Predict whether Income exceeds $50K/yr based on census data

Daphne Chen

7/08/2021

#Introduction

Data Set Information Adult Census Income: https://www.kaggle.com/uciml/adult-census-income This data was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0)).

Objective: The prediction task is to determine whether a person makes over $50K a year.

The methods we will be using in this project to predict income will be Logistic Regression and Decision Tree.

#Download Data and library This dataset has 32,561 entries with 15 variables.

Understand Data

There are some missing data in this dataset. Missing data is showing up as '?'. We will replace missing data with NA.

Capital_gain and capital_loss are investment income or loss. fnlwgt represents final weight. education_num is the number of years of education in total. relationship is the member's role in the family.

```
head(rawData)
```

```
## # 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>
```

```
summary(rawData)
```

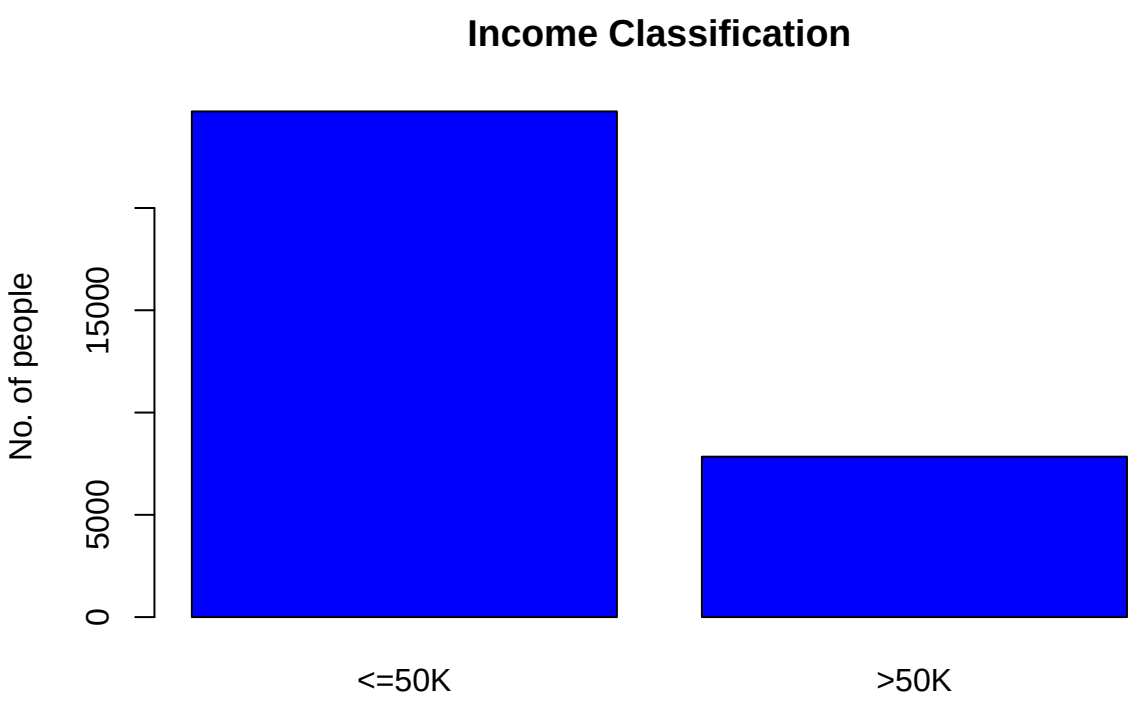
```
##       age         workclass             fnlwgt         education
##  Min.   :17.00   Length:32561       Min.   :  12285   Length:32561
##  1st Qu.:28.00   Class :character   1st Qu.: 117827   Class :character
##  Median :37.00   Mode  :character   Median : 178356   Mode  :character
##  Mean   :38.58                      Mean   : 189778
##  3rd Qu.:48.00                      3rd Qu.: 237051
##  Max.   :90.00                      Max.   :1484705
```

```
##  education.num   marital.status      occupation        relationship
##  Min.   : 1.00   Length:32561       Length:32561       Length:32561
##  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               sex             capital.gain    capital.loss
##  Length:32561       Length:32561       Min.   :    0   Min.   :   0.0
##  Class :character   Class :character   1st Qu.:    0   1st Qu.:   0.0
##  Mode  :character   Mode  :character   Median :    0   Median :   0.0
##                                        Mean   : 1078   Mean   :  87.3
##                                        3rd Qu.:    0   3rd Qu.:   0.0
##                                        Max.   :99999   Max.   :4356.0
##  hours.per.week  native.country        income
##  Min.   : 1.00   Length:32561       Length:32561
##  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
```

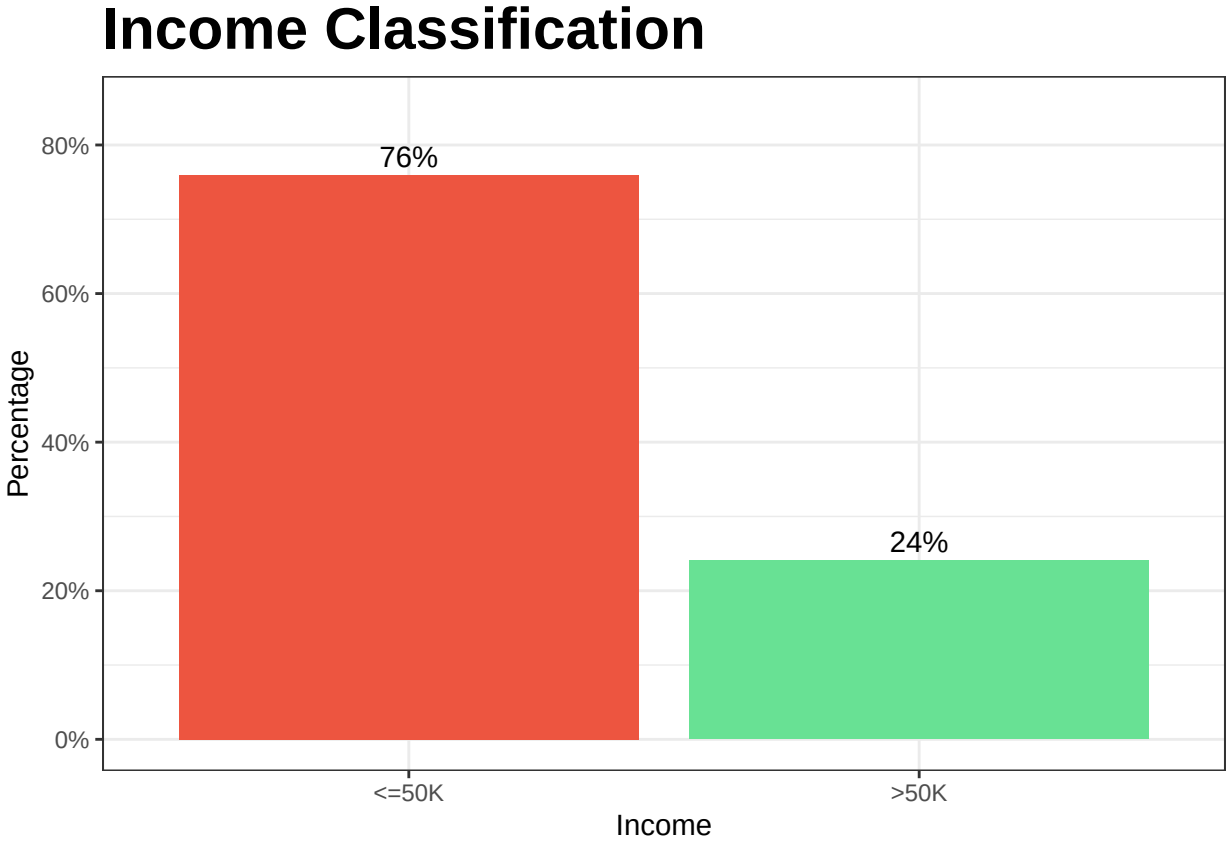
```
dim(rawData)
```

```
## [1] 32561    15
```

```
#Income
barplot(table(rawData$income),main = 'Income Classification',col='blue',ylab ='No. of people')
```

