project_final_632

Sagar Soneji fx9706

2023-04-25

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(car)
```

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

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

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

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

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

```
library(MASS)
```

```
##
## Attaching package: 'MASS'

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

```
library(randomForest)
```

```
## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:gridExtra':
##
##     combine

## The following object is masked from 'package:dplyr':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
library(rpart)
library(caret)
```

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

```
project <- read.csv('/Users/sagarsoneji/Downloads/KAG_conversion_data.csv')
str(project)
```

```
## 'data.frame':    1143 obs. of  11 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                 : chr  "30-34" "30-34" "30-34" "30-34" ...
##  $ gender              : chr  "M" "M" "M" "M" ...
##  $ 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 ...
```

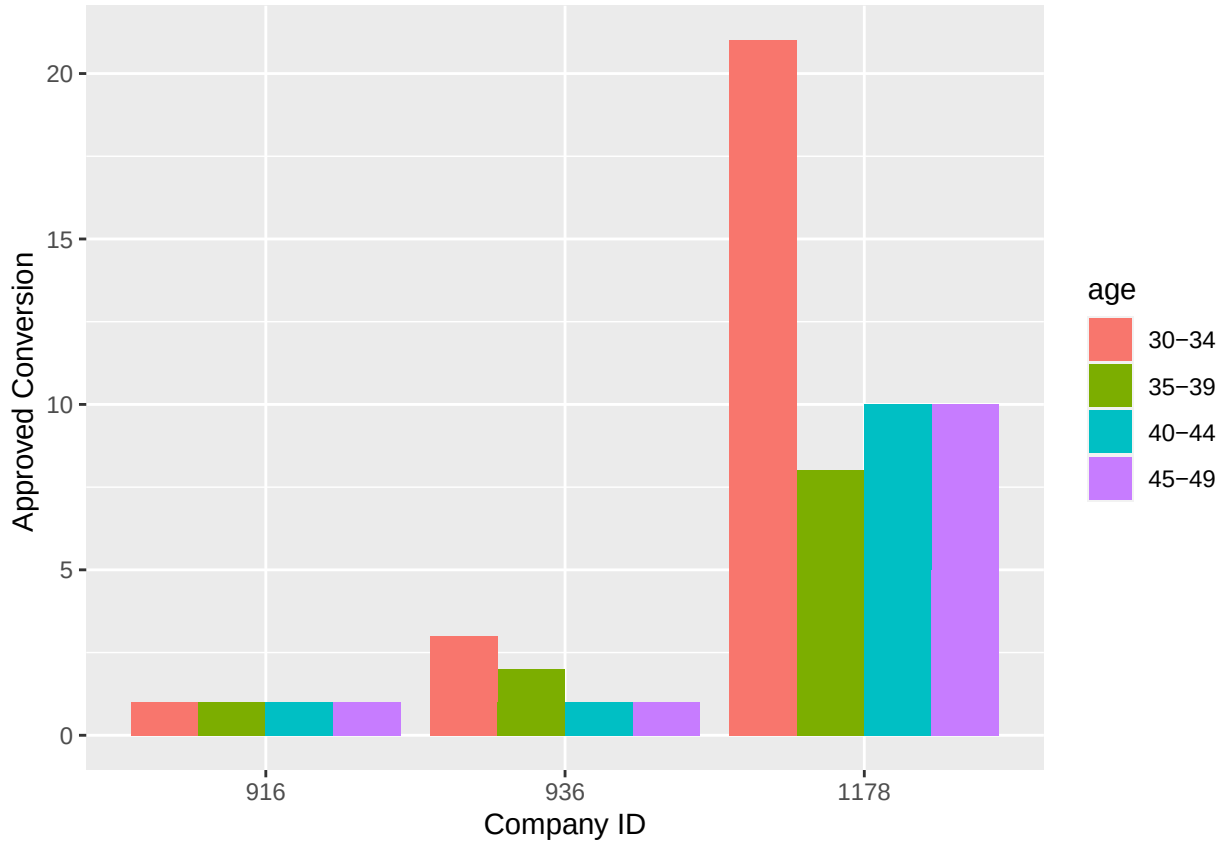
```
project$xyz_campaign_id <- as.factor(project$xyz_campaign_id)
head(project)
```

```
##    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
## 1      1  1.43                2                   1
## 2      2  1.82                2                   0
## 3      0  0.00                1                   0
## 4      1  1.25                1                   0
## 5      1  1.29                1                   1
## 6      0  0.00                1                   1
```

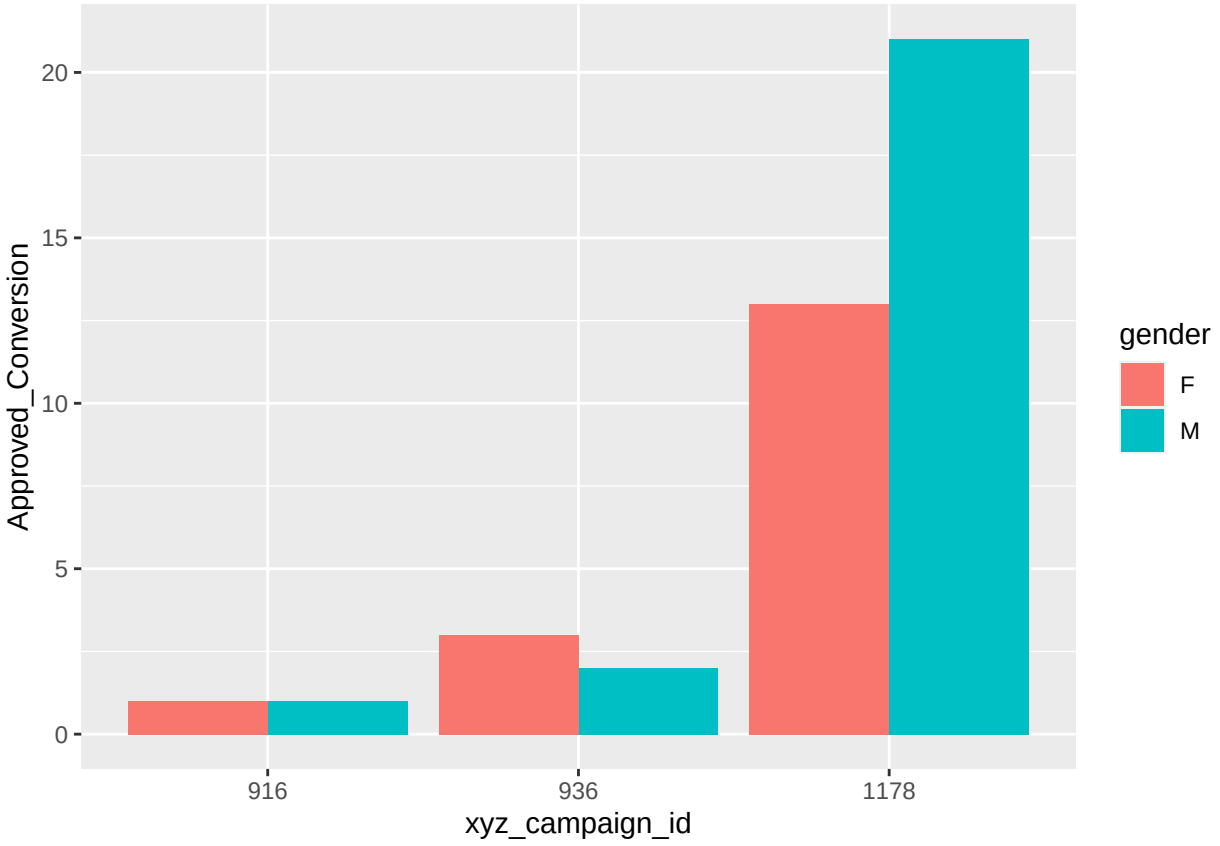
```
project$age <- as.factor(project$age)
project$gender <- as.factor(project$gender)
project$interest <- as.factor(project$interest)
head(project)
```

```
##    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
## 1      1  1.43                2                   1
## 2      2  1.82                2                   0
## 3      0  0.00                1                   0
## 4      1  1.25                1                   0
## 5      1  1.29                1                   1
## 6      0  0.00                1                   1
```

