Adult Census Income Prediction

Ajish Bhaskar

June 5, 2019

1. Introduction

This documentation is my interpretaton of the data analysis and predictions done on kaggle dataset - 'Adult Census Income'. This dataset was extracted from the 1994 Census bureau database. The prediction task was to determine whether a person makes over $50K a year. I have performed three different model fitting techniques such as 'Logistic Regression', 'Classification and Regression Tree' and 'K-Nearest Neighbors' and compared the accuracy of each prediction.

1.1. Preparations for Analysis

```r
#Packages and Libraries Used
if(!require(tidyverse)) install.packages("tidyverse")
if(!require(caret)) install.packages("caret")
if(!require(ROCR)) install.packages("ROCR")
library(plyr)
library(dplyr)
library(ggplot2)
library(caret)
library(tidyr)
library(rpart)
library(ROCR)
```

1.2. Read in the Dataset

Prior to running the project, it is assumed that the data set 'adult.csv' is downloaded from kaggle with column header and saved in '/data' folder of the current working directory.

```r
adult_full <- read_csv("data/adult.csv")
```

2. Data Exploration

Let us take a look at the structure and few sample records of the dataset 'adult_full' that we just read in.

```r
str(adult_full)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 32561 obs. of  15 variables:
##  $ age           : num  90 82 66 54 41 34 38 74 68 41 ...
##  $ workclass     : chr  "?" "Private" "?" "Private" ...
##  $ fnlwgt        : num  77053 132870 186061 140359 264663 ...
##  $ education     : chr  "HS-grad" "HS-grad" "Some-college" "7th-8th" ...
##  $ education.num : num  9 9 10 4 10 9 6 16 9 10 ...
##  $ marital.status: chr  "Widowed" "Widowed" "Widowed" "Divorced" ...
##  $ occupation    : chr  "?" "Exec-managerial" "?" "Machine-op-inspct" ...
##  $ relationship  : chr  "Not-in-family" "Not-in-family" "Unmarried" "Unmarried" ...
##  $ race          : chr  "White" "White" "Black" "White" ...
##  $ sex           : chr  "Female" "Female" "Female" "Female" ...
##  $ capital.gain  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ capital.loss  : num  4356 4356 4356 3900 3900 ...
##  $ hours.per.week: num  40 18 40 40 40 45 40 20 40 60 ...
##  $ native.country: chr  "United-States" "United-States" "United-States" "United-States" ...
##  $ income        : chr  "<=50K" "<=50K" "<=50K" "<=50K" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   age = col_double(),
##   ..   workclass = col_character(),
##   ..   fnlwgt = col_double(),
##   ..   education = col_character(),
##   ..   education.num = col_double(),
##   ..   marital.status = col_character(),
##   ..   occupation = col_character(),
##   ..   relationship = col_character(),
##   ..   race = col_character(),
##   ..   sex = col_character(),
##   ..   capital.gain = col_double(),
##   ..   capital.loss = col_double(),
##   ..   hours.per.week = col_double(),
##   ..   native.country = col_character(),
##   ..   income = col_character()
##   .. )
```

```
head(adult_full)
```

```
## # A tibble: 6 x 15
##     age workclass fnlwgt education education.num marital.status occupation
##   <dbl> <chr>      <dbl> <chr>             <dbl> <chr>          <chr>
## 1    90 ?          77053 HS-grad               9 Widowed        ?
## 2    82 Private   132870 HS-grad               9 Widowed        Exec-mana~
## 3    66 ?         186061 Some-col~            10 Widowed        ?
## 4    54 Private   140359 7th-8th               4 Divorced       Machine-o~
## 5    41 Private   264663 Some-col~            10 Separated      Prof-spec~
## 6    34 Private   216864 HS-grad               9 Divorced       Other-ser~
## # ... with 8 more variables: relationship <chr>, race <chr>, sex <chr>,
## #   capital.gain <dbl>, capital.loss <dbl>, hours.per.week <dbl>,
## #   native.country <chr>, income <chr>
```

There are 32561 observation of 15 variables.

3. Data Wrangling

After a brief look at the dataset, it appears that we have a few data problems that we need to address.

3.1 Missing Data

We can see that there are some missing data represented by "?" in the dataset. As a first set replace the "?" with NA.

```
for (i in 1:ncol(adult_full)) {adult_full[,i][adult_full[,i] == '?',] = NA}
```

After the replace, we have to find out where the missing values are using this simple function.

```r
sapply(adult_full, FUN = function(adult_full_cols) {
  table(is.na(adult_full_cols))
})
```

```
## $age
##
## FALSE
## 32561
##
## $workclass
##
## FALSE  TRUE
## 30725  1836
##
## $fnlwgt
##
## FALSE
## 32561
##
## $education
##
## FALSE
## 32561
##
## $education.num
##
## FALSE
## 32561
##
## $marital.status
##
## FALSE
## 32561
##
## $occupation
##
## FALSE  TRUE
## 30718  1843
##
## $relationship
##
## FALSE
## 32561
##
## $race
##
## FALSE
## 32561
##
## $sex
##
## FALSE
## 32561
##
```

```
## $capital.gain
##
## FALSE
## 32561
##
## $capital.loss
##
## FALSE
## 32561
##
## $hours.per.week
##
## FALSE
## 32561
##
## $native.country
##
## FALSE  TRUE
## 31978   583
##
## $income
##
## FALSE
## 32561
```

We see that the missing values are in the columns workclass, occupation, native.country. Now let us see if the missing data in those columns are related. In census domain when we miss an attribute, it is very much possible that we miss one or more related attributes. So let us first start with workclass and occupation.

```
table(which(is.na(adult_full['workclass'])) == which(is.na(adult_full['occupation'])))
```

```
##
## FALSE  TRUE
##  1342   501
```

```
table(which(is.na(adult_full['workclass'])) %in% which(is.na(adult_full['occupation'])))
```

```
##
## TRUE
## 1836
```

So whenever we have a NA in the workclass, we have NA in occupation most of the time.

We can see that most of the missing data is from native.country 'United-States' and most of them have income less than $50K.

```
table(adult_full[is.na(adult_full['workclass']),]$native.country)
```

```
##
##           Cambodia             Canada              China
##                  1                 14                  7
##           Columbia               Cuba Dominican-Republic
##                  3                  3                  3
##            Ecuador        El-Salvador            England
##                  1                  6                  4
##             France            Germany          Guatemala
##                  2                  9                  1
##              Haiti           Honduras               Hong
##                  2                  1                  1
##               Iran              Italy            Jamaica
##                  1                  5                  1
##              Japan               Laos             Mexico
##                  3                  1                 33
##          Nicaragua               Peru        Philippines
##                  1                  1                 10
##             Poland           Portugal        Puerto-Rico
##                  4                  3                  5
##           Scotland              South             Taiwan
##                  1                  9                  9
##           Thailand    Trinadad&Tobago      United-States
##                  1                  1               1659
##            Vietnam
##                  3
```

```
table(adult_full[is.na(adult_full['workclass']),]$income)
```

```
##
## <=50K  >50K
##  1645   191
```

Let us see how any such missing values that we have to address.

```
sum(is.na(adult_full))
```

```
## [1] 4262
```

Some models that we apply need that the predictors must not have missing values. So we are removing those from the data.

```
adult_full <- na.omit(adult_full)
adult_full <- data.frame(adult_full)
```

3.2 Collapsing Levels

Having looked closely at workclass, we see that we can collapse workclass to more meaningful categories.

```
adult_full[adult_full$workclass %in% c("Federal-gov","Local-gov","State-gov"),"workclass"] <- "Governme
adult_full[adult_full$workclass %in% c("Self-emp-inc","Self-emp-not-inc"),"workclass"] <- "Self-Employe
adult_full[adult_full$workclass %in%  c("Never-worked","Without-pay","Other","Unknown"),"workclass"] <-
```

Similarly we can collapse the native.country field to smaller levels.

```
adult_full[adult_full$native.country %in% c("Vietnam","Laos","Cambodia","Thailand"), "native.country"]
adult_full[adult_full$native.country %in% c("South"), "native.country"] <- "unknown"
adult_full[adult_full$native.country %in% c("China","India","HongKong","Iran","Philippines","Taiwan", "
adult_full[adult_full$native.country %in% c("Canada","Mexico","Puerto-Rico","United-States"), "native.c
adult_full[adult_full$native.country %in% c("Ecuador","Peru","Columbia","Trinadad&Tobago", "Cuba","Domi
adult_full[adult_full$native.country %in% c("France","Germany","Greece","Holand-Netherlands","Italy","Hu
adult_full[adult_full$native.country %in% c("Outlying-US(Guam-USVI-etc)"), "native.country"] <- "Oceania
```

We can tabularize the country data as below

```
table(adult_full$native.country)
```

```
##
##         Asia       Europe         Hong NorthAmerica      Oceania
##          499           493            19         28330             14
##        SEAsia SouthAmerica       unknown
##          116           620            71
```

3.3 Recoding of Variable

For ease of analysis and prediction we will also recode ‘income’ to be 0 if ‘<=50K’ or 1 otherwise.

```
adult_full$income <- ifelse(adult_full$income == "<=50K", 0, 1)
adult_full$income <- as.factor(adult_full$income)

levels(adult_full$income)
```

```
## [1] "0" "1"
```