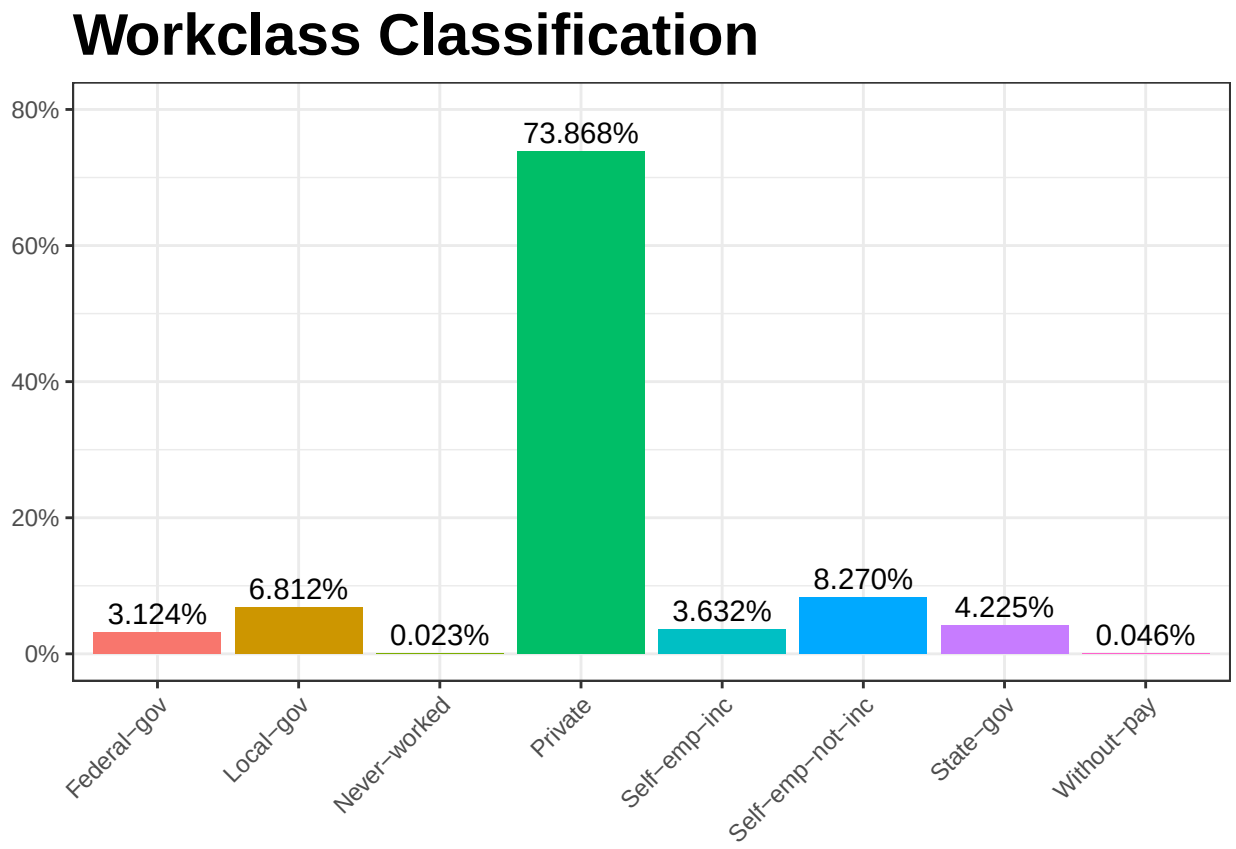



```r
#Income Classifciation
rawData %>%
ggplot(aes(income, group = 1)) +
  geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count") + labs(title="Income Classification")
  geom_text(aes( label = scales::percent(..prop..), y= ..prop.. ), size = 4, stat= "count", vjust = -0.4
theme_bw() +
theme(legend.position="none")+
scale_fill_manual("income", values = c("1" = "#ED5540", "2" = "#68E194"))+
scale_y_continuous(labels=scales::percent) +
ylab("Percentage") +
xlab("Income") +
coord_cartesian(ylim = c(0, 0.85)) +
theme(plot.title = element_text(color="black", face="bold", size=22, hjust=0))
```



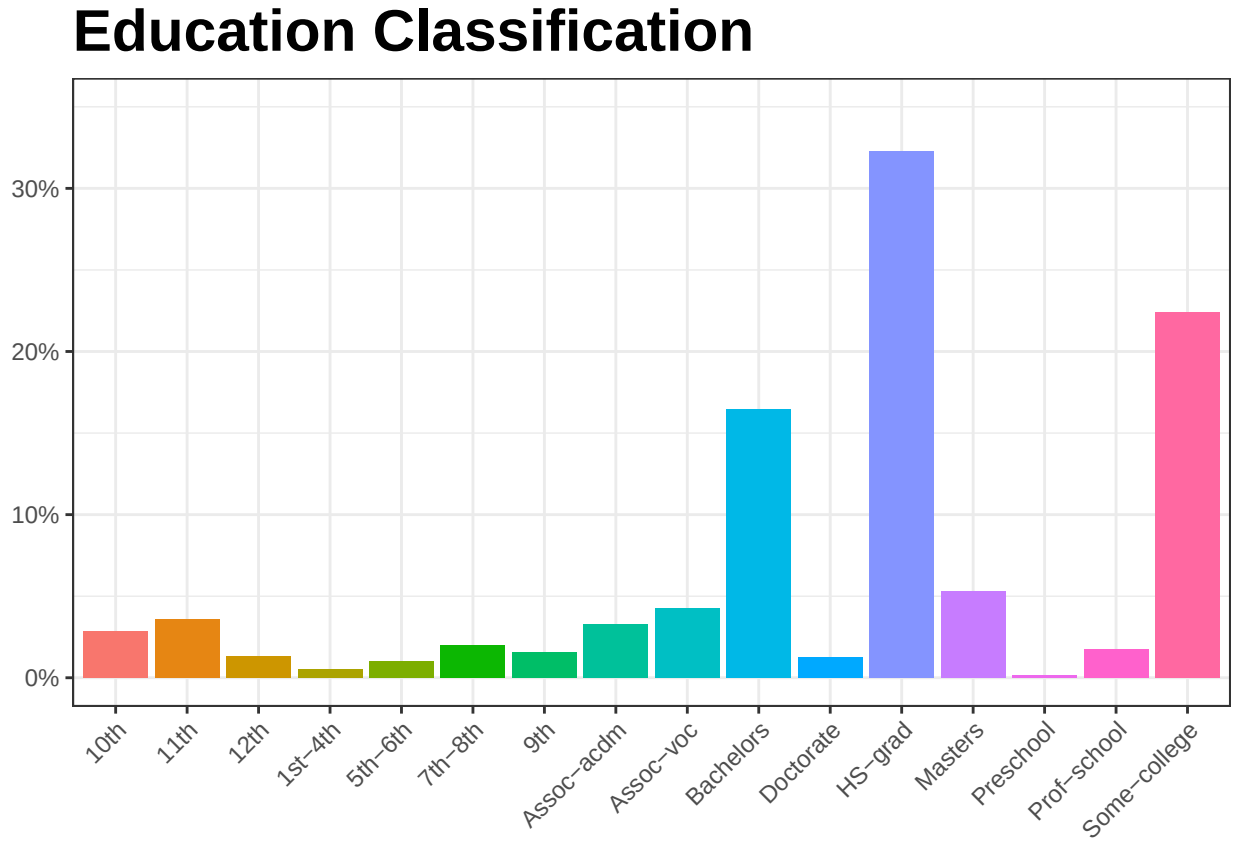

```
#family = "Circular Std",
```

```
#Workclass Classifciation
rawData %>% filter(workclass != "?") %>%
ggplot(aes(workclass, group = 1)) +geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count") + ge
theme(legend.position="none") +
scale_y_continuous(labels=scales::percent, limits = c(0, 1)) +
labs(title="Workclass Classification")+
coord_cartesian(ylim = c(0, 0.8)) +theme(plot.title = element_text(color="black", face="bold", size=22,
```

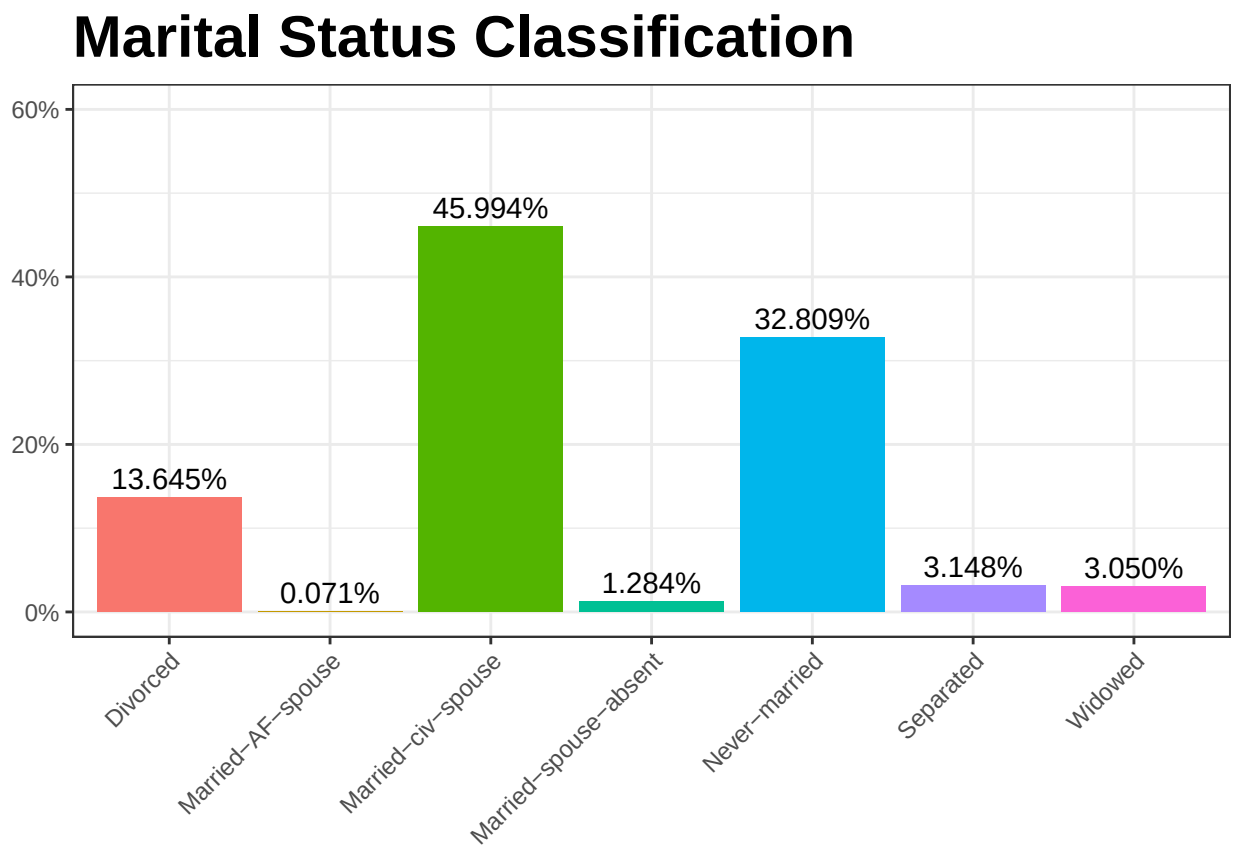


Looking at the table, it seems like male, people who are married, with more than 10 years of education, in exec-managerial, prof-specialty, or protective-service occupation, and work in the federal-government, local-government or self-employed are more likely to make more than 50K per year.

