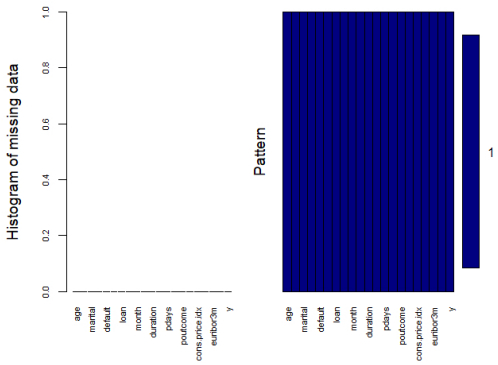
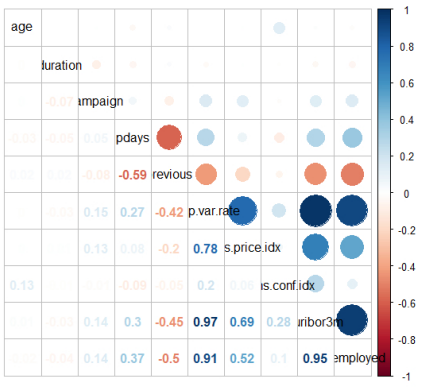
```r
###LOAD###
library(knitr)      # web widget
library(tidyverse)  # data manipulation
#> Warning: package 'tidyverse' was built under R version 4.0.3
#> Warning: package 'tidyr' was built under R version 4.0.3
library(data.table) # fast file reading
#>
#> Attaching package: 'data.table'
#> The following objects are masked from 'package:dplyr':
#>
#>     between, first, last
#> The following object is masked from 'package:purrr':
#>
#>     transpose
library(caret)      # rocr analysis
#> Warning: package 'caret' was built under R version 4.0.3
#> Loading required package: lattice
#>
#> Attaching package: 'caret'
#> The following object is masked from 'package:purrr':
#>
#>     lift
library(ROCR)       # rocr analysis
#> Warning: package 'ROCR' was built under R version 4.0.3
library(kableExtra) # nice table html formating
#> Warning: package 'kableExtra' was built under R version 4.0.3
#>
#> Attaching package: 'kableExtra'
#> The following object is masked from 'package:dplyr':
#>
#>     group_rows
library(gridExtra)  # arranging ggplot in grid
#> Warning: package 'gridExtra' was built under R version 4.0.3
#>
#> Attaching package: 'gridExtra'
#> The following object is masked from 'package:dplyr':
#>
#>     combine
library(rpart)      # decision tree
library(rpart.plot) # decision tree plotting
#> Warning: package 'rpart.plot' was built under R version 4.0.3
library(caTools)    # split
#> Warning: package 'caTools' was built under R version 4.0.3
library(colorspace)
library(grid)
library(VIM)    # split
#> Warning: package 'VIM' was built under R version 4.0.3
#> VIM is ready to use.
#> Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
#>
#> Attaching package: 'VIM'
#> The following object is masked from 'package:datasets':
#>
#>     sleep
library(corrplot)
#> Warning: package 'corrplot' was built under R version 4.0.3
#> corrplot 0.84 loaded
library(ggplot2)
library(plyr)
#> ------------------------------------------------------------------------------
#> You have loaded plyr after dplyr - this is likely to cause problems.
#> If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
#> library(plyr); library(dplyr)
#> ------------------------------------------------------------------------------
#>
#> Attaching package: 'plyr'
#> The following objects are masked from 'package:dplyr':
#>
#>     arrange, count, desc, failwith, id, mutate, rename, summarise,
#>     summarize
#> The following object is masked from 'package:purrr':
#>
#>     compact
library(psych)
#> Warning: package 'psych' was built under R version 4.0.3
#>
#> Attaching package: 'psych'
#> The following objects are masked from 'package:ggplot2':
#>
#>     %+%, alpha
library(dplyr)
require(tidyr)
require(lazyeval)
#> Loading required package: lazyeval
#>
#> Attaching package: 'lazyeval'
#> The following objects are masked from 'package:purrr':
#>
#>     is_atomic, is_formula
library(Epi)#ROC Curve
#> Warning: package 'Epi' was built under R version 4.0.3
library(nnet)
####DESCRIPTIVE STATISTICS####
#Read dataset
bank_full = read.csv(file = 'bank-additional-full.csv', sep =';', header = TRUE)

#Size of dataset
nrow(bank_full)
#> [1] 41188
#[1] 41188

ncol(bank_full)
#> [1] 21
#[1] 20

#view statistical summary
summary(bank_full)
#>       age            job              marital           education
#>  Min.   :17.00   Length:41188       Length:41188       Length:41188
#>  1st Qu.:32.00   Class :character   Class :character   Class :character
#>  Median :38.00   Mode  :character   Mode  :character   Mode  :character
#>  Mean   :40.02
#>  3rd Qu.:47.00
#>  Max.   :98.00
#>    default            housing              loan              contact
#>  Length:41188       Length:41188       Length:41188       Length:41188
#>  Class :character   Class :character   Class :character   Class :character
#>  Mode  :character   Mode  :character   Mode  :character   Mode  :character
#>
#>
#>
#>     month           day_of_week          duration         campaign
#>  Length:41188       Length:41188       Min.   :   0.0   Min.   : 1.000
#>  Class :character   Class :character   1st Qu.: 102.0   1st Qu.: 1.000
