# Adult Income Distribution

Duong Nguyen

28 April 2019

# Content



# 1. Executive Summary

From the UCI machine learning repository website, the adult database was chosen. The dataset is a **panel**, i.e. at one point in time and across all respondents data were collected. The goal is to develop a model to **predict** whether someone likely earns **over 50k USD**. For such a categorial question, the **logistic regression** was chosen. The best performing model is:

**Dependent variable**: *income*

**Independent variables**: *age, worklclass, education_num, occupation, relationship, race, sex, capital_gains, capital_loss, hours_per_week, native_region*

*omitted variables*: *fnlwgt, education, marial_status*

The features someone has which make her/him likely to earn over 50k:

The person in question is not young. If she/he works for the Federal Government, her/his odds is better than other people. The more time somebody has invested in her/his education the more likely this person will earn over 50k. If a person works in areas of Exec-managerial, Prof-specialty, Protective-service, Sales, or Tech-support, her/his chances are high to have over 50k.

As a husband the odds are higher than most people, except if someone is a working wife. In this case her chances are even higher to earn over 50k. If someone is not from any American-Indian-Eskimo minorities, his likelihood increases also. The person to earn over 50k is likely a male. If somebody has any capital gains or losses, they also indicate a high probability.

If she/he works part-time, she/he is less likely to earn more than other people. And if someone is born from Central America, she/he is less likely to make 50k compared to others who are born in Western Europe or in the US.

The predicted value from the logistic regression is a probability of earning over 50k. For decision purpose, a **threshold for the probability** of **p=0.3** was chosen under considering the balance of true and false positive rates.

The accuracy of the model using the train dataset results in *0.8254797*.

The accuracy of the model using the test dataset for validation results in **0.8269518**.

# 2. Data

# 2.1. Import and Cleaning

From the UCI machine learning repository website, I chose the adult database for my capstone project. https://archive.ics.uci.edu/ml/datasets/Adult (https://archive.ics.uci.edu/ml/datasets/Adult)

Using a word editor or note pad, it looks like as if the data are saved as text file separated by commas. After importing into R, I run through the dataset to look for missing values. They are denoted as *" ?“* The final code for importing the data looks like this:

```
#downloading dataset

adult <- read.table("https://archive.ics.uci.edu/ml/machine-learning-databases/adult/a
dult.data",
                    header = FALSE, sep = ",", na=" ?")
dim(adult)
```

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

The dataset has a dimension of 32561 observations and 15 variables. As column names the following names are used:

```r
#giving names for each column

names(adult) = c("age","workclass","fnlwgt","education","education_num","marial_statu
s","occupation",
                 "relationship","race","sex","capital_gain","capital_loss","hours_per_
week",
                 "native_country", "income")
```

The structure of the dataset is given by a mixture of integer and factor variables:

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

```
## 'data.frame':    32561 obs. of  15 variables:
##  $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ workclass     : Factor w/ 8 levels " Federal-gov",..: 7 6 4 4 4 4 4 6 4 4 ...
##  $ fnlwgt        : int  77516 83311 215646 234721 338409 284582 160187 209642 4578
1 159449 ...
##  $ education     : Factor w/ 16 levels " 10th"," 11th",..: 10 10 12 2 10 13 7 12 1
3 10 ...
##  $ education_num : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ marial_status : Factor w/ 7 levels " Divorced"," Married-AF-spouse",..: 5 3 1 3
3 3 4 3 5 3 ...
##  $ occupation    : Factor w/ 14 levels " Adm-clerical",..: 1 4 6 6 10 4 8 4 10
4 ...
##  $ relationship  : Factor w/ 6 levels " Husband"," Not-in-family",..: 2 1 2 1 6 6
2 1 2 1 ...
##  $ race          : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5
5 ...
##  $ sex           : Factor w/ 2 levels " Female"," Male": 2 2 2 2 1 1 1 2 1 2 ...
##  $ 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: Factor w/ 41 levels " Cambodia"," Canada",..: 39 39 39 39 5 39 2
3 39 39 39 ...
##  $ income        : Factor w/ 2 levels " <=50K"," >50K": 1 1 1 1 1 1 1 2 2 2 ...
```

All missing values are removed

```r
# ommitting all observations with nas and checking the new dimension
adult <- na.omit(adult)
dim(adult)
```

```
## [1] 30162    15
```

So over 2399 rows have been removed. (32561-30162).

The variable *fnlwgt* (final weighting) is a control variable for the data collection and will be removed from the dataset.

```
# variable fnlwgt will be dropped out, weighting variable
adult[[ "fnlwgt"]] <- NULL
```

The variable *age* is mapped into subcategories *Young, Middle_aged, Senior, and Old*:

```
#mapping age of 15, 25, 45, 65, 100 into Young, Middle_aged, Senior, Old

adult[[ "age"]] <- ordered(cut(adult[[ "age"]], c(15,25,45,65,100)),
                           labels= c("Young","Middle_age","Senior","Old"))
```

With regards to *education*, the factors are reordered:

```
#education reordering
adult$education<-ordered(adult$education,levels=c(" Preschool"," 1st-4th"," 5th-6th",
                                                  " 7th-8th"," 9th"," 10th"," 11th","
12th",
                                                  " HS-grad"," Some-college"," Assoc-a
cdm",
                                                  " Assoc-voc"," Bachelors"," Master
s",
                                                  " Prof-school"," Doctorate"))

adult$education<-as.factor(adult$education)
```

The variable *hours_per_week* is mapped into *Part_time, Full_time, Over_time, Workaholic*:

```
#mapping hours-per-week into Part_time, Full_time, Over_time, Workaholic

adult[[ "hours_per_week"]] <- ordered(cut(adult[[ "hours_per_week"]],
                                          c(0,25,40,60,110)),
                                      labels = c("Part_time", "Full_time", "Over_tim
e", "Workaholic"))
```

*Capital gains* and losses are mapped into *None, Low, High* respectively:

```
#mapping capital_gain

adult[[ "capital_gain"]] <- ordered(cut(adult[[ "capital_gain"]],
                                        c(-Inf,0,median(adult[[ "capital_gain"]][adult
[[ "capital_gain"]]>0]),
                                          Inf)), labels = c("None", "Low", "High"))

#mapping capital_loss

adult[[ "capital_loss"]] <- ordered(cut(adult[[ "capital_loss"]],
                                        c(-Inf,0, median(adult[[ "capital_loss"]][adul
t[[ "capital_loss"]]>0]),
                                          Inf)), labels = c("None", "Low", "High"))
```

With regards to the column *native_country*, there are 41 countries that will be mapped into 8 *native_region*. Afterwards the column *native _country* will be removed.

```
#native-country, there are 41 different countries
#mapping them into smaller groups
levels(adult$`native_country`)
```

```
##  [1] " Cambodia"                   " Canada"
##  [3] " China"                      " Columbia"
##  [5] " Cuba"                       " Dominican-Republic"
##  [7] " Ecuador"                    " El-Salvador"
##  [9] " England"                    " France"
## [11] " Germany"                    " Greece"
## [13] " Guatemala"                  " Haiti"
## [15] " Holand-Netherlands"         " Honduras"
## [17] " Hong"                       " Hungary"
## [19] " India"                      " Iran"
## [21] " Ireland"                    " Italy"
## [23] " Jamaica"                    " Japan"
## [25] " Laos"                       " Mexico"
## [27] " Nicaragua"                  " Outlying-US(Guam-USVI-etc)"
## [29] " Peru"                       " Philippines"
## [31] " Poland"                     " Portugal"
## [33] " Puerto-Rico"                " Scotland"
## [35] " South"                      " Taiwan"
## [37] " Thailand"                   " Trinadad&Tobago"
## [39] " United-States"              " Vietnam"
## [41] " Yugoslavia"
```

```
# defining different regions
east_asia <- c(" Cambodia", " China", " Hong", " Laos", " Thailand",
               " Japan", " Taiwan", " Vietnam")

central_asia <- c(" India", " Iran")

central_america <- c(" Cuba", " Guatemala", " Jamaica", " Nicaragua",
                     " Puerto-Rico",  " Dominican-Republic", " El-Salvador",
                     " Haiti", " Honduras", " Mexico", " Trinadad&Tobago")

south_america <- c(" Ecuador", " Peru", " Columbia")


west_europe <- c(" England", " Germany", " Holand-Netherlands", " Ireland",
                 " France", " Greece", " Italy", " Portugal", " Scotland")

east_europe <- c(" Poland", " Yugoslavia", " Hungary")

adult <- mutate(adult,
                native_region = ifelse(native_country %in% east_asia, " East-Asia",
                                       ifelse(native_country %in% central_asia, " Cent
ral-Asia",
                                              ifelse(native_country %in% central_ameri
ca, " Central-America",
                                                     ifelse(native_country %in% south_
america, " South-America",
                                                            ifelse(native_country %i
n% west_europe, " Europe-West",
                                                                   ifelse(native_count
ry %in% east_europe, " Europe-East",
                                                                          ifelse(nativ
e_country == " United-States", " United-States",
                                                                                 " Oth
ers" ))))))))


adult$native_region <- factor(adult$native_region, ordered = FALSE)

# dropping native_country column, as it has become absolete

adult[[ "native_country"]] <- NULL
```

The final structure of the modified dataset looks like this:

```
# Checking the final structure of the dataset
str(adult)
```

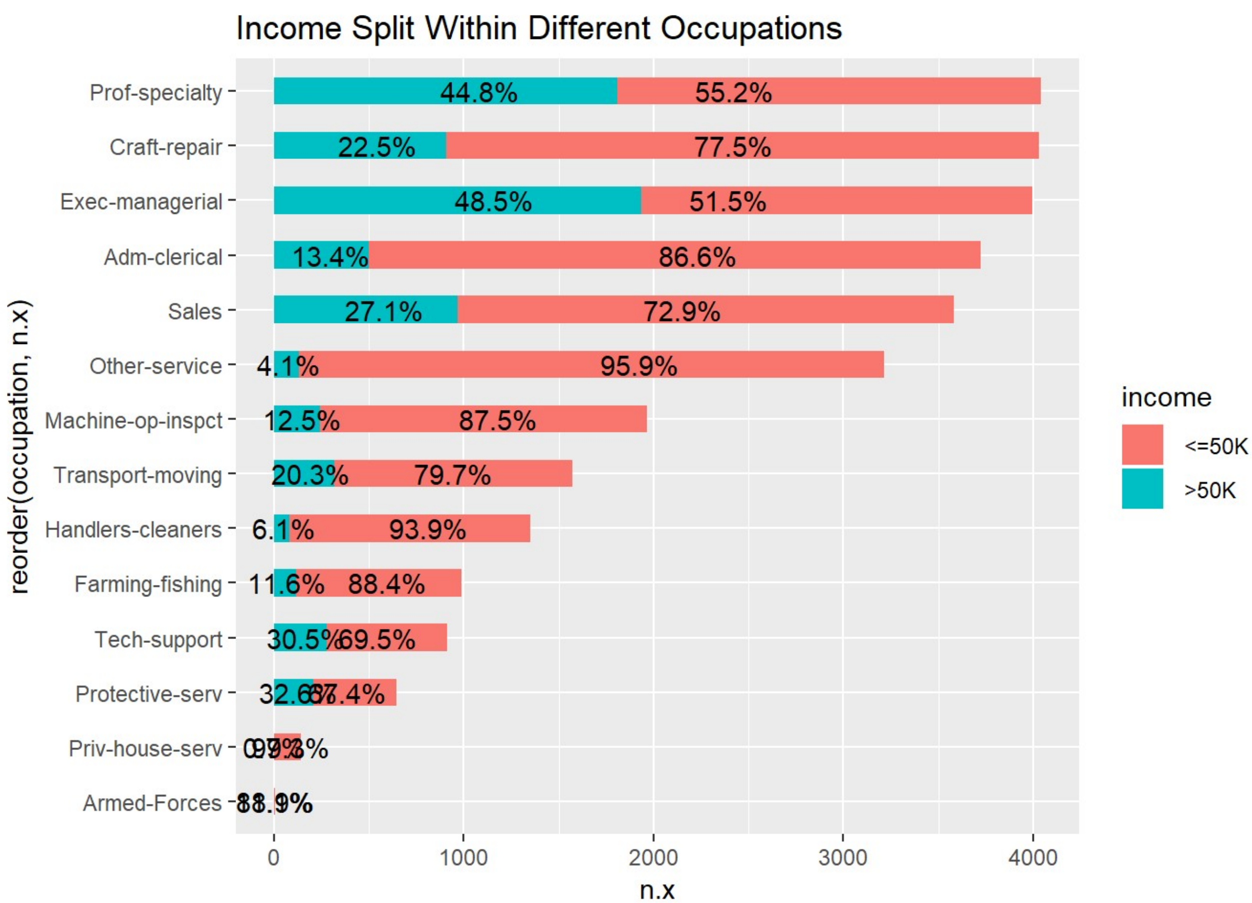
```
## 'data.frame':    30162 obs. of  14 variables:
##  $ age           : Ord.factor w/ 4 levels "Young"<"Middle_age"<..: 2 3 2 3 2 2 3 3
2 2 ...
##  $ workclass     : Factor w/ 8 levels " Federal-gov",..: 7 6 4 4 4 4 4 6 4 4 ...
##  $ education     : Ord.factor w/ 16 levels " Preschool"<" 1st-4th"<..: 13 13 9 7 1
3 14 5 9 14 13 ...
##  $ education_num : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ marial_status : Factor w/ 7 levels " Divorced"," Married-AF-spouse",..: 5 3 1 3
3 3 4 3 5 3 ...
##  $ occupation    : Factor w/ 14 levels " Adm-clerical",..: 1 4 6 6 10 4 8 4 10
4 ...
##  $ relationship  : Factor w/ 6 levels " Husband"," Not-in-family",..: 2 1 2 1 6 6
2 1 2 1 ...
##  $ race          : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5
5 ...
##  $ sex           : Factor w/ 2 levels " Female"," Male": 2 2 2 2 1 1 1 2 1 2 ...
##  $ capital_gain  : Ord.factor w/ 3 levels "None"<"Low"<"High": 2 1 1 1 1 1 1 1 3
2 ...
##  $ capital_loss  : Ord.factor w/ 3 levels "None"<"Low"<"High": 1 1 1 1 1 1 1 1 1
1 ...
##  $ hours_per_week: Ord.factor w/ 4 levels "Part_time"<"Full_time"<..: 2 1 2 2 2 2
1 3 3 2 ...
##  $ income        : Factor w/ 2 levels " <=50K"," >50K": 1 1 1 1 1 1 1 2 2 2 ...
##  $ native_region : Factor w/ 8 levels " Central-America",..: 8 8 8 8 1 8 1 8 8
8 ...
```

The dataset given is a panel, i.e. at one point in time and across all respondents data were collected. If the same respondents were examined over a long period, then we would have a time series.

