Adult Dataset Prediction

Alexandru Damian, Nickolai Petersen

12/1/2021

Overview

The Adult data set was extracted in 1994 from census data of the United States. It contains continuous and nominal attributes, describing some social information (age, race, sex, marital status, ...) about the citizens registered. This dataset is split into 2 parts: adult.data for training and adult.test for testing the model performance.

The task is to predict whether the citizen's income exceeds fifty thousand dollars a year.
For this report, the focus will be on inference, drawing logical conclusions from the dataset.

The number of observations is 32561 and the number of variables is 15.

```
colnames(db.adult) <- c("age", "workclass", "fnlwgt",
                        "education", "education_num",
                        "marital_status", "occupation",
                        "relationship", "race", "sex",
                        "capital_gain", "capital_loss",
                        "hours_per_week", "native_country", "income")
```

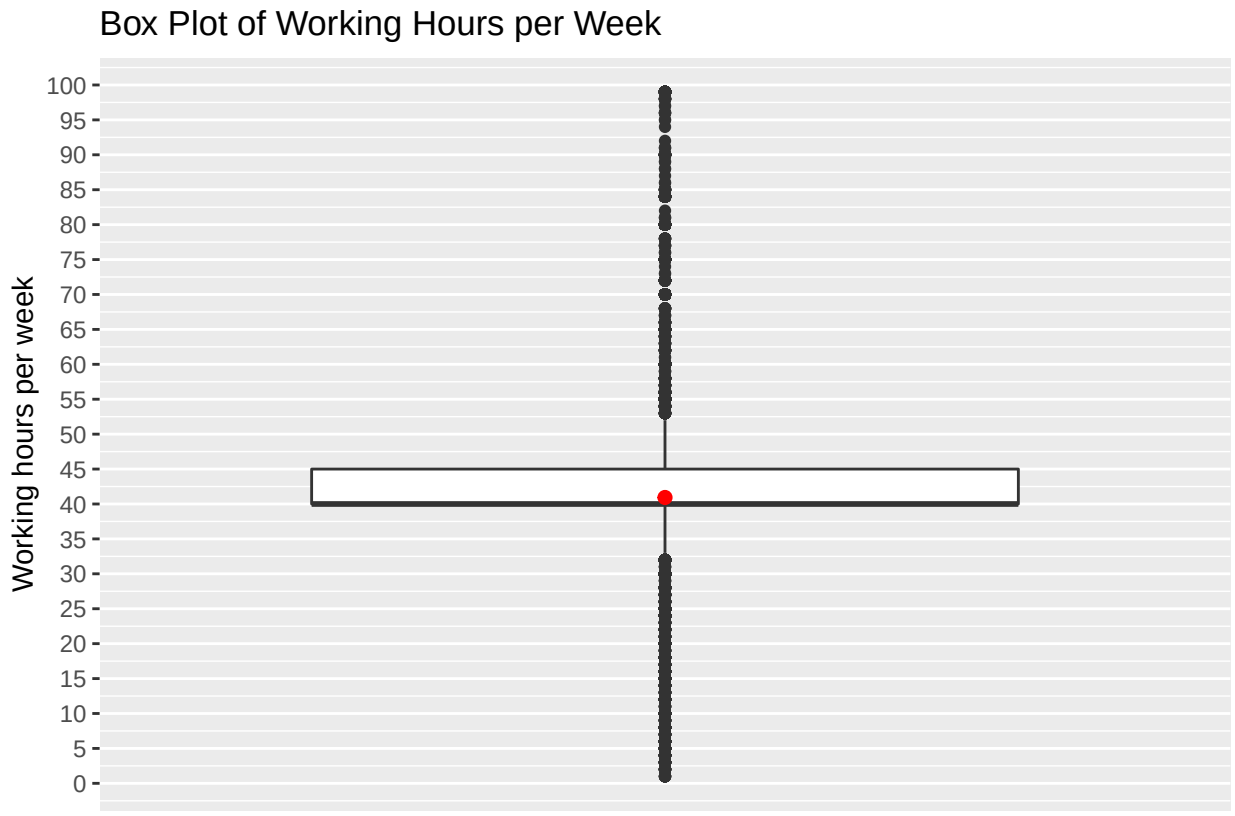
```
##   age         workclass fnlwgt  education education_num      marital_status
## 1  39         State-gov  77516  Bachelors            13       Never-married
## 2  50  Self-emp-not-inc  83311  Bachelors            13  Married-civ-spouse
## 3  38           Private 215646    HS-grad             9            Divorced
##           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
##   hours_per_week native_country income
## 1             40  United-States  <=50K
## 2             13  United-States  <=50K
## 3             40  United-States  <=50K
```

Data Transformations

1. hours_per_week

The mean number of working hours per week is 41 and around 50% of the people that responded the survey work between 40 and 35 hours per week.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   40.00   40.00   40.93   45.00   99.00
```



The width of the boxplot is equal to the Interquartile range. It is also noticeable that there are many outliers in the data represented by the black dots.
In order to work with the data to make predictions, the working hours will be grouped in 5 categories:
1. less than 40 hours per week
2. between 40 and 45 hours per week
3. between 45 and 60 hours per week
4. between 60 and 80 hours per week
5. more than 80 hours per week

```
db.adult$hours_w <- factor(db.adult$hours_w,
                           ordered = FALSE,
                           levels = c(" less_than_40",
                                      " between_40_and_45",
                                      " between_45_and_60",
                                      " between_60_and_80",
                                      " more_than_80"))
```

Percentages

24% of people work less than 40 hours/week
54% work between 40 and 45 hours/week
3% work between 60 and 80 hours/week
0.6% work more than 80 hours/week

2. native_country

There are 61 countries listed in the survey. This complicated the anlysis and might lead to overfitting. The solution is to group native_country into native_region.

```
Asia_East <- c(" Cambodia", " China", " Hong", " Laos", " Thailand",
               " Japan", " Taiwan", " Vietnam")

Asia_Central <- 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")


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

Europe_East <- c(" Poland", " Yugoslavia", " Hungary")
```

3. capital_gain and capital_loss

91% of capital_gain consists of 0
95% of capital_loss consists of 0

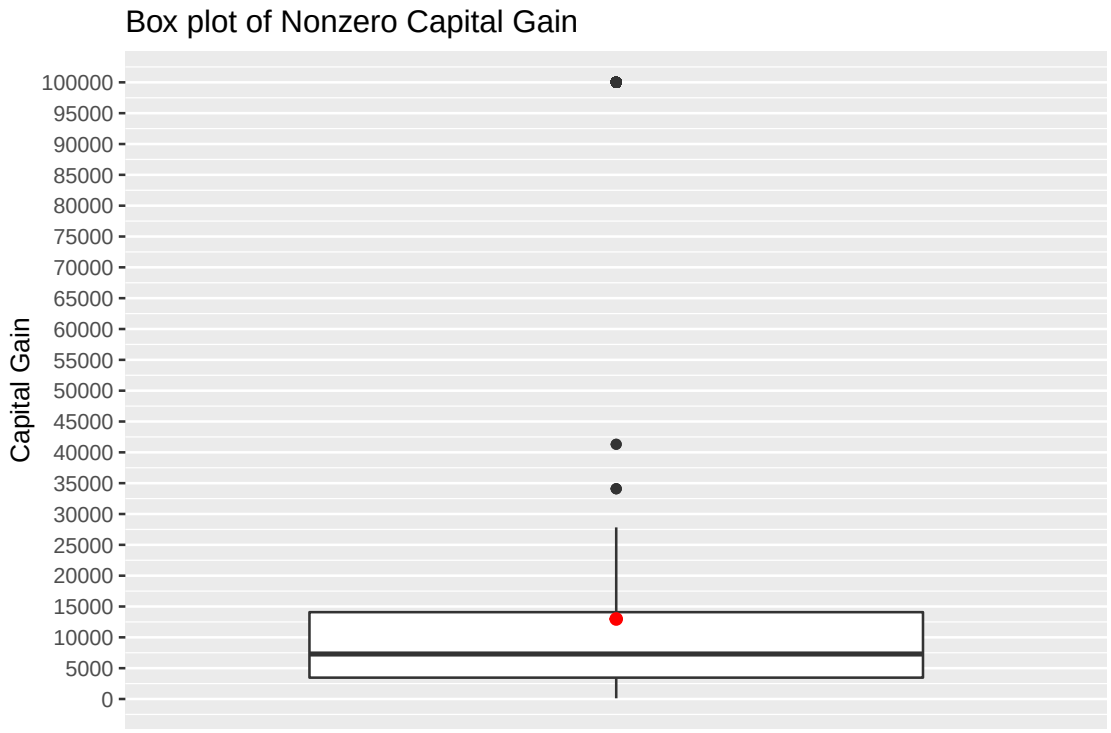
This can disrupt the analysis and as a result they need to be removed. In case they are removed, these values will be replaced with the mean.
The next table provides an overview of the resulting data, divided into quantiles. .

Table 1: Quantiles of the Nonzero Capital

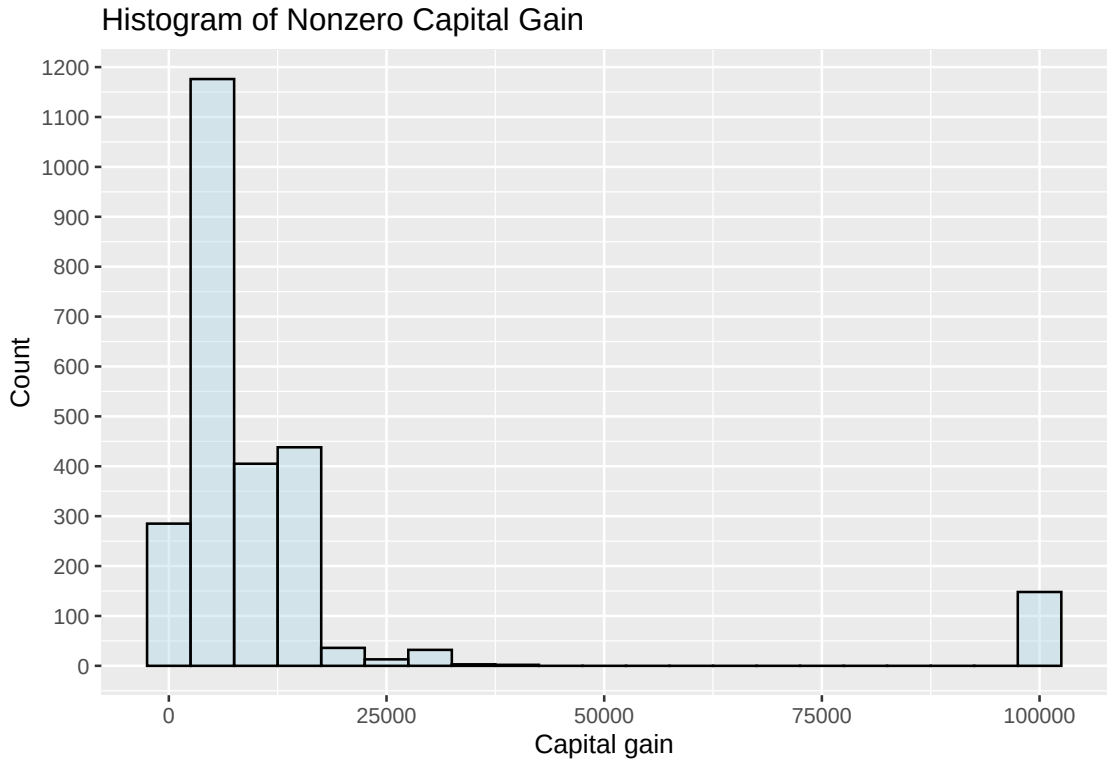
	Capital_Gain	Capital_Loss
0%	114	155
25%	3464	1672
50%	7298	1887
75%	14084	1977
100%	99999	4356

Table 2: IQR of the Nonzero Capital

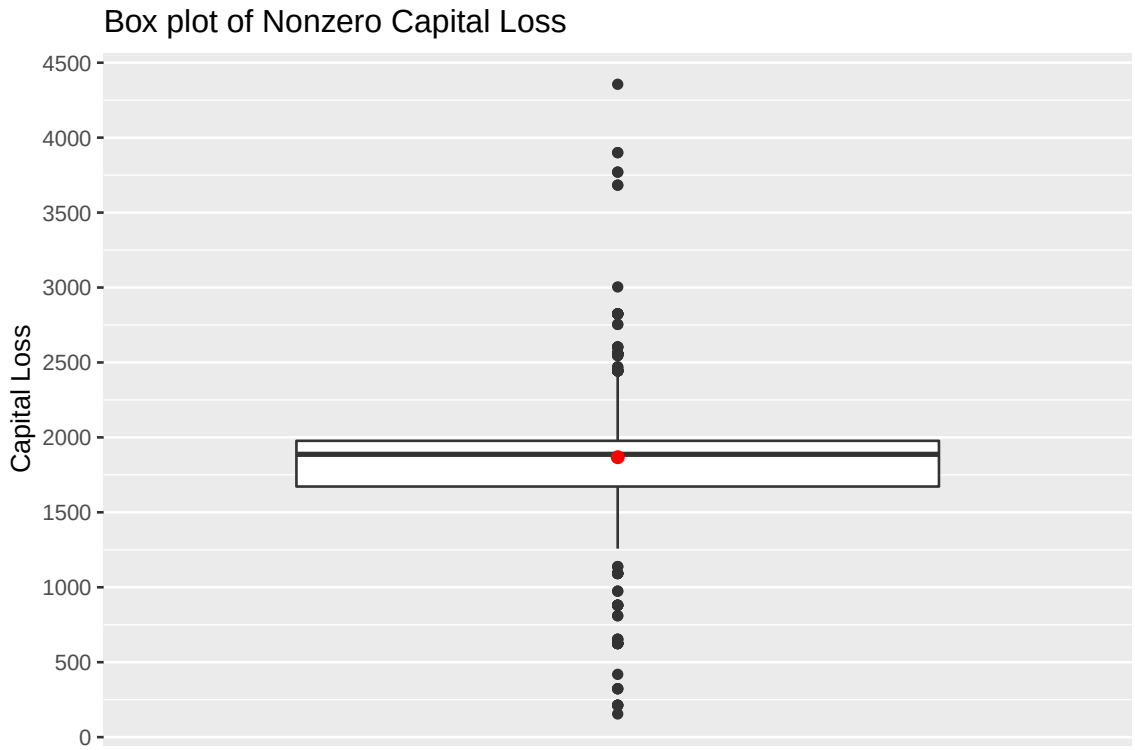
IQR_Capital_Gain	IQR_Capital_Loss
10620	305



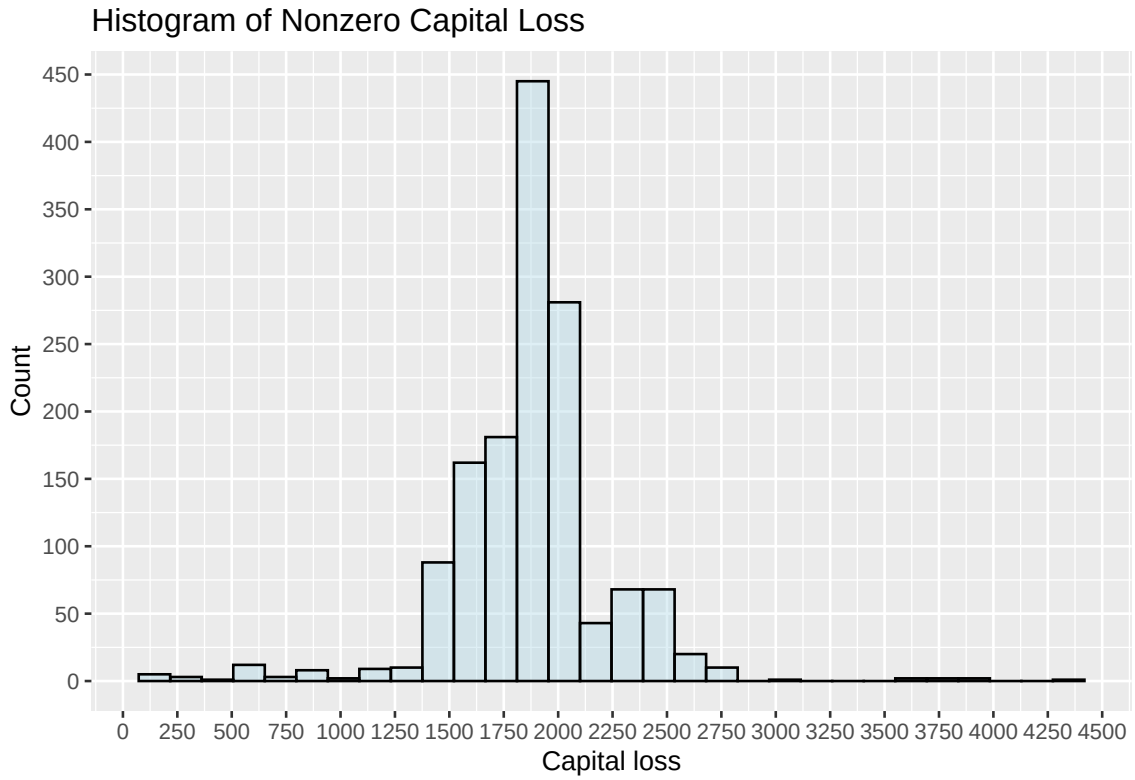
From the box plot it is noticeable that most capital gain is between \$3,000 and \$15,000.



From the histogram it is noticeable that most people with capital gain are situated between 0 and \$25,000 and the majority gains around \$5,000.



Most values lie between \$1,700 and \$2,000 and there are also many outliers.



The biggest number of people have a capital loss of approximately \$1,800.

Now when it comes to data transformation, using *Table 1: Quantiles of the Nonzero Capital* capital_loss and capital_gain will be divided based on quintiles.

