Data Exploration

Zicheng (Stone) Shi

2/29/2020

Pre-processing

Read data and load packages

```
## read in data and load library
library(tidyverse)

raw <- read.csv("/Volumes/GoogleDrive/My Drive/University of Notre Dame/MSBA Spring Semester/career/inte
glimpse(raw)
```

```
## Observations: 32,561
## Variables: 15
## $ age            <int> 39, 50, 38, 53, 28, 37, 49, 52, 31, 42, 37, 30, 23, ...
## $ workclass      <fct>  State-gov,  Self-emp-not-inc,  Private,  Private,  ...
## $ fnlwgt         <int> 77516, 83311, 215646, 234721, 338409, 284582, 160187...
## $ education      <fct>  Bachelors,  Bachelors,  HS-grad,  11th,  Bachelors,...
## $ education_num  <int> 13, 13, 9, 7, 13, 14, 5, 9, 14, 13, 10, 13, 13, 12, ...
## $ marital_status <fct>  Never-married,  Married-civ-spouse,  Divorced,  Mar...
## $ occupation     <fct>  Adm-clerical,  Exec-managerial,  Handlers-cleaners,...
## $ relationship   <fct>  Not-in-family,  Husband,  Not-in-family,  Husband,  ...
## $ race           <fct>  White,  White,  White,  Black,  Black,  White,  Bla...
## $ sex            <fct>  Male,  Male,  Male,  Male,  Female,  Female,  Femal...
## $ capital_gain   <int> 2174, 0, 0, 0, 0, 0, 0, 0, 14084, 5178, 0, 0, 0, 0, ...
## $ capital_loss   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ hours_per_week <int> 40, 13, 40, 40, 40, 40, 16, 45, 50, 40, 80, 40, 30, ...
## $ native_country <fct>  United-States,  United-States,  United-States,  Uni...
## $ income         <fct>  <=50K,  <=50K,  <=50K,  <=50K,  <=50K,  <=50K,  <=5...
```

Data cleansing

First, I will do a quick summary of the data set to check outliers and NAs.

```
summary(raw)
```

```
##       age                    workclass          fnlwgt
##  Min.   :17.00    Private         :22696   Min.   :  12285
##  1st Qu.:28.00    Self-emp-not-inc: 2541   1st Qu.: 117827
```

```
##  Median :37.00    Local-gov       : 2093   Median : 178356
##  Mean   :38.58    ?               : 1836   Mean   : 189778
##  3rd Qu.:48.00    State-gov       : 1298   3rd Qu.: 237051
##  Max.   :90.00    Self-emp-inc    : 1116   Max.   :1484705
##                  (Other)          :  981
##          education     education_num                 marital_status
##   HS-grad     :10501   Min.   : 1.00    Divorced             : 4443
##   Some-college: 7291   1st Qu.: 9.00    Married-AF-spouse    :   23
##   Bachelors   : 5355   Median :10.00    Married-civ-spouse   :14976
##   Masters     : 1723   Mean   :10.08    Married-spouse-absent:  418
##   Assoc-voc   : 1382   3rd Qu.:12.00    Never-married        :10683
##   11th        : 1175   Max.   :16.00    Separated            : 1025
##  (Other)      : 5134                    Widowed              :  993
##             occupation           relationship                  race
##   Prof-specialty :4140    Husband       :13193    Amer-Indian-Eskimo:  311
##   Craft-repair   :4099    Not-in-family : 8305    Asian-Pac-Islander: 1039
##   Exec-managerial:4066    Other-relative:  981    Black             : 3124
##   Adm-clerical   :3770    Own-child     : 5068    Other             :  271
##   Sales          :3650    Unmarried     : 3446    White             :27816
##   Other-service  :3295    Wife          : 1568
##  (Other)         :9541
##       sex          capital_gain    capital_loss    hours_per_week
##   Female:10771   Min.   :    0   Min.   :   0.0   Min.   : 1.00
##   Male  :21790   1st Qu.:    0   1st Qu.:   0.0   1st Qu.:40.00
##                  Median :    0   Median :   0.0   Median :40.00
##                  Mean   : 1078   Mean   :  87.3   Mean   :40.44
##                  3rd Qu.:    0   3rd Qu.:   0.0   3rd Qu.:45.00
##                  Max.   :99999   Max.   :4356.0   Max.   :99.00
##
##         native_country     income
##   United-States:29170    <=50K:24720
##   Mexico       :  643    >50K : 7841
##   ?            :  583
##   Philippines  :  198
##   Germany      :  137
##   Canada       :  121
##  (Other)       : 1709
```

```r
## replace the question marks with NA
raw$workclass <- gsub("?", NA, raw$workclass, fixed = TRUE)
raw$native_country <- gsub("?", NA, raw$native_country, fixed = TRUE)
```

I'll remove those NAs because they only account for 7% of our data. It is safe to drop them.

