Adult Income Data Project

CKME 136 Final Project Shabbir Yousuf Ali #[syali@ryerson.ca](mailto:syali@ryerson.ca) #<https://github.com/shabbiryousufali/CKME136> Winter 2019

1. Load requied libraries. Install package install.packages(“caret”) Install package install.packages(“corrplot”) Install package install.packages(‘Boruta’)

library(ggplot2)  
library(corrplot)

## Warning: package 'corrplot' was built under R version 3.4.4

## corrplot 0.84 loaded

library(Boruta)

## Warning: package 'Boruta' was built under R version 3.4.4

## Loading required package: ranger

## Warning: package 'ranger' was built under R version 3.4.4

library(randomForest)

## Warning: package 'randomForest' was built under R version 3.4.4

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:ranger':  
##   
## importance

## The following object is masked from 'package:ggplot2':  
##   
## margin

library(ROCR)

## Warning: package 'ROCR' was built under R version 3.4.4

## Loading required package: gplots

## Warning: package 'gplots' was built under R version 3.4.4

##   
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':  
##   
## lowess

library(caret)

## Warning: package 'caret' was built under R version 3.4.4

## Loading required package: lattice

library(rpart)

## Warning: package 'rpart' was built under R version 3.4.4

1. Load data.

setwd("C:/Ryerson/ckme136/project/rawdata")  
loc<-getwd()  
censusdata <- read.csv(file="adult.data",header=TRUE,sep=",", na.string = "?")  
  
#Add header to the columns  
names(censusdata) <- c('age',  
 'workclass',  
 'fnlwgt',  
 'education',  
 'educationnum',  
 'maritalstatus',  
 'occupation',  
 'relationship',  
 'race',  
 'sex',  
 'capitalgain',  
 'capitalloss',  
 'hoursperweek',  
 'nativecountry',  
 'income')

2.1. Split the data into train and test data.

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

1. Display dimensions, summary of data, names and overall structure of the data.

data <- training  
dim(data)

## [1] 24421 15

nrow(data)

## [1] 24421

ncol(data)

## [1] 15

dim(testing)

## [1] 8139 15

summary(data)

## age workclass fnlwgt   
## Min. :17.0 Private :17019 Min. : 12285   
## 1st Qu.:28.0 Self-emp-not-inc: 1906 1st Qu.: 117849   
## Median :37.0 Local-gov : 1563 Median : 178272   
## Mean :38.6 ? : 1378 Mean : 189664   
## 3rd Qu.:48.0 State-gov : 960 3rd Qu.: 236696   
## Max. :90.0 Self-emp-inc : 859 Max. :1484705   
## (Other) : 736   
## education educationnum maritalstatus   
## HS-grad :7844 Min. : 1.00 Divorced : 3362   
## Some-college:5508 1st Qu.: 9.00 Married-AF-spouse : 17   
## Bachelors :4024 Median :10.00 Married-civ-spouse :11184   
## Masters :1287 Mean :10.09 Married-spouse-absent: 310   
## Assoc-voc :1047 3rd Qu.:12.00 Never-married : 8015   
## 11th : 891 Max. :16.00 Separated : 763   
## (Other) :3820 Widowed : 770   
## occupation relationship   
## Prof-specialty :3116 Husband :9876   
## Craft-repair :3066 Not-in-family :6241   
## Exec-managerial:3042 Other-relative: 748   
## Adm-clerical :2853 Own-child :3818   
## Sales :2715 Unmarried :2585   
## Other-service :2479 Wife :1153   
## (Other) :7150   
## race sex capitalgain   
## Amer-Indian-Eskimo: 226 Female: 8143 Min. : 0   
## Asian-Pac-Islander: 769 Male :16278 1st Qu.: 0   
## Black : 2348 Median : 0   
## Other : 202 Mean : 1090   
## White :20876 3rd Qu.: 0   
## Max. :99999   
##   
## capitalloss hoursperweek nativecountry income   
## Min. : 0.00 Min. : 1.0 United-States:21883 <=50K:18540   
## 1st Qu.: 0.00 1st Qu.:40.0 Mexico : 489 >50K : 5881   
## Median : 0.00 Median :40.0 ? : 424   
## Mean : 87.23 Mean :40.4 Philippines : 140   
## 3rd Qu.: 0.00 3rd Qu.:45.0 Germany : 106   
## Max. :4356.00 Max. :99.0 Puerto-Rico : 91   
## (Other) : 1288

names(data)

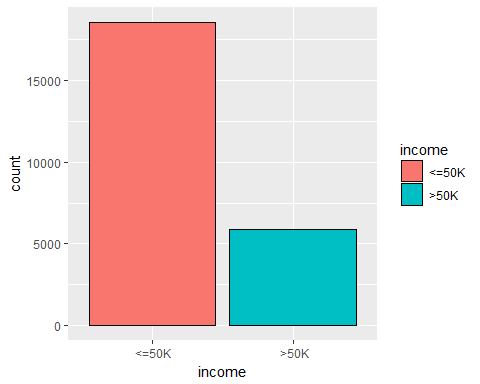
## [1] "age" "workclass" "fnlwgt" "education"   
## [5] "educationnum" "maritalstatus" "occupation" "relationship"   
## [9] "race" "sex" "capitalgain" "capitalloss"   
## [13] "hoursperweek" "nativecountry" "income"

str(data)

