Adult Census Income

Data Science: Capstone Proyect Report

Guillermo Ortiz

3/7/2022

Executive Summary

The Adult Census Income dataset was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker. It stores the information of adults from the entire US on their socioeconomic and demographic characteristics. The task is to predict whether or not adults earn more than 50K US dollars a year, based on the aforementioned characteristics. I choose the best model using the accuracy metric. Finally, I use a completely different set of observations for validation purposes. After exploring the data and developing some models, I reach an accuracy of 0.857.

Introduction

As Chakrabarty & Biswas (2018) point out, inequality in wealth and income is a major source of worry, particularly in the United States. Thus, one reasonable motivation to lessen the world's rising level of economic disparity is the possibility of reducing poverty. The notion of universal moral equality promotes long-term development and improves a country's economic stability. Governments in several countries have been working hard to address this issue and find the best answer possible.

In this project, the goal is to utilize machine learning and data mining techniques to help to properly diagnose the problem of income inequality. The Adult Dataset from UCI is utilized for doing this. The Adult Census Income dataset was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker. It stores the information of adults from the entire US on their socioeconomic and demographic characteristics. The classification task of this project is to determine if a person's annual income in the United States is greater than 50K or less than 50K, using several socioeconomic and demographic characteristics of the population as predictors.

In the first section I will load the packages that will be used. In the second section I will download the data both for modeling and for validation purposes. In the third section I start exploring the data and briefly evaluating each of the variables. In the fourth section I will prepare the data for modeling and analysis. The fifth section will focus on modeling and evaluation of algorithms, after which the best algorithm will be chosen. The sixth section will apply the best algorithm on all the data and evaluate its predictions on the dataset available for validation purposes only. Finally, some concluding remarks will finish this report.

Section 1: Packages & Options

In this section, I load the packages that will be used and set the preferable options for analysis and output printing.

```r
if(!require(tidyverse)) install.packages("tidyverse")
if(!require(caret)) install.packages("caret")
if(!require(fastDummies)) install.packages("fastDummies")
if(!require(gridExtra)) install.packages("gridExtra")
if(!require(broom)) install.packages("broom")

library(tidyverse)
library(caret)
library(fastDummies)
library(gridExtra)
library(broom)
options(digits = 4)
options(scipen = 999)
```

Section 2: Data Loading

The data is downloaded directly from the University of california Irvine's (UCI) Machine Learning Repository. It consists of two separate datasets: the *adult* dataset (used for modeling and analysis) and the *adult.test* dataset (used for validation and evaluation purposes). I decided to change the name of the *adult.test* dataset to *adult.validation*, in order to avoid confusion with the *test* set I will use to train the algorithm.

```r
adult <- read.csv(url("https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data"),
         header=FALSE, col.names = c("age", "workclass", "fnlwgt","education",
                                     "education.num", "marital.status", "occupation",
                                     "relationship", "race", "sex", "capital.gain",
                                     "capital.loss", "hours.per.week",
                                     "native.country", "income"))

adult.validation <- read.csv(url("https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult
                  header=FALSE, col.names = c("age", "workclass", "fnlwgt","education",
                                              "education.num","marital.status",
                                              "occupation", "relationship", "race",
                                              "sex", "capital.gain", "capital.loss",
                                              "hours.per.week","native.country",
                                              "income"))

adult.validation <- adult.validation[-1,] %>% mutate(age=as.integer(age))
```

Section 3: Data Exploration

In this section, I will briefly explore the adult dataset and try to find some patterns and data characteristics that can be useful for prediction.

I start observing the structure of the data.

```r
# Structure of dataset
str(adult)
```

```
## 'data.frame':    32561 obs. of  15 variables:
##  $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ workclass     : chr  " State-gov" " Self-emp-not-inc" " Private" " Private" ...
```

```
##  $ fnlwgt        : int  77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...
##  $ education     : chr  " Bachelors" " Bachelors" " HS-grad" " 11th" ...
##  $ education.num : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ marital.status: chr  " Never-married" " Married-civ-spouse" " Divorced" " Married-civ-spouse" ...
##  $ occupation    : chr  " Adm-clerical" " Exec-managerial" " Handlers-cleaners" " Handlers-cleaners"
##  $ relationship  : chr  " Not-in-family" " Husband" " Not-in-family" " Husband" ...
##  $ race          : chr  " White" " White" " White" " Black" ...
##  $ sex           : chr  " Male" " Male" " Male" " Male" ...
##  $ 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: chr  " United-States" " United-States" " United-States" " United-States" ...
##  $ income        : chr  " <=50K" " <=50K" " <=50K" " <=50K" ...
```

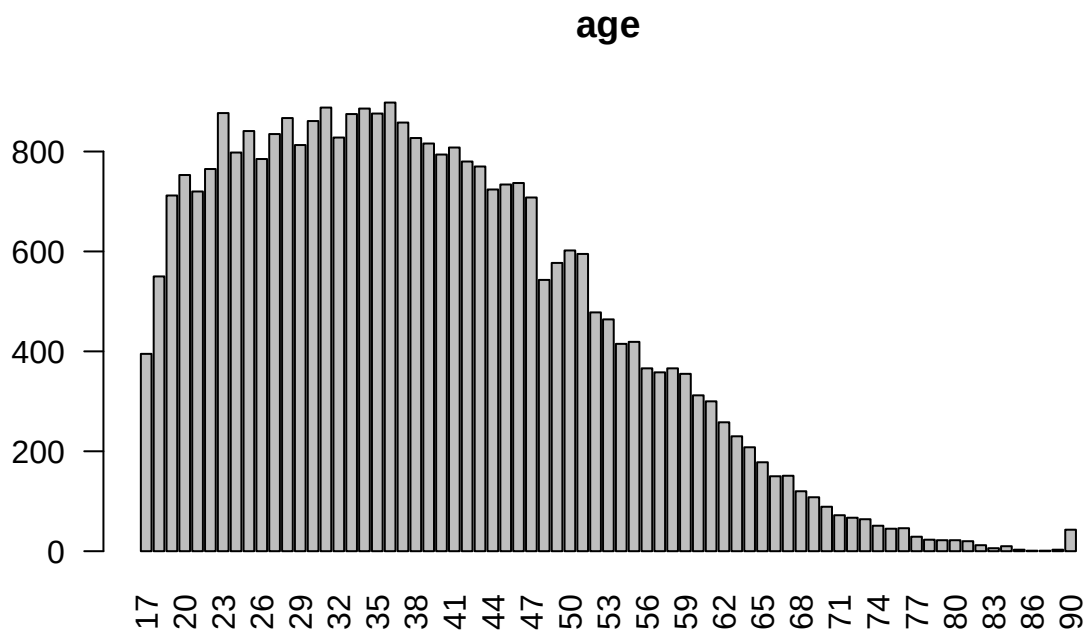
There are 15 variables in the adult dataset and 32,561 observations. The variables are: age, workclass, fnlwgt, education, education.num, marital.status, occupation, relationship, race, sex, capital.gain, capital.loss, hour.per.week, native.country and income. There are 8 character variables and 7 numeric (integer) variables. The goal is to predict (classify) the variable income and to determine whether a person earns more or less than 50K US dollars a year.

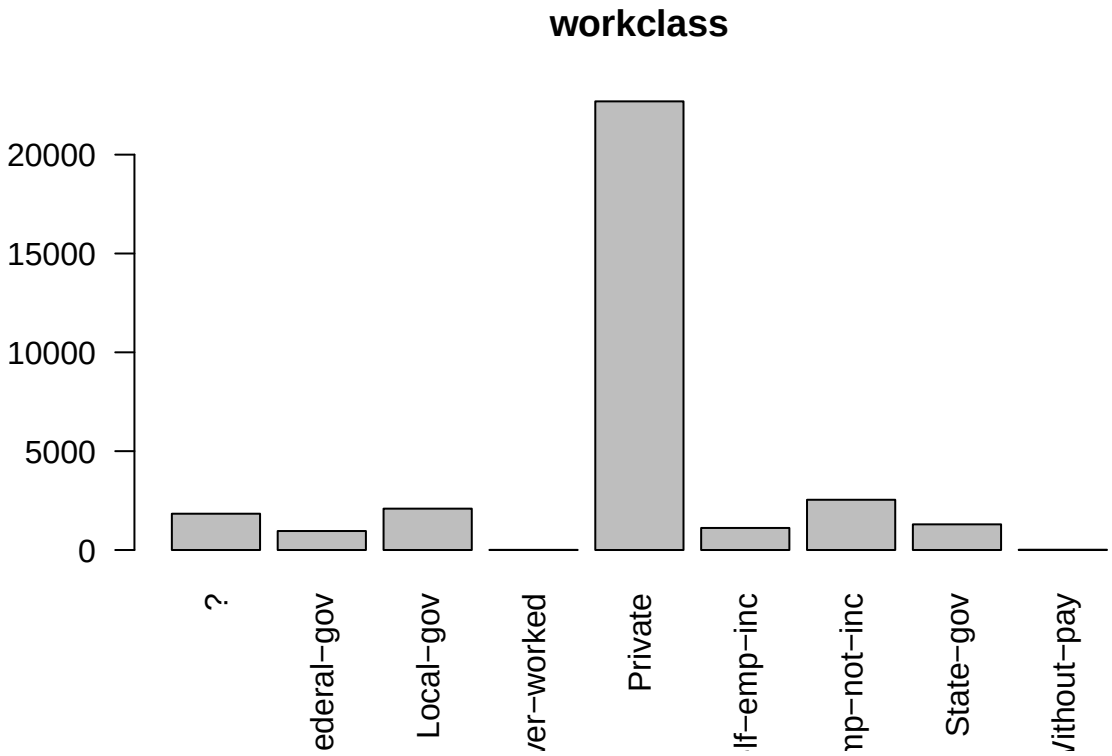
Next, I proceed to print the summary of the data, and show barplots/histograms for each variable to have a sense of what is the information contained in them.

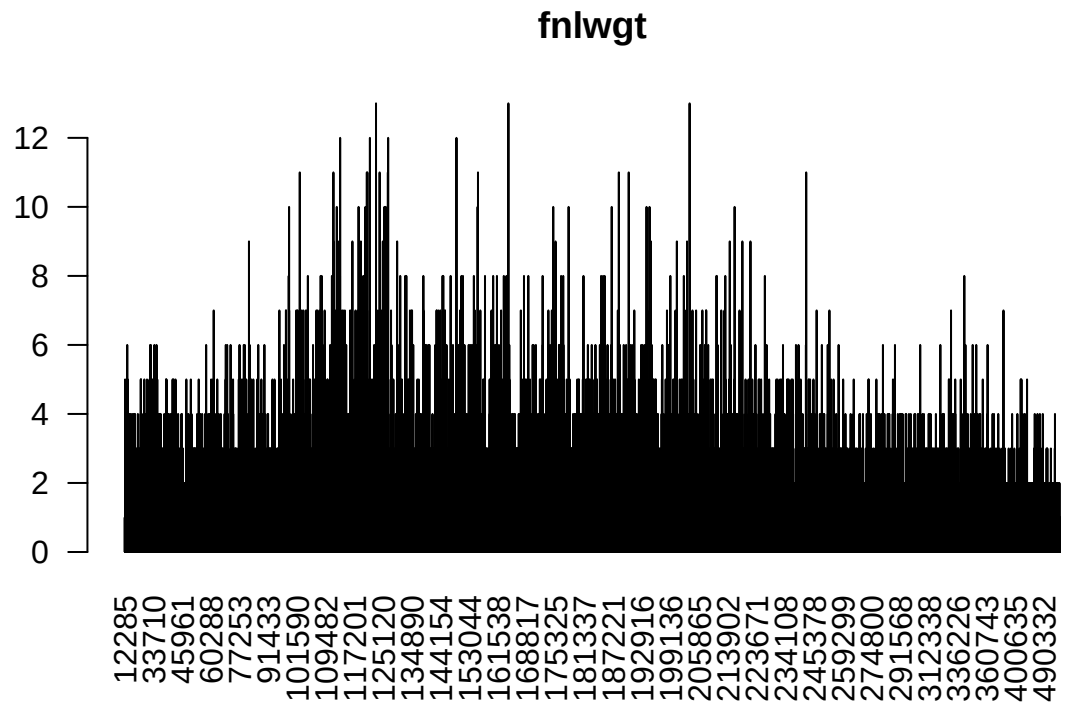
```
# Summary of dataset
summary(adult)
```

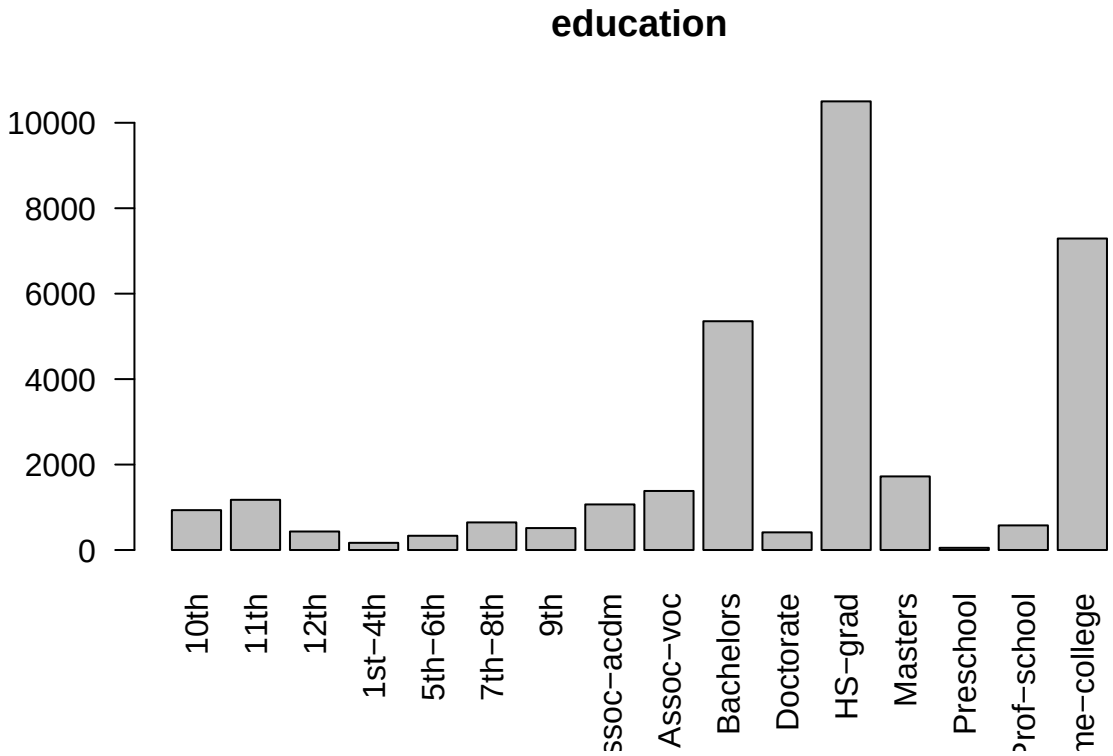
```
##       age        workclass            fnlwgt         education
##  Min.   :17.0   Length:32561       Min.   :  12285   Length:32561
##  1st Qu.:28.0   Class :character   1st Qu.: 117827   Class :character
##  Median :37.0   Mode  :character   Median : 178356   Mode  :character
##  Mean   :38.6                      Mean   : 189778
##  3rd Qu.:48.0                      3rd Qu.: 237051
##  Max.   :90.0                      Max.   :1484705
##  education.num  marital.status      occupation        relationship
##  Min.   : 1.0   Length:32561       Length:32561       Length:32561
##  1st Qu.: 9.0   Class :character   Class :character   Class :character
##  Median :10.0   Mode  :character   Mode  :character   Mode  :character
##  Mean   :10.1
##  3rd Qu.:12.0
##  Max.   :16.0
##      race               sex             capital.gain    capital.loss
##  Length:32561       Length:32561       Min.   :    0   Min.   :   0
##  Class :character   Class :character   1st Qu.:    0   1st Qu.:   0
##  Mode  :character   Mode  :character   Median :    0   Median :   0
##                                        Mean   : 1078   Mean   :  87
##                                        3rd Qu.:    0   3rd Qu.:   0
##                                        Max.   :99999   Max.   :4356
##  hours.per.week native.country        income
##  Min.   : 1.0   Length:32561       Length:32561
##  1st Qu.:40.0   Class :character   Class :character
##  Median :40.0   Mode  :character   Mode  :character
##  Mean   :40.4
##  3rd Qu.:45.0
##  Max.   :99.0
```

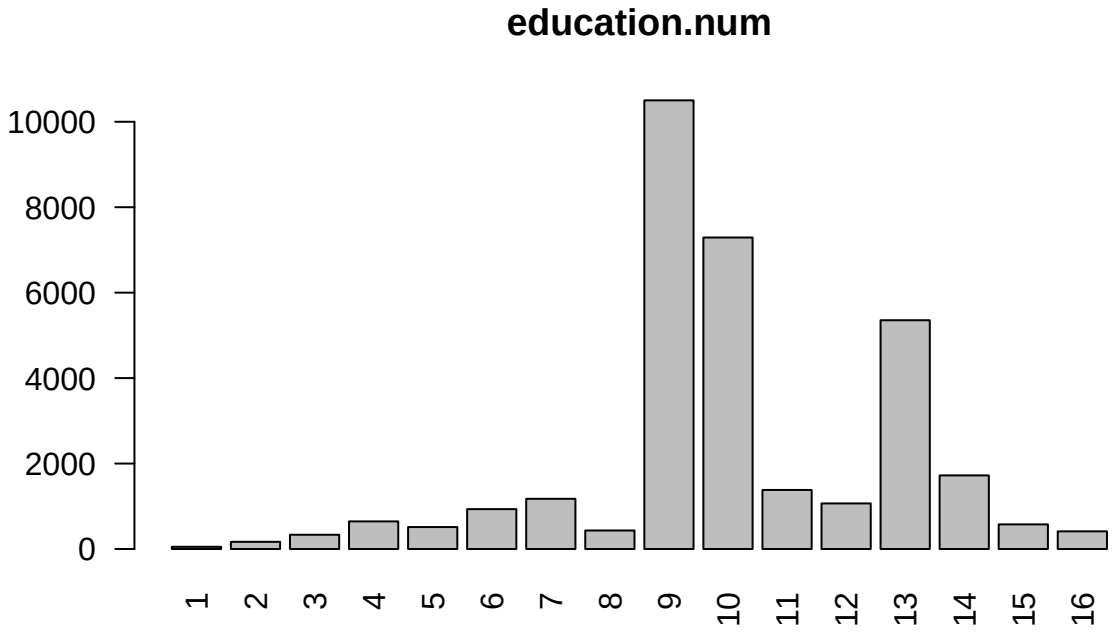
```r
# Barplot for each variable in dataset
for (i in 1:15) {
  table(adult[i]) %>% barplot(las=2, main=colnames(adult)[i])
}
```