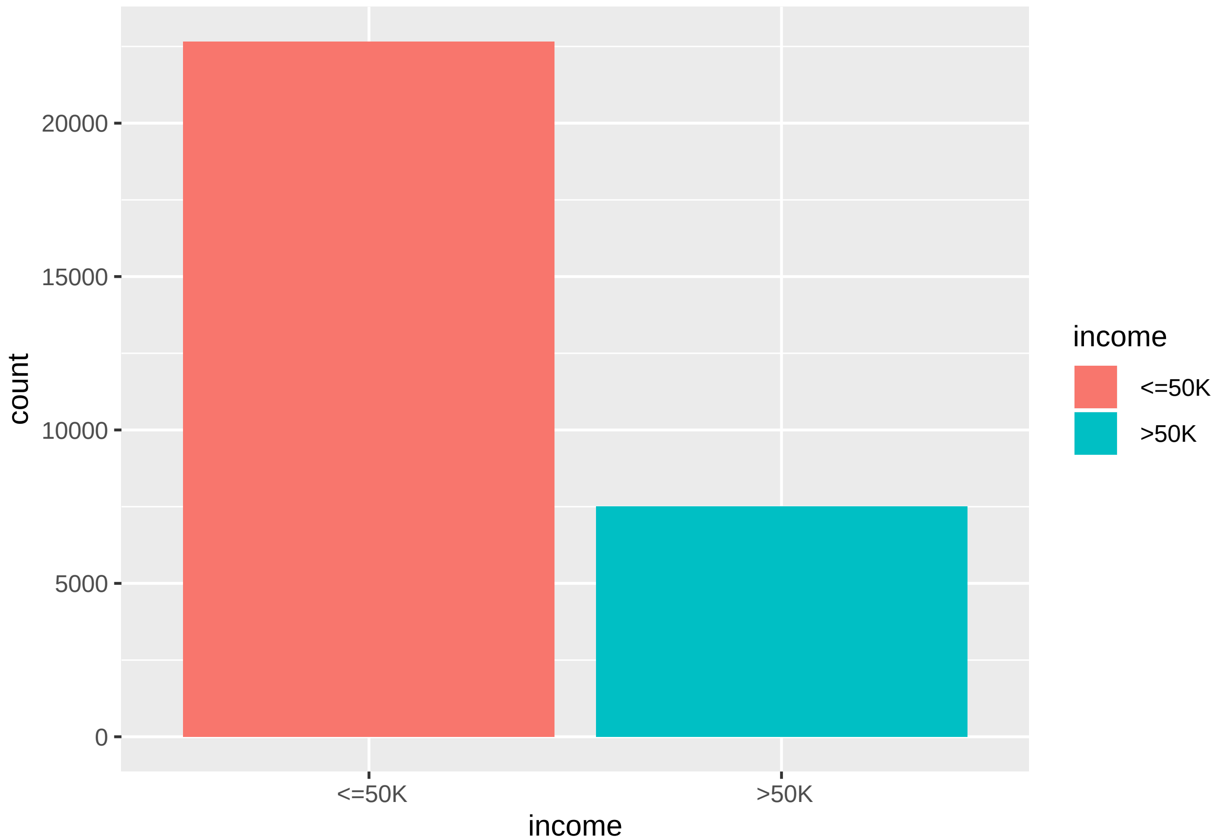
```r
raw_2 <- raw %>%
  filter(!is.na(workclass)) %>%
  filter(!is.na(native_country)) %>%
  mutate_at(c("workclass", "native_country"), as.factor)

summary(raw_2)
```

```
##       age                      workclass          fnlwgt
##  Min.   :17.00    Private         :22286   Min.   :  13769
```

```
##  1st Qu.:28.00    Self-emp-not-inc: 2499   1st Qu.: 117634  
##  Median :37.00    Local-gov       : 2067   Median : 178429  
##  Mean   :38.43    State-gov       : 1279   Mean   : 189802  
##  3rd Qu.:47.00    Self-emp-inc    : 1074   3rd Qu.: 237624  
##  Max.   :90.00    Federal-gov     :  943   Max.   :1484705  
##                  (Other)          :   21                    
##          education     education_num                  marital_status 
##   HS-grad     :9841   Min.   : 1.00    Divorced             : 4215  
##   Some-college:6680   1st Qu.: 9.00    Married-AF-spouse    :   21  
##   Bachelors   :5044   Median :10.00    Married-civ-spouse   :14066  
##   Masters     :1627   Mean   :10.12    Married-spouse-absent:  370  
##   Assoc-voc   :1307   3rd Qu.:13.00    Never-married        : 9731  
##   11th        :1049   Max.   :16.00    Separated            :  939  
##  (Other)      :4621                    Widowed              :  827  
##             occupation           relationship                   race      
##   Prof-specialty :4038    Husband        :12463    Amer-Indian-Eskimo:  286  
##   Craft-repair   :4030    Not-in-family : 7727    Asian-Pac-Islander:  895  
##   Exec-managerial:3992    Other-relative:  889    Black             : 2819  
##   Adm-clerical   :3721    Own-child     : 4471    Other             :  231  
##   Sales          :3584    Unmarried     : 3212    White             :25938  
##   Other-service  :3212    Wife          : 1407                            
##  (Other)         :7592                                                    
##       sex          capital_gain    capital_loss     hours_per_week 
##   Female: 9784   Min.   :    0   Min.   :   0.00   Min.   : 1.00  
##   Male  :20385   1st Qu.:    0   1st Qu.:   0.00   1st Qu.:40.00  
##                  Median :    0   Median :   0.00   Median :40.00  
##                  Mean   : 1092   Mean   :  88.35   Mean   :40.93  
##                  3rd Qu.:    0   3rd Qu.:   0.00   3rd Qu.:45.00  
##                  Max.   :99999   Max.   :4356.00   Max.   :99.00  
##                                                                    
##         native_country     income     
##   United-States:27511    <=50K:22661  
##   Mexico       :  610    >50K : 7508  
##   Philippines  :  188                 
##   Germany      :  128                 
##   Puerto-Rico  :  109                 
##   Canada       :  107                 
##  (Other)       : 1516
```