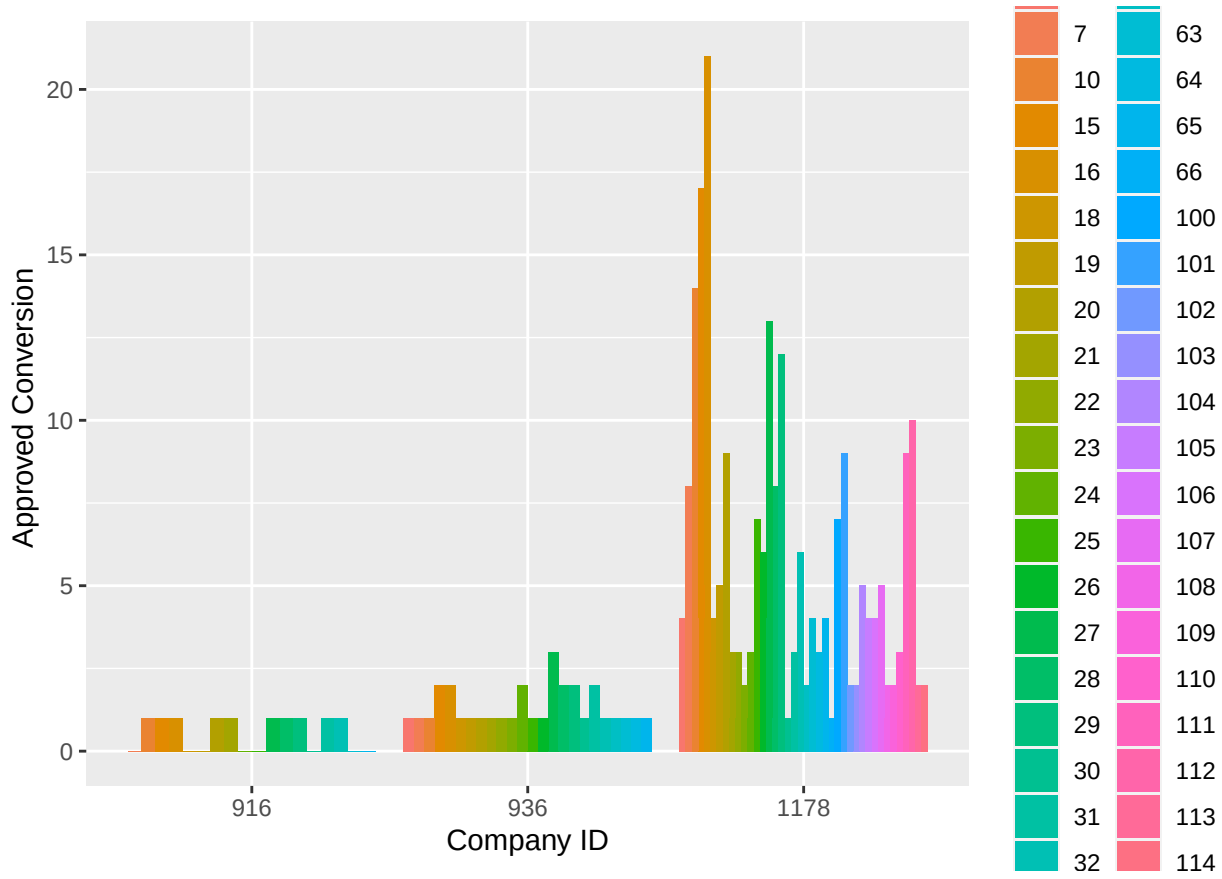
```
#bar plot of bar plot of approved conversion seperated by age and companies
ggplot(data = project, aes(xyz_campaign_id, Approved_Conversion, fill = age))+ geom_col(position = 'dod
```




```
#bar plot of bar plot of approved conversion seperated by gender and companies
ggplot(data = project, aes(xyz_campaign_id, Approved_Conversion, fill = gender))+ geom_col(position = '
```

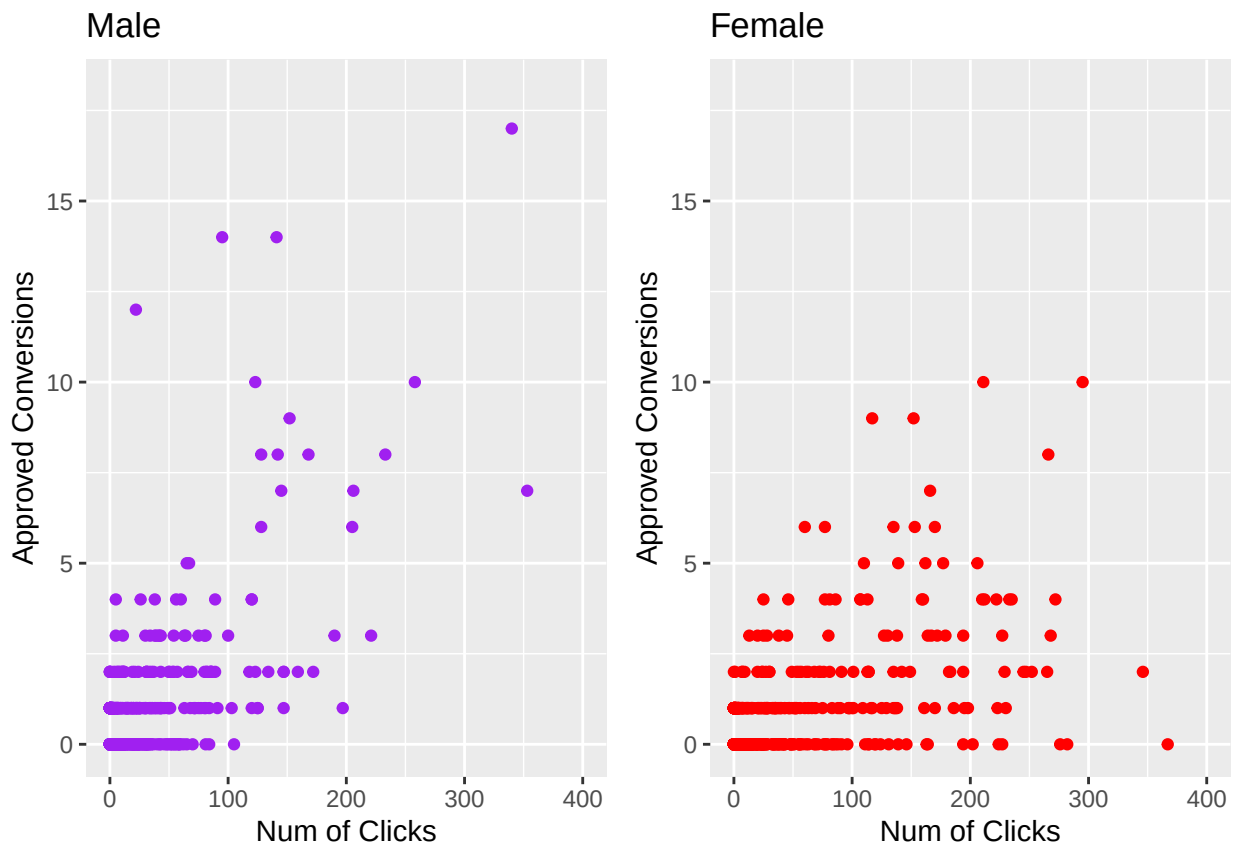



```
#by interests
#bar plot of bar plot of approved conversion seperated by age and companies
ggplot(data = project, aes(xyz_campaign_id, Approved_Conversion, fill = interest))+ geom_col(position =
```




```
#plot of clicks vs approved conversions factored by Sex to see how many people actully bought the produ
#creating subset of the dataset to compare variables better
df_female <- subset(project, project$gender == 'F')
df_male <- subset(project, project$gender == 'M')
#plotting each of the factors one by one
p_male <- ggplot(data = df_male, aes(Clicks, Approved_Conversion))+ geom_point(col = "purple") + ylim(0

p_female <- ggplot(data= df_female , aes(Clicks, Approved_Conversion))+ geom_point(col = "red") + ylim(

grid.arrange(p_male,p_female, ncol =2)
```

```
## Warning: Removed 1 rows containing missing values (‘geom_point()‘).
## Removed 1 rows containing missing values (‘geom_point()‘).
```




```
#plot of clicks vs approved conversions factored by age to see how mmay people actully bought the produ
df_32 <- subset(project, project$age == '30-34')
df_37 <- subset(project, project$age == '35-39')
df_42 <- subset(project, project$age == '40-44')
df_47 <- subset(project, project$age == '45-49')

nrow(df_32)
```

```
## [1] 426
```

```
nrow(df_37)
```

```
## [1] 248
```

```
nrow(df_42)
```

