Compstat FinalProject - Team15

2022-12-14

```
pacman::p_load(pacman,party,psych,rio,tidyverse,ggpubr)
#Reading the Dataset
df = read.csv('../../Downloads/Data/ads15.csv')

# adding new column "profit" to DataFrame
library(dplyr)
df <- df %>% mutate(profit = df$adrevenue - df$adcost)

df['ROI']<-df$profit/df$adcost

# split into DataFrames based on socialmedia platforms
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

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

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
### Splitting Age into 3 categories to evaluate it as a categorical variable
bin_age <- ceiling(log(max(df$age), 2)) + 1
df<-df %>%
  mutate(
    # Create categories
    age_group = dplyr::case_when(
      age <= 19            ~ "teen",
      age > 19 & age <= 35 ~ "youngadult",
      age > 35  ~             "oldadult",
    ),
    # Convert to factor
    age_group = factor(
      age_group,
      level = c("teen", "youngadult","oldadult")
    )
  )
fac_df <- df %>% filter(df$socialmedia=="Facebook")
Inst_df <- df %>% filter(df$socialmedia=="Instagram")
tk_df <- df %>% filter(df$socialmedia=="TikTok")
tw_df <- df %>% filter(df$socialmedia=="Twitter")
y_df <- df %>% filter(df$socialmedia=="YouTube")

#create relative frequency table
######Q1 a) ####
### We know that relative frequency means number of values of a particular category d
ivided by total number of values.
## there are 2 ways to this first using function to apply to all columns but that way
continuous value column also gets divided.
## Second, is to divide each categorical variable separately.
#approach 2

t1<- table(df$socialmedia)
t1
```

```
##
##  Facebook Instagram    TikTok   Twitter   YouTube
##        60        93       141        22       156
```

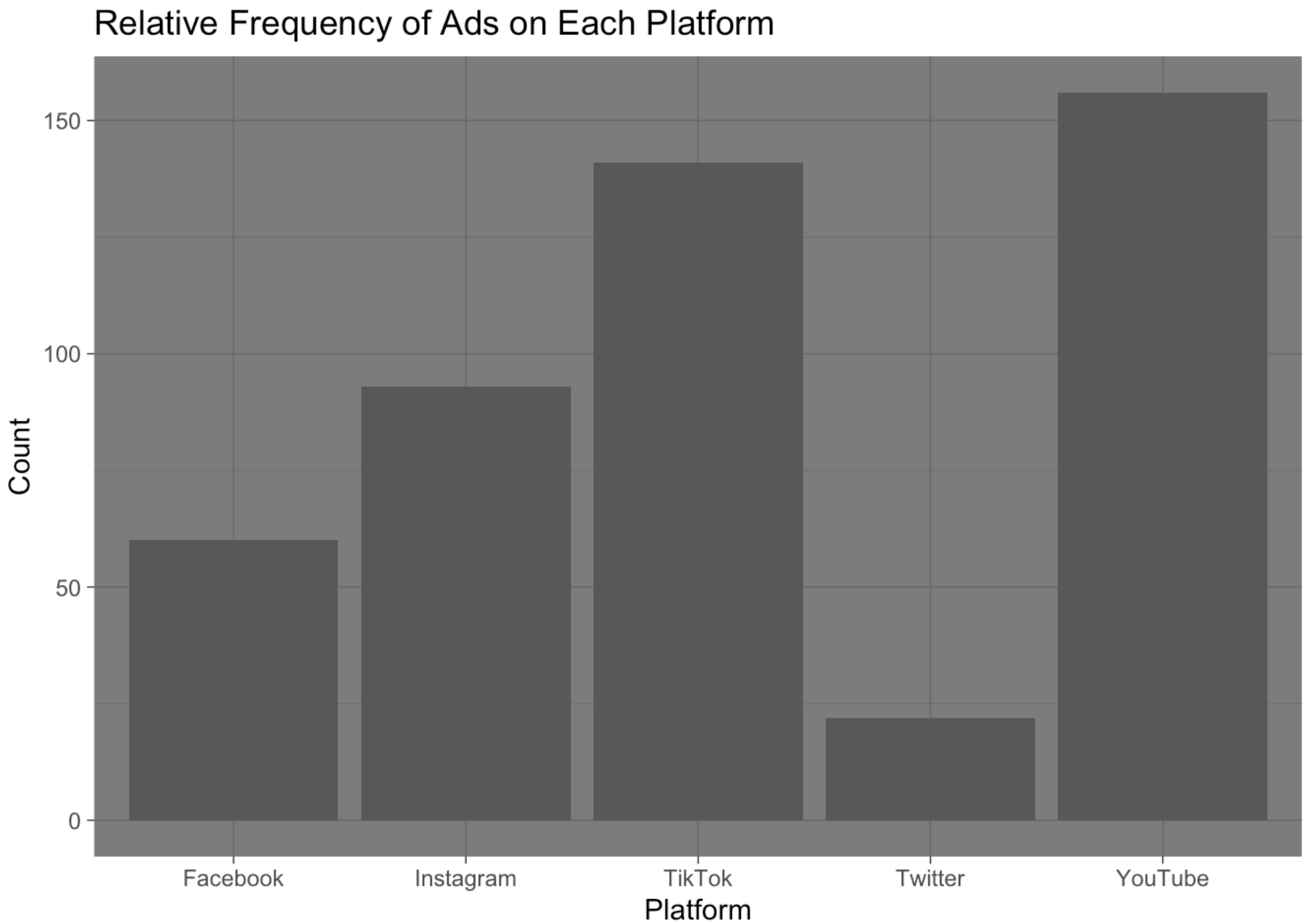
```
rel_table = prop.table(t1)
rel_table
```

```
##
##   Facebook  Instagram     TikTok    Twitter    YouTube
## 0.12711864 0.19703390 0.29872881 0.04661017 0.33050847
```

```
##  Facebook  Instagram     TikTok    Twitter    YouTube
##`0.12711864 0.19703390 0.29872881 0.04661017 0.33050847
rel_freq <- data.frame ("social Media"  = c("Facebook", "Instagram","TikTok","Twitter
","YouTube"),
                  "Frequency" = c("0.12711864", "0.19703390", "0.29872881", "0.046610
17", "0.33050847"),
                  "Count" = c("60","93","141","22","156"))
rel_freq
```

```
##   social.Media  Frequency Count
## 1     Facebook 0.12711864    60
## 2    Instagram 0.19703390    93
## 3       TikTok 0.29872881   141
## 4      Twitter 0.04661017    22
## 5      YouTube 0.33050847   156
```

```
# Barplot
library(ggplot2)

ggplot(df, aes(x = socialmedia),fill=socialmedia) +
  geom_bar(stat = "count") +
  scale_fill_manual(values = c("Facebook"="dark blue","Instagram"="purple","TikTok"="
black","Twitter"="light blue","YouTube"="red")) +
  labs(title = "Relative Frequency of Ads on Each Platform", x = "Platform", y = "Cou
nt") +
  theme_dark()
```


Relative Frequency of Ads on Each Platform
150
100
50
0
Count
Facebook
Instagram
TikTok
Twitter
YouTube
Platform

```
#### Q1 b) ####
## Using Goodness-of-Fit test because we are working with proportions and with multip
le categories.
#We are comparing true proportions with the expected proportions.
## Also, we have been given the expected proportions and we just calculated the obser
ved proportions above in the relative frequency table.

# H0 = PropFacebook = 0.1,PropInstagram= 0.2,PropTikTok= 0.3, PropTwitter= 0.1, PropY
ouTube= 0.3
# H1 = Atleast one of these proportions does not hold

library(stats)

# Specify the observed proportions of ads on each platform
obs <- c(0.12711864, 0.19703390, 0.29872881, 0.04661017, 0.33050847)

# Specify the expected proportions of ads on each platform
exp <- c(0.1, 0.2, 0.3, 0.1, 0.3)

# Conduct the Goodness-of-Fit test
chisq.test(obs, p=exp)
```

```
## Warning in chisq.test(obs, p = exp): Chi-squared approximation may be incorrect
```

```
##
##  Chi-squared test for given probabilities
##
## data:  obs
## X-squared = 0.039011, df = 4, p-value = 0.9998
```

```
# Observed p-value = 0.99 which is more than 0.05. We fail to reject NULL Hypothesis.
#There is no sufficient evidence to conclude that marketing department is not followi
ng the strategy.
```

```
#### Q1 c) ####
## Variance of each social media platform with respect to age.

cat(varF<- var(df[df$socialmedia == "Facebook", "age"]))
```

```
## 71.90141
```

```
cat(varI<- var(df[df$socialmedia == "Instagram", "age"]))
```

```
## 33.01987
```

```
cat(varTk<- var(df[df$socialmedia == "TikTok", "age"]))
```

```
## 10.60527
```

```
cat(varTw<- var(df[df$socialmedia == "Twitter", "age"]))
```

```
## 35.04762
```

```
cat(varY<- var(df[df$socialmedia == "YouTube", "age"]))
```

```
## 82.92055
```

```
## Standard Deviation of each social media platform with respect to age.

sdF<- sd(df[df$socialmedia == "Facebook", "age"])
sdI<- sd(df[df$socialmedia == "Instagram", "age"])
sdTk<- sd(df[df$socialmedia == "TikTok", "age"])
sdTw<- sd(df[df$socialmedia == "Twitter", "age"])
sdY<- sd(df[df$socialmedia == "YouTube", "age"])
sdF
```

```
## [1] 8.47947
```

```
sdI
```

```
## [1] 5.746292
```

```
sdTk
```

```
## [1] 3.256573
```

```
sdTw
```

```
## [1] 5.920103
```

```
sdY
```

```
## [1] 9.106072
```

```
## Coefficient of Variation of each social media platform with respect to age?

cvF = sd(df[df$socialmedia == "Facebook", "age"])/mean(df[df$socialmedia == "Facebook
", "age"])
cvF
```

```
## [1] 0.2760544
```

```
cvI= sd(df[df$socialmedia == "Instagram", "age"])/mean(df[df$socialmedia == "Instagra
m", "age"])
cvI
```

```
## [1] 0.2206462
```

```
cvTk = sd(df[df$socialmedia == "TikTok", "age"])/mean(df[df$socialmedia == "TikTok",
"age"])
cvTk
```

```
## [1] 0.1766065
```

```
cvTw = sd(df[df$socialmedia == "Twitter", "age"])/mean(df[df$socialmedia == "Twitter"
, "age"])
cvTw
```

```
## [1] 0.1644473
```

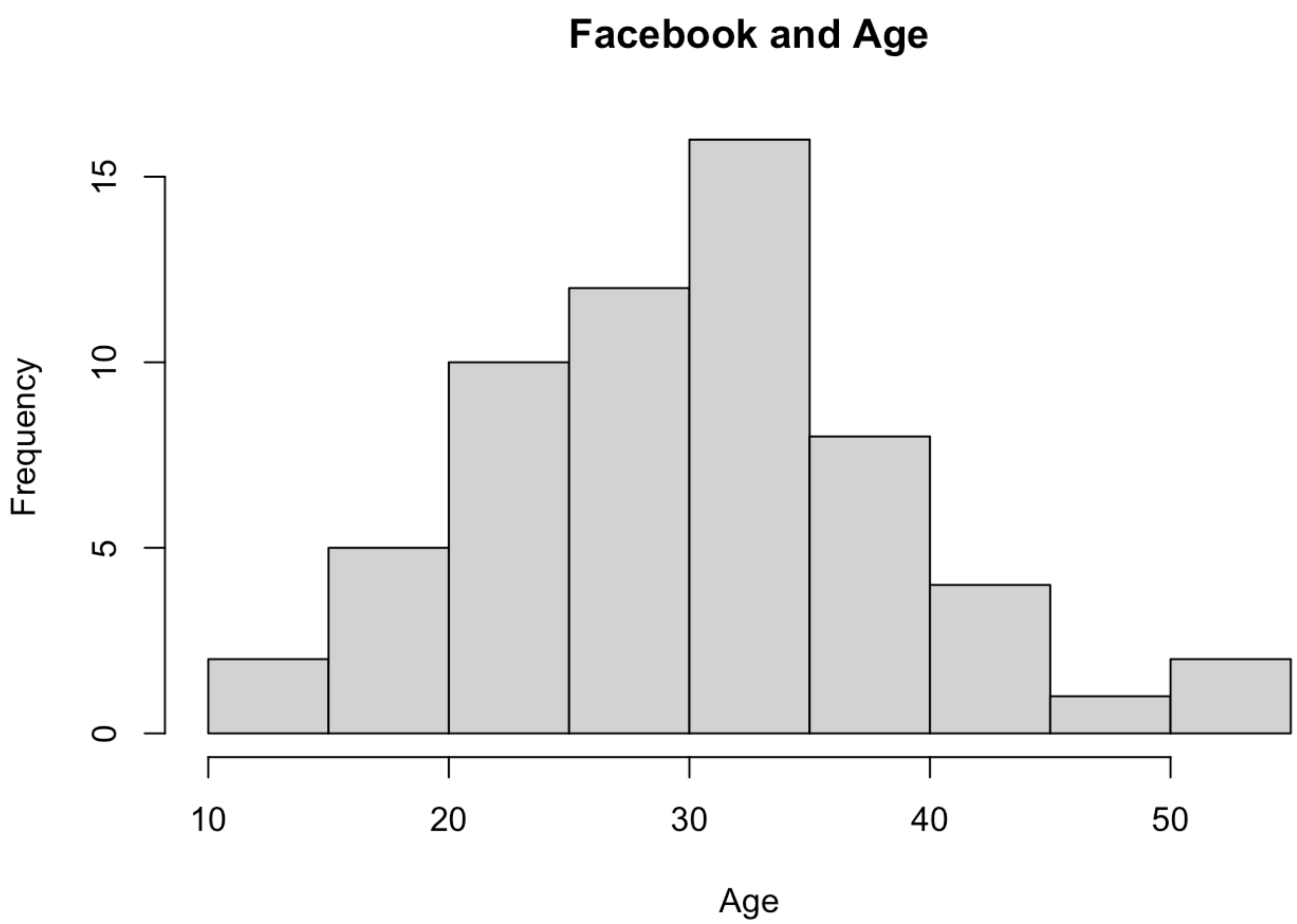
```
cvY = sd(df[df$socialmedia == "YouTube", "age"])/mean(df[df$socialmedia == "YouTube",
"age"])
cvY
```

```
## [1] 0.2879682
```

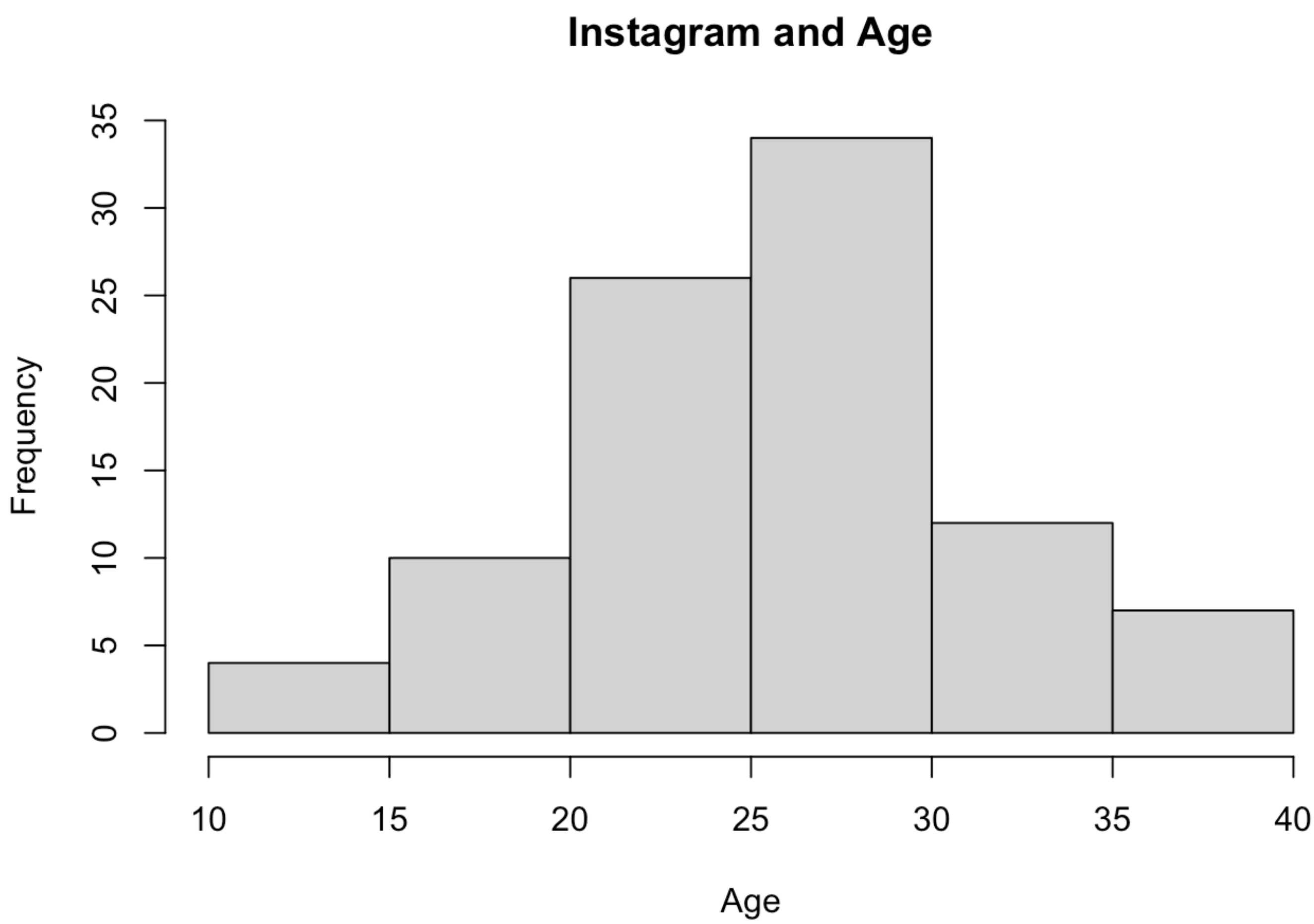
```
## skew of Age
library(moments)
skew<-skewness(df$age)
skew
```

```
## [1] 0.539083
```

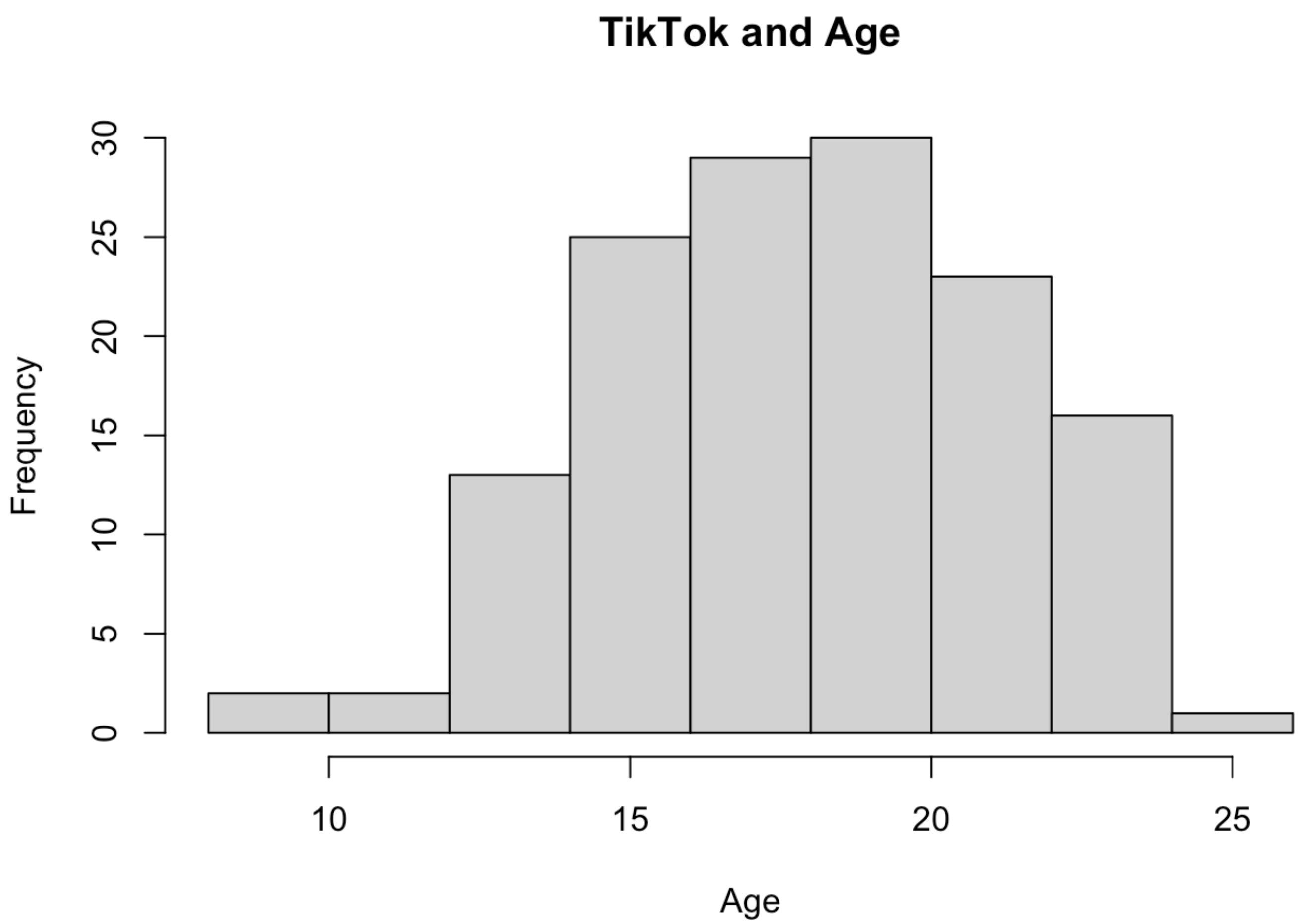
```
## Plot for distribution of Age variable.

hist(df[df$socialmedia == "Facebook", "age"],xlab ="Age",main = "Facebook and Age")
```



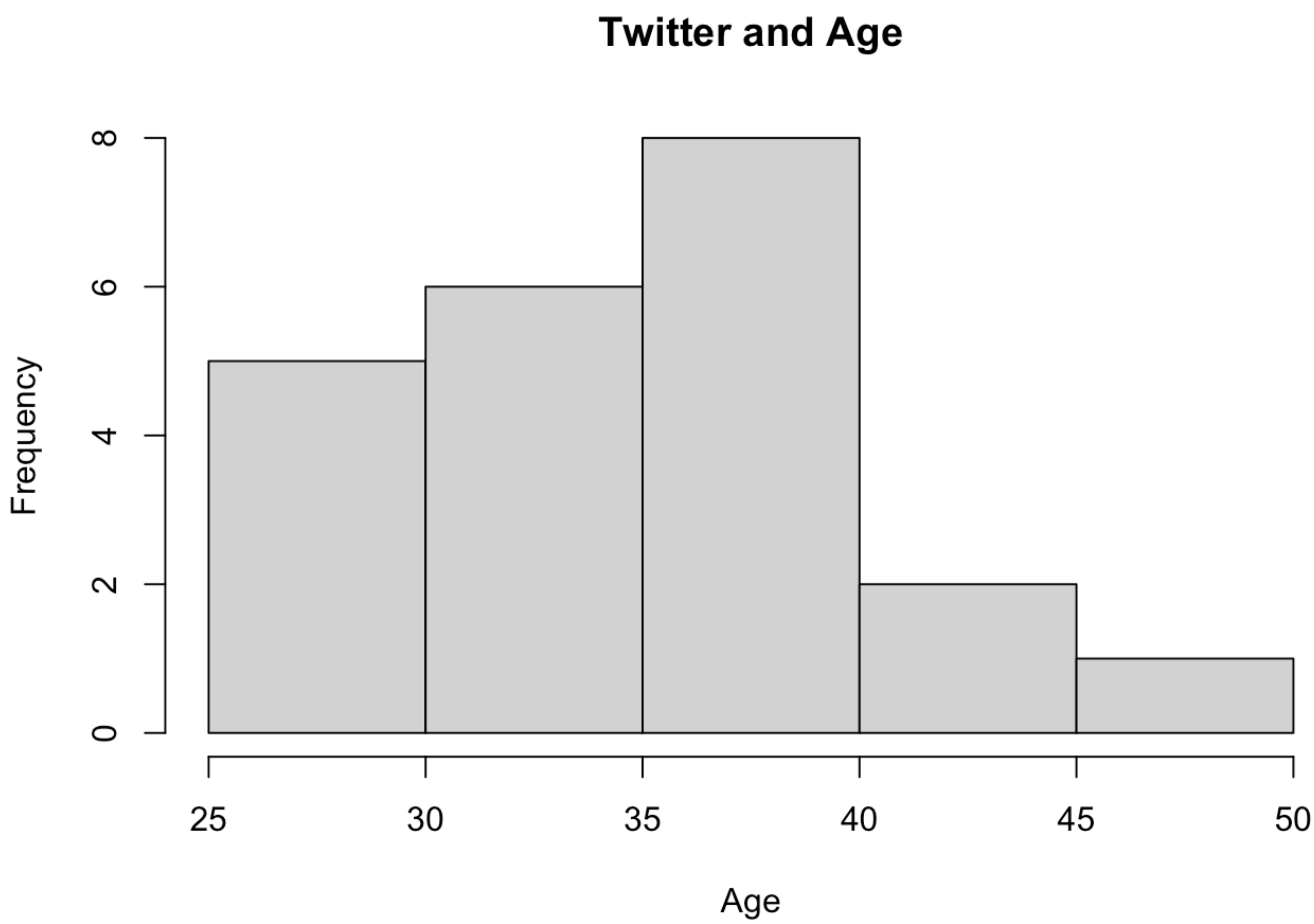

```
hist(df[df$socialmedia == "Instagram", "age"],xlab ="Age",main = "Instagram and Age")
```



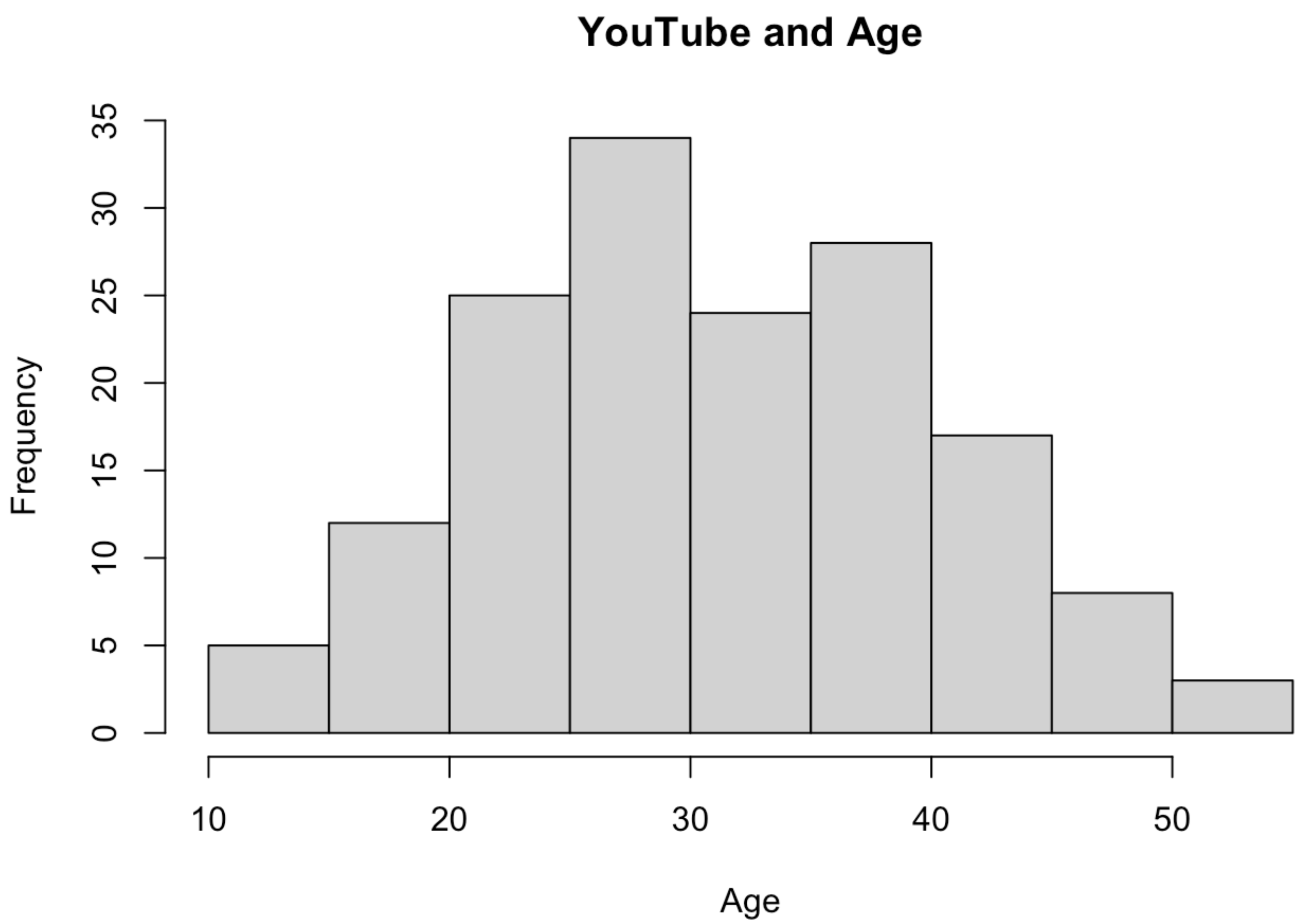

```
hist(df[df$socialmedia == "TikTok", "age"],xlab ="Age",main = "TikTok and Age")
```




```
hist(df[df$socialmedia == "Twitter", "age"],xlab ="Age",main = "Twitter and Age")
```




```
hist(df[df$socialmedia == "YouTube", "age"],xlab ="Age",main = "YouTube and Age")
```




```
## For YouTube maximum frequency of people are in  25-30 Age group
## For Twitter maximum frequency of people are in  35-40 Age group
## For TikTok maximum frequency of people are in 18-20 Age group
## For Instagram maximum frequency of people are in  25-30 Age group
## For Facebook maximum frequency of people are in 30-35 Age group
```