```
db.adult <- mutate(db.adult,
            cap_gain = ifelse(db.adult$capital_gain < 3464, " Low",
                       ifelse(db.adult$capital_gain >= 3464 &
                              db.adult$capital_gain <= 14080, " Medium", " High")))
```

```
db.adult <- mutate(db.adult,
            cap_loss = ifelse(db.adult$capital_loss < 1672, " Low",
                       ifelse(db.adult$capital_loss >= 1672 &
                              db.adult$capital_loss <= 1977, " Medium", " High")))
```

Pre-processing the Test Dataset

It is required to apply the same steps as those for training dataset.

The number of observations is 16281.

```
colnames(db.test) <- c("age", "workclass", "fnlwgt", "education",
                       "education_num", "marital_status", "occupation",
                       "relationship", "race", "sex", "capital_gain",
                       "capital_loss", "hours_per_week",
                       "native_country", "income")

db.test <- na.omit(db.test)
row.names(db.test) <- 1:nrow(db.test)
head(db.test, 5)
```

```
##   age  workclass fnlwgt     education education_num      marital_status
## 1  25    Private 226802          11th             7       Never-married
## 2  38    Private  89814       HS-grad             9  Married-civ-spouse
## 3  28  Local-gov 336951    Assoc-acdm            12  Married-civ-spouse
## 4  44    Private 160323  Some-college            10  Married-civ-spouse
## 5  34    Private 198693          10th             6       Never-married
##           occupation   relationship   race   sex capital_gain capital_loss
## 1  Machine-op-inspct      Own-child  Black  Male            0            0
## 2    Farming-fishing        Husband  White  Male            0            0
## 3    Protective-serv        Husband  White  Male            0            0
## 4  Machine-op-inspct        Husband  Black  Male         7688            0
## 5      Other-service  Not-in-family  White  Male            0            0
##   hours_per_week native_country  income
## 1             40  United-States  <=50K.
## 2             50  United-States  <=50K.
## 3             40  United-States   >50K.
## 4             40  United-States   >50K.
## 5             30  United-States  <=50K.
```

It is noticeable that the naming convention for income is not consistent across the 2 datasets.The column will values will have to be renamed. After renaming them, the steps from the adult dataset need to be applied to the test dataset.

Exporting the processed datasets

The changed datasets are ready to be analyzed and exported.

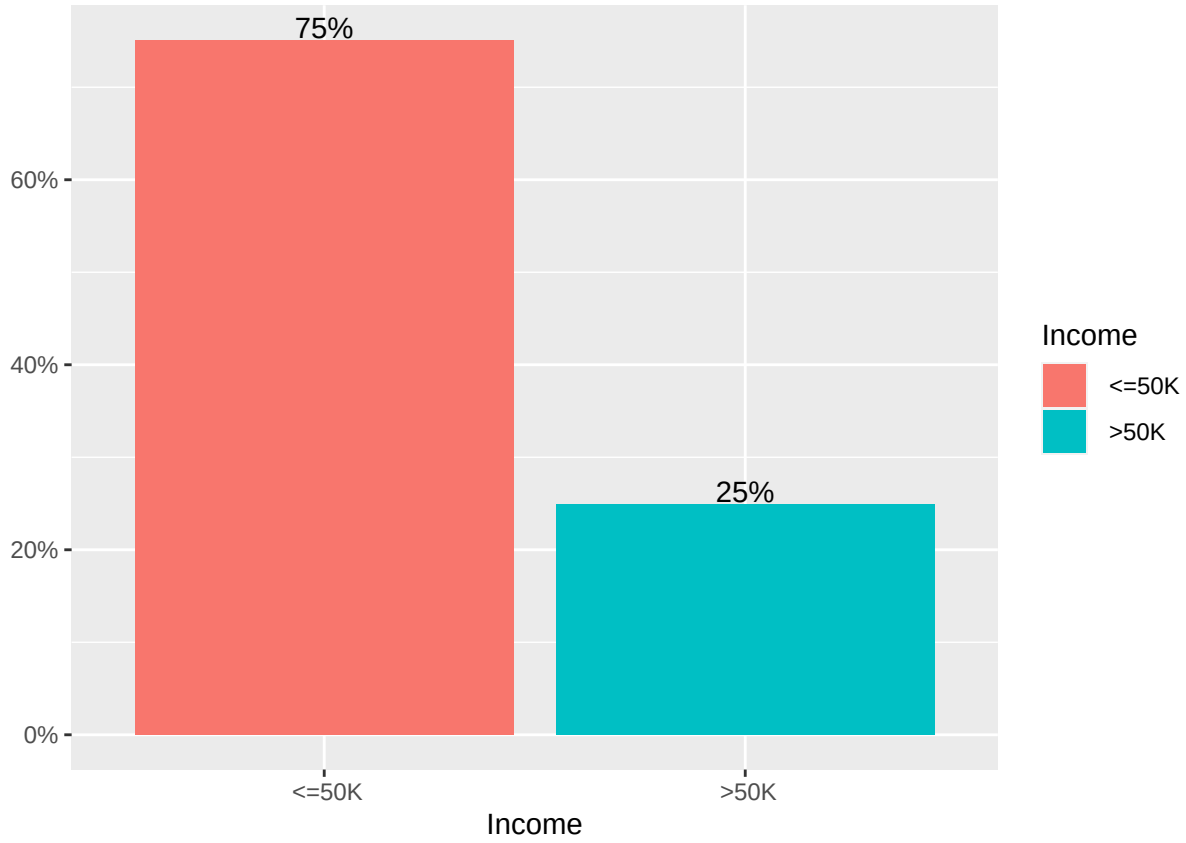
```
write.csv(db.adult, "adult_df.csv", row.names = FALSE)