#>  Mode  :character   Mode  :character   Median : 180.0   Median : 2.000
#>                                        Mean   : 258.3   Mean   : 2.568
#>                                        3rd Qu.: 319.0   3rd Qu.: 3.000
#>                                        Max.   :4918.0   Max.   :56.000
#>      pdays          previous       poutcome          emp.var.rate
#>  Min.   :  0.0   Min.   :0.000   Length:41188       Min.   :-3.40000
#>  1st Qu.:999.0   1st Qu.:0.000   Class :character   1st Qu.:-1.80000
#>  Median :999.0   Median :0.000   Mode  :character   Median : 1.10000
#>  Mean   :962.5   Mean   :0.173                      Mean   : 0.08189
#>  3rd Qu.:999.0   3rd Qu.:0.000                      3rd Qu.: 1.40000
#>  Max.   :999.0   Max.   :7.000                      Max.   : 1.40000
#>  cons.price.idx  cons.conf.idx     euribor3m      nr.employed
#>  Min.   :92.20   Min.   :-50.8   Min.   :0.634   Min.   :4964
#>  1st Qu.:93.08   1st Qu.:-42.7   1st Qu.:1.344   1st Qu.:5099
#>  Median :93.75   Median :-41.8   Median :4.857   Median :5191
#>  Mean   :93.58   Mean   :-40.5   Mean   :3.621   Mean   :5167
```

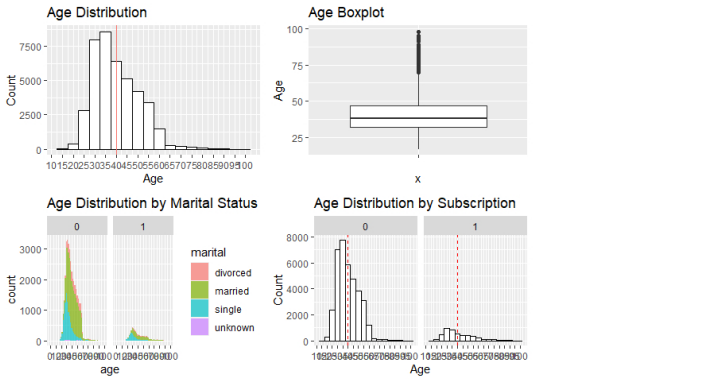
```
#>  3rd Qu.:93.99   3rd Qu.:-36.4   3rd Qu.:4.961   3rd Qu.:5228
#>  Max.   :94.77   Max.   :-26.9   Max.   :5.045   Max.   :5228
#>       y
#>  Length:41188
#>  Class :character
#>  Mode  :character
#>
#>
#>

#sample
head(bank_full)
#>   age       job marital    education default housing loan    contact month
#> 1  56 housemaid married     basic.4y      no      no   no  telephone   may
#> 2  57  services married high.school unknown      no   no  telephone   may
#> 3  37  services married high.school      no     yes   no  telephone   may
#> 4  40   admin. married     basic.6y      no      no   no  telephone   may
#> 5  56  services married high.school      no      no  yes  telephone   may
#> 6  45  services married     basic.9y unknown      no   no  telephone   may
#>   day_of_week duration campaign pdays previous     poutcome emp.var.rate
#> 1         mon      261        1   999        0 nonexistent          1.1
#> 2         mon      149        1   999        0 nonexistent          1.1
#> 3         mon      226        1   999        0 nonexistent          1.1
#> 4         mon      151        1   999        0 nonexistent          1.1
#> 5         mon      307        1   999        0 nonexistent          1.1
#> 6         mon      198        1   999        0 nonexistent          1.1
#>   cons.price.idx cons.conf.idx euribor3m nr.employed  y
#> 1         93.994         -36.4     4.857       5191 no
#> 2         93.994         -36.4     4.857       5191 no
#> 3         93.994         -36.4     4.857       5191 no
#> 4         93.994         -36.4     4.857       5191 no
#> 5         93.994         -36.4     4.857       5191 no
#> 6         93.994         -36.4     4.857       5191 no

#Replace the target variables yes = 1, no = 0
bank_full$y<-ifelse(bank_full$y =='yes', 1,0)
bank_full$y<-as.factor(bank_full$y)

#Check for duplicate rows
sum(duplicated(bank_full))
#> [1] 12
#[1] 1784

sum(!complete.cases(bank_full))
#> [1] 0
#[1] 0

#Plot the missing values
aggr_plot <- aggr(bank_full, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(bank_full), cex.axis=.7, gap
=3, ylab=c("Histogram of missing data","Pattern"))
```



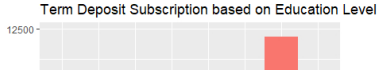

```
#>
#>  Variables sorted by number of missings:
#>        Variable Count
#>             age     0
#>             job     0
#>         marital     0
#>       education     0
#>         default     0
#>         housing     0
#>            loan     0
#>         contact     0
#>           month     0
#>     day_of_week     0
#>        duration     0
#>        campaign     0
#>           pdays     0
#>        previous     0
#>        poutcome     0
#>    emp.var.rate     0
#>  cons.price.idx     0
#>   cons.conf.idx     0
#>       euribor3m     0
#>     nr.employed     0
#>               y     0

#get numerical
new_df <- bank_full[sapply(bank_full,is.numeric)]
#Plot correlation plot
corrplot.mixed(cor(new_df),
               lower = "number",
               upper = "circle",
               tl.col = "black")
```

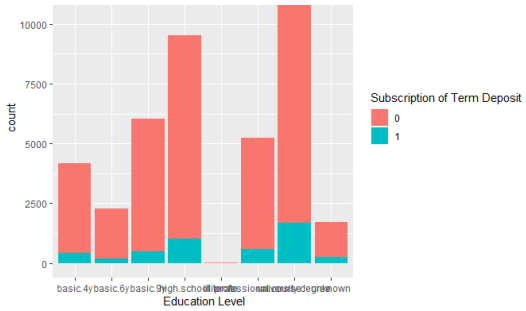


```
#rename y to Target
