Capstone: Predict Adult Income

Introduction

The goal of this project is to predict whether the income of adults is greater than $50k/yr. This is a classification problem. We use a couple of algorithms to train the model and then choose the most suitable one.

Dataset

Adult Data Set is used for the prediction purpose. The dataset can be downloaded using this URL: [https://archive.ics.uci.edu/ml/datasets/Adult]. As mentioned in this website - "Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0)). Prediction task is to determine whether a person makes over 50K a year."

The following files are available:

1. adult.data
2. adult.test
3. adult.names

adult.names file provides us with information about the data. It is mentioned that there are a total of 48842 instances which is split into train-test datasets.Train dataset is available in adult.data file, it has 32561 instances and test dataset is available in adult.test file, it has 16281 instances.

The variables of this dataset are as follows:

age, workclass, fnlwgt, education, education_num, marital_status, occupation, relationship, race, sex, capital_gain, capital_loss, hours_per_week, native_country, income - variable to be predicted.

Procedure

First we download data and then preprocess it. Data exploration is done to identify which of the above mentioned variables are to be used for income prediction. Then the following algorithms are applied to train the model. For training the model, data is split into train(90%) and validation(10%) datasets.

Algorithms used -
1. Logistic Regression
2. Support Vector Classifier
3. Random Forest Classifier
4. Gradient Boosting Classifier

Model performance is evaluated based on Overall Accuracy and F1 score. Confusion Matrix is used to obtain Accuracy and F1 score.

Accuracy - This is defined as the no. of correct predictions divided by the total number of the dataset.

F1-score - This is the harmonic mean of precision and recall.

$$F1 - score = \frac{2 * precision * recall}{precision + recall}$$

where, precision is Positive Predictive Value recall is sensitivity

If results of confusion Matrix is stored in "cm", we can obtain accuracy using the command cm$overall["Accuracy"] and F1-score using the command cm$byClass["F1"]

Downloading Data:

The required libraries and data are downloaded. Data is read into the file "income_data" and column names are assigned to the data.

```
#__________________________
#Download required packages
#__________________________

if(!require(tidyverse))
  install.packages("tidyverse", repos = "http://cran.us.r-project.org")
if(!require(caret))
  install.packages("caret", repos = "http://cran.us.r-project.org")
if(!require(data.table))
  install.packages("data.table", repos = "http://cran.us.r-project.org")
if(!require(e1071))
  install.packages("e1071", repos = "http://cran.us.r-project.org")
if(!require(randomForest))
  install.packages("randomForest", repos = "http://cran.us.r-project.org")

#_____________________________________________________________
# Part 1: Downloading data, preprocessing and Data exploration
#_____________________________________________________________

# Downloading and Reading data

if(!file.exists("adult.data")){
  download.file("http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data","adult.data
}
income_data <- read.csv("adult.data",header = FALSE)

# Set column names
colnames(income_data) <- c("age","workclass","fnlwgt","education","education_num","marital_status",
                           "occupation","relationship","race","sex","capital_gain","capital_loss",
                           "hours_per_week","native_country","income")
```

Methods/Analysis

Data preprocessing

First we check the data dimensions. There are 32561 rows and 15 columns. The structure of data reveals that there are 6 columns of type integer and 9 columns of type factor. Also, we can notice some columns have data in the form " ?" which is unknown/missing data(mentioned in adult.names) file. We can further confirm this using summary() and levels command. The no. of rows with missing data is few in number, so we remove these rows.

We re-read the income_data file and convert " ?" to NA and then omit rows containing NA. We set the column names again. Now the no. of rows is 30162 and 15 columns.

```
# Explore data

# To get the dimension of data
dim(income_data)
```

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

```r
# To check the structure of data
str(income_data)
```

```
## 'data.frame':    32561 obs. of  15 variables:
##  $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ workclass     : Factor w/ 9 levels " ?"," Federal-gov",..: 8 7 5 5 5 5 5 7 5 5 ...
##  $ fnlwgt        : int  77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...
##  $ education     : Factor w/ 16 levels " 10th"," 11th",..: 10 10 12 2 10 13 7 12 13 10 ...
##  $ education_num : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ marital_status: Factor w/ 7 levels " Divorced"," Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...
##  $ occupation    : Factor w/ 15 levels " ?"," Adm-clerical",..: 2 5 7 7 11 5 9 5 11 5 ...
##  $ relationship  : Factor w/ 6 levels " Husband"," Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...
##  $ race          : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...
##  $ sex           : Factor w/ 2 levels " Female"," Male": 2 2 2 2 1 1 1 2 1 2 ...
##  $ capital_gain  : int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ capital_loss  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hours_per_week: int  40 13 40 40 40 40 16 45 50 40 ...
##  $ native_country: Factor w/ 42 levels " ?"," Cambodia",..: 40 40 40 40 6 40 24 40 40 40 ...
##  $ income        : Factor w/ 2 levels " <=50K"," >50K": 1 1 1 1 1 1 1 2 2 2 ...
```

```r
# To get summary of data
summary(income_data)
```

```
##       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
##   Prof-specialty :4140   Husband       :13193
##   Craft-repair   :4099   Not-in-family : 8305
##   Exec-managerial:4066   Other-relative:  981
##   Adm-clerical   :3770   Own-child     : 5068
##   Sales          :3650   Unmarried     : 3446
##   Other-service  :3295   Wife          : 1568
##  (Other)         :9541
##                  race              sex          capital_gain
##   Amer-Indian-Eskimo:  311   Female:10771   Min.   :    0
##   Asian-Pac-Islander: 1039   Male  :21790   1st Qu.:    0
##   Black             : 3124                  Median :    0
##   Other             :  271                  Mean   : 1078
##   White             :27816                  3rd Qu.:    0
##                                             Max.   :99999
##
```

```
##   capital_loss    hours_per_week         native_country     income
##  Min.   :   0.0   Min.   : 1.00    United-States:29170   <=50K:24720
##  1st Qu.:   0.0   1st Qu.:40.00    Mexico       :  643   >50K : 7841
##  Median :   0.0   Median :40.00    ?            :  583
##  Mean   :  87.3   Mean   :40.44    Philippines  :  198
##  3rd Qu.:   0.0   3rd Qu.:45.00    Germany      :  137
##  Max.   :4356.0   Max.   :99.00    Canada       :  121
##                                   (Other)       : 1709
```

```
# To explore the levels in each column
sapply(income_data, levels)
```

```
## $age
## NULL
##
## $workclass
## [1] " ?"                " Federal-gov"      " Local-gov"
## [4] " Never-worked"     " Private"          " Self-emp-inc"
## [7] " Self-emp-not-inc" " State-gov"        " Without-pay"
##
## $fnlwgt
## NULL
##
## $education
##  [1] " 10th"         " 11th"         " 12th"         " 1st-4th"
##  [5] " 5th-6th"      " 7th-8th"      " 9th"          " Assoc-acdm"
##  [9] " Assoc-voc"    " Bachelors"    " Doctorate"    " HS-grad"
## [13] " Masters"      " Preschool"    " Prof-school"  " Some-college"
##
## $education_num
## NULL
##
## $marital_status
## [1] " Divorced"              " Married-AF-spouse"
## [3] " Married-civ-spouse"    " Married-spouse-absent"
## [5] " Never-married"         " Separated"
## [7] " Widowed"
##
## $occupation
##  [1] " ?"                 " Adm-clerical"      " Armed-Forces"
##  [4] " Craft-repair"      " Exec-managerial"   " Farming-fishing"
##  [7] " Handlers-cleaners" " Machine-op-inspct" " Other-service"
## [10] " Priv-house-serv"   " Prof-specialty"    " Protective-serv"
## [13] " Sales"             " Tech-support"      " Transport-moving"
##
## $relationship
## [1] " Husband"        " Not-in-family"  " Other-relative" " Own-child"
## [5] " Unmarried"      " Wife"
##
## $race
## [1] " Amer-Indian-Eskimo" " Asian-Pac-Islander" " Black"
## [4] " Other"              " White"
##
## $sex
## [1] " Female" " Male"
```

```
##
## $capital_gain
## NULL
##
## $capital_loss
## NULL
##
## $hours_per_week
## NULL
##
## $native_country
##  [1] " ?"                          " Cambodia"
##  [3] " Canada"                     " China"
##  [5] " Columbia"                   " Cuba"
##  [7] " Dominican-Republic"         " Ecuador"
##  [9] " El-Salvador"                " England"
## [11] " France"                     " Germany"
## [13] " Greece"                     " Guatemala"
## [15] " Haiti"                      " Holand-Netherlands"
## [17] " Honduras"                   " Hong"
## [19] " Hungary"                    " India"
## [21] " Iran"                       " Ireland"
## [23] " Italy"                      " Jamaica"
## [25] " Japan"                      " Laos"
## [27] " Mexico"                     " Nicaragua"
## [29] " Outlying-US(Guam-USVI-etc)" " Peru"
## [31] " Philippines"                " Poland"
## [33] " Portugal"                   " Puerto-Rico"
## [35] " Scotland"                   " South"
## [37] " Taiwan"                     " Thailand"
## [39] " Trinadad&Tobago"            " United-States"
## [41] " Vietnam"                    " Yugoslavia"
##
## $income
## [1] " <=50K" " >50K"
```

```r
# Convert " ?" data to NA and the remove rows with NA
income_data <- read.csv("adult.data",na.strings = c(" ?"),header = FALSE)

income_data <- na.omit(income_data)

# Set column names again
colnames(income_data) <- c("age","workclass","fnlwgt","education","education_num","marital_status",
                           "occupation","relationship","race","sex","capital_gain","capital_loss",
                           "hours_per_week","native_country","income")

# To get the dimension of new data
dim(income_data)
```

```
## [1] 30162    15
```

Splitting data into Train and validation sets:

We split the data into train and validation sets where validation set is 10% of the income_data. Train set has 27145 rows and validation has 3017 rows.

```r
# Create Train and validation set with validation set having 10% data
set.seed(1, sample.kind="Rounding")
# if using R 3.5 or earlier, use `set.seed(1)` instead

test_index <- createDataPartition(income_data$income,p = 0.1,
                                  times = 1,list = FALSE)
trainset <- income_data[-test_index, ]
validation <- income_data[test_index,]

# To get dimension of train and validation data
dim(trainset)
```

```
## [1] 27145    15
```

```r
dim(validation)
```

```
## [1] 3017   15
```

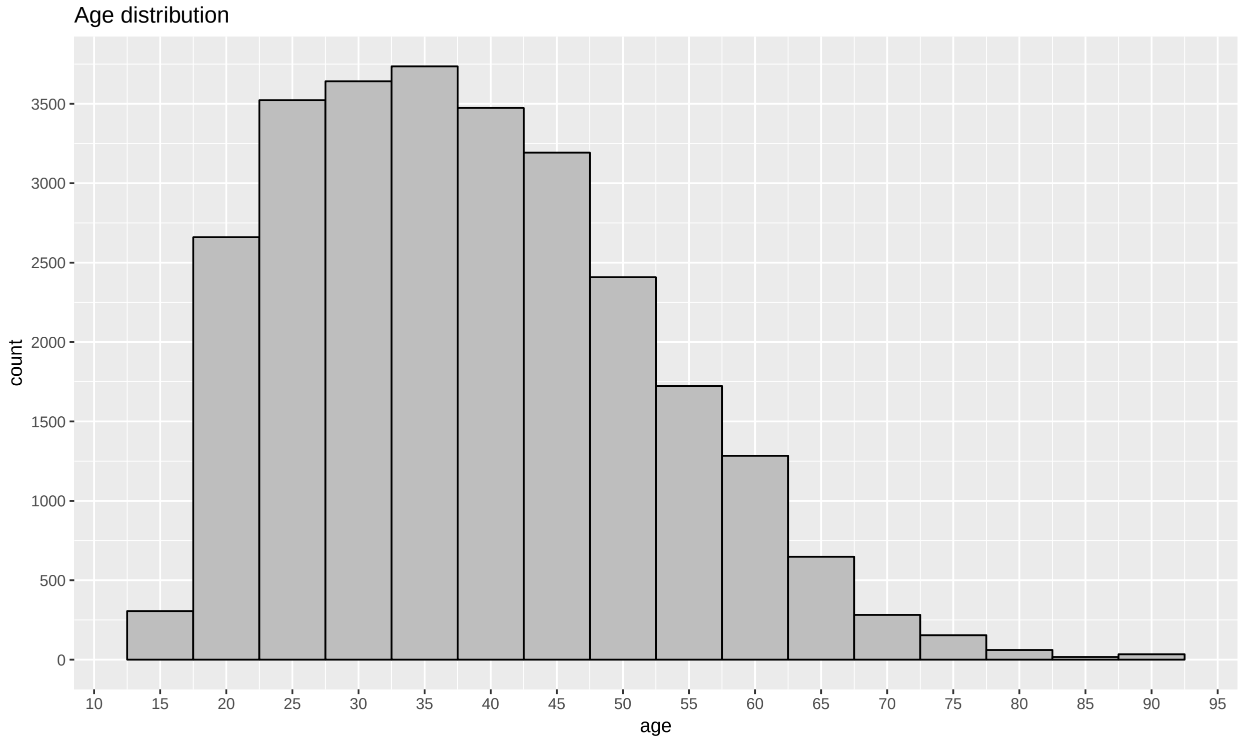
Data Exploration

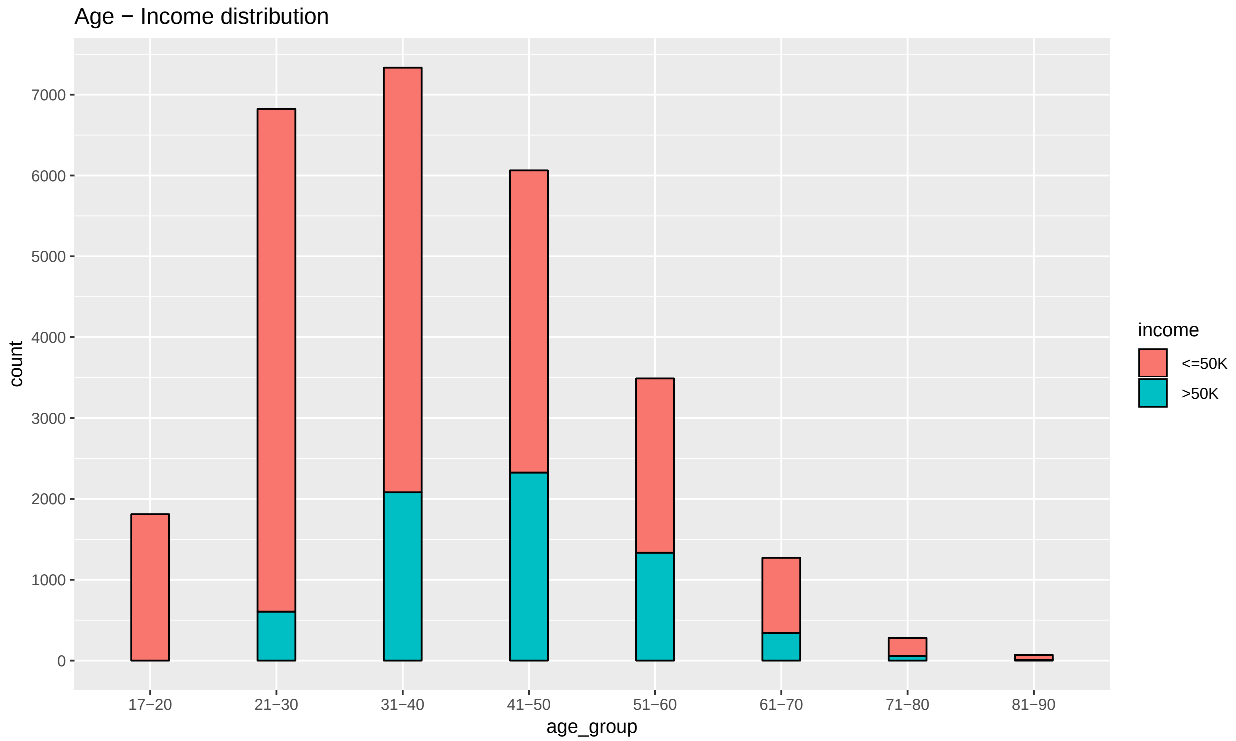
There are 14 variables, so we explore each variable further.

