**Session 11 Assignment2**

output:

word\_document: default

html\_document:

df\_print: paged

---

```{r}

#Variable Description

#age: age of client

#job : type of job

#marital : marital status

#education: highest educational achievement

#default: has credit in default?

#housing: has housing loan?

#loan: has personal loan?

#contact: contact communication type

#month: last contact month of year

#day\_of\_week: last contact day of the week

#duration: last contact duration, in seconds

#campaign: number of contacts performed during this campaign and for this client

#pdays: number of days that passed by after the client was last contacted from a previous campaign

(999 means client was not previously contacted)

#previous: number of contacts performed before this campaign and for this client

#poutcome: outcome of the previous marketing campaign

#emp.var.rate: employment variation rate - quarterly indicator

#cons.price.idx: consumer price index - monthly indicator

#cons.conf.idx: consumer confidence index - monthly indicator

#euribor3m: euribor 3 month rate - daily indicator

#nr.employed: number of employees - quarterly indicator

#y - has the client subscribed a term deposit?

## The data set can be obtained from http://archive.ics.uci.edu/ml/datasets/Bank+Marketing

## DATASET UNDERSTANDING

library(readr)

bank\_full <- read\_delim("C:/Users/Seshan/Desktop/Bank/bank-full.csv",

";", escape\_double = FALSE, trim\_ws = TRUE)

#Lets look at dataset and generate initial understanding about the column types

str(bank\_full)

#A deep check for NA in a particular column let say age

if(length(which(is.na(bank\_full$age)==TRUE)>0)){

print("Missing Value found in the specified column")

} else

print("All okay: No Missing Value found in the specified column")

# Check another example say

if(length(which(is.na(bank\_full$campaign)==TRUE)>0)){print("Missing Value found in the specified

column")} else

print("All okay: No Missing Value found in the specified column")

head(bank\_full) ## Displays first 6 rows for each variable

str(bank\_full) ## Describes each variables

summary(bank\_full) ## Provides basic statistical information of each variable

## DATA EXPLORATION - Check for Missing Data

## Option 1

is.na(bank\_full) ## Displays True for a missing value

## Since it is a large dataset, graphical display of missing values will prove to be easier

##Option 2

require(Amelia)

missmap(bank\_full,main="Missing Data - Bank ", col=c("red","grey"),legend=FALSE)

## No red colour stripes are visible. hence no missing values.

summary(bank\_full) ## displays missing values if any under every variable

#The Pearson’s chi-squared test of independence is one of the most basic and common hypothesis tests

in the statistical analysis of categorical data. It is a significance test. Given two categorical random

variables, X and Y, the chi-squared test of independence determines whether or not there exists a

statistical dependence between them. Formally, it is a hypothesis test. The chi-squared test assumes a

null hypothesis and an alternate hypothesis. The general practice is, if the p-value that comes out in the

result is less than a pre-determined significance level, which is 0.05 usually, then we reject the null

hypothesis.

#H0: The The two variables are independent

#H1: The The two variables are dependent

#The null hypothesis of the chi-squared test is that the two variables are independent and the alternate

hypothesis is that they are related.

#To establish that two categorical variables (or predictors) are dependent, the chi-squared statistic must

have a certain cutoff. This cutoff increases as the number of classes within the variable (or predictor)

increases.

#i. Pearson’s chi-squared test of independence (significance test)

**Is there any association between Job and default?**

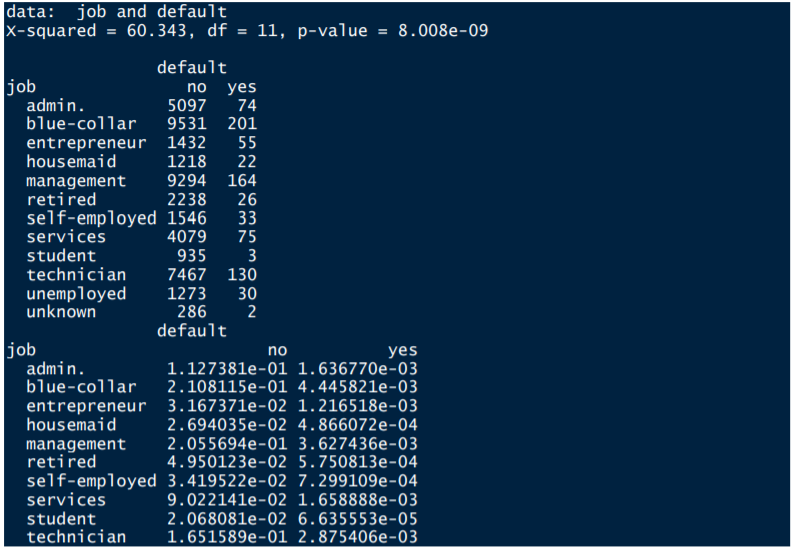
with(bank\_full, chisq.test( job, default))

with(bank\_full, table( job, default) )

# OR

with(bank\_full, prop.table(table( job,default)))

#Pearson's Chi-squared test



p-value = 8.008e-09

#Pearson's Chi-squared test

#since the p-value is < 2.2e-16 is less than the cu$t-off value of 0.05, we can reject the null hypothesis in

favor of alternative hypothesis and conclude, that the variables,( job & default- p-value = 8.008e-09) are

dependent to each other.

