Census Project

1. Load requied libraries.
2. Install packages install.packages("caret")
3. Install packaged install.packages("corrplot")
4. 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         :17001   Min.   :  14878
##  1st Qu.:28.00   Self-emp-not-inc: 1922   1st Qu.: 117959
##  Median :37.00   Local-gov       : 1594   Median : 178244
##  Mean   :38.59   State-gov       :  960   Mean   : 189756
##  3rd Qu.:48.00   Self-emp-inc    :  834   3rd Qu.: 236992
##  Max.   :90.00   (Other)         :  734   Max.   :1484705
##                  NA's            : 1376
```

```
##         education    education.num                marital.status
##  HS-grad     :7851   Min.   : 1.0   Divorced             : 3335
##  Some-college:5496   1st Qu.: 9.0   Married-AF-spouse    :   16
##  Bachelors   :4008   Median :10.0   Married-civ-spouse   :11211
##  Masters     :1308   Mean   :10.1   Married-spouse-absent:  315
##  Assoc-voc   :1044   3rd Qu.:12.0   Never-married        : 8026
##  11th        : 848   Max.   :16.0   Separated            :  763
##  (Other)     :3866                  Widowed              :  755
##            occupation           relationship                   race
##  Prof-specialty :3147   Husband       :9867   Amer-Indian-Eskimo:  234
##  Craft-repair   :3103   Not-in-family :6217   Asian-Pac-Islander:  796
##  Exec-managerial:3030   Other-relative: 746   Black             : 2349
##  Adm-clerical   :2848   Own-child     :3798   Other             :  207
##  Sales          :2728   Unmarried     :2612   White             :20835
##  (Other)        :8185   Wife          :1181
##  NA's           :1380
##      sex          capital.gain    capital.loss     hours.per.week
##  Female: 8055   Min.   :    0   Min.   :   0.00   Min.   : 1.0
##  Male  :16366   1st Qu.:    0   1st Qu.:   0.00   1st Qu.:40.0
##                 Median :    0   Median :   0.00   Median :40.0
##                 Mean   : 1106   Mean   :  89.57   Mean   :40.4
##                 3rd Qu.:    0   3rd Qu.:   0.00   3rd Qu.:45.0
##                 Max.   :99999   Max.   :4356.00   Max.   :99.0
##
##        native.country    income
##  United-States:21851   <=50K:18540
##  Mexico       :  475   >50K : 5881
##  Philippines  :  151
##  Germany      :  105
##  Canada       :   99
##  (Other)      : 1284
##  NA's         :  456
```

```
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"
```

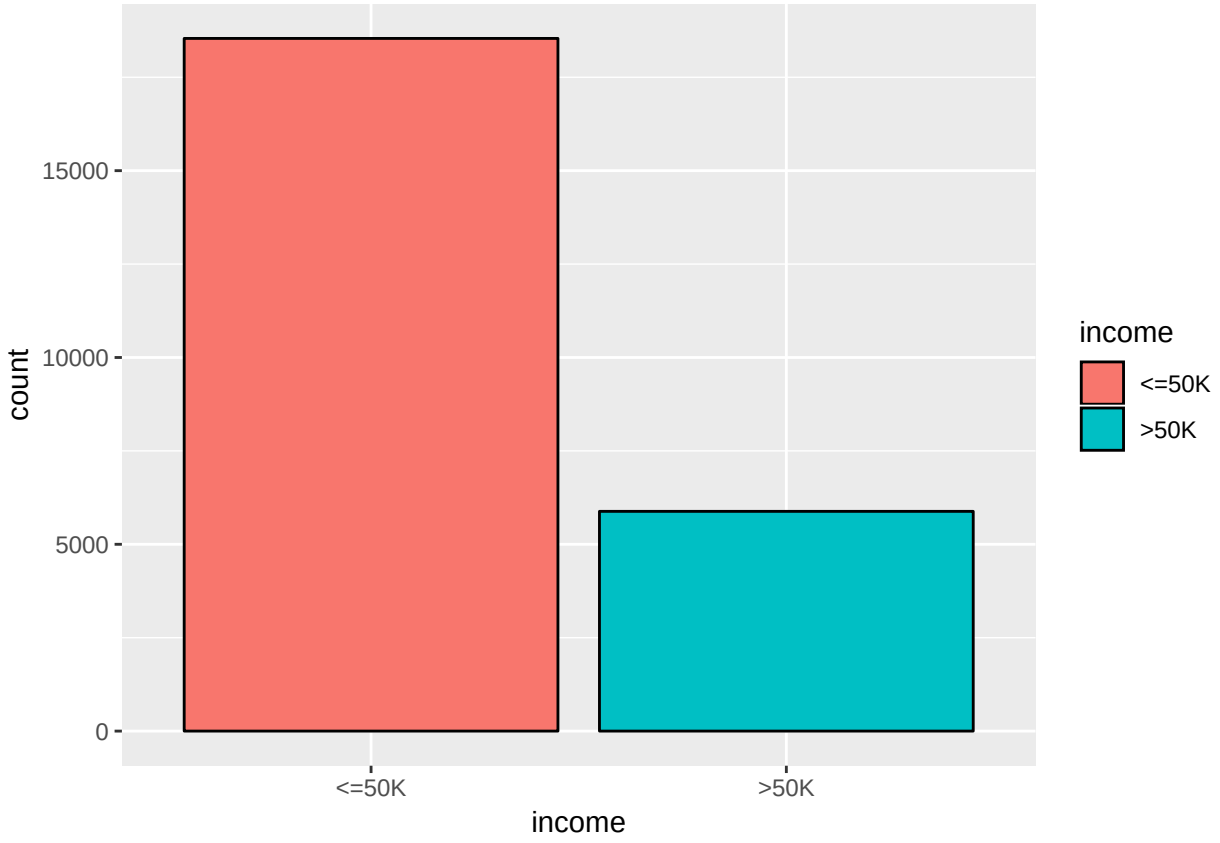
```
str(data)
```

```
## 'data.frame':    24421 obs. of  15 variables:
##  $ age           : int  90 82 66 54 41 74 41 45 52 32 ...
##  $ workclass     : Factor w/ 8 levels "Federal-gov",..: NA 4 NA 4 4 7 4 4 4 4 ...
##  $ fnlwgt        : int  77053 132870 186061 140359 264663 88638 70037 172274 129177 136204 ...
##  $ education     : Factor w/ 16 levels "10th","11th",..: 12 12 16 6 16 11 16 11 10 13 ...
##  $ education.num : int  9 9 10 4 10 16 10 16 13 14 ...
##  $ marital.status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 7 7 7 1 6 5 5 1 7 6 ...
##  $ occupation    : Factor w/ 14 levels "Adm-clerical",..: NA 4 NA 7 10 10 3 10 8 4 ...
##  $ relationship  : Factor w/ 6 levels "Husband","Not-in-family",..: 2 2 5 5 4 3 5 5 2 2 ...
##  $ race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 3 5 5 5 5 3 5 5 ...
##  $ sex           : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 2 1 1 2 ...
##  $ capital.gain  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ capital.loss  : int  4356 4356 4356 3900 3900 3683 3004 3004 2824 2824 ...
##  $ hours.per.week: int  40 18 40 40 40 20 60 35 20 55 ...
```

```
##  $ 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 1 2 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: 3212
```

```
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           1376
## fnlwgt                 0
## education              0
## education.num          0
```

```
## marital.status        0
## occupation         1380
## relationship          0
## race                  0
## sex                   0
## capital.gain          0
## capital.loss          0
## hours.per.week        0
## native.country      456
## income                0
```

```
nrow(data)
```

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

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

```
## [1] 22604
```

```
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: 1050
```

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

```
##            age      workclass         fnlwgt      education  education.num
##              0            460              0              0              0
## marital.status     occupation   relationship           race            sex
##              0            463              0              0              0
##   capital.gain   capital.loss hours.per.week native.country         income
##              0              0              0            127              0
```

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