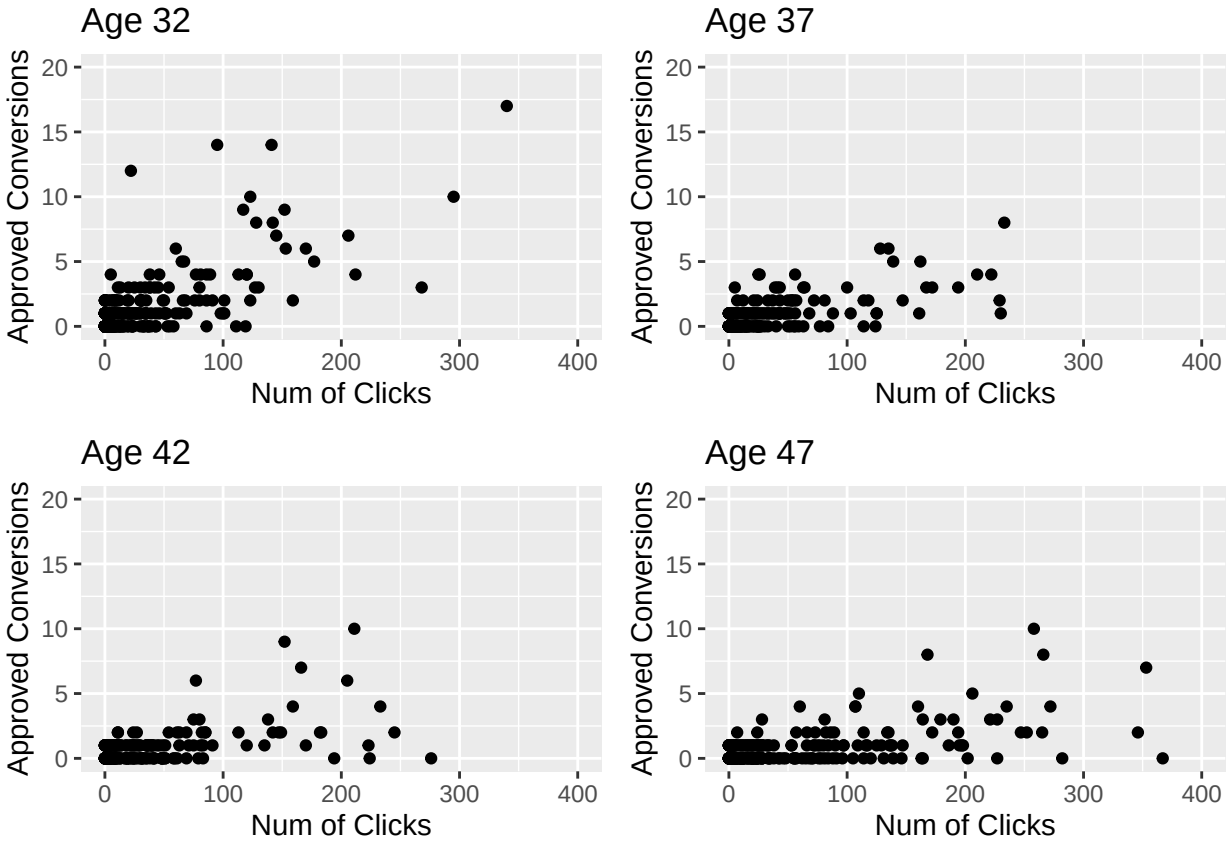
```
## [1] 210
```

```
nrow(df_47)
```

```
## [1] 259
```

