Census Project

1. Load requied libraries. Install package install.packages("caret") Install package install.packages("corrplot") Install package install.packages('Boruta')

```
library(ggplot2)
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(Boruta)
```

```
## Loading required package: ranger
```

```
library(caret)
```

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

2. Load census data.

```
setwd("c:/Ryerson University/Semester 4/ProjectCode")
loc<-getwd()
censusdata <- read.csv(file="census.csv",header=TRUE,sep=",", na.string = "?")
```

2.1. Divide the data into train and test data.

```
inTrain <- createDataPartition(y=censusdata$income, p= 0.75, list=FALSE)
training <- censusdata[inTrain,]
testing <- censusdata[-inTrain,]
```

3. Display dimensions, summary of data, names and structure of data.

```
data <- training
dim(data)
```

```
## [1] 24421    15
```

```
nrow(data)
```

```
## [1] 24421
```

```
ncol(data)
```

```
## [1] 15
```

```
dim(testing)
```

```
## [1] 8140   15
```

```
summary(data)
```

```
##       age                    workclass         fnlwgt
##  Min.   :17.00   Private         :17021   Min.   :  12285
##  1st Qu.:28.00   Self-emp-not-inc: 1895   1st Qu.: 117789
##  Median :37.00   Local-gov       : 1571   Median : 178147
##  Mean   :38.58   State-gov       :  958   Mean   : 189651
##  3rd Qu.:48.00   Self-emp-inc    :  865   3rd Qu.: 236873
##  Max.   :90.00   (Other)         :  708   Max.   :1484705
##                  NA's            : 1403
##         education    education.num                  marital.status
##  HS-grad     :7902   Min.   : 1.00   Divorced             : 3339
```

```
##  Some-college:5479   1st Qu.: 9.00   Married-AF-spouse    :   15
##  Bachelors   :3964   Median :10.00   Married-civ-spouse   :11268
##  Masters     :1282   Mean   :10.07   Married-spouse-absent:  310
##  Assoc-voc   :1048   3rd Qu.:12.00   Never-married        : 7986
##  11th        : 876   Max.   :16.00   Separated            :  759
##  (Other)     :3870                   Widowed              :  744
##            occupation           relationship                   race
##  Craft-repair   :3102   Husband       :9925   Amer-Indian-Eskimo:  238
##  Prof-specialty :3093   Not-in-family :6148   Asian-Pac-Islander:  775
##  Exec-managerial:3035   Other-relative: 727   Black              : 2357
##  Adm-clerical   :2850   Own-child     :3861   Other              :  206
##  Sales          :2727   Unmarried     :2588   White              :20845
##  (Other)        :8206   Wife          :1172
##  NA's           :1408
##      sex          capital.gain    capital.loss     hours.per.week
##  Female: 8059   Min.   :    0   Min.   :   0.00   Min.   : 1.00
##  Male  :16362   1st Qu.:    0   1st Qu.:   0.00   1st Qu.:40.00
##                 Median :    0   Median :   0.00   Median :40.00
##                 Mean   : 1086   Mean   :  89.63   Mean   :40.41
##                 3rd Qu.:    0   3rd Qu.:   0.00   3rd Qu.:45.00
##                 Max.   :99999   Max.   :4356.00   Max.   :99.00
##
##        native.country    income
##  United-States:21902   <=50K:18540
##  Mexico       :  493   >50K : 5881
##  Philippines  :  149
##  Germany      :   98
##  Canada       :   94
##  (Other)      : 1245
##  NA's         :  440
```

```
names(data)
```

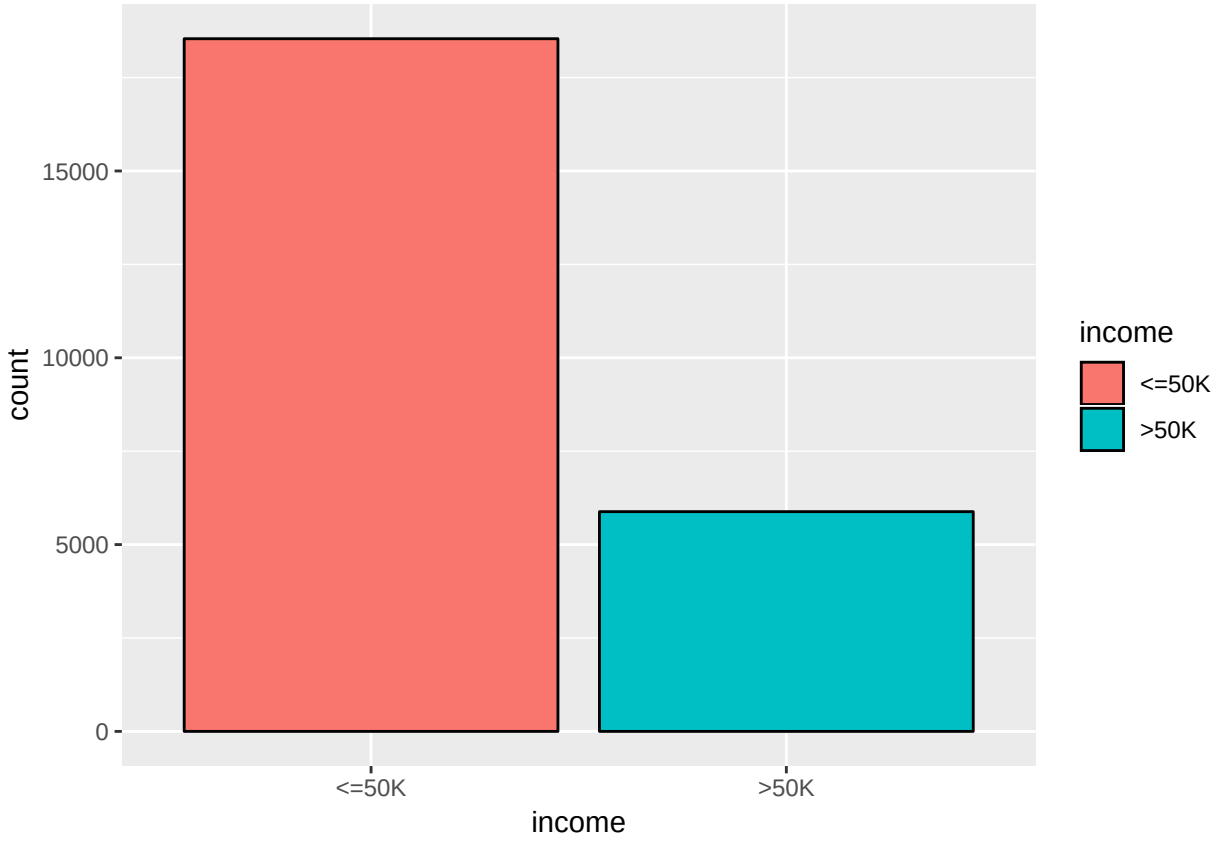
```
##  [1] "age"            "workclass"      "fnlwgt"         "education"
##  [5] "education.num"  "marital.status" "occupation"     "relationship"
##  [9] "race"           "sex"            "capital.gain"   "capital.loss"
## [13] "hours.per.week" "native.country" "income"
```

```
str(data)
```

```
## 'data.frame':    24421 obs. of  15 variables:
##  $ age           : int  66 54 41 34 74 68 41 45 38 52 ...
##  $ workclass     : Factor w/ 8 levels "Federal-gov",..: NA 4 4 4 7 1 4 4 6 4 ...
##  $ fnlwgt        : int  186061 140359 264663 216864 88638 422013 70037 172274 164526 129177 ...
##  $ education     : Factor w/ 16 levels "10th","11th",..: 16 6 16 12 11 12 16 11 15 10 ...
##  $ education.num : int  10 4 10 9 16 9 10 16 15 13 ...
##  $ marital.status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 7 1 6 1 5 1 5 1 5 7 ...
##  $ occupation    : Factor w/ 14 levels "Adm-clerical",..: NA 7 10 8 10 10 3 10 10 8 ...
##  $ relationship  : Factor w/ 6 levels "Husband","Not-in-family",..: 5 5 4 5 3 2 5 5 2 2 ...
##  $ race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 3 5 5 5 5 5 5 3 5 5 ...
##  $ sex           : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 2 1 2 1 ...
##  $ capital.gain  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ capital.loss  : int  4356 3900 3900 3770 3683 3683 3004 3004 2824 2824 ...
##  $ hours.per.week: int  40 40 40 45 20 40 60 35 45 20 ...
##  $ native.country: Factor w/ 41 levels "Cambodia","Canada",..: 39 39 39 39 39 39 NA 39 39 39 ...
##  $ income        : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 2 1 2 2 2 2 ...
```

4. Display Class Distributions.

```
# Imbalance data
result = summary(data$income)/nrow(data) * 100
ggplot(data=data,aes(income)) + geom_bar(aes(fill = income), color = "black")
```




```
result
```

```
##    <=50K     >50K
## 75.91827 24.08173
```