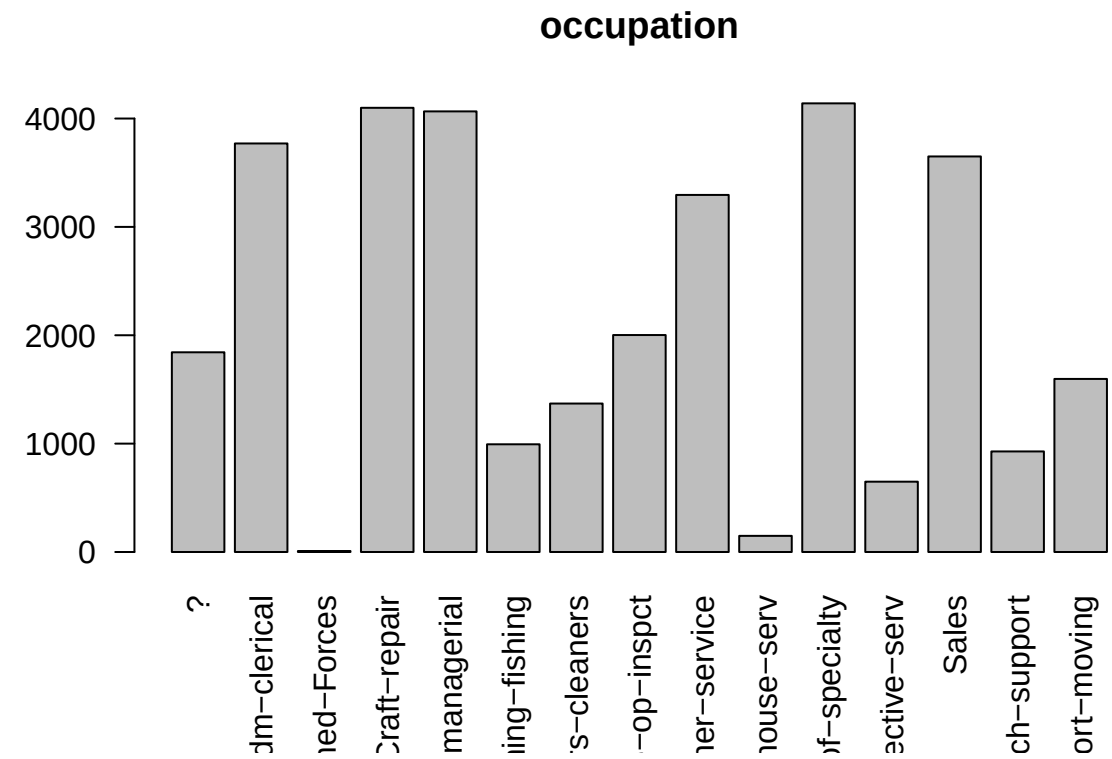

workclass
20000
15000
10000
5000
0
?
ederal−gov
Local−gov
/er−worked
Private
lf−emp−inc
np−not−inc
State−gov
/ithout−pay

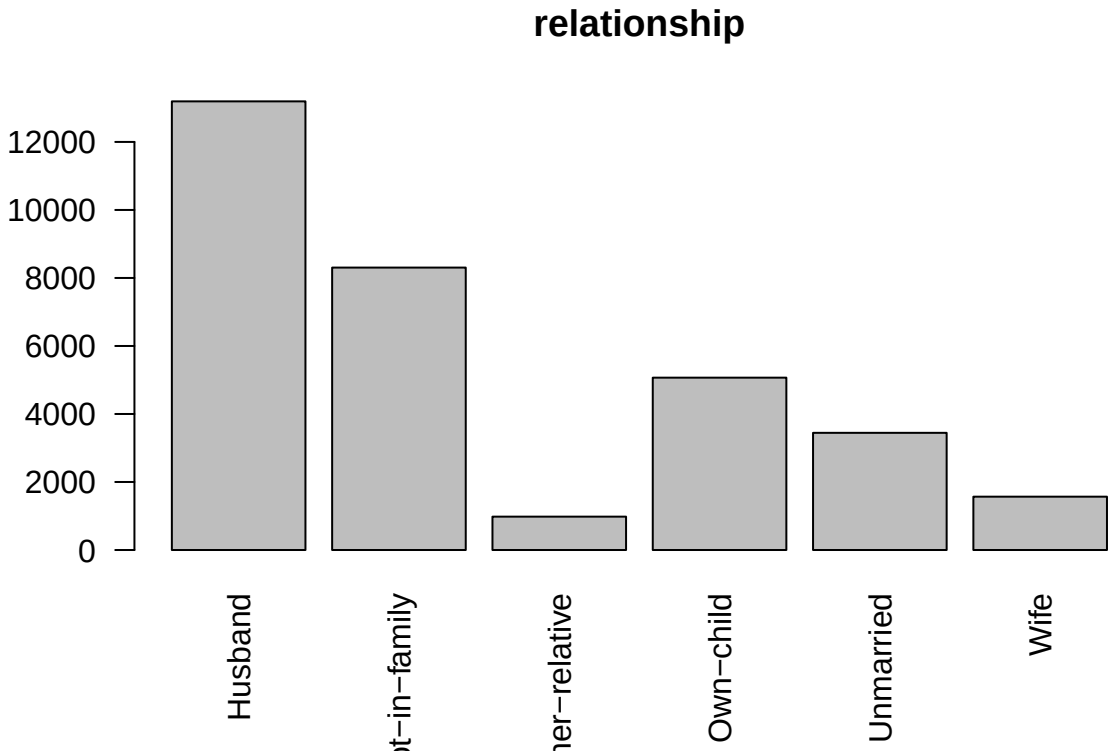

fnlwgt
12
10
8
6
4
2
0
12285
33710
45961
60288
77253
91433
101590
109482
117201
125120
134890
144154
153044
161538
168817
175325
181337
187221
192916
199136
205865
213902
223671
234108
245378
259299
274800
291568
312338
336226
360743
400635
490332

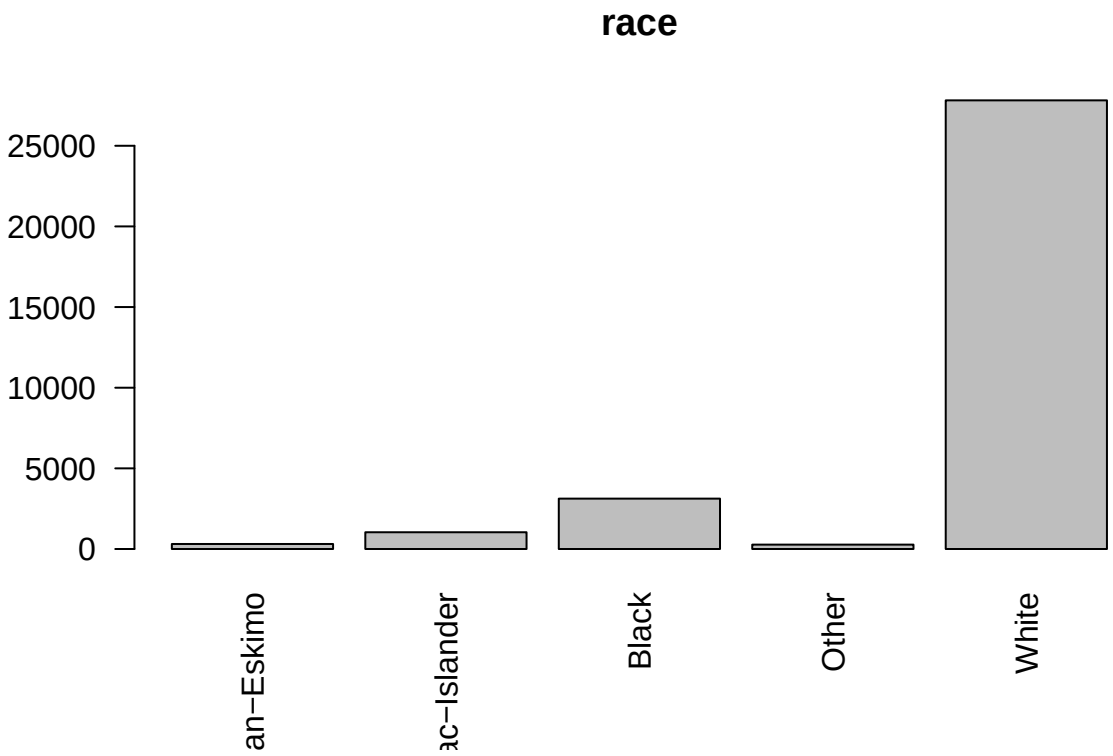

education
10000
8000
6000
4000
2000
0
10th
11th
12th
1st−4th
5th−6th
7th−8th
9th
ssoc−acdm
Assoc−voc
Bachelors
Doctorate
HS−grad
Masters
Preschool
rof−school
ne−college

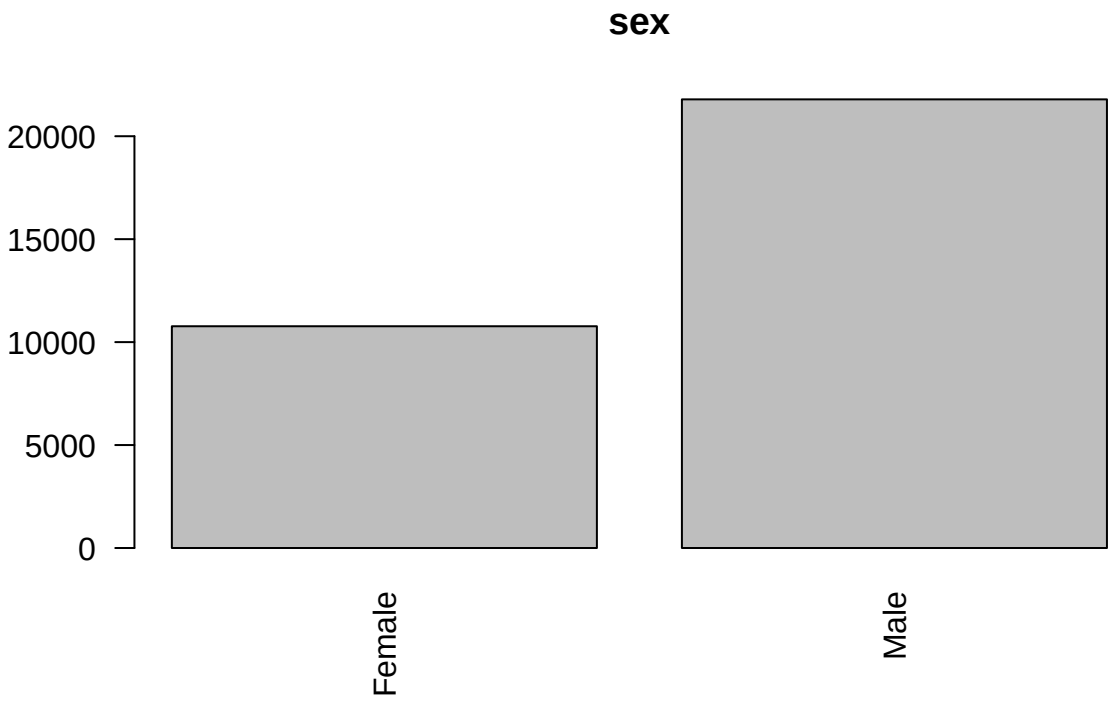

education.num
10000
8000
6000
4000
2000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

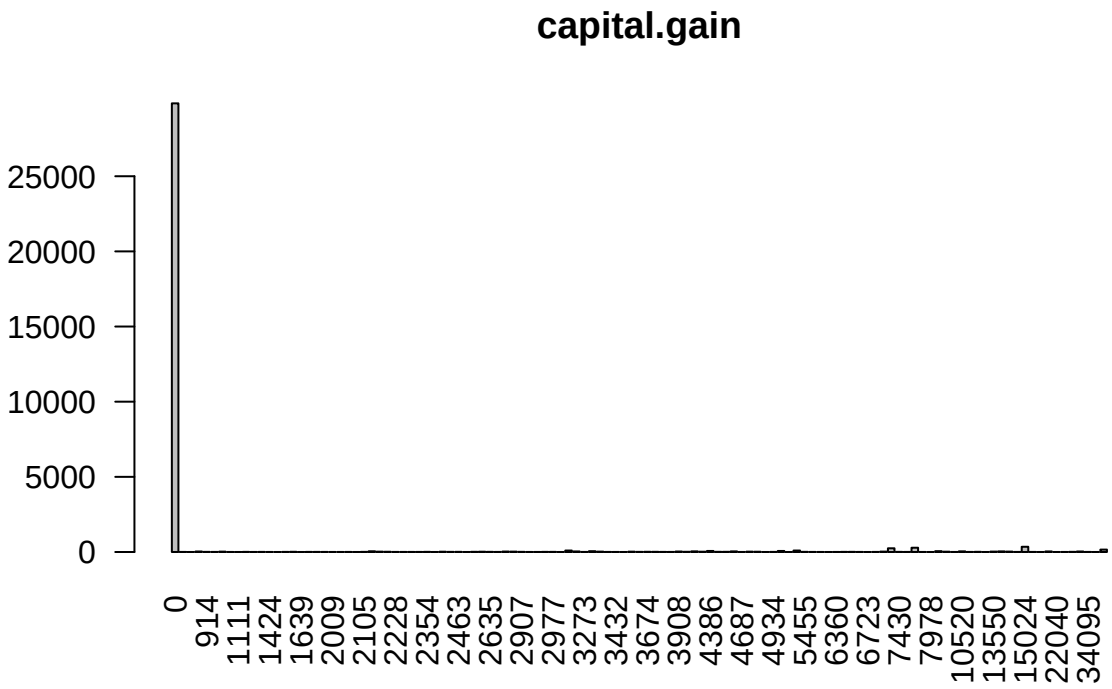

marital.status
14000
12000
10000
8000
6000
4000
2000
0
Divorced
AF−spouse
civ−spouse
se−absent
er−married
Separated
Widowed

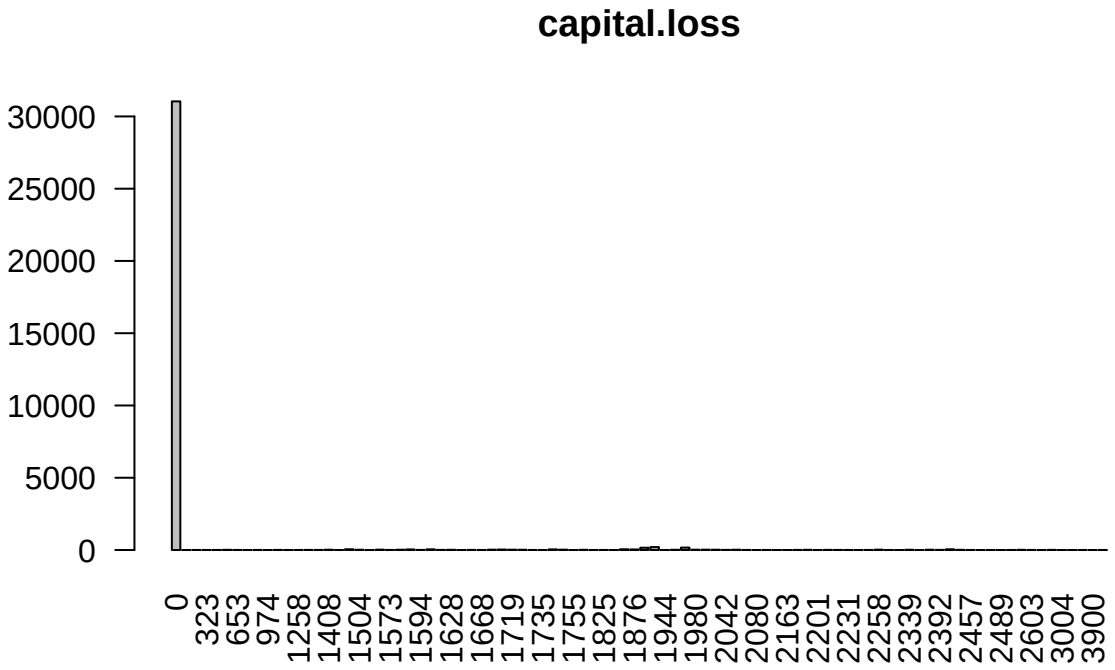

occupation
4000
3000
2000
1000
0
?
dm−clerical
ed−Forces
raft−repair
managerial
ing−fishing
s−cleaners
op−inspct
er−service
ouse−serv
f−specialty
ective−serv
Sales
ch−support
ort−moving

