Adult Census Income

This dataset is Adult Census income collected by the U.S. Census Bureau in 1994 and 1995. Data can be found on kaggle: https://www.kaggle.com/uciml/adult-census-income

Some of the steps that had to be made to cleanup the data was to re-classify most of the columns as factors in order for me to predict and test on them. In addition to that there were many rows that had NA or certain variables that had missing values and those needed to be filled in. I also had to re-classify if a person made more than 50k or less than 50k to make it easier to factor the income. In addition to that, there was 2 columns that I removed because I deemed them not important to the data - These were the fnlwgt (final weight of that census believes the entry represents) and education.num (number of years of education).

```r
# Reading in data
df <- read.csv("adult.csv", header = TRUE)

# DATA CLEANING
df$workclass <- as.factor(df$workclass)
df$education <- as.factor(df$education)
df$marital.status <- as.factor(df$marital.status)
df$occupation <- as.factor(df$occupation)
df$relationship <- as.factor(df$relationship)
df$race <- as.factor(df$race)
df$sex <- as.factor(df$sex)
df$native.country <- as.factor(df$native.country)
df$income <- as.factor(df$income)

# re-classifying
df$income<-ifelse(df$income=='>50K',1,0)
df$workclass<-ifelse(df$workclass=='?','Unknown',as.character(df$workclass))
df$income <- as.factor(df$income)

# Removing columns fnlwgt and education.num
df <- df[,-3]
df <- df[,-4]

df[df == "?"] <- NA
df <- na.omit(df)

colSums(is.na(df))
```

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

DATA EXPLORATION

```
# DATA EXPLORATION #
str(df)
```

```
## 'data.frame':    30162 obs. of  13 variables:
##  $ age           : int  82 54 41 34 38 74 68 45 38 52 ...
##  $ workclass     : chr  "Private" "Private" "Private" "Private" ...
##  $ education     : Factor w/ 16 levels "10th","11th",..: 12 6 16 12 1 11 12 11 15 10 ...
##  $ marital.status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 7 1 6 1 6 5 1 1 5 7 ...
##  $ occupation    : Factor w/ 15 levels "?","Adm-clerical",..: 5 8 11 9 2 11 11 11 11 9 ...
##  $ relationship  : Factor w/ 6 levels "Husband","Not-in-family",..: 2 5 4 5 5 3 2 5 2 2 ...
##  $ race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 5 5 5 5 5 3 5 5 ...
##  $ sex           : Factor w/ 2 levels "Female","Male": 1 1 1 1 2 1 1 1 2 1 ...
##  $ capital.gain  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ capital.loss  : int  4356 3900 3900 3770 3770 3683 3683 3004 2824 2824 ...
##  $ hours.per.week: int  18 40 40 45 40 20 40 35 45 20 ...
##  $ native.country: Factor w/ 42 levels "?","Cambodia",..: 40 40 40 40 40 40 40 40 40 40 ...
##  $ income        : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 2 2 2 ...
##  - attr(*, "na.action")= 'omit' Named int [1:2399] 1 3 10 15 19 25 45 49 50 66 ...
##   ..- attr(*, "names")= chr [1:2399] "1" "3" "10" "15" ...
```

```
names(df)
```

