Adult Income Predict

Yun Wu

1/8/2020

Introduction

The adult dataset is from the 1994 Census database. It is also known as "Census Income" dataset. this dataset can be found at http://files.grouplens.org/datasets/movielens/ml-10m.zip. This project is related to the course Data Science: Capstone from HarvardX's Data Science Professional Certificate. Thanks to Dr. Rafael Irizarry, I really learn something important to me. The income project is predicting adult income via variables such as age, education, race, workplace, etc. In this project, I have to clean and reduce the dimension first, then used several machine learning algorithm and compared the accuracy. the purpose is to get maximum possible accuracy in prediction.

preprocess the dataset

The adult income dataset is automatically downloaded

[adult income dataset] https://archive.ics.uci.edu/ml/datasets/adult

[adult income dataset -file]https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data

Download the dataset

```
library(tidyr)
library(tidyverse)
library(caret)
library(magrittr)
library(randomForest)
options(digits = 6)

dl <- tempfile()
download.file("https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data", dl)
adult_income <- read.table(dl, sep = ',', fill = F, strip.white = T) %>% set_colnames(
  c('age', 'workclass', 'fnlwgt', 'education', 'education_num', 'marital_status', 'occupation', 'relati

head(adult_income)
```

```
##   age        workclass fnlwgt education education_num     marital_status
## 1  39        State-gov  77516 Bachelors            13      Never-married
## 2  50 Self-emp-not-inc  83311 Bachelors            13 Married-civ-spouse
## 3  38          Private 215646   HS-grad             9           Divorced
## 4  53          Private 234721      11th             7 Married-civ-spouse
## 5  28          Private 338409 Bachelors            13 Married-civ-spouse
## 6  37          Private 284582   Masters            14 Married-civ-spouse
##          occupation  relationship  race    sex capital_gain capital_loss
## 1      Adm-clerical Not-in-family White   Male         2174            0
```

```
## 2   Exec-managerial       Husband White   Male            0            0
## 3 Handlers-cleaners Not-in-family White   Male            0            0
## 4 Handlers-cleaners       Husband Black   Male            0            0
## 5    Prof-specialty          Wife Black Female            0            0
## 6   Exec-managerial          Wife White Female            0            0
##   hours_per_week native_country income
## 1             40  United-States  <=50K
## 2             13  United-States  <=50K
## 3             40  United-States  <=50K
## 4             40  United-States  <=50K
## 5             40           Cuba  <=50K
## 6             40  United-States  <=50K
```

Dimension reducing

```r
mean(adult_income$capital_gain == 0)
```

```
## [1] 0.91671
```

```r
mean(adult_income$capital_loss == 0)
```

```
## [1] 0.953349
```

```r
mean(adult_income$native_country == "United-States")
```

```
## [1] 0.895857
```

I can observe that above 90% adult have zero capital_gain and capital_loss, and about 90% adult come from united states. Therefore,these three variables are skew. so I decide to delete them. regarding to education, it means same as education_num, and relationship is same as marital status. Fnlwgt is not related to our goal. So I delete education, relationship, and fnlwgt. So far, I finish the dimension reducing.

```
##   age        workclass education_num     marital_status        occupation
## 1  39        State-gov            13      Never-married      Adm-clerical
## 2  50 Self-emp-not-inc            13 Married-civ-spouse   Exec-managerial
## 3  38          Private             9           Divorced Handlers-cleaners
## 4  53          Private             7 Married-civ-spouse Handlers-cleaners
## 5  28          Private            13 Married-civ-spouse    Prof-specialty
## 6  37          Private            14 Married-civ-spouse   Exec-managerial
##    race    sex hours_per_week income
## 1 White   Male             40  <=50K
## 2 White   Male             13  <=50K
## 3 White   Male             40  <=50K
## 4 Black   Male             40  <=50K
## 5 Black Female             40  <=50K
## 6 White Female             40  <=50K
```

Clean dataset

Trim workclass column

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

The above summary of the subset shows that the variable of workclass has too many levels. I found 'Never-worked' and 'Without-pay' have a few data so I combine them to self-employed; combine federal-gov, state-gov, and local-gov levels to government. combine self-emp-inc and self-emp-not-inc to self-employed.

```
##
##       Unknown    Government Self_Employed       Private
##          1836          4351          3678         22696
```