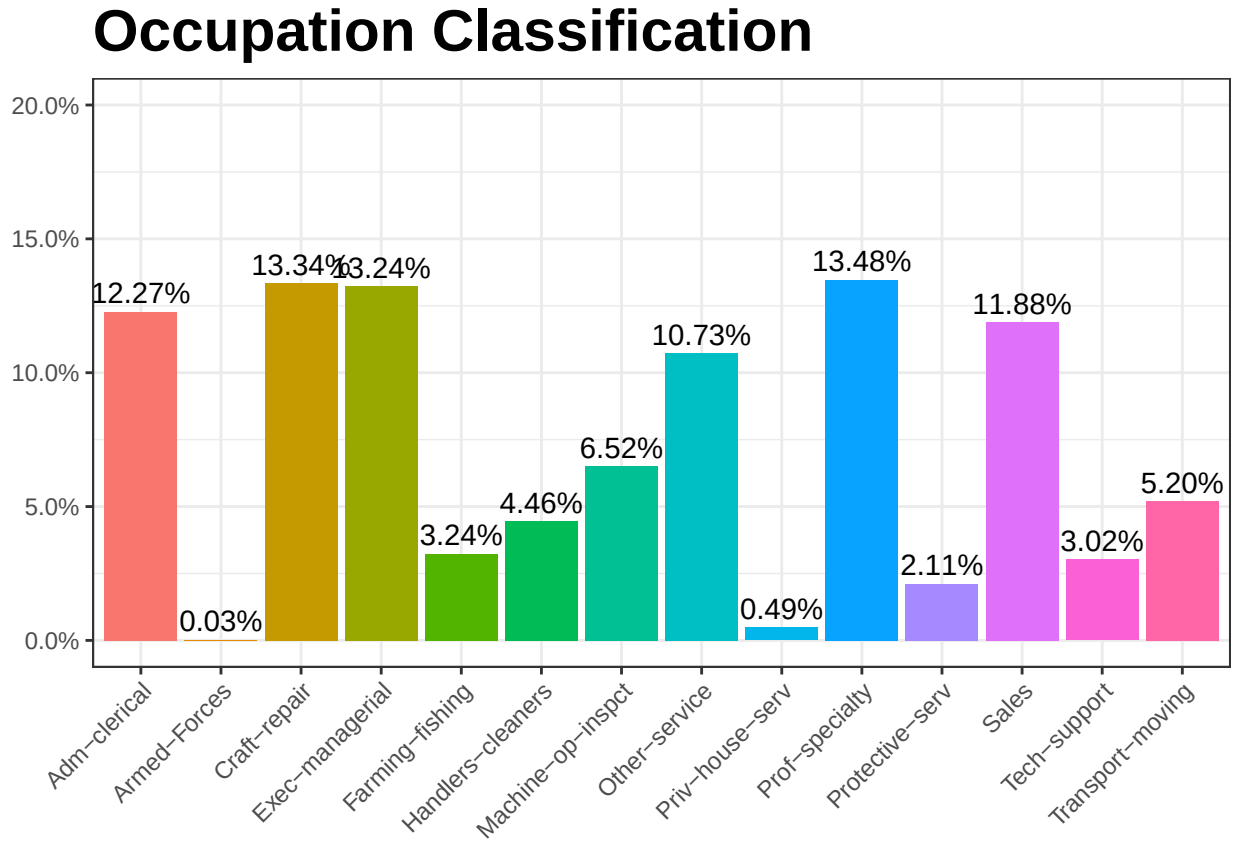
```
#Education Classifciation
rawData %>% filter(education != "?") %>%
ggplot(aes(education, group = 1)) +geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count") + th
theme(legend.position="none")+
scale_y_continuous(labels=scales::percent, limits = c(0, 1)) +
labs(title="Education Classification")+
coord_cartesian(ylim = c(0, 0.35)) +theme(plot.title = element_text(color="black", face="bold", size=22
```



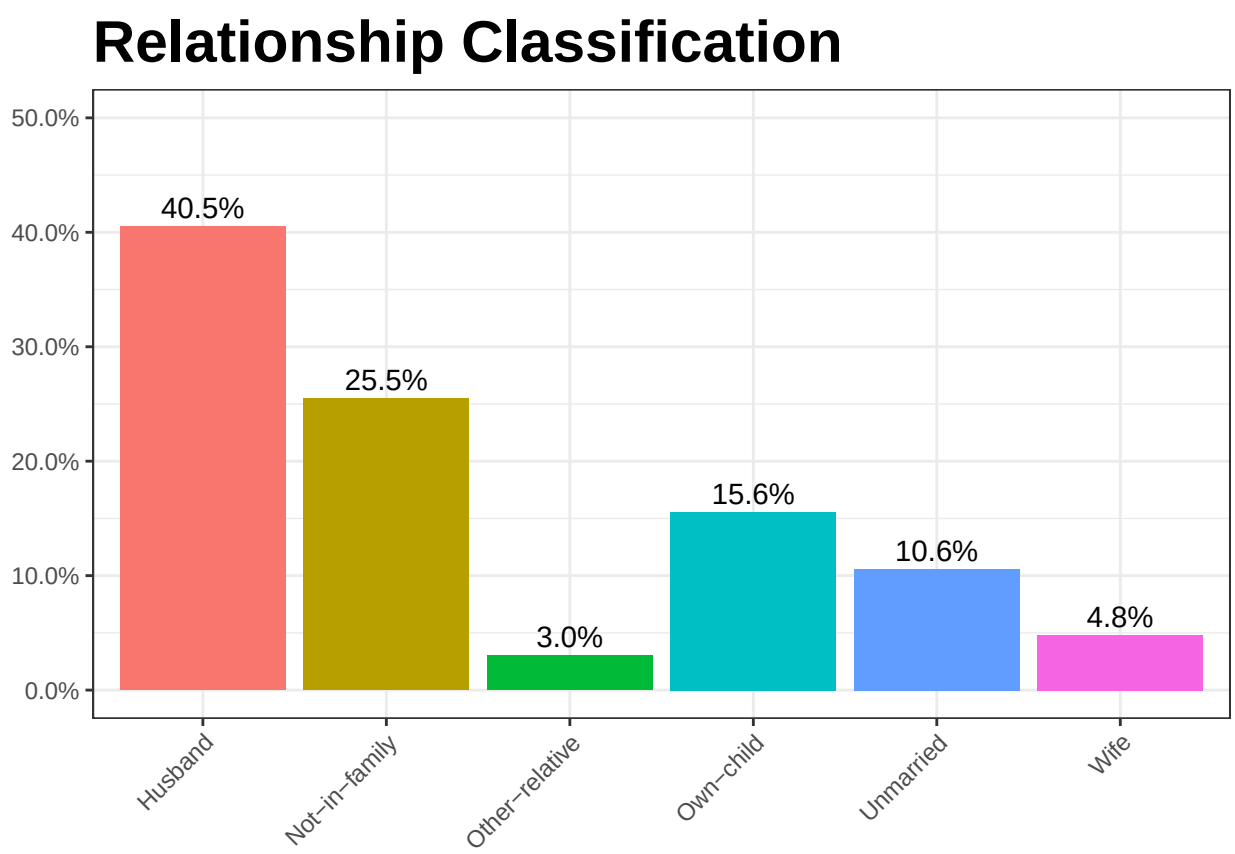

```
#marital.status
rawData %>% filter(marital.status != "?") %>%
ggplot(aes(marital.status, group = 1)) +geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count")
theme(legend.position="none")+
scale_y_continuous(labels=scales::percent, limits = c(0, 1)) +
labs(title="Marital Status Classification")+
coord_cartesian(ylim = c(0, 0.6)) +theme(plot.title = element_text(color="black", face="bold", size=22,
```



