BUAN 6356.004/MIS 6356.004
BUSINESS ANALYTICS WITH R

ANALYSIS ON 1994 Adult US CENSUS DATA

Rudra Abhishek (rxa200019)

Executive Summary

The project has been done on U.S. census data from 1994 taken from the UC Irvine Machine Learning Repository website. The project is divided into three parts: cleaning and preprocessing the data, exploratory data analysis, and predictive analysis. The first part of cleaning and preprocessing the data includes eliminating the non-important variables and transforming the data. The second part of the exploratory data analysis includes data visualization by creating graphs to understand the representations of some demographics in the data and also to understand the relationship between the response variable, income, and other variables. The third part of predictive analysis includes conducting classification analysis on the dataset by using techniques such as decision trees, neural networks, and logistic regression. The predictive analysis will help with building a model which can be used to predict whether the annual income is above $50,000 or equal to or below $50,000 based on a particular set of attributes. After implementing the classification techniques, the results will be used to arrive at the best model for the predictive task with the highest accuracy and lowest misclassification rate.

Motivation/Background

The level of disparity among people has been growing all around the world. A nation's main motive is to keep this disparity as minimal as possible. The disparity has been increased further by the COVID-19 pandemic, in which many people have lost their jobs or have had minimal means of earning an income. This study is very relevant in today's time as the comprehensive analysis will help understand the key attributes that play a role in improving the income levels of individuals. Economic equality helps to bring down poverty and also promotes economic growth.

Data Description

These are the following set of variables that are present in the dataset:

1. Age (numerical): the age of an individual
2. Workclass (categorical): employment status of an individual
3. Fnlwgt (numerical): This is the number of people census believes the entry represents
4. Education (categorical): the highest level of education of an individual
5. Education_num (numerical): number of years of education
6. Martial_status (categorical): marital status of an individual. The word Married-civ which is mentioned in this variable corresponds to a civilian spouse. Married-AF-Spouse corresponds to a spouse in the Armed Forces.
7. Occupation (categorical): the occupation of an individual
8. Relationship (categorical): represents what this individual is relative to others
9. Race (categorical)
10. Sex (categorical)
11. Capital_gain (numerical)
12. Capital_loss (numerical)

13. Hours_per_week (numerical): the number of hours an individual works per week
14. Native_country (categorical): country of origin for an individual
15. Income (categorical): whether or not an individual makes more than $50,000 annually.

There are 48842 records, out of which 32561 belong to the training dataset, and 16281 belong to the test dataset. The response variable is ‘income’, which indicates whether a person makes over 50k annually.

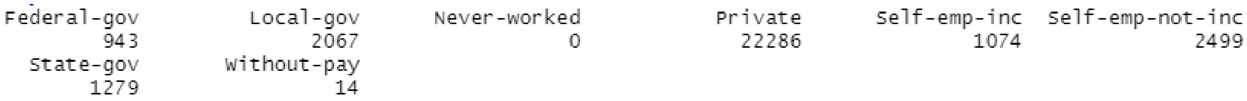
Part 1: Cleaning and Preprocessing

Training Data set: Cleaning missing values

The total number of records in the training dataset before conducting any preprocessing or transformation was 32561. The missing data were recognized by identifying the symbol ‘?’ and converting it into null values, then deleting those records from the data set. The variables in which this symbol was found are workclass, occupation, and native country. After removing missing data, the count of the records in the training data set was reduced to 30162 as there were 4262 missing records. The ‘fnlwgt’ variable was removed from the dataset as it is irrelevant in the analysis

Training Data set: Transformation

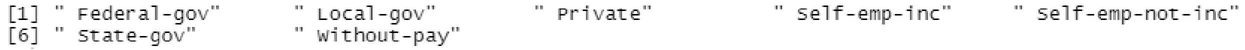
“workclass”:



```
Federal-gov        Local-gov     Never-worked          Private     Self-emp-inc  Self-emp-not-inc
        943             2067                0            22286             1074              2499
  State-gov      Without-pay
       1279               14
```

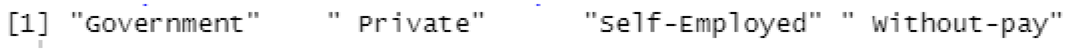
It is noted that there are no values under the level “Never-worked”. So, it was removed. Then groups were created to store similar levels under one group. For example, ‘Government’ was created to store ‘Federal-gov’, ‘Local-gov’, and ‘State-gov’. Another group named ‘Self-Employed’ was created to store ‘Self-emp-inc’, and ‘Self-emp-not-inc’.