## 'data.frame': 24421 obs. of 15 variables:  
## $ age : int 50 38 53 31 42 37 30 23 40 25 ...  
## $ workclass : Factor w/ 9 levels " ?"," Federal-gov",..: 7 5 5 5 5 5 8 5 5 7 ...  
## $ fnlwgt : int 83311 215646 234721 45781 159449 280464 141297 122272 121772 176756 ...  
## $ education : Factor w/ 16 levels " 10th"," 11th",..: 10 12 2 13 10 16 10 10 9 12 ...  
## $ educationnum : int 13 9 7 14 13 10 13 13 11 9 ...  
## $ maritalstatus: Factor w/ 7 levels " Divorced"," Married-AF-spouse",..: 3 1 3 5 3 3 3 5 3 5 ...  
## $ occupation : Factor w/ 15 levels " ?"," Adm-clerical",..: 5 7 7 11 5 5 11 2 4 6 ...  
## $ relationship : Factor w/ 6 levels " Husband"," Not-in-family",..: 1 2 1 2 1 1 1 4 1 4 ...  
## $ race : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 5 5 3 5 5 3 2 5 2 5 ...  
## $ sex : Factor w/ 2 levels " Female"," Male": 2 2 2 1 2 2 2 1 2 2 ...  
## $ capitalgain : int 0 0 0 14084 5178 0 0 0 0 0 ...  
## $ capitalloss : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ hoursperweek : int 13 40 40 50 40 80 40 30 40 35 ...  
## $ nativecountry: Factor w/ 42 levels " ?"," Cambodia",..: 40 40 40 40 40 40 20 40 1 40 ...  
## $ income : Factor w/ 2 levels " <=50K"," >50K": 1 1 1 2 2 2 2 1 2 1 ...

1. Display Class Distributions.

# Use the ggplot to find the income distribution <=50K VS >50K based on the training 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

1. Check and remove the missing values.

cat("Missing values in training set:", sum(is.na(data)), "\n")

## Missing values in training set: 0

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 0  
## fnlwgt 0  
## education 0  
## educationnum 0  
## maritalstatus 0  
## occupation 0  
## relationship 0  
## race 0  
## sex 0  
## capitalgain 0  
## capitalloss 0  
## hoursperweek 0  
## nativecountry 0  
## income 0

nrow(data)

## [1] 24421

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

## [1] 24421

nrow(testing)

## [1] 8139

cat("Missing values in testing set:", sum(is.na(testing)), "\n")

## Missing values in testing set: 0

na\_count1 <-sapply(testing, function(y) sum(length(which(is.na(y)))))  
na\_count1

## age workclass fnlwgt education educationnum   
## 0 0 0 0 0   
## maritalstatus occupation relationship race sex   
## 0 0 0 0 0   
## capitalgain capitalloss hoursperweek nativecountry income   
## 0 0 0 0 0

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

## [1] 8139

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$native.country <- factor(data$nativecountry)

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$native.country <- factor(testingdata$nativecountry)

1. Statistics of Numerical attributes

#find the Min, Max, Mean, Median, 1st and 3rd Quarter of the numerical attributes  
summary(data$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 17.0 28.0 37.0 38.6 48.0 90.0

summary(data$educationnum)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 9.00 10.00 10.09 12.00 16.00

summary(data$capitalgain)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 0 0 1090 0 99999

summary(data$capitalloss)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 0.00 87.23 0.00 4356.00

summary(data$hoursperweek)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.0 40.0 40.0 40.4 45.0 99.0

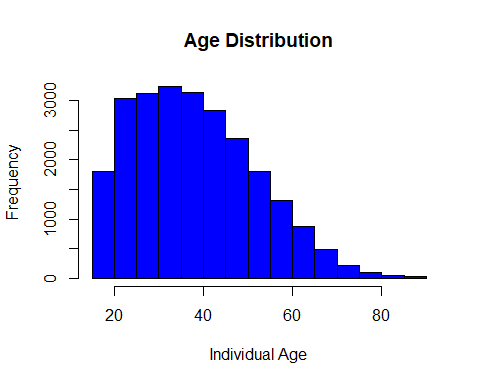
# statistics of numerical attributes  
summary(data$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 17.0 28.0 37.0 38.6 48.0 90.0

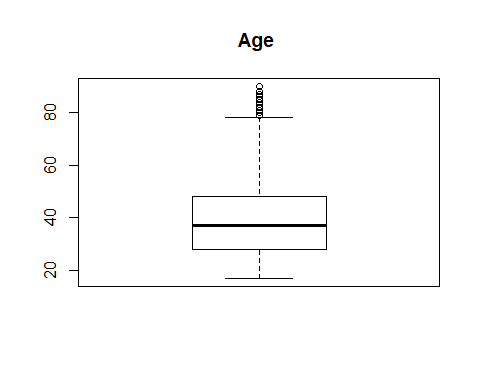
sd(data$age)

## [1] 13.69495

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



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



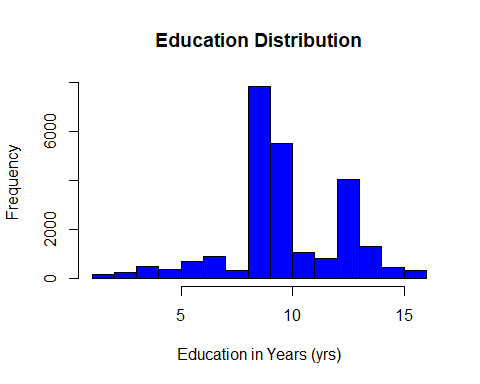
summary(data$education.num)

## Length Class Mode   
## 0 NULL NULL

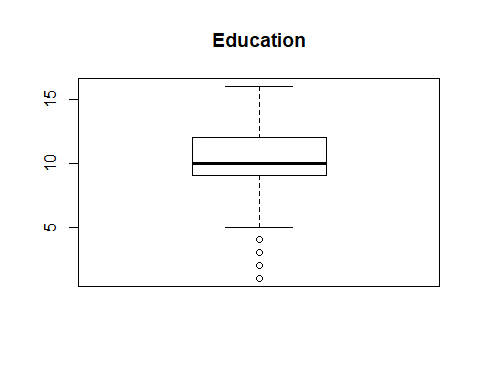
sd(data$education.num)