Class Variable Distribution

```r
raw_2 %>%
  ggplot(aes(x = income, fill = income)) +
  geom_bar()
```



The class distribution is quite imbalanced, I'll handle this in the later part of analysis.

Questions 1

Which race, sex combination is most represented in this data set? Which race, sex combination is least likely to make more than $50K?

```
## combine the race and sex columns
raw_2$Race_Sex <- as.factor(paste(raw_2$race, raw_2$sex, sep = ""))
table(raw_2$Race_Sex)
```

```
##
##  Amer-Indian-Eskimo Female    Amer-Indian-Eskimo Male
##                        107                        179
##  Asian-Pac-Islander Female    Asian-Pac-Islander Male
##                        294                        601
##               Black Female                 Black Male
##                       1400                       1419
##               Other Female                 Other Male
##                         87                        144
##               White Female                 White Male
##                       7896                      18042
```

Based on the table above, I can see the **White Male** combination is most represented in the data set.

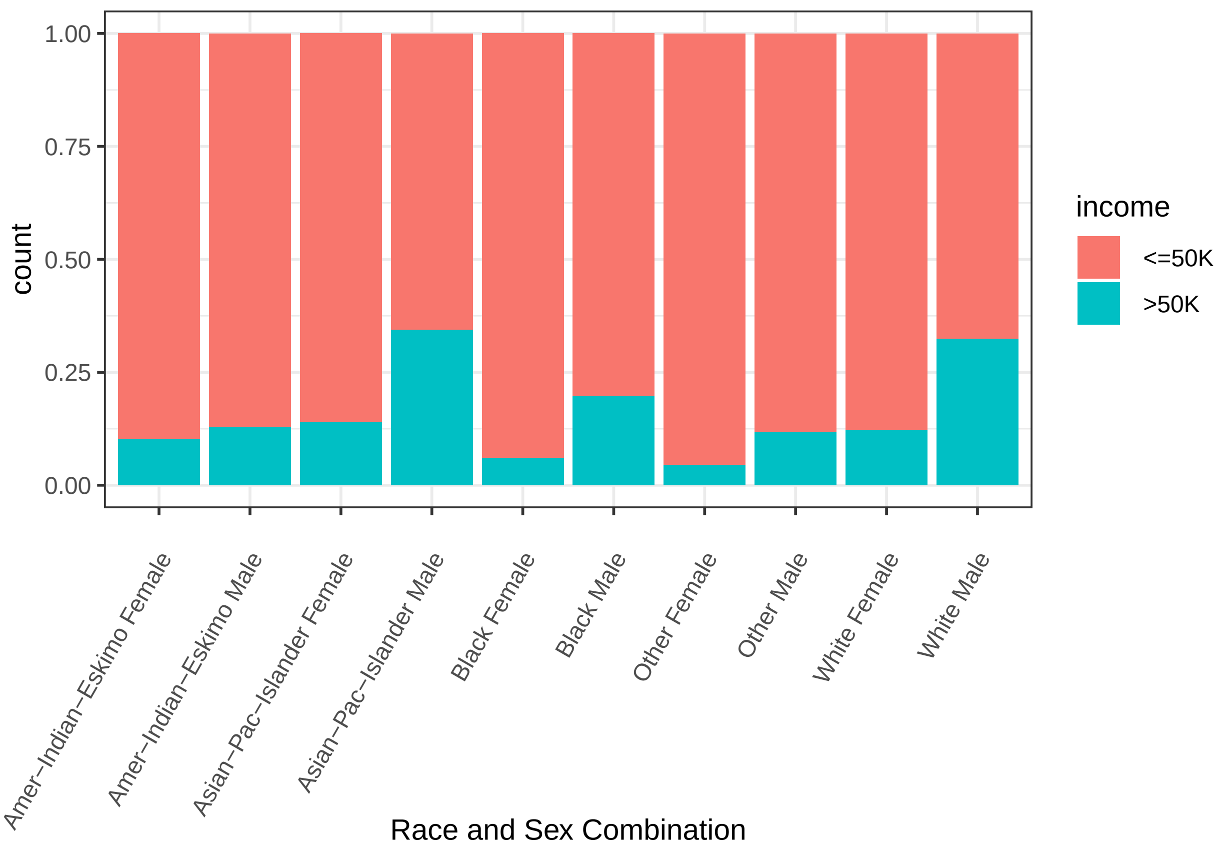
```r
table(raw_2$income)
```

```
##
##  <=50K   >50K
##  22661   7508
```

```r
table(raw_2$education)
```

```
##
##          10th          11th          12th       1st-4th       5th-6th
##           822          1049           377           151           288
##       7th-8th           9th    Assoc-acdm     Assoc-voc     Bachelors
##           558           455          1008          1307          5044
##     Doctorate       HS-grad       Masters     Preschool   Prof-school
##           375          9841          1627            45           542
##  Some-college
##          6680
```

```r
ggplot(raw_2, aes(x = Race_Sex, fill = income)) +
  geom_bar(position = 'fill') +
  theme_bw() +
  theme(axis.text.x  = element_text(angle=60, hjust=1, vjust=0.9)) +
  labs(x = "Race and Sex Combination")
```



