choose-your-own-project

Robert Young

2/8/2021

This program predicts whether an income exceeds $50K per year

based on census data extracted from the 1994 Census Bureau database

https://www.kaggle.com/uciml/adult-census-income

The three project files and the data files in .csv and .xlsx format

are available in the following GitHub repository:

https://github.com/musician60/choose_your_own_project

Introduction/Overview

Suppress warning messages:

```
options(warn = -1)
```

Install packages and load needed libraries:

```
if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us
.r-project.org")

## Loading required package: tidyverse
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.
3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse_conflict
s() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(tidyverse)
```

Read in the data:

```
savedwd <- getwd()
setwd("C:/Users/ryoung/Desktop")
adult <- read.csv("adult.csv")
setwd(savedwd)
```

Some of the variables in the dataset contain a "?" for missing data. Replace each "?" with "NA"

```
adult[adult == "?"] <- NA
```

Select the variables that will be needed for the program. Replace the NAs in the occupation variable with the word "Other". Select only the observations for the United States.

```
adult <- adult %>%
  select(age, education, marital.status, occupation, race,
         sex, hours.per.week, native.country, income) %>%
  mutate(occupation = ifelse(is.na(occupation), "Other", occupation)) %>%
  filter(native.country == "United-States")
```

Describe the dataset

```
str(adult)

## 'data.frame':    29170 obs. of  9 variables:
##  $ age           : int  90 82 66 54 41 34 38 74 68 45 ...
##  $ education     : chr  "HS-grad" "HS-grad" "Some-college" "7th-8th" ...
##  $ marital.status: chr  "Widowed" "Widowed" "Widowed" "Divorced" ...
##  $ occupation    : chr  "Other" "Exec-managerial" "Other" "Machine-op-insp
ct" ...
##  $ race          : chr  "White" "White" "Black" "White" ...
##  $ sex           : chr  "Female" "Female" "Female" "Female" ...
##  $ hours.per.week: int  40 18 40 40 40 45 40 20 40 35 ...
##  $ native.country: chr  "United-States" "United-States" "United-States" "U
nited-States" ...
##  $ income        : chr  "<=50K" "<=50K" "<=50K" "<=50K" ...
```

The dataset consists of 29170 observations of 9 variables. Variables age and hours.per.week are of type integer and the remaining variables are all of type character. We will be predicting income which has two levels: “<=50K” and “>50K”.

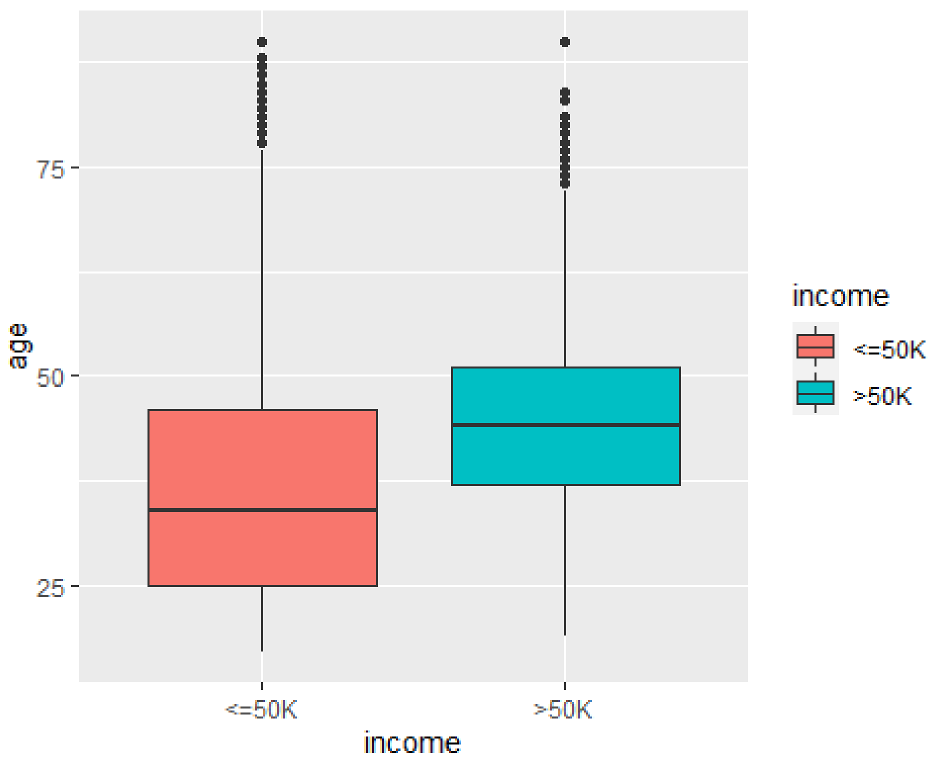
In this program, our goal is to predict whether or not a person earns over $50K, depending on age, education, marital status, occupation, race, sex, and hours worked each week. The variable native.country has already been used to select the observations for the United States and will not be used as a predictor.