write.csv(db.test, "test_df.csv", row.names = FALSE)
```

EDA - Exploratory Data Analysis

Now having pre-processed both datasets, it is possible to start analyzing the data.

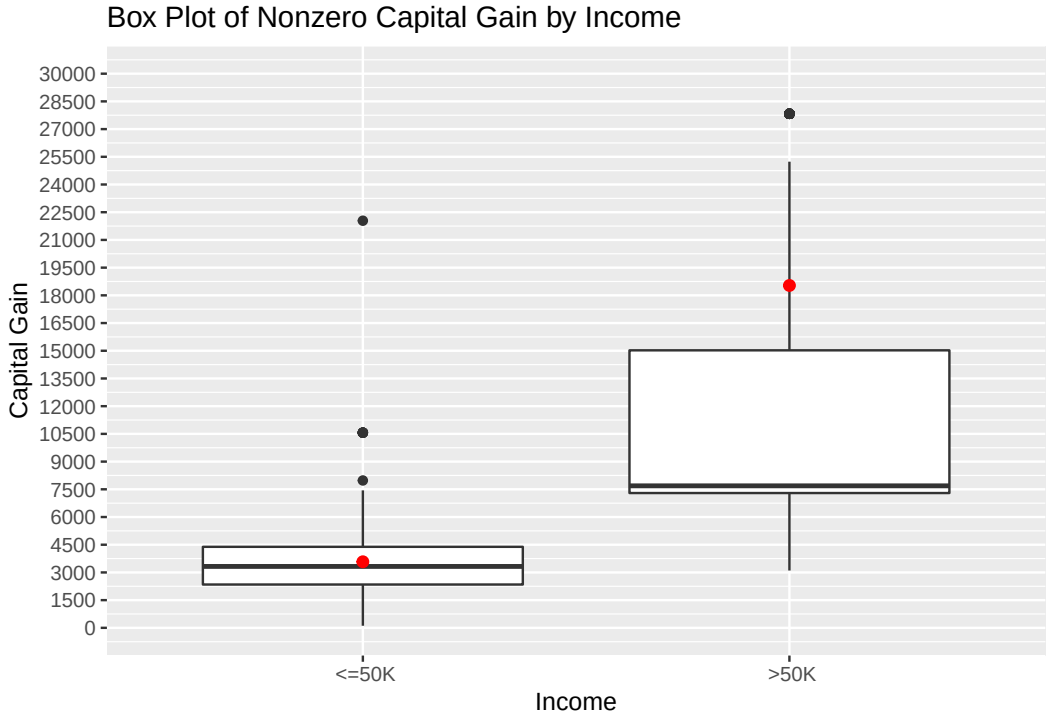
Summary statistic

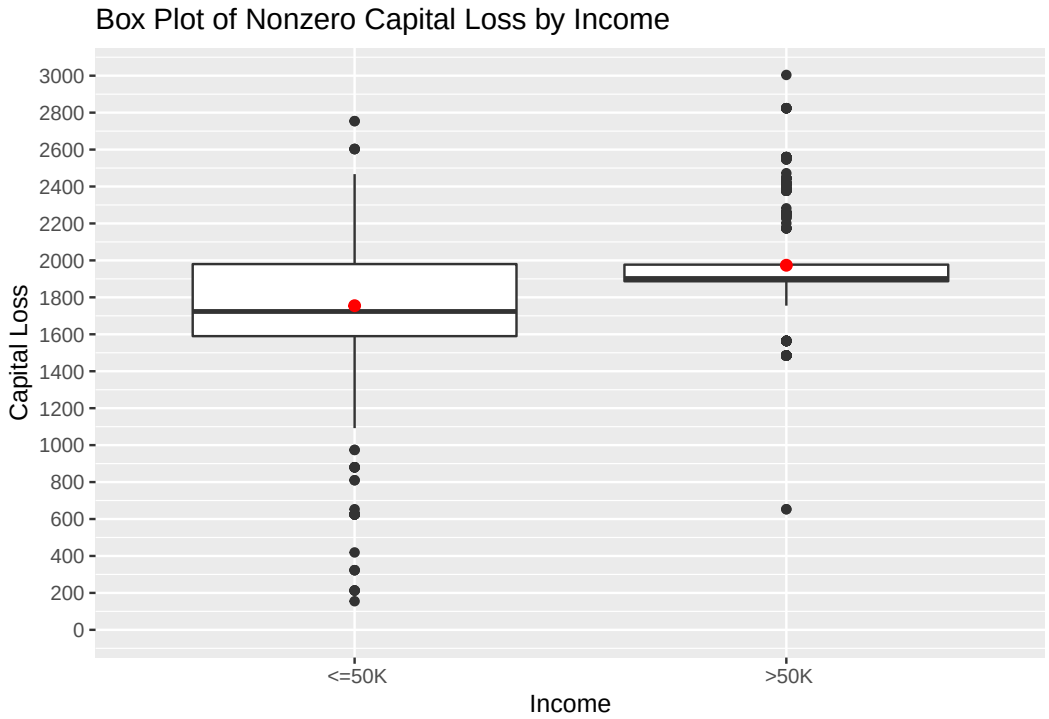
```
##  <=50K   >50K
##  22654   7508
```



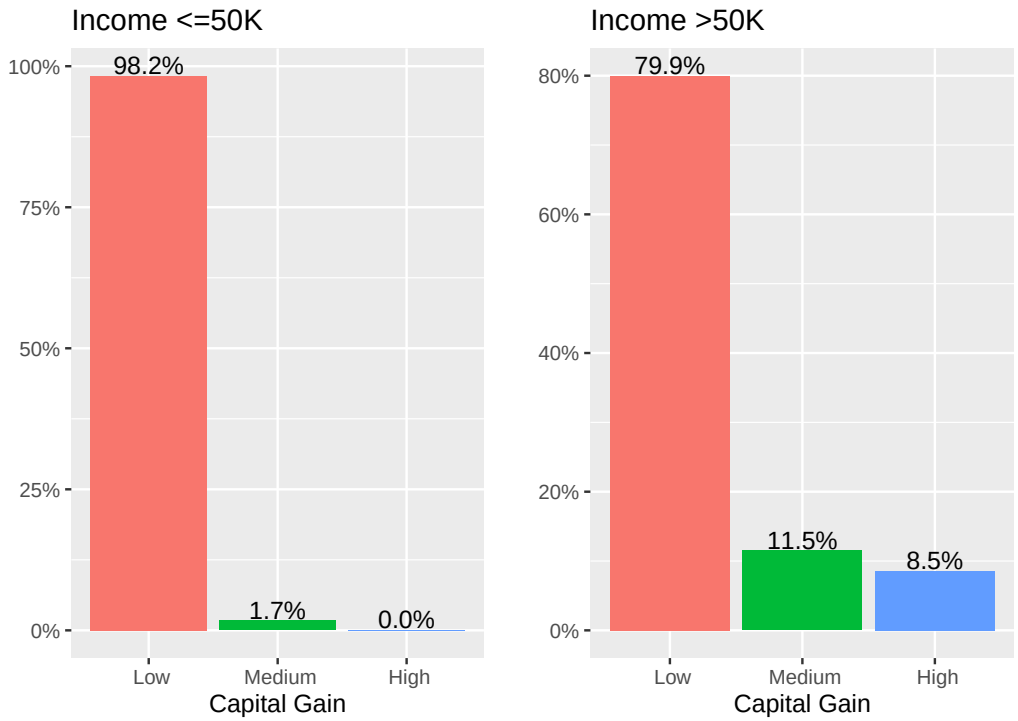
The graph above shows that 3/4 of people earn less than 50K a year and 1/4 of people earn above 50k a year.

Capital gain and capital loss

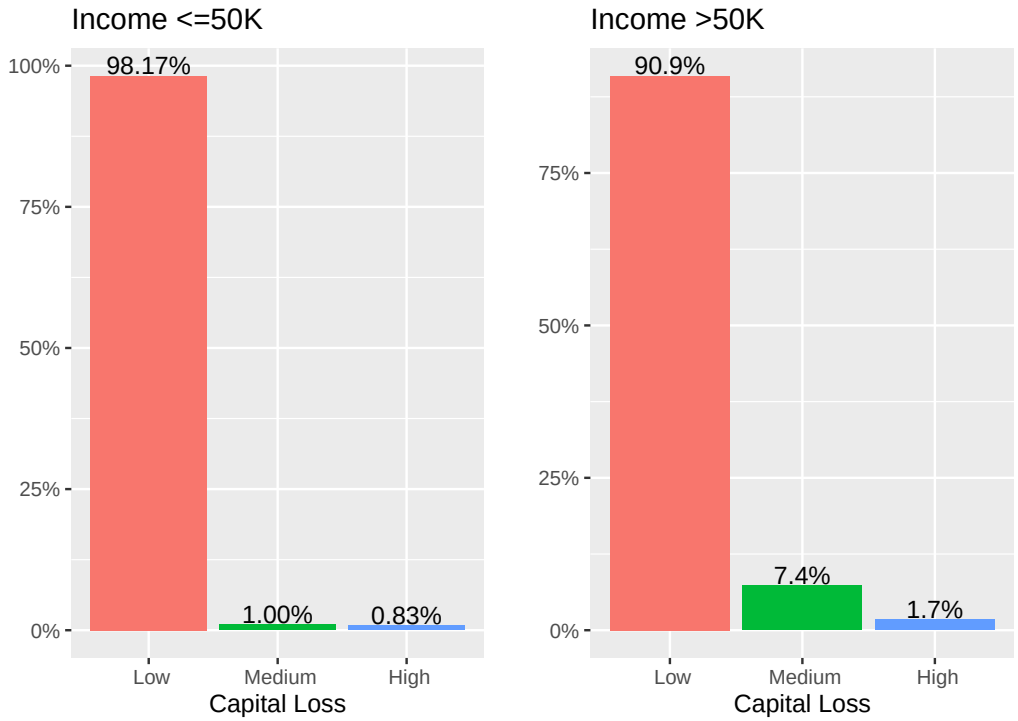




There is evidence of a strong relationship between the nonzero values of *capital_gain*, *capital_loss* and *income*. Despite the correlation, these variables will not be included in the predictive model due do the high number of 0 values that have been removed during pre-processing. The observation that 90% of the values were 0's means that less than 10% of the survey participants make any investments.



From the bar plot it is noticeable that the proportion of people with medium and high capital gain is significantly bigger within the group of people that earn more than 50K a year.



Combining the results from the above graphs, it is possible to conclude that the differences in capital gains and capital losses between the 2 groups could be explained reasoning that wealthier people tend to invest more often and into more volatile assets.

Age

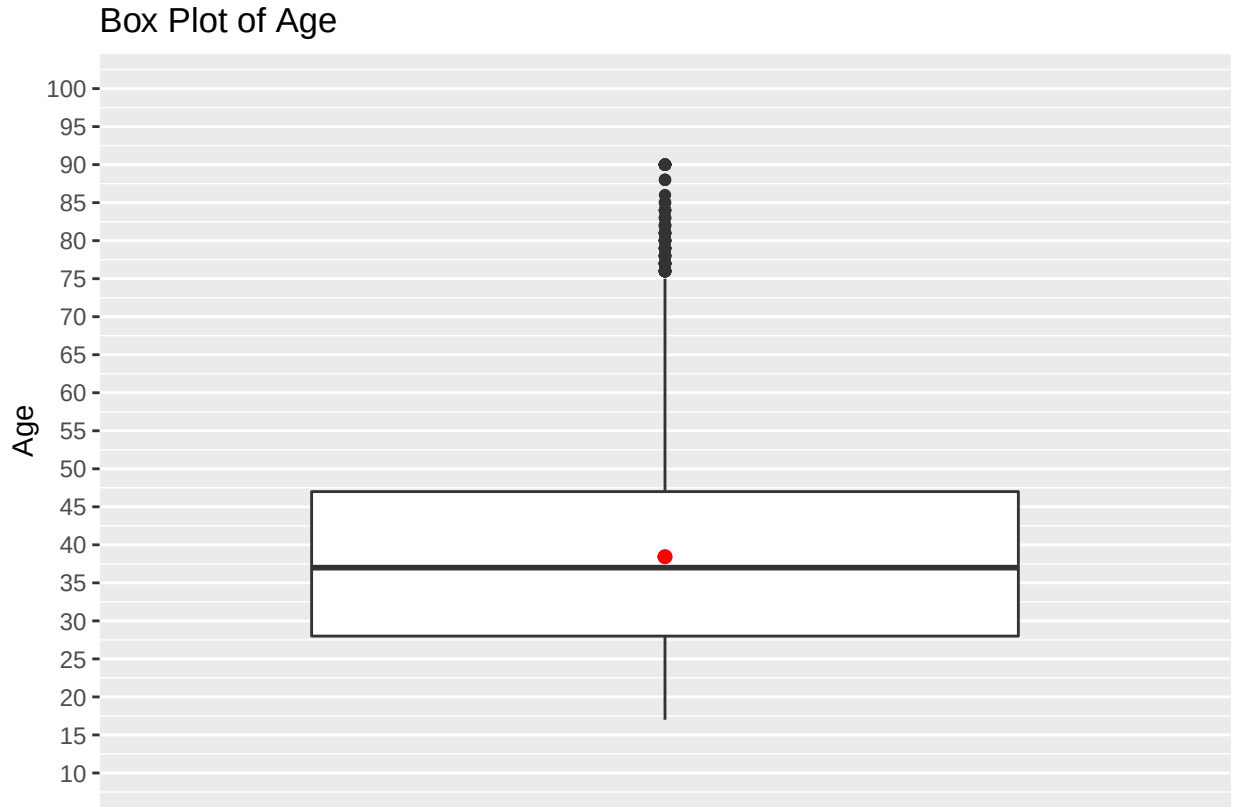
At least 50% of the people in the survey are between 28 and 47 years old. The dataset also has outliers, as there are people between 75 and 90 years old still working.
Most individuals are between 20 and 50 years old.

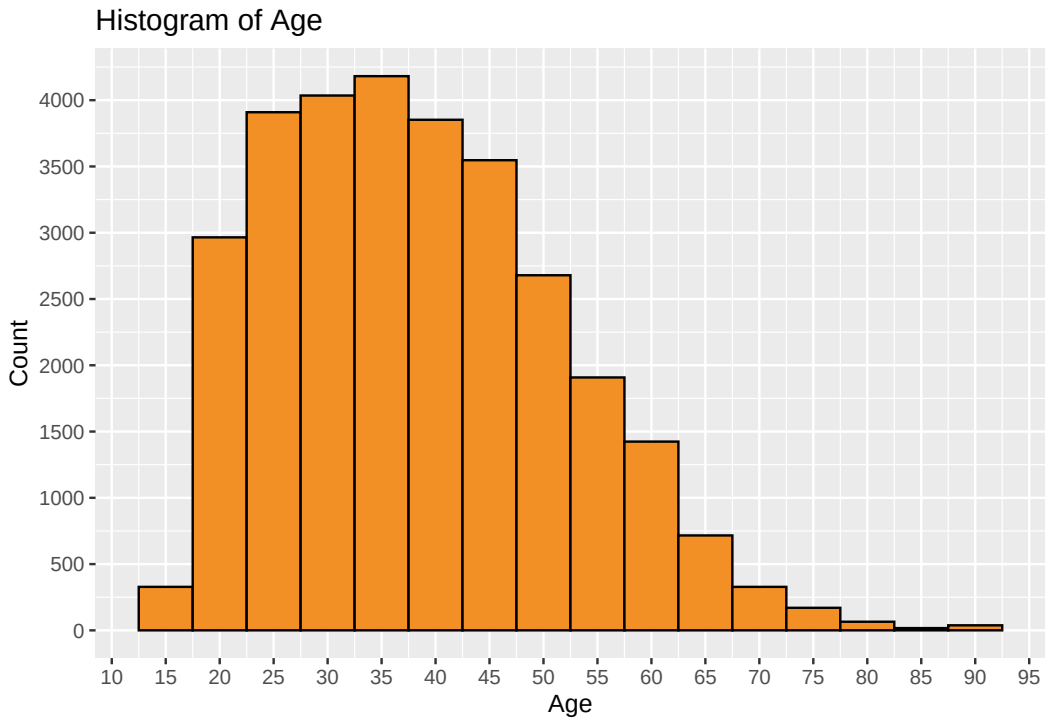
```
summary(db.adult$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   28.00   37.00   38.44   47.00   90.00
```

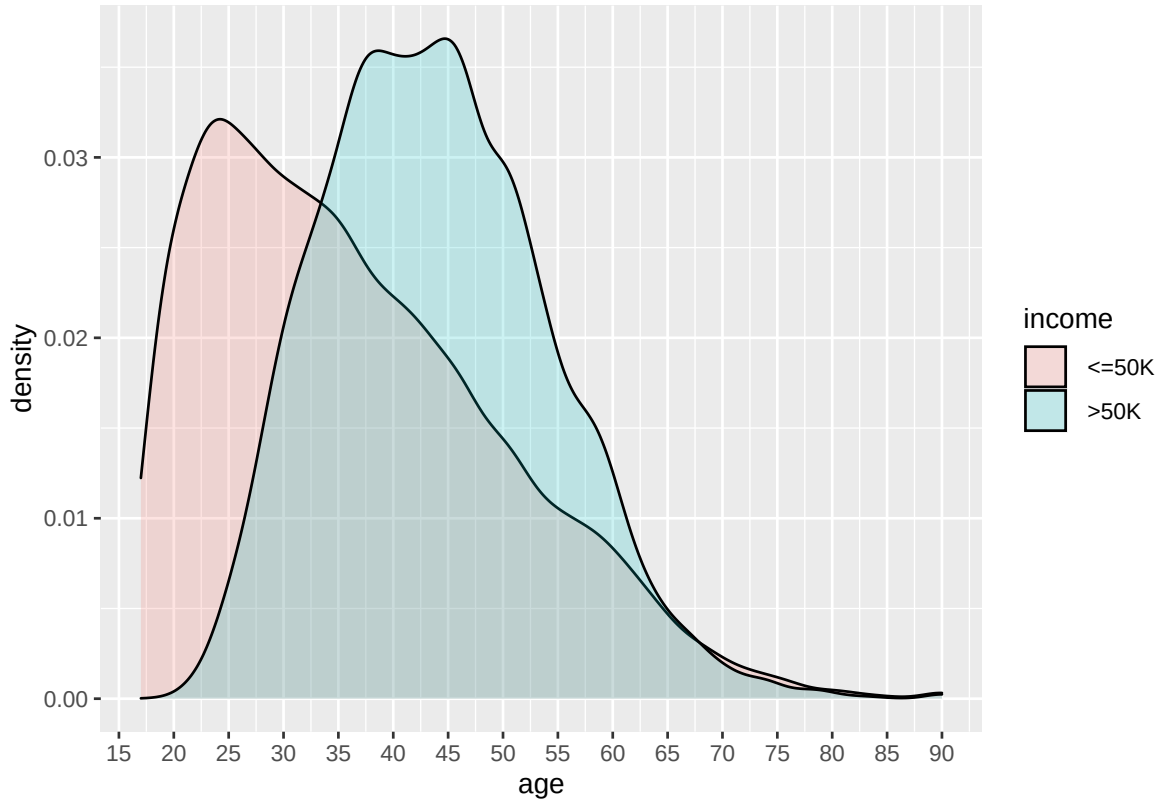
```
IQR(db.adult$age)
```

