2010 FEMALE  31         basic           0       en
## 23          1/21/2010 FEMALE  30         basic           0       en
##    affiliate_channel affiliate_provider first_affiliate_tracked signup_app
## 1             direct             direct               untracked        Web
## 3             direct             direct                  linked        Web
## 7              other         craigslist               untracked        Web
## 9              other         craigslist               untracked        Web
## 18            direct             direct                  linked        Web
## 23            direct             direct               untracked        Web
##    first_device_type first_browser country_destination
## 1        Mac Desktop        Chrome                  FR
## 3        Mac Desktop        Chrome                  US
## 7        Mac Desktop        Chrome                  FR
## 9    Desktop (Other)        Chrome                  US
## 18       Mac Desktop        Safari                  US
## 23       Mac Desktop        Chrome                  US
```

Observations: All 3 columns namely - gender, first_affiliate_tracked, first_browser contain the value '-unknown-' which has been deleted.

2c) Data cleaning (handle redundancy)

```
#identify duplicated data
which(duplicated(data_without_na))
```

```
## [1] 5014
```

```
#delete away duplicated data
data <- data_without_na[!duplicated(data_without_na),]
head(data)
```

```
##           id date_account_created timestamp_first_active
## 1  gdka1q5ktd            1/10/2010               2.01e+13
## 3  qsibmuz9sx            1/10/2010               2.01e+13
## 7  al8bcetz0g            1/12/2010               2.01e+13
## 9  hfrl5gle36            1/12/2010               2.01e+13
## 18 7my0vrljxc            1/15/2010               2.01e+13
## 23 k15j7mbny0            1/19/2010               2.01e+13
##    date_first_booking gender age signup_method signup_flow language
## 1           1/10/2010 FEMALE  29         basic           0       en
## 3           1/11/2010   MALE  30         basic           0       en
## 7           1/15/2010 FEMALE  26         basic           0       en
## 9           1/22/2010 FEMALE  32         basic           0       en
## 18          1/15/2010 FEMALE  31         basic           0       en
## 23          1/21/2010 FEMALE  30         basic           0       en
##    affiliate_channel affiliate_provider first_affiliate_tracked signup_app
## 1             direct             direct               untracked        Web
## 3             direct             direct                  linked        Web
## 7              other         craigslist               untracked        Web
## 9              other         craigslist               untracked        Web
## 18            direct             direct                  linked        Web
## 23            direct             direct               untracked        Web
##    first_device_type first_browser country_destination
## 1        Mac Desktop        Chrome                  FR
## 3        Mac Desktop        Chrome                  US
## 7        Mac Desktop        Chrome                  FR
## 9     Desktop (Other)        Chrome                  US
## 18       Mac Desktop        Safari                  US
## 23       Mac Desktop        Chrome                  US
```

Observations: All the data that were duplicated in the dataset has been deleted from the dataset.

3) Summarise the cleaned data set

```
#summarise cleaned data set
summary(data)
```

```
##           id       date_account_created timestamp_first_active
##  001xf4efvm:    1   2/22/2012:  33       Min.   :2.01e+13
```

