Capstone Adult Census Income Report

Wanjun Ling

7/12/2021

Contents

1 EXECUTIVE SUMMARY **1**

1.1 Introduction 1

1.2 Data Set 1

1.3 Objective 2

1.4 Key Steps 2

2 METHODS **2**

2.1 Data Preparation 2

2.2 Data Exploration 2

2.3 Data Pre-processing 9

2.4 Data Analysis 10

3 CONCLUSION **52**

3.1 Conclusion 52

3.2 Future Work 53

1 EXECUTIVE SUMMARY

1.1 Introduction

This project aims to tackle an income classification problem on Adult Census Data. The 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)). The prediction goal is to build a model that predicts whether a person made over $50K a year as best accuracy as possible.

1.2 Data Set

In this project adult.csv with 32561 observations and 15 columns downloaded from kaggle.com will be used as data source.

1.3 Objective

My goal is to build the model that predicts whether someone earned more than $50k with best accuracy as possible.

1.4 Key Steps

The key steps executed in this project includes:

1. DATA PREPARATION: create data frame "adult" from adult.csv by read.csv.
2. EXPLORATORY ANALYSIS: Collect data set statistics, analyze and visualize collarations between multiple continous & categorical variables, then predict target "income".
3. MODELING: Build and evaluate multiple models (logistic regression, classification (decision) tree and random forest) on predicting whether a given adult makes more than 50k.
4. CONCLUSION: Draw a conclusion based on modeling results and provide future research recommendations.

2 METHODS

2.1 Data Preparation

Download adult.csv through "https://www.kaggle.com/uciml/adult-census-income" and create data set "adult" through read.csv().

2.2 Data Exploration

```
##########################################################
# Explore basic statistics, Split it into train set adult_train and validation test set adult_test
##########################################################

# take a first glance of data set "adult" to understand total observations, variables and what are they
str(adult)
```

```
## 'data.frame':    32561 obs. of  15 variables:
##  $ age           : int  90 82 66 54 41 34 38 74 68 41 ...
##  $ workclass     : chr  "?" "Private" "?" "Private" ...
##  $ fnlwgt        : int  77053 132870 186061 140359 264663 216864 150601 88638 422013 70037 ...
##  $ education     : chr  "HS-grad" "HS-grad" "Some-college" "7th-8th" ...
##  $ education.num : int  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  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ capital.loss  : int  4356 4356 4356 3900 3900 3770 3770 3683 3683 3004 ...
##  $ hours.per.week: int  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" ...
```

```r
head(adult)
```

```
##   age workclass fnlwgt    education education.num marital.status
## 1  90         ?  77053      HS-grad             9        Widowed
## 2  82   Private 132870      HS-grad             9        Widowed
## 3  66         ? 186061 Some-college            10        Widowed
## 4  54   Private 140359      7th-8th             4       Divorced
## 5  41   Private 264663 Some-college            10      Separated
## 6  34   Private 216864      HS-grad             9       Divorced
##          occupation  relationship  race    sex capital.gain capital.loss
## 1                 ? Not-in-family White Female            0         4356
## 2   Exec-managerial Not-in-family White Female            0         4356
## 3                 ?     Unmarried Black Female            0         4356
## 4 Machine-op-inspct     Unmarried White Female            0         3900
## 5    Prof-specialty     Own-child White Female            0         3900
## 6     Other-service     Unmarried White Female            0         3770
##   hours.per.week native.country income
## 1             40  United-States  <=50K
## 2             18  United-States  <=50K
## 3             40  United-States  <=50K
## 4             40  United-States  <=50K
## 5             40  United-States  <=50K
## 6             45  United-States  <=50K
```

```r
dim(adult)
```

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

```r
# get the basic statistics of data set "adult"
summary(adult)
```

```
##       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
```

```
# check any variables with null value
anyNA(adult)
```

```
## [1] FALSE
```

From above exploration output, we know there are total 32561 observations of 15 variables and no null value in all 15 variables. While we observe there are some variables with character of "?" which need to be handled before further data analysis and processing.

Now we check all variables one by one to see any invalid value exists as well as the proportion of it.

```
unique(adult$age)
```

```
##  [1] 90 82 66 54 41 34 38 74 68 45 52 32 51 46 57 22 37 29 61 21 33 49 23 59 60
## [26] 63 53 44 43 71 48 73 67 40 50 42 39 55 47 31 58 62 36 72 78 83 26 70 27 35
## [51] 81 65 25 28 56 69 20 30 24 64 75 19 77 80 18 17 76 79 88 84 85 86 87
```

Variable "age" is a continuous one.

```
unique(adult$workclass)
```

```
## [1] "?"                "Private"          "State-gov"        "Federal-gov"
## [5] "Self-emp-not-inc" "Self-emp-inc"     "Local-gov"        "Without-pay"
## [9] "Never-worked"
```

Variable "workclass" is a categorical one which contains invalid value "?".

```
unique(adult$education)
```

```
##  [1] "HS-grad"      "Some-college" "7th-8th"      "10th"         "Doctorate"
##  [6] "Prof-school"  "Bachelors"    "Masters"      "11th"         "Assoc-acdm"
## [11] "Assoc-voc"    "1st-4th"      "5th-6th"      "12th"         "9th"
## [16] "Preschool"
```

Variable "education" is a categorical one.

```
unique(adult$education.num)
```

```
##  [1]  9 10  4  6 16 15 13 14  7 12 11  2  3  8  5  1
```

Variable "education.num" is a continuous one.

```r
unique(adult$marital.status)
```

```
## [1] "Widowed"               "Divorced"              "Separated"            
## [4] "Never-married"         "Married-civ-spouse"    "Married-spouse-absent"
## [7] "Married-AF-spouse"
```

Variable "marital.status" is a categorical one.

```r
unique(adult$occupation)
```

```
##  [1] "?"                 "Exec-managerial"   "Machine-op-inspct"
##  [4] "Prof-specialty"    "Other-service"     "Adm-clerical"     
##  [7] "Craft-repair"      "Transport-moving"  "Handlers-cleaners"
## [10] "Sales"             "Farming-fishing"   "Tech-support"     
## [13] "Protective-serv"   "Armed-Forces"      "Priv-house-serv"
```

Variable "occupation" is a categorical variables which contains invalid value "?".

```r
unique(adult$relationship)
```

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

Variable "relationship" is a categorical one.

```r
unique(adult$race)
```

```
## [1] "White"              "Black"              "Asian-Pac-Islander"
## [4] "Other"              "Amer-Indian-Eskimo"
```

Variable "race" is a categorical one.

```r
unique(adult$sex)
```

```
## [1] "Female" "Male"
```

Variable "sex" is a categorical one.

```r
unique(adult$capital.gain)
```

```
##   [1]     0 99999 41310 34095 27828 25236 25124 22040 20051 18481 15831 15024
##  [13] 15020 14344 14084 13550 11678 10605 10566 10520  9562  9386  8614  7978
##  [25]  7896  7688  7443  7430  7298  6849  6767  6723  6514  6497  6418  6360
##  [37]  6097  5721  5556  5455  5178  5060  5013  4934  4931  4865  4787  4687
##  [49]  4650  4508  4416  4386  4101  4064  3942  3908  3887  3818  3781  3674
##  [61]  3471  3464  3456  3432  3418  3411  3325  3273  3137  3103  2993  2977
##  [73]  2964  2961  2936  2907  2885  2829  2653  2635  2597  2580  2538  2463
##  [85]  2414  2407  2387  2354  2346  2329  2290  2228  2202  2176  2174  2105
##  [97]  2062  2050  2036  2009  1848  1831  1797  1639  1506  1471  1455  1424
## [109]  1409  1173  1151  1111  1086  1055   991   914   594   401   114
```

Variable "capital.gain" is a continuous one.

```
unique(adult$capital.loss)
```

```
##  [1] 4356 3900 3770 3683 3004 2824 2754 2603 2559 2547 2489 2472 2467 2457 2444
## [16] 2415 2392 2377 2352 2339 2282 2267 2258 2246 2238 2231 2206 2205 2201 2179
## [31] 2174 2163 2149 2129 2080 2057 2051 2042 2002 2001 1980 1977 1974 1944 1902
## [46] 1887 1876 1848 1844 1825 1816 1762 1755 1741 1740 1735 1726 1721 1719 1672
## [61] 1669 1668 1651 1648 1628 1617 1602 1594 1590 1579 1573 1564 1539 1504 1485
## [76] 1411 1408 1380 1340 1258 1138 1092  974  880  810  653  625  419  323  213
## [91]  155    0
```

Variable “capital.loss” is a continuous one.

```
unique(adult$hours.per.week)
```

```
##  [1] 40 18 45 20 60 35 55 76 50 42 25 32 90 48 15 70 52 72 39  6 65 12 80 67 99
## [26] 30 75 26 36 10 84 38 62 44  8 28 59  5 24 57 34 37 46 56 41 98 43 63  1 47
## [51] 68 54  2 16  9  3  4 33 23 22 64 51 19 58 53 96 66 21  7 13 27 11 14 77 31
## [76] 78 49 17 85 87 88 73 89 97 94 29 82 86 91 81 92 61 74 95
```

Variable “hours.per.week” is a continuous one.

```
unique(adult$native.country)
```

```
##  [1] "United-States"              "?"
##  [3] "Mexico"                     "Greece"
##  [5] "Vietnam"                    "China"
##  [7] "Taiwan"                     "India"
##  [9] "Philippines"                "Trinadad&Tobago"
## [11] "Canada"                     "South"
## [13] "Holand-Netherlands"         "Puerto-Rico"
## [15] "Poland"                     "Iran"
## [17] "England"                    "Germany"
## [19] "Italy"                      "Japan"
## [21] "Hong"                       "Honduras"
## [23] "Cuba"                       "Ireland"
## [25] "Cambodia"                   "Peru"
## [27] "Nicaragua"                  "Dominican-Republic"
## [29] "Haiti"                      "El-Salvador"
## [31] "Hungary"                    "Columbia"
## [33] "Guatemala"                  "Jamaica"
## [35] "Ecuador"                    "France"
## [37] "Yugoslavia"                 "Scotland"
## [39] "Portugal"                   "Laos"
## [41] "Thailand"                   "Outlying-US(Guam-USVI-etc)"
```

Variable “native.country” is a categorical one which contains invalid value “?”.

```
unique(adult$income)
```

```
## [1] "<=50K" ">50K"
```

Variable “income” is a categorical one.

Check all variables with invalid value “?”.

```
#Count the invalid value "?"
colSums(adult =="?")
```

```
##            age      workclass         fnlwgt      education  education.num
##              0           1836              0              0              0
## marital.status     occupation   relationship           race            sex
##              0           1843              0              0              0
##   capital.gain   capital.loss hours.per.week native.country         income
##              0              0              0            583              0
```

Now we have an idea of each variable and get the amount of invalid value “?” of variables “workclass”, “occupation” , “native.country”. We will calculate the percentage of invalid values to determine whether there will be significant impact on our prediction if we remove these observations.

```
sum(adult$workclass == "?")/nrow(adult)
```

```
## [1] 0.05638647
```

```
sum(adult$occupation == "?")/nrow(adult)
```

```
## [1] 0.05660146
```

```
sum(adult$native.country == "?")/nrow(adult)
```

```
## [1] 0.01790486
```

We can see the percentage of observations with invalid value “?” are much less (6% and 2%) than the ones with valid values. So we will remove these observations with invalid values in “workclass”, “occupation” and “native.country”.

```
# convert "?" to "NA"
adult[adult == "?"] <- NA

# Omitting NA values
adult <- na.omit(adult)
# Check again to make sure all observations are valid
colSums(adult =="?")
```

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

```r
anyNA(adult)
```

```
## [1] FALSE
```

Now we are fully aware of continuous variables. Befor we move to exploration of categorical variables, we'll split the data set "adult" to two data sets with 80:20 portion. One is training set "adult_train" and the other is validation set "adult_test".

```r
#split data set "adult" into "adult_train" and "adult_test" with percentage 80% and 20%
set.seed(20)
split <- sample.split(adult, SplitRatio = 0.8) # 80:20
adult_train <- subset(adult, split == TRUE)
adult_test <- subset(adult, split == FALSE)

str(adult_train)
```

```
## 'data.frame':    24129 obs. of  15 variables:
##  $ age           : int  54 41 34 68 45 38 52 32 46 45 ...
##  $ workclass     : Factor w/ 7 levels "Federal-gov",..: 3 3 3 1 3 5 3 3 3 3 ...
##  $ fnlwgt        : int  140359 264663 216864 422013 172274 164526 129177 136204 45363 172822 ...
##  $ education     : Factor w/ 16 levels "10th","11th",..: 6 16 12 12 11 15 10 13 15 2 ...
##  $ education.num : int  4 10 9 9 16 15 13 14 15 7 ...
##  $ marital.status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 1 6 1 1 1 5 7 6 1 1 ...
##  $ occupation    : Factor w/ 14 levels "Adm-clerical",..: 7 10 8 10 10 10 8 4 10 14 ...
##  $ relationship  : Factor w/ 6 levels "Husband","Not-in-family",..: 5 4 5 2 5 2 2 2 2 2 ...
##  $ race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 5 5 3 5 5 5 5 5 ...
##  $ sex           : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 2 1 2 2 2 ...
##  $ capital.gain  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ capital.loss  : int  3900 3900 3770 3683 3004 2824 2824 2824 2824 2824 ...
##  $ hours.per.week: int  40 40 45 40 35 45 20 55 40 76 ...
##  $ native.country: Factor w/ 41 levels "Cambodia","Canada",..: 39 39 39 39 39 39 39 39 39 39 ...
##  $ income        : chr  "<=50K" "<=50K" "<=50K" "<=50K" ...
```

```r
str(adult_test)
```

```
## 'data.frame':    6033 obs. of  15 variables:
##  $ age           : int  82 38 74 37 21 33 53 44 43 39 ...
##  $ workclass     : Factor w/ 7 levels "Federal-gov",..: 3 3 6 3 3 3 3 3 3 3 ...
##  $ fnlwgt        : int  132870 150601 88638 188774 34310 228696 149650 326232 115806 141584 ...
##  $ education     : Factor w/ 16 levels "10th","11th",..: 12 1 11 10 9 4 12 10 13 13 ...
##  $ education.num : int  9 6 16 13 11 2 9 13 14 14 ...
##  $ marital.status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 7 6 5 5 3 3 5 1 1 5 ...
##  $ occupation    : Factor w/ 14 levels "Adm-clerical",..: 4 1 10 4 3 3 12 4 4 12 ...
##  $ relationship  : Factor w/ 6 levels "Husband","Not-in-family",..: 2 5 3 2 1 2 2 5 5 2 ...
##  $ race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ sex           : Factor w/ 2 levels "Female","Male": 1 2 1 2 2 2 2 2 1 2 ...
##  $ capital.gain  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ capital.loss  : int  4356 3770 3683 2824 2603 2603 2559 2547 2547 2444 ...
##  $ hours.per.week: int  18 40 20 40 40 32 48 50 40 45 ...
##  $ native.country: Factor w/ 41 levels "Cambodia","Canada",..: 39 39 39 39 39 26 39 39 39 39 ...
##  $ income        : chr  "<=50K" "<=50K" ">50K" ">50K" ...
```

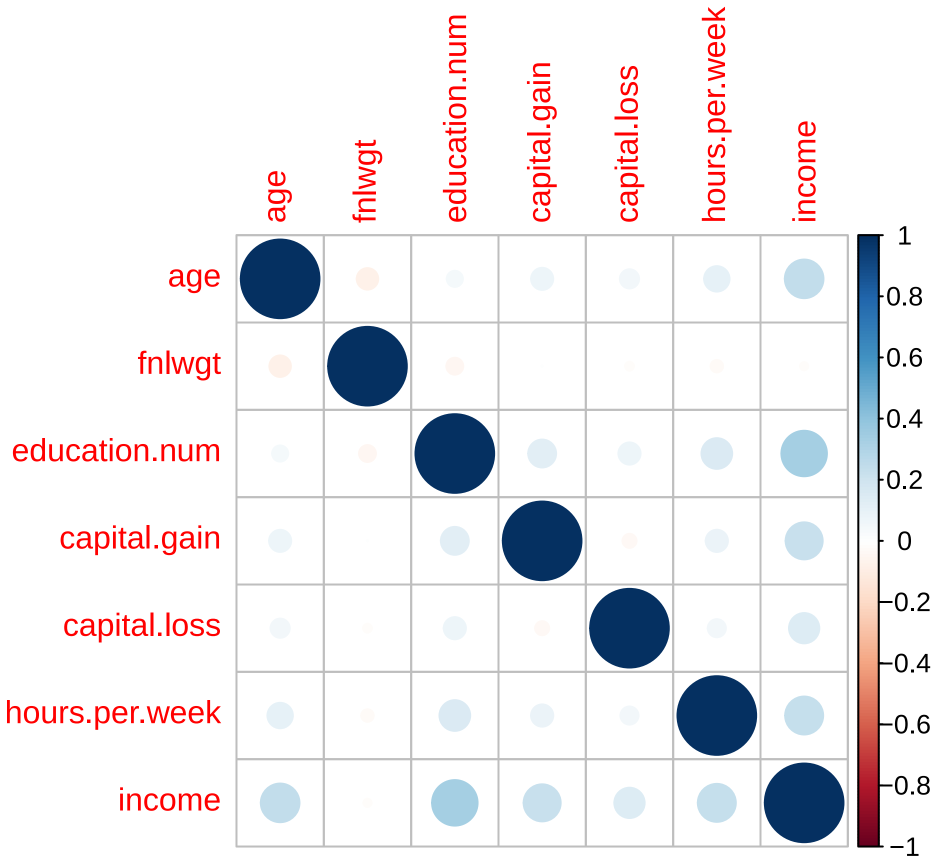
2.3 Data Pre-processing

Before we go to detailed data analysis, we will conduct minor data pre-processing on "adult_train" data sets. The data pre-processing tasks include converting feature "income" to numeric, omit ir-relative continuous variable "fnlwgt" .

```
# First check the correlation between continous variables "age", "fnlwgt", "capital.gain", "capital.los
# Before we check the correlation, we need convert "income" to numeric variable

adult_train$income<-ifelse(adult_train$income=='<=50K',0,1)

# list the correlations between continous variables "age", "fnlwgt",  "capital.gain", "capital.loss", "

continous_factors_cor <- cor(adult_train %>% select_if(is.numeric))
as.matrix(round(continous_factors_cor,3))
```

```
##                   age fnlwgt education.num capital.gain capital.loss
## age             1.000 -0.075         0.042        0.079        0.060
## fnlwgt         -0.075  1.000        -0.046        0.001       -0.013
## education.num   0.042 -0.046         1.000        0.126        0.080
## capital.gain    0.079  0.001         0.126        1.000       -0.032
## capital.loss    0.060 -0.013         0.080       -0.032        1.000
## hours.per.week  0.105 -0.026         0.152        0.080        0.053
## income          0.241 -0.011         0.335        0.222        0.149
##                hours.per.week income
## age                     0.105  0.241
## fnlwgt                 -0.026 -0.011
## education.num           0.152  0.335
## capital.gain            0.080  0.222
## capital.loss            0.053  0.149
## hours.per.week          1.000  0.233
## income                  0.233  1.000
```

```
# visualize the correlation between continous variables and income

columns <- c(1, 3, 5, 11, 12, 13, 15)
corrplot(cor(adult_train[,columns]))
```



From the correlation plot, we can see that these numerical variables do not seem to be highly correlated with target "income". However we still see that "education.num" is somehow correlated with target "income" with correlation 0.335. Then followed by "age" with correlation 0.241, "hours.per.week" with correlation 0.233 and "capital.gain" with correlation 0.222. The "fnlwgt", which may be some kind of weighting factor by guessing, is the lowest correlated with target "income" by correlation -0.011. Therefore we think "fnlwgt" can be ignored and dropped.

```
# drop variable "fnlwgt" from adult_train
adult_train <- adult_train[,-3]
```

```
adult_train$income <- mapvalues(adult_train$income, from = c(0,1), to = c('<=50K','>50K'))
adult_train$income <- as.factor(as.character(adult_train$income))
```

2.4 Data Analysis

After data pre-processing, we start further analysis on the distribution of target "income" in data set "adult_train".