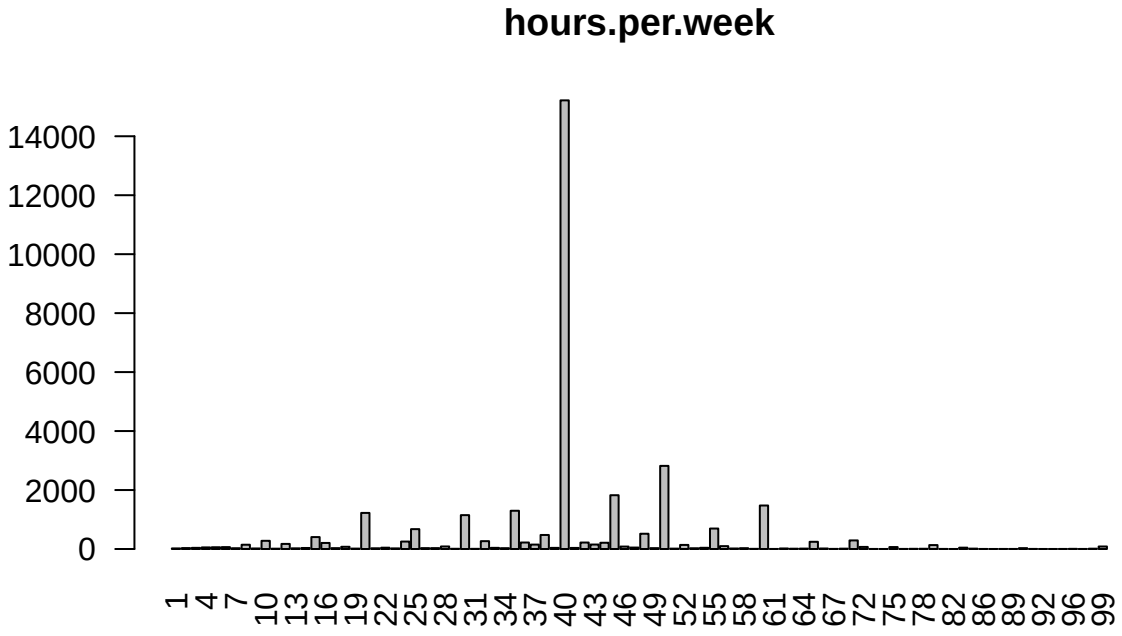

relationship
12000
10000
8000
6000
4000
2000
0
Husband
Own−child
Unmarried
Wife

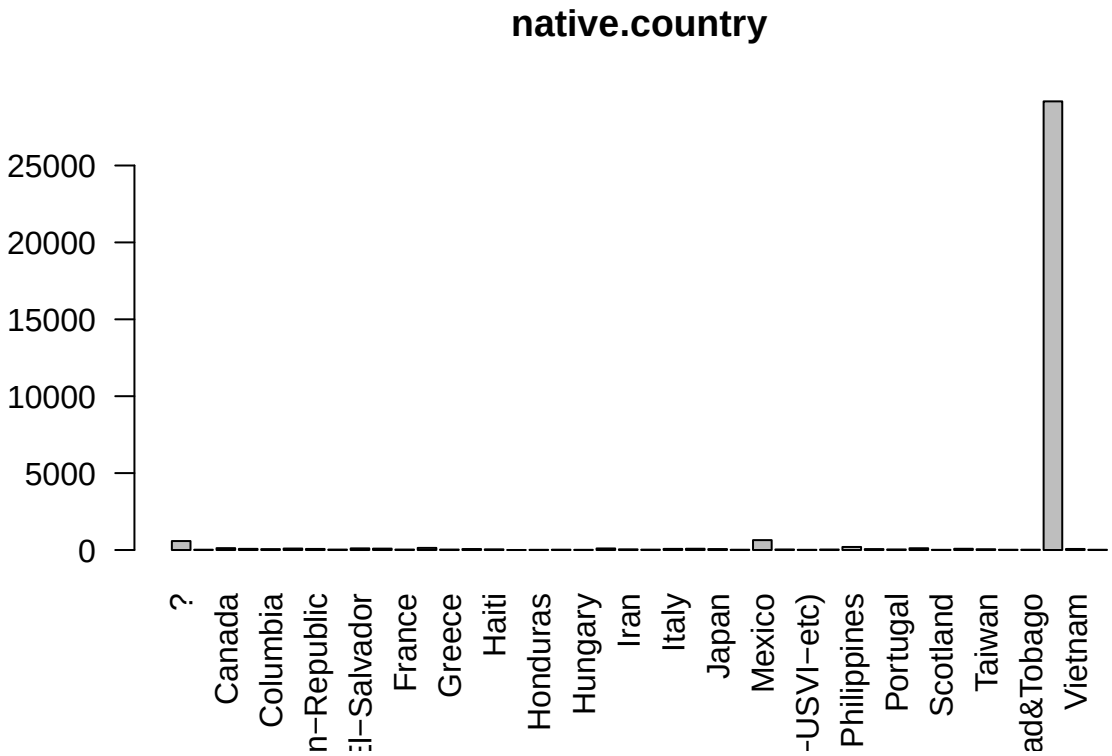

race
25000
20000
15000
10000
5000
0
Black
Other
White


sex
20000
15000
10000
5000
0
Female
Male


capital.gain
25000
20000
15000
10000
5000
0
0
914
1111
1424
1639
2009
2105
2228
2354
2463
2635
2907
2977
3273
3432
3674
3908
4386
4687
4934
5455
6360
6723
7430
7978
10520
13550
15024
22040
34095


capital.loss
30000
25000
20000
15000
10000
5000
0
0
323
653
974
1258
1408
1504
1573
1594
1628
1668
1719
1735
1755
1825
1876
1944
1980
2042
2080
2163
2201
2231
2258
2339
2392
2457
2489
2603
3004
3900


hours.per.week
14000
12000
10000
8000
6000
4000
2000
0
1
4
7
10
13
16
19
22
25
28
31
34
37
40
43
46
49
52
55
58
61
64
67
72
75
78
82
86
89
92
96
99


native.country
25000
20000
15000
10000
5000
0
?
Canada
Columbia
n−Republic
l−Salvador
France
Greece
Haiti
Honduras
Hungary
Iran
Italy
Japan
Mexico
−USVI−etc)
Philippines
Portugal
Scotland
Taiwan
ad&Tobago
Vietnam

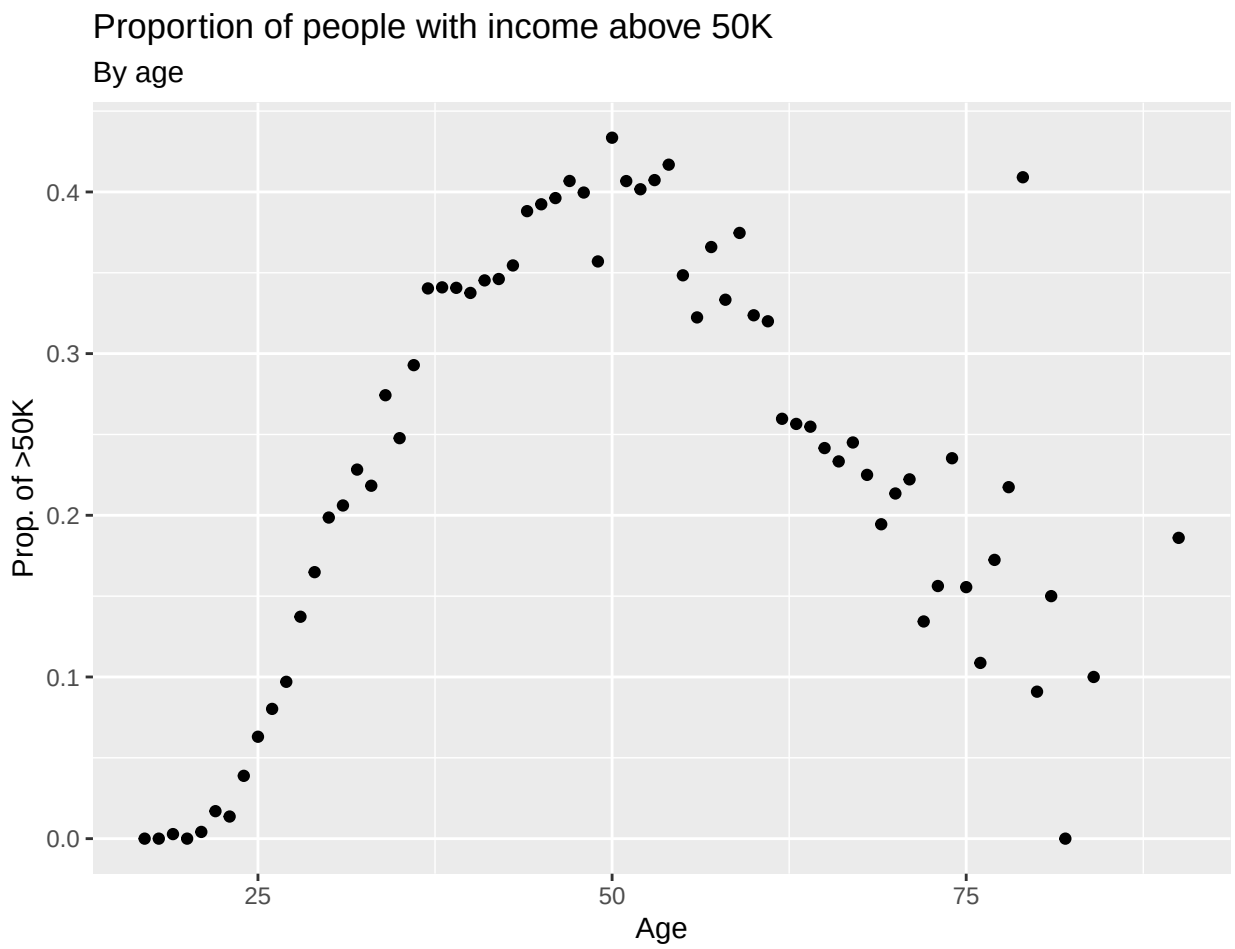

income
20000
15000
10000
5000
0
<=50K
>50K

After looking at the variables' distributions, I get a sense on what is the information they contain and how I can properly use them.

Relationship among variables and income

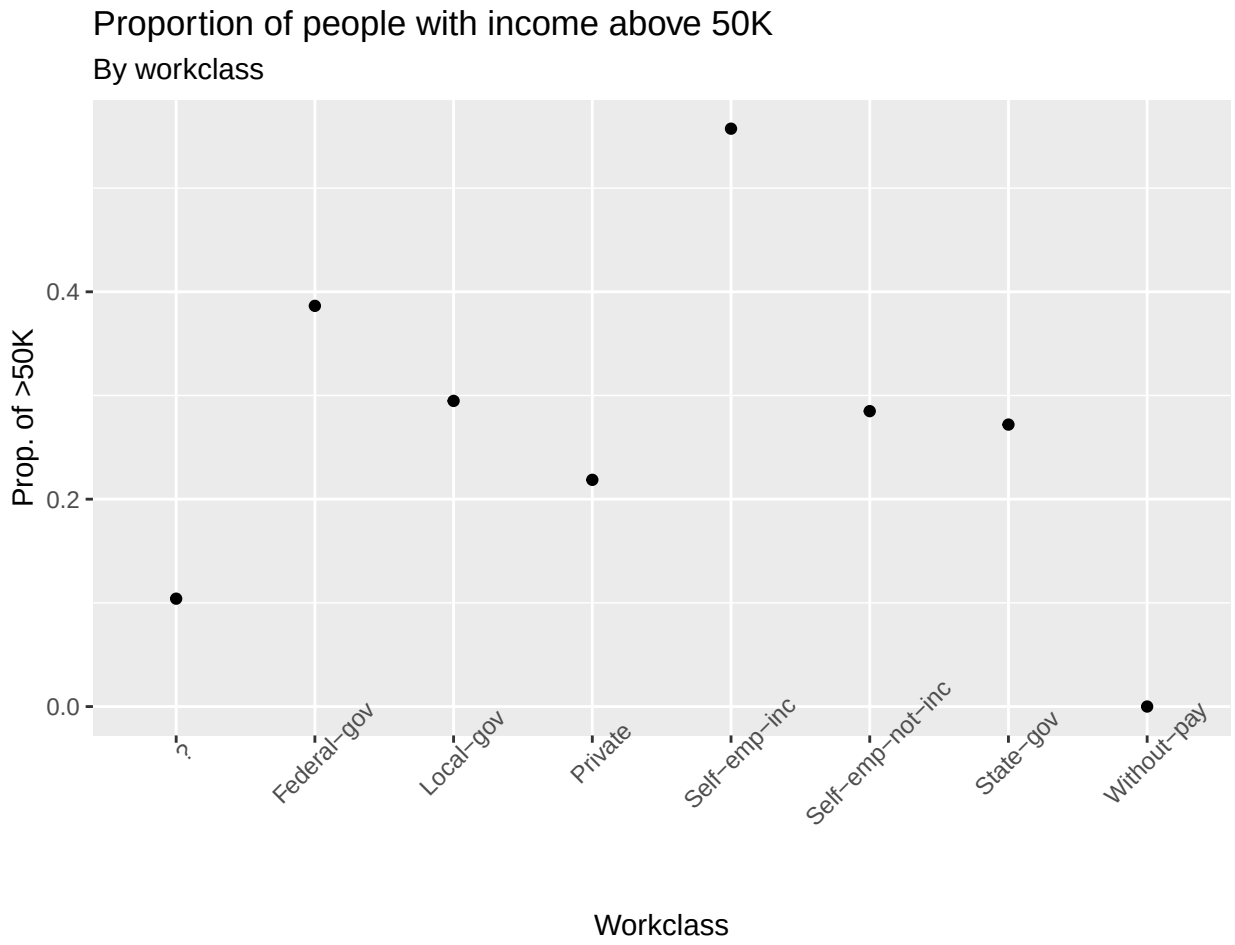
To deepen the analysis, I now make a plot of each variable and their relationship with the variable *income*, which is the one I will try to predict. More specifically, I plot the probability of income being greater than 50K, conditional on taking a specific value on each variable.

```
# Age
adult %>%
  mutate(x = round(age)) %>%
  group_by(x) %>%
  filter(n() >= 10) %>%
  summarise(prop = mean(income == " >50K")) %>%
  ggplot(aes(x, prop)) +
  geom_point() +
  ggtitle("Proportion of people with income above 50K",
          "By age") +
  labs(x="Age", y="Prop. of >50K")
```



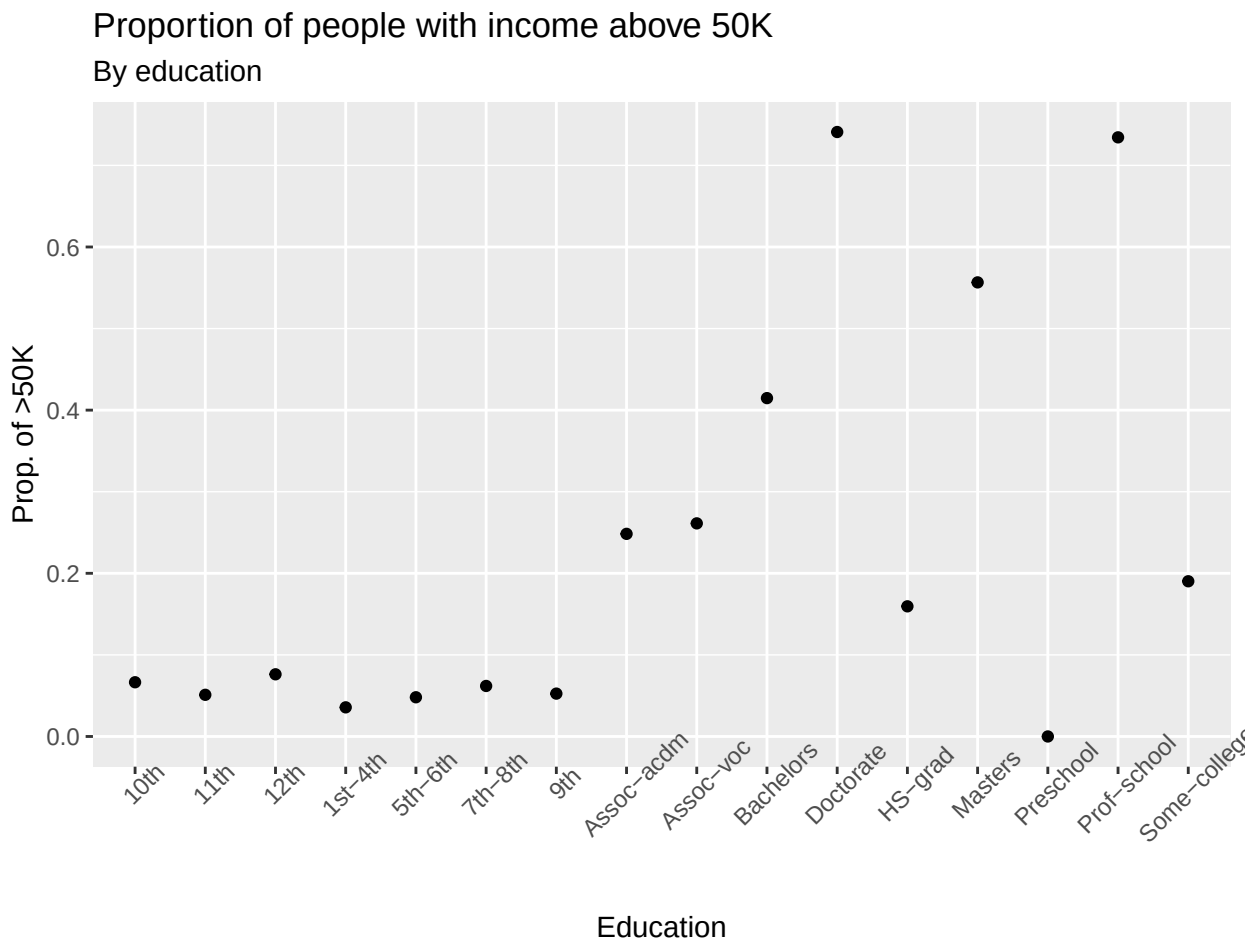
Apparently, the relationship between age and income seems to be quadratic.

```r
# Workclass
adult %>%
  mutate(x = workclass) %>%
  group_by(x) %>%
  filter(n() >= 10) %>%
  summarise(prop = mean(income == " >50K")) %>%
  ggplot(aes(x, prop)) +
  geom_point() +
  ggtitle("Proportion of people with income above 50K",
          "By workclass") +
  labs(x="Workclass", y="Prop. of >50K") +
  theme(axis.text.x = element_text(angle = 45))
```