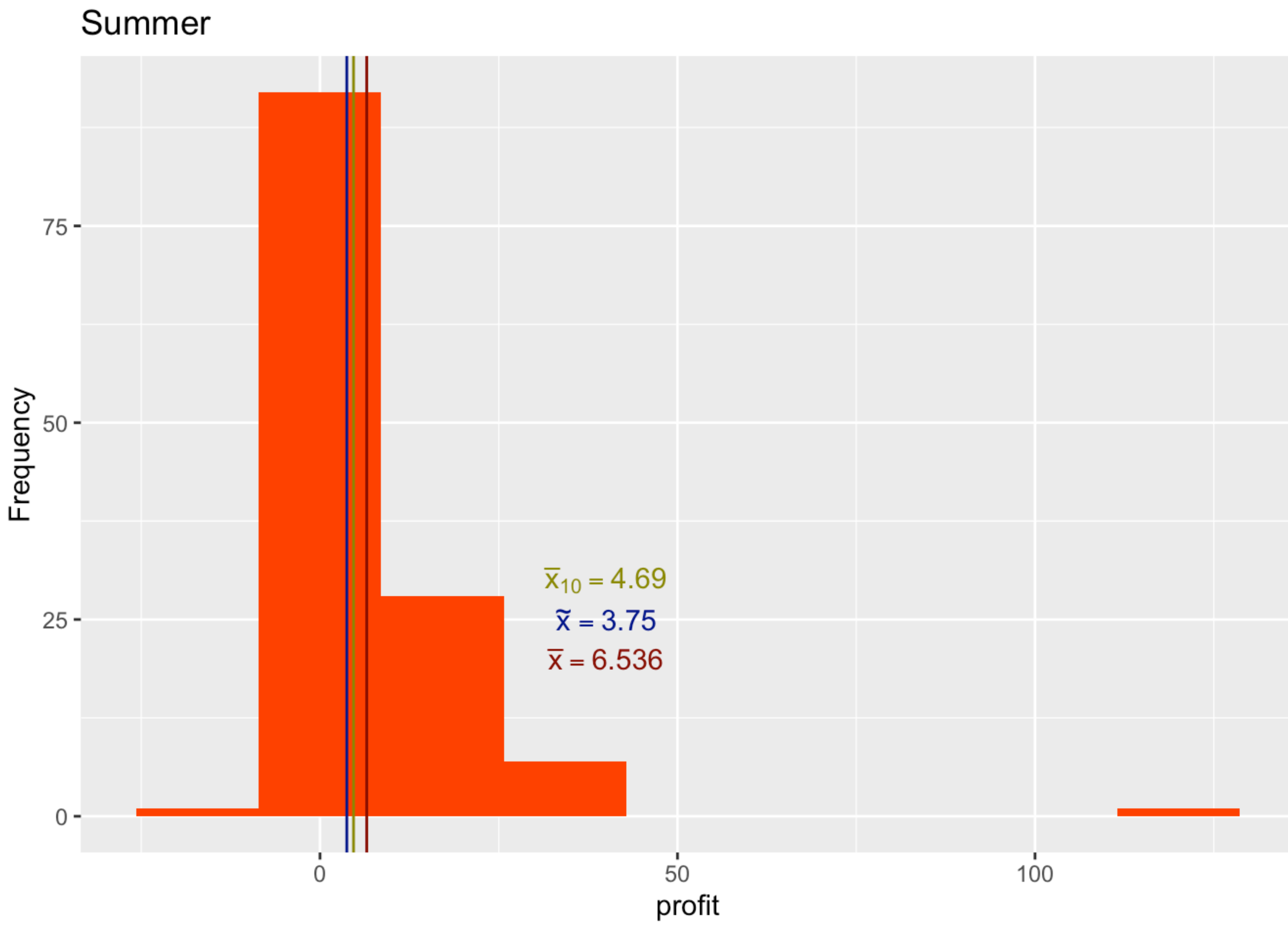
```
#Q2 A ####

sumr_df<-df %>%
  filter(season=="summer")
mean(sumr_df$adrevenue)
```

```
## [1] 13.01915
```

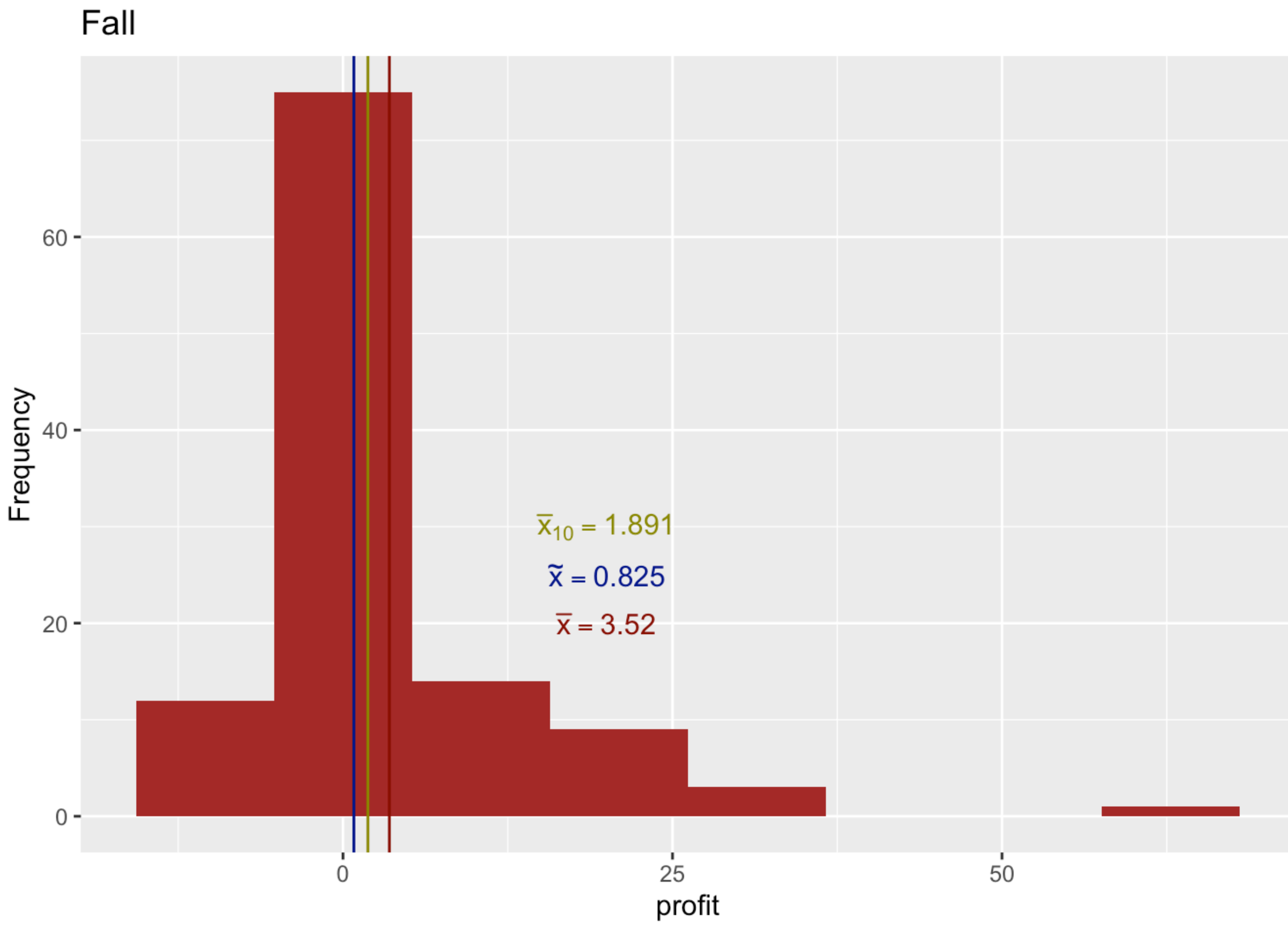
```
wntr_df<-df %>%
  filter(season=="winter")
```

```
sprng_df<-df %>%
  filter(season=="spring")

fll_df<-df %>%
  filter(season=="fall")

# Calculating Bin Width according to Sturges Formula
bin_sumr_plt <- ceiling(log(length(sumr_df$profit), 2)) + 1
bin_fll_plt <- ceiling(log(length(fll_df$profit), 2)) + 1
bin_sprng_plt <- ceiling(log(length(sprng_df$profit), 2)) + 1
bin_wntr_plt <- ceiling(log(length(wntr_df$profit), 2)) + 1



sumr_plt <- ggplot(sumr_df,aes(x=profit)) +
  ggtitle("Summer") +
  ylab("Frequency")+
  geom_histogram(fill="orangered1",bins=bin_sumr_plt) +
  geom_vline(aes(xintercept = mean(profit)), color = "darkred") +
  geom_vline(aes(xintercept = median(profit)),color = "darkblue") +
  geom_vline(aes(xintercept = mean(profit, trim = 0.1)),color = "yellow4") +
  annotate("text", x = 40, y = 20, label = paste("bar(x)==",round(mean(sumr_df$profit
), 3)), parse = T, color = "darkred") +
  annotate("text", x = 40, y = 25, label = paste("tilde(x)==",round(median(sumr_df$pr
ofit), 3)), parse = T, color = "darkblue") +
  annotate("text", x = 40, y = 30, label = paste("bar(x)[10]==",round(mean(sumr_df$pr
ofit,0.1), 3)), parse = T, color = "yellow4")




wntr_plt <- ggplot(wntr_df,aes(x=profit)) +
  ggtitle("Winter") +
  ylab("Frequency")+
  geom_histogram(fill="lightblue",bins=bin_wntr_plt) +
  geom_vline(aes(xintercept = mean(profit)), color = "darkred") +
  geom_vline(aes(xintercept = median(profit)),color = "darkblue") +
  geom_vline(aes(xintercept = mean(profit, trim = 0.1)),color = "yellow4") +
  annotate("text", x = 20, y = 20, label = paste("bar(x)==",round(mean(wntr_df$profit
), 3)), parse = T, color = "darkred") +
  annotate("text", x = 20, y = 25, label = paste("tilde(x)==",round(median(wntr_df$pr
ofit), 3)), parse = T, color = "darkblue") +
  annotate("text", x = 20, y = 30, label = paste("bar(x)[10]==",round(mean(wntr_df$pr
ofit,0.1), 3)), parse = T, color = "yellow4")
```

```
sprng_plt <- ggplot(sprng_df,aes(x=profit)) +
  ggtitle("Spring") +
  ylab("Frequency")+
  geom_histogram(fill="pink",bins=bin_sprng_plt) +
  geom_vline(aes(xintercept = mean(profit)), color = "darkred") +
  geom_vline(aes(xintercept = median(profit)),color = "darkblue") +
  geom_vline(aes(xintercept = mean(profit, trim = 0.1)),color = "yellow4") +
  annotate("text", x = 10, y = 20, label = paste("bar(x)==",round(mean(sprng_df$profi
t), 3)), parse = T, color = "darkred") +
  annotate("text", x = 10, y = 25, label = paste("tilde(x)==",round(median(sprng_df$p
rofit), 3)), parse = T, color = "darkblue") +
  annotate("text", x = 10, y = 30, label = paste("bar(x)[10]==",round(mean(sprng_df$p
rofit,0.1), 3)), parse = T, color = "yellow4")



fll_plt <- ggplot(fll_df,aes(x=profit)) +
  ggtitle("Fall") +
  ylab("Frequency")+
  geom_histogram(fill="brown",bins=bin_fll_plt) +
  geom_vline(aes(xintercept = mean(profit)), color = "darkred") +
  geom_vline(aes(xintercept = median(profit)),color = "darkblue") +
  geom_vline(aes(xintercept = mean(profit, trim = 0.1)),color = "yellow4") +
  annotate("text", x = 20, y = 20, label = paste("bar(x)==",round(mean(fll_df$profit)
, 3)), parse = T, color = "darkred") +
  annotate("text", x = 20, y = 25, label = paste("tilde(x)==",round(median(fll_df$pro
fit), 3)), parse = T, color = "darkblue") +
  annotate("text", x = 20, y = 30, label = paste("bar(x)[10]==",round(mean(fll_df$pro
fit,0.1), 3)), parse = T, color = "yellow4")


sumr_plt
```

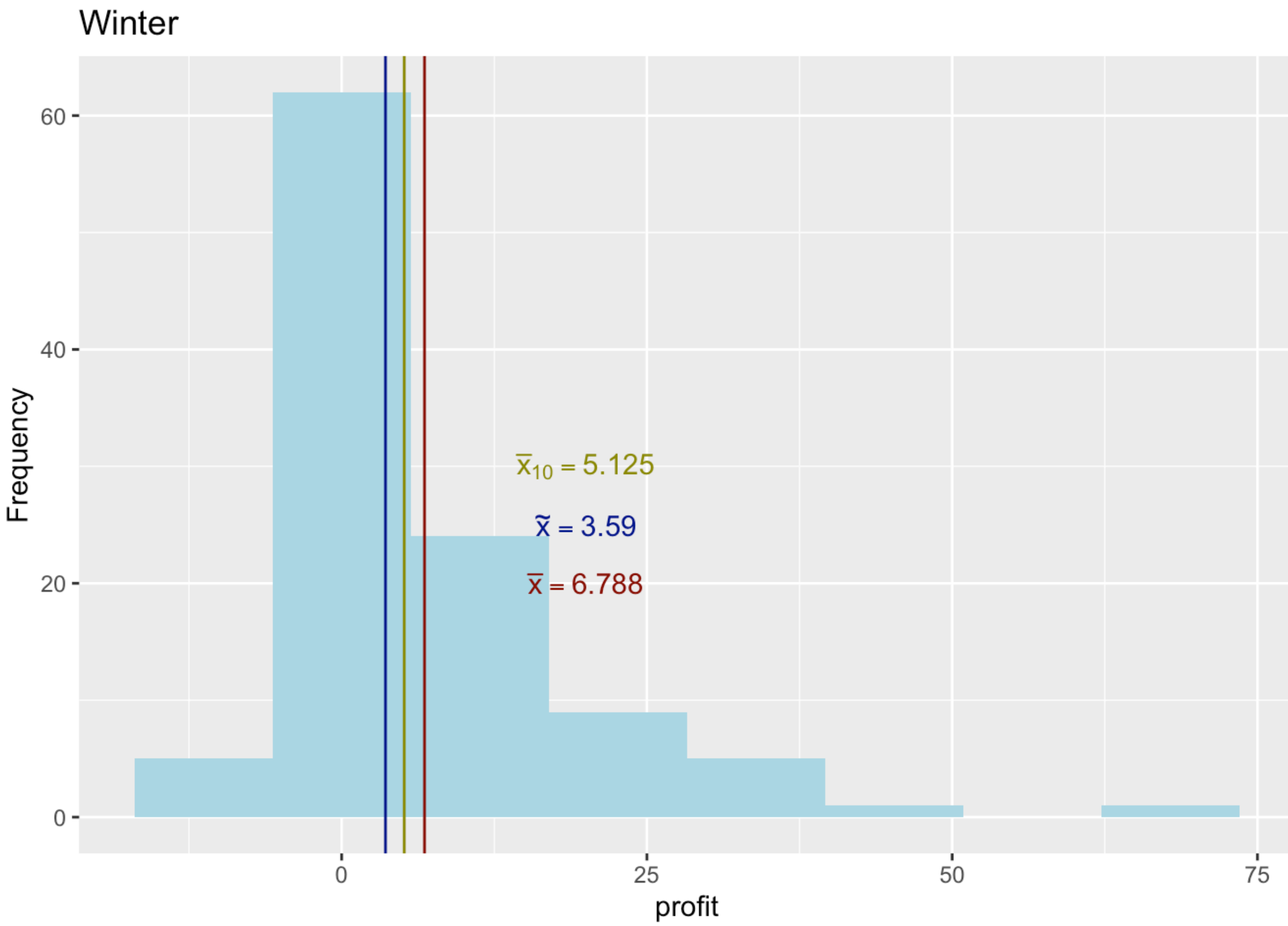

Summer
75
50
25
0
Frequency
$\bar{x}_{10} = 4.69$
$\tilde{x} = 3.75$
$\bar{x} = 6.536$
0
50
100
profit

```
fll_plt
```

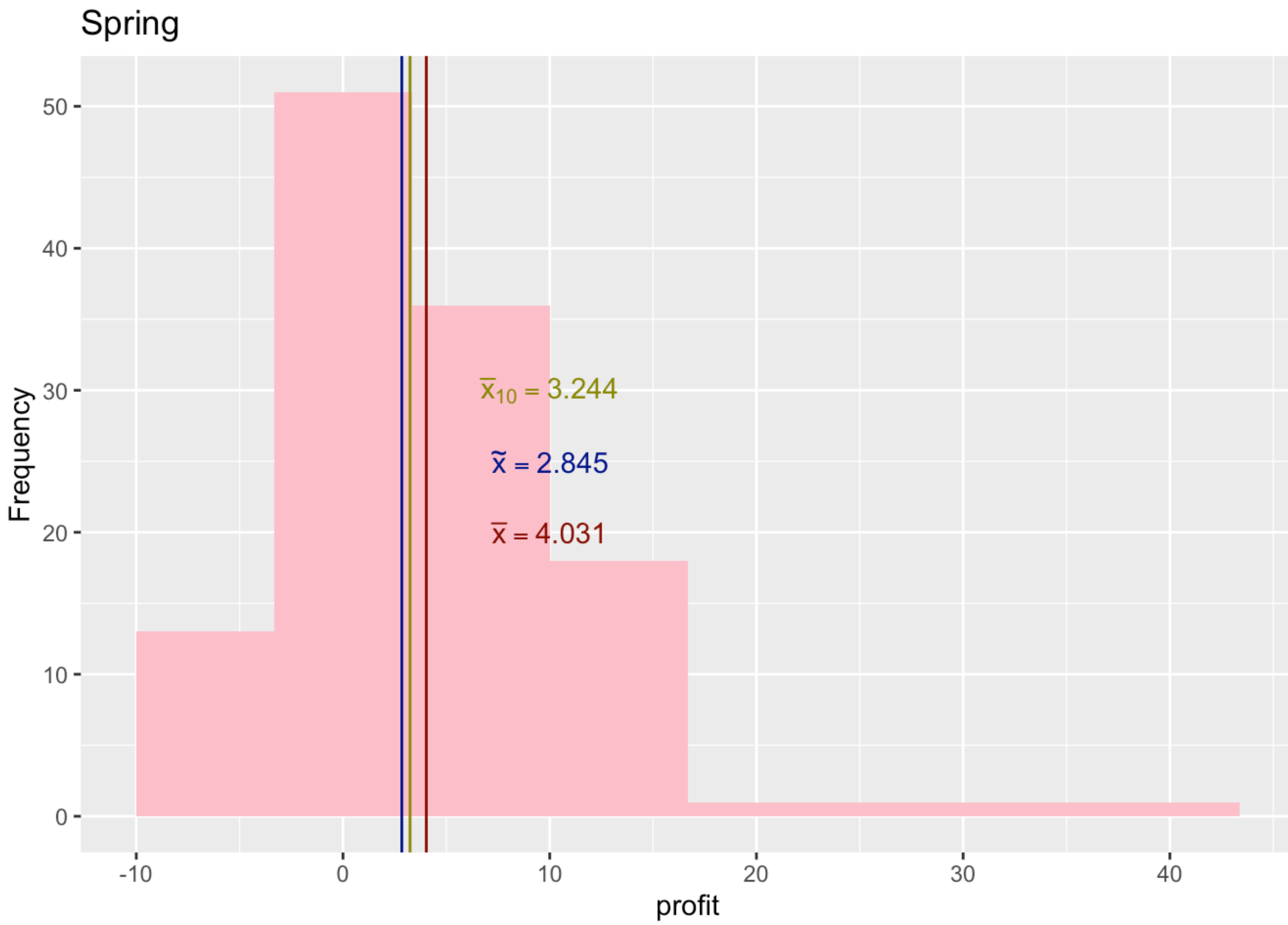


Fall
Frequency
60
40
20
0
0
25
50
profit
$\overline{x}_{10} = 1.891$
$\tilde{x} = 0.825$
$\overline{x} = 3.52$


```
wntr_plt
```


Winter
Frequency
profit
60
40
20
0
0
25
50
75
$\bar{x}_{10} = 5.125$
$\tilde{x} = 3.59$
$\bar{x} = 6.788$

```
sprng_plt
```

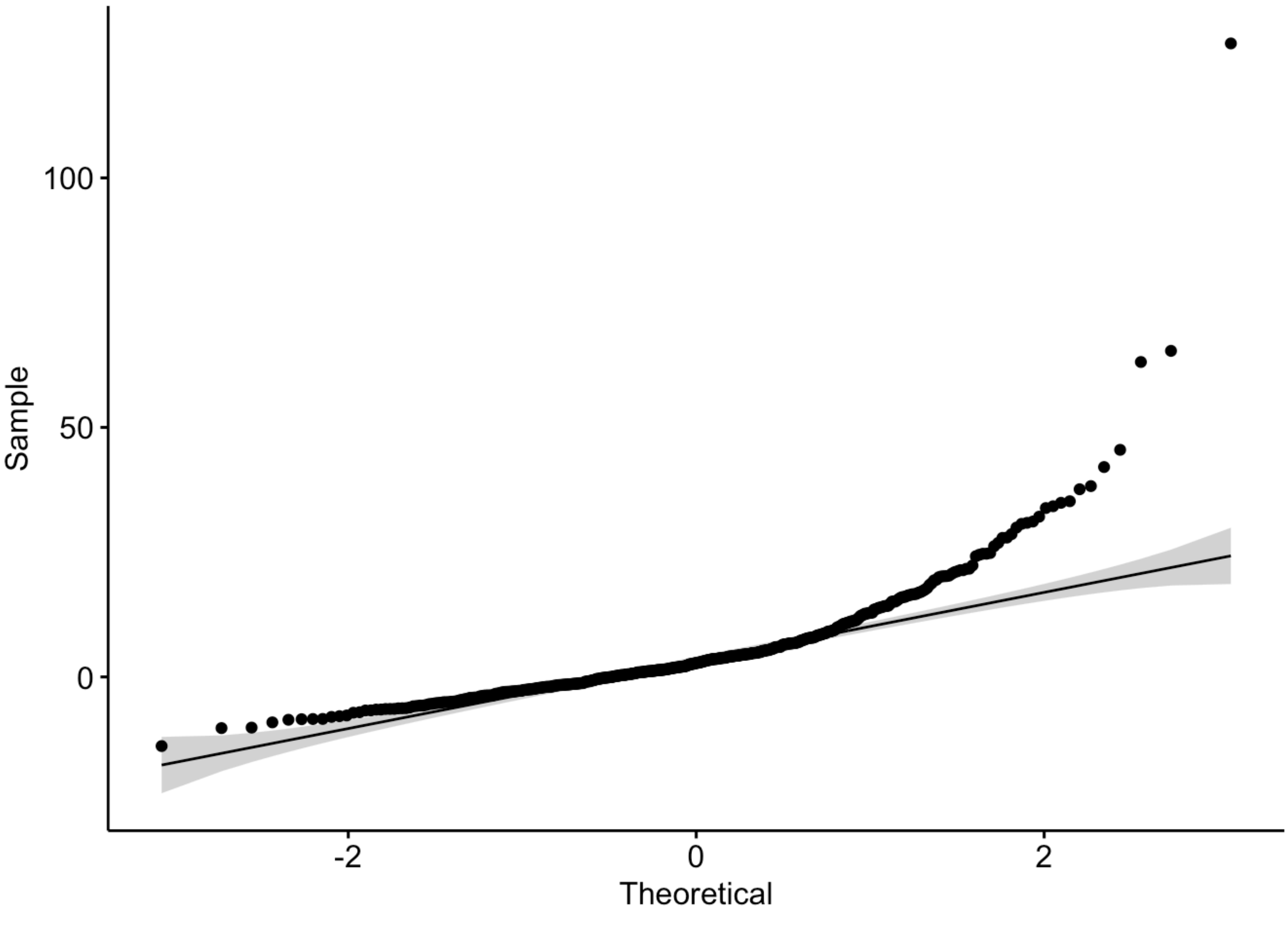



```
#### Q2 b) ####
# Checking for the normality at significance level 0.05
# H0 = sample distribution is normal
# H1 = sample is not normally distributed
shapiro.test(df$profit)
```

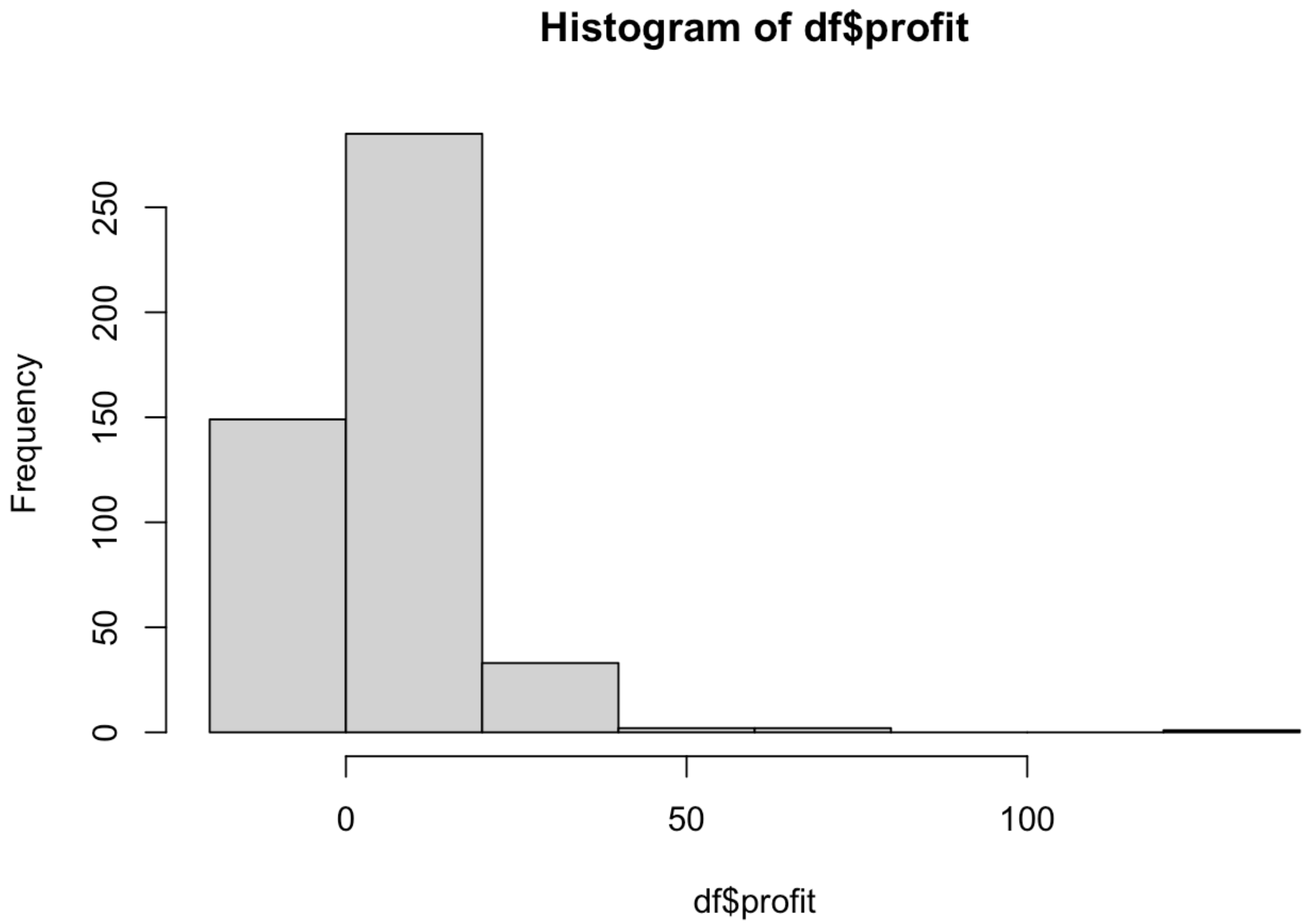
```
##
##  Shapiro-Wilk normality test
##
## data:  df$profit
## W = 0.73561, p-value < 2.2e-16
```

```
ggqqplot(df$profit)
```

```
## Warning: The following aesthetics were dropped during statistical transformation:
sample
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
## The following aesthetics were dropped during statistical transformation: sample
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```