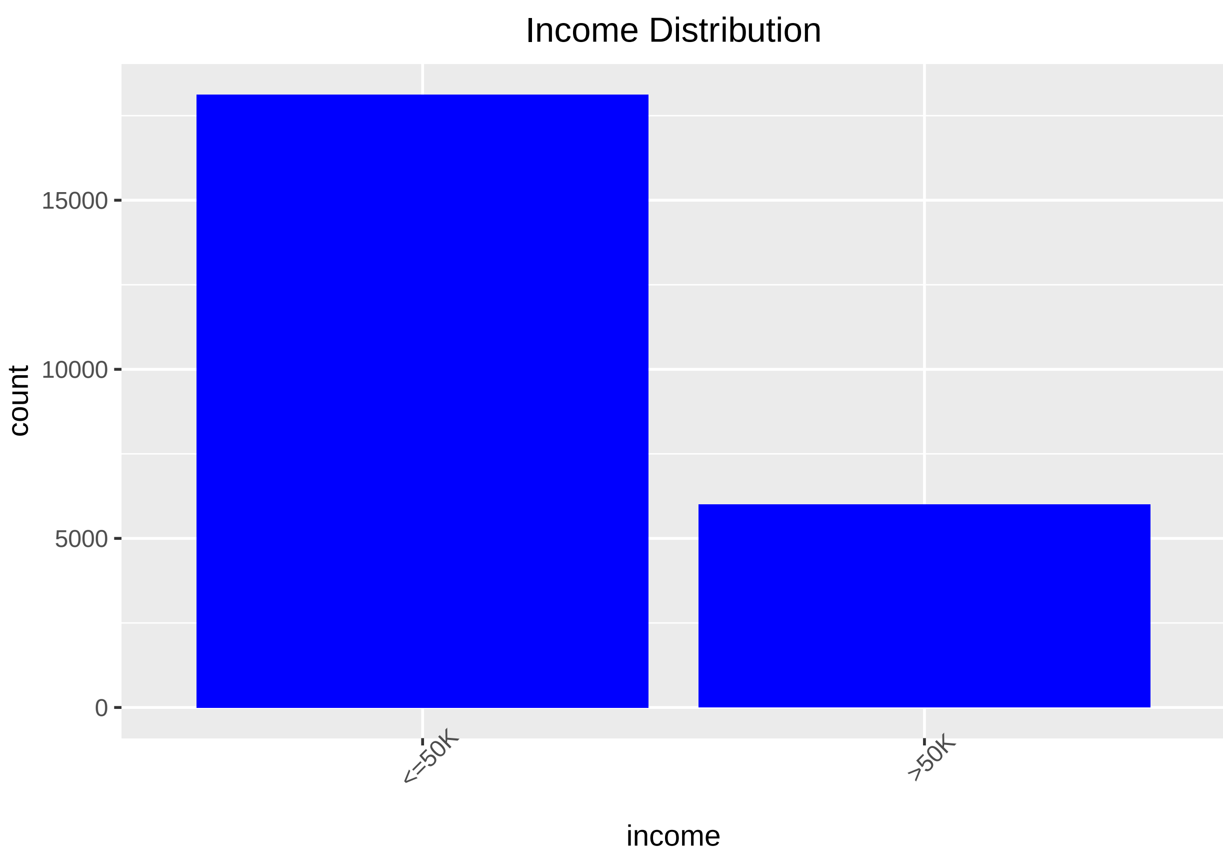
```
# Explore target "income" distribution
sum(adult_train$income == "<=50K")/nrow(adult_train)
```

```
## [1] 0.7510879
```

```r
sum(adult_train$income == ">50K")/nrow(adult_train)
```

```
## [1] 0.2489121
```

```r
#hist(adult_train$income)
```



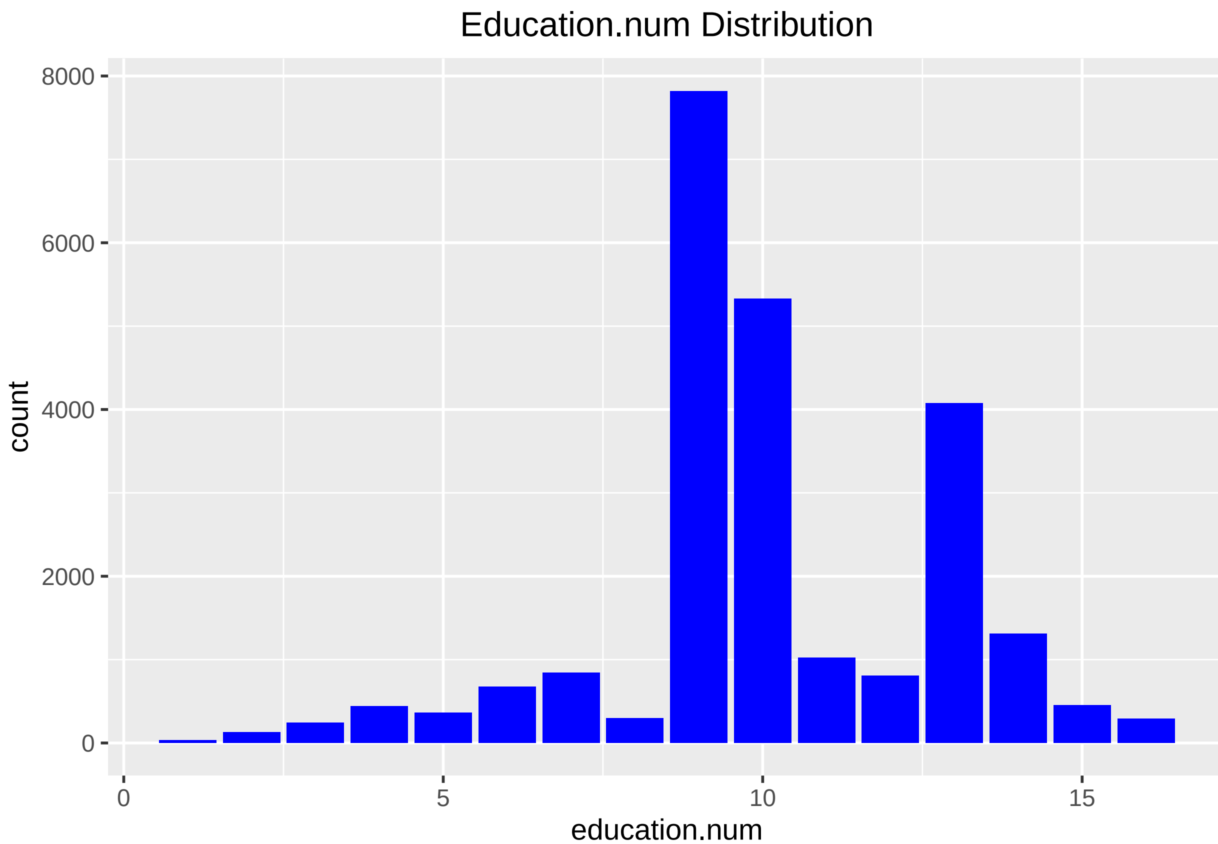
we can see that the target income has a very imbalanced distribution in data set "adult_train". Almost around 75% observations are below 50k income.This imbalanced feature could be a challenge to our predication. This will be verified in modeling session.

Since we have already known that there are multiple continuous variables in "adult_train" may be somehow relative with target income, we analyze the possible correlated variables first.

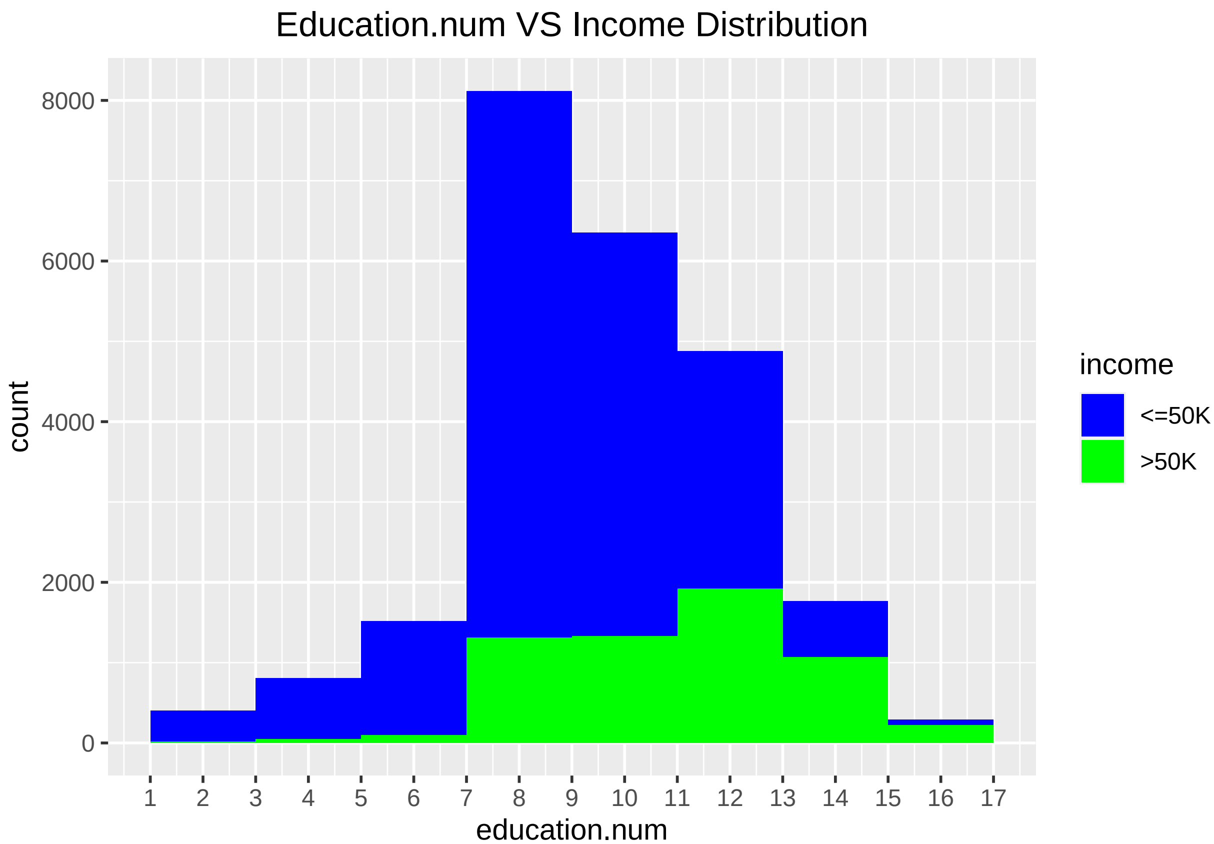
```r
#Further analysis on correlation between continuous variables and target "income"
# 1. Education.num
summary(adult_train$education.num)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    9.00   10.00   10.12   13.00   16.00
```

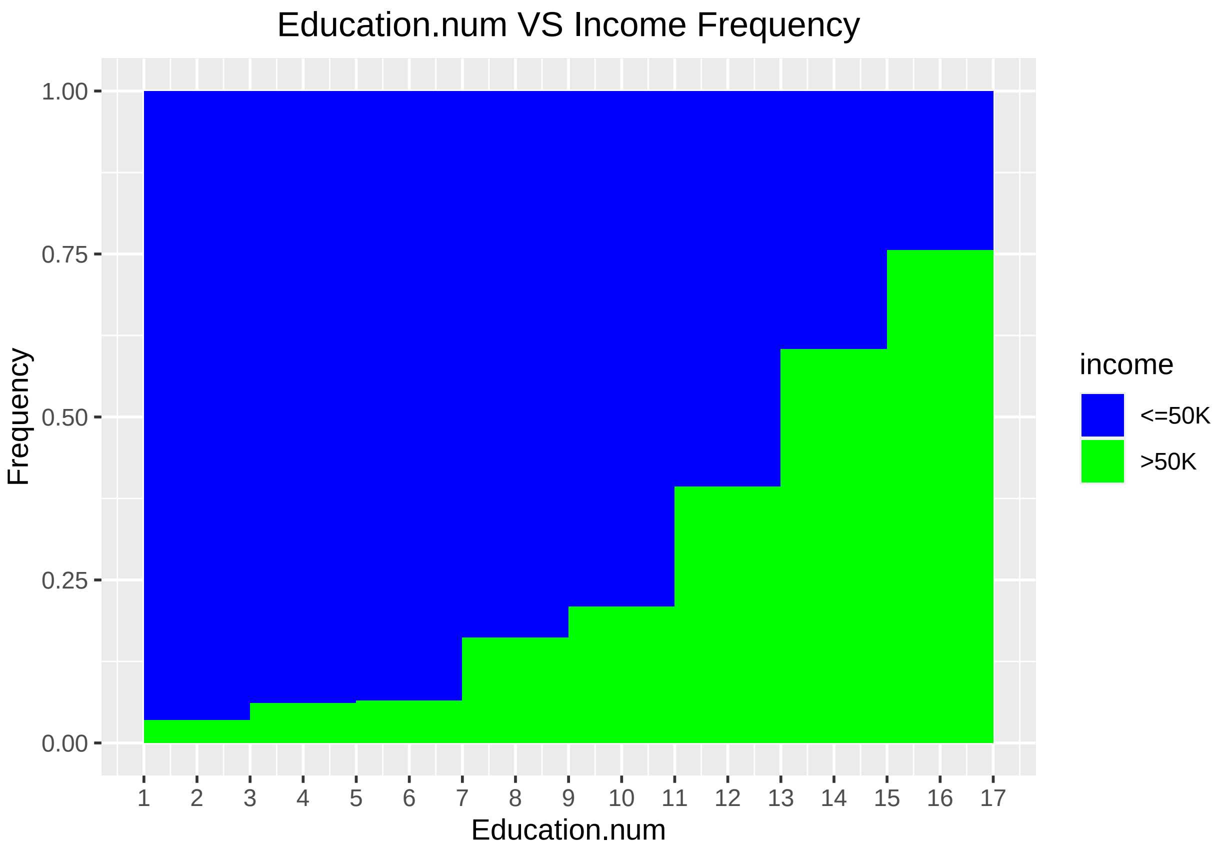
Now let us visualize the distribution of continuous variable "education.num" and distributions with target "income" together.



Plot “education.num” VS “income” by amount



Plot “education.num” VS “income” by frequency

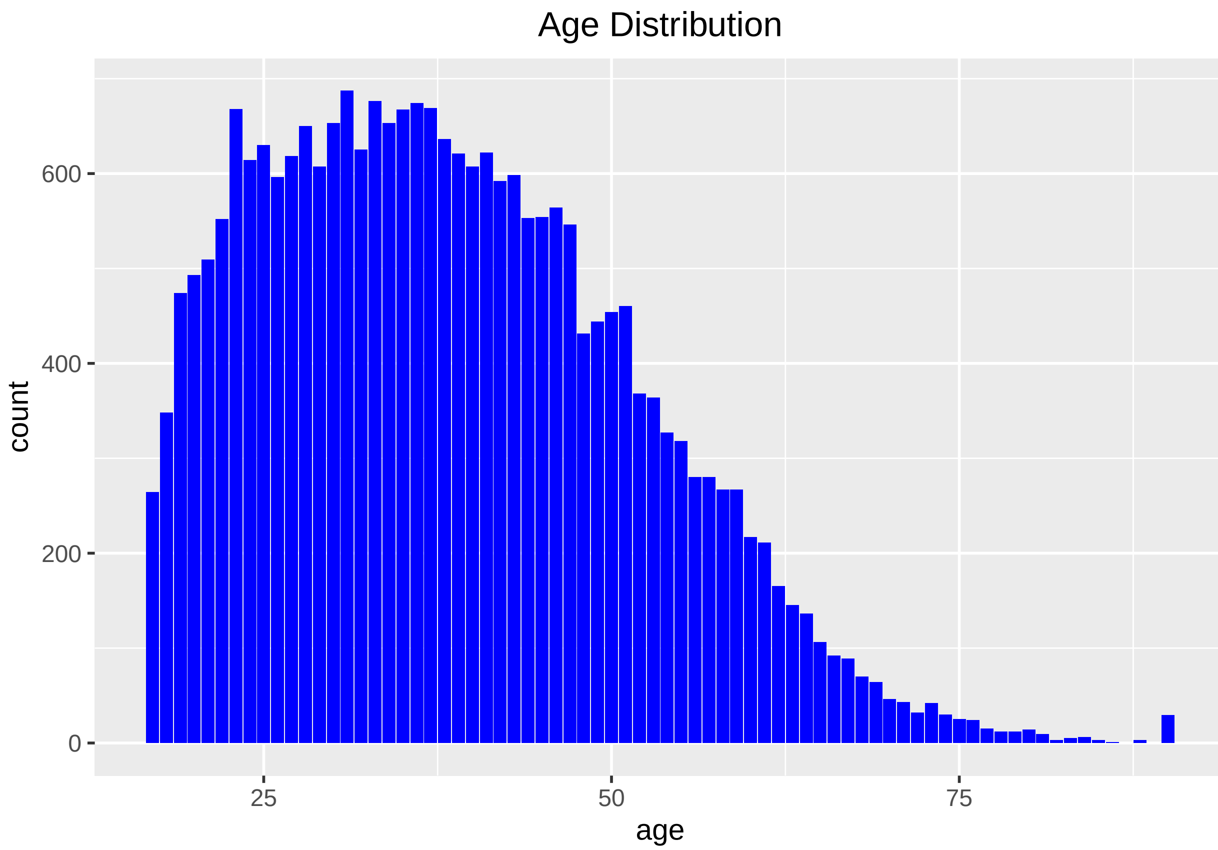


From above analysis and visualized plots we can see that both median and mean values of education.num are around 10. Most of people have education.num over 7. The plots also show the bigger number of the education.num the more likely to earn annual income over 50k.

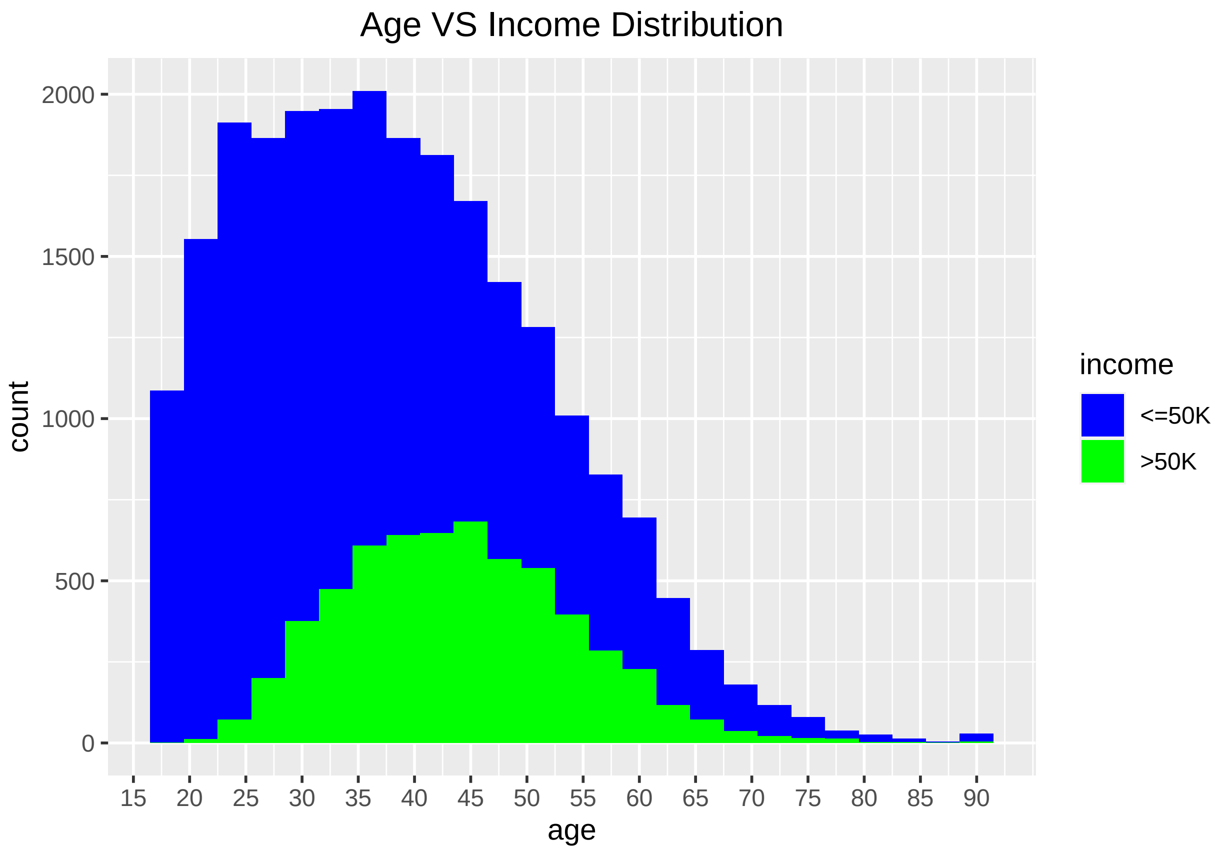
```
# 2. Age
summary(adult_train$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   28.00   37.00   38.44   47.00   90.00
```

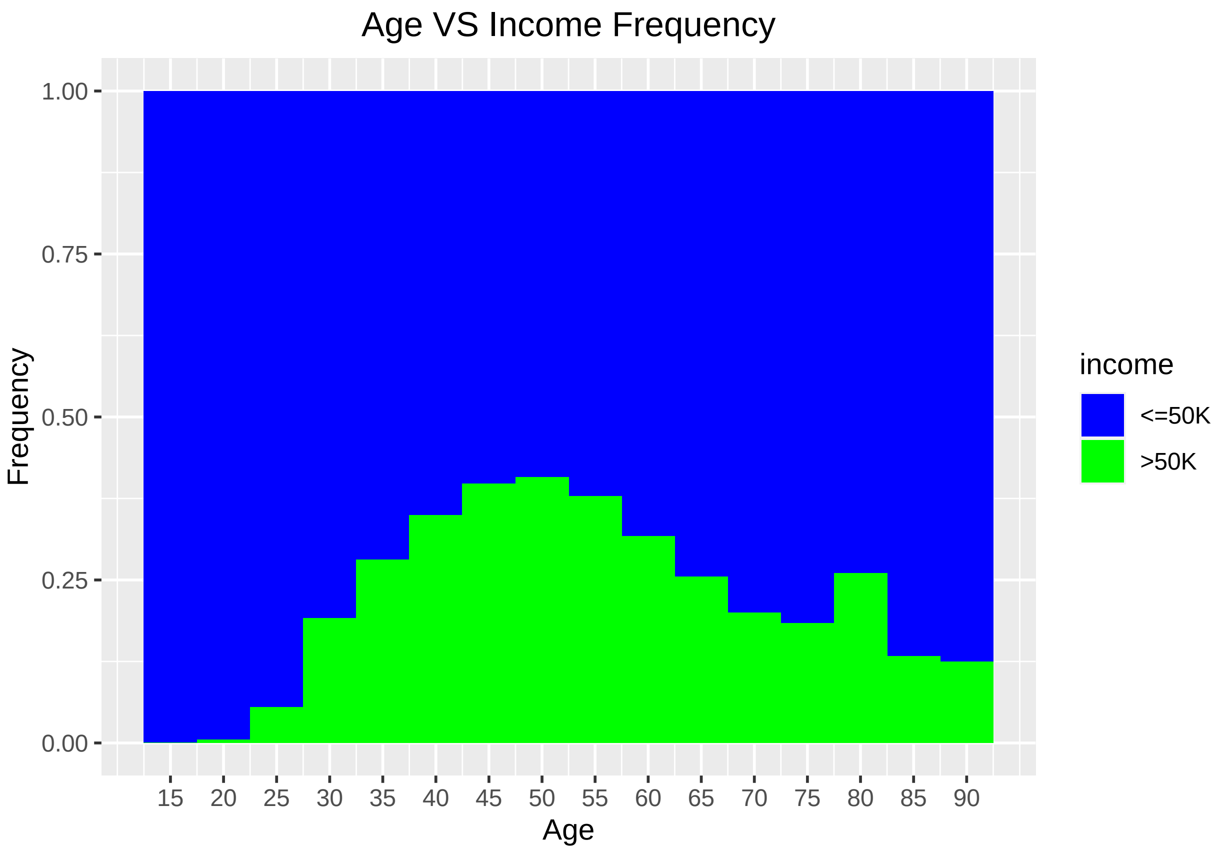
Now let us visualize the distribution of continuous variable "age" and distributions with target "income" together.



Plot "age" VS "income" by amount



Plot “education.num” VS “income” by frequency

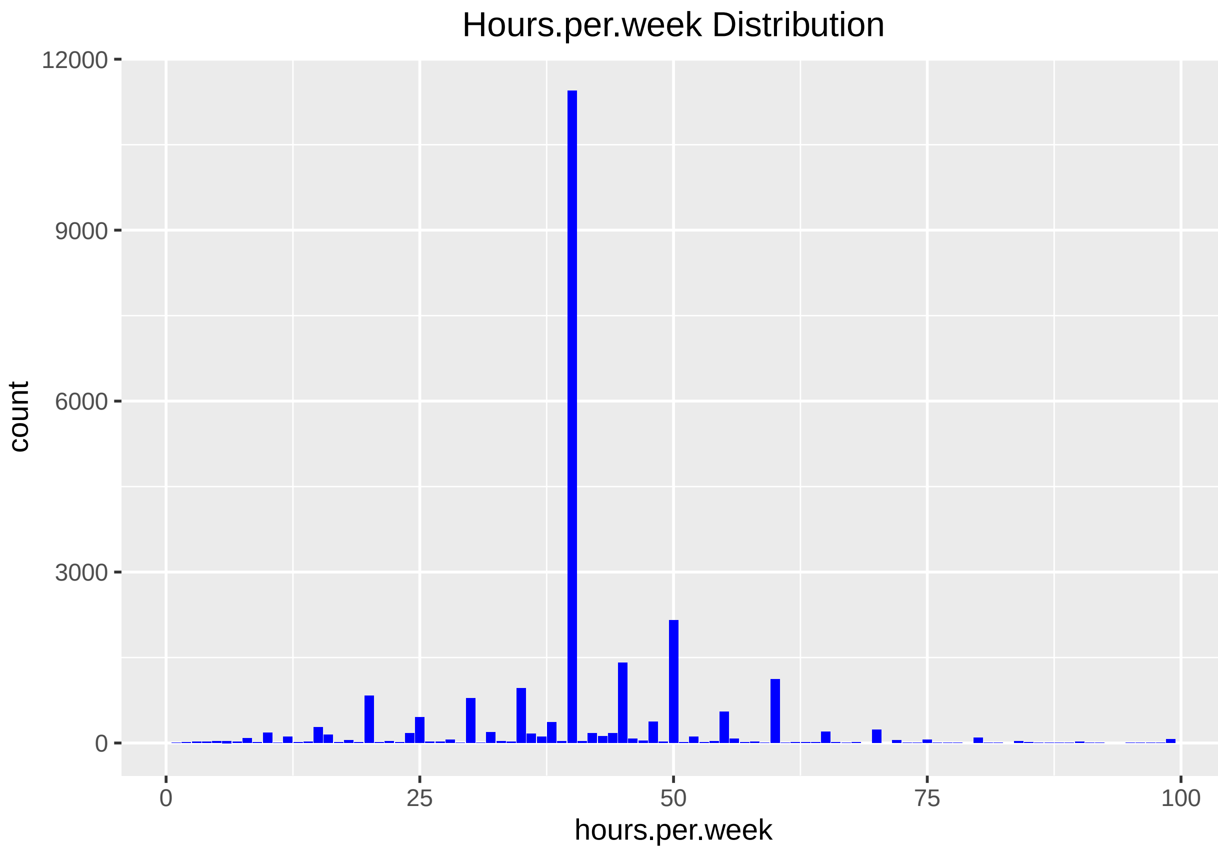


From above analysis and visualized plots we can see that the minimum and maximum values of age are 17 and 90. The mean value of age is around 38. The majority age distribution are between 28 and 47. From the charts above we can see that people who aged from 35 to 65 are more likely to have a income over 50k.

