Capstone Project

Sharik Baig

28/07/2022

Setting working directory and importing the dataset.

```
setwd("C:/Users/shari/OneDrive/Desktop/Business Analytics/Capstone")

adult_income <- read.csv("C:/Users/shari/OneDrive/Desktop/Business
Analytics/Capstone/adult.csv")
```

Importing the required libraries

```
library(ggplot2)

## Warning in register(): Can't find generic `scale_type` in package ggplot2
to
## register S3 method.

library(caTools)
library(caret)

## Loading required package: lattice
```

#Preview the data

```
head(adult_income)

##   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

str(adult_income)

## '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" ...
```

cleaning the data and combining factors of the working class

```
table(adult_income$workclass)

##
##                 ?      Federal-gov        Local-gov     Never-worked
##              1836              960             2093                7
##           Private     Self-emp-inc Self-emp-not-inc        State-gov
##             22696             1116             2541             1298
##      Without-pay
##                14

adult_income$workclass <- as.character(adult_income$workclass)

# combining work class of without-pay and never-worked as Unemployed
adult_income$workclass[adult_income$workclass == "Without-pay" |
adult_income$workclass == "Never-worked"] <- "Unemployed"

# combining work class of state-gov and local-gov as State/Local-gov
adult_income$workclass[adult_income$workclass == "State-gov" |
adult_income$workclass == "Local-gov"] <- "State/Local-gov"

# combining work class of self-emp-inc and self-emp-not-inc as Self-employed
```

```
adult_income$workclass[adult_income$workclass == "Self-emp-inc" |
adult_income$workclass == "Self-emp-not-inc"] <- "Self-employed"

# we are not combining federal work class and private work class because both
are different work classes
table(adult_income$workclass)

##
##                ?     Federal-gov         Private   Self-employed
State/Local-gov
##             1836             960           22696            3657
3391
##       Unemployed
##               21
```

Combining factors of marital status

```
table(adult_income$marital.status)

##
##              Divorced     Married-AF-spouse    Married-civ-spouse
##                  4443                    23                 14976
## Married-spouse-absent         Never-married             Separated
##                   418                 10683                  1025
##               Widowed
##                   993

adult_income$marital.status <- as.character(adult_income$marital.status)

# Combining Married-AF-spouse, Married-civ-spouse and Married-spouse-absent
as Married
adult_income$marital.status[adult_income$marital.status == "Married-AF-
spouse" | adult_income$marital.status == "Married-civ-spouse" |
adult_income$marital.status == "Married-spouse-absent"] <- "Married"

# Combining Divorced, Separated and Widowed as Not-Married
adult_income$marital.status[adult_income$marital.status == "Divorced" |
adult_income$marital.status == "Separated" | adult_income$marital.status ==
"Widowed"] <- "Not-Married"

table(adult_income$marital.status)

##
##       Married Never-married   Not-Married
##         15417         10683          6461
```

Combining factors of Country

```
adult_income$native.country <- as.character(adult_income$native.country)
# combining the below countries to North.America
North.America <- c("Canada","Cuba","Dominican-Republic","El-
Salvador","Guatemala","Haiti","Honduras","Jamaica","Mexico","Nicaragua","Outl
```

```
ying-US(Guam-USVI-etc)","Puerto-Rico","Trinadad&Tobago","United-States")

# combining the below countries to Asia
Asia <-
c("Cambodia","China","Hong","India","Iran","Japan","Laos","Philippines","Taiw
an","Thailand","Vietnam")

# combining the below countries to South.America
South.America <- c("Columbia","Ecuador","Peru")

# combining the below countries to Europe
Europe <- c("England", "France", "Germany", "Greece", "Holand-Netherlands",
"Hungary", "Ireland", "Italy", "Poland", "Portugal", "Scotland",
"Yugoslavia")

# combining the below countries to others
Others <- c("South","?")
adult_income$native.country[adult_income$native.country %in% North.America]
<- "North_America"
adult_income$native.country[adult_income$native.country %in% Asia] <- "Asia"
adult_income$native.country[adult_income$native.country %in% South.America]
<- "South_America"
adult_income$native.country[adult_income$native.country %in% Europe] <-
"Europe"
adult_income$native.country[adult_income$native.country %in% Others] <-
"Others"
table(adult_income$native.country)

##
##          Asia        Europe North_America        Others South_America
##           671           521         30588           663           118

# converting the below variables into factors
adult_income$workclass <- as.factor(adult_income$workclass)
adult_income$marital.status <- as.factor(adult_income$marital.status)
adult_income$native.country <- as.factor(adult_income$native.country)
str(adult_income)

## 'data.frame':    32561 obs. of  15 variables:
##  $ age           : int  90 82 66 54 41 34 38 74 68 41 ...
##  $ workclass     : Factor w/ 6 levels "?","Federal-gov",..: 1 3 1 3 3 3 3
5 2 3 ...
##  $ 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: Factor w/ 3 levels "Married","Never-married",..: 3 3 3
3 3 3 3 2 3 2 ...
##  $ 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: Factor w/ 5 levels "Asia","Europe",..: 3 3 3 3 3 3 3 3
3 4 ...
##  $ income        : chr  "<=50K" "<=50K" "<=50K" "<=50K" ...
```

