Project: Adult Census Income

Adult Data Set also known as Census Income from the UCI Machine Learning Repository is used for a classification problem. The goal is to determine whether a person makes more or less than $50K a year using multiple machine learning algorithms.

### Exploratory Data Analysis

library(readr)  
adult <- read\_csv("C:/Users/lamiae/Downloads/adult.data")  
dim(adult)

## [1] 32560 15

The dataset has 32560 observations and 15 columns.

#### Input/output variables

Changing the column names of the adult dataset and checking their data format :

colnames(adult) <- c("Age", "WorkClass","FinalWeight", "Education", "EdYear", "MaritalStatus",  
 "Occupation", "Relationship", "Race", "Gender","CapitalGain",  
 "CapitalLoss","HoursperWeek","NativeCountry","Income")  
sapply(adult,class)

## Age WorkClass FinalWeight Education EdYear   
## "numeric" "character" "numeric" "character" "numeric"   
## MaritalStatus Occupation Relationship Race Gender   
## "character" "character" "character" "character" "character"   
## CapitalGain CapitalLoss HoursperWeek NativeCountry Income   
## "numeric" "numeric" "numeric" "character" "character"

* Age: Age of the individual
  + continuous
* WorkClass: Employment status of the individual
  + Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
* FinalWeight: Census number representing the individual (might not be used for the analyis)
  + continuous.
* Education: Highest level of education the individual achieved
  + Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
* EdYear: Number of education years
  + continuous.
* MaritalStatus: Marital status of the individual
  + Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
* 0ccupation: Type of occupation of the individual
  + Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
* Relationship: relation of the individual to others
  + Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
* Race: Race of the individual
  + White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
* Gender: Gender of the individual
  + Female, Male.
* CapitalGain: : Capital gain for the individual
  + continuous.
* CapitalLoss: Capital loss for the individual
  + continuous.
* HoursperWeek: number of hours the individual works per week
  + continuous.
* NativeCountry: Country of origin
  + United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
* Income: If the individual makes more or less than $50,000 annually. This is the response variable
  + <=50k, >50k

The summary of the variables is given below:

summary(adult)

## Age WorkClass FinalWeight Education   
## Min. :17.00 Length:32560 Min. : 12285 Length:32560   
## 1st Qu.:28.00 Class :character 1st Qu.: 117832 Class :character   
## Median :37.00 Mode :character Median : 178363 Mode :character   
## Mean :38.58 Mean : 189782   
## 3rd Qu.:48.00 3rd Qu.: 237055   
## Max. :90.00 Max. :1484705   
## EdYear MaritalStatus Occupation Relationship   
## Min. : 1.00 Length:32560 Length:32560 Length:32560   
## 1st Qu.: 9.00 Class :character Class :character Class :character   
## Median :10.00 Mode :character Mode :character Mode :character   
## Mean :10.08   
## 3rd Qu.:12.00   
## Max. :16.00   
## Race Gender CapitalGain CapitalLoss   
## Length:32560 Length:32560 Min. : 0 Min. : 0.00   
## Class :character Class :character 1st Qu.: 0 1st Qu.: 0.00   
## Mode :character Mode :character Median : 0 Median : 0.00   
## Mean : 1078 Mean : 87.31   
## 3rd Qu.: 0 3rd Qu.: 0.00   
## Max. :99999 Max. :4356.00   
## HoursperWeek NativeCountry Income   
## Min. : 1.00 Length:32560 Length:32560   
## 1st Qu.:40.00 Class :character Class :character   
## Median :40.00 Mode :character Mode :character   
## Mean :40.44   
## 3rd Qu.:45.00   
## Max. :99.00

No missing values or NA exist in the dataset.

Looking at the distribution of the response variable income, we find 76% to 24% split for less than 50K and more than 50K

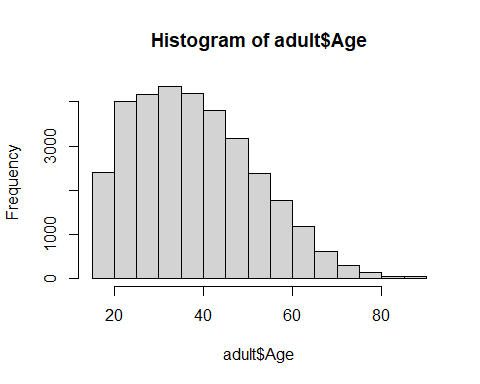
# distribution of income   
cbind(freq= table(adult$Income), perecentage = prop.table(table(adult$Income))\*100)

## freq perecentage  
## <=50K 24719 75.9183  
## >50K 7841 24.0817

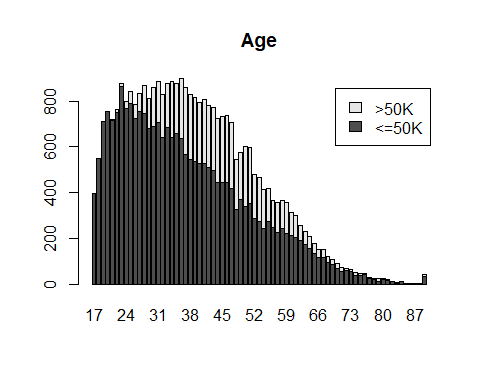
### Data Visualization

Looking at the variables ditributions

# Age distribution   
hist(adult$Age)

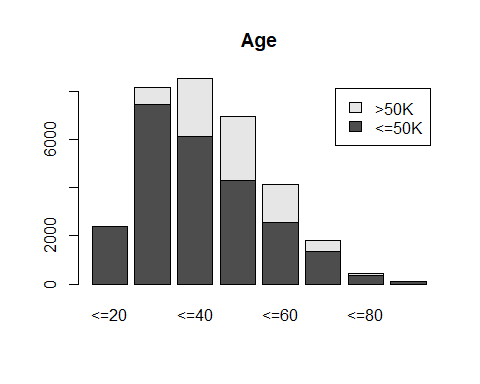


barplot(table(adult$Income, adult$Age) , main=names(adult)[1],  
 legend.text=unique(adult$Income))