After describing how suitable each variable is as a predictor, we will develop a series of models and evaluate them to see how effective each model is in predicting the income level.

**Methods/Analysis

First, we will examine income by age and display a numerical summary.

```
ggplot(adult, aes(income, age, fill = income))+
  geom_boxplot()
```



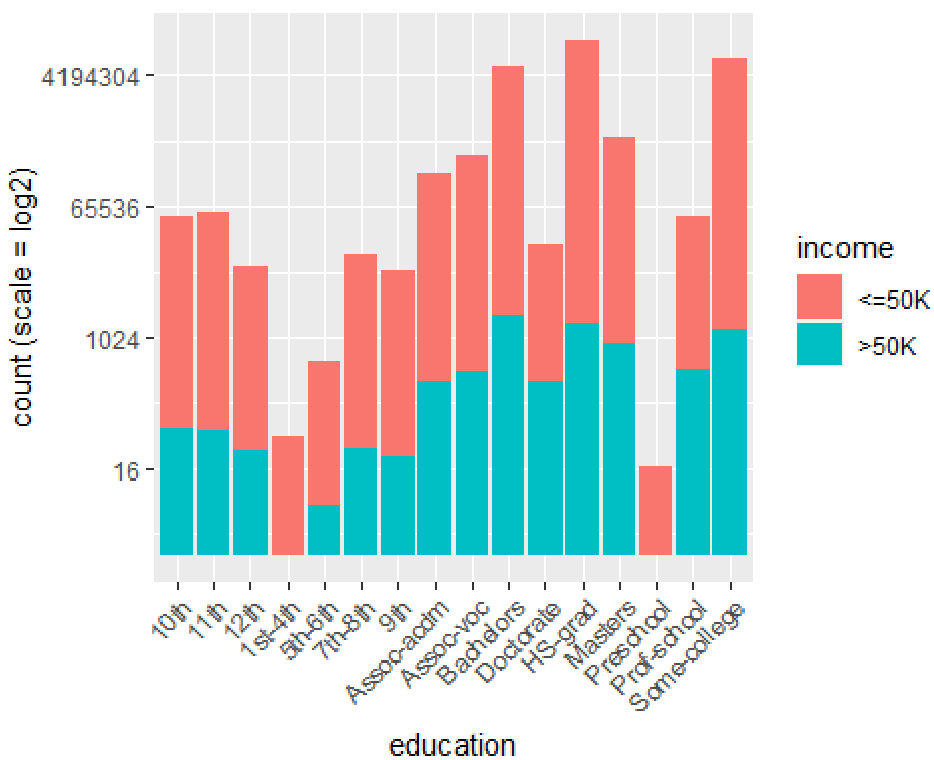
```r
adult %>% group_by(income) %>%
          summarize(min_age = min(age),
                    Q1_age = quantile(age, 0.25),
                    median_age = median(age),
                    mean_age = mean(age),
                    Q3_age = quantile(age, 0.75),
                    max_age = max(age), .groups = "drop")
```

```
## # A tibble: 2 x 7
##   income min_age Q1_age median_age mean_age Q3_age max_age
##   <chr>    <int>  <dbl>      <int>    <dbl>  <dbl>   <int>
## 1 <=50K       17     25         34     36.8     46      90
## 2 >50K        19     37         44     44.3     51      90
```