colnames(bank_full)[colnames(bank_full) == 'y'] <- 'target'

#####CATEGORICAL VARIABLES PLOT######
#plot the categorical variables change job to the relevant column to plot
ggplot(bank_full %>% count(job, target) %>%
         mutate(pct=n/sum(n),
                ypos = cumsum(n) - 0.5*n),
       aes(job, n, fill=target)) +
  geom_bar(stat="identity")
#> Error in count(., job, target): object 'job' not found


ggplot(bank_full %>% count(day_of_week, target) %>%    # Group by region and species, then count number in each group
         mutate(pct=n/sum(n),                # Calculate percent within each region
                ypos = cumsum(n) - 0.5*n),  # Calculate label positions
       aes(day_of_week, n, fill=target)) +
  geom_bar(stat="identity")
#> Error in count(., day_of_week, target): object 'day_of_week' not found

ggplot(bank_full %>% count(housing, target) %>%    # Group by region and species, then count number in each group
         mutate(pct=n/sum(n),                # Calculate percent within each region
                ypos = cumsum(n) - 0.5*n),  # Calculate label positions
       aes(housing, n, fill=target)) +
  geom_bar(stat="identity")
#> Error in count(., housing, target): object 'housing' not found

ggplot(bank_full %>% count(marital, target) %>%    # Group by region and species, then count number in each group
         mutate(pct=n/sum(n),                # Calculate percent within each region
                ypos = cumsum(n) - 0.5*n),  # Calculate label positions
       aes(marital, n, fill=target)) +
  geom_bar(stat="identity")
#> Error in count(., marital, target): object 'marital' not found

ggplot(bank_full %>% count(poutcome, target) %>%    # Group by region and species, then count number in each group
         mutate(pct=n/sum(n),                # Calculate percent within each region
                ypos = cumsum(n) - 0.5*n),  # Calculate label positions
       aes(poutcome, n, fill=target)) +
  geom_bar(stat="identity")
#> Error in count(., poutcome, target): object 'poutcome' not found

ggplot(bank_full %>% count(default, target) %>%    # Group by region and species, then count number in each group
         mutate(pct=n/sum(n),                # Calculate percent within each region
                ypos = cumsum(n) - 0.5*n),  # Calculate label positions
       aes(default, n, fill=target)) +
  geom_bar(stat="identity")
#> Error in count(., default, target): object 'default' not found
#####CATEGORICAL VARIABLES PLOT END######


#####PLOT OTHER VARIABLES AGE####
##Check age distribution
summary(bank_full$age)
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>   17.00   32.00   38.00   40.02   47.00   98.00
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 17.00   32.00   38.00   40.02   47.00   98.00

#Plot the age distribution
gg = ggplot (bank_full)
age_dis = gg + geom_histogram(aes(x=age),color="black", fill="white", binwidth = 5) +
  ggtitle('Age Distribution') +
  ylab('Count') +
  xlab('Age') +
  geom_vline(aes(xintercept = mean(age), color = "red")) +
  scale_x_continuous(breaks = seq(0,100,5)) +
  theme(legend.position = "none")

#Plot the age through boxplot
age_box_plot = gg + geom_boxplot(aes(x='', y=age)) +
  ggtitle('Age Boxplot') +
  ylab('Age')

##Plot Age distribution X Marital Status
age_marital <- ggplot(bank_full, aes(x=age, fill=marital)) +
  geom_histogram(binwidth = 2, alpha=0.7) +
  facet_grid(cols = vars(target)) +
  expand_limits(x=c(0,100)) +
  scale_x_continuous(breaks = seq(0,100,10)) +
  ggtitle("Age Distribution by Marital Status")

#Get the mean age
mean_age <- bank_full %>% group_by(target) %>% summarise(grp.mean=mean(age))

#Plot age and mean
age_sub <- ggplot (bank_full, aes(x=age)) +
  geom_histogram(color = "black", fill = "white", binwidth = 5) +
  facet_grid(cols=vars(target)) +
  ggtitle('Age Distribution by Subscription') + ylab('Count') + xlab('Age') +
  scale_x_continuous(breaks = seq(0,100,5)) +
  geom_vline(data=mean_age, aes(xintercept=grp.mean), color="red", linetype="dashed")

#Print all the plots
grid.arrange(age_dis,age_box_plot,age_marital,age_sub, ncol = 2, nrow = 2)
```

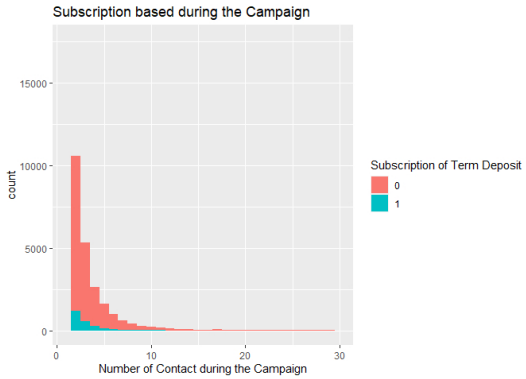



```
#####PLOT OTHER VARIABLES AGE END####