**b. Is there any significant difference in duration of last call between people having housing loan or not?**

with(bank\_additional\_full, chisq.test(duration,housing))

with(bank\_additional\_full, table( duration,housing) )

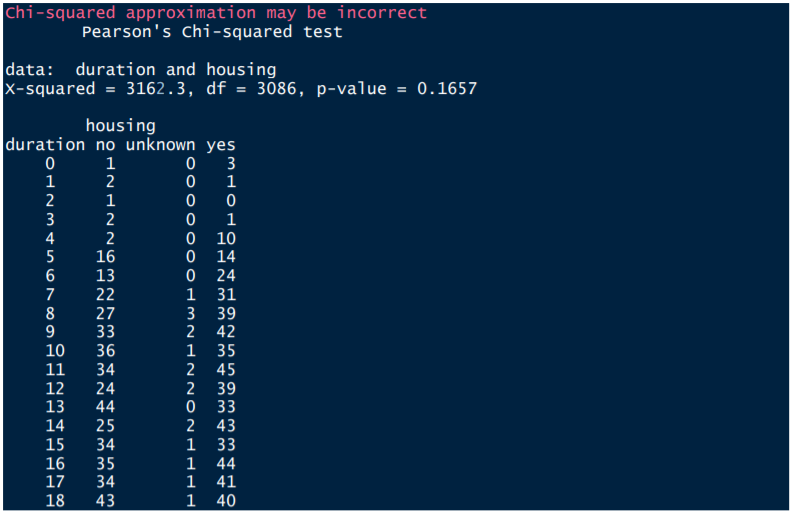
# OR

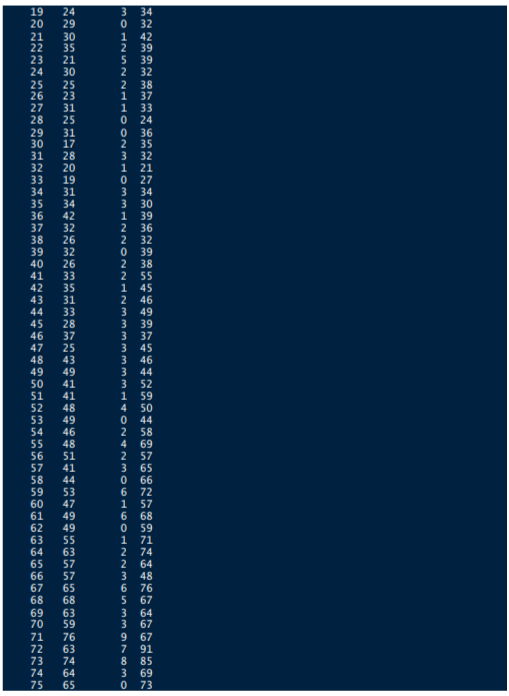
with(bank\_additional\_full, prop.table(table(duration, housing)))

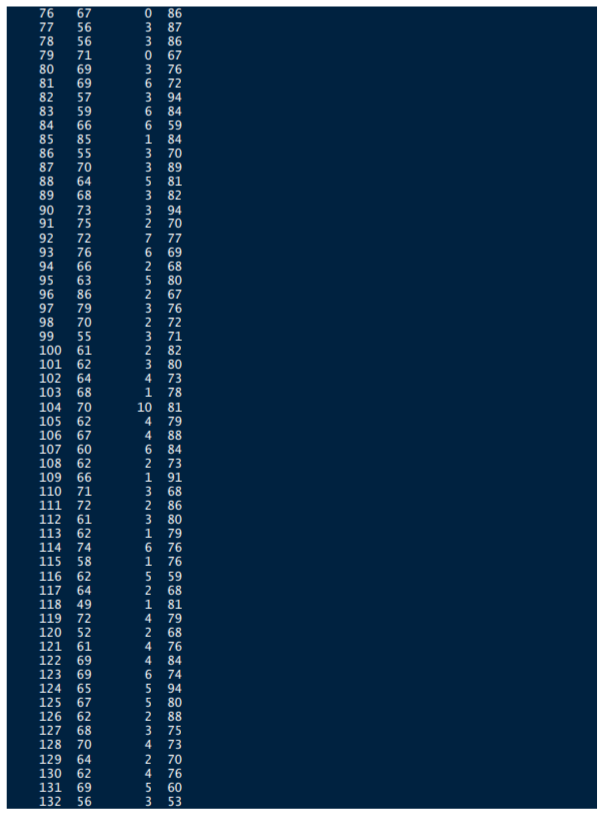
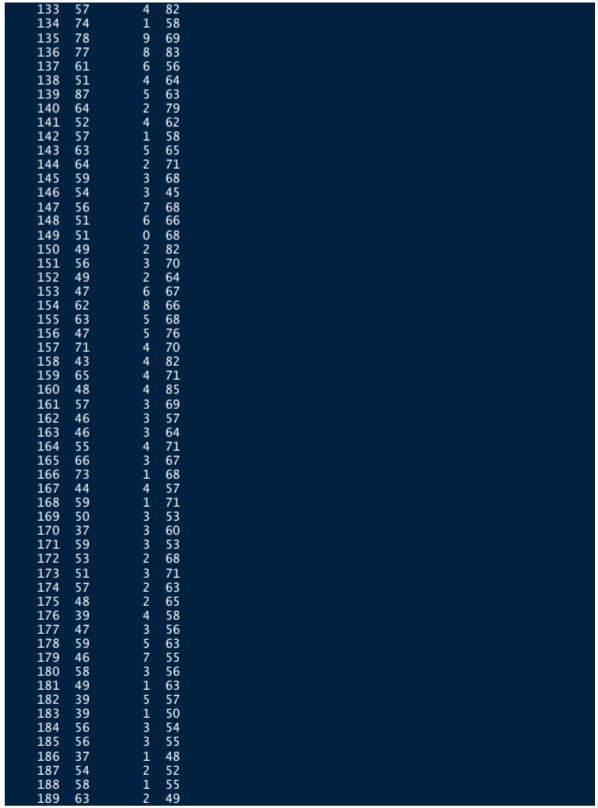