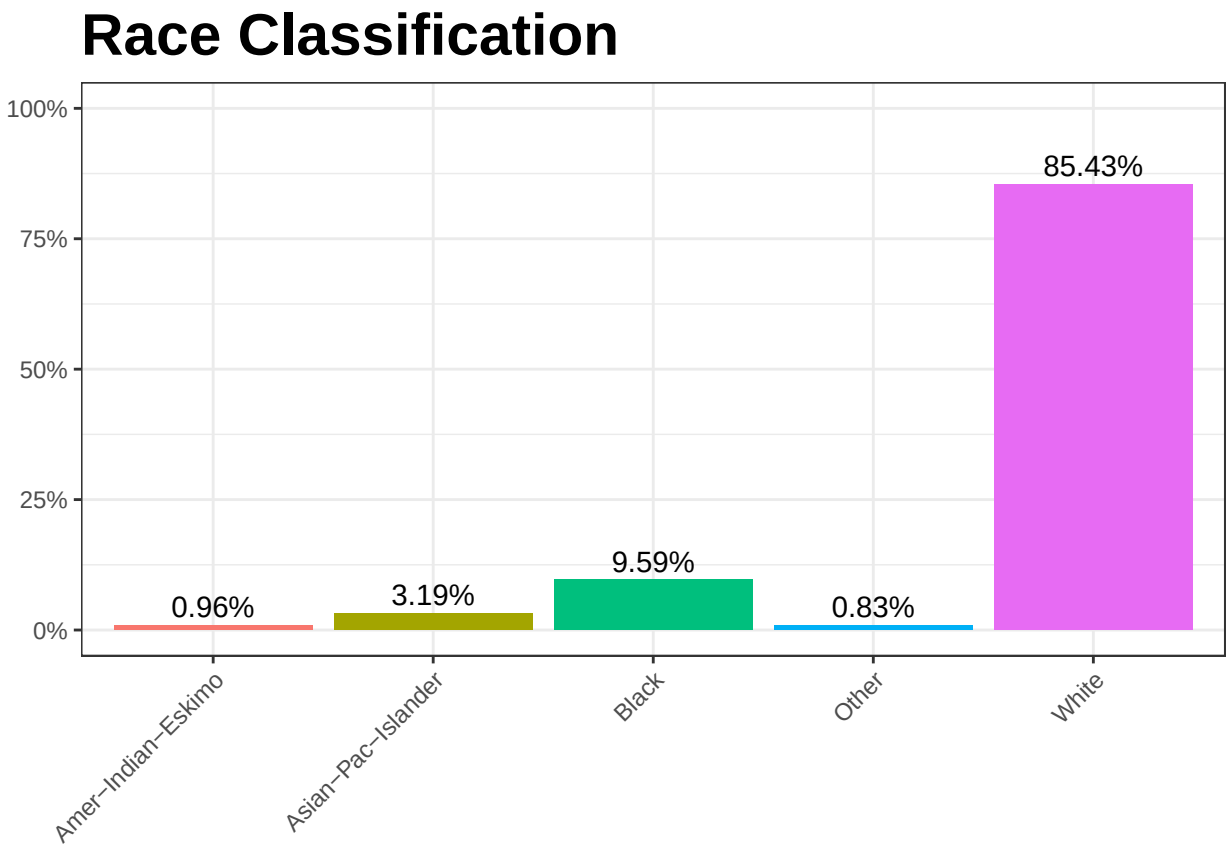

```
#occupation Classifciation
rawData %>% filter(occupation != "?") %>%
ggplot(aes(occupation, group = 1)) +geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count") + g
theme(legend.position="none")+
scale_y_continuous(labels=scales::percent, limits = c(0, 1)) +
labs(title="Occupation Classification")+
coord_cartesian(ylim = c(0, 0.2)) +theme(plot.title = element_text(color="black", face="bold", size=22,
```



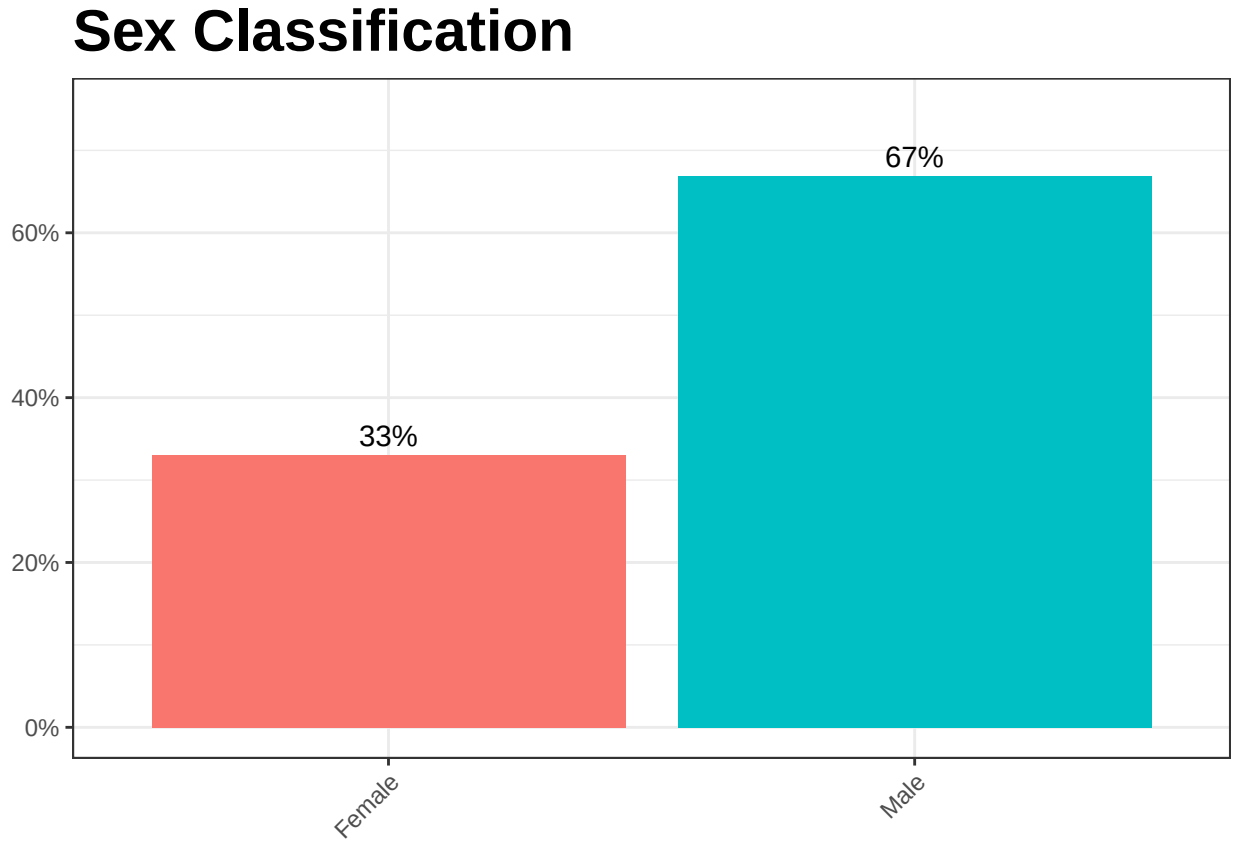

```
#relationship
rawData %>% filter(relationship != "?") %>%
ggplot(aes(relationship, group = 1)) +geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count") +
theme(legend.position="none")+
scale_y_continuous(labels=scales::percent, limits = c(0, 1)) +
labs(title="Relationship Classification")+
theme(legend.position = "none") + coord_cartesian(ylim = c(0, 0.5)) +theme(plot.title = element_text(co
```



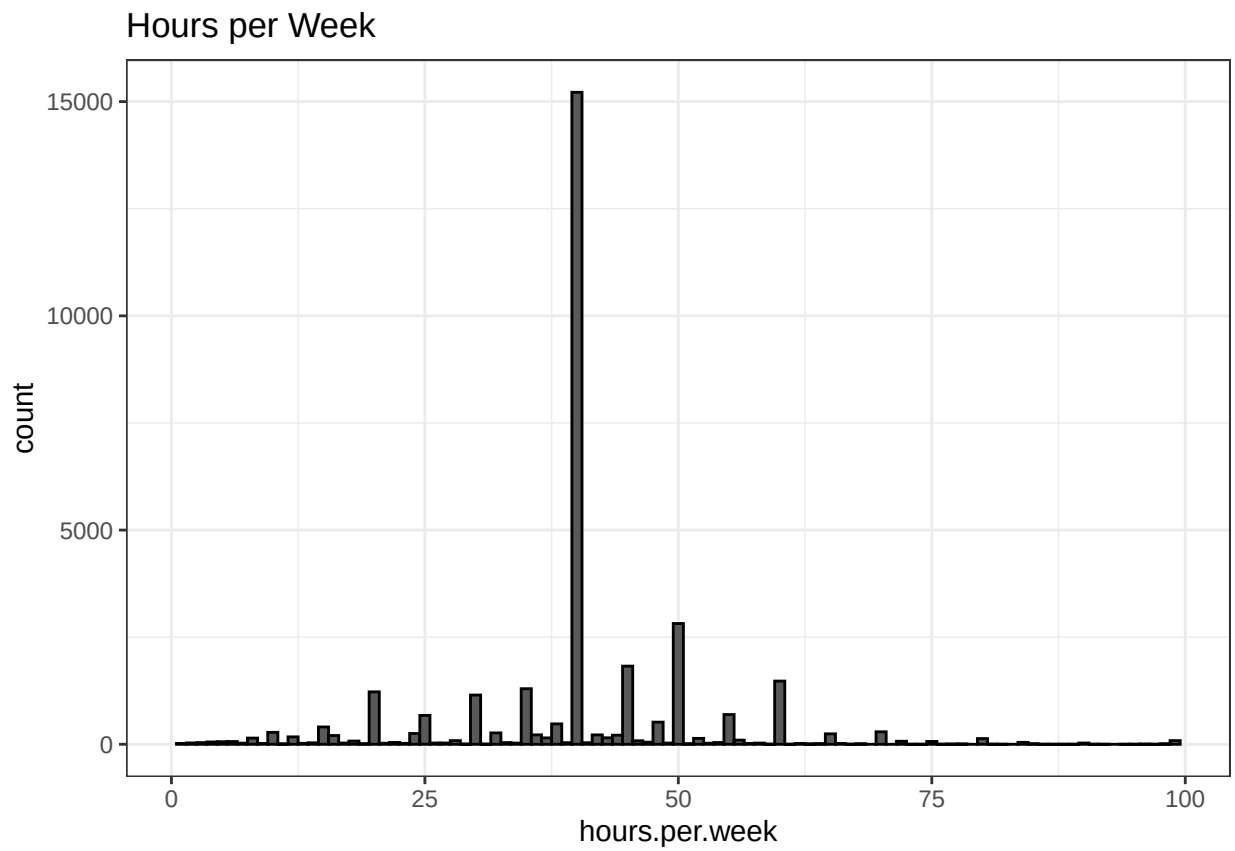

```
#race
rawData %>% filter(race != "?") %>%
ggplot(aes(race, group = 1)) +geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count") + geom_te
theme(legend.position="none")+
scale_y_continuous(labels=scales::percent, limits = c(0, 1)) +
labs(title="Race Classification")+
theme(legend.position = "none") + coord_cartesian(ylim = c(0, 1)) +theme(plot.title = element_text(colo
```