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

```
summary(df)
```

```
##       age         workclass               education   
##  Min.   :17.00   Length:30162       HS-grad     :9840  
##  1st Qu.:28.00   Class :character   Some-college:6678  
##  Median :37.00   Mode  :character   Bachelors   :5044  
##  Mean   :38.44                      Masters     :1627  
##  3rd Qu.:47.00                      Assoc-voc   :1307  
##  Max.   :90.00                      11th        :1048  
##                                     (Other)     :4618  
##                marital.status            occupation           relationship 
##  Divorced             : 4214   Prof-specialty :4038   Husband       :12463  
##  Married-AF-spouse    :   21   Craft-repair   :4030   Not-in-family : 7726  
##  Married-civ-spouse   :14065   Exec-managerial:3992   Other-relative:  889  
##  Married-spouse-absent:  370   Adm-clerical   :3721   Own-child     : 4466  
##  Never-married        : 9726   Sales          :3584   Unmarried     : 3212  
##  Separated            :  939   Other-service  :3212   Wife          : 1406  
##  Widowed              :  827   (Other)        :7585                         
##                  race            sex         capital.gain    capital.loss   
##  Amer-Indian-Eskimo:  286   Female: 9782   Min.   :    0   Min.   :   0.00  
##  Asian-Pac-Islander:  895   Male  :20380   1st Qu.:    0   1st Qu.:   0.00  
##  Black             : 2817                  Median :    0   Median :   0.00  
##  Other             :  231                  Mean   : 1092   Mean   :  88.37
##  White             :25933                  3rd Qu.:    0   3rd Qu.:   0.00
##                                            Max.   :99999   Max.   :4356.00
##
##  hours.per.week        native.country  income
##  Min.   : 1.00   United-States:27504   0:22654
##  1st Qu.:40.00   Mexico       :  610   1: 7508
##  Median :40.00   Philippines  :  188
##  Mean   :40.93   Germany      :  128
##  3rd Qu.:45.00   Puerto-Rico  :  109
##  Max.   :99.00   Canada       :  107
##                  (Other)      : 1516
```

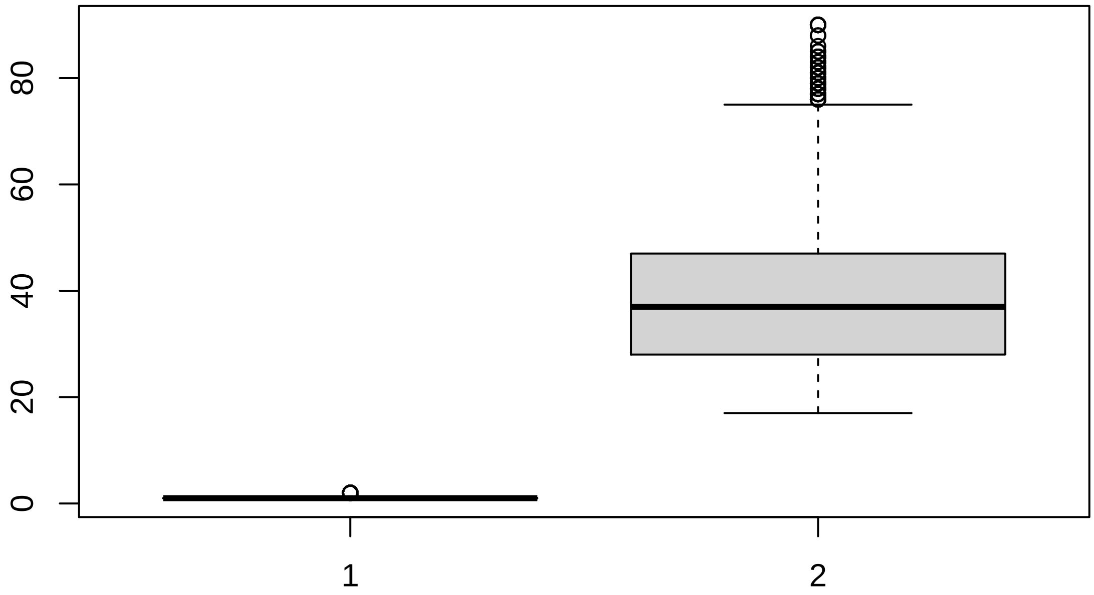
```
dim(df)
```

```
## [1] 30162    13
```