1. age

The minimum and maximum age is 17 and 90 yrs respectively. Mean age is 38 yrs. Most data lies in the range 20 to 50 yrs. Age vs income plot shows that the proportion of income ">50K" is greater for age >= 30 and age <=60 compared to other age groups.

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    17.00   28.00   37.00   38.43   47.00   90.00
```



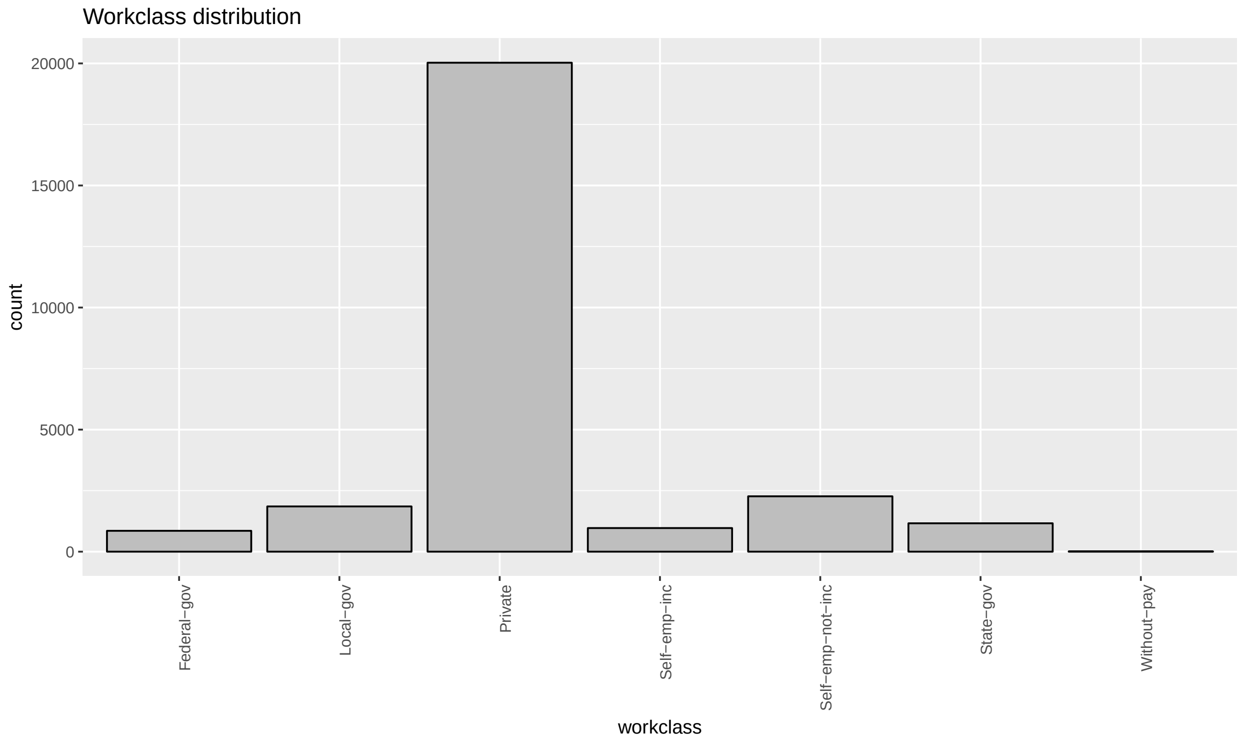


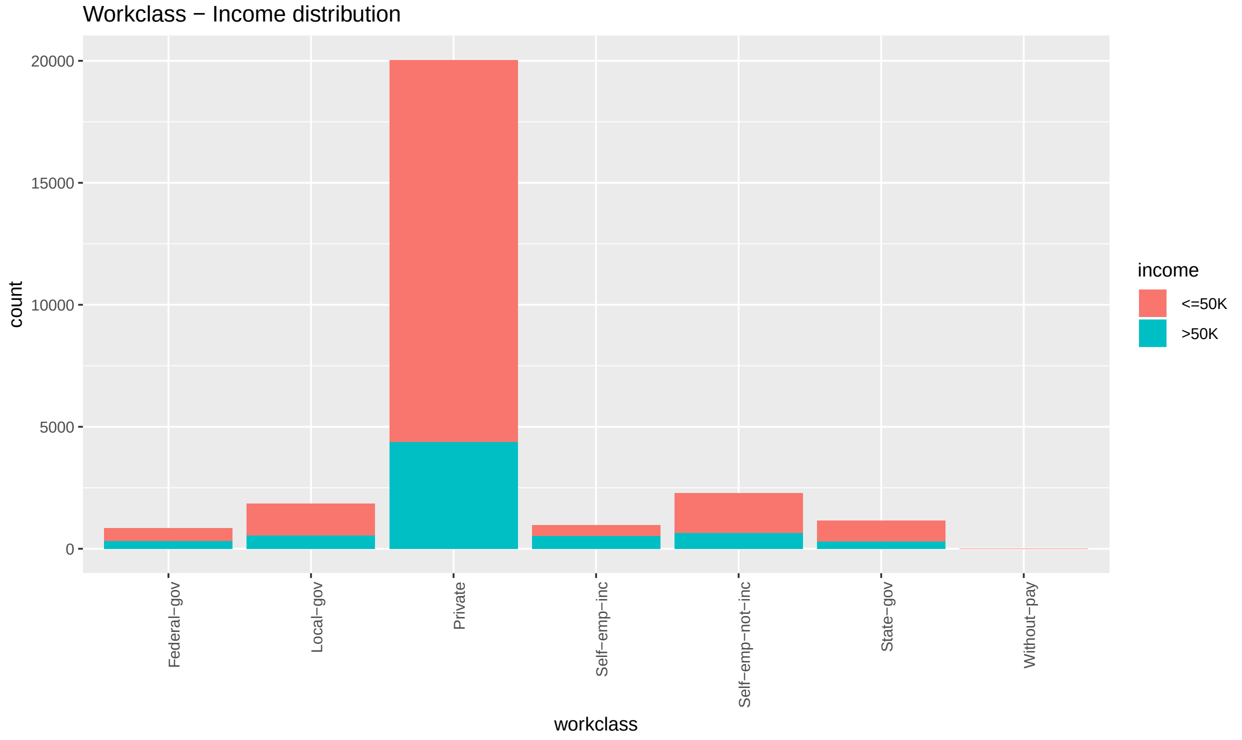
2. Workclass

Private sector constitutes majority of workclass. The level "never-worked" has 0 entries whereas without pay is negligible. The same can be seen in histogram as well.

The barplot with workclass and income also shows that income ">50k" is maximum in private workclass.

```
##       Federal-gov         Local-gov      Never-worked           Private
##               855              1853                 0             20028
##      Self-emp-inc  Self-emp-not-inc         State-gov       Without-pay
##               967              2268              1162                12
```

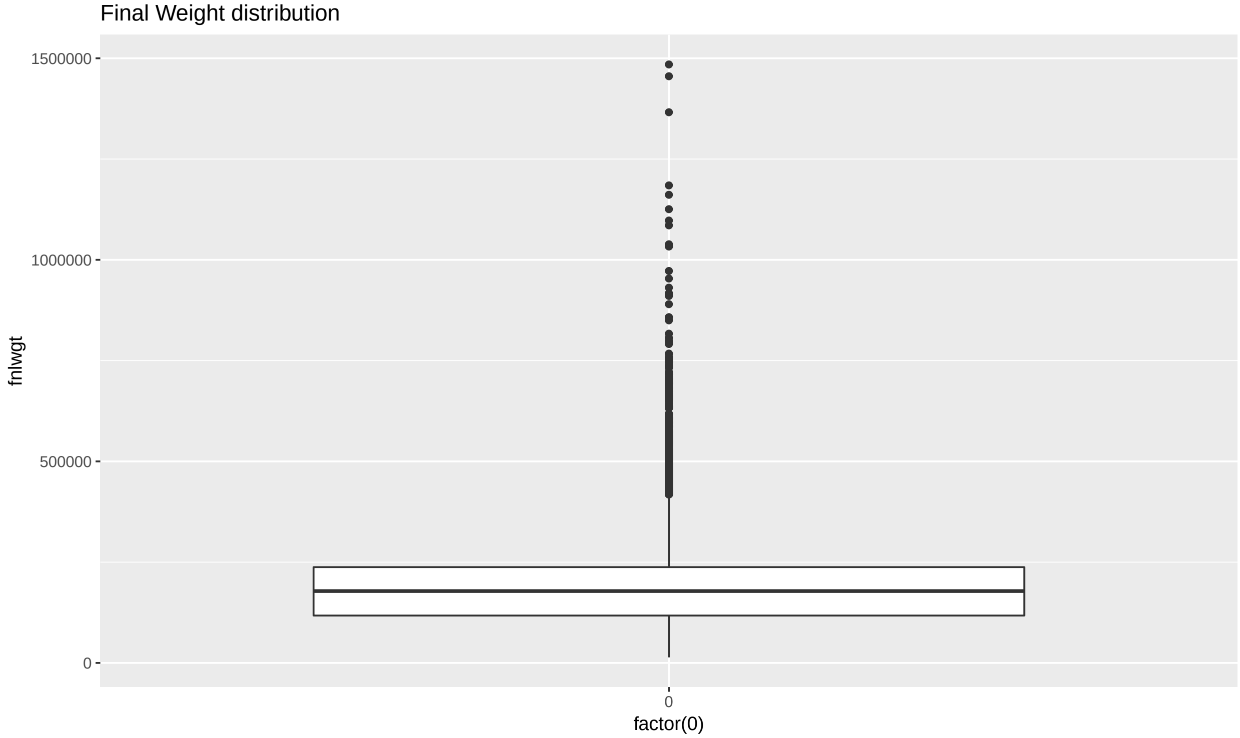


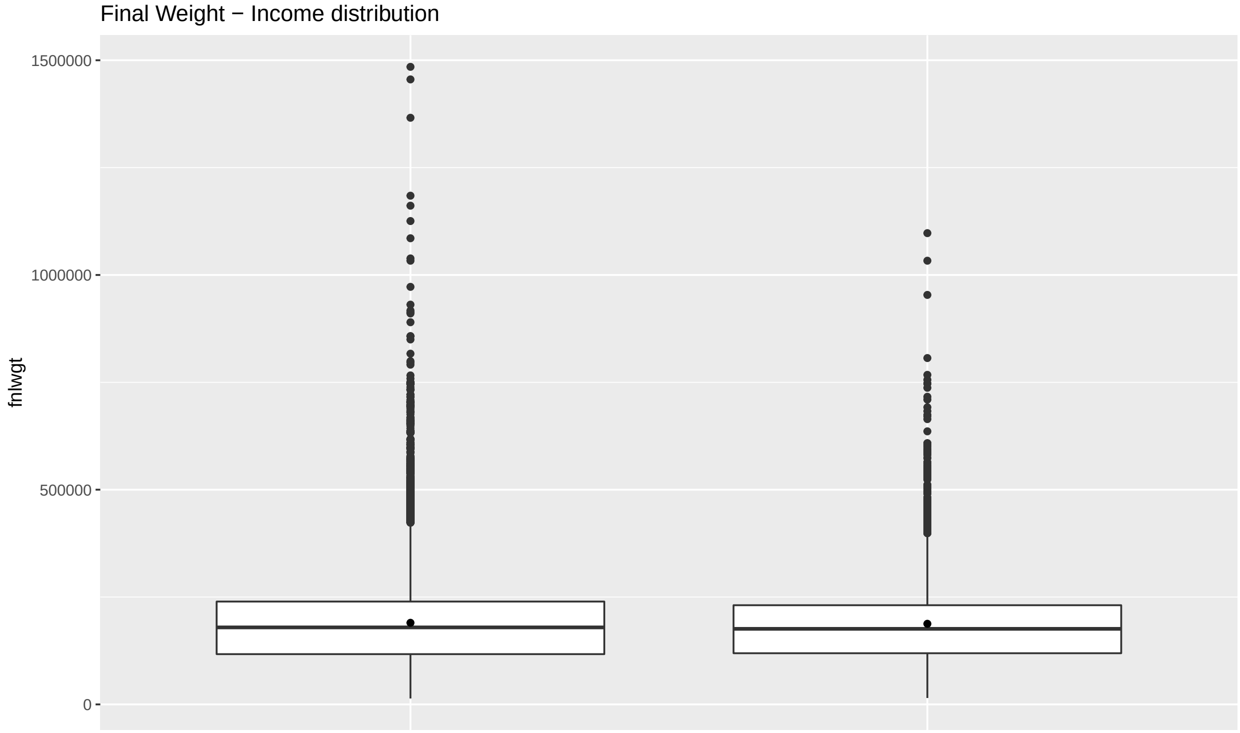


3. Final weight

Weights are assigned to participants based on demographic characteristics. For more details about fnlwgt, adult.names file can be referred. The minimum and maximum fnlwgt is 13769 and 1484705. As we can see in the boxplots there are many outliers and this variable is not very helpful in predicting income. Therefore we drop this variable.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13769  117526  178215  189453  237608 1484705
```