```
## [1] 7558
```

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

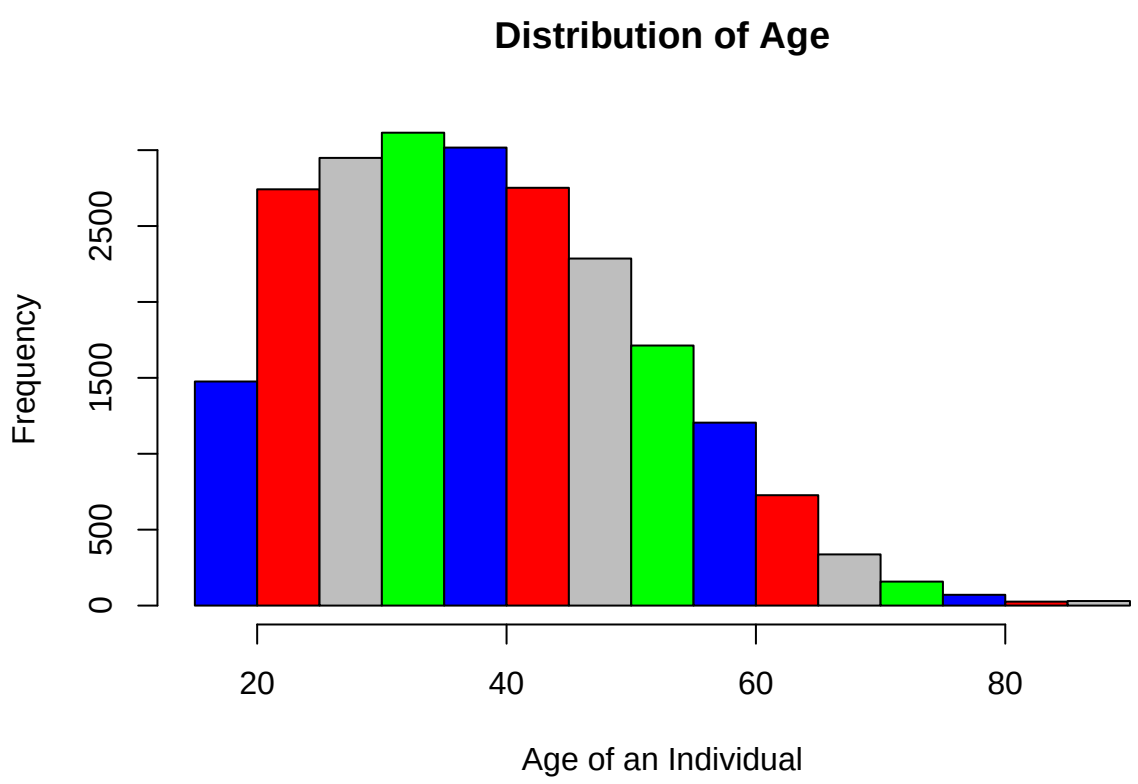
```
# statistics of numerical attributes
summary(data$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   28.00   37.00   38.45   47.00   90.00
```

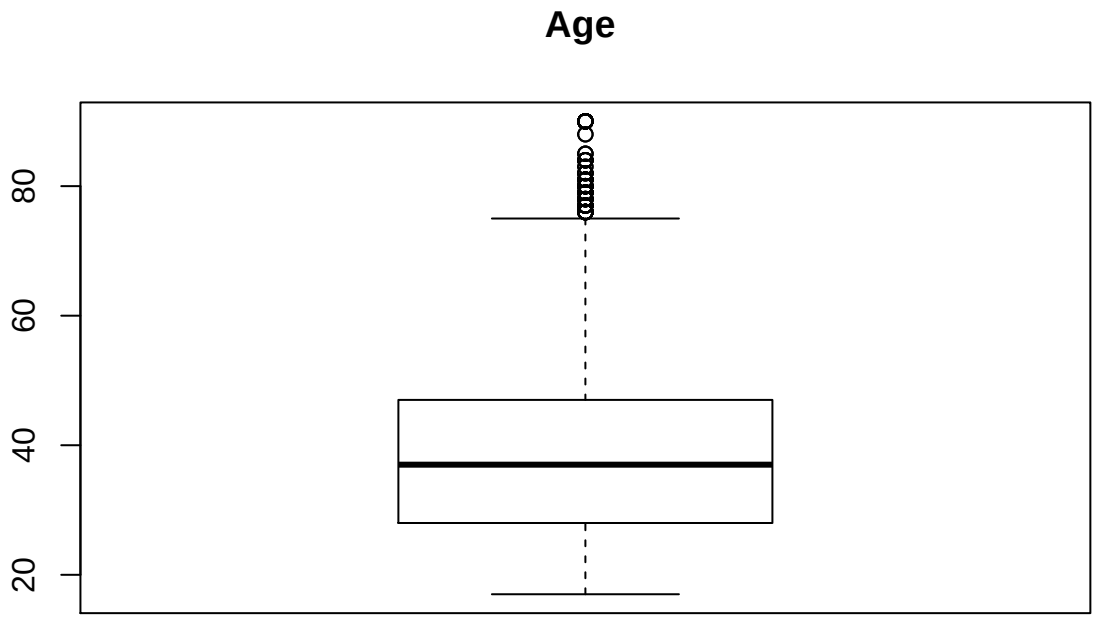
```
sd(data$age)
```

```
## [1] 13.10742
```

```
hist(data$age, main = "Distribution of Age",xlab = "Age of an Individual" ,col = c("blue","red","gray",
```




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



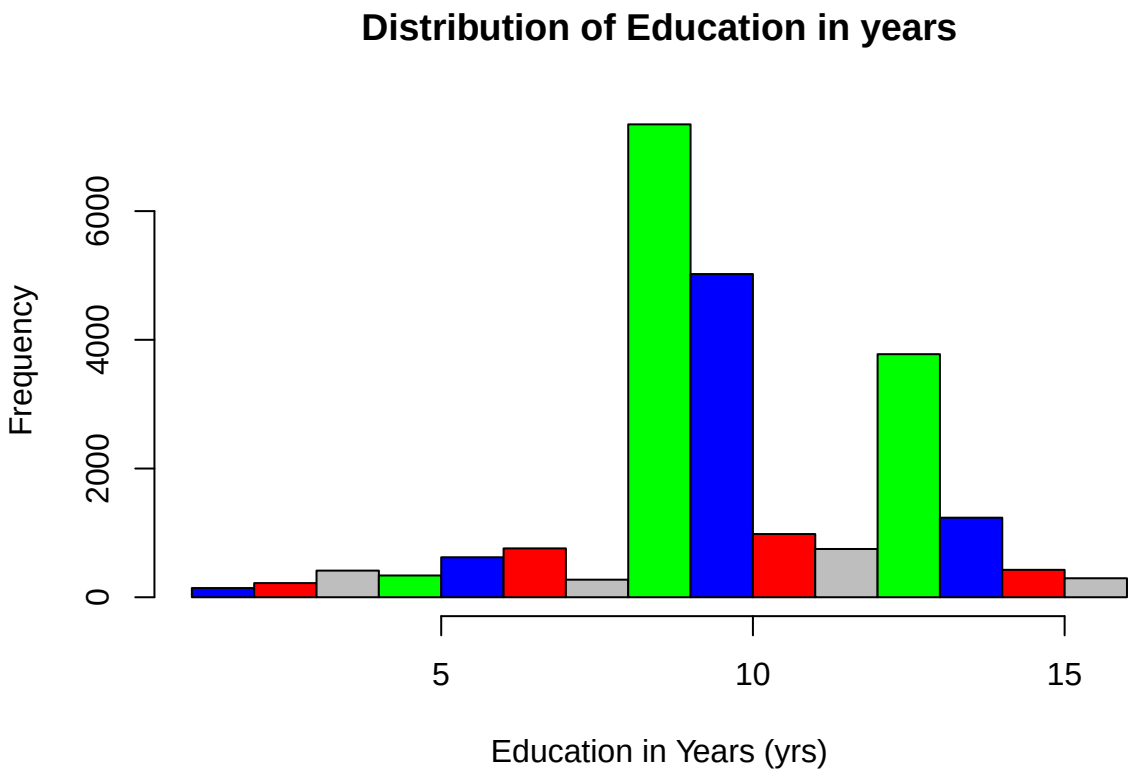