## [1] NA

hist(data$educationnum,main = "Education Distribution",xlab="Education in Years (yrs)",col = "blue")



boxplot(data$educationnum,main="Education")



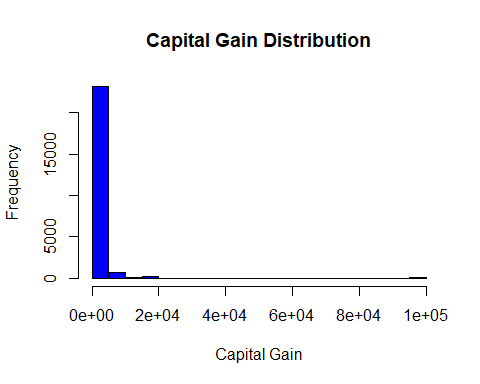
summary(data$capitalgain)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 0 0 1090 0 99999

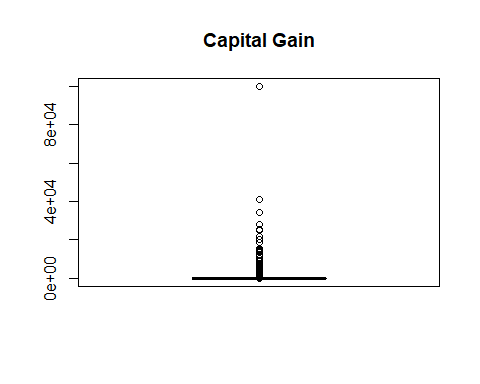
sd(data$capitalgain)

## [1] 7440.626

hist(data$capitalgain,main = "Capital Gain Distribution",xlab="Capital Gain",col = "blue")



boxplot(data$capitalgain,main="Capital Gain")



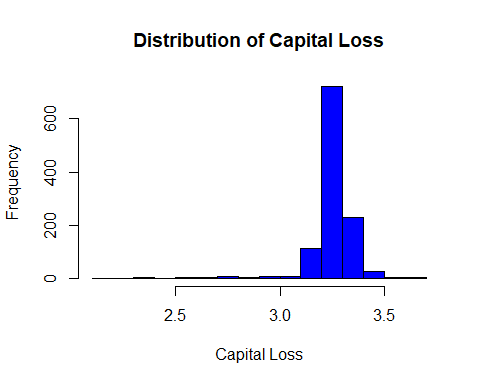
summary(data$capitalloss)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 0.00 87.23 0.00 4356.00

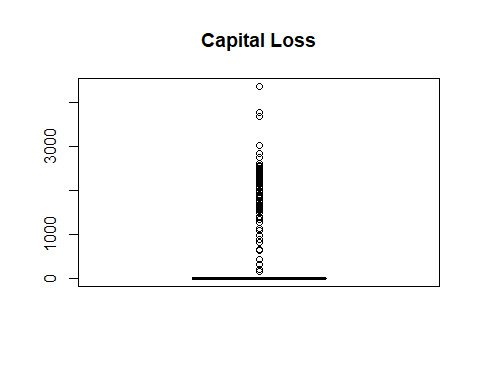
sd(data$capitalloss)

## [1] 403.7928

hist(log10(data$capitalloss),main = "Distribution of Capital Loss",xlab="Capital Loss",col = "blue")



boxplot(data$capitalloss,main="Capital Loss")



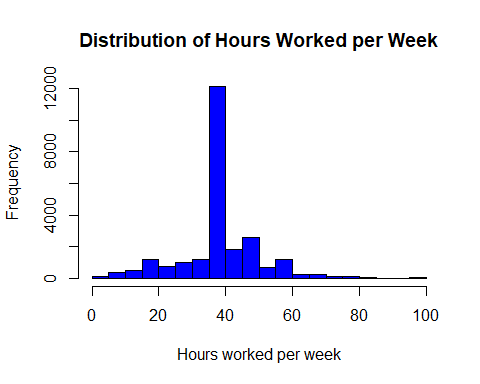
summary(data$hoursperweek)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.0 40.0 40.0 40.4 45.0 99.0

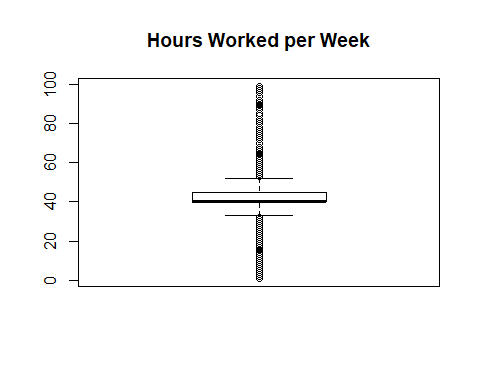
sd(data$`hours.per.week`)

## [1] NA

hist(data$hoursperweek,main = "Distribution of Hours Worked per Week",xlab="Hours worked per week",col = "blue")

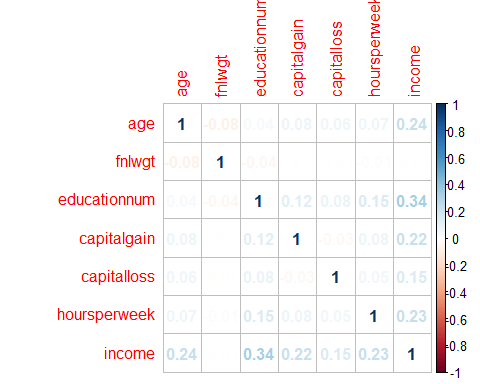


boxplot(data$hoursperweek,main="Hours Worked per Week")



7a. Find the Correlation between numerical attributes.