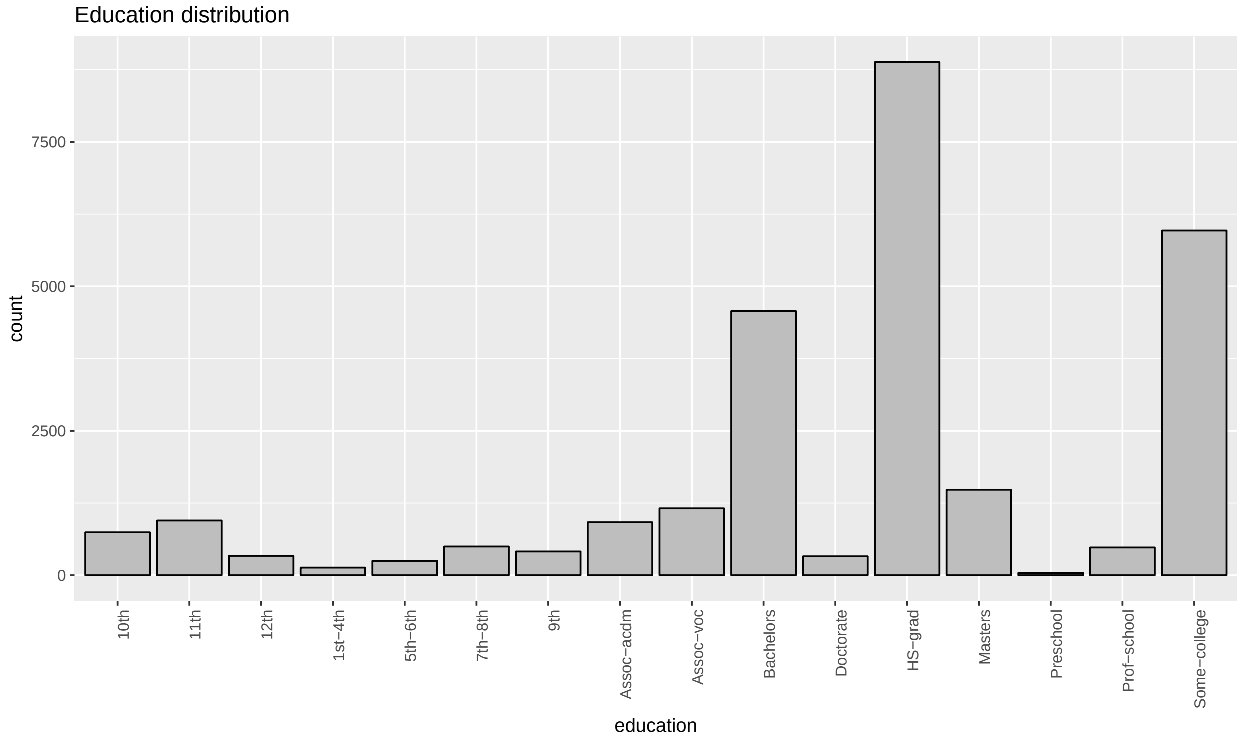


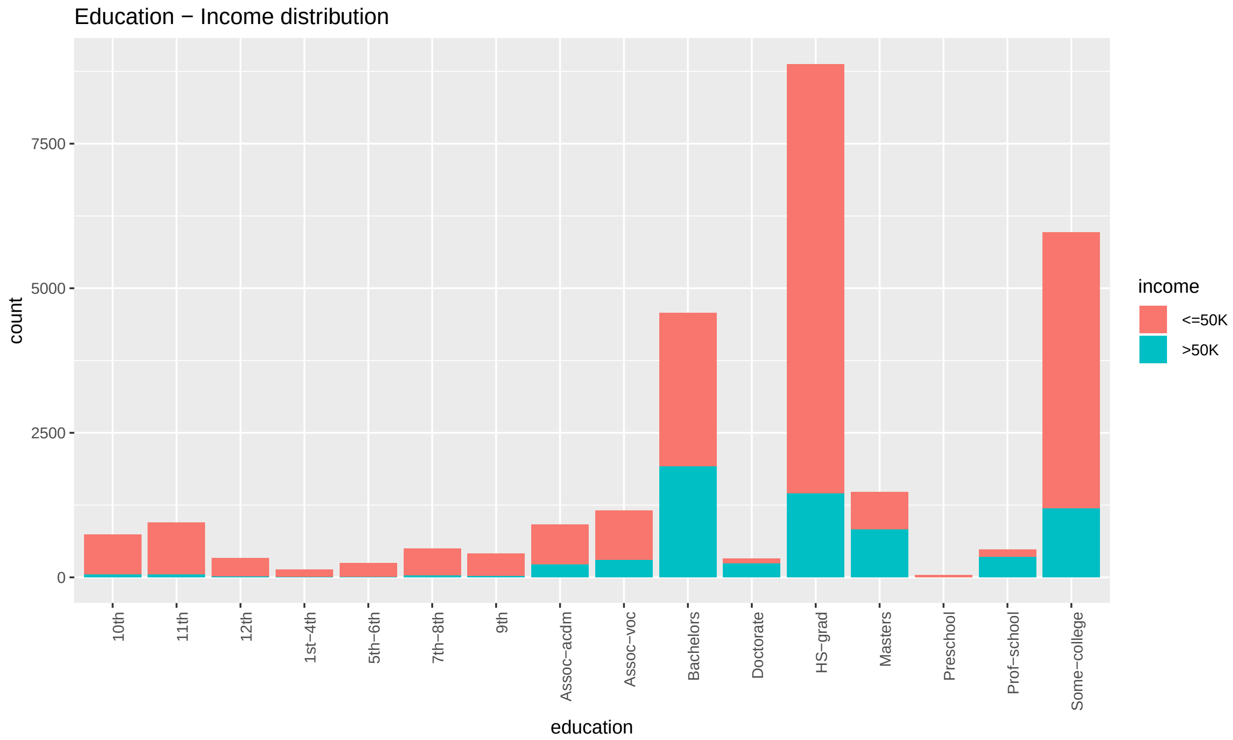


4. Education

The education of each individual can be seen in the histogram plot. Income “>50k” is more in the cases of bachelors, HS-grad, masters and some college. Doctorate and Prof-school have more cases of income “>50k” compared to the total no. of cases in the respective classes.

```
##          10th          11th          12th       1st-4th       5th-6th
##           743           948           337           133           250
##       7th-8th           9th    Assoc-acdm     Assoc-voc     Bachelors
##           498           412           917          1158          4572
##     Doctorate       HS-grad       Masters     Preschool   Prof-school
##           328          8879          1481            42           481
##  Some-college
##          5966
```





5. Education Number

Education_num is the numerical form of education. Therefore this is a redundant variable and we drop this variable.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    9.00   10.00   10.13   13.00   16.00
```

```
## # A tibble: 16 x 3
## # Groups:   education_num [16]
##    education_num education          n
##            <int> <fct>          <int>
##  1             1 " Preschool"      42
##  2             2 " 1st-4th"       133
##  3             3 " 5th-6th"       250
##  4             4 " 7th-8th"       498
##  5             5 " 9th"           412
##  6             6 " 10th"          743
##  7             7 " 11th"          948
##  8             8 " 12th"          337
##  9             9 " HS-grad"      8879
## 10            10 " Some-college"  5966
## 11            11 " Assoc-voc"     1158
## 12            12 " Assoc-acdm"     917
## 13            13 " Bachelors"     4572
## 14            14 " Masters"       1481
## 15            15 " Prof-school"    481
## 16            16 " Doctorate"      328
```

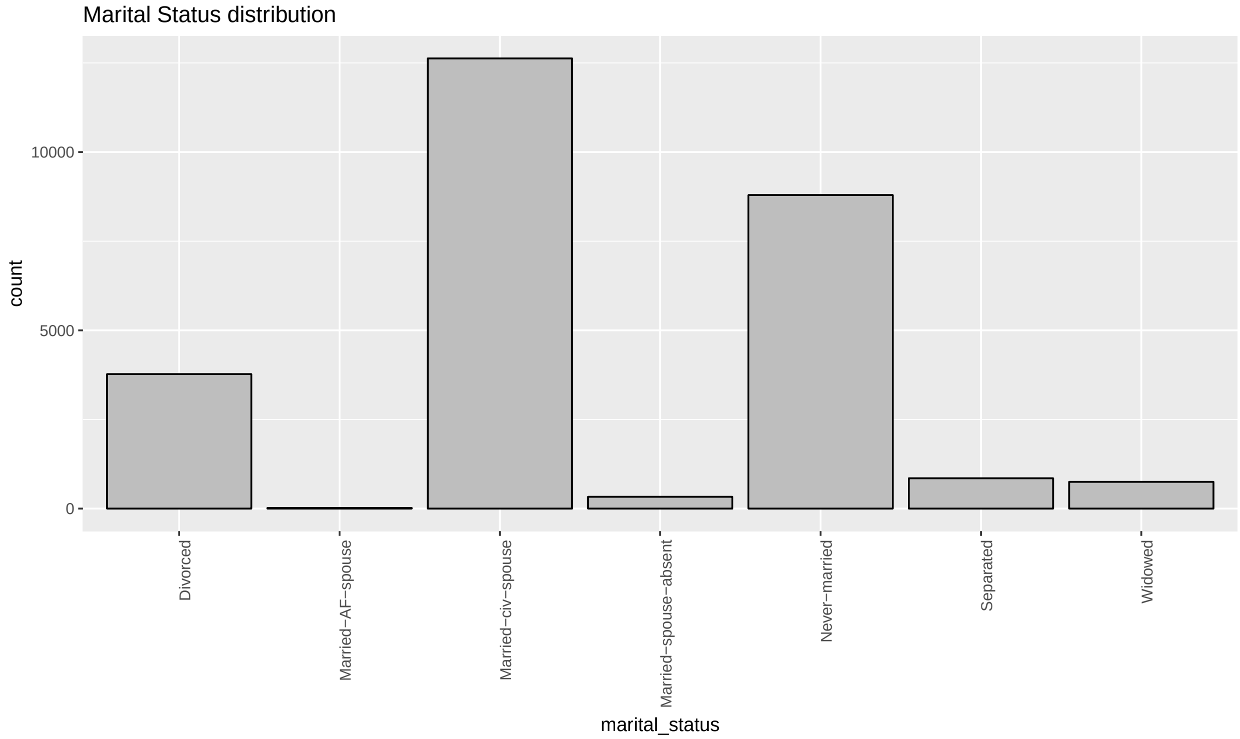
6. Marital Status

Summary and barplots indicate that majority of individuals are of the category married-civ-spouse. This same category has more no. of individuals with income ">50K".

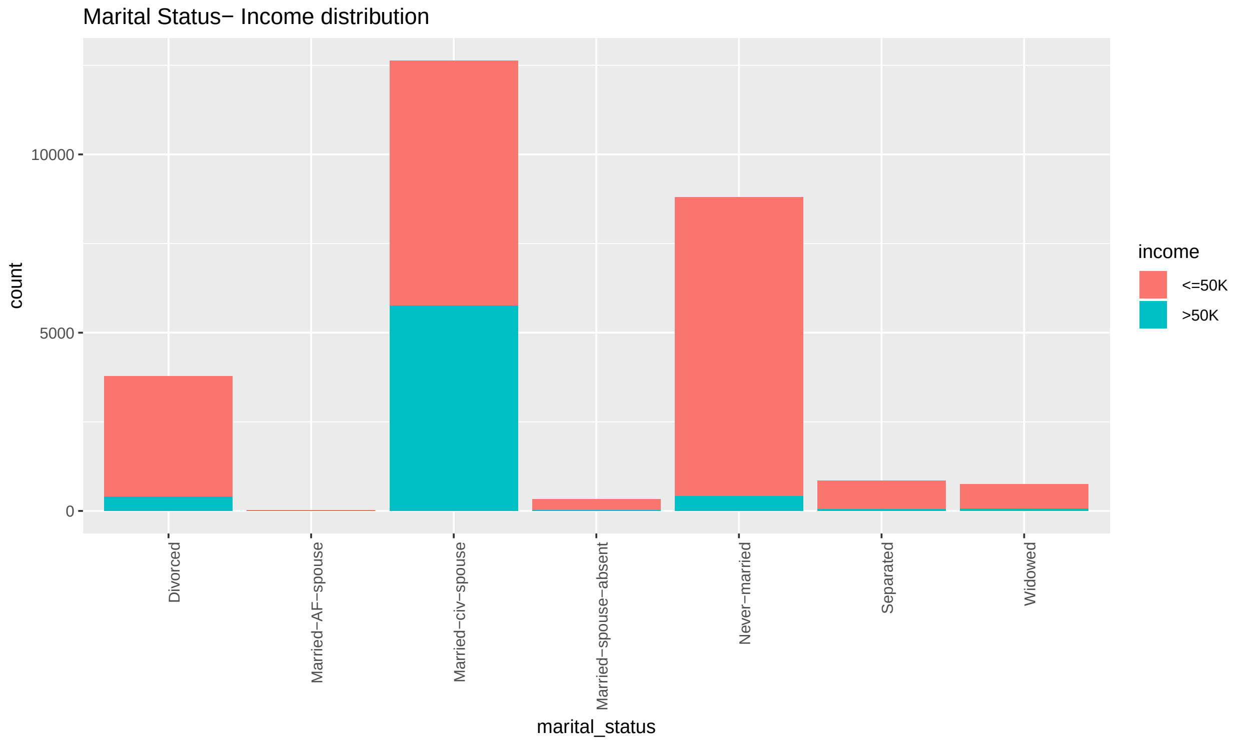
```
# 6. Marital status