Notice that the median age for those earning more than $50K is greater than the median age for those earning less than $50K. Also, there is more variability in the ages for those earning less than $50K. Also note that the minimum age for those earning more than $50K is 19 years, which is pretty remarkable. We will use age as one of our predictors.

Now we will look at income by education and display a numerical summary.

```r
ggplot(adult, aes(education, fill = income))+
  geom_bar(position = "stack") +
  scale_y_continuous(trans = "log2") +
  ylab("count (scale = log2)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



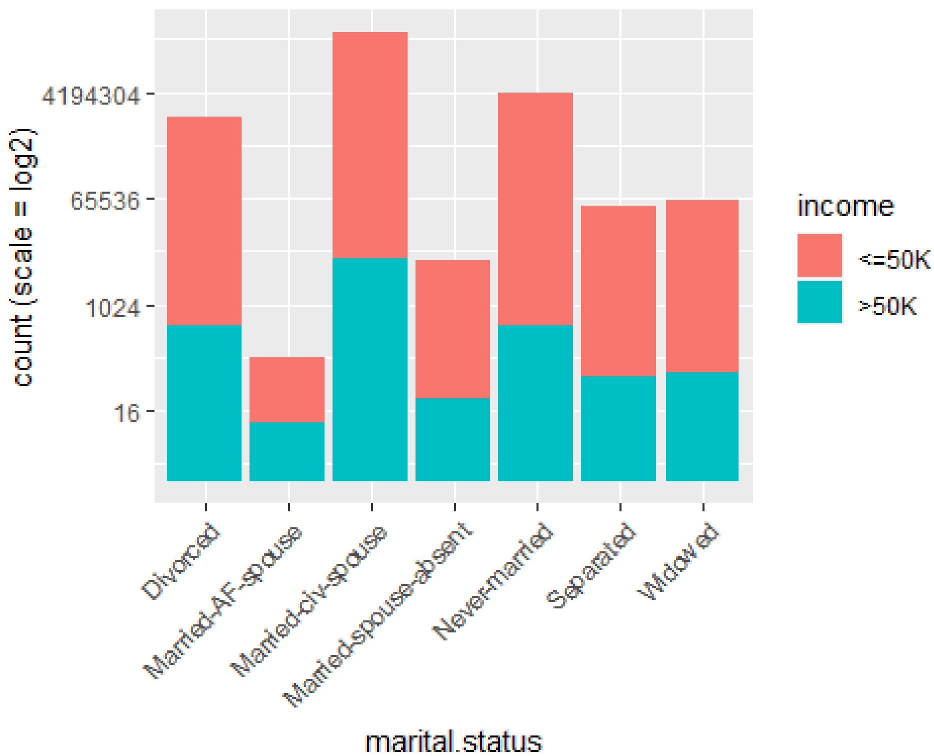
```r
table(adult$education, adult$income)
```

```
##
##                 <=50K >50K
##   10th            789   59
##   11th           1012   55
##   12th            337   28
##   1st-4th          45    1
##   5th-6th          92    5
##   7th-8th         468   31
##   9th             372   23
##   Assoc-acdm      735  247
##   Assoc-voc       953  336
##   Bachelors      2750 2016
##   Doctorate        79  249
##   HS-grad        8119 1583
##   Masters         661  866
##   Preschool        17    0
##   Prof-school     128  374
##   Some-college   5442 1298
```

Notice that there are three categories where there is a larger number of people earning over $50K, namely Doctorate, Masters, and Prof-school. We will use occupation as one of our predictors.

Next, we will look at income by marital status and display a numerical summary.