```
##  006b76pgvn:   1   9/13/2011:  32       1st Qu.:2.01e+13
##  00pyv1alvj:   1   3/13/2012:  29       Median :2.01e+13
##  00xhnwrb5b:   1   9/22/2011:  28       Mean   :2.01e+13
##  01geg3we7v:   1   1/31/2012:  27       3rd Qu.:2.01e+13
##  01i8kuelur:   1   1/18/2012:  26       Max.   :2.01e+13
##  (Other)   :5007   (Other)  :4838
##  date_first_booking       gender          age         signup_method
##  3/13/2012:  29     -unknown-:   0   Min.   :26.00   basic   :3010
##  2/22/2012:  28     FEMALE   :2653   1st Qu.:31.00   facebook:2003
##  2/23/2012:  26     MALE     :2337   Median :35.00
##  3/14/2012:  26     OTHER    :  23   Mean   :37.04
##  1/31/2012:  25                      3rd Qu.:41.00
##  3/6/2012 :  24                      Max.   :60.00
##  (Other)  :4855
##   signup_flow       language        affiliate_channel  affiliate_provider
##  Min.   :0.000   en     :4915   direct       :3098    direct    :3048
##  1st Qu.:0.000   zh     :  21   other        : 685    google    :1104
##  Median :2.000   es     :  15   sem-non-brand: 572    craigslist: 539
##  Mean   :1.469   fr     :  15   seo          : 292    other     : 224
##  3rd Qu.:3.000   de     :  12   sem-brand    : 271    vast      :  46
##  Max.   :6.000   it     :   8   content      :  63    bing      :  30
##                  (Other):  27   (Other)      :  32    (Other)   :  22
##   first_affiliate_tracked signup_app         first_device_type
##  linked       :1161       iOS  :  31   Mac Desktop    :2946
##  local ops    :   0       Moweb:   0   Windows Desktop:1645
##  marketing    :   4       Web  :4982   iPad           : 252
##  omg          : 450                    iPhone         : 103
##  product      :   9                    Android Phone  :  28
##  tracked-other: 177                    Desktop (Other):  20
##  untracked    :3212                    (Other)        :  19
##         first_browser  country_destination
##  Chrome         :1861   US     :3524
##  Safari         :1194   other  : 461
##  Firefox        :1139   FR     : 336
##  IE             : 404   IT     : 150
##  Mobile Safari  : 346   GB     : 146
##  Android Browser:  22   ES     : 144
##  (Other)        :  47   (Other): 252
```

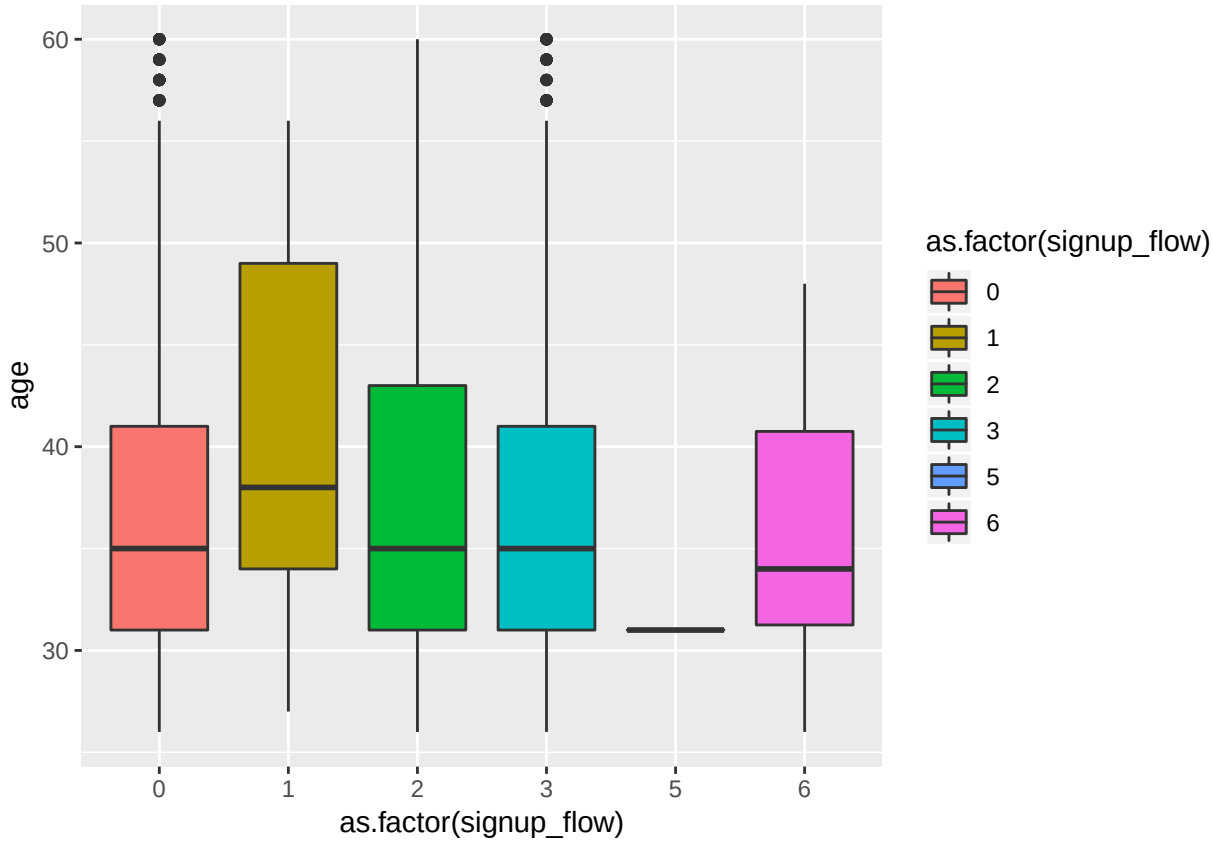
Observations: For the attributes/variables that are factors, the occurence of each data will be counted. However for attributes/variables that are integers, basic mathematical calculations such as 'min', 'Q1' (first quartile), 'Median', 'Q3' (third quartile), 'max' and 'mean' has been performed on the data.

4a) Draw graph for numeric variables