#Summary of marital_status
summary(trainset$marital_status)
```

```
##              Divorced     Married-AF-spouse    Married-civ-spouse 
##                  3772                    19                 12627 
##  Married-spouse-absent         Never-married             Separated 
##                   331                  8795                   851 
##               Widowed 
##                   750
```

```
# This plot displays marital_status distribution
trainset  %>% ggplot(aes(marital_status)) +
  geom_bar(col = "black", fill = "grey") +
  theme(axis.text.x = element_text(angle=90,hjust = 1)) +
  ggtitle("Marital Status distribution")
```



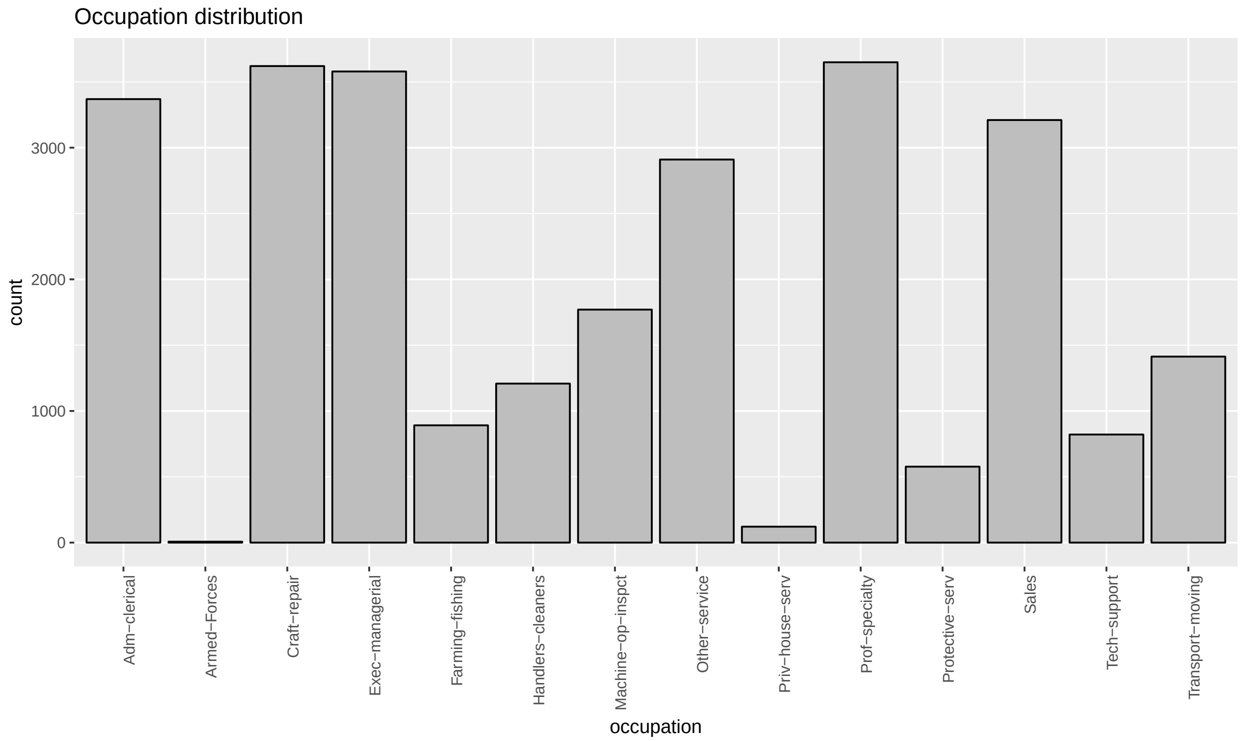
```
# This plot displays marital_status vs income
trainset  %>% ggplot(aes(marital_status,fill = income)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle=90,hjust = 1)) +
  ggtitle("Marital Status- Income distribution")
```



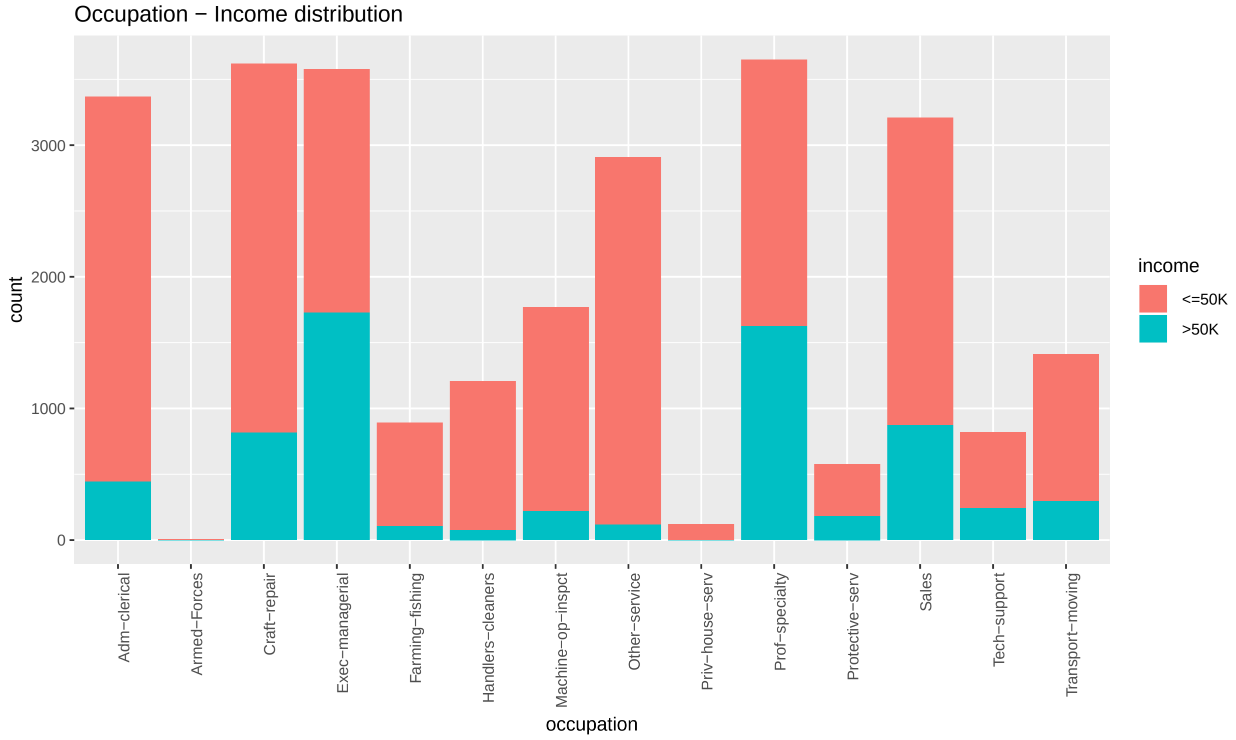
7. Occupation

People earning income “>50k” are mostly in the categories “Exec -Managerial” and “Prof-specialty”.

```
##       Adm-clerical       Armed-Forces       Craft-repair
##               3369                  7               3620
##    Exec-managerial    Farming-fishing  Handlers-cleaners
##               3579                891               1208
##  Machine-op-inspct      Other-service    Priv-house-serv
##               1770               2910                121
##     Prof-specialty    Protective-serv              Sales
##               3649                577               3210
##       Tech-support   Transport-moving
##                821               1413
```



Occupation distribution

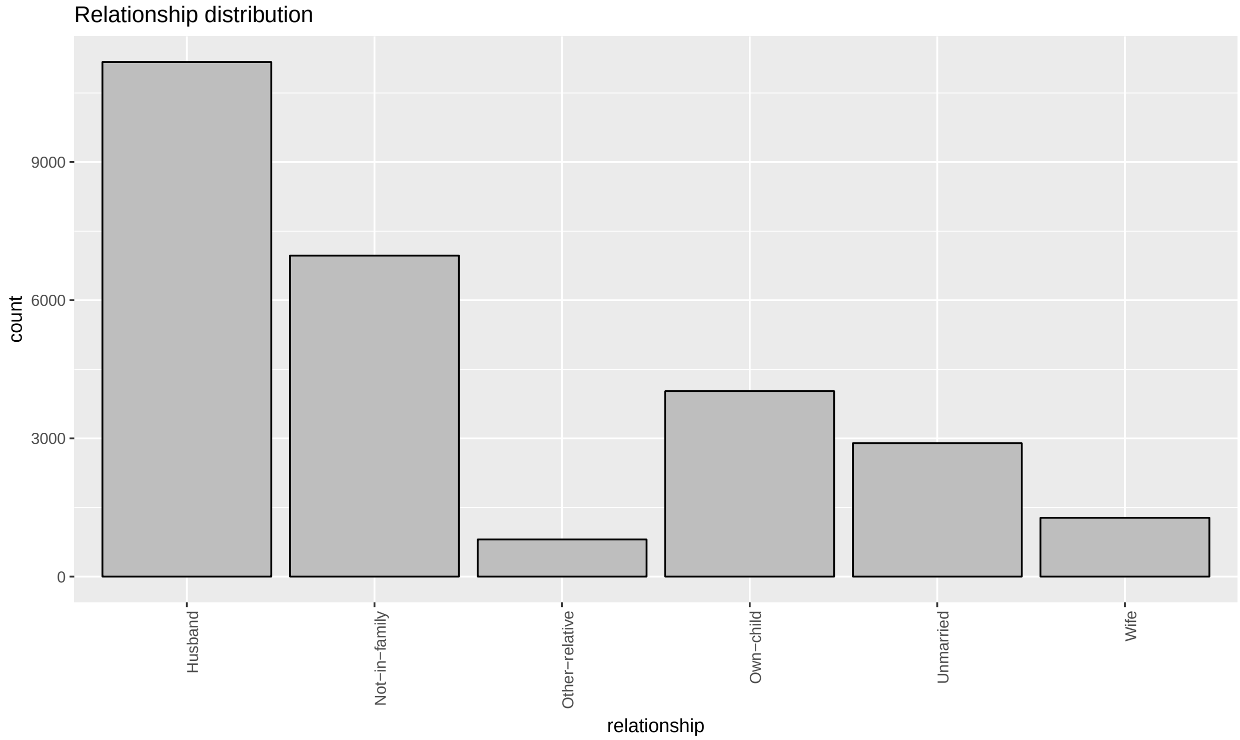


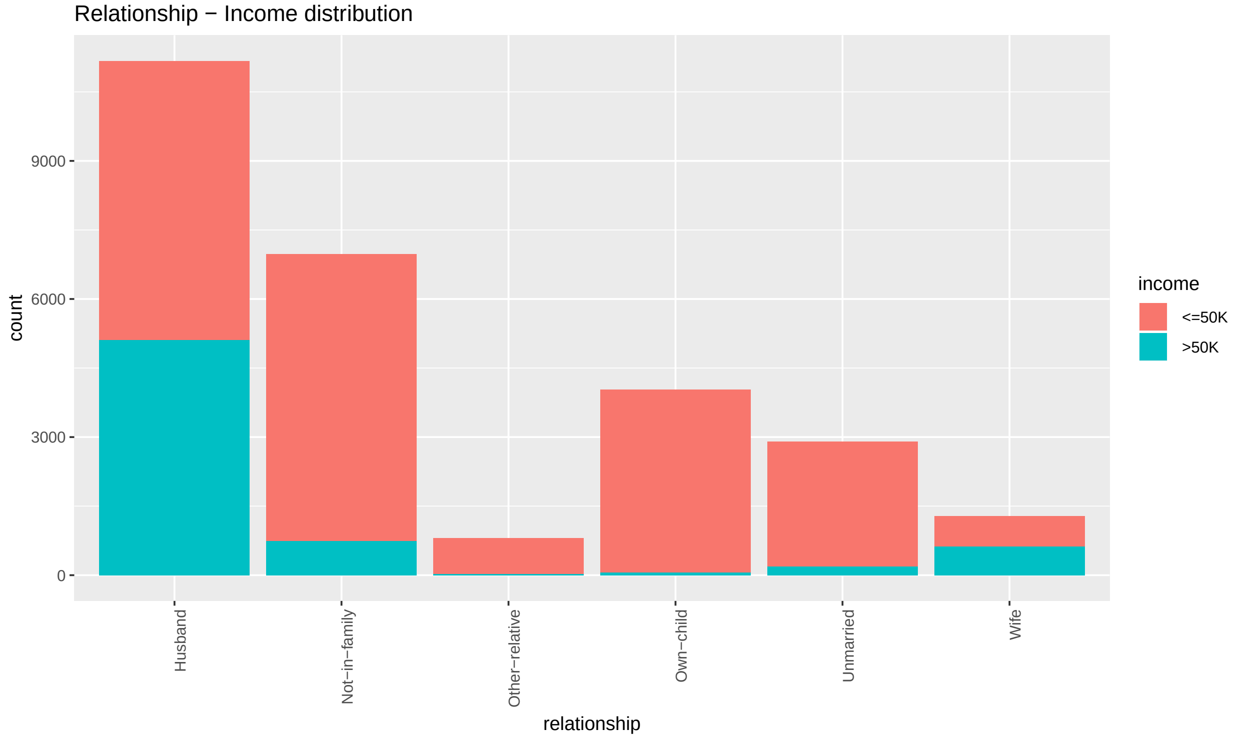
Occupation − Income distribution

8. Relationship

“Husband” category has most no. of entries and also income “>50k” is mostly seen in husband category.