Trim occupation column

```
##
##                 ?      Adm-clerical      Armed-Forces      Craft-repair
##              1843              3770                 9              4099
##   Exec-managerial   Farming-fishing Handlers-cleaners Machine-op-inspct
##              4066               994              1370              2002
##     Other-service   Priv-house-serv    Prof-specialty   Protective-serv
##              3295               149              4140               649
##             Sales      Tech-support  Transport-moving
##              3650               928              1597
```

There are too many levels here, but I can block the occupation into several groups:Blue-Collar, Professional, Sales, Service, and White-Collar.

```
##
##      Unknown White_Collar Professional  Blue_Collar      Service
##         1843         7836         5077        10062         4093
##        Sales
##         3650
```

Trim marital_status column

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

Block the marital_status into Divorced, married, seperated, single, and widowed.

```r
table(adult$marital_status)
```

```
##
## Bad married     Married      Single     Widowed
##        5468       15417       10683         993
```

So far, I complete the data preprocess.

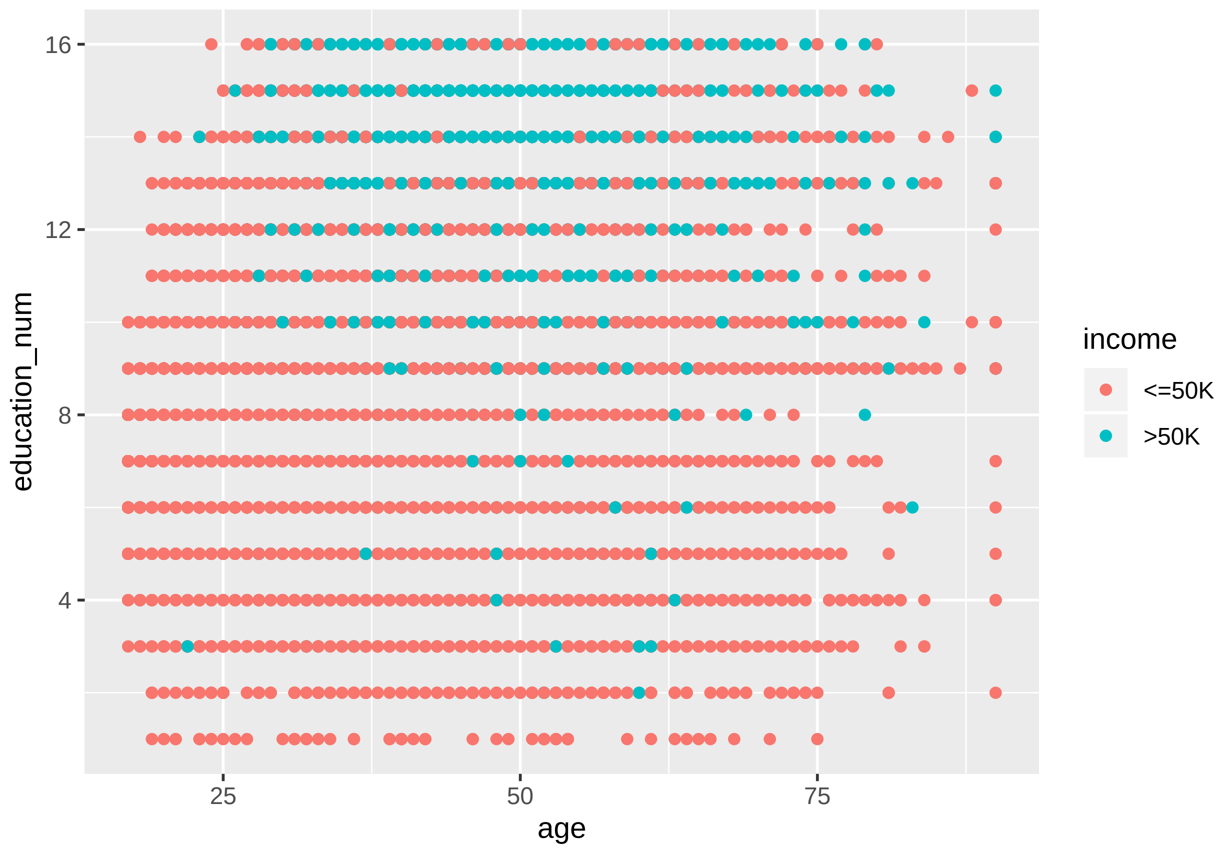
Methods and Analysis

Data Analysis

Explore the variables can help us to understand this dataset.