```
#Install ggplot
#install.packages("ggplot2")
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
#Plot the boxplot for signup_flow and age
ggplot(data, aes(x = as.factor(signup_flow), y = age, fill = as.factor(signup_flow))) +
  geom_boxplot()
```




```
#standard deviation for the different categories for signup_flow
sd(data$signup_flow==0)
```

```
## [1] 0.4919432
```

```
sd(data$signup_flow==1)
```

```
## [1] 0.07319875
```

```
sd(data$signup_flow==2)
```

```
## [1] 0.4665104
```

```
sd(data$signup_flow==3)
```

```
## [1] 0.4354197
```

```
sd(data$signup_flow==5)
```

```
## [1] 0.01412379
```

```r
sd(data$signup_flow==6)
```

```
## [1] 0.09938086
```

Note: At first glance, there seems to be some data cleaning issues with signup_flow - 0 and signup_flow - 5. However, they are not exactly classified as dirty data for which needs to be deleted.

Signup_flow = 0 Looking at the nature of signup_flow which is the steps that the users took to signup, there is no particular reason as to why this was recorded at 0 instead of an actual count of steps. Hence, if predictions are correct, those data with signup_flow - 0 is considered as missing data completely at random and the actual value for the signup_flow could be lost by chance. However, they will not be thus deleted and there are a couple of reasons to this.

1) Not all other fields of the same record (for signup_flow=0) has invalid data.
2) There are many data with signup_flow = 0, deleting them will thus affect the accuracy of analysis.