# 2.2. Exploration

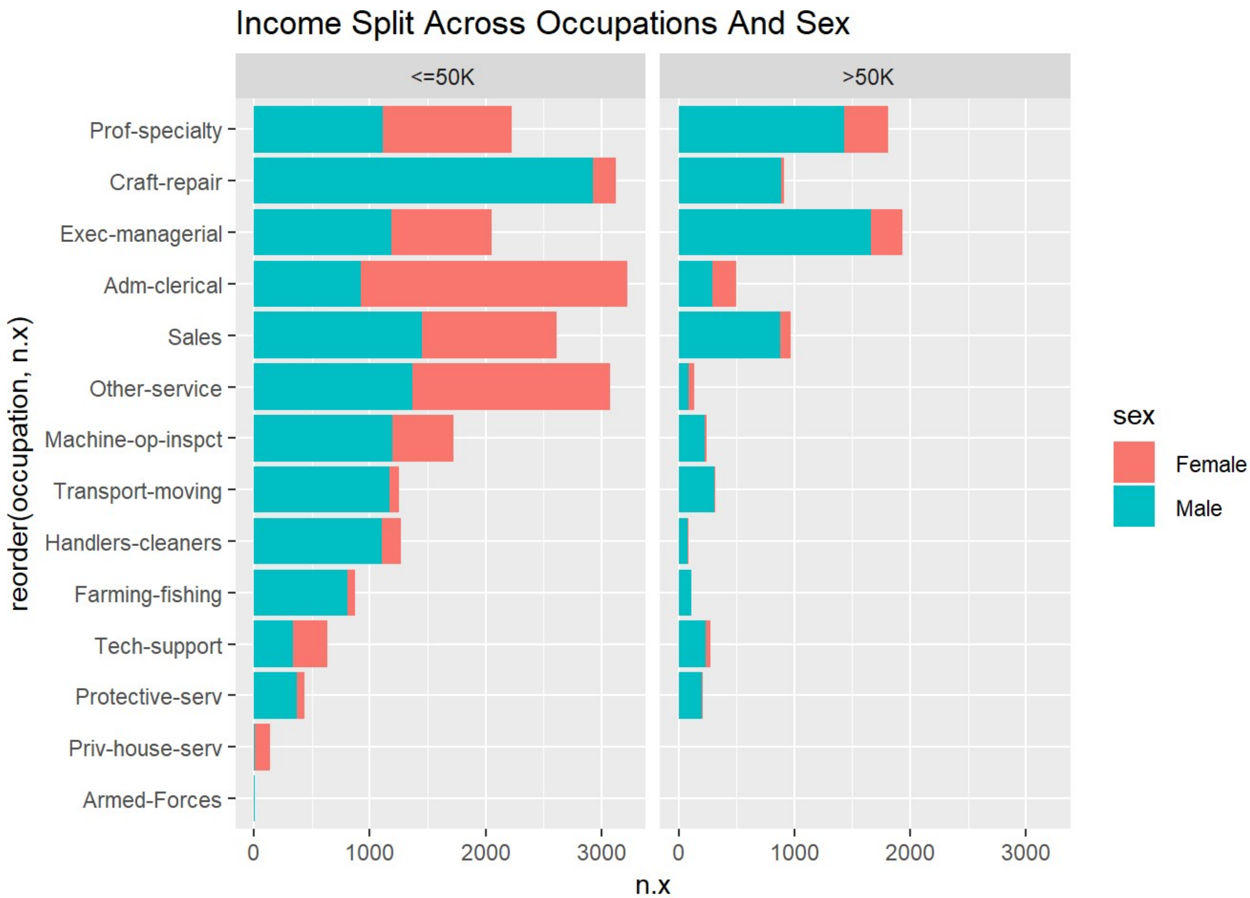
The following charts shows over different variables how the income distribution is split:

```
# Percentage distribution of income across different occupations

adult %>% count(occupation, income) %>%
  left_join(adult %>% count(occupation), by="occupation") %>%
  mutate(pct=(n.x/n.y)*100, ypos=0.6*n.x) %>%
  ggplot(aes(x=reorder(occupation,n.x), n.x, fill=income)) +
  geom_bar(stat = "identity", width = 0.5) +
  geom_text(position="stack",aes(label=paste0(sprintf ("%1.1f", pct),"%"), y=ypos))+
  coord_flip()+
  labs(title = "Income Split Within Different Occupations ")
```

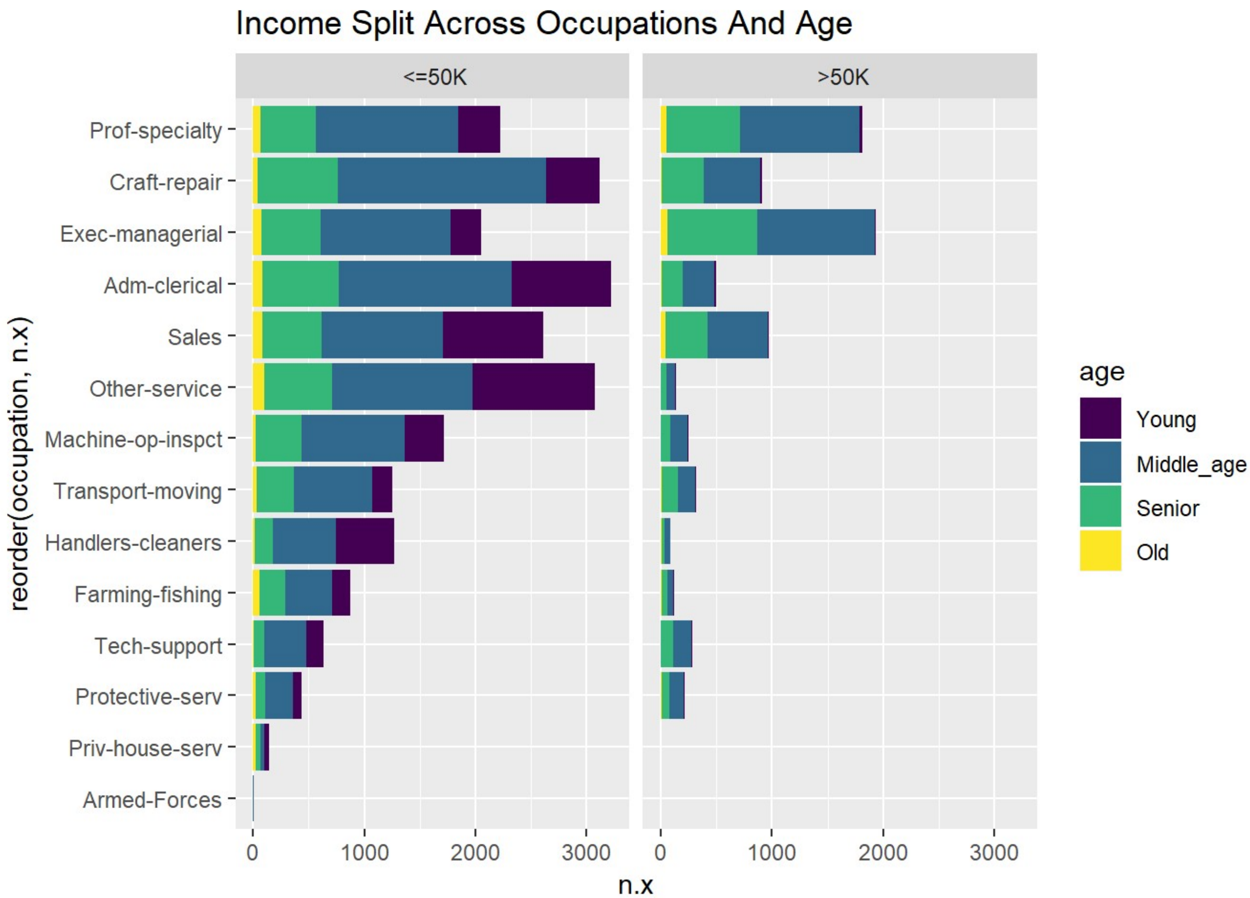



The chart *Income Split Within Different Occupations* shows the income distribution over different occupations. The three occupations with the most respondents are *Prof-specialty, Craft-repair,* and *Exec-managerial*. The occupations with the least respondents are *Protective-serv, Priv-house-serv,* and *Armed-Forces*. The three occupations with the highest proportions of people with income over 50k are *Exec-managerial, Prof-specialty*, and *Protective-serv*.

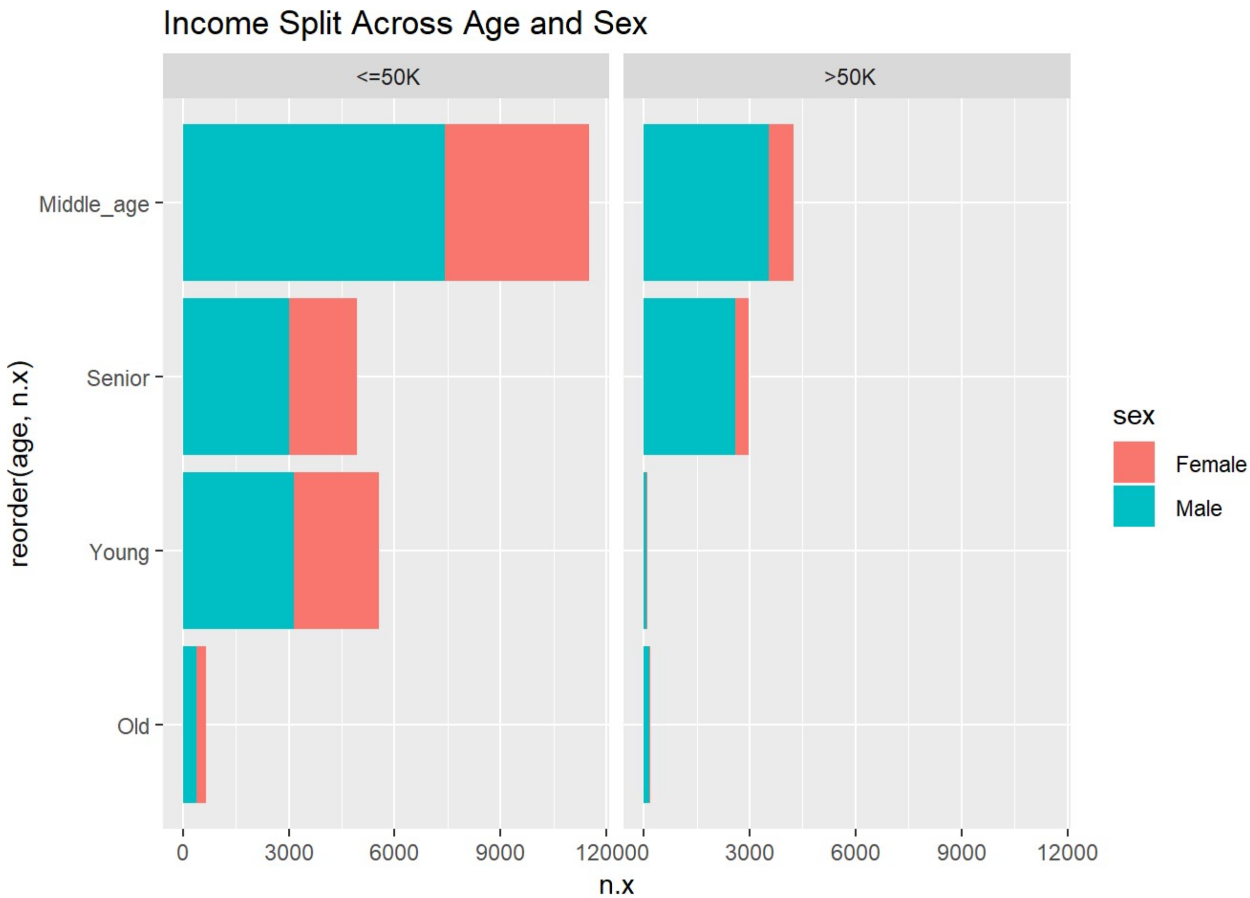
```
#Distribution across occupations and sex

adult %>% count(occupation, income, sex) %>%
  left_join(adult %>% count(occupation), by="occupation") %>%
  ggplot(aes(x=reorder(occupation,n.x), n.x, fill=sex)) +
  geom_bar(stat = "identity") +
  facet_grid(cols = vars(income))+
  coord_flip()+
  labs(title = "Income Split Across Occupations And Sex")
```




The chart *Income Split Across Occupations And Sex* outlines the dominance of male in having incomes over 50k. Most women earn less than 50k. A high number of women are working in *Adm-clerical, Other-service*, and *Sales*. It also shows that most respondents are male.