```
##         Husband   Not-in-family  Other-relative       Own-child
##           11171            6970             806            4025
##       Unmarried            Wife
##            2895            1278
```

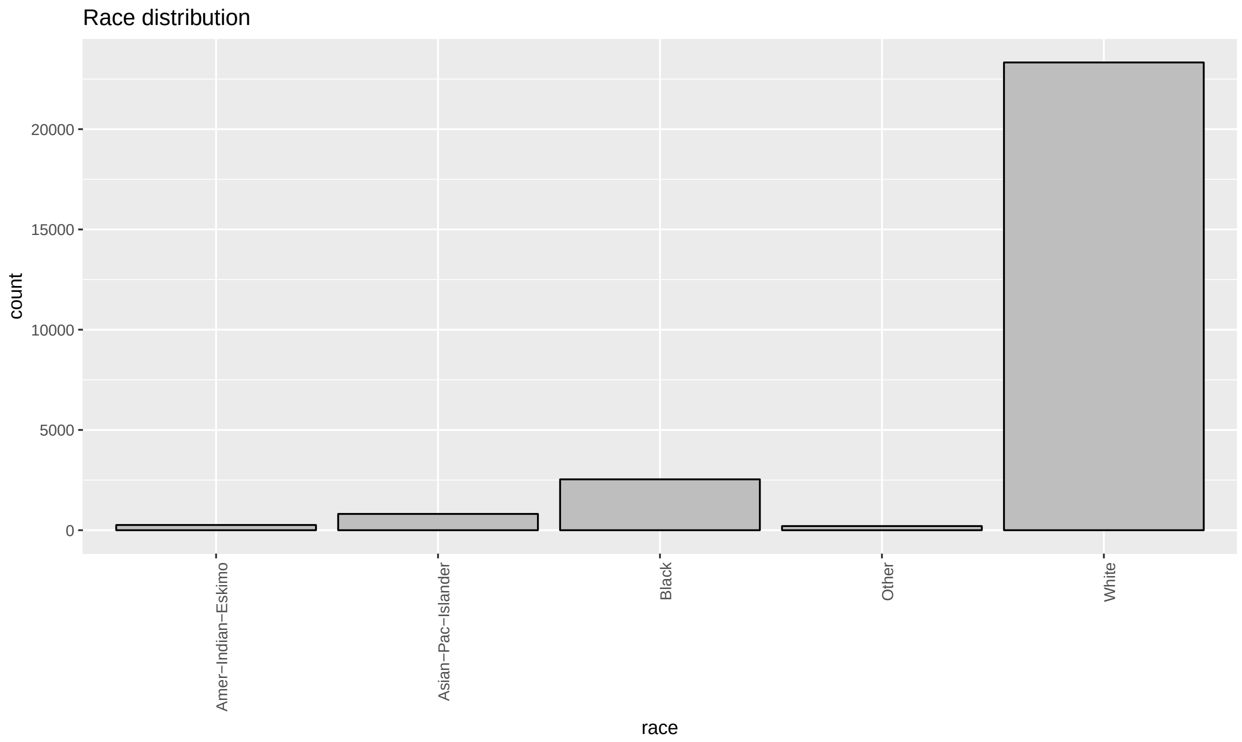


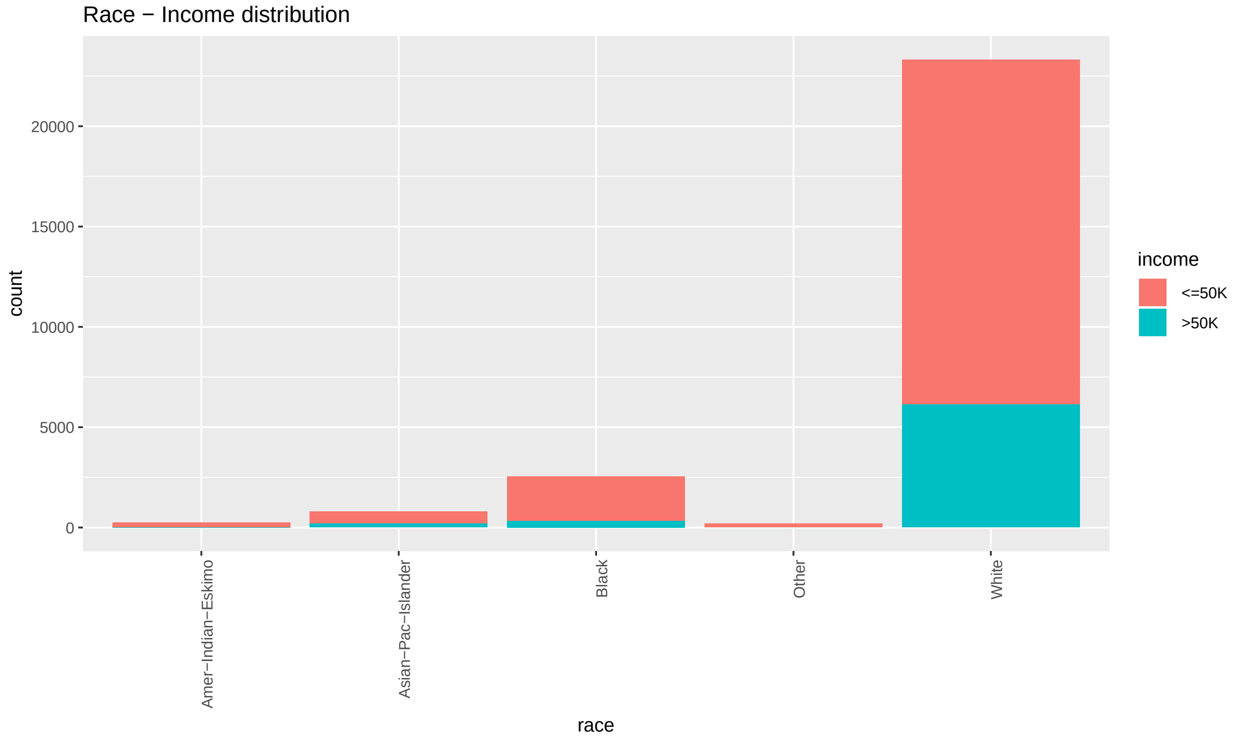


9. Race

Most of the participants belong to “white” category followed by “black”.

```
##  Amer-Indian-Eskimo  Asian-Pac-Islander              Black
##                 260                 813               2537
##               Other               White
##                 206               23329
```

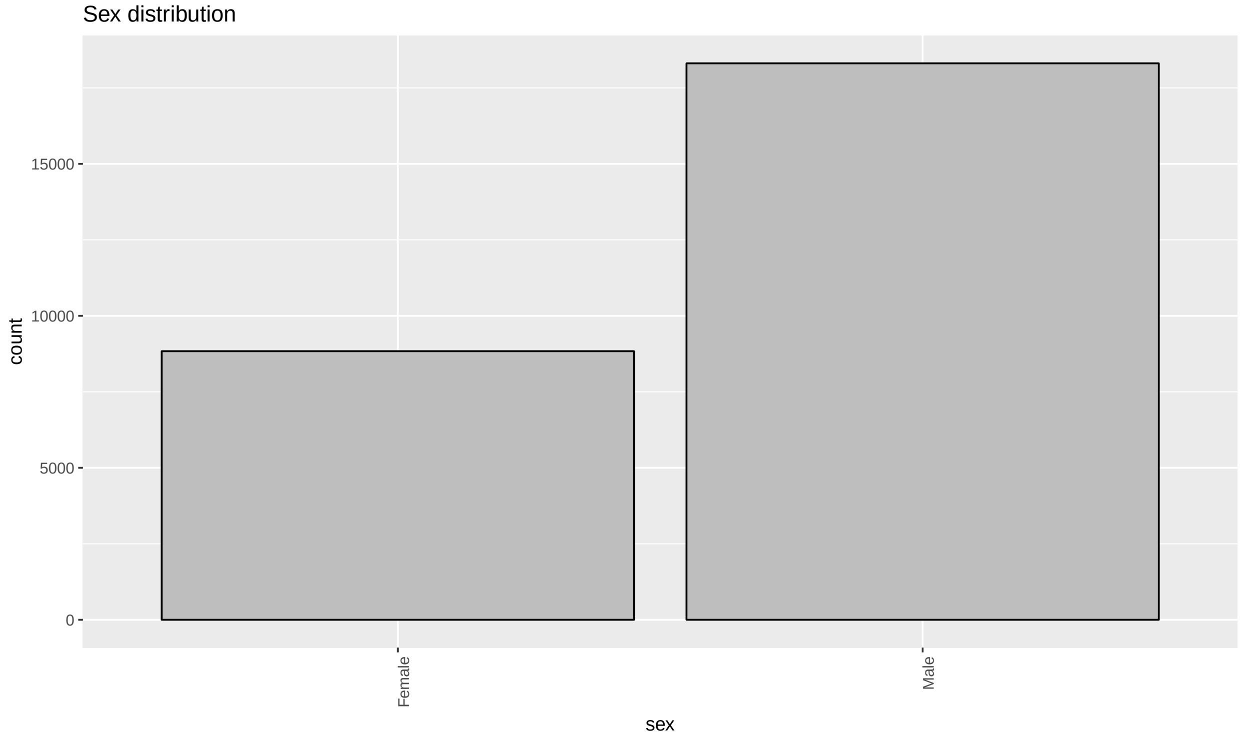


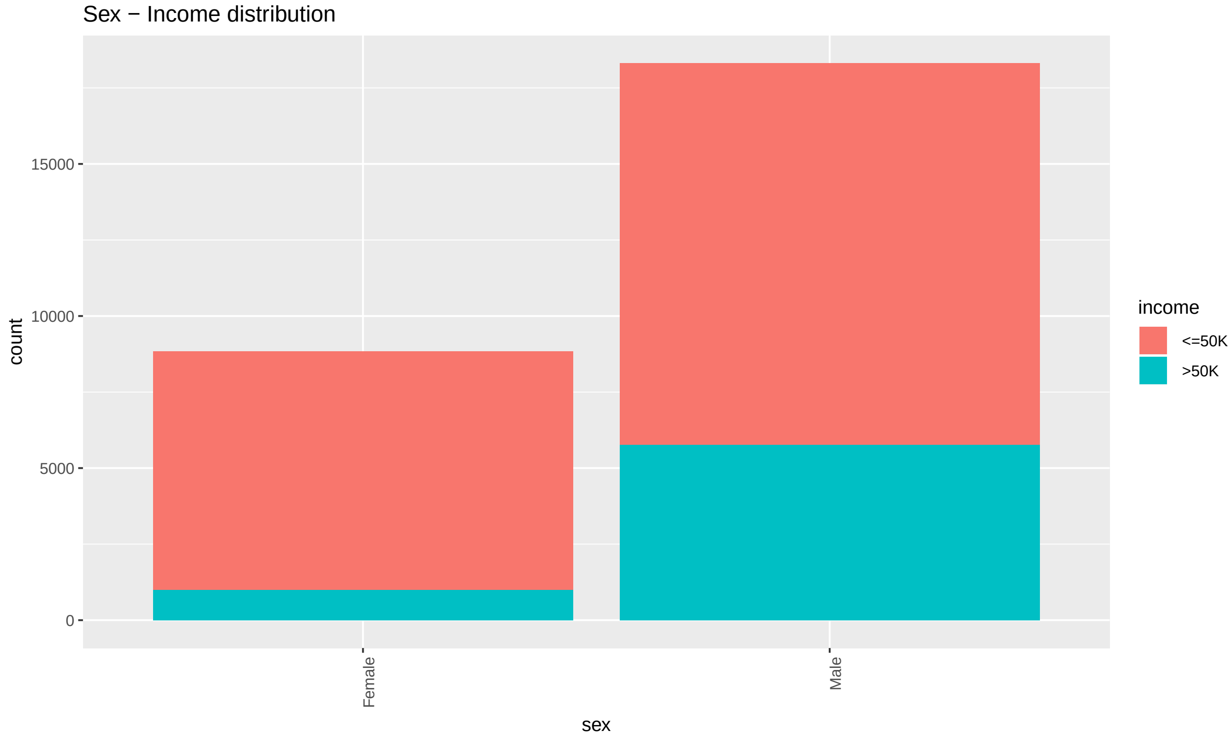


10. Sex

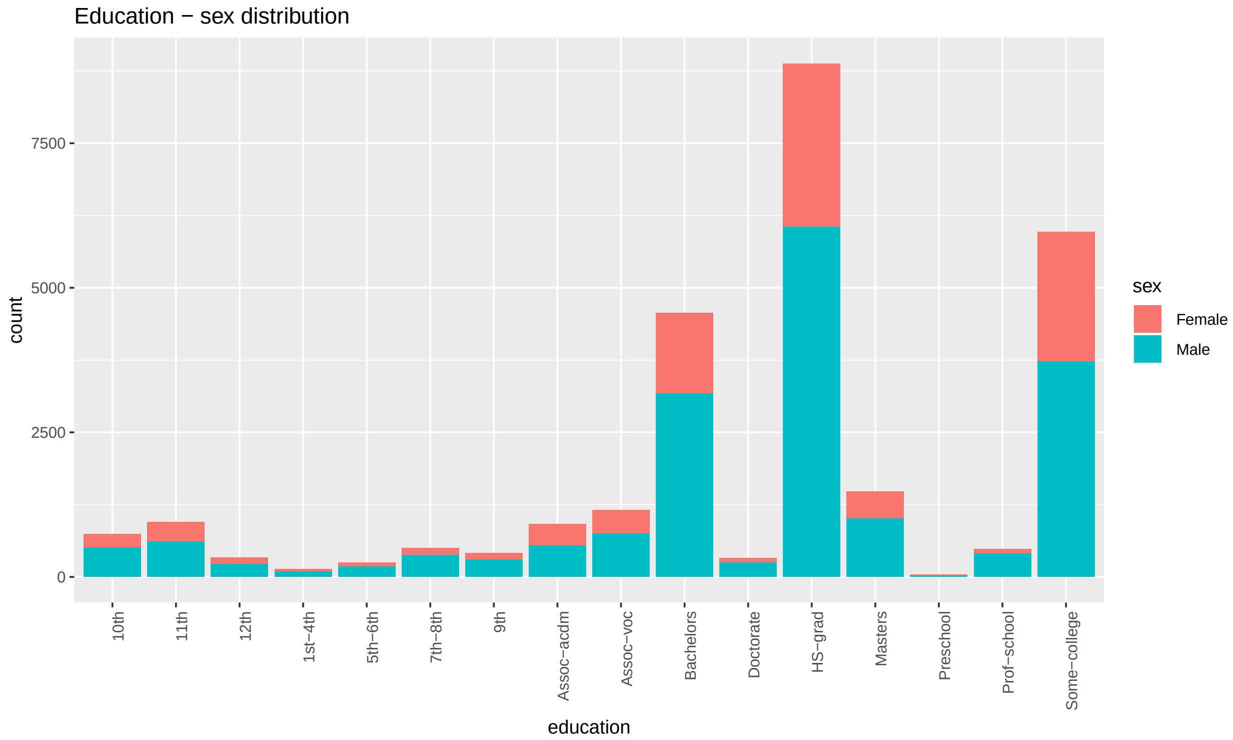
Around 67% are males and 33% are females. No. of females with income ">50k" is less compared to males.