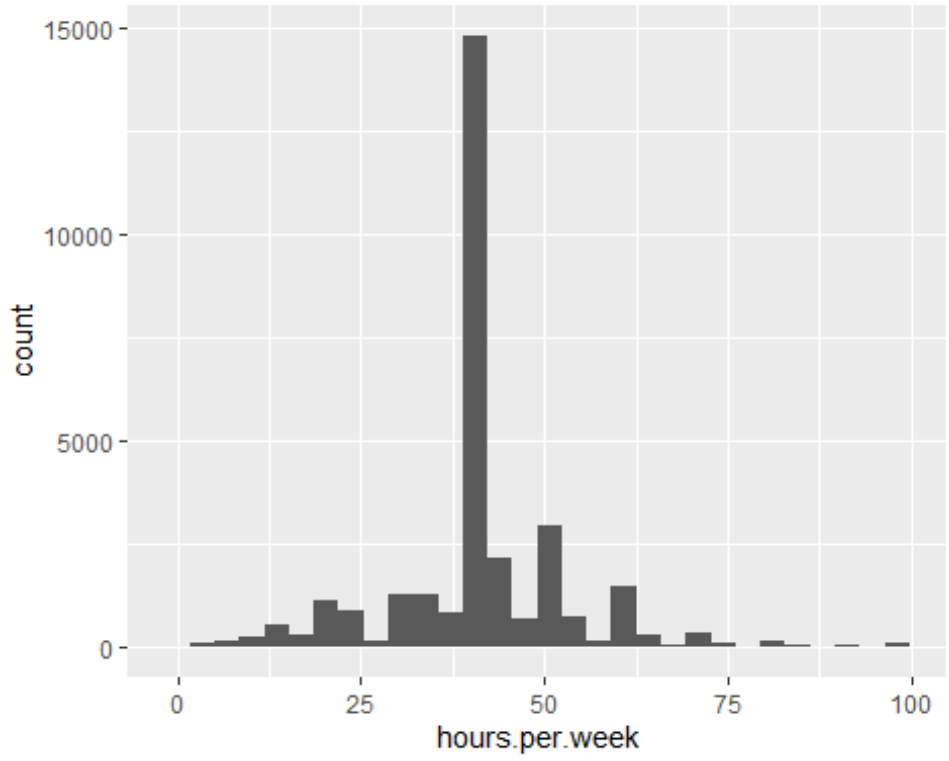
Now we deal with missing data

```
table(adult_income$workclass)

##
##               ?     Federal-gov         Private   Self-employed
State/Local-gov
##            1836             960           22696            3657
3391
##      Unemployed
##              21

# assigning NA to the missing values
adult_income[adult_income == "?"] <- NA

# Converting Income to factors as well
adult_income$income <- as.factor(adult_income$income)
adult_income[adult_income$income == "<=50k"] <- "0"
adult_income[adult_income$income == ">50k"] <- "1"
table(adult_income$workclass)

##
##               ?     Federal-gov         Private   Self-employed
State/Local-gov
##               0             960           22696            3657
3391
##      Unemployed
##              21

# omitting the NA values
adult_income <- na.omit(adult_income)
```