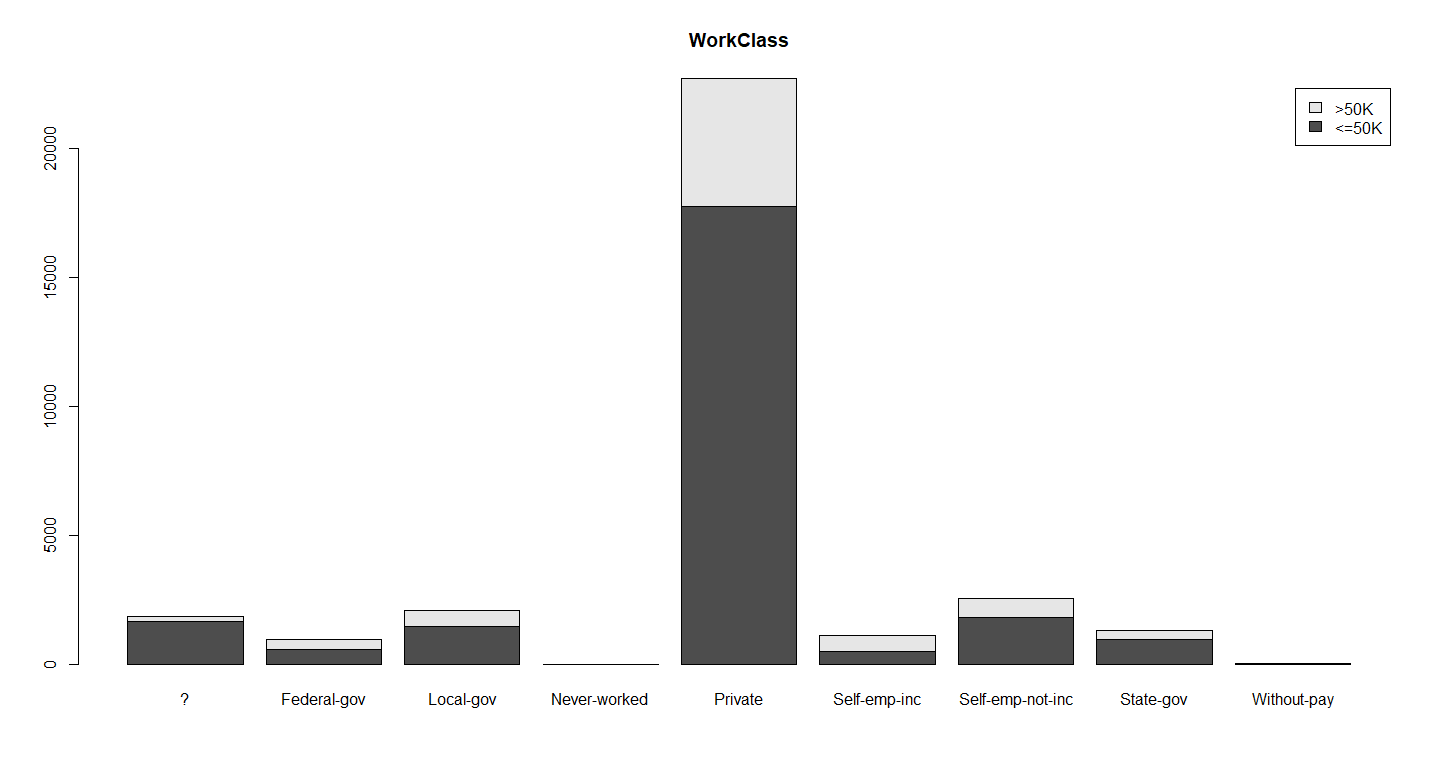
A better way to use the Age attribute in the analysis of income is by creating some age groups

adult$Age<- as.factor(ifelse(adult$Age <=20, '<=20',  
 ifelse(adult$Age <=30, '<=30',  
 ifelse(adult$Age <=40, '<=40',  
 ifelse(adult$Age <=50, '<=50',  
 ifelse(adult$Age <=60, '<=60',  
 ifelse(adult$Age <=70, '<=70',  
 ifelse(adult$Age <=80, '<=80','<=90'))))))))  
barplot(table(adult$Income, adult$Age) , main=names(adult)[1],  
 legend.text=unique(adult$Income))