Years of education can be discarded as we get the info from the education.num. Similarly the fnlwgt is a continous variable and is not very helpful. Same is the case with relationship as we can infer using marital status.

```
adult_full$education <- NULL
adult_full$fnlwgt <- NULL
adult_full$relationship <- NULL
```

Let’s take a look at the adult_full dataset after the data wrangling steps.

```
str(adult_full)
```

```
## 'data.frame':    30162 obs. of  12 variables:
##  $ age           : num  82 54 41 34 38 74 68 45 38 52 ...
##  $ workclass     : chr  "Private" "Private" "Private" "Private" ...
##  $ education.num : num  9 4 10 9 6 16 9 16 15 13 ...
##  $ marital.status: chr  "Widowed" "Divorced" "Separated" "Divorced" ...
##  $ occupation    : chr  "Exec-managerial" "Machine-op-inspct" "Prof-specialty" "Other-service" ...
##  $ race          : chr  "White" "White" "White" "White" ...
##  $ sex           : chr  "Female" "Female" "Female" "Female" ...
##  $ capital.gain  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ capital.loss  : num  4356 3900 3900 3770 3770 ...
##  $ hours.per.week: num  18 40 40 45 40 20 40 35 45 20 ...
##  $ native.country: chr  "NorthAmerica" "NorthAmerica" "NorthAmerica" "NorthAmerica" ...
##  $ income        : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 2 2 2 ...
```

```
head(adult_full)
```

```
##   age  workclass education.num marital.status        occupation  race
## 1  82    Private             9        Widowed   Exec-managerial White
## 2  54    Private             4       Divorced Machine-op-inspct White
## 3  41    Private            10      Separated    Prof-specialty White
## 4  34    Private             9       Divorced     Other-service White
## 5  38    Private             6      Separated      Adm-clerical White
## 6  74 Government            16  Never-married    Prof-specialty White
##      sex capital.gain capital.loss hours.per.week native.country income
## 1 Female            0         4356             18   NorthAmerica      0
## 2 Female            0         3900             40   NorthAmerica      0
## 3 Female            0         3900             40   NorthAmerica      0
## 4 Female            0         3770             45   NorthAmerica      0
## 5   Male            0         3770             40   NorthAmerica      0
## 6 Female            0         3683             20   NorthAmerica      1
```

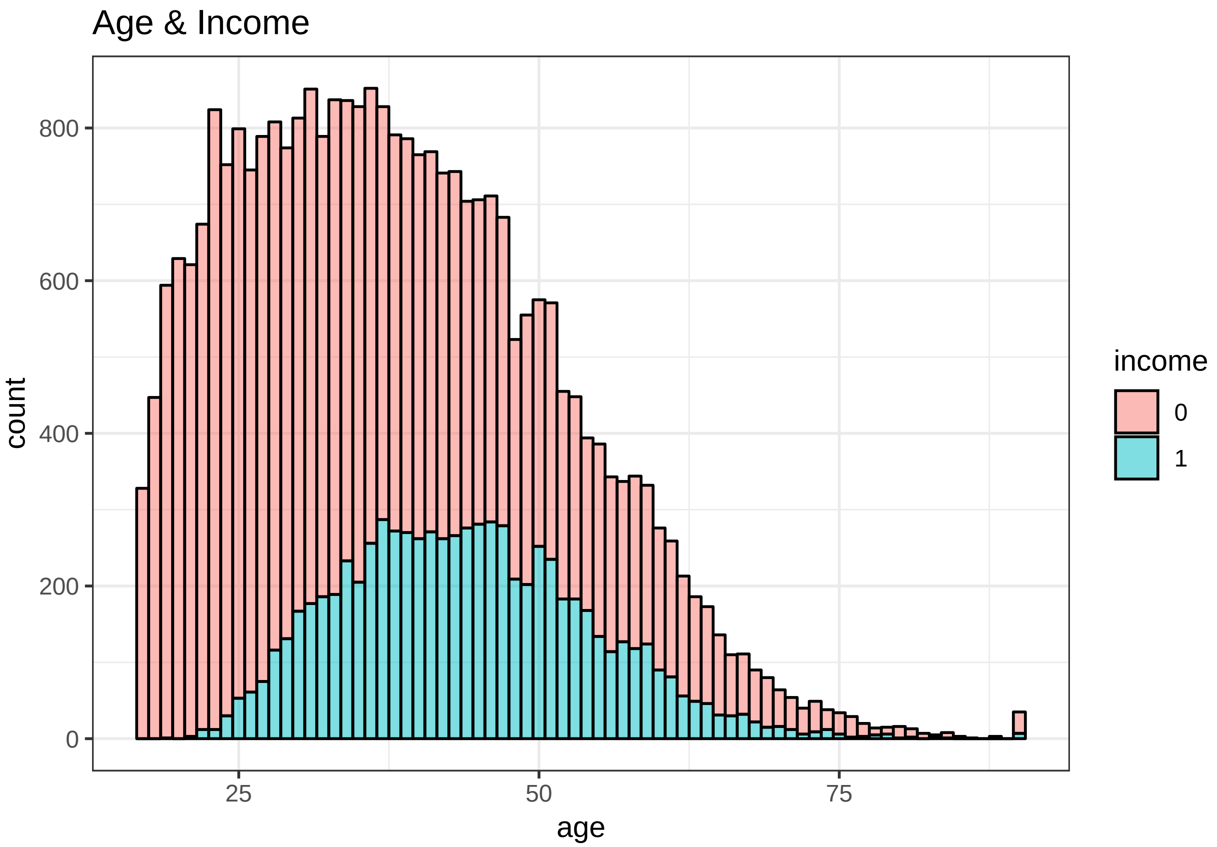
4. Data Analysis and Visualization

As income is the outcome that we want to predict, we will explore the relationship between each variables with income.

4.1 Age and Income

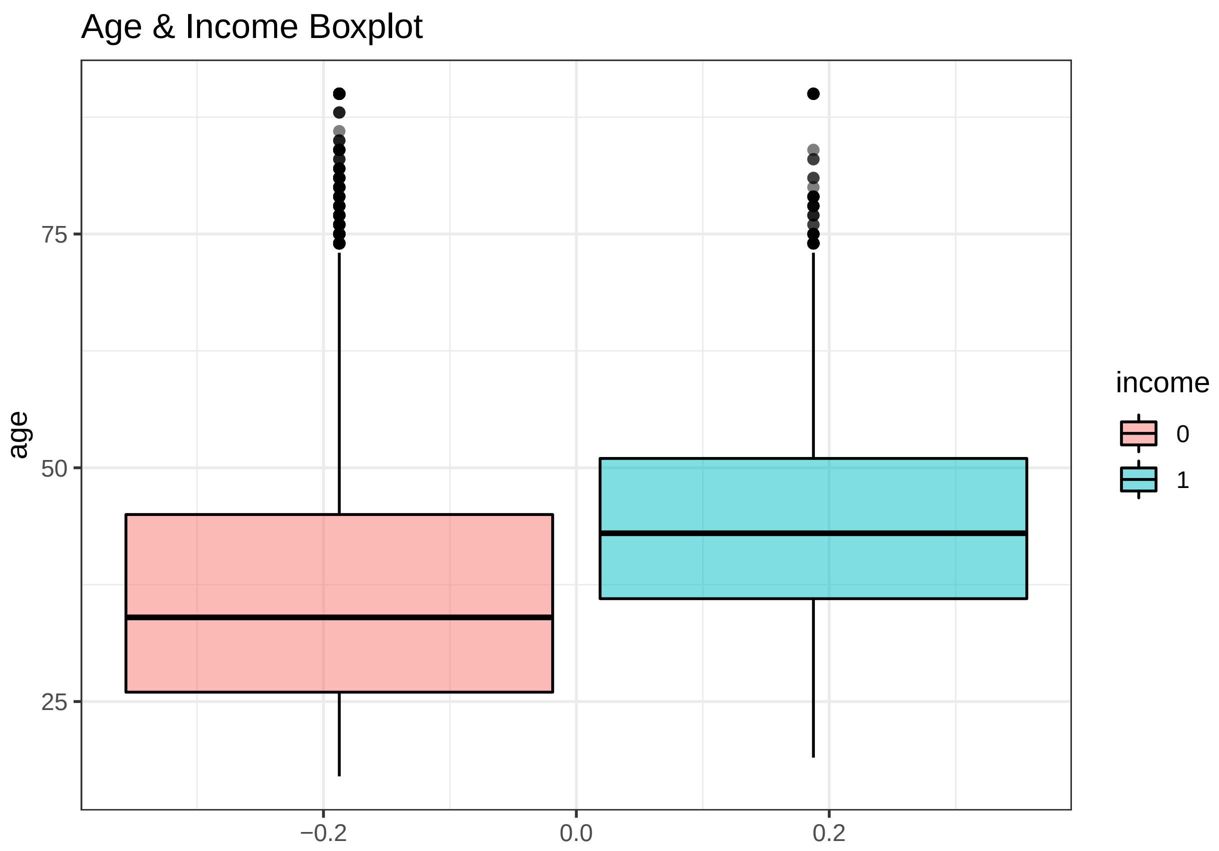
A simple histogram of age and income reveals that yonger people appear to have lower income and the spread of higher income appears mostly between people at 40’s - 50’s.

```
adult_full %>% ggplot(aes(age)) +
  geom_histogram(aes(fill=income),color='black',binwidth=1,alpha=0.5) +
  theme_bw() + ggtitle("Age & Income")
```



We can confirm it with a box plot.