```
## Female   Male
##   8837  18308
```





Further we can observe that no. of males are more educated than females which explains the income distribution difference in males and females.

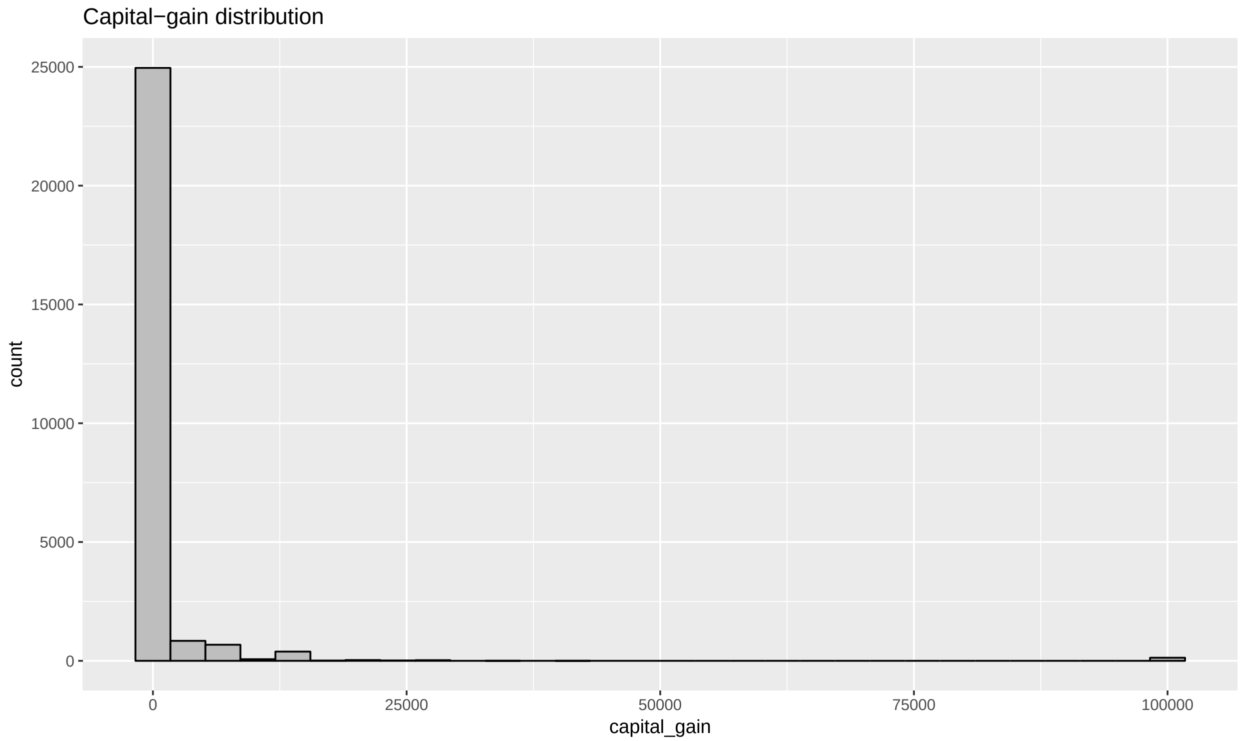


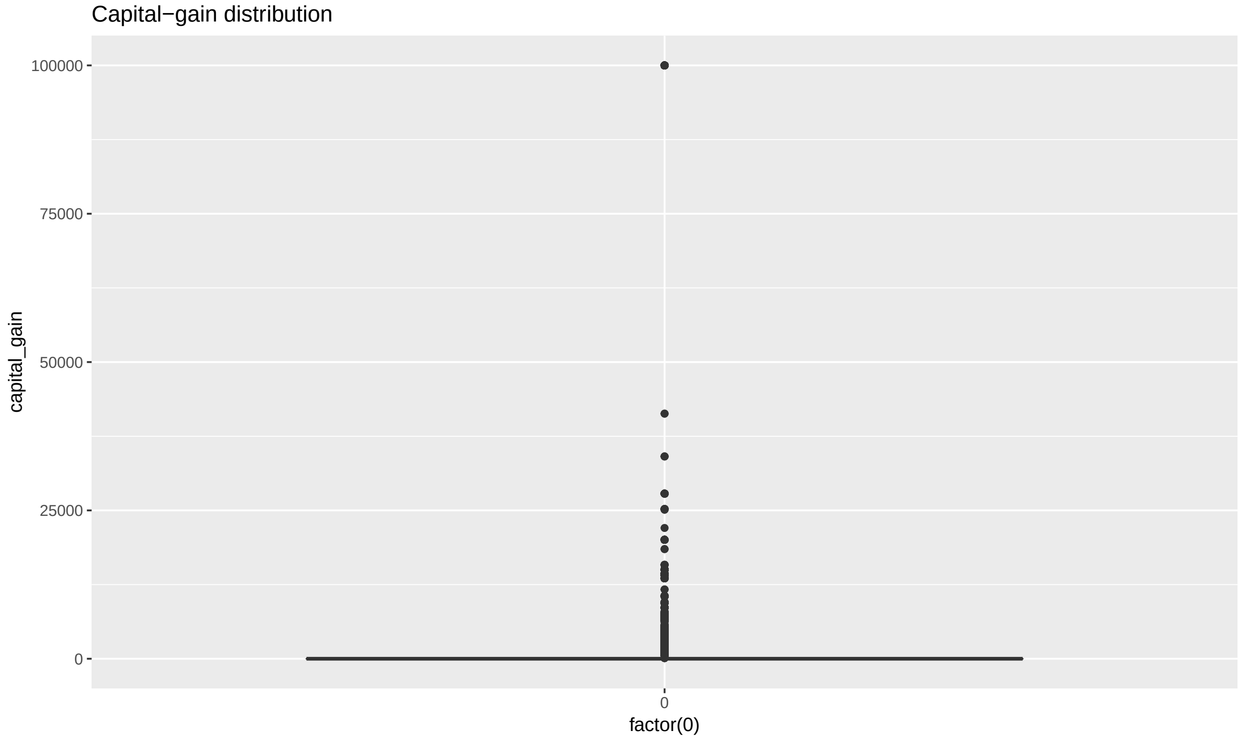
11. Capital Gain and Capital Loss

Capital gain and Capital loss many values as "0" and there are many outliers as well. This is because not everybody invests in stock markets.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0    1081       0   99999
```

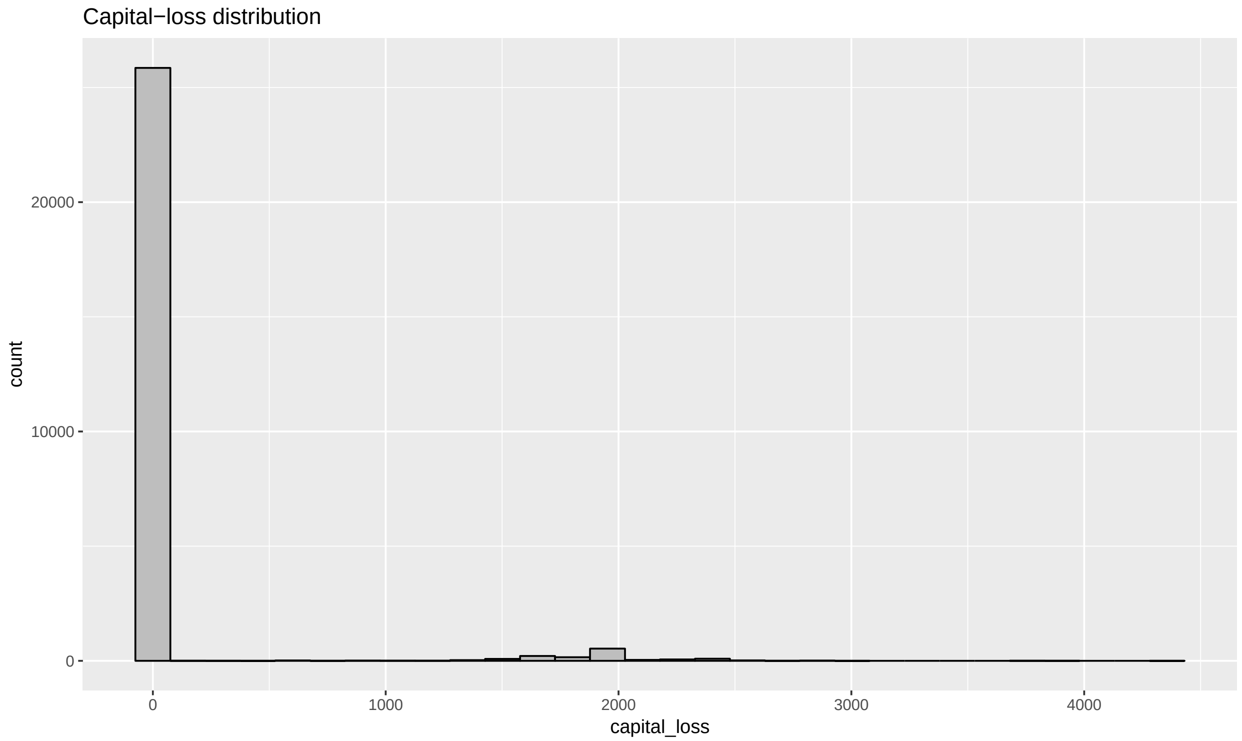
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