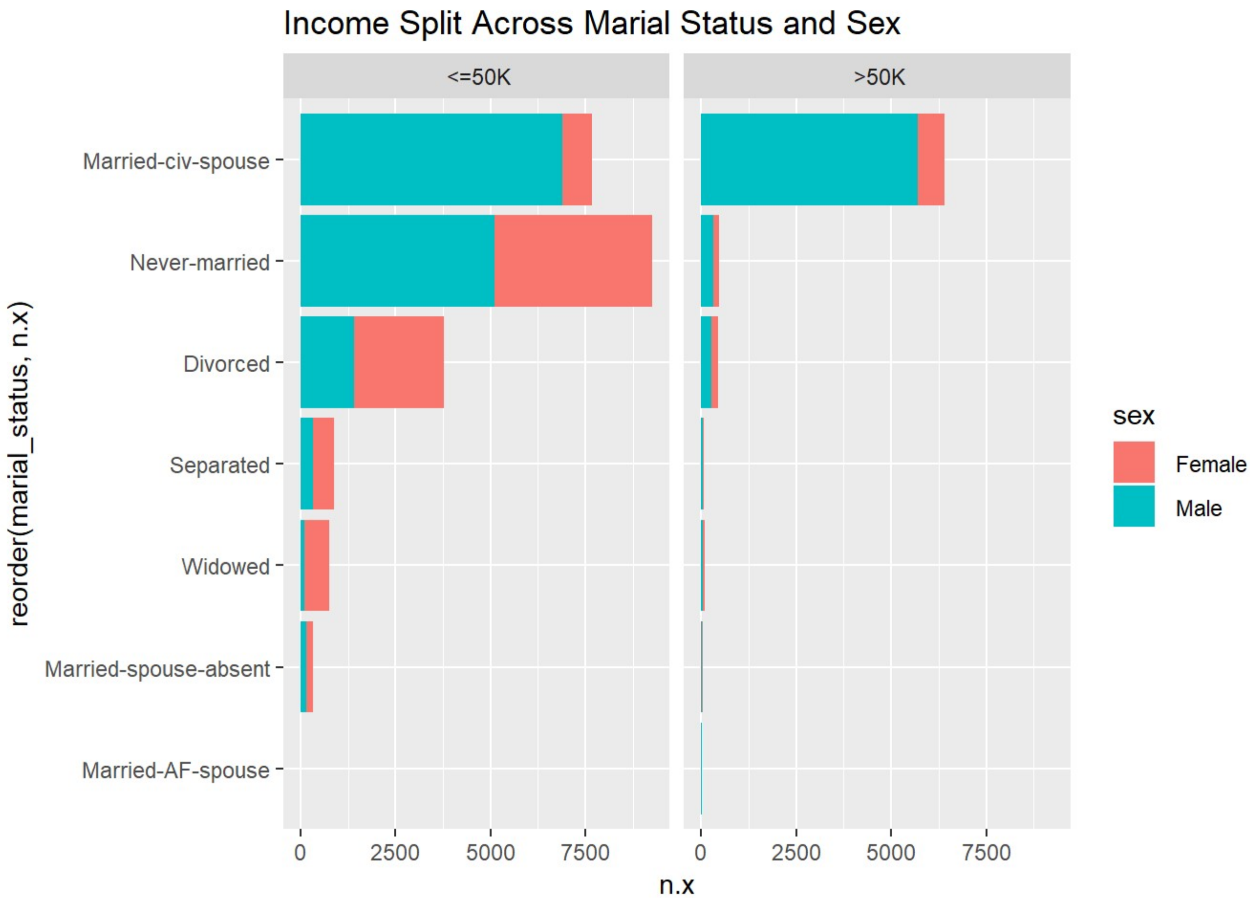



The chart *Income Split Across Occupations and Age* describes that young people are less likely to earn over 50k.




The chart *Income Split Across Age and Sex* shows the majority of respondents are in the groups of *Middle_age* and *Senior*. These two groups also have the highest number of people with income over 50k.

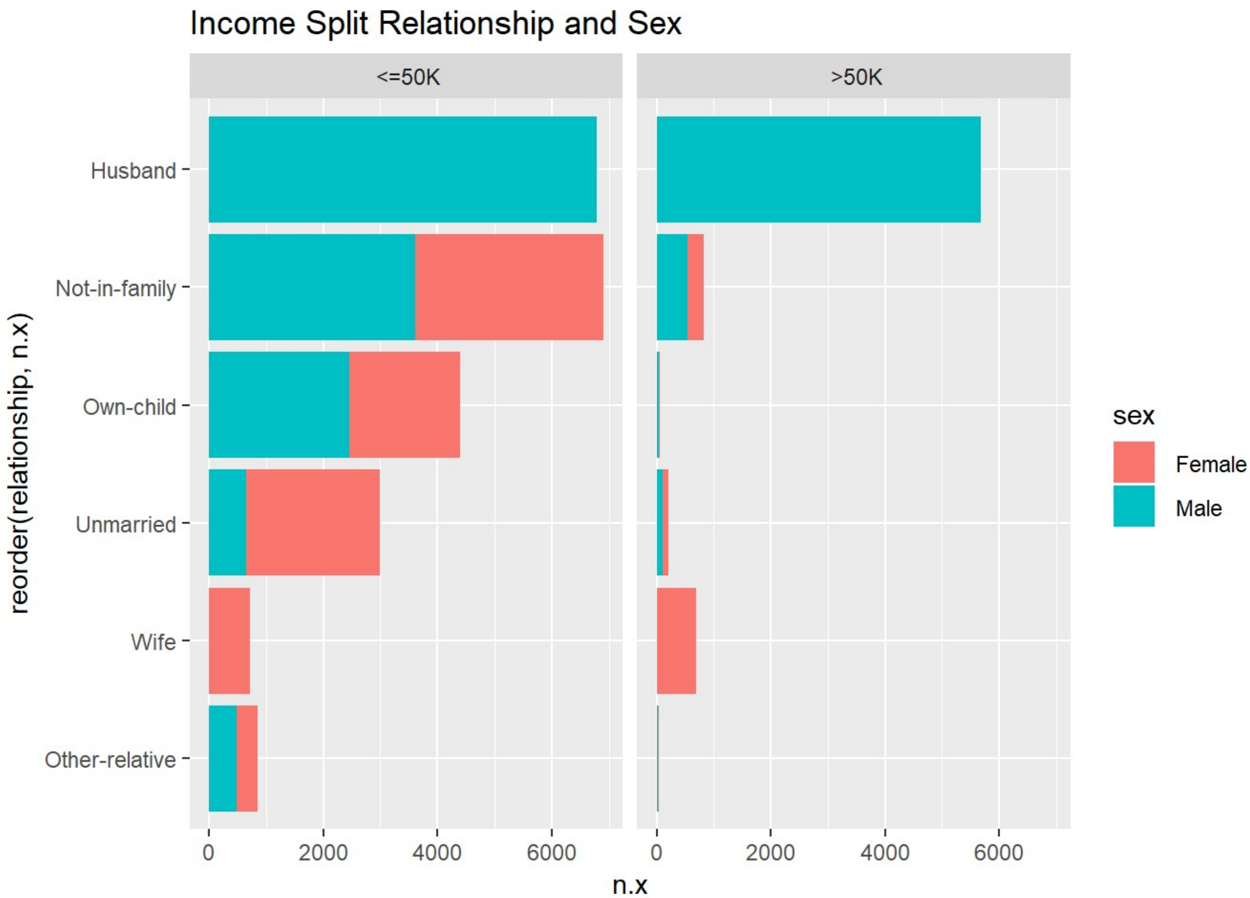
Regarding gender, the proportion of female and male in the *Young* age group is fairly even. But within the groups of *Middle_age* and *Senior* this proportion changes in favour of male.



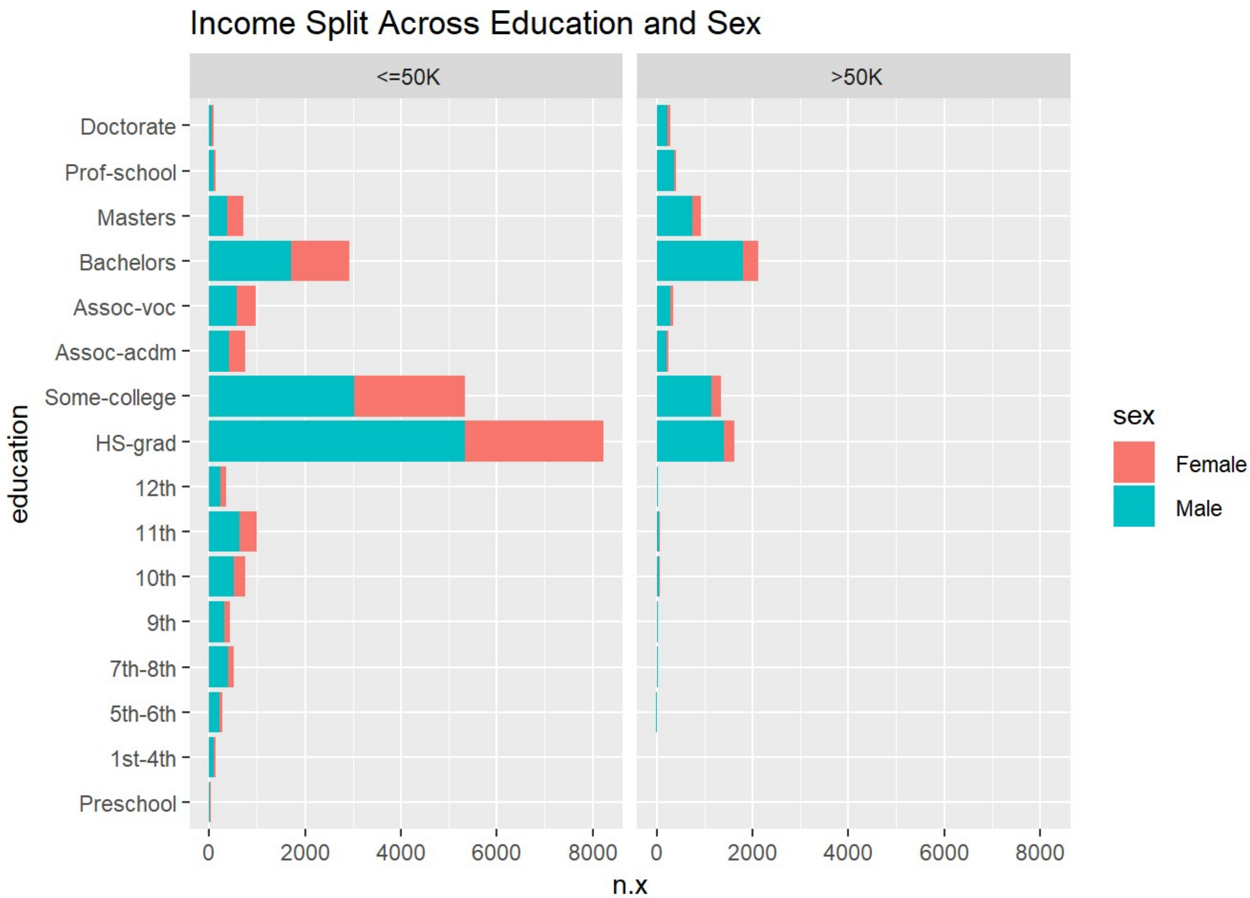

Could it be that because most married women decide to stay at home or choose to work part-time?

The chart *Income Split Across Marrial Status* reveals that the group *Married-civ-spouse* is dominated by *Male*. The male members of that group also have the highest numbers of people earning over 50k. The second largest group is the *Never-married* group. The proportion of female and male is quite even. It could be that this group consists mostly of young people.

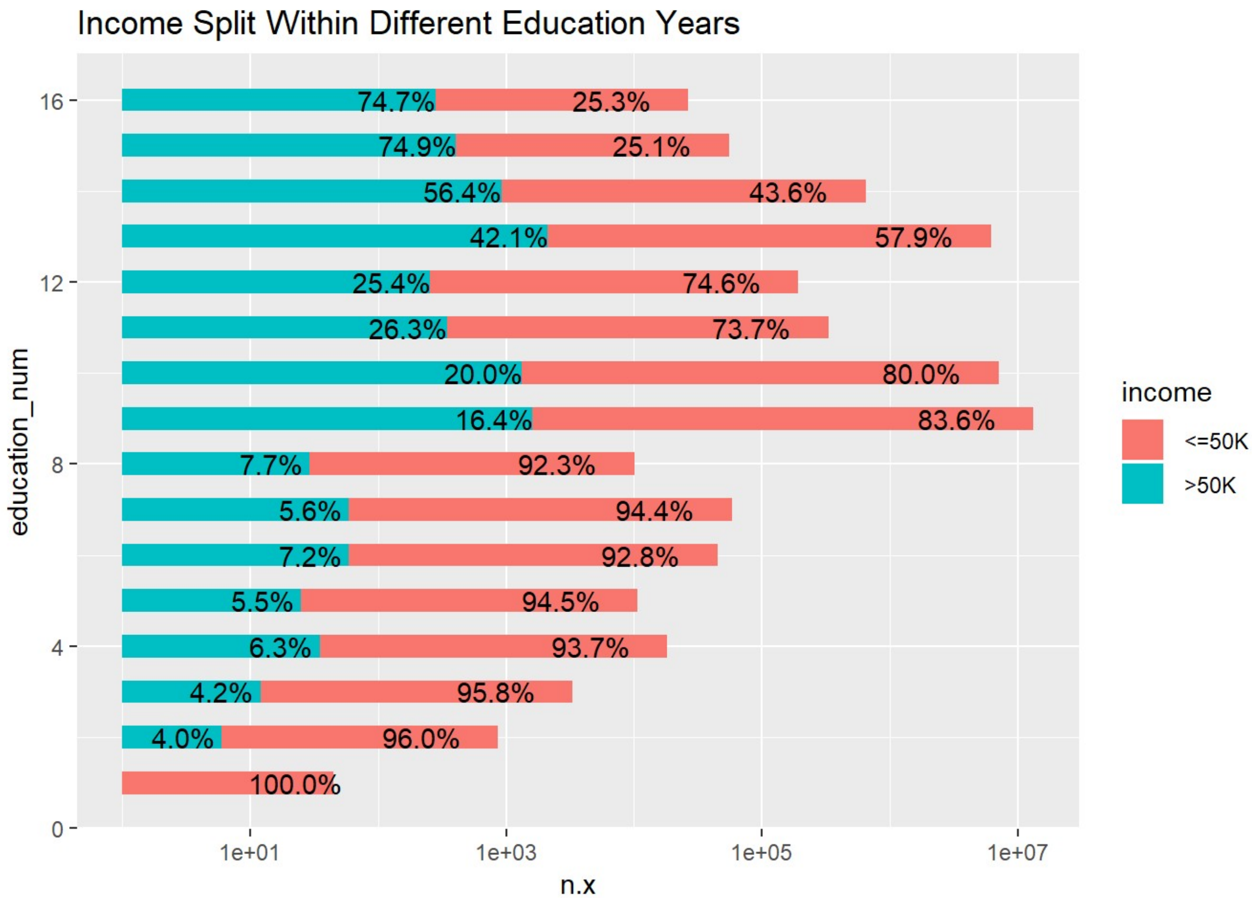

The chart *Income Split Across Work Hours Per Week and Sex* reveals that most women work full-time and over-time. Within the part-time jobs the proportion between female and male is quite even. I would have expected it to be dominated by females. Also here the females that choose to stay at home are not recognized.