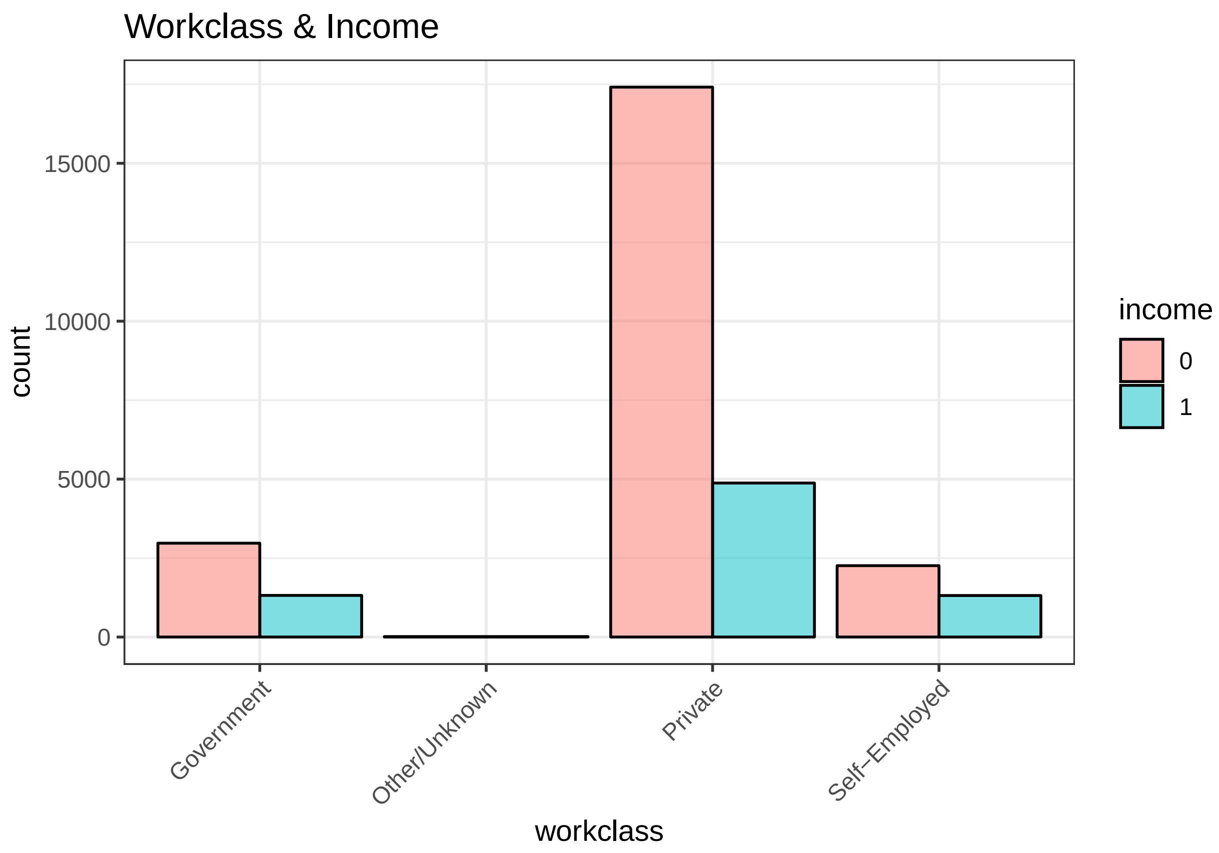
```
adult_full %>% ggplot() + geom_boxplot(aes(y=age,fill=income),color='black',alpha=0.5) +
theme_bw() + ggtitle("Age & Income Boxplot")
```



4.2 Workclass and Income

We can see that the people in workclass 'Private' clearly stand out, using the below plot.

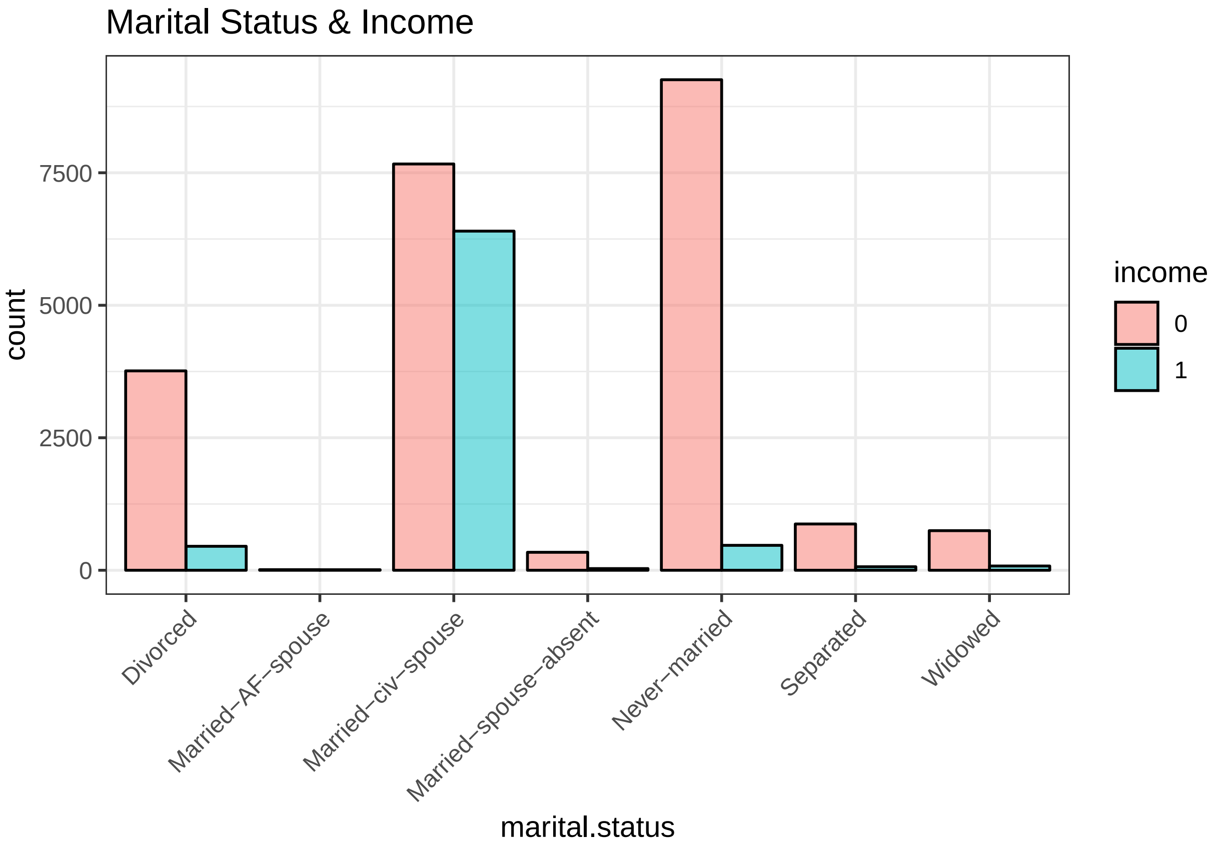
```
ggplot(adult_full,aes(workclass,group=income)) + geom_bar(aes(fill=income),color='black',alpha=0.5, posi
theme_bw()+   theme(axis.text.x = element_text(angle = 45, hjust = 1)) + ggtitle("Workclass & Income")
```



4.3 Marital Status and Income

People with marital status ‘Married-civ-spouse’ clearly shows a higher propotion of income.

