PREDICTING IF INCOME EXCEEDS $50,000 OR NOT

“MSCT32 : DATA SCIENCE AND ANALYTICS”
“PROJECT - TEAM 03”

Beulah Evanjalin, Aleena Prakash, Sridurga, Gundeti Veena, Dhavakumar

07 November 2020

Contents

1 Problem Statement:

To build a model that will predict if the income of any individual in the US is greater than or less than USD 50,000 based on the data available about that individual.

2 Objective:

To perform data analytics and machine-learning techniques on the given dataset "Project 3 adult.csv" to predict whether an individual's annual income exceeds $50,000 or not based on the given attributes like Age, Work-class, Final-weight, Education, Education-num (Number of years of education), Marital-status, Occupation, Relationship, Race, Sex, Capital-gain, Capital-loss, Hours-per-week and Native-country.

3 Introduction

To start off, the working directory should have to be set to the location we have our dataset "Project 3 adult.csv" and where we want our program to be saved.

```
# clearing out my work-space
rm(list = ls())
```

```
# reading the current working directory
getwd()
```

```
## [1] "/home/beulah/Data_Science/Project"
```

```
# set working directory
# setwd("./Data_Science/Project")
# loading in the dataset
adult <- read.table(file = "Project_3_adult.csv",
                    header = TRUE, sep = ",", stringsAsFactors = TRUE)
```

4 Exploratory Analysis

Exploratory data analysis or "EDA" is the first and foremost step in analyzing the data to summarize their main characteristics from an experiment.

4.1 About the Dataset

```
dim(adult)
```

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

Our given dataset "Project_3_adult.csv" contains 32,561 entries with a total of 15 columns representing different attributes of the people.
The list of attributes is as follows:

- **age**: the age of an individual
- **workclass**: represents the employment status of an individual
- **fnlwgt**: the number of people in the target population that the corresponding individual represents
- **education**: the highest level of education achieved by an individual
- **education-num**: the number of years of education in total
- **marital.status**: marital status of an individual
- **occupation**: the general type of occupation of an individual
- **relationship**: describes what this individual is relative to others
- **race**: descriptions of an individual's race
- **sex**: the biological sex of the individual
- **capital.gain**: income from investment sources other than salary which is a gain for an individual
- **capital.loss**: income from investment sources other than salary which is a loss for an individual
- **hours.per.week**: the hours an individual has reported to work per week
- **native.country**: country of origin for an individual
- **income**: whether or not an individual makes more than $50,000 annually

```r
# Returns the first few parts of our  data frame.
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
# Displaying compactly the internal structure of our adult dataset.
str(adult)
```

```
## 'data.frame':    32561 obs. of  15 variables:
##  $ age           : int  90 82 66 54 41 34 38 74 68 41 ...
##  $ workclass     : Factor w/ 9 levels "?","Federal-gov",..: 1 5 1 5 5 5 5 8 2 5 ...
##  $ fnlwgt        : int  77053 132870 186061 140359 264663 216864 150601 88638 422013 70037 ...
##  $ education     : Factor w/ 16 levels "10th","11th",..: 12 12 16 6 16 12 1 11 12 16 ...
##  $ education.num : int  9 9 10 4 10 9 6 16 9 10 ...
##  $ marital.status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 7 7 7 1 6 1 6 5 1 5 ...
##  $ occupation    : Factor w/ 15 levels "?","Adm-clerical",..: 1 5 1 8 11 9 2 11 11 4 ...
##  $ relationship  : Factor w/ 6 levels "Husband","Not-in-family",..: 2 2 5 5 4 5 5 3 2 5 ...
##  $ race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 3 5 5 5 5 5 5 5 ...
##  $ sex           : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 2 1 1 2 ...
##  $ 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: Factor w/ 42 levels "?","Cambodia",..: 40 40 40 40 40 40 40 40 40 1 ...
##  $ income        : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 1 2 1 2 ...
```

```
# Producing a summary of all records in our data set
summary(adult)
```

```
##       age                   workclass         fnlwgt
##  Min.   :17.00   Private         :22696   Min.   :  12285
##  1st Qu.:28.00   Self-emp-not-inc: 2541   1st Qu.: 117827
##  Median :37.00   Local-gov       : 2093   Median : 178356
##  Mean   :38.58   ?               : 1836   Mean   : 189778
##  3rd Qu.:48.00   State-gov       : 1298   3rd Qu.: 237051
##  Max.   :90.00   Self-emp-inc    : 1116   Max.   :1484705
##                  (Other)         :  981
##         education     education.num                marital.status
##  HS-grad     :10501   Min.   : 1.00   Divorced             : 4443
##  Some-college: 7291   1st Qu.: 9.00   Married-AF-spouse    :   23
##  Bachelors   : 5355   Median :10.00   Married-civ-spouse   :14976
##  Masters     : 1723   Mean   :10.08   Married-spouse-absent:  418
##  Assoc-voc   : 1382   3rd Qu.:12.00   Never-married        :10683
##  11th        : 1175   Max.   :16.00   Separated            : 1025
##  (Other)     : 5134                   Widowed              :  993
##            occupation          relationship                  race
##  Prof-specialty :4140   Husband       :13193   Amer-Indian-Eskimo:  311
##  Craft-repair   :4099   Not-in-family : 8305   Asian-Pac-Islander: 1039
##  Exec-managerial:4066   Other-relative:  981   Black             : 3124
##  Adm-clerical   :3770   Own-child     : 5068   Other             :  271
##  Sales          :3650   Unmarried     : 3446   White             :27816
##  Other-service  :3295   Wife          : 1568
##  (Other)        :9541
##      sex          capital.gain    capital.loss    hours.per.week
##  Female:10771   Min.   :    0   Min.   :   0.0   Min.   : 1.00
##  Male  :21790   1st Qu.:    0   1st Qu.:   0.0   1st Qu.:40.00
##                 Median :    0   Median :   0.0   Median :40.00
##                 Mean   : 1078   Mean   :  87.3   Mean   :40.44
##                 3rd Qu.:    0   3rd Qu.:   0.0   3rd Qu.:45.00
##                 Max.   :99999   Max.   :4356.0   Max.   :99.00
##
##       native.country    income
##  United-States:29170   <=50K:24720
```

```
## Mexico         :  643   >50K : 7841
## ?              :  583
## Philippines    :  198
## Germany        :  137
## Canada         :  121
## (Other)        : 1709
```

4.2 Data Cleaning

Since the missing values are denoted by a question mark ("?") and also there are no null values (NULL)in any of the columns in our dataset, this seems that our dataset has been pre-processed already.
Missing Values:
So, Missing values are represented by "?" in our dataset.
Firstly, checking how many of those "?"s each column has.

```r
# Number of missing values in each columns
sapply(adult, function(x) sum(x == "?"))
```

```
##            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
```

```r
# Total percentage of missing values
(sum(adult == "?") / nrow(adult))*100
```

```
## [1] 13.08928
```

Therefore, approximately 13% of our given dataset have missing values. And there are only three columns with some missing values. viz.,

- workclass = 1836 missing
- occupation = 1843 missing
- native.country = 583 missing

Since these three variables are qualitative, we refine our dataset to the one with no missing values.

