MVA_FacebookAnalysis.R

Soukhyada

Fri Feb 15 15:43:51 2019

```
#Soukhyada Vaidya
#Assignment 1: Facebook Ad Analysis
#Loading the data
dat <- read.csv("C:/Users/Soukhyada/Desktop/KAG_conversion_data.csv")
#View the data
View(dat)
#Displays the first few rows of the dataset
head(dat)
```

```
##    ad_id xyz_campaign_id fb_campaign_id   age gender interest Impressions
## 1 708746             916         103916 30-34      M       15        7350
## 2 708749             916         103917 30-34      M       16       17861
## 3 708771             916         103920 30-34      M       20         693
## 4 708815             916         103928 30-34      M       28        4259
## 5 708818             916         103928 30-34      M       28        4133
## 6 708820             916         103929 30-34      M       29        1915
##   Clicks Spent Total_Conversion Approved_Conversion click_rate conv_rate
## 1      1  1.43                2                   1          1    40-50%
## 2      2  1.82                2                   0          1     0-10%
## 3      0  0.00                1                   0          0     0-10%
## 4      1  1.25                1                   0          2     0-10%
## 5      1  1.29                1                   1          2   90-100%
## 6      0  0.00                1                   1          0   90-100%
##   loss_amount loss_rate
## 1        0.00     0-10%
## 2        1.82   90-100%
## 3        0.00     0-10%
## 4        1.25   90-100%
## 5        0.00     0-10%
## 6        0.00     0-10%
```

```
#Display the structure of the attributes
str(dat)
```

```
## 'data.frame':    1143 obs. of  15 variables:
##  $ ad_id              : int  708746 708749 708771 708815 708818 708820 708889 708895 708953 708958 ...
##  $ xyz_campaign_id    : int  916 916 916 916 916 916 916 916 916 916 ...
##  $ fb_campaign_id     : int  103916 103917 103920 103928 103928 103929 103940 103941 103951 103952 ...
##  $ age                : Factor w/ 4 levels "30-34","35-39",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ gender             : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ interest           : int  15 16 20 28 28 29 15 16 27 28 ...
##  $ Impressions        : int  7350 17861 693 4259 4133 1915 15615 10951 2355 9502 ...
##  $ Clicks             : int  1 2 0 1 1 0 3 1 1 3 ...
##  $ Spent              : num  1.43 1.82 0 1.25 1.29 ...
##  $ Total_Conversion   : int  2 2 1 1 1 1 1 1 1 1 ...
##  $ Approved_Conversion: int  1 0 0 0 1 1 0 1 0 0 ...
##  $ click_rate         : int  1 1 0 2 2 0 2 1 4 3 ...
##  $ conv_rate          : Factor w/ 9 levels "0-10%","10-20%",..: 5 1 1 1 9 9 1 9 1 1 ...
##  $ loss_amount        : num  0 1.82 0 1.25 0 ...
##  $ loss_rate          : Factor w/ 2 levels "0-10%","90-100%": 1 2 1 2 1 1 2 1 2 2 ...
```

```
#Loading required packages
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------- tidyverse 1
.2.1 --
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.3.0
```

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

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

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

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

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

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

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

```
## -- Conflicts ---------------------------------------------------------------------- tidyverse_conflic
ts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#Gives an overview of the data
glimpse(dat)
```

```
## Observations: 1,143
## Variables: 15
## $ ad_id               <int> 708746, 708749, 708771, 708815, 708818, 70...
## $ xyz_campaign_id     <int> 916, 916, 916, 916, 916, 916, 916, 916, 91...
## $ fb_campaign_id      <int> 103916, 103917, 103920, 103928, 103928, 10...
## $ age                 <fct> 30-34, 30-34, 30-34, 30-34, 30-34, 30-34, ...
## $ gender              <fct> M, M, M, M, M, M, M, M, M, M, M, M, M, M, ...
## $ interest            <int> 15, 16, 20, 28, 28, 29, 15, 16, 27, 28, 31...
## $ Impressions         <int> 7350, 17861, 693, 4259, 4133, 1915, 15615,...
## $ Clicks              <int> 1, 2, 0, 1, 1, 0, 3, 1, 1, 3, 0, 0, 0, 0, ...
## $ Spent               <dbl> 1.43, 1.82, 0.00, 1.25, 1.29, 0.00, 4.77, ...
## $ Total_Conversion    <int> 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ Approved_Conversion <int> 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, ...
## $ click_rate          <int> 1, 1, 0, 2, 2, 0, 2, 1, 4, 3, 0, 0, 0, 0, ...
## $ conv_rate           <fct> 40-50%, 0-10%, 0-10%, 0-10%, 90-100%, 90-1...
## $ loss_amount         <dbl> 0.000000, 1.820000, 0.000000, 1.250000, 0....
## $ loss_rate           <fct> 0-10%, 90-100%, 0-10%, 90-100%, 0-10%, 0-1...
```