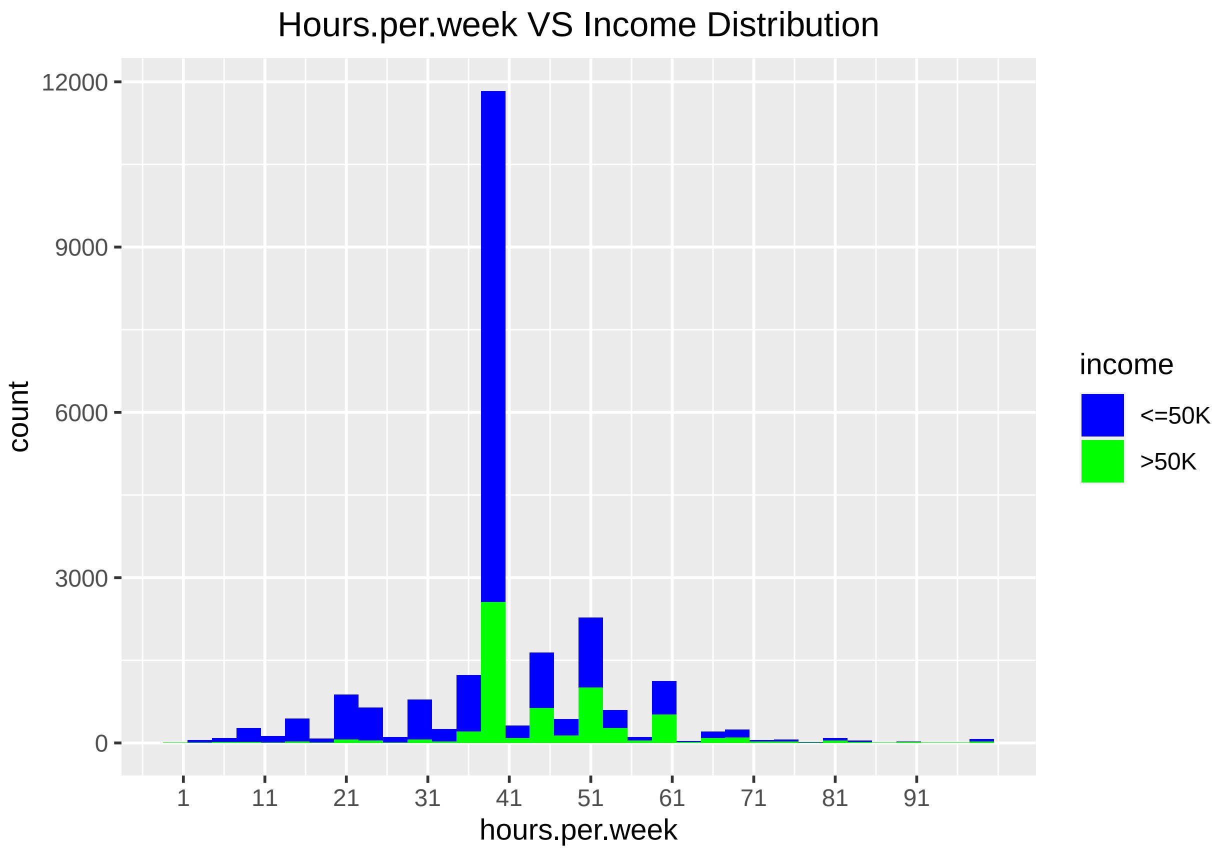
```
# 3. Hours.per.week
summary(adult_train$hours.per.week)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   40.00   40.00   40.94   45.00   99.00
```

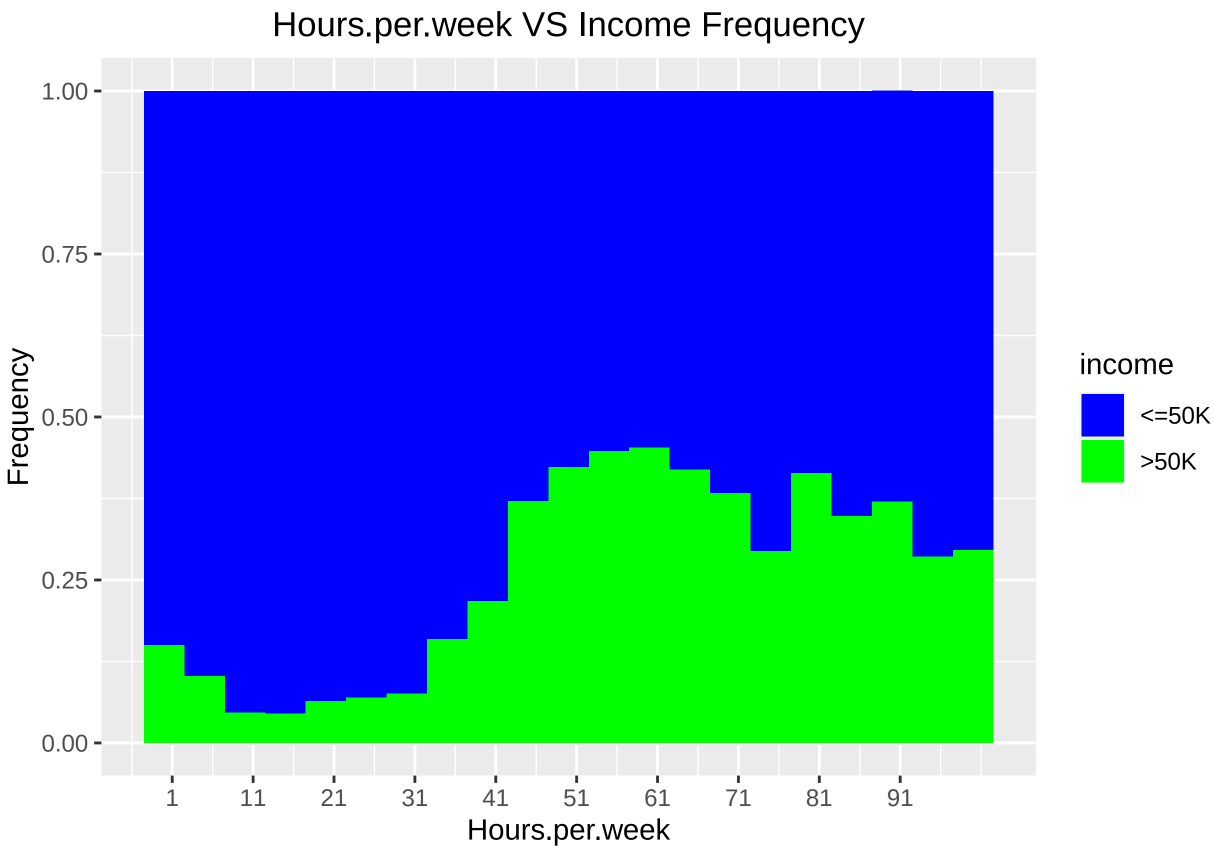
Now let us visualize the distribution of continuous variable “hours.per.week” and distributions with target “income” together.



Plot “hours.per.week” VS “income” by amount



Plot “hours.per.week” VS “income” by frequency

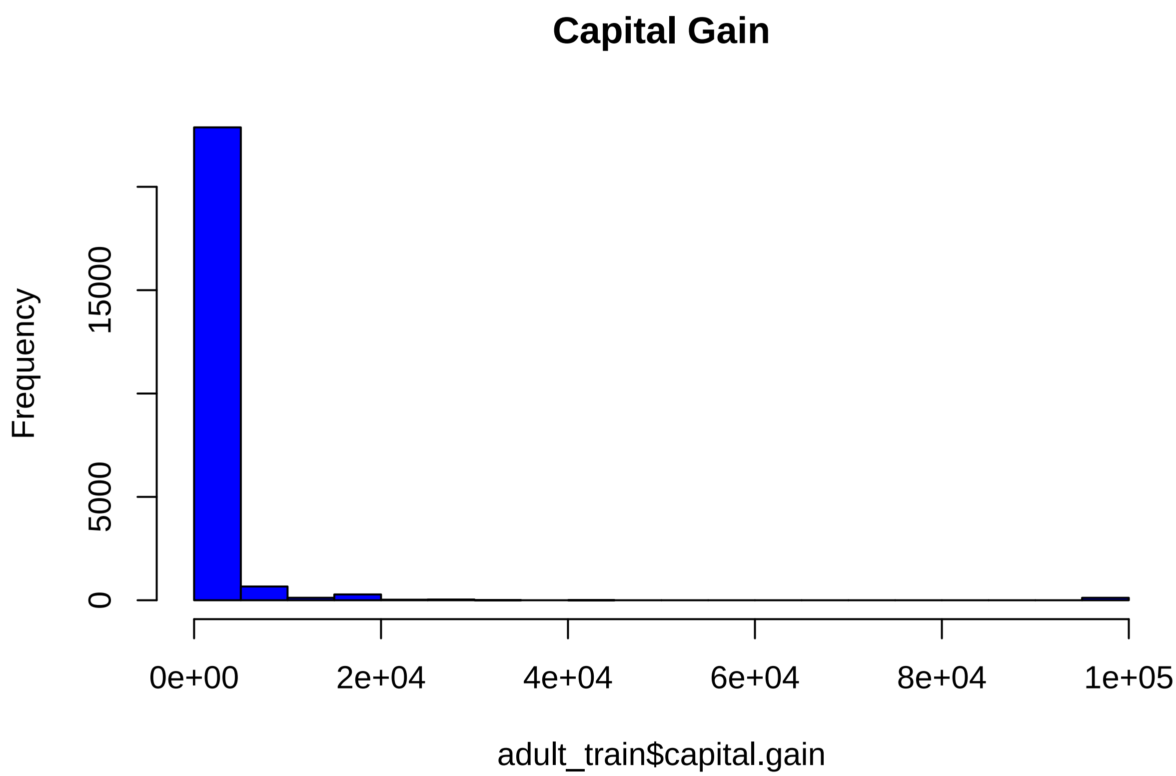


We can see that people who worked over 35 hours per week are more likely to have a income over 50k.

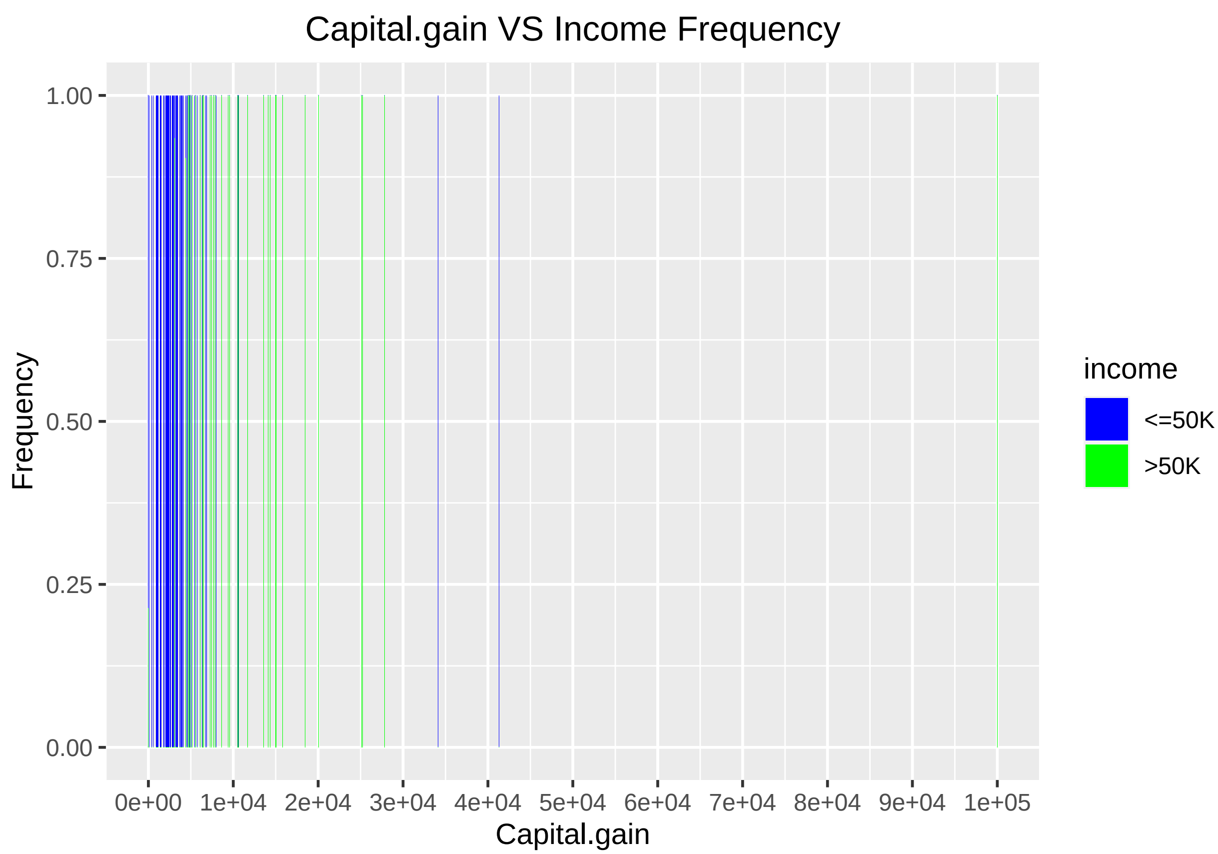
```
# 4. capital.gain & capital.loss & capital.gain - capital.loss
summary(adult_train$capital.gain)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0    1093       0   99999
```

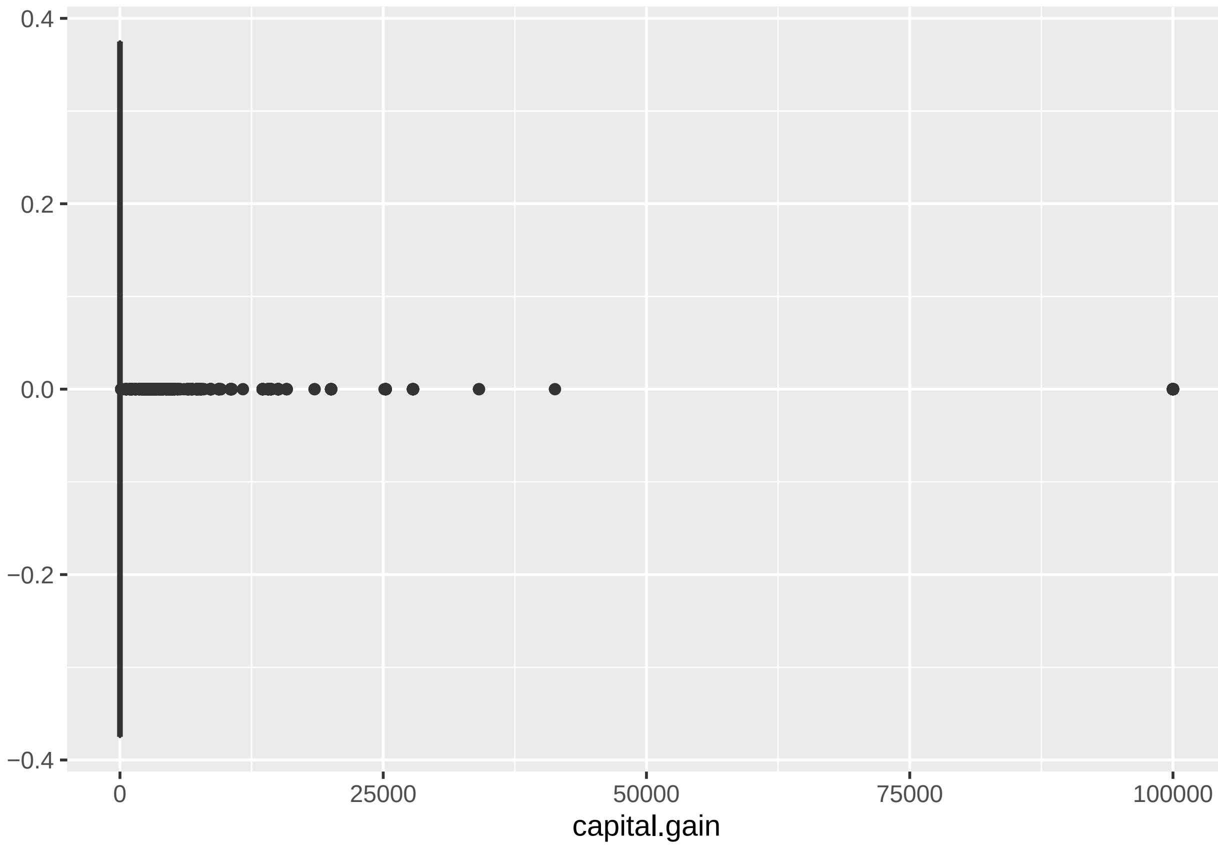
Now let us visualize the distribution of continuous variable "capital.gain" and distributions with target "income" together.



Plot “capital.gain” VS “income” by frequency



Boxplot of capital.gain

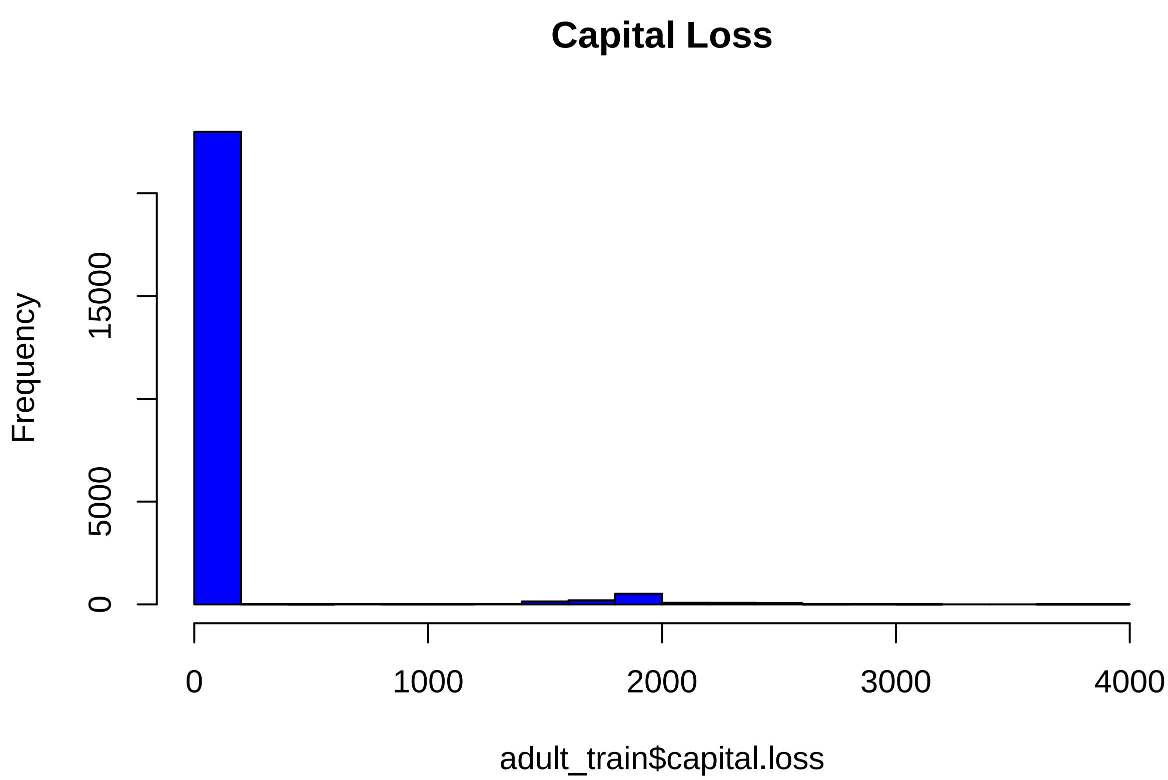


The mean value of capital.gain is 1093. The minimum, 1st quarter,median and 3rd quarter values are all 0 which means a person either has no capital gain or have capital gains with a large amount. The majority people don't have capital gain. The distribution of capital.gain is right skewed. Also from above boxplot we can see the max value 99999 of capital gain would be a potential outlier.