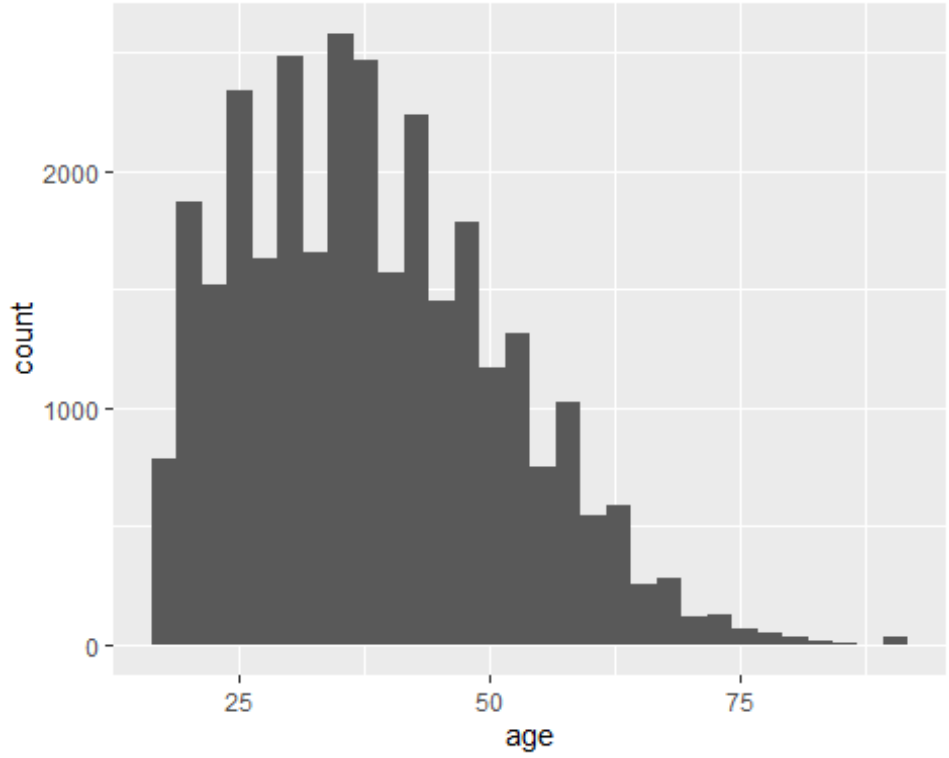
Exploring and analysing data

```
# We have to understand the trends and representations of certain
demographics
ggplot(adult_income, aes(age)) + geom_histogram(aes(fill = income), color =
"black", binwidth = 1)
```


```
ggplot(adult_income, aes(hours.per.week)) + geom_histogram()

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



```
ggplot(adult_income, aes(age)) + geom_histogram()

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



Now we build our Logistic Regression Model

The purpose of this model is to classify people into two groups, below 50K or above 50K

```
# now we split the data into 75% training and 25% testing
Adult_split <- sample.split(adult_income$income, SplitRatio = 0.8)

# we assigned training data to Adult_train
Adult_Train <- subset(adult_income, Adult_split == TRUE,)

# we assigned testing data to Adult_test
Adult_Test <- subset(adult_income, Adult_split == FALSE)

# Training the model
Adult_income_model <- glm(income ~., family = binomial(), data =
Adult_Train)

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

# Predicting the Salary class
Prediction <- predict(Adult_income_model, Adult_Test, type = "response")
```

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

Creating a confusion matrix

```
table(Adult_Test$income, Prediction >= 0.7)
```

```
##
##         FALSE TRUE
##   <=50K  4510  104
##   >50K    940  590
```

```
# Accuracy
(4510+576)/(4510+104+954+576)
```

```
## [1] 0.8277995
```

```
# REcall
4510/(4510+954)
```

```
## [1] 0.8254026
```

```
#precision
4510/(4510+104)
```

```
## [1] 0.9774599
```