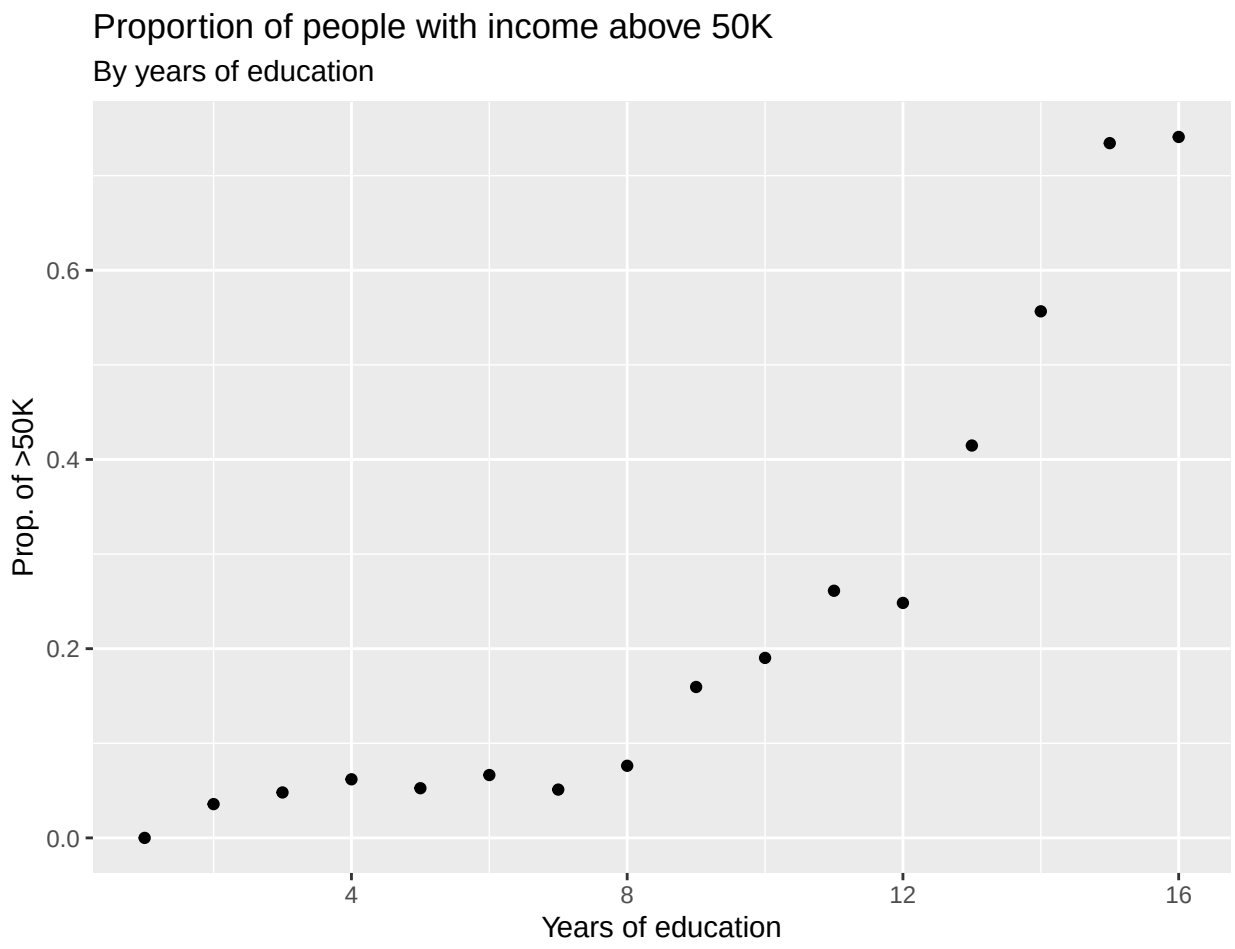
Apparently, the probability of earning more than 50K a year conditional on being self-employed is greater than conditioning on other workclasses.

```r
# Education
adult %>%
  mutate(x = education) %>%
  group_by(x) %>%
  filter(n() >= 10) %>%
  summarise(prop = mean(income == " >50K")) %>%
  ggplot(aes(x, prop)) +
  geom_point() +
  ggtitle("Proportion of people with income above 50K",
          "By education") +
  labs(x="Education", y="Prop. of >50K") +
  theme(axis.text.x = element_text(angle = 45))
```



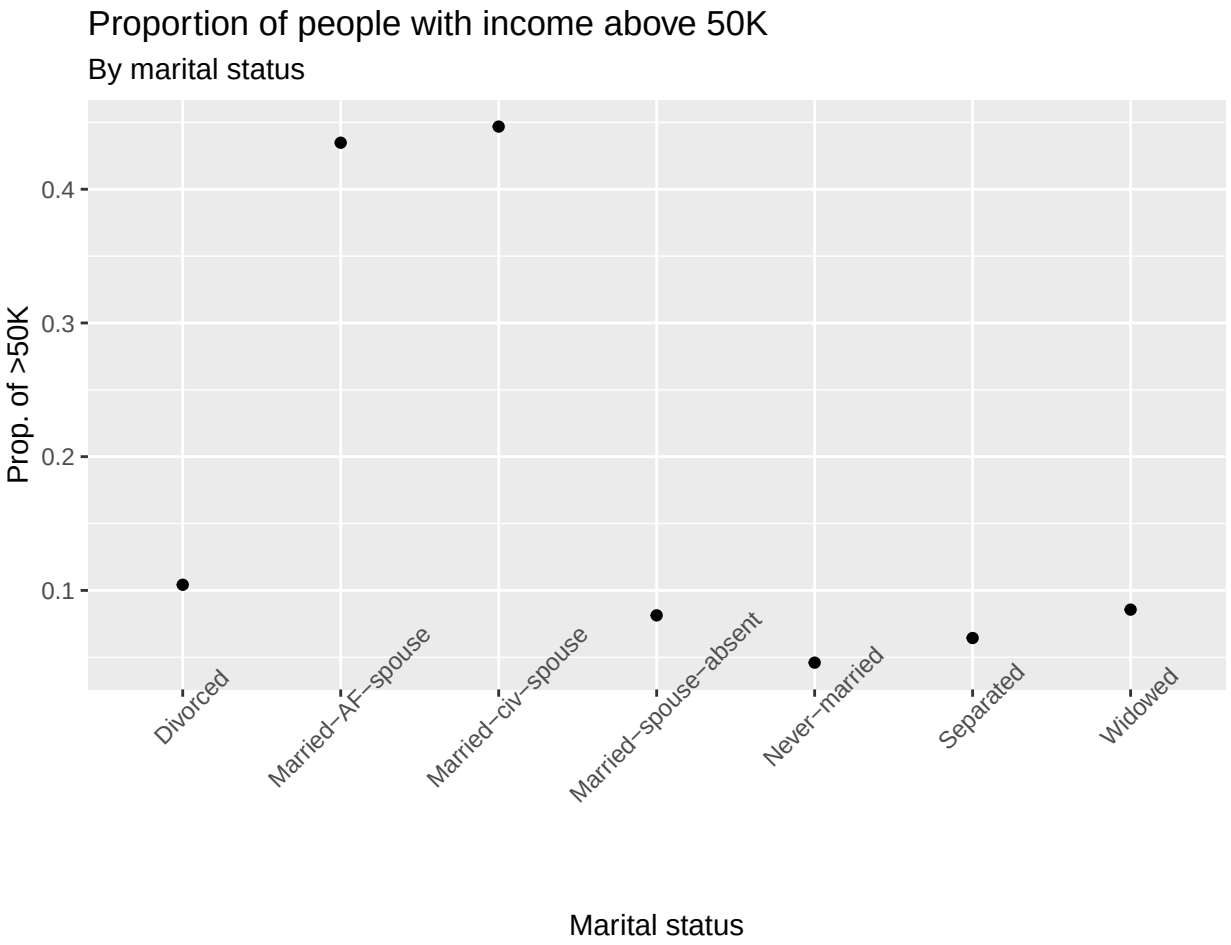
Apparently, the probability of earning more than 50K a year conditional on having a doctorate degree or being a professor is greater than conditioning on other education levels.

```r
# Years of education
adult %>%
  mutate(x = round(education.num)) %>%
  group_by(x) %>%
  filter(n() >= 10) %>%
  summarise(prop = mean(income == " >50K")) %>%
  ggplot(aes(x, prop)) +
  geom_point() +
  ggtitle("Proportion of people with income above 50K",
                        "By years of education") +
  labs(x="Years of education", y="Prop. of >50K")
```



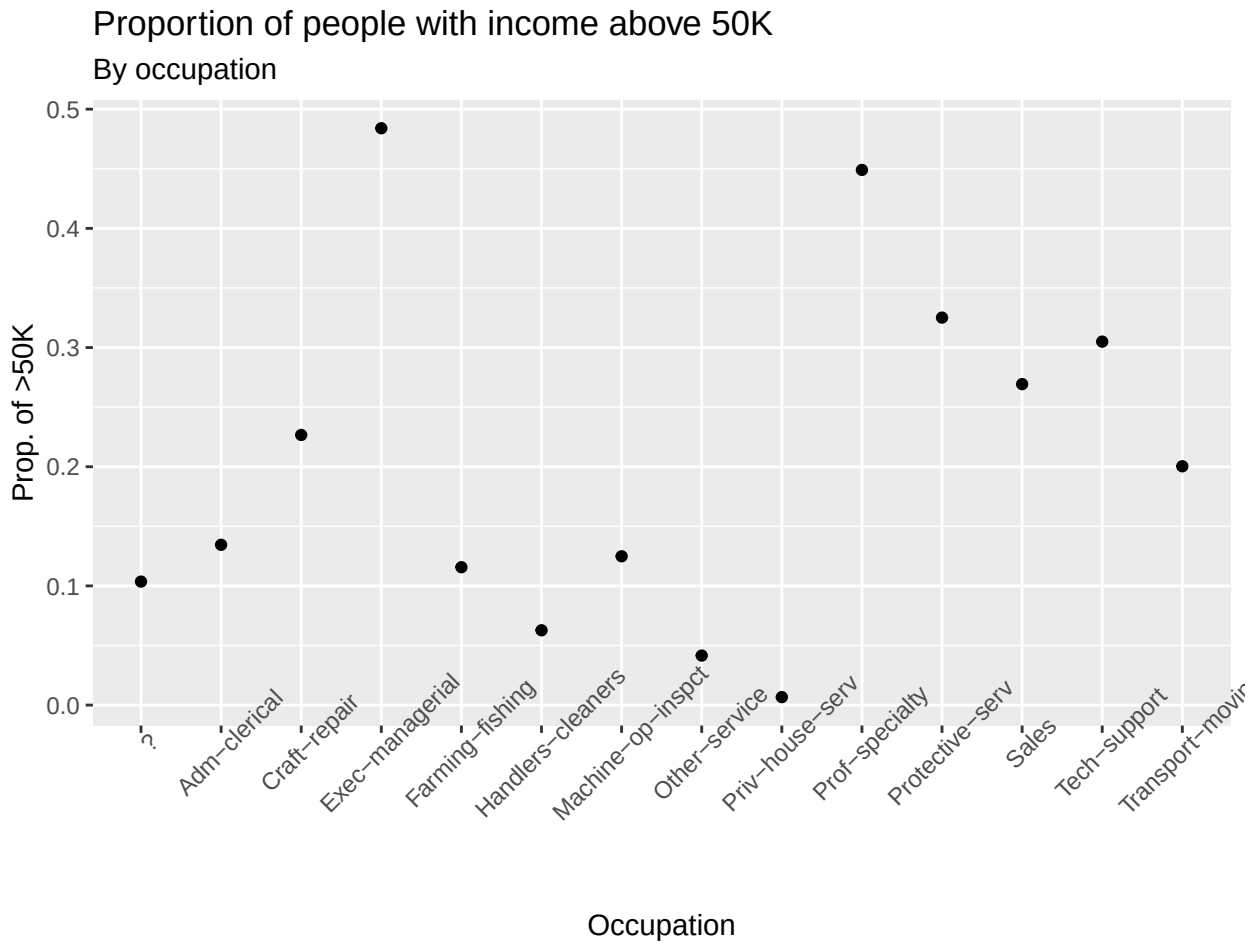
Apparently, more years of education correlates with greater probability of earning more than 50K a year. The relationship seems to be quadratic or exponential.

```r
# Marital status
adult %>%
  mutate(x = marital.status) %>%
  group_by(x) %>%
  filter(n() >= 10) %>%
  summarise(prop = mean(income == " >50K")) %>%
  ggplot(aes(x, prop)) +
  geom_point() +
  ggtitle("Proportion of people with income above 50K",
          "By marital status") +
  labs(x="Marital status", y="Prop. of >50K") +
  theme(axis.text.x = element_text(angle = 45))
```



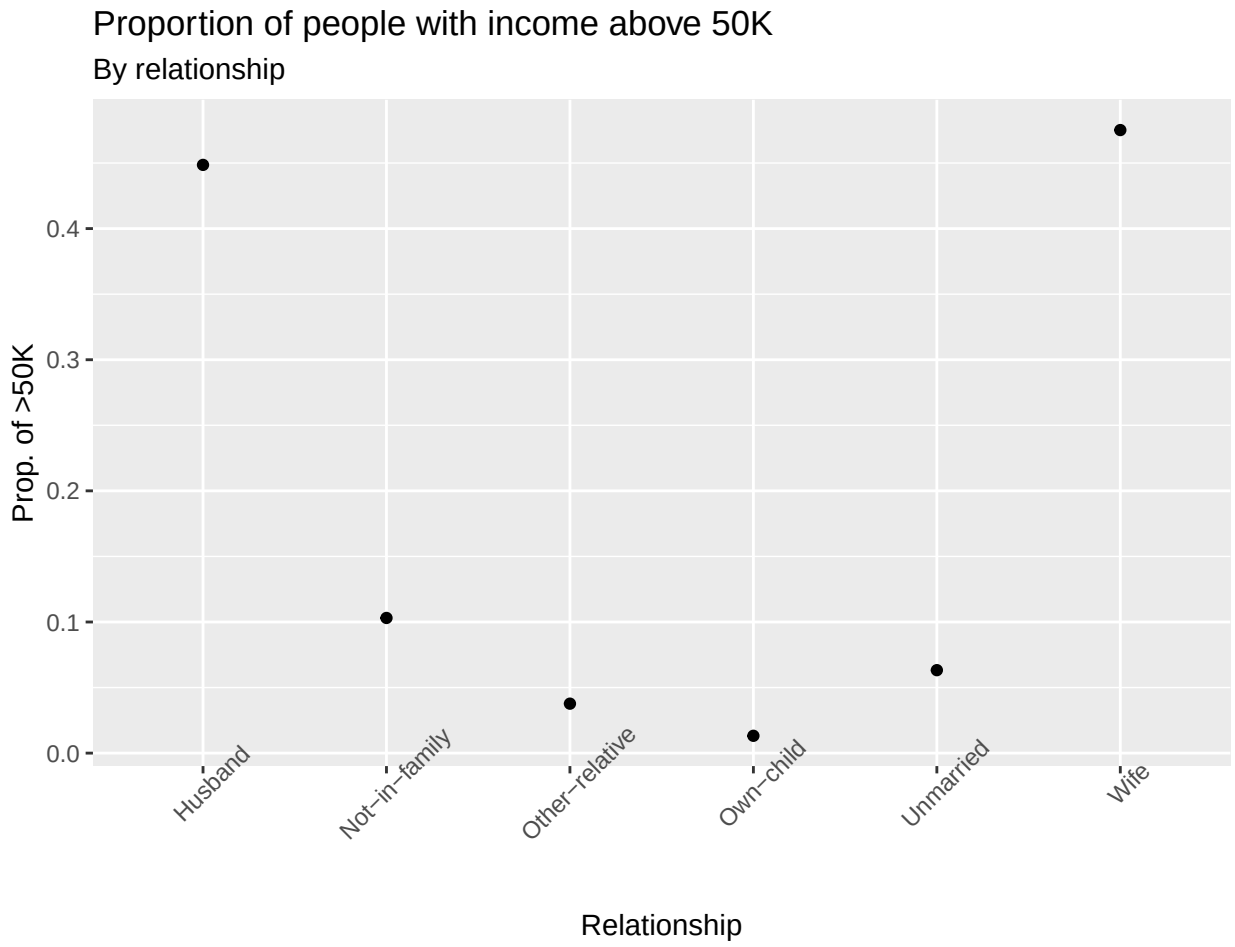
Apparently, married people tend to have a greater probability of earning 50K or more a year than other marital status.

```r
# Occupation
adult %>%
  mutate(x = occupation) %>%
  group_by(x) %>%
  filter(n() >= 10) %>%
  summarise(prop = mean(income == " >50K")) %>%
  ggplot(aes(x, prop)) +
  geom_point() +
  ggtitle("Proportion of people with income above 50K",
          "By occupation") +
  labs(x="Occupation", y="Prop. of >50K") +
  theme(axis.text.x = element_text(angle = 45))
```



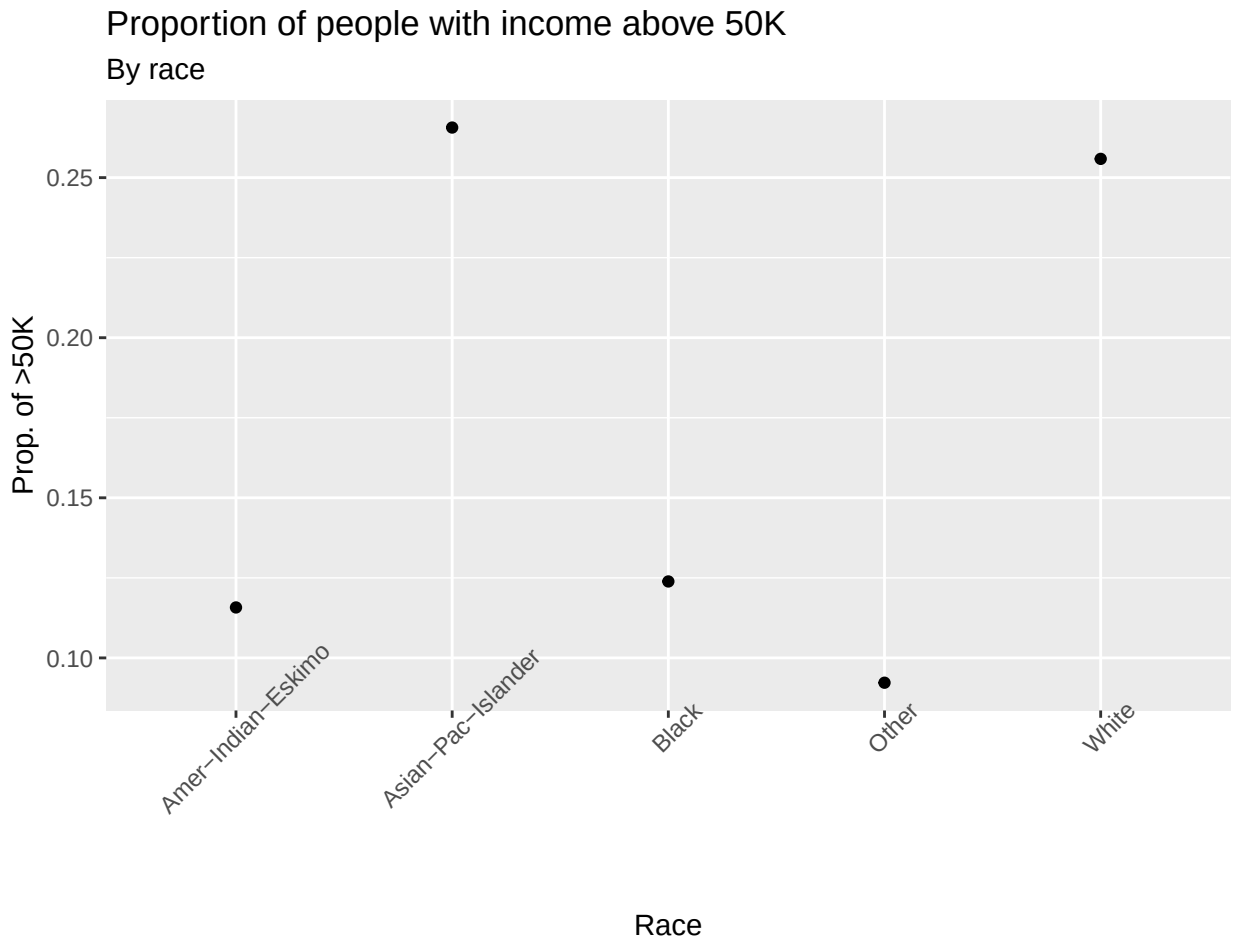
Apparently, the probability of earning more than 50K a year conditional on working on executive or managerial positions or as professionals is greater than conditioning on other occupations.

```r
# Relationship
adult %>%
  mutate(x = relationship) %>%
  group_by(x) %>%
  filter(n() >= 10) %>%
  summarise(prop = mean(income == " >50K")) %>%
  ggplot(aes(x, prop)) +
  geom_point() +
  ggtitle("Proportion of people with income above 50K",
          "By relationship") +
  labs(x="Relationship", y="Prop. of >50K") +
  theme(axis.text.x = element_text(angle = 45))
```