```r
ggplot(adult, aes(marital.status, fill = income))+
  geom_bar(position = "stack") +
  scale_y_continuous(trans = "log2") +
  ylab("count (scale = log2)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



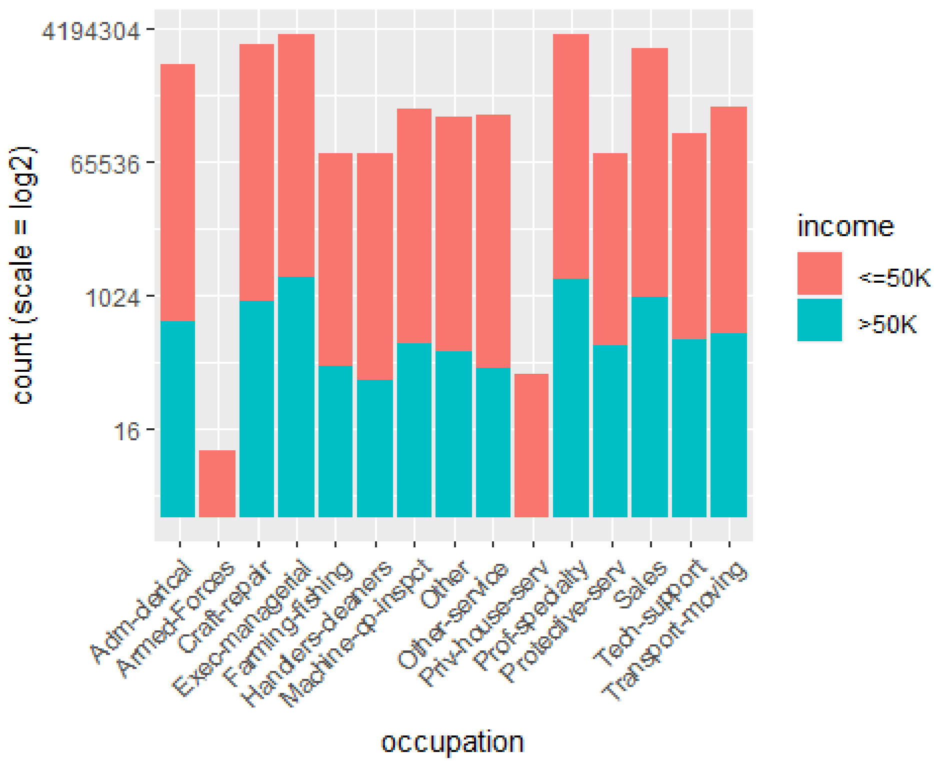

```
table(adult$marital.status, adult$income)
```

```
##
##                          <=50K  >50K
##   Divorced                3727   435
##   Married-AF-spouse         13    10
##   Married-civ-spouse      7251  6117
##   Married-spouse-absent    227    26
##   Never-married           9131   448
##   Separated                823    60
##   Widowed                  827    75
```

As can be seen from the graph and from the numerical summary, there are more people earning less than $50K than there are people earning more than $50K. In our models, we will only use variables of type character if at least one category has a larger number in the ">50K" column. We will use marital.status as one of our predictors.