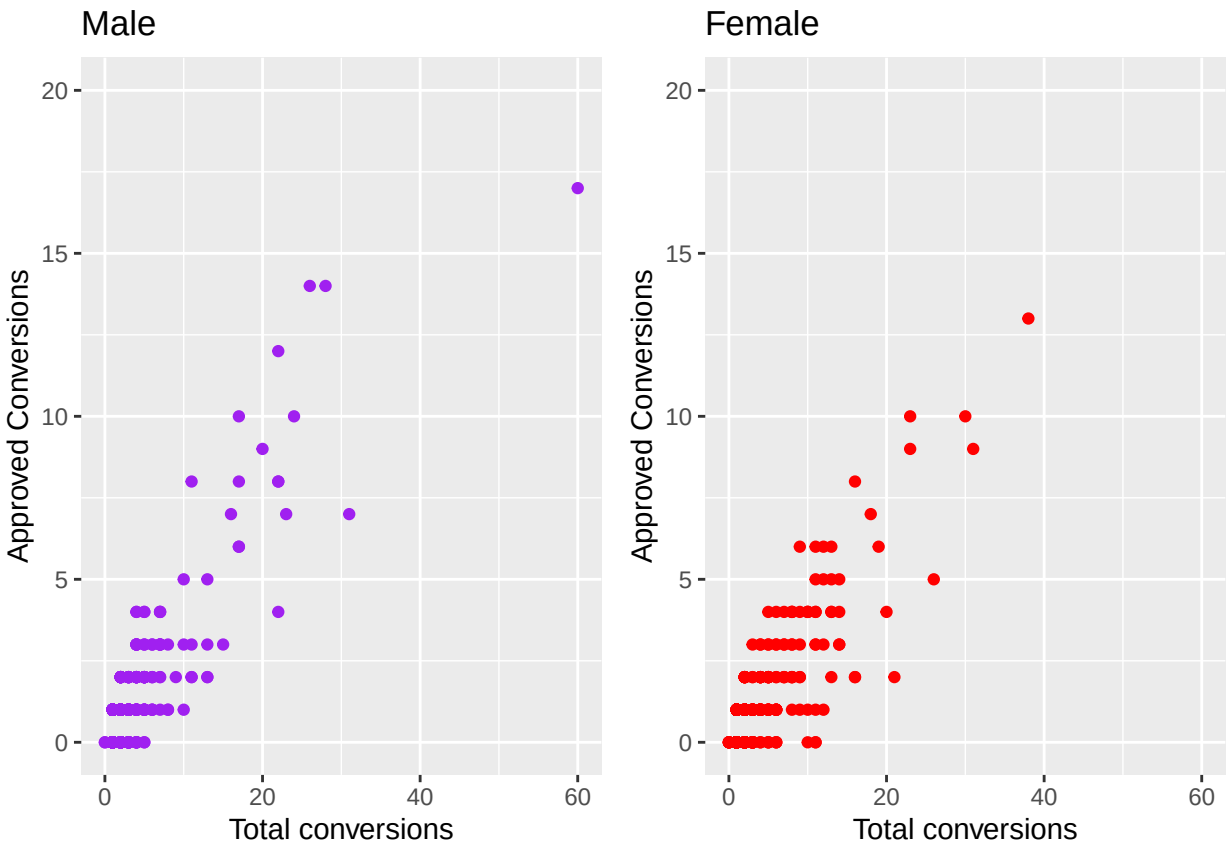
```
p_32 <- ggplot(data = df_32, aes(Clicks, Approved_Conversion))+ geom_point()+labs(title = 'Age 32', x =

p_37 <- ggplot(data = df_37, aes(Clicks, Approved_Conversion))+ geom_point()+labs(title = 'Age 37', x =

p_42 <- ggplot(data = df_42, aes(Clicks, Approved_Conversion))+ geom_point()+labs(title = 'Age 42', x =

p_47 <- ggplot(data = df_47, aes(Clicks, Approved_Conversion))+ geom_point()+labs(title = 'Age 47', x =

grid.arrange(p_32,p_37,p_42,p_47, ncol =2 , nrow = 2)
```

```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```



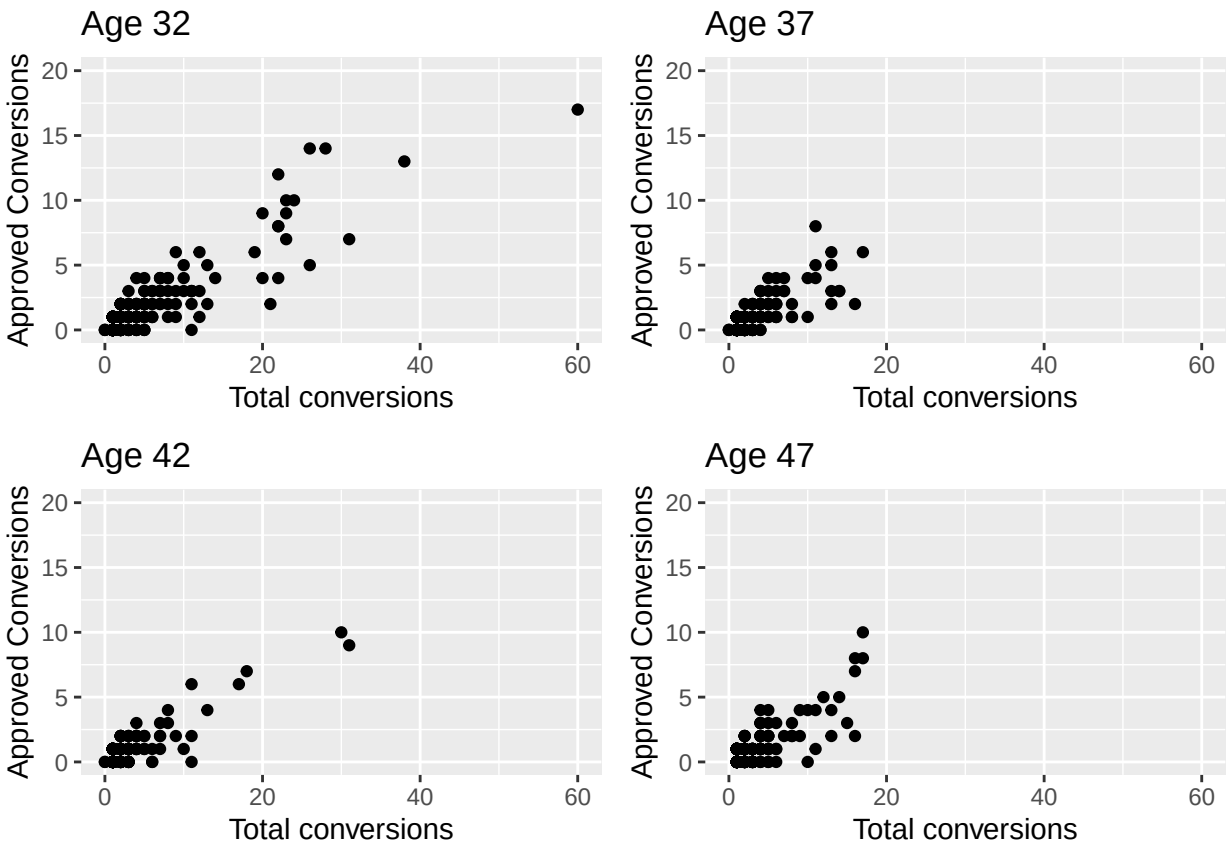

```
#plot of total conversion vs approved to see how many actually went from enquiring to buying
p_m <- ggplot(df_male , aes(Total_Conversion, Approved_Conversion)) + geom_point(col= "purple") + xlim(

p_f <- ggplot(df_female , aes(Total_Conversion, Approved_Conversion)) + geom_point(col= "red") + xlim(0

grid.arrange(p_m,p_f, ncol =2)
```

```
## Warning: Removed 1 rows containing missing values (`geom_point()`).
```




```
##plot of total conversion vs approved to see how many actually went from enquiring to buying by age
buy_32 <- ggplot(data = df_32, aes(Total_Conversion, Approved_Conversion))+ geom_point()+labs(title = '
buy_37 <- ggplot(data = df_37, aes(Total_Conversion, Approved_Conversion))+ geom_point()+labs(title = '
buy_42 <- ggplot(data = df_42, aes(Total_Conversion, Approved_Conversion))+ geom_point()+labs(title = '
buy_47 <- ggplot(data = df_47, aes(Total_Conversion, Approved_Conversion))+ geom_point()+labs(title = '

grid.arrange(buy_32,buy_37,buy_42,buy_47,nrow =2,ncol=2)
```

```
## Warning: Removed 1 rows containing missing values (‘geom_point()‘).
```



We can see three is a point way far out in the age

```
project[526,]
```

```
##       ad_id xyz_campaign_id fb_campaign_id   age gender interest Impressions
## 526 1121100            1178         144532 30-34      M       15     3052003
##     Clicks  Spent Total_Conversion Approved_Conversion
## 526    340 639.95               60                  17
```

```
# removing ad_id and fb_id
project <- project[,-1]
project <- project[,-2]

#Lets create some additional parametrics that will help in better data interpretation

#Click-through-rate (CTR): This is the percentage of how many of our impressions became clicks

project <- project %>%
  mutate(CTR = ((Clicks / Impressions)*100))

#cost per click : cpc : spent/click
project <- project %>%
  mutate(cpc = ((Spent / Clicks)))

#replacing NA in cpc by 0
project <- project %>%
```

```
  mutate(cpc = ifelse(is.na(cpc), 0 , cpc))

#Creating ROAS
project <- project %>%
  mutate(Conversion = Total_Conversion + Approved_Conversion,
         Conversion_Val = Total_Conversion * 2,
         Approved_ConVal = Approved_Conversion * 20) %>%
  mutate(Grand_ConVal = Conversion_Val + Approved_ConVal) %>%
  mutate(ROAS = round(Grand_ConVal / Spent, 2))

project <- project %>%
  mutate(ROAS = ifelse(is.na(ROAS), 0 , ROAS))

#removing inf
Project_good <- subset(project, project$Spent > 0)