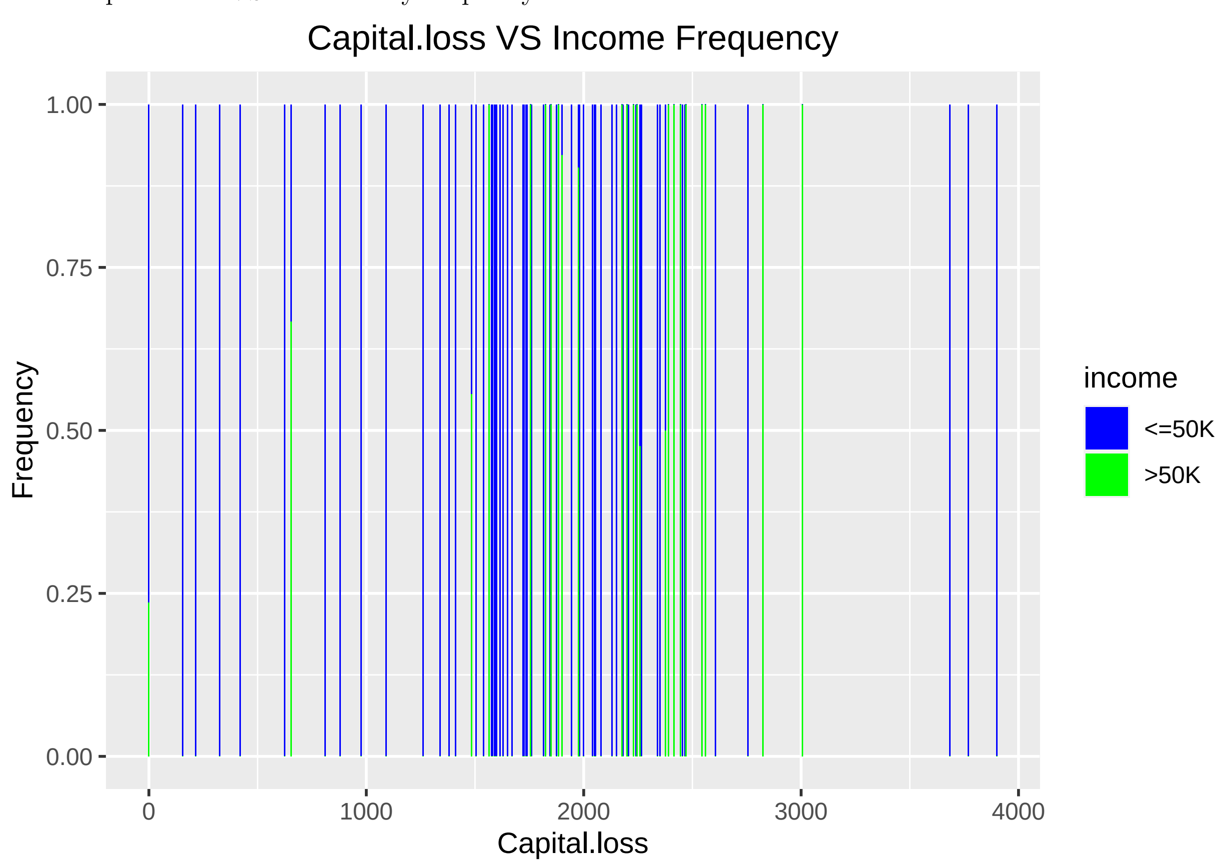
```
summary(adult_train$capital.loss)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.0     0.0    88.2     0.0  3900.0
```

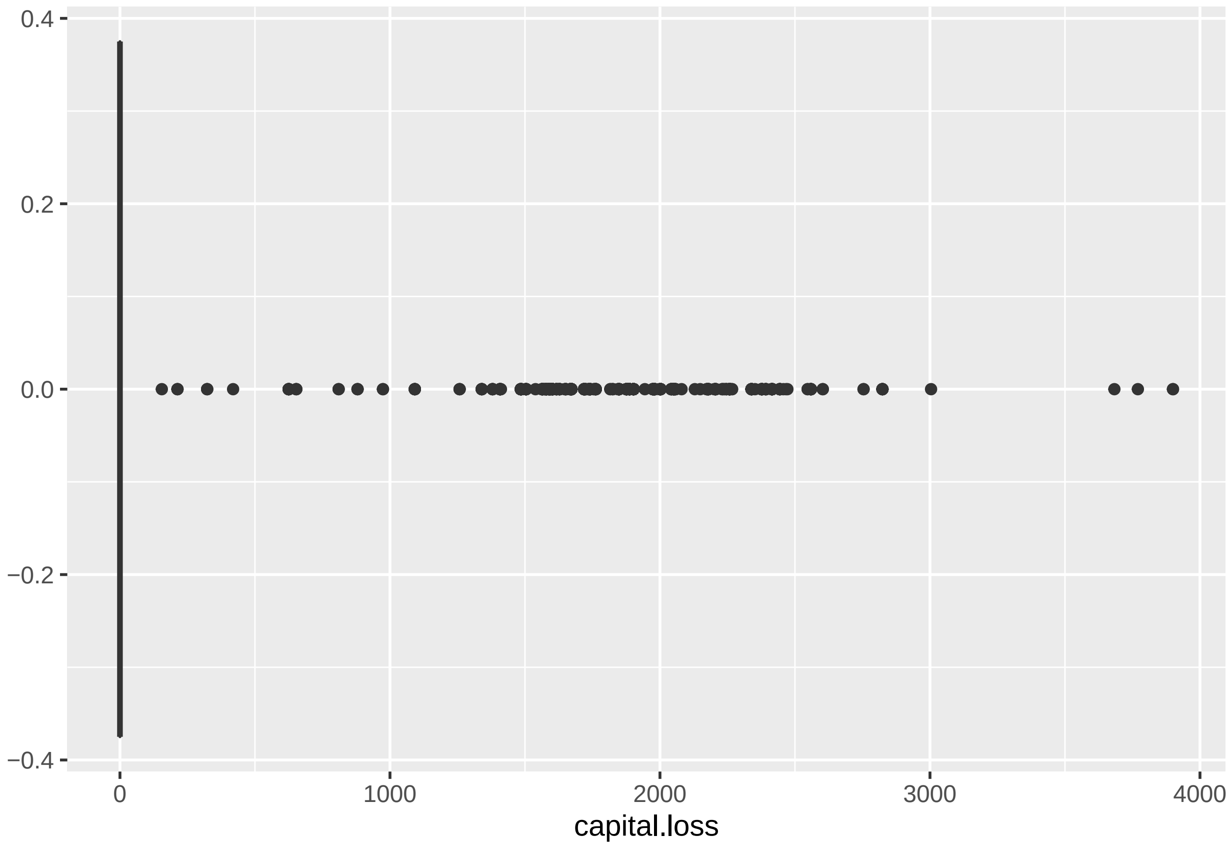
Now let us visualize the distribution of continuous variable "capital.loss" and distributions with target "income" together.



Plot “capital.loss” VS “income” by frequency



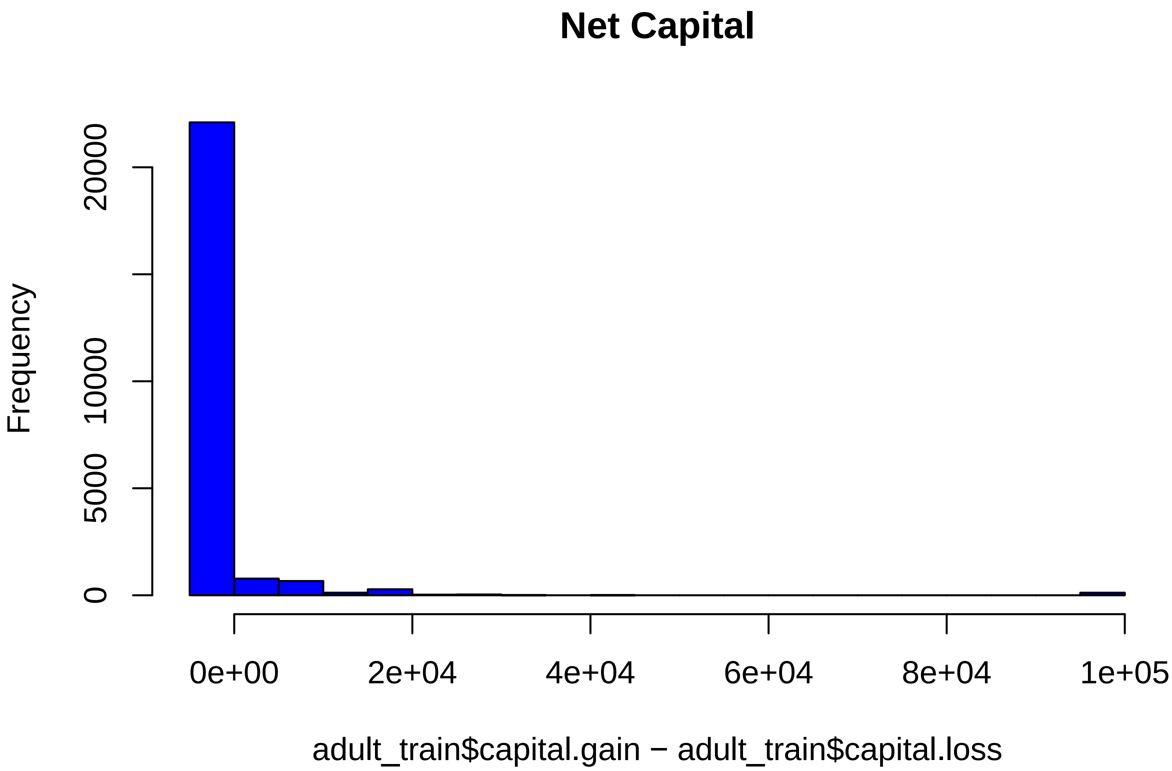
Boxplot of capital.loss



The mean value of capital.loss is 88. The minimum, 1st quarter,median and 3rd quarter values are all 0 which means a person either has no capital loss or have capital loss with a large amount.

Now let us check net capital.

```
#Net Capital
hist(adult_train$capital.gain-adult_train$capital.loss, col="blue", main="Net Capital")
```

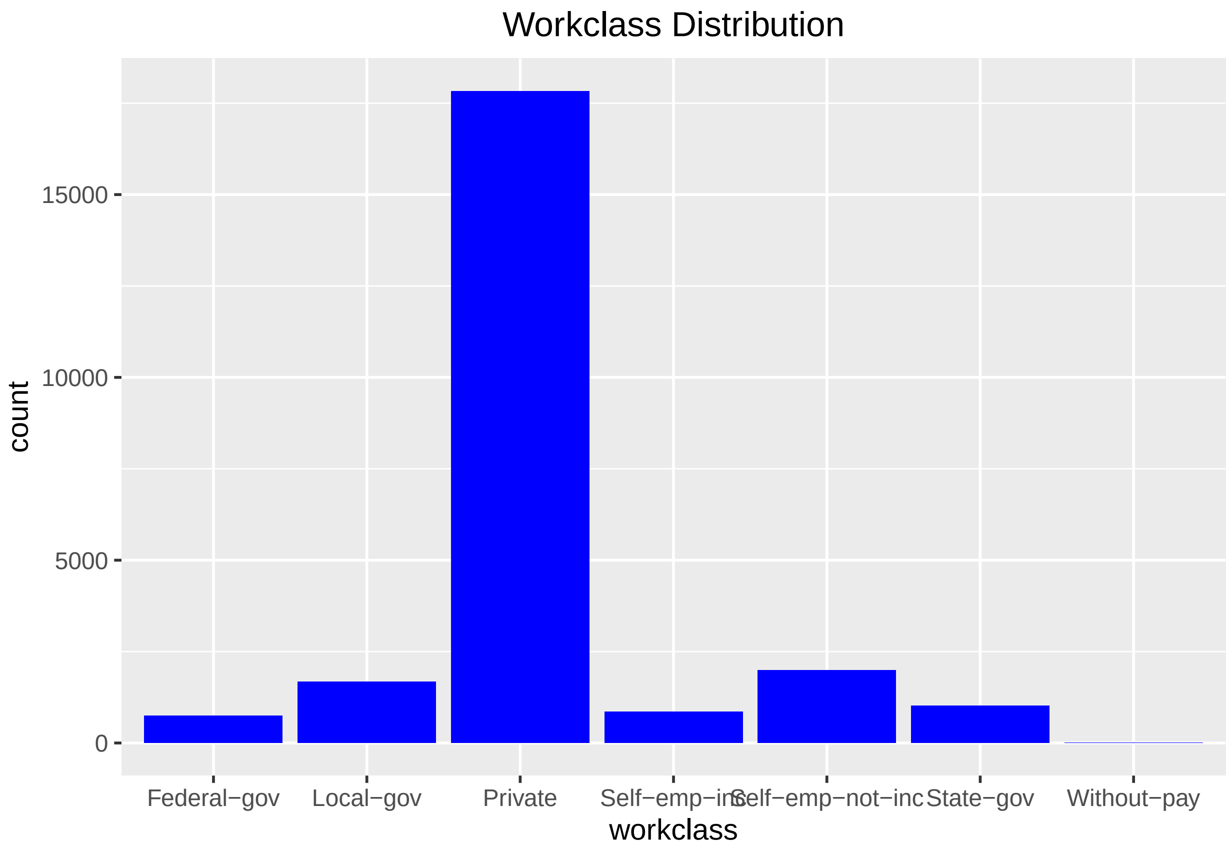


The majority net capital is below 0 which means most of people have net capital loss.

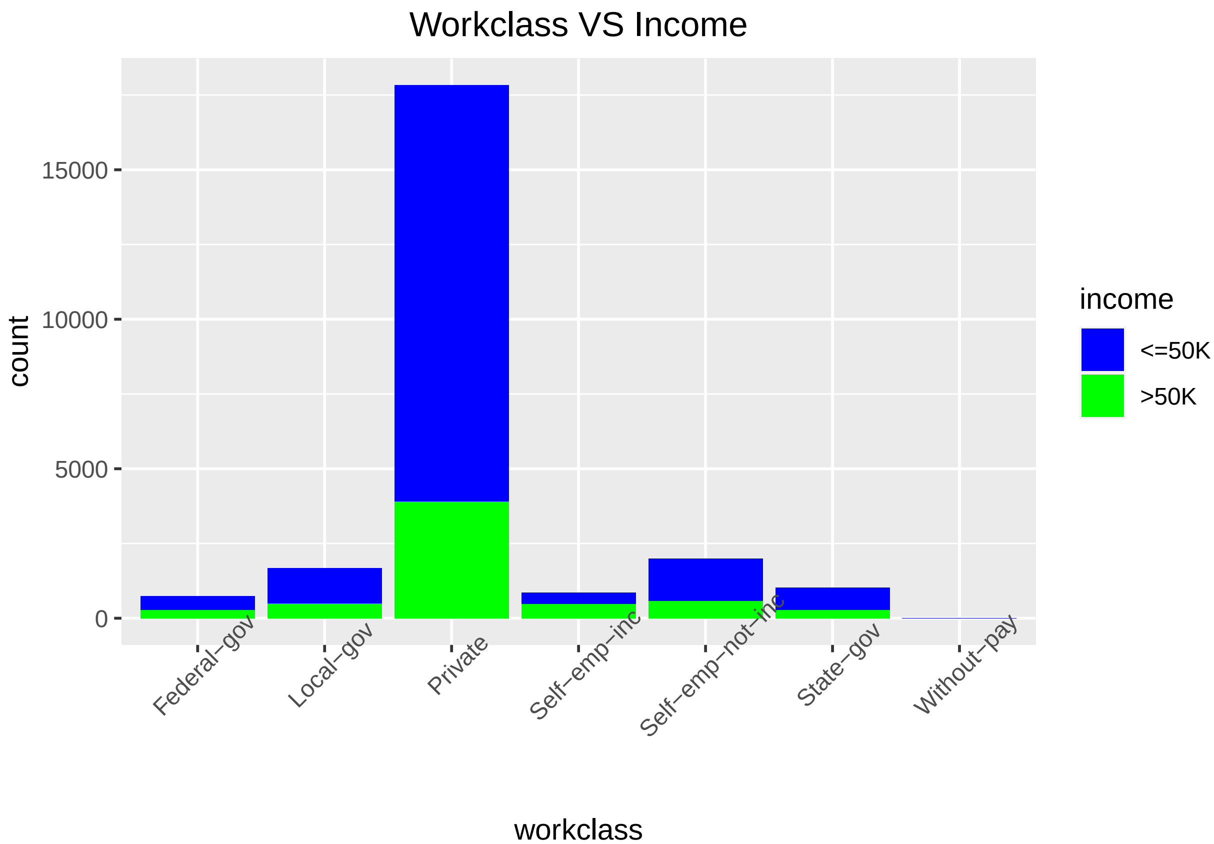
Besides continuous variables, we also need further check which categorical variables may impact final prediction of the target income and how it may impact.

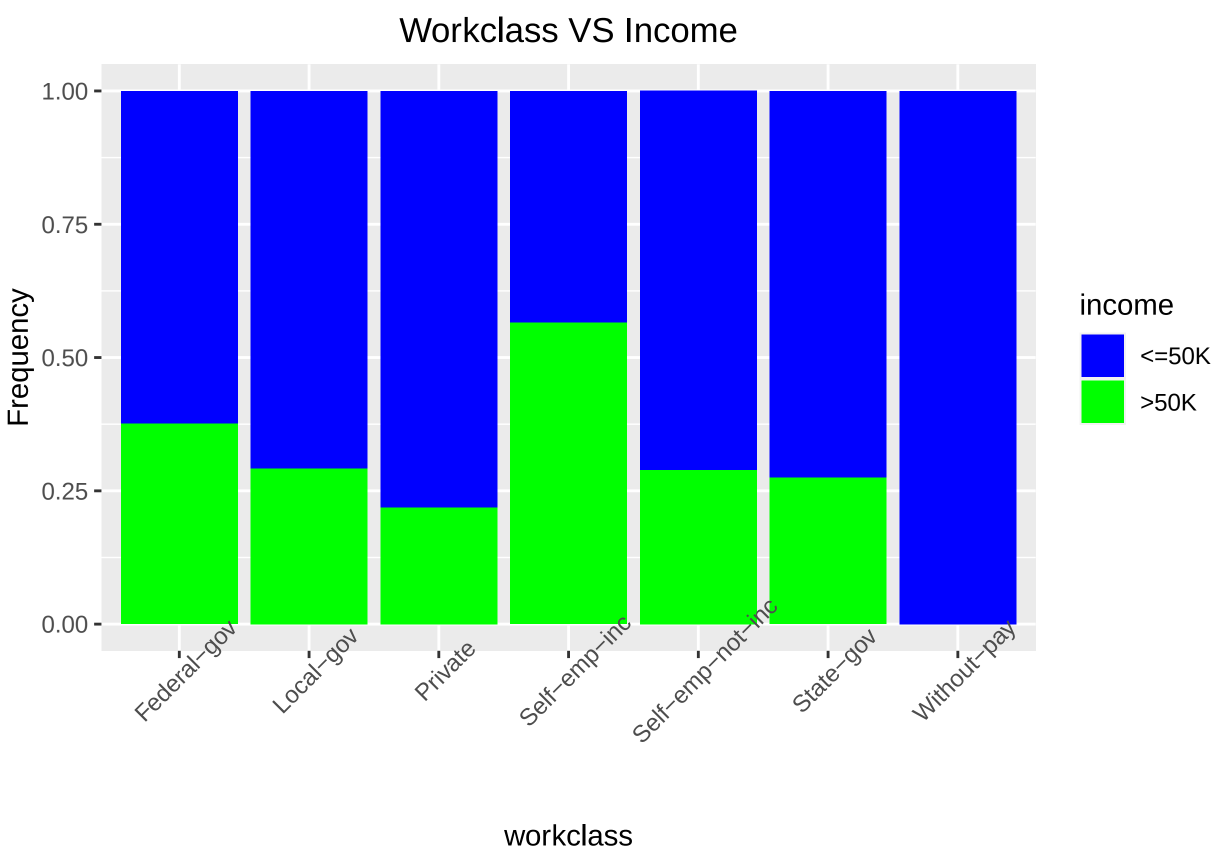
Now we will move to analysis of categorical variables.

Plot "workclass" VS "income" from both amount and frequency points of view



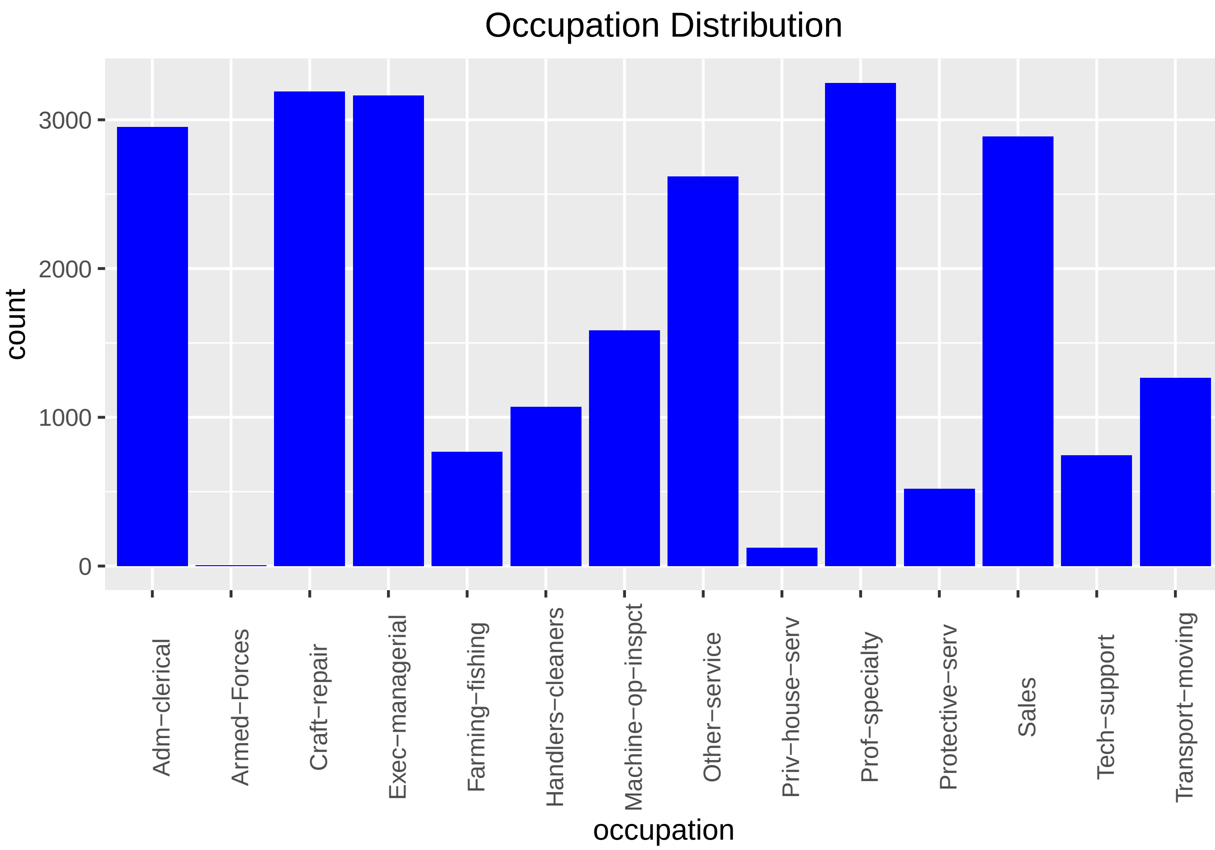

```
##
##                      <=50K  >50K
##    Federal-gov         468   282
##    Local-gov          1183   488
##    Private           13934  3899
##    Self-emp-inc        370   481
##    Self-emp-not-inc   1416   576
##    State-gov           740   280
##    Without-pay          12     0
```





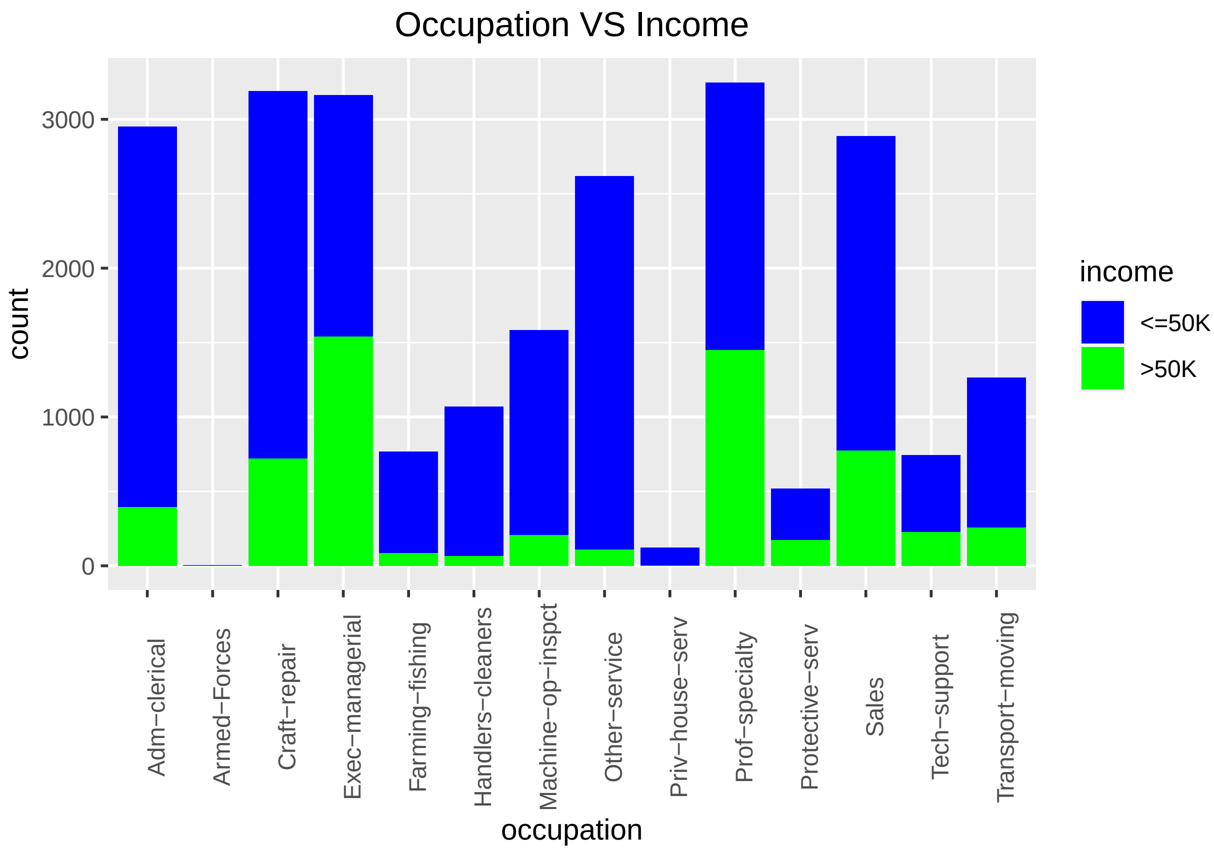
We can see people who work in the private sector has the largest number of population (3899) earning more than 50K per year. While in terms of the proportion, the people work in self-emp-inc have the biggest proportion over 56%.