5. Check and Cleaning missing values.

```
cat("Number of missing values in training set is:", sum(is.na(data)), "\n")
```

```
## Number of missing values in training set is: 3251
```

```
na_count <-sapply(data, function(y) sum(length(which(is.na(y)))))
na_count <- data.frame(na_count)
na_count
```

```
##                 na_count
## age                    0
## workclass           1403
## fnlwgt                 0
## education              0
## education.num          0
## marital.status         0
## occupation          1408
## relationship           0
```

```
## race                 0
## sex                  0
## capital.gain         0
## capital.loss         0
## hours.per.week       0
## native.country     440
## income               0
```

```
nrow(data)
```

```
## [1] 24421
```

```
data <- na.omit(data)
nrow(data)
```

```
## [1] 22597
```

```
nrow(testing)
```

```
## [1] 8140
```

```
cat("Number of missing values in test set is:", sum(is.na(testing)), "\n")
```

```
## Number of missing values in test set is: 1011
```

```
na_count1 <-sapply(testing, function(y) sum(length(which(is.na(y)))))
na_count1
```

```
##            age      workclass         fnlwgt      education  education.num
##              0            433              0              0              0
## marital.status     occupation   relationship           race            sex
##              0            435              0              0              0
##   capital.gain   capital.loss hours.per.week native.country         income
##              0              0              0            143              0
```

```
testingdata <- na.omit(testing)
nrow(testingdata)
```

```
## [1] 7565
```

5.1 Re-factoring the work class, occupation and native country after removing the NA values (exclude levels not required).

```
data$workclass <- factor(data$workclass)
data$occupation <- factor(data$occupation)
data$native.country <- factor(data$native.country)
```

5.1 Re-factoring the work class, occupation and native country after removing the NA values (exclude levels not required) for testing data also.

```
testingdata$workclass <- factor(testingdata$workclass)
testingdata$occupation <- factor(testingdata$occupation)
testingdata$native.country <- factor(testingdata$native.country)
```

6. Statistics of Numerical attributes

```
# statistics of numerical attributes
summary(data$age)
```

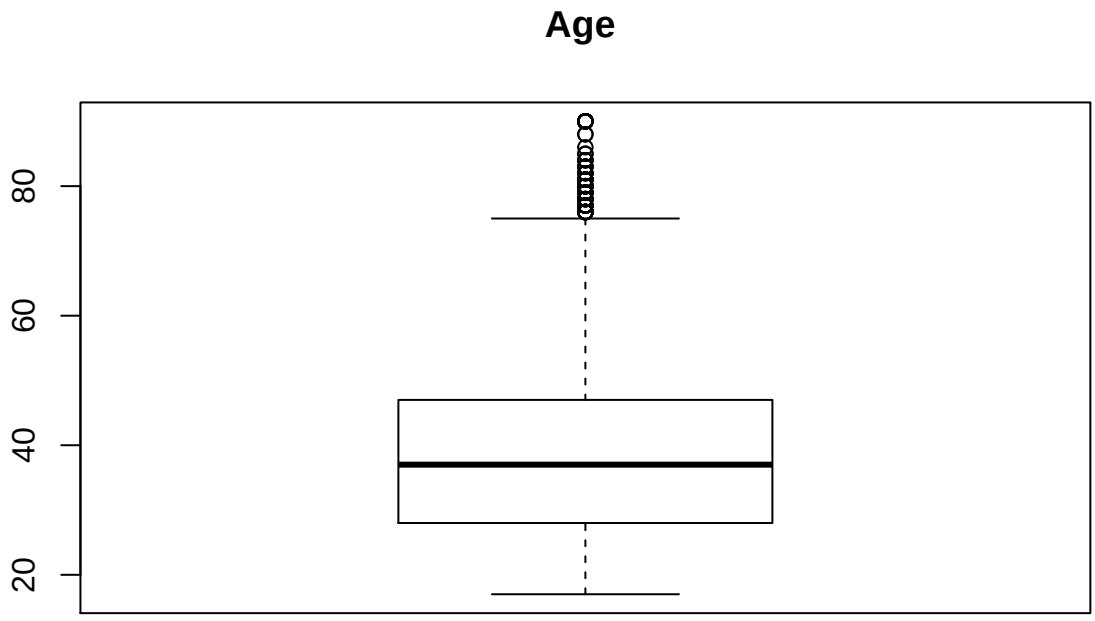
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   28.00   37.00   38.44   47.00   90.00
```

```r
sd(data$age)
```

```
## [1] 13.11207
```

```r
hist(data$age, main = "Distribution of Age",xlab = "Age of an Individual" ,col ="blue")
```


```r
boxplot(data$age,main="Age ")
```



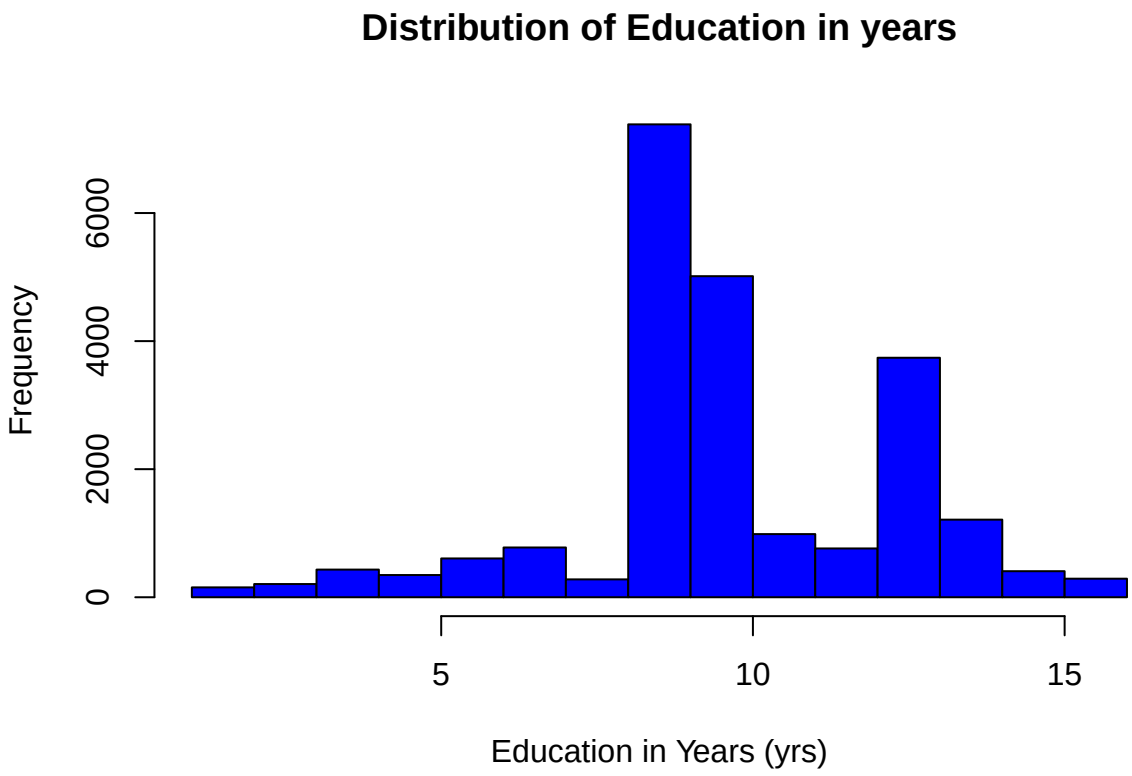

```
summary(data$education.num)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    9.00   10.00   10.12   12.00   16.00
```

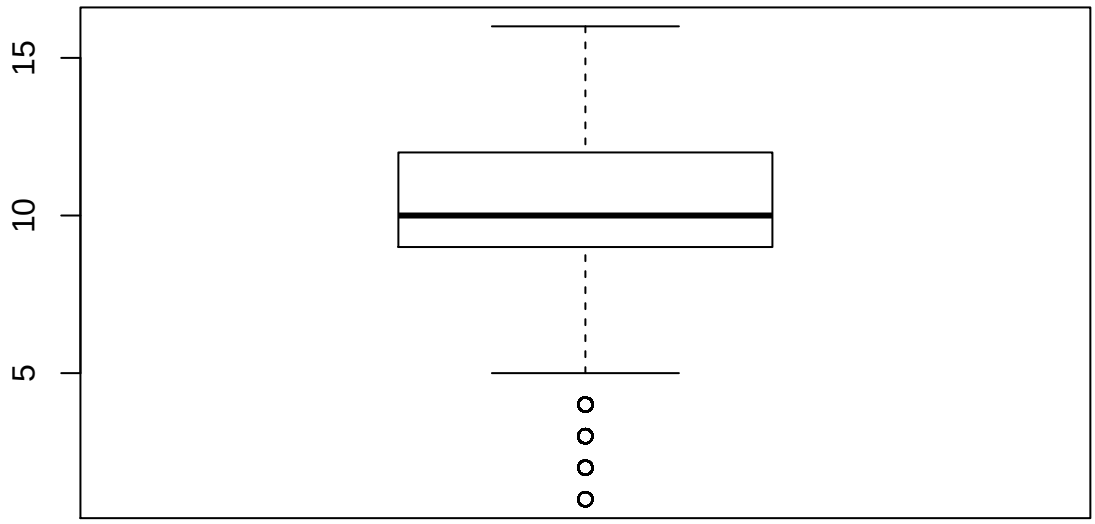
```
sd(data$education.num)
```

```
## [1] 2.553421
```

```
hist(data$education.num,main = "Distribution of Education in years",xlab="Education in Years (yrs)",col
```