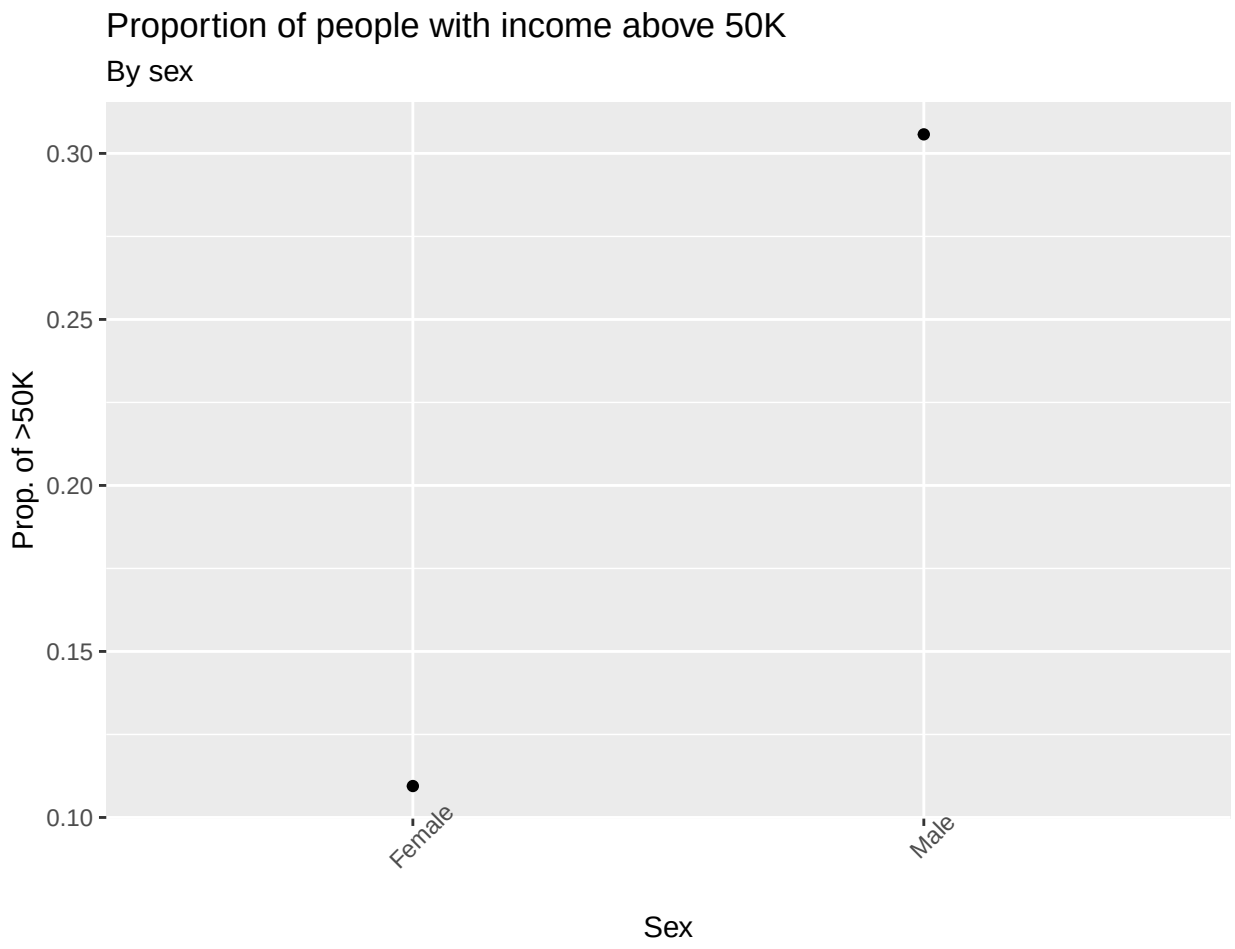
Again, married people seems to have greater probability of earning more than 50K a year.

```r
# Race
adult %>%
  mutate(x = race) %>%
  group_by(x) %>%
  filter(n() >= 10) %>%
  summarise(prop = mean(income == " >50K")) %>%
  ggplot(aes(x, prop)) +
  geom_point() +
  ggtitle("Proportion of people with income above 50K",
          "By race") +
  labs(x="Race", y="Prop. of >50K") +
  theme(axis.text.x = element_text(angle = 45))
```



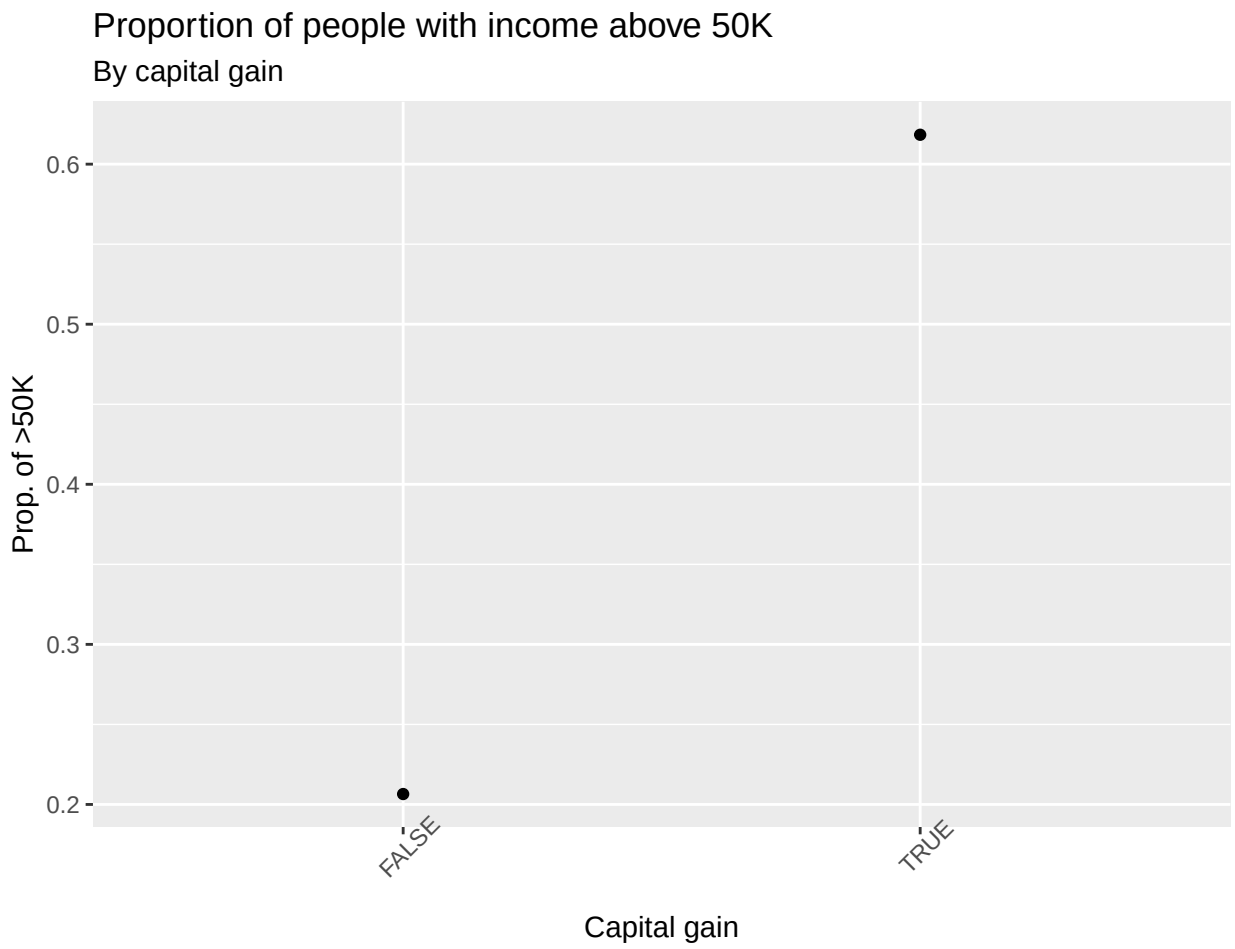
Apparently, the probability of earning more than 50K a year conditional on being white or asian is greater than conditioning on other races.

```r
# Sex
adult %>%
  mutate(x = sex) %>%
  group_by(x) %>%
  filter(n() >= 10) %>%
  summarise(prop = mean(income == " >50K")) %>%
  ggplot(aes(x, prop)) +
  geom_point() +
  ggtitle("Proportion of people with income above 50K",
          "By sex") +
  labs(x="Sex", y="Prop. of >50K") +
  theme(axis.text.x = element_text(angle = 45))
```



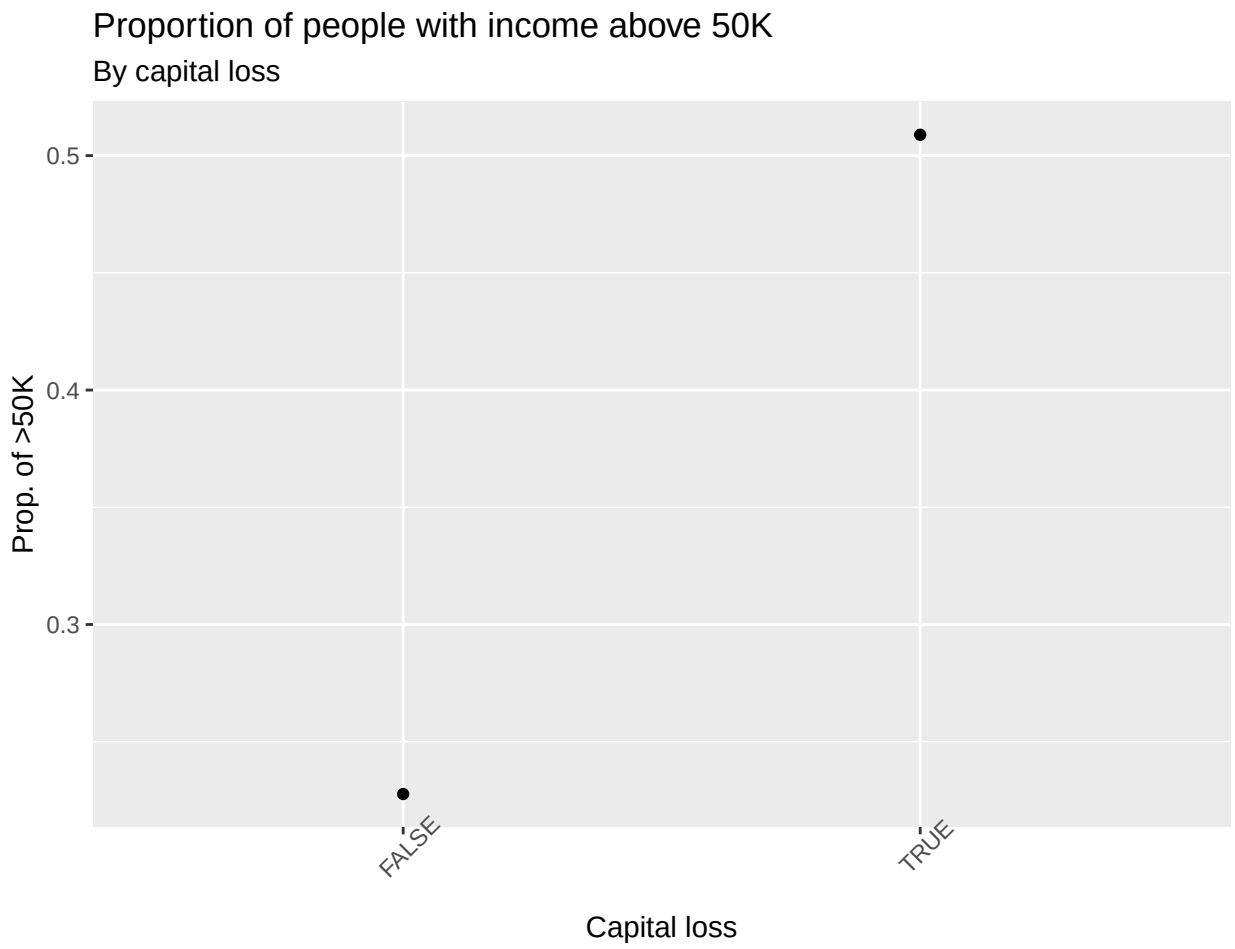
Apparently, males have greater probability of earning more than 50K a year than females.

```r
# Capital gains > 0
adult %>%
  mutate(x = capital.gain>0) %>%
  group_by(x) %>%
  filter(n() >= 10) %>%
  summarise(prop = mean(income == " >50K")) %>%
  ggplot(aes(x, prop)) +
  geom_point() +
  ggtitle("Proportion of people with income above 50K",
          "By capital gain") +
  labs(x="Capital gain", y="Prop. of >50K") +
  theme(axis.text.x = element_text(angle = 45))
```



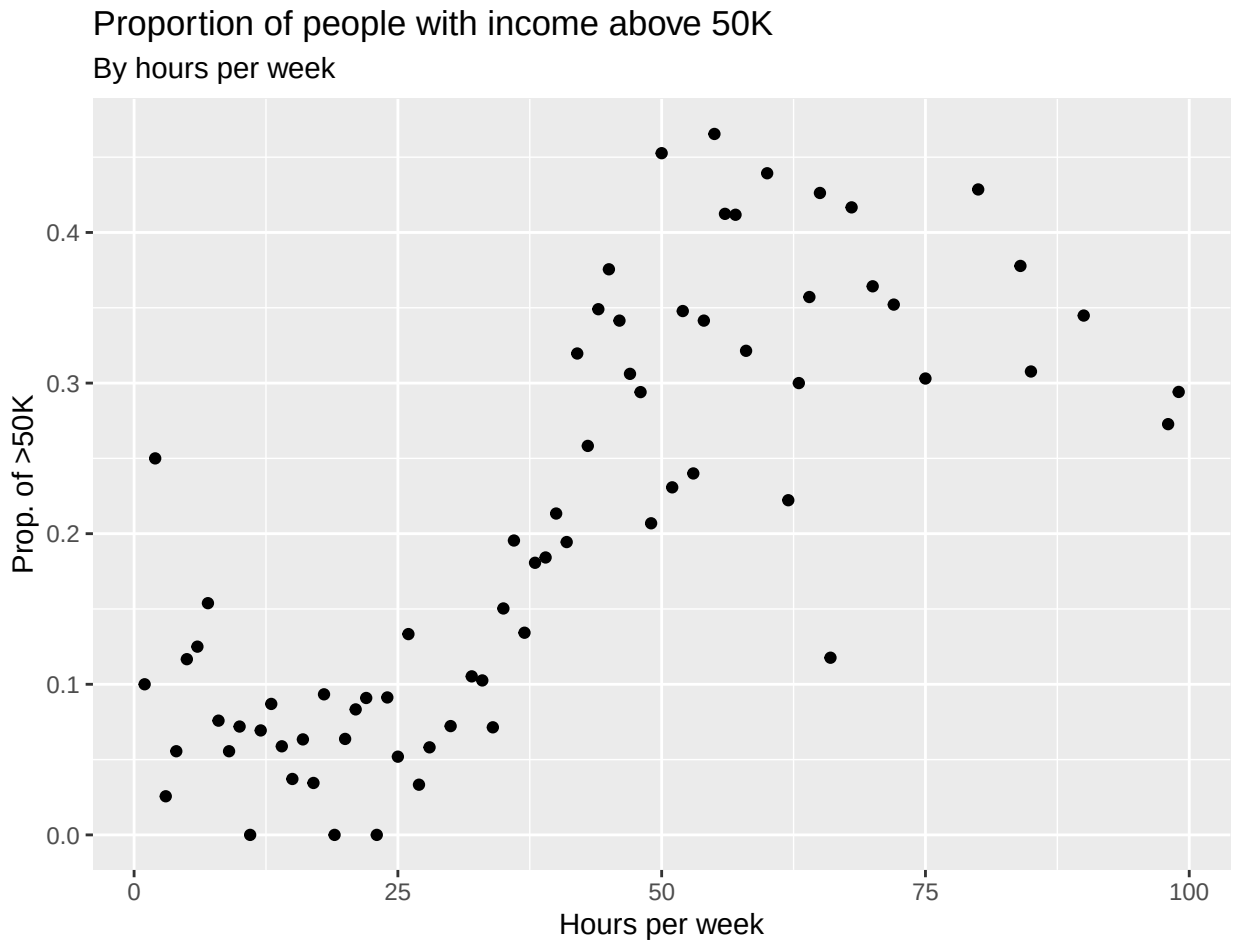
Intuitively, people with capital gains have greater probability of earning 50K or more a year than people without capital gains.

```r
# Capital loss > 0
adult %>%
  mutate(x = capital.loss>0) %>%
  group_by(x) %>%
  filter(n() >= 10) %>%
  summarise(prop = mean(income == " >50K")) %>%
  ggplot(aes(x, prop)) +
  geom_point() +
  ggtitle("Proportion of people with income above 50K",
          "By capital loss") +
  labs(x="Capital loss", y="Prop. of >50K") +
  theme(axis.text.x = element_text(angle = 45))
```