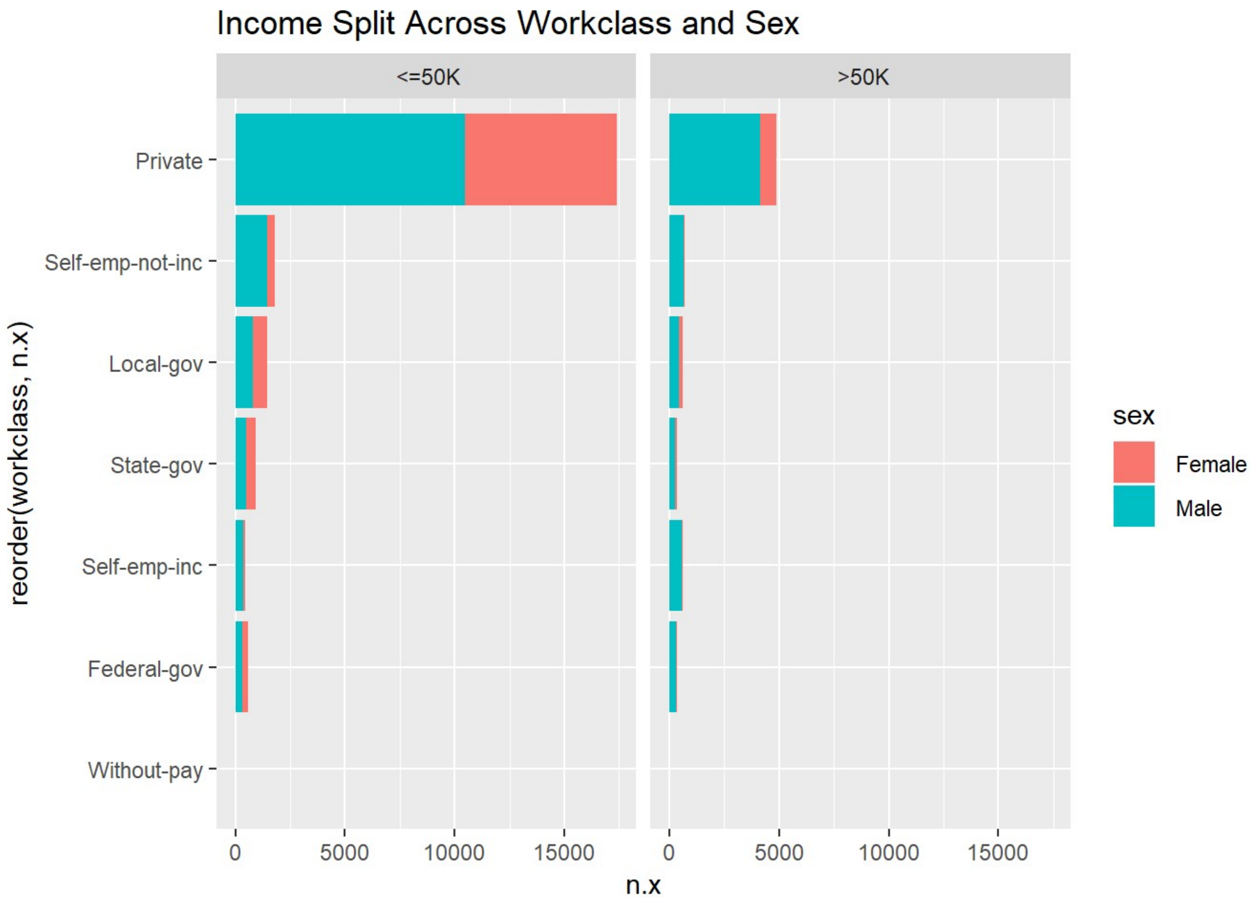

In the chart *Income Split Relationship and Sex*, the dominant group is *Husband*. But whom they are married to? The number of *Wife* is fairly small in the panel. It looks like married women decide to stay home. Looking at the groups *Not-in-family, Own-child, Unmarried*, and *Other relative*, the proportion of females and males are quite even. The group *Unmarried* shows a dominance of females.



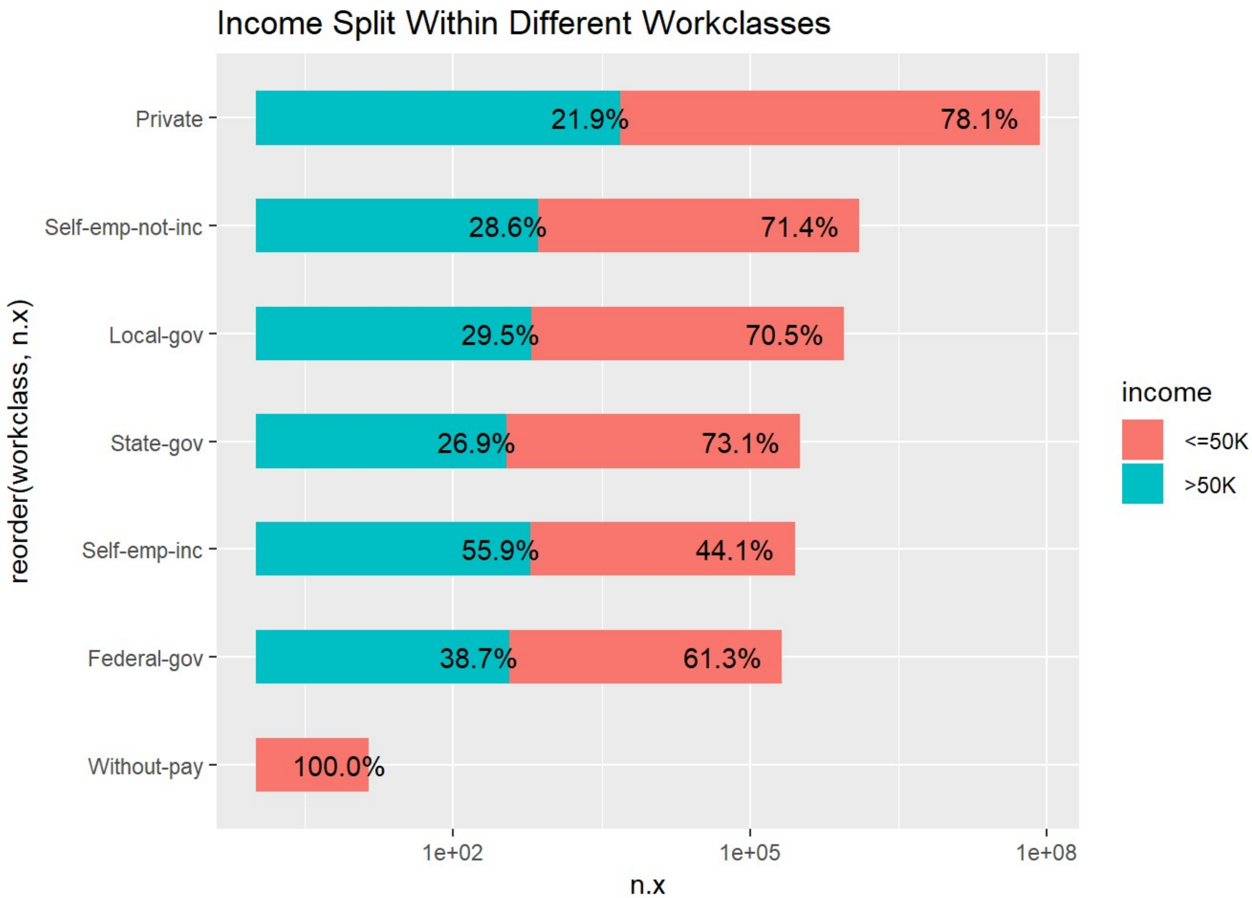

The chart *Income Split across Education and Sex* shows the majority of respondents have a *HS-grad, Some-college*, or *Bachelors*. People earning over 50k have at least a *HS-grad*. The higher the education , the higher proportion of people having earnings over 50k. Again, males dominate in every aspects.



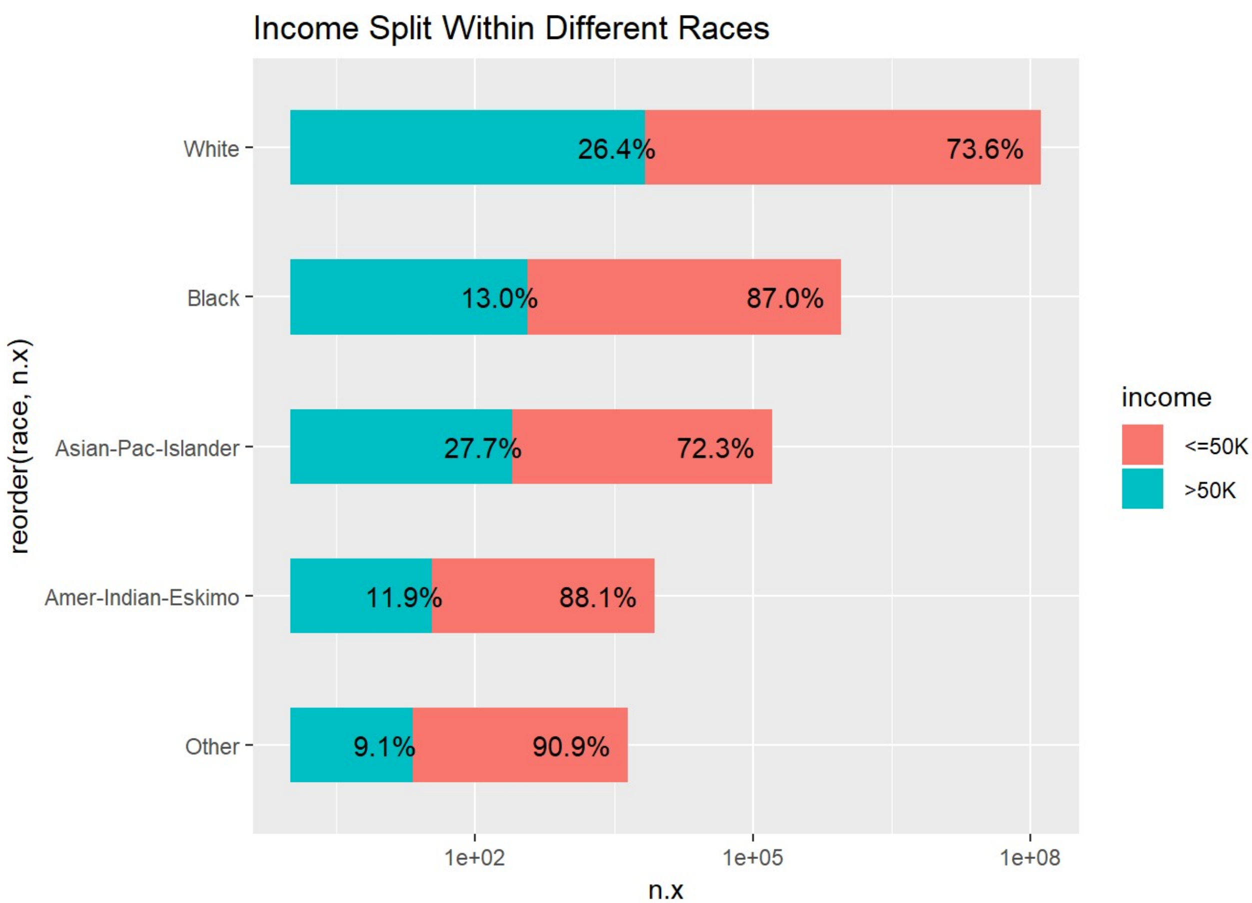

Another view on education provides the variable *education_num*. The chart *Income Split Within Different Education Years* demonstrates that the percentage of people earning over 50k are higher with more years of education spent. Note the horizontal axis is in log scale.