```
hist(df$profit)
```

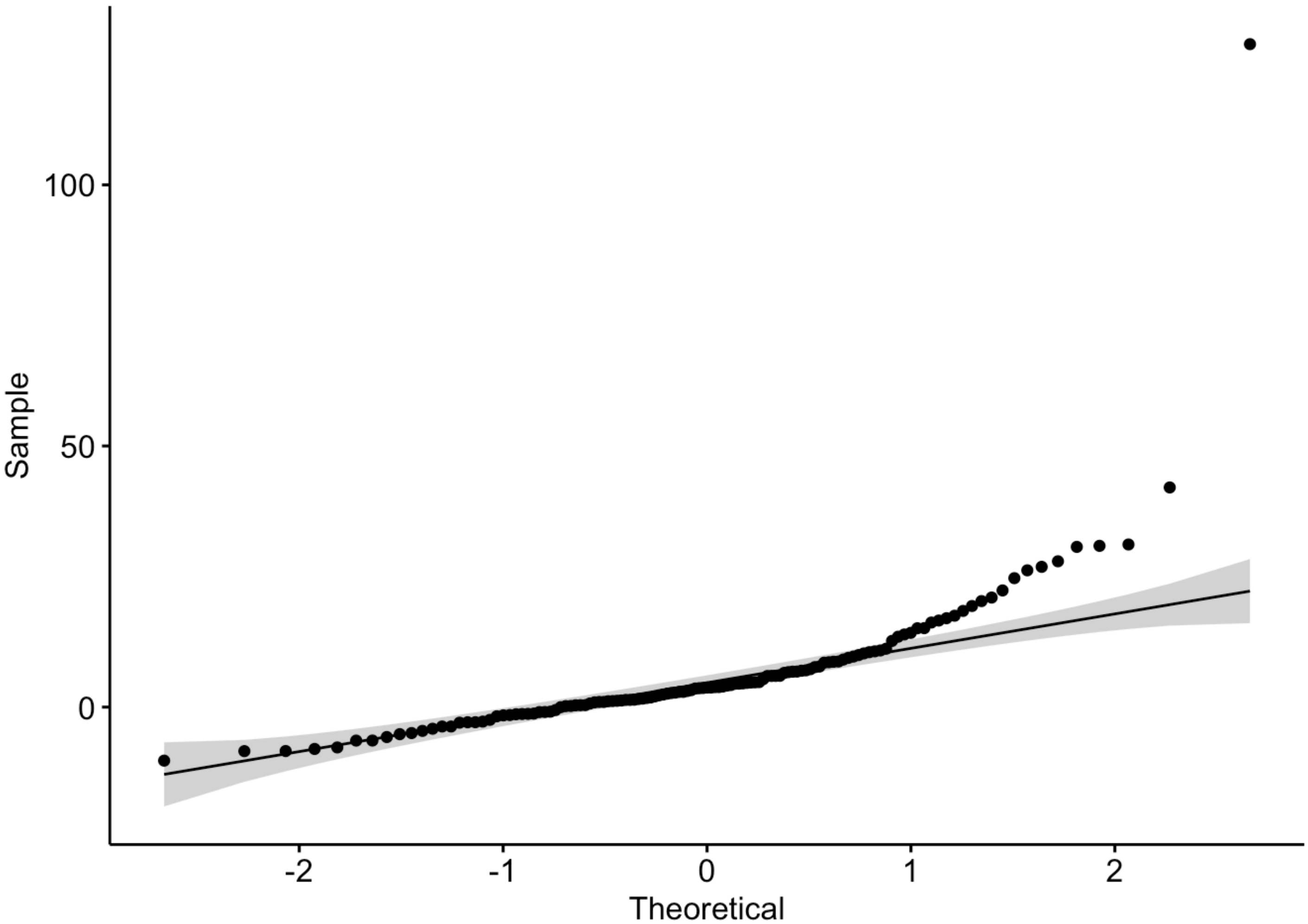



```
shapiro.test(sumr_df$profit)
```

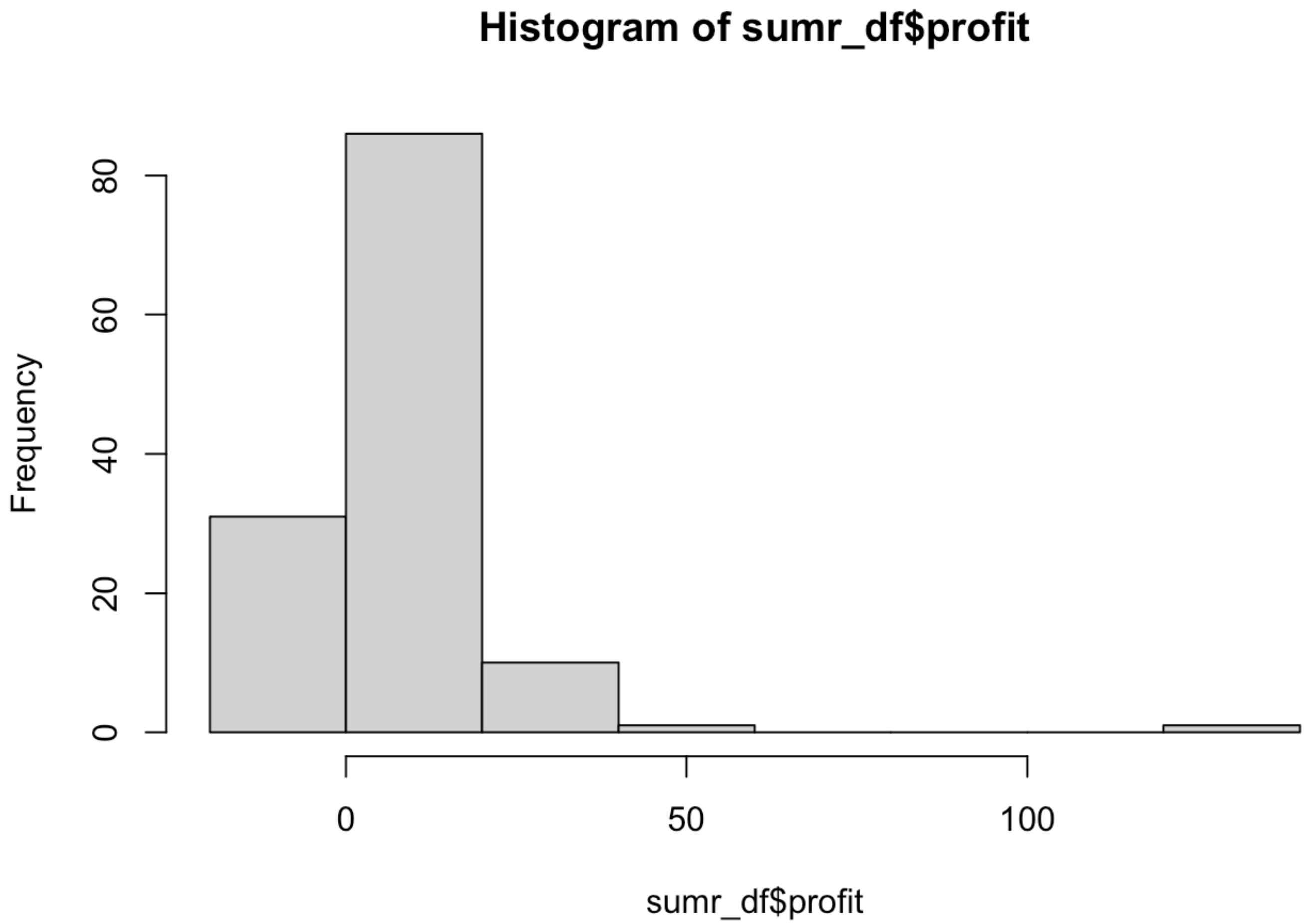
```
##
##  Shapiro-Wilk normality test
##
## data:  sumr_df$profit
## W = 0.61244, p-value < 2.2e-16
```

```
ggqqplot(sumr_df$profit)
```

```
## Warning: The following aesthetics were dropped during statistical transformation:
sample
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
## The following aesthetics were dropped during statistical transformation: sample
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```




```
hist(sumr_df$profit)
```

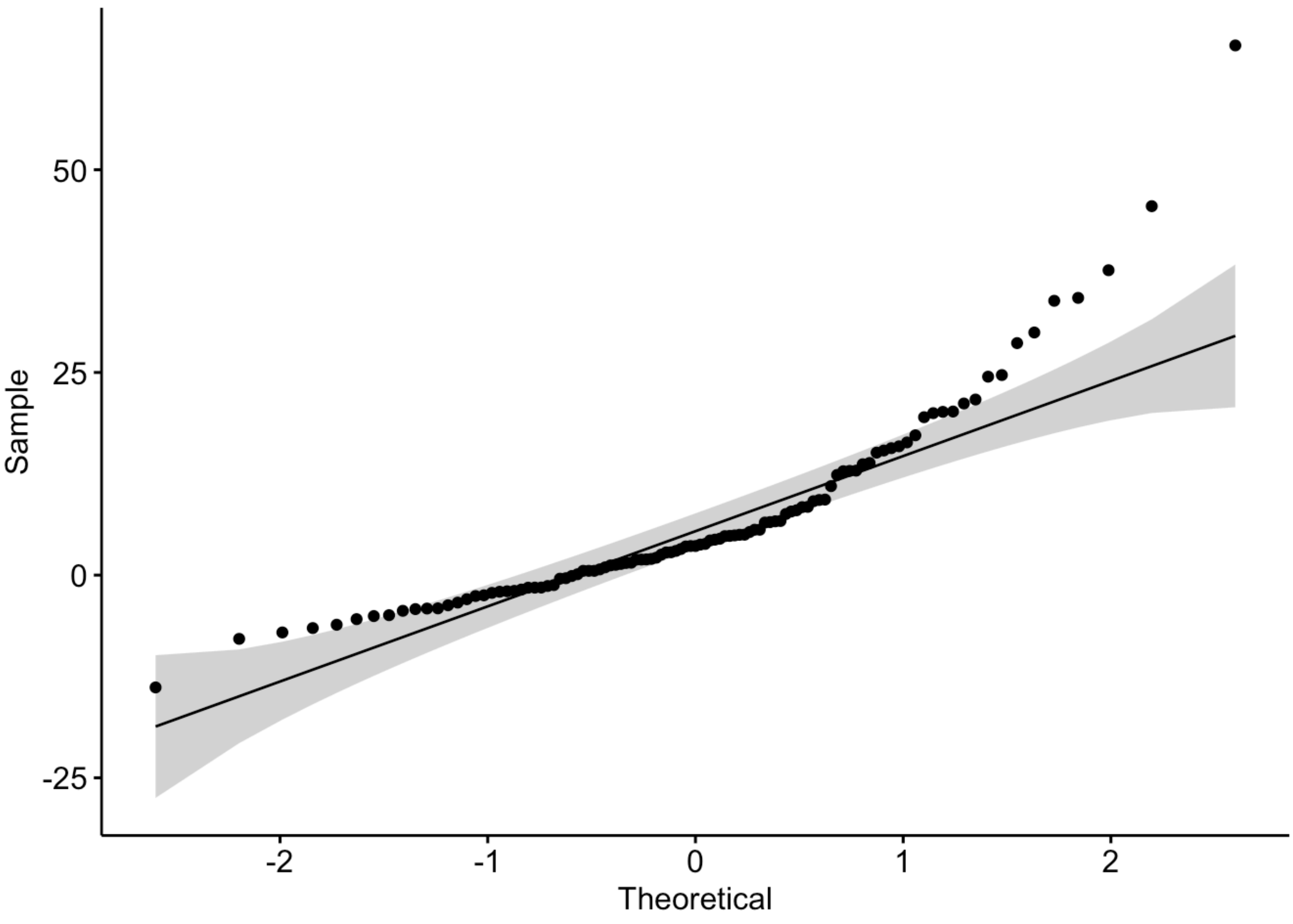



```
shapiro.test(wntr_df$profit)
```

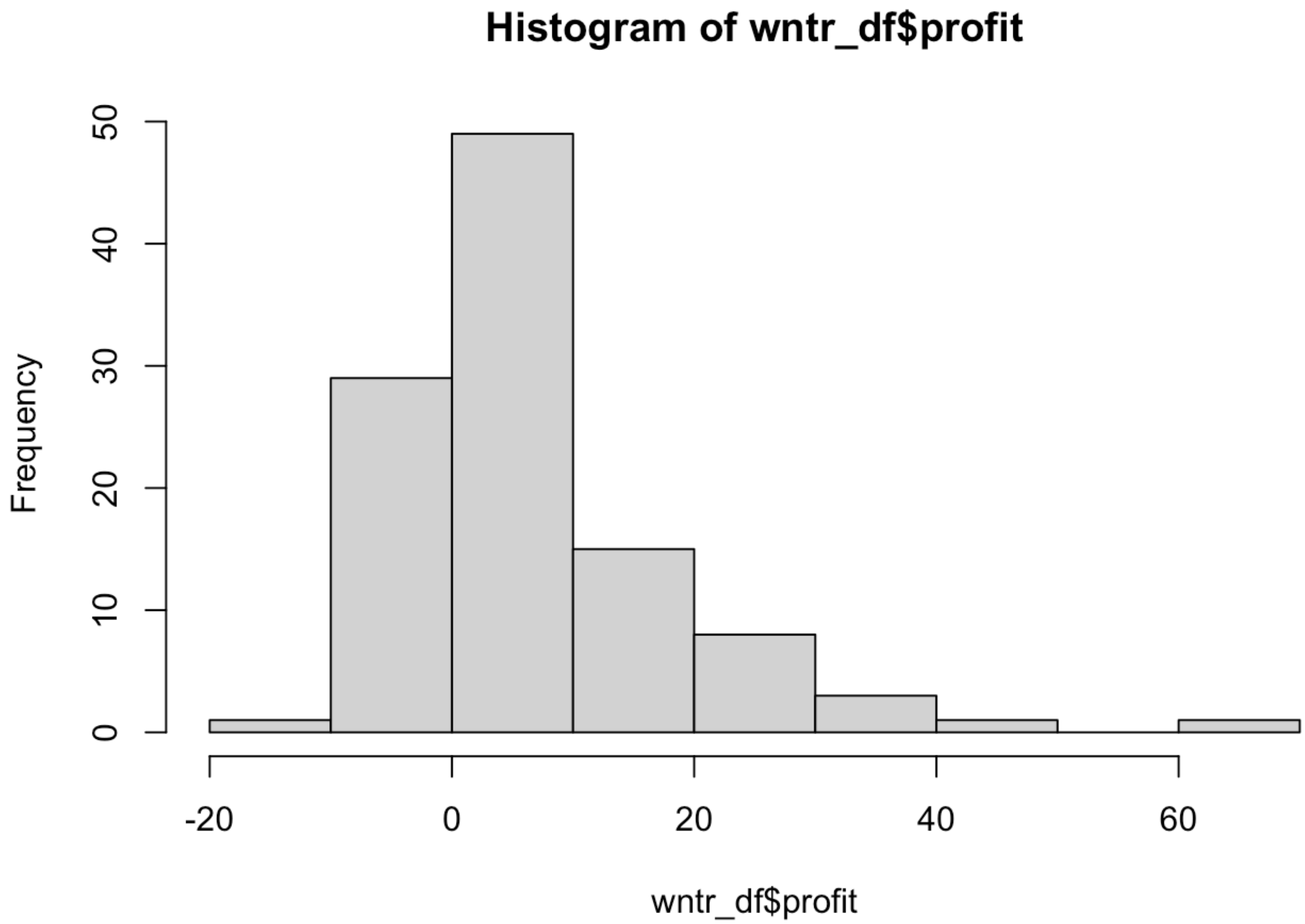
```
##
##  Shapiro-Wilk normality test
##
## data:  wntr_df$profit
## W = 0.84576, p-value = 3.495e-09
```

```
ggqqplot(wntr_df$profit)
```

```
## Warning: The following aesthetics were dropped during statistical transformation:
sample
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
## The following aesthetics were dropped during statistical transformation: sample
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```




```
hist(wntr_df$profit)
```




```
## Since sample is not normally distributed we chose to apply non-parametric inferenc
e
## We will use Wilcoxin Rank-Sum test since we don't know if the Two population Varia
nces are equal or not.
## At alpha significance level 0.05
## Hypotheses: H0 : μsumr ≤ μwntr vs. H1 : μsumr > μwntr, samples of sizes nsumr = 12
9 and nwntr = 107

wilcox.test(sumr_df$profit,wntr_df$profit,alternative = "greater",exact = F,correct =
F)
```

```
##
##  Wilcoxon rank sum test
##
## data:  sumr_df$profit and wntr_df$profit
## W = 6870, p-value = 0.5241
## alternative hypothesis: true location shift is greater than 0
```

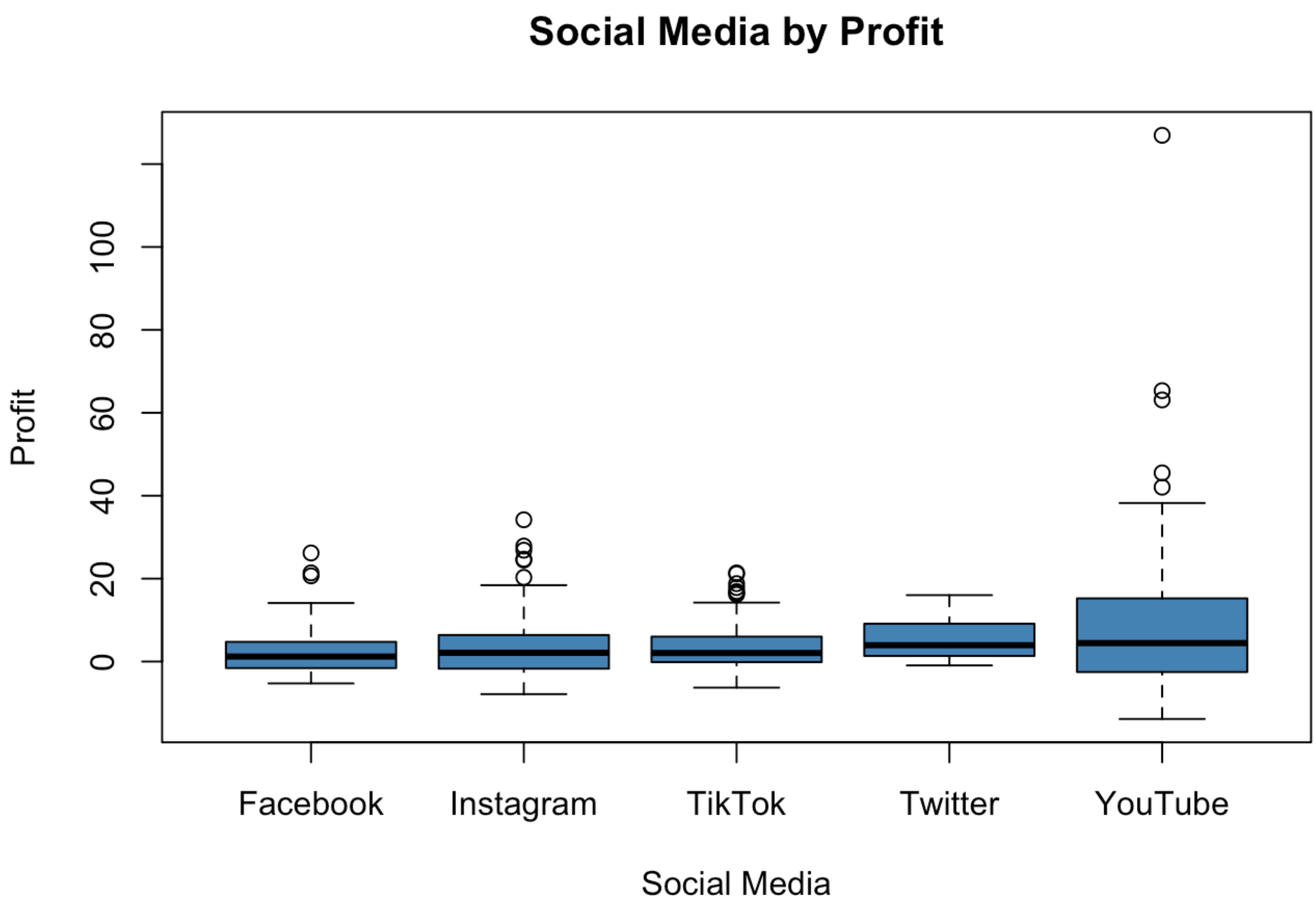
```
## p-value 0.5241 > 0.05 we fail to reject H0.
## Hence the Profit in winter is Greater than Profit in Summer.
## We don't have enough evidence to say that CEO is correct.
```

```
#### Q2 c) ####
#In order to test all the 4 season sample at once , we are going to use one way anova
. However, in order to conduct   one way anova test ,
# Three assumptions must hold:
# • Normality: Each group follows a normal distribution
# • Equal variances: Population variances for each group are equal
# • Independence: Observations are not correlate
#As we have seen earlier the normality doesn't hold true for summer and winter sample
.
## Since the underlying normality assumptions of ANOVA are violated we cannot go ahea
d with ANOVA test.
# We will perform Kruskal-wallis test which is non parametric equivalent of one-way A
NOVA.
kruskal.test(df$profit~df$season)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df$profit by df$season
## Kruskal-Wallis chi-squared = 10.831, df = 3, p-value = 0.01268
```

```
#H0 : μsumr = μwntr = μsprng = μfll
#H1 : at least one of the seasons has an average profit that is different from at lea
st one of the other seasons.

# As we can see that the p-value 0.01268 < 0.05 we reject H0.
#  there is a significant difference in the avg. Profit across the seasons.
```

```
#### Q3 a) ####
boxplot(df$profit ~ df$socialmedia,
        col='steelblue',
        main='Social Media by Profit',
        xlab='Social Media',
        ylab='Profit')
```




```
## From the boxplot we can observe that profit from YouTube is higher than other plat
forms
## Even though it's not a major difference profit is least in TikTok
## There are a few significant outliers in YouTube
```

```
#### Q3 b) ####

library("car")
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

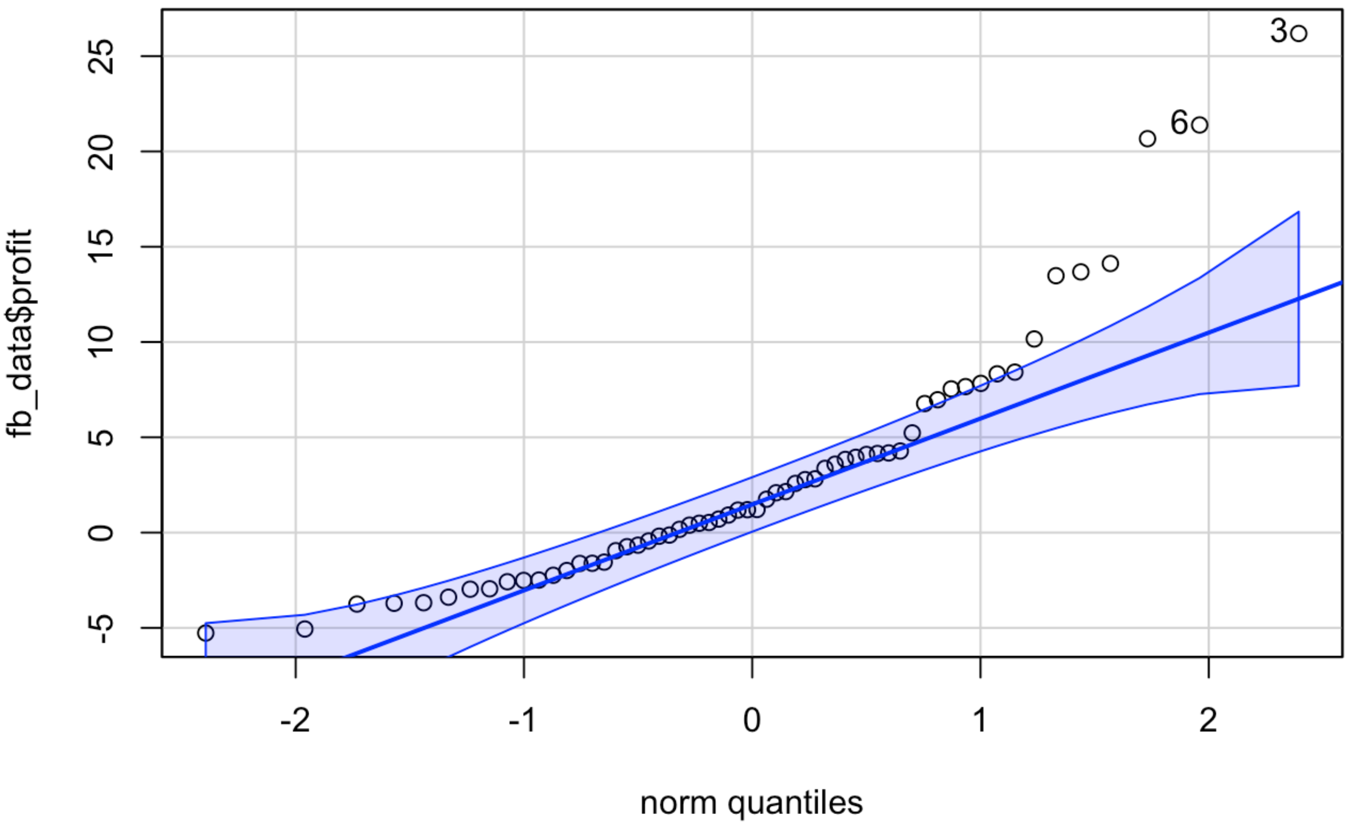
```
## The following object is masked from 'package:dplyr':
##
##     recode
```

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

```
## The following object is masked from 'package:psych':
##
##     logit
```

```
## The following object is masked from 'package:modeltools':
##
##     Predict
```

```
fb_data<-df[df$socialmedia == 'Facebook',]
Insta_data<-df[df$socialmedia == 'Instagram',]
Tk_data<-df[df$socialmedia == 'TikTok',]
Tw_data<-df[df$socialmedia == 'Twitter',]
YT_data<-df[df$socialmedia == 'YouTube',]
qqPlot(fb_data$profit)
```

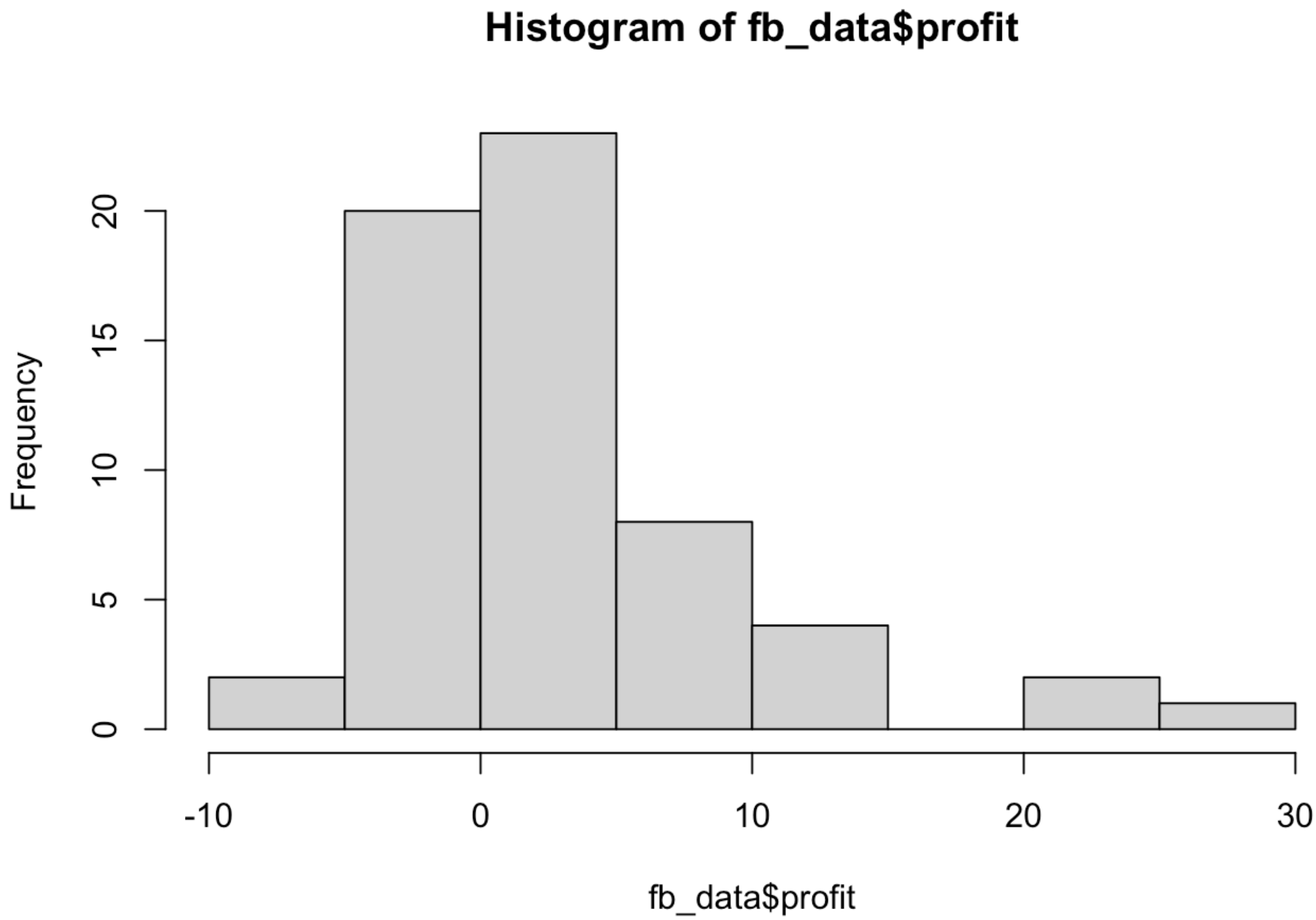



```
## [1] 3 6
```

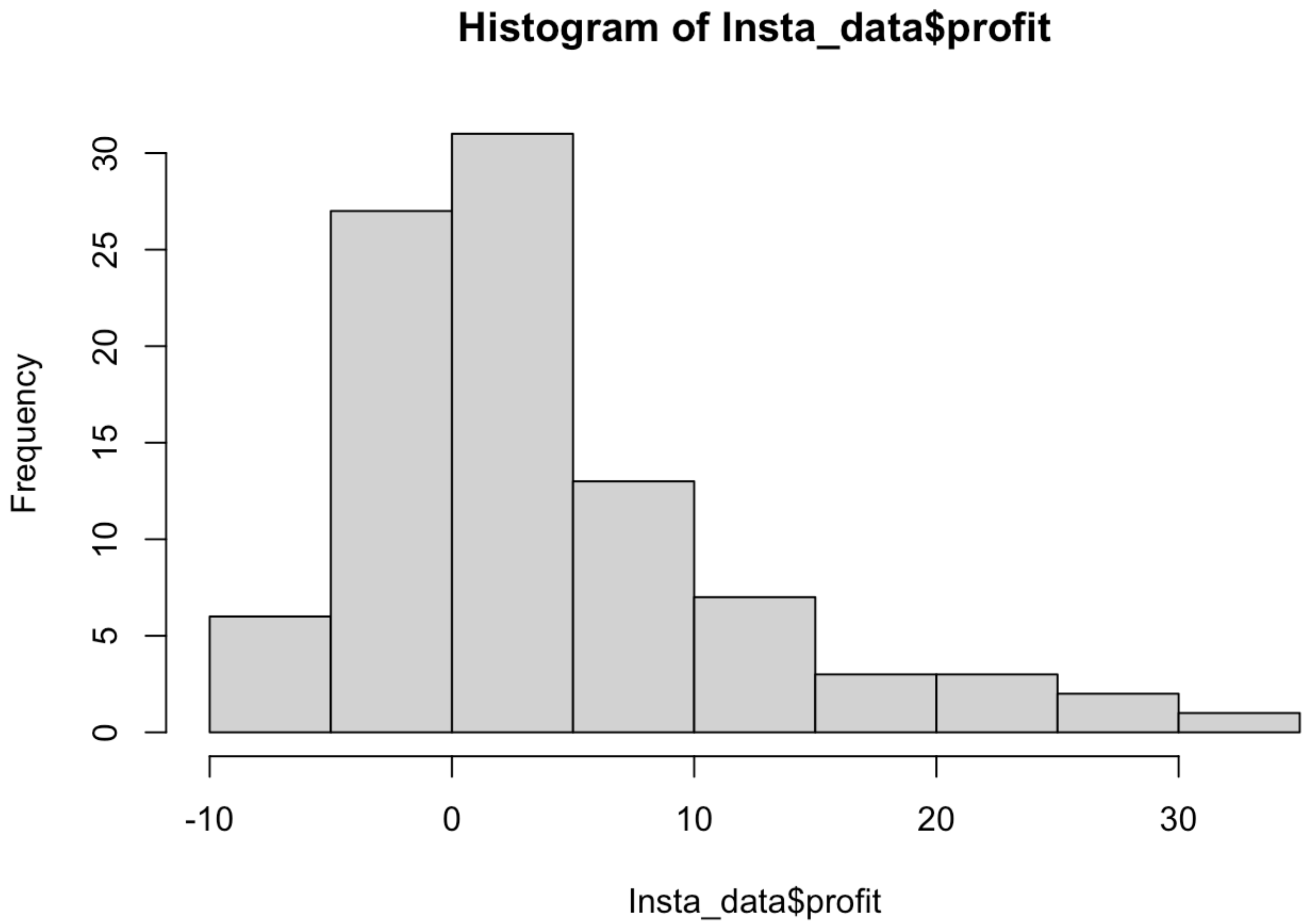
```
shapiro.test(fb_data$profit)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fb_data$profit
## W = 0.85936, p-value = 5.819e-06
```