```
## [1] 19
```

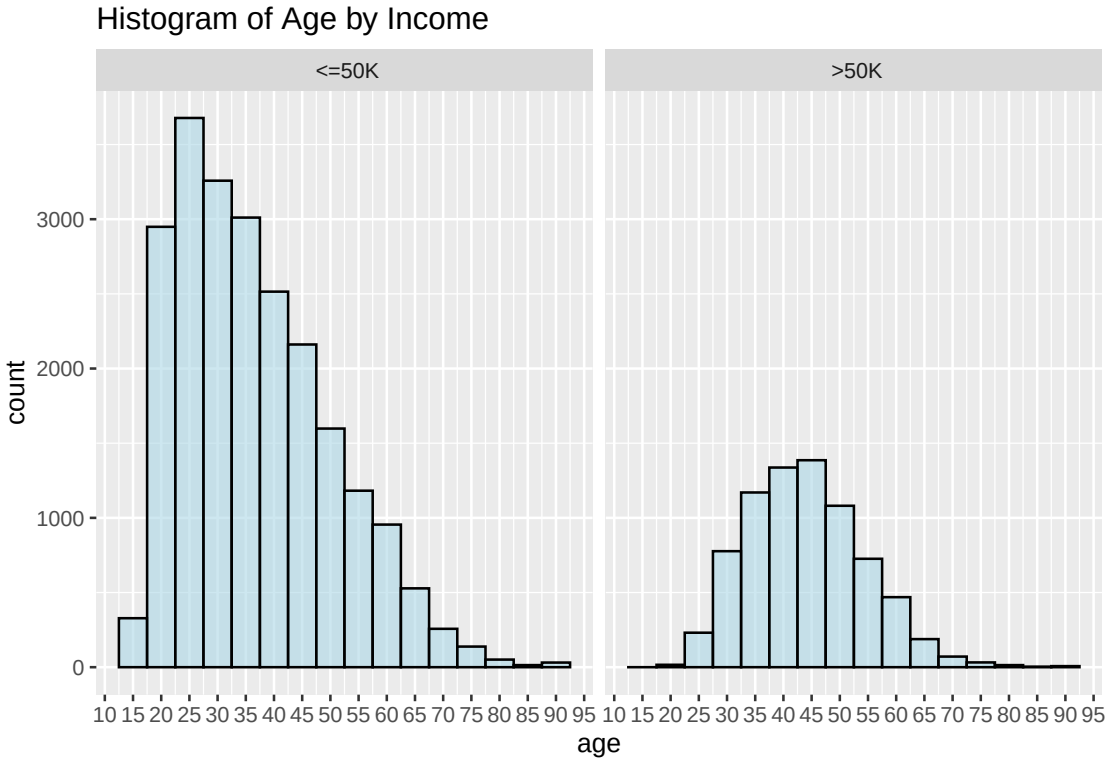


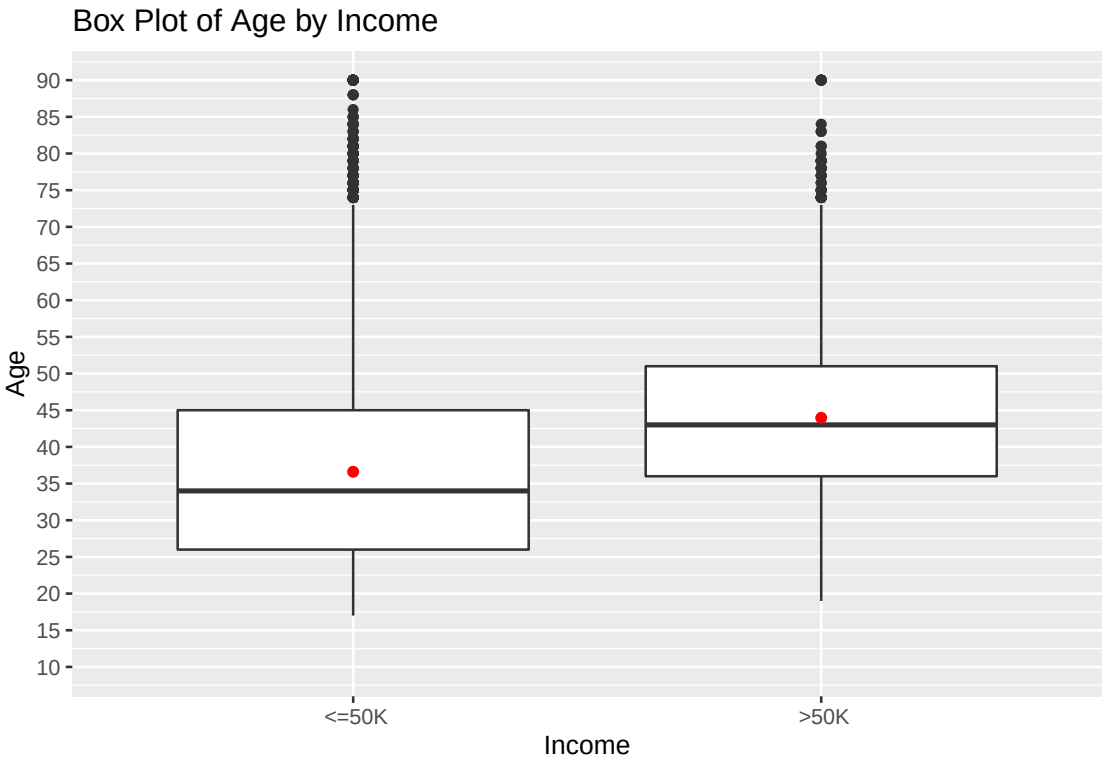


From the density graph it is noticeable that the majority of people earning more than 50K a year are between 30 and 55 years old and for the other group it is between the ages of 18 and 45.
There appears to be a strong correlation between age and income.

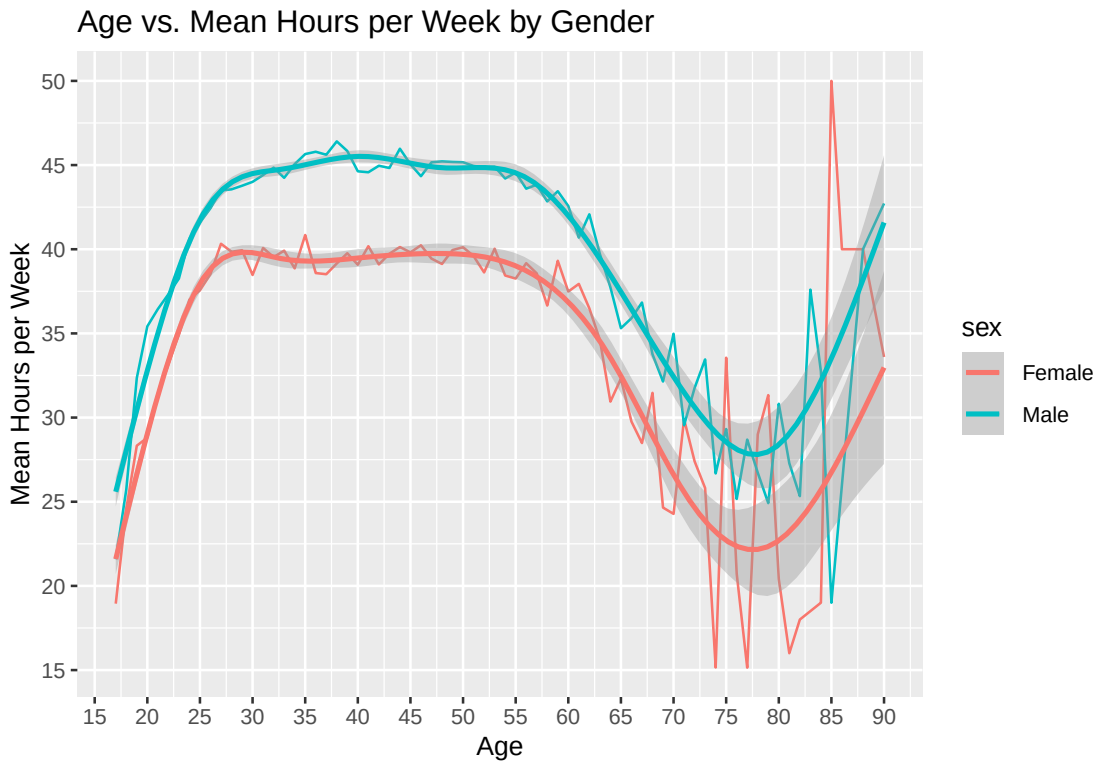


People who earn more than 50K a year are on average 43-44 years old and people who earn less than 50K are, on average, 34 to 37 years old.





Men tend to work more hours per week across all age groups. This relation changes after the age of 70 years, but that data should be neglected as it is extremely uncommon for people to work at that age.



Hours per week

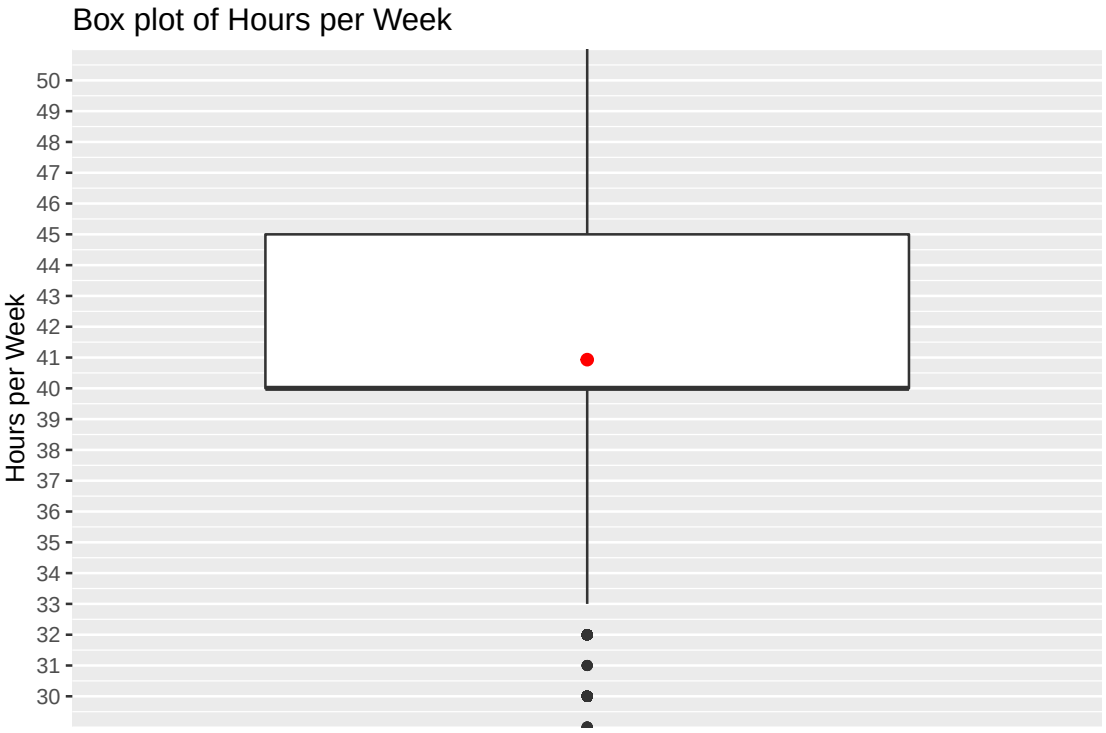
The mean number of hours per week is 41 and at least 50% of the survey participants work between 40 and 45 hours a week.

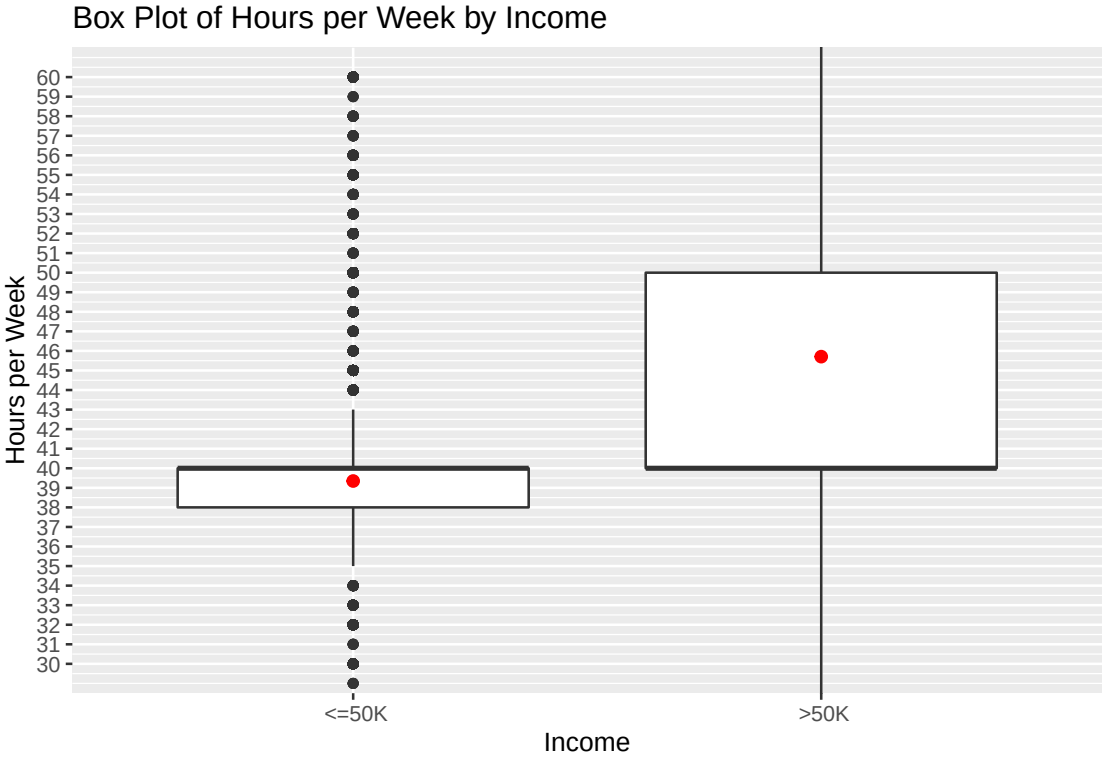
```
summary(db.adult$hours_per_week)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   40.00   40.00   40.93   45.00   99.00
```

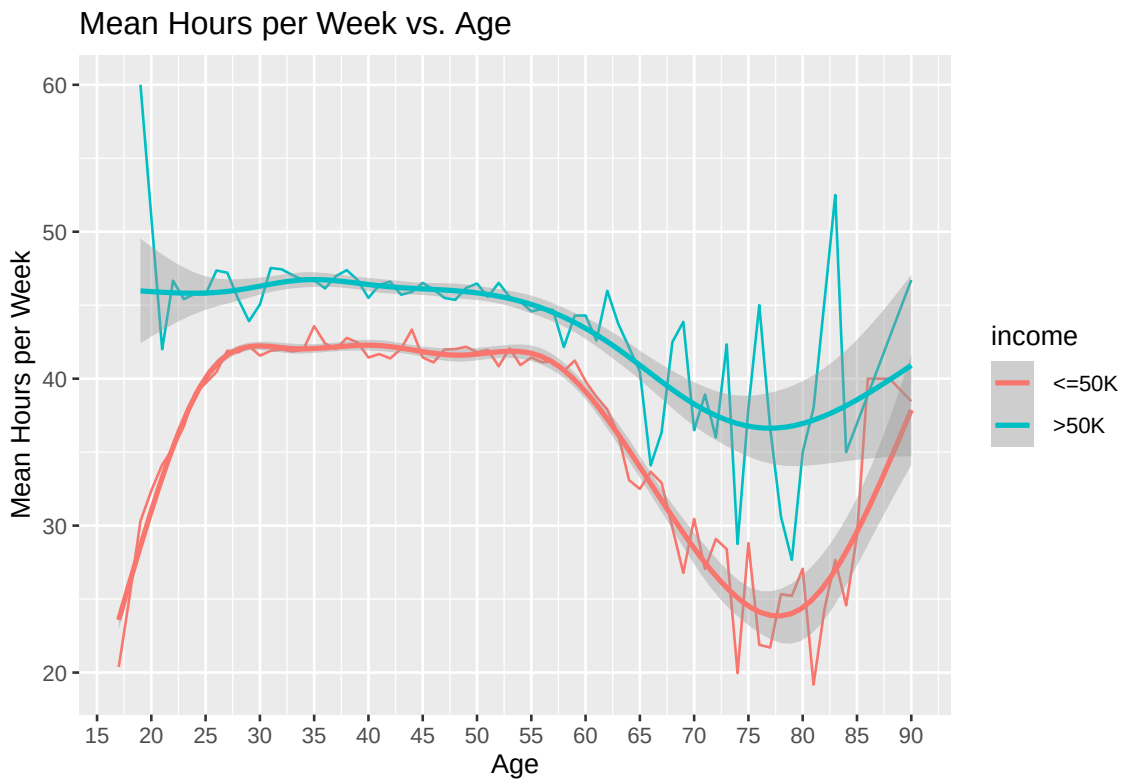
```
IQR(db.adult$hours_per_week)
```