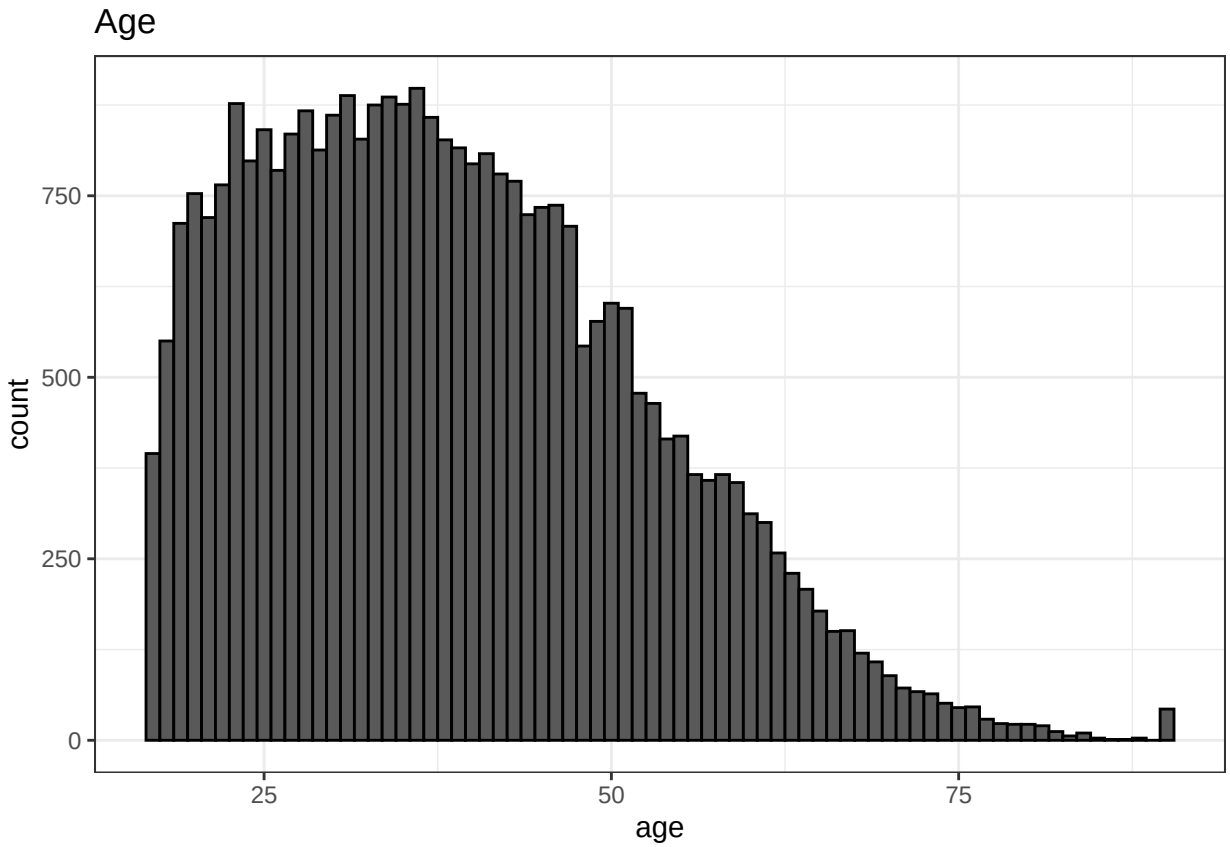

```
#sex
rawData %>% filter(sex != "?") %>%
ggplot(aes(sex, group = 1)) +geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count") + geom_tex
theme(legend.position="none")+
scale_y_continuous(labels=scales::percent, limits = c(0, 1)) +
labs(title="Sex Classification")+
theme(legend.position = "none") + coord_cartesian(ylim = c(0, 0.75)) +theme(plot.title = element_text(c
```



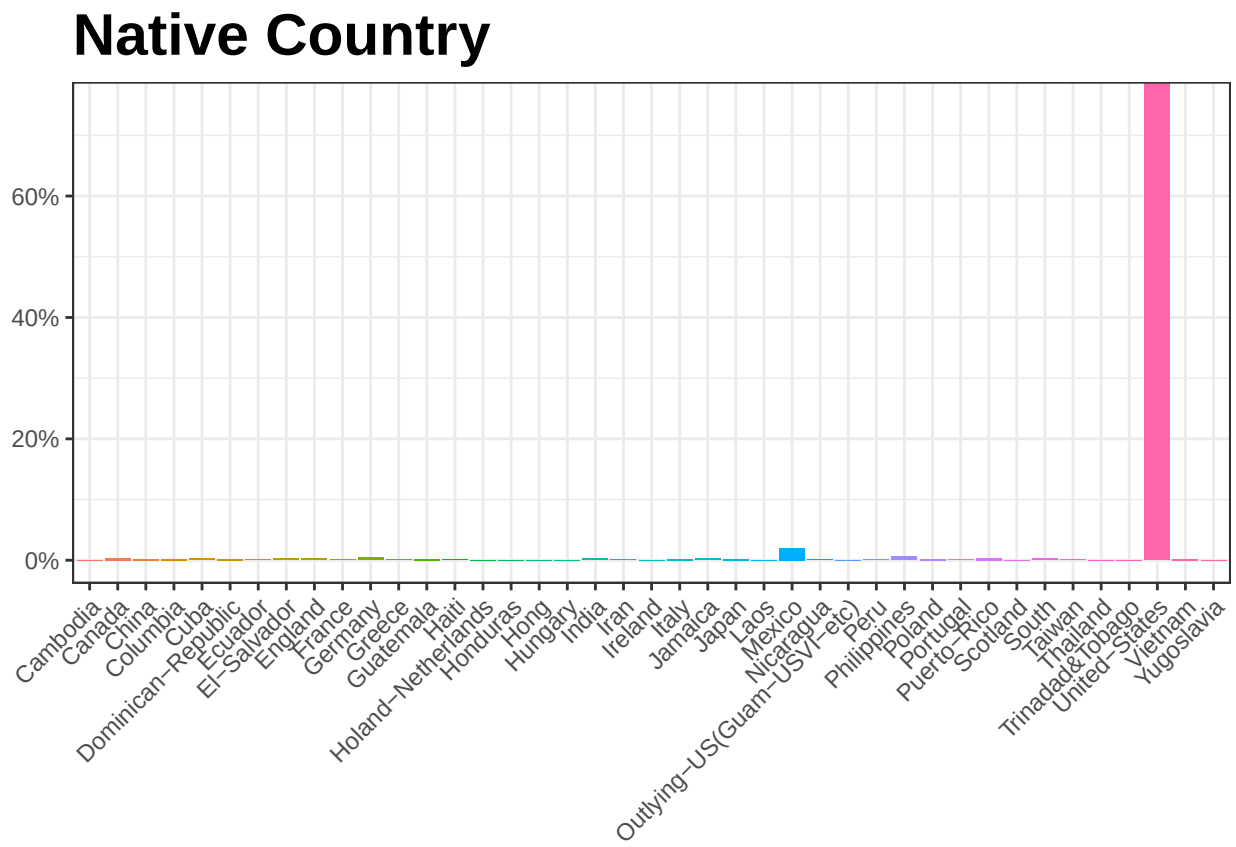

```
#hours.per.week
rawData %>% filter(hours.per.week != "?") %>%
ggplot(aes(hours.per.week, group = 1)) +geom_histogram(binwidth = 1, col="black") + theme_bw() + labs(t
```