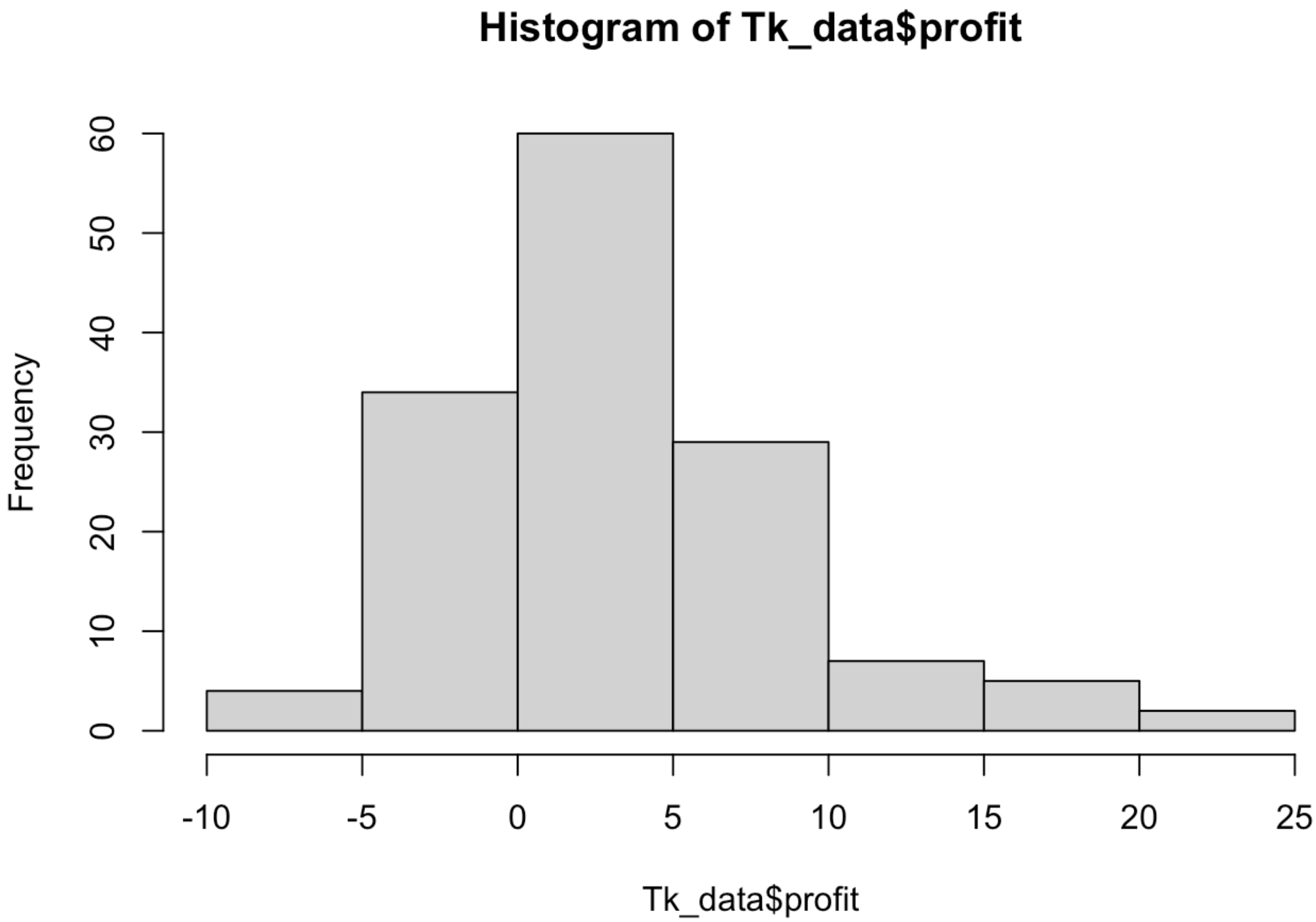
```
hist(fb_data$profit)
```



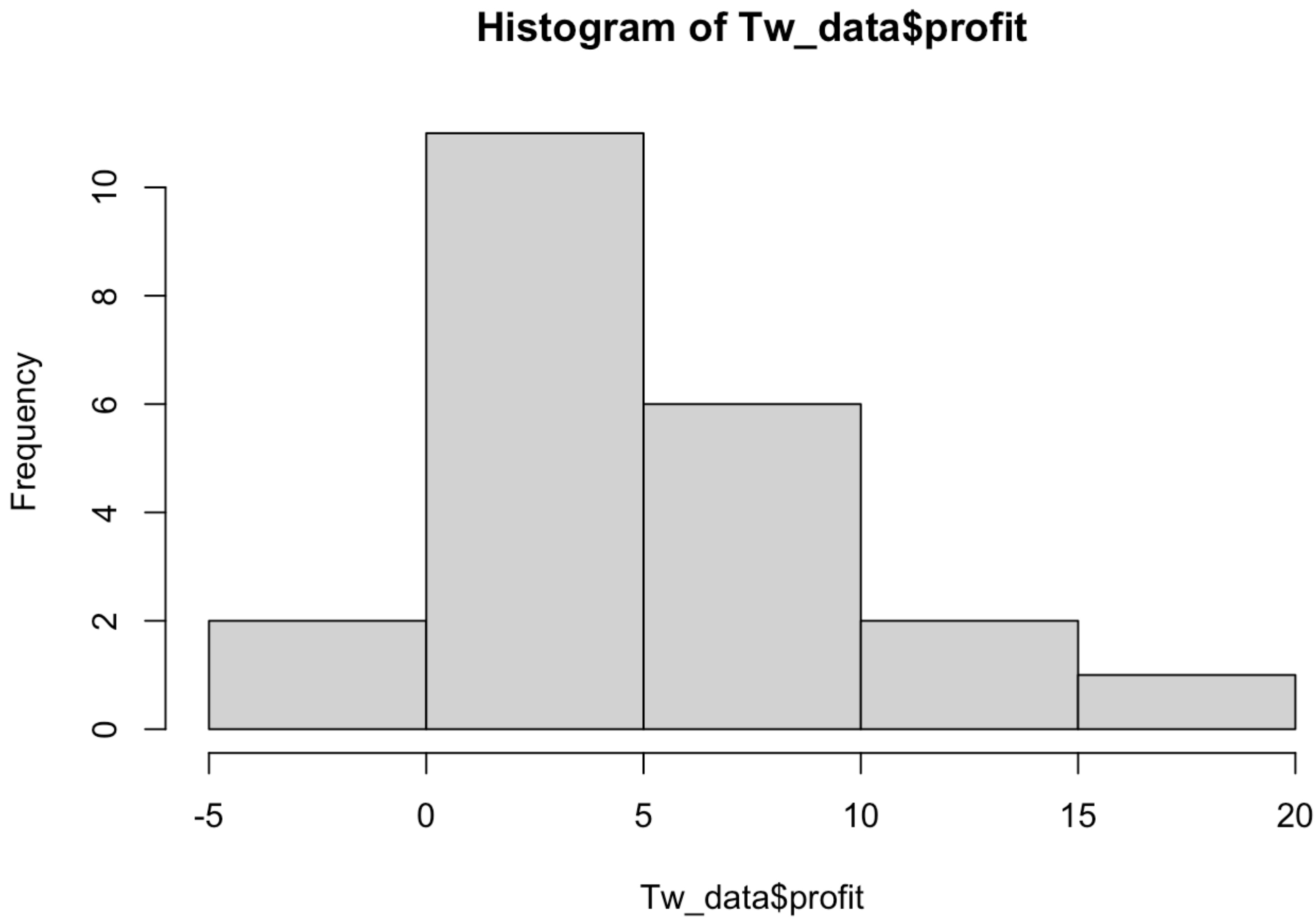

```
hist(Insta_data$profit)
```



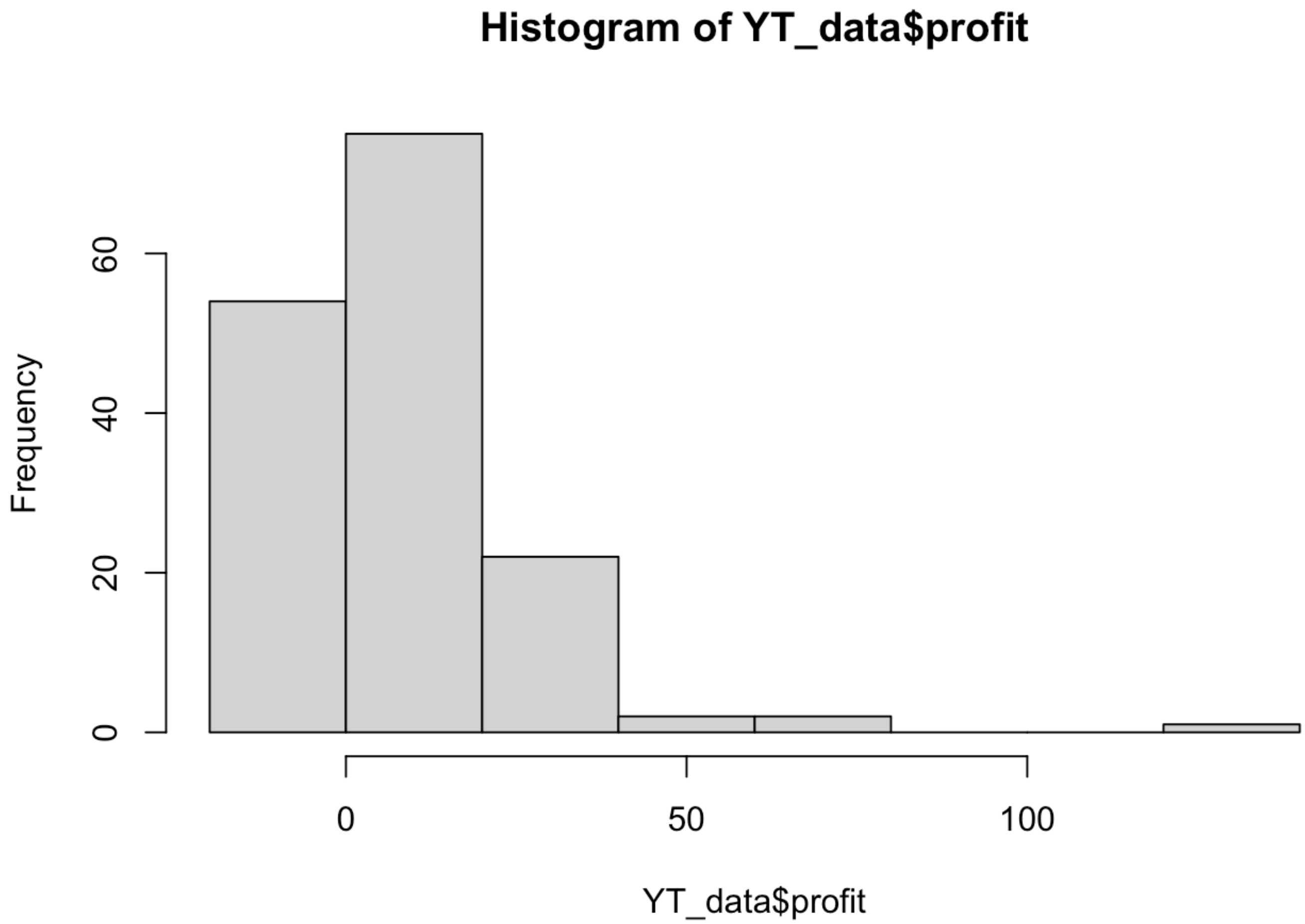

```
hist(Tk_data$profit)
```



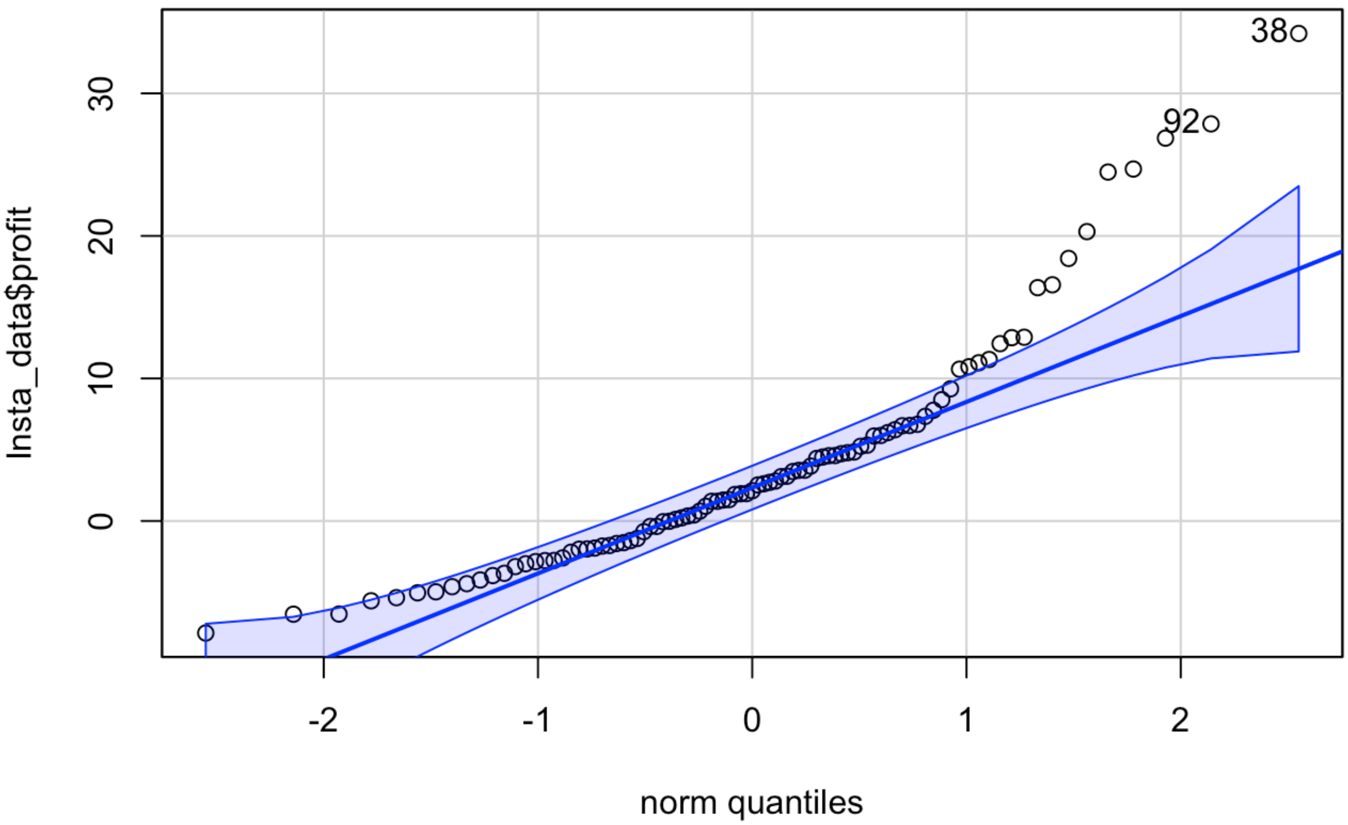

```
hist(Tw_data$profit)
```




```
hist(YT_data$profit)
```




```
qqPlot(Insta_data$profit)
```

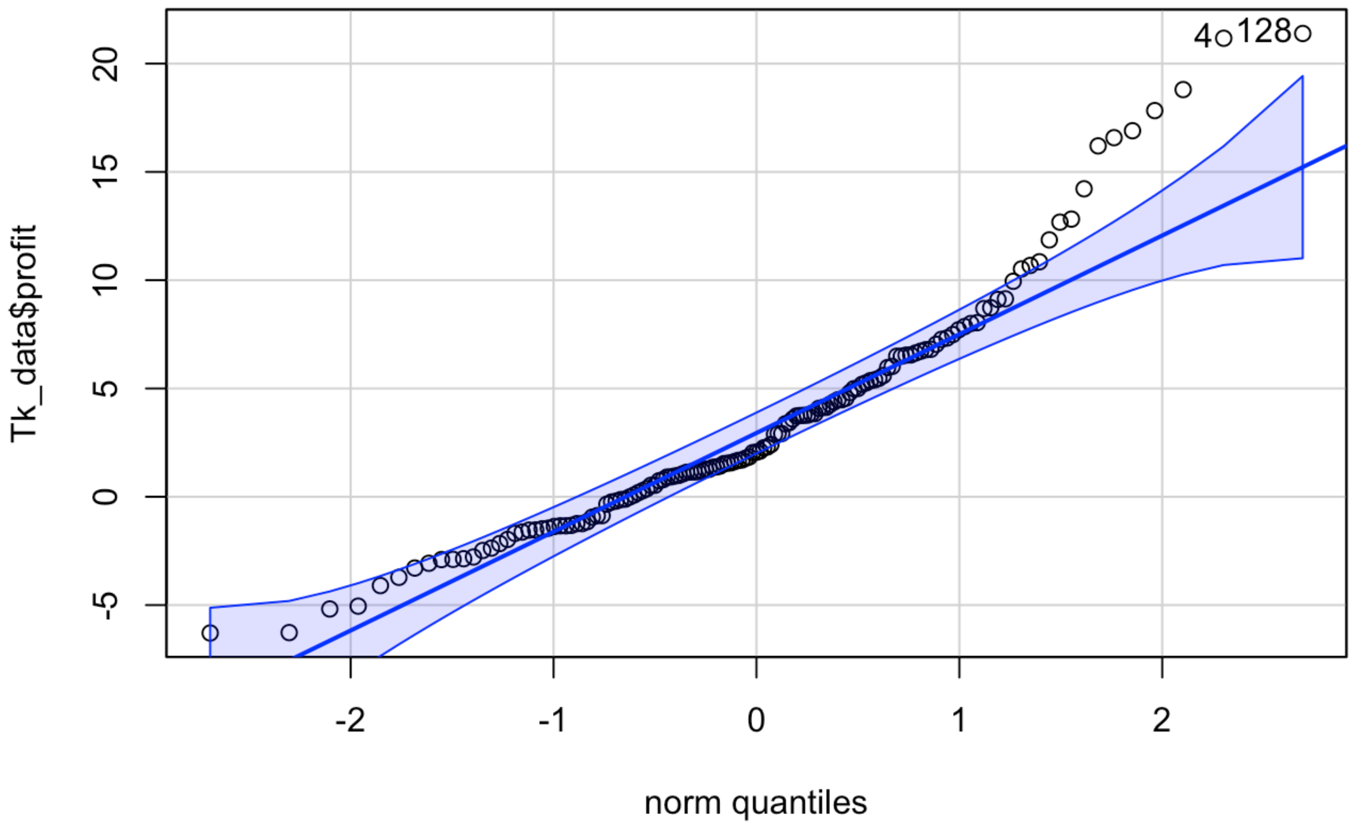



```
## [1] 38 92
```

```
shapiro.test(Insta_data$profit)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Insta_data$profit
## W = 0.87404, p-value = 2.318e-07
```

```
qqPlot(Tk_data$profit)
```




```
## [1] 128   4
```

```
shapiro.test(Tk_data$profit)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Tk_data$profit
## W = 0.92833, p-value = 1.483e-06
```

```
shapiro.test(Tw_data$profit)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Tw_data$profit
## W = 0.93059, p-value = 0.1262
```

```
shapiro.test(YT_data$profit)
```

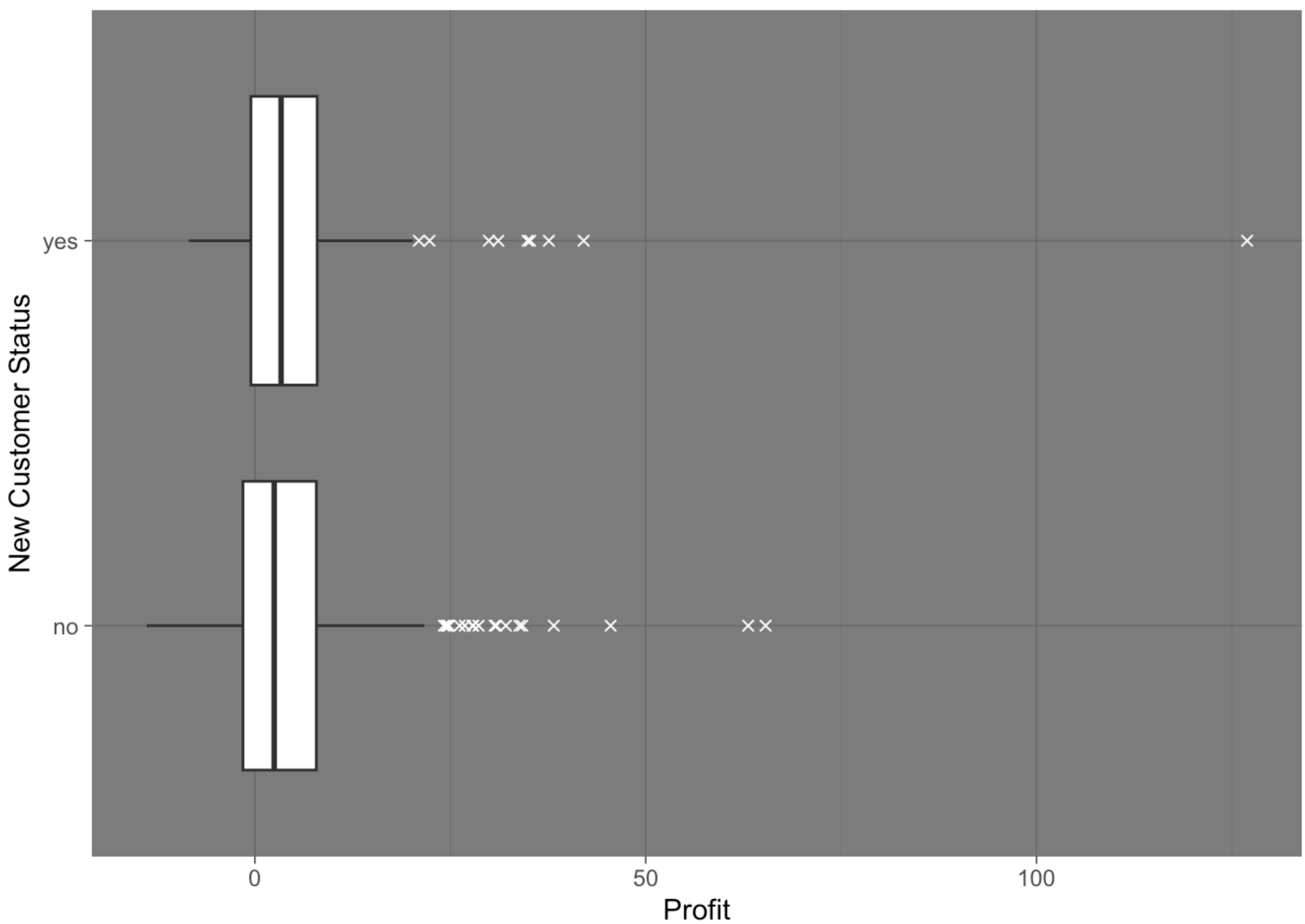
```
##
##  Shapiro-Wilk normality test
##
## data:  YT_data$profit
## W = 0.78247, p-value = 5.999e-14
```

```
## Since the underlying normality assumptions of ANOVA are violated we cannot go ahea
d with ANOVA test.
# We will perform Kruskal-wallis test which is non parametric equivalent of one-way A
NOVA.

## alpha is 0.05
#H0 : μfb = μInsta = μTk = μTw = μYT
#H1 : at least one of the social media platforms has an average profit that is differ
ent from at least one of the other social media platforms.


kruskal.test(df$profit~df$socialmedia)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df$profit by df$socialmedia
## Kruskal-Wallis chi-squared = 7.5755, df = 4, p-value = 0.1084
```

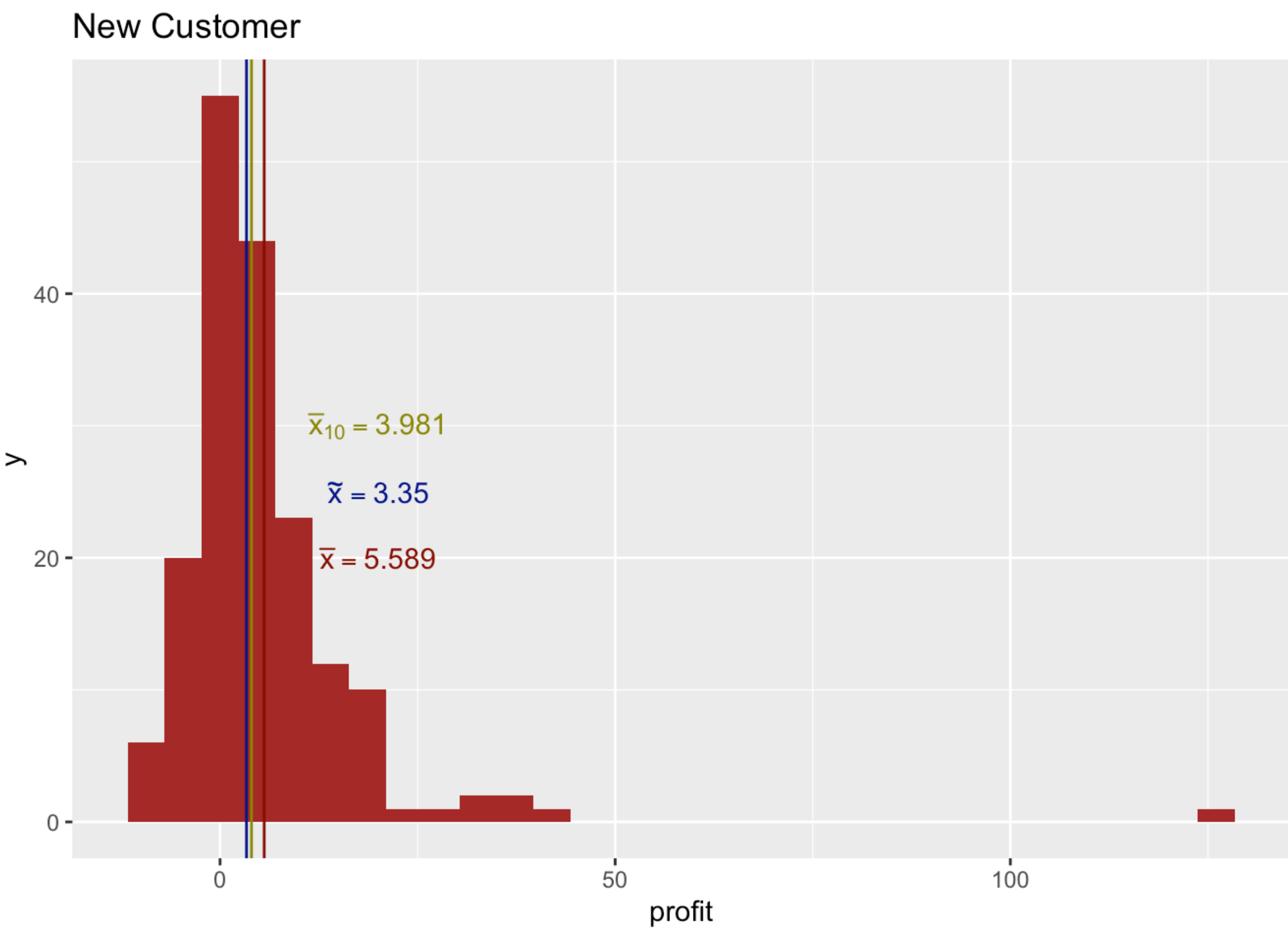
```
## As we can see that the p-value 0.1084 > 0.05 we fail to reject H0.
#  there is a no significant difference in the avg. Profit across the social media pl
atforms.
```

```
#Q4 A####
#Let's do analysis for new customer

new_df<-df %>%
  filter(newcustomer=="yes") %>%
  subset(select= -newcustomer)

old_df<-df %>%
  filter(newcustomer=="no") %>%
  subset(select= -newcustomer)


ggplot(df,aes(x=profit,y=newcustomer))+
  geom_boxplot(outlier.colour ="white" ,outlier.shape = 4)+
  xlab("Profit")+
  ylab("New Customer Status")+
  theme_dark()
```