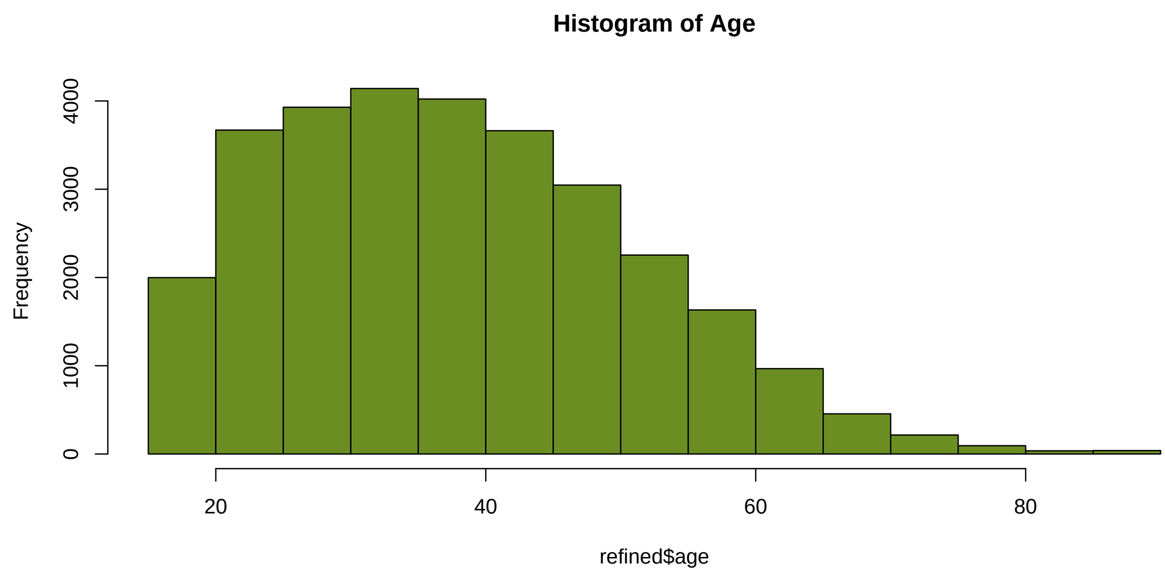
```r
# Dataset with no missing values
refined = adult[apply(adult != "?", 1, all),]
```

4.3 Data Visualization & Data Exploration

```r
# Installation of packages
# Loading the neccessary libraries
installPackage <- function(nameOfThePackage) {
  install.packages(nameOfThePackage)
  library(nameOfThePackage)
  return ("installed")
}
```

4.3.1 Age

```r
hist(refined$age, main = paste("Histogram of Age"), col = "olivedrab")
```




```r
min(refined$age)
```

```
## [1] 17
```

```r
max(refined$age)
```

```
## [1] 90
```

The age feature describes the age of the individual. The figure1 shows the age distribution among the entries in our dataset. The ages range from 17 to 90 years old with the majority of entries lies between the ages of 25 and 45 years.

```r
# if needed, then install the packages
# if already installed, then ignore
if (!require(ggplot2)) do.call("installPackage", args = list("ggplot2"))
if (!require(dplyr)) do.call("installPackage", args = list("dplyr"))
```

```r
# plot(refined$age, refined$income)
agg <- count(refined, age, income)
agg_ord <- mutate(agg, age, income)
ggplot(agg_ord) + geom_col(aes(x = age, y = n, fill = income)) +
  ggtitle('Income Level with Age Level')
```

Figure tells us that the age groups between 17-20, 70-80, and 80-90 has relatively less chance to have an income greater than $50,000.

4.3.2 Workclass

```
ggplot(refined, aes(workclass)) + geom_bar(fill = "olivedrab") +
  ggtitle('Exploring workclass of the individual')
```