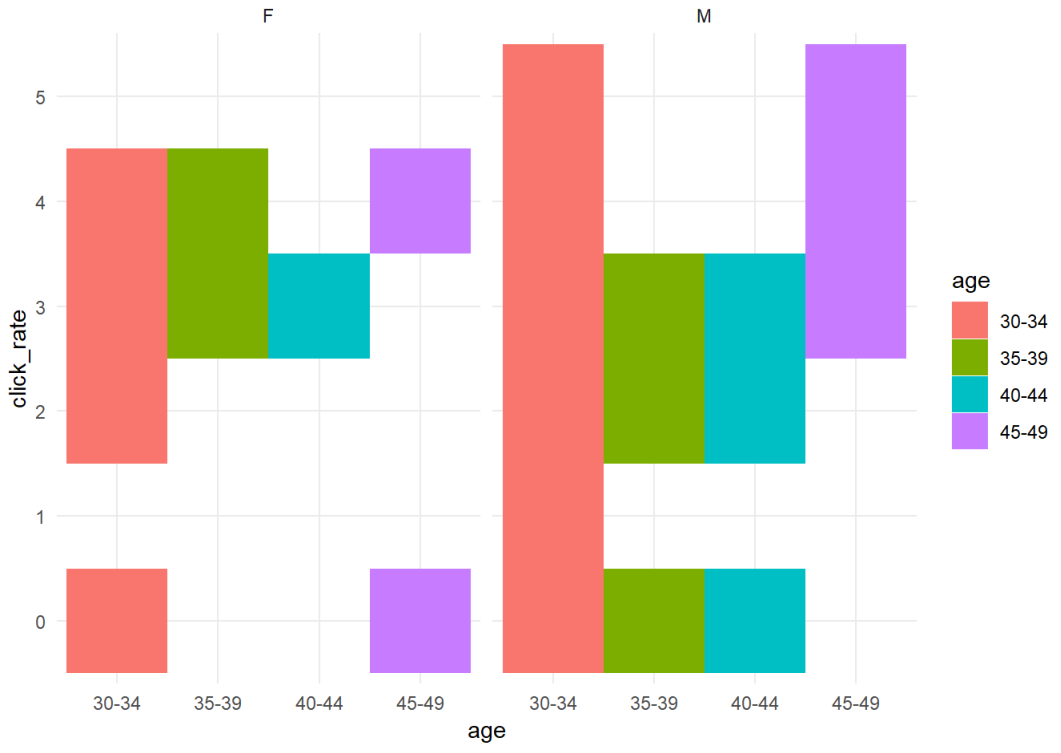
```
#Checking variable: age
unique(dat$age)
```

```
## [1] 30-34 35-39 40-44 45-49
## Levels: 30-34 35-39 40-44 45-49
```