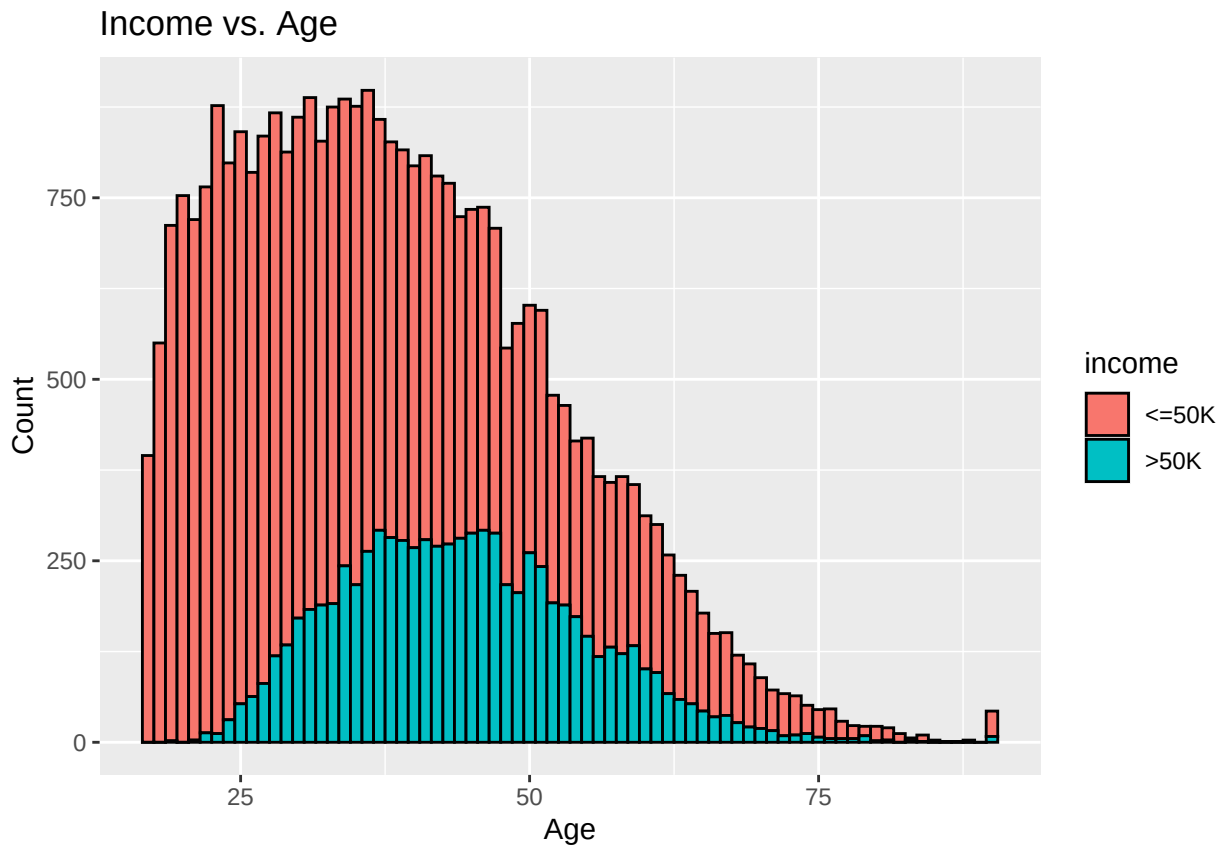

```
#Age
rawData %>% filter(age != "?") %>%
ggplot(aes(age, group = 1)) +geom_histogram(binwidth = 1, col="black") + theme_bw() + labs(title="Age")
```



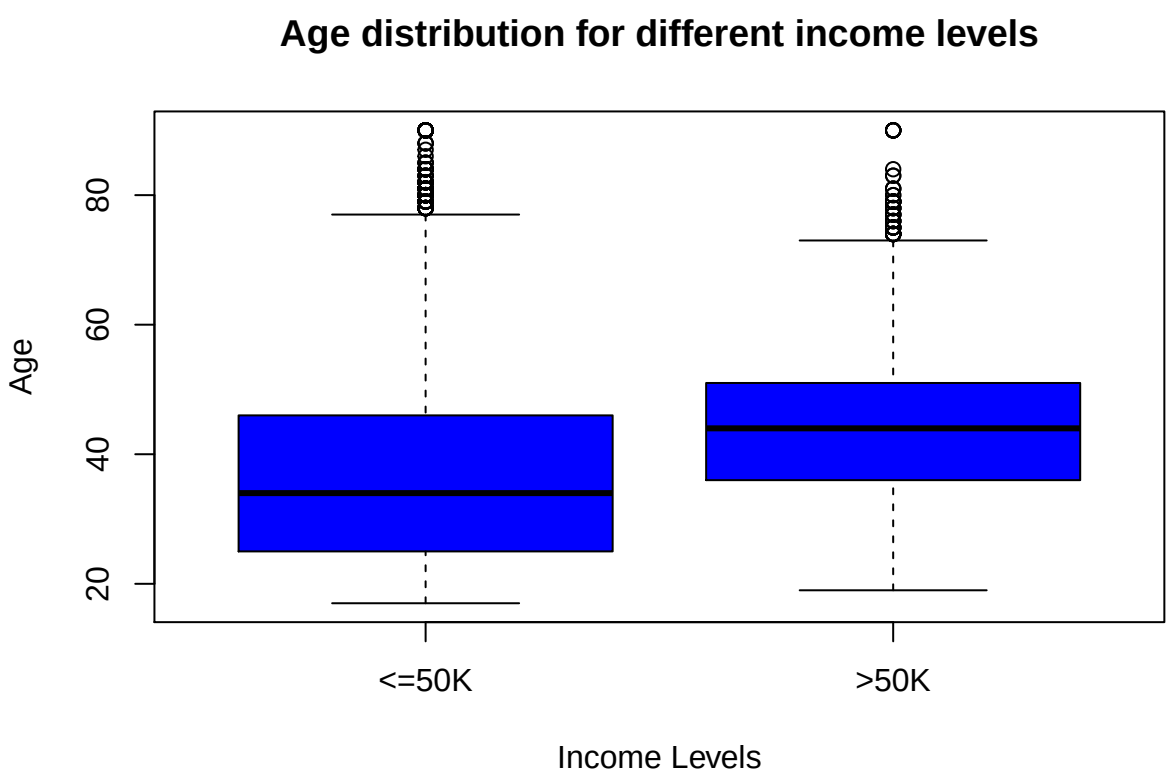

```
#native.country
rawData %>% filter(native.country != "?") %>%
ggplot(aes(native.country, group = 1)) +geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count")
theme(legend.position="none")+
scale_y_continuous(labels=scales::percent, limits = c(0, 1)) +
labs(title="Native Country")+
theme(legend.position = "none") + coord_cartesian(ylim = c(0, 0.75)) +theme(plot.title = element_text(c
```



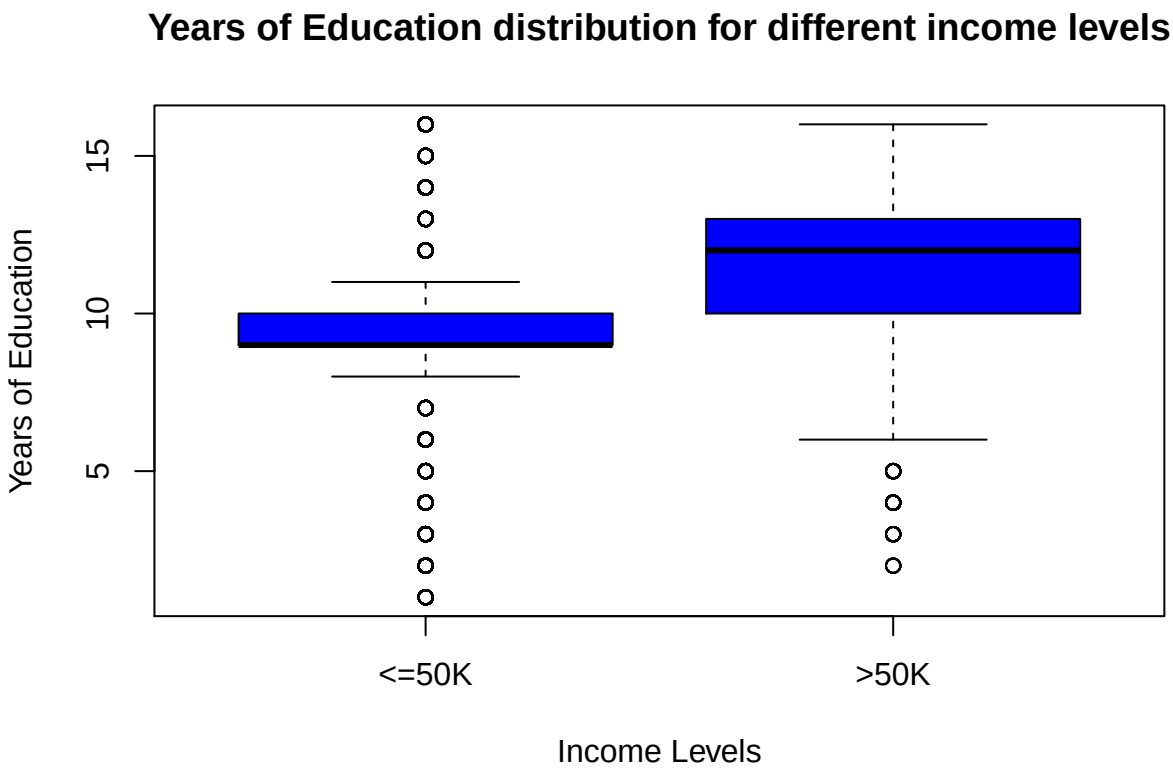

```
#Age vs. Income
ggplot(rawData) + aes(x=as.numeric(age), group=income, fill=income) +
  geom_histogram(binwidth=1, color='black')+
  labs(x="Age",y="Count",title = "Income vs. Age")
```