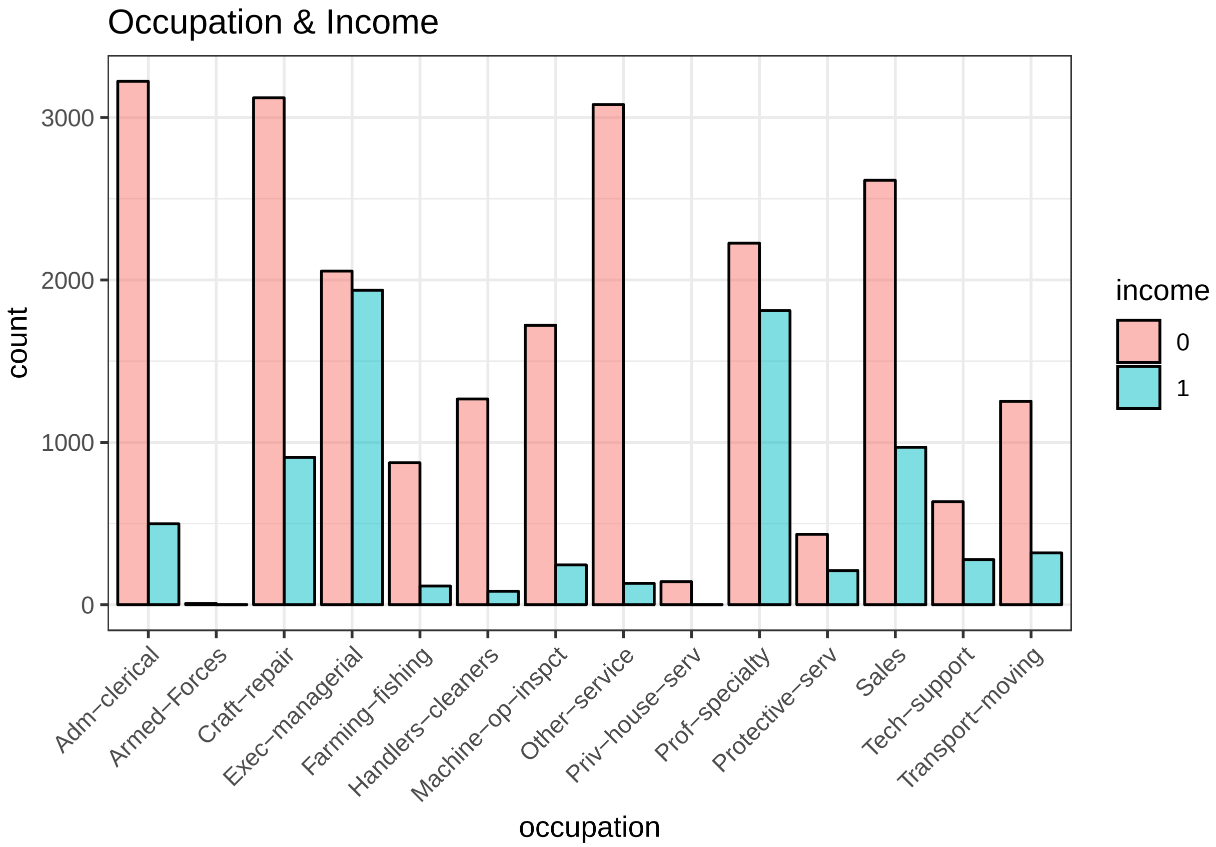
```
ggplot(adult_full,aes(marital.status,group=income)) + geom_bar(aes(fill=income),color='black',alpha=0.5
theme_bw()+ theme(axis.text.x = element_text(angle = 45, hjust = 1)) + ggtitle("Marital Status & Income
```



4.4 Occupation and Income

It is observed that people employed in ‘Exec-managerial’ and ‘Prof-specialty’ have a higher propotion of income more than 50K.

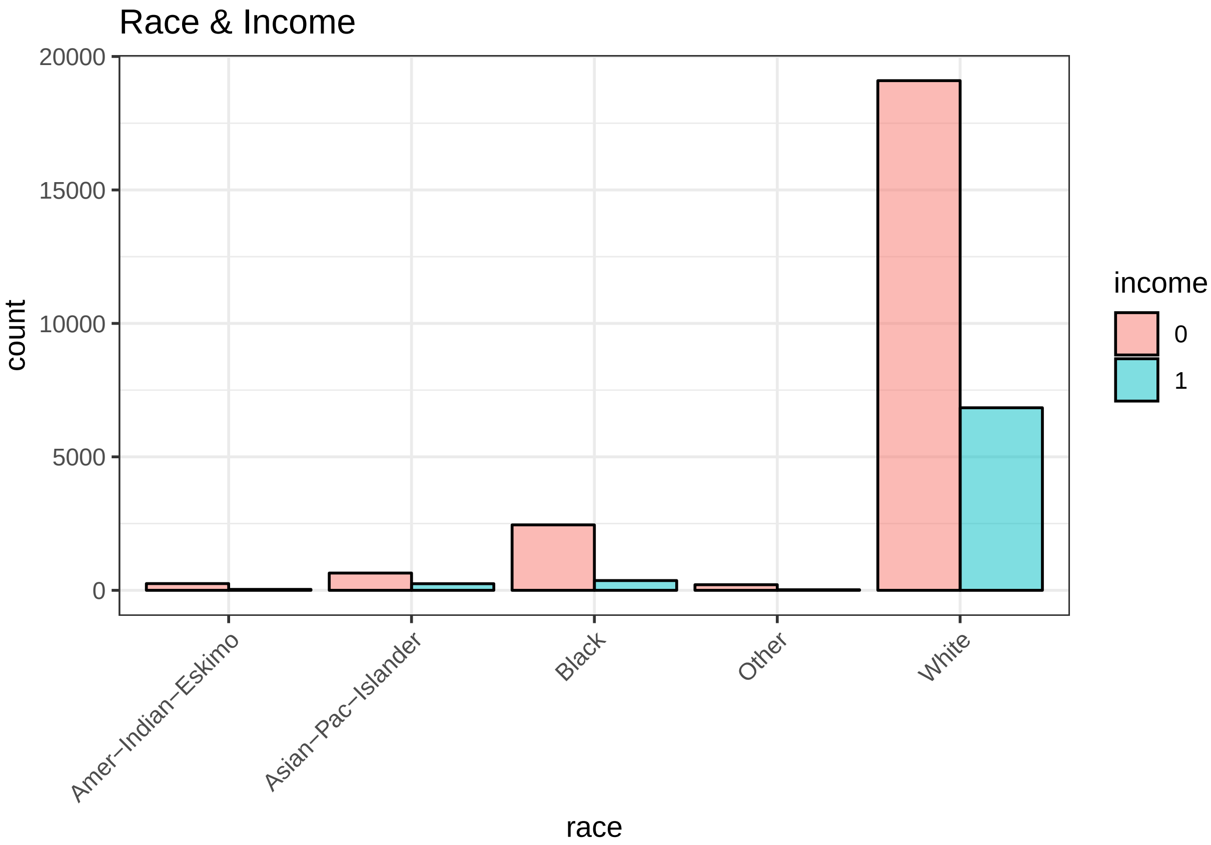
```
ggplot(adult_full,aes(occupation,group=income)) + geom_bar(aes(fill=income),color='black',alpha=0.5, pos
theme_bw()+ theme(axis.text.x = element_text(angle = 45, hjust = 1)) + ggtitle("Occupation & Income")
```



4.5 Race and Income

People identified as 'white' have higher representation in the data collected.

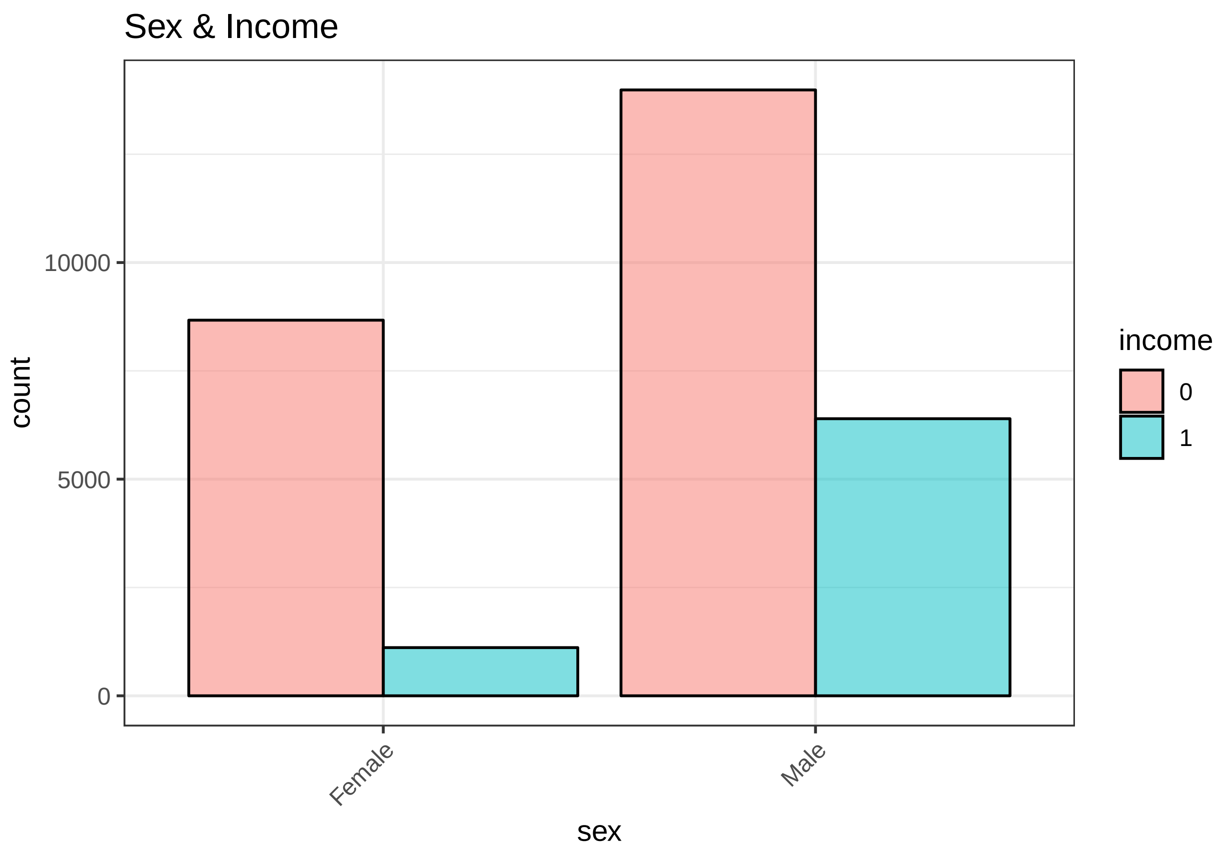
```
ggplot(adult_full,aes(race,group=income)) + geom_bar(aes(fill=income),color='black',alpha=0.5, position
theme_bw()+ theme(axis.text.x = element_text(angle = 45, hjust = 1)) + ggtitle("Race & Income")
```



4.6 Sex and Income

The gender male has a higher propotion of people earning more than 50K.