#####PLOT OTHER VARIABLES EDU + SUBS####

#Education X subscription
ggplot(data = bank_full, aes(x=education, fill=target)) +
  geom_bar() +
  ggtitle("Term Deposit Subscription based on Education Level") +
  xlab(" Education Level") +
  guides(fill=guide_legend(title="Subscription of Term Deposit"))
```

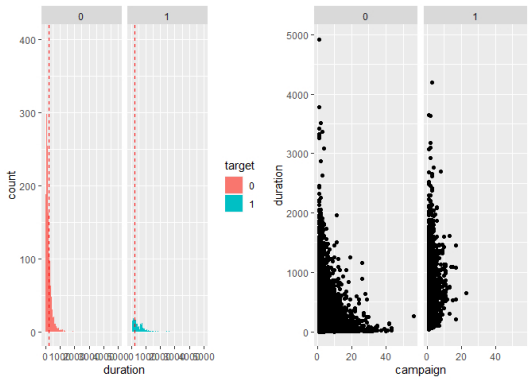




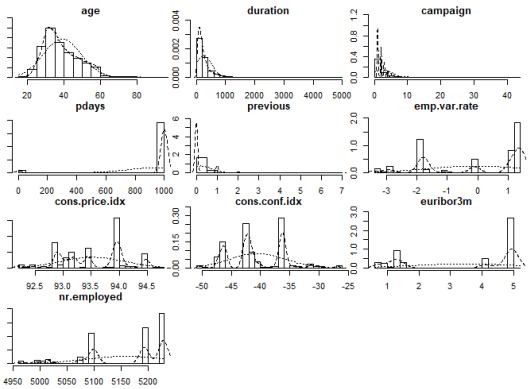

```
#####PLOT OTHER VARIABLES EDU + END####

#####PLOT OTHER VARIABLES CAMPAIGN####
#Subscription based on the campaign
ggplot(data=bank_full, aes(x=campaign, fill=target))+
  geom_histogram()+
  ggtitle("Subscription based during the Campaign")+
  xlab("Number of Contact during the Campaign")+
  xlim(c(min=1,max=30)) +
  guides(fill=guide_legend(title="Subscription of Term Deposit"))
#> `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
#> Warning: Removed 33 rows containing non-finite values (stat_bin).
#> Warning: Removed 4 rows containing missing values (geom_bar).
```



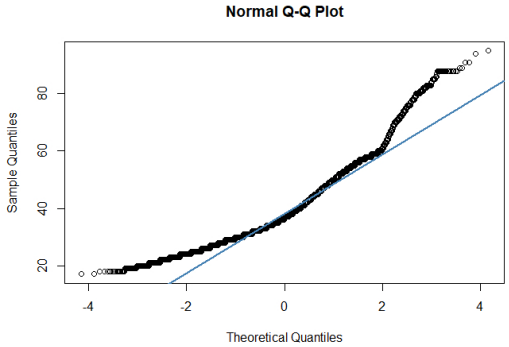

```
#####PLOT OTHER VARIABLES CAMPAIGN####


#####PLOT OTHER VARIABLES DURATION####

##Duration
range(bank_full$duration)
#> [1]    0 4918
#[1]    0 4918

summary(bank_full$duration)
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>     0.0   102.0   180.0   258.3   319.0  4918.0
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 0.0   102.0   180.0   258.3   319.0  4918.0

bank_full %>% select(duration) %>% arrange(desc(duration)) %>% head
#>   duration
#> 1     4918
#> 2     4199
#> 3     3785
#> 4     3643
#> 5     3631
#> 6     3509

##get the mean duration
mean_duration<- bank_full %>% group_by(target) %>% summarise(grp2.mean=mean(duration))

##plot the duration
duration_count <- ggplot(bank_full, aes(x=duration, fill = target)) +
  geom_histogram(binwidth = 2) +
  facet_grid(cols = vars(target)) +
  coord_cartesian(xlim = c(0,5000), ylim = c(0,400)) + geom_vline(data = mean_duration, aes(xintercept = grp2.mean), color =
"red", linetype = "dashed")

##duration x camp
duration_camp <- bank_full %>% filter(campaign < 63) %>%
  ggplot(aes(campaign, duration)) +
  geom_point() +
  facet_grid(cols = vars(target))

#print the plots
grid.arrange(duration_count,duration_camp, ncol = 2)
```



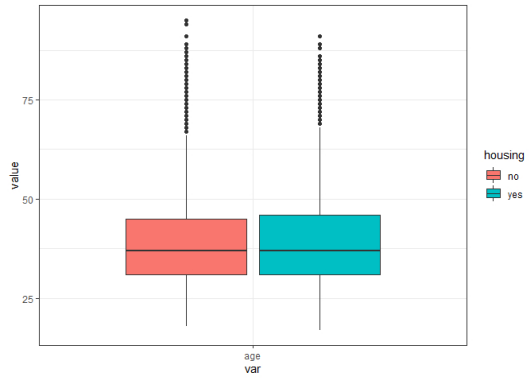
```
#####PLOT OTHER VARIABLES DURATION####

##############################STATISTICAL EVDENCE#####################################


#Replace all unknown with NA
bank_full[bank_full=="unknown"] <- NA

#remove na
bank_full <- bank_full[complete.cases(bank_full),]

#plot and make numeric
multi.hist(bank_full[,sapply(bank_full, is.numeric)])
```




```
pastecs::stat.desc(bank_full$age, basic=F)
#>       median         mean      SE.mean CI.mean.0.95          var      std.dev
#>  37.00000000  39.03001181   0.05918126   0.11599774 106.78182633  10.33352923
#>     coef.var
#>   0.26475855

#QQplot age
qqnorm(bank_full$age); qqline(bank_full$age,col ="steelblue", lwd = 2)
```




```
#Homogeneity of Variance
bank_full %>% gather(age, key = 'var', value = 'value') %>%
  ggplot(aes(x = var, y = value, fill = housing)) +
  geom_boxplot() +
  theme_bw()
```

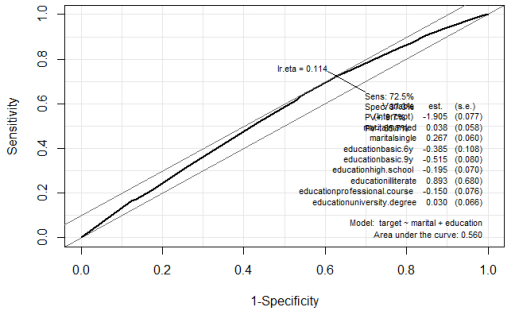



```
##LAVENE TEST
car::leveneTest(age ~ housing, data=bank_full)
#> Warning in leveneTest.default(y = y, group = group, ...): group coerced to
#> factor.
#> Levene's Test for Homogeneity of Variance (center = median)
#>          Df F value Pr(>F)
#> group     1  1.9359 0.1641
#>       30486
car::leveneTest(age ~ housing, data=bank_full, center = 'mean')
#> Warning in leveneTest.default(y = y, group = group, ...): group coerced to
#> factor.
#> Levene's Test for Homogeneity of Variance (center = "mean")
#>          Df F value  Pr(>F)
#> group     1  3.3772 0.06611 .
#>       30486
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##T-TEST
stats::t.test(age~housing,var.equal=TRUE,data=bank_full)
#>
#>  Two Sample t-test
#>
#> data:  age by housing
```

```
#> t = -0.81333, df = 30486, p-value = 0.416
#> alternative hypothesis: true difference in means is not equal to 0
#> 95 percent confidence interval:
#>  -0.3294226  0.1362076
#> sample estimates:
#>  mean in group no mean in group yes
#>          38.97766          39.07427