Explore age and education number

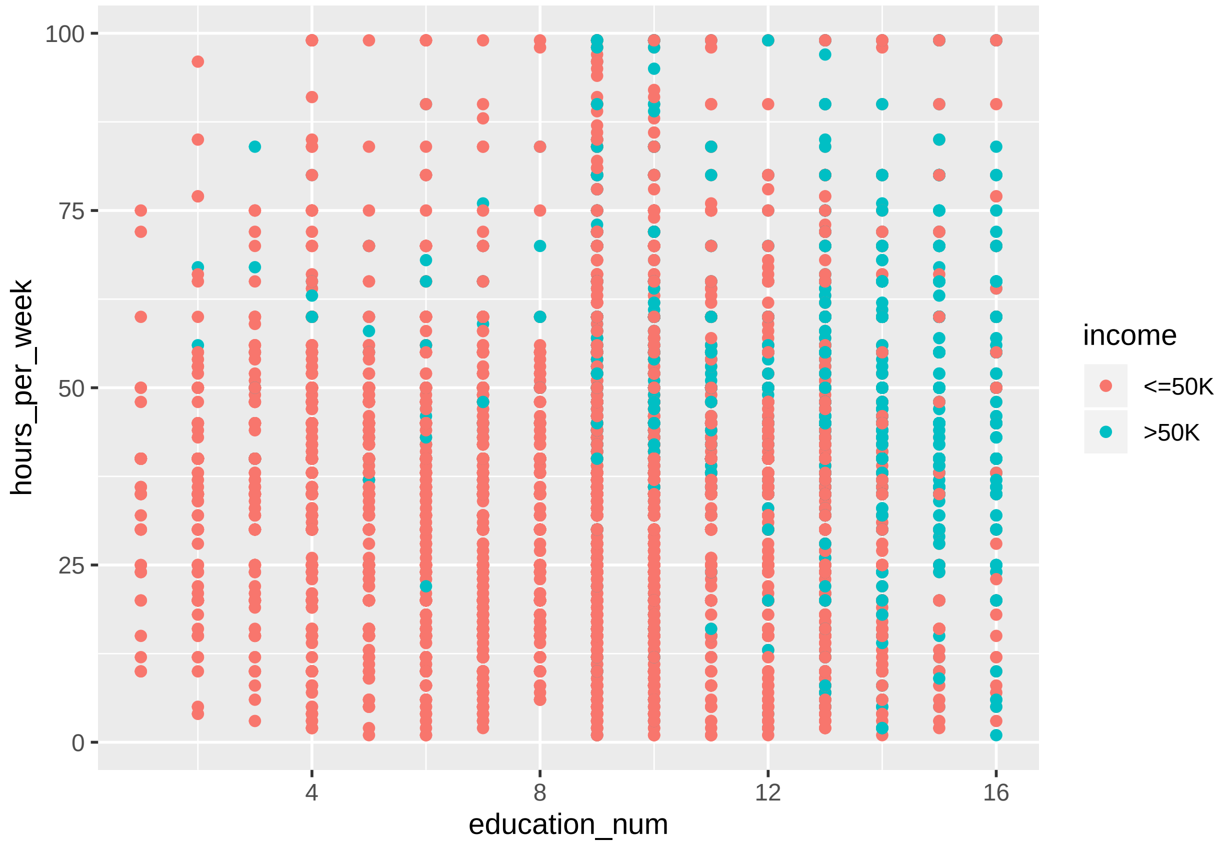
```
adult %>% ggplot(aes(age, education_num, color = income)) + geom_point()
```



by contrast of age, education number is more related to adult's income. The high education they have, the more money they make.