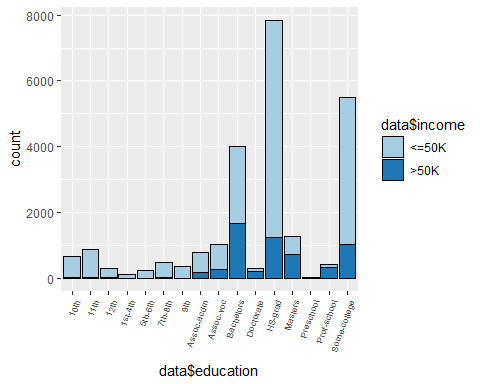
#Changing income to 0 <= 50k, 1 > 50k  
  
data1 <- data  
data1$income <- as.numeric(data1$income)-1  
#Correlation plot  
M <- c(1, 3, 5, 11:13, 15)  
corrplot(cor(data1[,M]),method = "number")



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

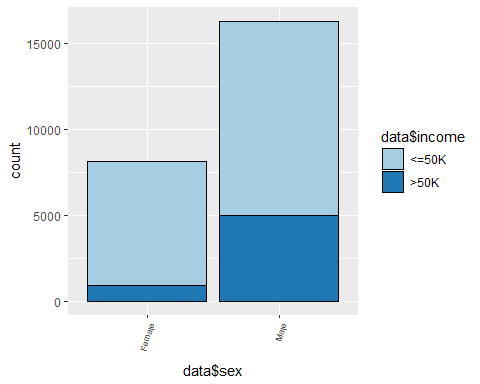
7b. Find the Correlation between categorical attributes with numerical attribute (income)

#based on the Education level  
ggplot(data, aes(x=data$education,fill=data$income)) + geom\_bar(position = "stack", color = "black") + theme(axis.text.x=element\_text(angle = 70 , hjust= 1, size=7)) + scale\_fill\_brewer(palette="Paired")



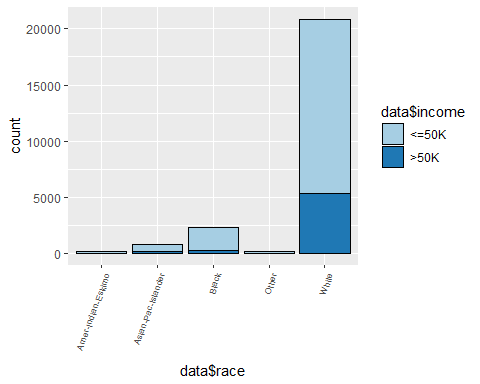
# Result shows adults with higher education has earning > 50K  
# Adults with Bachelors degree have maximum number of earnings > 50K, followed by doctorate and masters  
# Adults with lower education level have maximum portion of income <= 50K

#based on the sex  
ggplot(data, aes(x=data$sex,fill=data$income)) + geom\_bar(position = "stack", color = "black") + theme(axis.text.x=element\_text(angle = 70 , hjust= 1, size=7)) + scale\_fill\_brewer(palette="Paired")



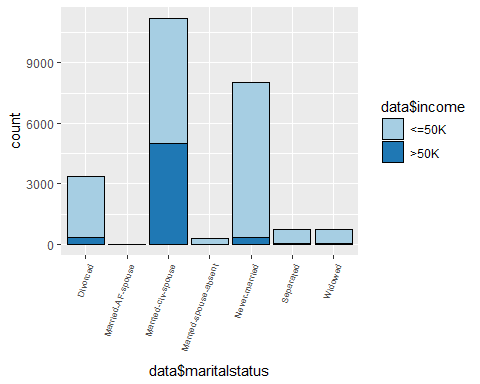
#Result shows the ratio of male earning income > 50K is more than female

#based on the race  
ggplot(data, aes(x=data$race,fill=data$income)) + geom\_bar(position = "stack", color = "black") + theme(axis.text.x=element\_text(angle = 70 , hjust= 1, size=7)) + scale\_fill\_brewer(palette="Paired")

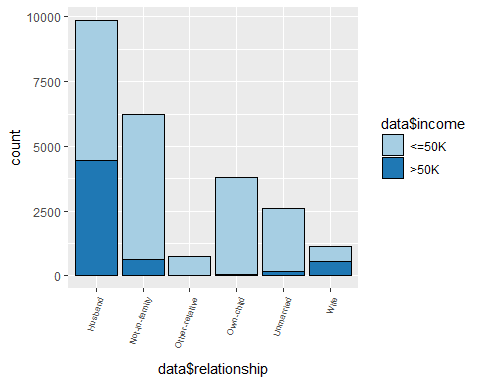


#Result shows the highest earning adults are white followed by Black and Asia pacific

#based on the marital status and relationship  
  
ggplot(data, aes(x=data$maritalstatus,fill=data$income)) + geom\_bar(position = "stack", color = "black") + theme(axis.text.x=element\_text(angle = 70 , hjust= 1, size=7)) + scale\_fill\_brewer(palette="Paired")

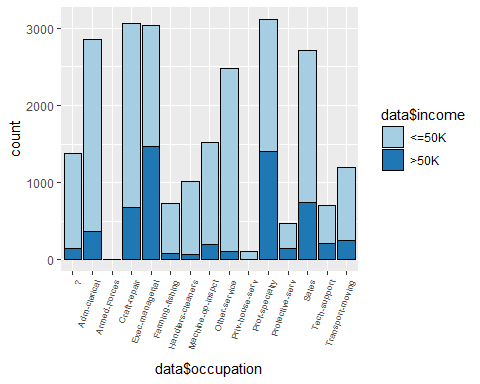


ggplot(data, aes(x=data$relationship,fill=data$income)) + geom\_bar(position = "stack", color = "black") + theme(axis.text.x=element\_text(angle = 70 , hjust= 1, size=7)) + scale\_fill\_brewer(palette="Paired")