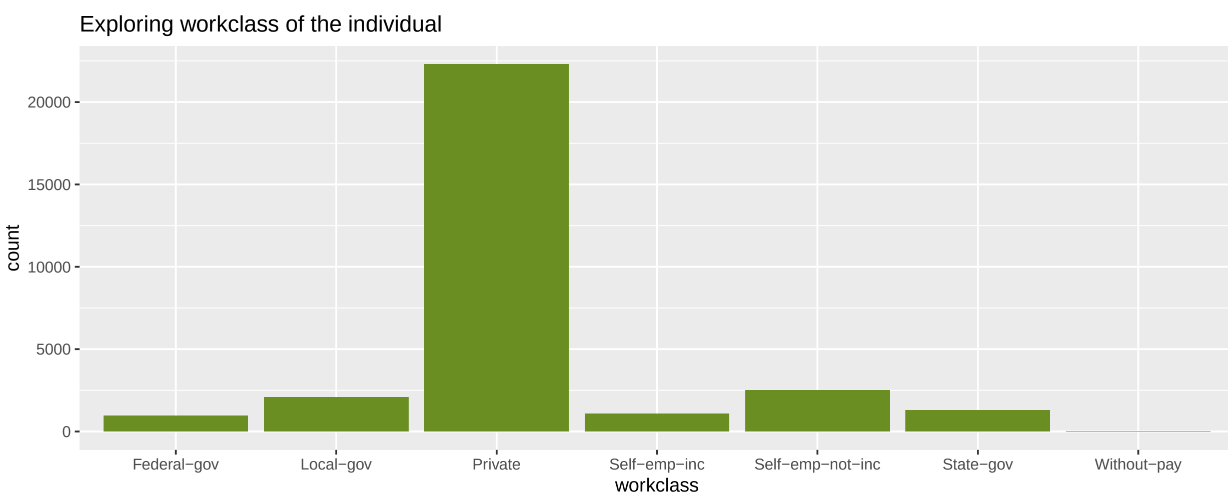


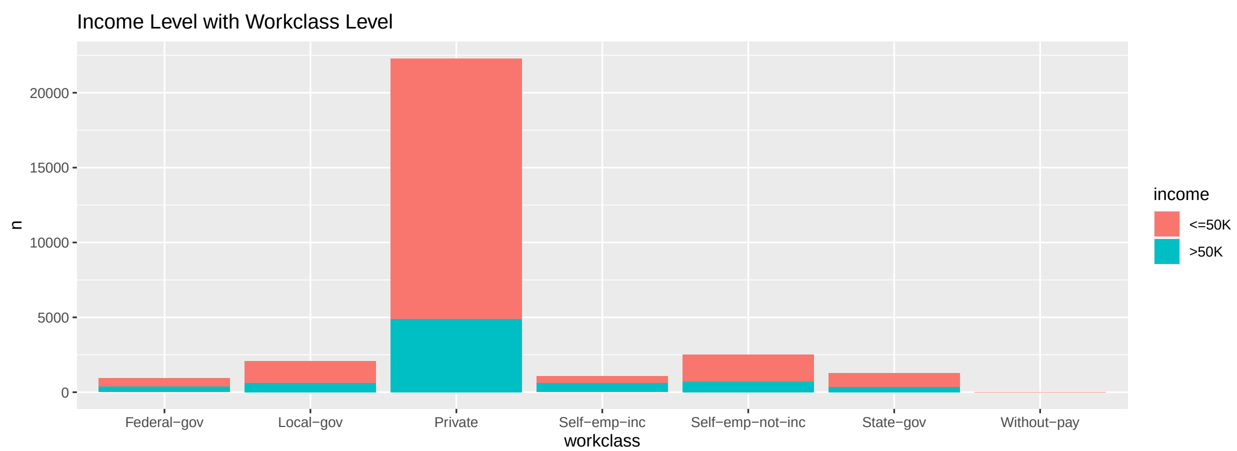
Figure tells us that the majority of the individuals work in the private sector

```
refined %>%
  group_by(workclass) %>%
  summarise(counts = n())
```

```
## # A tibble: 7 x 2
##   workclass        counts
##   <fct>             <int>
## 1 Federal-gov         943
## 2 Local-gov          2067
## 3 Private           22286
```

```
## 4 Self-emp-inc       1074
## 5 Self-emp-not-inc   2499
## 6 State-gov          1279
## 7 Without-pay          14
```

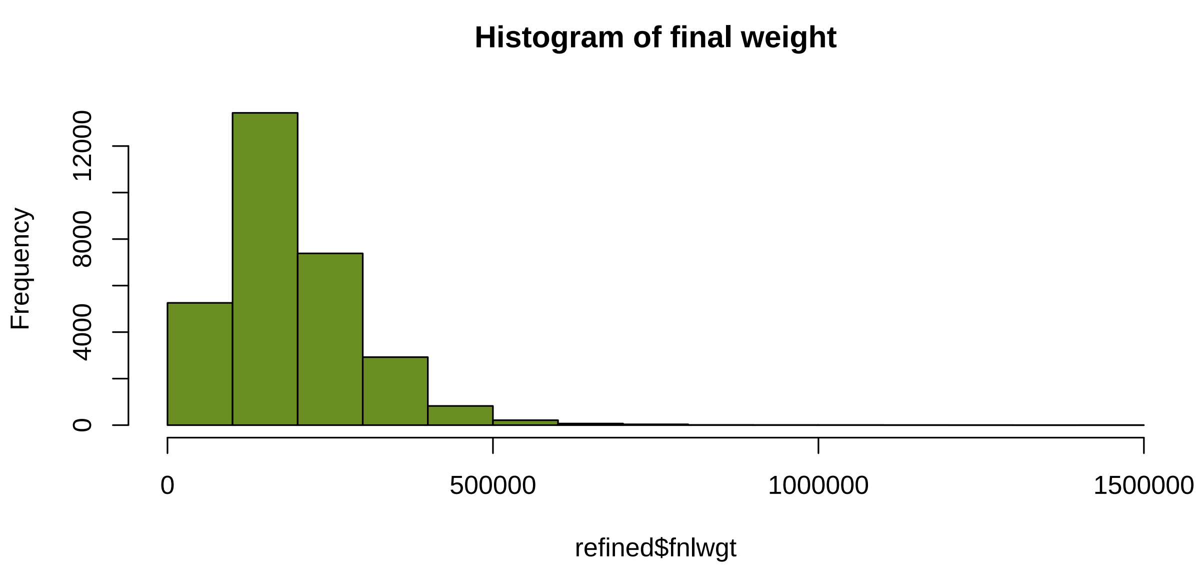
```
agg <- count(refined,workclass,income)
agg_ord <- mutate(agg, workclass, income = reorder(income, -n, sum))
ggplot(agg_ord) + geom_col(aes(x=workclass,y=n,fill=income)) +
  ggtitle('Income Level with Workclass Level')
```



From Figure the probabilities of making above $50,000 are similar among the work classes except for self.emp.inc and federal government.

4.3.3 Final weight

```
hist(refined$fnlwgt, main = paste("Histogram of final weight"), col = "olivedrab" )
```



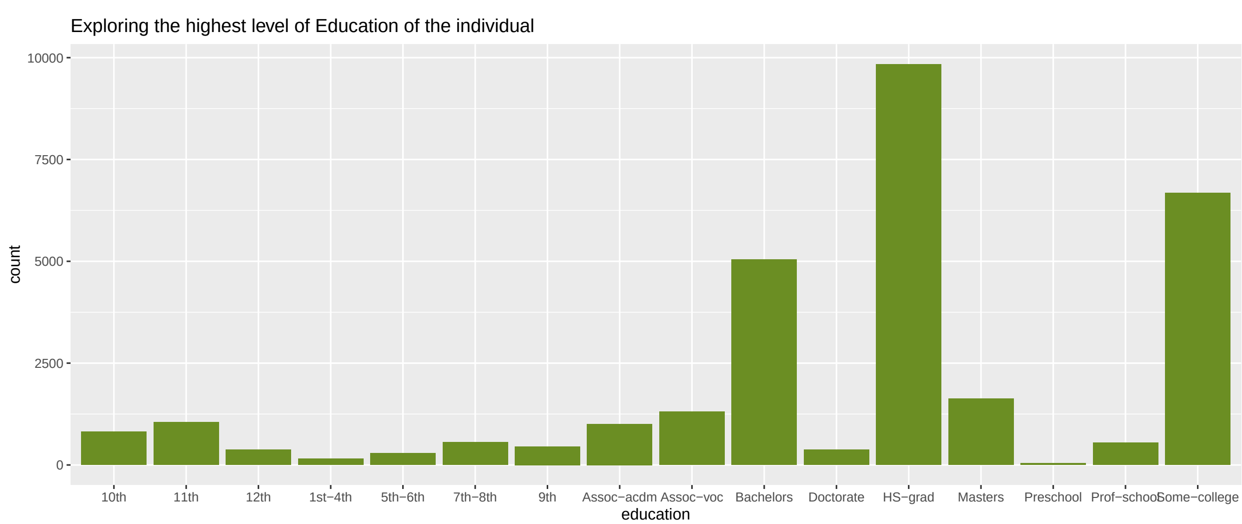
```r
refined %>%
  group_by(fnlwgt) %>%
  summarise(counts = n())
```

```
## # A tibble: 20,263 x 2
##    fnlwgt counts
##     <int>  <int>
##  1  13769      1
##  2  14878      1
##  3  18827      1
##  4  19214      1
##  5  19302      5
##  6  19395      2
##  7  19410      1
##  8  19491      1
##  9  19520      1
## 10  19700      1
## # ... with 20,253 more rows
```