Explore education number and hours per week

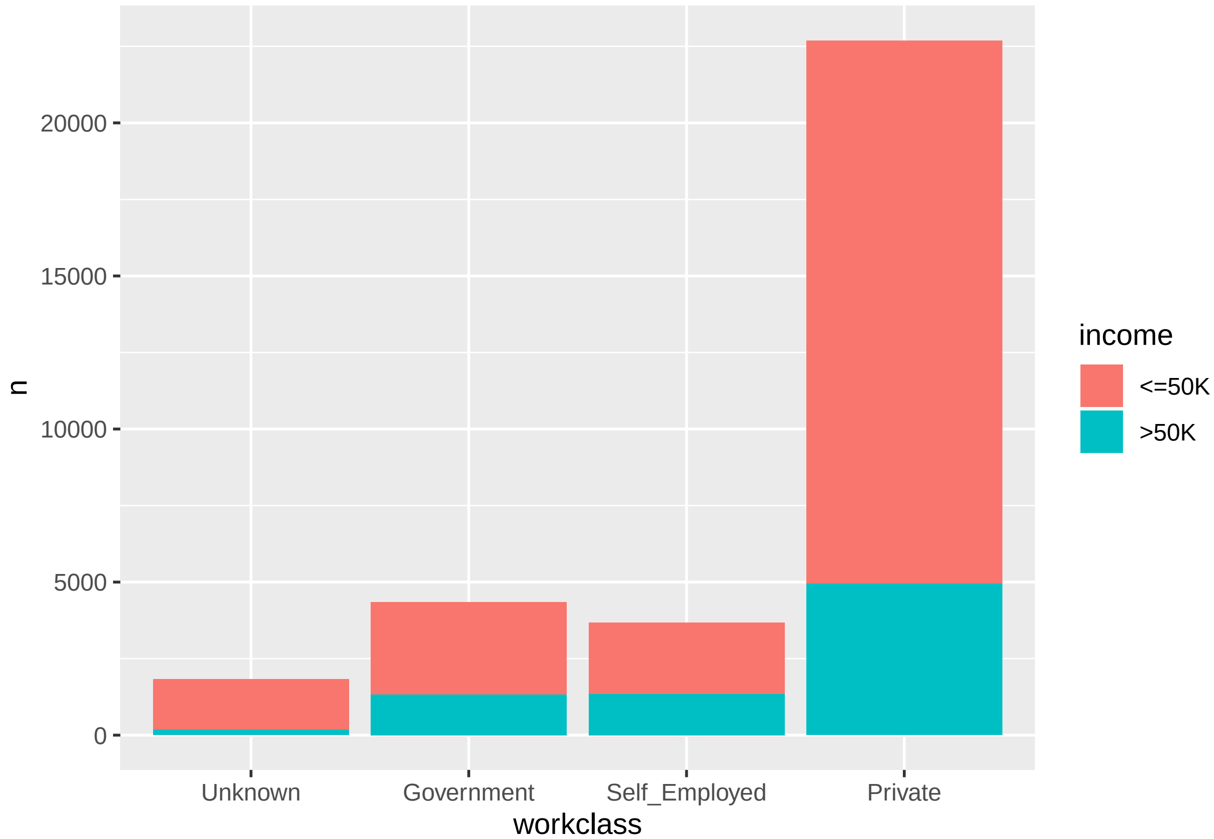
```
adult %>% ggplot(aes(education_num, hours_per_week, color = income)) + geom_point()
```



Those people who work more hours and have high education can make more money.