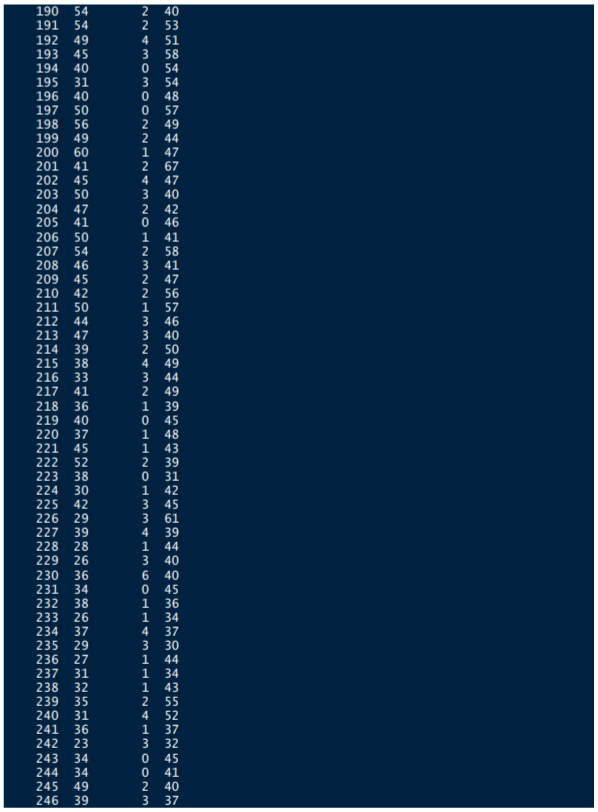
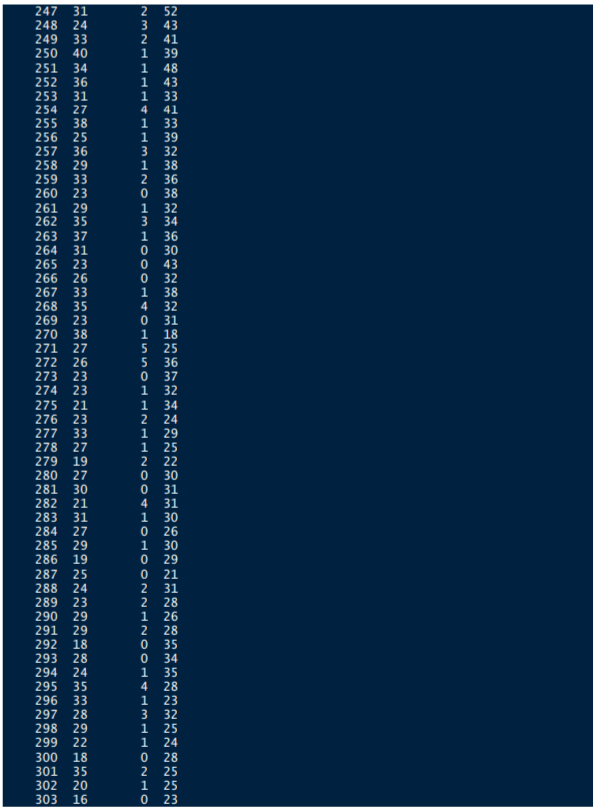
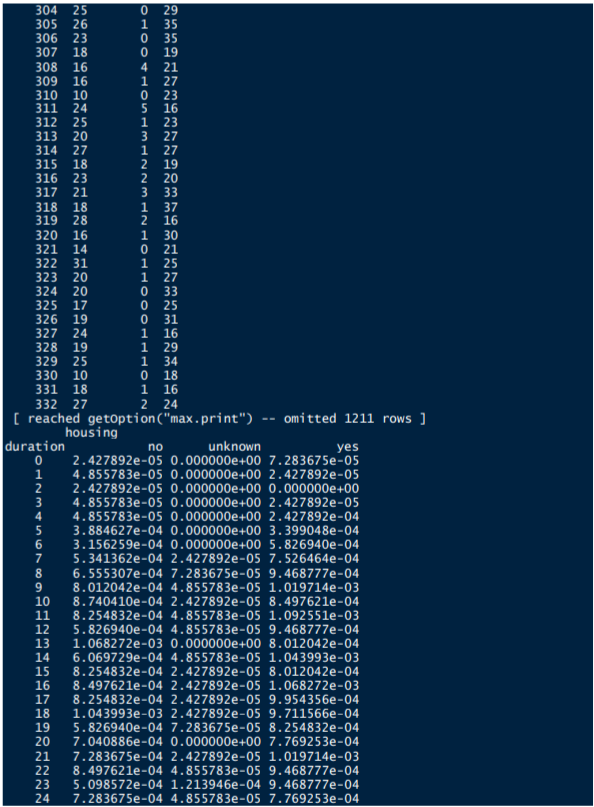
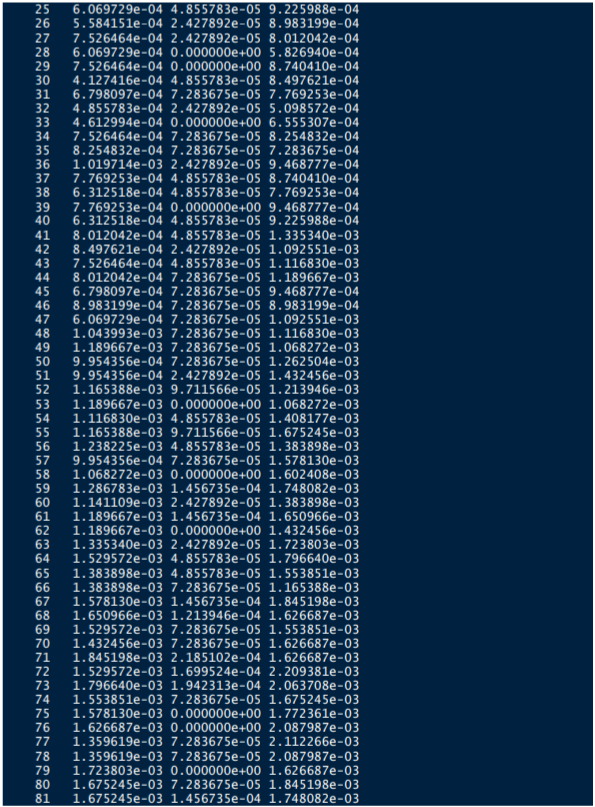
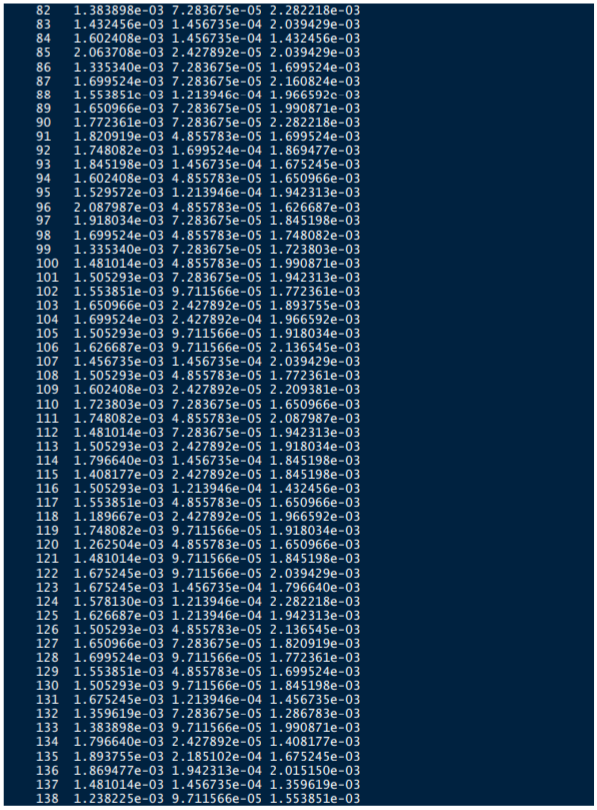
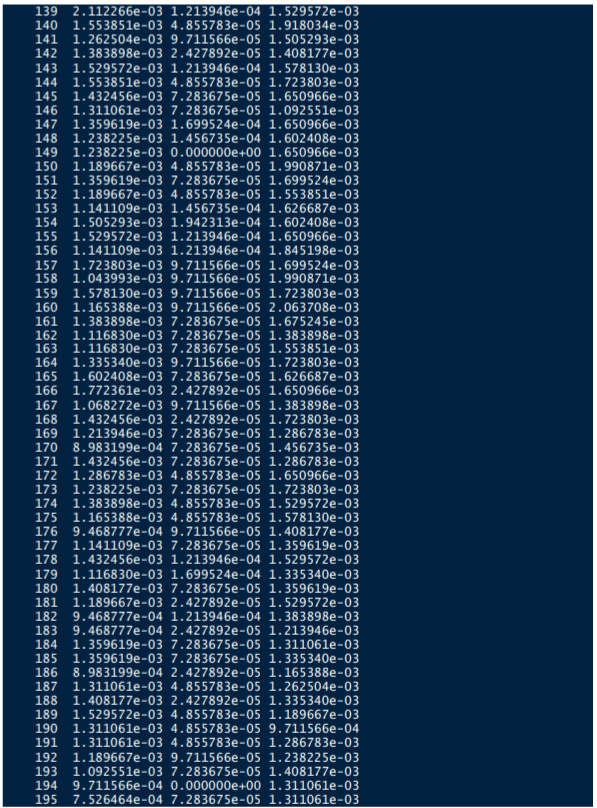
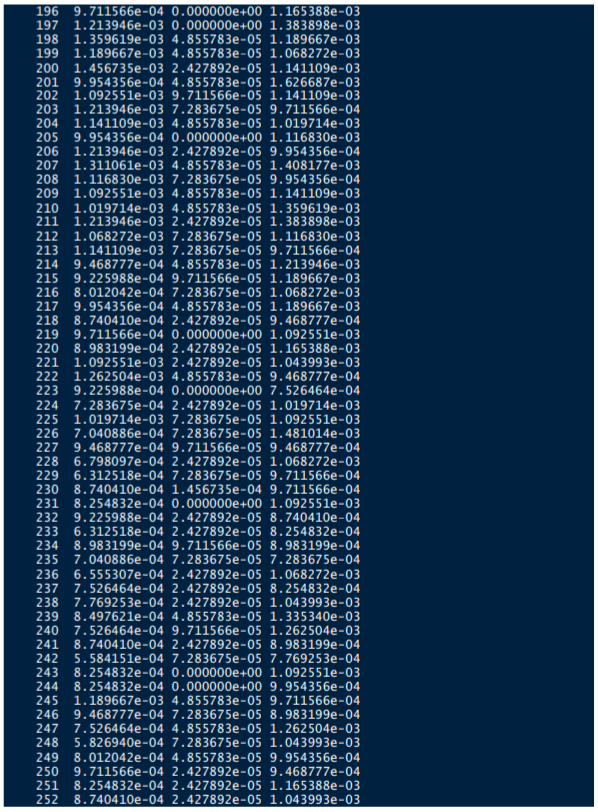
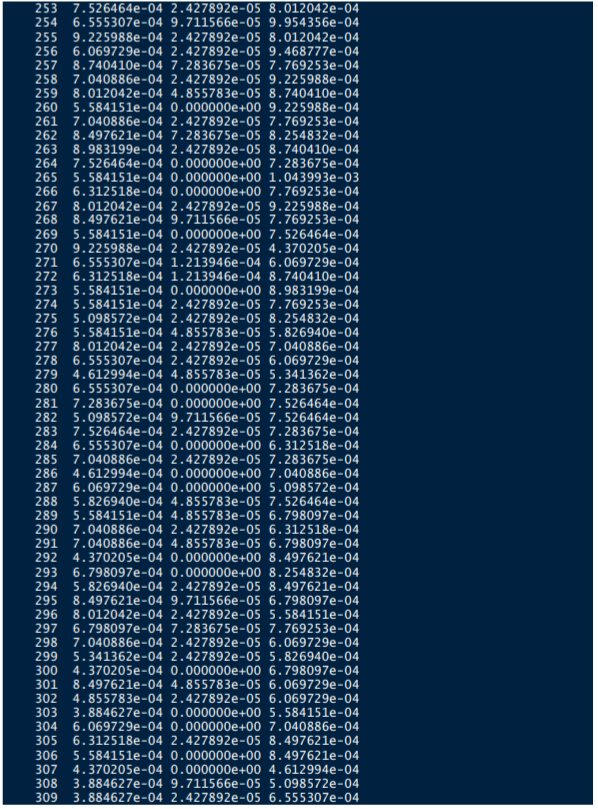
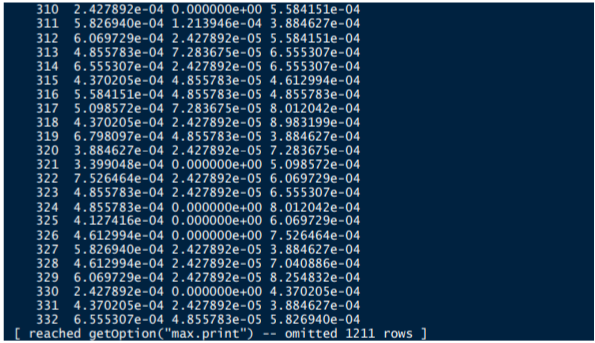
#data: duration and housing

#X-squared = 3162.3, df = 3086, p-value = 0.1657

#P value is above 0.05#