4.3.4 Education

The education feature describes the highest level of education of each individual in the dataset.

```r
ggplot(refined, aes(education)) +
  geom_bar(fill = "olivedrab") +
  ggtitle('Exploring the highest level of Education of the individual')
```




```r
refined %>%
  group_by(education) %>%
  summarise(counts = n())
```

```
## # A tibble: 16 x 2
##    education    counts
##    <fct>         <int>
```

```
##   1 10th           820
##   2 11th          1048
##   3 12th           377
##   4 1st-4th        151
##   5 5th-6th        288
##   6 7th-8th        557
##   7 9th            455
##   8 Assoc-acdm    1008
##   9 Assoc-voc     1307
## 10 Bachelors     5044
## 11 Doctorate      375
## 12 HS-grad       9840
## 13 Masters       1627
## 14 Preschool       45
## 15 Prof-school    542
## 16 Some-college  6678
```

Figure tells us that most of the individuals in the dataset have at most a high school education while only a small portion have a doctorate.

```
agg <- count(refined, education,income)
agg_ord <- mutate(agg, education, income = reorder(income, -n, sum))
ggplot(agg_ord) + geom_col(aes(x = education, y = n,fill = income)) +
  ggtitle('Income Level with Education Level')
```

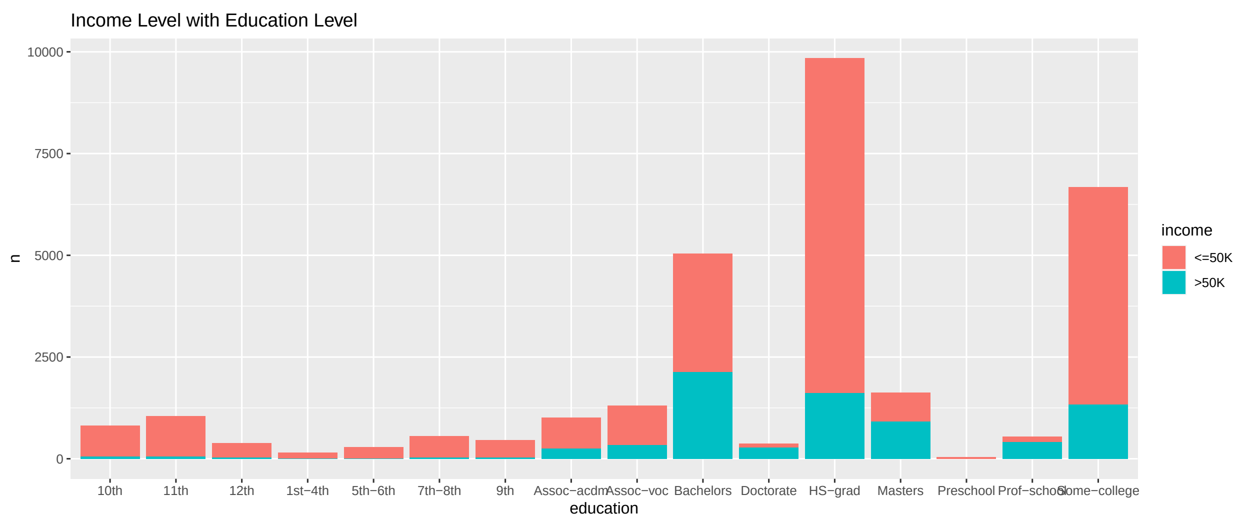
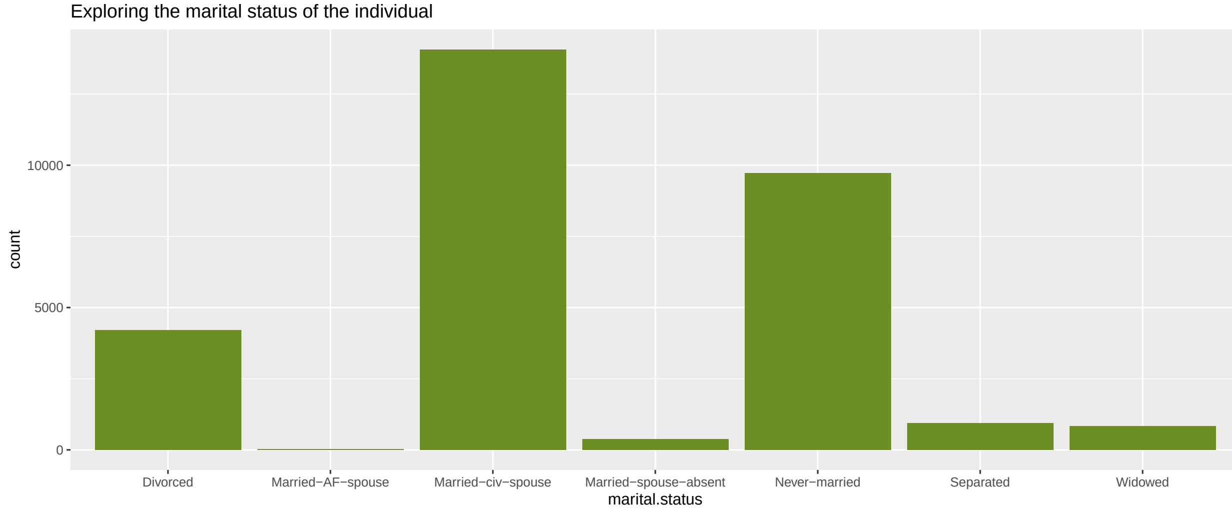


Figure shows the relationship between the highest level of education and the number of people labeled >50k and <=50k.

4.3.5 Marital Status

The education feature describes the highest level of education of each individual in the dataset.

```
ggplot(refined, aes(marital.status)) +
  geom_bar(fill = "olivedrab") +
  ggtitle('Exploring the marital status of the individual')
```




```r
refined %>%
  group_by(marital.status) %>%
  summarise(counts = n())
```

```
## # A tibble: 7 x 2
##   marital.status        counts
##   <fct>                  <int>
## 1 Divorced                4214
## 2 Married-AF-spouse         21
## 3 Married-civ-spouse     14065
## 4 Married-spouse-absent    370
## 5 Never-married           9726
## 6 Separated                939
## 7 Widowed                  827
```

```r
agg <- count(refined, marital.status,income)
agg_ord <- mutate(agg, marital.status, income = reorder(income, -n, sum))
ggplot(agg_ord) + geom_col(aes(x = marital.status, y = n,fill = income)) +
  ggtitle('Income Level with Marital status of the individual')
```