We will now examine income by occupation and display a numerical summary.

```
ggplot(adult, aes(occupation, fill = income))+
  geom_bar(position = "stack") +
  scale_y_continuous(trans = "log2") +
  ylab("count (scale = log2)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



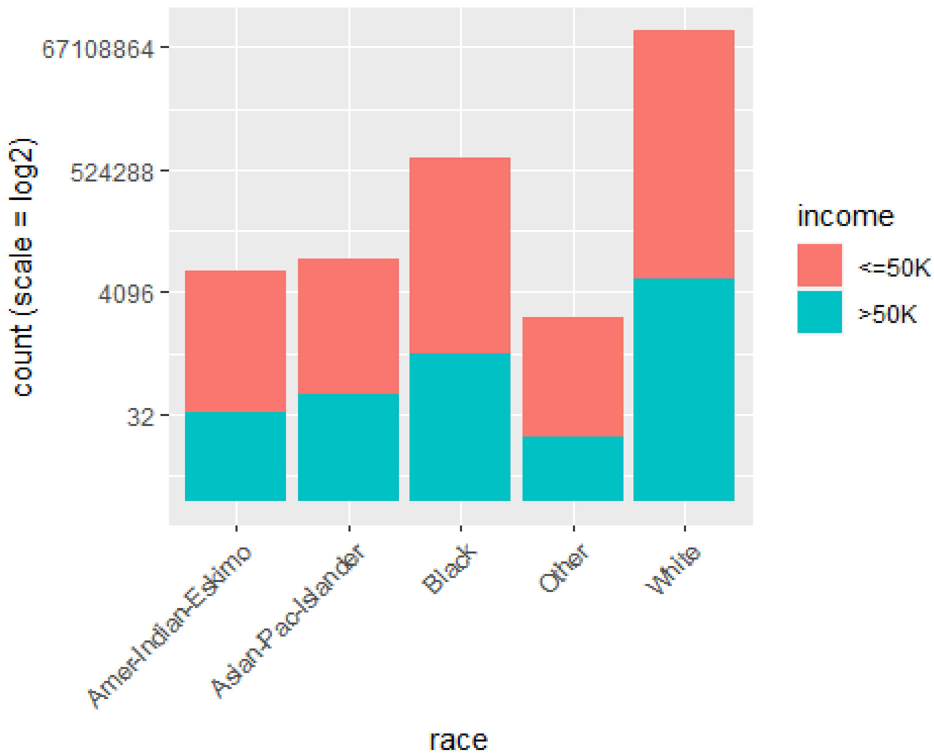

```
table(adult$occupation, adult$income)
```

```
##
##                     <=50K  >50K
##   Adm-clerical        2991   458
##   Armed-Forces           8     1
##   Craft-repair        2825   860
##   Exec-managerial     1917  1818
##   Farming-fishing      768   111
##   Handlers-cleaners   1116    73
##   Machine-op-inspct   1463   224
##   Other               1490   176
##   Other-service       2671   106
##   Priv-house-serv       89     1
##   Prof-specialty      2043  1650
##   Protective-serv      403   203
##   Sales               2436   928
##   Tech-support         593   257
##   Transport-moving    1186   305
```

As can be seen from the graph and the numerical summary, there are no categories where the larger number is in the “>50K” column. So, we will not use occupation as a predictor.

Next, we will look at income by race and display a numerical summary.

```r
ggplot(adult, aes(race, fill = income))+
  geom_bar(position = "stack") +
  scale_y_continuous(trans = "log2") +
  ylab("count (scale = log2)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