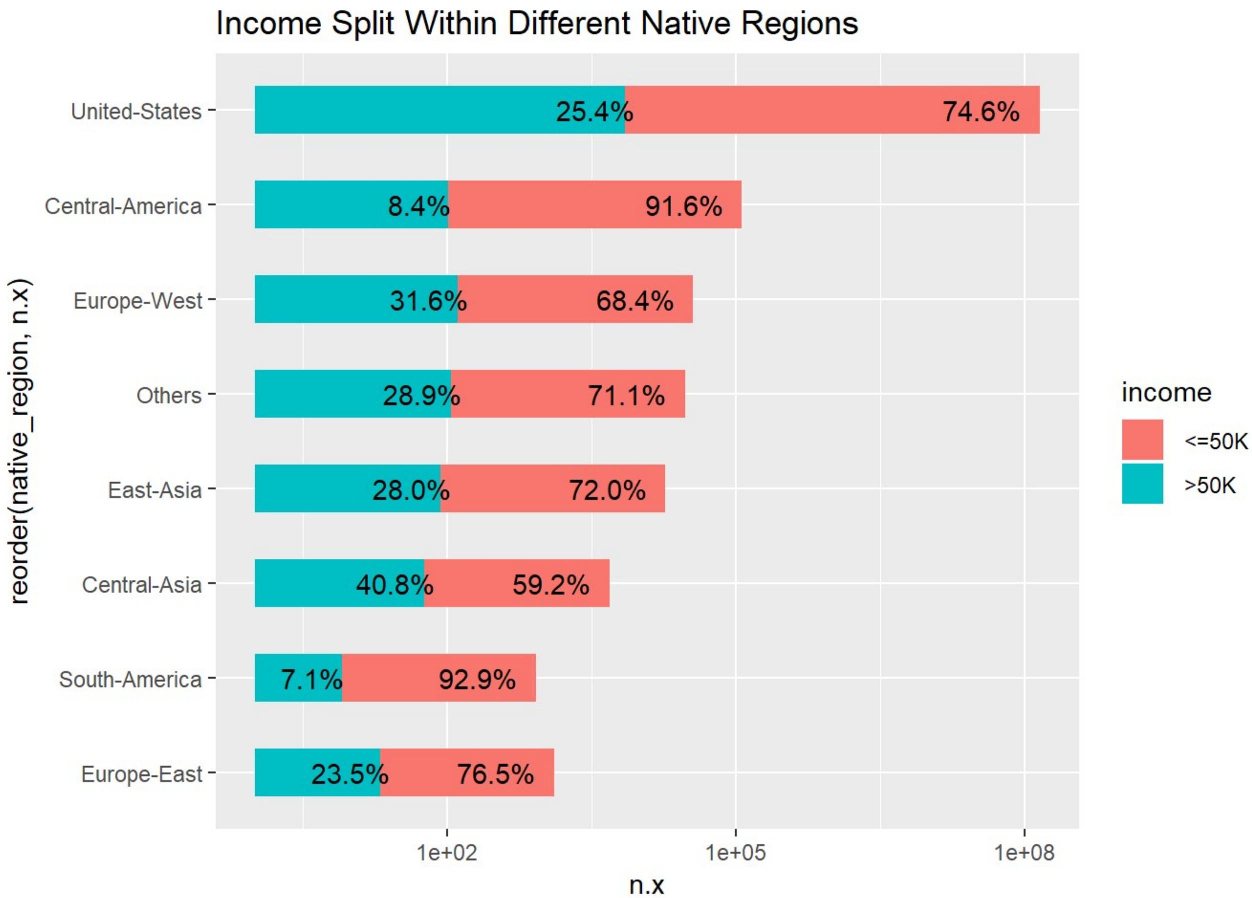

The chart *Income Split Across Workclass and Sex* reveals that most respondents are employed in the private sector.



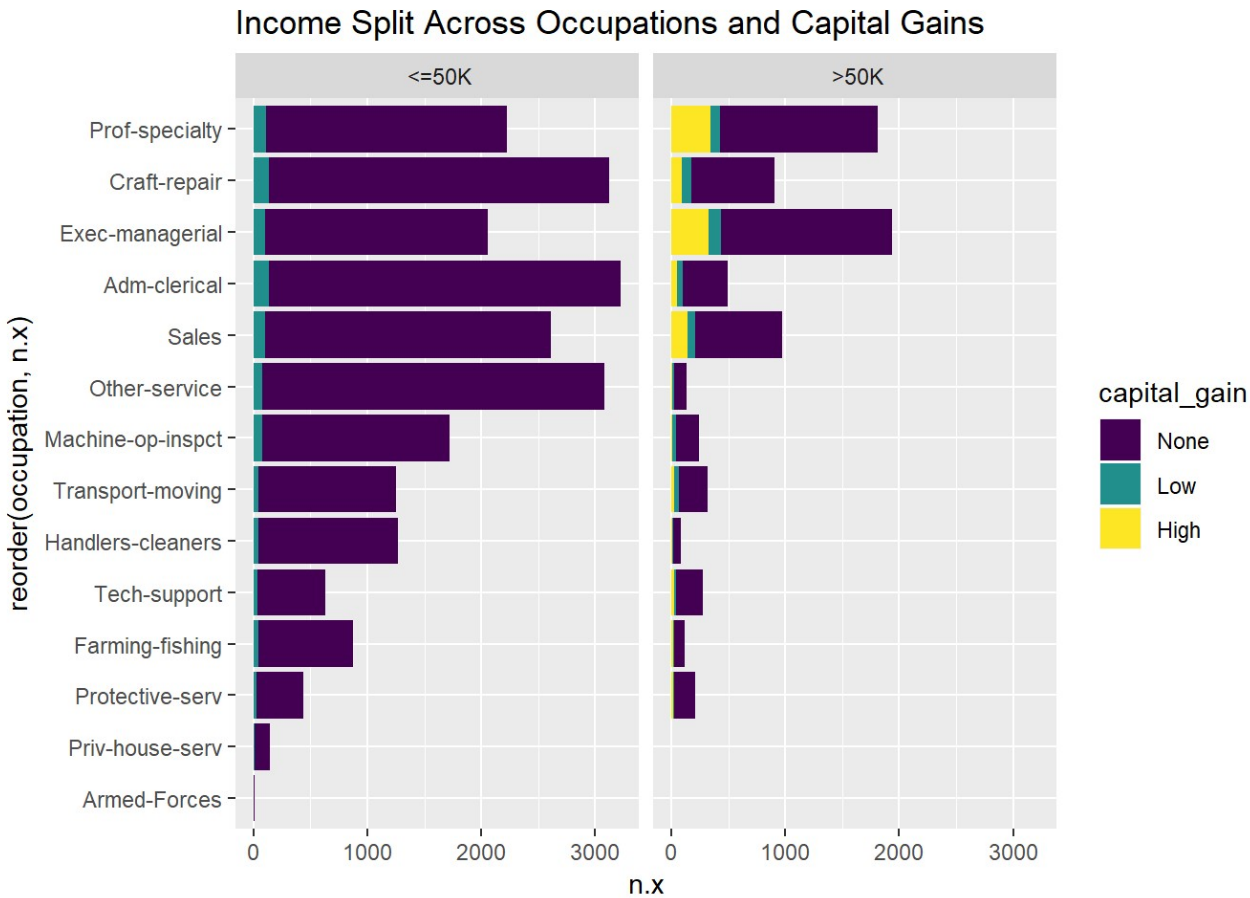

But the proportion of people earning over 50k is quite high, if they are self-employed or working in governmental sector, as outlined in the chart *Income Split Within Different Workclasses*. Note the horizontal axis is in log scale.



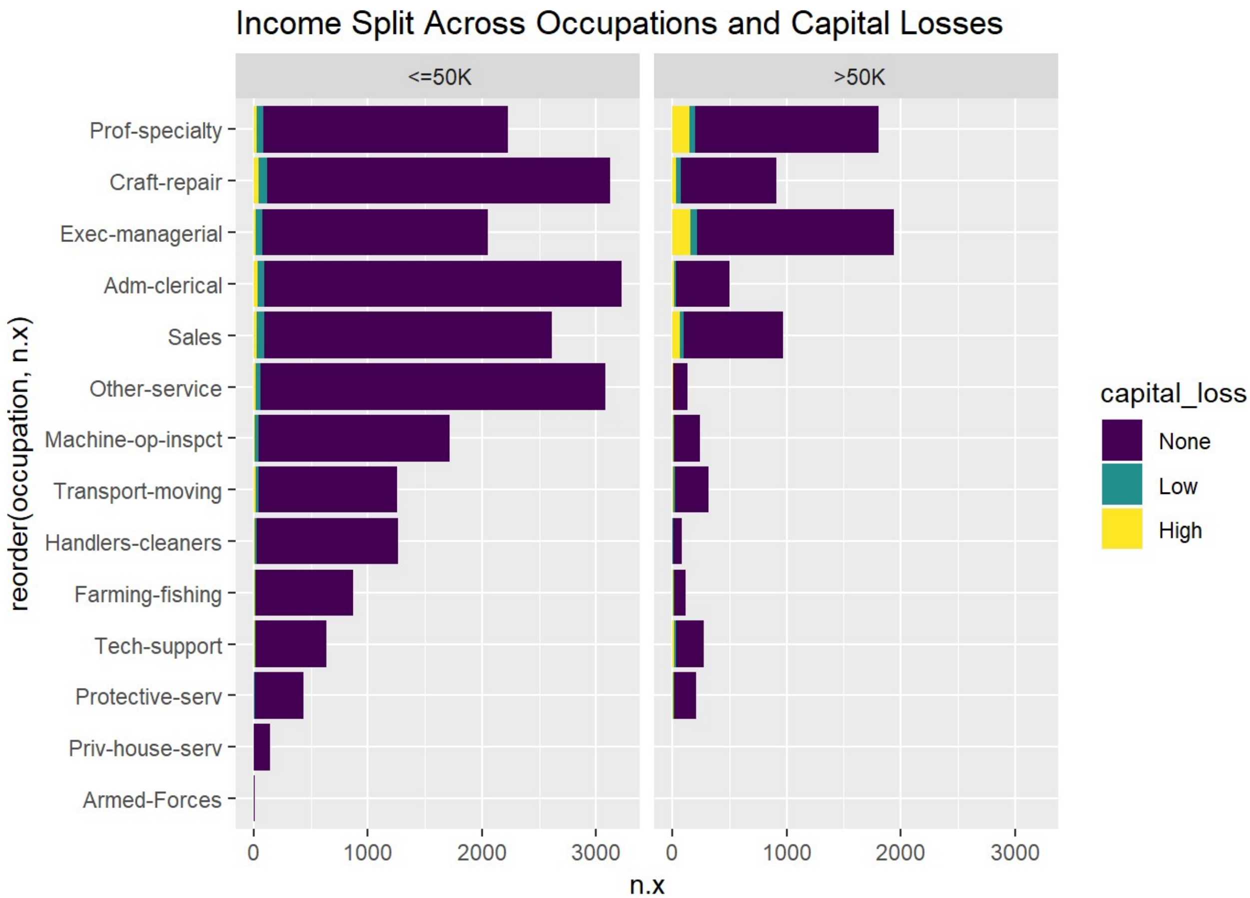

The dominant race in the panel is white. That is why the scale of the chart *Income Split Within Different Races* is in log scale. The income split within the *White* and *Asian-Pac-Islander* groups are fairly similar (26% to 74%). In contrast, the split within the *Black, Amer-Indian-Eskino*, and *Other* group is far more to 10% to 90%.




Depending where the respondents were born, there are differences in the proportion of the people earning over 50k. The groups *South-America* and *Central-America* have the lowest proportions of people with income over 50k compared to the group *United-States*. Whereas the groups *Central Asia, Europe-West, East-Asia*, and *Others* have higher proportions of high income earners. Note the horizontal axis is in log scale.




The chart *Income Split Across Occupations and Capital Gains* shows that most people do not have any capital gains. The few that have high capital gains are also likely to earn over 50k. They are the ones who can put some money for investments.




The chart *Income Split Across Occupations and Capital Losses* indicates like the chart before, that most people do not have any capital losses. The ones who earn more 50k, are likely to have capital loss, as they have money for investing.

# 3. Statistical Analysis

## 3.1. Logistic Regression

As mentioned before the dataset given is a panel with the dependent variable *income* as discrete variable. The task of the analysis is to predict whether someone earns over 50k. For such categorical issue the logistic regression approach is well suited.

The general idea of the approach is to give a probability if someone earns more than 50k based on the input of the dependent variables (note: *" >50k“*is at higher level in R).

$$Pr(income > 50k|IndependentVariables) = p$$

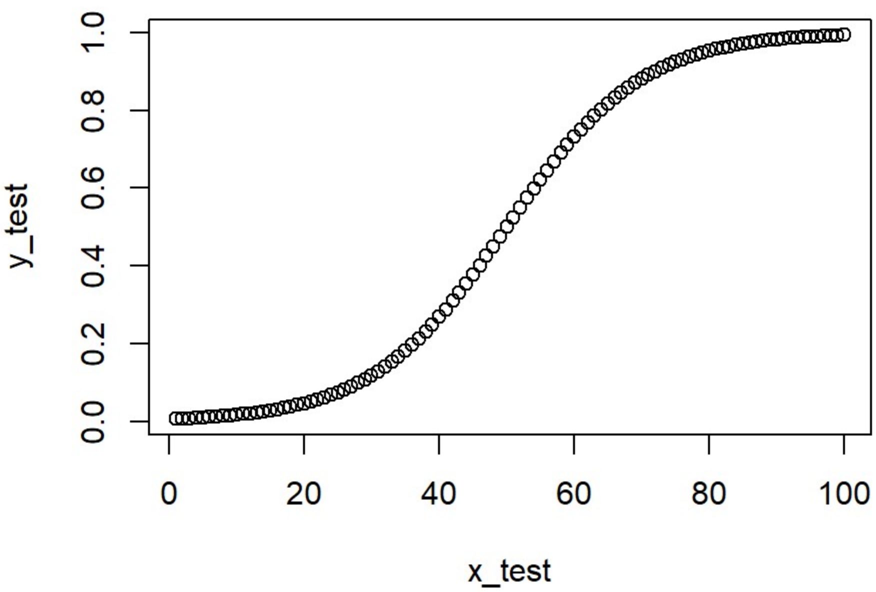
where

$$p = \frac{1}{1 + exp(-y)}$$

$$y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + .. + error$$

$$X_1, X_2, ... IndependentVariables$$

The probability function *p* has a s-shape curve and is bounded between 0 and 1. For illustration the probability function could look like this:




The graph suggests that the higher the x value the higher the probability of y. In this example, low x values (<40) implies y=0, and high x values (>60) implies y=1. What if we have observations with low x values but with y=1 or vice versa? So the shape of the probability function has to change. It could be flat, gradually increasing, or steep.