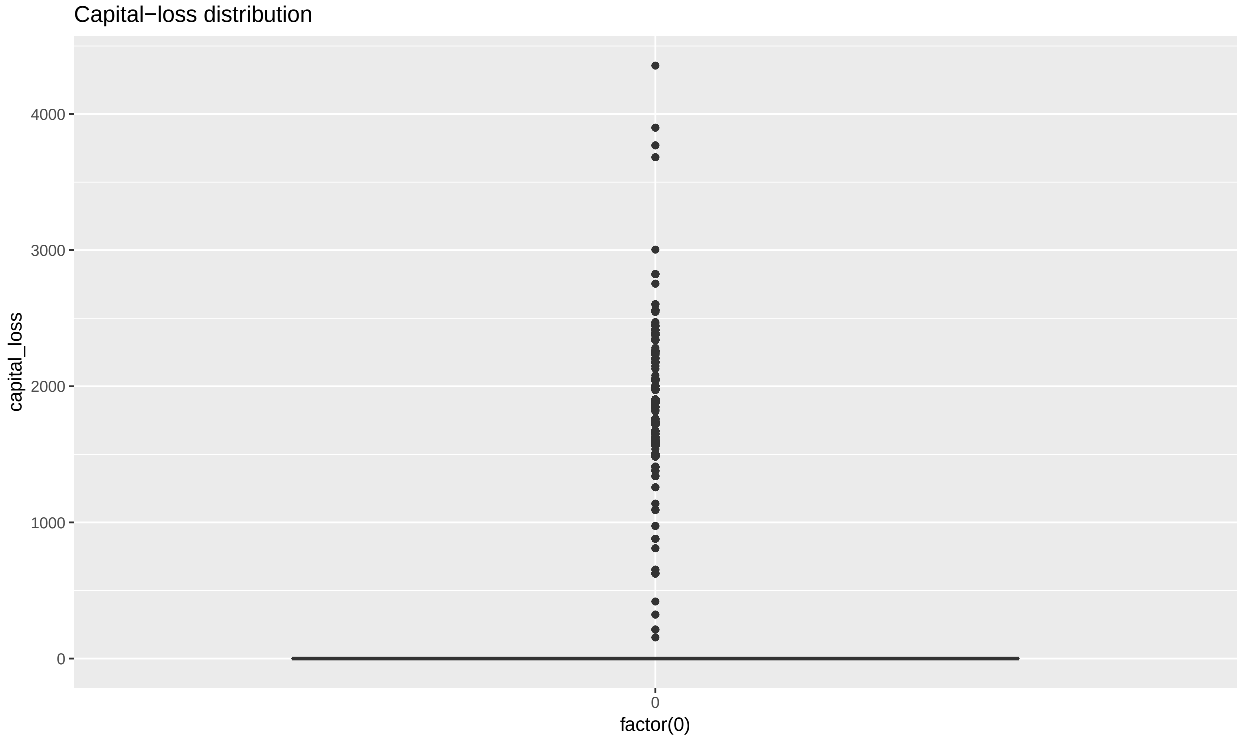




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00   88.89    0.00 4356.00
```

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

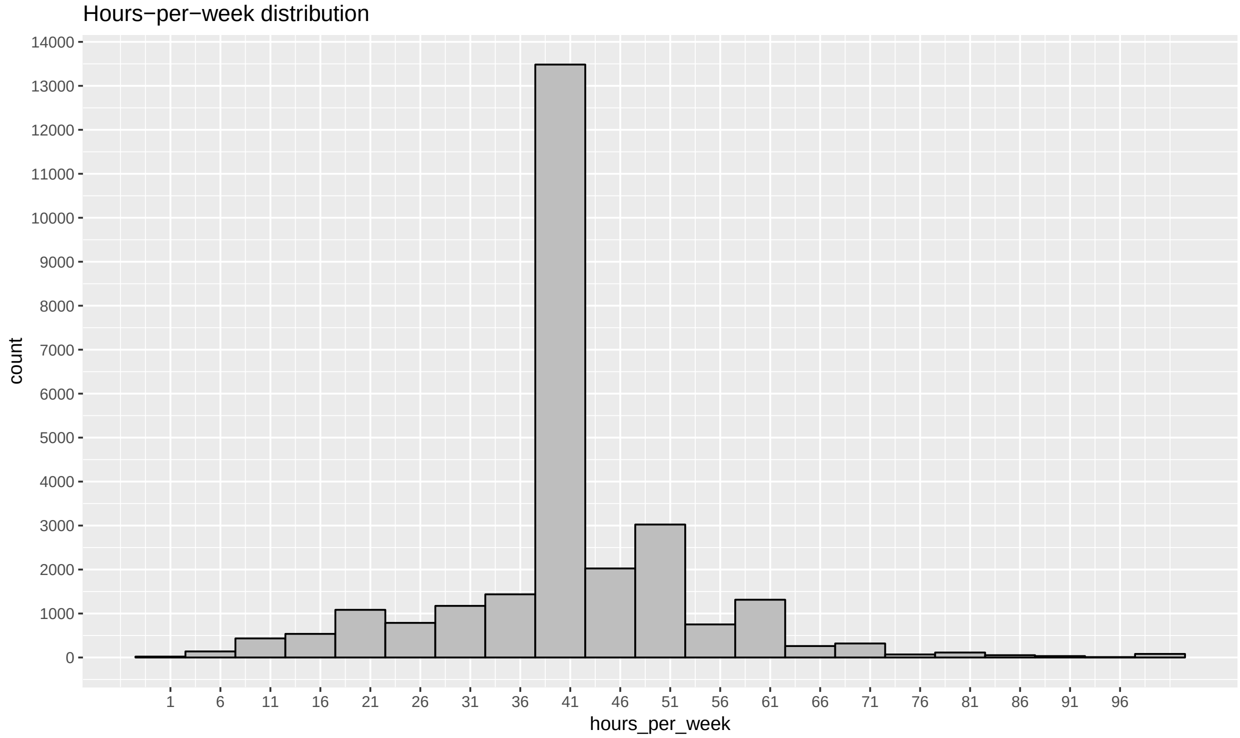


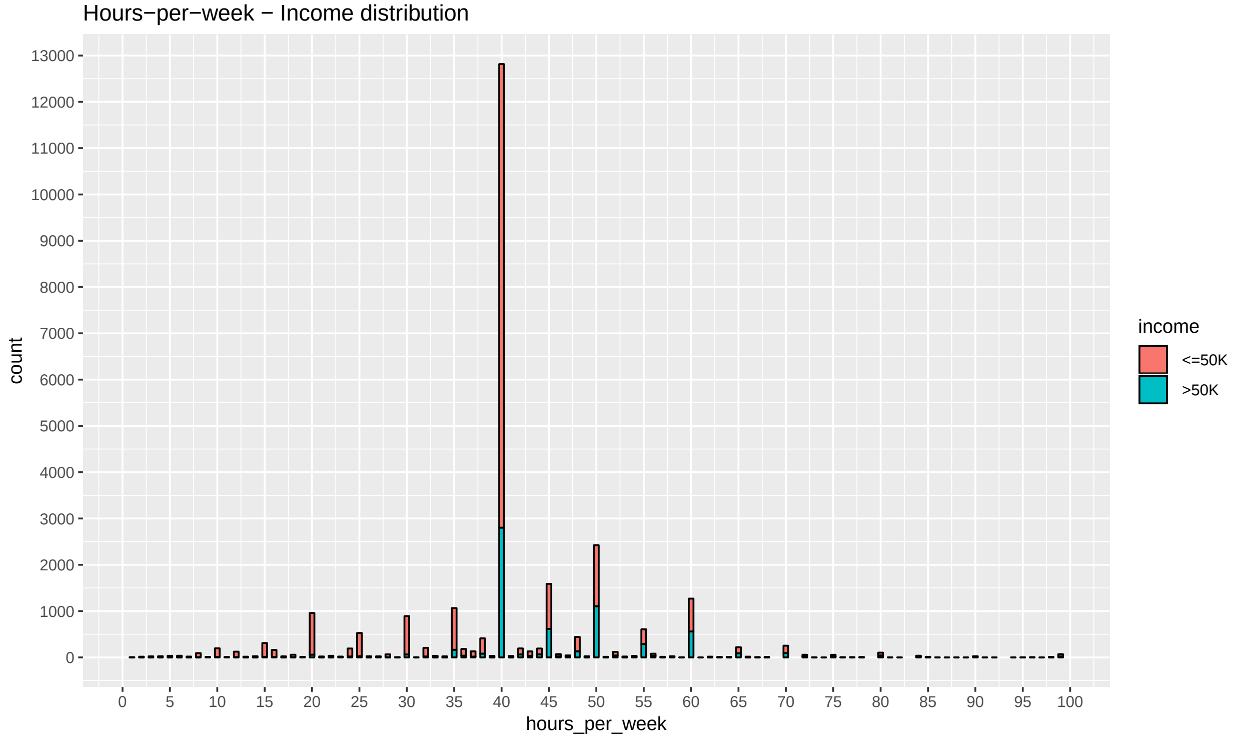


12. Hours-per-week

Minimum working hours is 1hr and maximum working hours is 99hrs. The mean no. of working hours is 40hrs. For income ">50k", the mean working hours is between 45 to 50 hrs.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.00   40.00   40.00   40.92   45.00   99.00
```

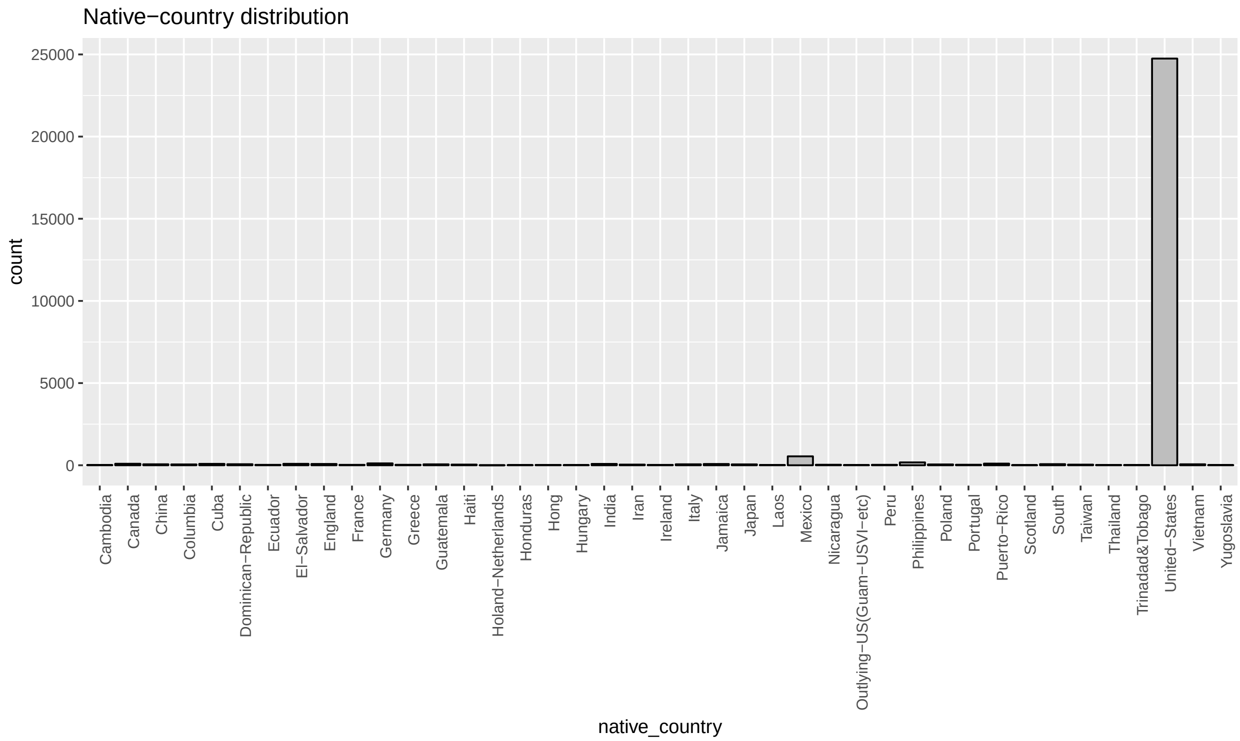


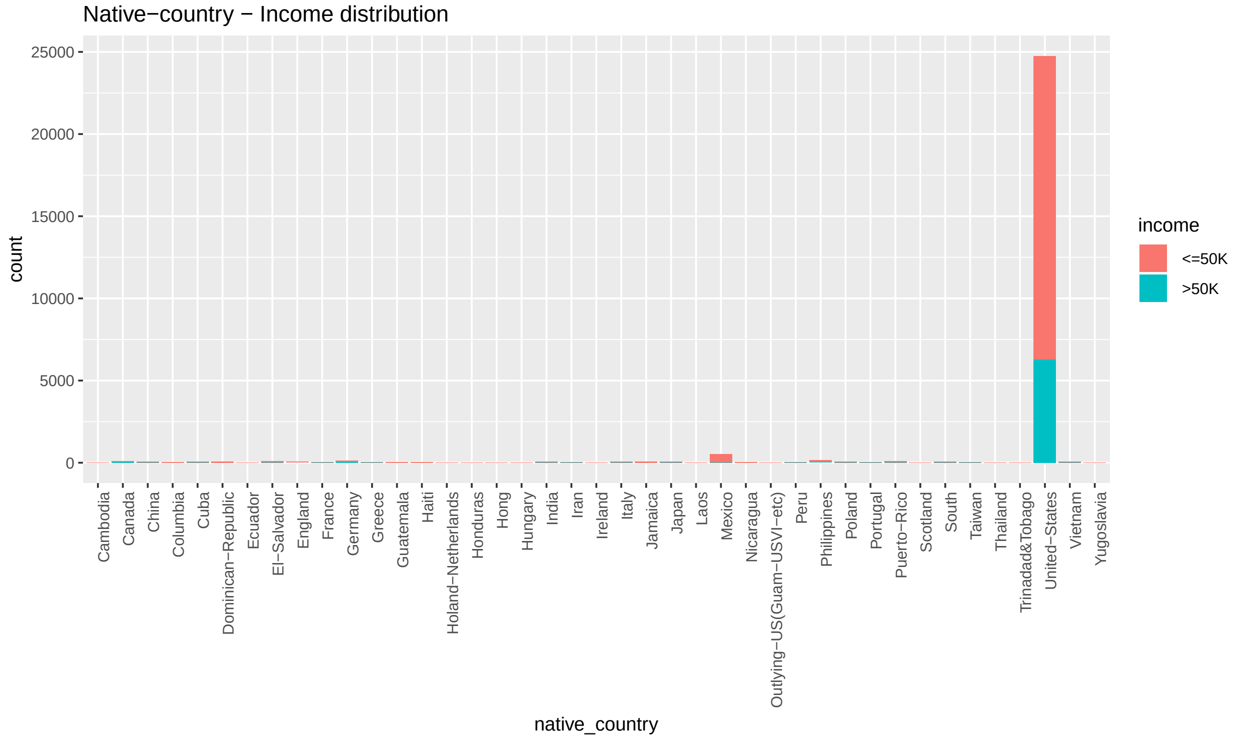


13. Native country

Summary shows that about 91% of the population belongs to United States.

```
##                     Cambodia                          Canada
##                           14                          91
##                        China                    Columbia
##                           59                          52
##                         Cuba          Dominican-Republic
##                           84                          63
##                      Ecuador                  El-Salvador
##                           25                          87
##                      England                      France
##                           78                          25
##                      Germany                      Greece
##                          115                          28
##                    Guatemala                       Haiti
##                           56                          41
##           Holand-Netherlands                    Honduras
##                            1                          11
##                         Hong                     Hungary
##                           18                          12
##                        India                        Iran
##                           81                          40
##                      Ireland                       Italy
##                           22                          58
##                      Jamaica                       Japan
##                           76                          54
##                         Laos                      Mexico
##                           16                         543
##                    Nicaragua  Outlying-US(Guam-USVI-etc)
##                           31                          14
##                         Peru                 Philippines
##                           27                         175
##                       Poland                    Portugal
##                           51                          32
##                  Puerto-Rico                    Scotland
##                          100                           8
##                        South                      Taiwan
##                           67                          38
##                     Thailand             Trinadad&Tobago
##                           17                          16
##                United-States                     Vietnam
##                        24747                          59
##                   Yugoslavia
##                           13
```





14. Income

This is the response variable. About 75% of the sample has income “<=50k”.

```
# 15. Income
# Summary of income
summary(trainset$income)
```

```
##  <=50K   >50K
##  20388   6757
```

```
# To calculate proportions for each factor level
percentage <- 100 * prop.table(table(trainset$income))

# Display in table format
cbind(freq = table(trainset$income), percentage = percentage)
```

```
##         freq percentage
##  <=50K 20388   75.10775
##  >50K   6757   24.89225
```

Data Modeling

We use all the variables except fnlwgt and education_num for data modeling. Algorithms used are as follows.

1. Logistic Regression

The response variable to be predicted is binary. So one of the best candidate algorithm is Logistic Regression.

Logistic Regression is suited for binary classification problems.Real-valued numbers are mapped to a value between 0 and 1 using Logistic Regression. Logistic Regression uses an equation where input values(x) are combined linearly using coefficient values(β) to predict an output value (y). Here, the output is in binary form(0 and 1). Each column used as input has an associated coefficient which is learned from training data.

We have used method as glm(generalized linear model) and family as binomial since this is Logistic Regression.

```
# 1. Logistic Regression

set.seed(1, sample.kind="Rounding")

# Fit data using caret package, method - glm, family - binomial
train_lr <- train(income ~ age + workclass + education + occupation +
                    relationship+ hours_per_week + native_country +
                    race + sex + marital_status + capital_gain + capital_loss,
                  data=trainset,
                  method = "glm",
                  family="binomial")

# Predict income using the above fitted model
pred_lr <- predict(train_lr,validation)

# Save results of Confusion Matrix
lr_acc <- confusionMatrix(pred_lr,validation$income)

# Add results to a table
results <- tibble(Method="Logistic Regression",
                  Accuracy_Train = lr_acc$overall["Accuracy"],
                  F1_Train = lr_acc$byClass[7])
results %>% knitr::kable()
```

Method	Accuracy_Train	F1_Train
Logistic Regression	0.8578058	0.9081567

Accuracy obtained using logistic regression is 0.8578058 and F1 score is 0.9081567.

2. Support Vector Classification

Next we try with Support Vector Classification.

Support vector machine(SVM) algorithm finds a hyperplane in an N-dimensional space where N is the no. of features. This hyperplane distinctly classifies the data points. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between the data points of both classes. The dimension of the hyperplane depends upon the number of features. Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier.These are the points that help us build our SVM.

SVM from caret package takes a long time for computing, therefore we use svm from base package.

```r
# 2. Support Vector Classifier

set.seed(3, sample.kind="Rounding")

## Fit data using svm from base package
svc <- svm(income ~ age + workclass + education + capital_gain +
             occupation + relationship + race + sex + capital_loss +
             hours_per_week + native_country + marital_status,
           data=trainset)

# Predict income using the above fitted model
pred_svc <- predict(svc,validation)

# Save results of Confusion Matrix
cm_svc <- confusionMatrix(pred_svc,validation$income)

# Add results to the results table
results <- bind_rows(results, tibble(Method="Support Vector Classifier",
                                     Accuracy_Train = cm_svc$overall["Accuracy"],
                                     F1_Train = cm_svc$byClass[7]))
results %>% knitr::kable()
```

Method	Accuracy_Train	F1_Train
Logistic Regression	0.8578058	0.9081567
Support Vector Classifier	0.8544912	0.9066950

Accuracy obtained using support vector classification is 0.8544912 and F1 score is 0.9066950.