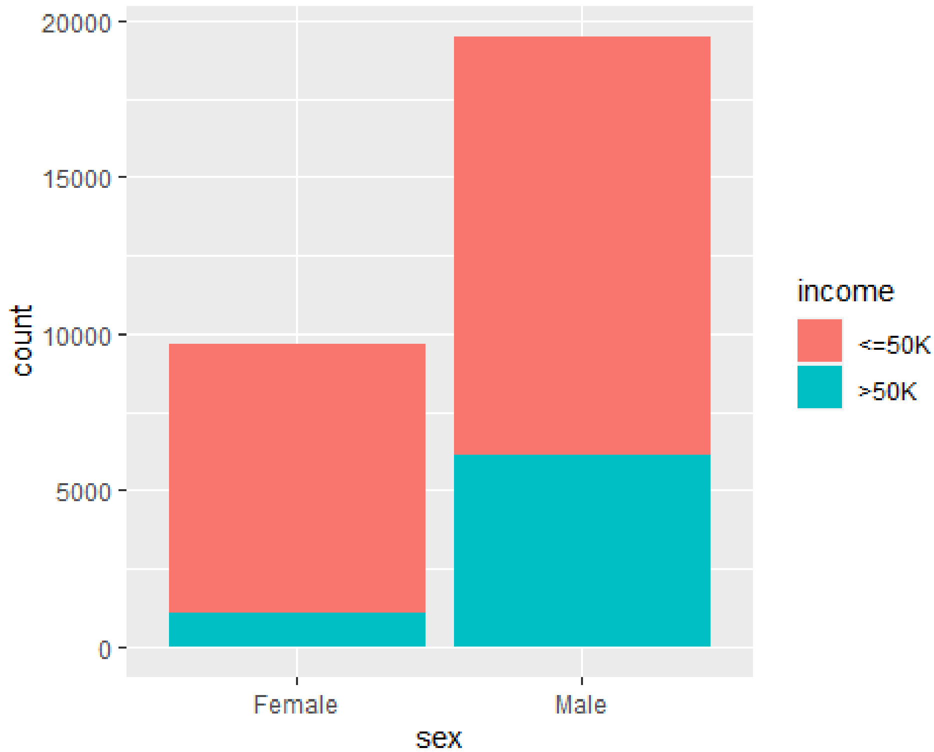

```r
table(adult$race, adult$income)

##
##                       <=50K  >50K
##   Amer-Indian-Eskimo    261    35
##   Asian-Pac-Islander    224    68
##   Black                2481   351
##   Other                 116    13
##   White               18917  6704
```

As can be seen from the graph and the numerical summary, there are no categories where the larger number is in the ">50K" column. So, we will not use race as a predictor.

Next, we will look at income by sex and display a numerical summary.

```
ggplot(adult, aes(sex, fill = income))+
  geom_bar(position = "stack")
```



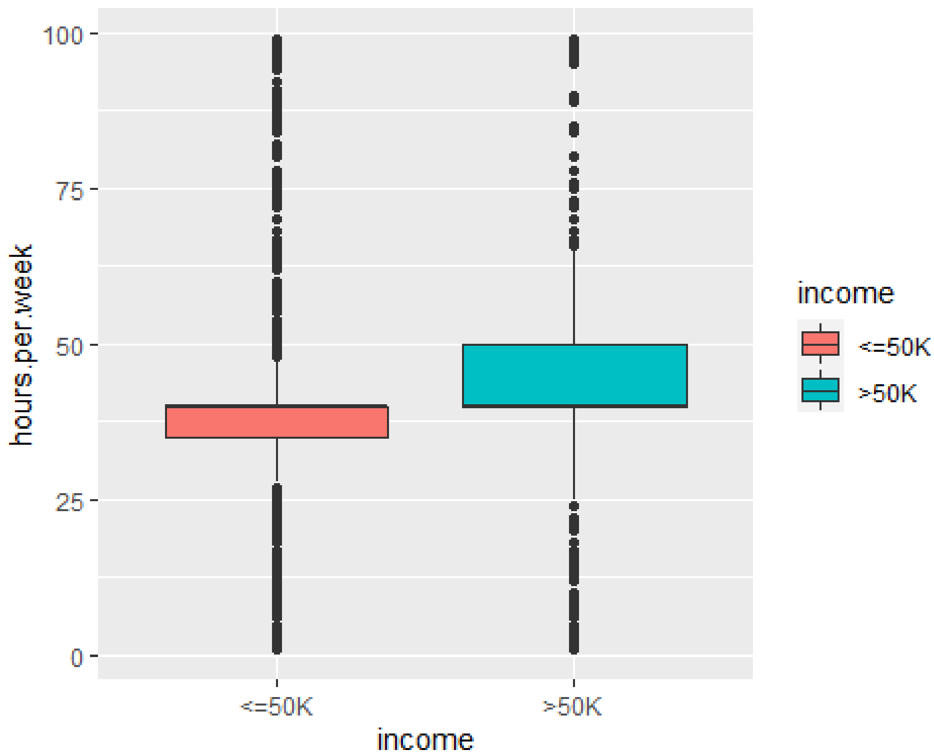

```
table(adult$sex, adult$income)

##
##          <=50K  >50K
##   Female  8610  1072
##   Male   13389  6099
```

As can be seen from the graph and the numerical summary, there are no categories where there is a larger number in the ">50K" column, so we will not use sex as a predictor.