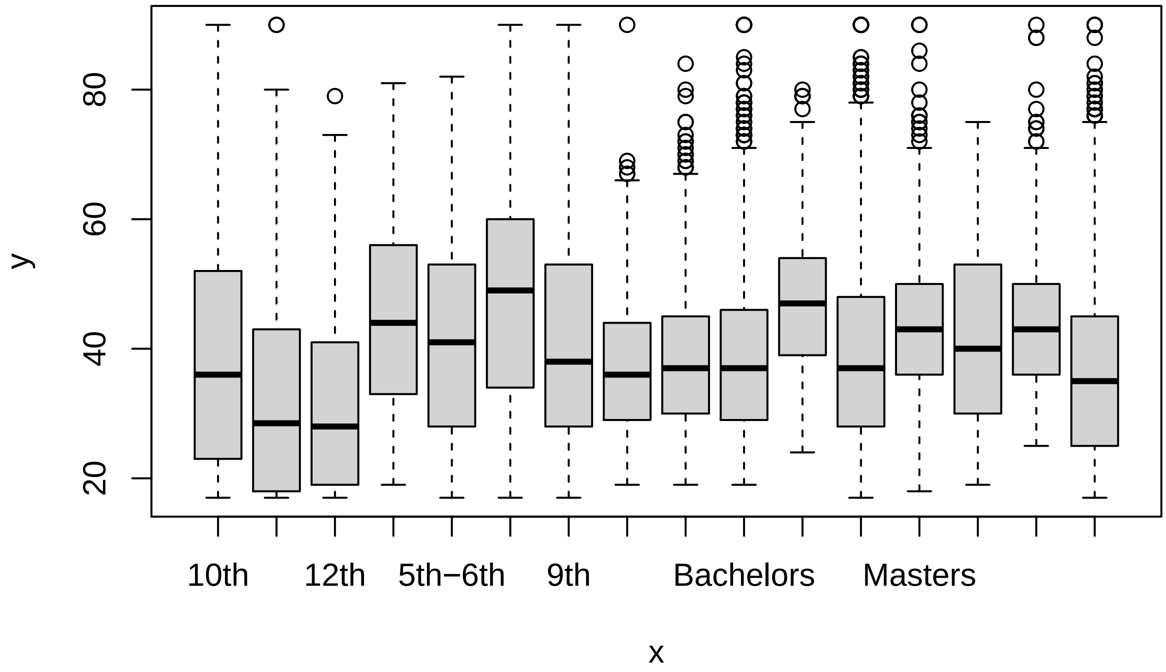
```
head(df)
```

```
##   age workclass    education marital.status        occupation   relationship
## 2  82   Private      HS-grad        Widowed   Exec-managerial  Not-in-family
## 4  54   Private      7th-8th       Divorced Machine-op-inspct      Unmarried
## 5  41   Private Some-college      Separated    Prof-specialty      Own-child
## 6  34   Private      HS-grad       Divorced     Other-service      Unmarried
## 7  38   Private         10th      Separated      Adm-clerical      Unmarried
## 8  74 State-gov    Doctorate  Never-married    Prof-specialty Other-relative
##    race    sex capital.gain capital.loss hours.per.week native.country income
## 2 White Female            0         4356             18  United-States      0
## 4 White Female            0         3900             40  United-States      0
## 5 White Female            0         3900             40  United-States      0
## 6 White Female            0         3770             45  United-States      0
## 7 White   Male            0         3770             40  United-States      0
## 8 White Female            0         3683             20  United-States      1
```