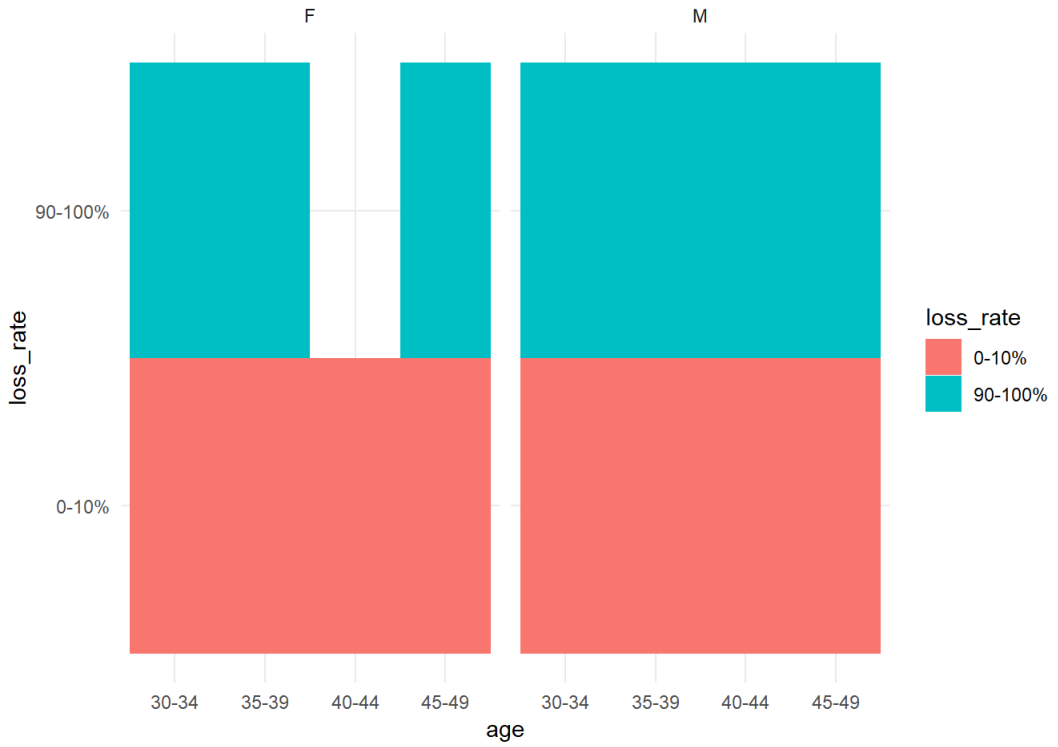
```r
#Adding New Columns for Analysis :
#Click Rate(click_rate)
#clicks per 10000 impressions
dat<- dat %>% mutate(click_rate = as.factor(ifelse(dat$Impressions != 0 , round(dat$Clicks/dat$Impressions*10000) , 0)))

#Conversion Rate(conv_rate)
#Approved conversions as a percentage of total conversions
dat<- dat %>% mutate(conv_rate = ifelse(dat$Total_Conversion != 0 , round(dat$Approved_Conversion/dat$Total_Conversion*100), 0))

#Conversion rate as factor
dat$conv_rate <-cut(dat$conv_rate, seq(0,100,10), right=TRUE, labels=c("0-10%","10-20%","20-30%","30-40%","40-50%","50-60%","60-70%","70-80%","80-90%","90-100%"))
dat$conv_rate[is.na(dat$conv_rate)] <- "0-10%"

#Loss amount
dat<- dat %>%  mutate(loss_amount = ifelse(dat$Clicks !=0, dat$Spent*(1 - dat$Approved_Conversion / dat$Clicks),0))

#Loss rate
dat<- dat %>% mutate(loss_rate =  ifelse(dat$Spent>0, round(dat$loss_amount/dat$Spent)*100,0))
dat$loss_rate <-cut(dat$loss_rate, seq(0,100,10), right=TRUE, labels=c("0-10%","10-20%","20-30%","30-40%","40-50%","50-60%","60-70%","70-80%","80-90%","90-100%"))
dat$loss_rate[is.na(dat$loss_rate)] <- "0-10%"

##Data Analysis
#Gives the details of xyz_campaign_id in tabular format
table(dat$xyz_campaign_id)
```

```
##
##  916  936 1178
##   54  464  625
```