**Is there any association between consumer price index and consumer?**

#Is there any association between consumer price index and consumer?

with(bank\_additional\_full, chisq.test(cons.price.idx,cons.conf.idx))

with(bank\_additional\_full, table(cons.price.idx,cons.conf.idx))

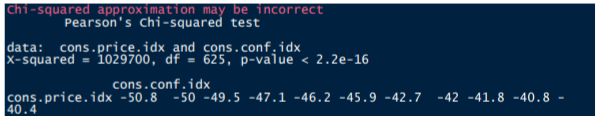
# OR

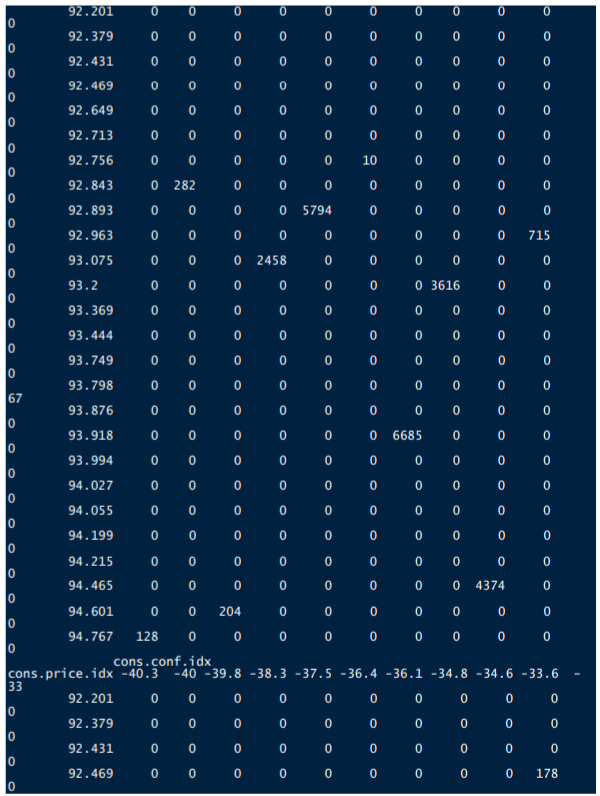
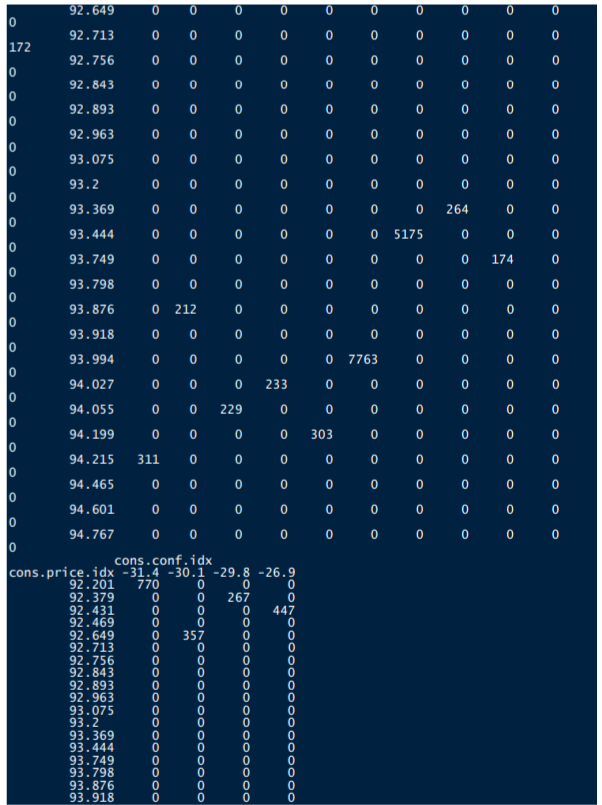
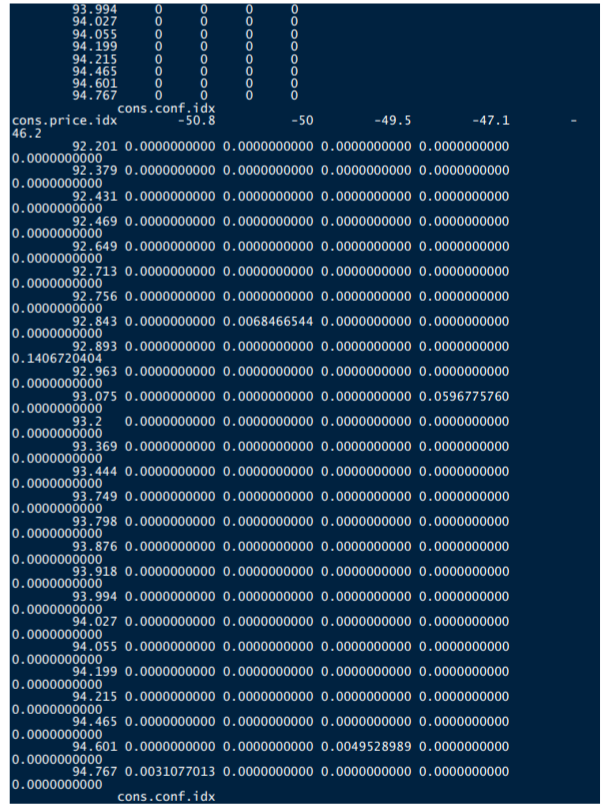