Explore workclass

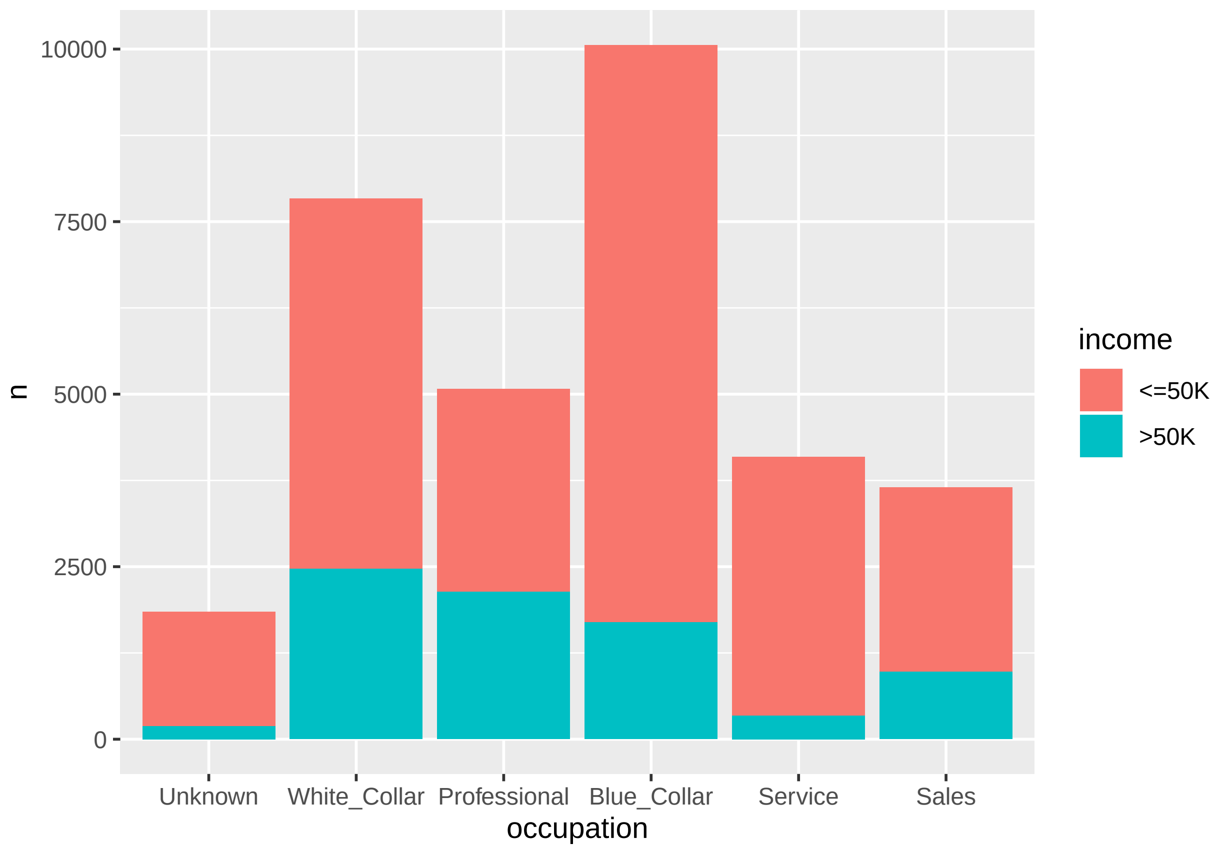
```
adult %>% group_by(workclass, income) %>% summarize(n = n()) %>% ggplot(aes(workclass, n, fill = income
```



From the figure, those who are self employed have the highest tendency of making greater than $50,000 a year.