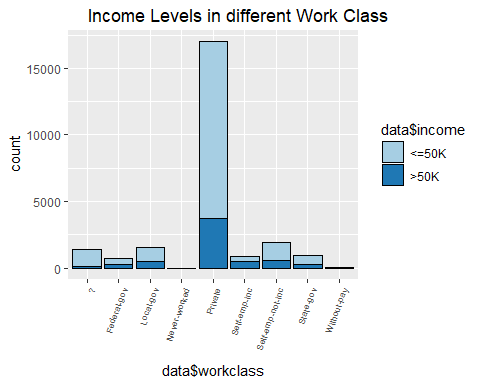
#Results in both the graphs show that Male and married people are earning more than 50K, as compared to female and unmarried people

#based on the occupation  
ggplot(data, aes(x=data$occupation,fill=data$income)) + geom\_bar(position = "stack", color = "black") + theme(axis.text.x=element\_text(angle = 70 , hjust= 1, size=7)) + scale\_fill\_brewer(palette="Paired")



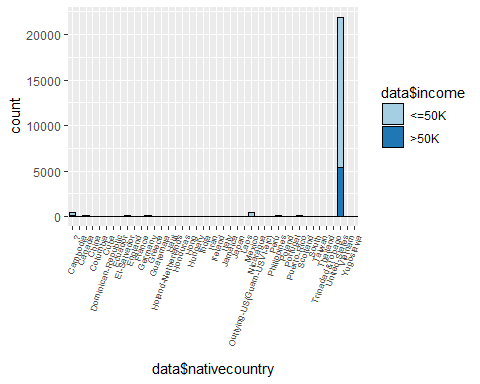
#Result shows adults with higher position like Manager, Professor are earning > 50K

#based on the work class  
ggplot(data, aes(x=data$workclass,fill=data$income)) + geom\_bar(position = "stack", color = "black") + ggtitle(' Income Levels in different Work Class')+ theme(axis.text.x=element\_text(angle = 70 , hjust= 1, size=7)) + scale\_fill\_brewer(palette="Paired")



#Result shows adults in private sector have maximum number of earning of > 50K

ggplot(data, aes(x=data$nativecountry,fill=data$income)) + geom\_bar(position = "stack", color = "black") + theme(axis.text.x=element\_text(angle = 70 , hjust= 1, size=7)) + scale\_fill\_brewer(palette="Paired")



#Result shows marjority of the adults belongs to the United States

Save the clean test and train data testdata.csv and traindata.csv files respectively.

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

Now we predict the data based on the traindata

model <- glm(income ~ age+ workclass+ education+maritalstatus+ occupation+ sex +hoursperweek, data = traindata, family = binomial('logit'))  
summary(model)

##   
## Call:  
## glm(formula = income ~ age + workclass + education + maritalstatus +   
## occupation + sex + hoursperweek, family = binomial("logit"),   
## data = traindata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.8154 -0.5518 -0.2372 -0.0526 3.3876   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -6.971653 0.242925 -28.699 < 2e-16  
## age 0.029373 0.001760 16.693 < 2e-16  
## workclass Federal-gov 0.966112 0.166737 5.794 6.86e-09  
## workclass Local-gov 0.362486 0.151204 2.397 0.0165  
## workclass Never-worked -10.740772 333.392623 -0.032 0.9743  
## workclass Private 0.554583 0.134859 4.112 3.92e-05  
## workclass Self-emp-inc 0.825242 0.161448 5.112 3.20e-07  
## workclass Self-emp-not-inc 0.106586 0.148343 0.719 0.4724  
## workclass State-gov 0.156498 0.165129 0.948 0.3433  
## workclass Without-pay -12.229956 218.138278 -0.056 0.9553  
## education 11th -0.056833 0.233697 -0.243 0.8079  
## education 12th 0.607009 0.277971 2.184 0.0290  
## education 1st-4th -0.725741 0.464978 -1.561 0.1186  
## education 5th-6th -0.460474 0.348199 -1.322 0.1860  
## education 7th-8th -0.591074 0.254284 -2.324 0.0201  
## education 9th -0.555298 0.296418 -1.873 0.0610  
## education Assoc-acdm 1.276529 0.193340 6.602 4.04e-11  
## education Assoc-voc 1.313741 0.185128 7.096 1.28e-12  
## education Bachelors 1.971646 0.172237 11.447 < 2e-16  
## education Doctorate 2.955892 0.232274 12.726 < 2e-16  
## education HS-grad 0.738202 0.168210 4.389 1.14e-05  
## education Masters 2.353122 0.183330 12.835 < 2e-16  
## education Preschool -11.633386 129.576912 -0.090 0.9285  
## education Prof-school 3.043054 0.217199 14.010 < 2e-16  
## education Some-college 1.074970 0.170528 6.304 2.90e-10  
## maritalstatus Married-AF-spouse 2.419811 0.575338 4.206 2.60e-05  
## maritalstatus Married-civ-spouse 2.093609 0.070977 29.497 < 2e-16  
## maritalstatus Married-spouse-absent 0.019944 0.233382 0.085 0.9319  
## maritalstatus Never-married -0.466253 0.086986 -5.360 8.32e-08  
## maritalstatus Separated -0.245498 0.174889 -1.404 0.1604  
## maritalstatus Widowed -0.042687 0.154873 -0.276 0.7828  
## occupation Adm-clerical 0.104270 0.107293 0.972 0.3311  
## occupation Armed-Forces -0.519519 1.396471 -0.372 0.7099  
## occupation Craft-repair 0.141664 0.092960 1.524 0.1275  
## occupation Exec-managerial 0.904007 0.094987 9.517 < 2e-16  
## occupation Farming-fishing -0.922612 0.154606 -5.968 2.41e-09  
## occupation Handlers-cleaners -0.671411 0.162294 -4.137 3.52e-05  
## occupation Machine-op-inspct -0.173751 0.115034 -1.510 0.1309  
## occupation Other-service -0.793418 0.135386 -5.860 4.62e-09  
## occupation Priv-house-serv -2.547249 1.217294 -2.093 0.0364  
## occupation Prof-specialty 0.628490 0.101722 6.178 6.47e-10  
## occupation Protective-serv 0.596874 0.145475 4.103 4.08e-05  
## occupation Sales 0.406006 0.098008 4.143 3.43e-05  
## occupation Tech-support 0.753259 0.129899 5.799 6.68e-09  
## occupation Transport-moving NA NA NA NA  
## sex Male 0.113354 0.056383 2.010 0.0444  
## hoursperweek 0.031028 0.001761 17.615 < 2e-16  
##   
## (Intercept) \*\*\*  
## age \*\*\*  
## workclass Federal-gov \*\*\*  
## workclass Local-gov \*   
## workclass Never-worked   
## workclass Private \*\*\*  
## workclass Self-emp-inc \*\*\*  
## workclass Self-emp-not-inc   
## workclass State-gov   
## workclass Without-pay   
## education 11th   
## education 12th \*   
## education 1st-4th   
## education 5th-6th   
## education 7th-8th \*   
## education 9th .   
## education Assoc-acdm \*\*\*  
## education Assoc-voc \*\*\*  
## education Bachelors \*\*\*  
## education Doctorate \*\*\*  
## education HS-grad \*\*\*  
## education Masters \*\*\*  
## education Preschool   
## education Prof-school \*\*\*  
## education Some-college \*\*\*  
## maritalstatus Married-AF-spouse \*\*\*  
## maritalstatus Married-civ-spouse \*\*\*  
## maritalstatus Married-spouse-absent   
## maritalstatus Never-married \*\*\*  
## maritalstatus Separated   
## maritalstatus Widowed   
## occupation Adm-clerical   
## occupation Armed-Forces   
## occupation Craft-repair   
## occupation Exec-managerial \*\*\*  
## occupation Farming-fishing \*\*\*  
## occupation Handlers-cleaners \*\*\*  
## occupation Machine-op-inspct   
## occupation Other-service \*\*\*  
## occupation Priv-house-serv \*   
## occupation Prof-specialty \*\*\*  
## occupation Protective-serv \*\*\*  
## occupation Sales \*\*\*  
## occupation Tech-support \*\*\*  
## occupation Transport-moving   
## sex Male \*   
## hoursperweek \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 26962 on 24420 degrees of freedom  
## Residual deviance: 17240 on 24375 degrees of freedom  
## AIC: 17332  
##   
## Number of Fisher Scoring iterations: 13