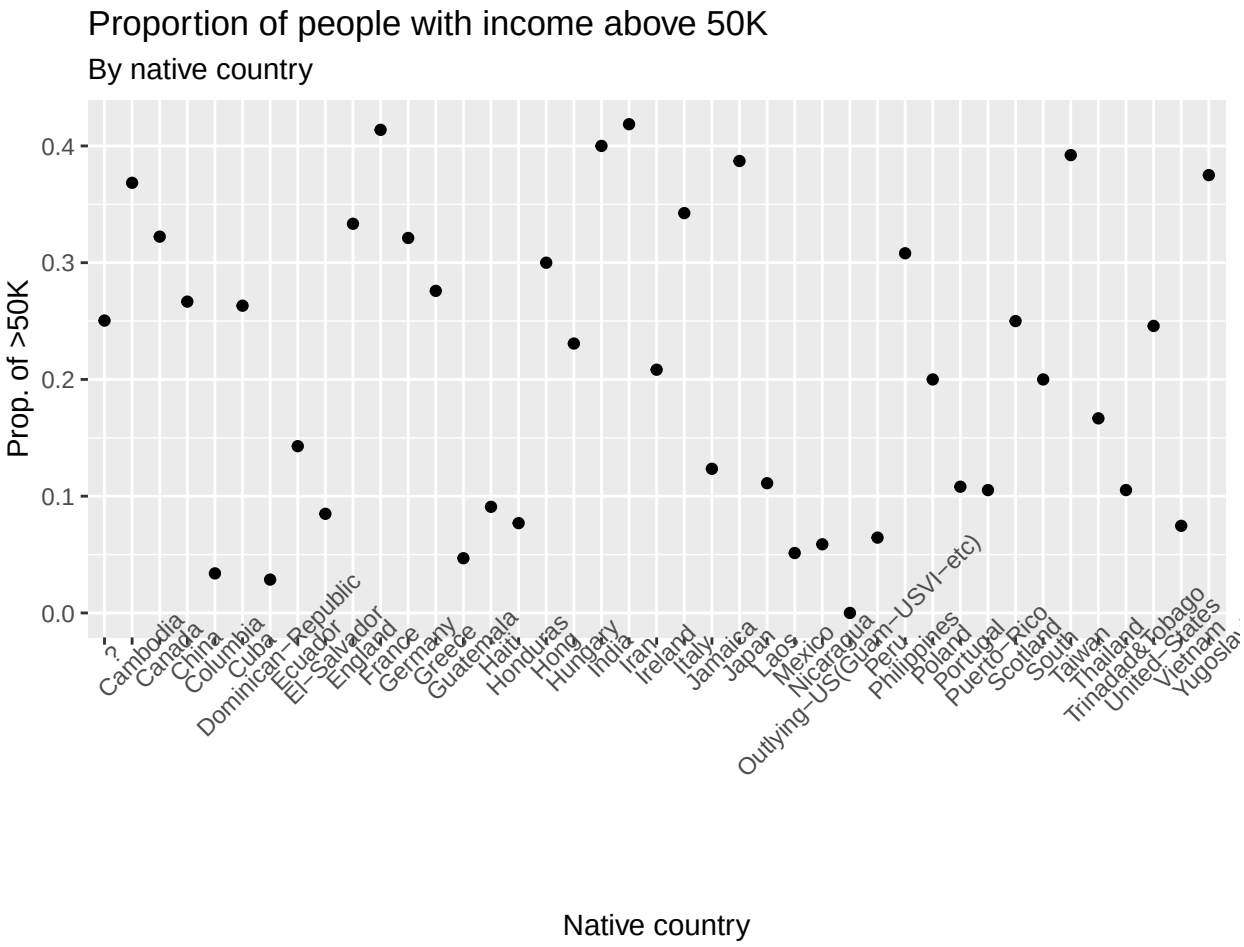
Similar to the previous variable, people with capital losses have greater probability of earning 50K or more a year than people without capital losses. These may reflect the fact that people involved with capital markets are the people with income in the higher percentiles.

```r
# Hours per week
adult %>%
  mutate(x = round(hours.per.week)) %>%
  group_by(x) %>%
  filter(n() >= 10) %>%
  summarise(prop = mean(income == " >50K")) %>%
  ggplot(aes(x, prop)) +
  geom_point() +
  ggtitle("Proportion of people with income above 50K",
          "By hours per week") +
  labs(x="Hours per week", y="Prop. of >50K")
```



Apparently, the relationship between hours per week and income seems to be cubed.

```r
# Native Country
adult %>%
  mutate(x = native.country) %>%
  group_by(x) %>%
  filter(n() >= 10) %>%
  summarise(prop = mean(income == " >50K")) %>%
  ggplot(aes(x, prop)) +
  geom_point() +
  ggtitle("Proportion of people with income above 50K",
          "By native country") +
  labs(x="Native country", y="Prop. of >50K") +
  theme(axis.text.x = element_text(angle = 45))
```



Finally, there is high variability in the probabilities of income above 50K a year controlled for native country. In addition, I find very few observations of people with some native countries for these probabilities to be stable.

Section 4: Data Preparation

In this section, I proceed to prepare the data for modeling and analysis.

First, I trim the string data to avoid problems in the code coming for misspelling spaces.

```
# Triming string data
adult <- adult %>% mutate(
  workclass = str_trim(workclass),
  education = str_trim(education),
  marital.status = str_trim(marital.status),
  occupation = str_trim(occupation),
  relationship = str_trim(relationship),
  race = str_trim(race),
  sex = str_trim(sex),
  native.country = str_trim(native.country),
  income = str_trim(income)
)
```

Secondly, I replace all hyphens with underscore to facilitate coding and avoid having to write backticks in every code line.

```
# Replacing all hyphens with underscore
adult <- adult %>% mutate(
  workclass = str_replace_all(workclass, "-", "_"),
  education = str_replace_all(education, "-", "_"),
  marital.status = str_replace_all(marital.status, "-", "_"),
  occupation = str_replace_all(occupation, "-", "_"),
  relationship = str_replace_all(relationship, "-", "_"),
  race = str_replace_all(race, "-", "_"),
  sex = str_replace_all(sex, "-", "_"),
  native.country = str_replace_all(native.country, "-", "_"),
  income = str_replace_all(income, "-", "_")
)
```

Finally, to start modeling, I create two sets of data from the adult dataset: the train set and the test set.

```
# Training set and test set
set.seed(1)

test_index <- createDataPartition(y = adult$income, times = 1, p = 0.2,
                                  list = FALSE)

train <- adult[-test_index,]
test <- adult[test_index,]

rm(test_index)
```

Section 5: Methods & Analysis

In this section I will proceed with the classification exercise.

After having evaluated the data and since the outcome variable is binary, I decided to use logistic regression as my main machine learning classification technique.

According to Belyadi and Haghighat (2021), Logistic regression is a very powerful supervised machine learning algorithm used for binary classification problems. The best way to think about logistic regression is that it is a linear regression but for classification problems. Logistic regression essentially uses a logistic function to model a binary output variable (Tolles & Meurer, 2016). The primary difference between linear regression and logistic regression is that logistic regression's range is bounded between 0 and 1. In addition, as opposed to linear regression, logistic regression does not require a linear relationship between inputs and output variables. This is due to applying a nonlinear log transformation to the odds ratio.

Method 1: Logistic regression with few variables

In this first attempt, which I will call the *judgement model* or the *parsimonious model*, I will use as predictors only the set of variables I consider may be the most useful for classification. These variables include: age, age squared, years of education, years of education squared, marital status (married or not), race (white/asian or not), capital gains, capital losses, sex (male or not), hours per week, hours per week squared and hours per week cubed.

First I prepare the datasets to include some transformed variables.

```
train1 <- train %>% mutate(
  agesq = age^2,
  education.numsq = education.num^2,
  married=ifelse(marital.status=="Married_civ_spouse"|
                   marital.status=="Married_AF_spouse", 1, 0),
  whiteasian=ifelse(race=="White"|
                      race=="Asian_Pac_Islander", 1, 0),
  male=ifelse(sex=="Male", 1, 0),
  hours.per.weeksq=hours.per.week^2,
  hours.per.weekcub=hours.per.week^3
)

test1 <- test %>% mutate(
  agesq = age^2,
  education.numsq = education.num^2,
  married=ifelse(marital.status=="Married_civ_spouse"|
                   marital.status=="Married_AF_spouse", 1, 0),
  whiteasian=ifelse(race=="White"|
                      race=="Asian_Pac_Islander", 1, 0),
  male=ifelse(sex=="Male", 1, 0),
  hours.per.weeksq=hours.per.week^2,
  hours.per.weekcub=hours.per.week^3
)
```

Then I proceed to run the regression.

```
fit_glm1 <- train1 %>% mutate(y=as.numeric(income==">50K")) %>%
  glm(y ~ age + agesq + education.num + education.numsq +
        married + whiteasian + capital.gain + capital.loss +
        male + hours.per.week + hours.per.weeksq +
        hours.per.weekcub,
      data=., family = "binomial")

summary(fit_glm1)
```

```
##
## Call:
## glm(formula = y ~ age + agesq + education.num + education.numsq +
##     married + whiteasian + capital.gain + capital.loss + male +
##     hours.per.week + hours.per.weeksq + hours.per.weekcub, family = "binomial",
##     data = .)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -4.099  -0.519  -0.220  -0.039   3.684
##
## Coefficients:
##                       Estimate   Std. Error z value             Pr(>|z|)
## (Intercept)       -13.46443209   0.43455501  -30.98 < 0.0000000000000002 ***
## age                 0.23420672   0.01141929   20.51 < 0.0000000000000002 ***
## agesq              -0.00229416   0.00012598  -18.21 < 0.0000000000000002 ***
## education.num       0.21856841   0.05072492    4.31            0.0000164 ***
## education.numsq     0.00577219   0.00237396    2.43              0.01504 *
## married             2.25781363   0.05204224   43.38 < 0.0000000000000002 ***
## whiteasian          0.34553164   0.07476298    4.62            0.0000038 ***
## capital.gain        0.00031409   0.00001108   28.34 < 0.0000000000000002 ***
## capital.loss        0.00063588   0.00004044   15.72 < 0.0000000000000002 ***
## male                0.07902494   0.05459833    1.45              0.14779
## hours.per.week      0.03463712   0.01593200    2.17              0.02970 *
## hours.per.weeksq    0.00057759   0.00034148    1.69              0.09076 .
## hours.per.weekcub  -0.00000834   0.00000228   -3.65              0.00026 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28756  on 26047  degrees of freedom
## Residual deviance: 17174  on 26035  degrees of freedom
## AIC: 17200
##
## Number of Fisher Scoring iterations: 7
```

```
p_hat_glm1 <- predict(fit_glm1, test1, type="response")

y_hat_glm1 <- factor(ifelse(p_hat_glm1 > 0.5, ">50K", "<=50K"))

confusionMatrix(y_hat_glm1, factor(test1$income))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction <=50K >50K
##      <=50K  4657  670
##      >50K    287  899
##
##                Accuracy : 0.853
##                  95% CI : (0.844, 0.862)
##     No Information Rate : 0.759
##     P-Value [Acc > NIR] : <0.0000000000000002
```

```
##
##                   Kappa : 0.562
##
##  Mcnemar's Test P-Value : <0.0000000000000002
##
##             Sensitivity : 0.942
##             Specificity : 0.573
##          Pos Pred Value : 0.874
##          Neg Pred Value : 0.758
##              Prevalence : 0.759
##          Detection Rate : 0.715
##    Detection Prevalence : 0.818
##       Balanced Accuracy : 0.757
##
##        'Positive' Class : <=50K
##
```

The accuracy I obtained with this first model is 0.853.

Method 2: Logistic regression with all variables of the dataset

In this second attempt, which I will call the *data driven model* or the *long model*, I will use as predictors all the variables in the dataset. In order to properly use the character variables I will create dummy variables for each category in each variable. These is done with the *dummy_cols* function from the *FastDummies* package.

First I prepare the data.

```
train2 <- dummy_cols(train,
                     select_columns = c("workclass", "education",
                                        "marital.status", "occupation",
                                        "relationship", "race", "sex",
                                        "native.country", "income"),
                     remove_first_dummy = TRUE) %>%
  select(-workclass, -fnlwgt, -education, -marital.status,
         -occupation, -relationship, -race, -sex,-native.country,
         -income)

test2 <- dummy_cols(test, select_columns = c("workclass", "education",
                                        "marital.status", "occupation",
                                        "relationship", "race", "sex",
                                        "native.country", "income"),
                     remove_first_dummy = TRUE) %>%
  select(-workclass, -fnlwgt, -education, -marital.status,
         -occupation, -relationship, -race, -sex, -native.country,
         -income)

train2 <- train2 %>% select(-`native.country_Holand_Netherlands`)
```