Explore occupation

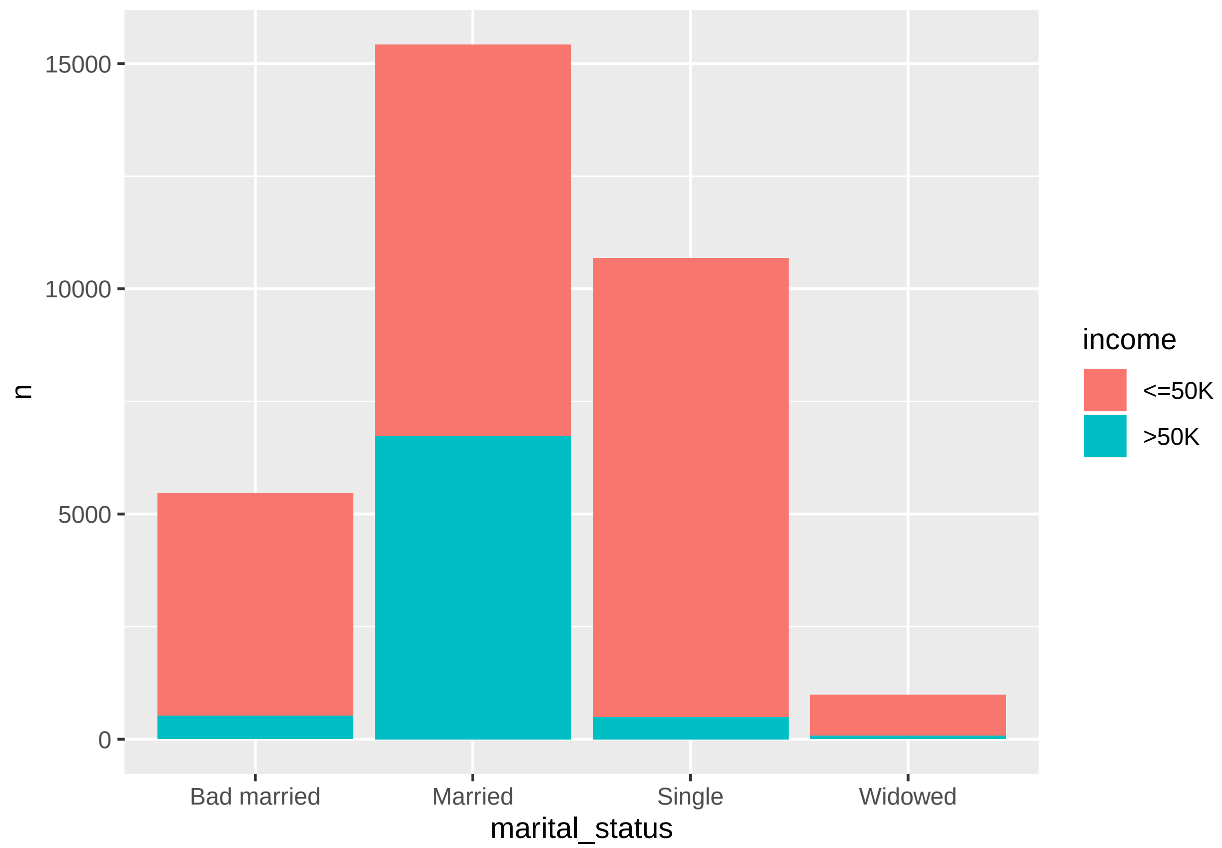
```
adult %>% group_by(occupation, income) %>% summarize(n = n()) %>% ggplot(aes(occupation, n, fill = incom
```



Nearly half of Professional occupation makes greater than $50,000 a year, while that percentage is only 13% for Service occupation.