```
summary(data$education.num)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    9.00   10.00   10.14   13.00   16.00
```

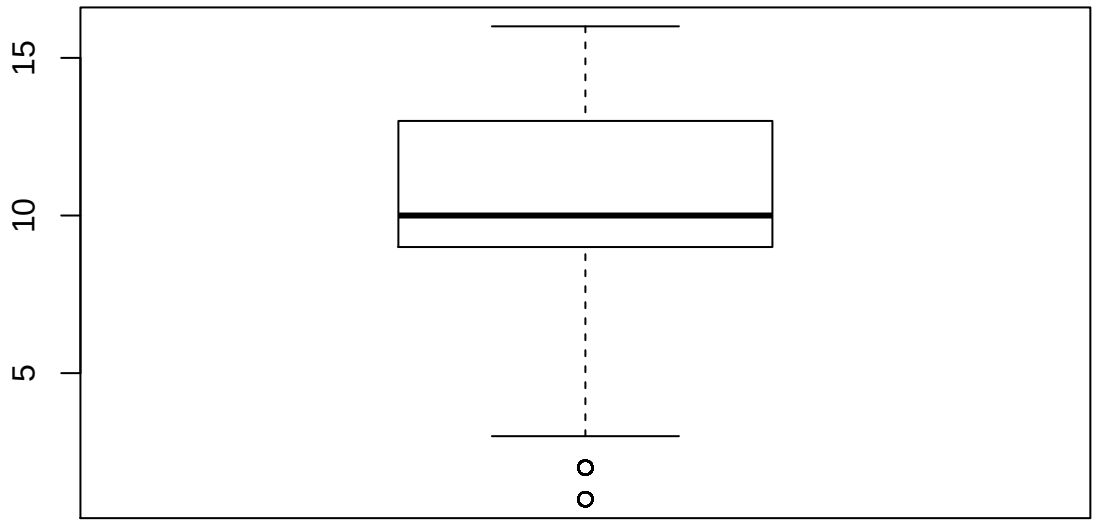
```
sd(data$education.num)
```

```
## [1] 2.555364
```

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




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



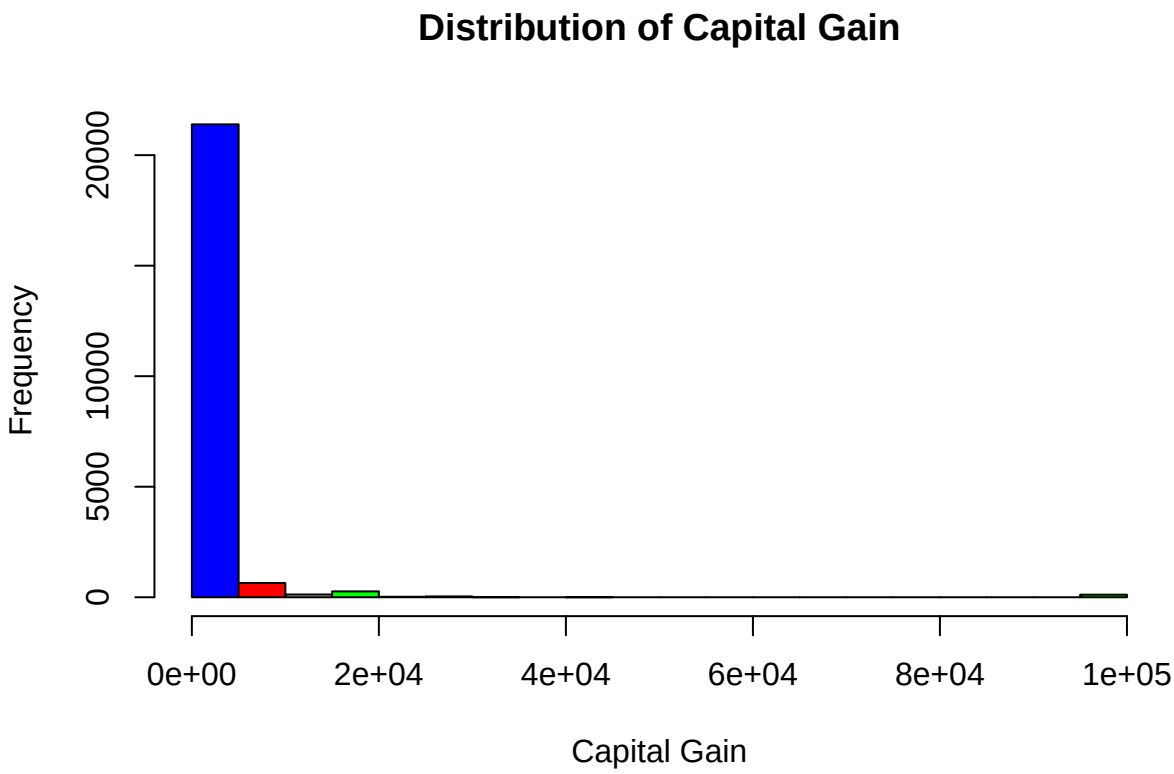

```
summary(data$capital.gain)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0    1129       0   99999
```

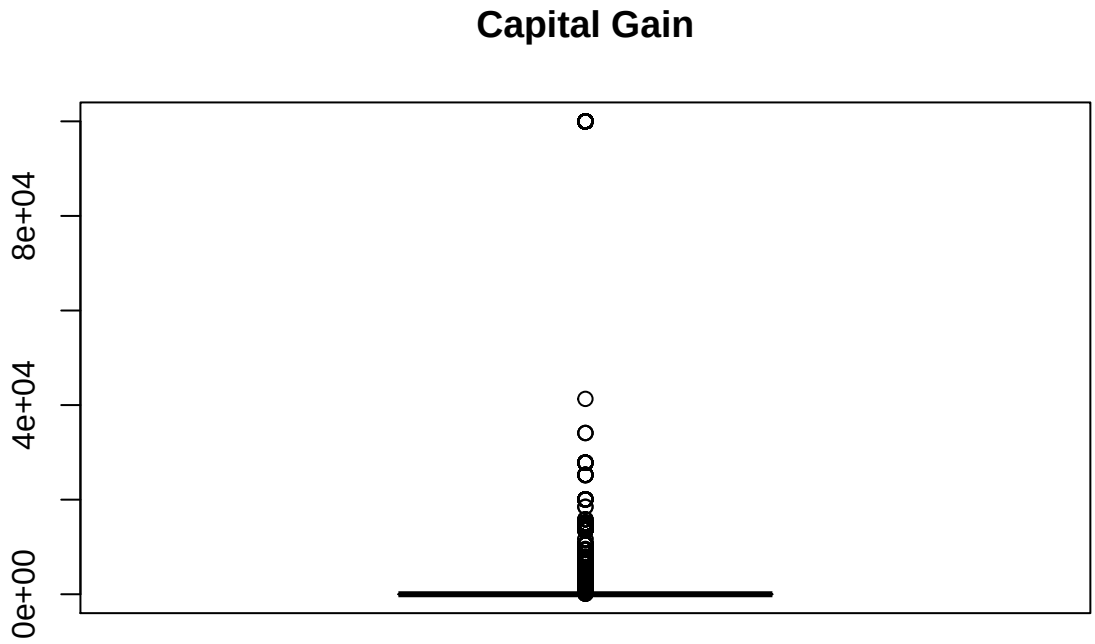
```
sd(data$capital.gain)
```

```
## [1] 7561.757
```

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




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



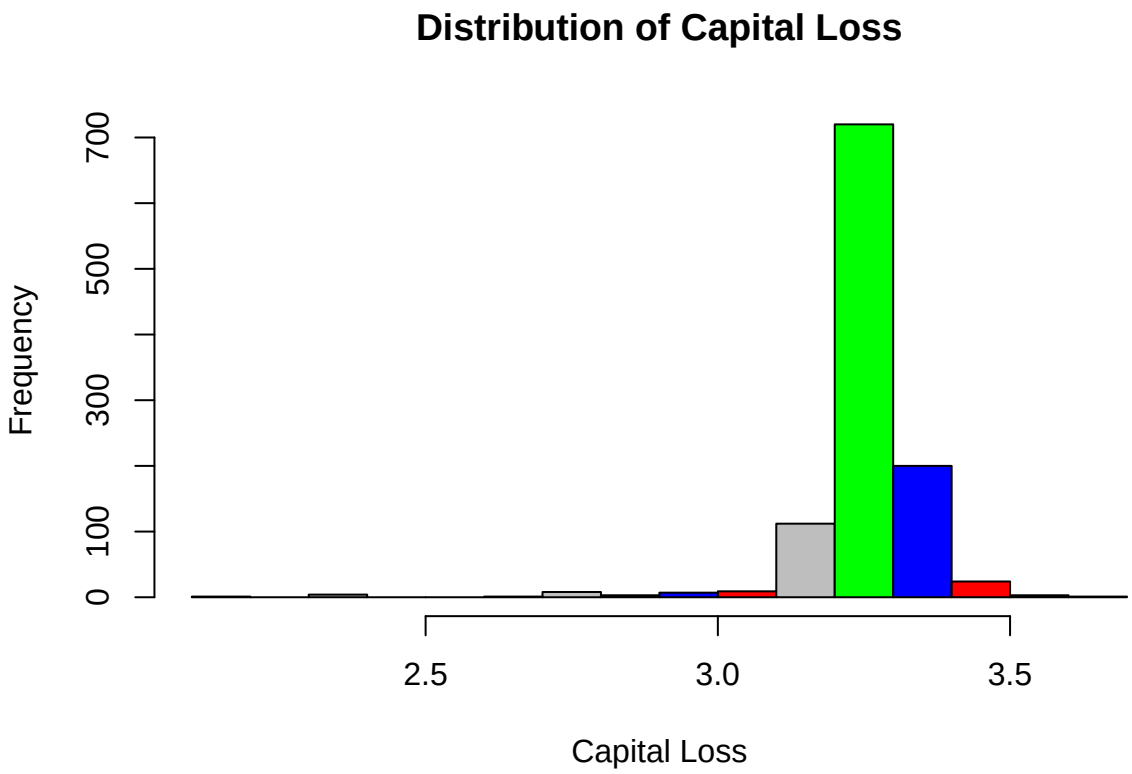