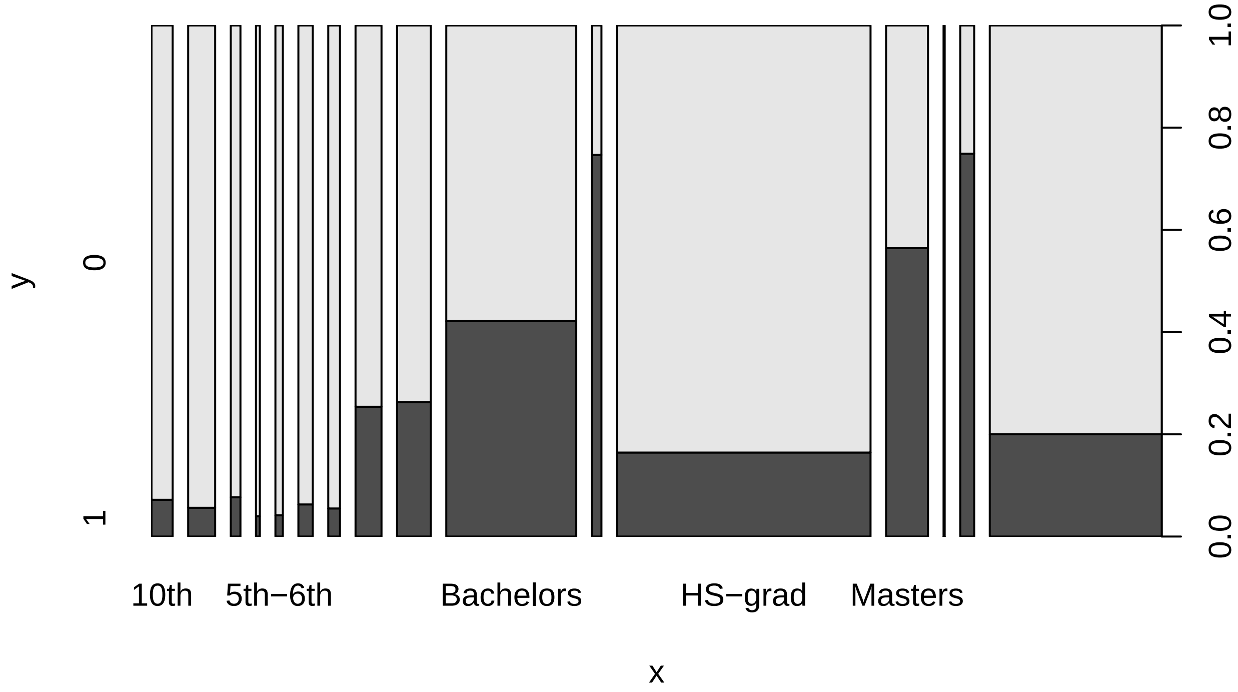
```
# GRAPHS #
boxplot(df$income, df$age)
```