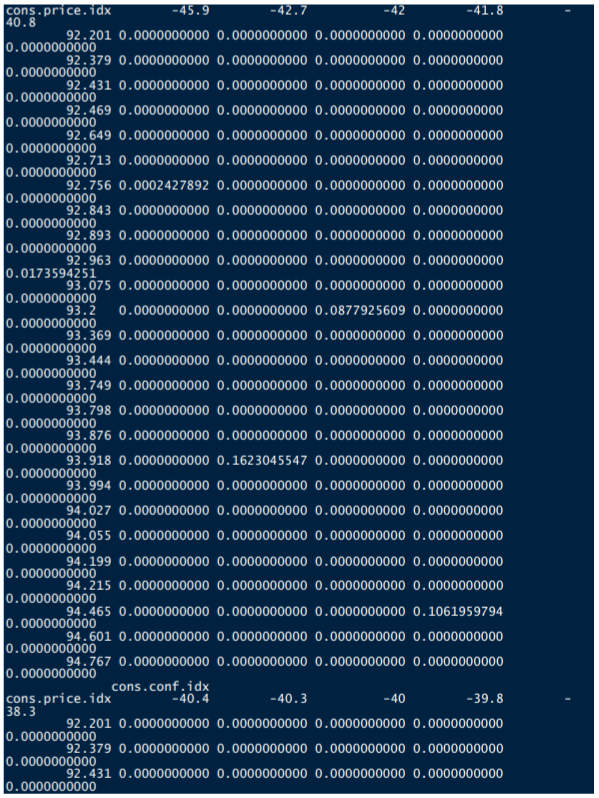
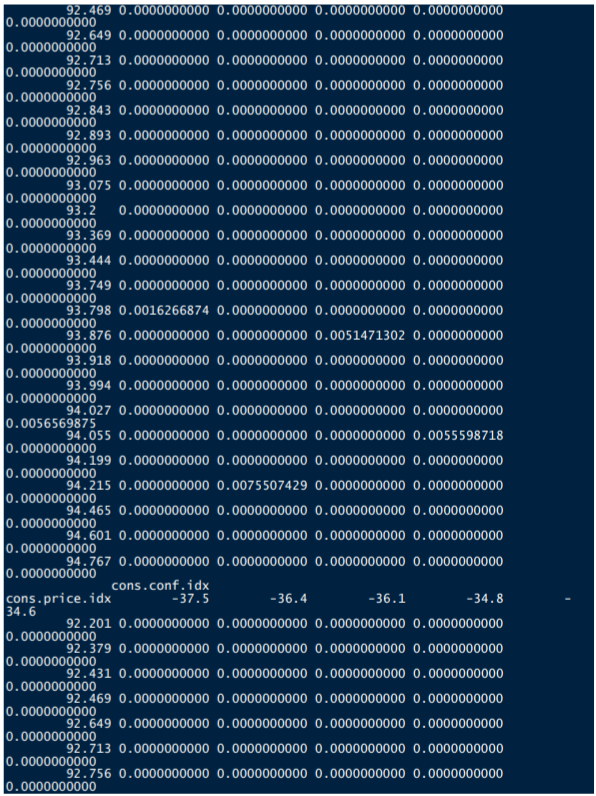
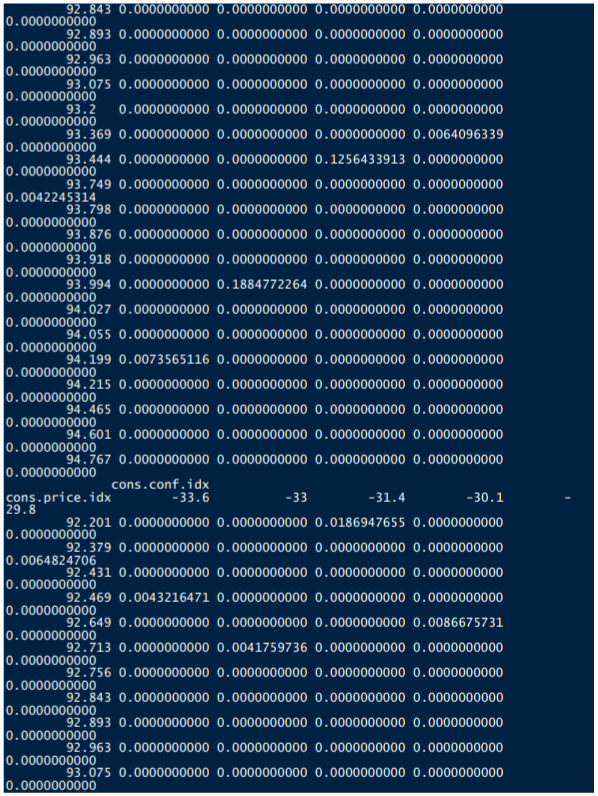
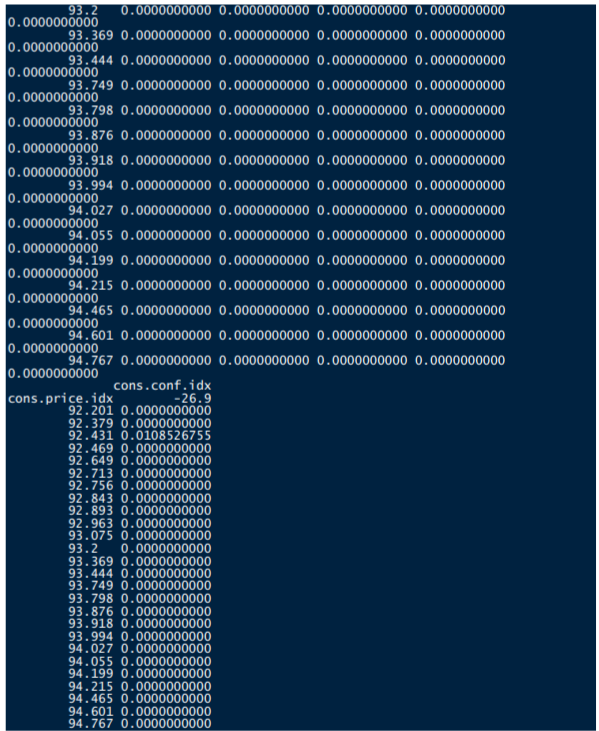
with(bank\_additional\_full, prop.table(table(cons.price.idx,cons.conf.idx)))

#p-value < 2.2e-16 and it is very much less than 0.05.we can reject the null hypothesis in favor

of alternative hypothesis and conclude, that the variables, (job & Marital-p-value < 2.2e16),(con.price.idx

, consumer- are dependent to each other.



