ML1000-Assignment-1

Crystal Zhu

09/02/2021

Data Exploration

Check the structure of the dataset

```
## 'data.frame':    32561 obs. of  15 variables:
##  $ age          : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ workclass    : Factor w/ 9 levels " ?"," Federal-gov",..: 8 7 5 5 5 5 5 7 5 5 ...
##  $ fnlwgt       : int  77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...
##  $ education    : Factor w/ 16 levels " 10th"," 11th",..: 10 10 12 2 10 13 7 12 13 10 ...
##  $ educationnum : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ maritalstatus: Factor w/ 7 levels " Divorced"," Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...
##  $ occupation   : Factor w/ 15 levels " ?"," Adm-clerical",..: 2 5 7 7 11 5 9 5 11 5 ...
##  $ 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 ...
##  $ capitalgain  : int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ capitalloss  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hoursperweek : int  40 13 40 40 40 40 16 45 50 40 ...
##  $ nativecountry: Factor w/ 42 levels " ?"," Cambodia",..: 40 40 40 40 6 40 24 40 40 40 ...
##  $ income       : Factor w/ 2 levels " <=50K"," >50K": 1 1 1 1 1 1 1 2 2 2 ...
```

```
## The dimension of the dataset is:
```

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

```
## The proportion of the classes in target variable are:
```

```
##
##     <=50K      >50K
## 0.7591904 0.2408096
```

There are 76% of people in the first class (income no higher than 50K), thus this is an imbalanced dataset.

Check duplicates

```
## [1] 24 15
```

Check missing values.

```
## The number of missing values for each variable are:
```

```
##  [1] "age-0 missing values"             "workclass-1836 missing values"
##  [3] "fnlwgt-0 missing values"          "education-0 missing values"
##  [5] "educationnum-0 missing values"    "maritalstatus-0 missing values"
##  [7] "occupation-1843 missing values"   "relationship-0 missing values"
##  [9] "race-0 missing values"            "sex-0 missing values"
## [11] "capitalgain-0 missing values"     "capitalloss-0 missing values"
```

```
## [13] "hoursperweek-0 missing values"    "nativecountry-582 missing values"
## [15] "income-0 missing values"
```

```
## The percentages of missing values for each variable are:
```

```
##           age     workclass        fnlwgt     education  educationnum
##      0.000000      5.642807      0.000000      0.000000      0.000000
## maritalstatus    occupation  relationship          race           sex
##      0.000000      5.664321      0.000000      0.000000      0.000000
##   capitalgain   capitalloss  hoursperweek nativecountry        income
##      0.000000      0.000000      0.000000      1.788733      0.000000
```

From the above, there are missing values in the data and all the missing values are from categorical variables. Thus decide to remove the records with missings.

Remove rows with missing values

```
## Now the dimension of the dataset becomes:
```

```
## [1] 30139    15
```

Recode the values of target variable

```
## Now the levels of the target variable are:
```

```
##     N     Y
## 22633  7506
```

Check data types

```
##           age     workclass        fnlwgt     education  educationnum
##     "integer"      "factor"     "integer"      "factor"     "integer"
## maritalstatus    occupation  relationship          race           sex
##      "factor"      "factor"      "factor"      "factor"      "factor"
##   capitalgain   capitalloss  hoursperweek nativecountry        income
##     "integer"     "integer"     "integer"      "factor"      "factor"
```

We can see there are both numeric and categorical variables in the dataset.

Check outliers for numeric variables

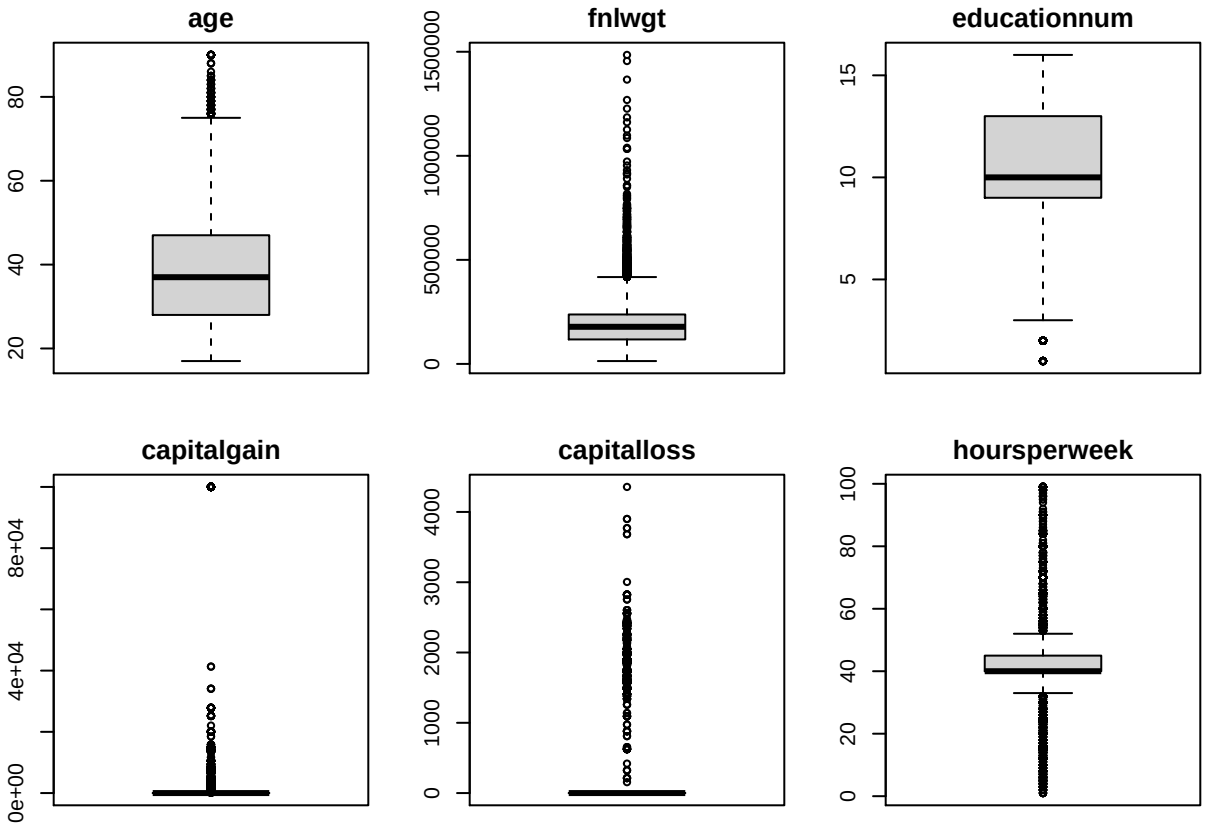
```
## age  -
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   28.00   37.00   38.44   47.00   90.00
```

```
## fnlwgt  -
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13769  117628  178417  189795  237605 1484705
```

```
## educationnum  -
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    9.00   10.00   10.12   13.00   16.00
```

```
## capitalgain  -
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0    1093       0   99999
```

```
## capitalloss  -
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00   88.44    0.00 4356.00
```