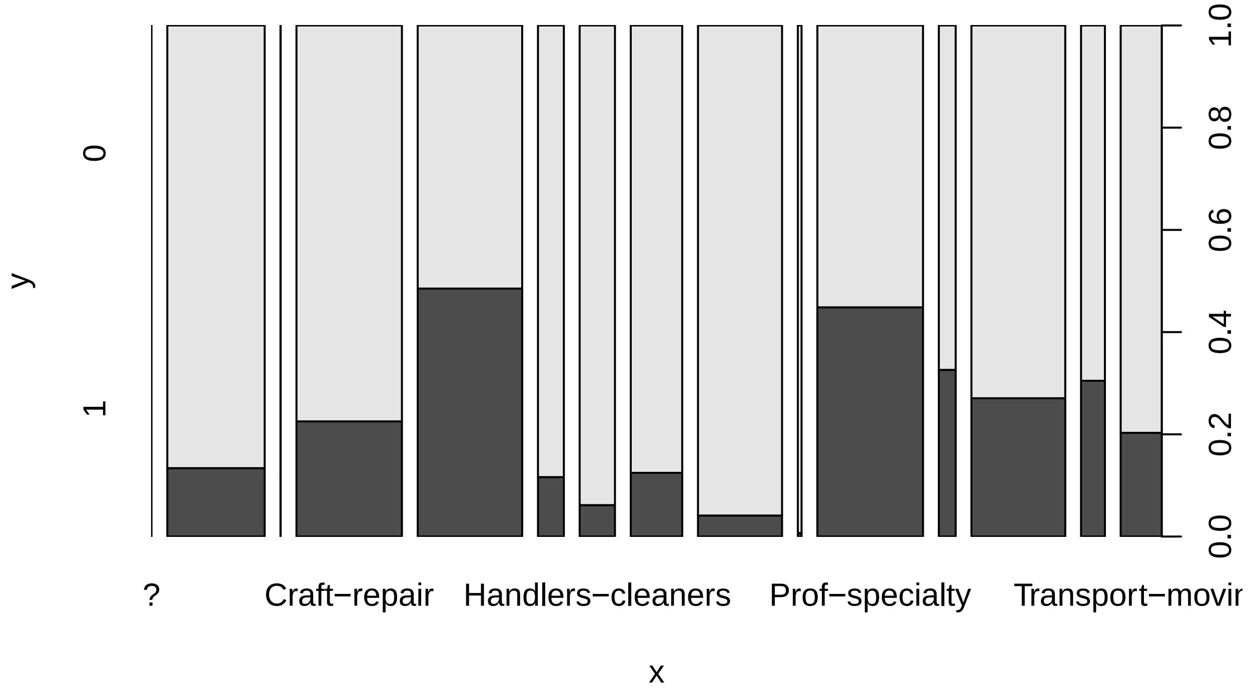
```
plot(df$education, df$age)
```



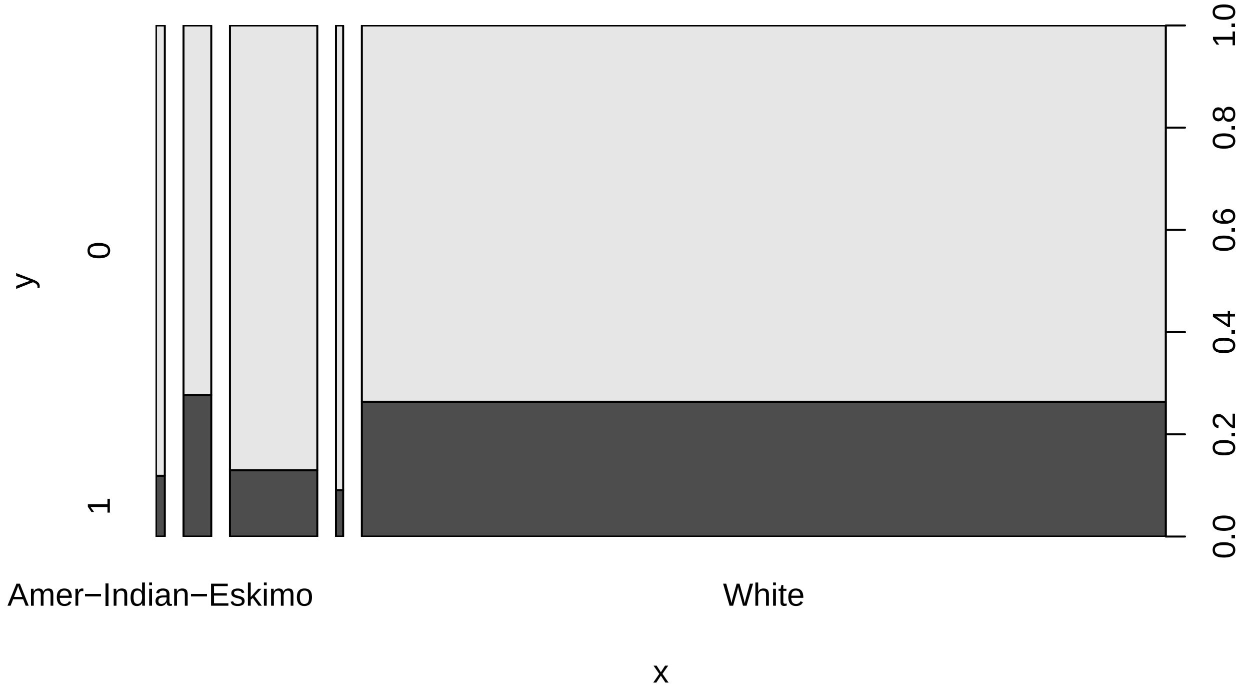
```
plot(df$education, df$income)
```



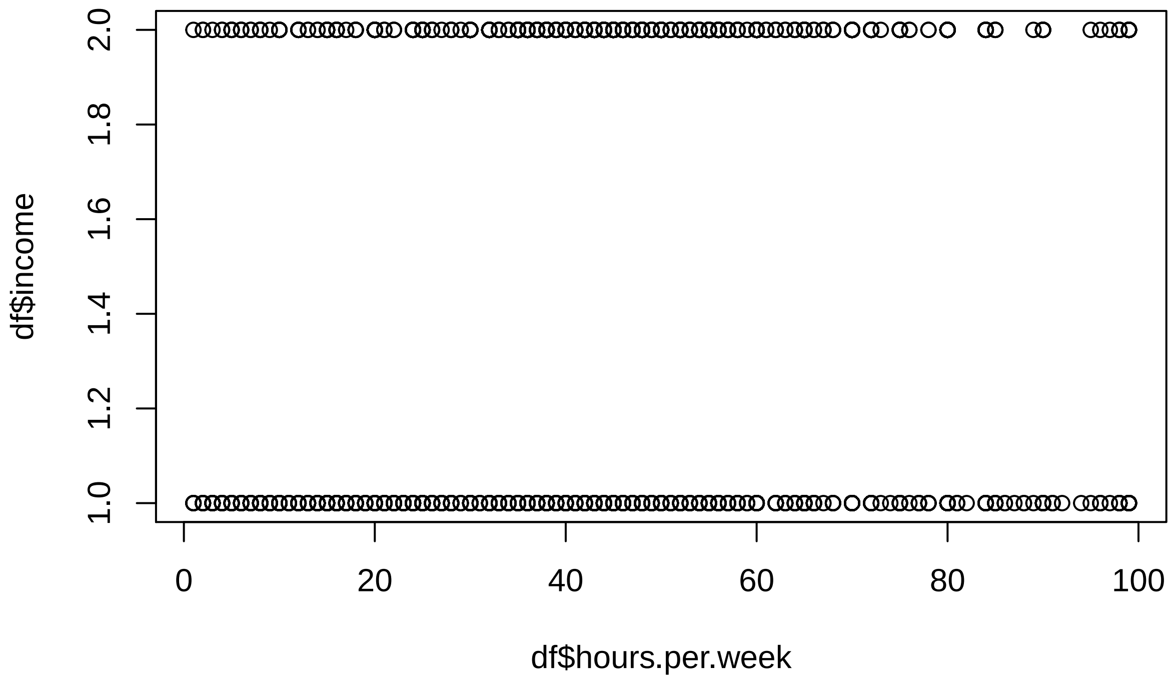
```
plot(df$occupation, df$income)
```



```
plot(df$race, df$income)
```



```
plot(df$hours.per.week, df$income)
```



ML Algorithms

```
set.seed(1234)
i <- sample(1:nrow(df), nrow(df)*.75, replace = FALSE)
train <- df[i,]
test <- df[-i,]

# Logistic Regression
glm1 <- glm(income~., data=train, family = "binomial")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
library(caret)
```

```
## Loading required package: lattice

## Loading required package: ggplot2
```

```
probs1 <- predict(glm1, data = train, family = "binomial")
pred1 <- ifelse(probs1>.5, "1", "0")
pred1 <- as.factor(pred1)
confusionMatrix(pred1, train$income)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 16362  2903
##          1   665  2691
##
##                Accuracy : 0.8423
##                  95% CI : (0.8375, 0.847)
##     No Information Rate : 0.7527
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5106
##
##  Mcnemar’s Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9609
##             Specificity : 0.4811
##          Pos Pred Value : 0.8493
##          Neg Pred Value : 0.8018
##              Prevalence : 0.7527
##          Detection Rate : 0.7233
##    Detection Prevalence : 0.8516
##       Balanced Accuracy : 0.7210
##
##        ’Positive’ Class : 0
##
```

I used Logistic Regression here because you could evaluate the amount a person could make based on the given information. In this case, it’s a simple if they made more than 50k or less than 50k, but using the information you could predict either. The results of the Logistic Regression were as followed: ACC: 84% The accuracy of the logistic regression was quite high at 84%. Taking a closer look, the sensitivity was at 96%, meaning that the model could find 96% of all the predicted incomes that are more than 50k. Specificity is at 48%, meaning that the model could find 48% of all predicted incomes less than 50k. This model was good at predicting those that made more than 50k, but not so much for those that makes less than 50k.