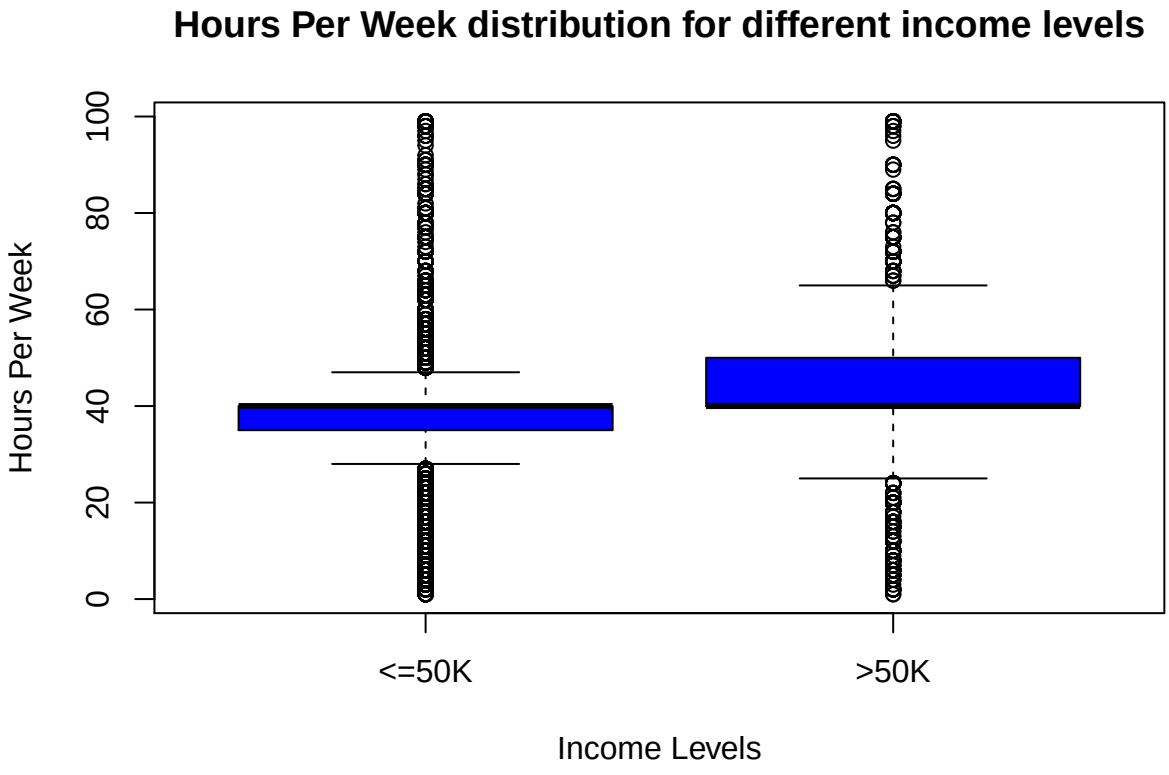

```
#Age vs. Income boxplot
boxplot (age ~ income, data = rawData,
         main = "Age distribution for different income levels",
         xlab = "Income Levels", ylab = "Age", col = "blue")
```




```r
#Years of Education
boxplot (education.num ~ income, data = rawData,
         main = "Years of Education distribution for different income levels",
         xlab = "Income Levels", ylab = "Years of Education", col = "blue")
```




```r
#Hours per week vs. Income level boxplot
boxplot (hours.per.week ~ income, data = rawData,
         main = "Hours Per Week distribution for different income levels",
         xlab = "Income Levels", ylab = "Hours Per Week", col = "blue")
```




```
#Put education number in ranges
rawData <- rawData %>% mutate(edu.range = case_when(education.num %in% c(0:5) ~ "0 - 5 years", educatio

#Create a table
table1(~ edu.range + native.country + sex + race + relationship + occupation + marital.status + educati
         , data = rawData)
```

```
## [1] "<table class=\"Rtable1\">\n<thead>\n<tr>\n<th class='rowlabel firstrow lastrow'></th>\n<th class
```

```
##   age workclass    education education.num marital.status        occupation
## 1  82   Private      HS-grad             9        Widowed   Exec-managerial
## 2  54   Private      7th-8th             4       Divorced Machine-op-inspct
## 3  41   Private Some-college            10      Separated    Prof-specialty
## 4  34   Private      HS-grad             9       Divorced     Other-service
## 5  38   Private         10th             6      Separated      Adm-clerical
## 6  74 State-gov    Doctorate            16  Never-married    Prof-specialty
##     relationship  race    sex hours.per.week native.country income
## 1  Not-in-family White Female             18  United-States  <=50K
## 2      Unmarried White Female             40  United-States  <=50K
## 3      Own-child White Female             40  United-States  <=50K
## 4      Unmarried White Female             45  United-States  <=50K
## 5      Unmarried White   Male             40  United-States  <=50K
## 6 Other-relative White Female             20  United-States   >50K
```

```
## [1] 30162    12
```

#Logistic Regression We are going to split data into test and training set: 70% vs. 30% Accuracy of this model using all predictors is 82.7%, which is fairly good.

There are a lot of confounding variables in this dataset. After removing confounding variables, we only have relationship and years of education left as variables. Accuracy of this model of using only relationship and years of education as predictors is 81.4%, which is very close to using most of the varilables in the dataset to predict income. We also tried predicting this model using only sex and years of education, but the accurecy of this model is only at 76.6%.

Below are the conclusions from the model using only relationship and years of education as predictors.

People with more than 10 years of education are 21 times more likely to make more than 50K than people who had 5 or less years of education. People who are in the husband relationship status are 11 times more likely to make more than 50K a year than people who are unmarried. People who are in the wife relationship status are 13 times more likely to make more than 50K a year than people who are unmarried. People who are in the Not-in-family relationship status are 30% more likely to make more than 50K a year than people who are unmarried. People who are in the own-child relationship status are 79% less likely to make more than 50K than people who are unmarried.

```
#0 = <=50K
#1 = >50K
#Put education number in ranges
adult <- adult %>% mutate(edu.range = case_when(education.num %in% c(0:5) ~ "0 - 5 years", education.nu

adult <- adult %>% mutate(income1 = case_when(income == ">50K" ~ 1,TRUE ~ 0))

#Split data into test and training set: 70% vs. 30%
index<-createDataPartition(adult$income,p=0.7,list = F)

train<-adult[index,]
test<-adult[-index,]

dim(train)
```

```
## [1] 21114    14
```

```
dim(test)
```

```
## [1] 9048   14
```

```
#Model
adult_blr <- glm(income1 ~ sex + education + relationship + workclass + race + occupation + native.coun
```

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

```
income_hat_a <- ifelse(predict(adult_blr, test) >= 0, 1, 0)
```

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

```
#Accuracy of model using all predictors
mean(income_hat_a == test$income1)
```

```
## [1] 0.8257073
```

```
##############################################


#Using only years of education and sex to predict income
adult_blr1 <- glm(income1 ~ edu.range + sex, data = train,family = "binomial")
summary(adult_blr1)
```

```
##
## Call:
## glm(formula = income1 ~ edu.range + sex, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2370  -0.6919  -0.6919  -0.1968   2.8119
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -3.93402    0.13750 -28.612   <2e-16 ***
## edu.range11+ years     2.71797    0.13400  20.283   <2e-16 ***
## edu.range6 - 10 years  1.27125    0.13363   9.513   <2e-16 ***
## sexMale                1.35503    0.04423  30.635   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 23697  on 21113  degrees of freedom
## Residual deviance: 20526  on 21110  degrees of freedom
## AIC: 20534
##
## Number of Fisher Scoring iterations: 5
```

```
#exp(coef(adult_blr1))
exp(cbind(OR = coef(adult_blr1), confint(adult_blr1)))
```

```
## Waiting for profiling to be done...
```

```
##                                OR      2.5 %      97.5 %
## (Intercept)            0.01956485  0.0148043  0.02540062
## edu.range11+ years    15.14949014 11.7532748 19.89212394
## edu.range6 - 10 years  3.56531953  2.7681852  4.67822628
## sexMale                3.87686473  3.5569155  4.23040165
```

```r
income_hat_a1 <- ifelse(predict(adult_blr1, test) >= 0, 1, 0)

#Accuracy of model using only years of education and sex
mean(income_hat_a1 == test$income1)
```

```
## [1] 0.7610522
```

```r
##########################################

#Change relationship factor order
train$relationship <-factor(train$relationship, levels=c("Unmarried", "Husband", "Wife", "Other-relative
levels(train$relationship)
```

```
## [1] "Unmarried"      "Husband"        "Wife"           "Other-relative"
## [5] "Own-child"      "Not-in-family"
```

```r
#Using only years of education and sex to predict income
adult_blr2 <- glm(income1 ~ edu.range + relationship, data = train,family = "binomial")
summary(adult_blr2)
```

```
##
## Call:
## glm(formula = income1 ~ edu.range + relationship, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5954  -0.6716  -0.2990  -0.0708   3.1280
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -4.58279    0.16016 -28.615  < 2e-16 ***
## edu.range11+ years          2.90267    0.13748  21.113  < 2e-16 ***
## edu.range6 - 10 years       1.41861    0.13640  10.401  < 2e-16 ***
## relationshipHusband         2.46723    0.09077  27.180  < 2e-16 ***
## relationshipWife            2.62423    0.11139  23.559  < 2e-16 ***
## relationshipOther-relative -0.30199    0.22526  -1.341  0.18004
## relationshipOwn-child      -1.40494    0.16673  -8.427  < 2e-16 ***
## relationshipNot-in-family   0.30569    0.09860   3.100  0.00193 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 23697  on 21113  degrees of freedom
## Residual deviance: 17091  on 21106  degrees of freedom
## AIC: 17107
##
## Number of Fisher Scoring iterations: 6
```

```r
#exp(coef(adult_blr1))
exp(cbind(OR = coef(adult_blr2), confint(adult_blr2)))
```

```
## Waiting for profiling to be done...
```

```
##                                    OR        2.5 %      97.5 %
## (Intercept)                 0.01022631  0.007411601  0.01389423
```