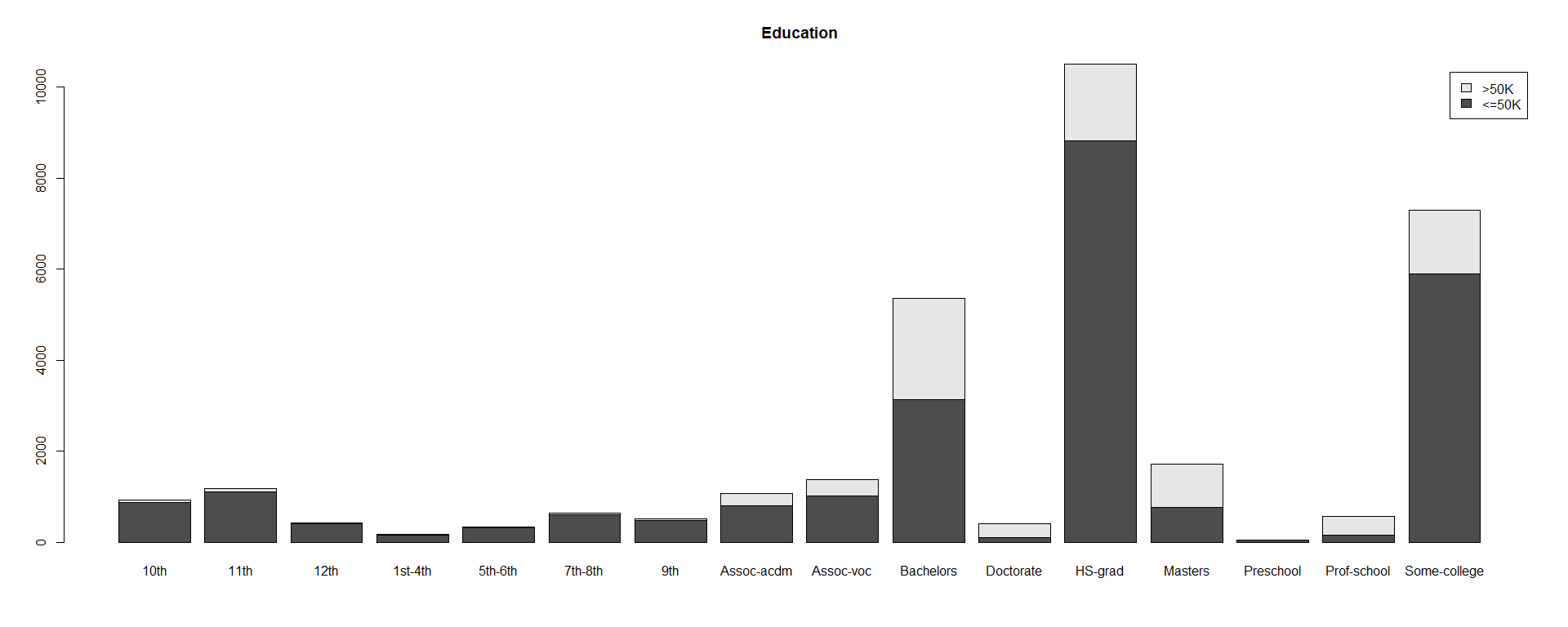
Looking at this bar plot, we can see that there is a variance between income >50K and <=50K and age of the individual. Individuals aged less than 20 and from 72 to 90 have an income lower than $5000, whereas individuals in their late 20’s, up to their 60’s have significantly different ratios of >50K and <=50K.

# Work Class   
barplot(table(adult$Income, adult$WorkClass) , main=names(adult)[2],  
 legend.text=unique(adult$Income))



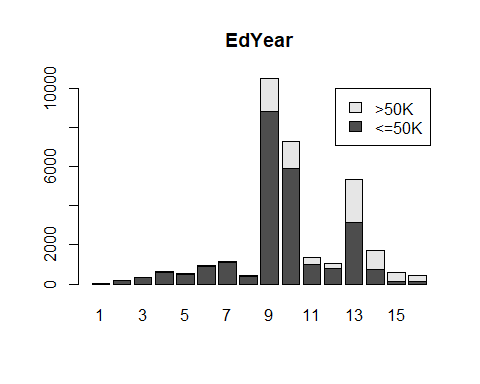
The ? means unknown work class. From the plot,we can see that most of the individuals work in the private sector. All the work classes have similar ratio of income except the self-emp-inc which represents individuals who own their own companies and have high incomes.

# Education   
barplot(table(adult$Income, adult$Education) , main=names(adult)[4],  
 legend.text=unique(adult$Income))

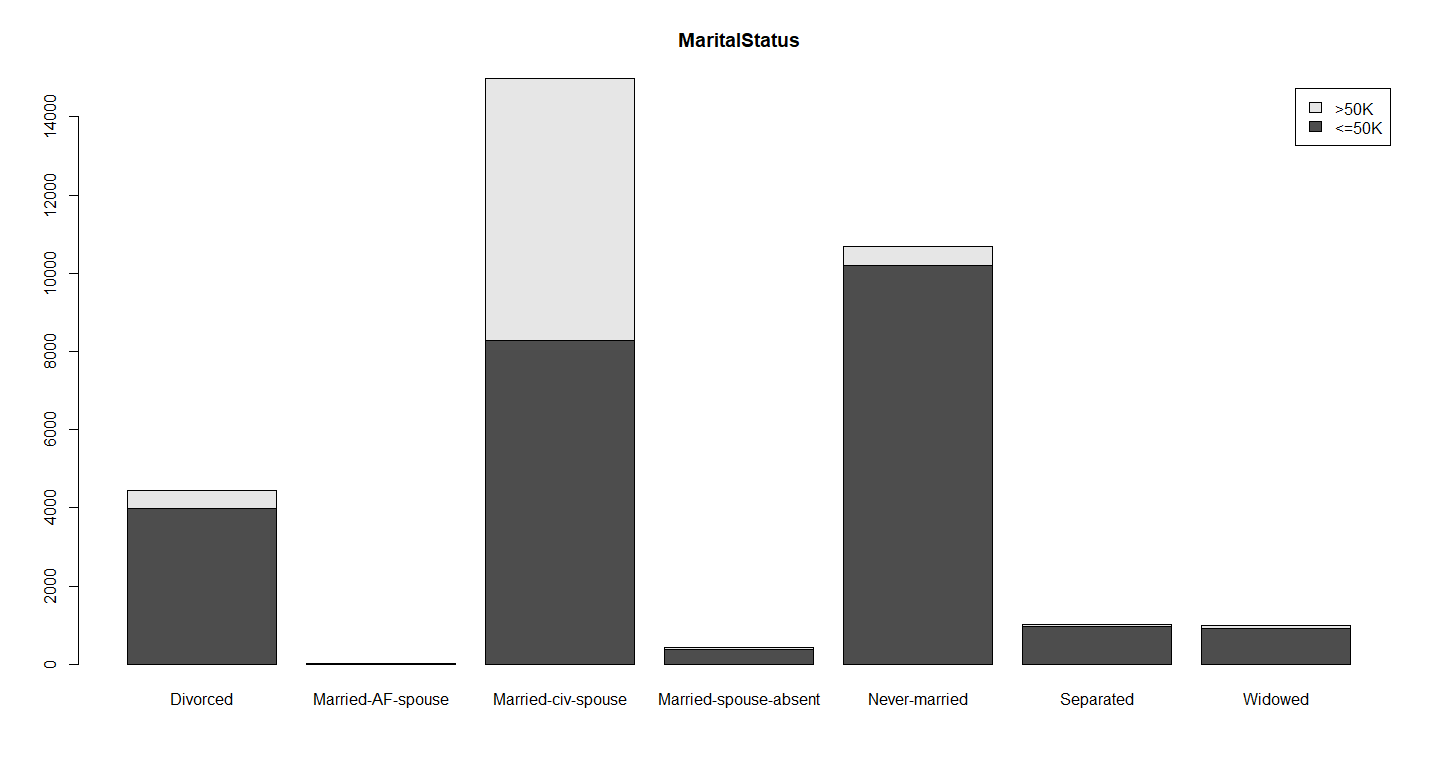


Looking at the distribution of the education level of the individuals in this census, most of them have at most a high school degree and only a few have a doctorate degree. Clearly, the higher the education the higher the income is >50K