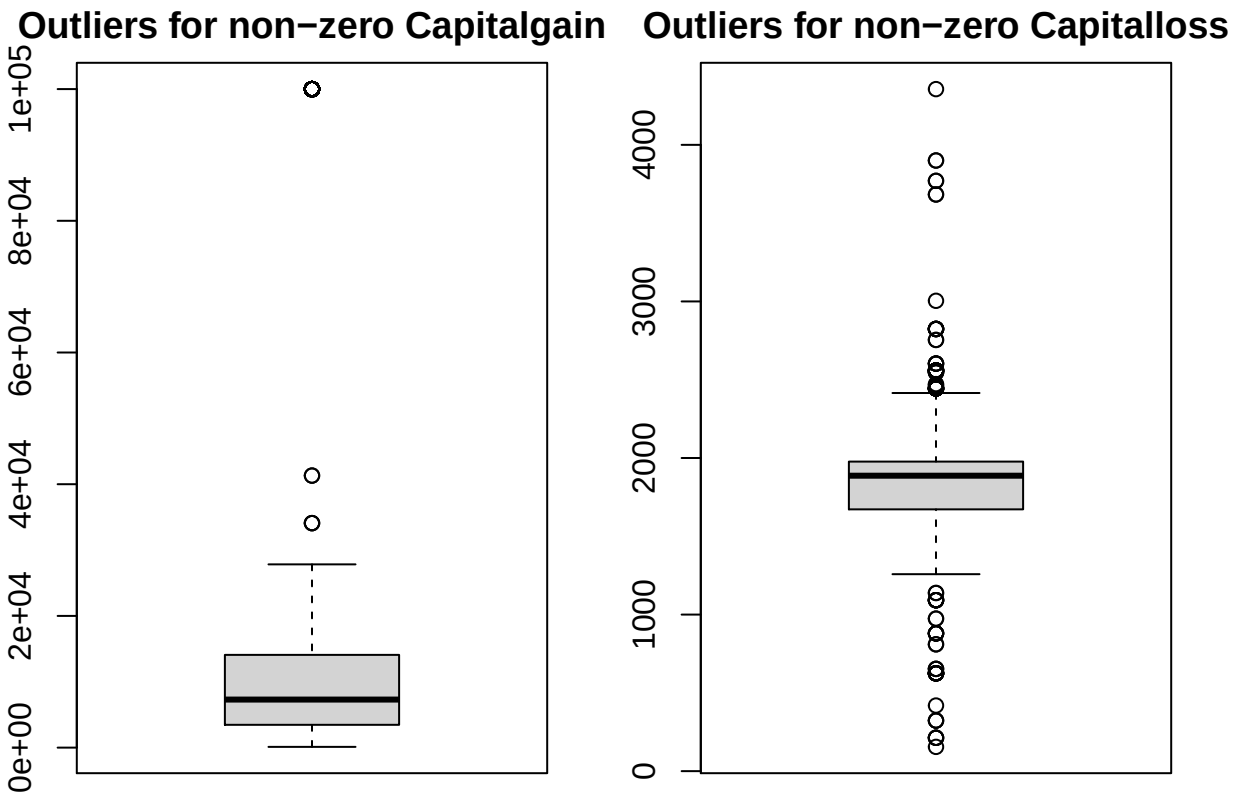



```
## hoursperweek  -
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   40.00   40.00   40.93   45.00   99.00
```

Since there are large number of zeros in capitalgain & capitalloss variables, let's check if there are outliers for non-zero values

```
par(mar = c(2,2,2,2))
par(mfrow=c(1,2))
boxplot(X$capitalgain[which(X$capitalgain!=0)],main="Outliers for non-zero Capitalgain")

boxplot(X$capitalloss[which(X$capitalloss!=0)],main="Outliers for non-zero Capitalloss")
```




```
#There are still outliers even excluding zeros for capitalgain and capitalloss variables.
```

There are many outliers for all numeric variables.

Check validality of column values

```
## [ 1 ] age - Numeric
##  Min 1st Qu. Mean 3rd Qu. Max:  17 28 38.44172 47 90
## [ 2 ] workclass - Categorical
##   Federal-gov  Local-gov  Never-worked  Private  Self-emp-inc  Self-emp-not-inc  State-gov  Without-
## [ 3 ] fnlwgt - Numeric
##  Min 1st Qu. Mean 3rd Qu. Max:  13769 117627.5 189795 237604.5 1484705
## [ 4 ] education - Categorical
##   10th  11th  12th  1st-4th  5th-6th  7th-8th  9th  Assoc-acdm  Assoc-voc  Bachelors  Doctorate  HS-
## [ 5 ] educationnum - Numeric
##  Min 1st Qu. Mean 3rd Qu. Max:  1 9 10.12253 13 16
## [ 6 ] maritalstatus - Categorical
##   Divorced  Married-AF-spouse  Married-civ-spouse  Married-spouse-absent  Never-married  Separated
## [ 7 ] occupation - Categorical
##   Adm-clerical  Armed-Forces  Craft-repair  Exec-managerial  Farming-fishing  Handlers-cleaners  Mac
## [ 8 ] relationship - Categorical
##   Husband  Not-in-family  Other-relative  Own-child  Unmarried  Wife
## [ 9 ] race - Categorical
##   Amer-Indian-Eskimo  Asian-Pac-Islander  Black  Other  White
## [ 10 ] sex - Categorical
##   Female  Male
## [ 11 ] capitalgain - Numeric
```

```
##  Min 1st Qu. Mean 3rd Qu. Max:  0 0 1092.841 0 99999
## [ 12 ] capitalloss - Numeric
##  Min 1st Qu. Mean 3rd Qu. Max:  0 0 88.43993 0 4356
## [ 13 ] hoursperweek - Numeric
##  Min 1st Qu. Mean 3rd Qu. Max:  1 40 40.9347 45 99
## [ 14 ] nativecountry - Categorical
##   Cambodia  Canada  China  Columbia  Cuba  Dominican-Republic  Ecuador  El-Salvador  England  France
## [ 15 ] income - Categorical
##  N Y
```