Secondly, I run the regression.

```
fit_glm2 <- train2 %>% glm(`income_>50K` ~ ., data=., family = "binomial")

summary(fit_glm2)
```

```
##
## Call:
## glm(formula = `income_>50K` ~ ., family = "binomial", data = .)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -4.281  -0.511  -0.187  -0.024   3.532
##
## Coefficients: (2 not defined because of singularities)
##                                          Estimate  Std. Error z value
## (Intercept)                            -9.6549839   0.7520319  -12.84
## age                                     0.0233788   0.0018389   12.71
## education.num                           0.1846635   0.0597514    3.09
## capital.gain                            0.0003165   0.0000116   27.40
## capital.loss                            0.0006207   0.0000411   15.10
## hours.per.week                          0.0289980   0.0017933   16.17
## workclass_Federal_gov                   1.1851303   0.1715633    6.91
## workclass_Local_gov                     0.4948912   0.1575982    3.14
## workclass_Never_worked                -11.9476916 593.9604756   -0.02
## workclass_Private                       0.6876457   0.1405871    4.89
## workclass_Self_emp_inc                  0.8523215   0.1677914    5.08
## workclass_Self_emp_not_inc              0.1641796   0.1537729    1.07
## workclass_State_gov                     0.4385043   0.1688853    2.60
## workclass_Without_pay                 -12.8955393 363.6908502   -0.04
## education_5th_6th                      -0.0164413   0.5095386   -0.03
## education_7th_8th                      -0.3812364   0.4052789   -0.94
## education_9th                          -0.4060953   0.3799152   -1.07
## education_10th                         -0.2885065   0.2889739   -1.00
## education_11th                         -0.5841009   0.2467584   -2.37
## education_12th                         -0.3438924   0.2768319   -1.24
## education_Assoc_acdm                   -0.1847198   0.1645059   -1.12
## education_Assoc_voc                     0.0237135   0.1148136    0.21
## education_Bachelors                     0.2174061   0.1929675    1.13
## education_Doctorate                     0.7323486   0.4004504    1.83
## education_HS_grad                      -0.1410084   0.0787274   -1.79
## education_Masters                       0.3615877   0.2586604    1.40
## education_Preschool                   -19.0994542 176.0884680   -0.11
## education_Prof_school                   0.6471995   0.3374521    1.92
## education_Some_college                         NA          NA      NA
## marital.status_Married_AF_spouse        2.8889961   0.6005337    4.81
## marital.status_Married_civ_spouse       2.2718582   0.2823468    8.05
## marital.status_Married_spouse_absent   -0.0092718   0.2616599   -0.04
## marital.status_Never_married           -0.4476439   0.0968512   -4.62
## marital.status_Separated               -0.0705996   0.1787230   -0.40
## marital.status_Widowed                  0.1998188   0.1697869    1.18
## occupation_Adm_clerical                 0.0527992   0.1101446    0.48
## occupation_Armed_Forces                -0.5370798   1.7424007   -0.31
## occupation_Craft_repair                 0.1448616   0.0944455    1.53
## occupation_Exec_managerial              0.8566367   0.0967816    8.85
## occupation_Farming_fishing             -0.9839398   0.1613372   -6.10
## occupation_Handlers_cleaners           -0.5853220   0.1599025   -3.66
## occupation_Machine_op_inspct           -0.2188106   0.1184886   -1.85
## occupation_Other_service               -0.7536962   0.1390243   -5.42
## occupation_Priv_house_serv             -4.0313626   1.7384615   -2.32
```

```
## occupation_Prof_specialty                       0.5823237   0.1037838    5.61
## occupation_Protective_serv                      0.5731710   0.1436604    3.99
## occupation_Sales                                0.3724403   0.1001568    3.72
## occupation_Tech_support                         0.7095995   0.1324824    5.36
## occupation_Transport_moving                            NA          NA      NA
## relationship_Not_in_family                      0.6171493   0.2790737    2.21
## relationship_Other_relative                    -0.4061228   0.2678252   -1.52
## relationship_Own_child                         -0.5675879   0.2724092   -2.08
## relationship_Unmarried                          0.5024325   0.2975421    1.69
## relationship_Wife                               1.2632229   0.1138223   11.10
## race_Asian_Pac_Islander                         0.5144198   0.3090431    1.66
## race_Black                                      0.3705851   0.2650855    1.40
## race_Other                                      0.2571858   0.3909034    0.66
## race_White                                      0.5418823   0.2533128    2.14
## sex_Male                                        0.8211634   0.0872213    9.41
## native.country_Cambodia                         1.2496493   0.7068038    1.77
## native.country_Canada                           0.4370243   0.3433646    1.27
## native.country_China                           -0.3491316   0.4252149   -0.82
## native.country_Columbia                        -2.4822383   1.1187920   -2.22
## native.country_Cuba                             0.7292349   0.3725434    1.96
## native.country_Dominican_Republic            -12.7229688 166.7582088   -0.08
## native.country_Ecuador                          0.0696125   0.7429601    0.09
## native.country_El_Salvador                     -0.1399527   0.5152831   -0.27
## native.country_England                          0.5889333   0.3599578    1.64
## native.country_France                           0.9940926   0.5749789    1.73
## native.country_Germany                          0.5666905   0.3261513    1.74
## native.country_Greece                          -0.4052179   0.6213157   -0.65
## native.country_Guatemala                       -1.7205715   1.3514929   -1.27
## native.country_Haiti                            0.1413198   0.7749915    0.18
## native.country_Honduras                        -0.8836845   2.3587321   -0.37
## native.country_Hong                             0.6710829   0.8634574    0.78
## native.country_Hungary                          0.8467189   0.8669004    0.98
## native.country_India                           -0.3950379   0.3712021   -1.06
## native.country_Iran                             0.1551248   0.5361149    0.29
## native.country_Ireland                          1.0338689   0.6698821    1.54
## native.country_Italy                            1.0960350   0.3680650    2.98
## native.country_Jamaica                          0.3456985   0.5055293    0.68
## native.country_Japan                            0.9174104   0.4496644    2.04
## native.country_Laos                             0.1042731   0.8682376    0.12
## native.country_Mexico                          -0.1070979   0.2767742   -0.39
## native.country_Nicaragua                       -0.1432833   0.7953188   -0.18
## `native.country_Outlying_US(Guam_USVI_etc)` -13.2159883 433.8433570   -0.03
## native.country_Peru                            -0.3462885   0.8753835   -0.40
## native.country_Philippines                      0.6815933   0.3140313    2.17
## native.country_Poland                           0.4692404   0.4645571    1.01
## native.country_Portugal                        -0.1989897   0.9066619   -0.22
## native.country_Puerto_Rico                     -0.1331590   0.4617211   -0.29
## native.country_Scotland                         0.0504497   0.8911914    0.06
## native.country_South                           -0.7178996   0.4936045   -1.45
## native.country_Taiwan                           0.4757430   0.5132390    0.93
## native.country_Thailand                      -12.3265334 405.3390077   -0.03
## `native.country_Trinadad&Tobago`               -0.8896502   1.1402048   -0.78
## native.country_United_States                    0.4814201   0.1555599    3.09
## native.country_Vietnam                         -0.6690706   0.7076770   -0.95
```

```
## native.country_Yugoslavia                    0.3059157   0.8007677    0.38
##                                                          Pr(>|z|)
## (Intercept)                                < 0.0000000000000002 ***
## age                                        < 0.0000000000000002 ***
## education.num                                           0.00200 **
## capital.gain                               < 0.0000000000000002 ***
## capital.loss                               < 0.0000000000000002 ***
## hours.per.week                             < 0.0000000000000002 ***
## workclass_Federal_gov                       0.00000000000492123 ***
## workclass_Local_gov                                     0.00169 **
## workclass_Never_worked                                  0.98395
## workclass_Private                           0.00000100201991006 ***
## workclass_Self_emp_inc                      0.00000037812993292 ***
## workclass_Self_emp_not_inc                              0.28567
## workclass_State_gov                                     0.00942 **
## workclass_Without_pay                                   0.97171
## education_5th_6th                                       0.97426
## education_7th_8th                                       0.34687
## education_9th                                           0.28511
## education_10th                                          0.31809
## education_11th                                          0.01793 *
## education_12th                                          0.21415
## education_Assoc_acdm                                    0.26149
## education_Assoc_voc                                     0.83637
## education_Bachelors                                     0.25989
## education_Doctorate                                     0.06743 .
## education_HS_grad                                       0.07328 .
## education_Masters                                       0.16214
## education_Preschool                                     0.91363
## education_Prof_school                                   0.05512 .
## education_Some_college                                      NA
## marital.status_Married_AF_spouse            0.00000150391789601 ***
## marital.status_Married_civ_spouse           0.00000000000000085 ***
## marital.status_Married_spouse_absent                    0.97173
## marital.status_Never_married                0.00000380100210820 ***
## marital.status_Separated                                0.69283
## marital.status_Widowed                                  0.23924
## occupation_Adm_clerical                                 0.63168
## occupation_Armed_Forces                                 0.75790
## occupation_Craft_repair                                 0.12508
## occupation_Exec_managerial                 < 0.0000000000000002 ***
## occupation_Farming_fishing                  0.00000000106965608 ***
## occupation_Handlers_cleaners                            0.00025 ***
## occupation_Machine_op_inspct                            0.06479 .
## occupation_Other_service                    0.00000005915798740 ***
## occupation_Priv_house_serv                              0.02040 *
## occupation_Prof_specialty                   0.00000002012397234 ***
## occupation_Protective_serv                  0.00006613944277651 ***
## occupation_Sales                                        0.00020 ***
## occupation_Tech_support                     0.00000008499961344 ***
## occupation_Transport_moving                                 NA
## relationship_Not_in_family                              0.02701 *
## relationship_Other_relative                             0.12943
## relationship_Own_child                                  0.03720 *
```

```
## relationship_Unmarried                                  0.09129 .
## relationship_Wife                              < 0.0000000000000002 ***
## race_Asian_Pac_Islander                                 0.09600 .
## race_Black                                              0.16212
## race_Other                                              0.51059
## race_White                                              0.03242 *
## sex_Male                                       < 0.0000000000000002 ***
## native.country_Cambodia                                 0.07706 .
## native.country_Canada                                   0.20310
## native.country_China                                    0.41161
## native.country_Columbia                                 0.02651 *
## native.country_Cuba                                     0.05029 .
## native.country_Dominican_Republic                       0.93918
## native.country_Ecuador                                  0.92535
## native.country_El_Salvador                              0.78593
## native.country_England                                  0.10181
## native.country_France                                   0.08382 .
## native.country_Germany                                  0.08230 .
## native.country_Greece                                   0.51428
## native.country_Guatemala                                0.20299
## native.country_Haiti                                    0.85531
## native.country_Honduras                                 0.70793
## native.country_Hong                                     0.43704
## native.country_Hungary                                  0.32871
## native.country_India                                    0.28723
## native.country_Iran                                     0.77231
## native.country_Ireland                                  0.12274
## native.country_Italy                                    0.00290 **
## native.country_Jamaica                                  0.49408
## native.country_Japan                                    0.04133 *
## native.country_Laos                                     0.90441
## native.country_Mexico                                   0.69879
## native.country_Nicaragua                                0.85703
## `native.country_Outlying_US(Guam_USVI_etc)`             0.97570
## native.country_Peru                                     0.69241
## native.country_Philippines                              0.02997 *
## native.country_Poland                                   0.31246
## native.country_Portugal                                 0.82628
## native.country_Puerto_Rico                              0.77304
## native.country_Scotland                                 0.95486
## native.country_South                                    0.14583
## native.country_Taiwan                                   0.35396
## native.country_Thailand                                 0.97574
## `native.country_Trinadad&Tobago`                        0.43524
## native.country_United_States                            0.00197 **
## native.country_Vietnam                                  0.34443
## native.country_Yugoslavia                               0.70244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28756  on 26047  degrees of freedom
## Residual deviance: 16655  on 25951  degrees of freedom
```

```
## AIC: 16849
##
## Number of Fisher Scoring iterations: 14
```

```
p_hat_glm2 <- predict(fit_glm2, test2, type="response")

y_hat_glm2 <- factor(ifelse(p_hat_glm2 > 0.5, 1, 0))

confusionMatrix(y_hat_glm2, factor(test2$`income_>50K`))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 4652  603
##          1  292  966
##
##                Accuracy : 0.863
##                  95% CI : (0.854, 0.871)
##     No Information Rate : 0.759
##     P-Value [Acc > NIR] : <0.0000000000000002
##
##                   Kappa : 0.597
##
##  Mcnemar's Test P-Value : <0.0000000000000002
##
##             Sensitivity : 0.941
##             Specificity : 0.616
##          Pos Pred Value : 0.885
##          Neg Pred Value : 0.768
##              Prevalence : 0.759
##          Detection Rate : 0.714
##    Detection Prevalence : 0.807
##       Balanced Accuracy : 0.778
##
##        'Positive' Class : 0
##
```

The accuracy I obtained with this second model is 0.863.

Method 3: Logistic regression using all information of the dataset and using some judgement

In this third attempt, I combine the methods 1 and 2. This means, to use all the information of the dataset and to include some transformed variables as well.

First I prepare the data.

```
train3 <- train2 %>% mutate(
  agesq = age^2,
  education.numsq = education.num^2,
  hours.per.weeksq=hours.per.week^2,
  hours.per.weekcub=hours.per.week^3
)
```

```
test3 <- test2 %>% mutate(
  agesq = age^2,
  education.numsq = education.num^2,
  hours.per.weeksq=hours.per.week^2,
  hours.per.weekcub=hours.per.week^3
)
```

Then, I run the regression.

```
fit_glm3 <- train3 %>% glm(`income_>50K` ~ ., data=., family = "binomial")

summary(fit_glm3)
```

```
##
## Call:
## glm(formula = `income_>50K` ~ ., family = "binomial", data = .)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -4.326  -0.493  -0.173  -0.015   3.508
##
## Coefficients: (3 not defined because of singularities)
##                                          Estimate   Std. Error z value
## (Intercept)                           -14.14835237   0.82392515  -17.17
## age                                     0.22537308   0.01217152   18.52
## education.num                           0.18181319   0.06016510    3.02
## capital.gain                            0.00031987   0.00001164   27.48
## capital.loss                            0.00061904   0.00004153   14.91
## hours.per.week                          0.02711707   0.01645484    1.65
## workclass_Federal_gov                   0.69325613   0.17477464    3.97
## workclass_Local_gov                     0.02425482   0.16073378    0.15
## workclass_Never_worked                -12.03820003 578.43652886   -0.02
## workclass_Private                       0.27425069   0.14303156    1.92
## workclass_Self_emp_inc                  0.47002676   0.17037108    2.76
## workclass_Self_emp_not_inc             -0.19385446   0.15587517   -1.24
## workclass_State_gov                    -0.00409837   0.17202424   -0.02
## workclass_Without_pay                 -12.88252391 362.82054052   -0.04
## education_5th_6th                       0.00394306   0.51029942    0.01
## education_7th_8th                      -0.26959192   0.40752816   -0.66
## education_9th                          -0.34225056   0.38057824   -0.90
## education_10th                         -0.21039389   0.29077483   -0.72
## education_11th                         -0.52521490   0.24875355   -2.11
## education_12th                         -0.34716937   0.27910633   -1.24
## education_Assoc_acdm                   -0.22073319   0.16641186   -1.33
## education_Assoc_voc                    -0.02149996   0.11615912   -0.19
## education_Bachelors                     0.20942585   0.19453587    1.08
## education_Doctorate                     0.78622168   0.40389817    1.95
## education_HS_grad                      -0.14814152   0.07953420   -1.86
## education_Masters                       0.30966481   0.26068710    1.19
## education_Preschool                   -19.05129408 185.56264049   -0.10
## education_Prof_school                   0.71451989   0.34046550    2.10
```

```
## education_Some_college                          NA          NA      NA
## marital.status_Married_AF_spouse          3.59639883   0.60299571    5.96
## marital.status_Married_civ_spouse         2.42559045   0.28586387    8.49
## marital.status_Married_spouse_absent      0.08753595   0.26337369    0.33
## marital.status_Never_married             -0.15593376   0.09787727   -1.59
## marital.status_Separated                 -0.01028481   0.17953651   -0.06
## marital.status_Widowed                    0.65758054   0.16986840    3.87
## occupation_Adm_clerical                   0.07436518   0.11161277    0.67
## occupation_Armed_Forces                  -0.49524669   1.87764938   -0.26
## occupation_Craft_repair                   0.11053833   0.09553393    1.16
## occupation_Exec_managerial                0.86383953   0.09790772    8.82
## occupation_Farming_fishing               -0.84005285   0.16122513   -5.21
## occupation_Handlers_cleaners             -0.53902377   0.16143310   -3.34
## occupation_Machine_op_inspct             -0.25824336   0.11987157   -2.15
## occupation_Other_service                 -0.68361794   0.14027226   -4.87
## occupation_Priv_house_serv               -3.80228325   2.39456528   -1.59
## occupation_Prof_specialty                 0.60581438   0.10494297    5.77
## occupation_Protective_serv                0.70115680   0.14577503    4.81
## occupation_Sales                          0.40446224   0.10147978    3.99
## occupation_Tech_support                   0.74052658   0.13460654    5.50
## occupation_Transport_moving                     NA          NA      NA
## relationship_Not_in_family                0.68969910   0.28318501    2.44
## relationship_Other_relative              -0.21431832   0.27456943   -0.78
## relationship_Own_child                   -0.28034200   0.27533014   -1.02
## relationship_Unmarried                    0.50751972   0.30159135    1.68
## relationship_Wife                         1.33765347   0.11584897   11.55
## race_Asian_Pac_Islander                   0.54731096   0.31354481    1.75
## race_Black                                0.37233661   0.26921239    1.38
## race_Other                                0.30964684   0.39852839    0.78
## race_White                                0.59422188   0.25747330    2.31
## sex_Male                                  0.84317725   0.08843933    9.53
## native.country_Cambodia                   1.17626326   0.71258571    1.65
## native.country_Canada                     0.49082121   0.34957610    1.40
## native.country_China                     -0.36340534   0.42103296   -0.86
## native.country_Columbia                  -2.29788243   1.09194434   -2.10
## native.country_Cuba                       0.83574925   0.37128526    2.25
## native.country_Dominican_Republic       -12.60920508 166.34224436   -0.08
## native.country_Ecuador                    0.39835523   0.69200408    0.58
## native.country_El_Salvador               -0.18692885   0.52743141   -0.35
## native.country_England                    0.67349601   0.36045752    1.87
## native.country_France                     0.96976850   0.57624026    1.68
## native.country_Germany                    0.58063979   0.32868054    1.77
## native.country_Greece                    -0.42587241   0.62028620   -0.69
## native.country_Guatemala                 -1.83170462   1.40482381   -1.30
## native.country_Haiti                      0.14652507   0.77402947    0.19
## native.country_Honduras                  -1.04166915   2.59601813   -0.40
## native.country_Hong                       0.85415297   0.86567488    0.99
## native.country_Hungary                    1.06128319   0.93689075    1.13
## native.country_India                     -0.44695051   0.37680367   -1.19
## native.country_Iran                       0.13154281   0.53741412    0.24
## native.country_Ireland                    1.25180976   0.66020543    1.90
## native.country_Italy                      1.05022954   0.37238478    2.82
## native.country_Jamaica                    0.39107987   0.50794927    0.77
## native.country_Japan                      0.92262588   0.44932873    2.05
```

```
## native.country_Laos                                0.32325674   0.89116696    0.36
## native.country_Mexico                             -0.07857773   0.27887133   -0.28
## native.country_Nicaragua                          -0.11045909   0.79239041   -0.14
## `native.country_Outlying_US(Guam_USVI_etc)` -12.87778123 432.01096689   -0.03
## native.country_Peru                               -0.32724987   0.88564652   -0.37
## native.country_Philippines                         0.69532167   0.31732658    2.19
## native.country_Poland                              0.62869895   0.46554206    1.35
## native.country_Portugal                           -0.16787882   0.90070312   -0.19
## native.country_Puerto_Rico                        -0.21542190   0.46720587   -0.46
## native.country_Scotland                           -0.07496038   0.91507622   -0.08
## native.country_South                              -0.67508749   0.49268935   -1.37
## native.country_Taiwan                              0.57081220   0.53548046    1.07
## native.country_Thailand                          -12.40554321 400.30977780   -0.03
## `native.country_Trinadad&Tobago`                  -1.00261615   1.13072020   -0.89
## native.country_United_States                       0.51007432   0.15679502    3.25
## native.country_Vietnam                            -0.52870602   0.70484386   -0.75
## native.country_Yugoslavia                          0.26012722   0.79449006    0.33
## agesq                                             -0.00220727   0.00013309  -16.58
## education.numsq                                            NA           NA      NA
## hours.per.weeksq                                   0.00056910   0.00035307    1.61
## hours.per.weekcub                                 -0.00000737   0.00000236   -3.13
##                                                             Pr(>|z|)
## (Intercept)                                    < 0.0000000000000002 ***
## age                                            < 0.0000000000000002 ***
## education.num                                               0.00251 **
## capital.gain                                   < 0.0000000000000002 ***
## capital.loss                                   < 0.0000000000000002 ***
## hours.per.week                                              0.09936 .
## workclass_Federal_gov                                  0.0000729139 ***
## workclass_Local_gov                                         0.88005
## workclass_Never_worked                                      0.98340
## workclass_Private                                           0.05519 .
## workclass_Self_emp_inc                                      0.00580 **
## workclass_Self_emp_not_inc                                  0.21363
## workclass_State_gov                                         0.98099
## workclass_Without_pay                                       0.97168
## education_5th_6th                                           0.99383
## education_7th_8th                                           0.50827
## education_9th                                               0.36850
## education_10th                                              0.46933
## education_11th                                              0.03474 *
## education_12th                                              0.21355
## education_Assoc_acdm                                        0.18470
## education_Assoc_voc                                         0.85316
## education_Bachelors                                         0.28169
## education_Doctorate                                         0.05158 .
## education_HS_grad                                           0.06252 .
## education_Masters                                           0.23488
## education_Preschool                                         0.91823
## education_Prof_school                                       0.03585 *
## education_Some_college                                         NA
## marital.status_Married_AF_spouse                       0.0000000025 ***
## marital.status_Married_civ_spouse              < 0.0000000000000002 ***
## marital.status_Married_spouse_absent                        0.73961
```

```
## marital.status_Never_married                          0.11113
## marital.status_Separated                              0.95432
## marital.status_Widowed                                0.00011 ***
## occupation_Adm_clerical                               0.50523
## occupation_Armed_Forces                               0.79197
## occupation_Craft_repair                               0.24725
## occupation_Exec_managerial                < 0.0000000000000002 ***
## occupation_Farming_fishing                       0.0000001884 ***
## occupation_Handlers_cleaners                          0.00084 ***
## occupation_Machine_op_inspct                          0.03121 *
## occupation_Other_service                         0.0000010963 ***
## occupation_Priv_house_serv                            0.11231
## occupation_Prof_specialty                        0.0000000078 ***
## occupation_Protective_serv                       0.0000015104 ***
## occupation_Sales                                 0.0000672974 ***
## occupation_Tech_support                          0.0000000377 ***
## occupation_Transport_moving                                NA
## relationship_Not_in_family                            0.01487 *
## relationship_Other_relative                           0.43506
## relationship_Own_child                                0.30858
## relationship_Unmarried                                0.09241 .
## relationship_Wife                         < 0.0000000000000002 ***
## race_Asian_Pac_Islander                               0.08089 .
## race_Black                                            0.16665
## race_Other                                            0.43717
## race_White                                            0.02100 *
## sex_Male                                  < 0.0000000000000002 ***
## native.country_Cambodia                               0.09880 .
## native.country_Canada                                 0.16030
## native.country_China                                  0.38807
## native.country_Columbia                               0.03534 *
## native.country_Cuba                                   0.02439 *
## native.country_Dominican_Republic                     0.93958
## native.country_Ecuador                                0.56485
## native.country_El_Salvador                            0.72303
## native.country_England                                0.06170 .
## native.country_France                                 0.09239 .
## native.country_Germany                                0.07730 .
## native.country_Greece                                 0.49235
## native.country_Guatemala                              0.19228
## native.country_Haiti                                  0.84986
## native.country_Honduras                               0.68823
## native.country_Hong                                   0.32379
## native.country_Hungary                                0.25731
## native.country_India                                  0.23556
## native.country_Iran                                   0.80663
## native.country_Ireland                                0.05795 .
## native.country_Italy                                  0.00480 **
## native.country_Jamaica                                0.44135
## native.country_Japan                                  0.04004 *
## native.country_Laos                                   0.71680
## native.country_Mexico                                 0.77812
## native.country_Nicaragua                              0.88913
## `native.country_Outlying_US(Guam_USVI_etc)`           0.97622
```

```
## native.country_Peru                                        0.71175
## native.country_Philippines                                 0.02844 *
## native.country_Poland                                      0.17687
## native.country_Portugal                                    0.85214
## native.country_Puerto_Rico                                 0.64474
## native.country_Scotland                                    0.93471
## native.country_South                                       0.17062
## native.country_Taiwan                                      0.28643
## native.country_Thailand                                    0.97528
## `native.country_Trinadad&Tobago`                           0.37524
## native.country_United_States                               0.00114 **
## native.country_Vietnam                                     0.45319
## native.country_Yugoslavia                                  0.74335
## agesq                                        < 0.0000000000000002 ***
## education.numsq                                              NA
## hours.per.weeksq                                           0.10699
## hours.per.weekcub                                          0.00177 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28756  on 26047  degrees of freedom
## Residual deviance: 16232  on 25948  degrees of freedom
## AIC: 16432
##
## Number of Fisher Scoring iterations: 14
```

```
p_hat_glm3 <- predict(fit_glm3, test3, type="response")

y_hat_glm3 <- factor(ifelse(p_hat_glm3 > 0.5, 1, 0))

confusionMatrix(y_hat_glm3, factor(test3$`income_>50K`))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 4668  595
##          1  276  974
##
##                Accuracy : 0.866
##                  95% CI : (0.858, 0.874)
##     No Information Rate : 0.759
##     P-Value [Acc > NIR] : <0.0000000000000002
##
##                   Kappa : 0.607
##
##  Mcnemar's Test P-Value : <0.0000000000000002
##
##             Sensitivity : 0.944
##             Specificity : 0.621
##          Pos Pred Value : 0.887
##          Neg Pred Value : 0.779
```

```
##              Prevalence : 0.759
##          Detection Rate : 0.717
##    Detection Prevalence : 0.808
##       Balanced Accuracy : 0.782
##
##        'Positive' Class : 0
##
```

The accuracy I obtained with this third model is 0.866.