predicttrain <- predict(model,traindata,type='response')

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =  
## ifelse(type == : prediction from a rank-deficient fit may be misleading

pred1 <- rep('<=50K', length(predicttrain))  
pred1[predicttrain>=.5] <- '>50K'  
tb1 <- table(pred1, traindata$income)  
tb1

##   
## pred1 <=50K >50K  
## <=50K 17168 2644  
## >50K 1372 3237

Apply different algorithm to predict the results using train and test data

1. DECISION TREE

Dectree<- rpart(income~ age+ workclass+ education+maritalstatus+ occupation+ sex +hoursperweek, data = traindata, method='class',cp =1e-3)  
  
#Result using traindata  
Dectree.Ptrain <- predict(Dectree,newdata= traindata, type = 'class')  
confusionMatrix(traindata$income,Dectree.Ptrain)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction <=50K >50K  
## <=50K 17269 1271  
## >50K 2508 3373  
##   
## Accuracy : 0.8453   
## 95% CI : (0.8407, 0.8498)  
## No Information Rate : 0.8098   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.5441   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.8732   
## Specificity : 0.7263   
## Pos Pred Value : 0.9314   
## Neg Pred Value : 0.5735   
## Prevalence : 0.8098   
## Detection Rate : 0.7071   
## Detection Prevalence : 0.7592   
## Balanced Accuracy : 0.7997   
##   
## 'Positive' Class : <=50K   
##

#Result using testdata  
Dectree.pred.prob <- predict(Dectree, newdata = testdata, type = 'prob')  
Dectree.pred <- predict(Dectree, newdata = testdata, type = 'class')  
confusionMatrix(testdata$income,Dectree.pred)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction <=50K >50K  
## <=50K 5700 479  
## >50K 888 1072  
##   
## Accuracy : 0.832   
## 95% CI : (0.8237, 0.8401)  
## No Information Rate : 0.8094   
## P-Value [Acc > NIR] : 7.256e-08   
##   
## Kappa : 0.5054   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.8652   
## Specificity : 0.6912   
## Pos Pred Value : 0.9225   
## Neg Pred Value : 0.5469   
## Prevalence : 0.8094   
## Detection Rate : 0.7003   
## Detection Prevalence : 0.7592   
## Balanced Accuracy : 0.7782   
##   
## 'Positive' Class : <=50K   
##

1. RANDOM FOREST

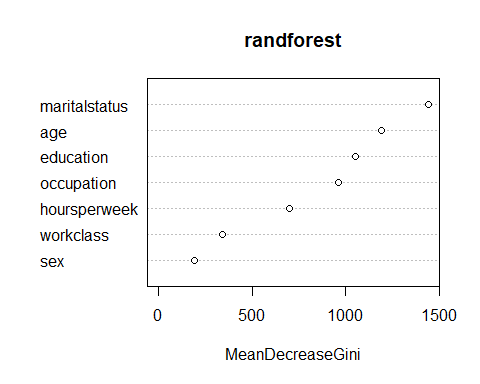
library(randomForest)  
levels(testdata$workclass) <- levels(traindata$workclass)  
randforest <- randomForest(income ~ age+ workclass+ education+maritalstatus+occupation+ sex+hoursperweek, data = traindata, ntree = 500)  
randforest.pred.prob <- predict(randforest, newdata = testdata, type = 'prob')  
randforest.pred <- predict(randforest, newdata = testdata, type = 'class')  
  