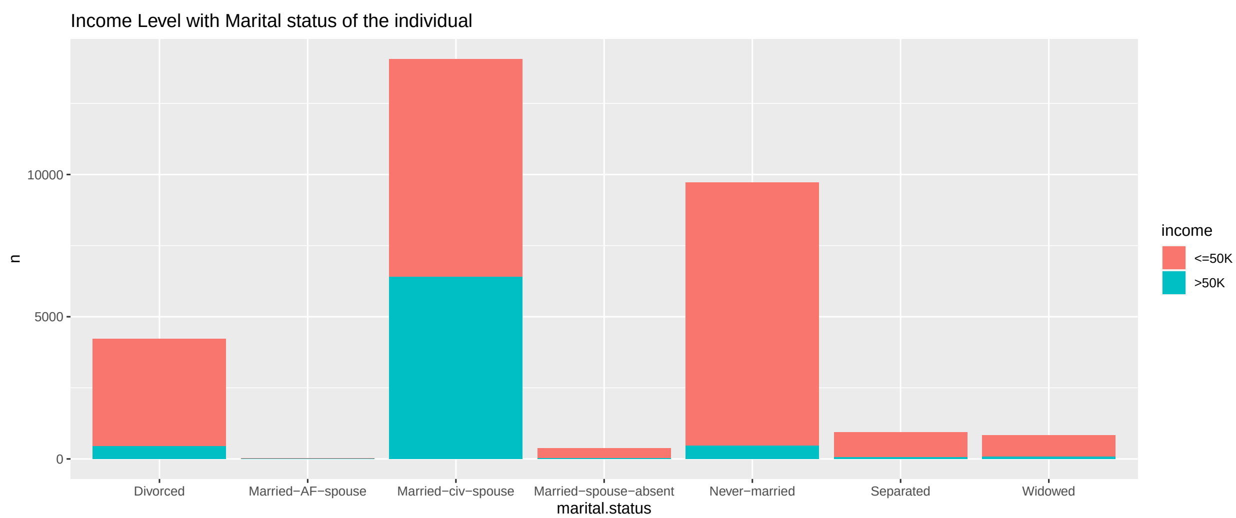


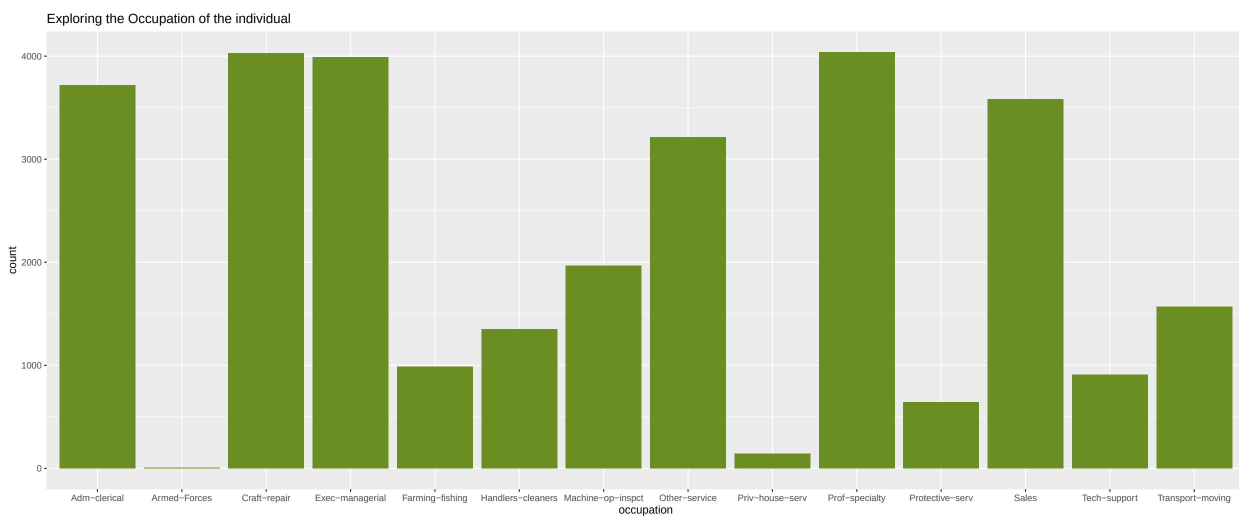
Figure shows the relationship between the marital status and the number of people labeled >50k and <=50k.

4.3.6 Occupation

```
refined %>%
  group_by(occupation) %>%
  summarise(counts = n())
```

```
## # A tibble: 14 x 2
##    occupation        counts
##    <fct>              <int>
##  1 Adm-clerical        3721
##  2 Armed-Forces           9
##  3 Craft-repair        4030
##  4 Exec-managerial     3992
##  5 Farming-fishing      989
##  6 Handlers-cleaners   1350
##  7 Machine-op-inspct   1966
##  8 Other-service       3212
##  9 Priv-house-serv      143
## 10 Prof-specialty      4038
## 11 Protective-serv      644
## 12 Sales               3584
## 13 Tech-support         912
## 14 Transport-moving    1572
```

```
ggplot(refined, aes(occupation)) +
  geom_bar(fill = "olivedrab") +
  ggtitle('Exploring the Occupation of the individual')
```




```
agg <- count(refined, occupation,income)
agg_ord <- mutate(agg, occupation, income = reorder(income, -n, sum))
ggplot(agg_ord) + geom_col(aes(x = occupation, y = n,fill = income)) +
  ggtitle('Income Level with Occupation of the individual')
```