Before the transformation:



```
[1] " Federal-gov"      " Local-gov"        " Private"          " Self-emp-inc"     " Self-emp-not-inc"
[6] " State-gov"        " Without-pay"
```

After the transformation:

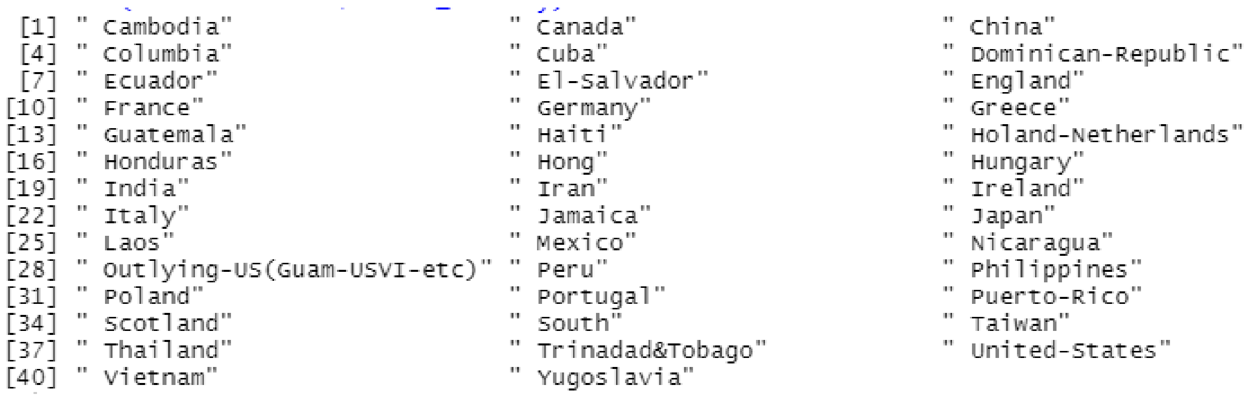


```
[1] "Government"    " Private"      "Self-Employed" " Without-pay"
```

“native_country”:

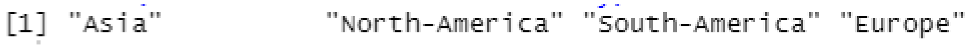
To simplify, the countries were put into categories according to their continent. The groups that were created were ‘Asia’, ‘South-America’, ‘North-America’, ‘Europe’.

Before the transformation:



```
 [1] " Cambodia"                   " Canada"                   " China"
 [4] " Columbia"                   " Cuba"                     " Dominican-Republic"
 [7] " Ecuador"                    " El-Salvador"              " England"
[10] " France"                     " Germany"                  " Greece"
[13] " Guatemala"                  " Haiti"                    " Holand-Netherlands"
[16] " Honduras"                   " Hong"                     " Hungary"
[19] " India"                      " Iran"                     " Ireland"
[22] " Italy"                      " Jamaica"                  " Japan"
[25] " Laos"                       " Mexico"                   " Nicaragua"
[28] " Outlying-US(Guam-USVI-etc)" " Peru"                     " Philippines"
[31] " Poland"                     " Portugal"                 " Puerto-Rico"
[34] " Scotland"                   " South"                    " Taiwan"
[37] " Thailand"                   " Trinadad&Tobago"          " United-States"
[40] " Vietnam"                    " Yugoslavia"
```

After the transformation:

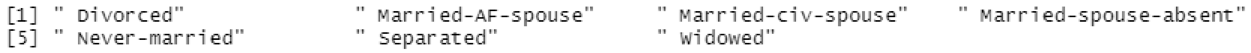


```
[1] "Asia"          "North-America" "South-America" "Europe"
```

<u>“Marital_status”</u>:

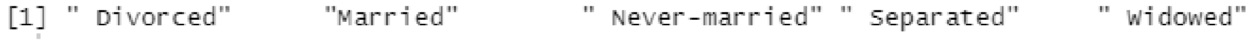
Group named “Married” was created under which the values such as “Married-AF-spouse”, “Married-civ-spouse”, and “Married-spouse-absent” were stored.