Explore marital_status

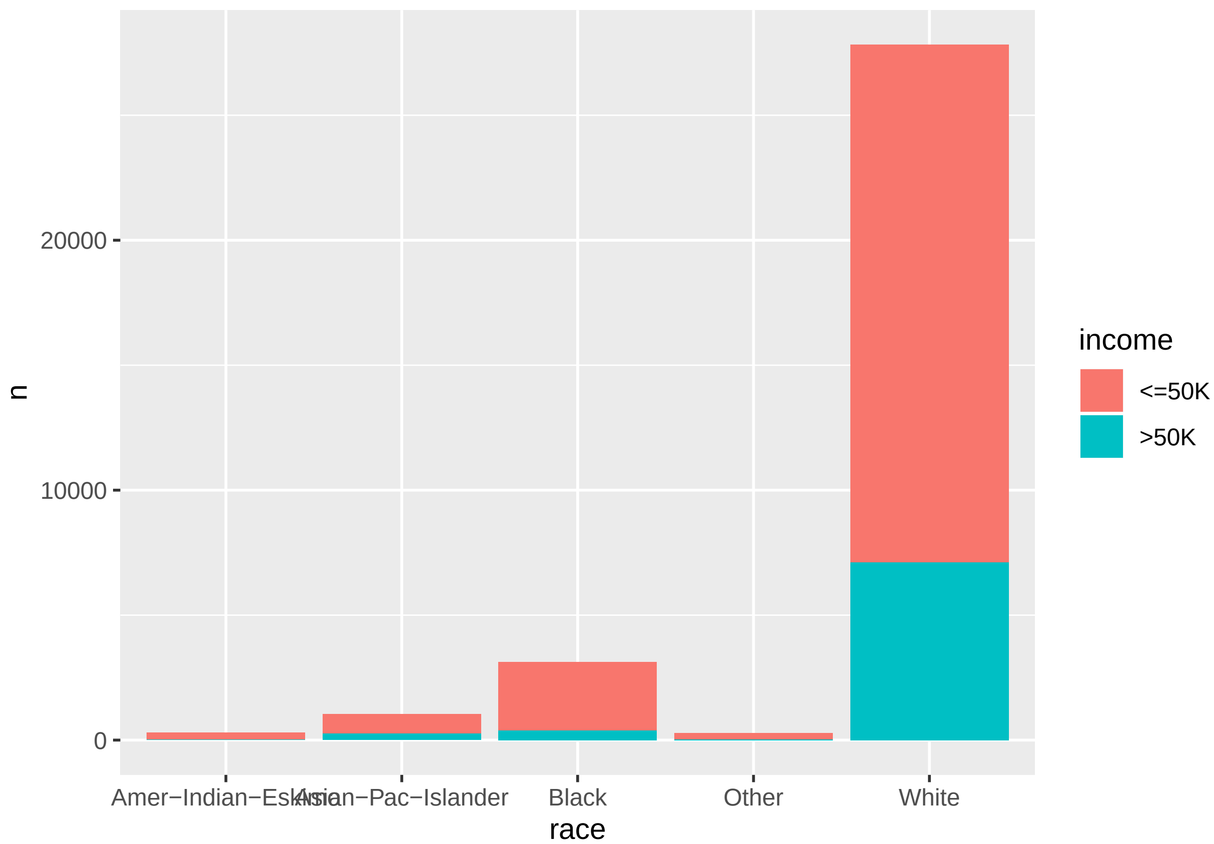
```
adult %>% group_by(marital_status, income) %>% summarize(n = n()) %>% ggplot(aes(marital_status, n, fil
```



For those who are married, nearly half of them are making greater than $50,000 a year.

Explore race

```
adult %>% group_by(race, income) %>% summarize(n = n()) %>% ggplot(aes(race, n, fill = income)) + geom_
```



White and Asian-Pacific Islander have high earning potentials.