From the bar chart above we can see, **Other Female** have the lowest percent of making income less than *$50k*, so they are least likely to make more than $50k among those race and sex combinations.

Question 2

Are there any columns that can be dropped from this data set without damaging the information contained within the data?

I'll remove the **education_num** column. The reason is that **education_num** contains the same information as **education**, we can see the more advanced the degree is, the larger the number of education years will be.

So it is safe for me to drop `education_num` column without damaging the information.

```
raw_2 <- raw_2 %>%
  dplyr::select(-education_num)
```

Question 3:

What steps did you take to prepare the data for your analysis and why did you need to do those steps? What tools did you use to do this data preparation and the associated analyses?

As I've done in my previous steps, before doing data analysis, we need to:

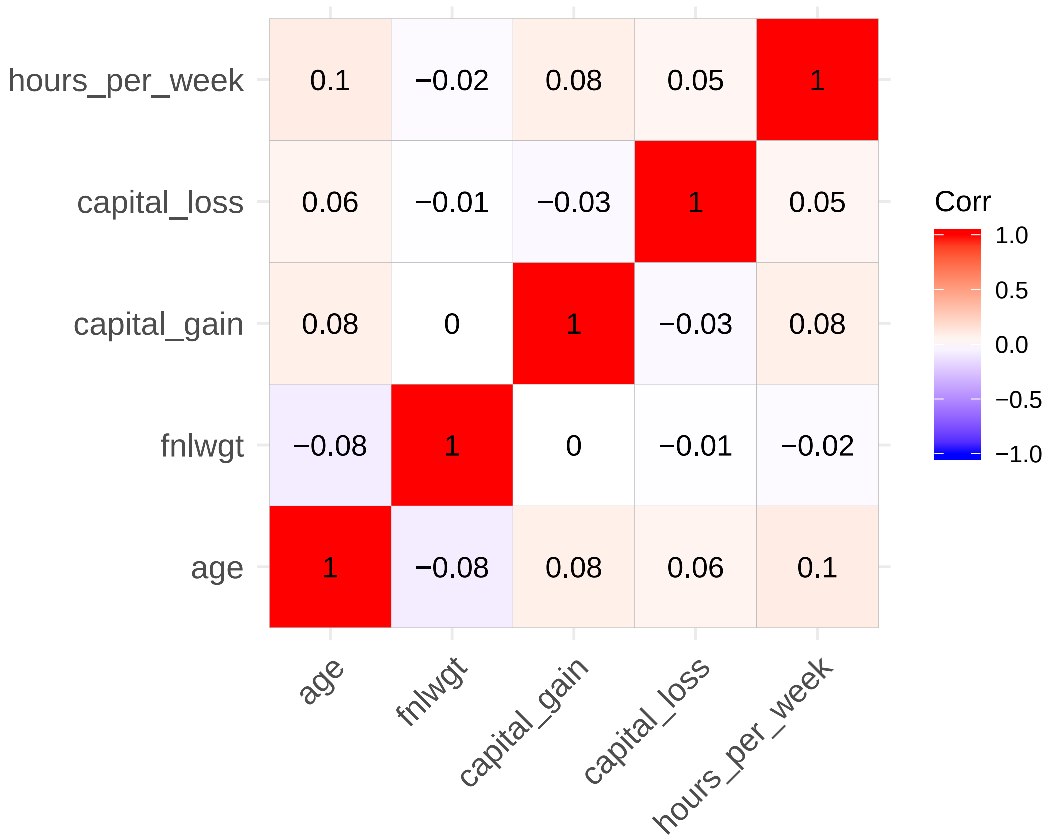
- Convert data to the proper types (e.g. from character to factor)
- Use various imputation techniques to handle missing data, such as mean imputation, predictive imputation, etc. In my previous preprocessing, I simply dropped these missing values because it only accounts for 7% of the entire data set
- Remove outliers. Outliers will have high leverage and might move the analysis towards another direction