```
summary(data$capital.loss)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00   90.51    0.00 4356.00
```

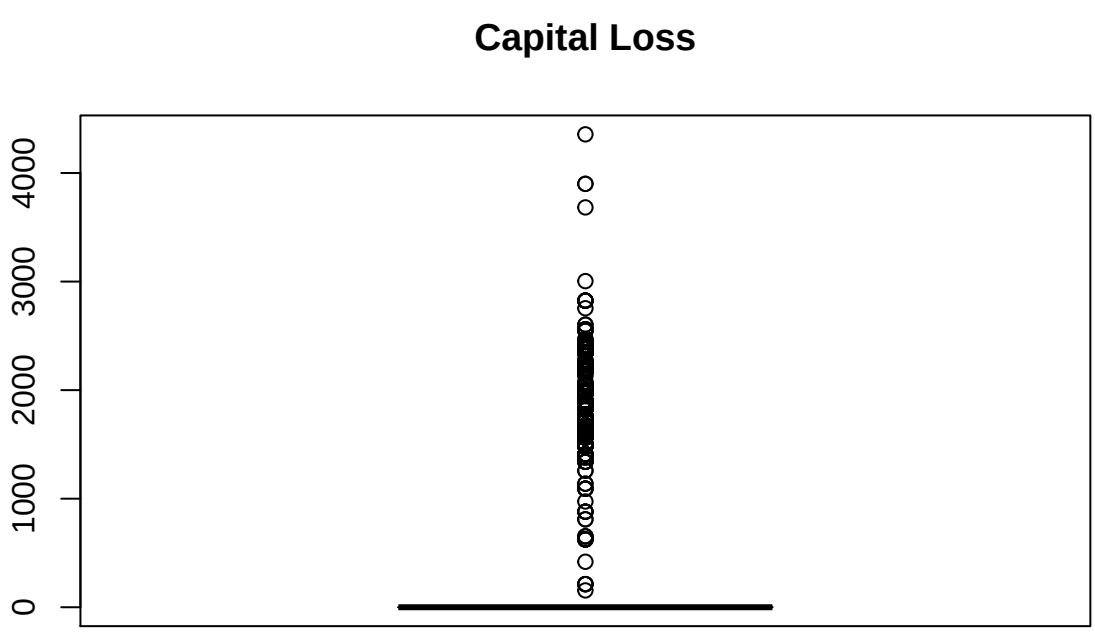
```
sd(data$capital.loss)
```

```
## [1] 409.0437
```

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




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



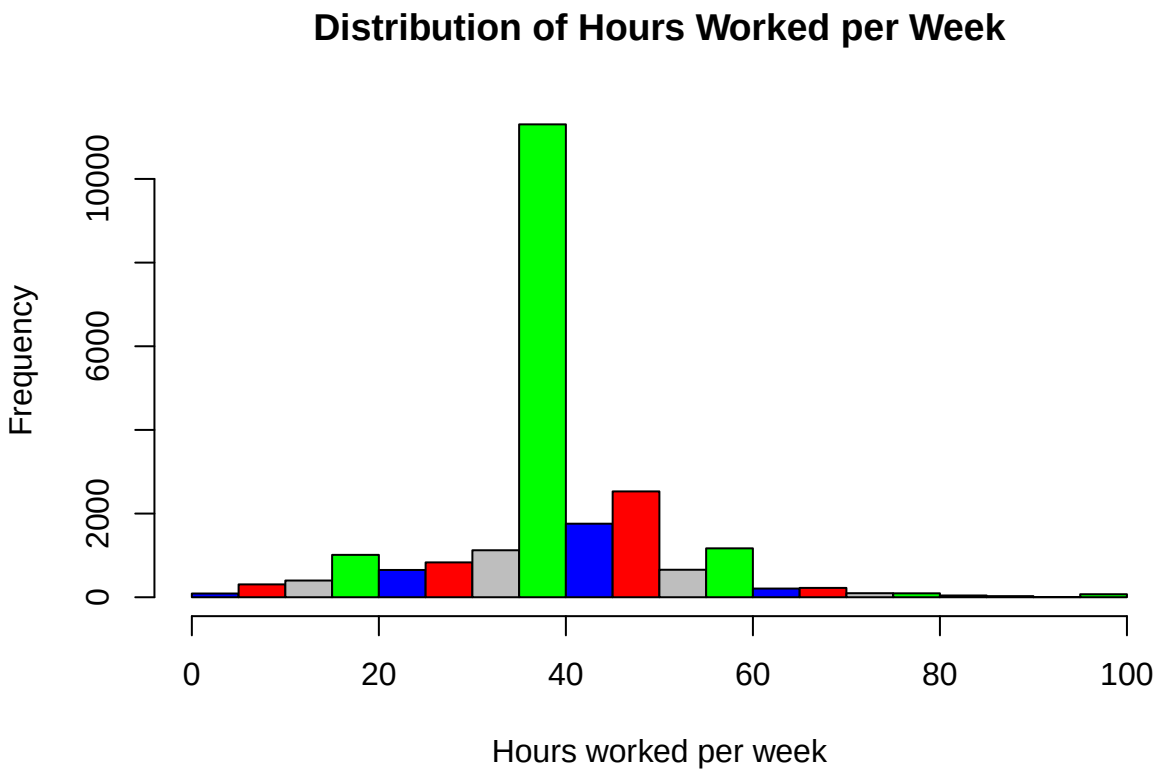

```
summary(data$hours.per.week)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   40.00   40.00   40.89   45.00   99.00
```

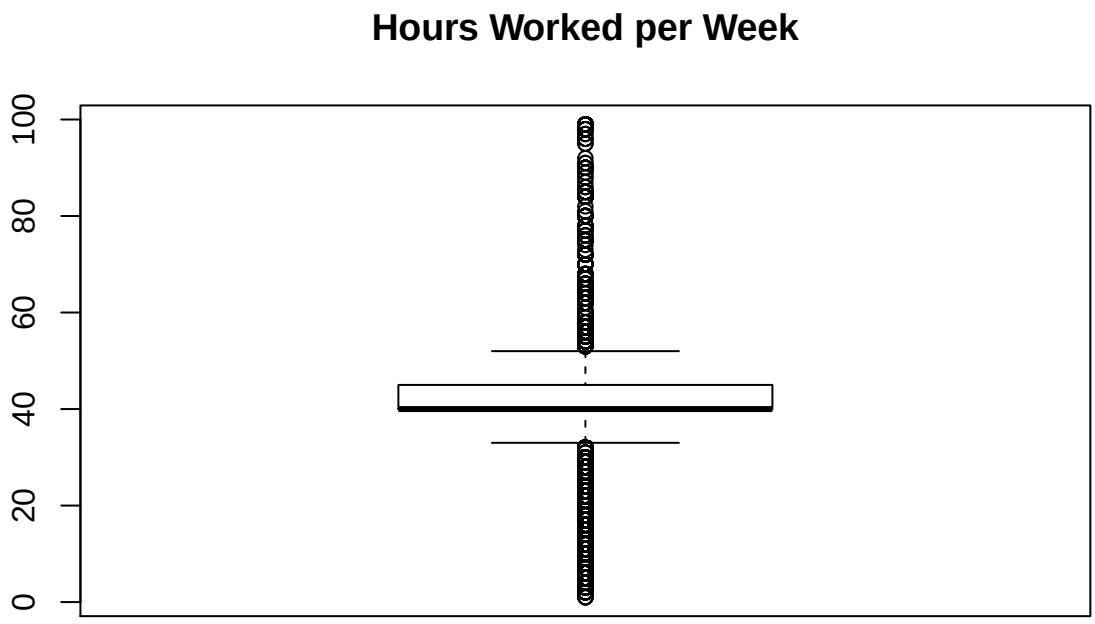
```
sd(data$hours.per.week)
```