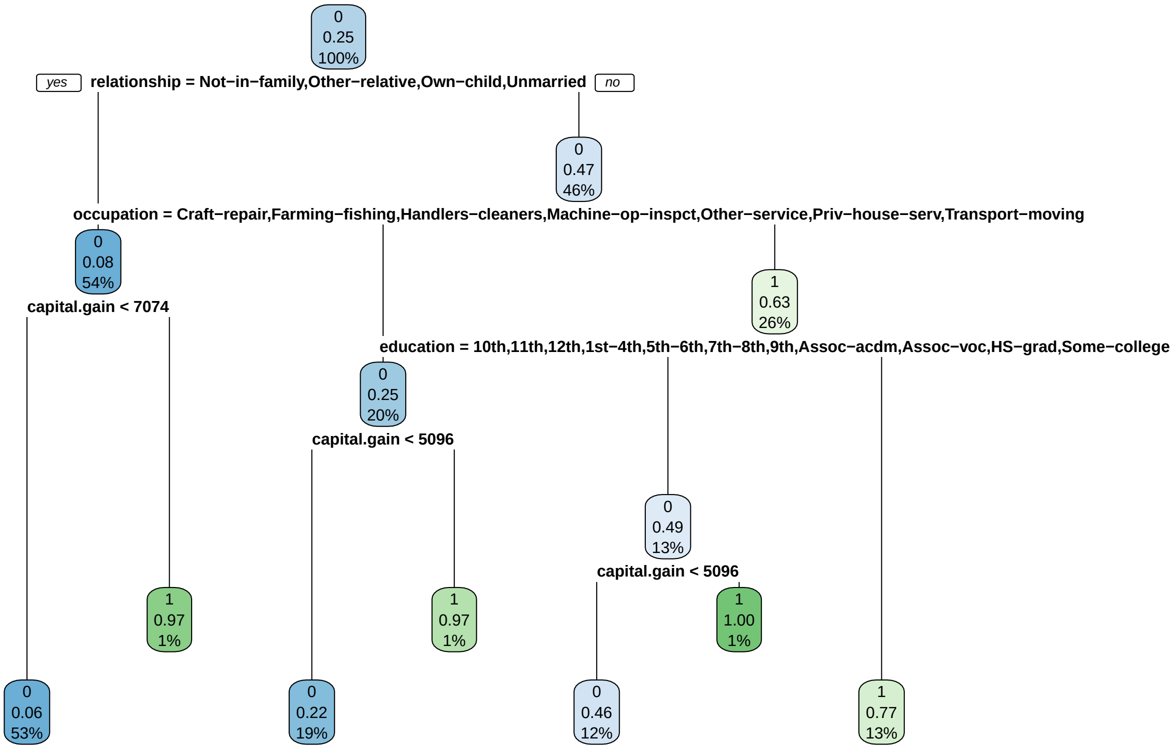
```r
# Naive Bayes
library(e1071)
nb1 <- naiveBayes(income~., data=train)
probs2 <- predict(nb1, newdata = test)
confusionMatrix(probs2, test$income)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 5300 1120
##          1  327  794
##
##                Accuracy : 0.8081
##                  95% CI : (0.799, 0.8169)
##     No Information Rate : 0.7462
##     P-Value [Acc > NIR] : < 2.2e-16
```

```
##
##                    Kappa : 0.4132
##
##  Mcnemar’s Test P-Value : < 2.2e-16
##
##              Sensitivity : 0.9419
##              Specificity : 0.4148
##           Pos Pred Value : 0.8255
##           Neg Pred Value : 0.7083
##               Prevalence : 0.7462
##           Detection Rate : 0.7028
##     Detection Prevalence : 0.8513
##        Balanced Accuracy : 0.6784
##
##         ’Positive’ Class : 0
##
```