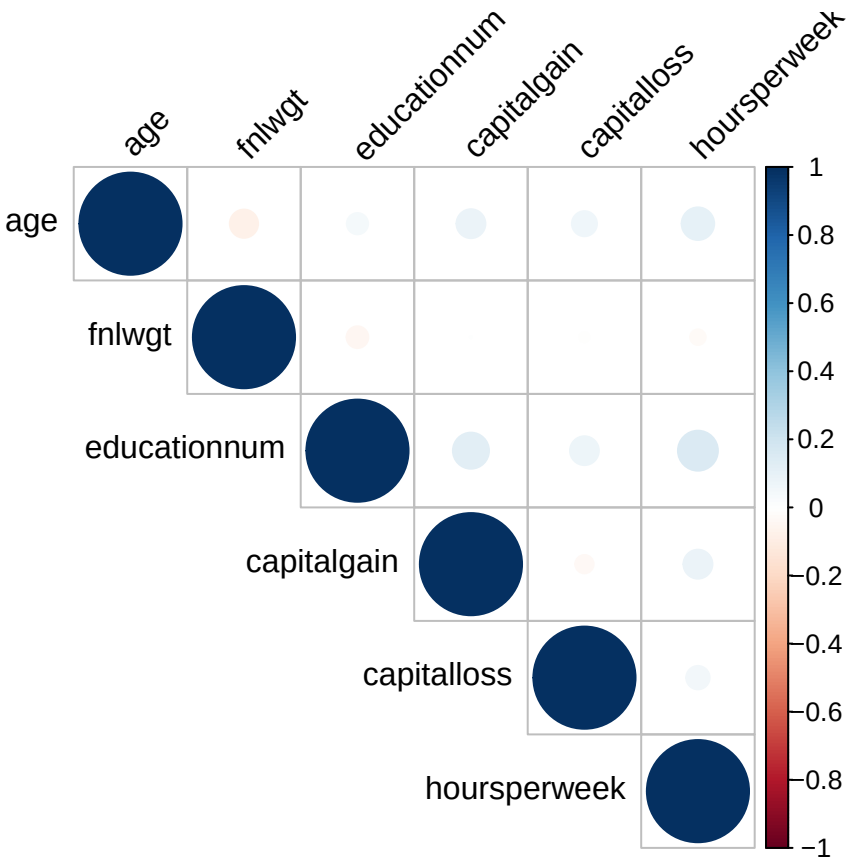
Exploratory data analysis

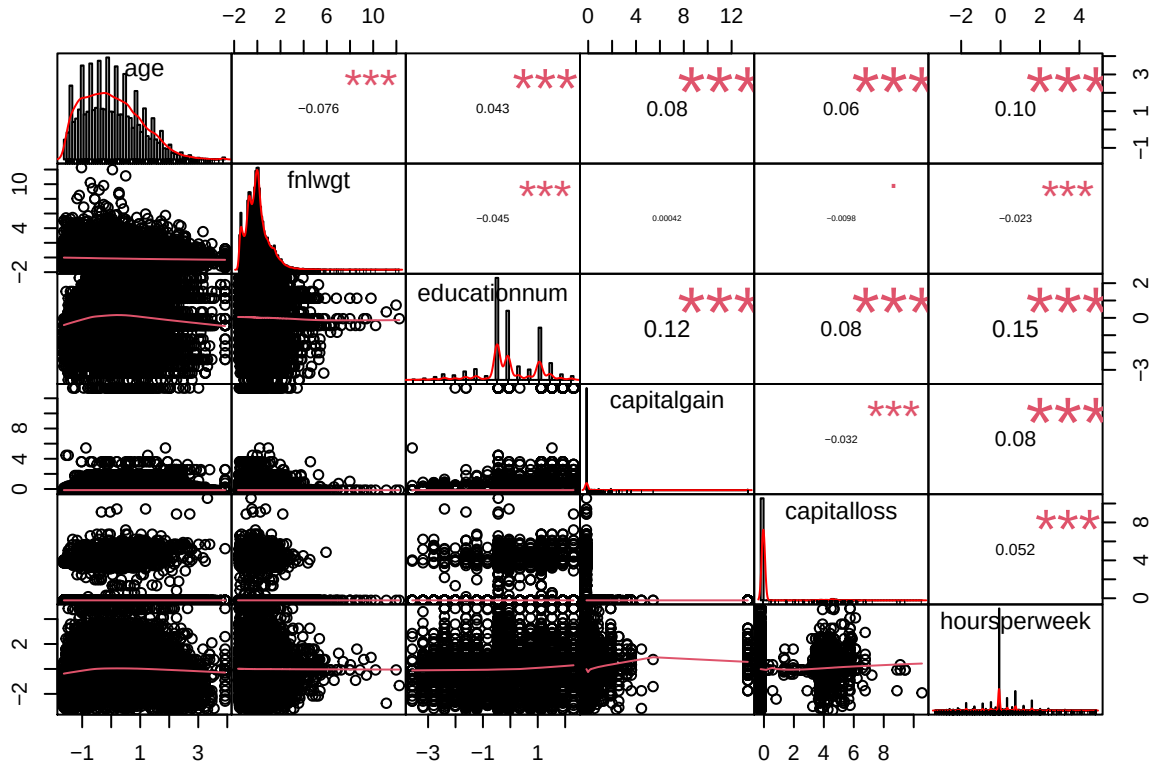
Check redundancy and correlations among variables - how one attribute's values vary from those of another

correlations

Pearson's correlation for numeric variables

```
##                age fnlwgt educationnum capitalgain capitalloss hoursperweek
## age           1.00  -0.08         0.04        0.08        0.06         0.10
## fnlwgt       -0.08   1.00        -0.05        0.00       -0.01        -0.02
## educationnum  0.04  -0.05         1.00        0.12        0.08         0.15
## capitalgain   0.08   0.00         0.12        1.00       -0.03         0.08
## capitalloss   0.06  -0.01         0.08       -0.03        1.00         0.05
## hoursperweek  0.10  -0.02         0.15        0.08        0.05         1.00
##
## n= 30139
##
##
## P
##              age    fnlwgt educationnum capitalgain capitalloss hoursperweek
## age                 0.0000 0.0000       0.0000      0.0000      0.0000
## fnlwgt       0.0000        0.0000       0.9419      0.0903      0.0000
## educationnum 0.0000 0.0000               0.0000      0.0000      0.0000
## capitalgain  0.0000 0.9419 0.0000                    0.0000      0.0000
## capitalloss  0.0000 0.0903 0.0000       0.0000                  0.0000
## hoursperweek 0.0000 0.0000 0.0000       0.0000      0.0000
```


age
fnlwgt
educationnum
capitalgain
capitalloss
hoursperweek
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1



Chi-square test & Cramer's V to show associations between categorical variables

```
## The chi-square test statistics for all combinations of categorical variables:

##                workclass education maritalstatus occupation relationship
## workclass             NA   2445.29     1720534.1   9314.397     9314.397
## education             NA        NA      127730.7   2182.093     2182.093
## maritalstatus         NA        NA            NA 321142.939   321142.939
## occupation            NA        NA            NA         NA   452085.000
## relationship          NA        NA            NA         NA           NA
## race                  NA        NA            NA         NA           NA
## sex                   NA        NA            NA         NA           NA
## nativecountry         NA        NA            NA         NA           NA
## income                NA        NA            NA         NA           NA
##                     race        sex nativecountry      income
## workclass      14712.967   4214.698      12388.388    357.3079
## education       1078.258   8534.112       1198.451    398.1371
## maritalstatus 137835.193 269434.626     115466.274 111620.4051
## occupation      1364.782  15299.116       2088.268    689.7322
## relationship    1364.782  15299.116       2088.268    689.7322
## race                  NA   3154.448      35767.888    844.6834
## sex                   NA         NA       4814.169    846.7848
## nativecountry         NA         NA             NA   1136.5222
## income                NA         NA             NA          NA

## The p-values of chi-square tests for all combinations of categorical variables:
```

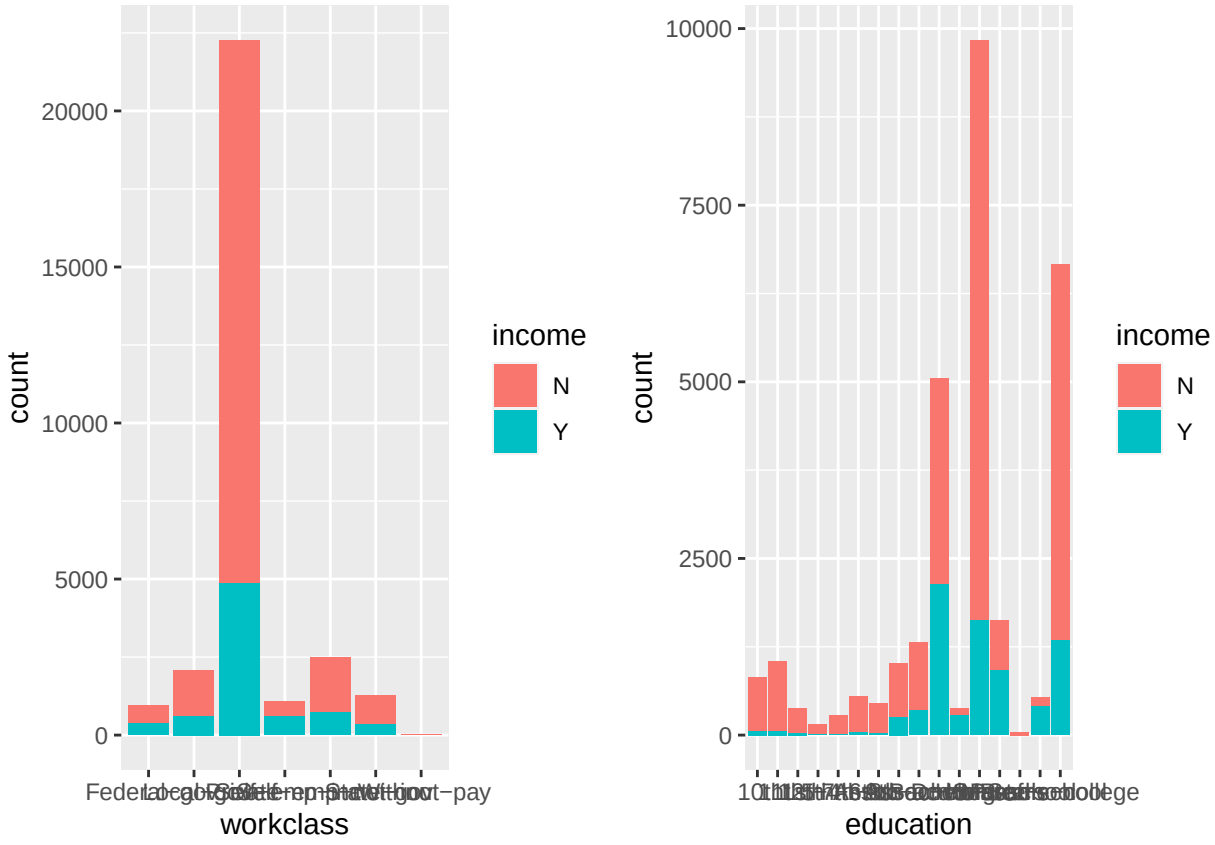
```
##               workclass     education maritalstatus    occupation  relationship
## workclass            NA 8.454723e-280  0.000000e+00  0.000000e+00  0.000000e+00
## education            NA            NA  5.362309e-35  0.000000e+00  0.000000e+00
## maritalstatus        NA            NA            NA 1.766781e-104 1.766781e-104
## occupation           NA            NA            NA            NA  0.000000e+00
## relationship         NA            NA            NA            NA            NA
## race                 NA            NA            NA            NA            NA
## sex                  NA            NA            NA            NA            NA
## nativecountry        NA            NA            NA            NA            NA
## income               NA            NA            NA            NA            NA
##                       race          sex nativecountry        income
## workclass      0.000000e+00 0.000000e+00  0.000000e+00  2.035856e-03
## education     5.765068e-203 0.000000e+00 5.196391e-233  1.811564e-69
## maritalstatus 1.061727e-219 8.302645e-17 2.848609e-199  0.000000e+00
## occupation    8.773627e-227 0.000000e+00  0.000000e+00 8.126362e-108
## relationship  8.773627e-227 0.000000e+00  0.000000e+00 8.126362e-108
## race                    NA 0.000000e+00  0.000000e+00 7.445756e-163
## sex                     NA           NA  0.000000e+00 4.239502e-144
## nativecountry           NA           NA            NA 2.788906e-228
## income                  NA           NA            NA            NA
```