Signup_flow = 5 Reference to the dataset after data cleaning has been completed, there seems to only be a single data left in the dataset. Hence, with only a cell of data, a boxplot can't be formed but this still does not give any reason for the data to be deleted.

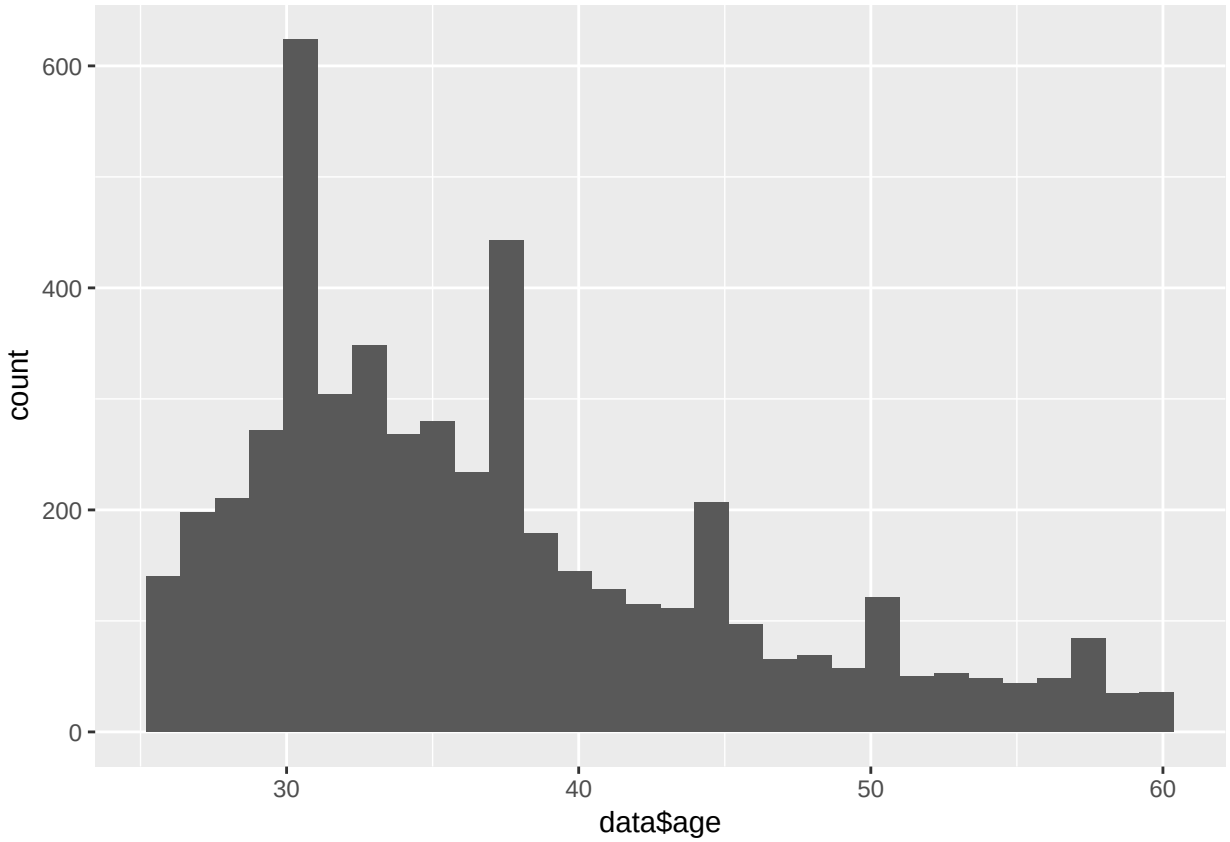
Observations: The above shows a cleaned boxplot for age against signup_flow and the relationship between both variables.

One comapison can be made from the signup_flow = 2 and signup_flow = 6 boxplot, where there exists some differences in observations. It can be seen that signup_flow=2 have data that are more widely spread out with a standard deviation of 0.4665104. Whereas, signup_flow=6 has a standard deviation of 0.09938086.

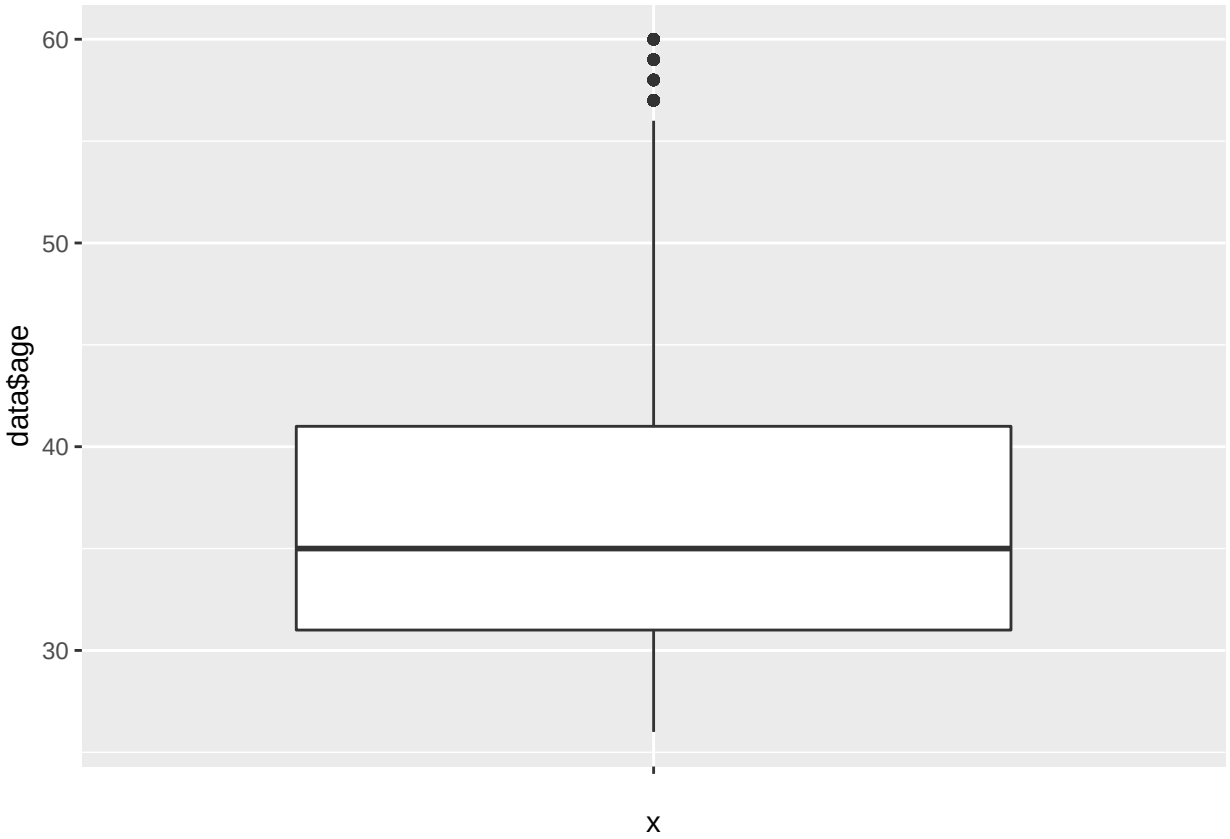
Thus, this means that signup_flow=2 has data that are more widely spread out while signup_flow=6 are more clustered together. In other words, there are more people with huge difference in age (from 26 to 60) who took 2 steps to signup while there are more people with similar ages who took 6 steps to signup (from age 26 to 48).Having said that, it is logically impossible for some people to take lesser steps to sign up than others. One possible reason for this would be the users left the website while signing up due to horrible web user interface and this has affected people from a variety of ages, not just older people, as can be seen from signup_flow=2. Hence, there could be initiatives made the owner of the website to improve the user interface.