Given the distribution of the *income* variable, the logistic approach tries to find the one shape of the probability curve that can cover most of the observations of the variable *income*. The optimization method behind is the maximum likelihood estimation.

We are interested in p, the probability or earning over 50k. For estimation purpose the probability function will be transformed into a log odds function, i.e

$$ln(\frac{p}{1-p}) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + .. + error$$

(Reference: A nice illustrative introduction into logistic regression and maximum likelihood could be found here: link (https://www.youtube.com/watch?v=vN5cNN2-HWE))

# 3.2. Model Estimation

Before proceeding the ordered variables *age, education, capital_gains, capital_loss, hours_per_week* will be changed into unordered factors. The dataset *adult2* is copy of *adult* but with unordered factors.( Otherwise, the output will show for ordered factors in x.L, x.Q, x.C (linear, quadratic, cubic parameters)).

```
#changing ordered variables age,education, capital_gains, capital_loss, hours_per_week
into unordered factors

adult2 <- adult
adult2$age <- factor(adult2$age, ordered = FALSE)
adult2$education <- factor(adult2$education, ordered = FALSE)
adult2$capital_gain <- factor(adult2$capital_gain, ordered = FALSE)
adult2$capital_loss <- factor(adult2$capital_loss, ordered = FALSE)
adult2$hours_per_week <- factor(adult2$hours_per_week, ordered = FALSE)
```

The dataset will be split into a *train_set* and *test_set*. The train set will be used for estimation and optimization, while the test set will be only used for validation. 80% of the data will be used for training, while 20% will be for testing.

```
test_index <- createDataPartition(y = adult2$income, times = 1, p = 0.2, list = FALSE)
train_set <- adult2[-test_index,]
test_set <- adult2[test_index,]
```

The first run of the logistic regression will be with all independent variables:

**Dependent variable**: *income*

**Independent variable**: *age, worklclass,education, education_num, marial_status, occupation, relationship, race, sex, capital_gains, capital_loss, hours_per_week, native_region*

```
# logistic regression
# first run

glmfit <- glm(income~., data=train_set, family=binomial)

summary(glmfit)
```

```
##
## Call:
## glm(formula = income ~ ., family = binomial, data = train_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7909  -0.5121  -0.1814  -0.0003   4.0290
##
## Coefficients: (1 not defined because of singularities)
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                        -20.247389 107.832470  -0.188 0.851059
## ageMiddle_age                        1.472935   0.127523  11.550  < 2e-16
## ageSenior                            1.899989   0.131150  14.487  < 2e-16
## ageOld                               1.013157   0.188816   5.366 8.06e-08
## workclass Local-gov                 -0.692195   0.125908  -5.498 3.85e-08
## workclass Private                   -0.498525   0.104978  -4.749 2.05e-06
## workclass Self-emp-inc              -0.198613   0.139594  -1.423 0.154795
## workclass Self-emp-not-inc          -0.943625   0.123701  -7.628 2.38e-14
## workclass State-gov                 -0.757405   0.140132  -5.405 6.48e-08
## workclass Without-pay              -12.976058 211.619312  -0.061 0.951106
## education 1st-4th                   11.822250 107.832333   0.110 0.912698
## education 5th-6th                   11.899242 107.831675   0.110 0.912132
## education 7th-8th                   11.869233 107.831330   0.110 0.912352
## education 9th                       12.049991 107.831373   0.112 0.911023
## education 10th                      12.366115 107.831266   0.115 0.908699
## education 11th                      12.118643 107.831278   0.112 0.910518
## education 12th                      12.800739 107.831427   0.119 0.905505
## education HS-grad                   13.016071 107.831143   0.121 0.903922
## education Some-college              13.374446 107.831146   0.124 0.901290
## education Assoc-acdm                13.459271 107.831189   0.125 0.900668
## education Assoc-voc                 13.503661 107.831171   0.125 0.900342
## education Bachelors                 14.134188 107.831149   0.131 0.895714
## education Masters                   14.488255 107.831174   0.134 0.893117
## education Prof-school               14.960880 107.831253   0.139 0.889653
## education Doctorate                 15.211175 107.831293   0.141 0.887819
## education_num                              NA         NA      NA       NA
## marial_status Married-AF-spouse      3.111079   0.638262   4.874 1.09e-06
## marial_status Married-civ-spouse     2.102872   0.327614   6.419 1.37e-10
## marial_status Married-spouse-absent  0.183167   0.273115   0.671 0.502439
## marial_status Never-married         -0.423234   0.100750  -4.201 2.66e-05
## marial_status Separated             -0.083734   0.185630  -0.451 0.651934
## marial_status Widowed                0.457361   0.180496   2.534 0.011280
## occupation Armed-Forces             -0.962261   1.456454  -0.661 0.508812
## occupation Craft-repair              0.029705   0.090162   0.329 0.741806
## occupation Exec-managerial           0.835766   0.087683   9.532  < 2e-16
## occupation Farming-fishing          -0.800150   0.155868  -5.134 2.84e-07
## occupation Handlers-cleaners        -0.583784   0.160042  -3.648 0.000265
## occupation Machine-op-inspct        -0.210333   0.114234  -1.841 0.065584
```

```
## occupation Other-service          -0.716844   0.131731  -5.442 5.28e-08
## occupation Priv-house-serv        -2.944451   1.715947  -1.716 0.086174
## occupation Prof-specialty          0.488785   0.093077   5.251 1.51e-07
## occupation Protective-serv         0.691513   0.140801   4.911 9.05e-07
## occupation Sales                   0.298896   0.093685   3.190 0.001421
## occupation Tech-support            0.646417   0.126422   5.113 3.17e-07
## occupation Transport-moving       -0.006837   0.112065  -0.061 0.951354
## relationship Not-in-family         0.395918   0.324266   1.221 0.222098
## relationship Other-relative       -0.464036   0.292377  -1.587 0.112487
## relationship Own-child            -0.568524   0.323073  -1.760 0.078452
## relationship Unmarried             0.230732   0.341226   0.676 0.498922
## relationship Wife                  1.317034   0.119971  10.978  < 2e-16
## race Asian-Pac-Islander            0.799440   0.304097   2.629 0.008566
## race Black                         0.494973   0.266514   1.857 0.063281
## race Other                        -0.266123   0.449462  -0.592 0.553789
## race White                         0.659775   0.253975   2.598 0.009382
## sex Male                           0.864354   0.092765   9.318  < 2e-16
## capital_gainLow                    0.612967   0.078763   7.782 7.11e-15
## capital_gainHigh                   6.327731   0.364286  17.370  < 2e-16
## capital_lossLow                    0.647660   0.110846   5.843 5.13e-09
## capital_lossHigh                   1.591137   0.125358  12.693  < 2e-16
## hours_per_weekFull_time            0.871249   0.108939   7.998 1.27e-15
## hours_per_weekOver_time            1.340297   0.111913  11.976  < 2e-16
## hours_per_weekWorkaholic           1.288037   0.143322   8.987  < 2e-16
## native_region Central-Asia        -0.173465   0.320791  -0.541 0.588685
## native_region East-Asia            0.034102   0.292666   0.117 0.907239
## native_region Europe-East          0.541019   0.381939   1.417 0.156627
## native_region Europe-West          0.644867   0.219731   2.935 0.003338
## native_region Others               0.608300   0.251415   2.419 0.015542
## native_region South-America       -0.974081   0.551584  -1.766 0.077400
## native_region United-States        0.541707   0.153871   3.521 0.000431
##
## (Intercept)
## ageMiddle_age                     ***
## ageSenior                         ***
## ageOld                            ***
## workclass Local-gov               ***
## workclass Private                 ***
## workclass Self-emp-inc
## workclass Self-emp-not-inc        ***
## workclass State-gov               ***
## workclass Without-pay
## education 1st-4th
## education 5th-6th
## education 7th-8th
## education 9th
## education 10th
## education 11th
## education 12th
```

```
## education HS-grad
## education Some-college
## education Assoc-acdm
## education Assoc-voc
## education Bachelors
## education Masters
## education Prof-school
## education Doctorate
## education_num
## marial_status Married-AF-spouse     ***
## marial_status Married-civ-spouse    ***
## marial_status Married-spouse-absent
## marial_status Never-married         ***
## marial_status Separated
## marial_status Widowed               *
## occupation Armed-Forces
## occupation Craft-repair
## occupation Exec-managerial          ***
## occupation Farming-fishing          ***
## occupation Handlers-cleaners        ***
## occupation Machine-op-inspct        .
## occupation Other-service            ***
## occupation Priv-house-serv          .
## occupation Prof-specialty           ***
## occupation Protective-serv          ***
## occupation Sales                    **
## occupation Tech-support             ***
## occupation Transport-moving
## relationship Not-in-family
## relationship Other-relative
## relationship Own-child              .
## relationship Unmarried
## relationship Wife                   ***
## race Asian-Pac-Islander             **
## race Black                          .
## race Other
## race White                          **
## sex Male                            ***
## capital_gainLow                     ***
## capital_gainHigh                    ***
## capital_lossLow                     ***
## capital_lossHigh                    ***
## hours_per_weekFull_time             ***
## hours_per_weekOver_time             ***
## hours_per_weekWorkaholic            ***
## native_region Central-Asia
## native_region East-Asia
## native_region Europe-East
## native_region Europe-West           **
```

```
## native_region Others                 *
## native_region South-America         .
## native_region United-States         ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27079  on 24128  degrees of freedom
## Residual deviance: 15398  on 24061  degrees of freedom
## AIC: 15534
##
## Number of Fisher Scoring iterations: 13
```

# 3.3. Output Reading

Above is the output of the logistic regression income versus all other variables. The first column *coefficient* indicates the name of each variables used. The second column shows the estimates for each coefficient.

Surprisingly, there are more coefficients than the number of independent variables. R looks at the different levels of each variables.

For example the variable *sex* has two levels, i.e. *female* and *male*. As reference to female, the coefficient estimate of 0.864354 means that relative to a female the odds for an income >50k increases with male. Note, the estimation is in a *ln(p/1-p)* world. The more positive the number the higher the odds.

Let`s have a look at the variable *relationship* for example. The one missing level is *Husband*. Relative to someone who is husband, if someone is *Not-in-family*, his or her odds of having an income >50k increases. Similarly, relative to a husband, if someone is a wife, the odds that this woman earns over 50k is high.

Quite strange when considering that most women in the panel earn less than 50k. But it could be, that married women who decide to work, choose to do so, because they can make over 50k.

On the far right column, the p-value for all coefficients are listed. Next to them are either several stars, or nothing.

The three stars $***$ means that with 99.90% confidence level, the respective coefficient is statistically significant different from zero. ($**$ = 99.00%; $*$ = 95% ; . = 90%). Put differently, the variables associated with the significant coefficients do have influence on the outcome of the variable income.

It looks like that *education* seems to be insignificant. Whereas with other variables, there are some or all levels that are significant. And with regards to the variable *education_num* there are **NAs** in the output.

Intuitively one would expect that education should have a significant influence whether someone earns over 50k. As shown in the data exploration part, there should be some correlation between education and the probability of earning over 50k. But the regression result does not support this thesis.

(Reference: A illustrative example how to run the logistic regression and read the output, could be be found here: link (https://www.youtube.com/watch?v=AVx7Wc1CQ7Y) )

# 3.4. Optimizing the Model

## 3.4.1. Multicollinearity

Multicollinearity exists if the independent variables are not independent from each other. The estimates of the model will have high standard errors resulting into insignificant estimates. (Reference: link (http://scg.sdsu.edu/logit_r/)).

My suspicion is that the variables *education* and *education_num* are highly correlated.