psych::describeBy(bank_full$age, bank_full$housing, mat=TRUE)
#>     item group1 vars     n     mean       sd median  trimmed    mad min max
#> X11    1     no    1 13967 38.97766 10.26829     37 38.05817 8.8956  18  95
#> X12    2    yes    1 16521 39.07427 10.38847     37 38.16343 8.8956  17  91
#>     range      skew kurtosis         se
#> X11    77 0.9910800 1.295258 0.08688538
#> X12    74 0.9707474 1.196006 0.08082268

res <- stats::t.test(age~housing,var.equal=TRUE,data=bank_full)

effectsize::t_to_d(t = res$statistic, res$parameter)
#>         d |         95% CI
#> ------------------------
#> -9.32e-03 | [-0.03, 0.01]


effes=round((res$statistic*res$statistic)/((res$statistic*res$statistic)+(res$parameter)),3)
effes
#> t
#> 0


############################STATISTICAL EVDENCE END################################

############################MODEL 1################################
#LOGISTIC MODEL
logmodel1 <- glm(target ~ marital + education, data = bank_full, na.action = na.exclude, family = binomial(link=logit))

#Full summary of the model
summary(logmodel1)
#>
#> Call:
#> glm(formula = target ~ marital + education, family = binomial(link = logit),
#>     data = bank_full, na.action = na.exclude)
#>
#> Deviance Residuals:
#>     Min       1Q   Median       3Q      Max
#> -0.8004  -0.5437  -0.4999  -0.4473   2.2382
#>
#> Coefficients:
#>                                Estimate Std. Error z value Pr(>|z|)
#> (Intercept)                    -1.90478    0.07713 -24.694  < 2e-16 ***
#> maritalmarried                  0.03799    0.05770   0.658 0.510325
#> maritalsingle                   0.26714    0.06032   4.428 9.49e-06 ***
#> educationbasic.6y              -0.38478    0.10826  -3.554 0.000379 ***
#> educationbasic.9y              -0.51481    0.08043  -6.401 1.54e-10 ***
#> educationhigh.school           -0.19509    0.06976  -2.797 0.005163 **
#> educationilliterate             0.89282    0.67966   1.314 0.188973
#> educationprofessional.course   -0.15002    0.07569  -1.982 0.047481 *
#> educationuniversity.degree      0.02990    0.06648   0.450 0.652946
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for binomial family taken to be 1)
#>
#>     Null deviance: 23160  on 30487  degrees of freedom
#> Residual deviance: 23000  on 30479  degrees of freedom
#> AIC: 23018
#>
#> Number of Fisher Scoring iterations: 4

#Chi-square plus significance
lmtest::lrtest(logmodel1)
#> likelihood ratio test
#>
#> Model 1: target ~ marital + education
#> Model 2: target ~ 1
#>   #Df  LogLik Df  Chisq Pr(>Chisq)
#> 1   9 -11500
#> 2   1 -11580 -8 160.25  < 2.2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## odds ratios
cbind(Estimate=round(coef(logmodel1),4),
      OR=round(exp(coef(logmodel1)),4))
#>                               Estimate     OR
#> (Intercept)                    -1.9048 0.1489
#> maritalmarried                  0.0380 1.0387
#> maritalsingle                   0.2671 1.3062
#> educationbasic.6y              -0.3848 0.6806
#> educationbasic.9y              -0.5148 0.5976
#> educationhigh.school           -0.1951 0.8228
#> educationilliterate             0.8928 2.4420
#> educationprofessional.course   -0.1500 0.8607
#> educationuniversity.degree      0.0299 1.0303

#1. Probability of answering yes when divorced
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*0)
#> Registered S3 methods overwritten by 'lme4':
#>   method                          from
#>   cooks.distance.influence.merMod car
#>   influence.merMod                car
#>   dfbeta.influence.merMod         car
#>   dfbetas.influence.merMod        car
#> (Intercept)
#>   0.1295686
#Probability of answering yes when married
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*1)
#> (Intercept)
#>   0.1339135
#Probability of answering yes when single
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*2)
#> (Intercept)
#>    0.138381
#Probability of answering yes when unknown
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*3)
#> (Intercept)
#>   0.1429729

##with the column education
#3.Probability of answering yes when divorced
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*0 +coef(logmodel1)[3]*0+coef(logmodel1)[3]*0)
#> (Intercept)
#>   0.1295686