```
## [1] 5
```

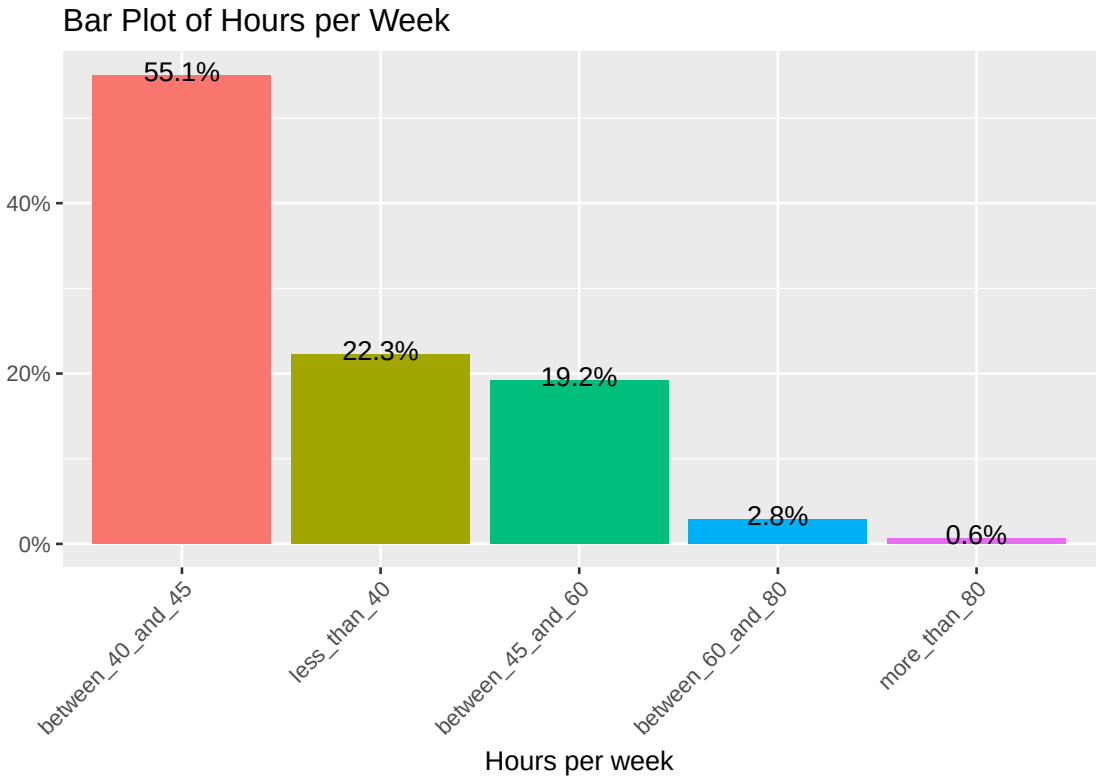




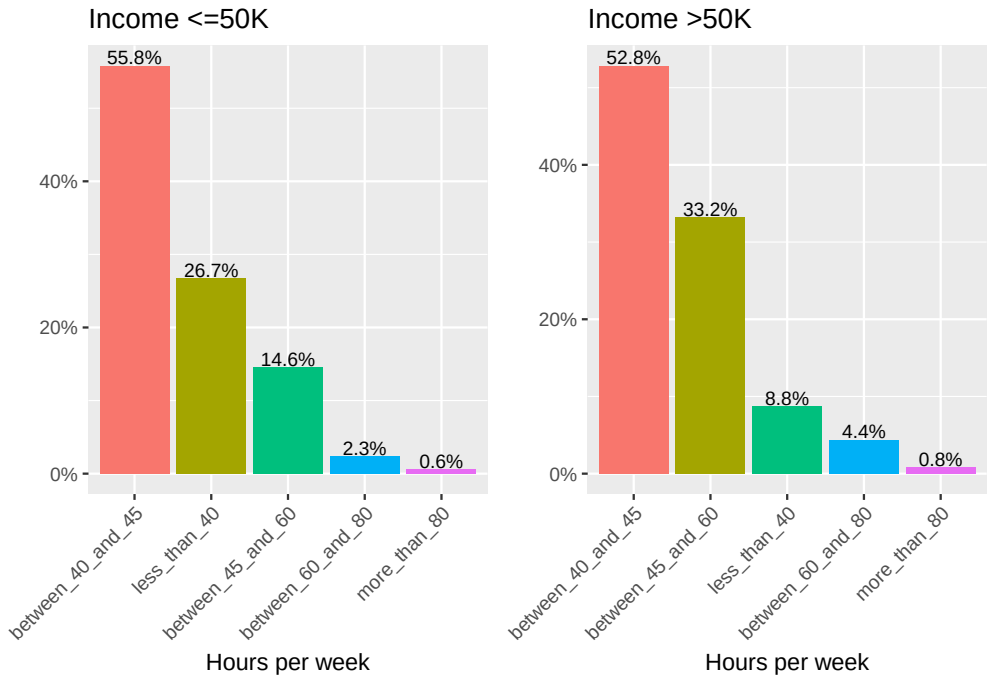
As it is observable from the box plot, the number of working hours for people that earn more than 50K is significantly higher than for people who earn less than 50K.
For all age groups, the number of working hours for the people that earn more than 50K a year is higher than for those that earn less than 50K.



Created variable *hours_w*

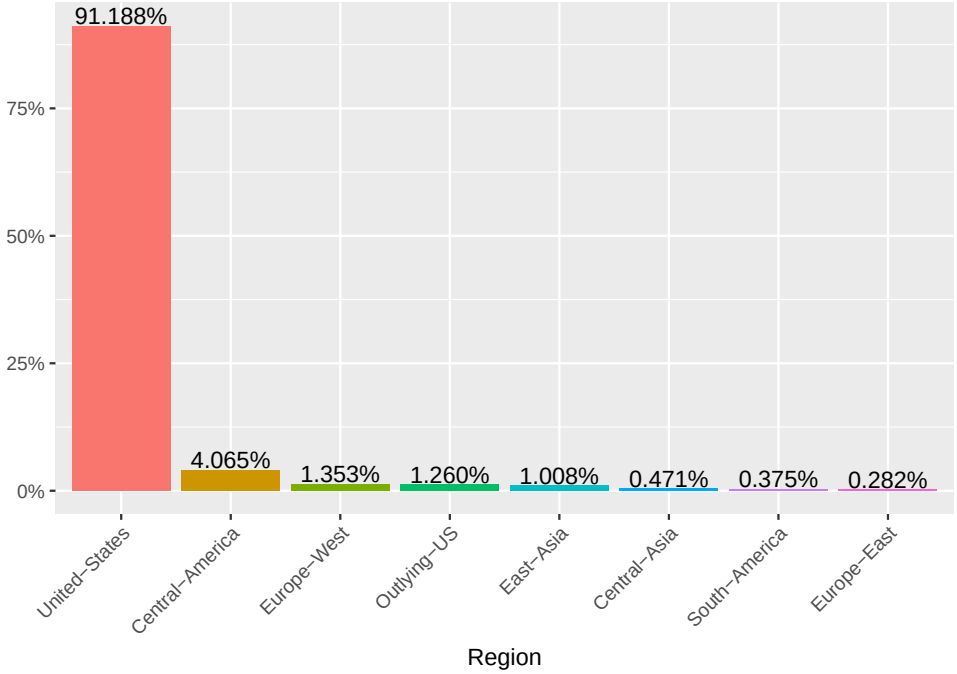


The percentage differences in hours per week between people in the two income categories are almost insignificant. The only exception is for the group between 45 and 60 hours a week, where 33.2% of people who earn more than 50K are situated compared to only 14.6% for the other income group.



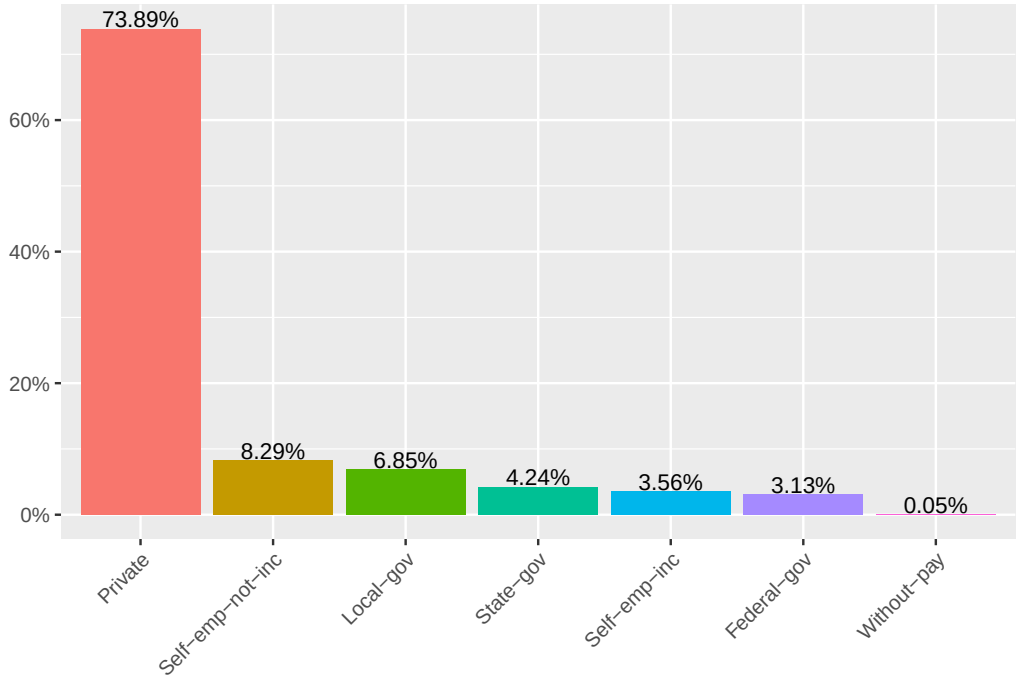
Created variable *native_region*

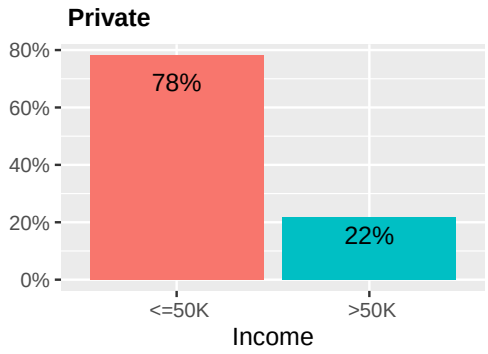
Most of the people that participated in the survey originate from America.Because the data is so limited for the other regions, conclusions made from this analysis can only be applied to America.

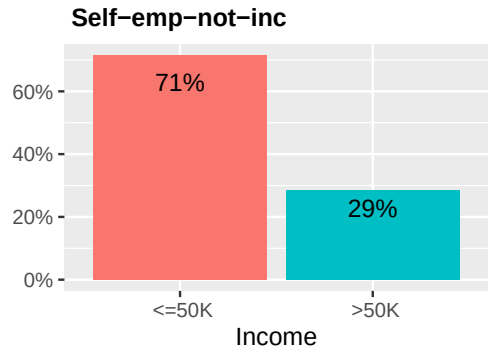


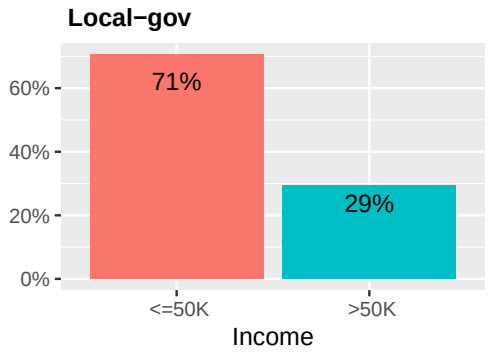
workclass

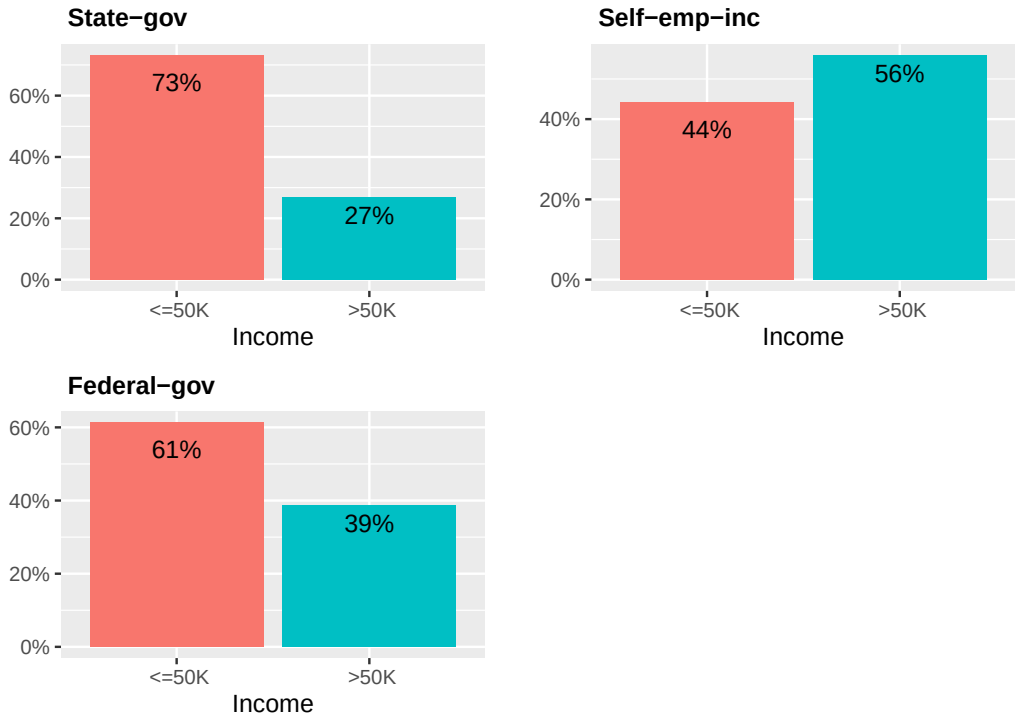
It is important to consider employment type compared to income. Without-pay will be removed from the dataset as it seems to be an outlier that will skew the results.







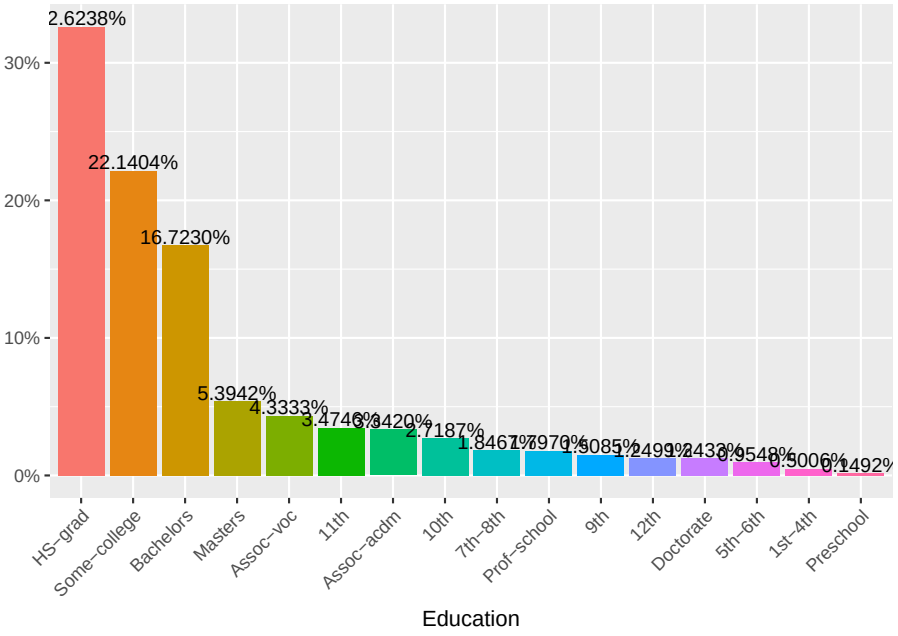




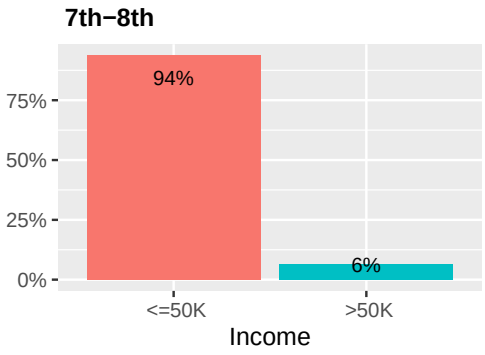
The percentage of people earning more than 50K a year is biggest for the *Self employed incorporated category*. The next category in the list for the highest percentage of people earning more than 50K is the *Federal government*. Except those first 2 categories, there aren't large differences in the distribution of the 2 income groups.