```
## [1] 11.97622
```

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




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



8. Exploratory analysis of the attribute native country.

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

```
##                    Cambodia                      Canada
##                          12                          87
##                       China                    Columbia
##                          51                          42
##                        Cuba          Dominican-Republic
##                          65                          51
##                     Ecuador                 El-Salvador
##                          16                          70
##                     England                      France
##                          73                          23
##                     Germany                      Greece
##                          96                          21
##                   Guatemala                       Haiti
##                          48                          27
##          Holand-Netherlands                    Honduras
##                           1                          10
##                        Hong                     Hungary
##                          12                           9
##                       India                        Iran
##                          83                          32
##                     Ireland                       Italy
##                          17                          54
##                     Jamaica                       Japan
```

```
##                         60                         43
##                       Laos                     Mexico
##                         10                        450
##                  Nicaragua Outlying-US(Guam-USVI-etc)
##                         26                         11
##                       Peru                Philippines
##                         24                        143
##                     Poland                   Portugal
##                         41                         23
##                Puerto-Rico                   Scotland
##                         79                          8
##                      South                     Taiwan
##                         56                         32
##                   Thailand            Trinadad&Tobago
##                         13                         17
##               United-States                    Vietnam
##                       20605                         50
##                  Yugoslavia
##                         13
```

9. 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
##           481           379         21606            56            82
```

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

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

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

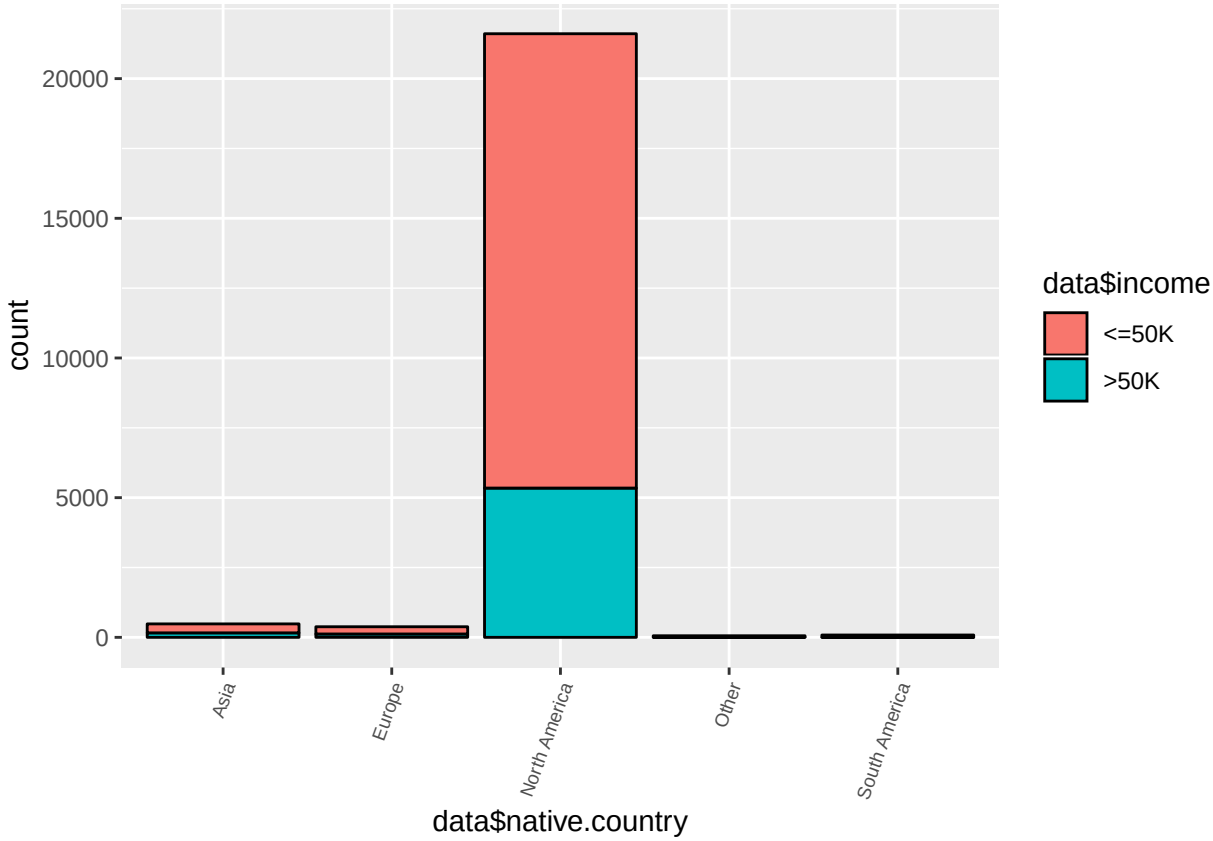
```
testingdata$native.country <- as.character(testingdata$native.country)
testingdata$native.country[testingdata$native.country %in% northAmerica] <- "North America"
testingdata$native.country[testingdata$native.country %in% asia] <- "Asia"
testingdata$native.country[testingdata$native.country %in% 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
##           153           114          7245            15             31
```

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

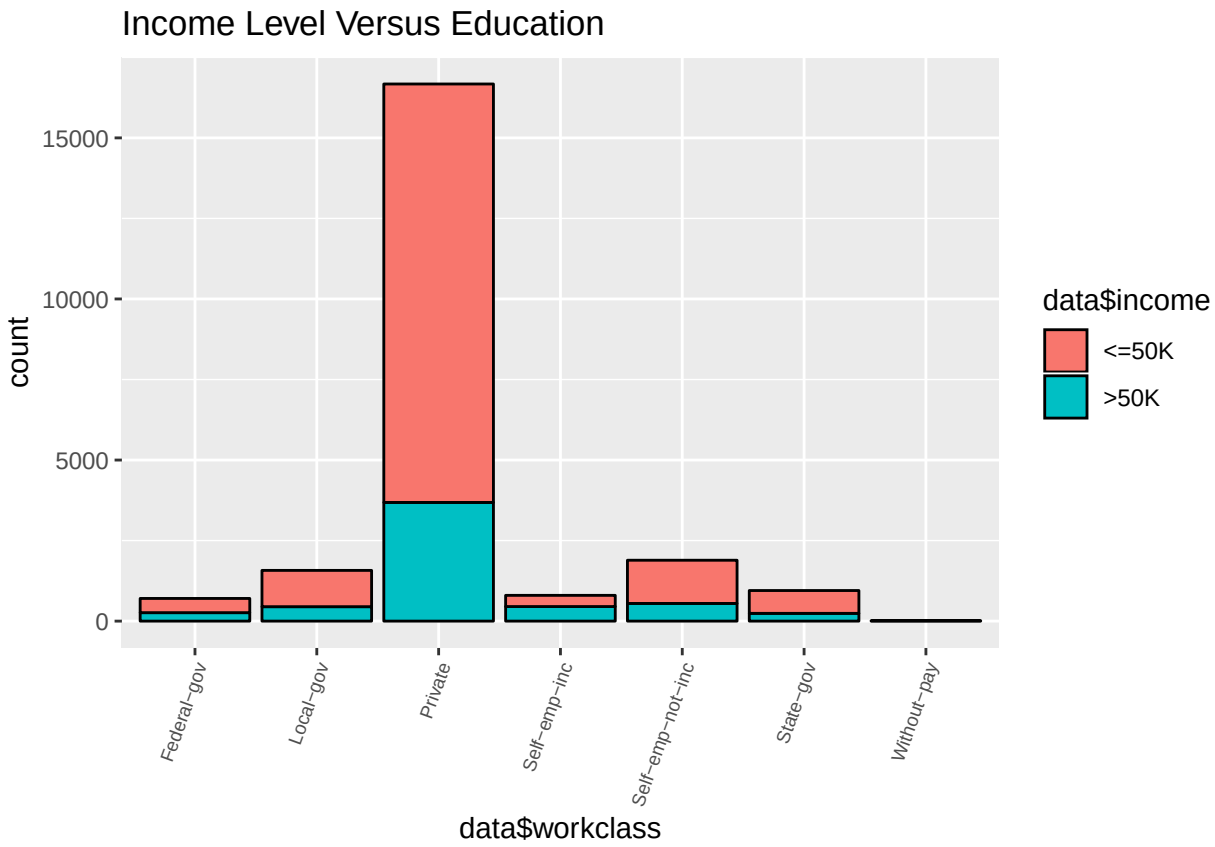
```
## [1] "Asia"          "Europe"        "North America" "Other"
## [5] "South America"
```