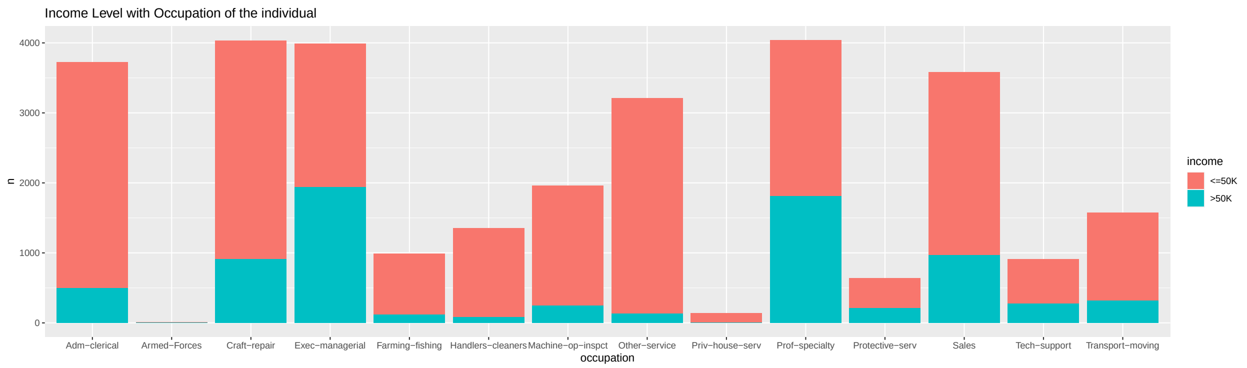


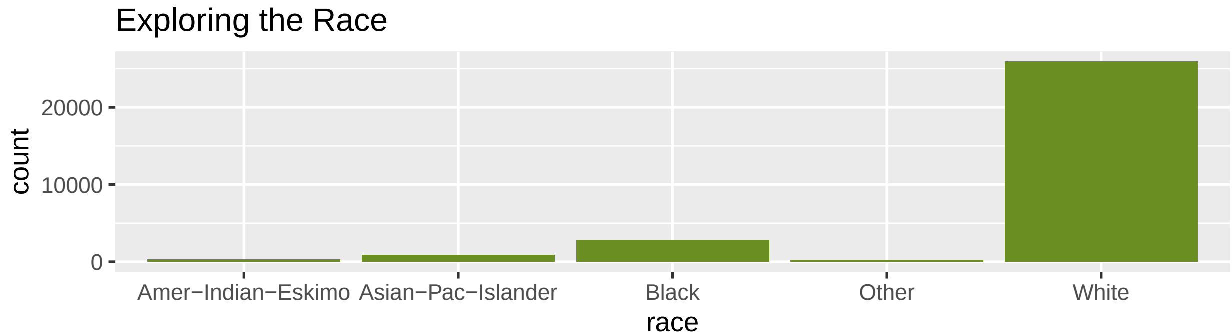
Figure shows the relationship between the occupation and the number of people labeled >50k and <=50k. Also exec.managerial and prof.specialty has a very high percentages of individuals making over $50,000. In addition, the percentages for Farming.fishing, Other.service and Handlers.cleaners are significantly lower than the rest of the distribution.

4.3.7 Race

```
refined %>%
  group_by(race) %>%
  summarise(counts = n())
```

```
## # A tibble: 5 x 2
##   race               counts
##   <fct>               <int>
## 1 Amer-Indian-Eskimo    286
## 2 Asian-Pac-Islander    895
## 3 Black                2817
## 4 Other                 231
## 5 White               25933
```

```
ggplot(refined, aes(race)) +
  geom_bar(fill = "olivedrab") +
  ggtitle('Exploring the Race')
```




```
agg <- count(refined, race,income)
agg_ord <- mutate(agg, race, income = reorder(income, -n, sum))
ggplot(agg_ord) + geom_col(aes(x = race, y = n,fill = income)) +
  ggtitle('Income Level with Race of the individual')
```

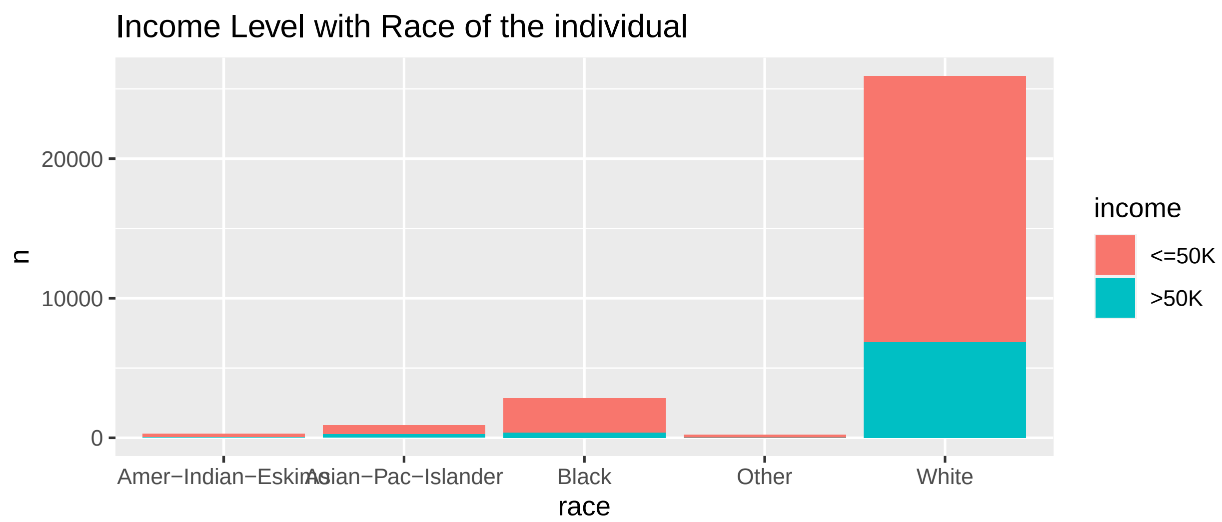
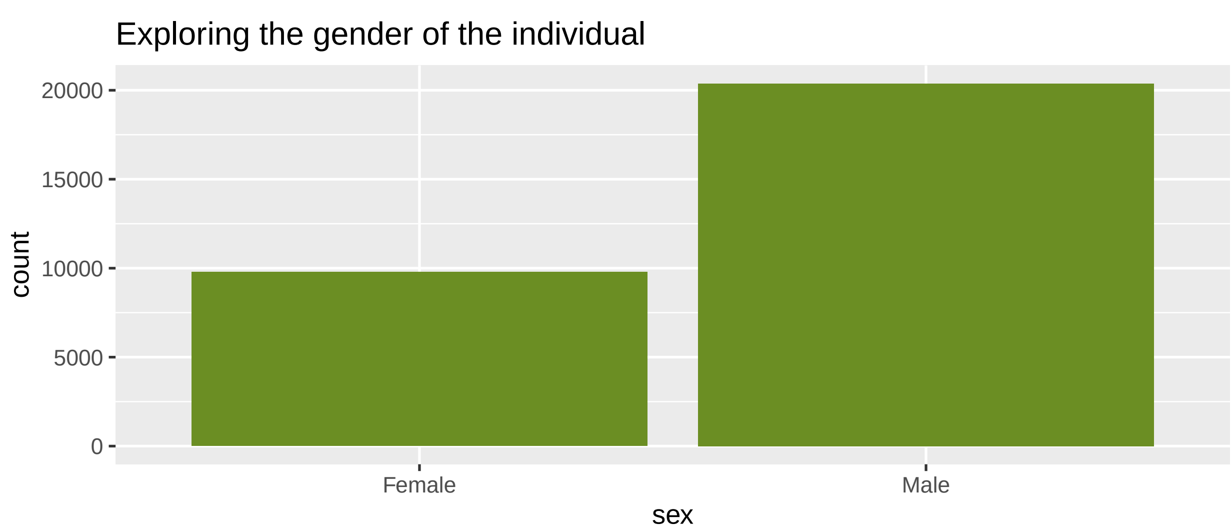


Figure shows the relationship between the race and the number of people labeled >50k and <=50k. It shows that Whites has larger percentage of entries greater than $50,000 than the rest of the races.