In the second run, the variable education will be removed

```
# first run with variables results into NA estimates for education_num,
# my suspiscion is education and education_num are highly correlated
# new run without education

glmfit <- glm(income~.-education, data=train_set, family=binomial)
summary(glmfit)
```

```
##
## Call:
## glm(formula = income ~ . - education, family = binomial, data = train_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7429  -0.5123  -0.1819  -0.0075   4.1421
##
## Coefficients:
##                                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)                           -9.632712   0.504742 -19.084  < 2e-16
## ageMiddle_age                          1.480872   0.127317  11.631  < 2e-16
## ageSenior                              1.923371   0.130749  14.710  < 2e-16
## ageOld                                 1.058426   0.187706   5.639 1.71e-08
## workclass Local-gov                   -0.692704   0.125358  -5.526 3.28e-08
## workclass Private                     -0.495936   0.104750  -4.734 2.20e-06
## workclass Self-emp-inc                -0.191600   0.139360  -1.375 0.169174
## workclass Self-emp-not-inc            -0.925507   0.123169  -7.514 5.73e-14
## workclass State-gov                   -0.732990   0.139096  -5.270 1.37e-07
## workclass Without-pay                -11.955786 128.027227  -0.093 0.925598
## education_num                          0.272686   0.010917  24.979  < 2e-16
## marial_status Married-AF-spouse        3.077141   0.640701   4.803 1.56e-06
## marial_status Married-civ-spouse       2.088674   0.328674   6.355 2.09e-10
## marial_status Married-spouse-absent    0.191776   0.272192   0.705 0.481082
## marial_status Never-married           -0.403302   0.100080  -4.030 5.58e-05
## marial_status Separated               -0.084115   0.185176  -0.454 0.649655
## marial_status Widowed                  0.457020   0.179808   2.542 0.011031
## occupation Armed-Forces               -0.896411   1.442013  -0.622 0.534179
## occupation Craft-repair                0.028588   0.089919   0.318 0.750541
## occupation Exec-managerial             0.851701   0.087097   9.779  < 2e-16
## occupation Farming-fishing            -0.807205   0.155754  -5.183 2.19e-07
## occupation Handlers-cleaners          -0.567911   0.159800  -3.554 0.000380
## occupation Machine-op-inspct          -0.202476   0.113875  -1.778 0.075397
## occupation Other-service              -0.715635   0.131639  -5.436 5.44e-08
## occupation Priv-house-serv            -2.884183   1.668255  -1.729 0.083834
## occupation Prof-specialty              0.555856   0.090609   6.135 8.53e-10
## occupation Protective-serv             0.689490   0.140672   4.901 9.51e-07
## occupation Sales                       0.307240   0.093351   3.291 0.000997
## occupation Tech-support                0.632966   0.125926   5.026 5.00e-07
## occupation Transport-moving           -0.001486   0.111694  -0.013 0.989388
## relationship Not-in-family             0.379048   0.325259   1.165 0.243868
## relationship Other-relative           -0.495570   0.291980  -1.697 0.089645
## relationship Own-child                -0.582925   0.324037  -1.799 0.072028
## relationship Unmarried                 0.202886   0.342077   0.593 0.553115
## relationship Wife                      1.304874   0.119636  10.907  < 2e-16
## race Asian-Pac-Islander                0.790527   0.303451   2.605 0.009184
## race Black                             0.493204   0.266165   1.853 0.063883
## race Other                            -0.281627   0.449679  -0.626 0.531129
```

```
## race White                            0.658082   0.253608   2.595 0.009462
## sex Male                              0.861969   0.092333   9.335  < 2e-16
## capital_gainLow                       0.610842   0.078689   7.763 8.31e-15
## capital_gainHigh                      6.210724   0.347898  17.852  < 2e-16
## capital_lossLow                       0.645110   0.110931   5.815 6.05e-09
## capital_lossHigh                      1.600824   0.125196  12.787  < 2e-16
## hours_per_weekFull_time               0.875709   0.108521   8.069 7.06e-16
## hours_per_weekOver_time               1.348473   0.111508  12.093  < 2e-16
## hours_per_weekWorkaholic              1.311258   0.142773   9.184  < 2e-16
## native_region Central-Asia           -0.181209   0.317510  -0.571 0.568189
## native_region East-Asia               0.030569   0.290513   0.105 0.916199
## native_region Europe-East             0.464070   0.380676   1.219 0.222818
## native_region Europe-West             0.606331   0.219846   2.758 0.005816
## native_region Others                  0.589719   0.250261   2.356 0.018452
## native_region South-America          -1.008816   0.545732  -1.849 0.064522
## native_region United-States           0.492533   0.151924   3.242 0.001187
##
## (Intercept)                         ***
## ageMiddle_age                       ***
## ageSenior                           ***
## ageOld                              ***
## workclass Local-gov                 ***
## workclass Private                   ***
## workclass Self-emp-inc
## workclass Self-emp-not-inc          ***
## workclass State-gov                 ***
## workclass Without-pay
## education_num                       ***
## marial_status Married-AF-spouse     ***
## marial_status Married-civ-spouse    ***
## marial_status Married-spouse-absent
## marial_status Never-married         ***
## marial_status Separated
## marial_status Widowed               *
## occupation Armed-Forces
## occupation Craft-repair
## occupation Exec-managerial          ***
## occupation Farming-fishing          ***
## occupation Handlers-cleaners        ***
## occupation Machine-op-inspct        .
## occupation Other-service            ***
## occupation Priv-house-serv          .
## occupation Prof-specialty           ***
## occupation Protective-serv          ***
## occupation Sales                    ***
## occupation Tech-support             ***
## occupation Transport-moving
## relationship Not-in-family
## relationship Other-relative         .
```

```
## relationship Own-child              .
## relationship Unmarried
## relationship Wife                   ***
## race Asian-Pac-Islander             **
## race Black                          .
## race Other
## race White                          **
## sex Male                            ***
## capital_gainLow                     ***
## capital_gainHigh                    ***
## capital_lossLow                     ***
## capital_lossHigh                    ***
## hours_per_weekFull_time             ***
## hours_per_weekOver_time             ***
## hours_per_weekWorkaholic            ***
## native_region Central-Asia
## native_region East-Asia
## native_region Europe-East
## native_region Europe-West           **
## native_region Others                *
## native_region South-America         .
## native_region United-States         **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27079  on 24128  degrees of freedom
## Residual deviance: 15433  on 24075  degrees of freedom
## AIC: 15541
##
## Number of Fisher Scoring iterations: 12
```

The new output shows no **NAs** for *education_num*.

I also omit the *education_num* variable, and keep *education* instead. The result shows no significance for all level coefficients of *education*, which does not make sense. So the preferred model includes the variable *education_num*.

To detect for further mulitcollinaerity, the function *vif()* will be used. Variables with values over 5 exhibit correlation with each other.

```
vif(glmfit)
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## age              1.207361  3        1.031905
## workclass        1.559614  6        1.037731
## education_num    1.521866  1        1.233640
## marial_status   55.008053  6        1.396483
## occupation       2.508688 13        1.036009
## relationship   133.146401  5        1.630921
## race             2.275787  4        1.108259
## sex              2.958475  1        1.720022
## capital_gain     1.045305  2        1.011139
## capital_loss     1.017902  2        1.004446
## hours_per_week   1.257477  3        1.038923
## native_region    2.259155  7        1.059941
```

The output of the *vif()* function indicates that *marial_status* and *relationship* are highly correlated. The next runs will be one without *relationship* and one without *marial_status*.

```
# attempt to eliminate variables with a GVIF higher than 5,
# the variables marial_status and relationship do have GVIF values >5.
# regression without variable relationship

glmfit <- glm(income~.-education-relationship, data=train_set, family=binomial)
summary(glmfit)
```

```
##
## Call:
## glm(formula = income ~ . - education - relationship, family = binomial,
##     data = train_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5745  -0.5162  -0.2032  -0.0097   4.0565
##
## Coefficients:
##                                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)                            -8.99162    0.37008 -24.296  < 2e-16
## ageMiddle_age                           1.55874    0.12543  12.427  < 2e-16
## ageSenior                               1.98798    0.12869  15.447  < 2e-16
## ageOld                                  1.09383    0.18571   5.890 3.86e-09
## workclass Local-gov                    -0.70788    0.12418  -5.700 1.19e-08
## workclass Private                      -0.50302    0.10405  -4.834 1.33e-06
## workclass Self-emp-inc                 -0.18812    0.13941  -1.349 0.177212
## workclass Self-emp-not-inc             -0.91820    0.12280  -7.477 7.59e-14
## workclass State-gov                    -0.74004    0.13840  -5.347 8.93e-08
## workclass Without-pay                 -11.81414  130.21912  -0.091 0.927711
## education_num                           0.27318    0.01089  25.080  < 2e-16
## marial_status Married-AF-spouse         3.38964    0.55173   6.144 8.07e-10
## marial_status Married-civ-spouse        2.22677    0.07644  29.132  < 2e-16
## marial_status Married-spouse-absent     0.25730    0.26696   0.964 0.335131
## marial_status Never-married            -0.37355    0.09453  -3.952 7.76e-05
## marial_status Separated                -0.09889    0.18124  -0.546 0.585315
## marial_status Widowed                   0.29502    0.17624   1.674 0.094145
## occupation Armed-Forces                -1.03940    1.43162  -0.726 0.467819
## occupation Craft-repair                -0.03542    0.08860  -0.400 0.689335
## occupation Exec-managerial              0.81700    0.08529   9.579  < 2e-16
## occupation Farming-fishing             -0.87998    0.15544  -5.661 1.50e-08
## occupation Handlers-cleaners           -0.61564    0.15956  -3.858 0.000114
## occupation Machine-op-inspct           -0.24858    0.11302  -2.200 0.027841
## occupation Other-service               -0.70945    0.13031  -5.444 5.20e-08
## occupation Priv-house-serv             -3.17469    1.62755  -1.951 0.051106
## occupation Prof-specialty               0.52829    0.08887   5.944 2.77e-09
## occupation Protective-serv              0.64802    0.14049   4.612 3.98e-06
## occupation Sales                        0.25858    0.09192   2.813 0.004906
## occupation Tech-support                 0.59852    0.12412   4.822 1.42e-06
## occupation Transport-moving            -0.05964    0.11087  -0.538 0.590632
## race Asian-Pac-Islander                 0.74514    0.30012   2.483 0.013033
## race Black                              0.49519    0.26340   1.880 0.060109
## race Other                             -0.29956    0.44981  -0.666 0.505439
## race White                              0.67028    0.25112   2.669 0.007604
## sex Male                                0.16311    0.06079   2.683 0.007289
## capital_gainLow                         0.61003    0.07867   7.755 8.87e-15
## capital_gainHigh                        6.15038    0.33494  18.363  < 2e-16
```

```
## capital_lossLow                     0.64747    0.11034   5.868 4.41e-09
## capital_lossHigh                    1.64180    0.12543  13.090  < 2e-16
## hours_per_weekFull_time             0.80854    0.10667   7.580 3.45e-14
## hours_per_weekOver_time             1.28738    0.10954  11.752  < 2e-16
## hours_per_weekWorkaholic            1.23486    0.14138   8.734  < 2e-16
## native_region Central-Asia         -0.17230    0.31777  -0.542 0.587665
## native_region East-Asia             0.04555    0.28788   0.158 0.874290
## native_region Europe-East           0.41575    0.38056   1.092 0.274637
## native_region Europe-West           0.62085    0.21882   2.837 0.004550
## native_region Others                0.55270    0.24800   2.229 0.025839
## native_region South-America        -1.02662    0.54803  -1.873 0.061028
## native_region United-States         0.48878    0.15116   3.234 0.001222
##
## (Intercept)                       ***
## ageMiddle_age                     ***
## ageSenior                         ***
## ageOld                            ***
## workclass Local-gov               ***
## workclass Private                 ***
## workclass Self-emp-inc
## workclass Self-emp-not-inc        ***
## workclass State-gov               ***
## workclass Without-pay
## education_num                     ***
## marial_status Married-AF-spouse   ***
## marial_status Married-civ-spouse  ***
## marial_status Married-spouse-absent
## marial_status Never-married       ***
## marial_status Separated
## marial_status Widowed             .
## occupation Armed-Forces
## occupation Craft-repair
## occupation Exec-managerial        ***
## occupation Farming-fishing        ***
## occupation Handlers-cleaners      ***
## occupation Machine-op-inspct      *
## occupation Other-service          ***
## occupation Priv-house-serv        .
## occupation Prof-specialty         ***
## occupation Protective-serv        ***
## occupation Sales                  **
## occupation Tech-support           ***
## occupation Transport-moving
## race Asian-Pac-Islander           *
## race Black                        .
## race Other
## race White                        **
## sex Male                          **
## capital_gainLow                   ***
```

```
## capital_gainHigh                   ***
## capital_lossLow                    ***
## capital_lossHigh                   ***
## hours_per_weekFull_time            ***
## hours_per_weekOver_time            ***
## hours_per_weekWorkaholic           ***
## native_region Central-Asia
## native_region East-Asia
## native_region Europe-East
## native_region Europe-West          **
## native_region Others               *
## native_region South-America        .
## native_region United-States        **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27079  on 24128  degrees of freedom
## Residual deviance: 15613  on 24080  degrees of freedom
## AIC: 15711
##
## Number of Fisher Scoring iterations: 12
```

```
vif(glmfit) # all remaining independent variables no longer correlated
```

```
##                   GVIF Df GVIF^(1/(2*Df))
## age           1.194513  3        1.030066
## workclass     1.556447  6        1.037555
## education_num 1.521066  1        1.233315
## marial_status 1.481719  6        1.033310
## occupation    2.481391 13        1.035573
## race          2.267818  4        1.107774
## sex           1.437121  1        1.198800
## capital_gain  1.041990  2        1.010336
## capital_loss  1.017195  2        1.004271
## hours_per_week 1.236800  3        1.036056
## native_region  2.240050  7        1.059299
```

```
# regression without variable marial_status
# residual deviance and AIC number are worse than witout marial_status instead
# all remaing variable independent variables no longer correlated

glmfit <- glm(income~.-education-marial_status, data=train_set, family=binomial)
summary(glmfit)
```

```
##
## Call:
## glm(formula = income ~ . - education - marial_status, family = binomial,
##     data = train_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7478  -0.5113  -0.1898  -0.0089   3.8949
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -7.523834   0.379308 -19.836  < 2e-16 ***
## ageMiddle_age                  1.544306   0.126048  12.252  < 2e-16 ***
## ageSenior                      2.018486   0.128923  15.656  < 2e-16 ***
## ageOld                         1.178886   0.186294   6.328 2.48e-10 ***
## workclass Local-gov           -0.693859   0.125100  -5.546 2.92e-08 ***
## workclass Private             -0.503106   0.104409  -4.819 1.45e-06 ***
## workclass Self-emp-inc        -0.181959   0.139184  -1.307 0.191101
## workclass Self-emp-not-inc    -0.927045   0.122904  -7.543 4.60e-14 ***
## workclass State-gov           -0.734604   0.138603  -5.300 1.16e-07 ***
## workclass Without-pay        -11.902160 125.899696  -0.095 0.924683
## education_num                  0.269208   0.010882  24.738  < 2e-16 ***
## occupation Armed-Forces       -0.777232   1.489822  -0.522 0.601883
## occupation Craft-repair        0.037314   0.089780   0.416 0.677690
## occupation Exec-managerial     0.857472   0.086902   9.867  < 2e-16 ***
## occupation Farming-fishing    -0.798932   0.155475  -5.139 2.77e-07 ***
## occupation Handlers-cleaners  -0.571943   0.159504  -3.586 0.000336 ***
## occupation Machine-op-inspct  -0.200211   0.113767  -1.760 0.078436 .
## occupation Other-service      -0.713978   0.131497  -5.430 5.65e-08 ***
## occupation Priv-house-serv    -2.544381   1.599622  -1.591 0.111697
## occupation Prof-specialty      0.552808   0.090302   6.122 9.25e-10 ***
## occupation Protective-serv     0.699438   0.140624   4.974 6.57e-07 ***
## occupation Sales               0.312945   0.093153   3.359 0.000781 ***
## occupation Tech-support        0.620257   0.125347   4.948 7.49e-07 ***
## occupation Transport-moving    0.003443   0.111618   0.031 0.975394
## relationship Not-in-family    -1.898352   0.064990 -29.210  < 2e-16 ***
## relationship Other-relative   -1.949966   0.234881  -8.302  < 2e-16 ***
## relationship Own-child        -2.790702   0.165457 -16.867  < 2e-16 ***
## relationship Unmarried        -1.915488   0.113653 -16.854  < 2e-16 ***
## relationship Wife              1.285768   0.118548  10.846  < 2e-16 ***
## race Asian-Pac-Islander        0.787969   0.304156   2.591 0.009579 **
## race Black                     0.467616   0.267257   1.750 0.080172 .
## race Other                    -0.303456   0.449868  -0.675 0.499965
## race White                     0.650173   0.254905   2.551 0.010752 *
## sex Male                       0.827695   0.090772   9.118  < 2e-16 ***
## capital_gainLow                0.621765   0.078587   7.912 2.54e-15 ***
## capital_gainHigh               6.193122   0.350836  17.652  < 2e-16 ***
## capital_lossLow                0.630744   0.110542   5.706 1.16e-08 ***
```

```
## capital_lossHigh              1.606539   0.125251  12.827  < 2e-16 ***
## hours_per_weekFull_time       0.869684   0.108444   8.020 1.06e-15 ***
## hours_per_weekOver_time       1.342282   0.111388  12.050  < 2e-16 ***
## hours_per_weekWorkaholic      1.315601   0.142620   9.224  < 2e-16 ***
## native_region Central-Asia   -0.141978   0.318446  -0.446 0.655707    
## native_region East-Asia       0.047483   0.289659   0.164 0.869789    
## native_region Europe-East     0.465345   0.382529   1.216 0.223795    
## native_region Europe-West     0.590041   0.219111   2.693 0.007084 ** 
## native_region Others          0.597875   0.250214   2.389 0.016873 *  
## native_region South-America  -1.015890   0.544282  -1.866 0.061975 .  
## native_region United-States   0.482154   0.151479   3.183 0.001458 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27079  on 24128  degrees of freedom
## Residual deviance: 15513  on 24081  degrees of freedom
## AIC: 15609
##
## Number of Fisher Scoring iterations: 12
```

```
vif(glmfit) # all remaining independent variables no longer correlated
```

```
##                    GVIF Df GVIF^(1/(2*Df))
## age            1.157519  3        1.024679
## workclass      1.557151  6        1.037594
## education_num  1.526267  1        1.235422
## occupation     2.500057 13        1.035871
## relationship   3.244141  5        1.124890
## race           2.260429  4        1.107322
## sex            2.867414  1        1.693344
## capital_gain   1.042535  2        1.010468
## capital_loss   1.016521  2        1.004105
## hours_per_week 1.253012  3        1.038307
## native_region  2.235458  7        1.059143
```