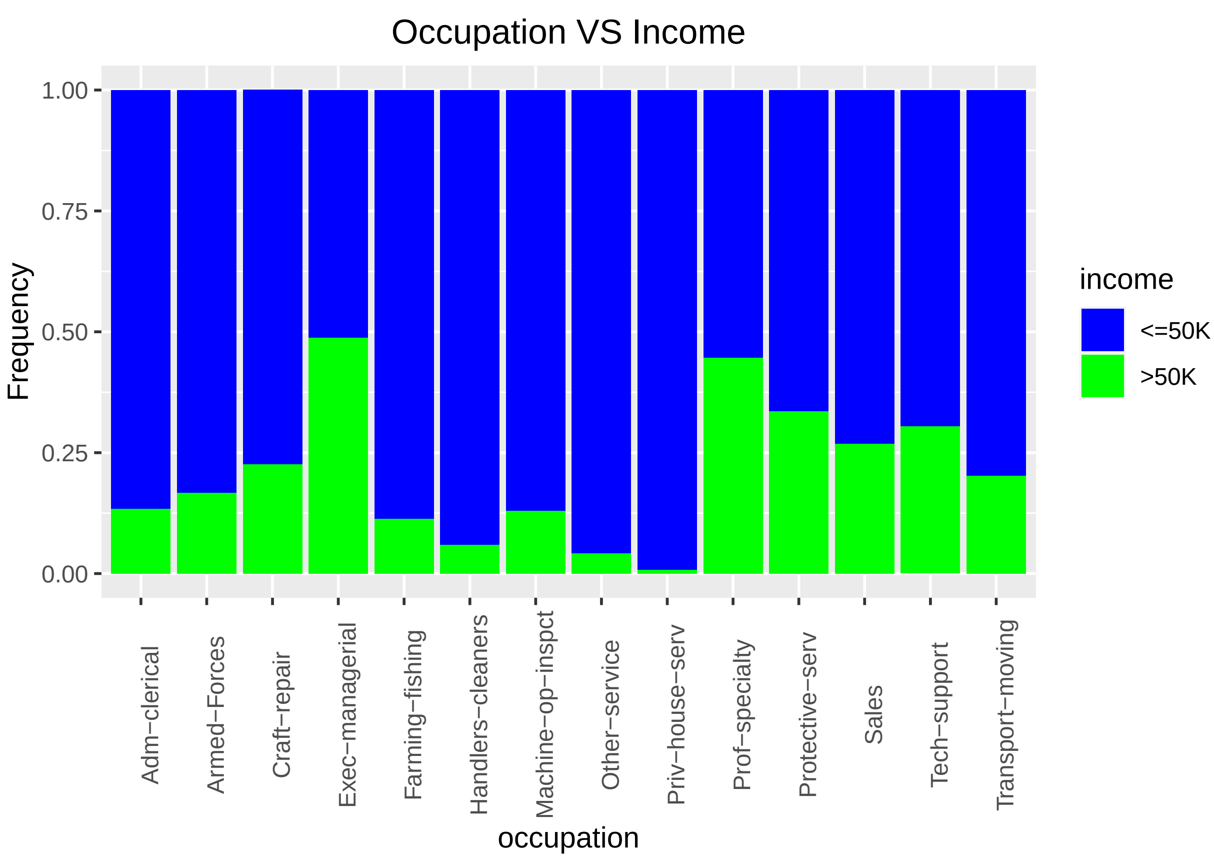
Plot "occupation" VS "income" from both amount and frequency points of view



we can see that the majority people are in private workclass.

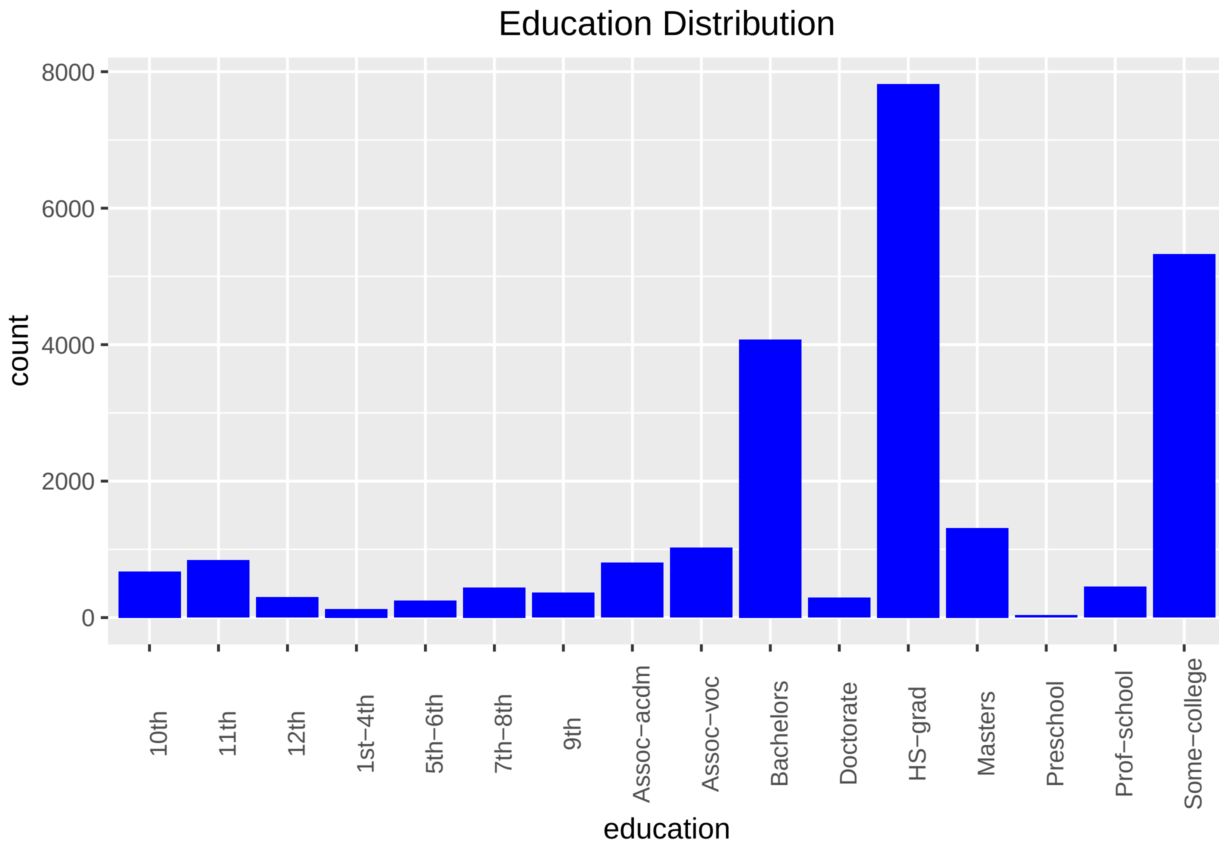
```
##
##                     <=50K >50K
##   Adm-clerical       2555  395
##   Armed-Forces          5    1
##   Craft-repair       2469  721
##   Exec-managerial    1620 1542
##   Farming-fishing     679   87
##   Handlers-cleaners 1005   64
##   Machine-op-inspct 1379  205
##   Other-service      2507  110
##   Priv-house-serv     122    1
##   Prof-specialty     1797 1450
##   Protective-serv     345  174
##   Sales              2113  774
##   Tech-support        517  226
##   Transport-moving   1010  256
```


Occupation VS Income
count
3000
2000
1000
0
income
<=50K
>50K
Adm−clerical
Armed−Forces
Craft−repair
Exec−managerial
Farming−fishing
Handlers−cleaners
Machine−op−inspct
Other−service
Priv−house−serv
Prof−specialty
Protective−serv
Sales
Tech−support
Transport−moving
occupation

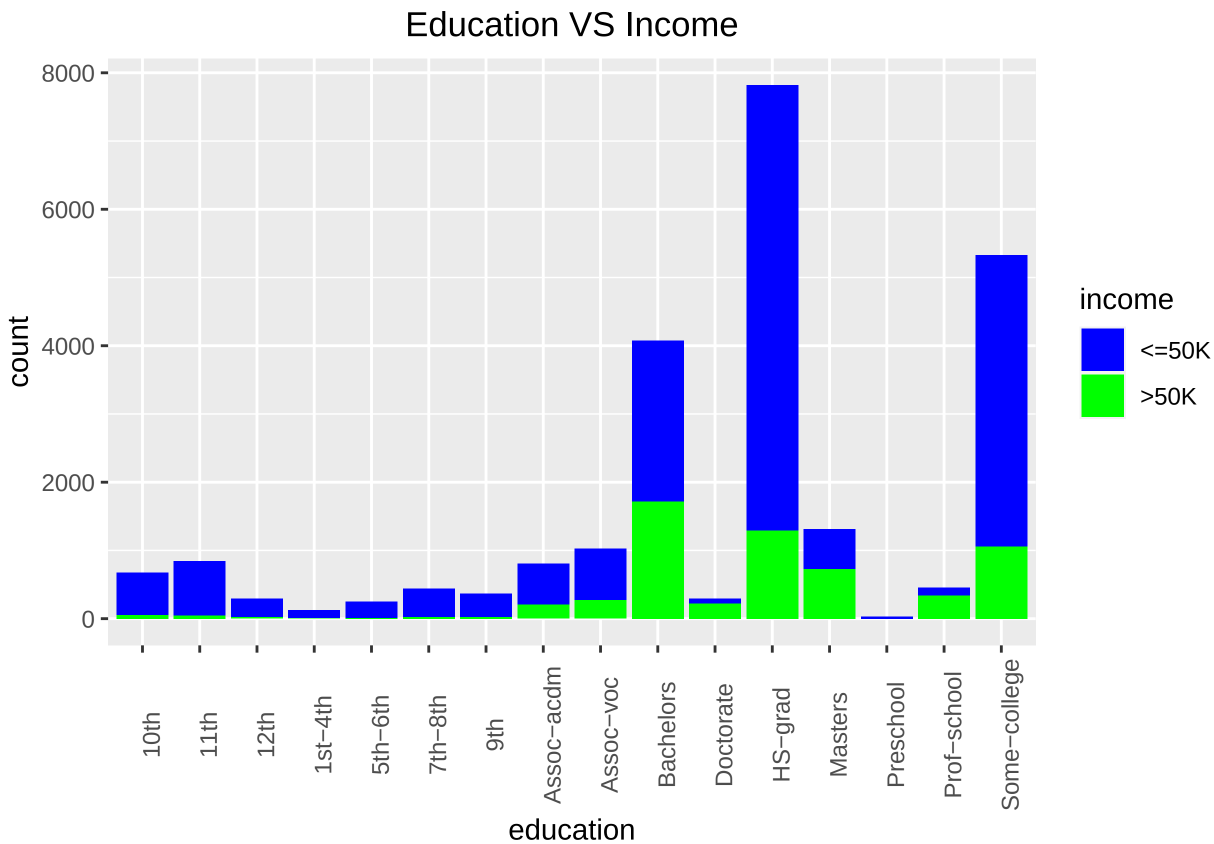

Occupation VS Income
Frequency
1.00
0.75
0.50
0.25
0.00
income
<=50K
>50K
Adm−clerical
Armed−Forces
Craft−repair
Exec−managerial
Farming−fishing
Handlers−cleaners
Machine−op−inspct
Other−service
Priv−house−serv
Prof−specialty
Protective−serv
Sales
Tech−support
Transport−moving
occupation

we can see that the majority people are in private workclass and people with occupation of executive management and professional specialty are more likely to have income over 50k from both amount and proportion point of view.

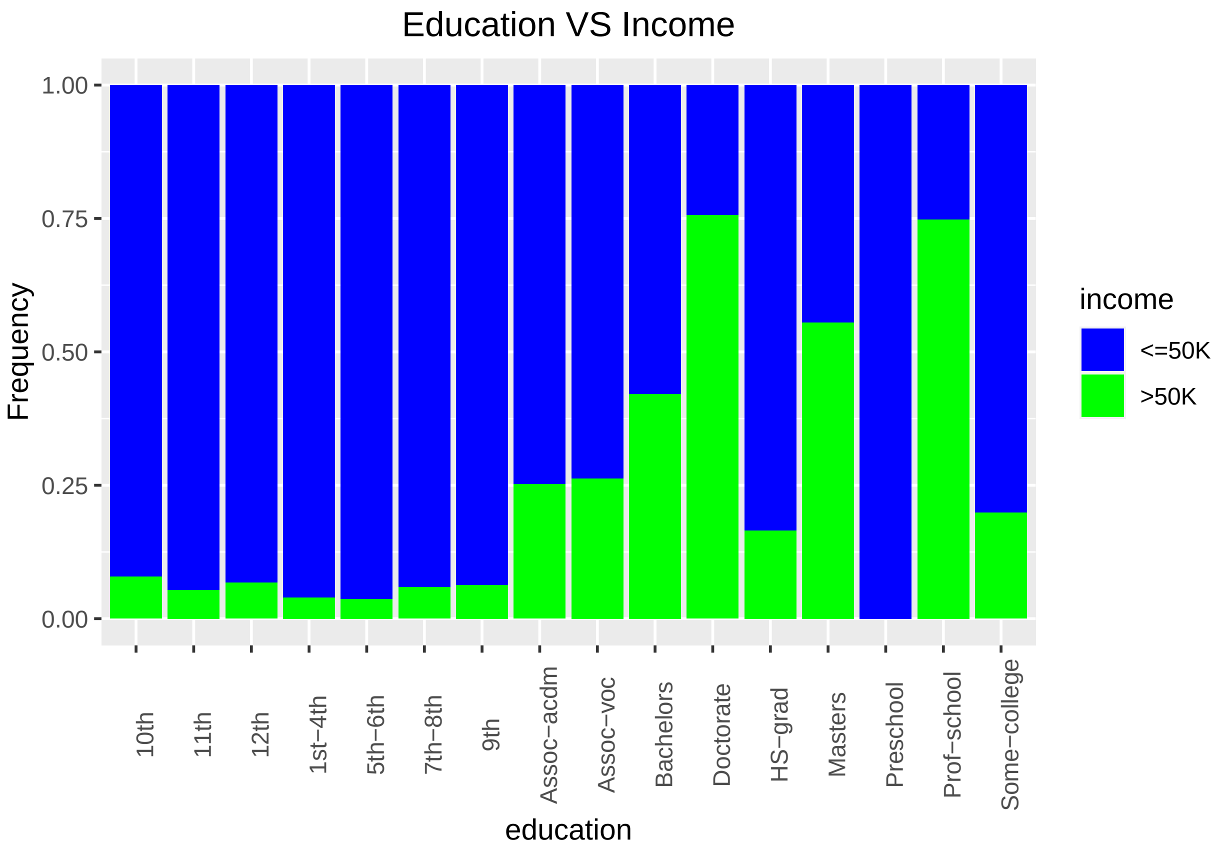
Plot "education" VS "income" from both amount and frequency points of view




```
##
##                <=50K >50K
##   10th           621   53
##   11th           795   45
##   12th           277   20
##   1st-4th        121    5
##   5th-6th        236    9
##   7th-8th        416   26
##   9th            341   23
##   Assoc-acdm     603  203
##   Assoc-voc      756  269
##   Bachelors     2359 1717
##   Doctorate       71  220
##   HS-grad       6527 1293
##   Masters        584  728
##   Preschool       33    0
##   Prof-school    114  338
##   Some-college  4269 1057
```

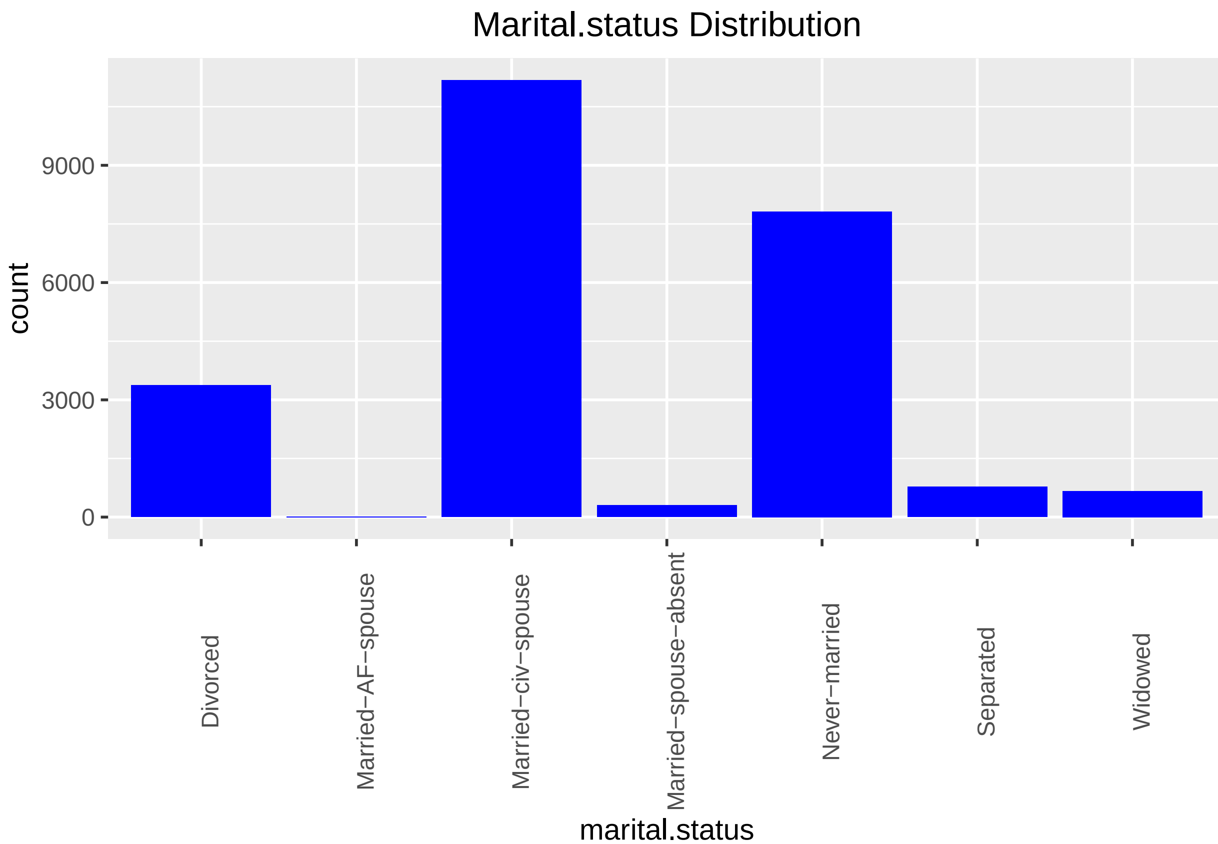


Plot Education VS Income from frequency points of view



we can see that the most of people are with education level of HS-grad followed by level of Some-college and Bachelors. People with Bachelors degree are more likely to have income over 50k from both amount and proportion point of view. People with doctorate education background have the best chance to earn over 50k. The charts above meet common sense that people with higher education background are more likely to have better income level even if we observed that very few people with education background lower than 12th also had income over 50k.