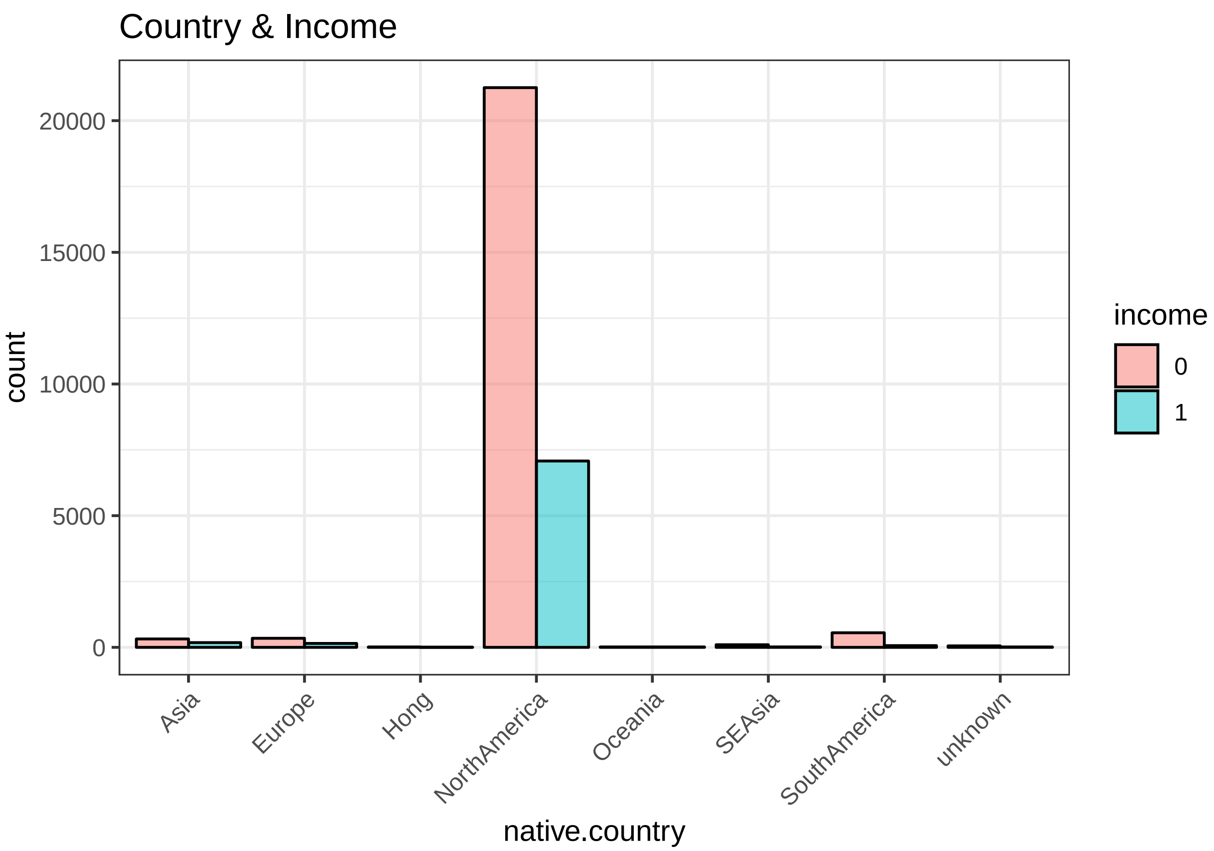
```
ggplot(adult_full,aes(sex,group=income)) + geom_bar(aes(fill=income),color='black',alpha=0.5, position=
theme_bw()+ theme(axis.text.x = element_text(angle = 45, hjust = 1)) + ggtitle("Sex & Income")
```



4.7 Native Country and Income

The data collected has a bias to NorthAmerica, as it is from US census. So this may not be a good predictor.

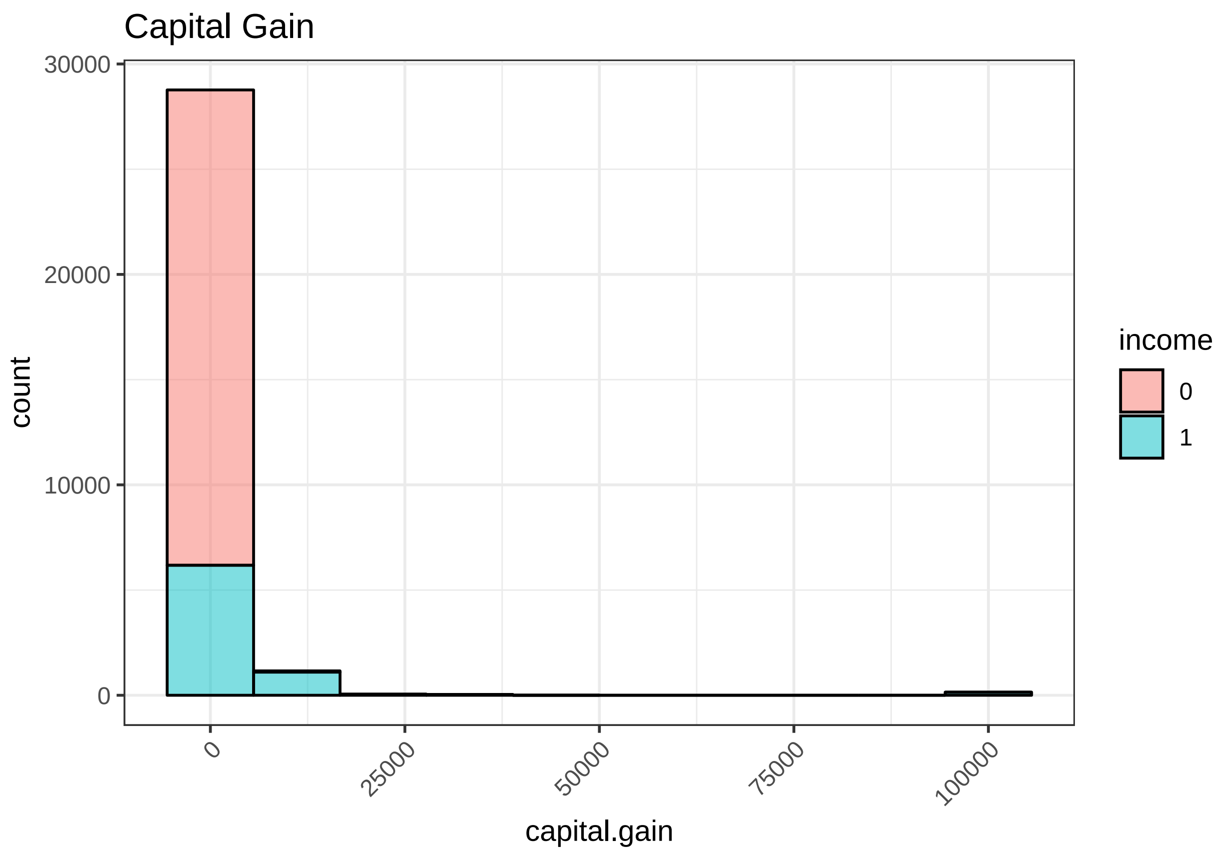
```
ggplot(adult_full,aes(native.country,group=income)) + geom_bar(aes(fill=income),color='black',alpha=0.5
theme_bw()+ theme(axis.text.x = element_text(angle = 45, hjust = 1)) + ggtitle("Country & Income")
```



4.8 Capital Gain and Income

A quick plot of the capital gain shows that the distribution is skewed.

```
ggplot(adult_full,aes(capital.gain,group=income)) + geom_histogram(aes(fill=income),bins=10, color='bla
theme_bw() + theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  ggtitle("Capital Gain")
```



Also it is worth noting that more than 90% of the people have reported 0 capital gain.