3. Random Forest Classification

Random forest consists of a large number of individual decision trees. Random Forest is an ensemble model. Each individual tree in the random forest gives a class prediction and the class with the most votes becomes the model's prediction.

Random Forest from caret package takes a long time for computing, therefore we use rf from base package.

```r
# 3. Random Forest Classifier

set.seed(4, sample.kind="Rounding")

## Fit data using rf from base package
raf <- randomForest(income ~ age + workclass + education + capital_gain +
                      occupation + relationship + race + sex + capital_loss +
                      hours_per_week + native_country + marital_status,
                    data = trainset)

# Predict income using the above fitted model
pred_raf <- predict(raf ,validation)

# Save results of Confusion Matrix
cm_raf <- confusionMatrix(pred_raf,validation$income)

# Add results to the results table
results <- bind_rows(results, tibble(Method="Random Forest Classifier",
                                     Accuracy_Train = cm_raf$overall["Accuracy"],
                                     F1_Train = cm_raf$byClass[7]))
results %>% knitr::kable()
```

Method	Accuracy_Train	F1_Train
Logistic Regression	0.8578058	0.9081567
Support Vector Classifier	0.8544912	0.9066950
Random Forest Classifier	0.8674180	0.9136069

Accuracy obtained using random forest classification is 0.8674180 and F1 score is 0.9136069. Here we see an improvement over logistic regression and svm.

4. Gradient Boosting Classifier

We use boosting method to convert weak learners into strong leaders. Here a new tree is fit based on the previous tree with modified weights. Gradient Boosting trains many models in a gradual, additive and sequential manner.

Accuracy obtained using gradient boosting classification is 0.8700696 and F1 score is 0.9169492. Here we see an improvement over logistic regression, svm and random forest classification.

Results

Now that we have trained our models, we apply them to test data. As noted earlier test data is available here: [https://archive.ics.uci.edu/ml/datasets/Adult] in adult.test file.

We follow the same preprocessing steps followed for train data set. The results are tabulated in the form of a table for easy comparison between models.

```r
#________________________________________________
# Part 4: Results / Check the models on test set
#________________________________________________

# Downloading and Reading test data
```

```
if(!file.exists("adult.test")){
  download.file("http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.test","adult.test
}
income_test <- read.csv("adult.test",header = FALSE,skip=1)

# Set column names
colnames(income_test) <- c("age","workclass","fnlwgt","education","education_num","marital_status",
                           "occupation","relationship","race","sex","capital_gain","capital_loss",
                           "hours_per_week","native_country","income")

# explore test data

# To check the structure of test data
str(income_test)
```

```
## 'data.frame':    16281 obs. of  15 variables:
##  $ age           : int  25 38 28 44 18 34 29 63 24 55 ...
##  $ workclass     : Factor w/ 9 levels " ?"," Federal-gov",..: 5 5 3 5 1 5 1 7 5 5 ...
##  $ fnlwgt        : int  226802 89814 336951 160323 103497 198693 227026 104626 369667 104996 ...
##  $ education     : Factor w/ 16 levels " 10th"," 11th",..: 2 12 8 16 16 1 12 15 16 6 ...
##  $ education_num : int  7 9 12 10 10 6 9 15 10 4 ...
##  $ marital_status: Factor w/ 7 levels " Divorced"," Married-AF-spouse",..: 5 3 3 3 5 5 5 3 5 3 ...
##  $ occupation    : Factor w/ 15 levels " ?"," Adm-clerical",..: 8 6 12 8 1 9 1 11 9 4 ...
##  $ relationship  : Factor w/ 6 levels " Husband"," Not-in-family",..: 4 1 1 1 4 2 5 1 5 1 ...
##  $ race          : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 3 5 5 3 5 5 3 5 5 5 ...
##  $ sex           : Factor w/ 2 levels " Female"," Male": 2 2 2 2 1 2 2 2 1 2 ...
##  $ capital_gain  : int  0 0 0 7688 0 0 0 3103 0 0 ...
##  $ capital_loss  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hours_per_week: int  40 50 40 40 30 30 40 32 40 10 ...
##  $ native_country: Factor w/ 41 levels " ?"," Cambodia",..: 39 39 39 39 39 39 39 39 39 39 ...
##  $ income        : Factor w/ 2 levels " <=50K."," >50K.": 1 1 2 2 1 1 1 2 1 1 ...
```

```
# To get the dimension of test data
dim(income_test)
```

```
## [1] 16281    15
```

```
# To get summary of test data
summary(income_test)
```

```
##       age                      workclass        fnlwgt
##  Min.   :17.00    Private         :11210   Min.   :  13492
##  1st Qu.:28.00    Self-emp-not-inc: 1321   1st Qu.: 116736
##  Median :37.00    Local-gov       : 1043   Median : 177831
##  Mean   :38.77    ?               :  963   Mean   : 189436
##  3rd Qu.:48.00    State-gov       :  683   3rd Qu.: 238384
##  Max.   :90.00    Self-emp-inc    :  579   Max.   :1490400
##                  (Other)          :  482
##          education     education_num                  marital_status
##   HS-grad     :5283   Min.   : 1.00    Divorced             :2190
##   Some-college:3587   1st Qu.: 9.00    Married-AF-spouse    :  14
##   Bachelors   :2670   Median :10.00    Married-civ-spouse   :7403
##   Masters     : 934   Mean   :10.07    Married-spouse-absent: 210
##   Assoc-voc   : 679   3rd Qu.:12.00    Never-married        :5434
##   11th        : 637   Max.   :16.00    Separated            : 505
##  (Other)      :2491                    Widowed              : 525
##               occupation            relationship
##   Prof-specialty :2032    Husband       :6523
##   Exec-managerial:2020    Not-in-family :4278
##   Craft-repair   :2013    Other-relative: 525
##   Sales          :1854    Own-child     :2513
##   Adm-clerical   :1841    Unmarried     :1679
##   Other-service  :1628    Wife          : 763
##  (Other)         :4893
##                    race             sex         capital_gain
##   Amer-Indian-Eskimo:  159    Female: 5421   Min.   :    0
##   Asian-Pac-Islander:  480    Male  :10860   1st Qu.:    0
##   Black             : 1561                   Median :    0
##   Other             :  135                   Mean   : 1082
##   White             :13946                   3rd Qu.:    0
##                                              Max.   :99999
##
##   capital_loss    hours_per_week        native_country      income
##  Min.   :   0.0   Min.   : 1.00   United-States:14662   <=50K.:12435
##  1st Qu.:   0.0   1st Qu.:40.00   Mexico       :  308   >50K. : 3846
##  Median :   0.0   Median :40.00   ?            :  274
##  Mean   :  87.9   Mean   :40.39   Philippines  :   97
##  3rd Qu.:   0.0   3rd Qu.:45.00   Puerto-Rico  :   70
##  Max.   :3770.0   Max.   :99.00   Germany      :   69
##                                  (Other)       :  801
```

```r
# To explore the levels in each column of test data
sapply(income_test, levels)
```

```
## $age
## NULL
##
## $workclass
## [1] " ?"                " Federal-gov"      " Local-gov"
## [4] " Never-worked"     " Private"          " Self-emp-inc"
## [7] " Self-emp-not-inc" " State-gov"        " Without-pay"
##
## $fnlwgt
## NULL
##
## $education
##  [1] " 10th"         " 11th"         " 12th"         " 1st-4th"
##  [5] " 5th-6th"      " 7th-8th"      " 9th"          " Assoc-acdm"
##  [9] " Assoc-voc"    " Bachelors"    " Doctorate"    " HS-grad"
## [13] " Masters"      " Preschool"    " Prof-school"  " Some-college"
##
## $education_num
## NULL
##
## $marital_status
## [1] " Divorced"              " Married-AF-spouse"
## [3] " Married-civ-spouse"    " Married-spouse-absent"
## [5] " Never-married"         " Separated"
## [7] " Widowed"
##
```

```
## $occupation
##  [1] " ?"                 " Adm-clerical"      " Armed-Forces"
##  [4] " Craft-repair"      " Exec-managerial"   " Farming-fishing"
##  [7] " Handlers-cleaners" " Machine-op-inspct" " Other-service"
## [10] " Priv-house-serv"   " Prof-specialty"    " Protective-serv"
## [13] " Sales"             " Tech-support"      " Transport-moving"
##
## $relationship
## [1] " Husband"        " Not-in-family"  " Other-relative" " Own-child"
## [5] " Unmarried"      " Wife"
##
## $race
## [1] " Amer-Indian-Eskimo" " Asian-Pac-Islander" " Black"
## [4] " Other"              " White"
##
## $sex
## [1] " Female" " Male"
##
## $capital_gain
## NULL
##
## $capital_loss
## NULL
##
## $hours_per_week
## NULL
##
## $native_country
##  [1] " ?"                          " Cambodia"
##  [3] " Canada"                     " China"
##  [5] " Columbia"                   " Cuba"
##  [7] " Dominican-Republic"         " Ecuador"
##  [9] " El-Salvador"                " England"
## [11] " France"                     " Germany"
## [13] " Greece"                     " Guatemala"
## [15] " Haiti"                      " Honduras"
## [17] " Hong"                       " Hungary"
## [19] " India"                      " Iran"
## [21] " Ireland"                    " Italy"
## [23] " Jamaica"                    " Japan"
## [25] " Laos"                       " Mexico"
## [27] " Nicaragua"                  " Outlying-US(Guam-USVI-etc)"
## [29] " Peru"                       " Philippines"
## [31] " Poland"                     " Portugal"
## [33] " Puerto-Rico"                " Scotland"
## [35] " South"                      " Taiwan"
## [37] " Thailand"                   " Trinadad&Tobago"
## [39] " United-States"              " Vietnam"
## [41] " Yugoslavia"
##
## $income
## [1] " <=50K." " >50K."
```

```r
# Convert " ?" data to NA and the remove rows with NA
income_test <- read.csv("adult.test",na.strings = c(" ?"),header = FALSE,skip=1)
income_test <- na.omit(income_test)

# Set column names again
colnames(income_test) <- c("age","workclass","fnlwgt","education","education_num","marital_status",
                           "occupation","relationship","race","sex","capital_gain","capital_loss",
                           "hours_per_week","native_country","income")

# Assign levels of income to test data
levels(income_test$income)[1] <- " <=50K"
levels(income_test$income)[2] <- " >50K"

# To ensure levels of native_country in train and test data are same
levels(income_test$native_country) <- levels(trainset$native_country)

# 1. Logistic Regression

# Use test data to predict income using the above fitted logistic regression model
pred_lrtest <- predict(train_lr,income_test)

# Save results of Confusion Matrix
lr_test <- confusionMatrix(pred_lrtest,income_test$income)

# Add results to the table
test_results <- tibble(Accuracy_Test = lr_test$overall["Accuracy"],
                       F1_Test = lr_test$byClass[7])
test_results %>% knitr::kable()
```

Accuracy_Test	F1_Test
0.8351262	0.8986985

```r
# 2. Support Vector Classifier

set.seed(3, sample.kind="Rounding")

# Use test data to predict income using the above fitted svm model
pred_svctest <- predict(svc,income_test)

# Save results of Confusion Matrix
svc_test <- confusionMatrix(pred_svctest,income_test$income)

# Add results to the table
test_results <- bind_rows(test_results, tibble(
  Accuracy_Test = svc_test$overall["Accuracy"],
  F1_Test = svc_test$byClass[7]))
test_results %>% knitr::kable()
```

Accuracy_Test	F1_Test
0.8351262	0.8986985
0.8485392	0.9042039

```r
# 3. Random Forest Classifier

set.seed(4, sample.kind="Rounding")

# Use test data to predict income using the above fitted random forest model
pred_raftest <- predict(raf,income_test)

# Save results of Confusion Matrix
raf_test <- confusionMatrix(pred_raftest,income_test$income)

# Add results to the table
test_results <- bind_rows(test_results, tibble(
  Accuracy_Test = raf_test$overall["Accuracy"],
  F1_Test = raf_test$byClass[7]))
test_results %>% knitr::kable()
```

Accuracy_Test	F1_Test
0.8351262	0.8986985
0.8485392	0.9042039
0.8324037	0.8852101

```r
# 4. Gradient Boosting Classifier

set.seed(6, sample.kind="Rounding")

# Use test data to predict income using the above fitted gradient boosting model
pred_gbctest <- predict(gbc,income_test)

# Save results of Confusion Matrix
gbc_test <- confusionMatrix(pred_gbctest,income_test$income)

# Add results to the table
test_results <- bind_rows(test_results, tibble(
  Accuracy_Test = gbc_test$overall["Accuracy"],
  F1_Test= gbc_test$byClass[7]))
test_results %>% knitr::kable()
```

Accuracy_Test	F1_Test
0.8351262	0.8986985
0.8485392	0.9042039
0.8324037	0.8852101
0.8608234	0.9117399

```r
# Add test results to train results table
results <- bind_cols(results,test_results)
results %>% knitr::kable()
```

Method	Accuracy_Train	F1_Train	Accuracy_Test	F1_Test
Logistic Regression	0.8578058	0.9081567	0.8351262	0.8986985
Support Vector Classifier	0.8544912	0.9066950	0.8485392	0.9042039

Method	Accuracy_Train	F1_Train	Accuracy_Test	F1_Test
Random Forest Classifier	0.8674180	0.9136069	0.8324037	0.8852101
Gradient Boosting Classifier	0.8700696	0.9169492	0.8608234	0.9117399

Out of the 4 models used on test data set, gradient boosting classifier has the best accuracy and F1 score followed by logistic regression. But gradient boosting classifier has more computation time compared to other models.

Conclusion

We use adult data set to predict whether income is "<=50k" or ">50k". After downloading data we removed all the rows with missing/unknown values. We explored data and realized final weight and education_num are not useful for the prediction. Therefore, we drop these variables.

Then we fit our data using the above models and on find that gradient boosing classifier has the best performance in this scenario but it also takes a longer time for the comptation.

The next comparable model is logistic regression which has better results along with lesser computaion time.

I would like to explore more on parameter tuning and also see if any processing can be done on the different levels of each variable. Maybe variables like capital gain and caplital loss can be fine tuned further and we can achieve even higher accuracy and F1 score.