Also, I can check the correlation between continuous variables to see if there is any high correlation.

```
library(ggcorrplot)

#index vector numeric variables
numericVars <- which(sapply(raw_2, FUN = is.numeric))

#saving names for use later on
numericVarNames <- names(numericVars)

cat("There are", length(numericVarNames), "numeric variables")
```

```
## There are 5 numeric variables
```

```
raw_numVar <- raw_2[, numericVars]
corr <- cor(raw_numVar, use = 'pairwise.complete.obs')

ggcorrplot(corr, lab = TRUE)
```



I don't see any high correlation here, so we are good to include them for future analysis.

Another thing I can do with categorical variables is to check their relationship with the class variable. If I see any variable that has low chi-square value and high p-value, I will know the categorical variable is independent on the class variable, so it is useless to include them in the model training stage.

```
chi.square <- vector()
p.value <- vector()
cateVar <- raw_2 %>%
  dplyr::select(-income) %>%
  purrr::keep(is.factor)

for (i in 1:length(cateVar)) {
 p.value[i] <- chisq.test(raw_2$income, unname(unlist(cateVar[i])), correct = FALSE)[3]$p.value
 chi.square[i] <- unname(chisq.test(raw_2$income, unname(unlist(cateVar[i])), correct = FALSE)[1]$stati
}

chi_sqaure_test <- tibble(variable = names(cateVar)) %>%
  add_column(chi.square = chi.square) %>%
  add_column(p.value = p.value)
knitr::kable(chi_sqaure_test)
```

variable	chi.square	p.value
workclass	806.6021	0
education	4072.6672	0
marital_status	6063.7772	0
occupation	3690.5126	0

variable	chi.square	p.value
relationship	6235.7589	0
race	304.5680	0
sex	1416.1243	0
native_country	317.6842	0
Race_Sex	1623.8228	0

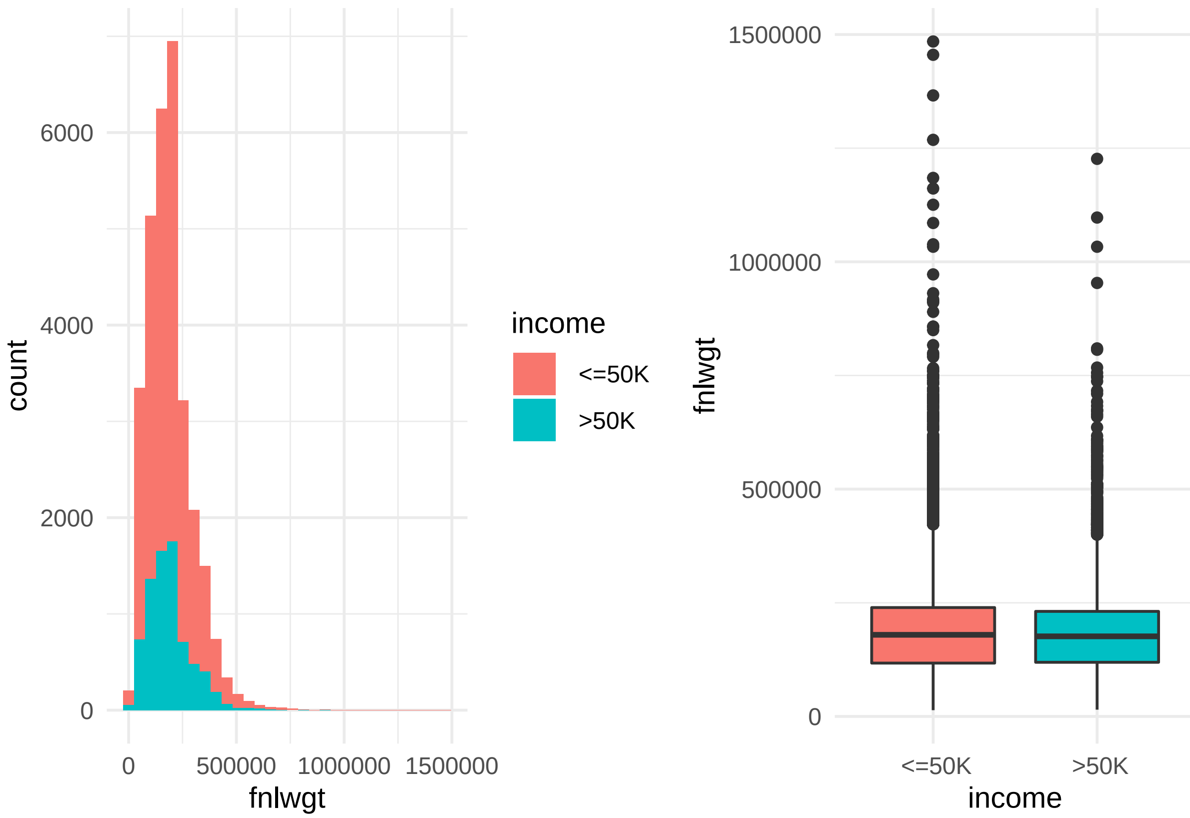
I'll keep all of the categorical variables because they are all dependent on the class variable based on the chi-square test.