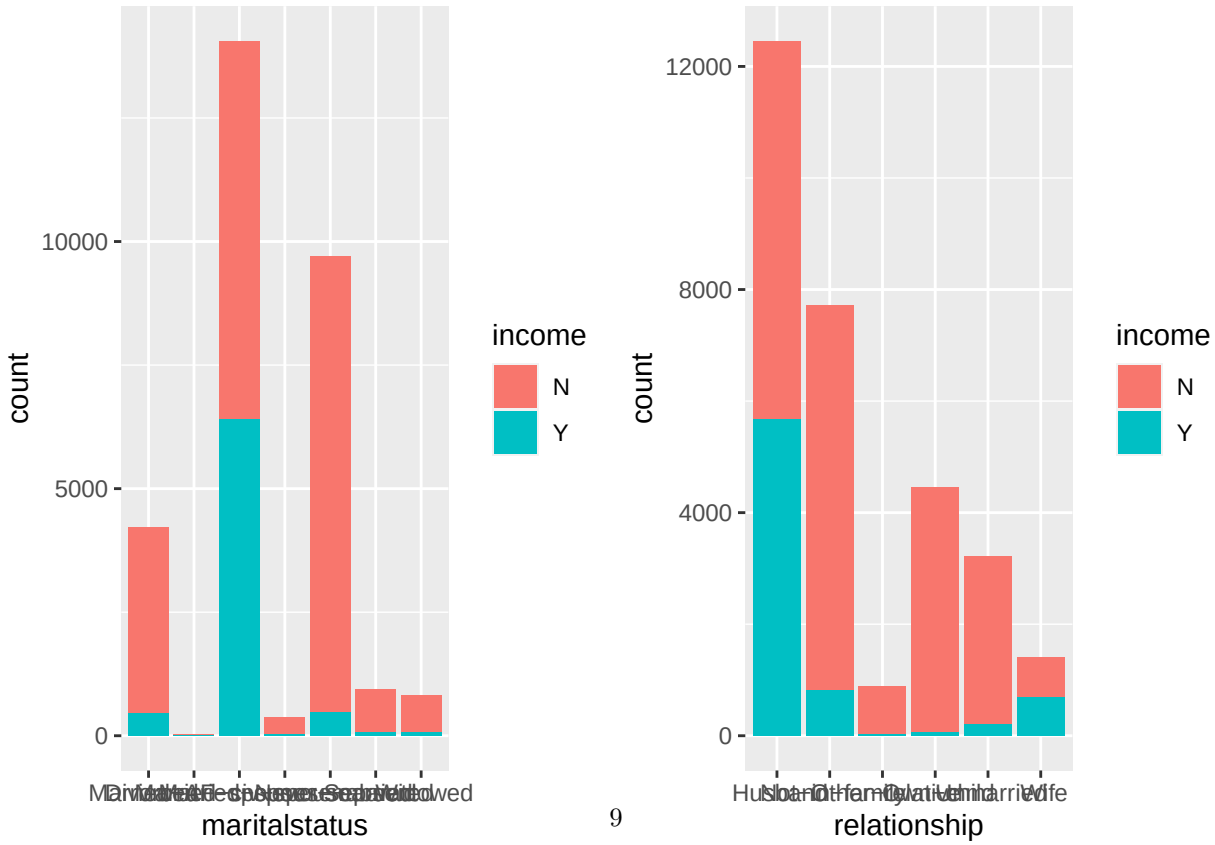
```
## The associations (Cramer's V) for all combinations of categorical variables:
```

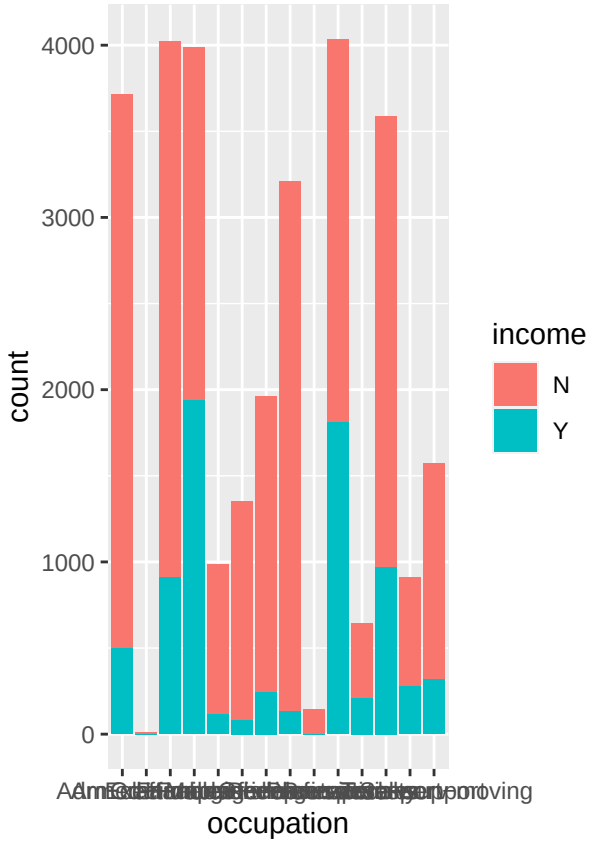
```
##               workclass education maritalstatus occupation relationship
## workclass            NA       NaN     0.8966812  0.1435381    0.1435381
## education            NA        NA           NaN        NaN          NaN
## maritalstatus        NA        NA            NA  0.8428284    0.8428284
## occupation           NA        NA            NA         NA    1.0000000
## relationship         NA        NA            NA         NA           NA
## race                 NA        NA            NA         NA           NA
## sex                  NA        NA            NA         NA           NA
## nativecountry        NA        NA            NA         NA           NA
## income               NA        NA            NA         NA           NA
##                     race       sex nativecountry     income
## workclass     0.28523978 0.1037163     0.2867200 0.05444110
## education            NaN       NaN           NaN        NaN
## maritalstatus 0.87305180 0.8292598     0.8753428 0.96222665
## occupation    0.08687435 0.1976046     0.1177182 0.07563901
## relationship  0.08687435 0.1976046     0.1177182 0.07563901
## race                  NA 0.1320753     0.4871887 0.08370522
## sex                   NA        NA     0.1787357 0.08380928
## nativecountry         NA        NA            NA 0.09709448
## income                NA        NA            NA         NA
```