# Education Years   
barplot(table(adult$Income, adult$EdYear) , main=names(adult)[5],  
 legend.text=unique(adult$Income))

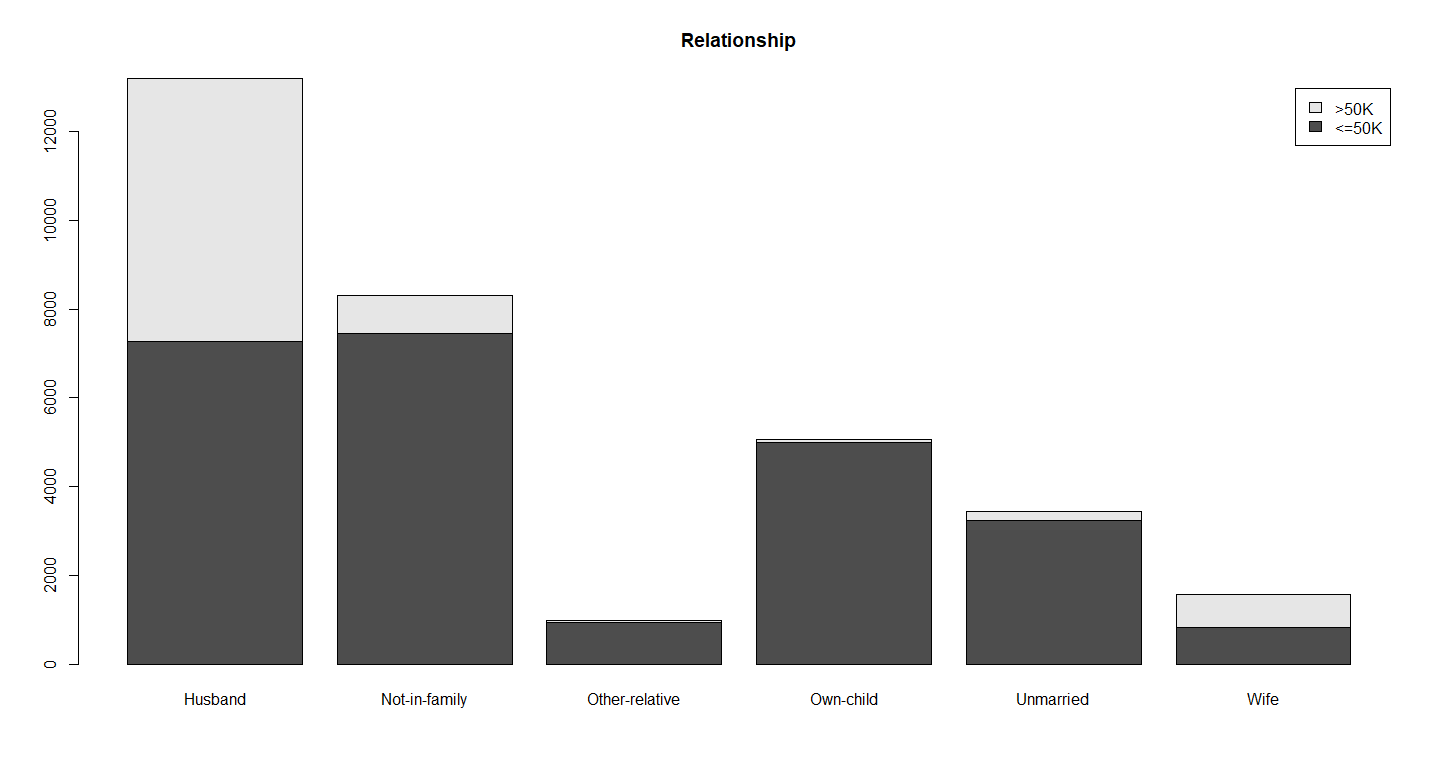
 Education and EdYear have similar distributions because EdYear is the number form of Education which represent the highest level of education achieved by the individual. Therefore, only one of these variables will be used in the analysis.

barplot(table(adult$Income, adult$MaritalStatus) , main=names(adult)[6],  
 legend.text=unique(adult$Income))