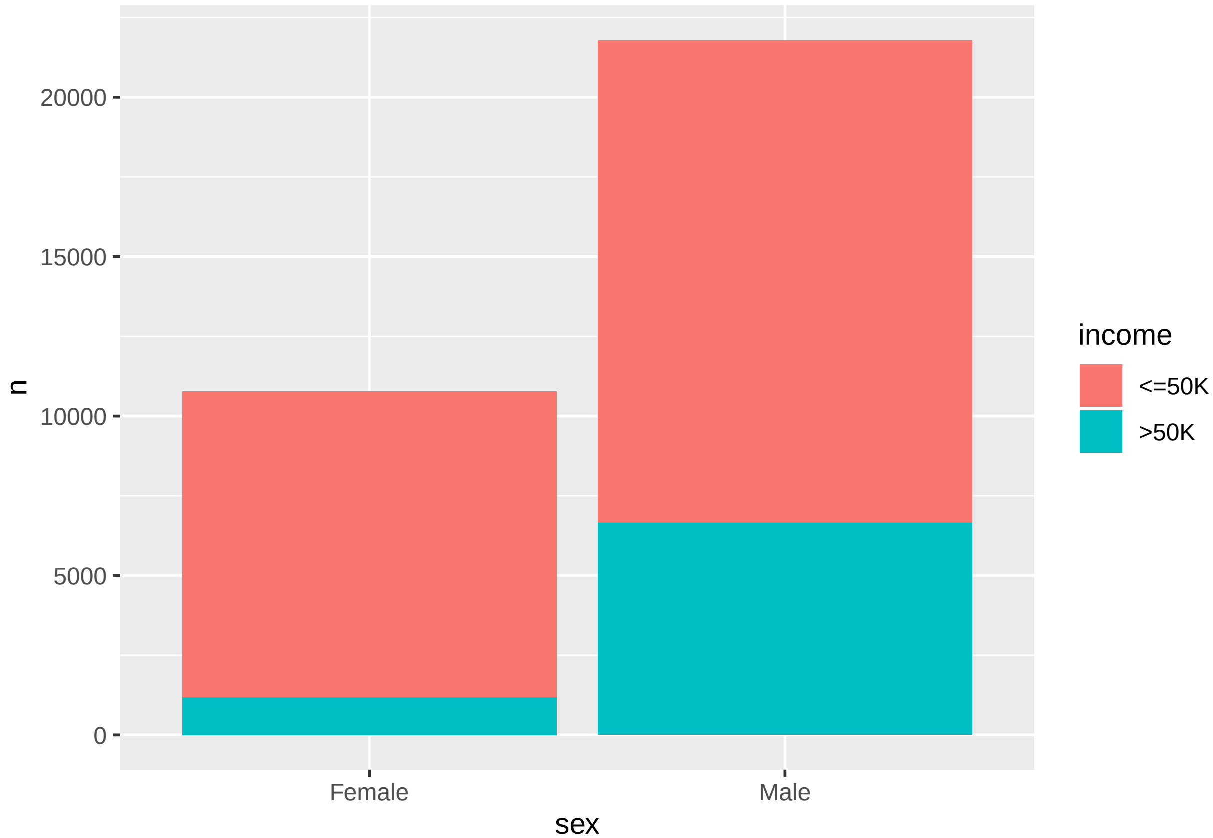
Explore age

```
adult %>% group_by(age, income) %>% summarize(n = n()) %>% ggplot(aes(age, n, fill = income)) + geom_ba
```

Those people at age between 50 to 60 have the highest tendency of making greater than $50,000 a year.

Explore sex

```
adult %>% group_by(sex, income) %>% summarize(n = n()) %>% ggplot(aes(sex, n, fill = income)) + geom_bar
```



Male adult is easier to make greater than $50,000 a year.

Modelling Approach

create train_set and test_set

```
set.seed(1)

test_index <- createDataPartition(y = adult$age, times = 1, p = 0.2, list = FALSE)
train_set <- adult[-test_index,]
test_set <- adult[test_index,]
```

Logistic regression

```
fit_glm <- train(income ~ ., method = "glm", data = train_set)
y_hat_glm <- predict(fit_glm, test_set)
Accuracy_glm <- confusionMatrix(y_hat_glm, test_set$income)$overall["Accuracy"]
Accuracy_results <- tibble(method = "logistic regression", Accuracy = Accuracy_glm)
print.data.frame(Accuracy_results)
```

```
##                method Accuracy
## 1 logistic regression 0.832361
```

Random forest classification

```
set.seed(1)
my_control <- trainControl(method = "cv", number = 5)
fit_rf <- train(y = train_set[,9], x =train_set[,-9], method = "rf", ntree = 1000, trControl = my_contro
y_hat_rf <- predict(fit_rf, test_set)
Accuracy_rf <- confusionMatrix(y_hat_rf, test_set$income)$overall["Accuracy"]
Accuracy_results <- bind_rows(Accuracy_results,
                          tibble(method="Random forest",
                                 Accuracy = Accuracy_rf))
```

Results

The Accuracy values are the following:

```
##                method Accuracy
## 1 logistic regression 0.832361
## 2       Random forest 0.843875
```

We therefore found the better method is random forest.

Discussion

We can see the variables of workclass, marital_status are categories. So the classification model is better then regression.

```
dim(adult)
```

```
## [1] 32561     9
```

Taking about the random forest, due to the high dimension which is 10 of columns and 32561 of rows, it spends a lot of time on processing. I can't do more research. for example I would should have research optimized the parameters such as 'mtry' and 'ntree'. I probably didn't have the best model.

Conclusion

According to the model, sometimes we can adjust an adult's income just by his race, workclass, marital status, occupation, education, etc. Because of the category variables, the regression is not good at it. On contrast, the classification method performs better.