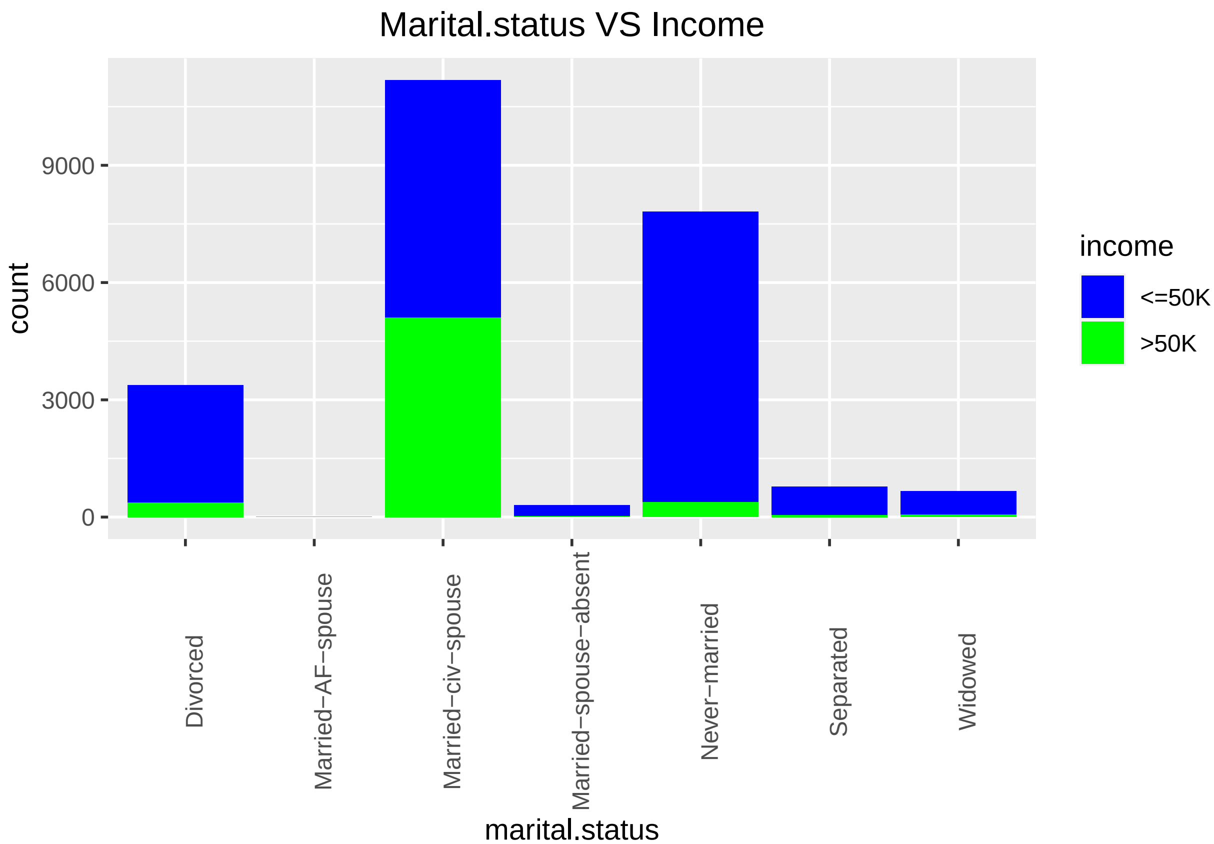
Plot “marital.status” VS “income” from both amount and frequency points of view

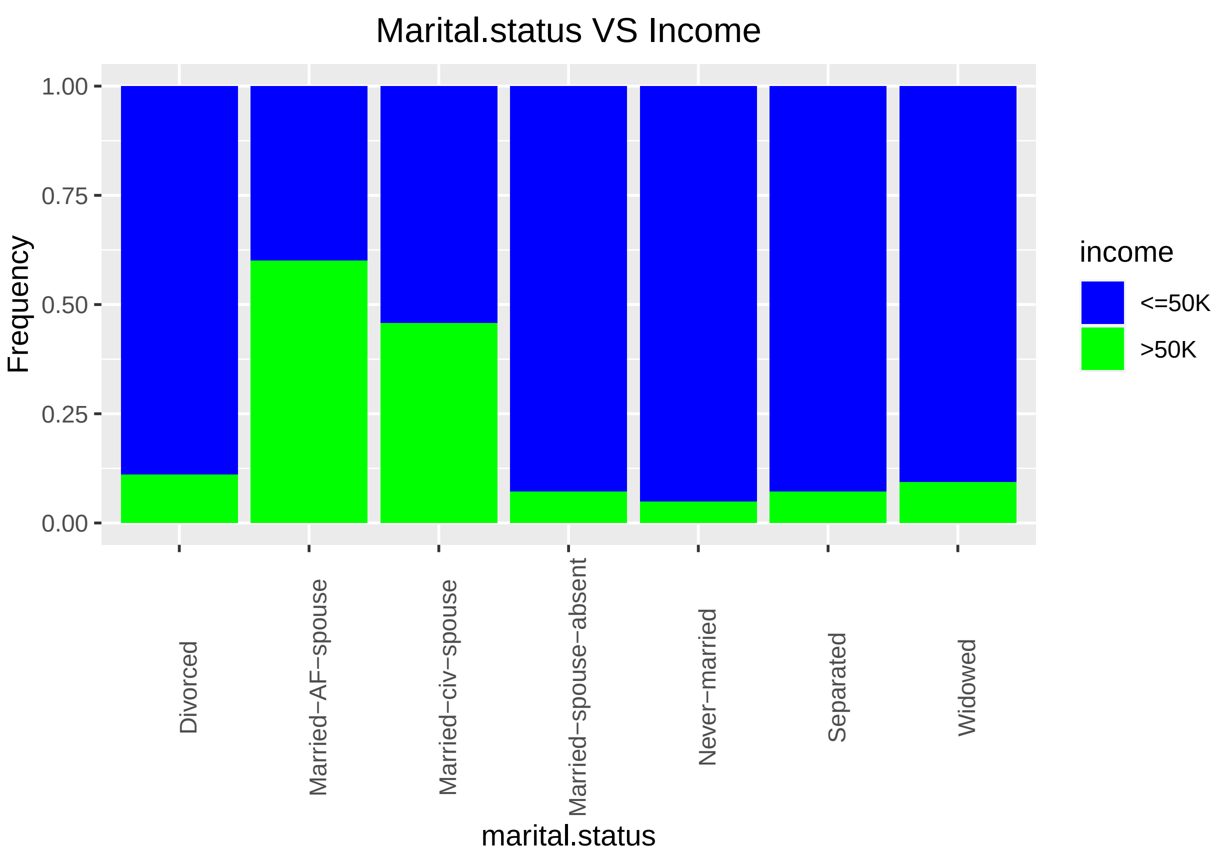


We can see that the most of people are with marital status of Married-civ-spouse followed by status of Never-married and Divorced

Now plot marital.status VS Income from amount and frequency points of view

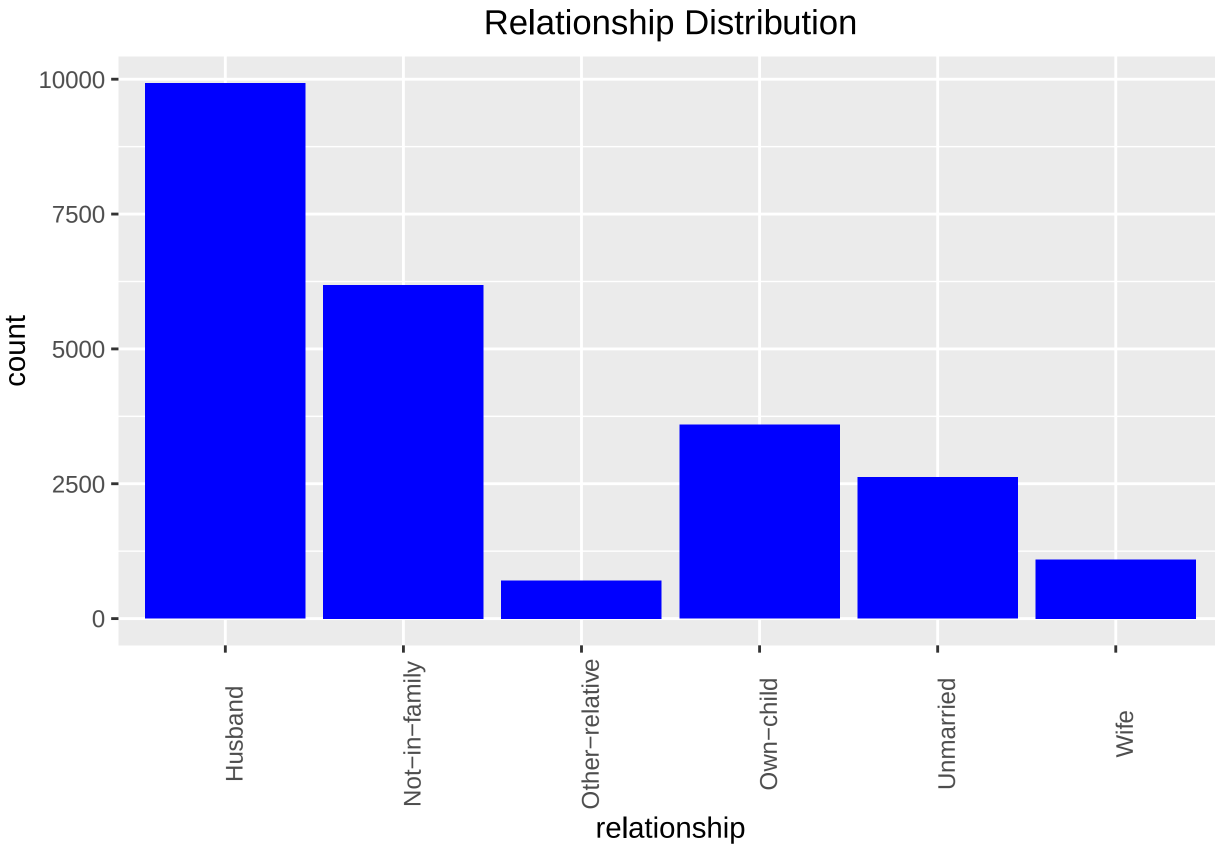
```
##
##                          <=50K >50K
##   Divorced                3002  373
##   Married-AF-spouse          6    9
##   Married-civ-spouse      6070 5106
##   Married-spouse-absent    284   22
##   Never-married           7438  379
##   Separated                720   55
##   Widowed                  603   62
```





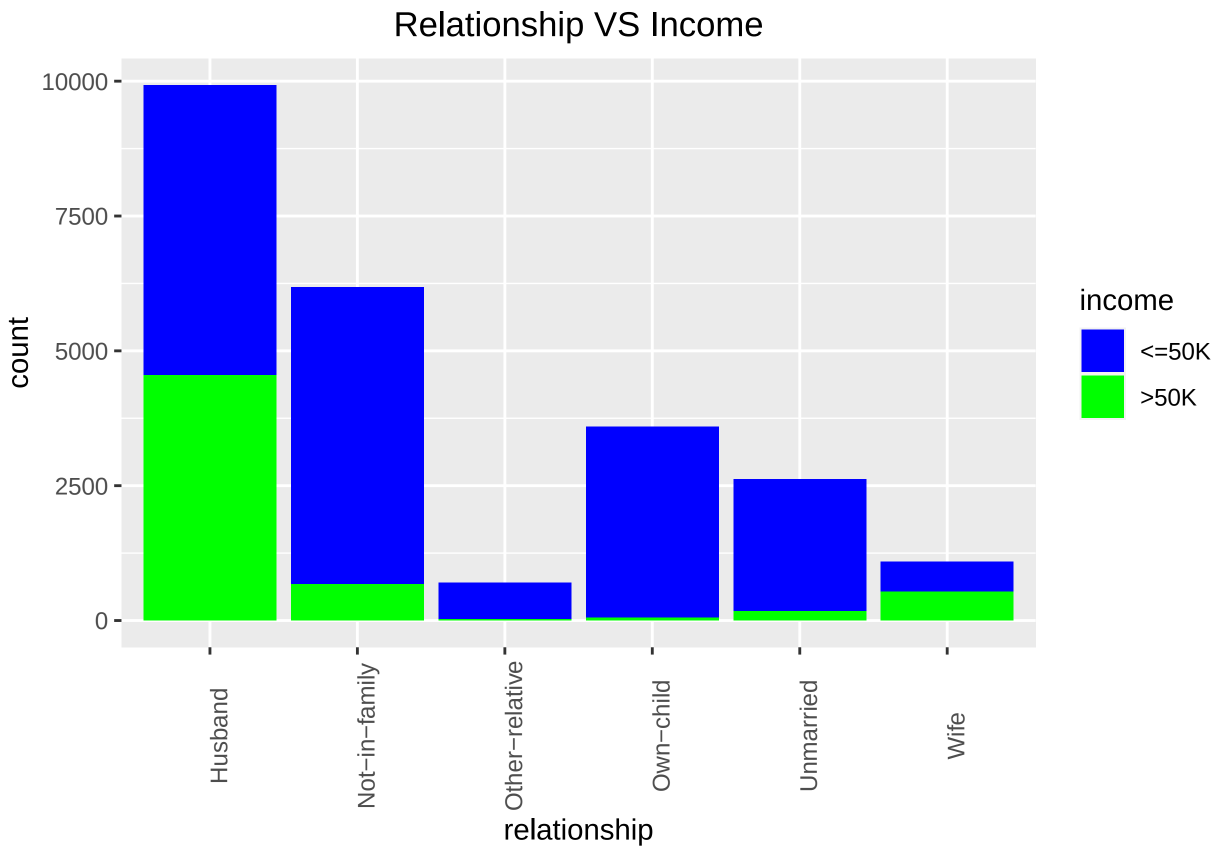
We can see that the most of people are with marital status of Married-civ-spouse followed by status of Never-married and Divorced . Meanwhile people with married status have much higher proportion of earning over 50k income per year.

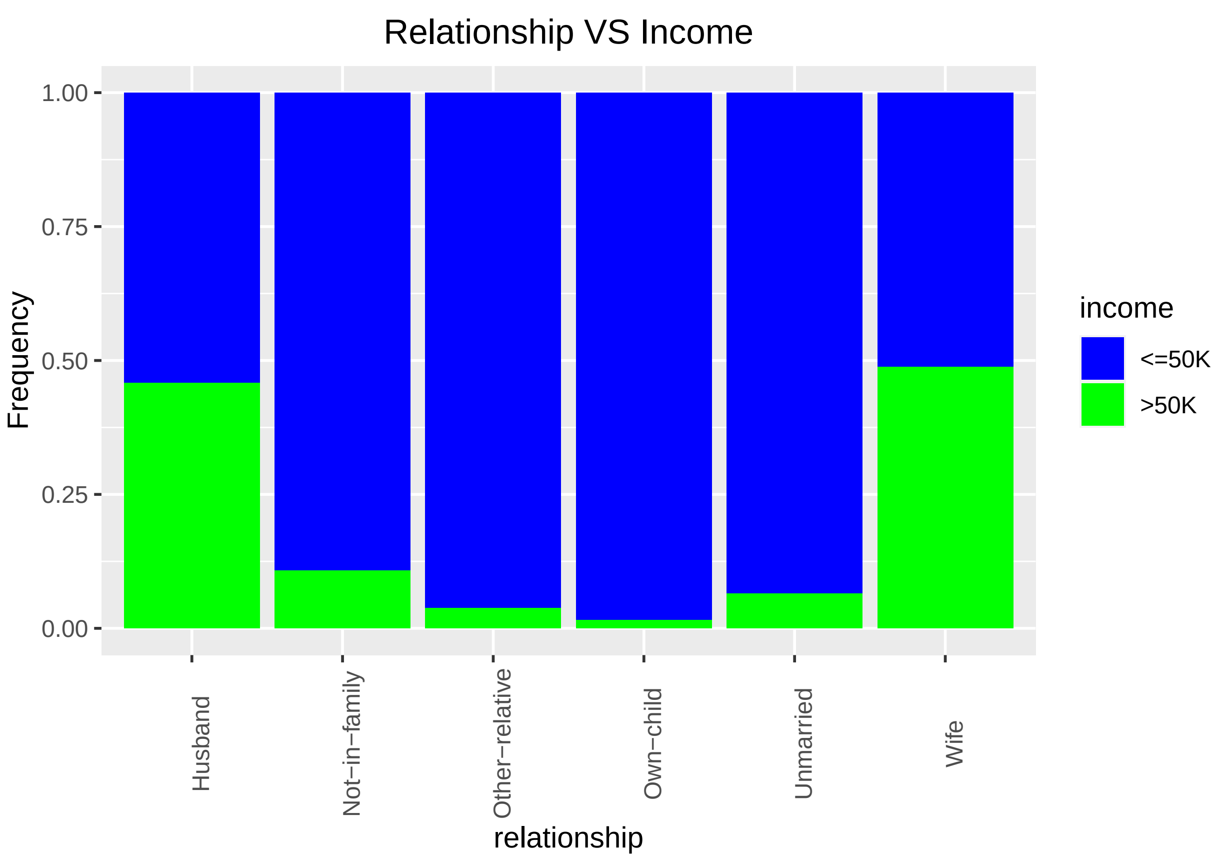
Plot "relationship" VS "income" from amount and frequency points of view.



We can see that the majority people in the sample data set are husband, followed by Not-in-family, Own-child and unmarried.

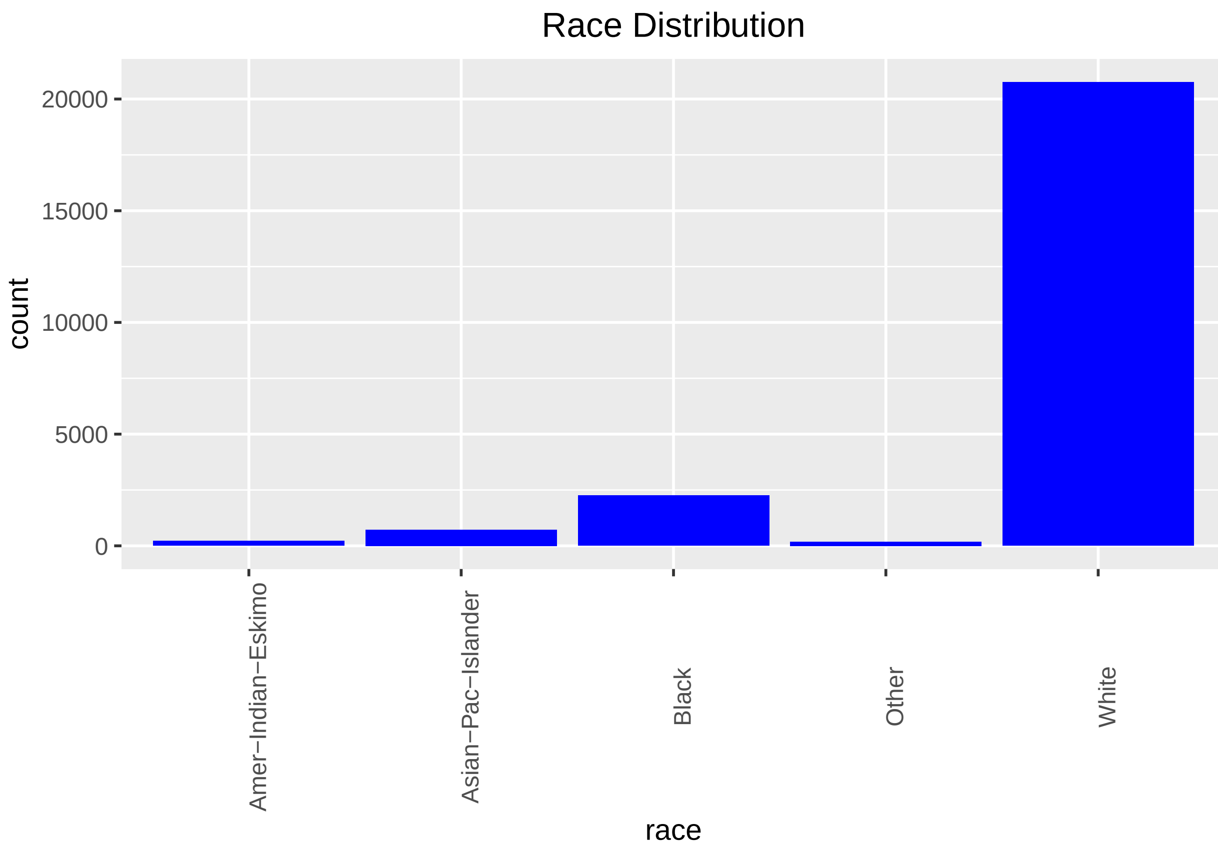
```
##
##                    <=50K >50K
##   Husband          5377 4549
##   Not-in-family    5515  669
##   Other-relative    679   27
##   Own-child        3539   55
##   Unmarried        2451  171
##   Wife              562  535
```


Relationship VS Income
count
10000
7500
5000
2500
0
Husband
Not−in−family
Other−relative
Own−child
Unmarried
Wife
relationship
income
<=50K
>50K

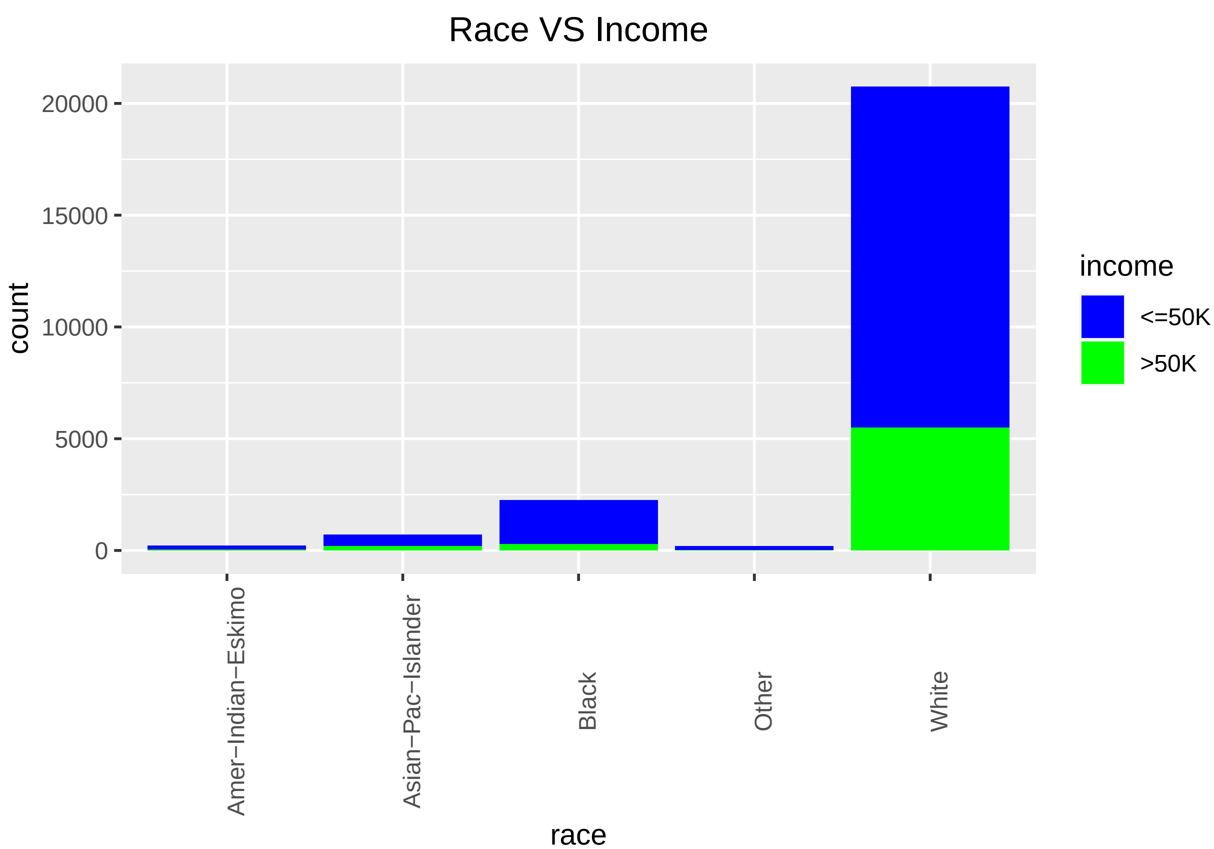

Relationship VS Income
Frequency
1.00
0.75
0.50
0.25
0.00
Husband
Not−in−family
Other−relative
Own−child
Unmarried
Wife
relationship
income
<=50K
>50K

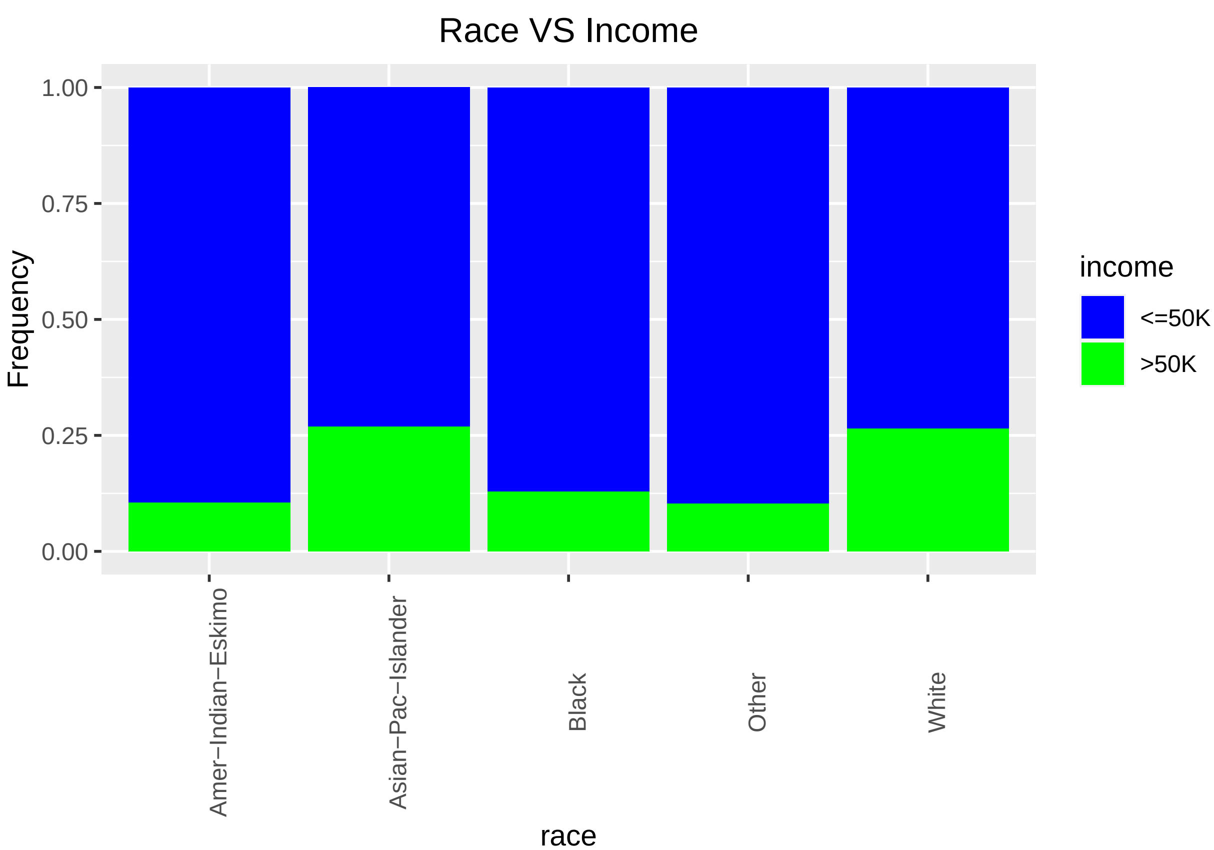
We can see that the majority people in data set "adult_train" are husband, followed by Not-in-family, Own-child and unmarried. Husbands have the largest number of population (4549) earning more than 50K per year. While in terms of the proportion, wives have the biggest proportion close to 50% earning over 50k income. This result is quite reasonable since we already know that married people are more likely to earn over 50k income.