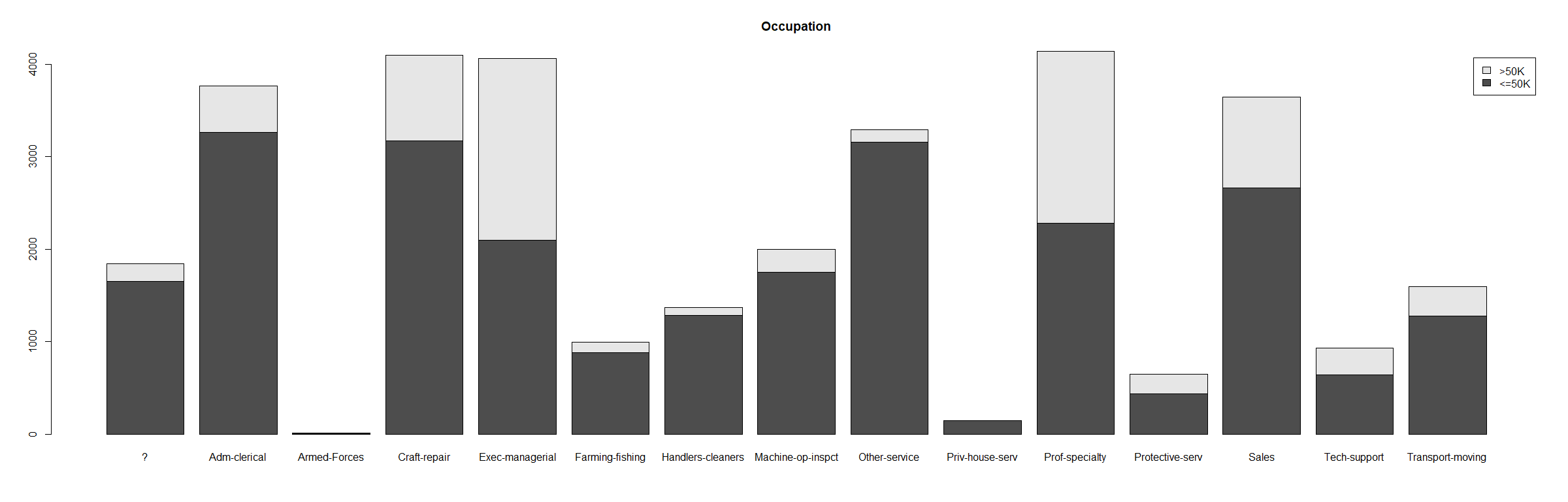
The plot shows that most of the individuals in the dataset are married followed by Nevermarried. The ratio of <=50K is very high for most of the marital status except for married individuals who can achieve an income greater than 50K.

barplot(table(adult$Income, adult$Relationship) , main=names(adult)[8],  
 legend.text=unique(adult$Income))



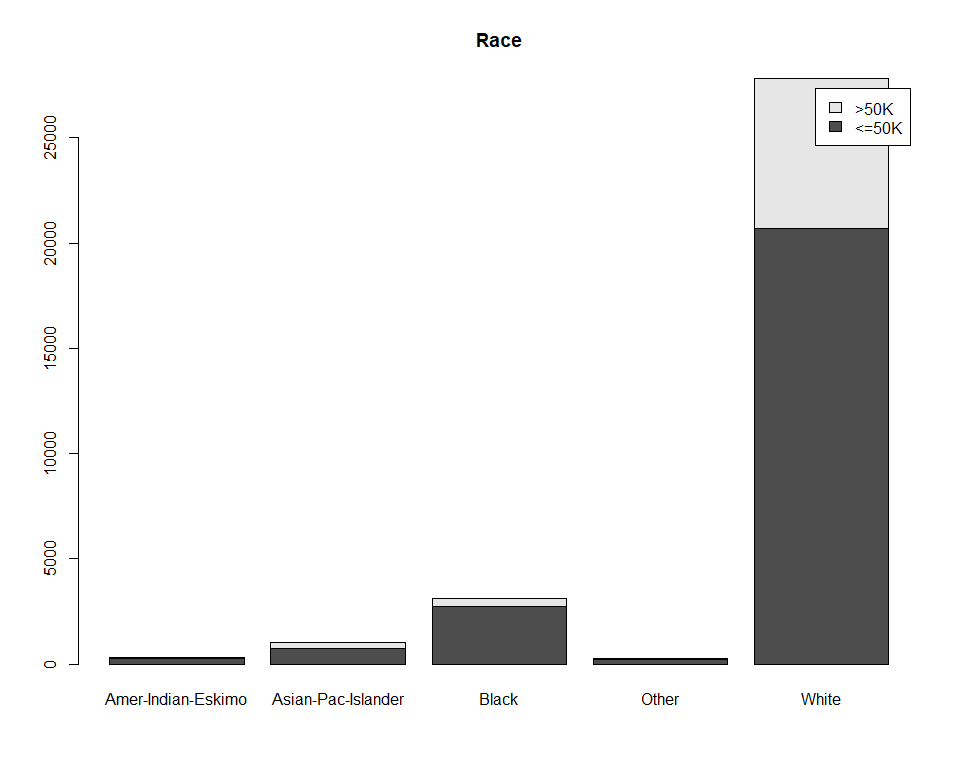
The variable relationship provides similar information as the MaritalStatus variable. Therefore it will not be included in the analysis of income.

barplot(table(adult$Income, adult$Occupation) , main=names(adult)[7],  
 legend.text=unique(adult$Income))



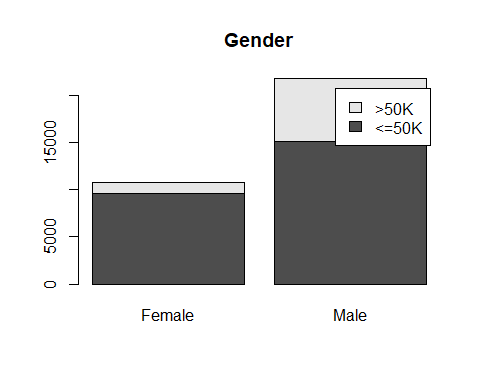
Looking at the occupation plot,The income vary significantly with different occupations. For instance, Exec-managerial and prof-specialty have higher percentages of incomes greater than 50K, whereas Farming-fishing, Other-service and Handlers-cleaners have significantly lower income than the other occupations.

# race  
barplot(table(adult$Income, adult$Race) , main=names(adult)[9],  
 legend.text=unique(adult$Income))

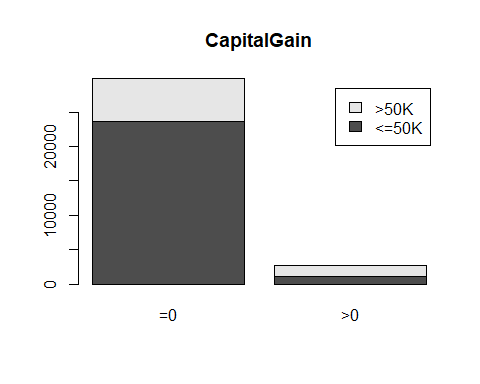


The plot shows that white is the largest represented race in the dataset followed by black and that white has the highest income greater than $50K followed by asian than black. This variable is not used for the analysis because of the disproportion of the distribution of the race attribute.