```
boxplot(data$education.num,main="Distribution of Education")
```



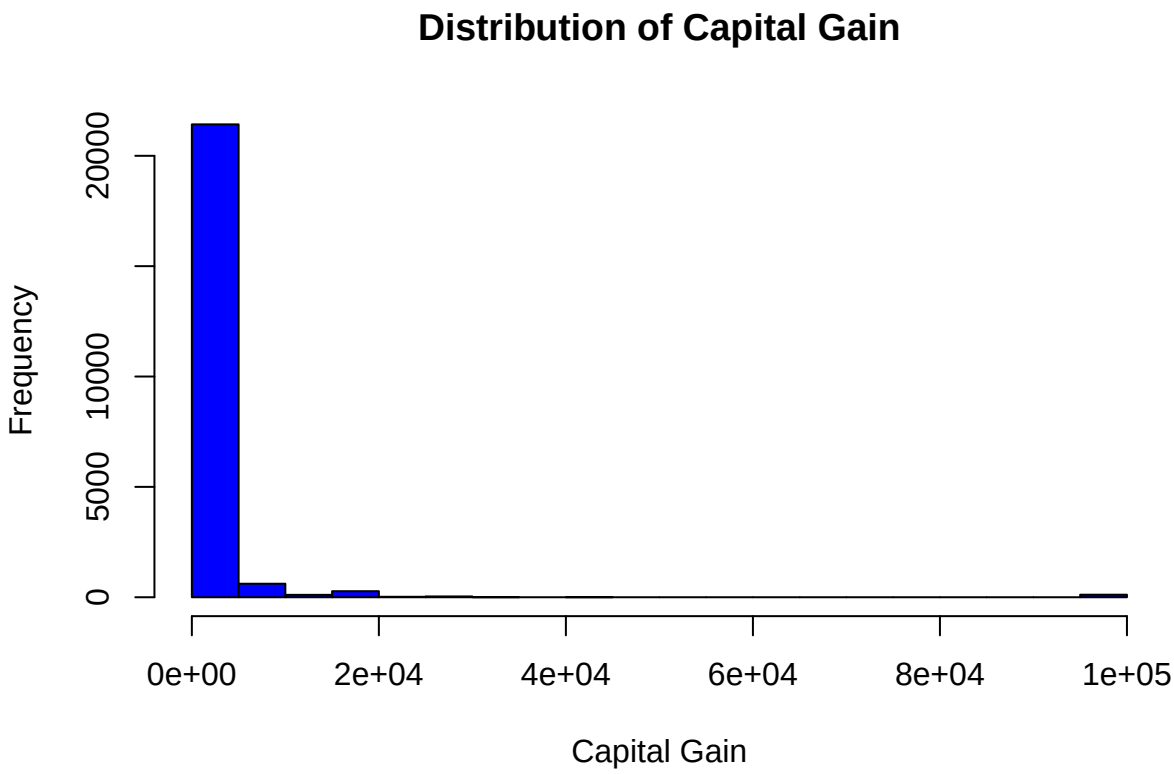

```
summary(data$capital.gain)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0    1110       0   99999
```

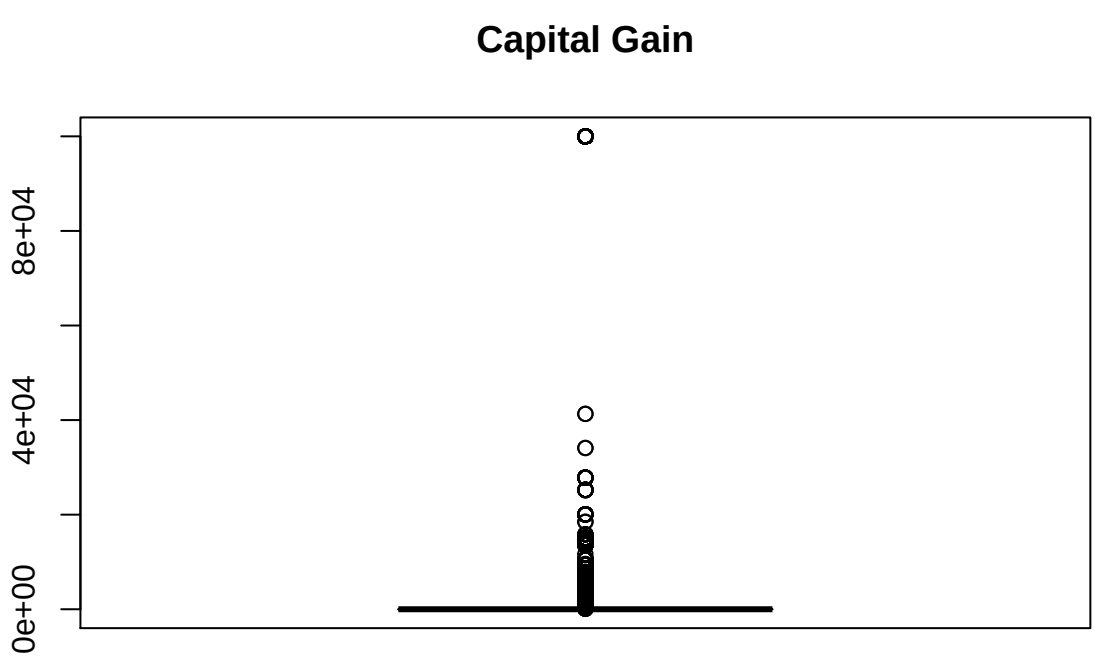
```
sd(data$capital.gain)
```

```
## [1] 7530.779
```

```
hist(data$capital.gain,main = "Distribution of Capital Gain",xlab="Capital Gain",col = "blue")
```




```
boxplot(data$capital.gain,main="Capital Gain")
```



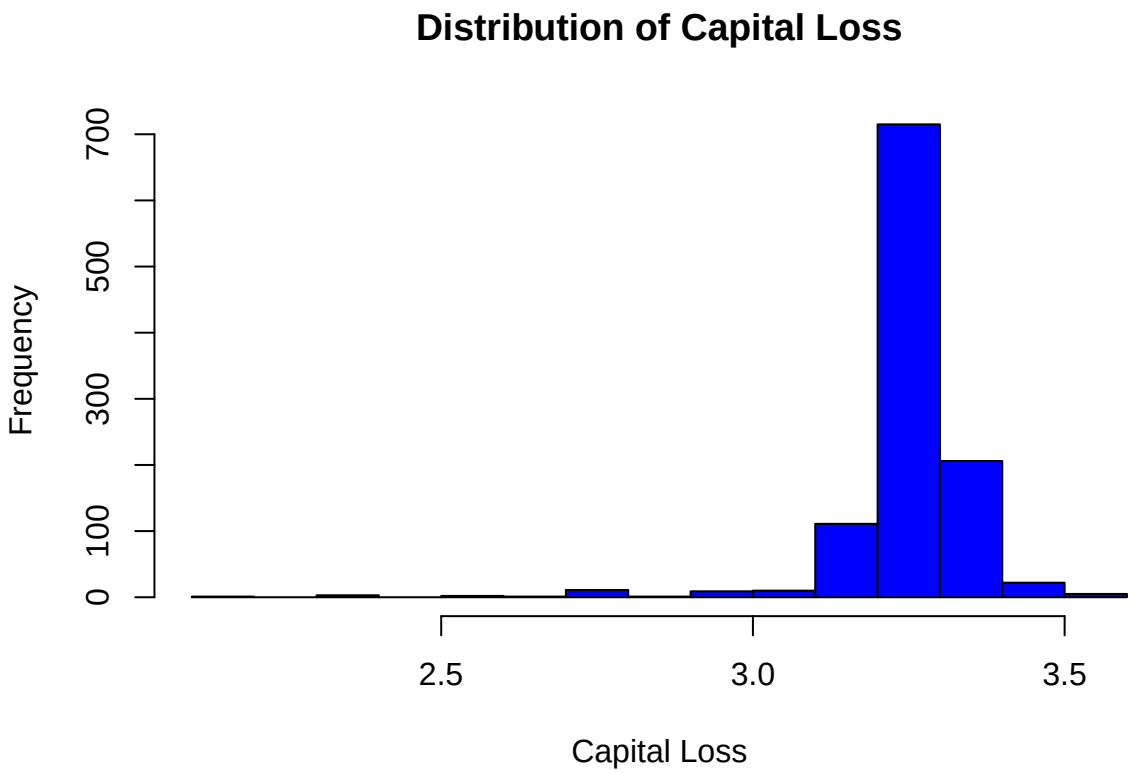

```
summary(data$capital.loss)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00   90.55    0.00 3900.00
```

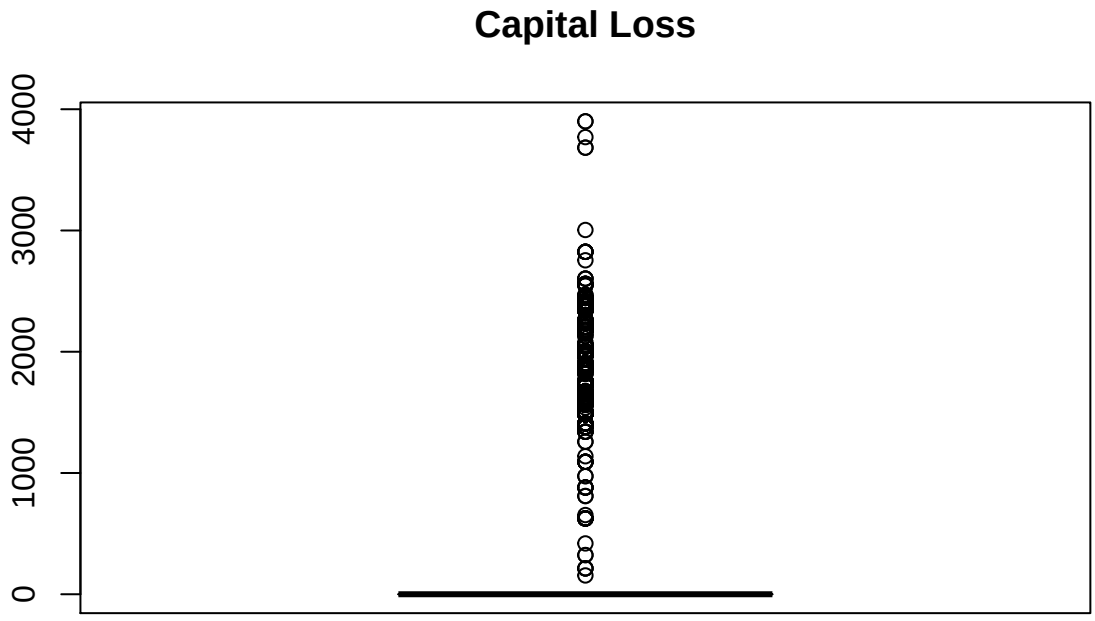
```
sd(data$capital.loss)
```