Question 4:

The column "fnlwgt" is a continuous variable that has a complicated, interconnected definition. For this column is a higher value or a lower value more likely to predict high income?

```
library(patchwork) # for displaying the plots

fnlwgt_histogram <- ggplot(raw_2, aes(x = fnlwgt, fill = income)) +
  geom_histogram() +
  theme_minimal()

fnlwgt_boxplot <- ggplot(raw_2, aes(x = income, y = fnlwgt, fill = income)) +
  geom_boxplot() +
  theme_minimal() +
  theme(legend.position = 'none')

fnlwgt_histogram | fnlwgt_boxplot
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The distribution of these two groups are quite similar. There is no big difference between the mean of two groups.

I'll do a t-test to check if the difference between two groups is significant.

- *Null hypothesis*: the difference between two groups is not significant
- *Alternative hypothesis*: the difference between the two groups is significant

```
t.test(raw_2$fnlwgt ~ raw_2$income, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  raw_2$fnlwgt by raw_2$income
## t = 1.5919, df = 13246, p-value = 0.1114
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -508.7977 4908.1920
## sample estimates:
## mean in group  <=50K  mean in group  >50K
##             190349.7             188150.0
```

The mean `fnlwgt` of higher income group is lower than the other group, so it is possible that lower fnlwgt value will lead to higher income, but the p-value is 0.11, which is greater than our pre-determined threshold 0.05, so we failed to reject the null hypothesis. We know the mean difference between two groups isn't significant. We won't be able to tell if lower fnlwgt value or higher fnlwgt value will lead to higher income.

Question 5:

If we could only have access to one of the columns (not the target column) and still needed to make an income prediction, which column would you choose and why? What if you could have access to 3 columns?

There are a lot of feature selection techniques in the wild, such as Lasso, random forest, and xgboost. Here I'll use decision tree to select the best predictor(s) becuase it is simple and fast.

The decision tree algorithm that I'm going to implement uses *Gini impurity* measure to determine the optimal feature to split upon.

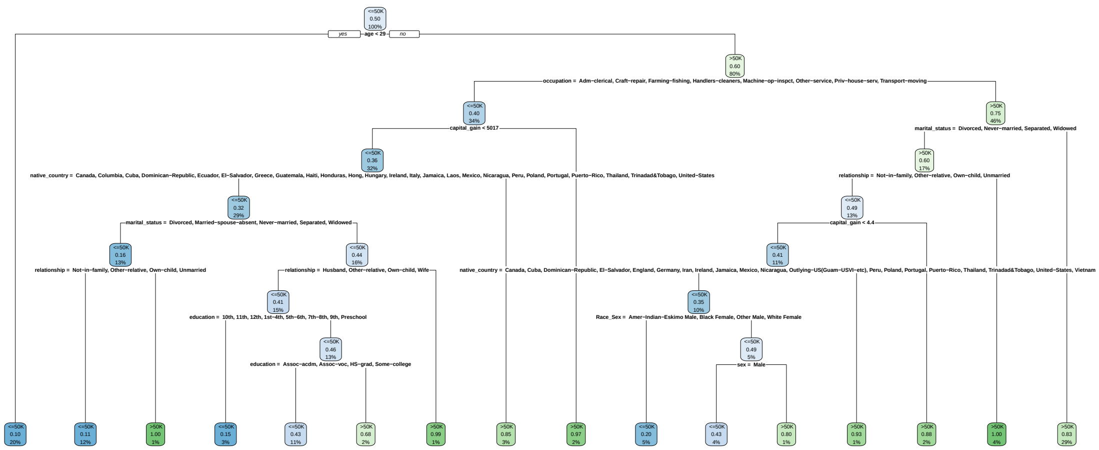
Even though we're not going to do the predictions here, but it is still a good idea to split the data and handle the class imbalanced problem before building the model.