Finally, we will examine income by hours.per.week and display a numerical summary.

```
ggplot(adult, aes(income, hours.per.week, fill = income))+
  geom_boxplot()
```




```
adult %>%
  group_by(income) %>%
  summarize(min_hours = min(hours.per.week),
            Q1_hour = quantile(hours.per.week, 0.25),
            median_hours = median(hours.per.week),
            mean_hours = mean(hours.per.week),
            Q3_hours = quantile(hours.per.week, 0.75),
            max_hours = max(hours.per.week), .groups = "drop")

## # A tibble: 2 x 7
##   income min_hours Q1_hour median_hours mean_hours Q3_hours max_hours
##   <chr>      <int>   <dbl>        <int>      <dbl>    <dbl>     <int>
## 1 <=50K          1      35           40       38.8       40        99
## 2 >50K           1      40           40       45.5       50        99
```

Both levels of income have the same median value, but 25% of those earning under $50K are working 40 hours or more, whereas 25% of those earning over 50K are working 50 hours or more. Also, there is more variability in the over $50K group. We will use hours.per.week as one of our predictors.

So, we have identified three variables that we will be able to use for our model building: age, education, and hours.per.week.

We are now ready to build the models. Note: in choosing the cutoffs for age and hours.per.week, we tried the following values: first quartile, median, mean, and third quartile. When we chose the first quartile as the cutoff, the models provided optimum performance.

Model 1 - using age to predict income

```
ages <- adult$age[adult$income == ">50K"]
ages_summary <- summary(ages)
ages_cutoff <- ages_summary["1st Qu."]
predicted_values <- ifelse(adult$age >= ages_cutoff, 1, 0)
true_values <- as.numeric(as.factor(adult$income)) - 1
accuracy <- mean(predicted_values == true_values)

#create a results table
results <- tibble(method = "model 1", accuracy = accuracy)
results

## # A tibble: 1 x 2
##   method  accuracy
##   <chr>      <dbl>
## 1 model 1    0.607
```

Model 2 - using education to predict income We will predict an income over %50K if the person has a Doctorate, a Masters, or attendance in a Prof-school.

```
education_table <- table(adult$education, adult$income)
education_table

##
##                 <=50K >50K
##   10th            789   59
##   11th           1012   55
##   12th            337   28
##   1st-4th          45    1
##   5th-6th          92    5
##   7th-8th         468   31
##   9th             372   23
##   Assoc-acdm      735  247
##   Assoc-voc       953  336
##   Bachelors      2750 2016
##   Doctorate        79  249
##   HS-grad        8119 1583
##   Masters         661  866
##   Preschool        17    0
##   Prof-school     128  374
##   Some-college   5442 1298

predicted_values <- ifelse(adult$education %in% c("Doctorate", "Masters", "Pr
of-school"), 1, 0)
accuracy <- mean(predicted_values == true_values)
```

```
#update the results table
results <- bind_rows(results, tibble(method = "model 2", accuracy = accuracy)
)
results
```

```
## # A tibble: 2 x 2
##   method  accuracy
##   <chr>      <dbl>
## 1 model 1    0.607
## 2 model 2    0.775
```

Model 3 - using hours.per.week to predict income We will use the first quartile of the hours worked as a cutoff for predicting incomes over $50K

```
hours <- adult$hours.per.week[adult$income == ">50K"]
hours_summary <- summary(hours)
hours_cutoff <- hours_summary["1st Qu."]
predicted_values <- ifelse(adult$hours.per.week >= hours_cutoff, 1, 0)
accuracy <- mean(predicted_values == true_values)

#update the results table
results <- bind_rows(results, tibble(method = "model 3", accuracy = accuracy)
)
results
```

```
## # A tibble: 3 x 2
##   method  accuracy
##   <chr>      <dbl>
## 1 model 1    0.607
## 2 model 2    0.775
## 3 model 3    0.439
```