# confusion matrix   
tb3 <- table(randforest.pred, testdata$income)  
tb3

##   
## randforest.pred <=50K >50K  
## <=50K 5654 820  
## >50K 525 1140

confusionMatrix(testdata$income,randforest.pred)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction <=50K >50K  
## <=50K 5654 525  
## >50K 820 1140  
##   
## Accuracy : 0.8347   
## 95% CI : (0.8265, 0.8428)  
## No Information Rate : 0.7954   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.5236   
## Mcnemar's Test P-Value : 1.088e-15   
##   
## Sensitivity : 0.8733   
## Specificity : 0.6847   
## Pos Pred Value : 0.9150   
## Neg Pred Value : 0.5816   
## Prevalence : 0.7954   
## Detection Rate : 0.6947   
## Detection Prevalence : 0.7592   
## Balanced Accuracy : 0.7790   
##   
## 'Positive' Class : <=50K   
##

varImpPlot (randforest)



1. LINEAR REGRESION

linReg <- glm(income ~ age+ workclass+ education+maritalstatus+ occupation+ sex +hoursperweek, data = traindata, family = binomial('logit'))  
summary(linReg)

##   
## Call:  
## glm(formula = income ~ age + workclass + education + maritalstatus +   
## occupation + sex + hoursperweek, family = binomial("logit"),   
## data = traindata)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.8154 -0.5518 -0.2372 -0.0526 3.3876   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -6.971653 0.242925 -28.699 < 2e-16  
## age 0.029373 0.001760 16.693 < 2e-16  
## workclass Federal-gov 0.966112 0.166737 5.794 6.86e-09  
## workclass Local-gov 0.362486 0.151204 2.397 0.0165  
## workclass Never-worked -10.740772 333.392623 -0.032 0.9743  
## workclass Private 0.554583 0.134859 4.112 3.92e-05  
## workclass Self-emp-inc 0.825242 0.161448 5.112 3.20e-07  
## workclass Self-emp-not-inc 0.106586 0.148343 0.719 0.4724  
## workclass State-gov 0.156498 0.165129 0.948 0.3433  
## workclass Without-pay -12.229956 218.138278 -0.056 0.9553  
## education 11th -0.056833 0.233697 -0.243 0.8079  
## education 12th 0.607009 0.277971 2.184 0.0290  
## education 1st-4th -0.725741 0.464978 -1.561 0.1186  
## education 5th-6th -0.460474 0.348199 -1.322 0.1860  
## education 7th-8th -0.591074 0.254284 -2.324 0.0201  
## education 9th -0.555298 0.296418 -1.873 0.0610  
## education Assoc-acdm 1.276529 0.193340 6.602 4.04e-11  
## education Assoc-voc 1.313741 0.185128 7.096 1.28e-12  
## education Bachelors 1.971646 0.172237 11.447 < 2e-16  
## education Doctorate 2.955892 0.232274 12.726 < 2e-16  
## education HS-grad 0.738202 0.168210 4.389 1.14e-05  
## education Masters 2.353122 0.183330 12.835 < 2e-16  
## education Preschool -11.633386 129.576912 -0.090 0.9285  
## education Prof-school 3.043054 0.217199 14.010 < 2e-16  
## education Some-college 1.074970 0.170528 6.304 2.90e-10  
## maritalstatus Married-AF-spouse 2.419811 0.575338 4.206 2.60e-05  
## maritalstatus Married-civ-spouse 2.093609 0.070977 29.497 < 2e-16  
## maritalstatus Married-spouse-absent 0.019944 0.233382 0.085 0.9319  
## maritalstatus Never-married -0.466253 0.086986 -5.360 8.32e-08  
## maritalstatus Separated -0.245498 0.174889 -1.404 0.1604  
## maritalstatus Widowed -0.042687 0.154873 -0.276 0.7828  
## occupation Adm-clerical 0.104270 0.107293 0.972 0.3311  
## occupation Armed-Forces -0.519519 1.396471 -0.372 0.7099  
## occupation Craft-repair 0.141664 0.092960 1.524 0.1275  
## occupation Exec-managerial 0.904007 0.094987 9.517 < 2e-16  
## occupation Farming-fishing -0.922612 0.154606 -5.968 2.41e-09  
## occupation Handlers-cleaners -0.671411 0.162294 -4.137 3.52e-05  
## occupation Machine-op-inspct -0.173751 0.115034 -1.510 0.1309  
## occupation Other-service -0.793418 0.135386 -5.860 4.62e-09  
## occupation Priv-house-serv -2.547249 1.217294 -2.093 0.0364  
## occupation Prof-specialty 0.628490 0.101722 6.178 6.47e-10  
## occupation Protective-serv 0.596874 0.145475 4.103 4.08e-05  
## occupation Sales 0.406006 0.098008 4.143 3.43e-05  
## occupation Tech-support 0.753259 0.129899 5.799 6.68e-09  
## occupation Transport-moving NA NA NA NA  
## sex Male 0.113354 0.056383 2.010 0.0444  
## hoursperweek 0.031028 0.001761 17.615 < 2e-16  
##   
## (Intercept) \*\*\*  
## age \*\*\*  
## workclass Federal-gov \*\*\*  
## workclass Local-gov \*   
## workclass Never-worked   
## workclass Private \*\*\*  
## workclass Self-emp-inc \*\*\*  
## workclass Self-emp-not-inc   
## workclass State-gov   
## workclass Without-pay   
## education 11th   
## education 12th \*   
## education 1st-4th   
## education 5th-6th   
## education 7th-8th \*   
## education 9th .   
## education Assoc-acdm \*\*\*  
## education Assoc-voc \*\*\*  
## education Bachelors \*\*\*  
## education Doctorate \*\*\*  
## education HS-grad \*\*\*  
## education Masters \*\*\*  
## education Preschool   
## education Prof-school \*\*\*  
## education Some-college \*\*\*  
## maritalstatus Married-AF-spouse \*\*\*  
## maritalstatus Married-civ-spouse \*\*\*  
## maritalstatus Married-spouse-absent   
## maritalstatus Never-married \*\*\*  
## maritalstatus Separated   
## maritalstatus Widowed   
## occupation Adm-clerical   
## occupation Armed-Forces   
## occupation Craft-repair   
## occupation Exec-managerial \*\*\*  
## occupation Farming-fishing \*\*\*  
## occupation Handlers-cleaners \*\*\*  
## occupation Machine-op-inspct   
## occupation Other-service \*\*\*  
## occupation Priv-house-serv \*   
## occupation Prof-specialty \*\*\*  
## occupation Protective-serv \*\*\*  
## occupation Sales \*\*\*  
## occupation Tech-support \*\*\*  
## occupation Transport-moving   
## sex Male \*   
## hoursperweek \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 26962 on 24420 degrees of freedom  
## Residual deviance: 17240 on 24375 degrees of freedom  
## AIC: 17332  
##   
## Number of Fisher Scoring iterations: 13