**Is the employment variation rate consistent across job types?**

#

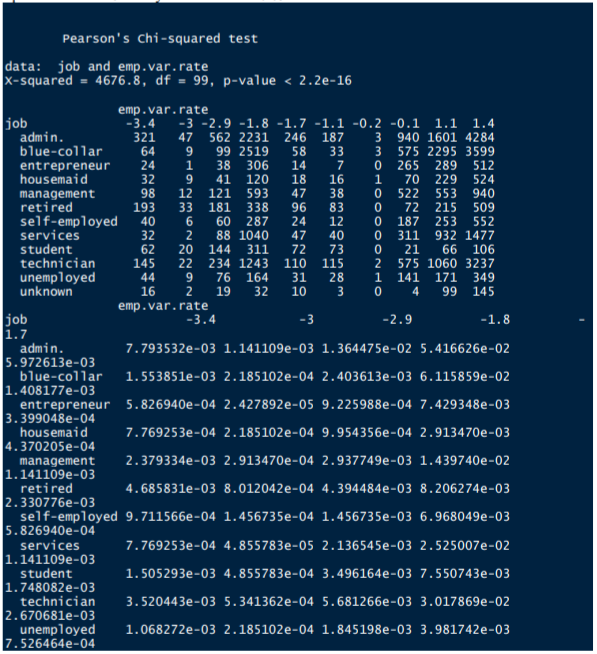
with(bank\_additional\_full, chisq.test( job,emp.var.rate))

with(bank\_additional\_full, table( job,emp.var.rate) )

# OR

with(bank\_additional\_full, prop.table(table( job,emp.var.rate)))

#p-value < 2.2e-16 is very much less than 0.05

**Is the employment variation rate same across education?**

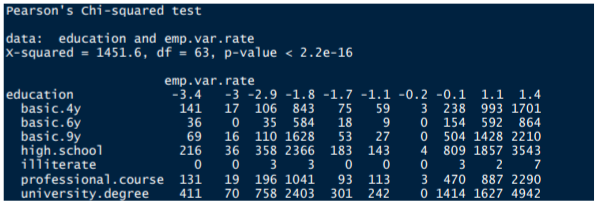
**Which group is more confident?**

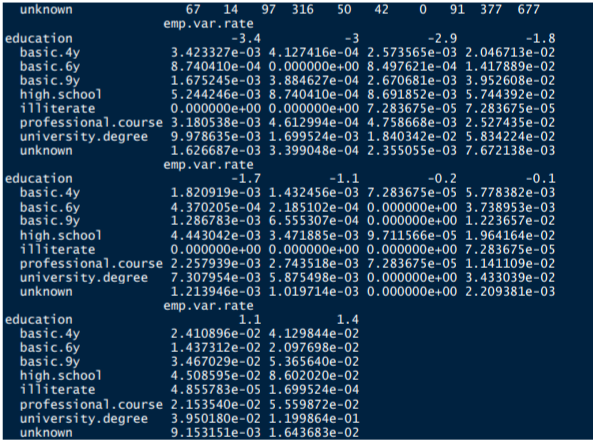
with(bank\_additional\_full, chisq.test( education,emp.var.rate))

with(bank\_additional\_full, table( education, emp.var.rate) )

# OR

with(bank\_additional\_full, prop.table(table( education,emp.var.rate)))





bank\_marketing\_data <- read\_delim("C:/Users/Seshan/Desktop/bank\_marketing\_data.csv",

";", escape\_double = FALSE, trim\_ws = TRUE)

head(bank\_marketing\_data)

# We look at difference between mean and median in summary if it's more there might be

outliers

boxplot(bank\_marketing\_data$age, main="Age Box plot",yaxt="n", xlab="Age",

horizontal=TRUE, col=terrain.colors(2))

# By plotting histogram we can ensure if there are outliers or not

**## DATA VISUALISATION**

## Use Box plots (Only for continuous variables)- To Check Ouliers

boxplot(bank\_marketing\_data$age~bank\_marketing\_data$contact, main=" AGE",ylab="age of

customers",xlab="contact")

boxplot(bank\_marketing\_data$age~bank\_marketing\_data$job, main=" AGE",ylab="age of

customers",xlab="job")

boxplot(bank\_marketing\_data$age~bank\_marketing\_data$education, main=" AGE",ylab="age

of customers",xlab="education")

boxplot(bank\_marketing\_data$age~bank\_marketing\_data$marital, main=" AGE",ylab="age of

customers",xlab="marital")

**## Barplots for Categorical Variables**

barplot(table(bank\_marketing\_data$job),col="red",main="JOB")