Model 4 - using age and education to predict income

```
predicted_values <- ifelse(adult$age >= ages_cutoff &
                           adult$education %in% c("Doctorate", "Masters", "Pr
of-school"), 1, 0)
accuracy <- mean(predicted_values == true_values)

#update the results table
results <- bind_rows(results, tibble(method = "model 4", accuracy = accuracy)
)
```

```r
results
```

```
## # A tibble: 4 x 2
##   method  accuracy
##   <chr>      <dbl>
## 1 model 1    0.607
## 2 model 2    0.775
## 3 model 3    0.439
## 4 model 4    0.777
```

Model 5 - using age and hours.per.week to predict income

```r
predicted_values <- ifelse(adult$age >= ages_cutoff & adult$hours.per.week >=
hours_cutoff, 1, 0)
accuracy <- mean(predicted_values == true_values)

#update the results table
results <- bind_rows(results, tibble(method = "model 5", accuracy = accuracy)
)
results
```

```
## # A tibble: 5 x 2
##   method  accuracy
##   <chr>      <dbl>
## 1 model 1    0.607
## 2 model 2    0.775
## 3 model 3    0.439
## 4 model 4    0.777
## 5 model 5    0.671
```

Model 6 - using hours.per.week and education to predict income

```r
predicted_values <- ifelse(adult$hours.per.week >= hours_cutoff &
                           adult$education %in% c("Doctorate", "Masters", "Pr
of-school"), 1, 0)
accuracy <- mean(predicted_values == true_values)

#update the results table
results <- bind_rows(results, tibble(method = "model 6", accuracy = accuracy)
)
results
```

```
## # A tibble: 6 x 2
##   method  accuracy
##   <chr>      <dbl>
## 1 model 1    0.607
## 2 model 2    0.775
## 3 model 3    0.439
## 4 model 4    0.777
## 5 model 5    0.671
## 6 model 6    0.777
```

Model 7 - using age, hours.per.week, and education to predict income

```
predicted_values <- ifelse(adult$age >= ages_cutoff & adult$hours.per.week >=
hours_cutoff &
                           adult$education %in% c("Doctorate", "Masters", "Pr
of-school"), 1, 0)
accuracy <- mean(predicted_values == true_values)

#update the results table
results <- bind_rows(results, tibble(method = "model 7", accuracy = accuracy)
)
results

## # A tibble: 7 x 2
##   method  accuracy
##   <chr>      <dbl>
## 1 model 1    0.607
## 2 model 2    0.775
## 3 model 3    0.439
## 4 model 4    0.777
## 5 model 5    0.671
## 6 model 6    0.777
## 7 model 7    0.776
```

Results

Notice that models 4 and 6 performed the best. Models 2 and 7 also performed well compared to the rest of the models. All four of these models used education as a predictor, which demonstrates the value of education for increasing future earnings. Model 2 performed well even though it used only one predictor, namely education. Interestingly, model 4 did not perform the best, even though it used all three predictors.

Conclusion

The report shows that it is possible for one, two, or three predictors to be used in making predictions. We began the report by explaining how the data was read and how it was prepared for analysis. We then examined all the variables in the dataset to determine which variables might make good predictors. Then we constructed all possible models from these three predictors. The surprising result from the analysis is that the number of predictors is not as important as the relationship of a predictor to the predicted value, namely income. The report shows that it's not necessary to use all the variables in a dataset to make a good prediction. The report also shows the key role that education plays in a person's income. It's not surprising that education, age, and hours worked have an impact on a person's future earnings. We know from experience that as a person gets older, his income generally tends to increase. Also, if a person works hard and puts in a lot of hours, his income will generally increase over time. We were limited in the study by the nature of the data provided. Certainly, gender, race, and occupation affect a person's income, but it did not seem possible to explore this using the data we had. These are certainly issues that could be explored at a future time.