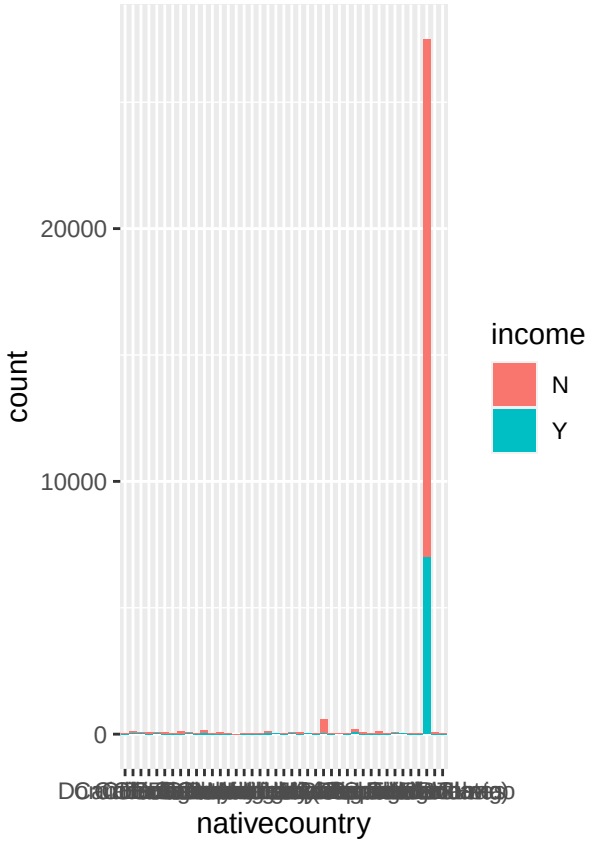
Barcharts for categorical variables

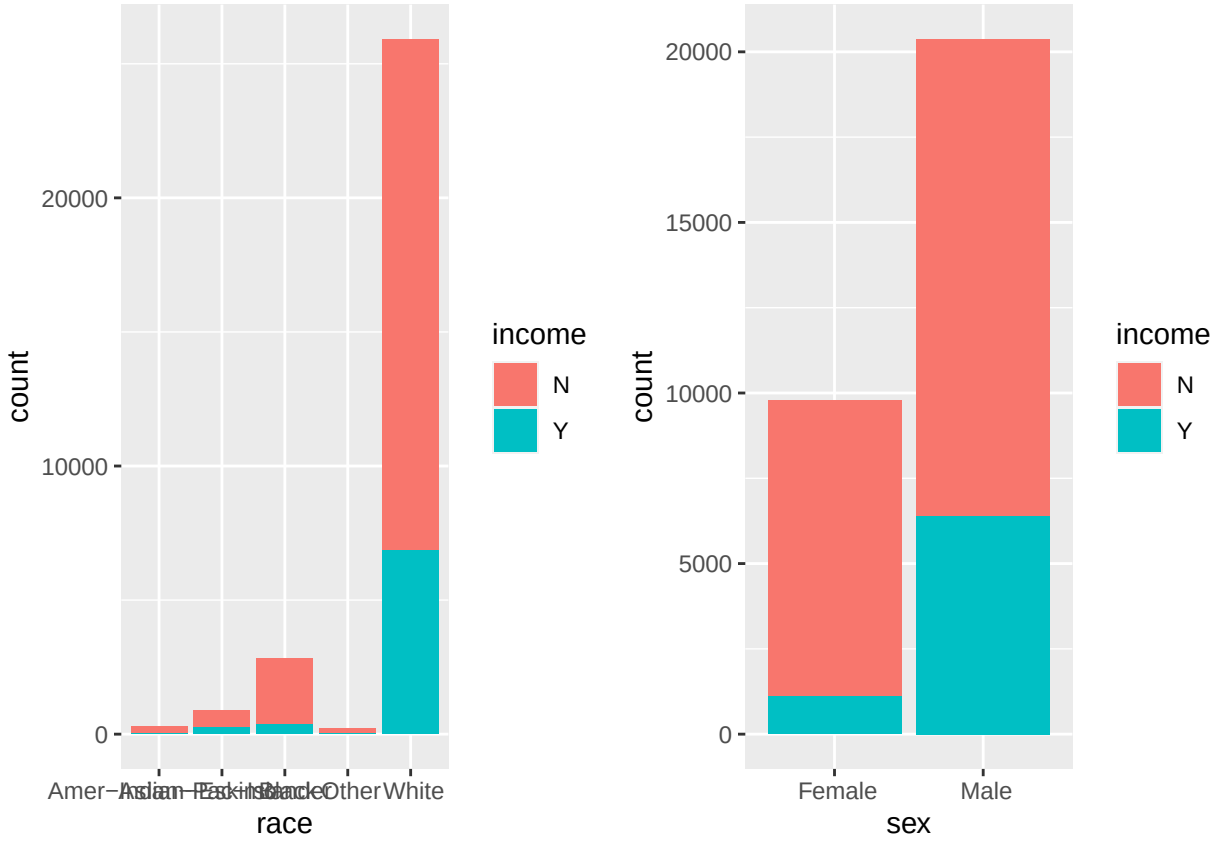
barplots for categorical variables by target variable






4000
3000
2000
1000
0
count
income
N
Y
occupation


20000
10000
0
count
income
N
Y
nativecountry



Feature Engineering

1.Education and educationnum are redundant

```
##
##                  1    2    3    4    5    6    7    8    9   10   11   12
##    10th          0    0    0    0    0  820    0    0    0    0    0    0
##    11th          0    0    0    0    0    0 1048    0    0    0    0    0
##    12th          0    0    0    0    0    0    0  377    0    0    0    0
##    1st-4th       0  149    0    0    0    0    0    0    0    0    0    0
##    5th-6th       0    0  287    0    0    0    0    0    0    0    0    0
##    7th-8th       0    0    0  556    0    0    0    0    0    0    0    0
##    9th           0    0    0    0  455    0    0    0    0    0    0    0
##    Assoc-acdm    0    0    0    0    0    0    0    0    0    0    0 1008
##    Assoc-voc     0    0    0    0    0    0    0    0    0    0 1307    0
##    Bachelors     0    0    0    0    0    0    0    0    0    0    0    0
##    Doctorate     0    0    0    0    0    0    0    0    0    0    0    0
##    HS-grad       0    0    0    0    0    0    0    0 9834    0    0    0
##    Masters       0    0    0    0    0    0    0    0    0    0    0    0
##    Preschool    44    0    0    0    0    0    0    0    0    0    0    0
##    Prof-school   0    0    0    0    0    0    0    0    0    0    0    0
##    Some-college  0    0    0    0    0    0    0    0    0 6669    0    0
##
##                  13   14   15   16
##    10th          0    0    0    0
##    11th          0    0    0    0
```

```
##    12th            0    0    0    0
##    1st-4th         0    0    0    0
##    5th-6th         0    0    0    0
##    7th-8th         0    0    0    0
##    9th             0    0    0    0
##    Assoc-acdm      0    0    0    0
##    Assoc-voc       0    0    0    0
##    Bachelors    5042    0    0    0
##    Doctorate       0    0    0  375
##    HS-grad         0    0    0    0
##    Masters         0 1626    0    0
##    Preschool       0    0    0    0
##    Prof-school     0    0  542    0
##    Some-college    0    0    0    0
```

From the above perfectly 1-1 relationship, we can see these two variabls are essentially exact the same. So we decide to remove the educationnum variable.

2. Nativecountry

There are 41 levels, namely 41 different countries, in the dataset. Since too many levels for a categorical variable could lead to overfitting, we decide to regroup the native countries into regions.

```
##
##  Amer-Indian-Eskimo  Asian-Pac-Islander               Black               Other
##                   2                  68                   0                   0
##               White
##                   1
```

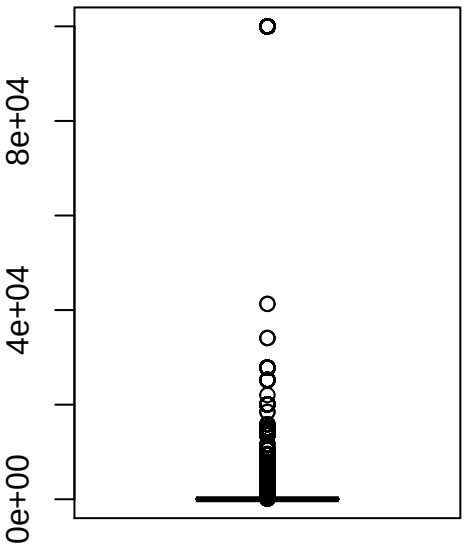
Since the race of almost all records with Native country "South" is "Asian-Pac-Islander", we think the country "South" is very likely to be South Korea. So we decided to group the country "South" into Asia_East