barplot(table(bank\_marketing\_data$marital),col="green",main="Marital")

barplot(table(bank\_marketing\_data$education),col="red",main="Education")

barplot(table(bank\_marketing\_data$emp.var.rate ),col="red",main="emp.var.rate")

hist(bank\_marketing\_data$age,col=terrain.colors(10))

#Correlation Analysis What we saw in the box plot can be emphasized by correlation plot, It can

tell if predictor is a good predictor or not a good predictor. This analysis can help us decide if we

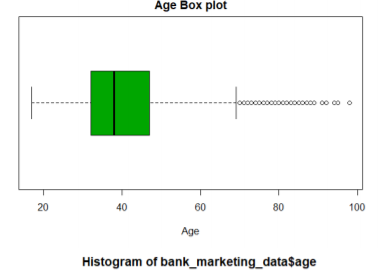
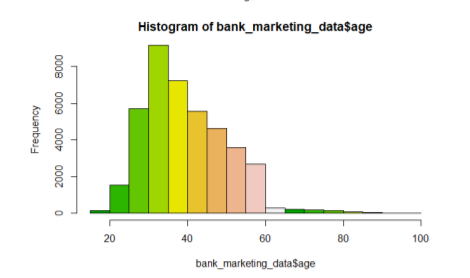
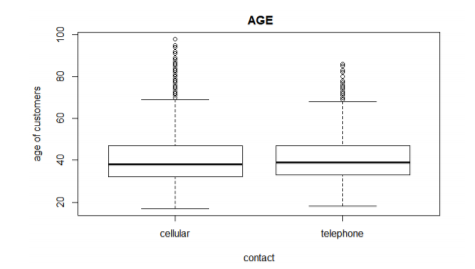
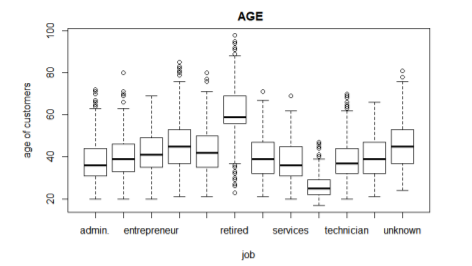
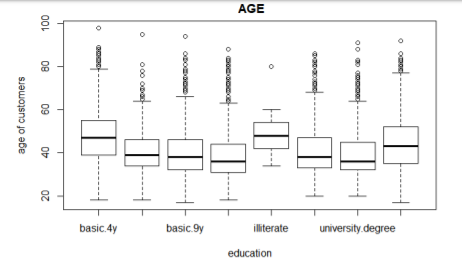
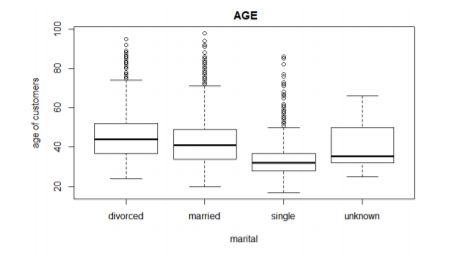
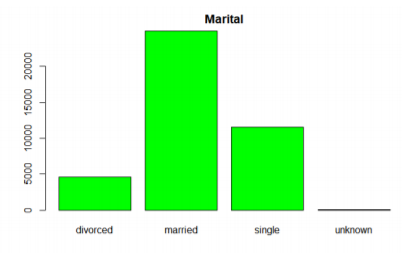
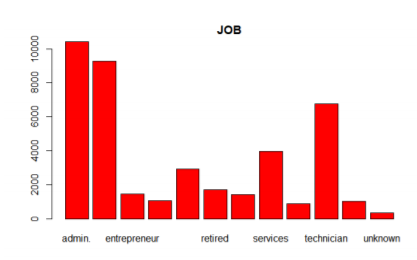
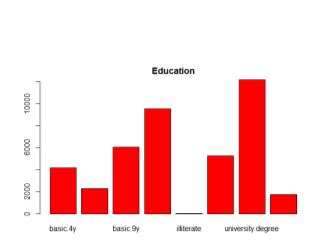
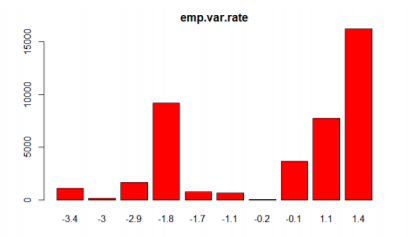
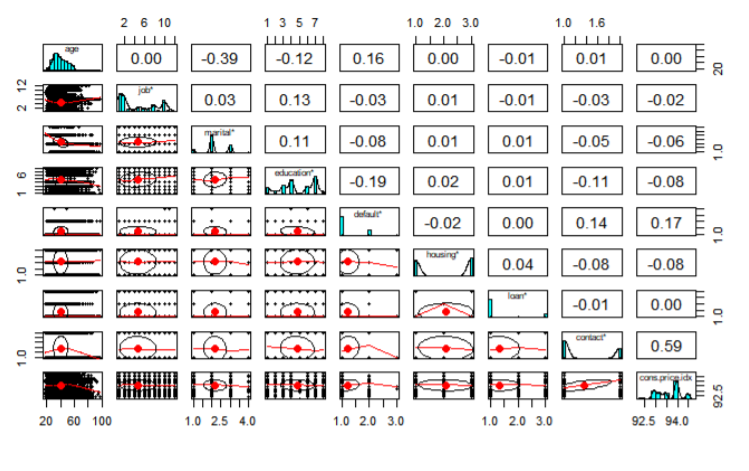
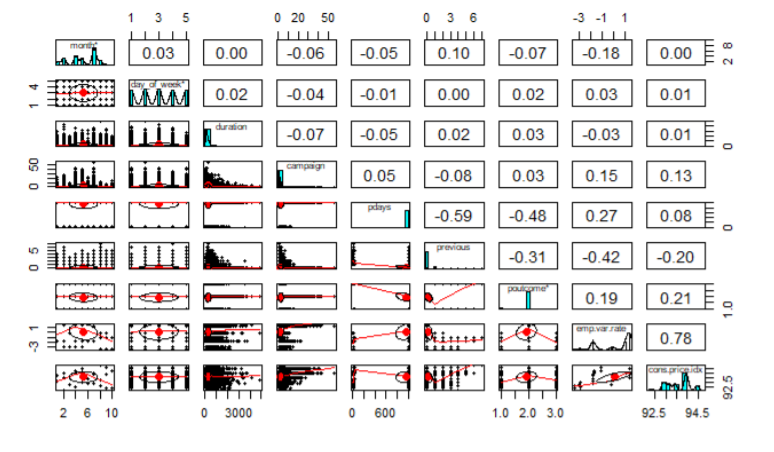
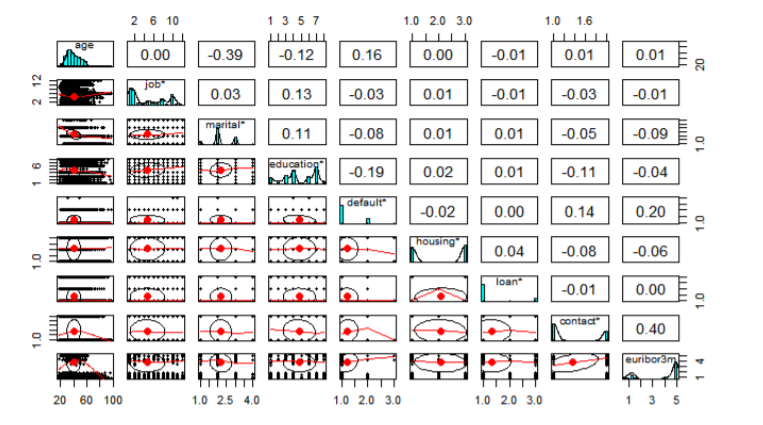
can drop some columns/predictors depending upon its correlation with the outcome variable.

library(psych)

pairs.panels(bank\_marketing\_data[, c(1:8,17)])

pairs.panels(bank\_marketing\_data[, c(9:17)])

pairs.panels(bank\_marketing\_data[, c(1:8,19)])

Subset Selection/ Feature-space reduction: Features-space can be reduced by selecting subsets

based upon correlation values obtained

#################Subset Selection################# lib

bank\_marketing\_data\_sub<-bank\_marketing\_data[, c(1:4,7:9,12,14,15,17)]

str(bank\_marketing\_data\_sub)

pairs.panels(bank\_marketing\_data\_sub)

#3.4. Data transformation and Binning We do data transformation and binning for better

modeling. We convert categorical variable into numerical using binning.