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

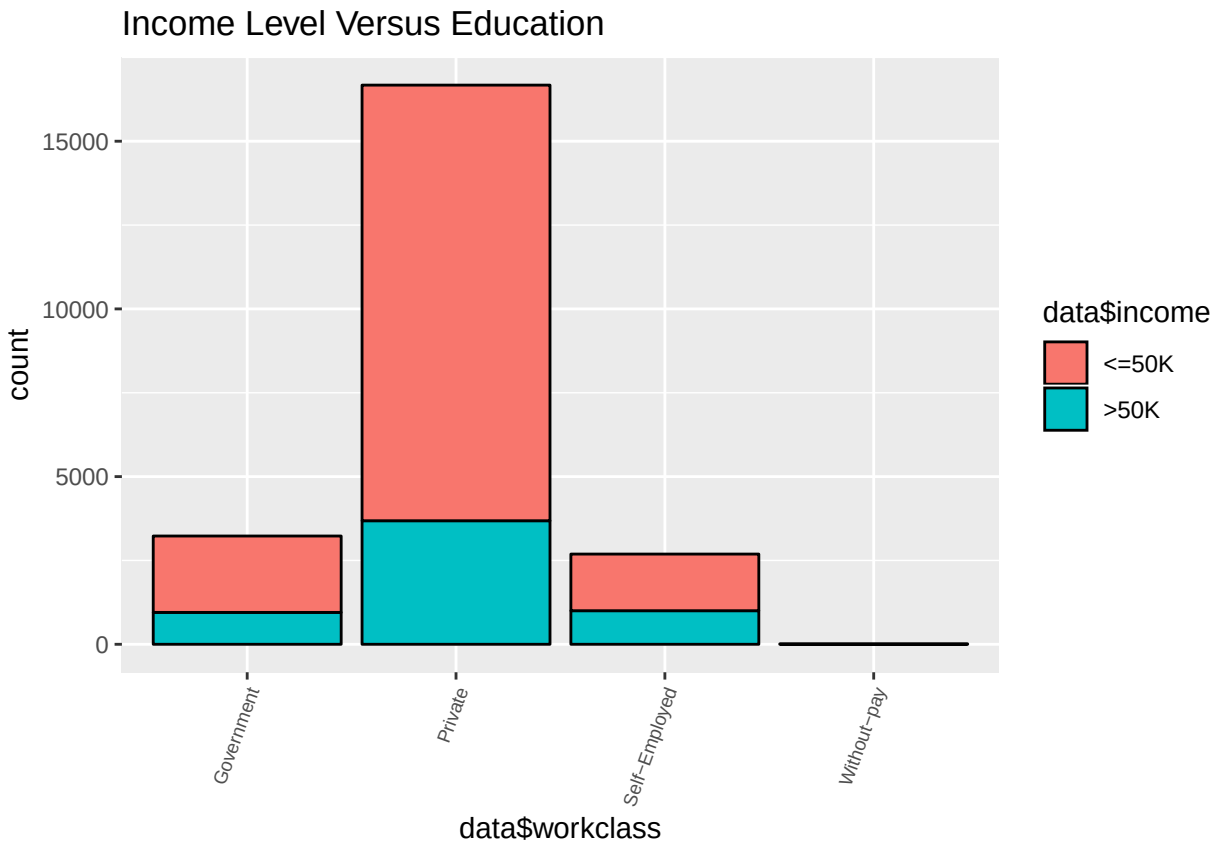


10.2. Reducing/ Combining levels of work class in train data set.

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

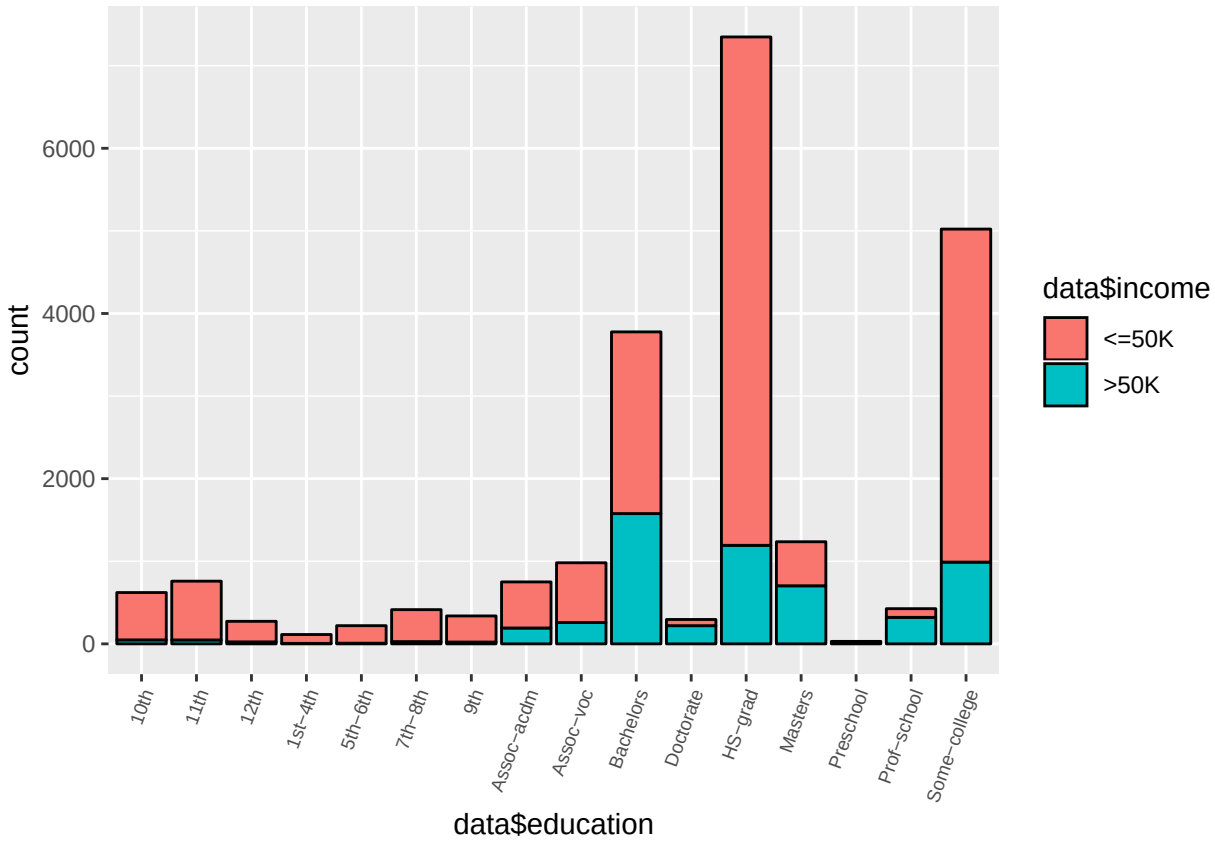



```
data$workclass <- gsub('^Federal-gov', 'Government', data$workclass)
data$workclass <- gsub('^Local-gov', 'Government', data$workclass)
data$workclass <- gsub('^State-gov', 'Government', data$workclass)


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
```