4.3.8 Sex

```
ggplot(refined, aes(sex)) +
  geom_bar(fill = "olivedrab") +
  ggtitle('Exploring the gender of the individual')
```




```
refined %>%
  group_by(sex) %>%
  summarise(counts = n())
```

```
## # A tibble: 2 x 2
##   sex    counts
##   <fct>   <int>
## 1 Female   9782
## 2 Male    20380
```

There is almost double the sample size of males in comparison to females in the dataset.

```
agg <- count(refined, sex,income)
agg_ord <- mutate(agg, sex, income = reorder(income, -n, sum))
ggplot(agg_ord) + geom_col(aes(x = sex, y = n,fill = income)) +
  ggtitle('Income Level with Biological Gender of the individual')
```

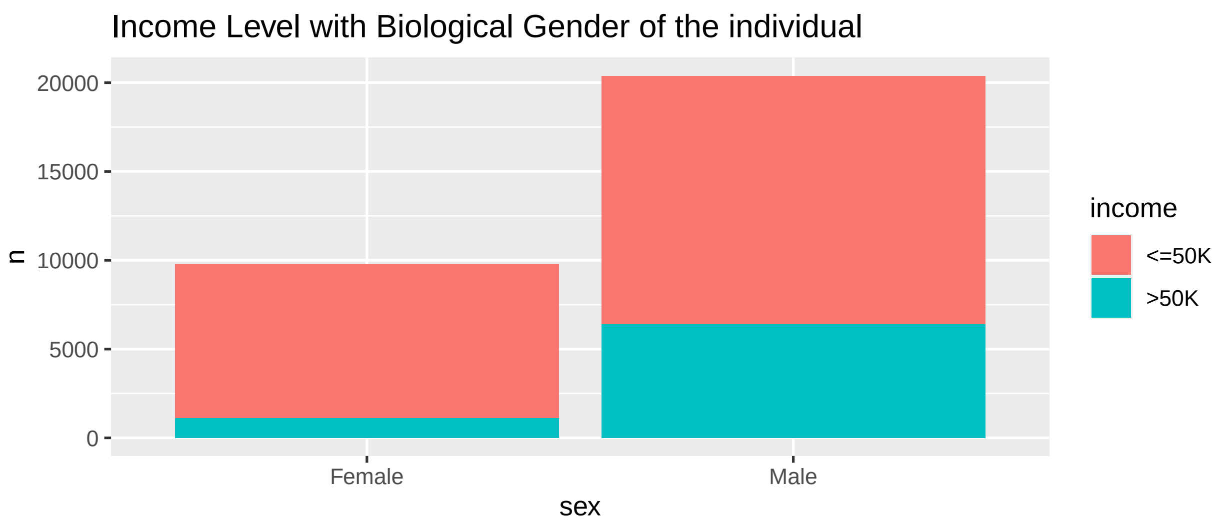


Figure shows the relationship between the sex and the number of people labeled >50k and <=50k. It tell us that the percentage of males who make greater than $50,000 is much greater than the percentage of females that make the same amount.

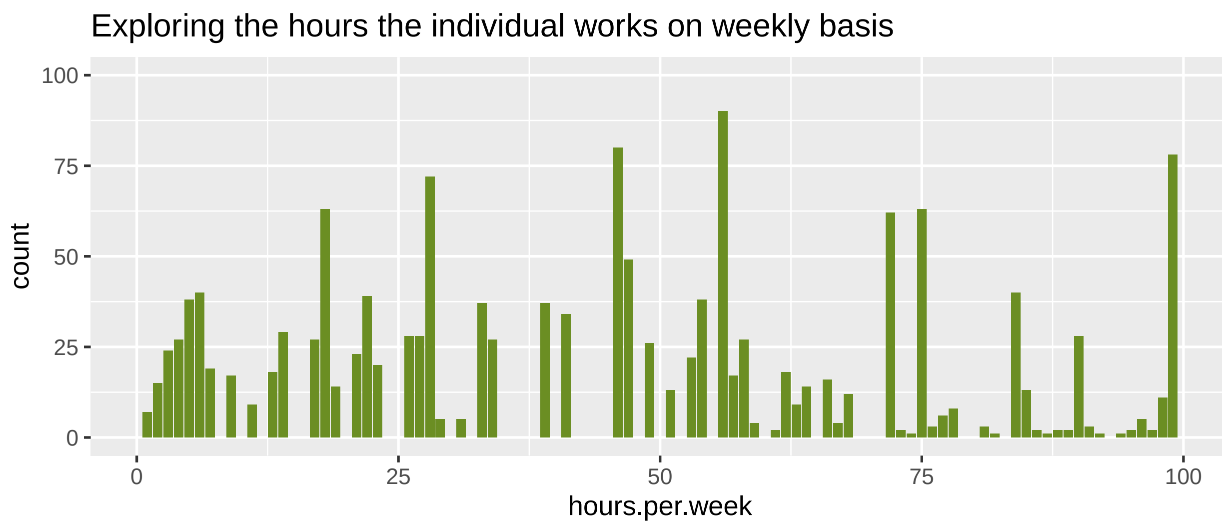
4.3.9 Hours Per Week

```
refined %>%
  group_by(hours.per.week) %>%
  summarise(counts = n())
```

```
## # A tibble: 94 x 2
##    hours.per.week counts
##             <int>  <int>
##  1              1      7
##  2              2     15
##  3              3     24
##  4              4     27
##  5              5     38
##  6              6     40
##  7              7     19
##  8              8    102
##  9              9     17
## 10             10    222
## # ... with 84 more rows
```

The vast majority of individuals are working 40 hourweeks.

```r
ggplot(refined, aes(hours.per.week)) +
  geom_bar(fill = "olivedrab") +
  ggtitle('Exploring the hours the individual works on weekly basis') +
  ylim(0, 100)
```




```r
agg <- count(refined, hours.per.week,income)
agg_ord <- mutate(agg, hours.per.week, income = reorder(income, -n, sum))
ggplot(agg_ord) + geom_col(aes(x = hours.per.week, y = n,fill = income)) +
  ggtitle('Income Level with hours per Week') +
  ylim(0, 100)
```

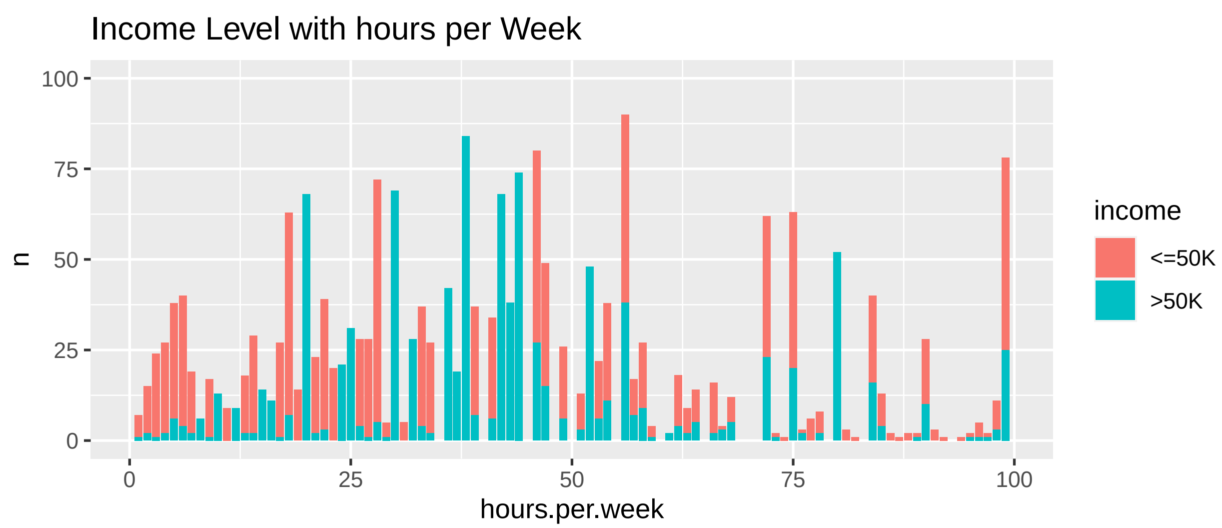


Figure shows the relationship between the sex and the number of people labeled >50k and <=50k. It tell us that the percentage of individuals making over $50,000 drastically decreases when less than 40 hours per week, and increases significantly when greater than 40 hours per week.