#Conversions / click
Project_good <- Project_good %>%
  mutate(conversion_percent = ((Approved_Conversion/Clicks)*100))

#if conversion percent is greater than 10 we will consider the ad to be succesull for the sake od this

Project_good <- Project_good %>%
  mutate(Success = ifelse(conversion_percent > 5, 1 , 0))

Project_good$Success <- factor(Project_good$Success, levels = c(0,1))

table(Project_good$Success)
```

```
##
##   0   1
## 674 262
```

```
head(Project_good)
```

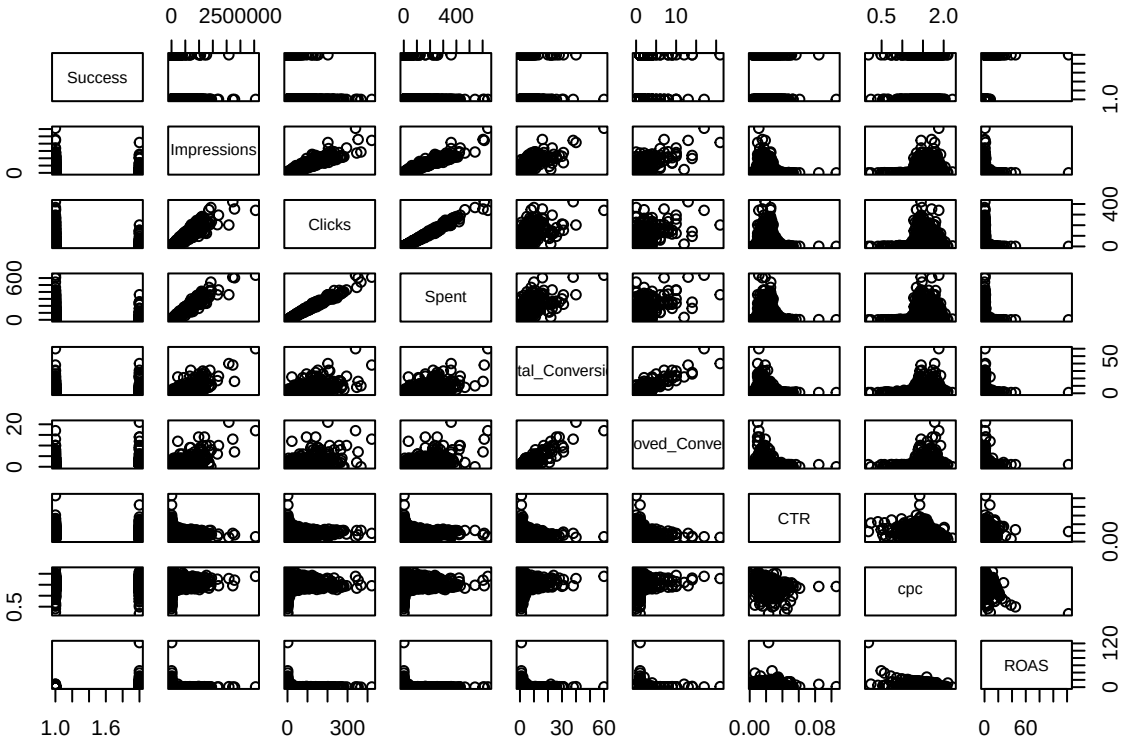
```
##   xyz_campaign_id   age gender interest Impressions Clicks Spent
## 1             916 30-34      M       15        7350      1  1.43
## 2             916 30-34      M       16       17861      2  1.82
## 4             916 30-34      M       28        4259      1  1.25
## 5             916 30-34      M       28        4133      1  1.29
## 7             916 30-34      M       15       15615      3  4.77
## 8             916 30-34      M       16       10951      1  1.27
##   Total_Conversion Approved_Conversion         CTR  cpc Conversion
## 1                2                   1 0.013605442 1.43          3
## 2                2                   0 0.011197581 0.91          2
## 4                1                   0 0.023479690 1.25          1
## 5                1                   1 0.024195500 1.29          2
## 7                1                   0 0.019212296 1.59          1
## 8                1                   1 0.009131586 1.27          2
##   Conversion_Val Approved_ConVal Grand_ConVal  ROAS conversion_percent Success
## 1              4              20           24 16.78               100       1
## 2              4               0            4  2.20                 0       0
```

```
## 4              2               0            2  1.60                  0       0
## 5              2              20           22 17.05                100       1
## 7              2               0            2  0.42                  0       0
## 8              2              20           22 17.32                100       1
```

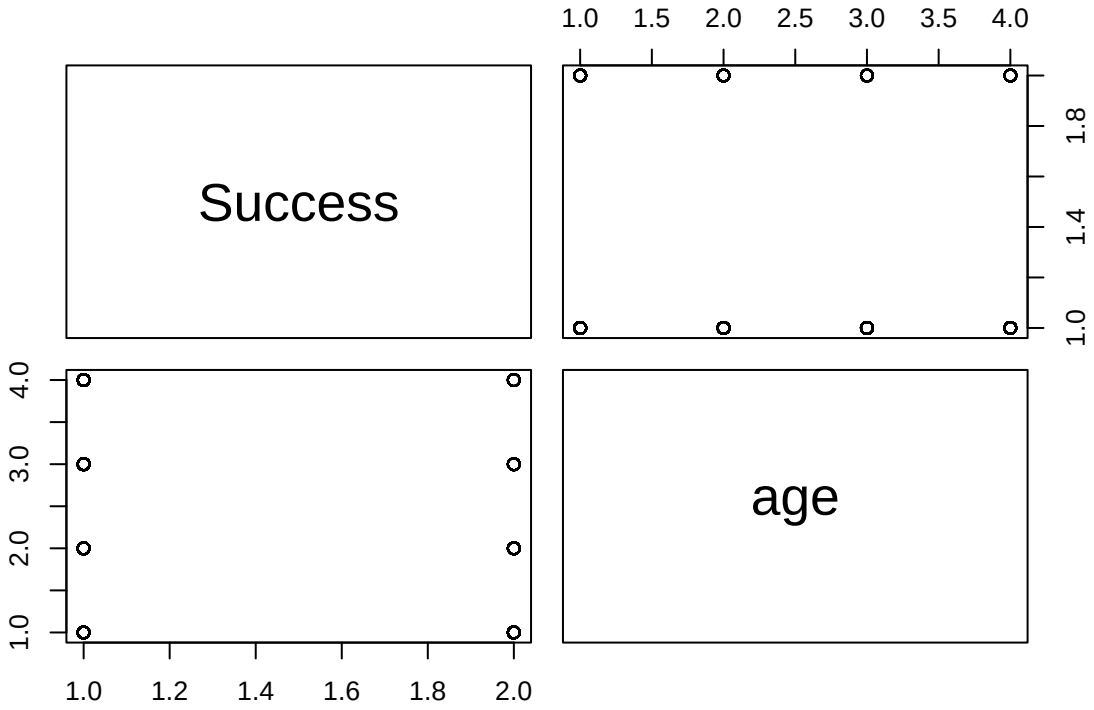
```
#remove variables that are npot needed
Project_good <- Project_good[,-12:-15]
head(Project_good)
```

```
##   xyz_campaign_id   age gender interest Impressions Clicks Spent
## 1             916 30-34      M       15        7350      1  1.43
## 2             916 30-34      M       16       17861      2  1.82
## 4             916 30-34      M       28        4259      1  1.25
## 5             916 30-34      M       28        4133      1  1.29
## 7             916 30-34      M       15       15615      3  4.77
## 8             916 30-34      M       16       10951      1  1.27
##   Total_Conversion Approved_Conversion         CTR  cpc  ROAS
## 1                2                   1 0.013605442 1.43 16.78
## 2                2                   0 0.011197581 0.91  2.20
## 4                1                   0 0.023479690 1.25  1.60
## 5                1                   1 0.024195500 1.29 17.05
## 7                1                   0 0.019212296 1.59  0.42
## 8                1                   1 0.009131586 1.27 17.32
##   conversion_percent Success
## 1                100       1
## 2                  0       0
## 4                  0       0
## 5                100       1
## 7                  0       0
## 8                100       1
```