```
## [1] 408.6882
```

```
hist(log10(data$capital.loss),main = "Distribution of Capital Loss",xlab="Capital Loss",col = "blue")
```




```
boxplot(data$capital.loss,main="Capital Loss")
```



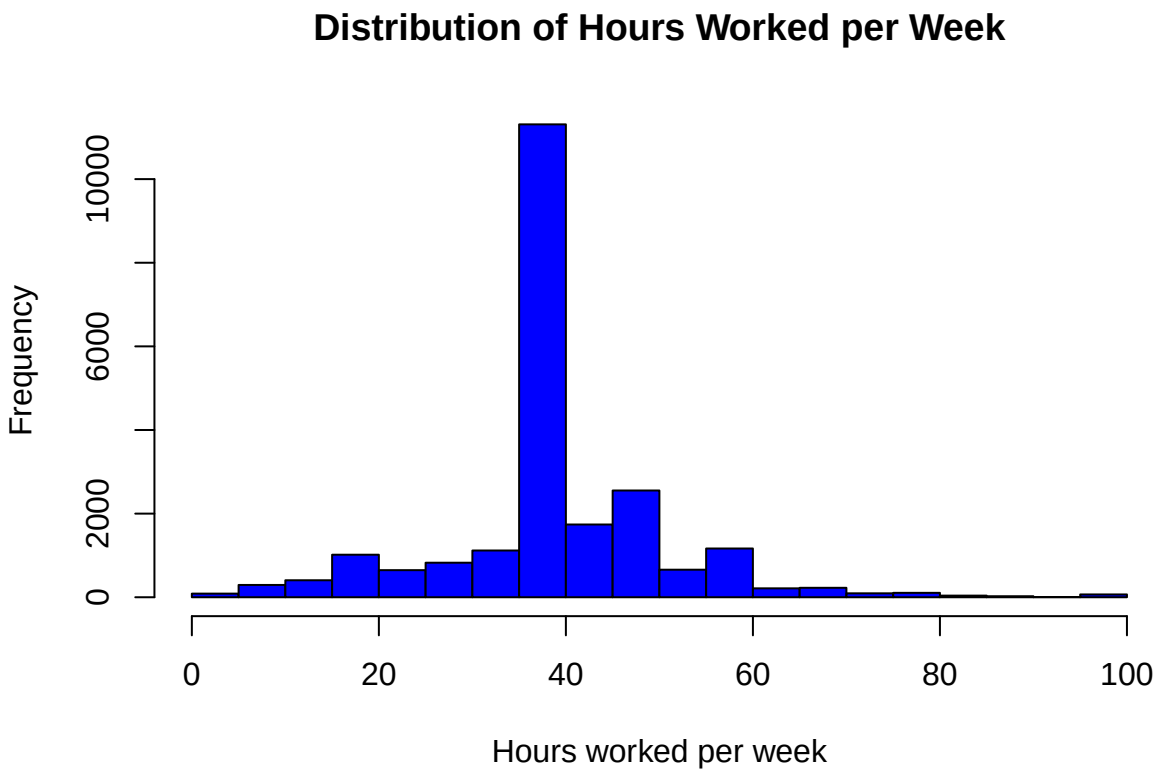

```
summary(data$hours.per.week)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   40.00   40.00   40.91   45.00   99.00
```

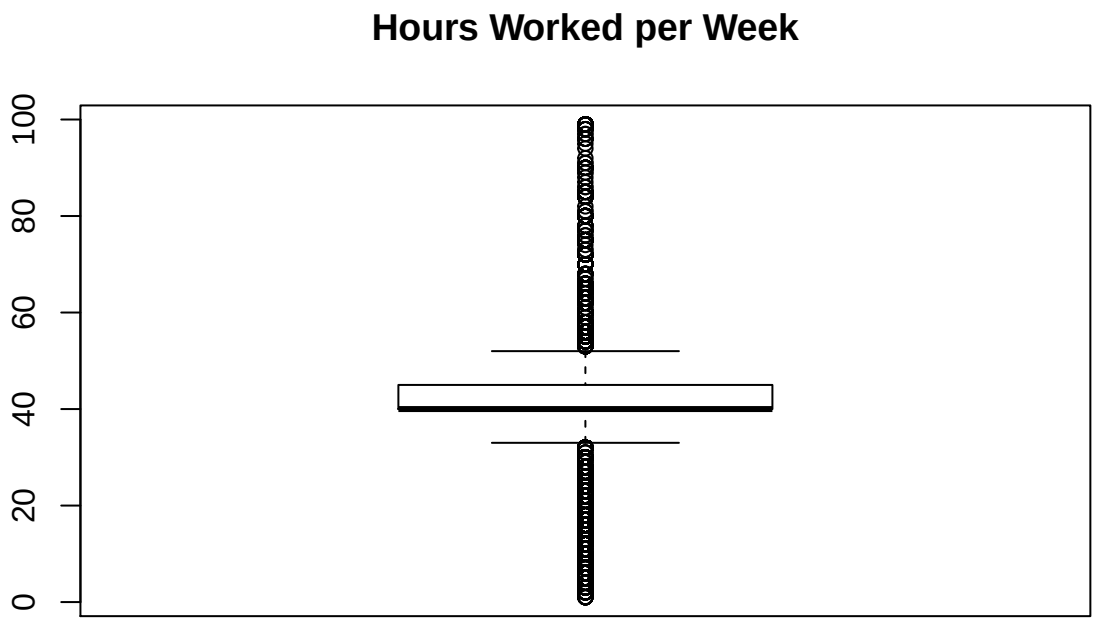
```
sd(data$hours.per.week)
```

```
## [1] 11.94349
```

```
hist(data$hours.per.week,main = "Distribution of Hours Worked per Week",xlab="Hours worked per week",co
```



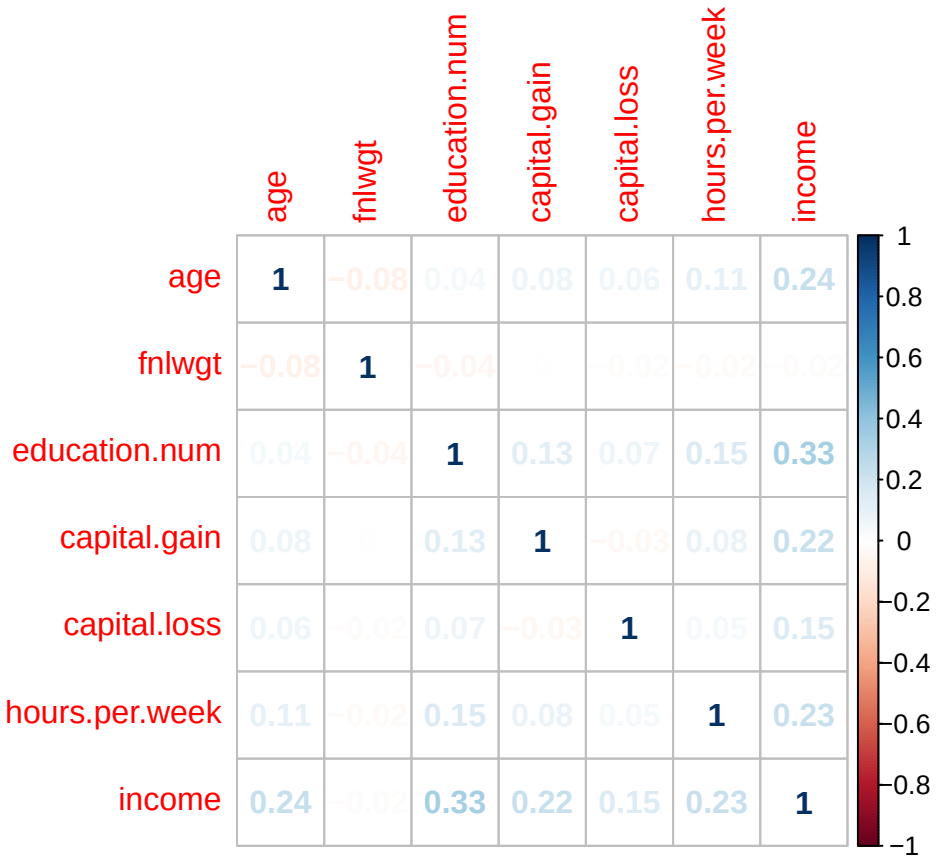

```r
boxplot(data$hours.per.week,main="Hours Worked per Week")
```



7. Correlation between numerical attributes.

```
#Changing income to 0 <= 50k, 1 > 50k

data1 <- data
data1$income <- as.numeric(data1$income)-1
#Correlation plot
M <- c(1, 3, 5, 11:13, 15)
corrplot(cor(data1[,M]),method = "number")
```




```
############################################################
# Correlations shows that numeric attributes are related #
# but are not strongly correlated. The variables are positively
# correlated. Education has the highest correlation 0.33 with income.
# Capital gain 0.22, age 0.24 and hours worked 0.23. The variables are positively correlted
# with each other.
############################################################
```

8. Exploratory analysis of the attribute native country.