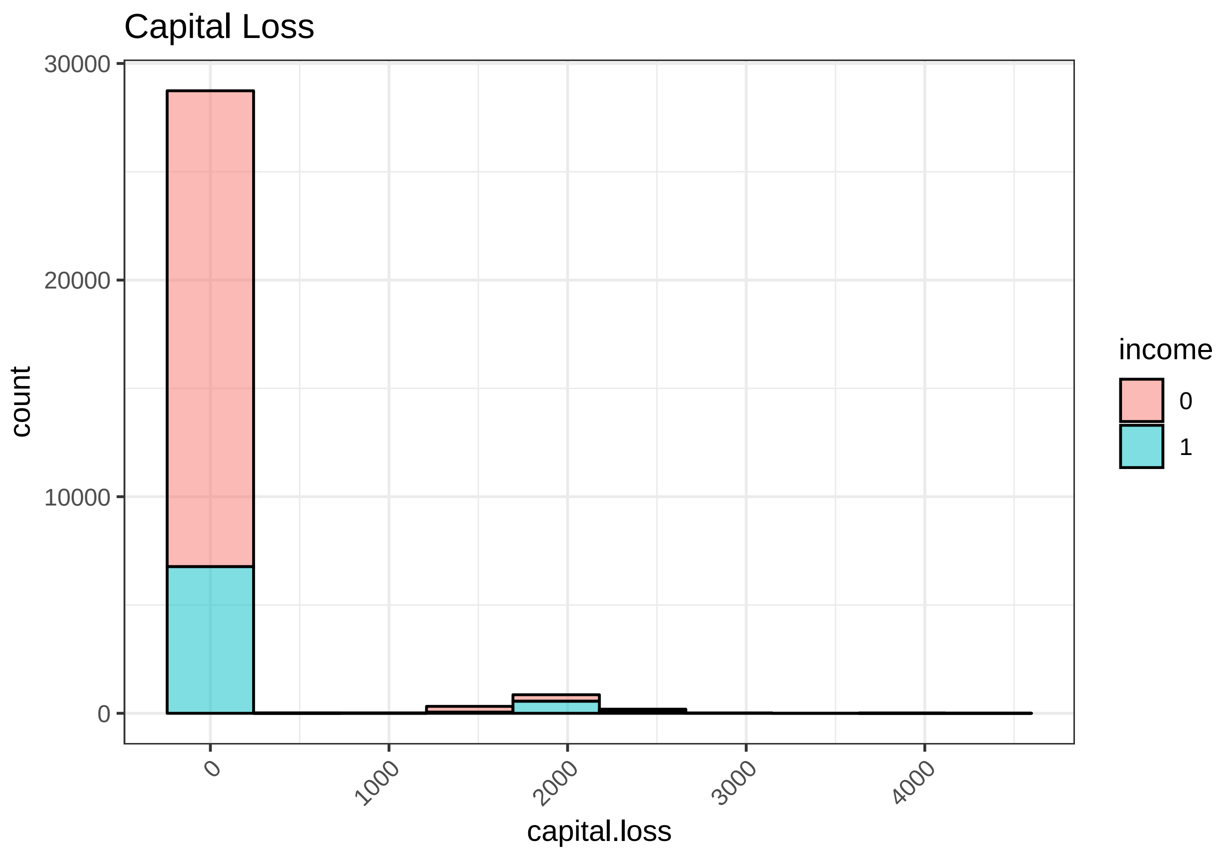
```r
sum(adult_full$capital.gain == 0)/length(adult_full$capital.gain) * 100
```

```
## [1] 91.58544
```

4.9 Capital Loss and Income

We see a similar observation on capital loss as capital gain.

```r
ggplot(adult_full,aes(capital.loss,group=income)) + geom_histogram(aes(fill=income),bins=10, color='bla
theme_bw()+ theme(axis.text.x = element_text(angle = 45, hjust = 1)) + ggtitle("Capital Loss")
```



Also more than 95% of the people have reported 0 capital loss

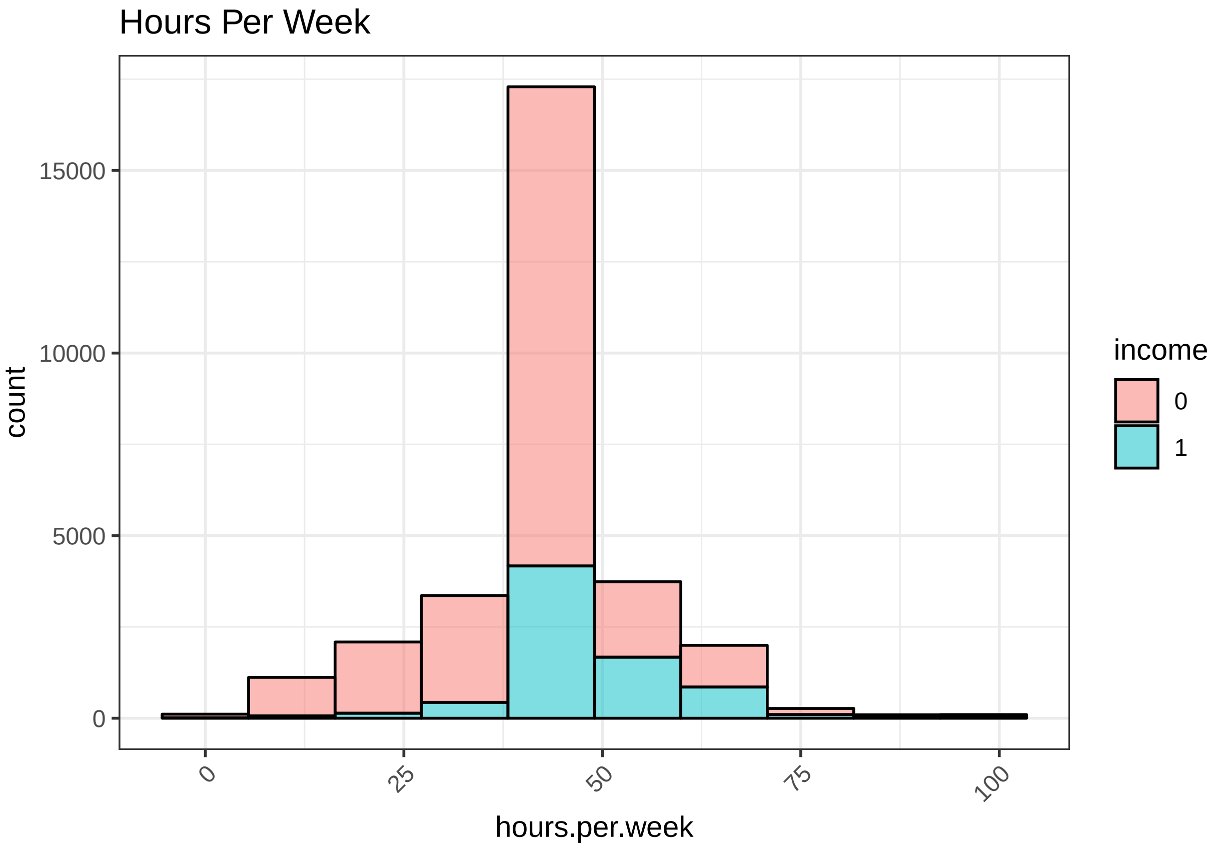
```
sum(adult_full$capital.loss == 0)/length(adult_full$capital.loss) * 100
```

```
## [1] 95.26888
```

4.10 Hours per week and Income

A histogram of hours per week reveals that majority of people are working 40-50 hours per week and hence the distibution is also accordingly.

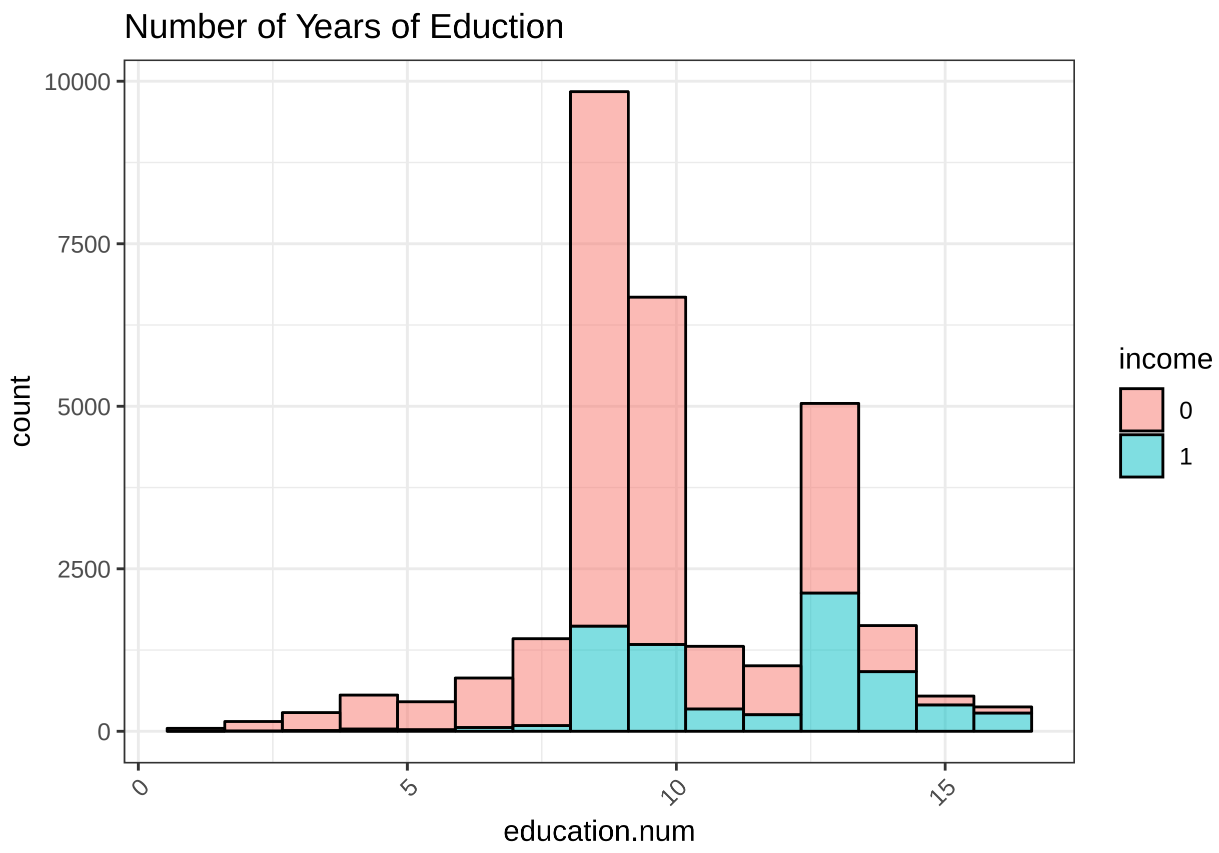
```
ggplot(adult_full,aes(hours.per.week,group=income)) + geom_histogram(aes(fill=income),bins=10, color='bl
theme_bw()+ theme(axis.text.x = element_text(angle = 45, hjust = 1)) + ggtitle("Hours Per Week")
```