```
#Creating Test and train data
set.seed(99)
train_index <- sample(nrow(Project_good), 0.7 * nrow(Project_good))
train_data <- Project_good[train_index, ]

test_data <- Project_good[-train_index,]

pairs(Success~ Impressions + Clicks + Spent + Total_Conversion + Approved_Conversion + CTR + cpc +ROAS
```




```
pairs(Success~ age  ,data = train_data)
```




```
model <- glm(Success ~ age + gender + interest + Clicks + CTR + cpc + ROAS   , data = train_data, family
summary(model)
```

```
##
## Call:
## glm(formula = Success ~ age + gender + interest + Clicks + CTR +
##     cpc + ROAS, family = "binomial", data = train_data)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.946e+01  3.079e+00  -6.319 2.64e-10 ***
## age35-39     9.233e-01  4.929e-01   1.873   0.0610 .
## age40-44     7.211e-01  5.538e-01   1.302   0.1929
## age45-49     5.831e-01  6.012e-01   0.970   0.3321
## genderM     -8.582e-01  4.641e-01  -1.849   0.0644 .
## interest7    3.124e-01  1.657e+00   0.188   0.8505
## interest10   1.837e-01  1.257e+00   0.146   0.8838
## interest15  -4.040e+00  1.697e+00  -2.381   0.0173 *
## interest16  -1.739e+00  1.310e+00  -1.328   0.1842
## interest18  -5.205e-01  1.343e+00  -0.387   0.6984
## interest19  -1.162e-01  1.494e+00  -0.078   0.9380
## interest20  -3.014e-01  1.367e+00  -0.221   0.8254
## interest21  -2.097e-01  1.379e+00  -0.152   0.8791
## interest22   2.474e-01  1.926e+00   0.128   0.8978
## interest23  -2.646e-01  1.420e+00  -0.186   0.8522
```

```
## interest24  -3.016e+00  1.865e+00  -1.618   0.1058
## interest25  -6.535e+00  7.176e+00  -0.911   0.3625
## interest26  -2.185e+00  1.706e+00  -1.280   0.2004
## interest27  -2.566e+00  1.695e+00  -1.514   0.1300
## interest28  -8.644e-01  1.558e+00  -0.555   0.5790
## interest29  -1.198e+00  1.447e+00  -0.828   0.4075
## interest30   4.680e-02  1.695e+00   0.028   0.9780
## interest31  -3.791e-01  1.784e+00  -0.213   0.8317
## interest32  -7.396e-01  1.519e+00  -0.487   0.6263
## interest36  -6.359e+00  1.221e+01  -0.521   0.6025
## interest63  -1.424e+00  1.374e+00  -1.036   0.3000
## interest64   1.504e-01  1.406e+00   0.107   0.9148
## interest65   2.087e+00  1.565e+00   1.333   0.1824
## interest66  -8.299e-02  1.654e+00  -0.050   0.9600
## interest100  4.420e-01  2.211e+00   0.200   0.8416
## interest101  1.457e+00  1.905e+00   0.765   0.4444
## interest102  2.440e+00  1.746e+00   1.397   0.1624
## interest103 -6.072e-01  4.160e+00  -0.146   0.8840
## interest104 -1.658e+01  7.912e+03  -0.002   0.9983
## interest105 -1.318e+01  1.590e+03  -0.008   0.9934
## interest106 -1.356e+01  1.628e+03  -0.008   0.9934
## interest107  1.147e+00  1.716e+00   0.669   0.5036
## interest108  1.062e+00  1.703e+00   0.623   0.5331
## interest109 -1.362e+01  1.466e+03  -0.009   0.9926
## interest110  1.326e+00  1.708e+00   0.776   0.4376
## interest111 -1.258e+01  2.121e+03  -0.006   0.9953
## interest112 -2.292e+00  5.987e+00  -0.383   0.7018
## interest113 -8.120e-03  2.338e+00  -0.003   0.9972
## interest114  2.549e-01  2.037e+00   0.125   0.9004
## Clicks      -2.581e-03  5.311e-03  -0.486   0.6270
## CTR          1.020e+01  2.722e+01   0.375   0.7079
## cpc          9.703e+00  1.532e+00   6.335 2.37e-10 ***
## ROAS         3.528e+00  4.150e-01   8.502  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 756.30  on 654  degrees of freedom
## Residual deviance: 228.86  on 607  degrees of freedom
## AIC: 324.86
##
## Number of Fisher Scoring iterations: 16
```

```
model_1 <- glm(Success ~ age + gender+ cpc + ROAS   , data = train_data, family = "binomial")
summary(model_1)
```

```
##
## Call:
## glm(formula = Success ~ age + gender + cpc + ROAS, family = "binomial",
##     data = train_data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -15.4106     1.8131  -8.500  < 2e-16 ***
## age35-39      0.7368     0.4081   1.806    0.071 .
## age40-44      0.6299     0.4466   1.411    0.158
## age45-49      0.3767     0.4824   0.781    0.435
## genderM      -0.5627     0.3468  -1.623    0.105
## cpc           7.1186     1.0137   7.022 2.18e-12 ***
## ROAS          2.9007     0.3098   9.363  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 756.30  on 654  degrees of freedom
## Residual deviance: 281.71  on 648  degrees of freedom
## AIC: 295.71
##
## Number of Fisher Scoring iterations: 9
```