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*0 +coef(logmodel1)[3]*1+coef(logmodel1)[3]*0)
#> (Intercept)
#>   0.1627865
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*0 +coef(logmodel1)[3]*2+coef(logmodel1)[3]*0)
#> (Intercept)
#>   0.2025391
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*0 +coef(logmodel1)[3]*3+coef(logmodel1)[3]*0)
#> (Intercept)
#>   0.2491108
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*0 +coef(logmodel1)[3]*4+coef(logmodel1)[3]*0)
#> (Intercept)
#>   0.3023313
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*0 +coef(logmodel1)[3]*5+coef(logmodel1)[3]*0)
#> (Intercept)
#>   0.3614488
#3.Probability of answering yes when married
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*1 +coef(logmodel1)[3]*0+coef(logmodel1)[3]*0)
```

```
#> (Intercept)
#>   0.1339135
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*1 +coef(logmodel1)[3]*1+coef(logmodel1)[3]*0)
#> (Intercept)
#>   0.1680304
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*1 +coef(logmodel1)[3]*2+coef(logmodel1)[3]*0)
#> (Intercept)
#>   0.2087442
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*1 +coef(logmodel1)[3]*3+coef(logmodel1)[3]*0)
#> (Intercept)
#>   0.2562842
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*1 +coef(logmodel1)[3]*4+coef(logmodel1)[3]*0)
#> (Intercept)
#>   0.3104037
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*1 +coef(logmodel1)[3]*5+coef(logmodel1)[3]*0)
#> (Intercept)
#>   0.3702619
#Probability of answering yes when single
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*2 +coef(logmodel1)[3]*0+coef(logmodel1)[3]*0)
#> (Intercept)
#>    0.138381
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*2 +coef(logmodel1)[3]*1+coef(logmodel1)[3]*0)
#> (Intercept)
#>   0.1734082
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*2 +coef(logmodel1)[3]*2+coef(logmodel1)[3]*0)
#> (Intercept)
#>   0.2150881
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*2 +coef(logmodel1)[3]*3+coef(logmodel1)[3]*0)
#> (Intercept)
#>   0.2635916
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*2 +coef(logmodel1)[3]*4+coef(logmodel1)[3]*0)
#> (Intercept)
#>   0.3185932
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*2 +coef(logmodel1)[3]*5+coef(logmodel1)[3]*0)
#> (Intercept)
#>   0.3791624

###with neither
#3.Probability of answering yes when divorced
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*0 +coef(logmodel1)[3]*0+coef(logmodel1)[3]*0)
#> (Intercept)
#>   0.1295686
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*0 +coef(logmodel1)[3]*0+coef(logmodel1)[3]*1)
#> (Intercept)
#>   0.1627865
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*0 +coef(logmodel1)[3]*0+coef(logmodel1)[3]*2)
#> (Intercept)
#>   0.2025391
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*0 +coef(logmodel1)[3]*0+coef(logmodel1)[3]*3)
#> (Intercept)
#>   0.2491108
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*0 +coef(logmodel1)[3]*0+coef(logmodel1)[3]*4)
#> (Intercept)
#>   0.3023313
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*0 +coef(logmodel1)[3]*0+coef(logmodel1)[3]*5)
#> (Intercept)
#>   0.3614488
#3.Probability of answering yes when married
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*1 +coef(logmodel1)[3]*0+coef(logmodel1)[3]*0)
#> (Intercept)
#>   0.1339135
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*1 +coef(logmodel1)[3]*0+coef(logmodel1)[3]*1)
#> (Intercept)
#>   0.1680304
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*1 +coef(logmodel1)[3]*0+coef(logmodel1)[3]*2)
#> (Intercept)
#>   0.2087442
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*1 +coef(logmodel1)[3]*0+coef(logmodel1)[3]*3)
#> (Intercept)
#>   0.2562842
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*1 +coef(logmodel1)[3]*0+coef(logmodel1)[3]*4)
#> (Intercept)
#>   0.3104037
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*1 +coef(logmodel1)[3]*0+coef(logmodel1)[3]*5)
#> (Intercept)
#>   0.3702619
#Probability of answering yes when single 0
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*2 +coef(logmodel1)[3]*0+coef(logmodel1)[3]*0)
#> (Intercept)
#>    0.138381
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*2 +coef(logmodel1)[3]*0+coef(logmodel1)[3]*1)
#> (Intercept)
#>   0.1734082