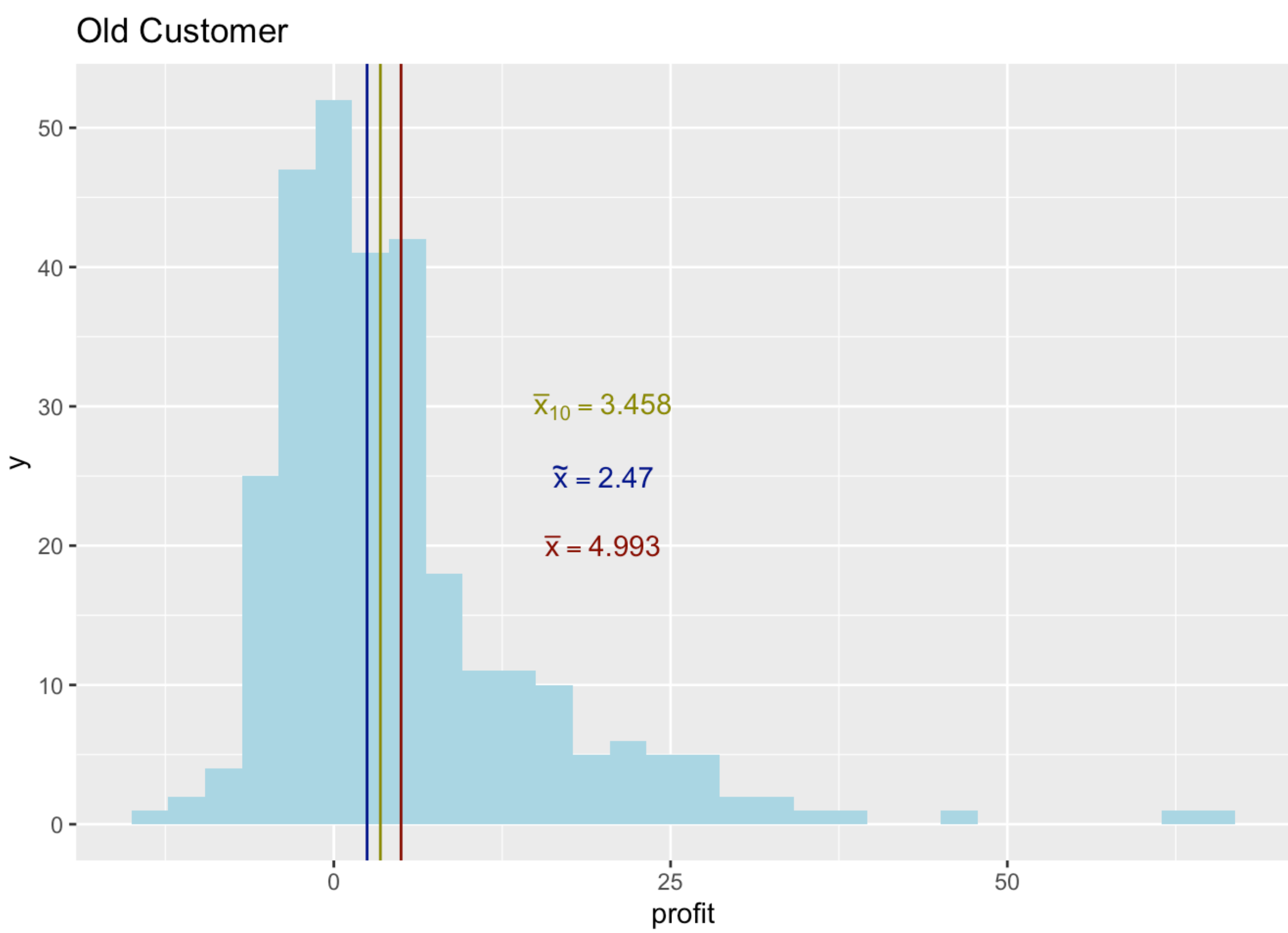
```
new_plt <- ggplot(new_df,aes(x=profit)) +
  ggtitle("New Customer") +
  geom_histogram(fill="brown") +
  geom_vline(aes(xintercept = mean(profit)), color = "darkred") +
  geom_vline(aes(xintercept = median(profit)),color = "darkblue") +
  geom_vline(aes(xintercept = mean(profit, trim = 0.1)),color = "yellow4") +
  annotate("text", x = 20, y = 20, label = paste("bar(x)==",round(mean(new_df$profit)
, 3)), parse = T, color = "darkred") +
  annotate("text", x = 20, y = 25, label = paste("tilde(x)==",round(median(new_df$pro
fit), 3)), parse = T, color = "darkblue") +
  annotate("text", x = 20, y = 30, label = paste("bar(x)[10]==",round(mean(new_df$pro
fit,0.1), 3)), parse = T, color = "yellow4")
new_plt
```

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



```
old_plt <- ggplot(old_df,aes(x=profit)) +
  ggtitle("Old Customer") +
  geom_histogram(fill="lightblue") +
  geom_vline(aes(xintercept = mean(profit)), color = "darkred") +
  geom_vline(aes(xintercept = median(profit)),color = "darkblue") +
  geom_vline(aes(xintercept = mean(profit, trim = 0.1)),color = "yellow4") +
  annotate("text", x = 20, y = 20, label = paste("bar(x)==",round(mean(old_df$profit)
, 3)), parse = T, color = "darkred") +
  annotate("text", x = 20, y = 25, label = paste("tilde(x)==",round(median(old_df$pro
fit), 3)), parse = T, color = "darkblue") +
  annotate("text", x = 20, y = 30, label = paste("bar(x)[10]==",round(mean(old_df$pro
fit,0.1), 3)), parse = T, color = "yellow4")
old_plt
```

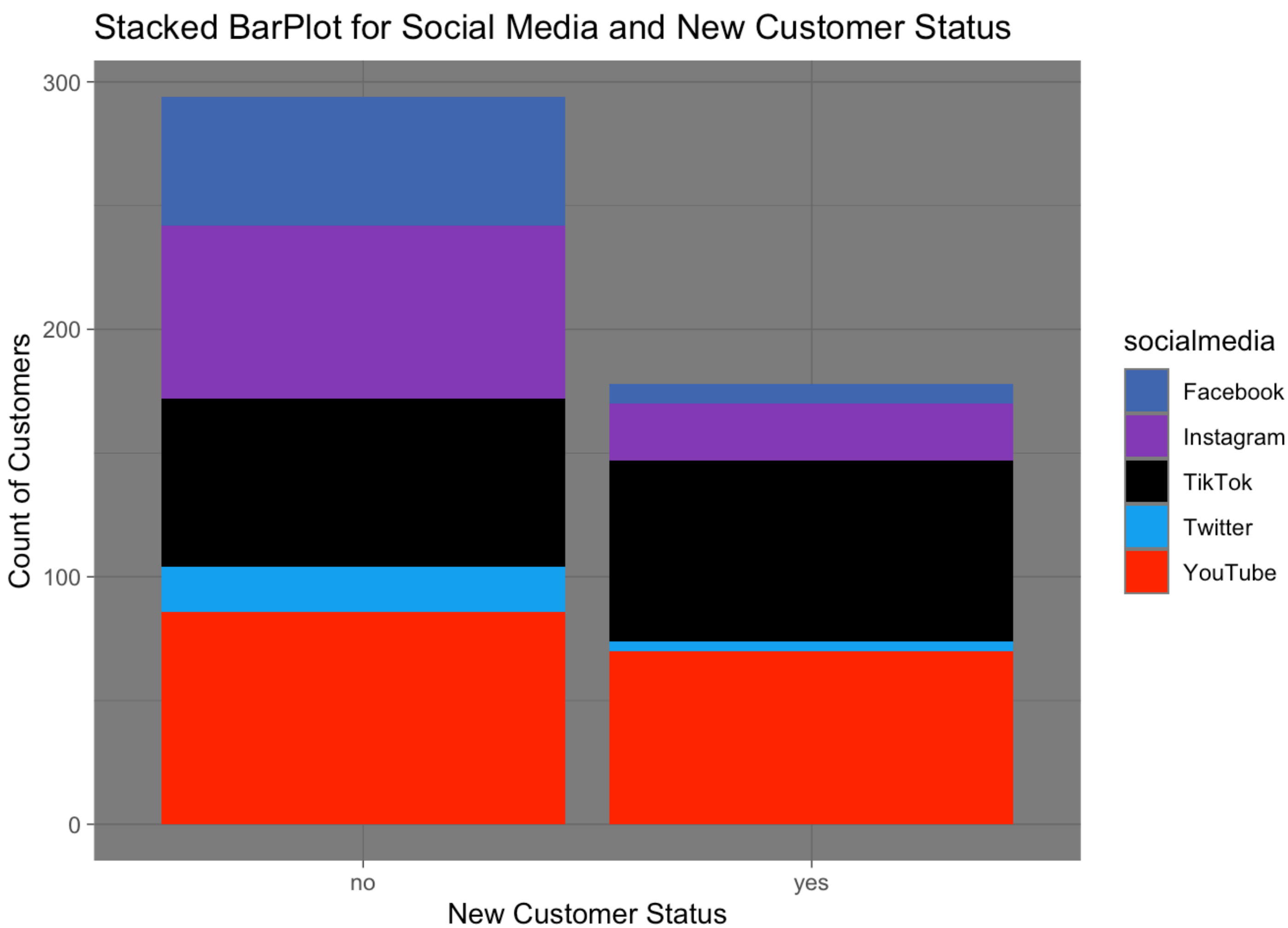
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
pacman::p_load(pivottabler)

pt <- PivotTable$new()
pt$addData(df)
pt$addColumnDataGroups("socialmedia")
pt$addRowDataGroups("newcustomer")
pt$defineCalculation(calculationName="TotalCustomers", summariseExpression="n()")
pt$renderPivot()
```

	Facebook	Instagram	TikTok	Twitter	YouTube	Total
no	52	70	68	18	86	294
yes	8	23	73	4	70	178
Total	60	93	141	22	156	472

```
#f09433 ,#e6683c ,#dc2743 ,#cc2366 ,#bc1888

#colfunc <- colorRampPalette(c("#f09433" ,"#e6683c" ,"#dc2743" ,"#cc2366" ,"#bc1888")
)
#colfunc(10)

ggplot(df, aes(fill = socialmedia, x = newcustomer)) +
  geom_bar(position = "stack", stat = "count") +
  ggtitle("Stacked BarPlot for Social Media and New Customer Status") +
  xlab("New Customer Status") +
  ylab("Count of Customers") +
  #scale_fill_brewer(palette = "Set1") +
  scale_fill_manual(values = c("#4267B2","#833AB4","#000000","#1DA1F2","#FF0000")) +
  theme_dark()
```




```
# Let's do chi squared test of Independence
# H0: Social media platform is independent to rate of acquiring new customers
# H1: Social media platform is associated to rate of acquiring new customers
# alpha = 0.05
```

```
group_by(new_df, socialmedia) %>%
  summarise(
    count = n(),
    mean = mean(adcost, na.rm = TRUE),
    sd = sd(adcost, na.rm = TRUE)
  )
```

```
## # A tibble: 5 × 4
##   socialmedia count  mean    sd
##   <chr>       <int> <dbl> <dbl>
## 1 Facebook        8  5.00  1.31
## 2 Instagram      23  5.86  1.75
## 3 TikTok         73  3.97  2.37
## 4 Twitter         4  2.88  1.27
## 5 YouTube        70  9.62  2.73
```

```
#### Q4 b) ####

new_cust_fb<-fb_data[fb_data$newcustomer == 'yes',]
new_cust_Insta<-Insta_data[Insta_data$newcustomer == 'yes',]
new_cust_Tk<-Tk_data[Tk_data$newcustomer == 'yes',]
new_cust_Tw<-Tw_data[Tw_data$newcustomer == 'yes',]
new_cust_YT<-YT_data[YT_data$newcustomer == 'yes',]


shapiro.test(new_cust_fb$adcost)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  new_cust_fb$adcost
## W = 0.88236, p-value = 0.1983
```

```
shapiro.test(new_cust_Insta$adcost)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  new_cust_Insta$adcost
## W = 0.94459, p-value = 0.2252
```

```
shapiro.test(new_cust_Tk$adcost)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  new_cust_Tk$adcost
## W = 0.94761, p-value = 0.004313
```

```
shapiro.test(new_cust_Tw$adcost)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  new_cust_Tw$adcost
## W = 0.78649, p-value = 0.08012
```

```
shapiro.test(new_cust_YT$adcost)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  new_cust_YT$adcost
## W = 0.96314, p-value = 0.03741
```

```
## Let's check for homoscedasticity

bartlett.test(adcost ~ socialmedia, data=new_df)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  adcost by socialmedia
## Bartlett's K-squared = 10.731, df = 4, p-value = 0.02975
```

```
##In the above case we see that p=0.002975 < 0.05, thus for bartlett test we reject N
ULL Hypothesis.
## therefore homoscedasticity assumption doesn't hold true
## We will proceed with Kruskal walis because normality , homoscedasticity doesn't ho
ld

#alpha is 0.05

#Ho: new customer's ad cost is same for all social media platform(µfb=µTk=µTw=µInst=µ
YT)
#H1: new customer's ad cost is different for at least one platform

kruskal.test(new_df$adcost~new_df$socialmedia)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  new_df$adcost by new_df$socialmedia
## Kruskal-Wallis chi-squared = 99.262, df = 4, p-value < 2.2e-16
```

```
# Since p-value is less than 0.05 we reject NULL Hypothesis.
# We can conclude that different rates are associated with aquiring new customers acr
oss various social media platforms

#This question is to check for interaction between SocialMedia and NewCustomers
# This approach is applicable if we are considering count of newcustomers

pt2 <- PivotTable$new()
pt2$addData(df)
pt2$addColumnDataGroups("socialmedia")
pt2$addRowDataGroups("newcustomer")
pt2$defineCalculation(calculationName="count", summariseExpression="n()")
pt2$renderPivot()
```

	Facebook	Instagram	TikTok	Twitter	YouTube	Total
no	52	70	68	18	86	294
yes	8	23	73	4	70	178
Total	60	93	141	22	156	472

```
x <- matrix(c(52,8,70,23,68,73,18,4,86,70),nrow = 2,ncol = 5)
x
```

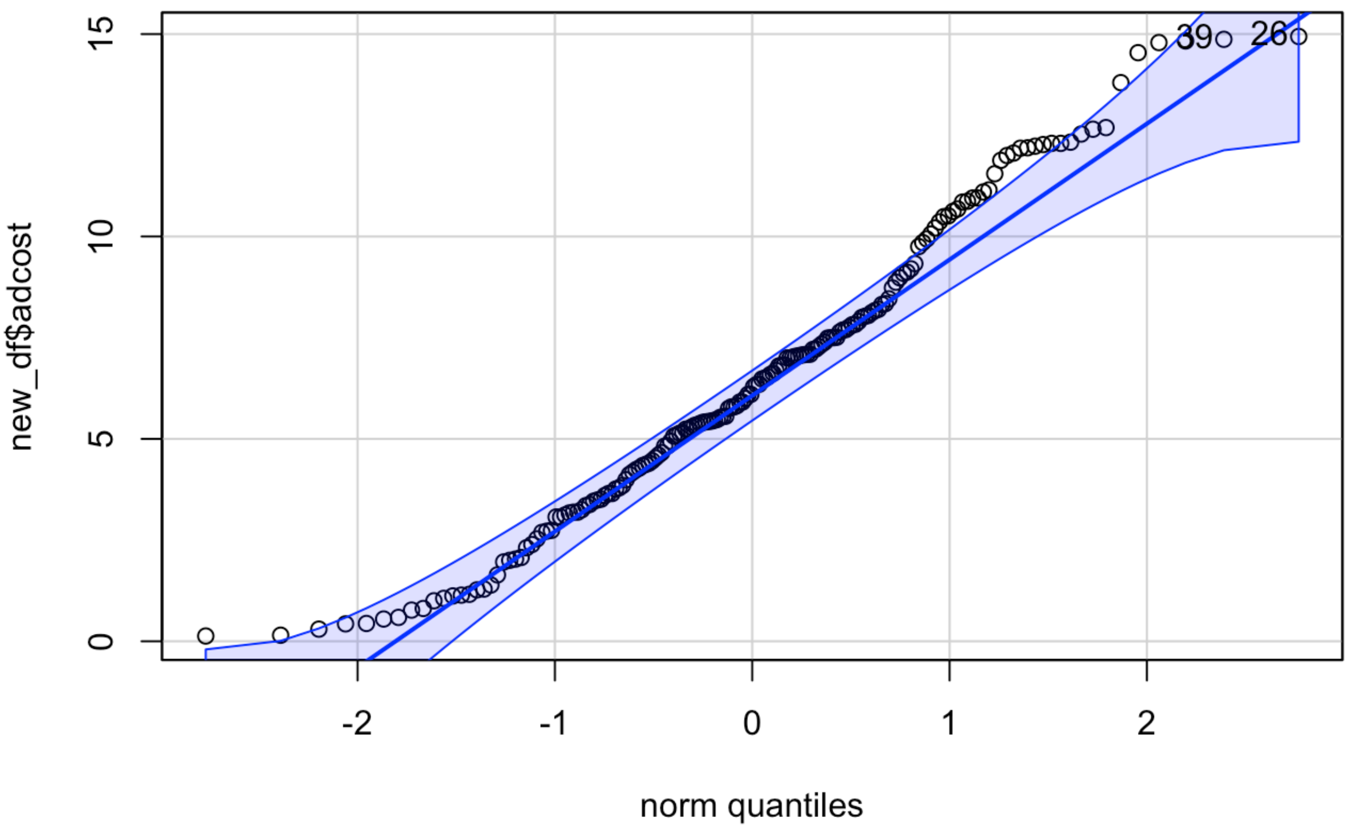
```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]   52   70   68   18   86
## [2,]    8   23   73    4   70
```

```
chisq.test(x,correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  x
## X-squared = 40.696, df = 4, p-value = 3.106e-08
```

```
# p value is less than 0.05 therefore we reject Null Hypothesis.
# This approach concluded->
#  that different rates are associated with aquiring new customers across various soc
ial media platforms
```

```
library("car")
qqPlot(new_df$adcost)
```




```
## [1] 26 39
```

```
#### Q4 c) ####
n=length(df$newcustomer)
x=length(new_df$adcost)
p<-x/n
#check normality assumptions
cat("check normality assumptions:\n")
```

```
## check normality assumptions:
```

```
cat("n*p>=5:",n*p>=5)
```

```
## n*p>=5: TRUE
```

```
cat("\nn*(1-p)>=5:",n*(1-p)>=5)
```

```
##
## n*(1-p)>=5: TRUE
```

```
q<-1-p
z_alpha<-1.96
CI_upper<-(p+(z_alpha*sqrt(p*q/n)))
CI_lower<-(p-(z_alpha*sqrt(p*q/n)))
cat("\nCI:(",CI_lower,",",CI_upper,")")
```

```
##
## CI:( 0.333394 , 0.4208433 )
```

```
cat("We are 95% confident that the interval ( 0.333394 , 0.4208433 ) contains the tru
e population proportion of ads that lead to new customer")
```

```
## We are 95% confident that the interval ( 0.333394 , 0.4208433 ) contains the true
population proportion of ads that lead to new customer
```

```
#approach 2
prop.test(x,n,correct=FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  x out of n, null probability 0.5
## X-squared = 28.508, df = 1, p-value = 9.329e-08
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3345523 0.4216690
## sample estimates:
##         p
## 0.3771186
```

```
binom.test(x,n,0.5,conf.level = 0.95)
```

```
##
##  Exact binomial test
##
## data:  x and n
## number of successes = 178, number of trials = 472, p-value = 1.043e-07
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.3332250 0.4225623
## sample estimates:
## probability of success
##              0.3771186
```

```
#### Q4 d) ####

# alpha is 0.05

#H0:new_cust$profit <= exist_cust$profit
#H1:new_cust$profit > exist_cust$profit
#check the variance of the samples

#existing_cust<-df[df$newcustomer == 'no',]
#ne<-length(existing_cust$newcustomer)
#nn<-length(new_cust$newcustomer)
shapiro.test(df$profit)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$profit
## W = 0.73561, p-value < 2.2e-16
```

```
shapiro.test(new_df$profit)
```

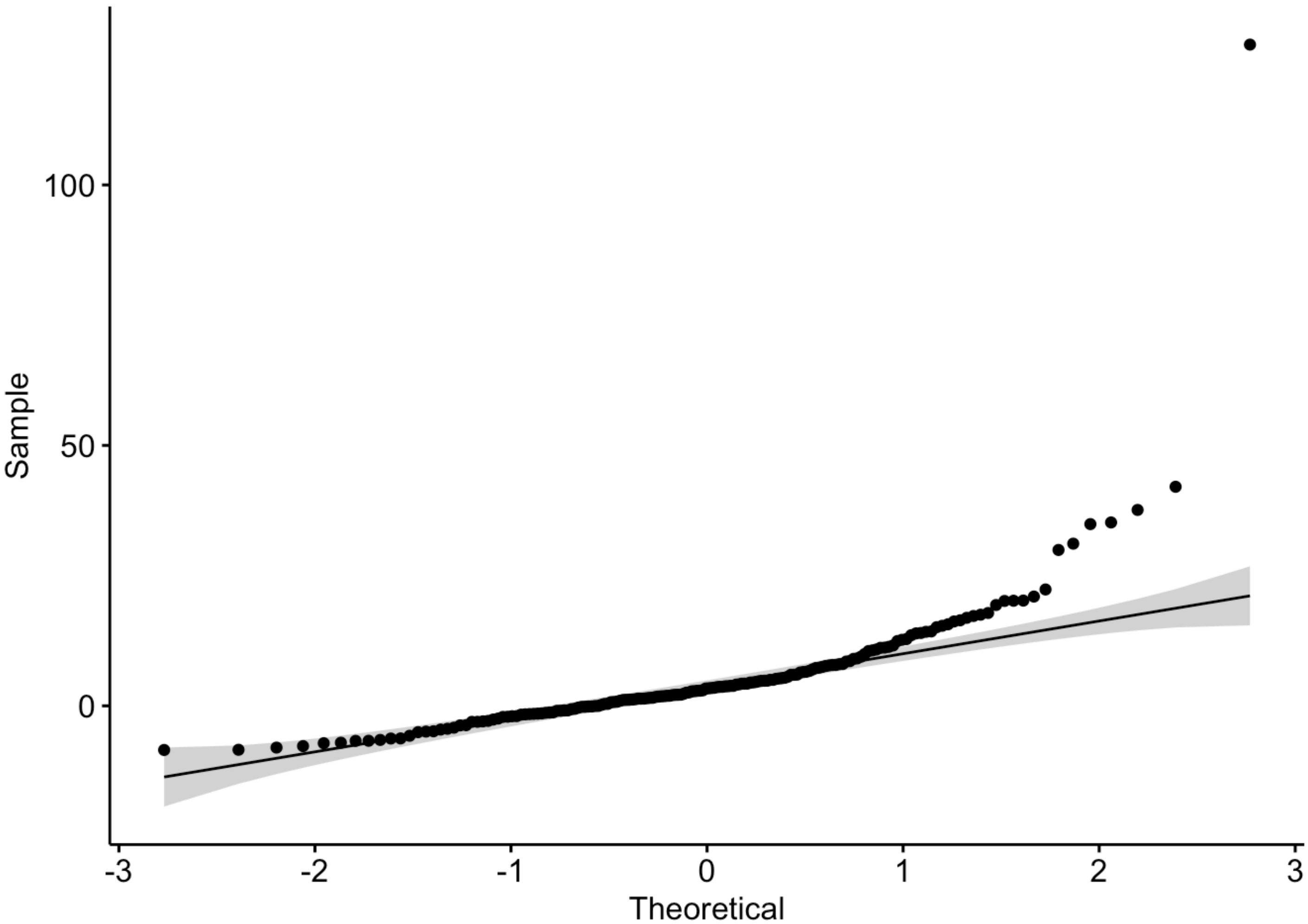
```
##
##  Shapiro-Wilk normality test
##
## data:  new_df$profit
## W = 0.61706, p-value < 2.2e-16
```

```
shapiro.test(old_df$profit)
```

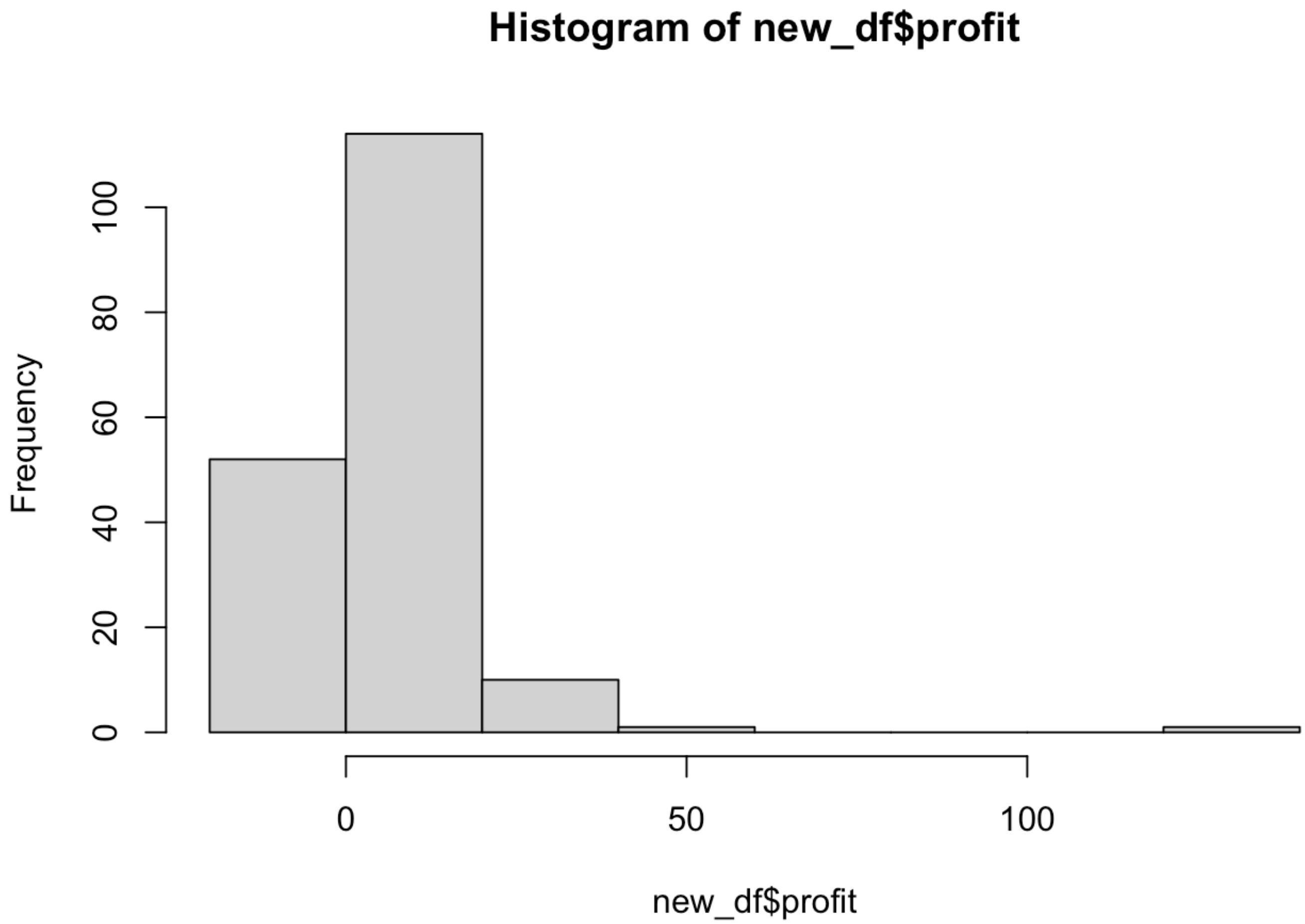
```
##
##  Shapiro-Wilk normality test
##
## data:  old_df$profit
## W = 0.82822, p-value < 2.2e-16
```

```
## Data not normally distributed

ggqqplot(new_df$profit)
```

```
## Warning: The following aesthetics were dropped during statistical transformation:
sample
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
## The following aesthetics were dropped during statistical transformation: sample
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```




```
hist(new_df$profit)
```