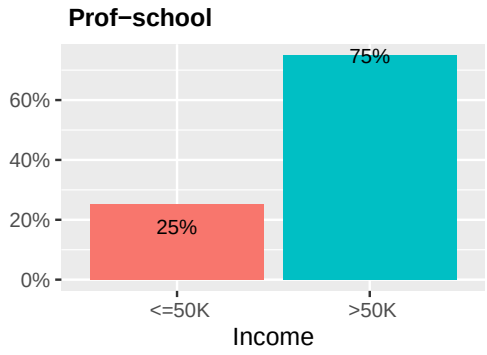
education

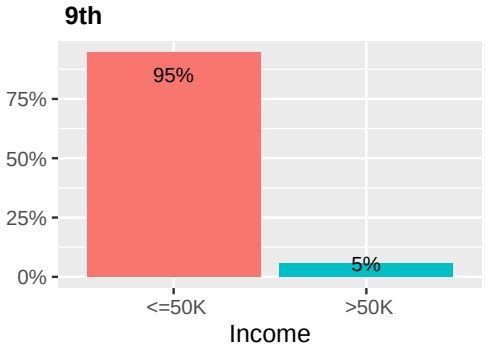
Since there are no people with Preschool level of education that earn more than 50K, this category will be removed.

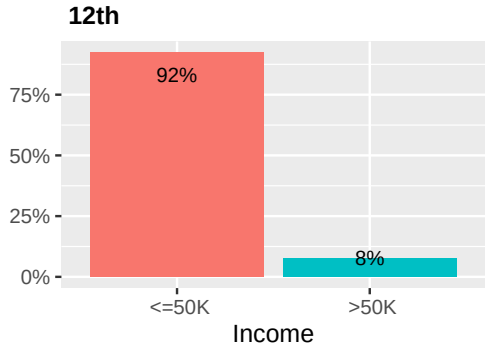


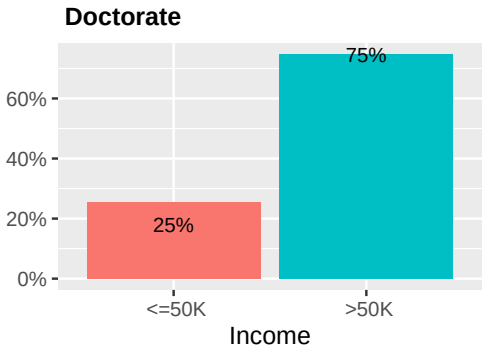
The categories spanning from 1st grade to 12th grade have a very limited percentage of people with income greater than 50K a year. For high school and college the percentages are also relatively small.The biggest percentage of people with an income over 50K belongs to the category " Prof-school" with 75%, followed by “Masters” with 56% and finally " Bachelors" with 42%. As of now, the data indicates the most important predictor of income category is the education level.

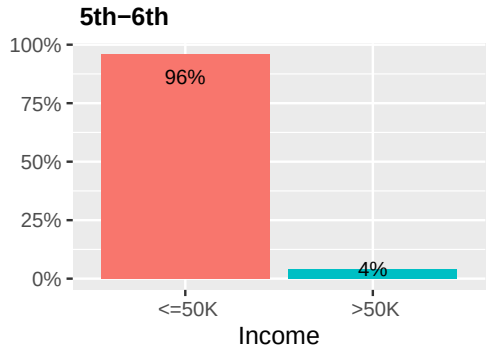

7th−8th
94%
6%
75%
50%
25%
0%
<=50K
>50K
Income

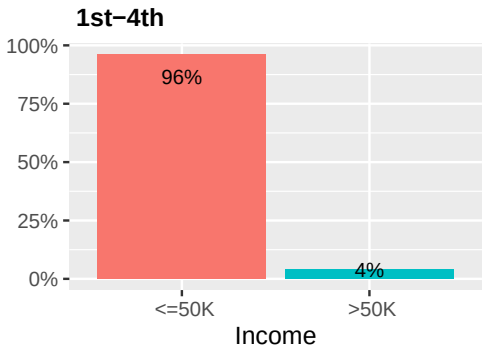

Prof−school
75%
25%
60%
40%
20%
0%
<=50K
>50K
Income


9th
95%
5%
75%
50%
25%
0%
<=50K
>50K
Income

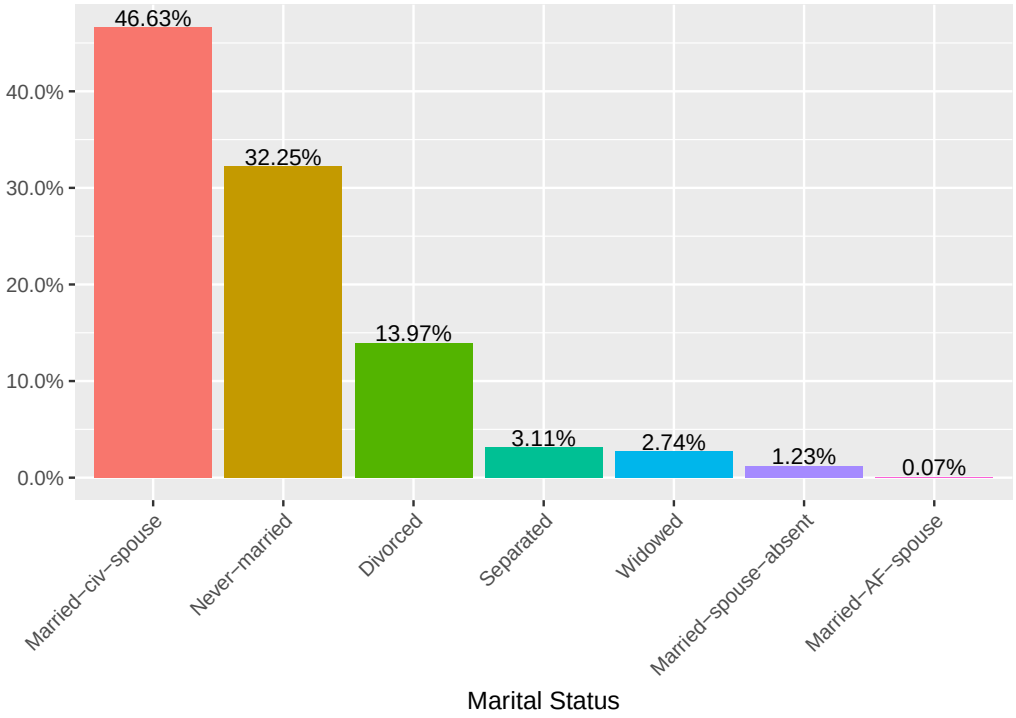

12th
92%
8%
75%
50%
25%
0%
<=50K
>50K
Income

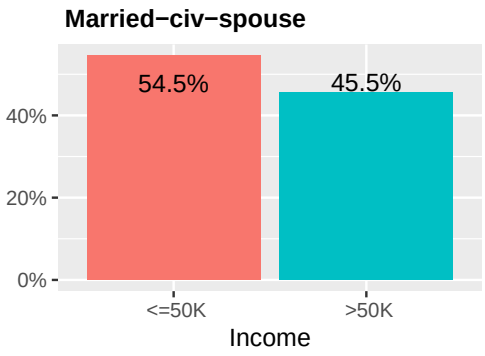

Doctorate
75%
25%
60%
40%
20%
0%
<=50K
>50K
Income

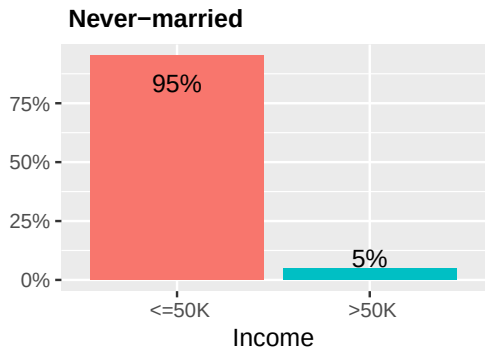

5th−6th
96%
4%
100%
75%
50%
25%
0%
<=50K
>50K
Income

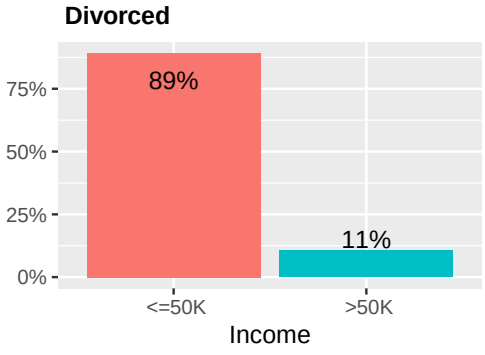

1st−4th
96%
4%
100%
75%
50%
25%
0%
<=50K
>50K
Income

marital_status









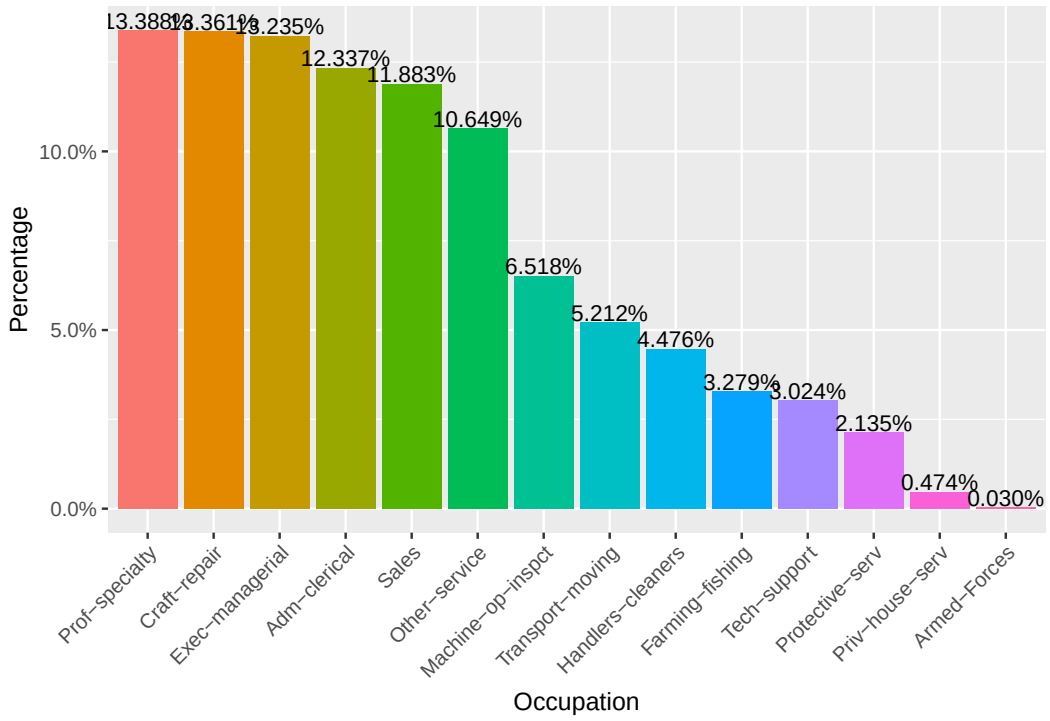