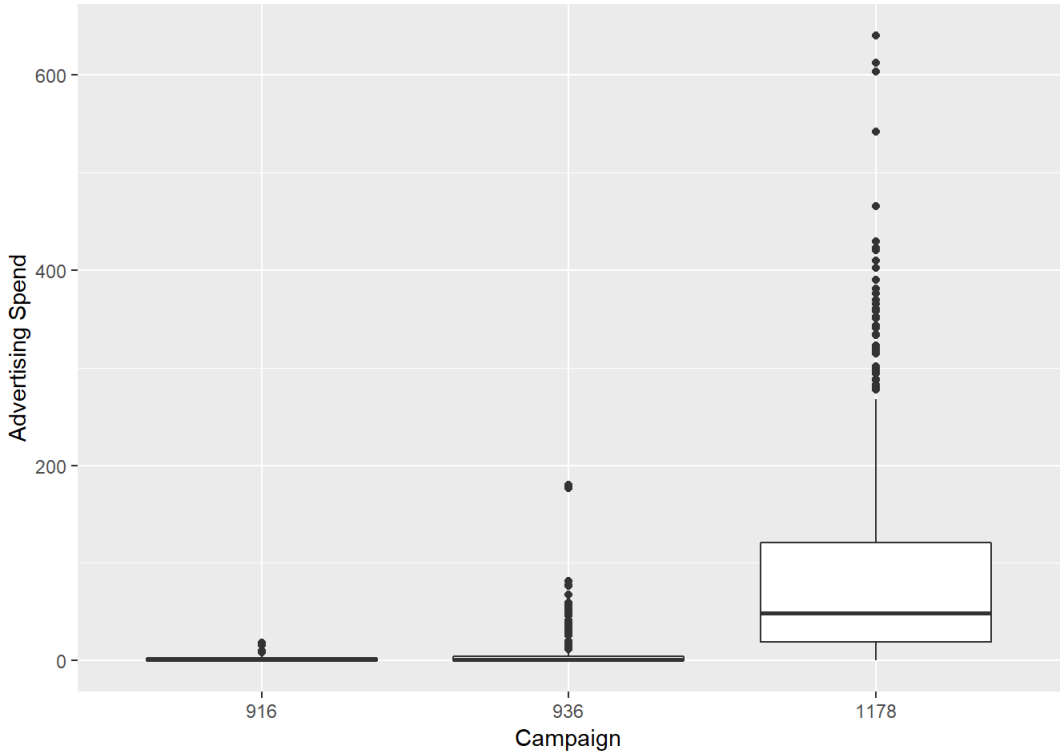
```r
#There are 3 companies and the analysis will be done seperately
dat_916<-dat %>% filter(xyz_campaign_id %in% c("916"))

#Loading package for plots.
library(ggplot2)

##Analysis done on ad campaign 916
ggplot(data = dat_916) +aes(x = age, y = click_rate, fill = age) +geom_tile() +theme_minimal() +facet_wrap(vars(gender))
```



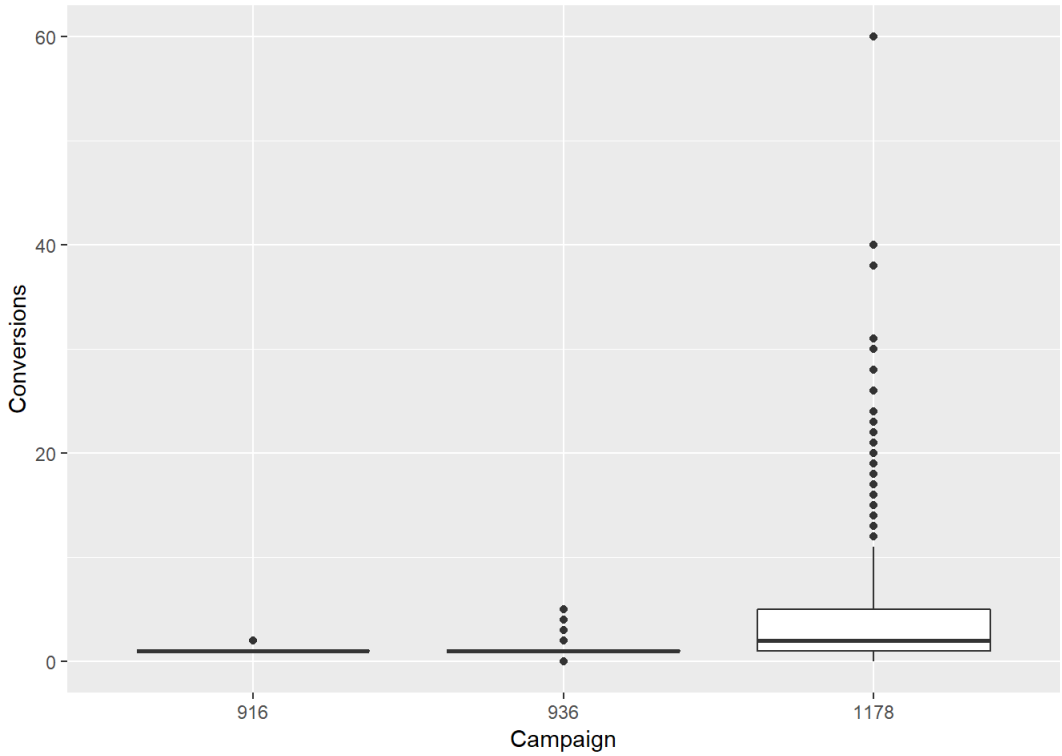
```
#Analysis:For both gender groups Male and Female the age group 30-34 seems to respond with higher click rate ranging from 2 to 5
#But we cannot say anything conclusive about other groups because we don't have sufficient data in these groups


# Analysis  using using "loss rate" column. It shows how many approved conversion a company gets as a return to spent money. If it's 0 , the return equals to spend. If it's 100%, all spend is a loss.
ggplot(data = dat_916) +aes(x = age, y = loss_rate, fill = loss_rate) +geom_tile() +theme_minimal() +facet_wrap(vars(gender))
```