```
summary(data$native.country)
```

```
##                    Cambodia                      Canada
##                          15                          83
##                       China                    Columbia
##                          47                          39
##                        Cuba              Dominican-Republic
##                          63                          50
##                     Ecuador                 El-Salvador
##                          22                          62
##                     England                      France
##                          66                          19
##                     Germany                      Greece
##                          91                          18
##                   Guatemala                       Haiti
##                          50                          31
##           Holand-Netherlands                    Honduras
```

```
##                          1                         10
##                       Hong                    Hungary
##                         13                         10
##                      India                       Iran
##                         72                         33
##                    Ireland                      Italy
##                         17                         44
##                    Jamaica                      Japan
##                         57                         42
##                       Laos                     Mexico
##                         15                        468
##                  Nicaragua Outlying-US(Guam-USVI-etc)
##                         21                          9
##                       Peru                Philippines
##                         23                        142
##                     Poland                   Portugal
##                         45                         24
##                 Puerto-Rico                   Scotland
##                         81                          7
##                      South                     Taiwan
##                         57                         32
##                   Thailand             Trinadad&Tobago
##                         13                         13
##               United-States                    Vietnam
##                      20631                         48
##                  Yugoslavia
##                         13
```

9.1 Reducing/Combining levels of native country in training data.

```
data$native.country <- as.character(data$native.country)
asia <- c("Cambodia", "China", "Hong", "India", "Iran", "Japan", "Laos","Philippines", "Taiwan", "Thaila
northAmerica <- c("Canada", "Cuba", "Dominican-Republic", "El-Salvador", "Guatemala","Haiti", "Honduras
southAmerica <- c("Columbia", "Ecuador", "Peru")
europe <- c("England", "France", "Germany", "Greece", "Holand-Netherlands","Hungary", "Ireland", "Italy
            "Yugoslavia")
other <- c("South")
data$native.country[data$native.country %in% northAmerica] <- "North America"
data$native.country[data$native.country %in% asia] <- "Asia"
data$native.country[data$native.country %in% southAmerica] <- "South America"
data$native.country[data$native.country %in% europe] <- "Europe"
data$native.country[data$native.country %in% other] <- "Other"
table(data$native.country)
```

```
##
##          Asia        Europe North America         Other South America
##           472           355         21629            57            84
```

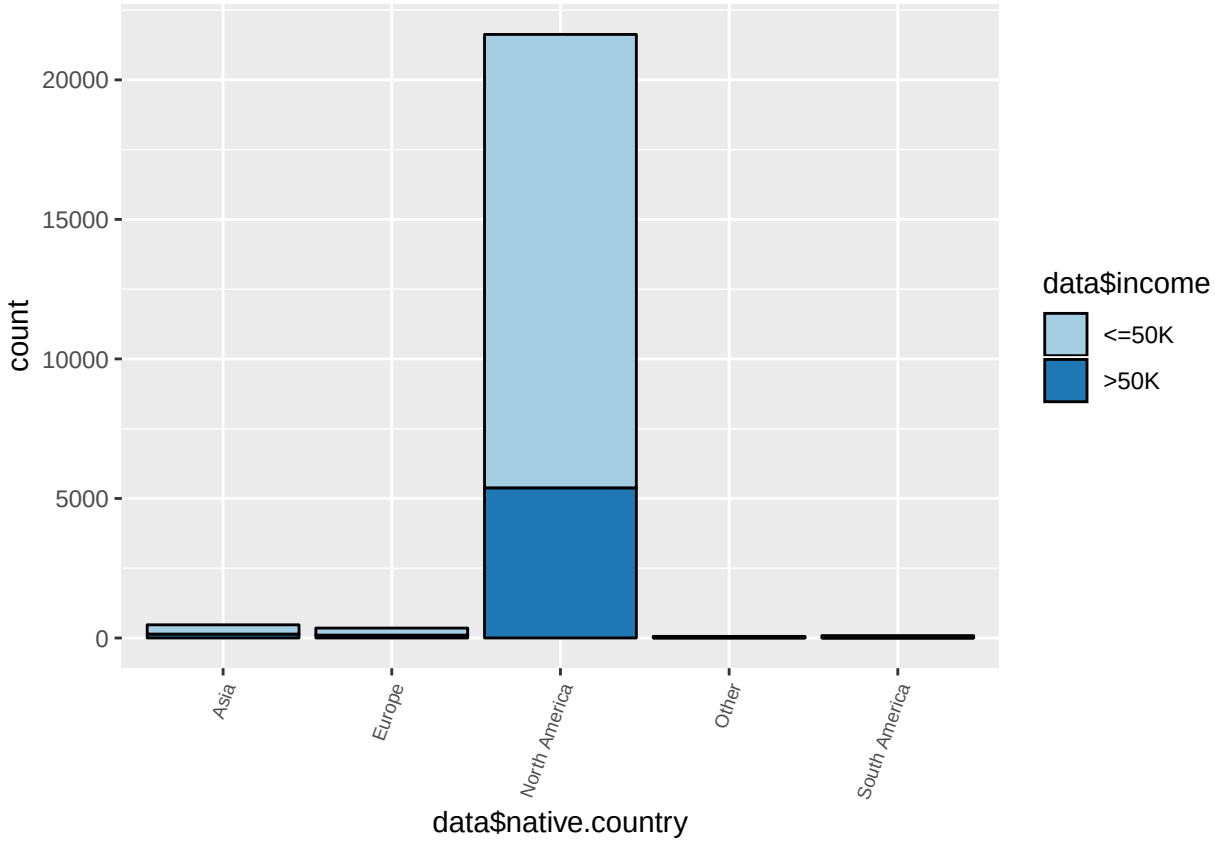
```
data$native.country <- as.factor(data$native.country)
levels(data$native.country)
```

```
## [1] "Asia"          "Europe"        "North America" "Other"
## [5] "South America"
```

```
## ------------------------------------------------------------------------------------
## Reduce the level of native country into 5 levels "Asia", "Europe", "North America"
```

```
## "Other", "South America"
## ------------------------------------------------------------------------------------------
```

```
ggplot(data, aes(x=data$native.country,fill=data$income)) + geom_bar(position = "stack", color = "black
```




```
## ------------------------------------------------------------------------------------------
## Native conutry of the Majority of the population is North America.
## ------------------------------------------------------------------------------------------
```

9.2. Reducing/Combining levels of native country in testing data.

```
testingdata$native.country <- as.character(testingdata$native.country)
testingdata$native.country[testingdata$native.country %in% northAmerica] <- "North America"
testingdata$native.country[testingdata$native.country %in% asia] <- "Asia"
testingdata$native.country[testingdata$native.country %in% southAmerica] <- "South America"
testingdata$native.country[testingdata$native.country %in% europe] <- "Europe"
testingdata$native.country[testingdata$native.country %in% other] <- "Other"
table(testingdata$native.country)
```

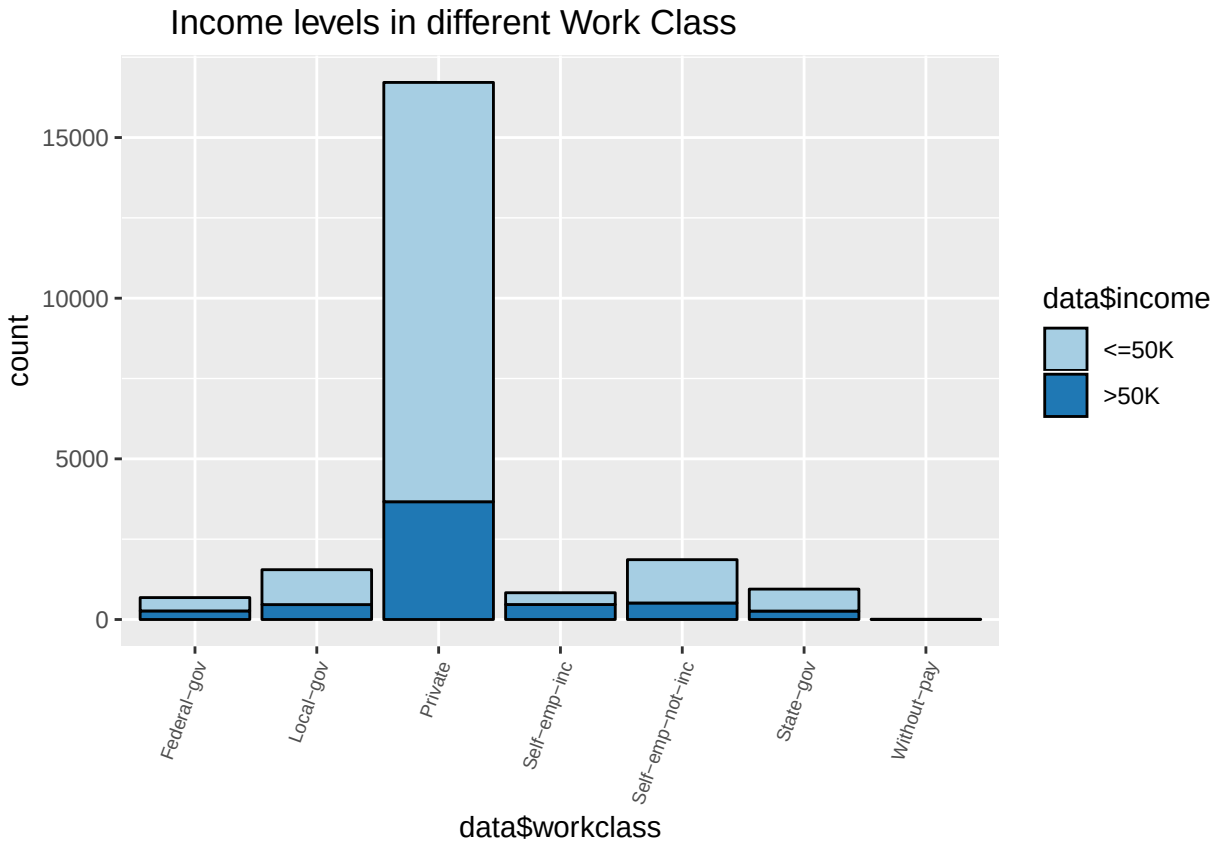
```
##
##          Asia        Europe North America         Other South America
##           162           138          7222            14            29
```

```
testingdata$native.country <- as.factor(testingdata$native.country)
levels(testingdata$native.country)
```

```
## [1] "Asia"          "Europe"        "North America" "Other"
## [5] "South America"
```

10.1. Combining categories of work class in training data.