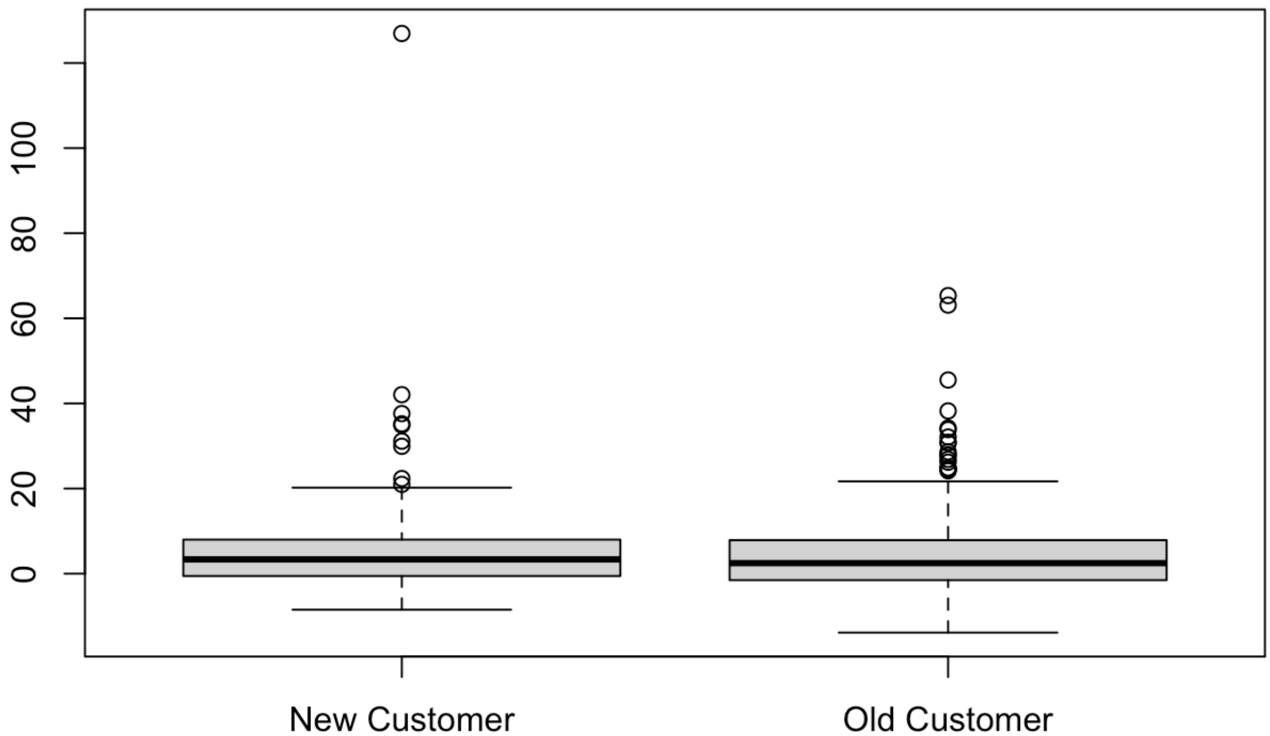



```
wilcox.test(new_df$profit,old_df$profit,alternative = "greater",exact = F,correct = F
)
```

```
##
##  Wilcoxon rank sum test
##
## data:  new_df$profit and old_df$profit
## W = 27502, p-value = 0.176
## alternative hypothesis: true location shift is greater than 0
```

```
## We fail to reject NULL Hypothesis.
## We can conclude that we don't have enough evidence to validate the analysts claim
that acquiring new customers is more profitable than trying to sell more products to
existing customers.
```

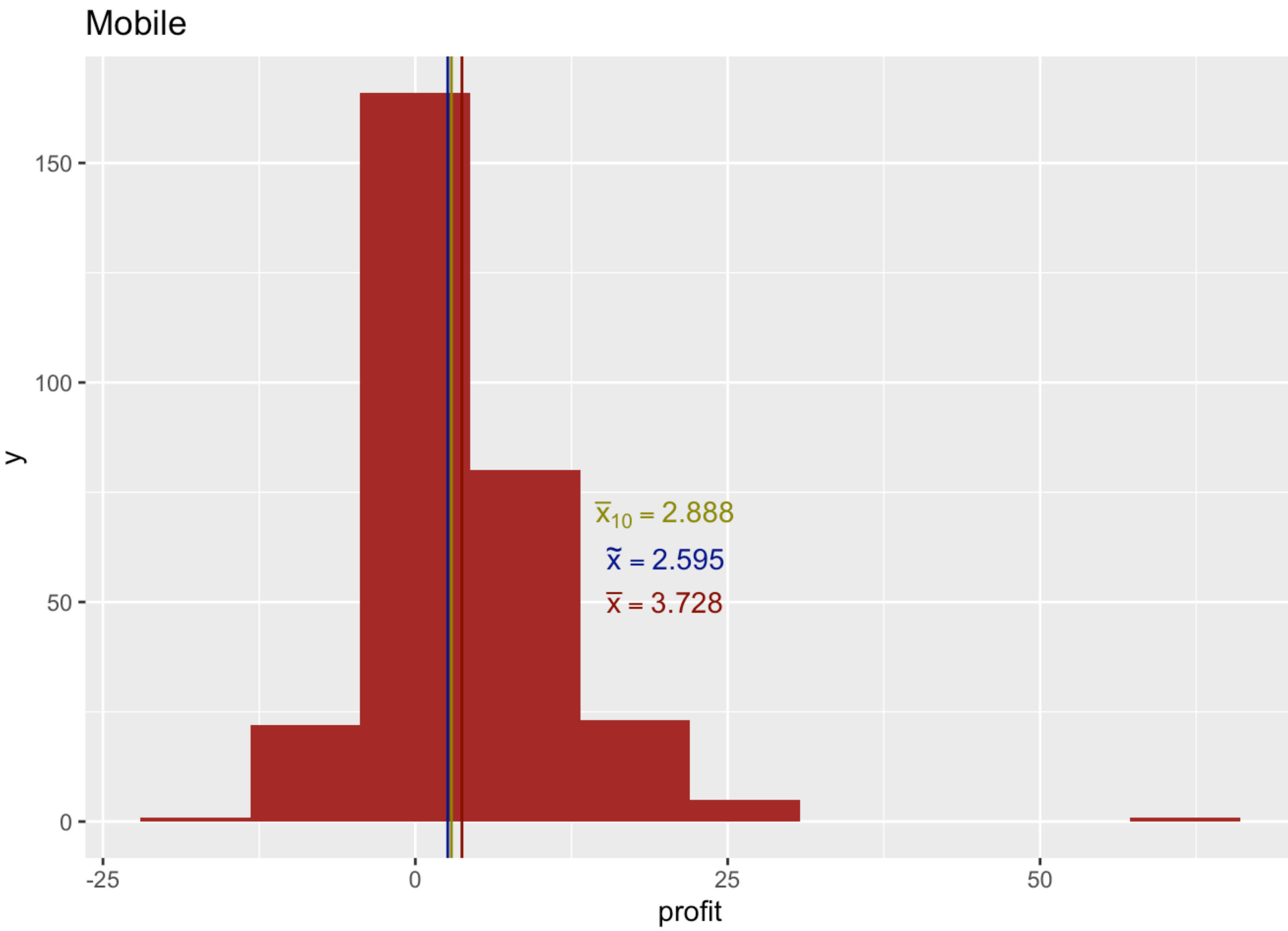
```
boxplot(new_df$profit, old_df$profit, names=c("New Customer","Old Customer"))
```




```
#since varinace are unequal, we choose Welch's test
t.test(new_df$profit, old_df$profit,alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  new_df$profit and old_df$profit
## t = 0.52988, df = 318.94, p-value = 0.2983
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -1.258935       Inf
## sample estimates:
## mean of x mean of y
##  5.588596  4.992857
```

```
#### Q5 a) ####
mob_df<-df %>%
  filter(mobile=="mobile") %>%
  subset(select= -mobile)

comp_df<-df %>%
  filter(mobile!="mobile") %>%
  subset(select= -mobile)

bin_mob_plt <- ceiling(log(length(mob_df$profit), 2)) + 1
bin_com_plt <- ceiling(log(length(comp_df$profit), 2)) + 1

mob_plt <- ggplot(mob_df,aes(x=profit)) +
  ggtitle("Mobile") +
  geom_histogram(fill="brown",bins = bin_mob_plt) +
  geom_vline(aes(xintercept = mean(profit)), color = "darkred") +
  geom_vline(aes(xintercept = median(profit)),color = "darkblue") +
  geom_vline(aes(xintercept = mean(profit, trim = 0.1)),color = "yellow4") +
  annotate("text", x = 20, y = 50, label = paste("bar(x)==",round(mean(mob_df$profit)
, 3)), parse = T, color = "darkred") +
  annotate("text", x = 20, y = 60, label = paste("tilde(x)==",round(median(mob_df$pro
fit), 3)), parse = T, color = "darkblue") +
  annotate("text", x = 20, y = 70, label = paste("bar(x)[10]==",round(mean(mob_df$pro
fit,0.1), 3)), parse = T, color = "yellow4")
mob_plt
```



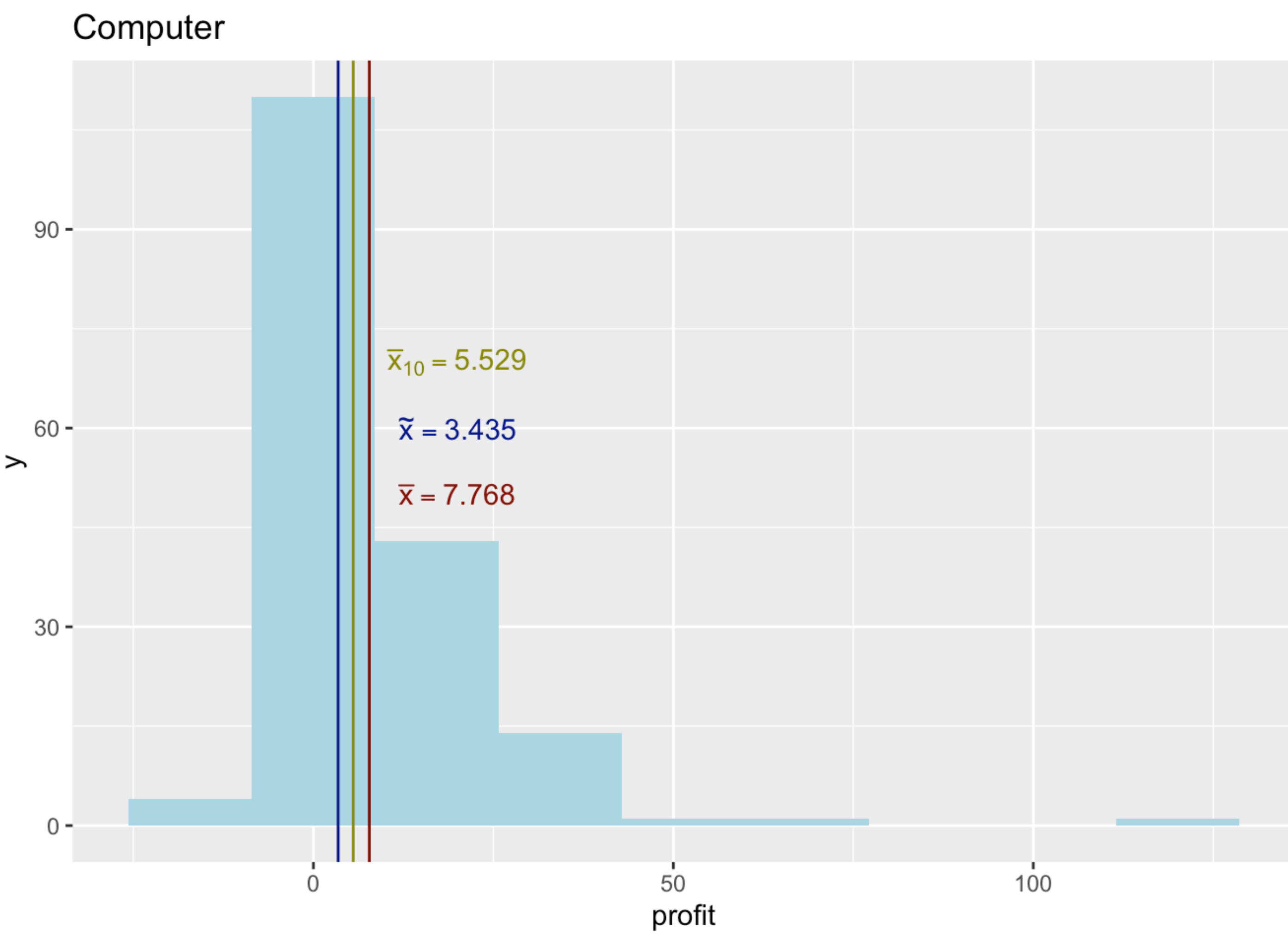

```
skewness(mob_df$profit)
```

```
## [1] 2.420909
```

```
skewness(comp_df$profit)
```

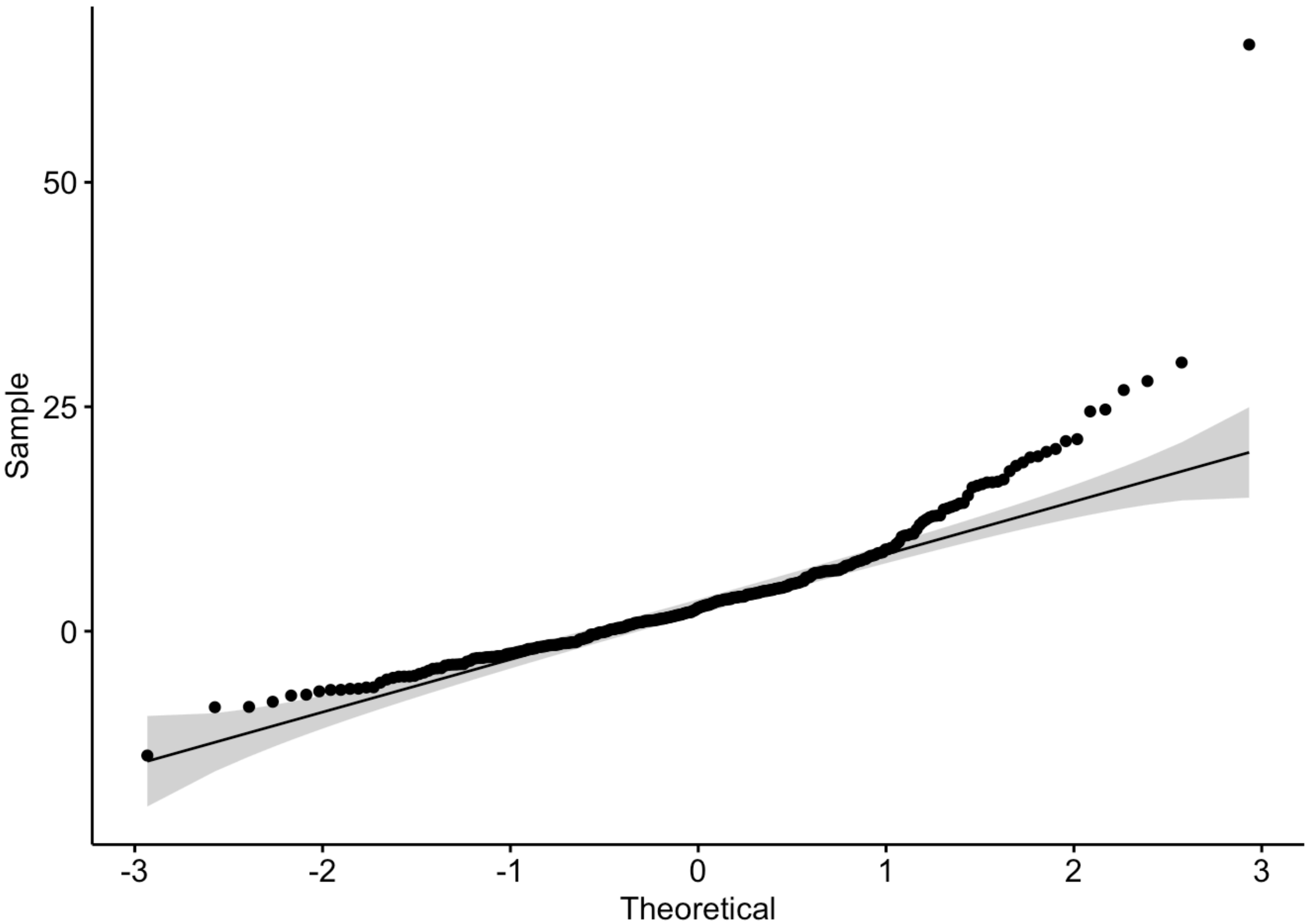
```
## [1] 3.374315
```

```
comp_plt <- ggplot(comp_df,aes(x=profit)) +
  ggtitle("Computer") +
  geom_histogram(fill="lightblue",bins=bin_com_plt) +
  geom_vline(aes(xintercept = mean(profit)), color = "darkred") +
  geom_vline(aes(xintercept = median(profit)),color = "darkblue") +
  geom_vline(aes(xintercept = mean(profit, trim = 0.1)),color = "yellow4") +
  annotate("text", x = 20, y = 50, label = paste("bar(x)==",round(mean(comp_df$profit
), 3)), parse = T, color = "darkred") +
  annotate("text", x = 20, y = 60, label = paste("tilde(x)==",round(median(comp_df$pr
ofit), 3)), parse = T, color = "darkblue") +
  annotate("text", x = 20, y = 70, label = paste("bar(x)[10]==",round(mean(comp_df$pr
ofit,0.1), 3)), parse = T, color = "yellow4")
comp_plt
```



```r
#### Q5 b) ####

# Checking for interaction between Social media and mobile phone

# we can check for normality first visually by qq plots

ggqqplot(mob_df$profit)
```

```
## Warning: The following aesthetics were dropped during statistical transformation:
sample
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
## The following aesthetics were dropped during statistical transformation: sample
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```




```
#not quite sure if it's a normal distribution or not.
# which normality test to use Kolmogorov-Smirnov (K-S) normality test and Shapiro-Wil
k’s test?
# conducting shapiro test at alpha 0.05 with H0:sample distribution is normal.
shapiro.test(mob_df$profit)
```

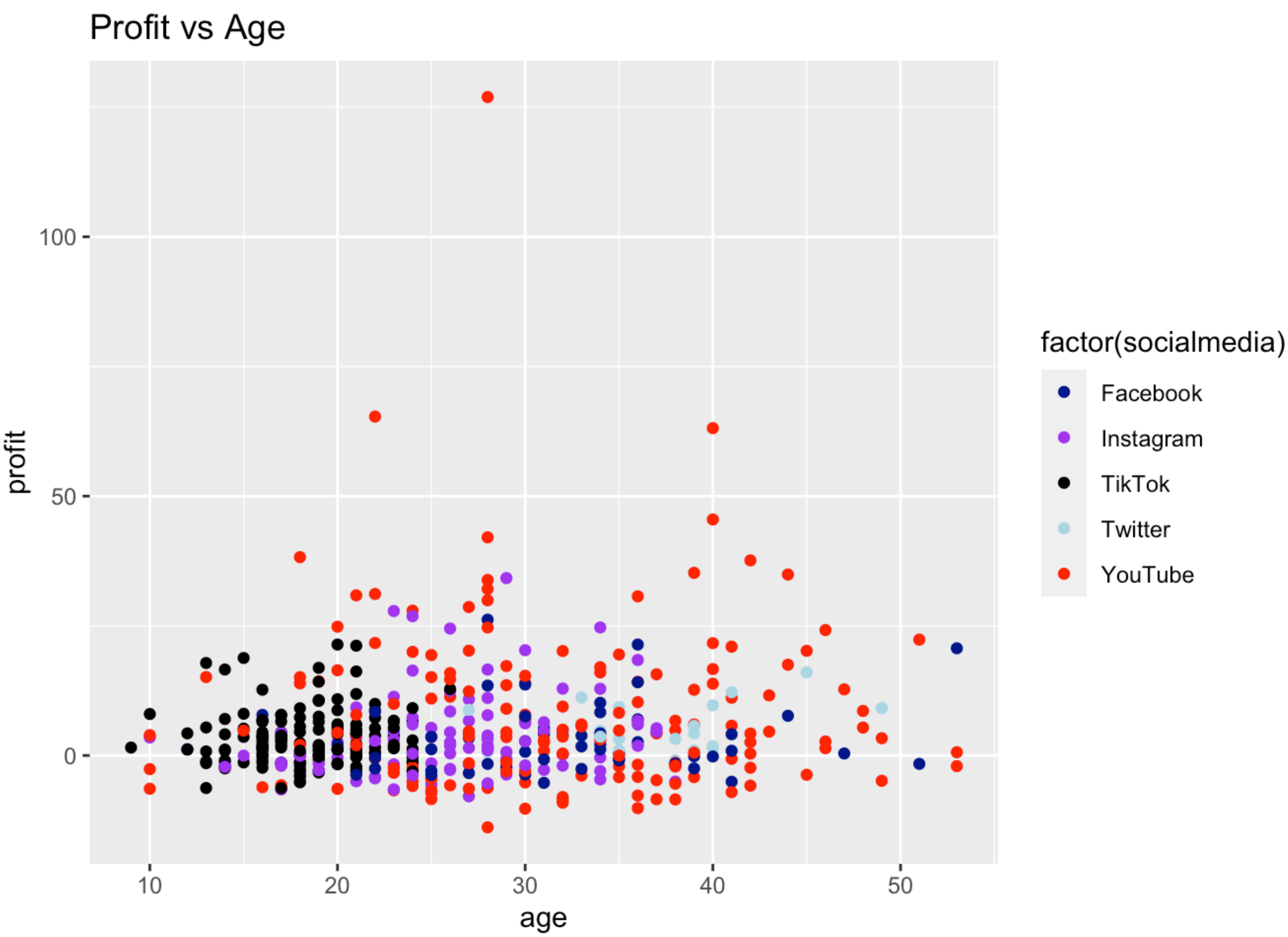
```
##
##  Shapiro-Wilk normality test
##
## data:  mob_df$profit
## W = 0.85035, p-value = 2.527e-16
```

```
# p value is less than 0.05 therefore we reject null hypothesis.
# normality doesn't hold in this scenario.
# hence we will use Wilcoxon Rank-Sum Test (Mann-Whitney U Test)
# H0 : The medians of the two populations are identical meaning profit doesn't depend
on whether we are on computer or mobile.

# Conducting Wilcoxon Rank-Sum test at alpha 0.05
wilcox.test(mob_df$profit,comp_df$profit,exact = F,correct = F)
```

```
##
##  Wilcoxon rank sum test
##
## data:  mob_df$profit and comp_df$profit
## W = 23293, p-value = 0.06551
## alternative hypothesis: true location shift is not equal to 0
```

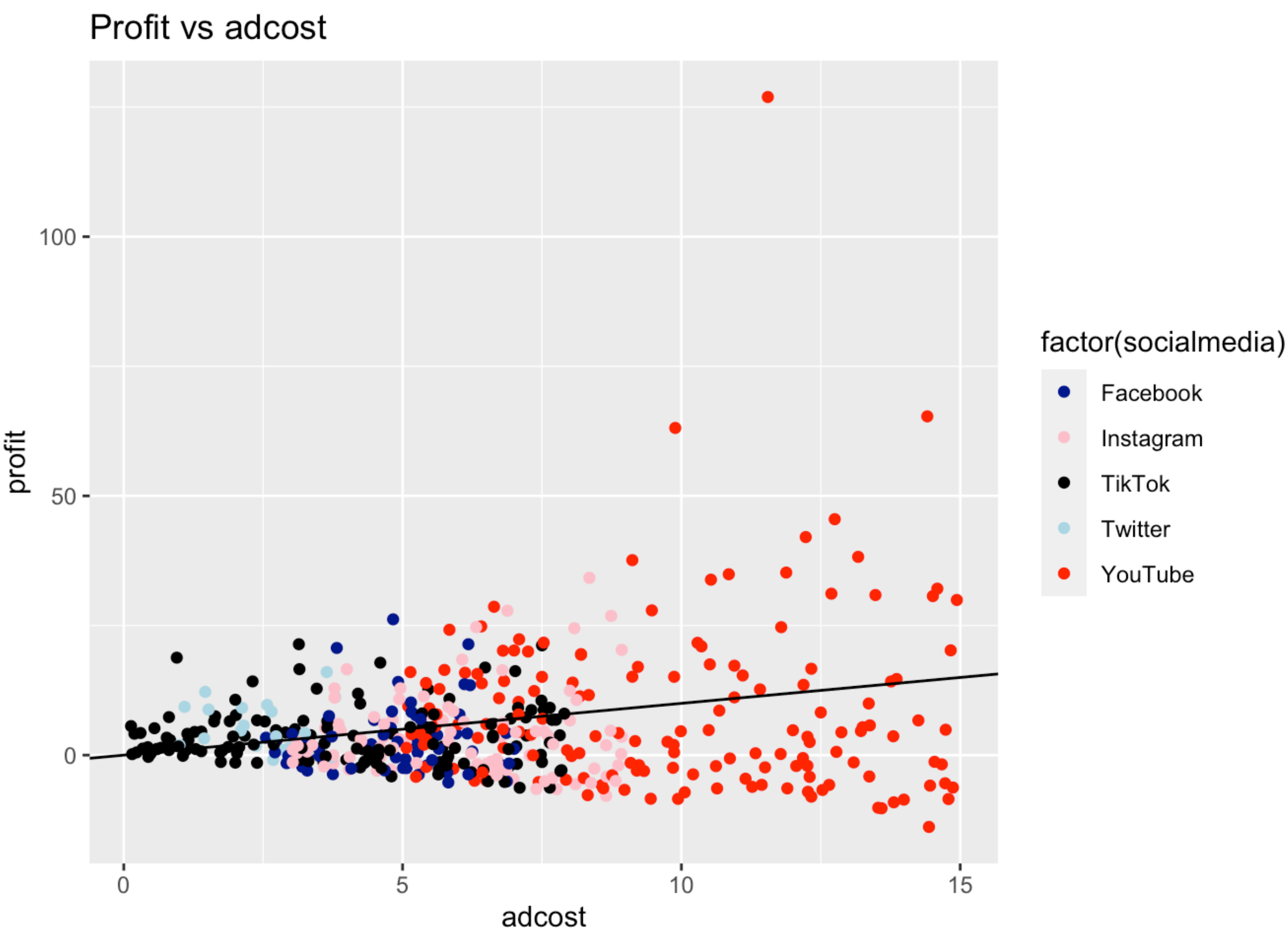
```
# since p value 0.06551 > 0.05 we fail to reject H0.
# hence Profit doesn't depend on whether we are on mobile or not.
```

```
#### Q6 a) ####
# Create the scatter plot
ggplot(df, aes(x = age, y =profit)) +
  ggtitle("Profit vs Age") +
  geom_point(aes(color = factor(socialmedia)))+ scale_color_manual(values = c(" dark
blue", " purple", "black","light blue","red"))
```




```
## We can observe that TikTok has young audience of less than 30 and the profit is al
so less.
## But tiktok doesn't give huge profits unlike YouTube which has few outliers.
## YouTube is used by all age groups
```

```
#### Q6 b) ####

ggplot(df, aes(x = adcost, y = profit)) +
  ggtitle("Profit vs adcost") +
  geom_point(aes(color = factor(socialmedia)))+ scale_color_manual(values = c("dark b
lue", " pink", "black","light blue","red")) + geom_abline()
```




```
## We observe Heteroscedasticity from the scatter plot
## Although it is a weak, there is a positive correlation between profit and adcost
## we noticed that YouTube's adcost is highest whereas, TikTok's is lowest
```

```
#### Q6 c) ####
# alpha is 0.05
# H0 = rho(Age, Profit) = 0
# H1 = rho(Age, Profit) !=0


cor.test(df$profit, df$age, method = "pearson", alternative = "two.sided", conf.level
= 0.95)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$profit and df$age
## t = 2.5263, df = 470, p-value = 0.01186
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.02575751 0.20387152
## sample estimates:
##       cor
## 0.1157449
```

```
# since p-value = 0.01 which is less than 0.05. we can reject null hypothesis.
# We can say that there is enough evidence to conclude that Age and Profit are correl
ated.
```

```
#### Q6 d) ####
# alpha is 0.05

# H0 = rho(Profit,adcost)=0
# H1 = rho(Profit,adcost)!=0

cor.test(df$profit, df$adcost, method = "spearman", alternative = "two.sided", conf.l
evel = 0.95,exact=FALSE)
```

```
##
##  Spearman's rank correlation rho
##
## data:  df$profit and df$adcost
## S = 17691357, p-value = 0.8376
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##          rho
## -0.009458226
```