```r
#plot histogram for age
ggplot(data=data, aes(data$age)) + geom_histogram()
```

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



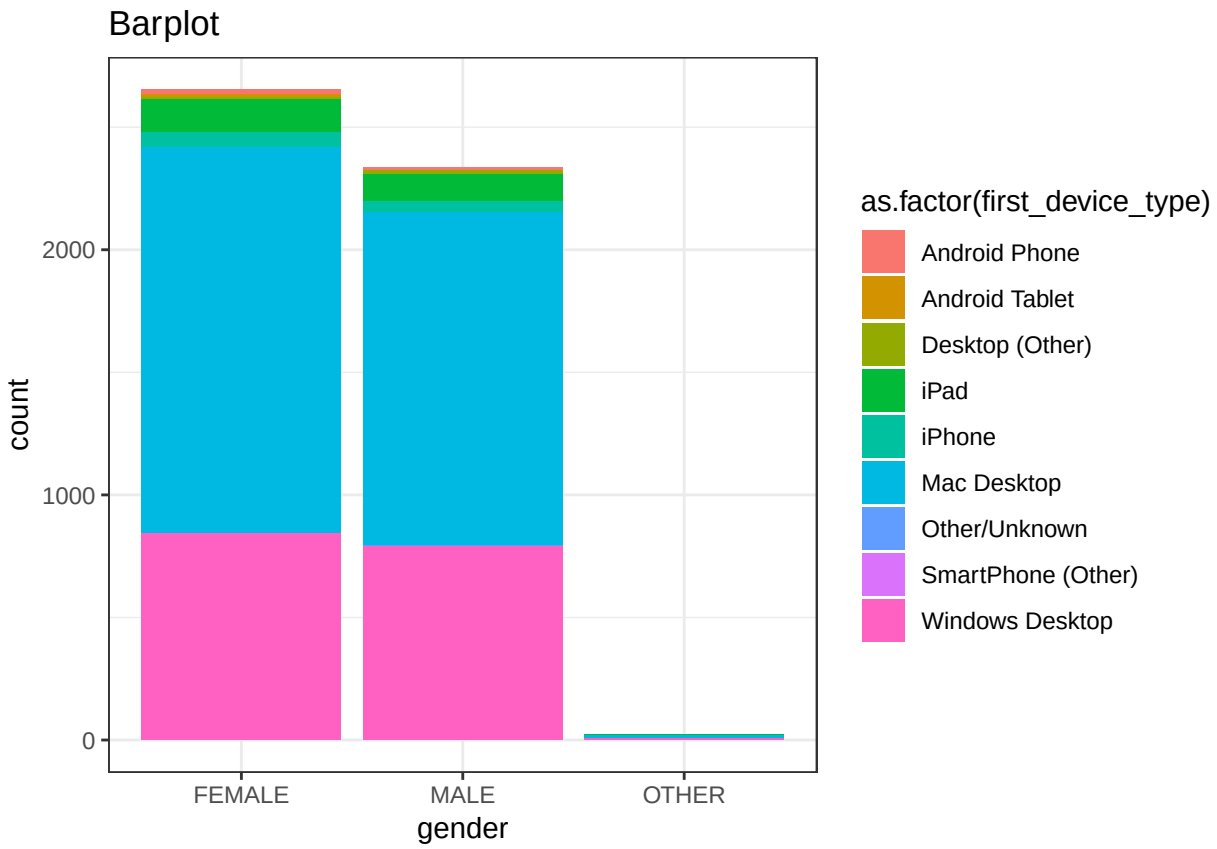