The biggest percentage of people with income higher than 50K belong to the categories *" Married-AF-spouse"* and *"Married-civ-spouse"*. Married people seem to earn more, but it's not possible to draw conclusion from this data alone. When taking into consideration the *age* variable, then it might be possible to explain the differences by pointing out that older people earn more and are more likely to be married than younger people.

occupation

When it comes to occupation, the data seems to be diversified enough to be considered a random sample.

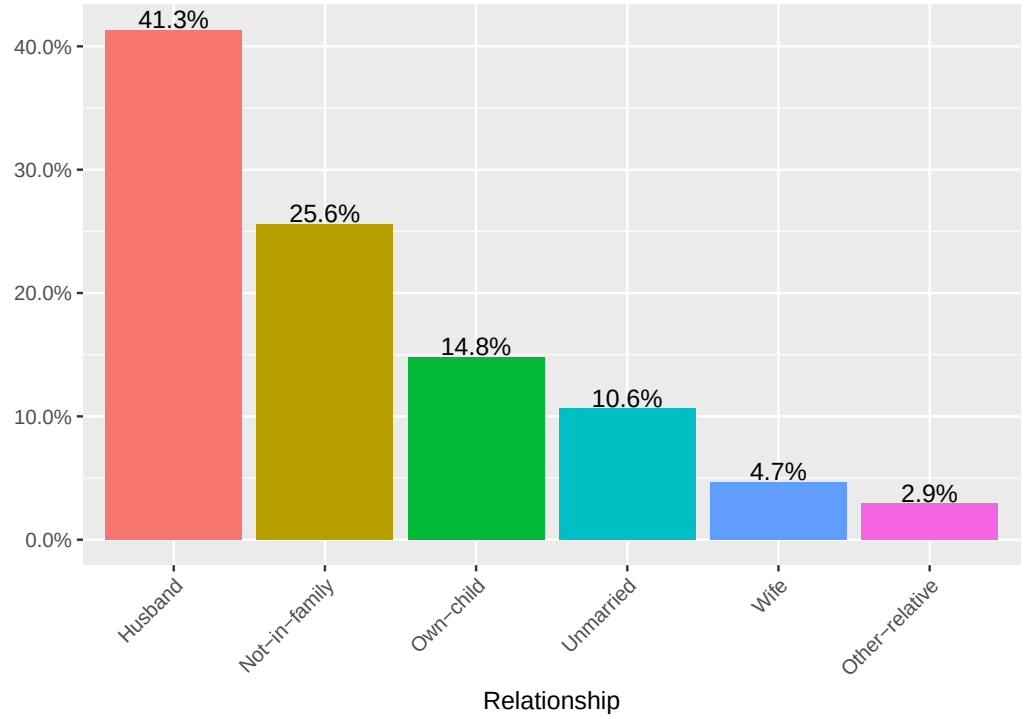
```
summary(db.adult$occupation)
```

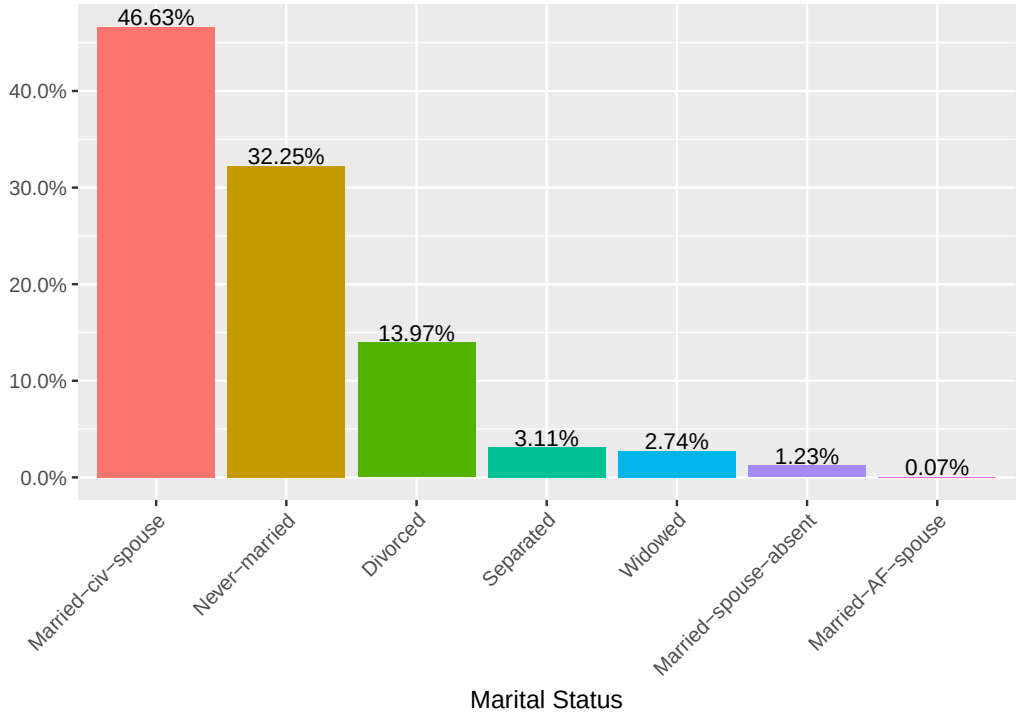
```
##       Adm-clerical       Armed-Forces       Craft-repair    Exec-managerial
##               3721                  9               4030               3992
##    Farming-fishing  Handlers-cleaners  Machine-op-inspct      Other-service
##                989               1350               1966               3212
##    Priv-house-serv     Prof-specialty    Protective-serv              Sales
##                143               4038                644               3584
##       Tech-support   Transport-moving
##                912               1572
```

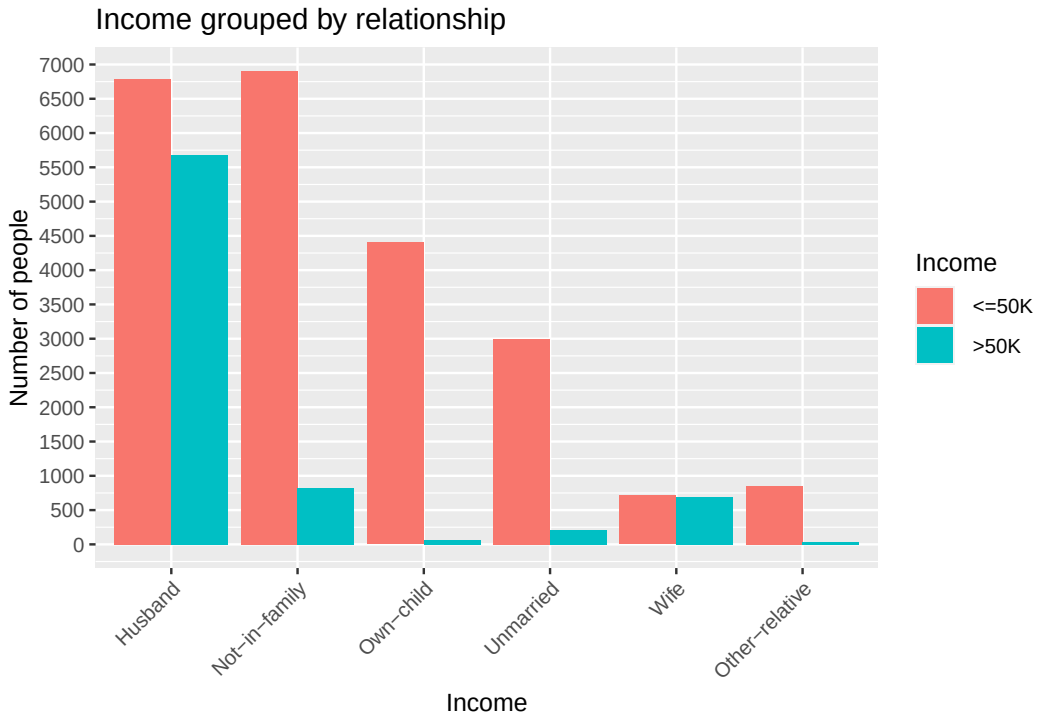


relationship

The relationship variable is closely related to *marital status* and supports the previous data.

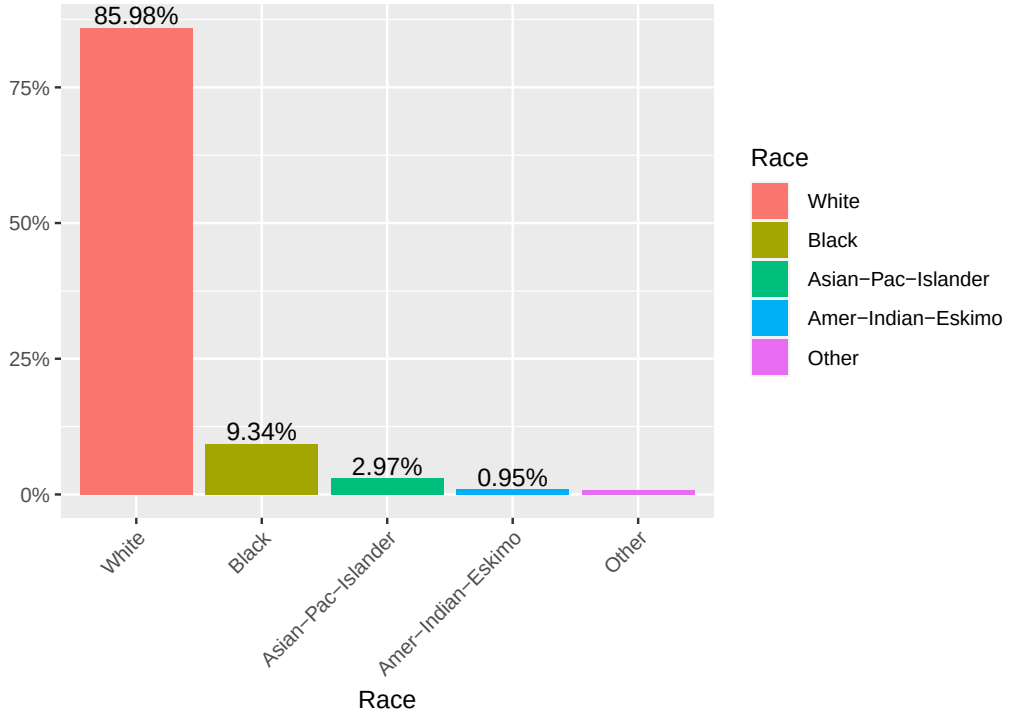


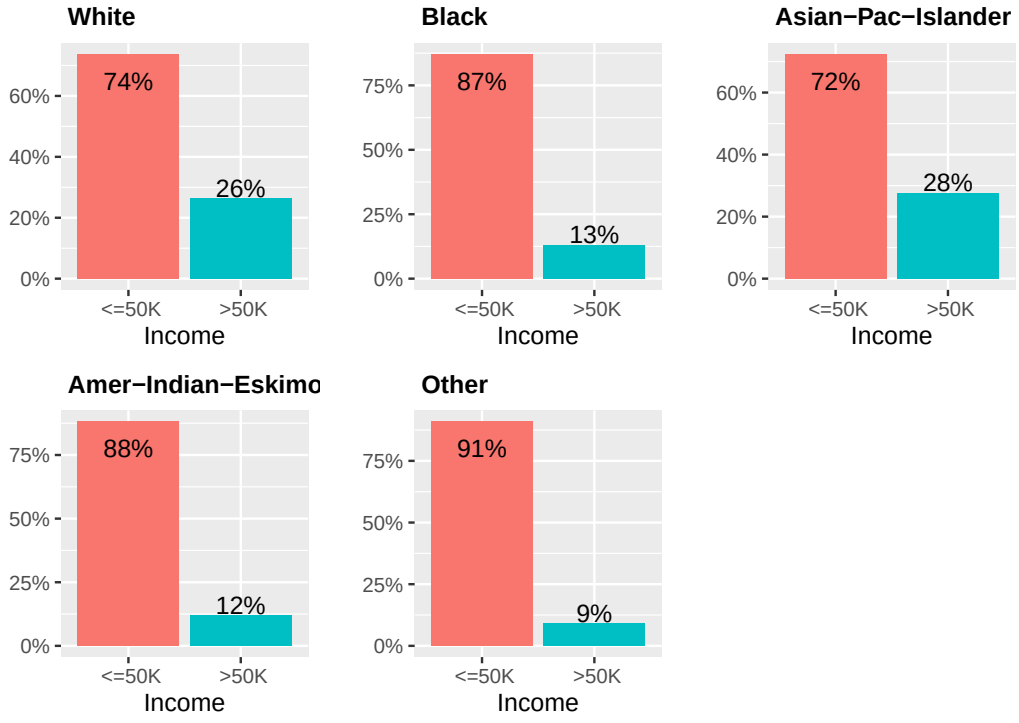




race

Most participants to the survey are white people, and because of this limitation it is not possible to extrapolate conclusions to the larger population.



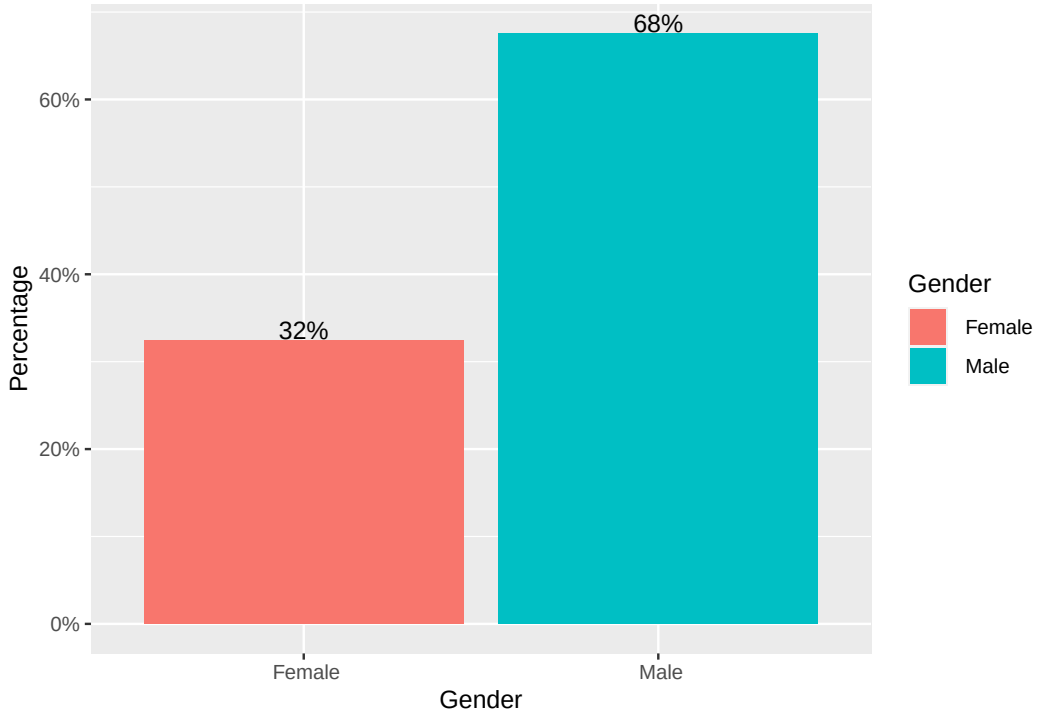


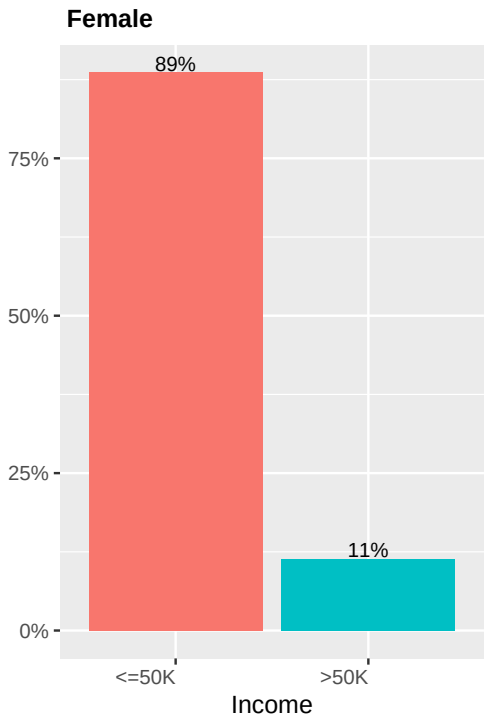
The variable *sex*

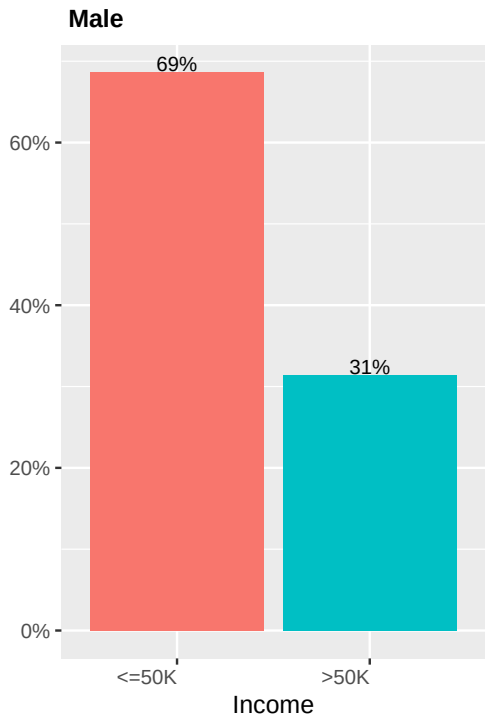
Taking into consideration the *race* variable it is possible to conclude that most survey respondents are white males.
There is a change the dataset might be biased.