```
# since p-value = 0.8376 which is more than 0.05 we fail to reject null Hypothesis.
# there is not enough evidence to conclude that there is a significant correlation be
tween Profit and adcost.
```

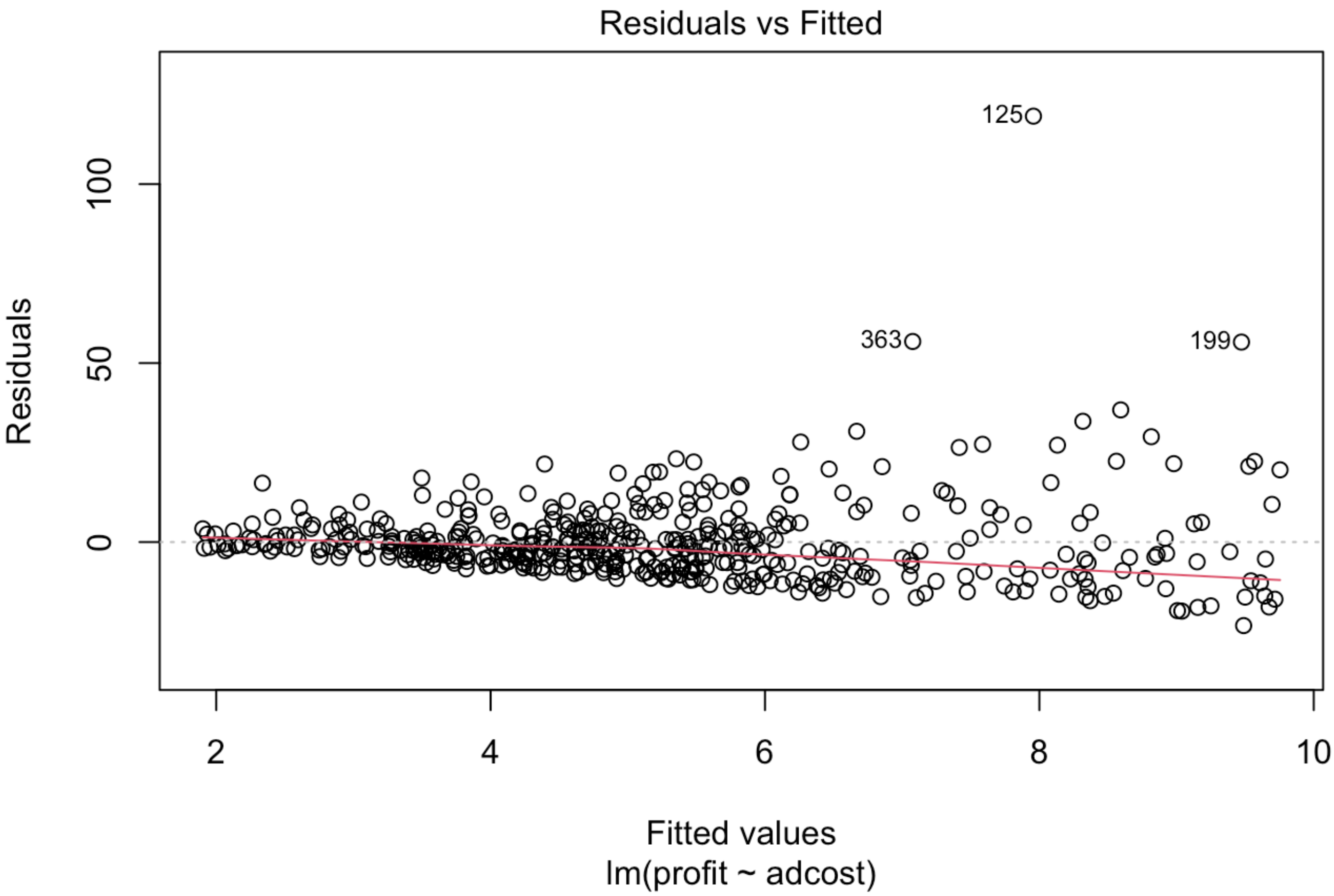
```
#### Q6 e) ####

## we know that simple linear regression hold multiple assumptions
## Shapiro test to check for normality
shapiro.test(df$profit)
```

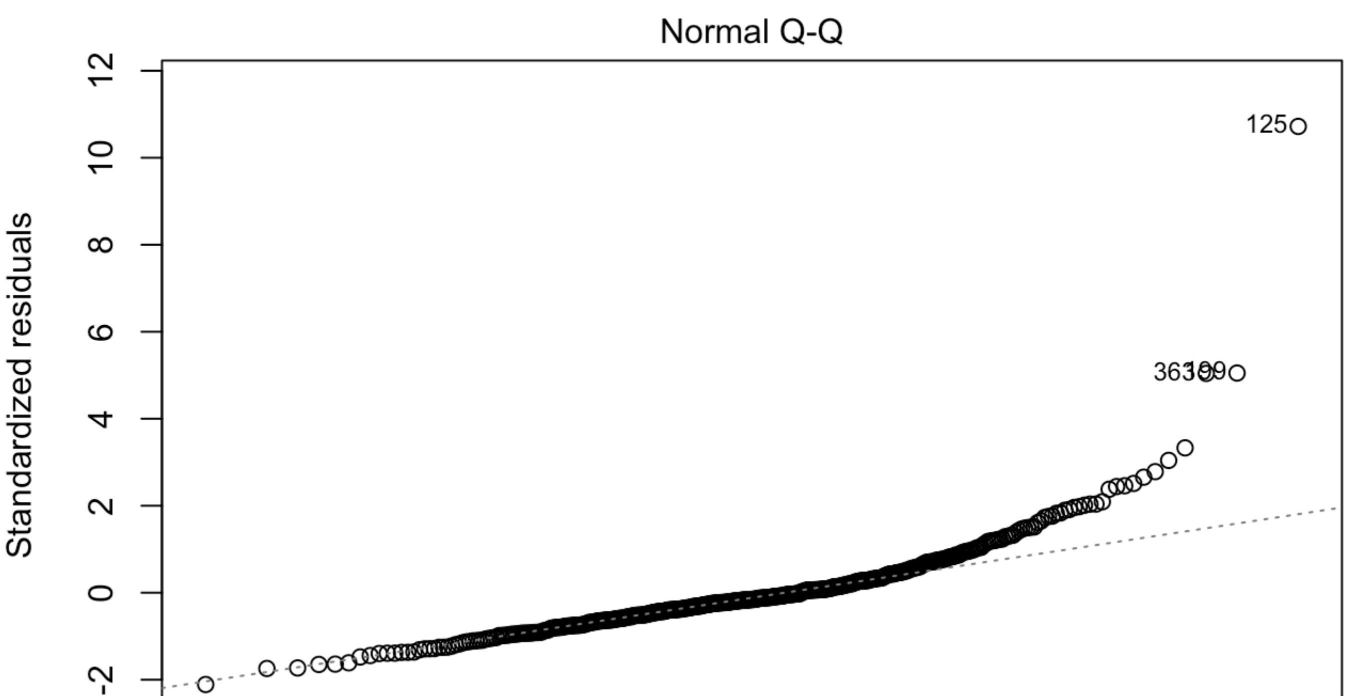
```
##
##  Shapiro-Wilk normality test
##
## data:  df$profit
## W = 0.73561, p-value < 2.2e-16
```

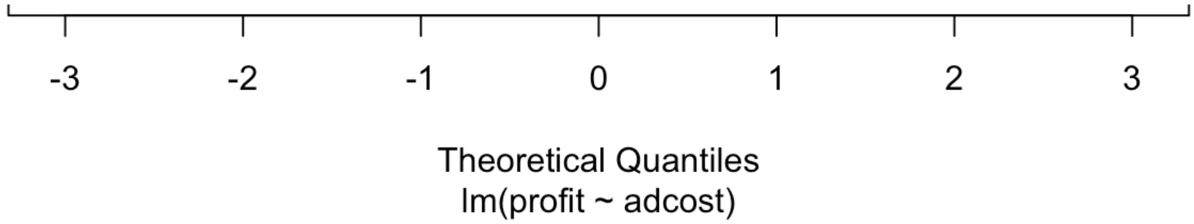
```
## Since y is not normally distributed for x we conclude that linear regression might not be optimal.
linear_model <- lm(profit ~ adcost, data = df)
summary(linear_model)
```

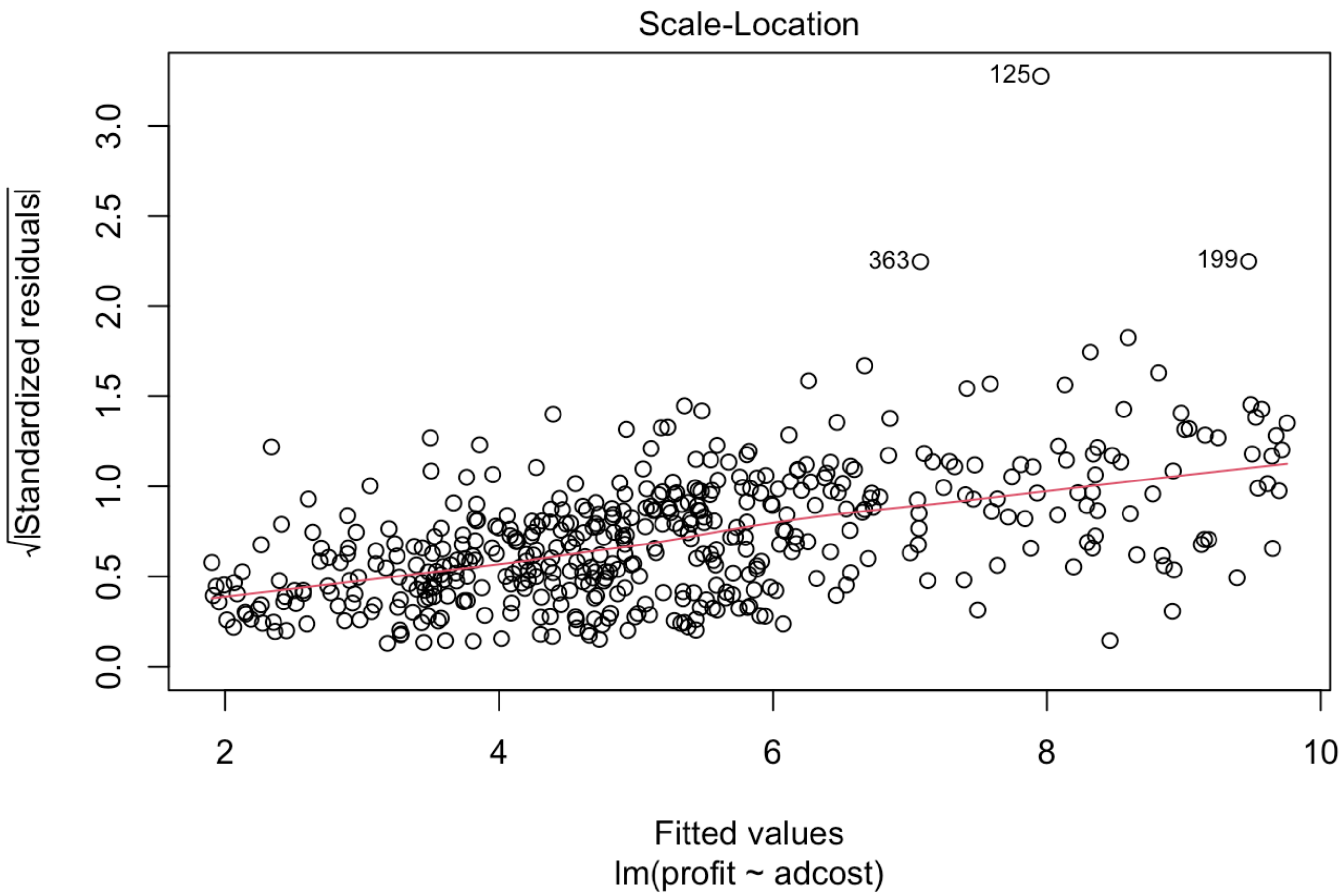
```
##
## Call:
## lm(formula = profit ~ adcost, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.339  -5.977  -1.722   3.403 118.983
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.8330     1.0736   1.707 0.088421 .
## adcost        0.5302     0.1478   3.588 0.000368 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.14 on 470 degrees of freedom
## Multiple R-squared:  0.02666,    Adjusted R-squared:  0.02459
## F-statistic: 12.87 on 1 and 470 DF,  p-value: 0.0003684
```

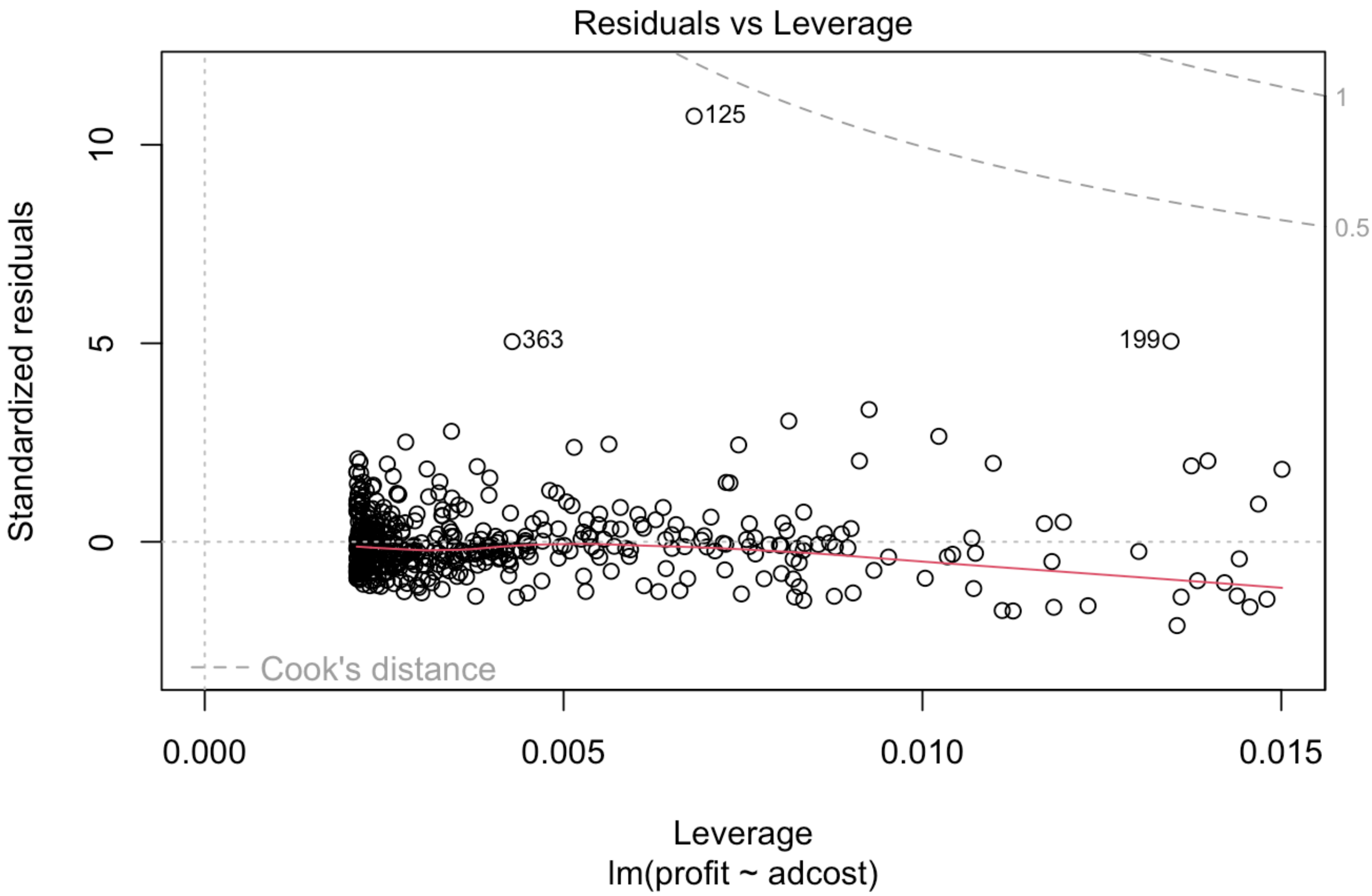
```
plot(linear_model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(profit ~ adcost)
125
363
199
100
50
0
2
4
6
8
10


Normal Q-Q
Standardized residuals
125
363
199
12
10
8
6
4
2
0
-2


-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(profit ~ adcost)


Scale-Location
125
363
199
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2
4
6
8
10
Fitted values
lm(profit ~ adcost)




```
## it can be observed that values are not evenly distributed among all values of pred
ictor variable.
## we observe that R^2 value is 0.02 which means this is not a good fit for the model
as expected.
#### equation->
#profit = 1.8330 + 0.53(adcost)

## inference -assumptions not followed, observed a very poor performing model

cor(df$profit,df$adcost)
```

```
## [1] 0.1632729
```

```
## coefficient of determination

r = cor(df$profit,df$adcost)
cod = r^2
print(paste("Coefficient of determination:", cod))
```

```
## [1] "Coefficient of determination: 0.0266580496771806"
```

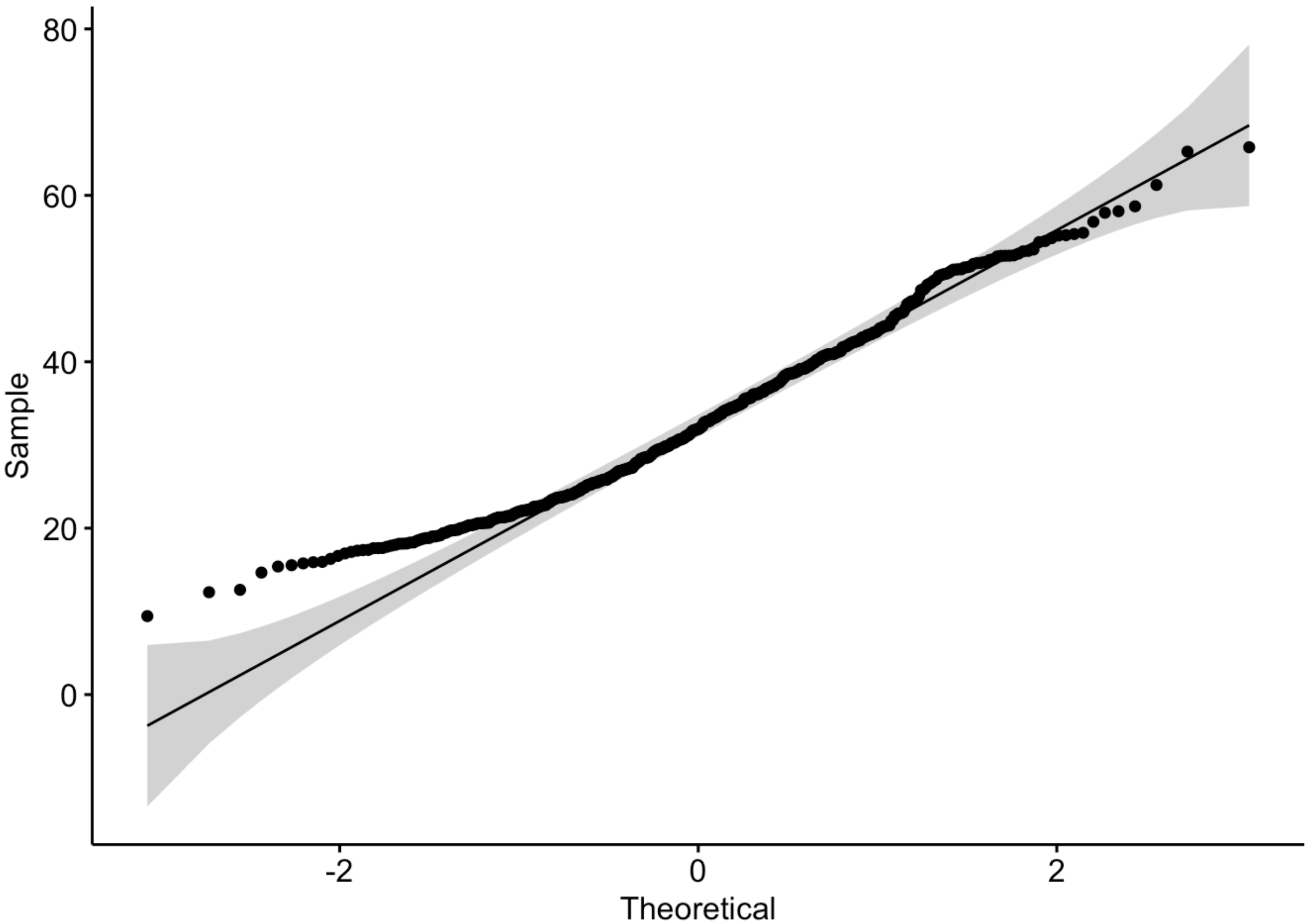
```
## We can observe from the residual plot there is a clear heteroscedasticity.
## This is a problem, in part, because the observations with larger errors will have
more pull or influence on the fitted model.
```

```
#### Q6 f) ####
## Shapiro test to check for normality
shapiro.test(df$profit)
```

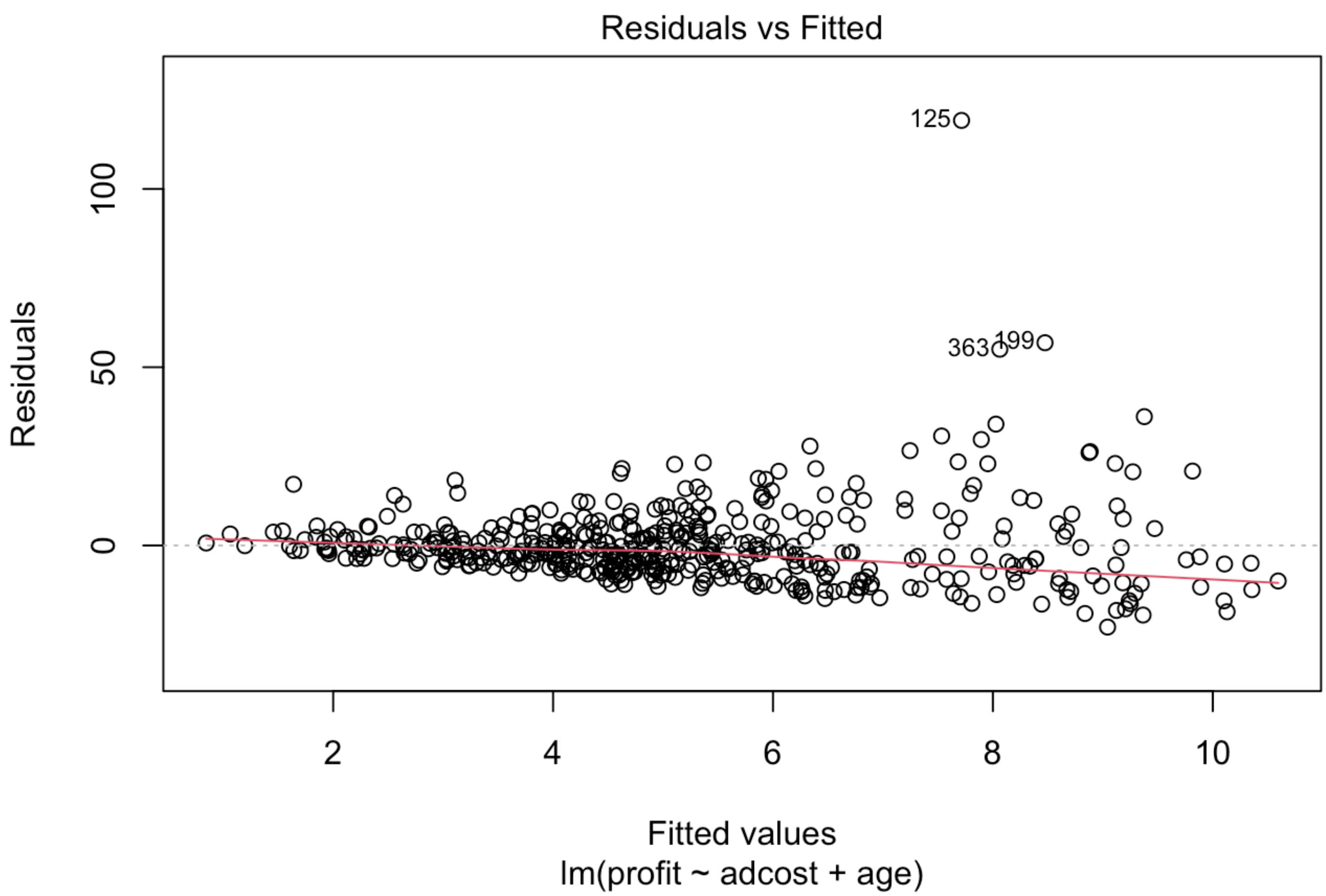
```
##
##  Shapiro-Wilk normality test
##
## data:  df$profit
## W = 0.73561, p-value < 2.2e-16
```

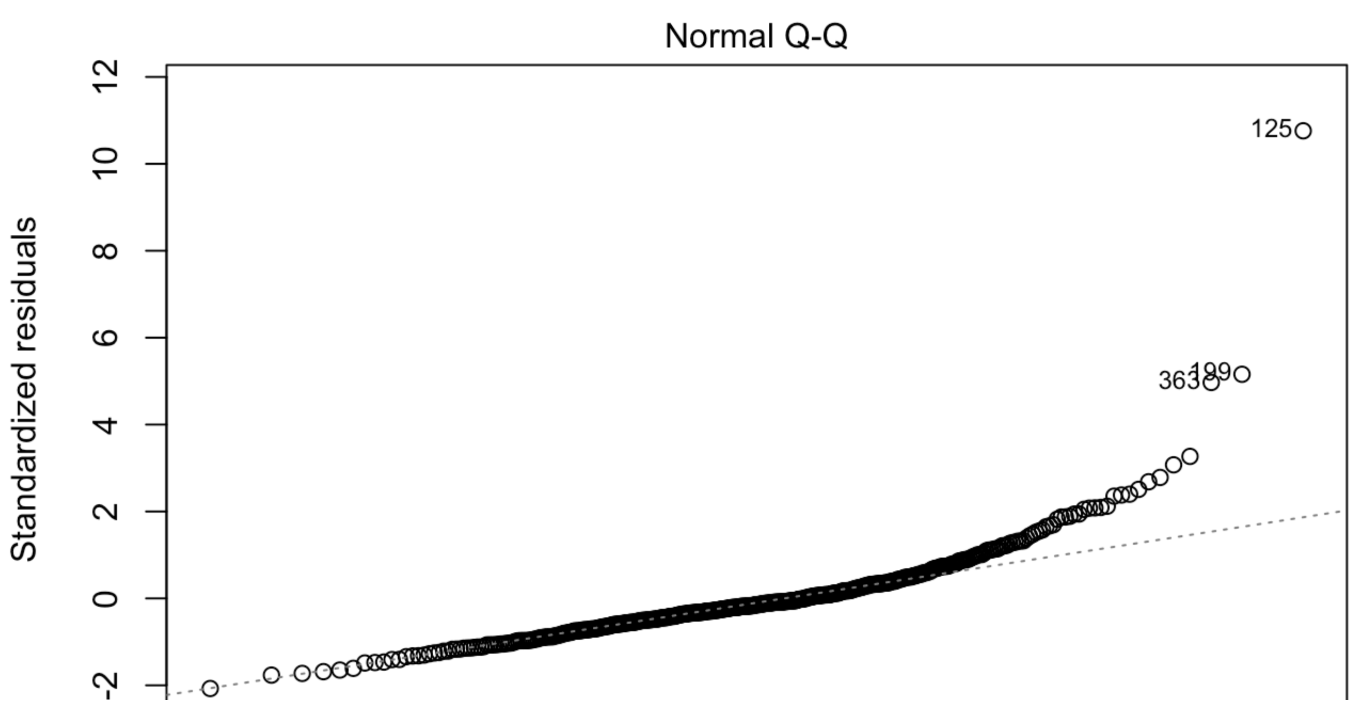
```
##assumptions not held
ggqqplot(df$adcost+df$age)
```