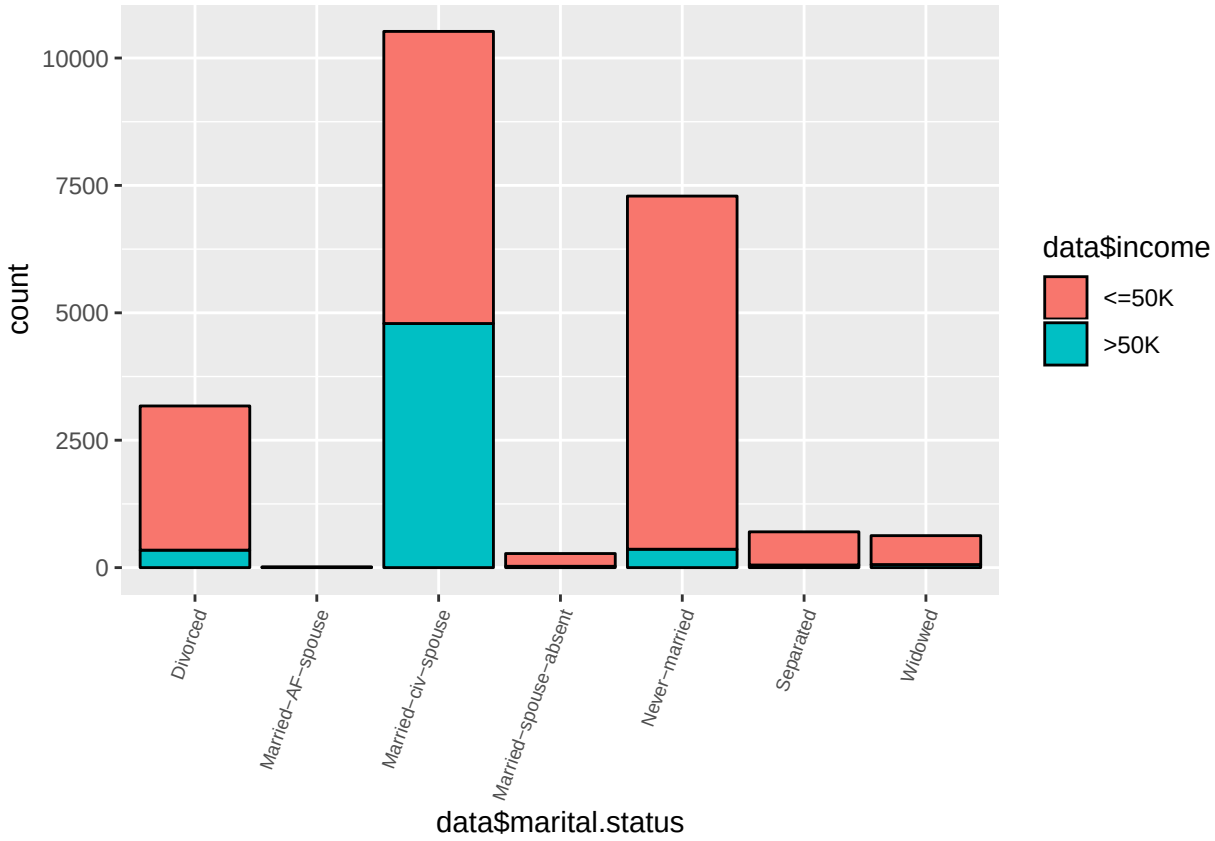


11. Relationship between categorical variables and income.

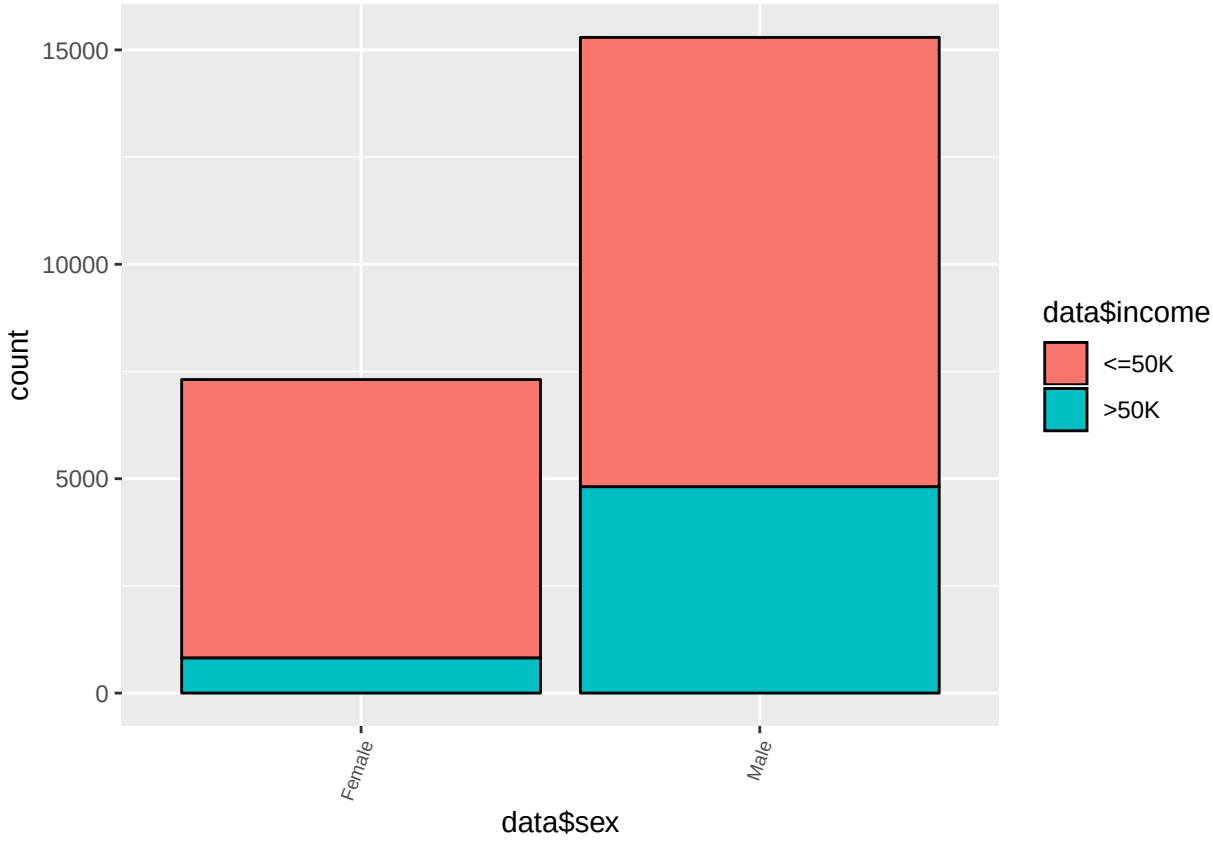
```
ggplot(data, aes(x=data$education,fill=data$income)) + geom_bar(position = "stack", color = "black") + t
```



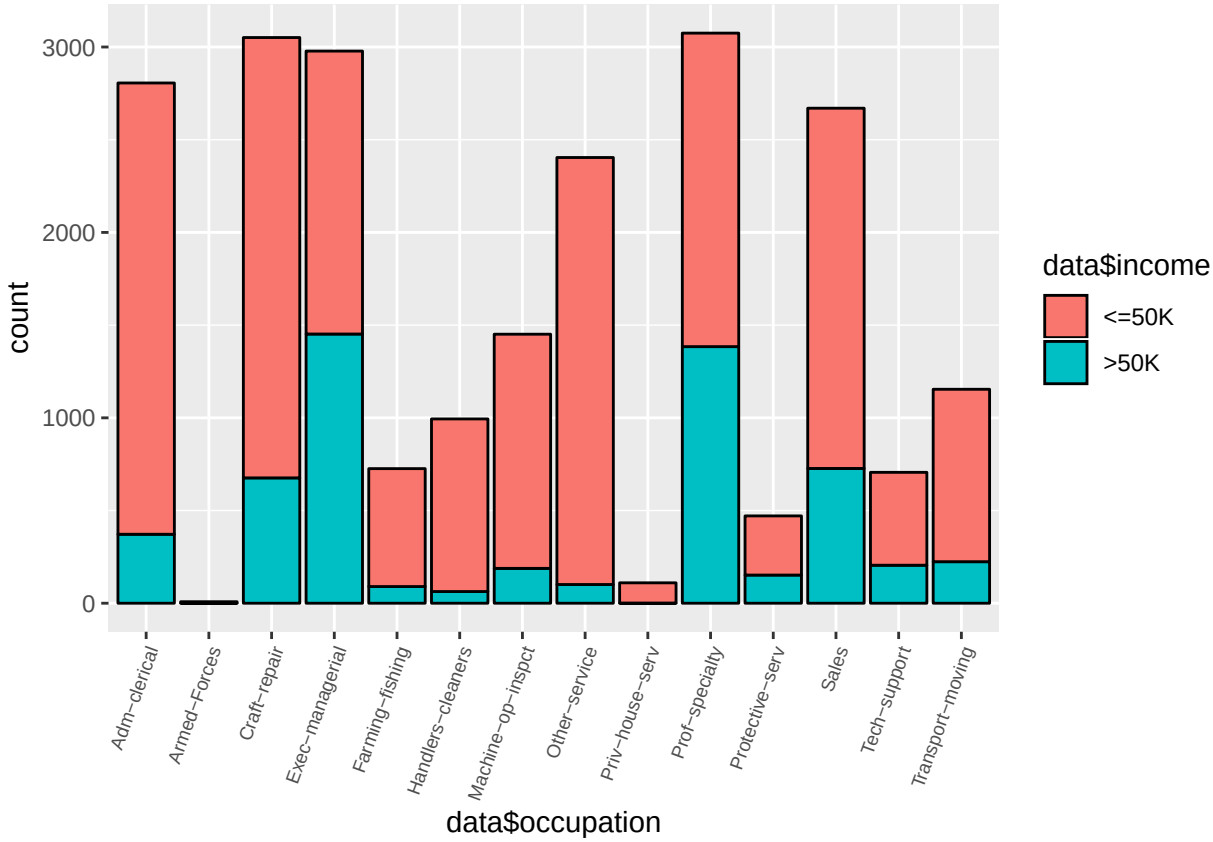

```
ggplot(data, aes(x=data$marital.status,fill=data$income)) + geom_bar(position = "stack", color = "black"
```