```
summary(db.adult$sex)
```

```
##  Female    Male
##    9782   20380
```







Tests for independence of variables

Pearson's Chi Square Test of Independence will be used to test the independence of the categorical variables two by two.
The following null hypothesis will be tested:
H1: The two categorical variables are independent in the considered population against the alternative hypothesis
H2: The two categorical variables are dependent (related) in the considered population.

variables *sex* and *income*

The p-value is less than 0.05 which means the null hypothesis that the two categorical variables are independent is rejected.

```
chisq.test(db.adult$sex, db.adult$income)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  db.adult$sex and db.adult$income
## X-squared = 1415.3, df = 1, p-value < 2.2e-16
```

variables *race* and *income*

The null hypothesis is rejected at the 0.05 significance level, the p-values are less than 0.05. This indicates there is a strong correlation between "race" and "income".

```
chisq.test(db.adult$race, db.adult$income)
```

```
##
##  Pearson's Chi-squared test
##
## data:  db.adult$race and db.adult$income
## X-squared = 304.24, df = 4, p-value < 2.2e-16
```

variables *workclass* and *income*

A warning message is displayed, most likely because there are cells with expected cell counts less than 5. The results are to be interpreted with caution or neglected.

```
chisq.test(table(db.adult$workclass, db.adult$income))
```

```
## Warning in chisq.test(table(db.adult$workclass, db.adult$income)): Chi-squared
## approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(db.adult$workclass, db.adult$income)
## X-squared = 804.16, df = 6, p-value < 2.2e-16
```

```
## Warning in chisq.test(table(db.adult$workclass, db.adult$income)): Chi-squared
## approximation may be incorrect
```