```
ggplot(data, aes(x=data$workclass,fill=data$income)) + geom_bar(position = "stack", color = "black") + g
```

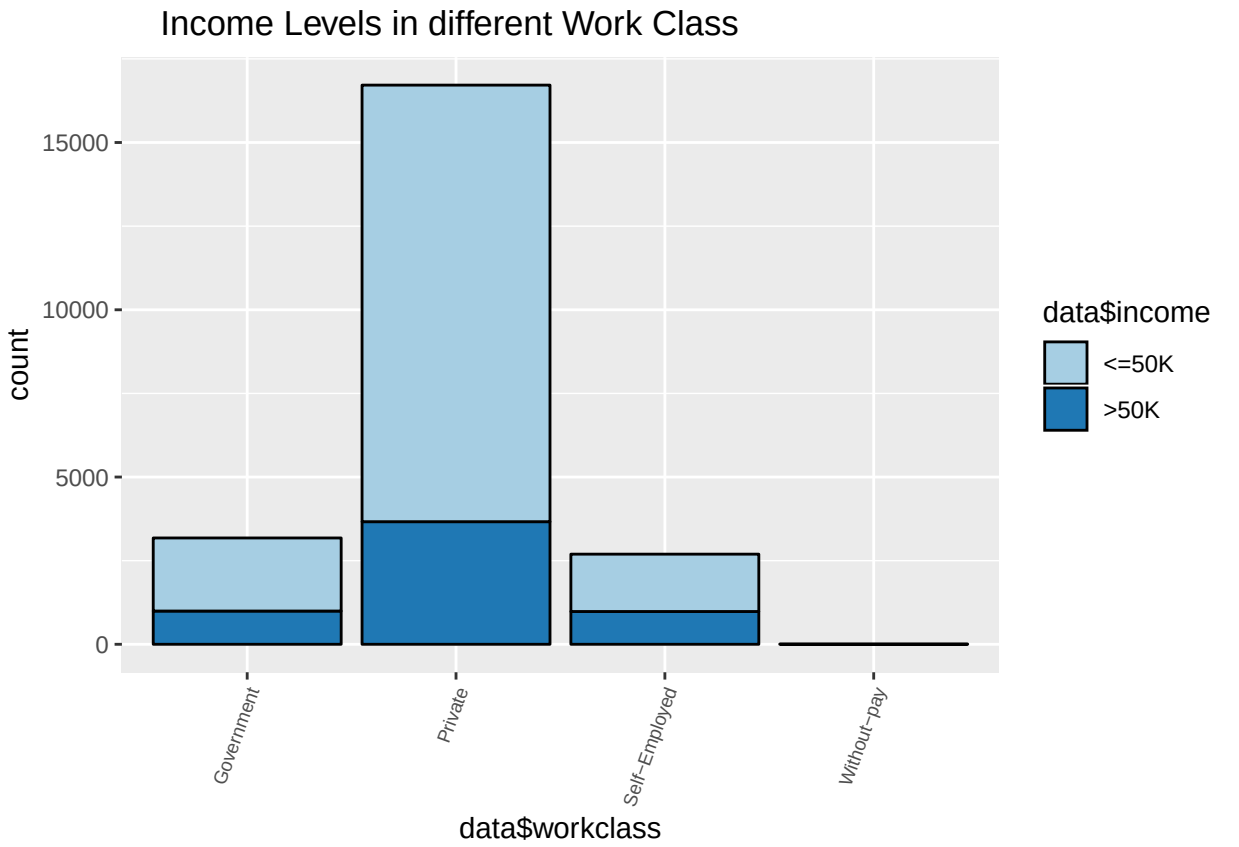



```
data$workclass <- gsub('^Federal-gov', 'Government', data$workclass)
data$workclass <- gsub('^Local-gov', 'Government', data$workclass)
data$workclass <- gsub('^State-gov', 'Government', data$workclass)


data$workclass <- gsub('^Self-emp-inc', 'Self-Employed', data$workclass)
data$workclass <- gsub('^Self-emp-not-inc', 'Self-Employed', data$workclass)

data$workclass <- gsub('^Other', 'Other', data$workclass)
data$workclass <- gsub('^Unknown', 'Other', data$workclass)

data$workclass <- as.factor(data$workclass)
ggplot(data, aes(x=data$workclass,fill=data$income)) + geom_bar(position = "stack", color = "black") + g
```




```
## ------------------------------------------------------------------
## Replace Federal-gov, Local_gov and State_gov into government.
## Self-emp-inc and self-emp-not-inc into Self-Employed.
## other and unknown into other.
##  ------------------------------------------------------------------

## Observations: Most of the people earning more than 50K are in private sector
## after that self employment and then Government.
```

10.2. Combining categories of work class in testing data.

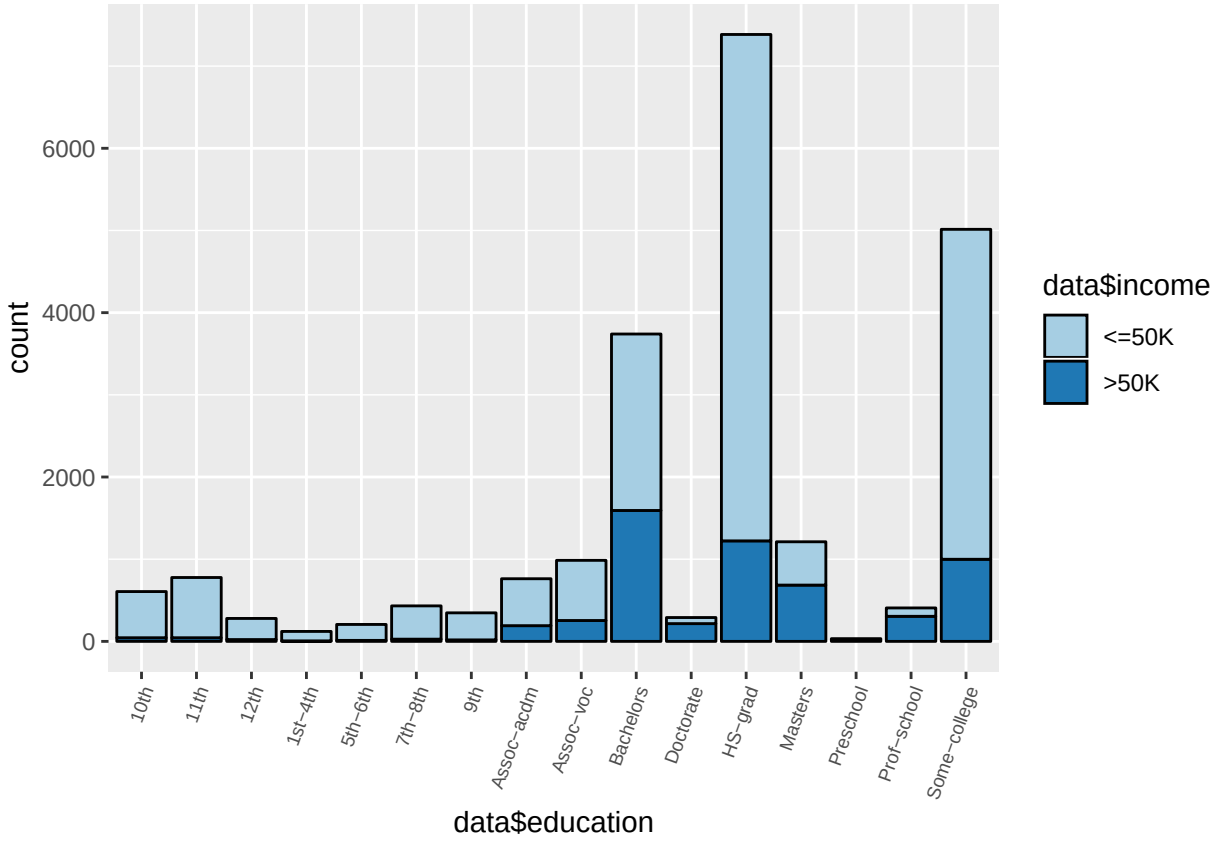
```
testingdata$workclass <- gsub('^Federal-gov', 'Government', testingdata$workclass)
testingdata$workclass <- gsub('^Local-gov', 'Government', testingdata$workclass)
testingdata$workclass <- gsub('^State-gov', 'Government', testingdata$workclass)


testingdata$workclass <- gsub('^Self-emp-inc', 'Self-Employed', testingdata$workclass)
testingdata$workclass <- gsub('^Self-emp-not-inc', 'Self-Employed', testingdata$workclass)