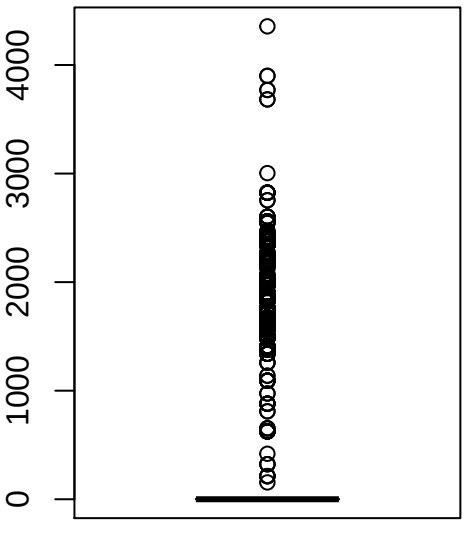
3. Capital Gain and Capital Loss

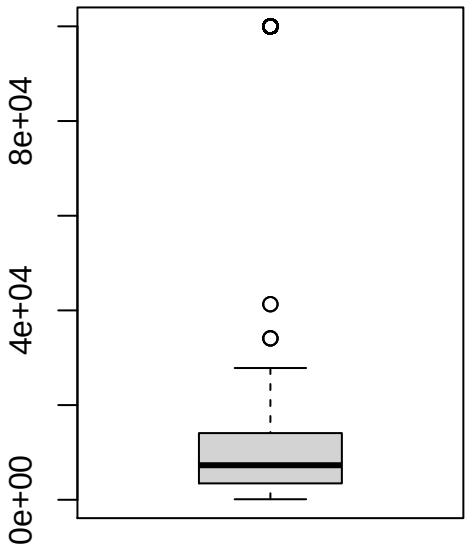
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0    1093       0   99999
```

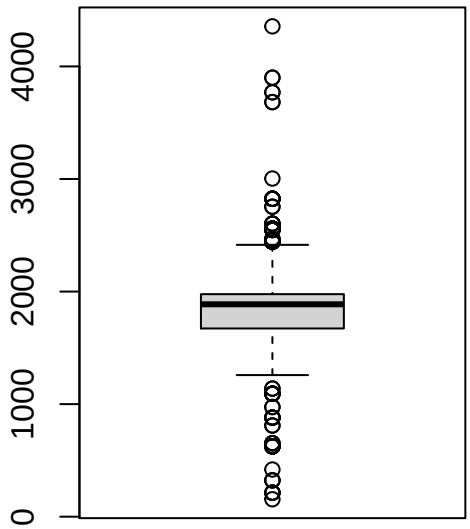
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00   88.44    0.00 4356.00
```

```
## the proportion of zeros in Capital Gain is 91.57902%
## the proportion of zeros in Capital Loss is 95.26527%
```


0e+00
4e+04
8e+04


0
1000
2000
3000
4000





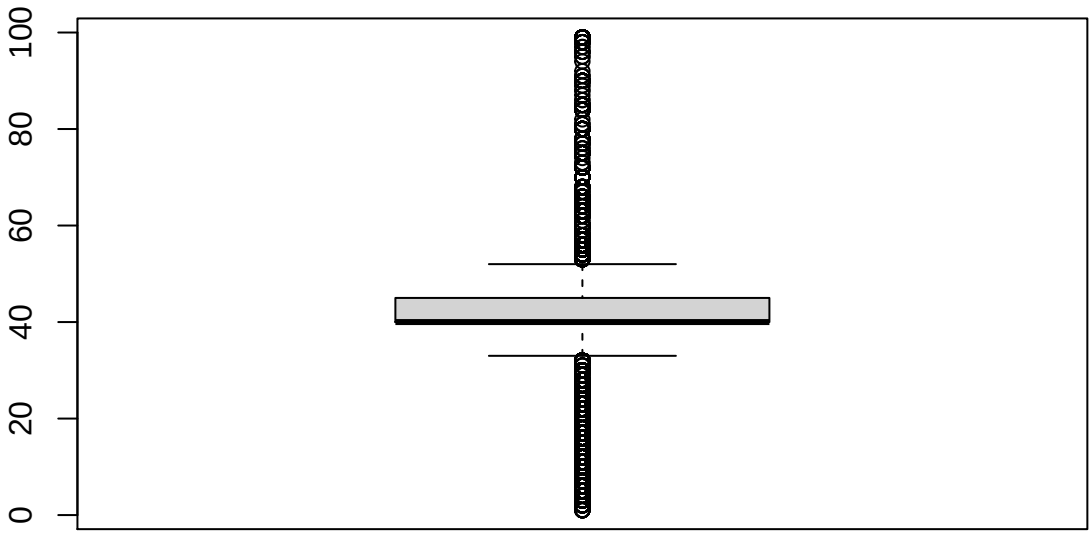

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     114    3464    7298   12978   14084   99999
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     155    1672    1887    1868    1977    4356
```

From the above boxplots and summary of Captital Gain and Capital Loss variables, we can see that over 90% of the two variables are 0. So we decided to categorize the two variables into groups - one group is for the zeros, and other groups based on quantiles of non-zeros.More specifically, if the value is 0, then it's grouped as "Zero". If the value is not zero and lower than 1st quantile, it's grouped as "Low". The values between 1st and 3rd quantitles are grouped into "Medium" and those higher than 3rd quantile are categorized as "High".

4. Hours per week

From the previous boxplot, we can see there are large number of outliers in the hours_per_week variable. Let's review.

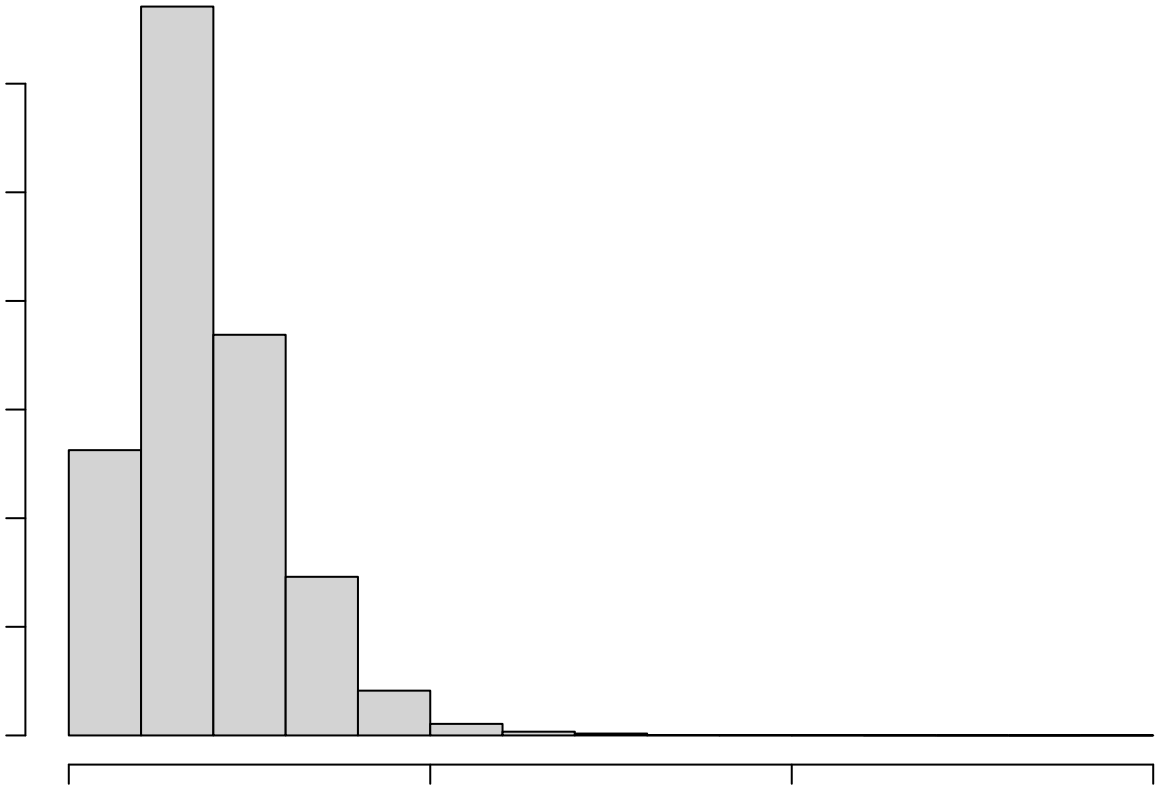


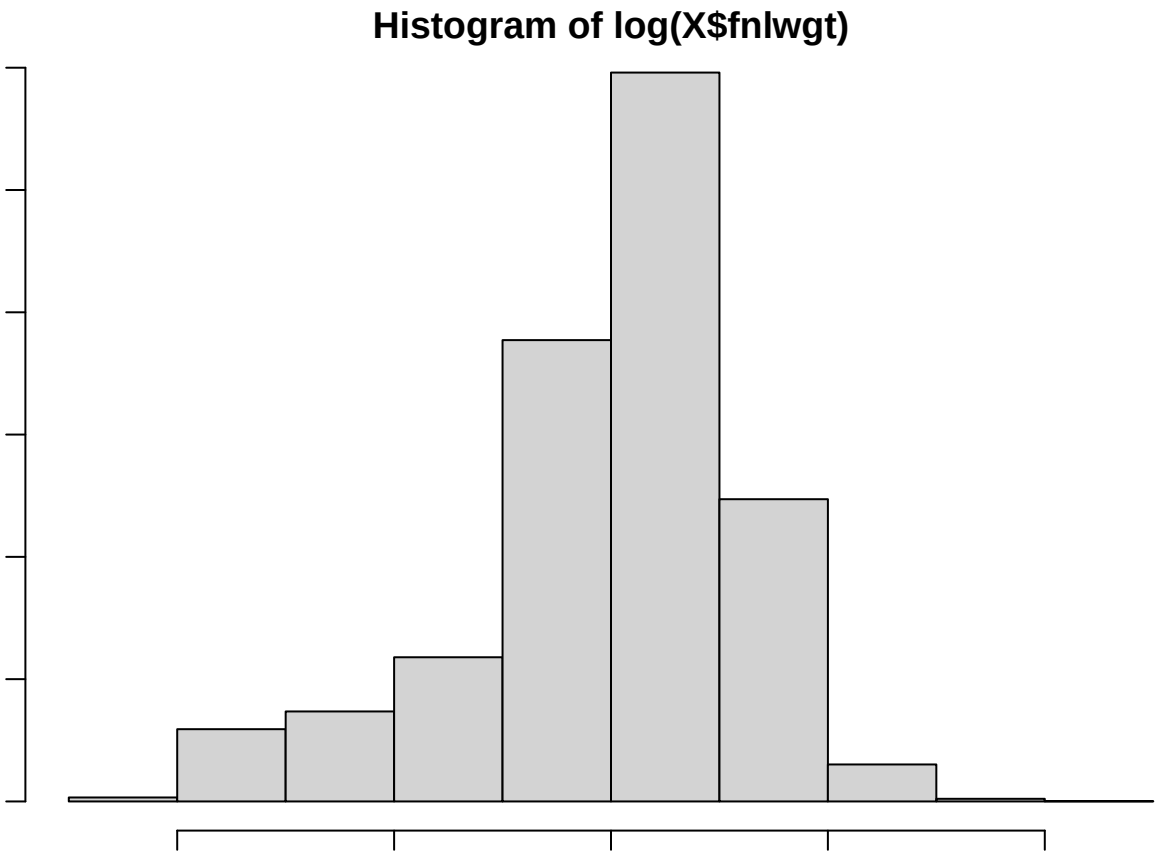

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   40.00   40.00   40.93   45.00   99.00
```