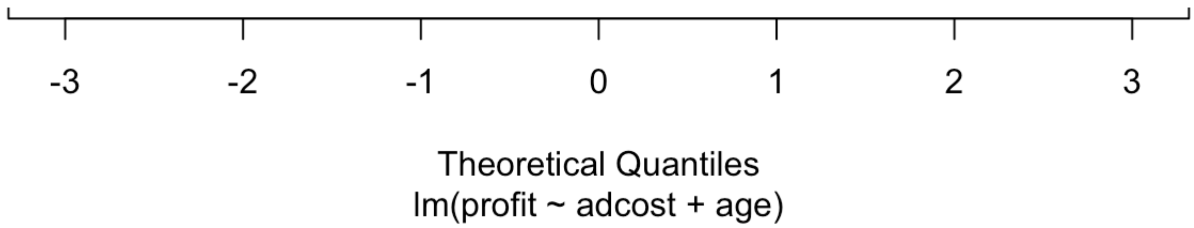
```
## Warning: The following aesthetics were dropped during statistical transformation:
sample
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
## The following aesthetics were dropped during statistical transformation: sample
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```

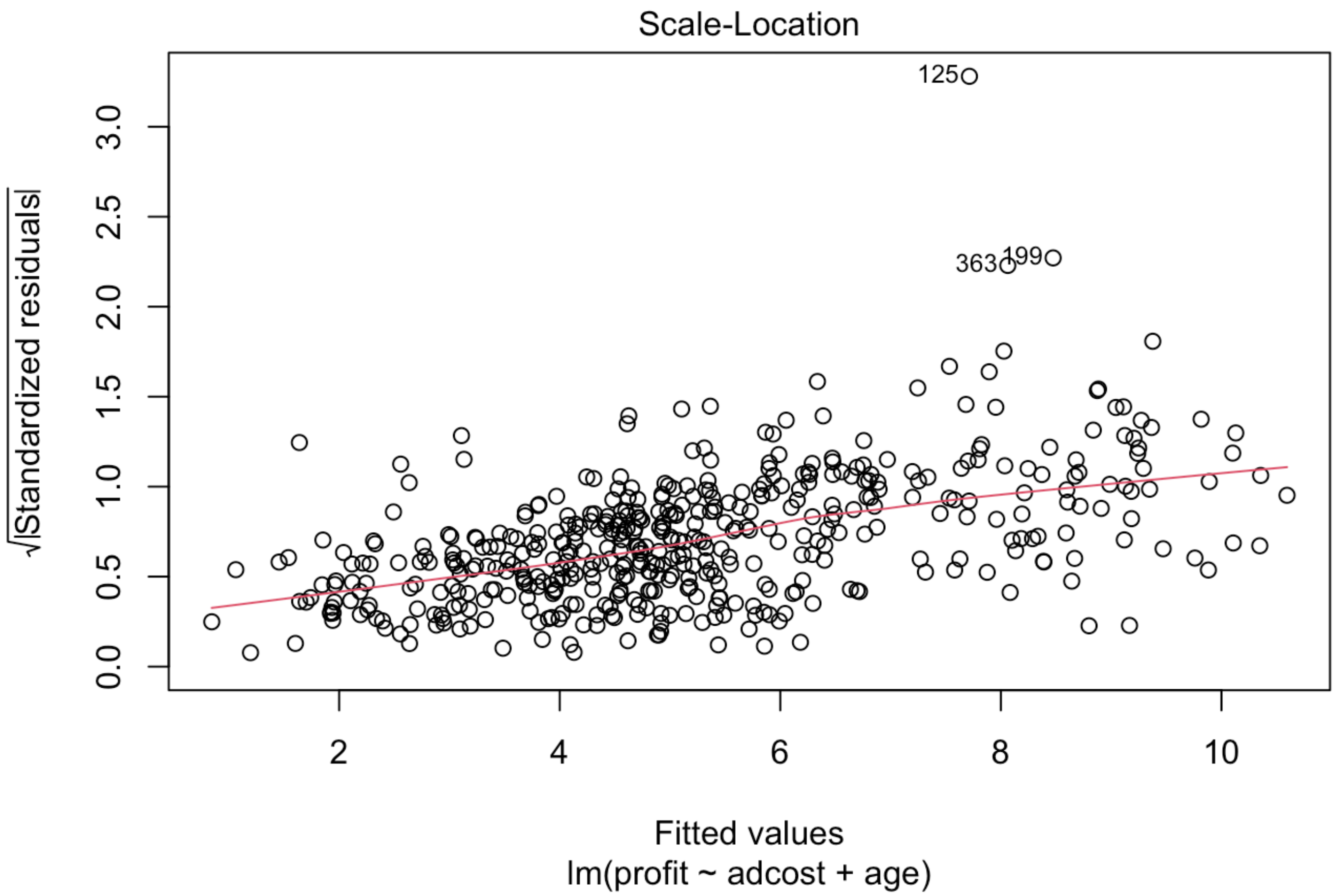


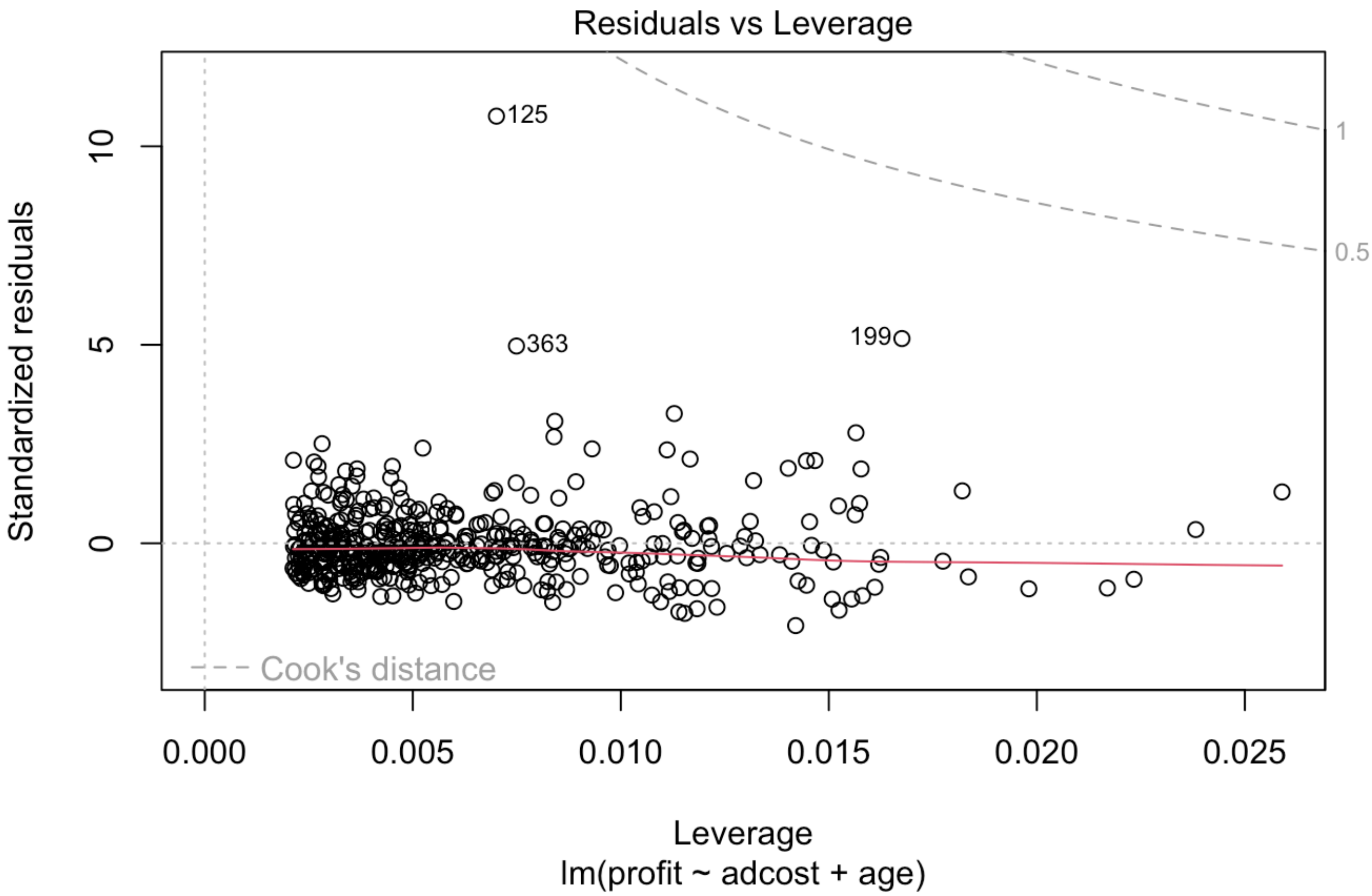

```
multi_model <- lm(profit ~ adcost + age, data = df)

plot(multi_model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(profit ~ adcost + age)
125
363
199
100
50
0
2
4
6
8
10


Normal Q-Q
Standardized residuals
125
363
199
12
10
8
6
4
2
0
-2


-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(profit ~ adcost + age)


Scale-Location
125
363
199
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2
4
6
8
10
Fitted values
lm(profit ~ adcost + age)




```
## values not evenly distributed for values of predictor variable.
summary(multi_model)
```

```
##
## Call:
## lm(formula = profit ~ adcost + age, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.893  -5.860  -1.616   3.689 119.225
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.18576    1.67682  -0.111  0.91184
## adcost       0.45961    0.15428   2.979  0.00304 **
## age          0.09258    0.05913   1.566  0.11811
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.12 on 469 degrees of freedom
## Multiple R-squared:  0.03172,    Adjusted R-squared:  0.02759
## F-statistic: 7.682 on 2 and 469 DF,  p-value: 0.0005216
```

```
## inference- Not a good fit since adjusted R^2 = 0.02
## We can observe from the residual plot there is a clear heteroscedasticity.
## This is a problem, in part, because the observations with larger errors will have
more pull or influence on the fitted model.
```

```
#### Q6 g) #####
# H0 = adcost does not improve the fit of the model
# H1 = adcost improves the fit of the model

fullmodel <- lm(profit ~ adcost + age, data = df)

summary(fullmodel)
```

```
##
## Call:
## lm(formula = profit ~ adcost + age, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.893  -5.860  -1.616   3.689 119.225
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.18576    1.67682  -0.111  0.91184
## adcost       0.45961    0.15428   2.979  0.00304 **
## age          0.09258    0.05913   1.566  0.11811
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.12 on 469 degrees of freedom
## Multiple R-squared:  0.03172,    Adjusted R-squared:  0.02759
## F-statistic: 7.682 on 2 and 469 DF,  p-value: 0.0005216
```

```
reducedmodel <- lm(profit ~ age, data = df)
summary(reducedmodel)
```

```
##
## Call:
## lm(formula = profit ~ age, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.259  -6.056  -2.264   3.293 121.531
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.37513    1.60615   0.856   0.3923
## age          0.14405    0.05702   2.526   0.0119 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.21 on 470 degrees of freedom
## Multiple R-squared:  0.0134, Adjusted R-squared:  0.0113
## F-statistic: 6.382 on 1 and 470 DF,  p-value: 0.01186
```

```
# Conduct the F-test to compare 2 models one with adcost to see if it improves the li
near regression model.
anova(reducedmodel, fullmodel, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: profit ~ age
## Model 2: profit ~ adcost + age
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1    470 59085
## 2    469 57988  1    1097.2 8.8744 0.003042 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#the p-value is 0.003042, which is less than alpha = 0.05.
#This means that the adcost variable significantly improves the fit of the model, and
we can reject the null hypothesis that it does not significantly predict the profit v
ariable.
#This suggests that the adcost variable is useful for predicting the profit variable,
and should be included in the model.
```

```
#### Q7 ####
## We know that ANOVA requires data distribution to be normal. We saw from our previo
us analysis that this assumption in our case is violated. Hence, we chose Kruskal-Wal
is test, which is a non-parametric equivalent test for ANOVA. The Kruskal Wallis test
will tell us if there is a significant difference between groups.

#We have come across multiple approaches to solve Q7
#Approach 1:  We conducted Kruskal-Wallis test on profit across each platform

## alpha is 0.05
#H0 : μfb = μInsta = μTk = μTw = μYT
#H1 : at least one of the social media platforms has an average profit that is differ
ent from at least one of the other social media platforms.


kruskal.test(df$profit~df$socialmedia)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df$profit by df$socialmedia
## Kruskal-Wallis chi-squared = 7.5755, df = 4, p-value = 0.1084
```

```
## As we can see that the p-value 0.1084 > 0.05 we fail to reject H0.
#  there is a no significant difference in the avg. Profit across the social media pl
atforms.
# Approach 1 suggested that we earn equal profit across each platform hence we planne
d to divide 100 dollars equally for all the platforms.

## Now approach 2

##### Approach 2 : In this approach we analyzed the effect of each variable on the pr
ofit across each platform. We considered each variable individually and performed sta
tistical test to understand their contribution towards profit.
#We know that ANOVA requires data distribution to be normal. We saw from our previous
analysis that this assumption in our case is violated. Hence, we chose Kruskal-Walis
test, which is a non-parametric equivalent test for ANOVA. The Kruskal Wallis test wi
ll tell us if there is a significant difference between groups.
#Season as a factor:
#To check if season was one of the contributor of the profit we ran Kruskal-Walis tes
t for season for each #platform individually.
# Three assumptions must hold:
# • Normality: Each group follows a normal distribution
# • Equal variances: Population variances for each group are equal
# • Independence: Observations are not correlate
#As we have seen earlier the normality doesn't hold true for summer and winter sample
.
## Since the underlying normality assumptions of ANOVA are violated we cannot go ahea
d with ANOVA test.
# We will perform Kruskal-wallis test which is non parametric equivalent of one-way A
NOVA.
kruskal.test(y_df$profit~y_df$season)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  y_df$profit by y_df$season
## Kruskal-Wallis chi-squared = 7.0296, df = 3, p-value = 0.07096
```

```
kruskal.test(fac_df$profit~fac_df$season)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  fac_df$profit by fac_df$season
## Kruskal-Wallis chi-squared = 2.9525, df = 3, p-value = 0.399
```

```
kruskal.test(Inst_df$profit~Inst_df$season)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Inst_df$profit by Inst_df$season
## Kruskal-Wallis chi-squared = 2.4791, df = 3, p-value = 0.4791
```

```
kruskal.test(tk_df$profit~tk_df$season)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  tk_df$profit by tk_df$season
## Kruskal-Wallis chi-squared = 5.0196, df = 3, p-value = 0.1704
```

```
kruskal.test(tw_df$profit~tw_df$season)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  tw_df$profit by tw_df$season
## Kruskal-Wallis chi-squared = 0.79269, df = 3, p-value = 0.8512
```

```
## Season doesn't affect the profit for individual platforms, so we can eliminate it
from our final equation


###Since we only had 2 categories under new customer, we conducted Wilcox test on cus
tomer variable across each platform.

#### To check how new_customer or old_customer affect the profits for YouTube
y_df_new<-y_df %>% filter(newcustomer=="yes") %>% select(profit)
y_df_old<-y_df %>% filter(newcustomer=="no") %>% select(profit)
wilcox.test(y_df_new$profit,y_df_old$profit,exact = F,correct = F)
```

```
##
##  Wilcoxon rank sum test
##
## data:  y_df_new$profit and y_df_old$profit
## W = 2956, p-value = 0.8474
## alternative hypothesis: true location shift is not equal to 0
```

```
## We observe that newcustomer feature doesn't affect profit for this platform.


## To check how new_customer or old_customer affect the profits for Facebook
fac_df_new<-fac_df %>% filter(newcustomer=="yes") %>% select(profit)
fac_df_old<-fac_df %>% filter(newcustomer=="no") %>% select(profit)
wilcox.test(fac_df_new$profit,fac_df_old$profit,exact = F,correct = F)
```

```
##
##  Wilcoxon rank sum test
##
## data:  fac_df_new$profit and fac_df_old$profit
## W = 170, p-value = 0.4086
## alternative hypothesis: true location shift is not equal to 0
```

```
## We observe that newcustomer feature doesn't affect profit for this platform.

### To check how new_customer or old_customer affect the profits for Instagram
Inst_df_new<-Inst_df %>% filter(newcustomer=="yes") %>% select(profit)
Inst_df_old<-Inst_df %>% filter(newcustomer=="no") %>% select(profit)
wilcox.test(Inst_df_new$profit,Inst_df_old$profit,exact = F,correct = F)
```

```
##
##  Wilcoxon rank sum test
##
## data:  Inst_df_new$profit and Inst_df_old$profit
## W = 980, p-value = 0.1192
## alternative hypothesis: true location shift is not equal to 0
```

```
## We observe that newcustomer feature doesn't affect profit for this platform.

## To check how new_customer or old_customer affect the profits for Twitter
tw_df_new<-tw_df %>% filter(newcustomer=="yes") %>% select(profit)
tw_df_old<-tw_df %>% filter(newcustomer=="no") %>% select(profit)
wilcox.test(tw_df_new$profit,tw_df_old$profit,exact = F,correct = F)
```

```
##
##  Wilcoxon rank sum test
##
## data:  tw_df_new$profit and tw_df_old$profit
## W = 33, p-value = 0.7984
## alternative hypothesis: true location shift is not equal to 0
```

```
## We observe that newcustomer feature doesn't affect profit for this platform.

## To check how new_customer or old_customer affect the profits for TikTok
tk_df_new<-tk_df %>% filter(newcustomer=="yes") %>% select(profit)
tk_df_old<-tk_df %>% filter(newcustomer=="no") %>% select(profit)
wilcox.test(tk_df_new$profit,tk_df_old$profit,exact = F,correct = F)
```

```
##
##  Wilcoxon rank sum test
##
## data:  tk_df_new$profit and tk_df_old$profit
## W = 2748, p-value = 0.2724
## alternative hypothesis: true location shift is not equal to 0
```

```
###The above results show that none of the platforms were affected by new customer va
riable.

## To check if age is affecting the profit for each social media platform
## Since the groups involved were more than two , we conducted Kruskal-Walis test on
age group for each      platform.
## As mentioned above age was divided into 3 categories
kruskal.test(y_df$profit~y_df$age_group)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  y_df$profit by y_df$age_group
## Kruskal-Wallis chi-squared = 0.091573, df = 2, p-value = 0.9552
```

```
kruskal.test(fac_df$profit~fac_df$age_group)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  fac_df$profit by fac_df$age_group
## Kruskal-Wallis chi-squared = 0.4519, df = 2, p-value = 0.7978
```

```
kruskal.test(Inst_df$profit~Inst_df$age_group)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Inst_df$profit by Inst_df$age_group
## Kruskal-Wallis chi-squared = 4.6295, df = 2, p-value = 0.09879
```

```
kruskal.test(tk_df$profit~tk_df$age_group)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  tk_df$profit by tk_df$age_group
## Kruskal-Wallis chi-squared = 1.316, df = 1, p-value = 0.2513
```

```
kruskal.test(tw_df$profit~tw_df$age_group)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  tw_df$profit by tw_df$age_group
## Kruskal-Wallis chi-squared = 0.13043, df = 1, p-value = 0.718
```

```
## We observed that age-group do not affect the profit for each social media platform
s.

###Since we only had 2 categories under mobile variable, we conducted Wilcox test on
mobile variable across each platform.

## To check if mobile feature affected profit for YouTube.
y_df_mob<-y_df %>% filter(mobile=="mobile") %>% select(profit)
y_df_comp<-y_df %>% filter(mobile=="computer") %>% select(profit)
wilcox.test(y_df_mob$profit,y_df_comp$profit,exact=F,correct=F)
```

```
##
##  Wilcoxon rank sum test
##
## data:  y_df_mob$profit and y_df_comp$profit
## W = 1923, p-value = 0.007654
## alternative hypothesis: true location shift is not equal to 0
```

```
## We observed that mobile feature significantly affected profit for YouTube.

## To check if mobile feature affected profit for Facebook.
fac_df_mob<-fac_df %>% filter(mobile=="mobile") %>% select(profit)
fac_df_comp<-fac_df %>% filter(mobile=="computer") %>% select(profit)
wilcox.test(fac_df_mob$profit,fac_df_comp$profit,exact=F,correct=F)
```

```
##
##  Wilcoxon rank sum test
##
## data:  fac_df_mob$profit and fac_df_comp$profit
## W = 376, p-value = 0.2739
## alternative hypothesis: true location shift is not equal to 0
```

```
## We observed that mobile feature doesn't have an impact on the profit for Facebook.

## To check if mobile feature affected profit for Instagram.
Inst_df_mob<-Inst_df %>% filter(mobile=="mobile") %>% select(profit)
Inst_df_comp<-Inst_df %>% filter(mobile=="computer") %>% select(profit)
wilcox.test(Inst_df_mob$profit,Inst_df_comp$profit,exact=F,correct=F)
```

```
##
##  Wilcoxon rank sum test
##
## data:  Inst_df_mob$profit and Inst_df_comp$profit
## W = 632, p-value = 0.6235
## alternative hypothesis: true location shift is not equal to 0
```

```
## We observed that mobile feature doesn't have an impact on the profit for Instagram
.

## To check if mobile feature affected profit for Twitter.
tw_df_mob<-tw_df %>% filter(mobile=="mobile") %>% select(profit)
tw_df_comp<-tw_df %>% filter(mobile=="computer") %>% select(profit)
wilcox.test(tw_df_mob$profit,tw_df_comp$profit,exact=F,correct=F)
```

```
##
##  Wilcoxon rank sum test
##
## data:  tw_df_mob$profit and tw_df_comp$profit
## W = 57, p-value = 0.9456
## alternative hypothesis: true location shift is not equal to 0
```

```
## We observed that mobile feature doesn't have an impact on the profit for Twitter.

## To check if mobile feature affected profit for TikTok.
tk_df_mob<-tk_df %>% filter(mobile=="mobile") %>% select(profit)
tk_df_comp<-tk_df %>% filter(mobile=="computer") %>% select(profit)
wilcox.test(tk_df_mob$profit,tk_df_comp$profit,exact=F,correct=F)
```

```
##
##  Wilcoxon rank sum test
##
## data:  tk_df_mob$profit and tk_df_comp$profit
## W = 1361, p-value = 0.001138
## alternative hypothesis: true location shift is not equal to 0
```

```
## We observed that mobile feature significantly affected profit for TikTok.

### Once we got to know what all features were having an impact or not on the profit
of each social media platform
## We implemented simple and multiple linear regression models just for those feature
s affecting profit.
y_lin<-lm(profit~ adrevenue+mobile,data = y_df)
summary(y_lin)
```

```
##
## Call:
## lm(formula = profit ~ adrevenue + mobile, data = y_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4916 -2.2486 -0.1514  2.5396  4.8168
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -8.8809     0.3920 -22.654   <2e-16 ***
## adrevenue      0.9616     0.0136  70.709   <2e-16 ***
## mobilemobile  -0.8792     0.5038  -1.745    0.083 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.882 on 153 degrees of freedom
## Multiple R-squared:  0.9713, Adjusted R-squared:  0.9709
## F-statistic:  2590 on 2 and 153 DF,  p-value: < 2.2e-16
```

```
Inst_lin<-lm(profit~ adrevenue,data = Inst_df)
summary(Inst_lin)
```

```
##
## Call:
## lm(formula = profit ~ adrevenue, data = Inst_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.284 -1.389  0.302  1.487  2.619
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.31571    0.28135  -18.89   <2e-16 ***
## adrevenue    0.93664    0.02176   43.05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.764 on 91 degrees of freedom
## Multiple R-squared:  0.9532, Adjusted R-squared:  0.9527
## F-statistic:  1853 on 1 and 91 DF,  p-value: < 2.2e-16
```

```
tk_lin<-lm(profit~ adrevenue+mobile,data = tk_df)
summary(tk_lin)
```

```
##
## Call:
## lm(formula = profit ~ adrevenue + mobile, data = tk_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7281 -1.7566 -0.0946  1.8828  5.1937
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.69715    0.59398  -6.224 5.44e-09 ***
## adrevenue     0.82417    0.03211  25.668  < 2e-16 ***
## mobilemobile  1.02614    0.62484   1.642    0.103
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.173 on 138 degrees of freedom
## Multiple R-squared:  0.8368, Adjusted R-squared:  0.8345
## F-statistic: 353.9 on 2 and 138 DF,  p-value: < 2.2e-16
```

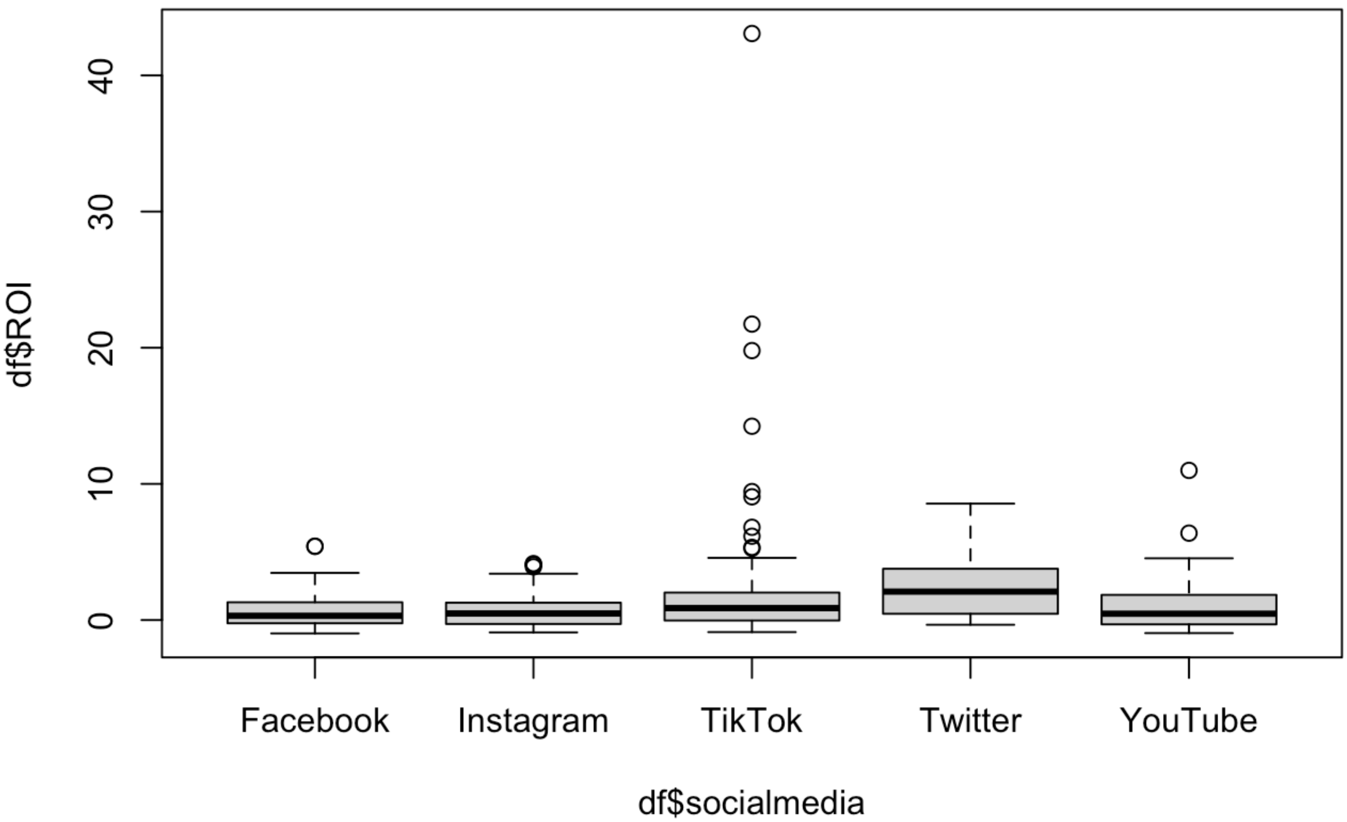
```
tw_lin<-lm(profit~ adrevenue,data = tw_df)
summary(tw_lin)
```

```
##
## Call:
## lm(formula = profit ~ adrevenue, data = tw_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2273 -0.6615 -0.1341  0.4507  1.4988
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.47155    0.37375  -6.613 1.93e-06 ***
## adrevenue    0.99072    0.04146  23.893 3.53e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8725 on 20 degrees of freedom
## Multiple R-squared:  0.9662, Adjusted R-squared:  0.9645
## F-statistic: 570.9 on 1 and 20 DF,  p-value: 3.533e-16
```

```
fac_lin<-lm(profit~ adrevenue,data = fac_df)
summary(fac_lin)
```

```
##
## Call:
## lm(formula = profit ~ adrevenue, data = fac_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3004 -0.8397 -0.1011  1.0948  2.2021
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.48924    0.23907  -18.78   <2e-16 ***
## adrevenue    0.95560    0.02333   40.96   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.192 on 58 degrees of freedom
## Multiple R-squared:  0.9666, Adjusted R-squared:  0.966
## F-statistic:  1678 on 1 and 58 DF,  p-value: < 2.2e-16
```

```
###Ultimately we got equations for each platform with only those features that affect
the profit for those platforms.

# P_youtube = -8.88 + 0.96(adrevenue) - 0.879(mobile)
# P_ Instagram = -5.315 + 0.93(adrevenue)
# P_ TikTok = -3.69 + 0.82(adrevenue) + 1.02(mobile)
# P_Twitter = -2.47 + 0.99(adrevenue)
# P_Facebook = -4.48 + 0.95(adrevenue)

## Solving the above equation, we came up the division of 100 as below:
# $41.5 to YouTube
# $11.7 to TikTok
# $7.3  to Twitter
# $17.5 to Facebook
# $22 to Instagram


#### Approach 3
#### We are considering Return on Investment
boxplot(df$ROI~df$socialmedia)
```




```
fb_model<-lm(profit~ROI,data=fac_df)
summary(fb_model)
```

```
##
## Call:
## lm(formula = profit ~ ROI, data = fac_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0249 -0.4297 -0.0127  0.5904  4.9446
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.02661    0.20143   0.132    0.895
## ROI          4.74373    0.13784  34.414   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.409 on 58 degrees of freedom
## Multiple R-squared:  0.9533, Adjusted R-squared:  0.9525
## F-statistic:  1184 on 1 and 58 DF,  p-value: < 2.2e-16
```

```
Insta_model<-lm(profit~ROI,data=Inst_df)
summary(Insta_model)
```

```
##
## Call:
## lm(formula = profit ~ ROI, data = Inst_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7980 -0.9169  0.0436  0.8356  9.1215
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.3176     0.2939  -1.081    0.283
## ROI           6.2005     0.2081  29.795   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.486 on 91 degrees of freedom
## Multiple R-squared:  0.907,  Adjusted R-squared:  0.906
## F-statistic: 887.8 on 1 and 91 DF,  p-value: < 2.2e-16
```

```
Tw_model<-lm(profit~ROI,data=tw_df)
summary(Tw_model)
```

```
##
## Call:
## lm(formula = profit ~ ROI, data = tw_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9376 -1.5435 -0.8245  0.6726  7.9559
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.4947     0.8484   1.762   0.0934 .
## ROI           1.4927     0.2404   6.208  4.6e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.772 on 20 degrees of freedom
## Multiple R-squared:  0.6584, Adjusted R-squared:  0.6413
## F-statistic: 38.54 on 1 and 20 DF,  p-value: 4.599e-06
```

```
Tk_model<-lm(profit~ROI+mobile,data=tk_df)
summary(Tk_model)
```

```
##
## Call:
## lm(formula = profit ~ ROI + mobile, data = tk_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.286  -3.076  -1.034   2.196  17.069
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.63005    1.30601  -0.482  0.63027
## ROI           0.39213    0.08854   4.429 1.91e-05 ***
## mobilemobile  3.62462    1.38366   2.620  0.00979 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.885 on 138 degrees of freedom
## Multiple R-squared:  0.1751, Adjusted R-squared:  0.1632
## F-statistic: 14.65 on 2 and 138 DF,  p-value: 1.7e-06
```

```
YT_model<-lm(profit~ROI+mobile,data=y_df)
summary(YT_model)
```

```
##
## Call:
## lm(formula = profit ~ ROI + mobile, data = y_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1509  -1.6247   0.3806   1.9436  20.3344
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.0254     0.5750  -1.783   0.0765 .
## ROI            9.9885     0.2572  38.836   <2e-16 ***
## mobilemobile   0.7398     0.8960   0.826   0.4103
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.075 on 153 degrees of freedom
## Multiple R-squared:  0.911,  Adjusted R-squared:  0.9099
## F-statistic: 783.3 on 2 and 153 DF,  p-value: < 2.2e-16
```

```
## P_fb = 0.0266 + 4.743(ROI)
## P_Inst = -0.317 + 6.20(ROI)
## P_Tw = 1.494 + 1.492(ROI)
## P_Tk = -0.63 + 0.392(ROI) + 3.62(mobile)
## P_YT = -1.02 + 9.98(ROI) + 0.73(mobile)

# We got the proportions as below->
## facebook-> 19%
## Instagram-> 24%
## Twitter-> 13%
## TikTok -> 6%
## YouTube -> 38%
```