In both regression, the *vif()* function no longer shows any multicollinearity issues. But which model is best?

## 3.4.2. Deviance

On the bottom of the output of the logistic regression, the numbers for null deviance, residual deviance, and the AIC are given.

These numbers show how well the model fits to the data. The lower the numbers the better the model. The null deviance number shows how well the model fits with only the intercept. The residual deviance number and AIC measure the fitness with inclusion of the independent variable.

(Reference: What deviance means, could be found here: link (https://www.youtube.com/watch?v=B2nJ3U4E1VA))

The model without the variable *relationship* shows the following deviance and AIC numbers:

*Residual deviance: 15613 on 24080 degrees of freedom*

*AIC: 15711*

The model without the variable *marial_status* has the following deviance and AIC numbers:

*Residual deviance: 15513 on 24081 degrees of freedom*

*AIC: 15609*

The model without *education* and *marial_status* is the best performing model. The residual deviance and the AIC number are less than the model with relationship instead.

The final model looks like this:

**Dependent variable**: *income*

**Independent variable**: *age, worklclass, education_num, occupation, relationship, race, sex, capital_gains, capital_loss, hours_per_week, native_region*

**Omitted variabels**: *fnlwgt, education, marial_status*

# 3.5 Interpretation of Model Estimates

In plain English, the features someone has which makes him likely to earn over 50k:

You are no longer a young person. If you work for the Federal Government, your odds is better than other people. The more time you have invested in your education the more likely you will earn over 50k.

If you work in areas of Exec-managerial, Prof-specialty, Protective-service, Sales, or Tech-support, your chances are high to have over 50k. As a husband your odds are higher than most people, except if you are a working wife. In this case your chances are even higher to earn over 50k.

If you are not from any American-Indian-Eskimo minorities, the likelihood increases also. You are likely a male. If you have any capital gains or losses also indicates a high probability. If you work part-time, you are less likely to earn more than other people.

And if you are born from Central America, you are less likely to make 50k compared to others who are born in Western Europe or in the US.

# 3.6. Finding the Right Threshold

Recall that the estimates were obtained in the *ln(p/1-p)* world. Using the estimates from that regression, the *predict()* function returns a vector of probabilities.

Intuitively, one would say that a probability over 0.5 could be a good indication if someone earns over 50k.

The table below shows for the predicted value *FALSE*, if someone earns less than 50k. If someone earns more than 50k, then the predicted value is *TRUE*.

```
# proposed probabilities given estimates from logistic model
glm_hat <- predict(glmfit, data = train_set, type = "response")


# As threshold whether someone earns over 50k, a probability of 0.5 is assumed
table(ActualValue=train_set$income, PredictedValue = glm_hat >0.5)
```

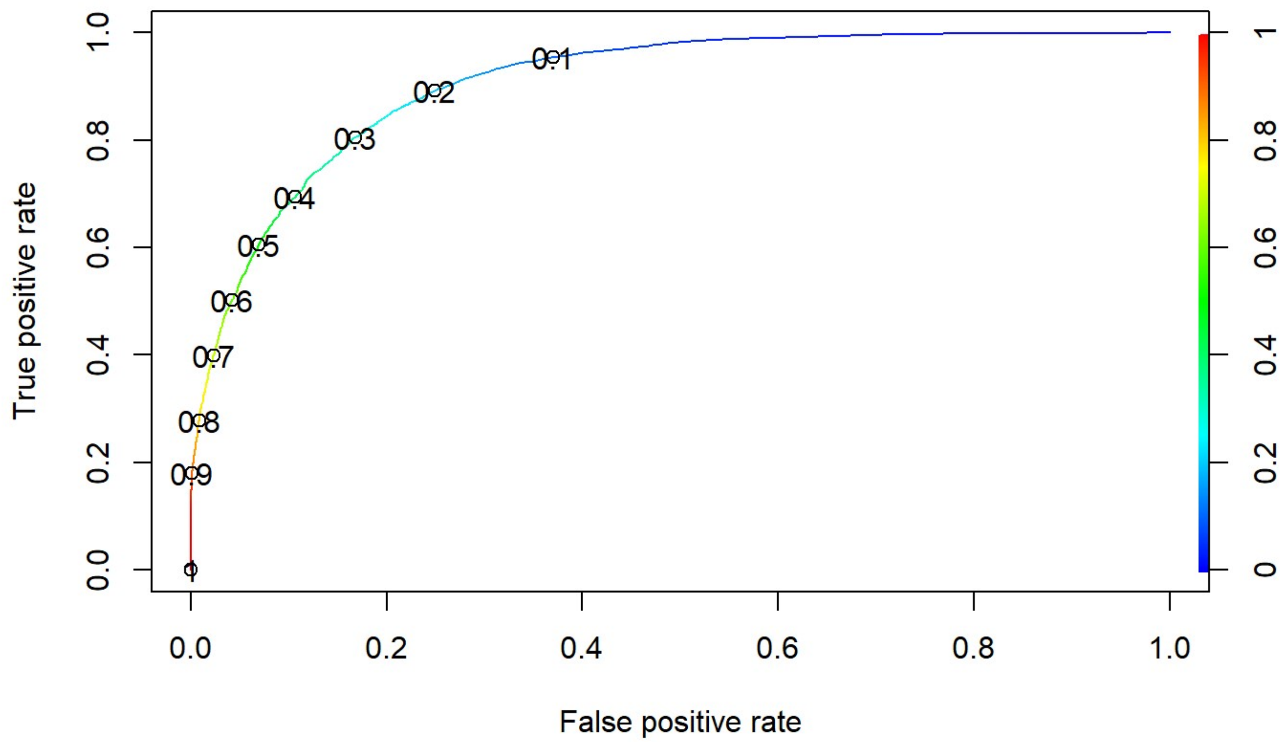
```
##            PredictedValue
## ActualValue FALSE  TRUE
##       <=50K 16866  1257
##       >50K   2371  3635
```

The accuracy is $(16866 + 3635)/(16866 + 1257 + 2371 + 3635) = 0.8496415$ .

Could the accuracy be improved by altering the threshold number?

That is to find the threshold number, that increases the number that the model predicts correctly. Or put differently the numbers in the confusion matrix along the down diagonal, i.e. (16847 and 3667) from left to right should be increased (true positive). Unfortunately there is a cost, the numbers of the false predictions increases as well (false positive rate). There is a trade off the be considered:

```
ROCRPred = prediction(glm_hat, train_set$income)
ROCRPerf <- performance(ROCRPred, "tpr","fpr")
plot(ROCRPerf, colorize=TRUE  , print.cutoffs.at=seq(0.1, by=0.1))
```




The chart plots the true positive rate against the false positive rate.

As we move along the curve, we decrease the threshold value *p*. The true positive rate increases sharply at the beginning, but the incremental gain diminishes more and more and the “cost” in form of the false positive increases sharply.

The optimum should be the one threshold value *p*, where one unit gain in true positive rate equals one unit in false positive rate lost.

From the chart above, a threshold value of 0.3 looks reasonable. The resulting confusion matrix looks like this:

```
# From the ROCR plot, the best ration is around 0.3

table(ActualValue=train_set$income, PredictedValue = glm_hat >0.3)
```

```
##             PredictedValue
## ActualValue FALSE  TRUE
##       <=50K 15082  3041
##       >50K   1170  4836
```

The accuracy is given by $(15082 + 4836)/(15082 + 3041 + 1170 + 4836) = 0.8254797$ .

The accuracy decreases with a *p* value of *0.30*. Although the accuracy is less than before, the threshold of 0.30 has a more balanced consideration of the true positive rate and false positive rate.

(Reference: link (https://www.datacamp.com/community/tutorials/confusion-matrix-calculation-r))

# 3.7. Validation

Using the dataset *test_set* for validation, the result looks as follows:

```
#  validation with test dataset

glm_hat_test <- predict(glmfit, newdata = test_set, type = "response")

table(ActualValue=test_set$income, PredictedValue = glm_hat_test >0.3)
```

```
##            PredictedValue
## ActualValue FALSE TRUE
##       <=50K  3764  767
##       >50K    277 1225
```

The accuracy of the model using the *test_set* dataset results $(3764 + 1225)/(3764 + 767 + 277 + 1225) = 0.8269518$ .

# 4. Summary and Conclusion

From the UCI machine learning repository website, the panel dataset adult was chosen. The goal was to develop a model to predict whether someone was likely to earn over 50k USD. For such a categorial question, the logistic regression was chosen. After eliminating multicollinearity issues, the best performing model is as follows:

**Dependent variable** : *income*

**Independent variables**: *age, worklclass, education_num, occupation, relationship race, sex, capital_gains, capital_loss, hours_per_week, native_region*

*omitted variables*: *fnlwgt, education, marial_status*

The features someone has which make her/him likely to earn over 50k:

The person in question is not young. If she/he works for the Federal Government, her/his odds is better than other people. The more time somebody has invested in her/his education the more likely this person will earn over 50k. If a person works in areas of Exec-managerial, Prof-specialty, Protective-service, Sales, or Tech-support, her/his chances are high to have over 50k.

As a husband the odds are higher than most people, except if someone is a working wife. In this case her chances are even higher to earn over 50k. If someone is not from any American-Indian-Eskimo minorities, his likelihood increases also. The person to earn over 50k is likely a male. If somebody has any capital gains or losses, they also indicate a high probability.

If she/he works part-time, she/he is less likely to earn more than other people. And if someone is born from Central America, she/he is less likely to make 50k compared to others who are born in Western Europe or in the US.

The predicted value from the logistic regression is a probability of earning over 50k. For decision purpose, a threshold for the probability of $p=0.3$ was chosen under considering the balance of true and false positive rates.

The accuracy of the model using the train dataset results in *0.8254797*.

The accuracy of the model using the test dataset for validation results in *0.8269518*.