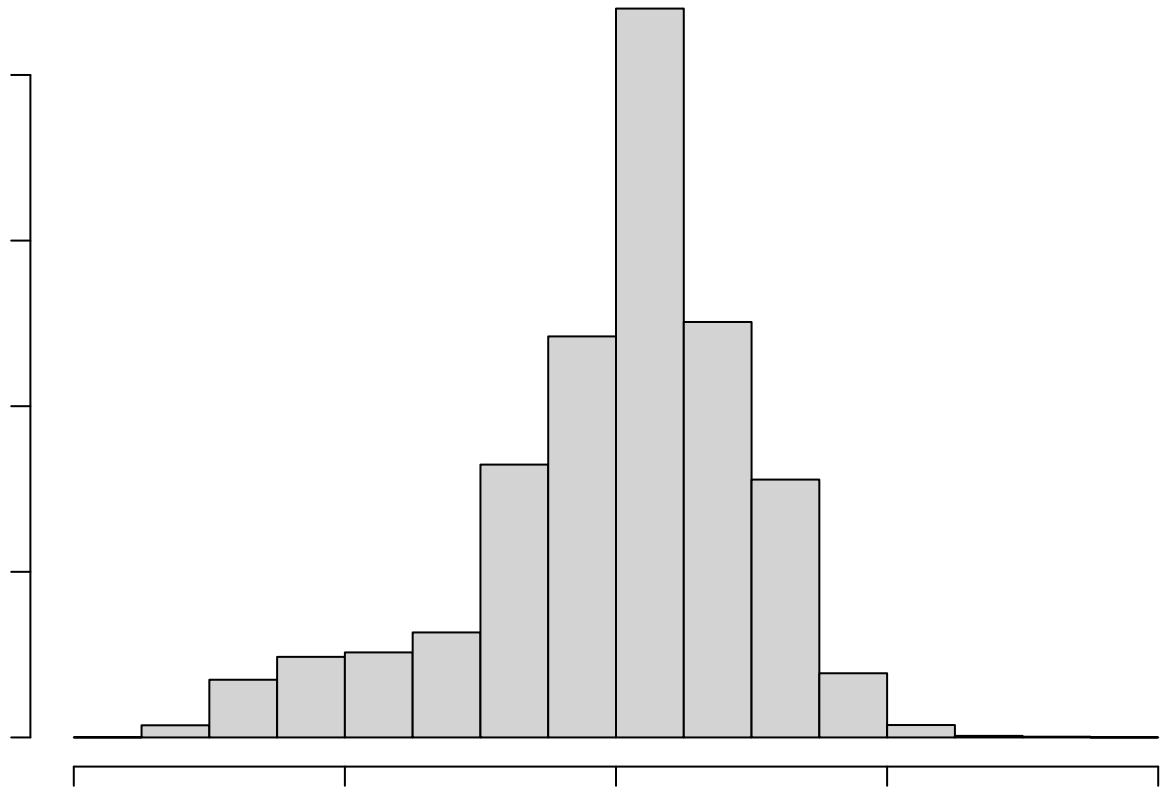
We decide to group this variable in the following way: if the value is lower than the 1st quantile (40), it's called "less_than_40". If a value is between the 1st and 3rd quantitle (45), it's called "between_40_and_45". If the value is higher than 3rd quantile, it's called "higher_than_45".

5. fnlwgt

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13769  117628  178417  189795  237605 1484705
```


Histogram of X$fnlwgt


Histogram of log(X$fnlwgt)



From the above plots, we can see that the fnlwgt variable generally follows a log-normal distribution and the values are generally large. So we decide to do a log transformation and then a standardization on it.

6. Age standardization
7. Drop empty level of workclass

Drop the original variables for the re-grouped variables.

```
## The current structure of the dataset is:

## 'data.frame':    30139 obs. of  14 variables:
##  $ workclass      : Factor w/ 7 levels " Federal-gov",..: 6 5 3 3 3 3 3 5 3 3 ...
##  $ education      : Factor w/ 16 levels " 10th"," 11th",..: 10 10 12 2 10 13 7 12 13 10 ...
##  $ maritalstatus  : 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 ...
##  $ income         : Factor w/ 2 levels "N","Y": 1 1 1 1 1 1 1 2 2 2 ...
##  $ native_region  : Factor w/ 9 levels " Central-America",..: 8 8 8 8 1 8 1 8 8 8 ...
##  $ cap_gain       : Factor w/ 4 levels "High","Low","Medium",..: 2 4 4 4 4 4 4 4 4 1 3 ...
##  $ cap_loss       : Factor w/ 4 levels "High","Low","Medium",..: 4 4 4 4 4 4 4 4 4 4 4 ...
##  $ hours_w        : Factor w/ 3 levels "between_40_and_45",..: 1 3 1 1 1 1 3 1 2 1 ...
##  $ fnlwgt_logstand: num [1:30139, 1] -1.151 -1.036 0.472 0.606 1.186 ...
##  $ age_stand      : num [1:30139, 1] 0.0425 0.8802 -0.0336 1.1087 -0.7952 ...
```

Train supervised learning models