Section 6: Final Validation

After the analysis conducted in the last section, I decide to choose the third model as the preferred one, given that it is the model that exerted the highest accuracy.

Now I will conduct the classification exercise using the validation set to find the final accuracy of this exercise.

First, I prepare the data.

```
adult <- dummy_cols(adult,
                    select_columns = c("workclass", "education",
                                       "marital.status", "occupation",
                                       "relationship", "race", "sex",
                                       "native.country", "income"),
                     remove_first_dummy = TRUE) %>%
  select(-workclass, -fnlwgt, -education, -marital.status,
         -occupation, -relationship, -race, -sex,-native.country,
         -income) %>%
  mutate(agesq = age^2,
         education.numsq = education.num^2,
         hours.per.weeksq=hours.per.week^2,
         hours.per.weekcub=hours.per.week^3) %>%
  select(-native.country_Holand_Netherlands)
```

Then, I run the regression.

```
fit_final <- adult %>% glm(`income_>50K` ~ ., data=., family = "binomial")

summary(fit_final)
```

```
##
## Call:
## glm(formula = `income_>50K` ~ ., family = "binomial", data = .)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -5.061  -0.485  -0.168  -0.016   3.898
##
## Coefficients: (3 not defined because of singularities)
##                                              Estimate   Std. Error z value
## (Intercept)                                -14.37608026   0.77983567  -18.43
## age                                          0.23193238   0.01093464   21.21
## education.num                                0.20444296   0.05908766    3.46
```

```
## capital.gain                          0.00032299   0.00001038   31.12
## capital.loss                          0.00065037   0.00003756   17.32
## hours.per.week                        0.02184492   0.01483384    1.47
## workclass_Federal_gov                 0.59340240   0.15669884    3.79
## workclass_Local_gov                  -0.06463412   0.14315389   -0.45
## workclass_Never_worked              -10.54791306 260.89313542   -0.04
## workclass_Private                     0.18060586   0.12737131    1.42
## workclass_Self_emp_inc                0.36742271   0.15185812    2.42
## workclass_Self_emp_not_inc           -0.26705658   0.13884321   -1.92
## workclass_State_gov                  -0.17258446   0.15452814   -1.12
## workclass_Without_pay               -12.27944563 197.04484851   -0.06
## education_5th_6th                     0.06403459   0.49386838    0.13
## education_7th_8th                    -0.26137262   0.39330311   -0.66
## education_9th                        -0.26100353   0.36203554   -0.72
## education_10th                       -0.25282276   0.27974847   -0.90
## education_11th                       -0.38273539   0.23284745   -1.64
## education_12th                       -0.25858277   0.25037122   -1.03
## education_Assoc_acdm                 -0.26811282   0.15665420   -1.71
## education_Assoc_voc                  -0.05713505   0.10681986   -0.53
## education_Bachelors                   0.14860329   0.18886322    0.79
## education_Doctorate                   0.65004861   0.39118750    1.66
## education_HS_grad                    -0.13992557   0.07532351   -1.86
## education_Masters                     0.24664420   0.25291187    0.98
## education_Preschool                 -20.01942965 320.47325985   -0.06
## education_Prof_school                 0.66247838   0.32855019    2.02
## education_Some_college                       NA           NA      NA
## marital.status_Married_AF_spouse      3.33054404   0.56226422    5.92
## marital.status_Married_civ_spouse     2.36835503   0.26958932    8.79
## marital.status_Married_spouse_absent  0.05091765   0.23192415    0.22
## marital.status_Never_married         -0.19322293   0.08831504   -2.19
## marital.status_Separated             -0.07214279   0.16476909   -0.44
## marital.status_Widowed                0.55842411   0.15361052    3.64
## occupation_Adm_clerical               0.13144714   0.10037260    1.31
## occupation_Armed_Forces              -0.99649178   1.60391960   -0.62
## occupation_Craft_repair               0.13851430   0.08572742    1.62
## occupation_Exec_managerial            0.89635899   0.08816400   10.17
## occupation_Farming_fishing           -0.77044074   0.14204638   -5.42
## occupation_Handlers_cleaners         -0.52983169   0.14712918   -3.60
## occupation_Machine_op_inspct         -0.21026190   0.10723692   -1.96
## occupation_Other_service             -0.66315284   0.12557270   -5.28
## occupation_Priv_house_serv           -3.92843377   2.19476354   -1.79
## occupation_Prof_specialty             0.64435152   0.09463632    6.81
## occupation_Protective_serv            0.80582220   0.13226064    6.09
## occupation_Sales                      0.41526670   0.09129045    4.55
## occupation_Tech_support               0.78868928   0.12117211    6.51
## occupation_Transport_moving                  NA           NA      NA
## relationship_Not_in_family            0.65542671   0.26731672    2.45
## relationship_Other_relative          -0.17775713   0.24921789   -0.71
## relationship_Own_child               -0.36572178   0.26390775   -1.39
## relationship_Unmarried                0.45208981   0.28317732    1.60
## relationship_Wife                     1.42884726   0.10426553   13.70
## race_Asian_Pac_Islander               0.72619213   0.27342784    2.66
## race_Black                            0.47819981   0.23525855    2.03
## race_Other                            0.25928729   0.36074120    0.72
```

```
## race_White                                      0.65960519   0.22444177    2.94
## sex_Male                                        0.88407167   0.08012823   11.03
## native.country_Cambodia                         1.41307756   0.63824545    2.21
## native.country_Canada                           0.55985518   0.30210095    1.85
## native.country_China                           -0.52034819   0.39125129   -1.33
## native.country_Columbia                        -1.81408275   0.80046991   -2.27
## native.country_Cuba                             0.64016705   0.33460875    1.91
## native.country_Dominican_Republic              -1.62152424   1.04498404   -1.55
## native.country_Ecuador                          0.25200221   0.67442163    0.37
## native.country_El_Salvador                     -0.44769808   0.50972092   -0.88
## native.country_England                          0.60322537   0.33386521    1.81
## native.country_France                           0.77054502   0.52259390    1.47
## native.country_Germany                          0.65248731   0.28686574    2.27
## native.country_Greece                          -0.85139919   0.56738151   -1.50
## native.country_Guatemala                        0.05972754   0.77377834    0.08
## native.country_Haiti                            0.07610107   0.68924396    0.11
## native.country_Honduras                        -1.26504116   2.69645145   -0.47
## native.country_Hong                             0.26356215   0.66518315    0.40
## native.country_Hungary                          0.24828510   0.80682381    0.31
## native.country_India                           -0.20240861   0.33432909   -0.61
## native.country_Iran                             0.17033446   0.45184125    0.38
## native.country_Ireland                          0.90649433   0.63515605    1.43
## native.country_Italy                            0.95100244   0.34889435    2.73
## native.country_Jamaica                          0.23144705   0.46556947    0.50
## native.country_Japan                            0.55599863   0.42173312    1.32
## native.country_Laos                            -0.20295513   0.88724987   -0.23
## native.country_Mexico                          -0.26074871   0.25611151   -1.02
## native.country_Nicaragua                       -0.42498446   0.79200530   -0.54
## `native.country_Outlying_US(Guam_USVI_etc)` -11.80090573 213.22037371   -0.06
## native.country_Peru                            -0.53619862   0.87077368   -0.62
## native.country_Philippines                      0.61611112   0.28382083    2.17
## native.country_Poland                           0.30516200   0.42463897    0.72
## native.country_Portugal                         0.22383244   0.64373493    0.35
## native.country_Puerto_Rico                     -0.22605269   0.40801645   -0.55
## native.country_Scotland                         0.06206425   0.81140570    0.08
## native.country_South                           -0.83586620   0.44221642   -1.89
## native.country_Taiwan                           0.31526085   0.48850333    0.65
## native.country_Thailand                        -0.20667438   0.82010316   -0.25
## `native.country_Trinadad&Tobago`               -0.31657505   0.86896262   -0.36
## native.country_United_States                    0.40356873   0.13951607    2.89
## native.country_Vietnam                         -0.87728344   0.61933096   -1.42
## native.country_Yugoslavia                       0.88449827   0.67009596    1.32
## agesq                                          -0.00225595   0.00011943  -18.89
## education.numsq                                         NA           NA      NA
## hours.per.weeksq                                0.00068301   0.00031842    2.15
## hours.per.weekcub                              -0.00000805   0.00000213   -3.78
##                                                           Pr(>|z|)
## (Intercept)                                 < 0.0000000000000002 ***
## age                                         < 0.0000000000000002 ***
## education.num                                             0.00054 ***
## capital.gain                                < 0.0000000000000002 ***
## capital.loss                                < 0.0000000000000002 ***
## hours.per.week                                            0.14085
## workclass_Federal_gov                                     0.00015 ***
```

```
## workclass_Local_gov                                0.65163
## workclass_Never_worked                             0.96775
## workclass_Private                                  0.15621
## workclass_Self_emp_inc                             0.01554 *
## workclass_Self_emp_not_inc                         0.05442 .
## workclass_State_gov                                0.26406
## workclass_Without_pay                              0.95031
## education_5th_6th                                  0.89684
## education_7th_8th                                  0.50633
## education_9th                                      0.47095
## education_10th                                     0.36613
## education_11th                                     0.10023
## education_12th                                     0.30170
## education_Assoc_acdm                               0.08699 .
## education_Assoc_voc                                0.59274
## education_Bachelors                                0.43138
## education_Doctorate                                0.09657 .
## education_HS_grad                                  0.06322 .
## education_Masters                                  0.32945
## education_Preschool                                0.95019
## education_Prof_school                              0.04376 *
## education_Some_college                                  NA
## marital.status_Married_AF_spouse                0.0000000031526 ***
## marital.status_Married_civ_spouse             < 0.0000000000000002 ***
## marital.status_Married_spouse_absent               0.82623
## marital.status_Never_married                       0.02868 *
## marital.status_Separated                           0.66150
## marital.status_Widowed                             0.00028 ***
## occupation_Adm_clerical                            0.19033
## occupation_Armed_Forces                            0.53441
## occupation_Craft_repair                            0.10615
## occupation_Exec_managerial                    < 0.0000000000000002 ***
## occupation_Farming_fishing                      0.0000000583231 ***
## occupation_Handlers_cleaners                       0.00032 ***
## occupation_Machine_op_inspct                       0.04991 *
## occupation_Other_service                        0.0000001284616 ***
## occupation_Priv_house_serv                         0.07347 .
## occupation_Prof_specialty                       0.0000000000098 ***
## occupation_Protective_serv                      0.0000000011103 ***
## occupation_Sales                                0.0000053939544 ***
## occupation_Tech_support                         0.0000000000757 ***
## occupation_Transport_moving                             NA
## relationship_Not_in_family                         0.01421 *
## relationship_Other_relative                        0.47568
## relationship_Own_child                             0.16581
## relationship_Unmarried                             0.11038
## relationship_Wife                             < 0.0000000000000002 ***
## race_Asian_Pac_Islander                            0.00791 **
## race_Black                                         0.04209 *
## race_Other                                         0.47229
## race_White                                         0.00329 **
## sex_Male                                      < 0.0000000000000002 ***
## native.country_Cambodia                            0.02683 *
## native.country_Canada                              0.06385 .
```

```
## native.country_China                                    0.18353
## native.country_Columbia                                 0.02343 *
## native.country_Cuba                                     0.05572 .
## native.country_Dominican_Republic                       0.12073
## native.country_Ecuador                                  0.70866
## native.country_El_Salvador                              0.37977
## native.country_England                                  0.07079 .
## native.country_France                                   0.14036
## native.country_Germany                                  0.02293 *
## native.country_Greece                                   0.13347
## native.country_Guatemala                                0.93847
## native.country_Haiti                                    0.91208
## native.country_Honduras                                 0.63896
## native.country_Hong                                     0.69194
## native.country_Hungary                                  0.75829
## native.country_India                                    0.54490
## native.country_Iran                                     0.70619
## native.country_Ireland                                  0.15352
## native.country_Italy                                    0.00642 **
## native.country_Jamaica                                  0.61910
## native.country_Japan                                    0.18738
## native.country_Laos                                     0.81907
## native.country_Mexico                                   0.30863
## native.country_Nicaragua                                0.59155
## `native.country_Outlying_US(Guam_USVI_etc)`             0.95586
## native.country_Peru                                     0.53804
## native.country_Philippines                              0.02995 *
## native.country_Poland                                   0.47236
## native.country_Portugal                                 0.72806
## native.country_Puerto_Rico                              0.57956
## native.country_Scotland                                 0.93903
## native.country_South                                    0.05873 .
## native.country_Taiwan                                   0.51869
## native.country_Thailand                                 0.80103
## `native.country_Trinadad&Tobago`                        0.71562
## native.country_United_States                            0.00382 **
## native.country_Vietnam                                  0.15663
## native.country_Yugoslavia                               0.18685
## agesq                                  < 0.0000000000000002 ***
## education.numsq                                           NA
## hours.per.weeksq                                        0.03195 *
## hours.per.weekcub                                       0.00016 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 35948  on 32560  degrees of freedom
## Residual deviance: 20039  on 32461  degrees of freedom
## AIC: 20239
##
## Number of Fisher Scoring iterations: 13
```

Finally, I validate the model using the validation set.

```
#### Preparing validation data

# Triming string data

adult.validation <- adult.validation %>% mutate(
  workclass = str_trim(workclass),
  education = str_trim(education),
  marital.status = str_trim(marital.status),
  occupation = str_trim(occupation),
  relationship = str_trim(relationship),
  race = str_trim(race),
  sex = str_trim(sex),
  native.country = str_trim(native.country),
  income = str_trim(income)
)

# Replacing all dots and hyphens with underscore

adult.validation <- adult.validation %>% mutate(
  workclass = str_replace_all(workclass, "-", "_"),
  education = str_replace_all(education, "-", "_"),
  marital.status = str_replace_all(marital.status, "-", "_"),
  occupation = str_replace_all(occupation, "-", "_"),
  relationship = str_replace_all(relationship, "-", "_"),
  race = str_replace_all(race, "-", "_"),
  sex = str_replace_all(sex, "-", "_"),
  native.country = str_replace_all(native.country, "-", "_"),
  income = str_replace_all(income, "-", "_")
)


# Adding relevant variables

adult.validation <- dummy_cols(adult.validation,
                               select_columns = c("workclass", "education",
                                                  "marital.status", "occupation",
                                                  "relationship", "race", "sex",
                                                  "native.country", "income"),
                    remove_first_dummy = TRUE) %>%
  select(-workclass, -fnlwgt, -education, -marital.status,
         -occupation, -relationship, -race, -sex,-native.country,
         -income) %>%
  mutate(agesq = age^2,
         education.numsq = education.num^2,
         hours.per.weeksq=hours.per.week^2,
         hours.per.weekcub=hours.per.week^3)

p_hat_final <- predict(fit_final, adult.validation, type="response")

y_hat_final <- factor(ifelse(p_hat_final > 0.5, 1, 0))

confusionMatrix(y_hat_final, factor(adult.validation$`income_>50K`))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 11595  1496
##          1   840  2350
##
##                Accuracy : 0.857
##                  95% CI : (0.851, 0.862)
##     No Information Rate : 0.764
##     P-Value [Acc > NIR] : <0.0000000000000002
##
##                   Kappa : 0.577
##
##  Mcnemar's Test P-Value : <0.0000000000000002
##
##             Sensitivity : 0.932
##             Specificity : 0.611
##          Pos Pred Value : 0.886
##          Neg Pred Value : 0.737
##              Prevalence : 0.764
##          Detection Rate : 0.712
##    Detection Prevalence : 0.804
##       Balanced Accuracy : 0.772
##
##        'Positive' Class : 0
##
```

The final accuracy of this exercise is 0.857.

Conclusion

The Adult Census Income dataset was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker. It stores the information of adults from the entire US on their socioeconomic and demographic characteristics.

The classification task of this project is to determine if a person's annual income in the United States is greater than 50K or less than 50K, using several socioeconomic and demographic characteristics of the population as predictors.

After loading, exploring and preparing the data, I test three different logistic regression models and choose the one that exerts the highest accuracy. The model chosen is a logistic regression that includes all available information of the dataset and also include some transformed variables in order to take advantage of their non-linear relationship with the outcome variable, income greater than 50K a year.

To finish this report, I validate the chosen model by using a completely different dataset and evaluating the accuracy obtained. The final accuracy of this exercise ends up being 0.857.

To further improve the classification exercise, using other types of models such as random forest, knn, lasso, lda or qda may be useful. However, having into account that the dataset includes several variables that can be used as predictors, a very high computational power may be needed to properly train these types of models.