```
AIC(model_1)
```

```
## [1] 295.7053
```

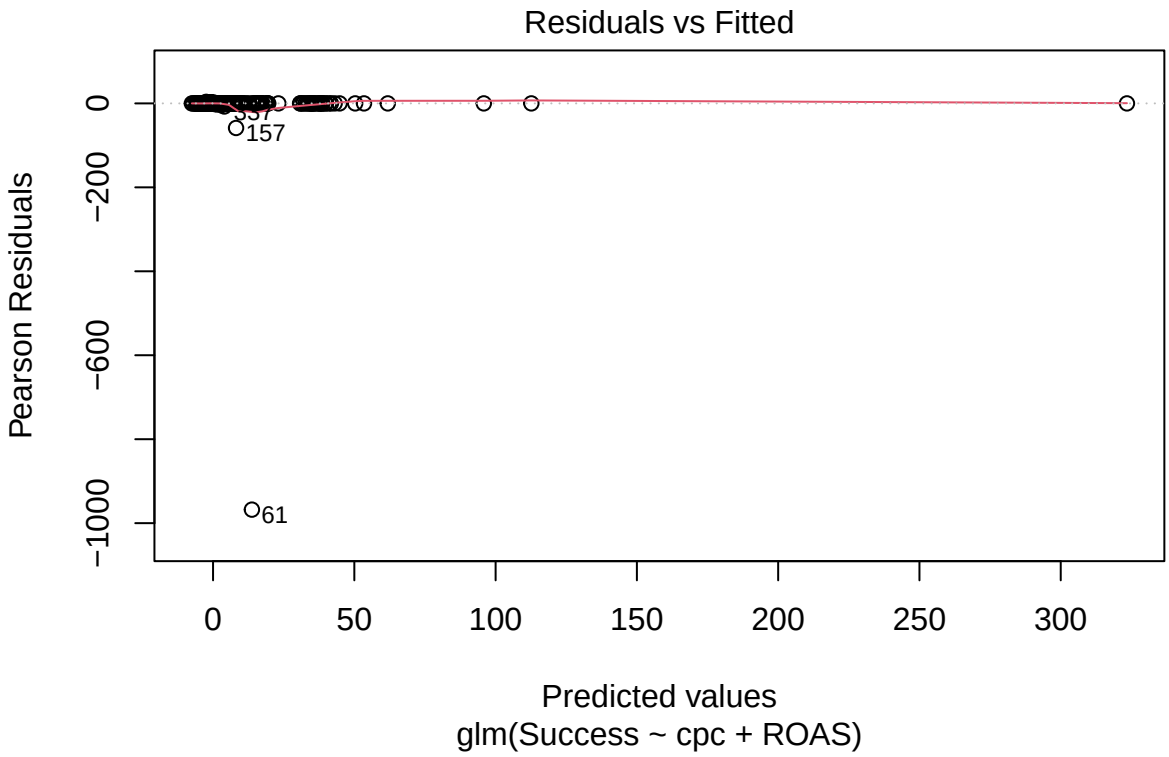
```
model_step <- step(model, trace = F)
AIC(model_step)
```

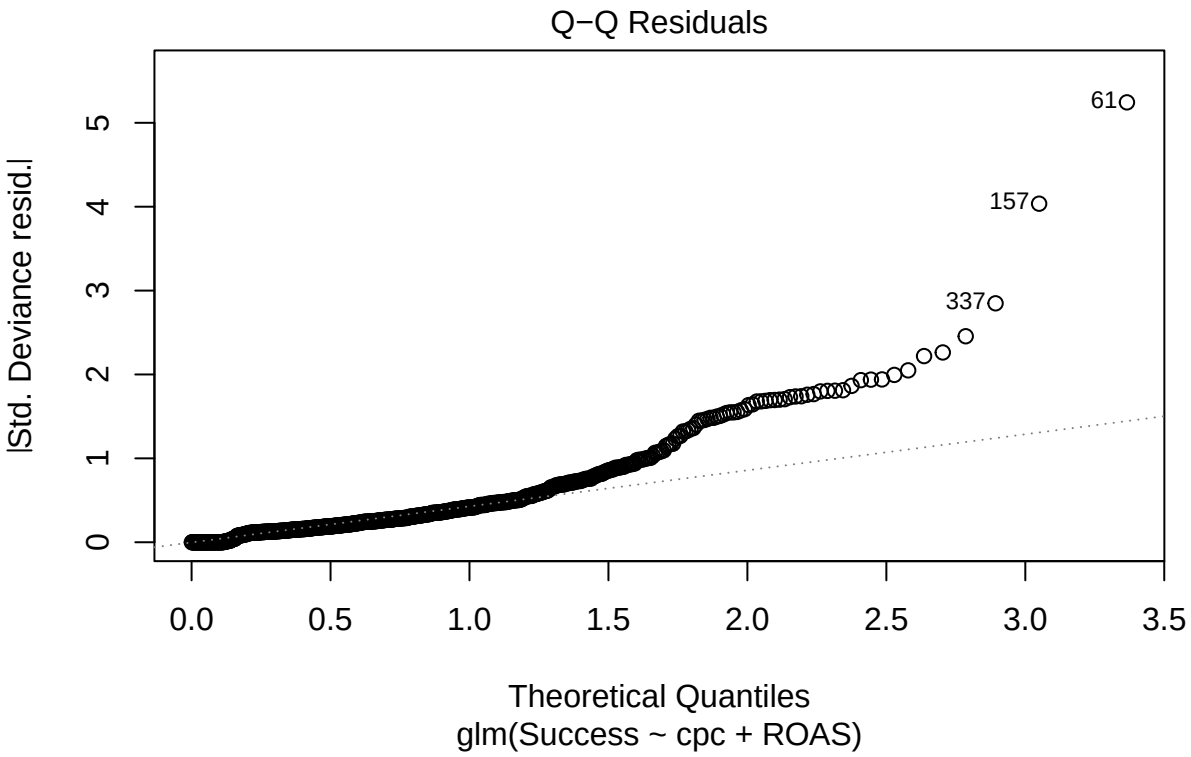
```
## [1] 293.5647
```

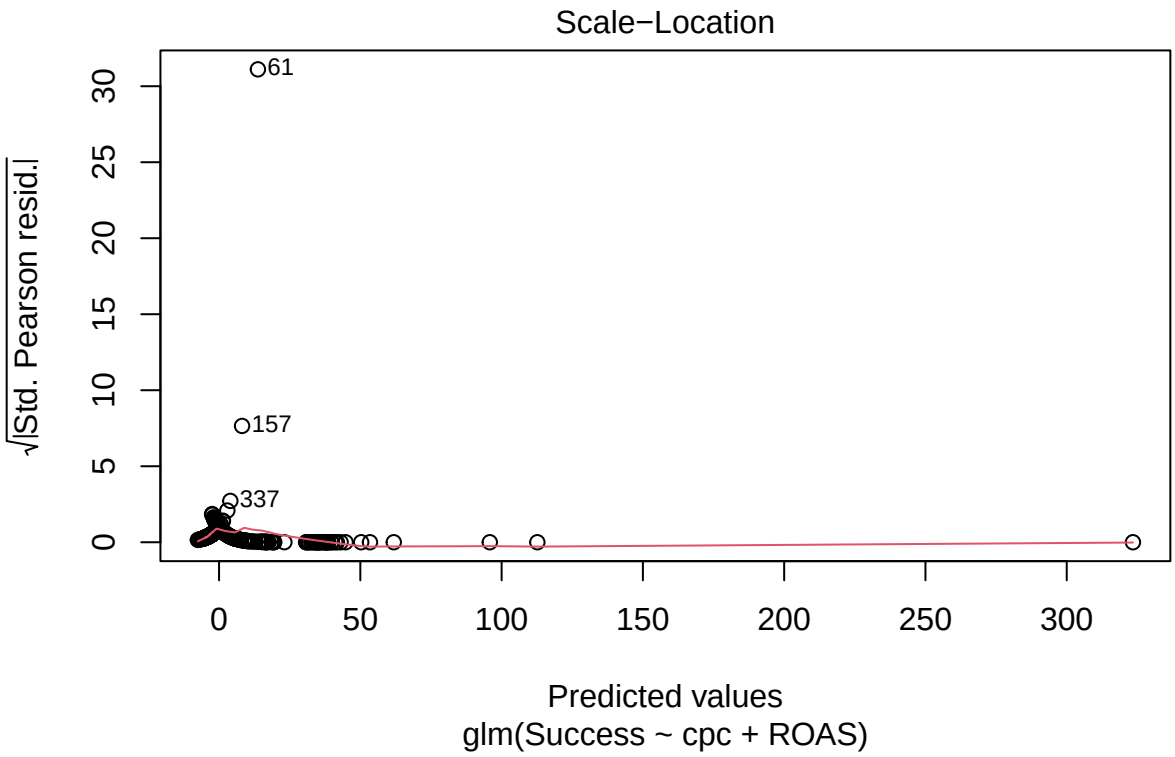
```
summary(model_step)
```

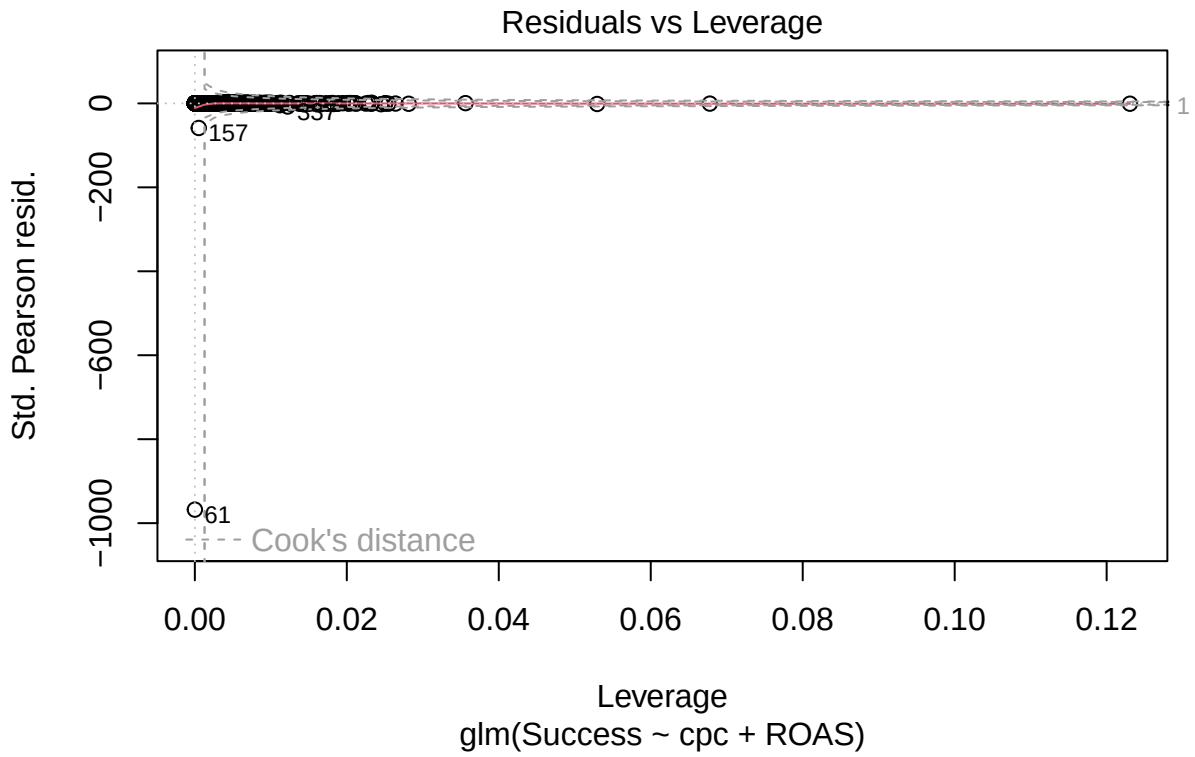
```
##
## Call:
## glm(formula = Success ~ cpc + ROAS, family = "binomial", data = train_data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -14.0394     1.6059  -8.742  < 2e-16 ***
## cpc           6.3726     0.8895   7.165  7.8e-13 ***
## ROAS          2.7519     0.2921   9.422  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 756.30  on 654  degrees of freedom
## Residual deviance: 287.56  on 652  degrees of freedom
## AIC: 293.56
##
## Number of Fisher Scoring iterations: 9
```