I used Naive Bayes here due to the large data set that is given. By comparing the 50k income with the different variety of classes that are provided. In this case, those that make more than 50k can be compared to those that make less to generate a prediction. The results of the Naive Bayes are as follows: ACC: 80% The accuracy is actually quite high and with the sensitivity at 94%, that means the model could find 94% of all the predicted incomes that are more than 50k. Specificity is at 41% meaning that the model could find 41% of all predicted incomes less than 50k. The Naive Bayes was good at predicting those that made more than 50k, but not so much for those that make less than 50k.

```r
# Decision Tree
library(rpart)
library(rpart.plot)
decision_tree <- rpart(income~., data = test, method = "class")
rpart.plot(decision_tree)
```


0
0.25
100%
yes
relationship = Not−in−family,Other−relative,Own−child,Unmarried
no
0
0.47
46%
occupation = Craft−repair,Farming−fishing,Handlers−cleaners,Machine−op−inspct,Other−service,Priv−house−serv,Transport−moving
0
0.08
54%
capital.gain < 7074
1
0.63
26%
education = 10th,11th,12th,1st−4th,5th−6th,7th−8th,9th,Assoc−acdm,Assoc−voc,HS−grad,Some−college
0
0.25
20%
capital.gain < 5096
0
0.49
13%
capital.gain < 5096
1
0.97
1%
1
0.97
1%
1
1.00
1%
0
0.06
53%
0
0.22
19%
0
0.46
12%
1
0.77
13%

```
test$predicted.income <- predict(decision_tree, test, type = "class")
confMat <- table(test$predicted.income, test$income)
accuracy <- sum(diag(confMat))/sum(confMat)
print("Confusion Matrix")
```

```
## [1] "Confusion Matrix"
```

```
confMat
```

```
##
##         0    1
##    0 5401  977
##    1  226  937
```

```
print("Accuracy: ")
```

```
## [1] "Accuracy: "
```

```
accuracy
```

```
## [1] 0.8404721
```

I used a Decision Tree because with many variables, the model could split the dataset into smaller models to evaluate the complexity if a person could make more than 50k or not. In this case, the decision tree split relationship into occupation into education followed by their income. ACC: 84% The model recorded an 84% accuracy, which is quite high.

RESULTS ANALYSIS

Ranking the algorithms from best to worst: 1. Logistic Regression 2. Decision Tree 3. Naive Bayes

The reason why Naive Bayes was ranked last was because it had the lowest accuracy. I think this attributed to the large dataset and Naive Bayes operates on strong assumptions, so as a result it had a lower accuracy compared to the other models. The decision tree was next best because the data was split up into a binary operation that allowed the decision tree to predict if they made greater than 50k or less than 50k. A binary predictor worked best for the decision tree and as a result the model was able to predict on either binary predictions. Logistic Regression worked the best because it was predicting an income of either >50k or <50k, both of these values converted to a binary to make it easier for the Logistic Regression to predict. The model only had to predict weather, yes they made greater than 50k, or no they didnt make greater than 50k, which is why I think it had such great success. Additionally the reason why I think all the models predicted a low specificity is because their was a greater quantity to predict on for those that made less than 50k than those that made more than 50k. There are fewer occupations and certain degrees that allow the population to make more than 50k, so in that sense this could also explain why there was a high sensitivity and a low specificity.

All the model scripts were able to learn from the data and this is useful to know because if you wanted to know on certain incomes in a demographic than using these algorithms would be best, if you stuck with a binary prediction for the income. The decision tree would be best if you wanted to observe the breakdown for each attribute, but the logistic regression would be best if you wanted data to tell you if a certain demographic is making a certain income amount.