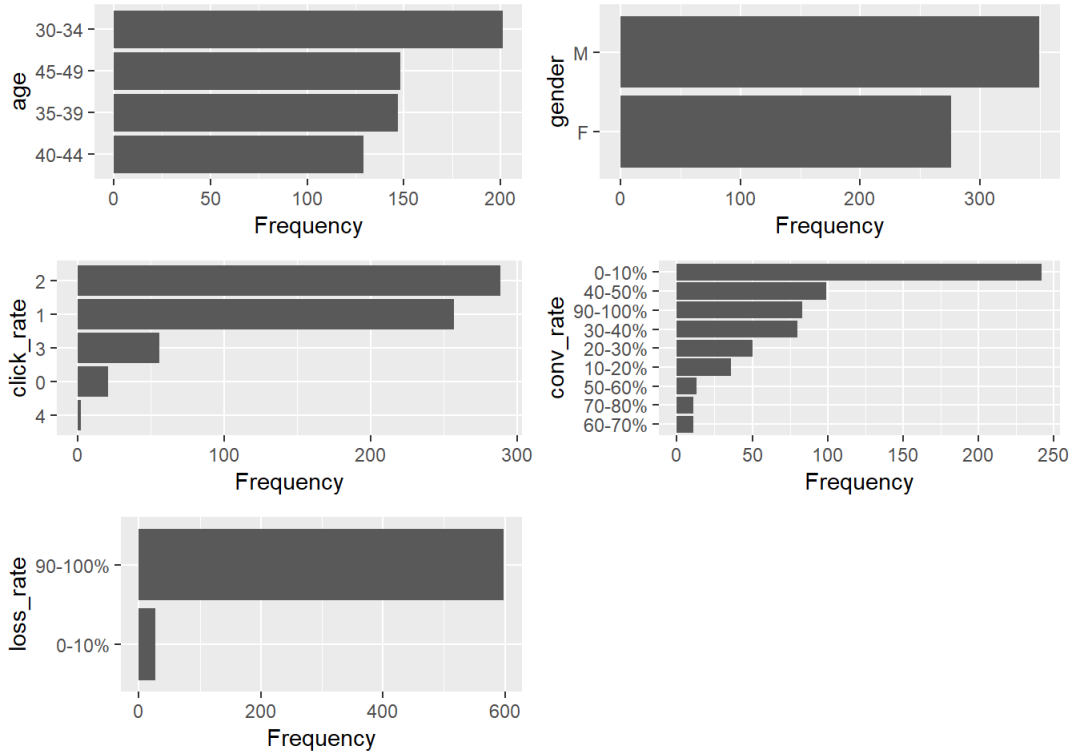

```
#Analysis:We cannot conclude anything from these plots because in all the age groups of both genders there is a 50% loss of 0-10


##Analysis on comparison of ad campaigns w.r.t Advertising Spend and Conversions
ggplot(dat, aes(as.factor(xyz_campaign_id), Spent)) + geom_boxplot() + labs(x = "Campaign", y = "Advertising Spend")
```