testingdata$workclass <- gsub('^Other', 'Other', testingdata$workclass)
testingdata$workclass <- gsub('^Unknown', 'Other', testingdata$workclass)
testingdata$workclass <- as.factor(testingdata$workclass)
```

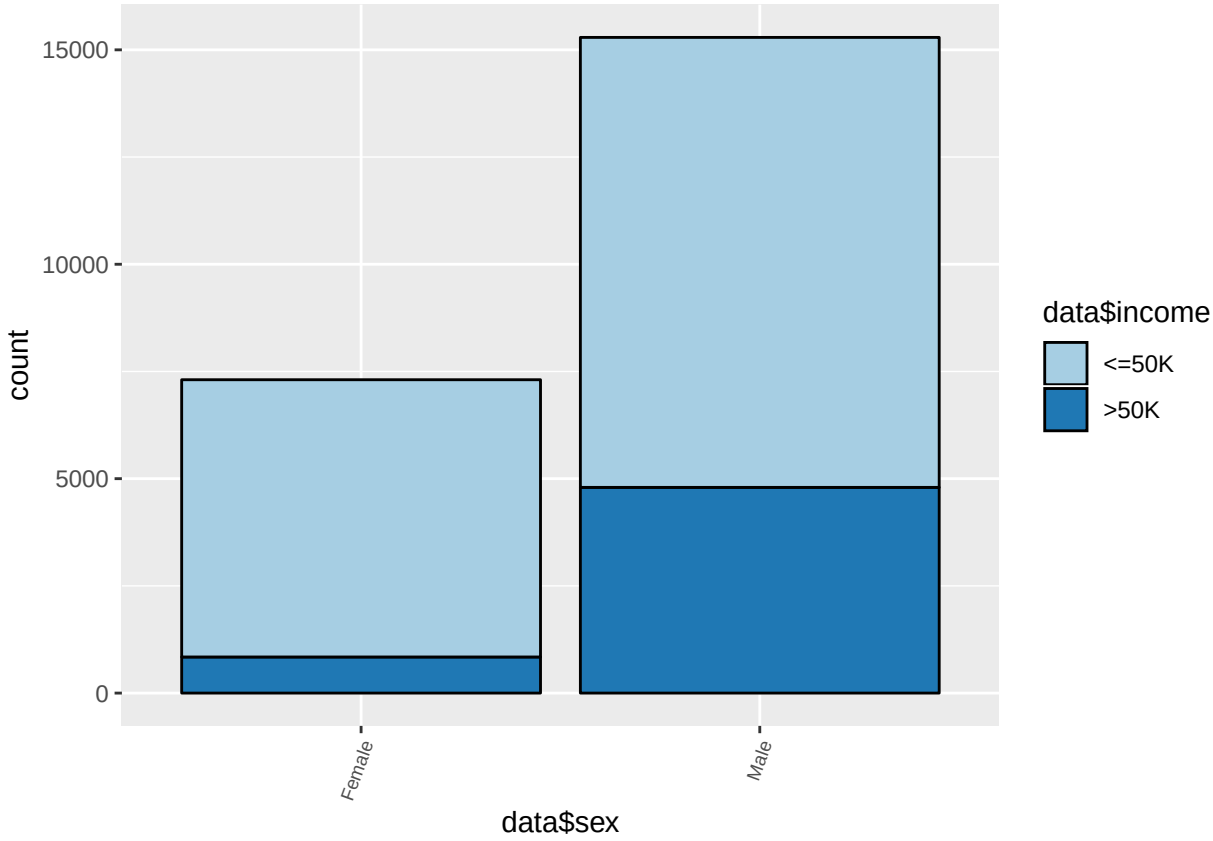
11. Relationship between categorical variables and income.

```
ggplot(data, aes(x=data$education,fill=data$income)) + geom_bar(position = "stack", color = "black") + t
```



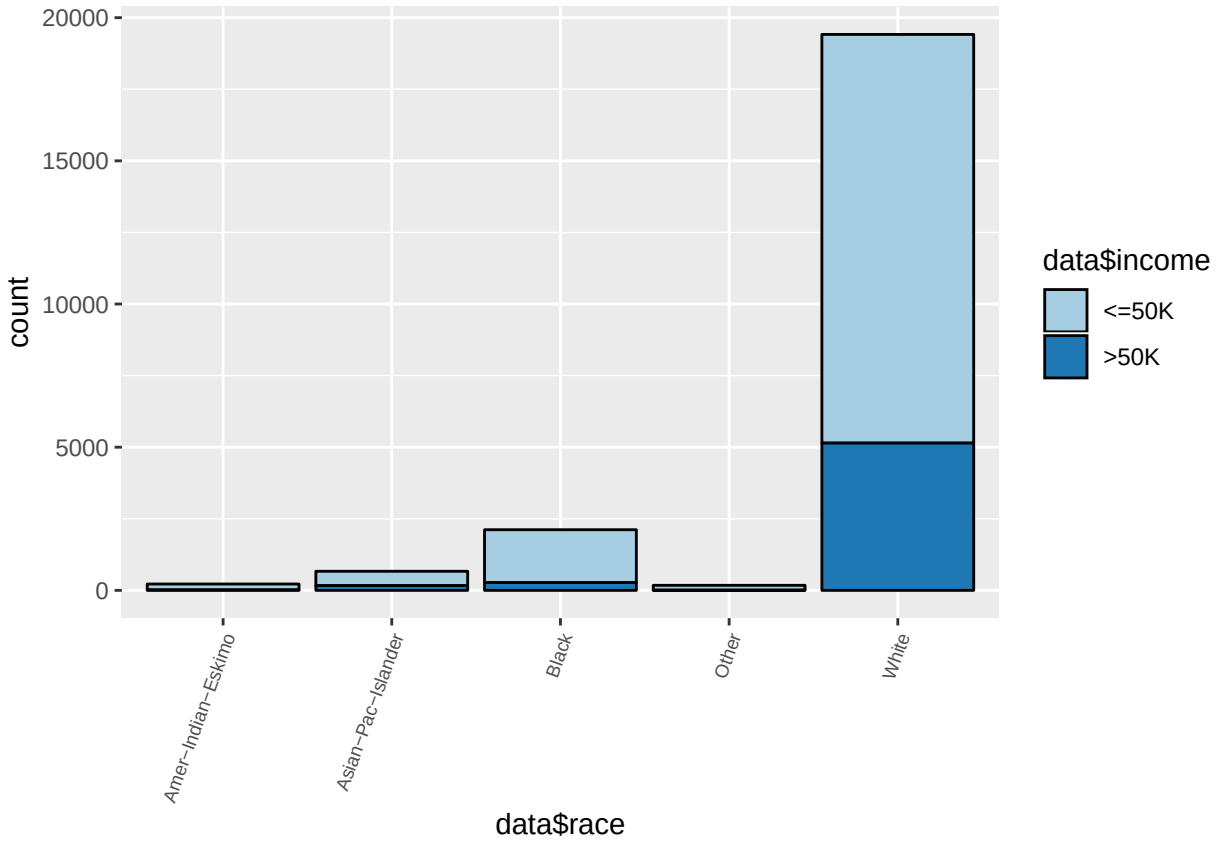

```
##------------------------------------------------------------------------------
## The plot shows that the maximum number of adults earning income greater than 50K
## have bachelor's degree.
## In doctorate and masters also, the largest proportion is earning greater than 50 K.
## In lower  education levels the largest proportion have income less than 50K.
## Higher education results in higher income.
##------------------------------------------------------------------------------


ggplot(data, aes(x=data$sex,fill=data$income)) + geom_bar(position = "stack", color = "black") + theme(a
```



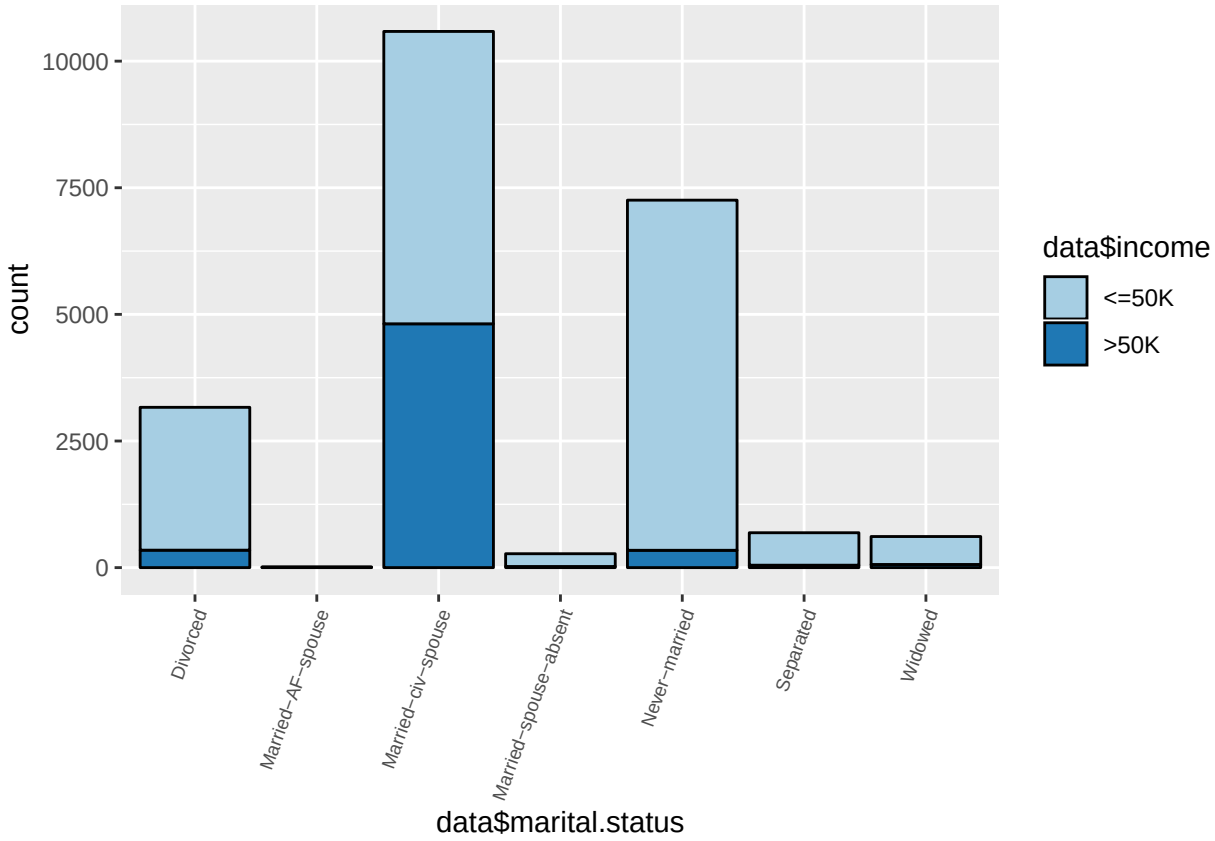