Plot "race" VS "income" from both amount and frequency angles



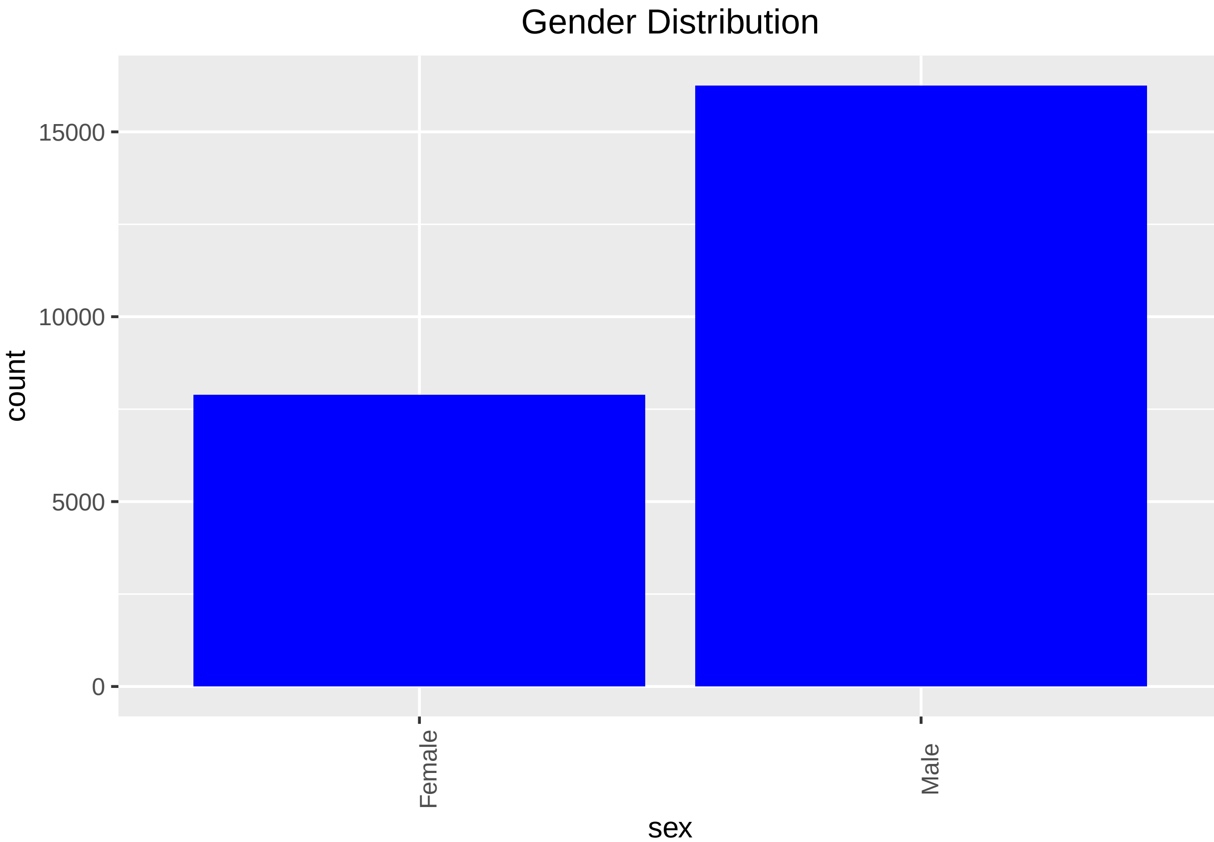

```
##
##                       <=50K  >50K
##   Amer-Indian-Eskimo    195    23
##   Asian-Pac-Islander    524   193
##   Black                1966   290
##   Other                 166    19
##   White               15272  5481
```



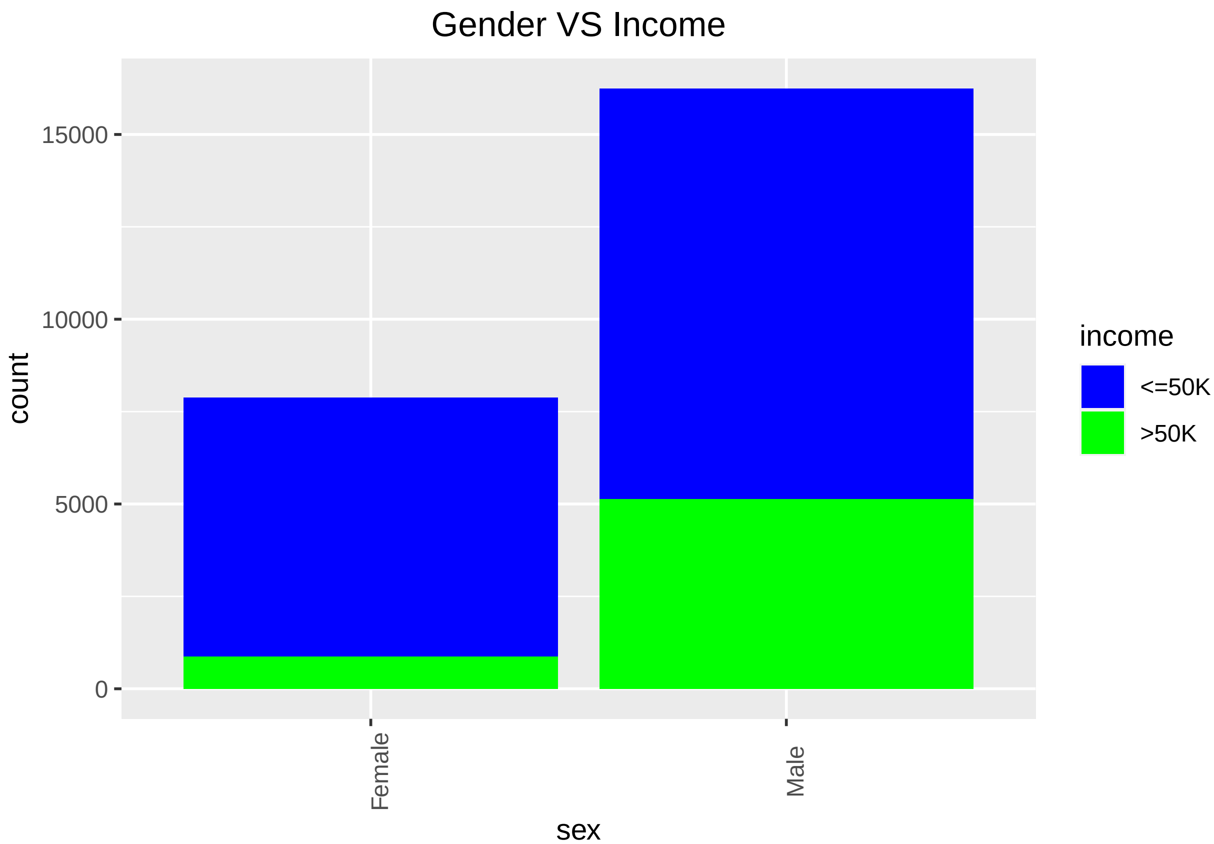


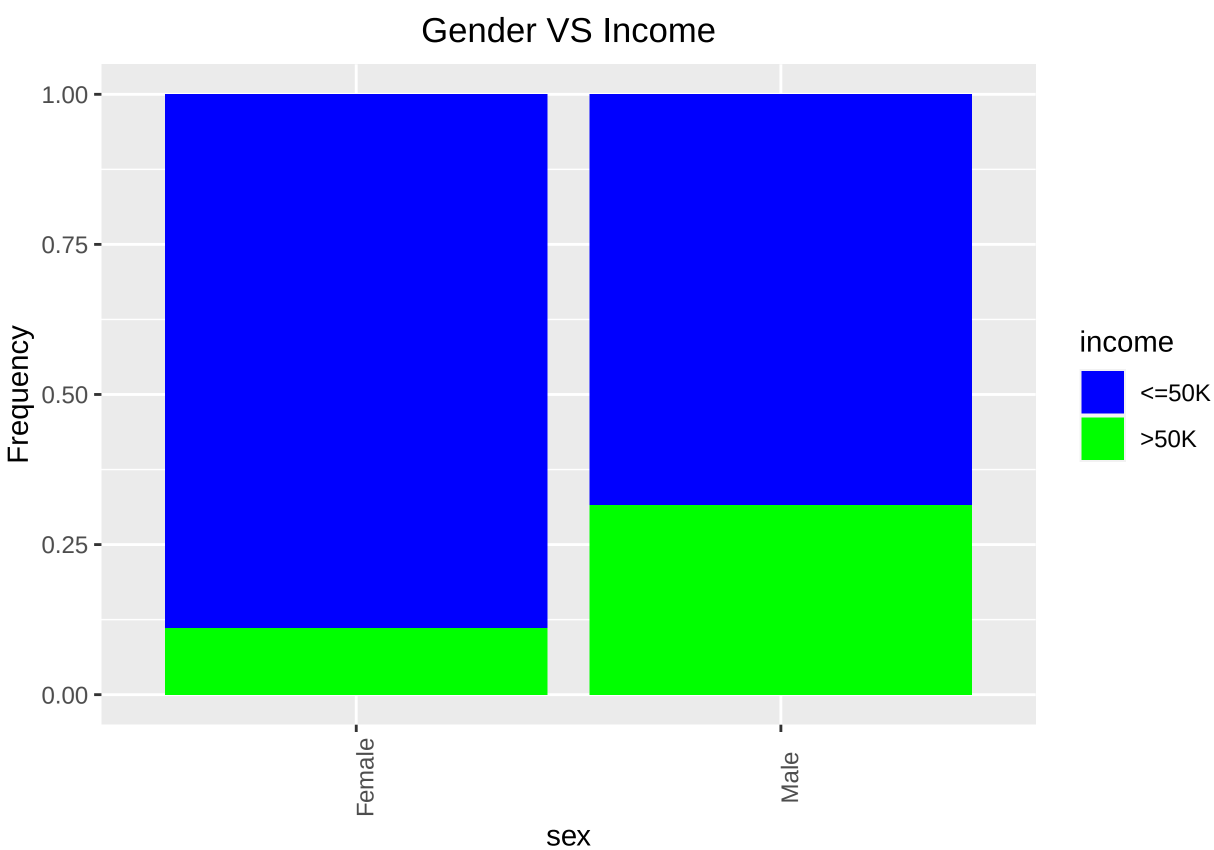
We can see that the majority people are White people followed by Black people and Asian-Pac-Islander. It is not surprising that white people have the most amount of earning over 50k income. While in terms of the proportion of earning over 50k income, Asian-Pac-Islander people have slightly bigger proportion of 27% than white people of 26%. This indicates that Asian-Pac-Islander and white people are more likely to have higher income.

Plot "sex" VS "income" from both amount and frequency angles



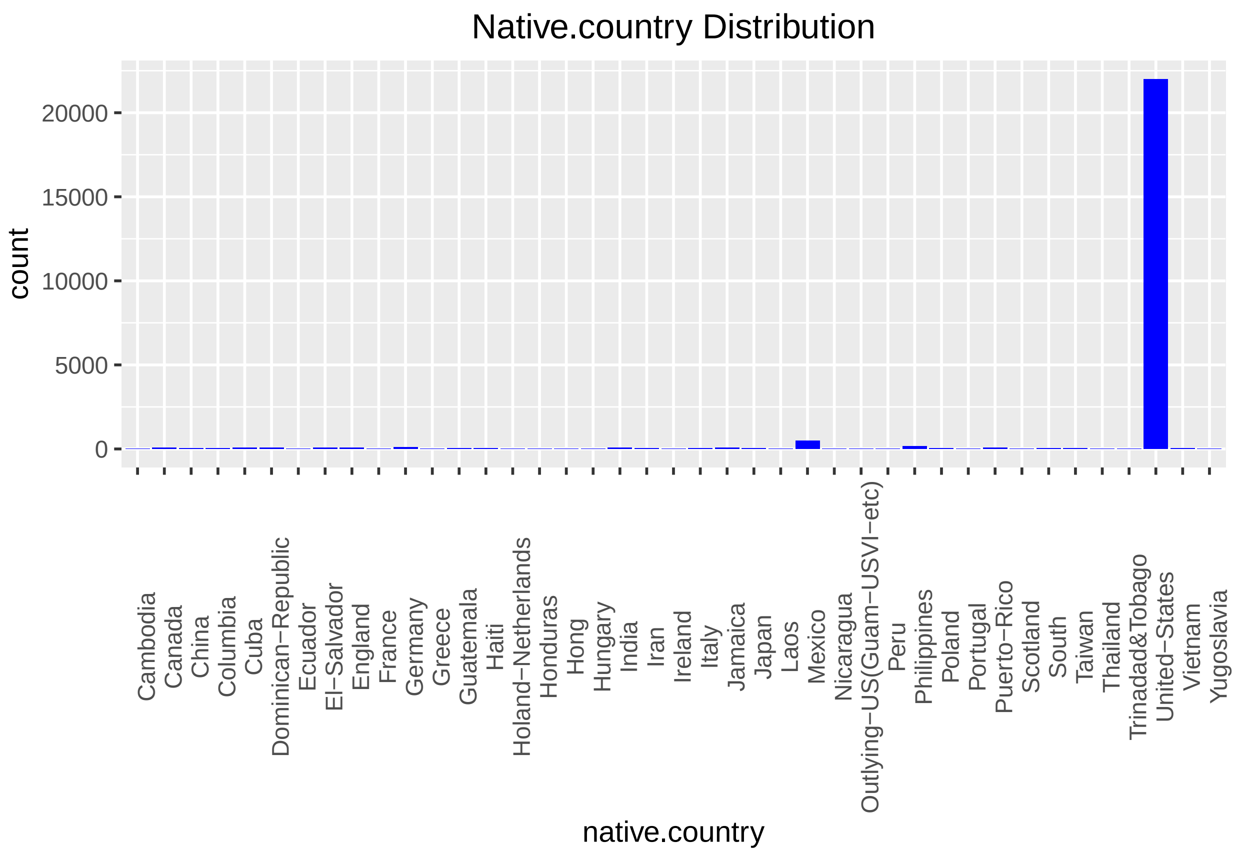

```
##
##          <=50K  >50K
##   Female  7006   877
##   Male   11117  5129
```





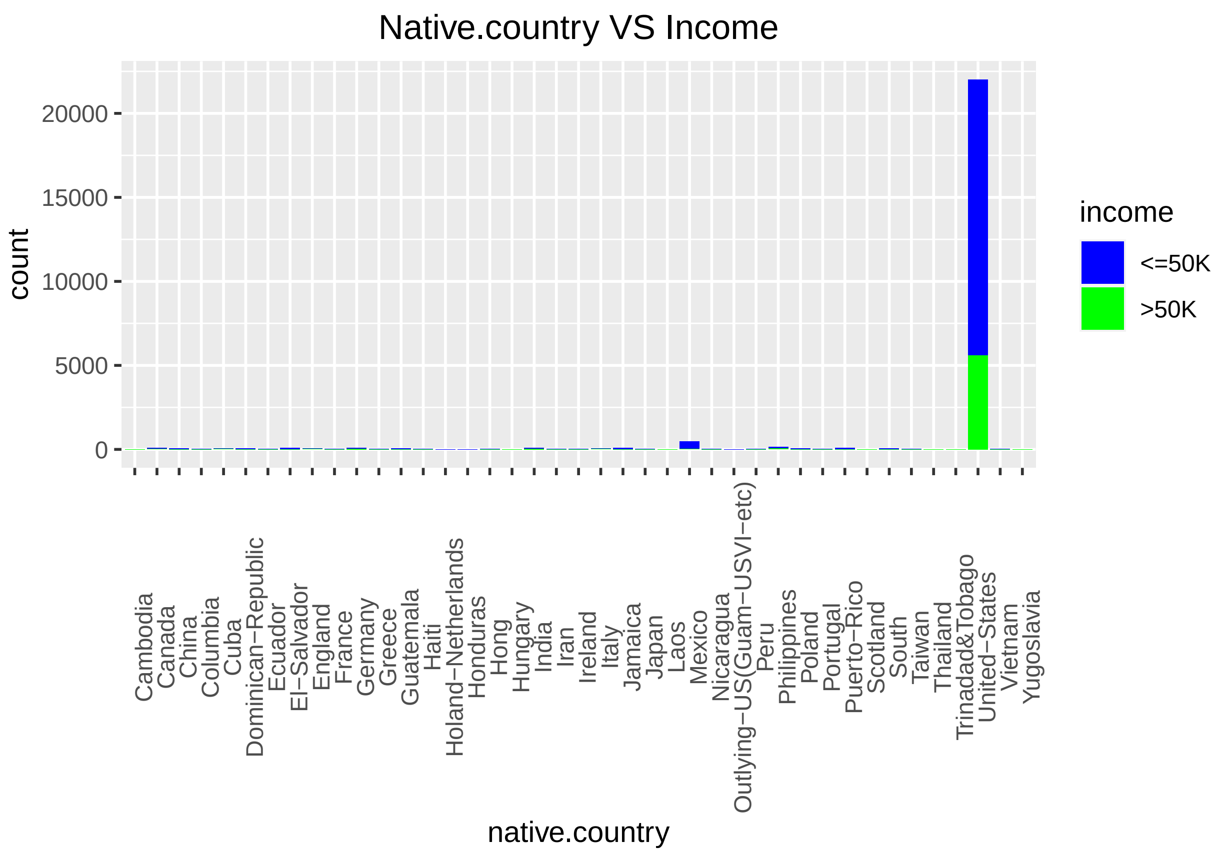
we can see that the majority people here are male. And male are more likely to earn over 50k income with over double proportion compared with female.

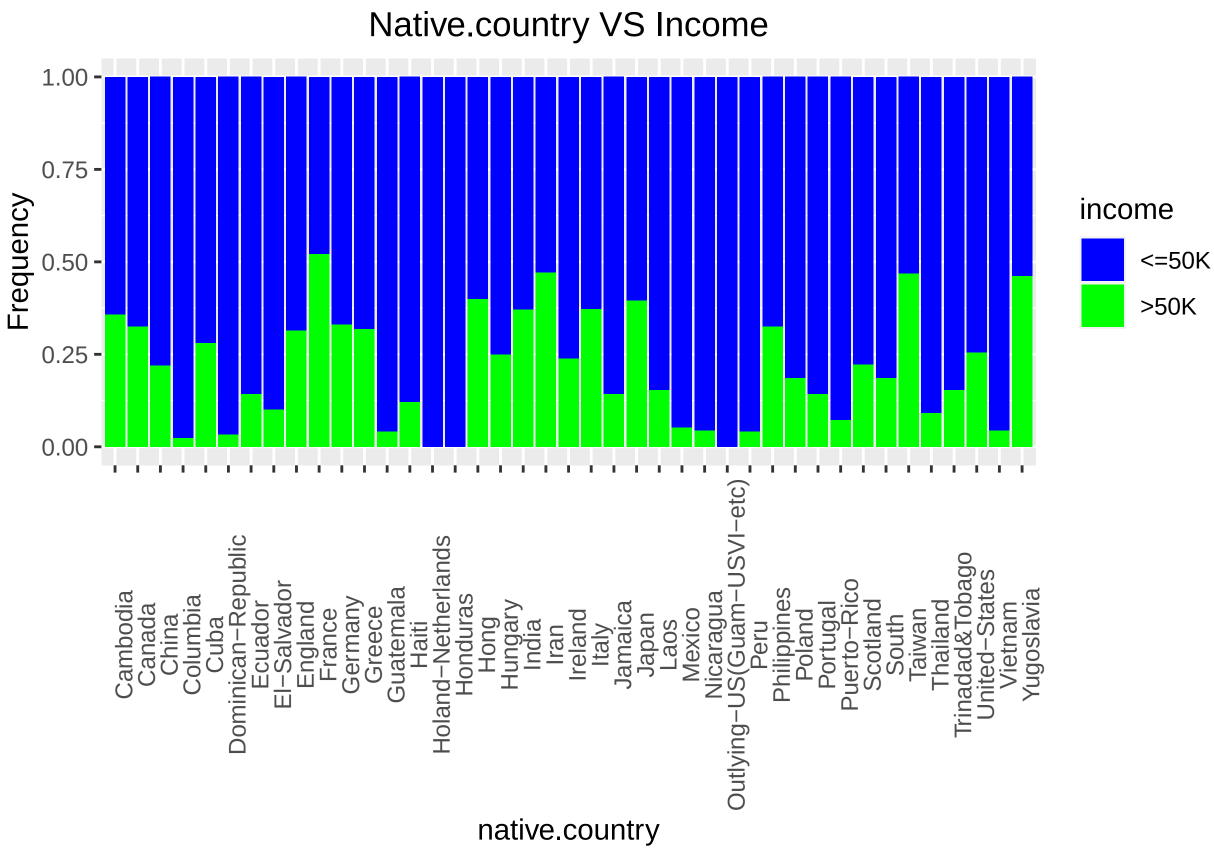
Plot “native.country” VS “income” from both amount and frequency points of view




```
##
##                          <=50K  >50K
##    Cambodia                 9     5
##    Canada                  58    28
##    China                   39    11
##    Columbia                42     1
##    Cuba                    54    21
##    Dominican-Republic      58     2
##    Ecuador                 18     3
##    El-Salvador             71     8
##    England                 48    22
##    France                  11    12
##    Germany                 69    34
##    Greece                  15     7
##    Guatemala               46     2
##    Haiti                   29     4
##    Holand-Netherlands       1     0
##    Honduras                 8     0
##    Hong                     9     6
##    Hungary                  6     2
##    India                   51    30
##    Iran                    18    16
##    Ireland                 16     5
```

```
##    Italy                          37     22
##    Jamaica                        60     10
##    Japan                          29     19
##    Laos                           11      2
##    Mexico                        457     25
##    Nicaragua                      22      1
##    Outlying-US(Guam-USVI-etc)     12      0
##    Peru                           23      1
##    Philippines                   110     53
##    Poland                         35      8
##    Portugal                       24      4
##    Puerto-Rico                    77      6
##    Scotland                        7      2
##    South                          48     11
##    Taiwan                         17     15
##    Thailand                       10      1
##    Trinadad&Tobago                11      2
##    United-States               16406   5597
##    Vietnam                        44      2
##    Yugoslavia                      7      6
```