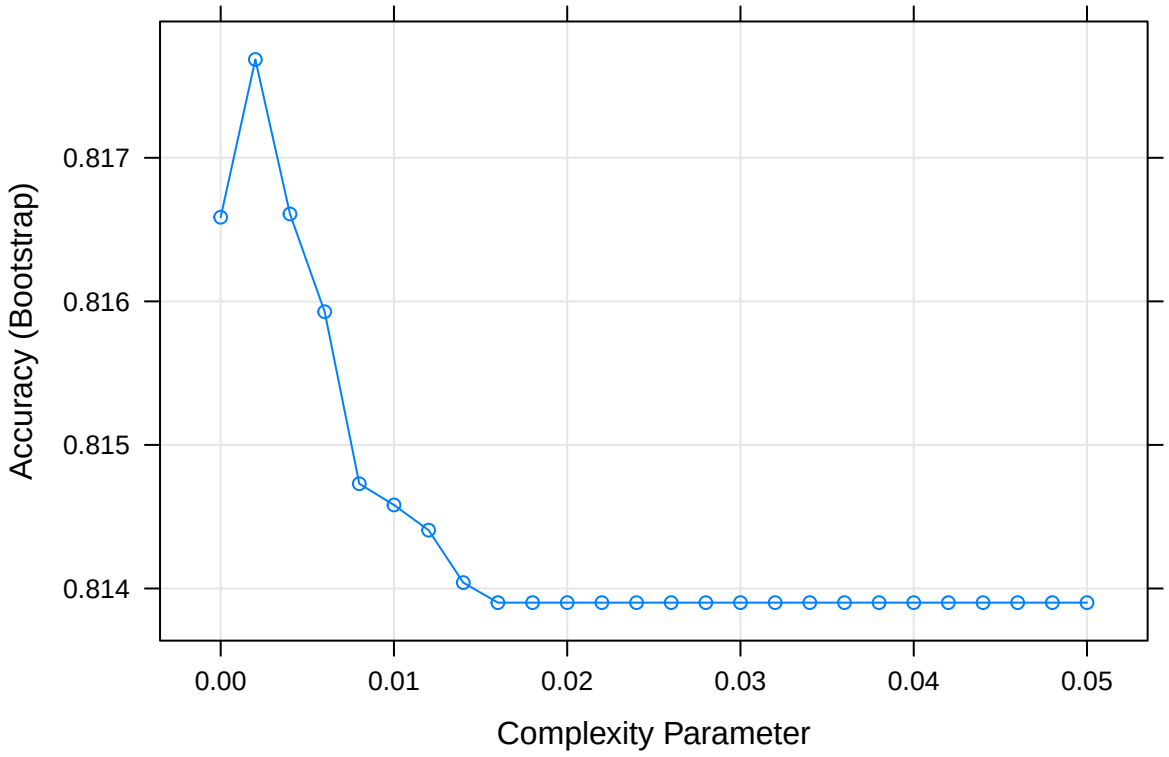
```
## edu.range11+ years          18.22266147 14.034691613 24.07912674
## edu.range6 - 10 years        4.13136488  3.189004447  5.44813095
## relationshipHusband         11.78970687  9.903233627 14.13890254
## relationshipWife            13.79394537 11.114591215 17.20314051
## relationshipOther-relative   0.73934206  0.465347905  1.12933384
## relationshipOwn-child        0.24538259  0.175465067  0.33784685
## relationshipNot-in-family    1.35755848  1.122066715  1.65192867
```

```
income_hat_a2 <- ifelse(predict(adult_blr2, test) >= 0, 1, 0)

#Accuracy of model using only years of education and relationship
mean(income_hat_a2== test$income1)
```

```
## [1] 0.811008
```

#Decision Tree Accuracy is 81.6%, which is very close to the results from our logistic regression model when using all varilables to predict income. Thre decision tree shows that relationship and education level are the most important varilables when it comes to predicting income.

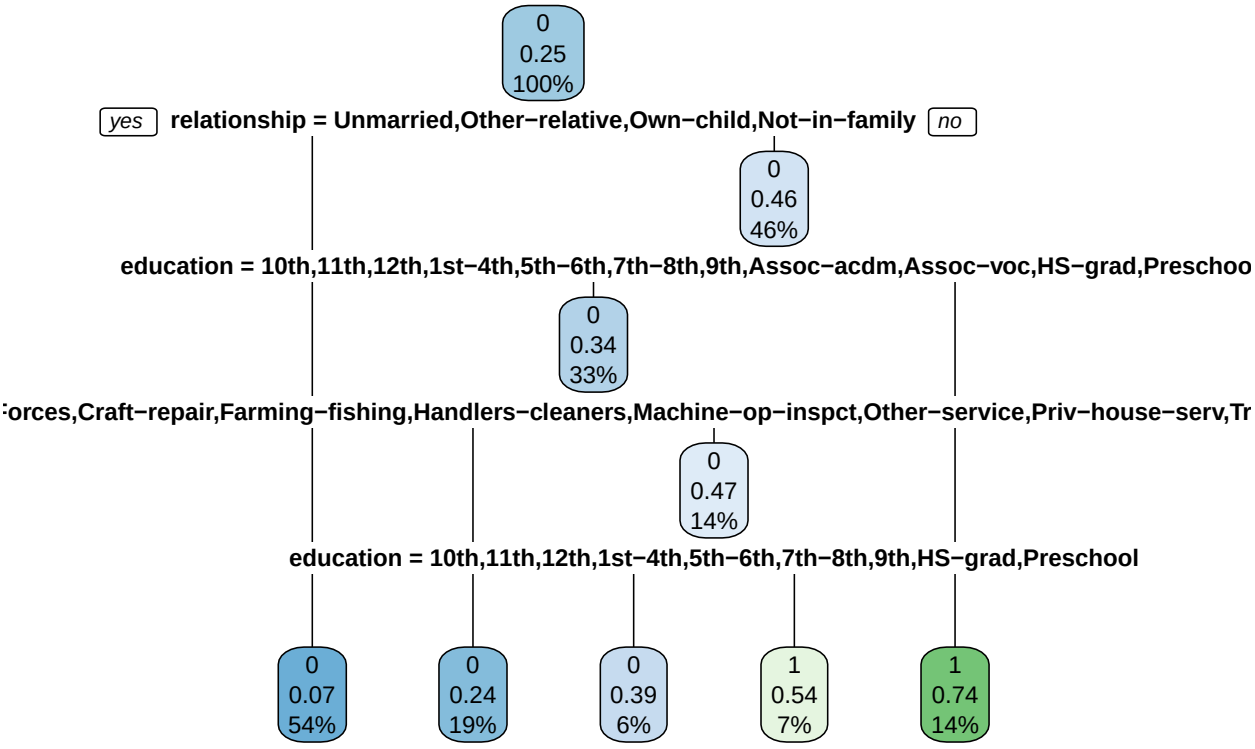
```
#0 = <=50K
#1 = >50K


fit_rpart11 <- train(income ~ sex + education + relationship + workclass + edu.range + race + occupatio
                     method = "rpart",
                     tuneGrid = data.frame(cp = seq(0, 0.05, 0.002)))
plot(fit_rpart11)
```



```
income_hat <- predict(fit_rpart11, test)
cm_test <- confusionMatrix(data = factor(income_hat), reference = factor(test$income))

cm_test$overall["Accuracy"]
```

```
##  Accuracy
## 0.8148762
```

```
##################################

tree_adult_model<-rpart(income1 ~ sex + education + relationship + workclass + edu.range + race + occupa

rpart.plot(tree_adult_model, extra = 106)
```



#Conclusion After performing logistic regression and decision tree classification techniques and taking into account their accuracies, we can conclude both models had an accuracy around 82% when using almost all varilables in the dataset to predict income. Logistic regression had a slighly higher accuracy at 82.7%.