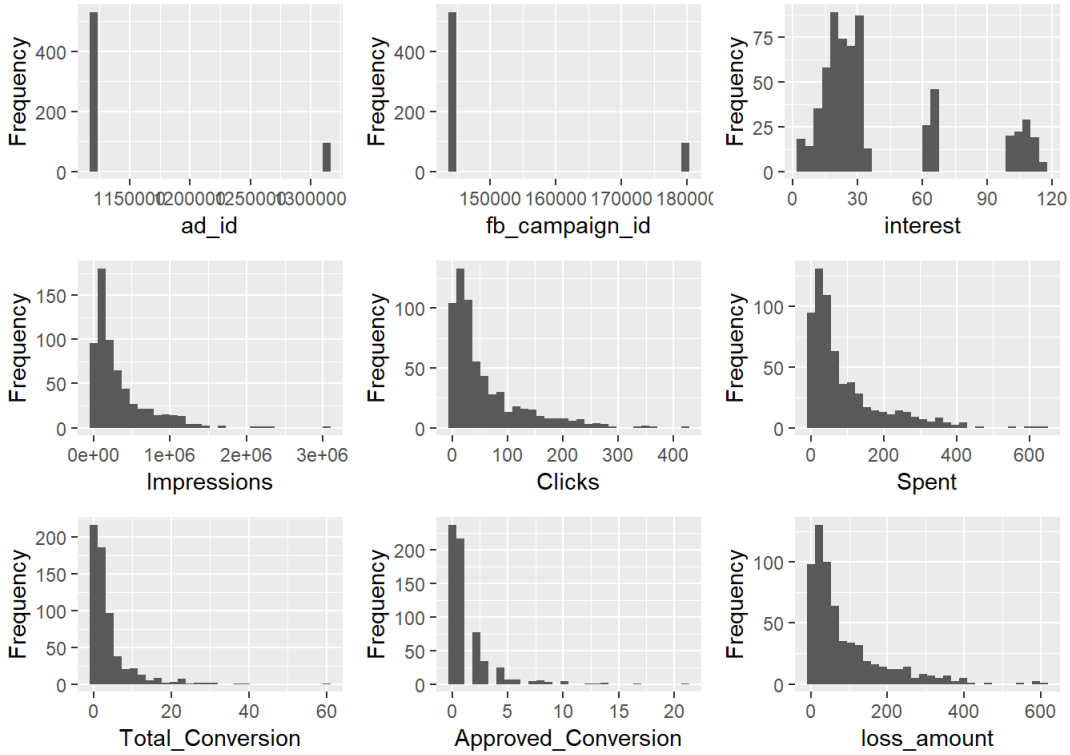

```
ggplot(dat, aes(as.factor(xyz_campaign_id), Total_Conversion)) + geom_boxplot() + labs(x = "Campaign", y = "
Conversions")
```