```r
## split the data using a stratified sampling approach.
library(caTools)
set.seed(888)
sample_set <- raw_2 %>%
  pull(.) %>%
  sample.split(SplitRatio = .7)

raw_train <- subset(raw_2, sample_set == TRUE)
raw_test <- subset(raw_2, sample_set == FALSE)
```

```r
## use SMOTE to handle class imbalance
library(DMwR)
set.seed(888)
raw_train <- SMOTE(income ~ ., data.frame(raw_train), perc.over = 100, perc.under = 200)
```

Then we can put the data into the decision tree model.

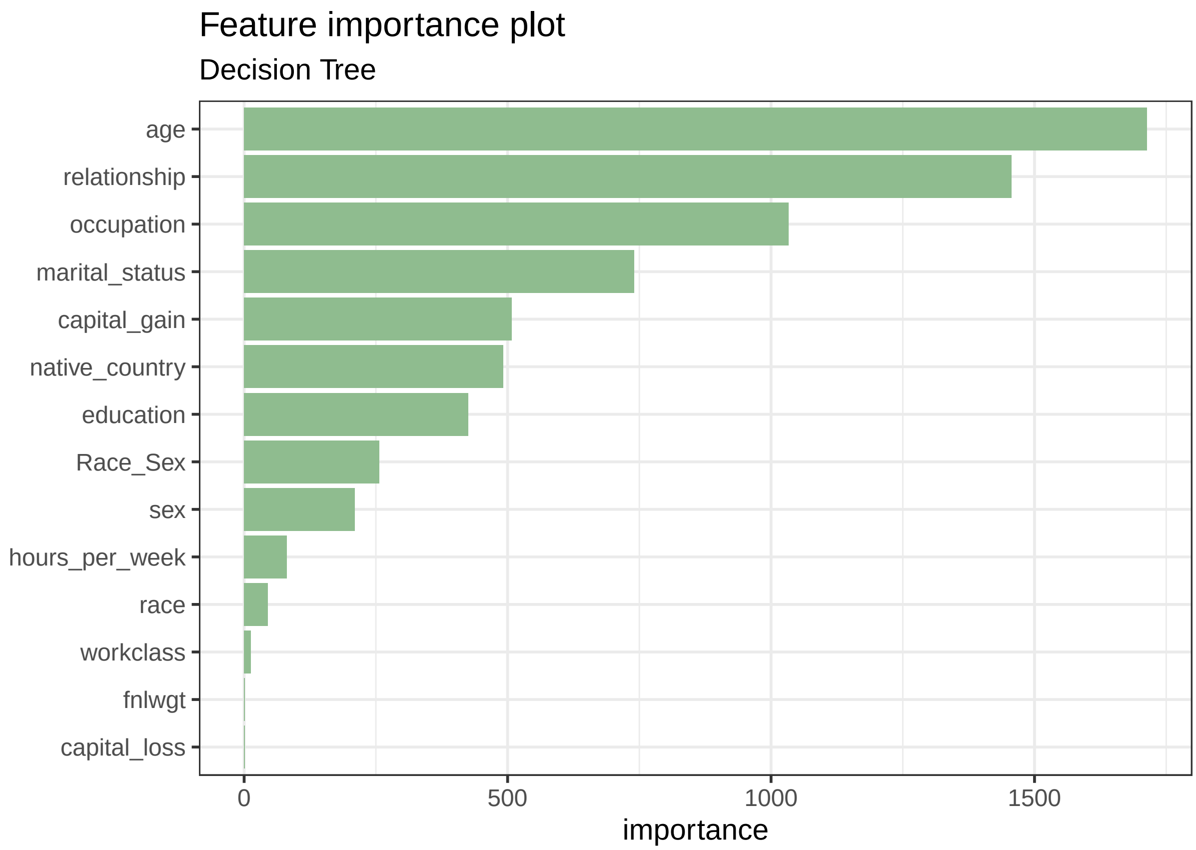
```r
library(rpart)
library(rpart.plot)
tree.mod <-
  rpart(
    income ~.,
    method = "class",
    data = raw_train,
    control = rpart.control(cp = 0.004)
  )

rpart.plot(tree.mod)
```



The root node is `age`. So if I can only get access to one column, I will use `age` to try to make the best predictions because it will lead to a best quality of a split.