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*2 +coef(logmodel1)[3]*0+coef(logmodel1)[3]*2)
#> (Intercept)
#>   0.2150881
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*2 +coef(logmodel1)[3]*0+coef(logmodel1)[3]*3)
#> (Intercept)
#>   0.2635916
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*2 +coef(logmodel1)[3]*0+coef(logmodel1)[3]*4)
#> (Intercept)
#>   0.3185932
arm::invlogit(coef(logmodel1)[1]+ coef(logmodel1)[2]*2 +coef(logmodel1)[3]*0+coef(logmodel1)[3]*5)
#> (Intercept)
#>   0.3791624

#Chi-square plus significance
lmtest::lrtest(logmodel1)
#> Likelihood ratio test
#>
#> Model 1: target ~ marital + education
#> Model 2: target ~ 1
#>   #Df LogLik Df  Chisq Pr(>Chisq)
#> 1   9 -11500
#> 2   1 -11580 -8 160.25  < 2.2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#Pseudo Rsquared
DescTools::PseudoR2(logmodel1, which="CoxSnell")
#>    CoxSnell
#> 0.005242304

DescTools::PseudoR2(logmodel1, which="Nagelkerke")
#>  Nagelkerke
#> 0.009850854

#Output the marital and education and ROC plot
ROC(form=target ~ marital+education, data=bank_full,plot="ROC")

#Check the assumption of linearity of independent variables and log odds using a Hosmer-Lemeshow test, if this is not statis
tically significant we are ok
generalhoslem::logitgof(bank_full$target,fitted(logmodel1))
#> Warning in generalhoslem::logitgof(bank_full$target, fitted(logmodel1)): Not
#> possible to compute 10 rows. There might be too few observations.
#>
#>  Hosmer and Lemeshow test (binary model)
#>
#> data:  bank_full$target, fitted(logmodel1)
#> X-squared = 14.036, df = 7, p-value = 0.05054

#Collinearity
vifmodel<-car::vif(logmodel1)
vifmodel
#>                GVIF Df GVIF^(1/(2*Df))
#> marital   1.039695  2        1.009779
#> education 1.039695  6        1.003249

#Tolerance
1/vifmodel
```

```
#> marital    0.9618207 0.5000000       0.9903154
#> education 0.9618207 0.1666667       0.9967613

##############################MODEL 1 END#################################

##############################MODEL 2 END#################################
#LOGISTIC MODEL
logmodel2 <- glm(target ~ marital + education, data = bank_full, na.action = na.exclude, family = binomial(link=logit))

#Full summary of the model
summary(logmodel2)
#>
#> Call:
#> glm(formula = target ~ marital + education, family = binomial(link = logit),
#>     data = bank_full, na.action = na.exclude)
#>
#> Deviance Residuals:
#>     Min       1Q   Median       3Q      Max
#> -0.8004  -0.5437  -0.4999  -0.4473   2.2382
#>
#> Coefficients:
#>                                Estimate Std. Error z value Pr(>|z|)
#> (Intercept)                   -1.90478    0.07713 -24.694  < 2e-16 ***
#> maritalmarried                 0.03799    0.05770   0.658 0.510325
#> maritalsingle                  0.26714    0.06032   4.428 9.49e-06 ***
#> educationbasic.6y             -0.38478    0.10826  -3.554 0.000379 ***
#> educationbasic.9y             -0.51481    0.08043  -6.401 1.54e-10 ***
#> educationhigh.school          -0.19509    0.06976  -2.797 0.005163 **
#> educationilliterate            0.89282    0.67966   1.314 0.188973
#> educationprofessional.course -0.15002    0.07569  -1.982 0.047481 *
#> educationuniversity.degree    0.02990    0.06648   0.450 0.652946
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for binomial family taken to be 1)
#>
#>     Null deviance: 23160  on 30487  degrees of freedom
#> Residual deviance: 23000  on 30479  degrees of freedom
#> AIC: 23018
#>
#> Number of Fisher Scoring iterations: 4

#Chi-square plus significance
lmtest::lrtest(logmodel2)
#> Likelihood ratio test
#>
#> Model 1: target ~ marital + education
#> Model 2: target ~ 1
#>   #Df LogLik Df  Chisq Pr(>Chisq)
#> 1   9 -11500
#> 2   1 -11580 -8 160.25  < 2.2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## odds ratios
cbind(Estimate=round(coef(logmodel2),4),
      OR=round(exp(coef(logmodel2)),4))
#>                               Estimate     OR
#> (Intercept)                   -1.9048 0.1489
#> maritalmarried                 0.0380 1.0387
#> maritalsingle                  0.2671 1.3062
#> educationbasic.6y             -0.3848 0.6806
#> educationbasic.9y             -0.5148 0.5976
#> educationhigh.school          -0.1951 0.8228