Before the transformation:



```
[1] " Divorced"              " Married-AF-spouse"     " Married-civ-spouse"    " Married-spouse-absent"
[5] " Never-married"         " Separated"             " Widowed"
```

After the transformation:

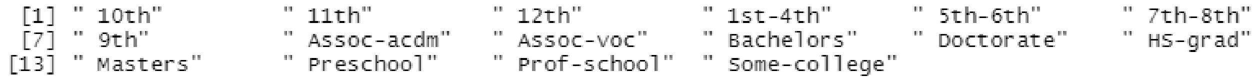


```
[1] " Divorced"      "Married"        " Never-married" " Separated"     " Widowed"
```

<u>“Education”</u>:

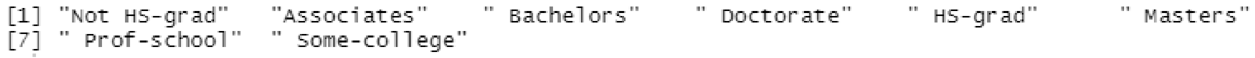
The groups such as “Not HS-grad” and “Associates” were created. Under “Not HS-grad” the values that related to education from Preschool to 12th grade were stored. Values such as “Assoc-acdm” and “Assoc-voc” were stored under ‘Associates’.

Before the transformation:



```
 [1] " 10th"         " 11th"         " 12th"         " 1st-4th"      " 5th-6th"      " 7th-8th"
 [7] " 9th"          " Assoc-acdm"   " Assoc-voc"    " Bachelors"    " Doctorate"    " HS-grad"
[13] " Masters"      " Preschool"    " Prof-school"  " Some-college"
```

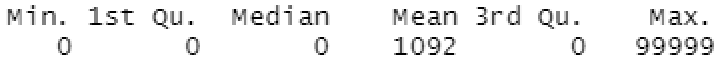
After the transformation:



```
[1] "Not HS-grad"   "Associates"    " Bachelors"    " Doctorate"    " HS-grad"      " Masters"
[7] " Prof-school"  " Some-college"
```

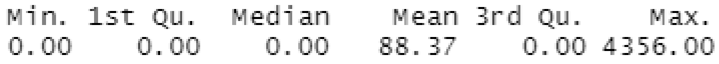
<u>“capital_gain” & “capital_loss”</u>:

Capital gain summary statistics:



```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0       0       0    1092       0   99999
```

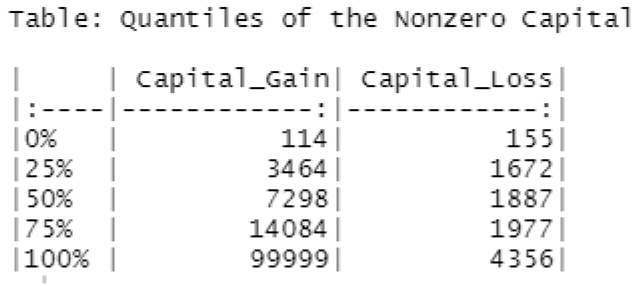
Capital loss summary statistics:



```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.00    0.00    0.00   88.37    0.00 4356.00
```

It can be seen that the median for capital gain and capital loss is 0. By digging deeper, it is found that 91.59% of the population don’t have any capital gain and 95.27% of the population have no capital loss.

To understand better, the values containing the zeroes have been excluded from both the capital gain and capital loss columns.



```
Table: Quantiles of the Nonzero Capital

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

Keeping the above observations in mind, the values of the variables “capital_loss” and “capital_gain” were grouped into categories and stored in the variables “cap_loss” and “cap_gain,” respectively. The groups for “cap_gain” were decided by marking all values of “capital_gain” which are less than the first quartile of the nonzero capital gain (which is equal to 3464) as “Low”; all values that are between the first and third quartile (between 3464 and 14080) - as “Medium”; and all values greater than or equal to the third quartile are marked as “High”. The groups for “cap_loss” were decided by marking all values of “capital_loss” which are less than the first quartile of the nonzero capital gain (which is equal to 1672) as “Low”; all values that are between the first and third quartile (between 1672 and 1977) - as “Medium”; and all values greater than or equal to the third quartile are marked as “High”.

Testing Data set: Cleaning missing values

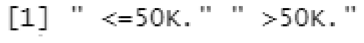
The values in the training dataset with the symbol ‘?’ were marked as null values. The number of records in the training data is 16281, and the missing values in the testing data set are 2203. After the removal of missing data, the number was reduced to 15060. The ‘fnlwgt’ variable was removed from the dataset as it is irrelevant in the analysis

Testing Dataset: Transformation

“Income”:

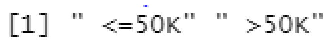
The values in the training dataset were like ‘<=50k’ and ‘>50k’. However, these values were like ‘<=50k.’ and ‘>50k’ in the testing dataset. As both of them differed, the format of the values present in the column “income” in the testing dataset was changed by removing the period to match with the training dataset.

Before the transformation:



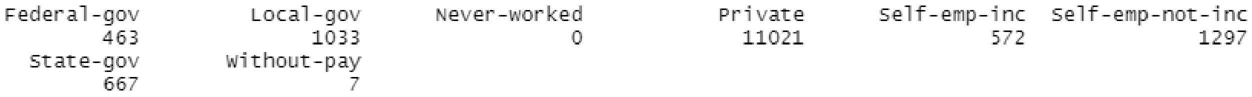
```
[1] " <=50K." " >50K."
```

After the transformation:



```
[1] " <=50K" " >50K"
```

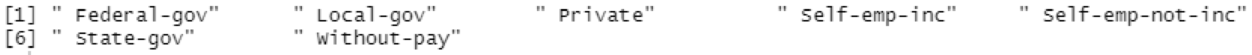
“workclass”:



```
 Federal-gov        Local-gov     Never-worked          Private     Self-emp-inc Self-emp-not-inc 
         463             1033                0            11021              572             1297 
   State-gov      Without-pay 
         667                7 
```

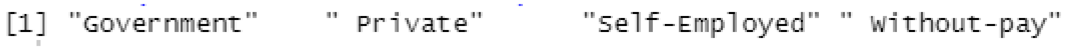
It is noted that there are no values under the level “Never-worked”. So, it was removed. Then groups were created to store similar levels under one group. For example, ‘Government’ was created to store ‘Federal-gov’, ‘Local-gov’, and ‘State-gov’. Another group named ‘Self-Employed’ was created to store ‘Self-emp-inc’, and ‘Self-emp-not-inc’.

Before the transformation:



```
[1] " Federal-gov"      " Local-gov"        " Private"          " Self-emp-inc"     " Self-emp-not-inc"
[6] " State-gov"        " Without-pay"     
```

After the transformation:

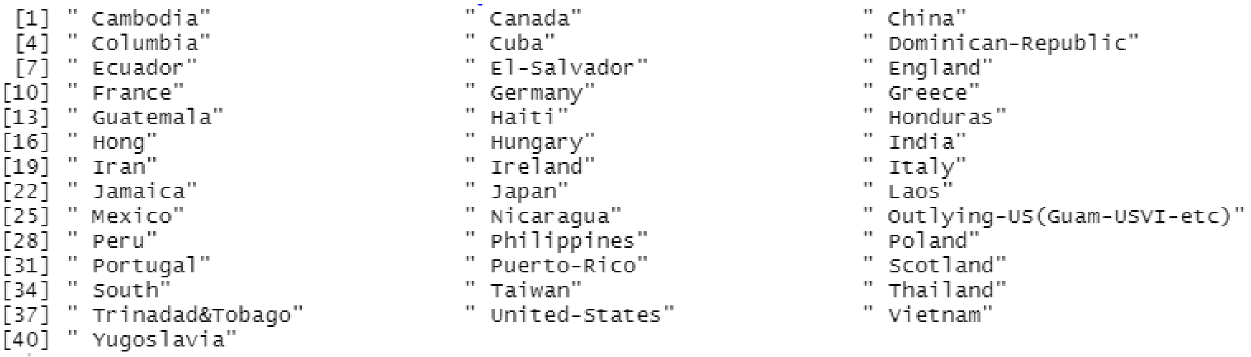


```
[1] "Government"    " Private"      "Self-Employed" " Without-pay" 
```

“native_country”:

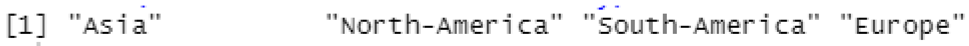
To simplify, the countries were put into categories according to their continent. The groups that were created were ‘Asia’, ‘South-America’, ‘North-America’, ‘Europe’.

Before the transformation:



```
 [1] " Cambodia"                   " Canada"                     " China"                     
 [4] " Columbia"                   " Cuba"                       " Dominican-Republic"        
 [7] " Ecuador"                    " El-Salvador"                " England"                   
[10] " France"                     " Germany"                    " Greece"                    
[13] " Guatemala"                  " Haiti"                      " Honduras"                  
[16] " Hong"                       " Hungary"                    " India"                     
[19] " Iran"                       " Ireland"                    " Italy"                     
[22] " Jamaica"                    " Japan"                      " Laos"                      
[25] " Mexico"                     " Nicaragua"                  " Outlying-US(Guam-USVI-etc)"
[28] " Peru"                       " Philippines"                " Poland"                    
[31] " Portugal"                   " Puerto-Rico"                " Scotland"                  
[34] " South"                      " Taiwan"                     " Thailand"                  
[37] " Trinadad&Tobago"            " United-States"              " Vietnam"                   
[40] " Yugoslavia"                
```

After the transformation:

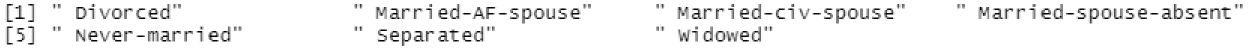


```
[1] "Asia"          "North-America" "South-America" "Europe"       
```

“Marital_status”:

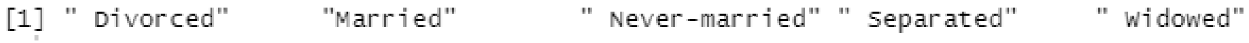
Group named “Married” was created under which the values such as “Married-AF-spouse”, “Married-civ-spouse”, and “Married-spouse-absent” were stored.

Before the transformation:



```
[1] " Divorced"              " Married-AF-spouse"     " Married-civ-spouse"    " Married-spouse-absent"
[5] " Never-married"         " Separated"             " Widowed"              
```

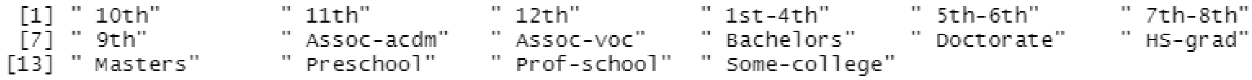
After the transformation:



```
[1] " Divorced"      "Married"        " Never-married" " Separated"     " Widowed"      
```

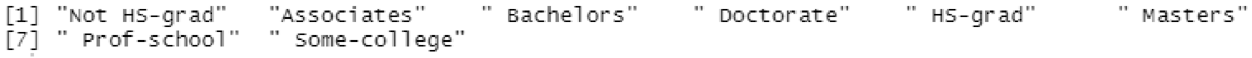
“Education”:

The groups such as “Not HS-grad” and “Associates” were created. Under “Not HS-grad” the values that related to education from Preschool to 12th grade were stored. Values such as “Assoc-acdm” and “Assoc-voc” were stored under ‘Associates’.
Before the transformation:



```
 [1] " 10th"         " 11th"         " 12th"         " 1st-4th"      " 5th-6th"      " 7th-8th"
 [7] " 9th"          " Assoc-acdm"   " Assoc-voc"    " Bachelors"    " Doctorate"    " HS-grad"
[13] " Masters"      " Preschool"    " Prof-school"  " Some-college"
```

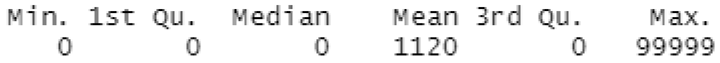
After the transformation:



```
[1] "Not HS-grad"   "Associates"    " Bachelors"    " Doctorate"    " HS-grad"      " Masters"
[7] " Prof-school"  " Some-college"
```

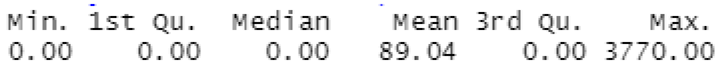
<u>“capital_gain” & “capital_loss”</u>:

Capital gain summary statistics:



```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0       0       0    1120       0   99999
```

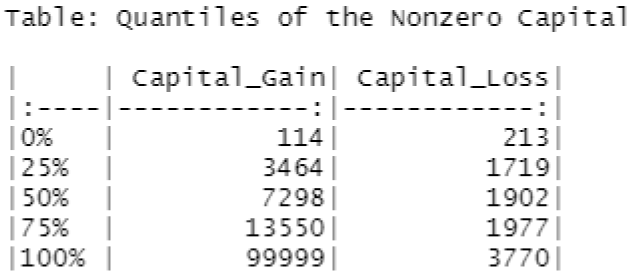
Capital loss summary statistics:



```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.00    0.00    0.00   89.04    0.00 3770.00
```

It can be seen that the median for capital gain and capital loss is 0. By digging deeper, it is found that 91.69% of the population don’t have any capital gain and 95.27% of the population have no capital loss.

To understand better, the values containing the zeroes have been excluded from both the capital gain and capital loss columns.



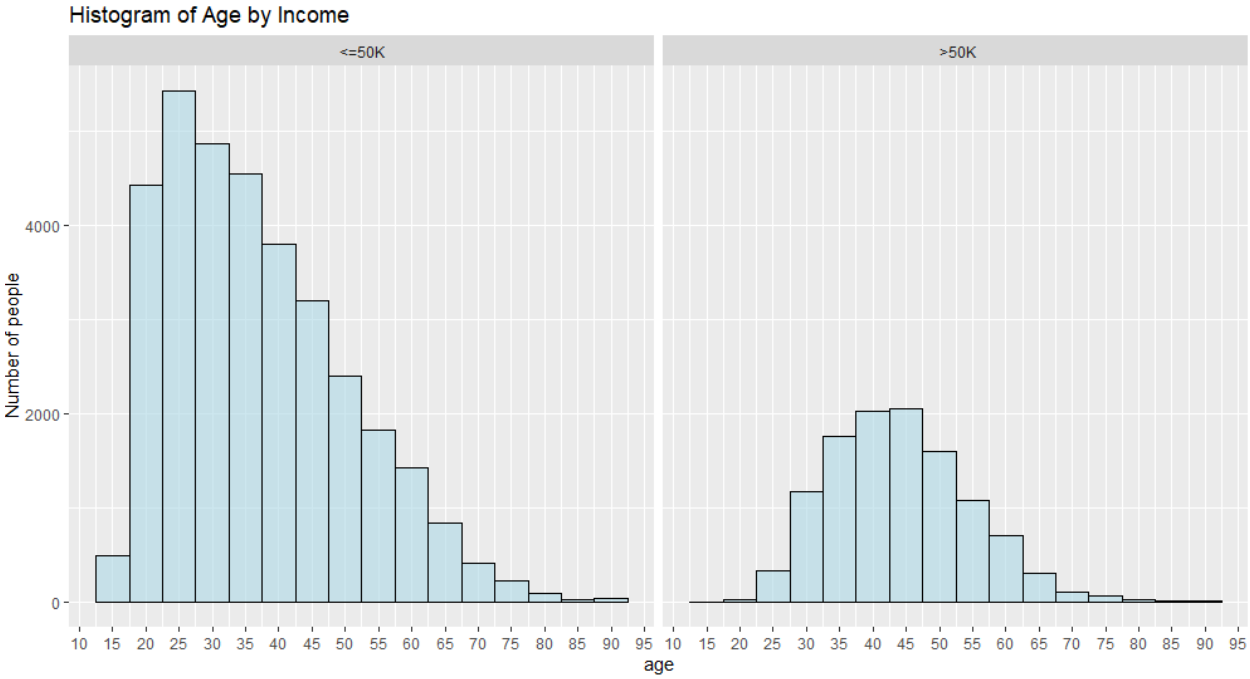
```
Table: Quantiles of the Nonzero Capital