5 Building the Model

The purpose of building the model is to classify the people into two groups, below 50k or above 50k in income. We are building the model using training data, and then predicting the salary class using the testing data.

5.1 Train-Test Split

We are using sample.split function to split the dataset into training set and Testing set for our model.

```r
# if needed, then install the packages
# if already installed, then ignore
if (!require(caTools)) do.call("installPackage", args = list("caTools"))
```

```r
set.seed(12345)
# Spliting data for Buildind the Model
SampleSplit <- sample.split(refined$income, SplitRatio = 0.7)
trainingSet <- subset(refined, SampleSplit == TRUE)
testingSet <- subset(refined,SampleSplit == FALSE)
```

5.2 Decision Tree Model

```r
# if needed, then install the packages
# if already installed, then ignore
if (!require(rpart)) do.call("installPackage", args = list("rpart"))
if (!require(rpart.plot)) do.call("installPackage", args = list("rpart.plot"))
```

```r
salaryTree <- rpart(income ~. , data = refined)
```

5.3 Naive Bayes Model

```r
# if needed, then install the packages
# if already installed, then ignore
if (!require(e1071)) do.call("installPackage", args = list("e1071"))
```

```r
salaryNaiveBayes <- naiveBayes(income ~. , data = refined)
```

5.4 Random Forest Model

```r
# if needed, then install the packages
# if already installed, then ignore
if (!require(randomForest)) do.call("installPackage", args = list("randomForest"))
```

```r
salaryForest <- randomForest(income ~. , data = refined)
```

6 Prediction

6.1 For the Decision Tree model

```
# Prediction for the Decision Tree model
PredictIncomeTree <- predict(salaryTree, testingSet, type = 'class')
```

6.2 For the Naive Bayes Model

```
#Prediction for the Naive Bayes model
PredictIncomeNaiveBayes <- predict(salaryNaiveBayes, testingSet, type = 'class')
```

6.3 For the Random Forest Model

```
# Prediction for the Random Forest model
PredictIncomeForest <- predict(salaryForest, testingSet, type = 'class')
```

7 Confusion Matrix

A confusion matrix is a technique for summarizing the performance of a classification algorithm.

```
# if needed, then install the packages
# if already installed, then ignore
if (!require(caret)) do.call("installPackage", args = list("caret"))
```

7.1 For the Decision Tree model

```
#checking The Accuracy of the Decision Tree model
confusionMatrix(PredictIncomeTree, testingSet $ income)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction <=50K >50K
##      <=50K  6459 1115
##      >50K    337 1137
##
##                Accuracy : 0.8395
##                  95% CI : (0.8318, 0.847)
##     No Information Rate : 0.7511
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5147
##
```

```
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9504
##             Specificity : 0.5049
##          Pos Pred Value : 0.8528
##          Neg Pred Value : 0.7714
##              Prevalence : 0.7511
##          Detection Rate : 0.7139
##    Detection Prevalence : 0.8371
##       Balanced Accuracy : 0.7276
##
##        'Positive' Class : <=50K
##
```

7.2 For the Naive Bayes model

```r
#checking The Accuracy of the Naive Bayes model
confusionMatrix(PredictIncomeNaiveBayes, testingSet $ income)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction <=50K >50K
##      <=50K  6313 1124
##      >50K    483 1128
##
##                Accuracy : 0.8224
##                  95% CI : (0.8144, 0.8302)
##     No Information Rate : 0.7511
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.475
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9289
##             Specificity : 0.5009
##          Pos Pred Value : 0.8489
##          Neg Pred Value : 0.7002
##              Prevalence : 0.7511
##          Detection Rate : 0.6977
##    Detection Prevalence : 0.8219
##       Balanced Accuracy : 0.7149
##
##        'Positive' Class : <=50K
##
```

7.3 For the Random Forest model

```r
#checking The Accuracy of the Random Forest model
confusionMatrix(PredictIncomeForest, testingSet $ income)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction <=50K >50K
##      <=50K  6661  272
##      >50K    135 1980
##
##                Accuracy : 0.955
##                  95% CI : (0.9505, 0.9592)
##     No Information Rate : 0.7511
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8772
##
##  Mcnemar's Test P-Value : 1.57e-11
##
##             Sensitivity : 0.9801
##             Specificity : 0.8792
##          Pos Pred Value : 0.9608
##          Neg Pred Value : 0.9362
##              Prevalence : 0.7511
##          Detection Rate : 0.7362
##    Detection Prevalence : 0.7662
##       Balanced Accuracy : 0.9297
##
##        'Positive' Class : <=50K
##
```

8 Result

8.1 Decision Tree

Worked rather well but is not consistently accurate because the features in the data set may not be completely independent. And it gives accuracy on test set as 0.8395.
i.e., 84% to predict the salary class of a person based upon the given information.

8.2 Naïve Bayes

This model was the least successful model as the data had a fatal flaw causing incompatibility with the way the classifier works. Though it gives accuracy on test set as 0.8224.
i.e., 82% to predict the salary class of a person based upon the given information.

8.3 Random Forest

This model worked the best out of all of our models giving the highest accuracy. And it gives accuracy on test set as 0.955.
i.e., 95% to predict the salary class of a person based upon the given information.

9 Conclusion

An individual chances of getting salary more than 50,000 USD strongly based on the individual's age, gender, education and occupation. As a group We came to understand that there various approach available for data exploration, data visualization and tools to play arround with data. And also we understood a lot about R language and its power, especially Rstudio paves us a great way to write our document using Rmarkdown in knitr.

Lastly, we now strongly agree that the Machine learning is no denying a powerful, but it should not be considered as a substitute of traditional statistical modeling.

10 Reference

- https://rstudio.com/wp-content/uploads/2016/03/rmarkdown-cheatsheet-2.0.pdf
- Mastering RStudio – Develop, Communicate, and Collaborate with R by Julian Hillebr and Maximilian H. Nierhoff https://learning.oreilly.com/library/view/mastering-rstudio/9781783982547/
- R for Everyone by Jared P. Lander https://learning.oreilly.com/library/view/r-for-everyone/9780133257182/
- https://bookdown.org/yihui/rmarkdown-cookbook/