#################Binning and Data Transformation#################

#bank\_marketing\_data\_sub$age <- cut(bank\_marketing\_data\_sub$age, c(1,20,40,60,100))

#bank\_marketing\_data\_sub$is\_divorced <- ifelse( bank\_marketing\_data\_sub$marital ==

"divorced", 1, 0)

bank\_marketing\_data\_sub$is\_nr.employed <- ifelse( bank\_marketing\_data\_sub$education ==

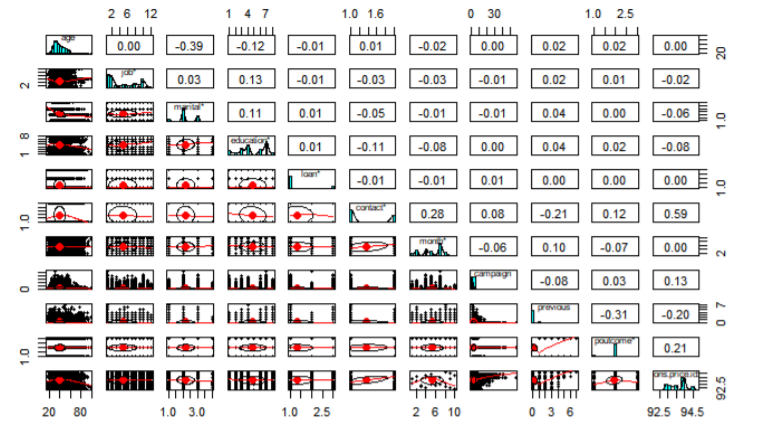
"employed", 1, 0)

#bank\_marketing\_data\_sub$is\_single <- ifelse( bank\_marketing\_data\_sub$marital == "single",

1, 0)

bank\_marketing\_data\_sub$nr.employed <- NULL

str(bank\_marketing\_data\_sub)



#scatter.smooth(x=bank\_marketing\_data$job, y=bank\_marketing\_data$emp.var.rate,

main="emp.var.rate ~ job") # scatterplot

# load library

library(corrplot)

# load the data

data<-bank\_marketing\_data

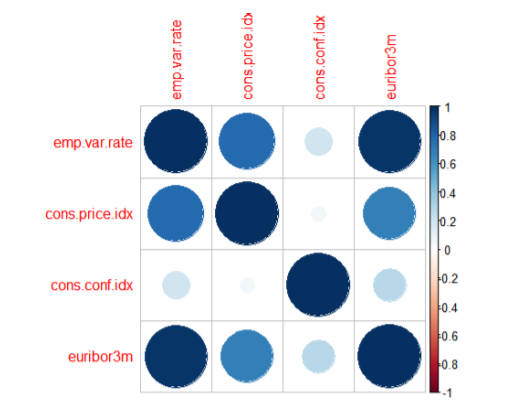
data(bank\_marketing\_data\_sub)

# calculate correlations

correlations <- cor(bank\_marketing\_data[,16:19])

**# create correlation plot**

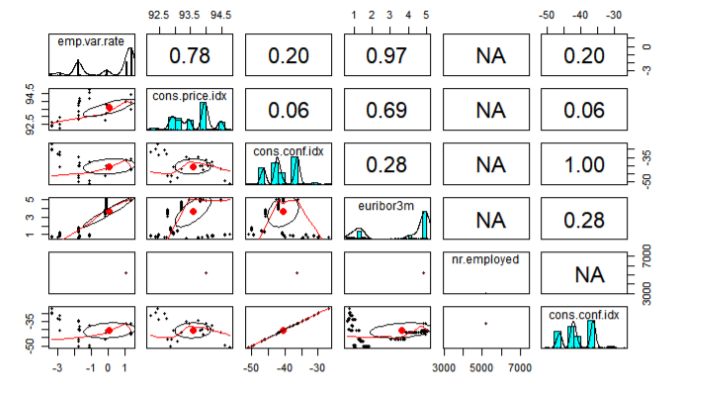
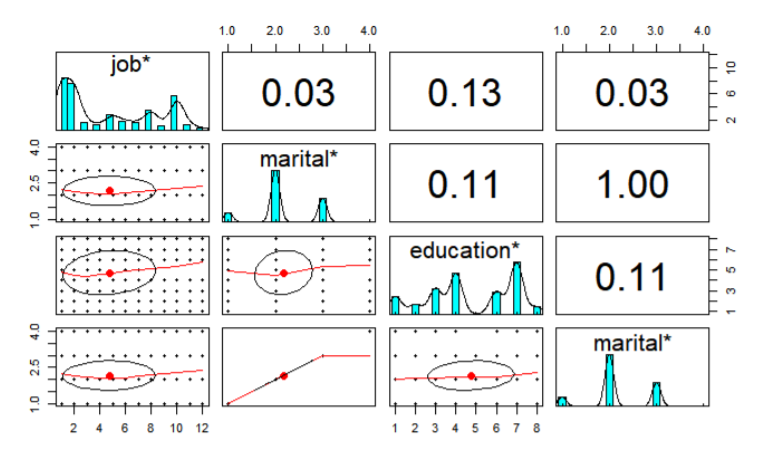
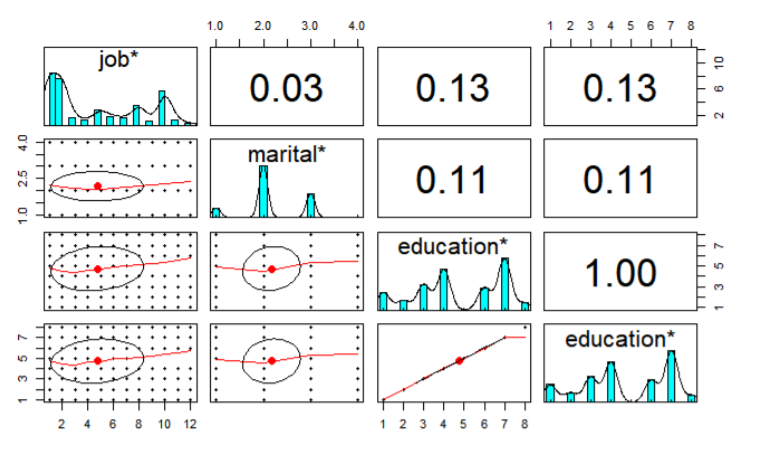
corrplot(correlations, method="circle")



airs.panels(bank\_marketing\_data[, c(16:20,18)])

pairs.panels(bank\_marketing\_data[, c(2:4,3)])

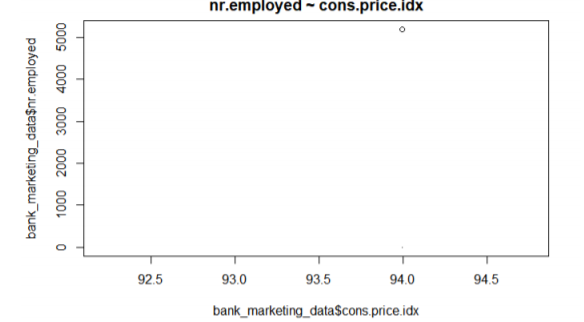
pairs.panels(bank\_marketing\_data[, c(2:4,4)])

head(bank\_marketing\_data)

scatter.smooth(x=bank\_marketing\_data$cons.price.idx, y=bank\_marketing\_data$nr.employed,

main="nr.employed ~ cons.price.idx")



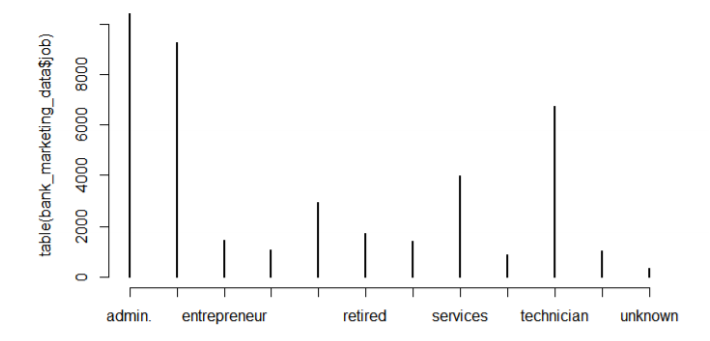
#cor(bank\_marketing\_data$age, bank\_marketing\_data$emp.var.rate)

head(bank\_marketing\_data)

table(bank\_marketing\_data$job)

table(bank\_marketing\_data$marital)

plot(table(bank\_marketing\_data$job))



library(psych)

pairs.panels(bank\_marketing\_data[,1:6])