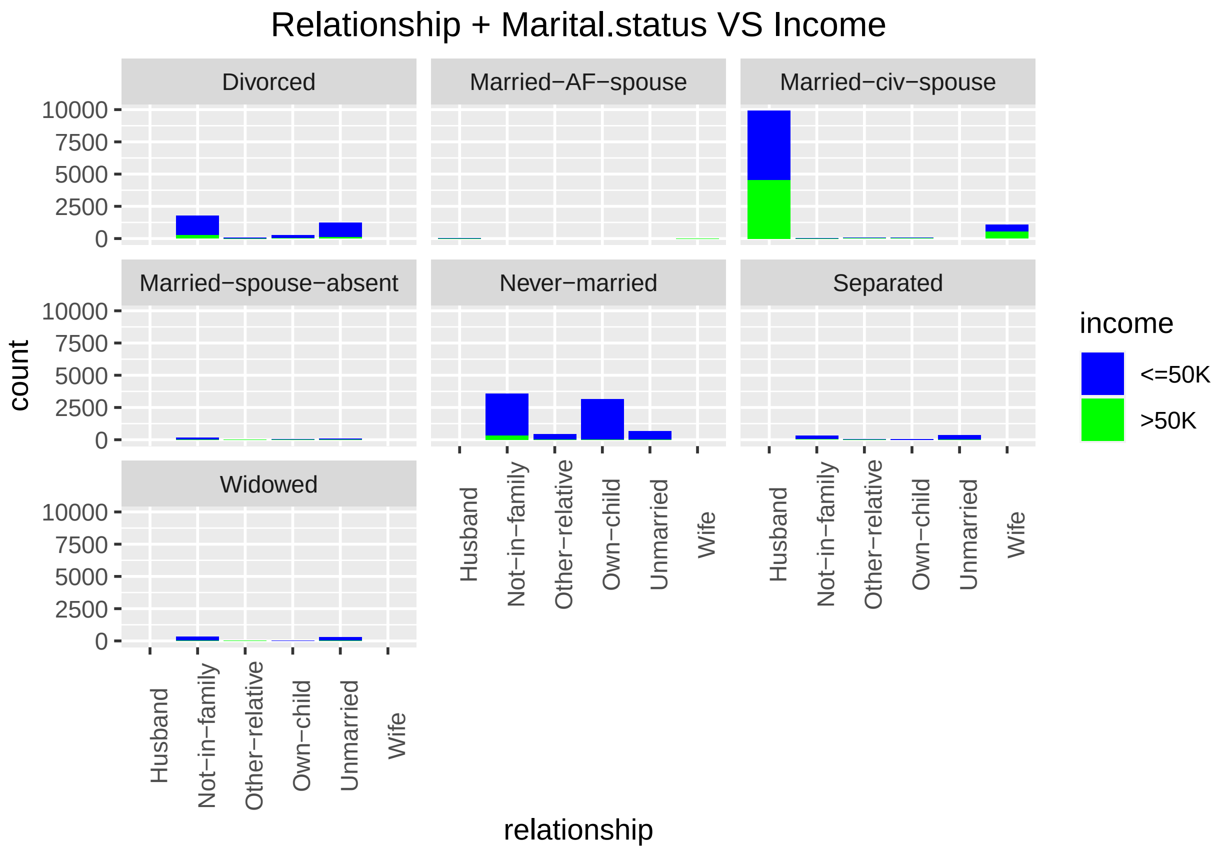


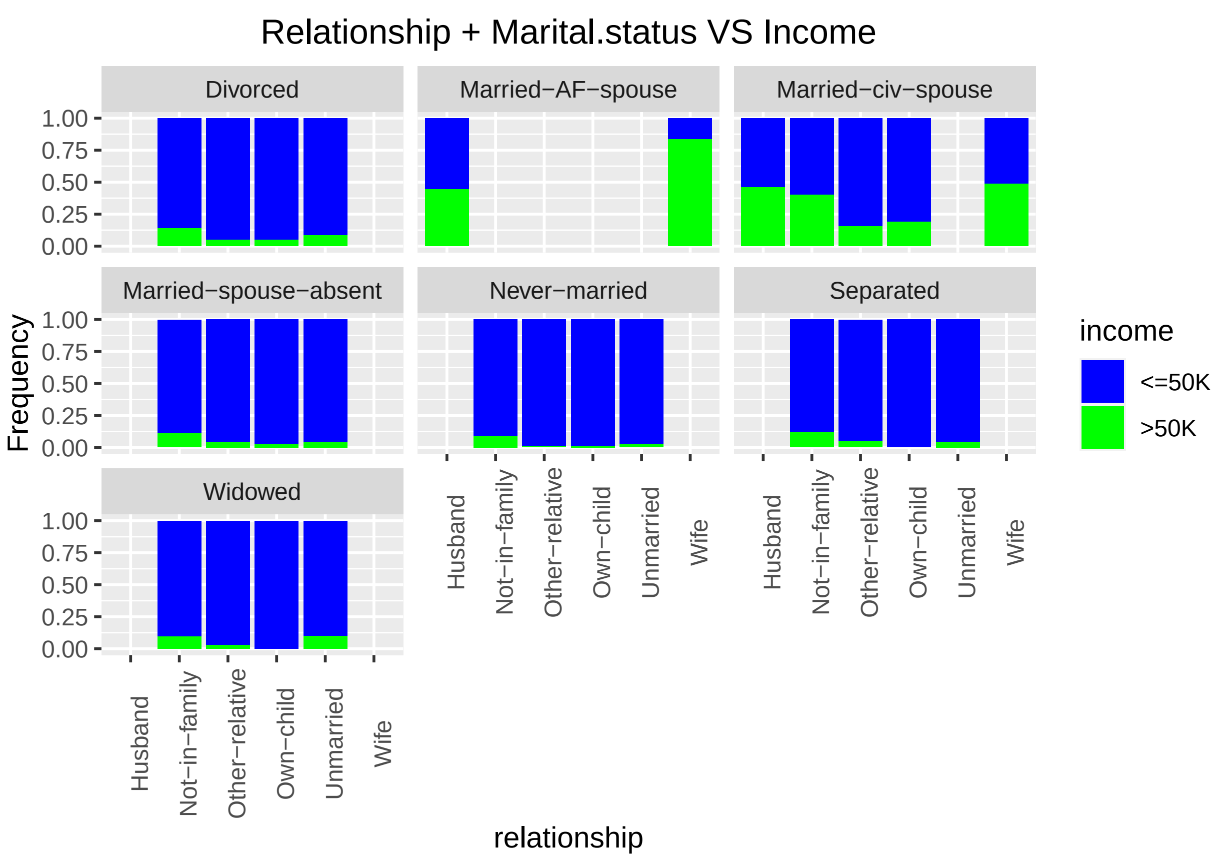


We can see that the native country of significant majority people is the United States. Although people from United States have the most amount of earning over 50k income, while in terms of the proportion of earning over 50k income, people from France, England, Taiwan, India, Japan, Cambodia, China etc have better chance to earn over 50k income.

Now let us further take a look of multiple variable combinations analysis.

Plot “relationship + marital.status” VS “income” from both amount and frequency angles


Relationship + Marital.status VS Income
Divorced
Married−AF−spouse
Married−civ−spouse
Married−spouse−absent
Never−married
Separated
Widowed
count
10000
7500
5000
2500
0
Husband
Not−in−family
Other−relative
Own−child
Unmarried
Wife
relationship
income
<=50K
>50K


Relationship + Marital.status VS Income
Divorced
Married−AF−spouse
Married−civ−spouse
Married−spouse−absent
Never−married
Separated
Widowed
Frequency
1.00
0.75
0.50
0.25
0.00
Husband
Not−in−family
Other−relative
Own−child
Unmarried
Wife
relationship
income
<=50K
>50K

Again, we can see married husband and wife have better chance to earn more than 50k income

Plot “workclass + sex” VS “income” from both amount and frequency points of view

From the chart above we can see that male in Self-emp-inc have the best chance to earn more than 50k income. we can use similar approach to analyze other combinations. The details will not be listed here.

Now we have completed a detailed data analysis for each variables in adult_train. The analysis indicates that a married male adult aging from 35 to 65 who works in self-emp-inc with education above bachelors and capital gains is more likely to earn more than 50k income. We'll verify our finding by variable importance of random forest model in modeling session.

#Modeling

In this session, we will use three models to predict whether a given adult will earn more than 50k income. There are logistic regression model, classification (decision tree) model and random forest model.

First, let us start from logistic regression model.

```
#drop irrelative "fnlwgt" from adult_test

adult_test <- adult_test[,-3]
adult_test$income <- as.factor(adult_test$income)

#Model 1: Logistic Regression

glmfit <- glm(income~., data=adult_train, family=binomial)
pred<- predict(glmfit,newdata=adult_test,type = 'response')
pred_lgr<- ifelse(pred>0.5,">50K","<=50K")
confusionMatrix(factor(pred_lgr),adult_test$income,positive = ">50K")
```

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

```
##           Reference
## Prediction <=50K >50K
##      <=50K  4202  598
##      >50K    329  904
##
##                Accuracy : 0.8463
##                  95% CI : (0.837, 0.8554)
##     No Information Rate : 0.751
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.563
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.6019
##             Specificity : 0.9274
##          Pos Pred Value : 0.7332
##          Neg Pred Value : 0.8754
##              Prevalence : 0.2490
##          Detection Rate : 0.1498
##    Detection Prevalence : 0.2044
##       Balanced Accuracy : 0.7646
##
##        'Positive' Class : >50K
##
```

```
Lgr_Accu <-confusionMatrix(factor(pred_lgr),adult_test$income,positive = ">50K")$overall["Accuracy"]
```

From consufionMatrix, we can see the prediction accuracy of logistic regression is 0.8463. It seems good. While if we take a look of Sensitivity and Specificity, we would not agree that the model is good. With Sensitivity 0.6019 and Specificity 0.9247, it means the model predicts better for income "<=50k" but not income ">50k". This low sensitivity may come from the highly imbalanced distribution of our target income as mentioned at the beginning. Since our target is to predict a given adult with income ">50k", we need a better model to achieve it. We will re-sample train set to check whether we can improve the sensitivity.

```
#resample train set to get a balanced one
adult_train_balanced <- ovun.sample(income~.,data=adult_train,method = "both")$data

glmfit_balanced <- glm(income~., data=adult_train_balanced, family=binomial)
pred_balanced<- predict(glmfit_balanced,newdata=adult_test,type = 'response')
pred_lgr_balanced<- ifelse(pred_balanced>0.5,">50K","<=50K")
confusionMatrix(factor(pred_lgr_balanced),adult_test$income,positive = ">50K")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction <=50K >50K
##      <=50K  3557  242
##      >50K    974 1260
##
##                Accuracy : 0.7984
##                  95% CI : (0.7881, 0.8085)
##     No Information Rate : 0.751
```

```
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5365
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.8389
##             Specificity : 0.7850
##          Pos Pred Value : 0.5640
##          Neg Pred Value : 0.9363
##              Prevalence : 0.2490
##          Detection Rate : 0.2089
##    Detection Prevalence : 0.3703
##       Balanced Accuracy : 0.8120
##
##        'Positive' Class : >50K
##
```

```
Lgr_Accu_Balanced <- confusionMatrix(factor(pred_lgr_balanced),adult_test$income,positive = ">50K")$over
```

Now we can see the accuracy drops to 0.7984 from 0.8463. However both Sensitivity and Specificity are more balanced with value 0.8389 and 0.7850. The sensitivity increases to favor our target of predicting a given adult with income ">50k".

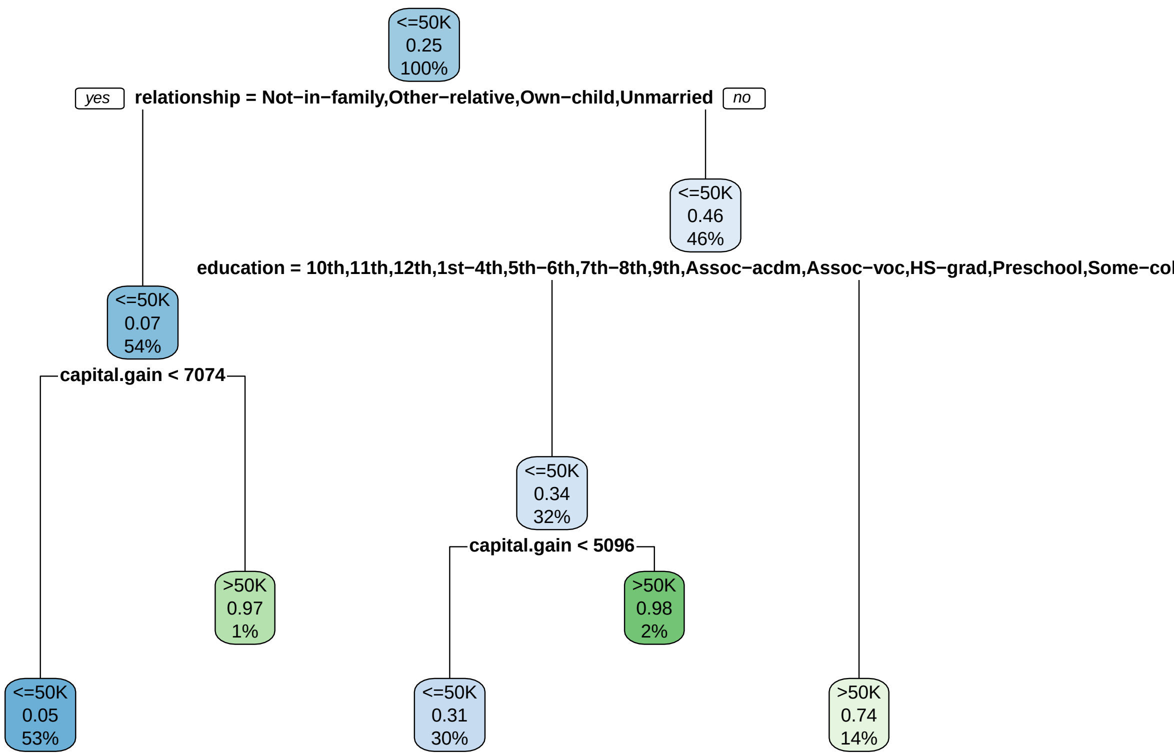
Then we will move to classification (decision) tree model.

```
#Model 2: Classification (Decision) Tree

Dctfit <- rpart(income ~., data = adult_train, method = "class")
Pred_Dct<- predict(Dctfit,newdata = adult_test,type = 'class')
confusionMatrix(Pred_Dct,adult_test$income,positive = ">50K")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction <=50K >50K
##      <=50K  4299  754
##      >50K    232  748
##
##                Accuracy : 0.8366
##                  95% CI : (0.827, 0.8458)
##     No Information Rate : 0.751
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5055
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.4980
##             Specificity : 0.9488
##          Pos Pred Value : 0.7633
##          Neg Pred Value : 0.8508
##              Prevalence : 0.2490
```

```
##          Detection Rate : 0.1240
##    Detection Prevalence : 0.1624
##       Balanced Accuracy : 0.7234
##
##        'Positive' Class : >50K
##
```

```
Dct_Accu <- confusionMatrix(Pred_Dct,adult_test$income,positive = ">50K")$overall["Accuracy"]

#Visualize the decision tree
rpart.plot(Dctfit)
```



From consufionMatrix, we can see the prediction accuracy of classification tree is 0.8366. The accuracy is a little bit worse than logistic regression. It may hit overfitting issue. While if we take a look of Sensitivity and Specificity, we would not agree that the model is good. With Sensitivity 0.4980 and Specificity 0.9488, it means the model predicts better for income "<=50k" instead of income ">50k".

Last, we will try random forest model. Random forest is a collection of decision trees on randomly selected samples. Random forest can mitigate over fitting issue which usually occurs in decision tree model. We expect random forest model will be more accurate than classification tree.

```
#Model 3: Random Forest

RfFit<- randomForest(income~.,data= adult_train, importance = TRUE)
Pred_Rf<- predict(RfFit,newdata = adult_test, type = 'class')
confusionMatrix(Pred_Rf,adult_test$income,positive = ">50K")
```

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

```
##           Reference
## Prediction <=50K >50K
##      <=50K  4197  520
##      >50K    334  982
##
##                Accuracy : 0.8584
##                  95% CI : (0.8494, 0.8671)
##     No Information Rate : 0.751
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6051
##
##  Mcnemar's Test P-Value : 2.443e-10
##
##             Sensitivity : 0.6538
##             Specificity : 0.9263
##          Pos Pred Value : 0.7462
##          Neg Pred Value : 0.8898
##              Prevalence : 0.2490
##          Detection Rate : 0.1628
##    Detection Prevalence : 0.2181
##       Balanced Accuracy : 0.7900
##
##        'Positive' Class : >50K
##
```

```
Rf_Accu <- confusionMatrix(Pred_Rf,adult_test$income,positive = ">50K")$overall["Accuracy"]
```

From consufionMatrix, we can see the prediction accuracy of random forest increases to 0.8584. While if we take a look of Sensitivity and Specificity, we would not agree that the model is a perfect one. With Sensitivity 0.6538 and Specificity 0.9263, it means the model predicts better for income "<=50k" but not income ">50k".

Now we will take a look of variable importance.

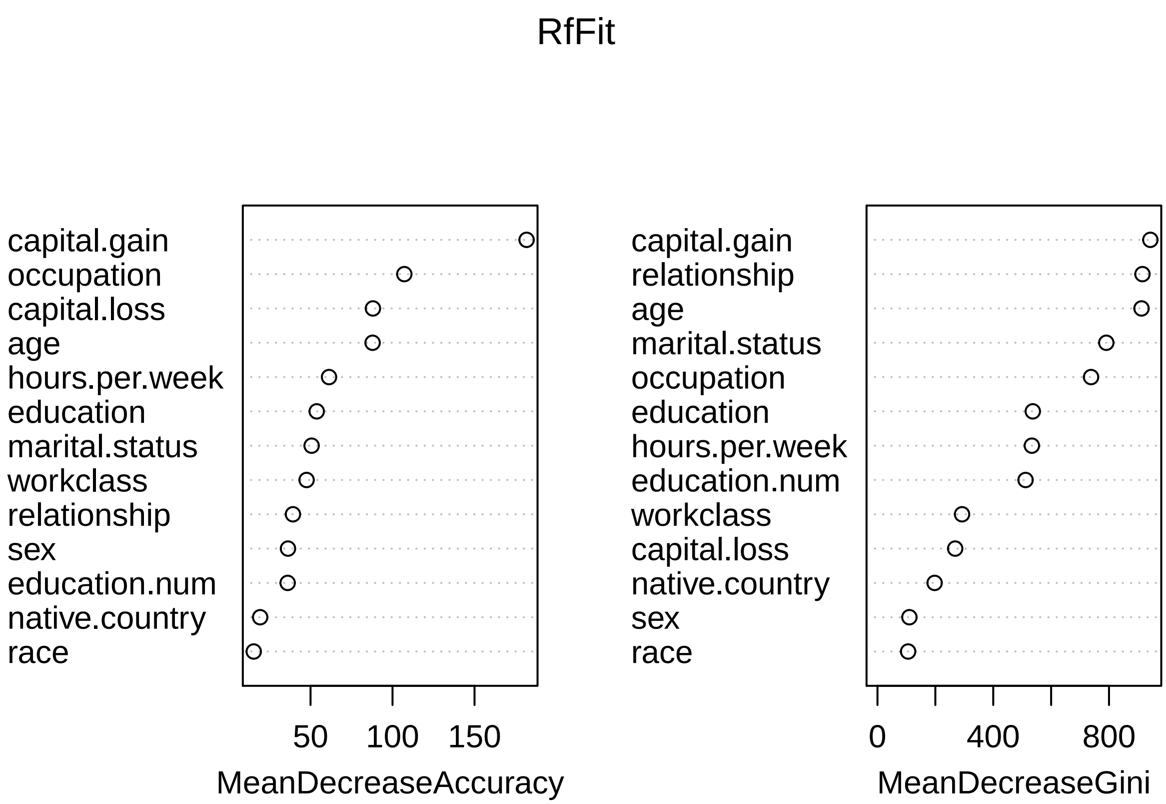
```
RfFit$importance
```

```
##                      <=50K          >50K MeanDecreaseAccuracy MeanDecreaseGini
## age            0.002037681  0.0682615308         0.0185278126         912.2928
## workclass      0.006519288  0.0029722907         0.0056382525         292.2871
## education      0.025176496  0.0115848385         0.0217862505         536.7119
## education.num  0.029438318  0.0137852138         0.0255480347         511.7493
## marital.status 0.035657179  0.0721046171         0.0447249445         790.6885
## occupation     0.015714461  0.0560837262         0.0257718767         737.9952
## relationship   0.030584061  0.0750832079         0.0416641460         915.5290
## race           0.001058692  0.0006168173         0.0009499212         105.9698
## sex            0.007602811  0.0021731845         0.0062516115         110.3120
## capital.gain   0.030399974  0.0622677662         0.0383330276         942.8526
## capital.loss   0.004264547  0.0202576775         0.0082483393         268.5162
## hours.per.week 0.003309421  0.0302975380         0.0100310069         533.4344
## native.country 0.002182078 -0.0006088506         0.0014873418         197.4860
```

```r
rf_imp = data.table(RfFit$importance,names = rownames(RfFit$importance))[order(-MeanDecreaseGini)]

rf_imp
```

```
##          <=50K          >50K MeanDecreaseAccuracy MeanDecreaseGini
##  1: 0.030399974  0.0622677662         0.0383330276         942.8526
##  2: 0.030584061  0.0750832079         0.0416641460         915.5290
##  3: 0.002037681  0.0682615308         0.0185278126         912.2928
##  4: 0.035657179  0.0721046171         0.0447249445         790.6885
##  5: 0.015714461  0.0560837262         0.0257718767         737.9952
##  6: 0.025176496  0.0115848385         0.0217862505         536.7119
##  7: 0.003309421  0.0302975380         0.0100310069         533.4344
##  8: 0.029438318  0.0137852138         0.0255480347         511.7493
##  9: 0.006519288  0.0029722907         0.0056382525         292.2871
## 10: 0.004264547  0.0202576775         0.0082483393         268.5162
## 11: 0.002182078 -0.0006088506         0.0014873418         197.4860
## 12: 0.007602811  0.0021731845         0.0062516115         110.3120
## 13: 0.001058692  0.0006168173         0.0009499212         105.9698
##              names
##  1:   capital.gain
##  2:   relationship
##  3:            age
##  4: marital.status
##  5:     occupation
##  6:      education
##  7: hours.per.week
##  8:  education.num
##  9:      workclass
## 10:   capital.loss
## 11: native.country
## 12:            sex
## 13:           race
```

```r
varImpPlot(RfFit)
```



From the above important variables plot, we can see that the most important variables are relationship and capital.gain.

Now we compare the accuracy of four predictions.

```
Accuracy_Comp<-data.frame(Model=c('Logistic Regression','Logistic Regression With Balanced Sample','Dec

Accuracy_Comp
```

```
##                                    Model  Accuracy
## 1                    Logistic Regression 0.8463451
## 2 Logistic Regression With Balanced Sample 0.7984419
## 3                          Decision Tree 0.8365656
## 4                          Random Forest 0.8584452
```

3 CONCLUSION

3.1 Conclusion

Now we know random forest has the best accuracy 0.8584 on predicting a given adult earning over 50k in four models. This result meets our expectation that random forest can mitigate over fitting issue which usually occurs in decision tree model therefore have better performance on predication. However all four models except the one "Logistic Regression With Balanced Sample" hit the same issue of low sensitivity.

3.2 Future Work

We need more powerful machine learning algorithms and techniques to predict this imbalanced classification target. We may further optimize random forest model by using Hyperparameters technique or other optimization models.