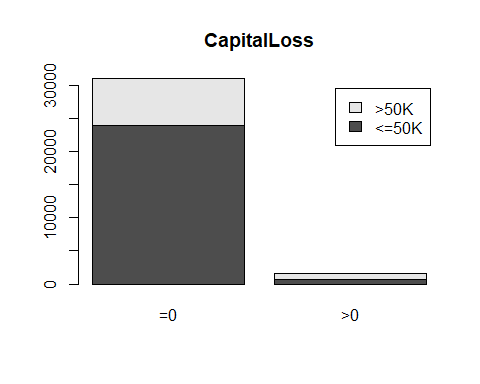
# Gender  
barplot(table(adult$Income, adult$Gender) , main=names(adult)[10],  
 legend.text=unique(adult$Income))

 Male individuals are represented almost twice more than females in the dataset and their percentage of high income is greater than that of females. This variation can be important for the income analysis.

#capital gain/loss  
adult$CapitalGain <- as.factor(ifelse(adult$CapitalGain >0, '>0', '=0'))  
  
barplot(table(adult$Income, adult$CapitalGain) , main=names(adult)[11],  
 legend.text=unique(adult$Income))

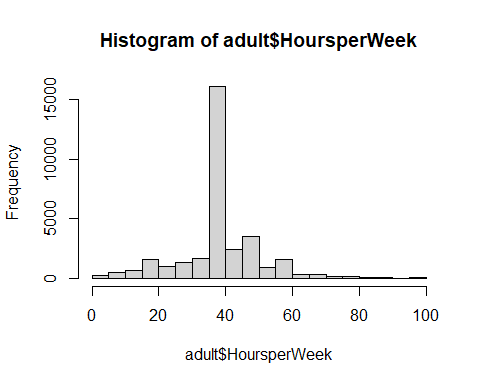


adult$CapitalLoss <- as.factor(ifelse(adult$CapitalLoss >0, '>0', '=0'))  
barplot(table(adult$Income, adult$CapitalLoss) , main=names(adult)[12],  
 legend.text=unique(adult$Income))

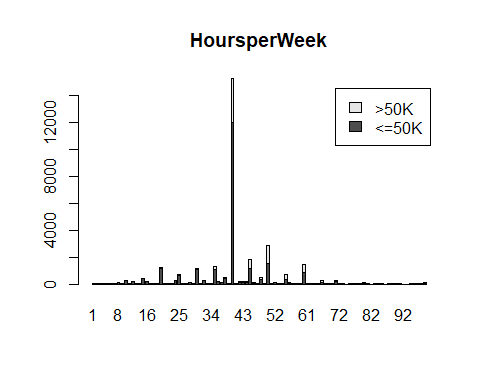


Both variables have high zero values, therefore Capital gain and Capital Loss will not be used in the analysis.

hist(adult$HoursperWeek)

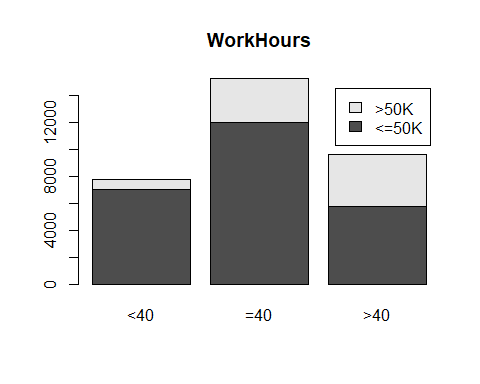


barplot(table(adult$Income, adult$HoursperWeek) , main=names(adult)[13],  
 legend.text=unique(adult$Income))



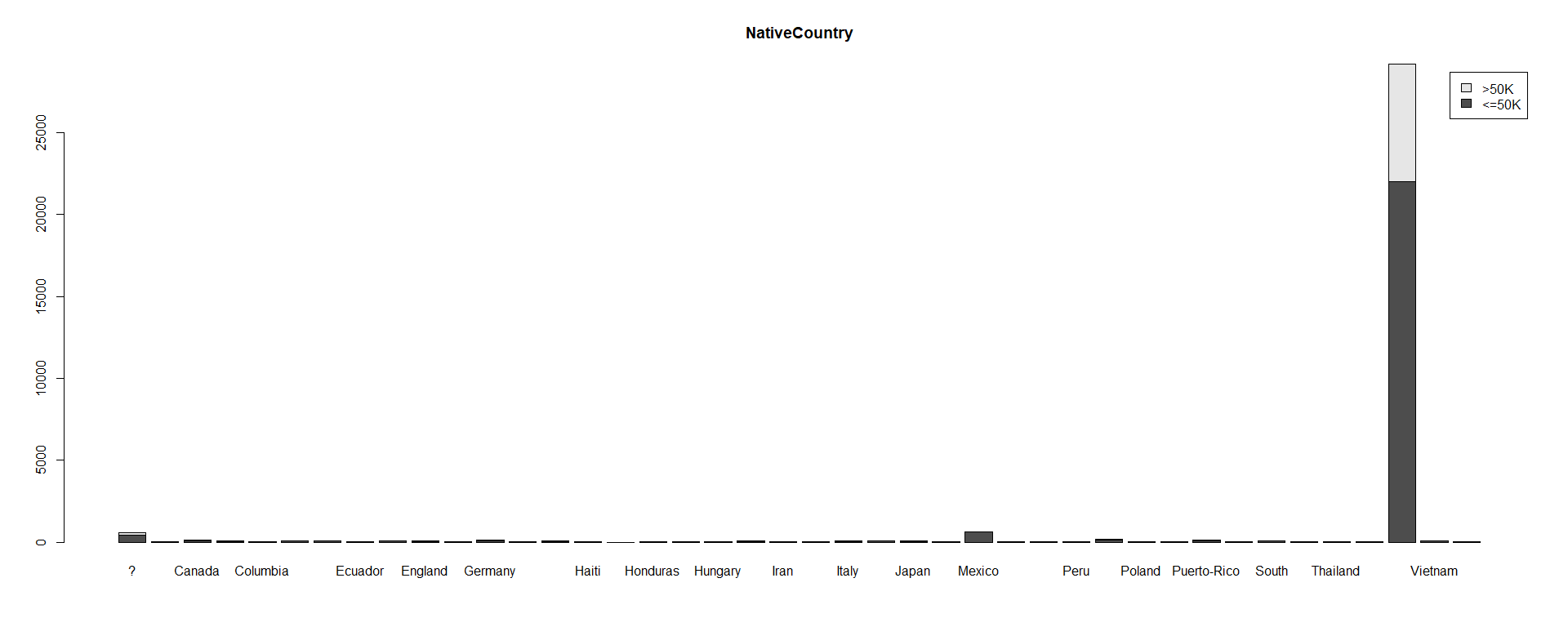
For better visualization and understanding of the hours per week variable, let’s create a new variable with 3 values: \* >40 : Individual works less than 40 hours per week \* =40 : Individual works exactly 40 hours per week \* <40 : Individual works more than 40 hours per week