```
## The percentages of the two levels of target variable are:
```

```
##     N     Y
## 18107  6005
```

so randomly select 1/3 rows from the majority group of the training data to balance the training dataset

```
## The percentages of the two levels of target variable at the randomly selected sample are:
```

```
##    N    Y
## 6337    0
```

Then combine the randomly down sampled majority group with the original minority group

```
## The percentages of the two levels of target variable at the combined balanced sample are:
```

```
##    N    Y
## 6337 6005
```

Then randomly shuffle the rows so that not all "N" incomes are on top and "Y"s are in the end

```
## The percentages of the two levels of target variable at the final training dataset are not changed!
```

```
##    N    Y
## 6337 6005
```

First, Let's fit a random forest model.

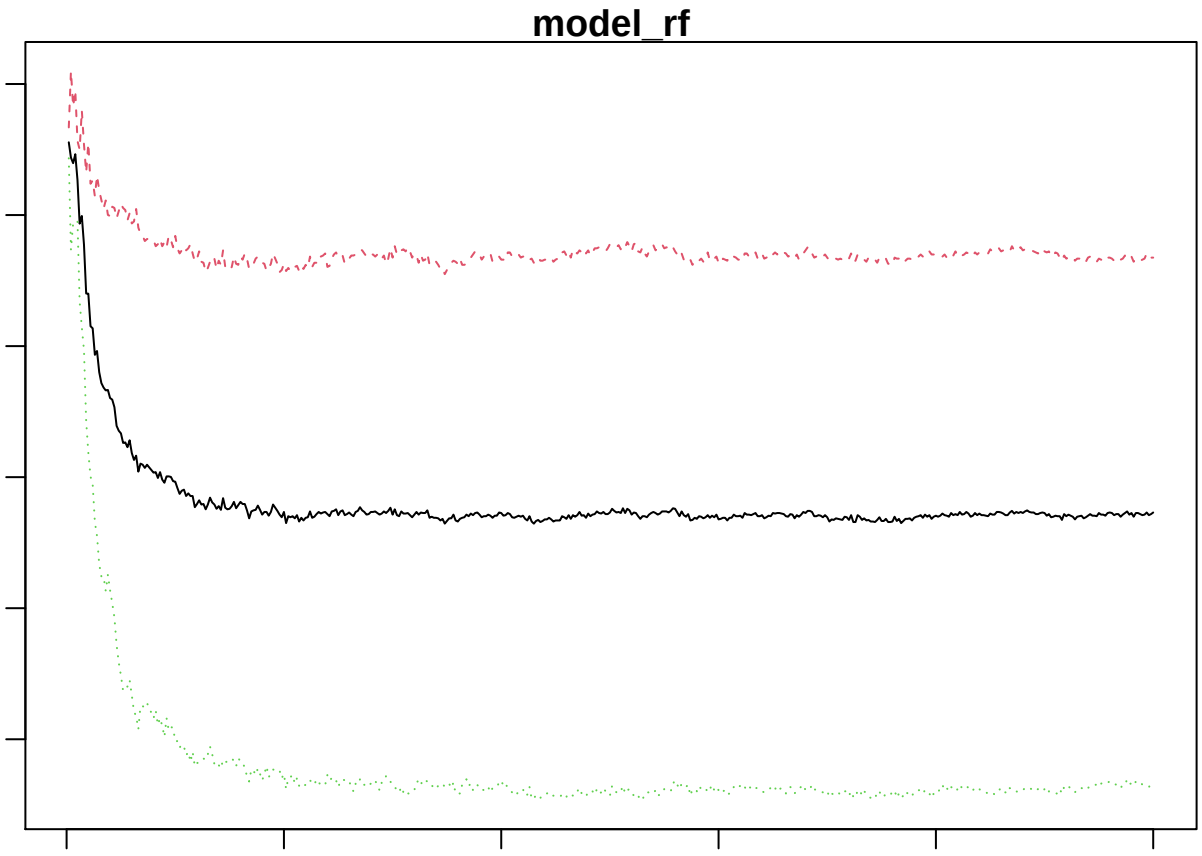
```
## The fitted random forest model:
```

```
##
## Call:
##  randomForest(formula = income ~ ., data = X_train_bal, trControl = cv_5)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 17.46%
## Confusion matrix:
##      N    Y class.error
## N 4984 1353   0.2135080
## Y  802 5203   0.1335554
```

```
## Importance of features based on the random forest model:
```

```
##                 MeanDecreaseGini
## workclass              203.34772
## education              657.74296
## maritalstatus          697.09625
## occupation             592.00506
## relationship           836.16462
## race                    80.28688
## sex                    103.37707
## native_region           80.83812
```

```
## cap_gain               351.70769
## cap_loss                93.19420
## hours_w                199.00989
## fnlwgt_logstand        554.62025
## age_stand              787.29905
```




```
## The confusion matrix of random forest model:

## Confusion Matrix and Statistics
##
##           Reference
## Prediction    N    Y
##          N 5797   79
##          Y  540 5926
##
##                Accuracy : 0.9498
##                  95% CI : (0.9458, 0.9536)
##     No Information Rate : 0.5135
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8998
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9868
##             Specificity : 0.9148
##          Pos Pred Value : 0.9165
```

```
##          Neg Pred Value : 0.9866
##              Prevalence : 0.4865
##          Detection Rate : 0.4801
##    Detection Prevalence : 0.5239
##       Balanced Accuracy : 0.9508
##
##        'Positive' Class : Y
##
```

Then let's try a knn model.

```
## The fitted knn model:
```

```
## k-Nearest Neighbors
##
## 12342 samples
##    13 predictor
##     2 classes: 'N', 'Y'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 9874, 9873, 9874, 9874, 9873
## Resampling results across tuning parameters:
##
##   k  Accuracy   Kappa
##   5  0.7943607  0.5891189
##   7  0.8011670  0.6028533
##   9  0.8071630  0.6150502
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 9.
```

```
## The confusion matrix of knn model:
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    N    Y
##          N 5111  712
##          Y 1226 5293
##
##                Accuracy : 0.843
##                  95% CI : (0.8364, 0.8494)
##     No Information Rate : 0.5135
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6864
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.8814
##             Specificity : 0.8065
##          Pos Pred Value : 0.8119
##          Neg Pred Value : 0.8777
##              Prevalence : 0.4865
##          Detection Rate : 0.4289
```

```
##    Detection Prevalence : 0.5282
##       Balanced Accuracy : 0.8440
##
##        'Positive' Class : Y
##
```

Next, let's try a glm model.

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## The fitted glm model:
```

```
## Generalized Linear Model
##
## 12342 samples
##    13 predictor
##     2 classes: 'N', 'Y'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 9873, 9874, 9874, 9874, 9873
## Resampling results:
##
##   Accuracy   Kappa
##   0.8211802  0.6424959
```

```
## The confusion matrix of glm model:
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    N    Y
##          N 5116  947
##          Y 1221 5058
##
##                Accuracy : 0.8243
##                  95% CI : (0.8175, 0.831)
##     No Information Rate : 0.5135
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6488
##
##  Mcnemar's Test P-Value : 4.541e-09
##
##             Sensitivity : 0.8423
##             Specificity : 0.8073
##          Pos Pred Value : 0.8055
##          Neg Pred Value : 0.8438
##              Prevalence : 0.4865
##          Detection Rate : 0.4098
##    Detection Prevalence : 0.5088
##       Balanced Accuracy : 0.8248
##
##        'Positive' Class : Y
##
```

Finally, let's try regularized logistic regression.

```
## The confusion matrix of logistic regression model:
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    N    Y
##          N 5125  941
##          Y 1212 5064
##
##                Accuracy : 0.8256
##                  95% CI : (0.8187, 0.8322)
##     No Information Rate : 0.5135
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6513
##
##  Mcnemar's Test P-Value : 5.923e-09
##
##             Sensitivity : 0.8433
##             Specificity : 0.8087
##          Pos Pred Value : 0.8069
##          Neg Pred Value : 0.8449
##              Prevalence : 0.4865
##          Detection Rate : 0.4103
##    Detection Prevalence : 0.5085
##       Balanced Accuracy : 0.8260
##
##        'Positive' Class : Y
##
```