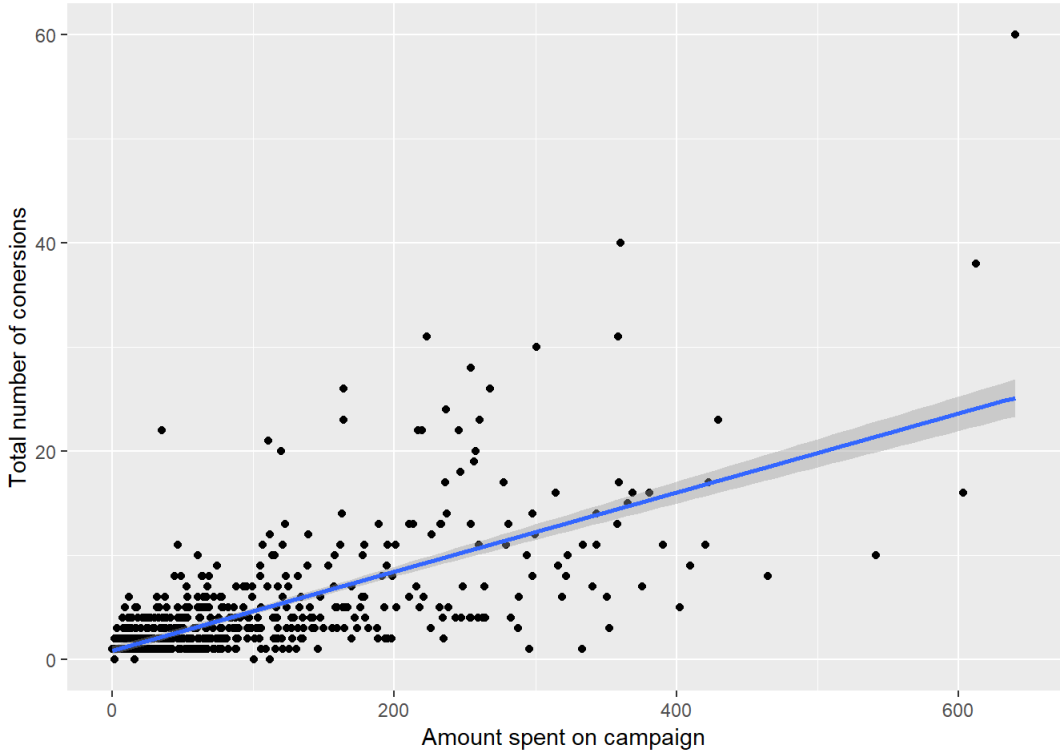
```
#From the above first box-plot we can see that company 1178 spends more on advertisements comparatively
#And the second box plot shows there is a more total conversion rate for company 1178

#Hence, from the above plots, let's do further analysis on ad campaign 1178

##Analysis done on ad campaign 1178
#Creating a new dataframe that just includes the data from that campaign.
dat1178 <- dat %>%filter(xyz_campaign_id == 1178)

#Loading package for plotting barplots and histograms
library(DataExplorer)

#Checking the data variable by variable
#Plotting the frequencies w.r.t each variable
options(repr.plot.width=4, repr.plot.height=4)
plot_bar(dat1178)
```



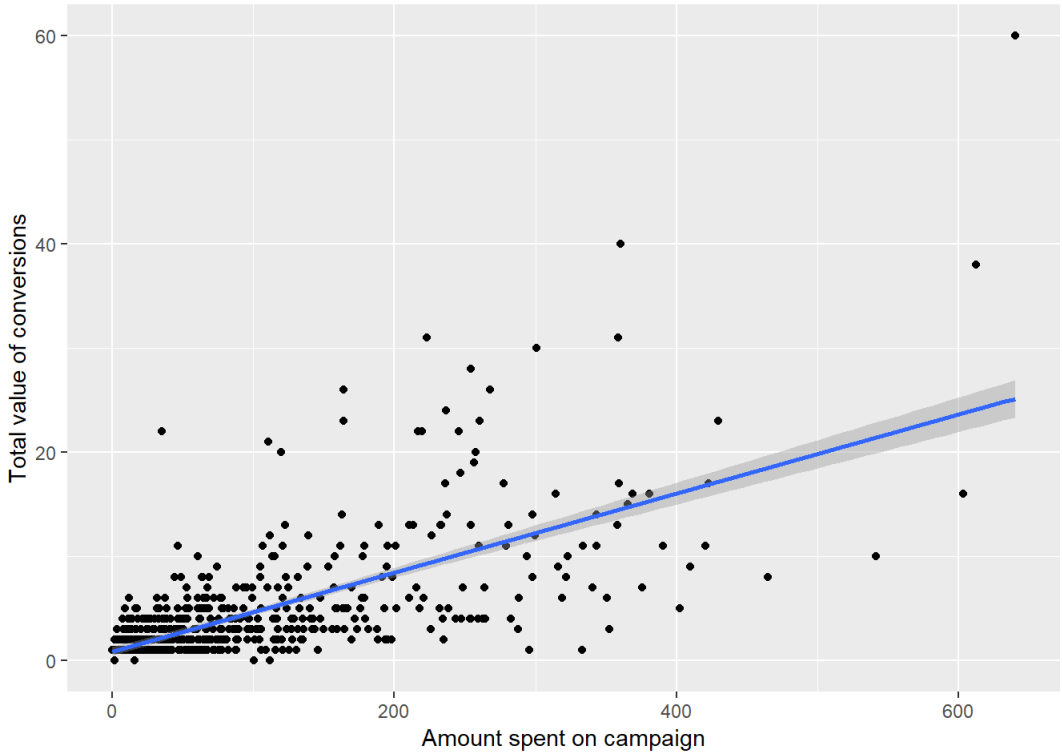

```
options(repr.plot.width=8, repr.plot.height=4)
plot_histogram(dat1178[-2])
```




```
#Preliminary analysis of campaign 1178
#Let's look at what happens to the number of conversions and the value of our conversions when we spend more money on our campaign.
options(repr.plot.width=6, repr.plot.height=3)
ggplot(dat1178, aes(Spent, Total_Conversion)) + geom_point() + geom_smooth(method = "lm") +
  labs(x = "Amount spent on campaign", y = "Total number of conersions")
```




```
ggplot(dat1178, aes(Spent, Total_Conversion)) + geom_point() + geom_smooth(method = "lm") +
  labs(x = "Amount spent on campaign", y = "Total value of conversions")
```




```
#Analysis:Here, it looks like the more we spend, the more we get back.
#But the amount of data is quite sparse at the right-hand side so we cannot say that the statement is accura
te without considering more analysis.

#Now, lets find correlation for 1178 Campaign
#Loading packages for correlation
library(caret)
```