4.11 Years of Eduction and Income

It is observed that propotion of people with higher education tend to earn more than 50K than the less educated people.

```
ggplot(adult_full,aes(education.num,group=income)) + geom_histogram(aes(fill=income),bins=15, color='bla
theme_bw()+ theme(axis.text.x = element_text(angle = 45, hjust = 1)) + ggtitle("Number of Years of Educ
```



5. Model Fitting

Before we start the actual model fitting, we will start by splitting the 'adult_full' dataset into two. 75% percent of the dataset will go to a training set known as 'adult_train' which we will use for model fitting and the remaining 25% will go to a validation set known as 'adult_test'.

```
set.seed(1234)

test_index <- createDataPartition(y = adult_full$income, times = 1, p = 0.25, list = FALSE)
adult_test <- adult_full[test_index,]
adult_train <- adult_full[-test_index,]
```

We can verify the dimensions of the two as below.

```
dim(adult_test)
```

```
## [1] 7541   12
```

```
dim(adult_train)
```

```
## [1] 22621    12
```

5.1 Classification And Regression Trees

Let's first build our prediction model using the Classification And Regression Trees (CART) technique. We will be using the adult_train dataset to build the model and the perform the prediction of income on adult_train dataset. We will be using all the available predictors for coming up with a model to predict income. We will also find the probability of the prediction using the type "prob" for later use.

```
tree_adult <- rpart(income ~ ., data = adult_train, method = 'class', minbucket=20)
tree_pred <- predict(tree_adult, newdata = adult_test, type = 'class')
tree_pred_prob <- predict(tree_adult, adult_test, type = "prob")
```

Let's verify the accuracy, sensitivity and specificity using confusion matrix.

```
confusionMatrix(adult_test$income, tree_pred)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 5360  304
##          1  941  936
##
##                Accuracy : 0.8349
##                  95% CI : (0.8263, 0.8432)
##     No Information Rate : 0.8356
##     P-Value [Acc > NIR] : 0.5692
##
##                   Kappa : 0.5019
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.8507
##             Specificity : 0.7548
##          Pos Pred Value : 0.9463
##          Neg Pred Value : 0.4987
##              Prevalence : 0.8356
##          Detection Rate : 0.7108
##    Detection Prevalence : 0.7511
##       Balanced Accuracy : 0.8027
##
##        'Positive' Class : 0
##
```

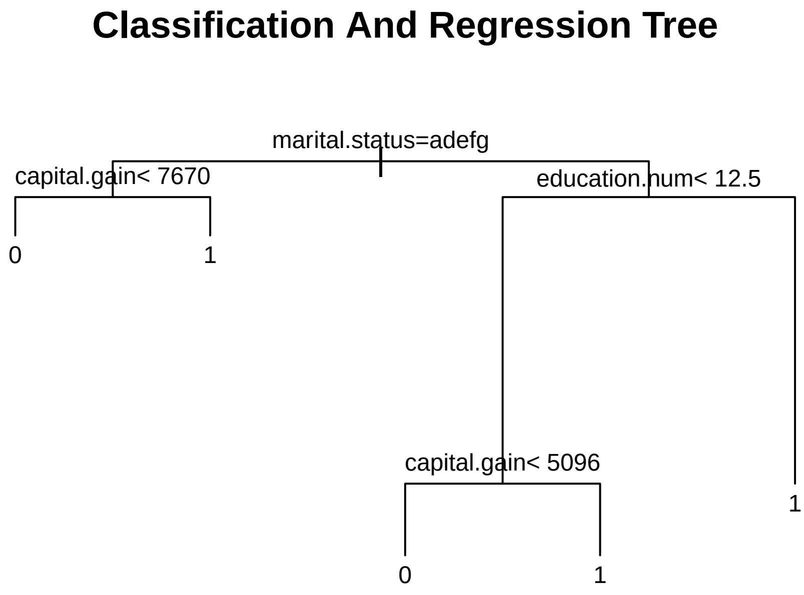
Also verify the error rate by a table function using predicted object and actual income found in the test dataset.

```
tab_tree <- table(tree_pred, adult_test$income)
error_tree = 1 - sum(tab_tree[row(tab_tree)==col(tab_tree)])/sum(tab_tree)
error_tree
```

```
## [1] 0.1650975
```

The above results look very good. Let's plot the tree and see the predictors and decisions made.

```
plot(tree_adult, margin = 0.1)
text(tree_adult, cex = 0.75)
title("Classification And Regression Tree")
```



5.2 Logistic Regression

Now use the logistic regression model fitting for predicting the income in the test dataset. We will use all the predictors like before.

```
glm_fit <- glm(income ~ ., family = binomial(logit), data = adult_train)
glm_pred<- predict(glm_fit, adult_test, type = "response")
```

Let's run a summary statistics and see what are the significant predictors

```
summary(glm_fit)
```

```
##
## Call:
## glm(formula = income ~ ., family = binomial(logit), data = adult_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.1170  -0.5180  -0.2111  -0.0096   3.8341
##
## Coefficients:
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                        -9.207e+00  3.945e-01 -23.337  < 2e-16
## age                                 2.727e-02  1.901e-03  14.343  < 2e-16
## workclassOther/Unknown             -1.189e+01  1.516e+02  -0.078  0.93750
## workclassPrivate                    3.924e-02  6.158e-02   0.637  0.52403
## workclassSelf-Employed             -2.263e-01  8.103e-02  -2.793  0.00523
## education.num                        2.874e-01  1.100e-02  26.116  < 2e-16
## marital.statusMarried-AF-spouse      3.324e+00  5.553e-01   5.985 2.16e-09
## marital.statusMarried-civ-spouse     2.167e+00  7.883e-02  27.486  < 2e-16
## marital.statusMarried-spouse-absent  3.172e-02  2.716e-01   0.117  0.90702
## marital.statusNever-married         -5.159e-01  9.664e-02  -5.338 9.38e-08
## marital.statusSeparated              7.236e-02  1.751e-01   0.413  0.67952
## marital.statusWidowed               -2.262e-01  1.875e-01  -1.206  0.22771
## occupationArmed-Forces               1.310e-01  1.972e+00   0.066  0.94704
## occupationCraft-repair               9.372e-03  9.034e-02   0.104  0.91738
## occupationExec-managerial            7.872e-01  8.680e-02   9.069  < 2e-16
## occupationFarming-fishing           -1.138e+00  1.582e-01  -7.191 6.41e-13
## occupationHandlers-cleaners         -7.135e-01  1.639e-01  -4.353 1.34e-05
## occupationMachine-op-inspct         -2.838e-01  1.153e-01  -2.462  0.01383
## occupationOther-service             -9.037e-01  1.359e-01  -6.650 2.92e-11
## occupationPriv-house-serv           -3.804e+00  1.976e+00  -1.925  0.05417
## occupationProf-specialty             5.380e-01  8.924e-02   6.029 1.65e-09
## occupationProtective-serv            3.101e-01  1.429e-01   2.171  0.02993
## occupationSales                      2.397e-01  9.331e-02   2.569  0.01021
## occupationTech-support               5.729e-01  1.260e-01   4.548 5.42e-06
## occupationTransport-moving          -7.093e-02  1.121e-01  -0.633  0.52673
## raceAsian-Pac-Islander               9.087e-01  3.133e-01   2.900  0.00373
## raceBlack                            4.627e-01  2.667e-01   1.735  0.08277
## raceOther                           -6.844e-02  4.402e-01  -0.155  0.87644
## raceWhite                            5.732e-01  2.538e-01   2.258  0.02393
## sexMale                              1.443e-01  6.212e-02   2.323  0.02017
## capital.gain                         3.210e-04  1.209e-05  26.543  < 2e-16
## capital.loss                         6.576e-04  4.396e-05  14.960  < 2e-16
## hours.per.week                       3.017e-02  1.937e-03  15.576  < 2e-16
## native.countryEurope                 4.180e-01  2.663e-01   1.570  0.11646
## native.countryHong                  -2.910e-01  8.127e-01  -0.358  0.72033
## native.countryNorthAmerica           2.698e-01  2.234e-01   1.207  0.22726
## native.countryOceania               -1.107e+01  1.709e+02  -0.065  0.94835
## native.countrySEAsia                -2.380e-01  3.760e-01  -0.633  0.52676
## native.countrySouthAmerica          -2.998e-01  2.954e-01  -1.015  0.31013
## native.countryunknown               -1.393e+00  5.136e-01  -2.712  0.00669
##
## (Intercept)                         ***
## age                                 ***
## workclassOther/Unknown
## workclassPrivate
## workclassSelf-Employed              **
## education.num                       ***
## marital.statusMarried-AF-spouse     ***
## marital.statusMarried-civ-spouse    ***
## marital.statusMarried-spouse-absent
## marital.statusNever-married         ***
## marital.statusSeparated
## marital.statusWidowed
## occupationArmed-Forces
## occupationCraft-repair
## occupationExec-managerial           ***
## occupationFarming-fishing           ***
## occupationHandlers-cleaners         ***
## occupationMachine-op-inspct         *
## occupationOther-service             ***
## occupationPriv-house-serv           .
## occupationProf-specialty            ***
## occupationProtective-serv           *
## occupationSales                     *
## occupationTech-support              ***
## occupationTransport-moving
## raceAsian-Pac-Islander              **
## raceBlack                           .
## raceOther
## raceWhite                           *
## sexMale                             *
## capital.gain                        ***
## capital.loss                        ***
## hours.per.week                      ***
## native.countryEurope
## native.countryHong
## native.countryNorthAmerica
## native.countryOceania
## native.countrySEAsia
## native.countrySouthAmerica
## native.countryunknown               **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 25388  on 22620  degrees of freedom
## Residual deviance: 14897  on 22581  degrees of freedom
## AIC: 14977
##
## Number of Fisher Scoring iterations: 12
```