```
plot(model_step)
```


Residuals vs Fitted
Pearson Residuals
0
−200
−600
−1000
157
61
0
50
100
150
200
250
300
Predicted values
glm(Success ~ cpc + ROAS)


Q−Q Residuals
|Std. Deviance resid.|
0
1
2
3
4
5
61
157
337
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Theoretical Quantiles
glm(Success ~ cpc + ROAS)


Scale−Location
√|Std. Pearson resid.|
61
157
337
0
5
10
15
20
25
30
0
50
100
150
200
250
300
Predicted values
glm(Success ~ cpc + ROAS)




```
probs_test <- predict(model_step, newdata = test_data,type = "response")
length(probs_test)
```

```
## [1] 281
```

```
preds_test <- rep(0, 281)
preds_test[probs_test > 0.5] <- 1

cm <- caret::confusionMatrix(table(preds_test, test_data$Success))
print(cm)
```

```
## Confusion Matrix and Statistics
##
##
## preds_test   0   1
##          0 186  22
##          1   6  67
##
##                Accuracy : 0.9004
##                  95% CI : (0.8592, 0.9328)
##     No Information Rate : 0.6833
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7581
```

```
##
##  Mcnemar's Test P-Value : 0.004586
##
##             Sensitivity : 0.9688
##             Specificity : 0.7528
##          Pos Pred Value : 0.8942
##          Neg Pred Value : 0.9178
##              Prevalence : 0.6833
##          Detection Rate : 0.6619
##    Detection Prevalence : 0.7402
##       Balanced Accuracy : 0.8608
##
##        'Positive' Class : 0
##
```

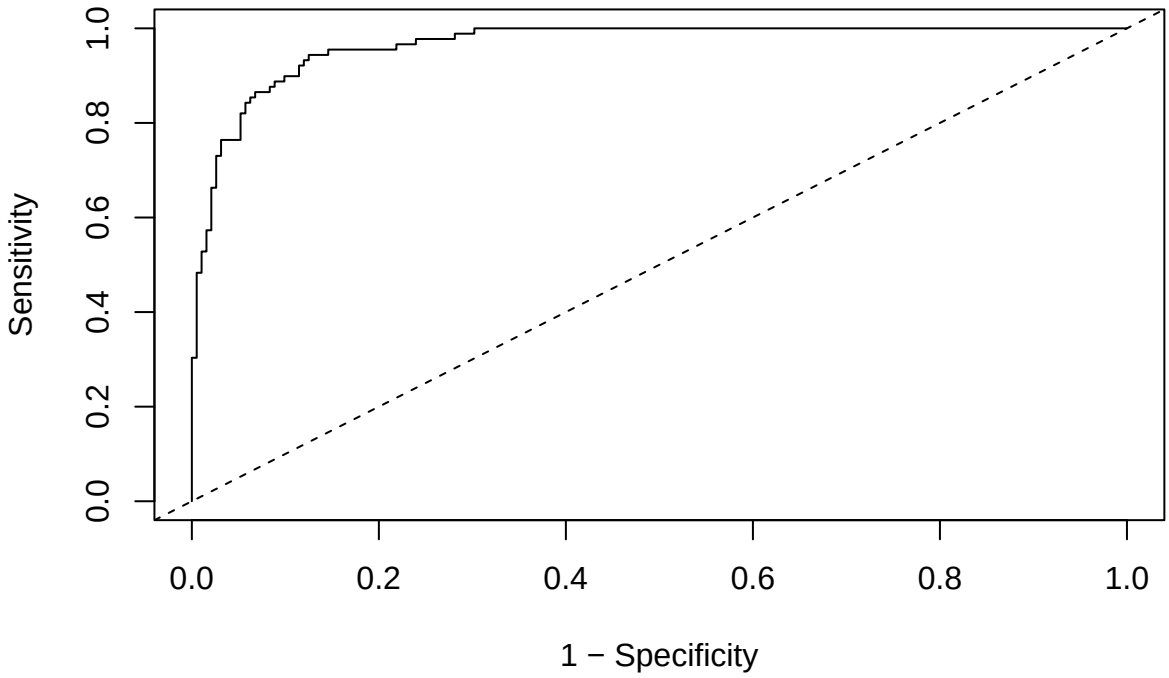
```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
roc_obj <- roc(test_data$Success, probs_test)
```

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases
```

```
plot(1 - roc_obj$specificities, roc_obj$sensitivities, type="l",
xlab = "1 - Specificity", ylab = "Sensitivity")
# plot red point corresponding to 0.5 threshold:
 points(x = 1-cm$specificity, y = cm$sensitivity, col="red", pch=19)
 abline(0, 1, lty=2) # 1-1 line
```




```
auc(roc_obj)
```

```
## Area under the curve: 0.9658
```