```
##-------------------------------------------------------------------------
## Ratio of Males earning income greater than 50K are more as compare to female.
## -------------------------------------------------------------------------


ggplot(data, aes(x=data$race,fill=data$income)) + geom_bar(position = "stack", color = "black") + theme
```

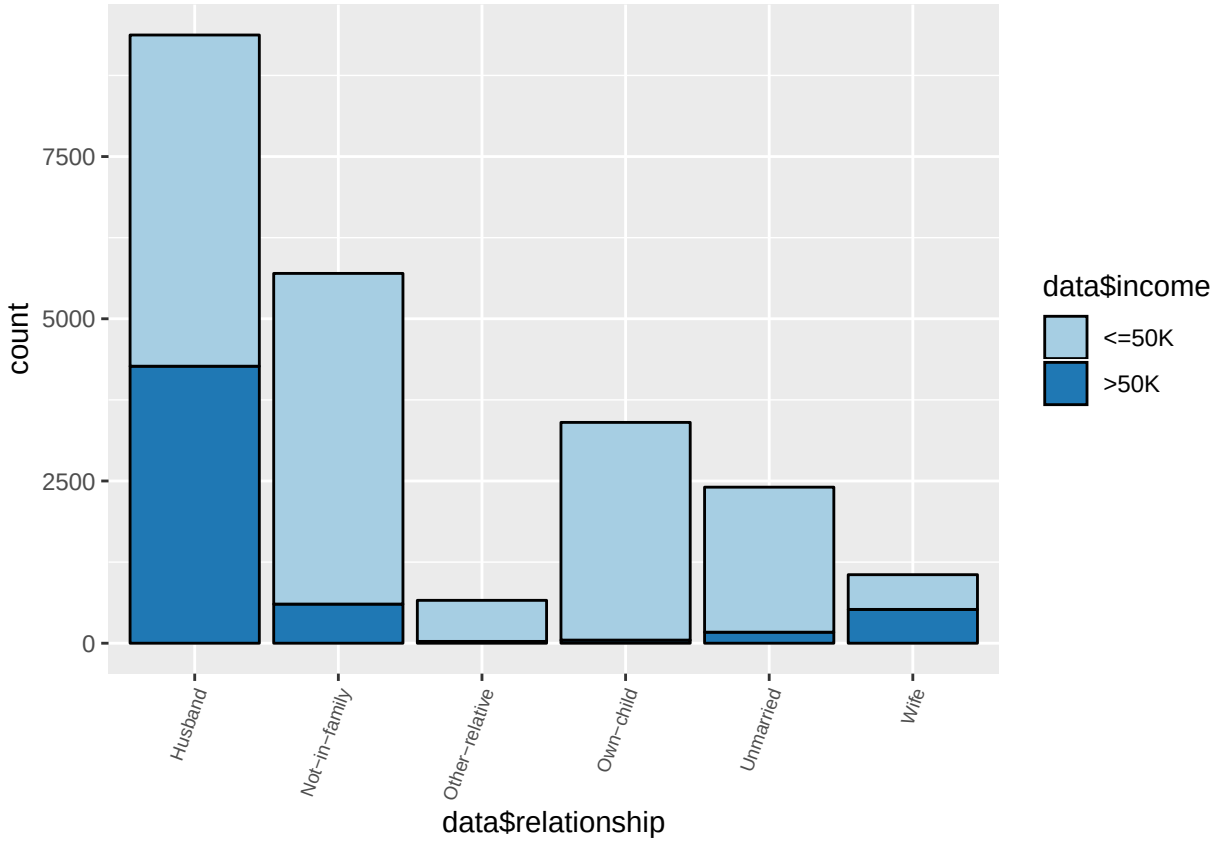



```
##--------------------------------------------------------------------------------
## Observations: Plot shows that in terms of race the highest earning people are
## from race White,then Asian-pacific and black.
##--------------------------------------------------------------------------------

ggplot(data, aes(x=data$marital.status,fill=data$income)) + geom_bar(position = "stack", color = "black
```

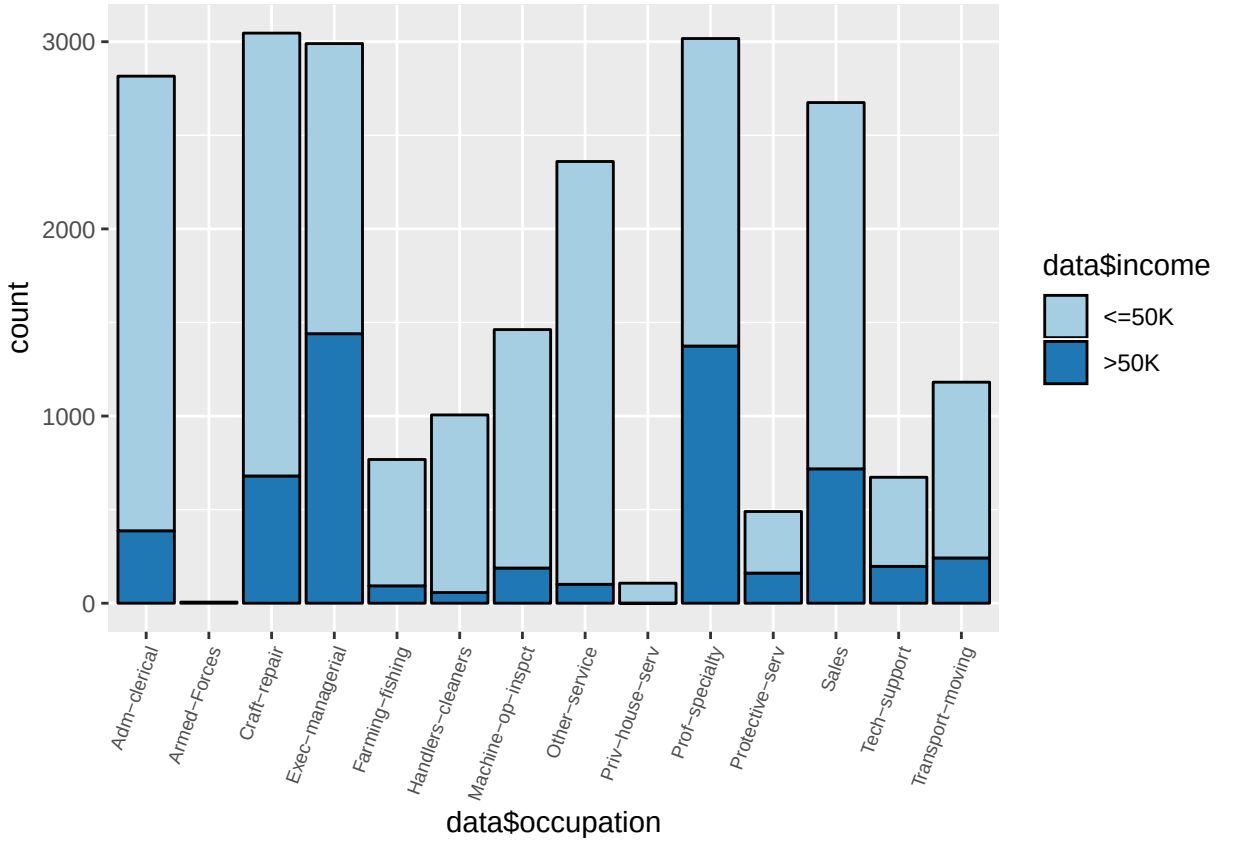



```
ggplot(data, aes(x=data$relationship,fill=data$income)) + geom_bar(position = "stack", color = "black")
```




```
##-------------------------------------------------------------------------------
## Observations: Plots shows that married people are earning more than 50k.
##-------------------------------------------------------------------------------

ggplot(data, aes(x=data$occupation,fill=data$income)) + geom_bar(position = "stack", color = "black") +
```




```
##-------------------------------------------------------------------------------
## Observations: Plots shows that in terms of occuupation people with managerical job and
## professors are earning more than 50 K in the highest ratio.
## Showing that people at highest post are earning more.
##-------------------------------------------------------------------------------
```

Saving the clean test and train data in testdata.csv and traindata.csv respectively.

```
write.csv(data, "traindata.csv", row.names = FALSE)
write.csv(testingdata, "testdata.csv", row.names = FALSE)
```