```
#plot barplot for age
ggplot(data = data, aes(x = "", y = data$age)) +
  geom_boxplot()
```



The above 2 graphs focusses on the variable age, using boxplot and histogram. The histogram basically tells more than the boxplot, showing more about the distribution of the different ages. From the histogram, we can tell that most users are clustered between age 30-40, with its peak at age 30.

4b) Draw graph for categorical variables

```
#visualise first_device_type data in a barplot
ggplot(data=data) + geom_bar(aes(x=gender,fill=as.factor(first_device_type))) +
  ggtitle(label="Barplot")+theme_bw()
```



Note: The gender 'other' has not been deleted previously as there can be a possibility that either the users are provided the choice of choosing 'other' as their gender or it is data that are missing at random/*missing completely at random*.

It could be missing data at random as perhaps either gender could find it less professional to say they use a Windows desktop as compared to Mac desktop. However, since this dataset has been identified to be transactional in nature, there is rare chances of users being allowed to choose what device they use but rather, this will recorded by the system itself (possibly).

This might also be an indicator of missing data (gender) completely at random. This means that the data that belongs in 'others' might belong to either gender instead and the actual gender that belong to 'others' might have been lost, which is a more valid prediction.

Hence, the 'other' gender will not be deleted as if so, the accuracy of the analysis will be affected.

Observations: From the above boxplot, it can be seen that there are more females than males who visited the website. For both genders, there seems to be a wide variety of device types used by the users. In particular, the device type that is most used is 'Mac Desktop'. Since this is the case, the website company can use this platform which the users prefers, for them to connect with the users.

For example, if more users uses the desktop, having any desktop-based campaign would increase the likelihood of them to receive any marketing messages etc. With that, the users can now be targetted at the correct platform for more business opportunities.