predictiontrain <- predict(linReg,traindata,type='response')

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =  
## ifelse(type == : prediction from a rank-deficient fit may be misleading

pred1 <- rep('<=50K', length(predictiontrain))  
pred1[predictiontrain>=.5] <- '>50K'  
tb1 <- table(pred1, traindata$income)  
tb1

##   
## pred1 <=50K >50K  
## <=50K 17168 2644  
## >50K 1372 3237

prob <- predict(linReg, testdata, type = 'response')

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =  
## ifelse(type == : prediction from a rank-deficient fit may be misleading

prediction <- predict(linReg,testdata,type='response')

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =  
## ifelse(type == : prediction from a rank-deficient fit may be misleading

########################################################################  
# P values shows that Age ,workclass, education, marital status, occupation,  
# race, sex, hours per week are the significant attributes.  
########################################################################  
pred <- rep('<=50K', length(prob))  
pred[prob>=.5] <- '>50K'  
tb <- table(pred, testdata$income)  
tb

##   
## pred <=50K >50K  
## <=50K 5684 904  
## >50K 495 1056

# Confusion matrix shows that it has an Accuracy of 83.01%  
# misclasification 17%.

Finally we have to compare the the Algorithm

###DECISION TREE  
prtree <- prediction(Dectree.pred.prob[,2],testdata$income)  
perftree <- performance(prtree,measure="tpr",x.measure="fpr")  
DTFrametree <- data.frame(FP=perftree@x.values[[1]],TP=perftree@y.values[[1]])  
auctree <- performance(prtree, measure='auc')@y.values[[1]]  
auctree

## [1] 0.8500693

###RANDOM FOREST  
prRForest <- prediction(randforest.pred.prob[,2],testdata$income)  
perfRForest <- performance(prRForest,measure="tpr",x.measure="fpr")  
DTFrameRForest <- data.frame(FP=perfRForest@x.values[[1]],TP=perfRForest@y.values[[1]])  
aucFtree <- performance(prRForest, measure='auc')@y.values[[1]]  
aucFtree

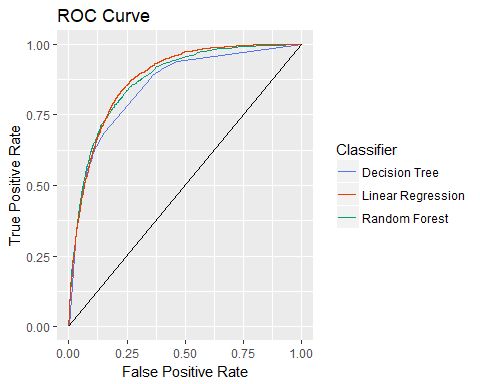
## [1] 0.8733921

## LINEAR REGRESION  
pr <- prediction(prob,testdata$income)  
perf <- performance(pr,measure="tpr", x.measure="fpr")  
DtFrameReg <- data.frame(FP=perf@x.values[[1]],TP=perf@y.values[[1]])  
aucRegresion <- performance(pr,measure='auc')@y.values[[1]]  
aucRegresion

## [1] 0.879603

Use of ROC curve

g <- ggplot() +   
 geom\_line(data = DTFrametree, aes(x = FP, y = TP, color = 'Decision Tree')) +   
 geom\_line(data = DTFrameRForest, aes(x = FP, y = TP, color = 'Random Forest')) +   
 geom\_line(data = DtFrameReg, aes(x = FP, y = TP, color = 'Linear Regression')) +   
 geom\_segment(aes(x = 0, xend = 1, y = 0, yend = 1)) +  
 ggtitle('ROC Curve') +   
 labs(x = 'False Positive Rate', y = 'True Positive Rate')   
  
g + scale\_colour\_manual(name = 'Classifier', values = c('Decision Tree'='#5674E9', 'Random Forest'='#009E73', 'Linear Regression'='#E63F00'))



auc <- rbind(aucRegresion,auctree,aucFtree)  
rownames(auc) <- (c('Decision Tree', 'Random Forest', 'Linear Regression'))  
colnames(auc) <- 'ROC Curve Area'  
round(auc, 6)

## ROC Curve Area  
## Decision Tree 0.879603  
## Random Forest 0.850069  
## Linear Regression 0.873392