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

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

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(corrplot)
```

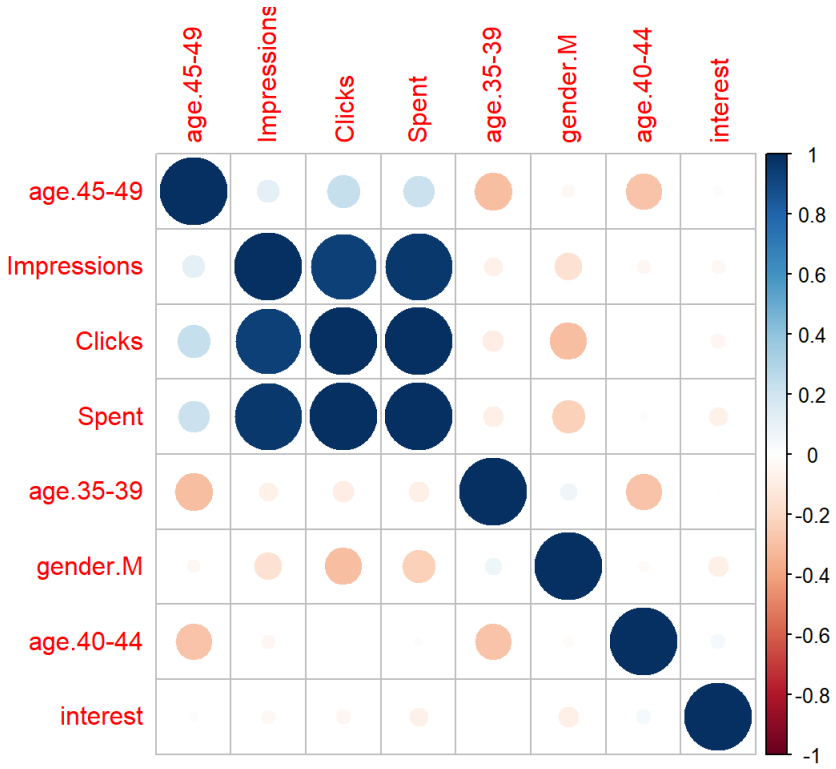
```
## Warning: package 'corrplot' was built under R version 3.5.2
```

```
## corrplot 0.84 loaded
```

```
ds_1178_predict<-dat1178 %>% select( age, gender,interest, Impressions,Clicks,Spent)
#Setting random number as seed
set.seed(1234)
ds_1178_predict_Data <- dummyVars("~.",data=ds_1178_predict, fullRank=T)
ds_1178_predict_final <- as.data.frame(predict(ds_1178_predict_Data,ds_1178_predict))
print(names(ds_1178_predict_final))
```

```
## [1] "age.35-39"   "age.40-44"   "age.45-49"   "gender.M"    "interest"
## [6] "Impressions" "Clicks"      "Spent"
```

```
corMatMy <- cor(ds_1178_predict_final)
corrplot(corMatMy, order = "hclust")
```




```
#Analysis: From the correlation matrix we can see that it's pretty obvious that the clicks are strongly corr
elated to impressions and spent.

#Conclusion: Data Analysis and Visualization done on different campaigns to find out the relation between mu
ltiple factors such as Click rate, conversion rate and loss rate.
```