adult$WorkHours <- as.factor(ifelse(adult$HoursperWeek < 40, '<40',  
 ifelse(adult$HoursperWeek == 40, '=40', '>40')))  
barplot(table(adult$Income, adult$WorkHours) , main=names(adult)[16],  
 legend.text=unique(adult$Income))



Most individuals in the dataset work exactly 40 hours per week. From this plot, we can see that the more hours an individual works the higher the proportion of income <50K is.

barplot(table(adult$Income, adult$NativeCountry) , main=names(adult)[14],  
 legend.text=unique(adult$Income))



The NativeCountry variable is not going to be included in the analysis because of the disproportional distribution of its values.

Therefore, the unnecessary variables are removed from the dataset.

# removing some attributes  
adult<-adult[,c(-3,-5,-8,-9,-11,-12,-13,-14)]  
colnames(adult)

## [1] "Age" "WorkClass" "Education" "MaritalStatus"  
## [5] "Occupation" "Gender" "Income" "WorkHours"

The variables listed above are the only ones used in the analysis.

### Classification Analysis

The algorithms used in this classification project are linear discriminant analysis, logistic regression, support vector machines and Classification trees.

#### Data Partition

Splitting the adult dataset to 75% training and 25% testing datasets

#Splitting datasets  
library(caret)  
set.seed(123)  
inTrain = createDataPartition(adult$Income, p = .75, list = FALSE)  
train.data=adult[inTrain,]  
test.data=adult[-inTrain,]

#### Linear Discriminant Analysis

library(MASS)  
  
# fit lda model  
model.lda<- lda(Income~.,data=train.data)  
  
# Using the testing set to evaluate the classification performance  
predictions.lda = data.frame(predict(model.lda, test.data))  
confusionM.lda<-confusionMatrix(predictions.lda$class,as.factor(test.data$Income))  
#print(confusionM.lda$table)  
print(confusionM.lda)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction <=50K >50K  
## <=50K 5687 847  
## >50K 492 1113  
##   
## Accuracy : 0.8355   
## 95% CI : (0.8272, 0.8435)  
## No Information Rate : 0.7592   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.5204   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.9204   
## Specificity : 0.5679   
## Pos Pred Value : 0.8704   
## Neg Pred Value : 0.6935   
## Prevalence : 0.7592   
## Detection Rate : 0.6987   
## Detection Prevalence : 0.8028   
## Balanced Accuracy : 0.7441   
##   
## 'Positive' Class : <=50K   
##

#### Logistic Regression

library(nnet)  
library(tidyverse)  
# Fit the model  
model.logReg <- nnet::multinom(Income ~., data = train.data)

## # weights: 55 (54 variable)  
## initial value 16927.347296   
## iter 10 value 9500.177168  
## iter 20 value 9226.519386  
## iter 30 value 8809.385213  
## iter 40 value 8648.876548  
## iter 50 value 8536.513125  
## iter 60 value 8526.588685  
## iter 70 value 8526.500473  
## final value 8526.498085   
## converged

# Summarize the model  
summary(model.logReg)

## Call:  
## nnet::multinom(formula = Income ~ ., data = train.data)  
##   
## Coefficients:  
## Values Std. Err.  
## (Intercept) -8.09566393 7.383652e-01  
## Age<=30 2.46895645 7.133655e-01  
## Age<=40 3.34681302 7.129417e-01  
## Age<=50 3.70792054 7.132396e-01  
## Age<=60 3.82915096 7.137792e-01  
## Age<=70 3.42304428 7.164342e-01  
## Age<=80 3.28544803 7.335391e-01  
## Age<=90 2.92320109 8.053970e-01  
## WorkClassFederal-gov 1.06057169 1.472151e-01  
## WorkClassLocal-gov 0.42214807 1.370483e-01  
## WorkClassNever-worked -8.02904599 1.512237e-06  
## WorkClassPrivate 0.65440877 1.200437e-01  
## WorkClassSelf-emp-inc 0.97489446 1.442218e-01  
## WorkClassSelf-emp-not-inc 0.27914791 1.319389e-01  
## WorkClassState-gov 0.26914190 1.488576e-01  
## WorkClassWithout-pay -10.43153104 4.936937e-06  
## Education11th -0.10123843 2.357630e-01  
## Education12th 0.41163197 2.820481e-01  
## Education1st-4th -1.30591703 6.214079e-01  
## Education5th-6th -0.49128764 3.272563e-01  
## Education7th-8th -0.49048976 2.460904e-01  
## Education9th -0.56841939 2.974783e-01  
## EducationAssoc-acdm 1.19779376 1.871047e-01  
## EducationAssoc-voc 1.20742151 1.791449e-01  
## EducationBachelors 1.77180209 1.659410e-01  
## EducationDoctorate 2.83102168 2.297846e-01  
## EducationHS-grad 0.66780176 1.612825e-01  
## EducationMasters 2.15075182 1.780975e-01  
## EducationPreschool -15.26303603 NaN  
## EducationProf-school 2.86561908 2.135169e-01  
## EducationSome-college 0.98340966 1.639149e-01  
## MaritalStatusMarried-AF-spouse 3.10526309 5.865234e-01  
## MaritalStatusMarried-civ-spouse 2.17379187 7.184913e-02  
## MaritalStatusMarried-spouse-absent -0.09334579 2.439811e-01  
## MaritalStatusNever-married -0.18053372 8.856276e-02  
## MaritalStatusSeparated -0.06547489 1.675898e-01  
## MaritalStatusWidowed 0.34070237 1.551238e-01  
## OccupationAdm-clerical -0.19823282 1.142585e-01  
## OccupationArmed-Forces -0.85368916 1.367211e+00  
## OccupationCraft-repair -0.31266774 1.082067e-01  
## OccupationExec-managerial 0.48634971 1.058189e-01  
## OccupationFarming-fishing -1.20184064 1.566444e-01  
## OccupationHandlers-cleaners -1.10663831 1.694670e-01  
## OccupationMachine-op-inspct -0.65855452 1.279039e-01  
## OccupationOther-service -1.20217129 1.425878e-01  
## OccupationPriv-house-serv -2.10142521 9.726523e-01  
## OccupationProf-specialty 0.24239610 1.094353e-01  
## OccupationProtective-serv 0.32301804 1.499551e-01  
## OccupationSales 0.01510097 1.095677e-01  
## OccupationTech-support 0.28953572 1.361863e-01  
## OccupationTransport-moving -0.49239910 1.242165e-01  
## GenderMale 0.16872124 5.692601e-02  
## WorkHours=40 0.48311487 6.423222e-02  
## WorkHours>40 0.92455410 6.653260e-02  
##   
## Residual Deviance: 17053   
## AIC: 17159