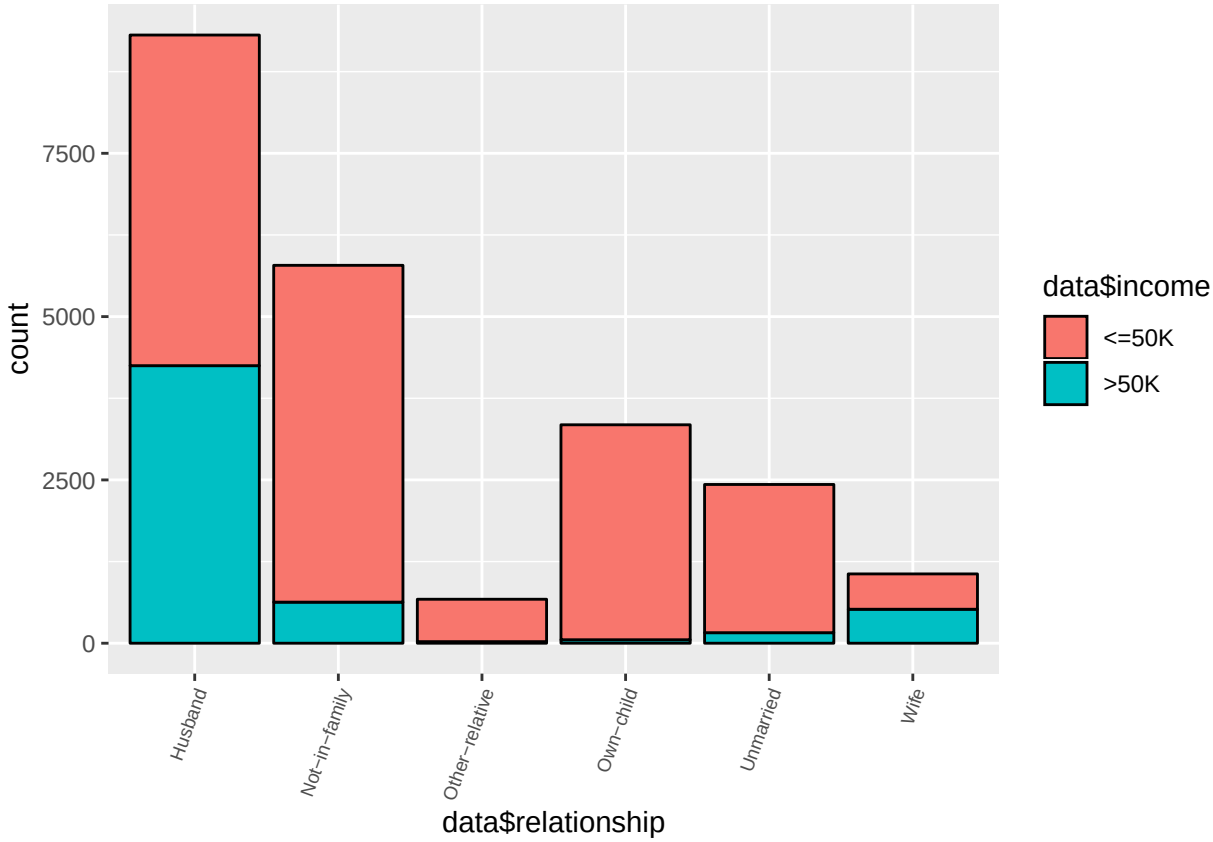

```
ggplot(data, aes(x=data$sex,fill=data$income)) + geom_bar(position = "stack", color = "black") + theme(a
```



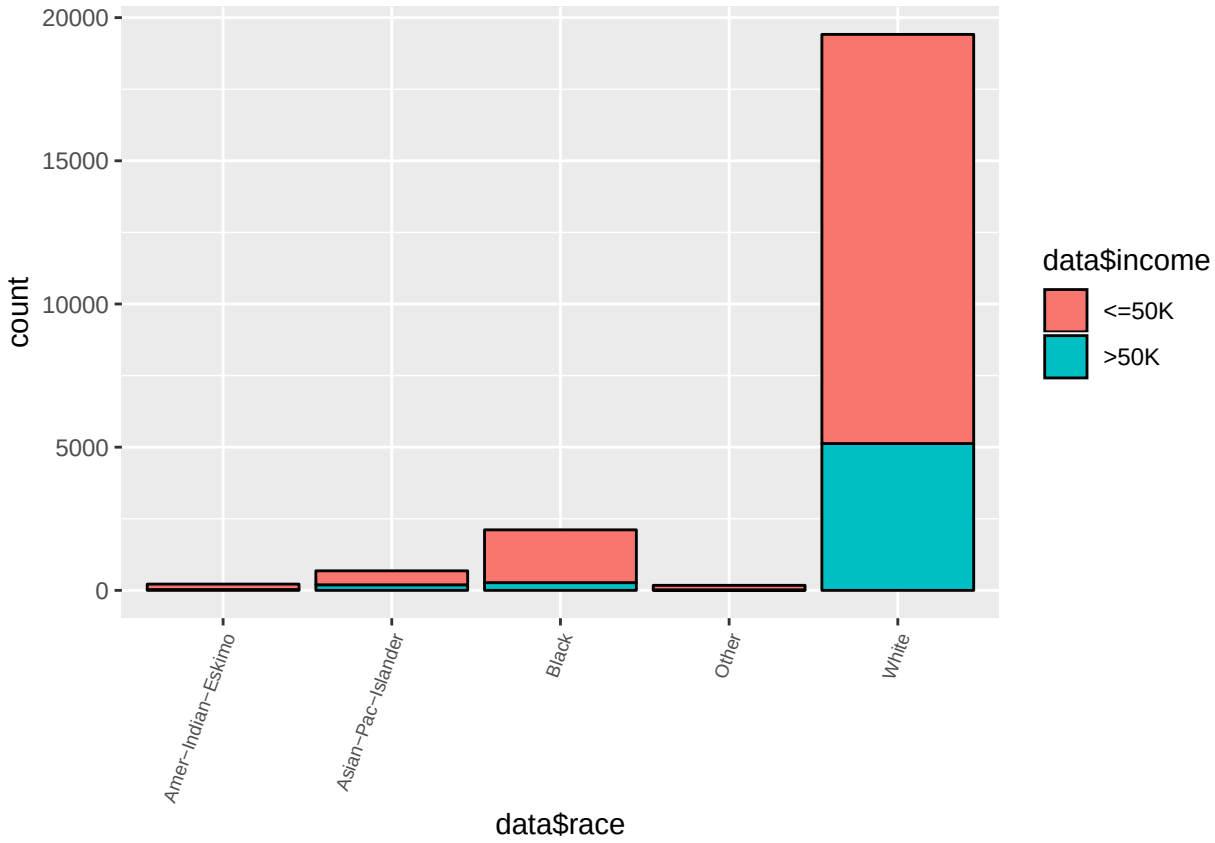

```
ggplot(data, aes(x=data$occupation,fill=data$income)) + geom_bar(position = "stack", color = "black") +
```




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




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


20000
15000
10000
5000
0
count
Amer−Indian−Eskimo
Asian−Pac−Islander
Black
Other
White
data$race
data$income
<=50K
>50K