#> educationilliterate            0.8928 2.4420
#> educationprofessional.course  -0.1500 0.8607
#> educationuniversity.degree     0.0299 1.0303

# Probability of answering yes when the loan is no
arm::invlogit(coef(logmodel2)[1]+ coef(logmodel2)[2]*0)
#> (Intercept)
#>   0.1295686

#Probability of answering yes when the Loan is no
arm::invlogit(coef(logmodel2)[1]+ coef(logmodel2)[2]*1)
#> (Intercept)
#>   0.1339135

#Probability of answering yes when the Loan is no and age
arm::invlogit(coef(logmodel2)[1]+ coef(logmodel2)[2]*0 +coef(logmodel2)[3]*0+coef(logmodel2)[3]*1)
#> (Intercept)
#>   0.1627865

#Probability of answering yes when the Loan is yes and age
arm::invlogit(coef(logmodel2)[1]+ coef(logmodel2)[2]*1 +coef(logmodel2)[3]*0+coef(logmodel2)[3]*1)
#> (Intercept)
#>   0.1680304

#Chi-square plus significance
lmtest::lrtest(logmodel2)
#> Likelihood ratio test
#>
#> Model 1: target ~ marital + education
#> Model 2: target ~ 1
#>   #Df LogLik Df  Chisq Pr(>Chisq)
#> 1   9 -11500
#> 2   1 -11580 -8 160.25  < 2.2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#Pseudo Rsquared
DescTools::PseudoR2(logmodel2, which="CoxSnell")
#>    CoxSnell
#> 0.005242304

DescTools::PseudoR2(logmodel2, which="Nagelkerke")
#>  Nagelkerke
#> 0.009850854

#Output the marital and education and ROC plot
ROC(form=target ~ marital+education, data=bank_full,plot="ROC")
```



```r
#Check the assumption of linearity of independent variables and log odds using a Hosmer-Lemeshow test, if this is not statis
tically significant we are ok
generalhoslem::logitgof(bank_full$target,fitted(logmodel2))
#> Warning in generalhoslem::logitgof(bank_full$target, fitted(logmodel2)): Not
#> possible to compute 10 rows. There might be too few observations.
#>
#>  Hosmer and Lemeshow test (binary model)
#>
#> data:  bank_full$target, fitted(logmodel2)
#> X-squared = 14.036, df = 7, p-value = 0.05054

#Collinearity
vifmodel<-car::vif(logmodel2)
vifmodel
#>               GVIF Df GVIF^(1/(2*Df))
#> marital   1.039695  2        1.009779
#> education 1.039695  6        1.003249

#Tolerance
1/vifmodel
#>                 GVIF        Df GVIF^(1/(2*Df))
#> marital   0.9618207 0.5000000       0.9903154
#> education 0.9618207 0.1666667       0.9967613
################################MODEL 2 END#################################


#general assumption to count and the frequency
count(bank_full$marital)
#>         x  freq
#> 1 divorced  3553
#> 2  married 17492
#> 3   single  9443

count(bank_full$education)
#>                     x  freq
#> 1           basic.4y  2380
#> 2           basic.6y  1389
#> 3           basic.9y  4276
#> 4        high.school  7699
#> 5         illiterate    11
#> 6 professional.course  4321
#> 7   university.degree 10412

count(bank_full$housing)
#>     x  freq
#> 1  no 13967
#> 2 yes 16521

count(bank_full$age)
#>     x freq
#> 1  17    2
#> 2  18   15
#> 3  19   21
#> 4  20   47
#> 5  21   84
#> 6  22  115
#> 7  23  205
#> 8  24  375
#> 9  25  500
#> 10 26  599
#> 11 27  699
#> 12 28  809
#> 13 29 1263
#> 14 30 1441
#> 15 31 1643
#> 16 32 1555
#> 17 33 1524
#> 18 34 1431
#> 19 35 1399
#> 20 36 1391
#> 21 37 1140
#> 22 38 1033
#> 23 39 1035
#> 24 40  807
#> 25 41  906
#> 26 42  793
#> 27 43  741
#> 28 44  677
#> 29 45  664
#> 30 46  659
#> 31 47  600
#> 32 48  670
#> 33 49  516
#> 34 50  552
#> 35 51  460
#> 36 52  506
#> 37 53  469
#> 38 54  432
#> 39 55  400
#> 40 56  407
#> 41 57  372
#> 42 58  383
#> 43 59  266
#> 44 60  156
#> 45 61   64
#> 46 62   42
#> 47 63   41
#> 48 64   45
#> 49 65   36
#> 50 66   40
#> 51 67   21
#> 52 68   30
#> 53 69   30
#> 54 70   41
#> 55 71   46
#> 56 72   28
#> 57 73   28
#> 58 74   29
#> 59 75   20
#> 60 76   28
#> 61 77   12
#> 62 78   20
#> 63 79   12
#> 64 80   26
#> 65 81   16
#> 66 82   13
#> 67 83   15
#> 68 84    4
#> 69 85    7
#> 70 86    3
#> 71 87    1
#> 72 88   22
#> 73 89    2
#> 74 91    2
#> 75 94    1
#> 76 95    1




library(reprex)
#> Warning: package 'reprex' was built under R version 4.0.3
r.file <- "PSI_INP_TU060_DS_c15339871_Erika-Secillano.r" ## path to your R script file
reprex(input = r.file, outfile = NA)
#> Warning in readLines(path): incomplete final line found on
#> 'PSI_INP_TU060_DS_c15339871_Erika-Secillano.r'
```