# Make predictions  
predicted.logReg <- model.logReg %>% predict(test.data)  
# Model accuracy  
mean(predicted.logReg == test.data$Income)

## [1] 0.8343777

# confusion matrix  
confusionM.logReg<-confusionMatrix(predicted.logReg, as.factor(test.data$Income))  
print(confusionM.logReg)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction <=50K >50K  
## <=50K 5716 885  
## >50K 463 1075  
##   
## Accuracy : 0.8344   
## 95% CI : (0.8261, 0.8424)  
## No Information Rate : 0.7592   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.5111   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.9251   
## Specificity : 0.5485   
## Pos Pred Value : 0.8659   
## Neg Pred Value : 0.6990   
## Prevalence : 0.7592   
## Detection Rate : 0.7023   
## Detection Prevalence : 0.8110   
## Balanced Accuracy : 0.7368   
##   
## 'Positive' Class : <=50K   
##

#### Support Vector Machines

library(e1071)  
  
svmfit = svm(as.factor(Income) ~ ., data = test.data, kernel = "radial", cost = 10, scale = FALSE)  
print(svmfit)

##   
## Call:  
## svm(formula = as.factor(Income) ~ ., data = test.data, kernel = "radial",   
## cost = 10, scale = FALSE)  
##   
##   
## Parameters:  
## SVM-Type: C-classification   
## SVM-Kernel: radial   
## cost: 10   
##   
## Number of Support Vectors: 3115

test.svm<-predict(svmfit,data=test.data)  
ConfusionM.svm<-confusionMatrix(test.svm,as.factor(test.data$Income))  
print(ConfusionM.svm)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction <=50K >50K  
## <=50K 5796 917  
## >50K 383 1043  
##   
## Accuracy : 0.8403   
## 95% CI : (0.8321, 0.8482)  
## No Information Rate : 0.7592   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.5184   
##   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.9380   
## Specificity : 0.5321   
## Pos Pred Value : 0.8634   
## Neg Pred Value : 0.7314   
## Prevalence : 0.7592   
## Detection Rate : 0.7121   
## Detection Prevalence : 0.8248   
## Balanced Accuracy : 0.7351   
##   
## 'Positive' Class : <=50K   
##

#### Naive Bayes Classifier

NBclassfier=naiveBayes(as.factor(Income)~., data=train.data)  
test.NB=predict(NBclassfier, newdata = test.data, type = "class")  
confusionM.NB<-confusionMatrix(test.NB,as.factor(test.data$Income))  
print(confusionM.NB)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction <=50K >50K  
## <=50K 5343 623  
## >50K 836 1337  
##   
## Accuracy : 0.8207   
## 95% CI : (0.8122, 0.829)  
## No Information Rate : 0.7592   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.5273   
##   
## Mcnemar's Test P-Value : 2.854e-08   
##   
## Sensitivity : 0.8647   
## Specificity : 0.6821   
## Pos Pred Value : 0.8956   
## Neg Pred Value : 0.6153   
## Prevalence : 0.7592   
## Detection Rate : 0.6565   
## Detection Prevalence : 0.7330   
## Balanced Accuracy : 0.7734   
##   
## 'Positive' Class : <=50K   
##

### Conclusion

In this analysis census data is used to analyze the income of adults based on the following characteristics: individual’s age, gender, work class, education, marital status, occupation, and work hours.  
The methods used are linear discriminant, logistic regression, support vector machines, and naive bayes classifier to classify the income of individuals into two categories: less than 50K or greater than or equal to 50K. The accuracy of the models used in this analysis is as follows:

# Model performance metrics  
data.frame(  
 Model = c("LDA", "Log Reg", "SVM", "NBC"),  
 Accuracy = c(confusionM.lda$overall[1], confusionM.logReg$overall[1], ConfusionM.svm$overall[1], confusionM.NB$overall[1])  
)

## Model Accuracy  
## 1 LDA 0.8354835  
## 2 Log Reg 0.8343777  
## 3 SVM 0.8402752  
## 4 NBC 0.8207396

From these results, it can be seen that all the models have a high accuracy in classifying the adult’s income. However, support vector machines is more accurate. For possible improvements, parameter tuning can be applied in order to achieve a higher accuracy, also other methods can be used to model the response variable with less attributes.