```
tree.importance <- tree.mod$variable.importance

#barplot(t(tree.mod$variable.importance),horiz=TRUE)

importance <- as.data.frame(tree.mod$variable.importance)

names(importance) <- c("importance")

importance <- cbind(feature = rownames(importance), importance)

rownames(importance) <- 1:nrow(importance)

importance %>%
  arrange(desc(importance)) %>%
  top_n(20) %>%
  ggplot(aes(x = reorder(feature, importance), y = importance)) +
  geom_col(fill = "darkseagreen") +
  coord_flip() +
  theme_bw() +
  labs(title = "Feature importance plot", subtitle = "Decision Tree", x= "")
```

```
## Selecting by importance
```



If I can use three columns, I will use: `age`, `occupation` and `relationship` based on the feature importance (gini impurity measure).

Question 6:

What level of education should you achieve if you want to have a better than 50% chance of making more than $50K (per this data set)?

I'll run `education` on `income` with logistic regression because the coefficent outputs can be converted to probabilities.

```
logit_mod <- glm(income ~ education, family = binomial(link = "logit"), data = raw_2)
summary(logit_mod)
```

```
##
## Call:
## glm(formula = income ~ education, family = binomial(link = "logit"),
##     data = raw_2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6629  -0.6681  -0.5992  -0.0016   2.5399
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -2.55972    0.13513 -18.943  < 2e-16 ***
```

```
## education 11th          -0.26045    0.19031  -1.369   0.1711
## education 12th           0.07481    0.23583   0.317   0.7511
## education 1st-4th       -0.62525    0.43798  -1.428   0.1534
## education 5th-6th       -0.57577    0.32437  -1.775   0.0759 .
## education 7th-8th       -0.14451    0.22078  -0.655   0.5128
## education 9th           -0.28519    0.24614  -1.159   0.2466
## education Assoc-acdm     1.48216    0.15328   9.669  < 2e-16 ***
## education Assoc-voc      1.53031    0.14901  10.270  < 2e-16 ***
## education Bachelors      2.24306    0.13810  16.242  < 2e-16 ***
## education Doctorate      3.64063    0.17988  20.239  < 2e-16 ***
## education HS-grad        0.93324    0.13784   6.770 1.28e-11 ***
## education Masters        2.81806    0.14408  19.559  < 2e-16 ***
## education Preschool    -11.00635   79.81450  -0.138   0.8903
## education Prof-school    3.65342    0.16756  21.804  < 2e-16 ***
## education Some-college   1.17343    0.13855   8.469  < 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: 33855  on 30168  degrees of freedom
## Residual deviance: 29947  on 30153  degrees of freedom
## AIC: 29979
##
## Number of Fisher Scoring iterations: 12
```

The formula can be expressed as:

$$income = -2.556 - 0.254*11th + 0.098*12th - 0.663*1st-4thgrade + ... + 3.621*Prof-school + 1.167*Some-college$$

I'll convert the log odds to probabilities.

```
library(gdata) #for trim
odds_value <- vector() #store odds
prob_value <- vector() #store probs

#remove extra space
education_levels <- trim(levels(raw_2$education)[-1])

#add eduction before corresponding grade name
education_levels <- paste("education", education_levels)


for (i in 1:length(education_levels)) {
  #calculate odds
  odds_value[i] <- exp(coef(logit_mod)["(Intercept)"] + coef(logit_mod)[education_levels[i]])
  #prob = odds / (1+odds)
  prob_value[i] <- odds_value[i] / (odds_value[i] + 1)
}

#make it as a data frame
result <- tibble(variable = trim(levels(raw_2$education)[-1])) %>%
```

```r
  add_column(odds_value = odds_value) %>%
  add_column(prob_value = prob_value) %>%
  arrange(desc(prob_value))

knitr::kable(result)
```

variable	odds_value	prob_value
Prof-school	2.9852941	0.7490775
Doctorate	2.9473684	0.7466667
Masters	1.2947814	0.5642286
Bachelors	0.7285812	0.4214909
Assoc-voc	0.3572170	0.2631982
Assoc-acdm	0.3404255	0.2539683
Some-college	0.2500000	0.2000000
HS-grad	0.1966196	0.1643126
12th	0.0833333	0.0769231
7th-8th	0.0669216	0.0627240
11th	0.0595960	0.0562440
9th	0.0581395	0.0549451
5th-6th	0.0434783	0.0416667
1st-4th	0.0413793	0.0397351
Preschool	0.0000013	0.0000013

Based on the result table, we can see if you want to have a better than 50% chance of making more than $50k, you should achieve at least a master degree.