We will also find the error rate using a table function like before.

```
tab_glm <- table(actual= adult_test$income, predicted= glm_pred>0.5)
error_glm = 1 - sum(tab_glm[row(tab_glm)==col(tab_glm)])/sum(tab_glm)
error_glm
```

```
## [1] 0.1589975
```

As we can see, we get an improved results using logistic modeling techniques.

5.3 K-Nearest Neighbors

Now for the last modeling, we will use the K-Nearest Neighbor (KNN). As with the previous methods we will be using all available predictors.

```
fit_knn <- knn3(income ~ ., adult_train,  k = 5, prob=TRUE)
knn_pred <- predict(fit_knn, adult_test, type="class")
knn_pred_prob <- predict(fit_knn, adult_test, type = "prob")
```

Let's verify the accuracy, sensitivity and specificity using confusion matrix.

```
confusionMatrix(adult_test$income, knn_pred)
```

```
## Confusion Matrix and Statistics
```

```
##
##           Reference
## Prediction    0    1
##          0 5164  500
##          1  732 1145
##
##                Accuracy : 0.8366
##                  95% CI : (0.8281, 0.8449)
##     No Information Rate : 0.7819
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5442
##
##  Mcnemar's Test P-Value : 4.666e-11
##
##             Sensitivity : 0.8758
##             Specificity : 0.6960
##          Pos Pred Value : 0.9117
##          Neg Pred Value : 0.6100
##              Prevalence : 0.7819
##          Detection Rate : 0.6848
##    Detection Prevalence : 0.7511
##       Balanced Accuracy : 0.7859
##
##        'Positive' Class : 0
##
```

Finally, verify the error rate using the table function like below. The results look pretty close to that of logistic regression.

```
tab_knn <- table(knn_pred, adult_test$income)
error_knn = 1 - sum(tab_knn[row(tab_knn)==col(tab_knn)])/sum(tab_knn)
error_knn
```

```
## [1] 0.1633736
```

6 Results

As we saw, the majority of observations in the data set has income less than $50,000 a year, sensitivity and specificity contribute to the overall accuracy. we can get different sensitivity and specificity for each model. For this reason, a very common approach to evaluating methods is to compare them graphically by plotting them. A commonly used plot that does this is the Receiver Operating Characteristic (ROC) curve.ROC curve is a plot of true positive rate against false positive rate under all threshold values. The different classifiers are compared using ROC curve.

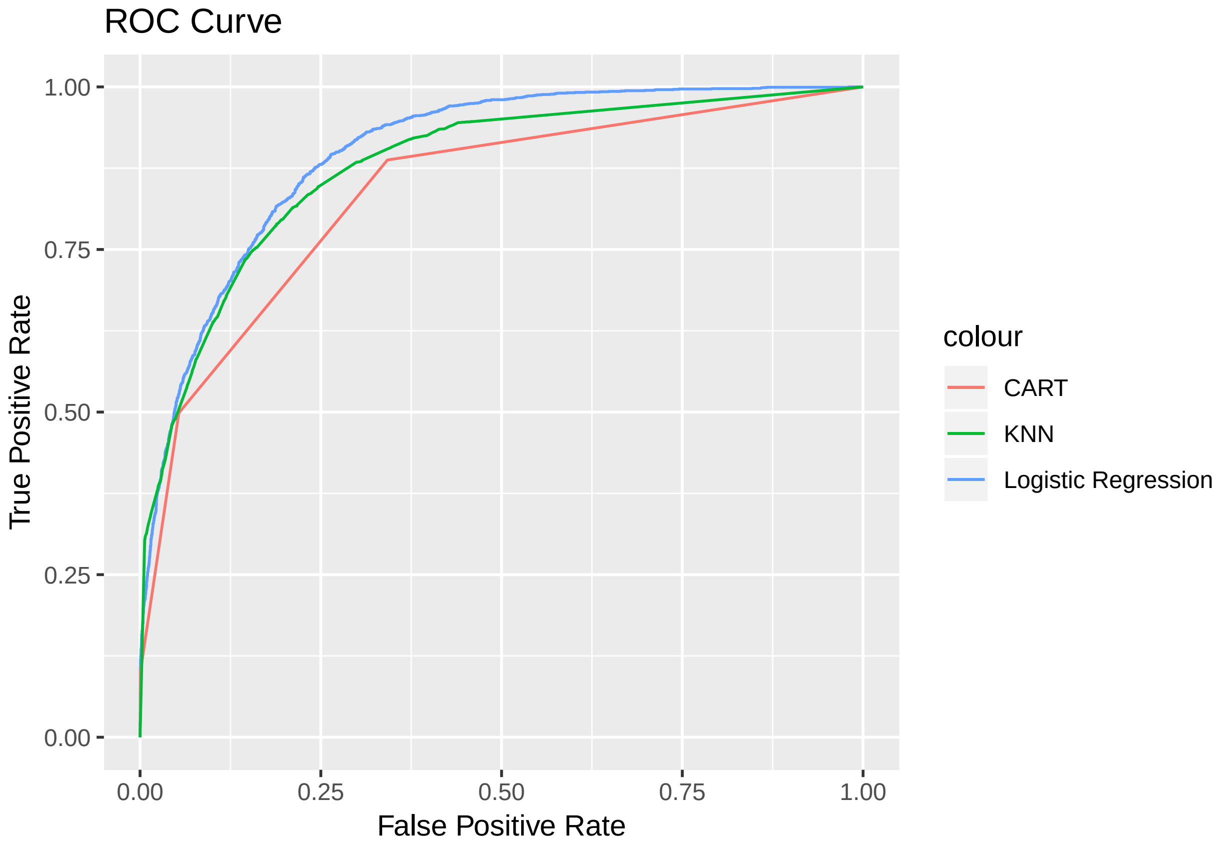
Inorder to plot the curves, let's create prediction objects and dataframes for True Positives (TP) and False Positives (FP) for the three models.

```
#Create a prediction object & dataframe for Logistic
pr <- prediction(glm_pred, adult_test$income)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
dd <- data.frame(FP = prf@x.values[[1]], TP = prf@y.values[[1]])

#Create a prediction object & dataframe for CART
pr2 <- prediction(tree_pred_prob[,2], adult_test$income)
prf2 <- performance(pr2, measure = "tpr", x.measure = "fpr")
```

```
dd2 <- data.frame(FP = prf2@x.values[[1]], TP = prf2@y.values[[1]])

#Create a prediction object & dataframe for KNN
pr3 <- prediction(knn_pred_prob[,2], adult_test$income)
prf3 <- performance(pr3, measure = "tpr", x.measure = "fpr")
dd3 <- data.frame(FP = prf3@x.values[[1]], TP = prf3@y.values[[1]])
```

Plot the curves using geom_line with the data frames built above.

```
# plot ROC curves for the three prediction objects
ggplot() +
  geom_line(data = dd, aes(x = FP, y = TP, color = 'Logistic Regression')) +
  geom_line(data = dd2, aes(x = FP, y = TP, color = 'CART')) +
  geom_line(data = dd3, aes(x = FP, y = TP, color = 'KNN')) +
  ggtitle('ROC Curve') +
  labs(x = 'False Positive Rate', y = 'True Positive Rate')
```



Now we will find the Area Under the ROCs Curve (AUC) which is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is.

```
# Area Under ROC curve represents the accuracy
auc <- rbind( performance(pr, measure = 'auc')@y.values[[1]],
              performance(pr2, measure = 'auc')@y.values[[1]],
              performance(pr3, measure = 'auc')@y.values[[1]]
              )
```

```r
rownames(auc) <- (c('Logistic Regression', 'CART', 'KNN'))
colnames(auc) <- 'Area Under ROC Curve'
round(auc, 4)
```

```
##                     Area Under ROC Curve
## Logistic Regression               0.8986
## CART                              0.8368
## KNN                               0.8772
```

7. Conclusion

As seen from the AUC, the three models performed relatively close to each other. The model built using the Logistic Regression performed best followed by KNN. The CART method gave the least accurate prediction among the three model that we did. The goal of this project was to find a model that accurately predicts if an individual makes more than $50K a year. With a higher AUC the Logistic Regression wins as the best method that we applied. Thank you!