|     | Capital_Gain| Capital_Loss|
|:----|------------:|------------:|
|0%   |          114|          213|
|25%  |         3464|         1719|
|50%  |         7298|         1902|
|75%  |        13550|         1977|
|100% |        99999|         3770|
```

Keeping the above observations in mind, the values of the variables “capital_loss” and “capital_gain” were grouped into categories and stored in the variables “cap_loss” and “cap_gain,” respectively. The groups for “cap_gain” were decided by marking all values of “capital_gain” which are less than the first quartile of the nonzero capital gain (which is equal to 3464) as “Low”; all values that are between the first and third quartile (between 3464 and 13550) - as “Medium”; and all values greater than or equal to the third quartile are marked as “High”. The groups for “cap_loss” were decided by marking all values of “capital_loss” which are less than the first quartile of the nonzero capital gain (which is equal to 1719) as “Low”; all values that are between the first and third quartile (between 1719 and 1977) - as “Medium”; and all values greater than or equal to the third quartile are marked as “High”.

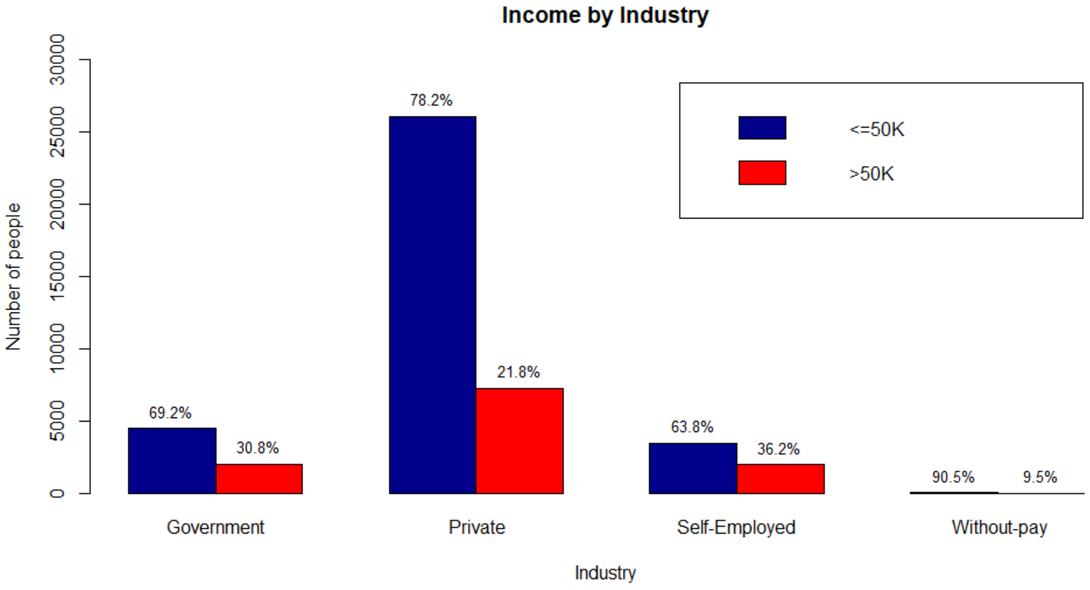
Part 2: Exploratory Data Analysis

“age”:



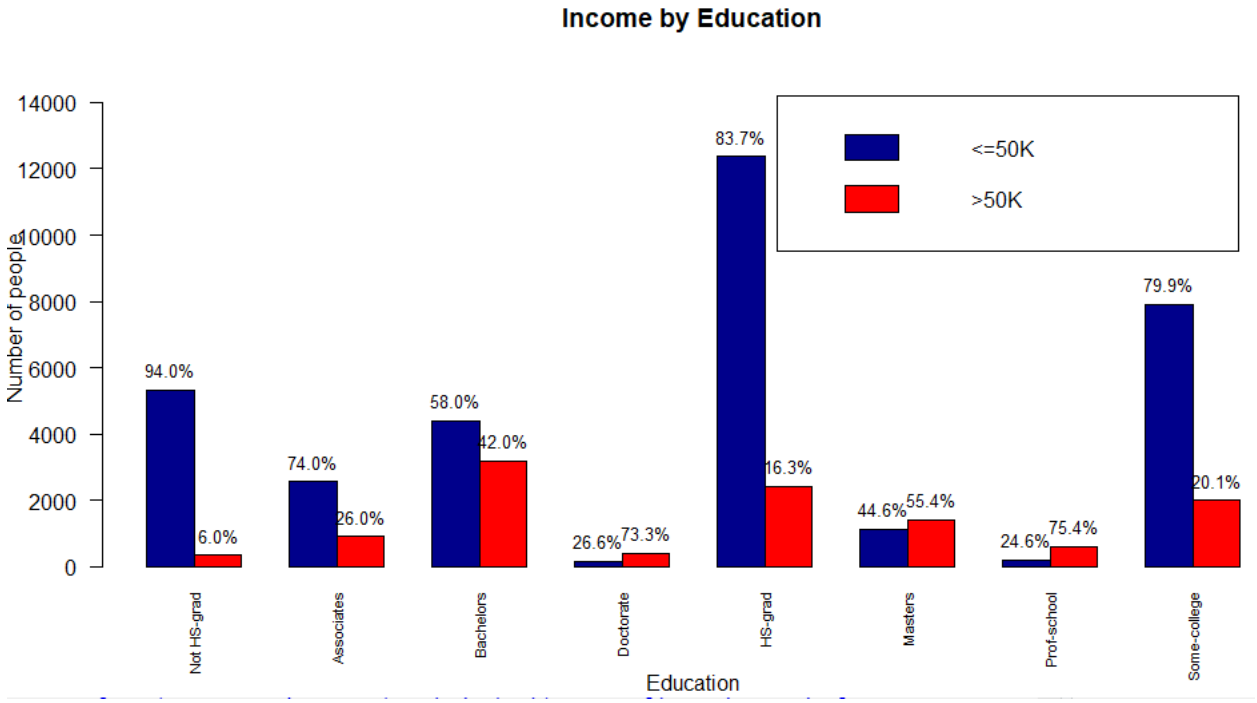
As it can be seen from one of the graphs, most of the people who fall into the category of having annual income less than or equal to 50 thousand dollars are the ones whose ages are from 20 to 30. Another graph shows that most of the people who fall into the category of having annual income greater than 50 thousand dollars are the ones whose ages are between 35 and 50.

“workclass”:



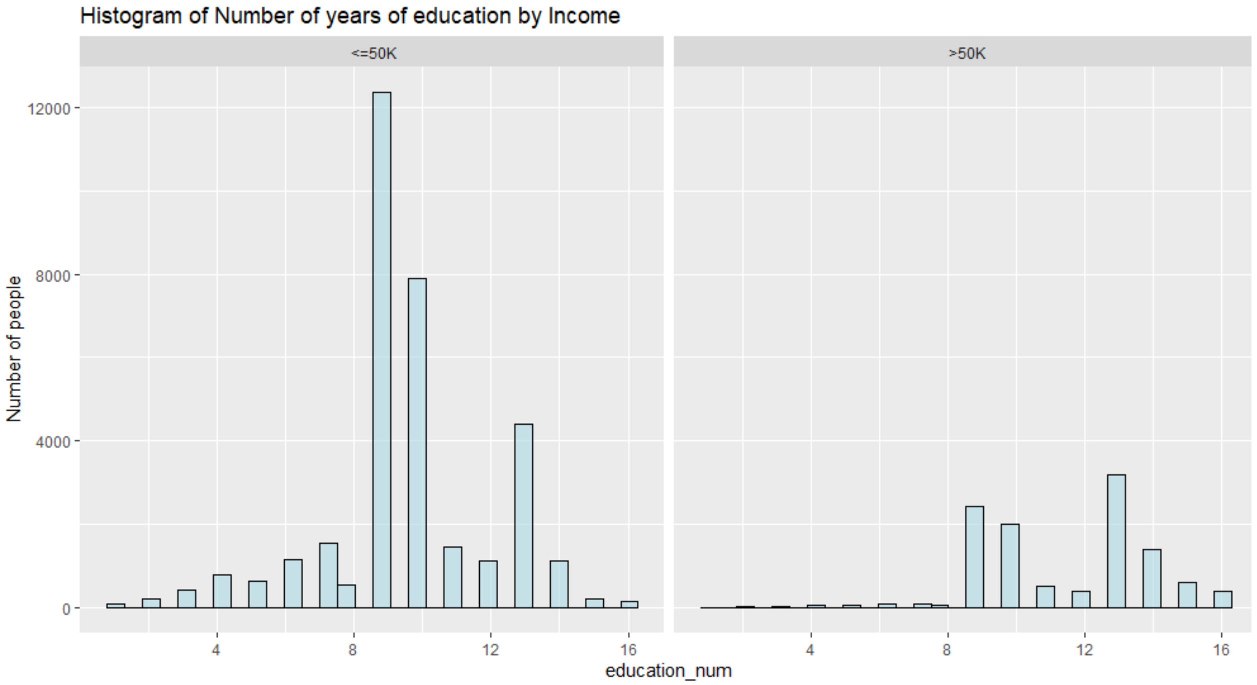
The graph shows that most people whose information is present in the census dataset work in the private industry. Furthermore, it can be observed that self-employed people have the highest income since 36.2% of the self-employed population have annual income higher than 50 thousand dollars, which is the highest among the people who work in other industries. Also, people who work without pay or work for a minimal amount are the ones that consist of the highest percentage of people having an annual income of fewer than 50 thousand dollars.

<u>“education”</u>:



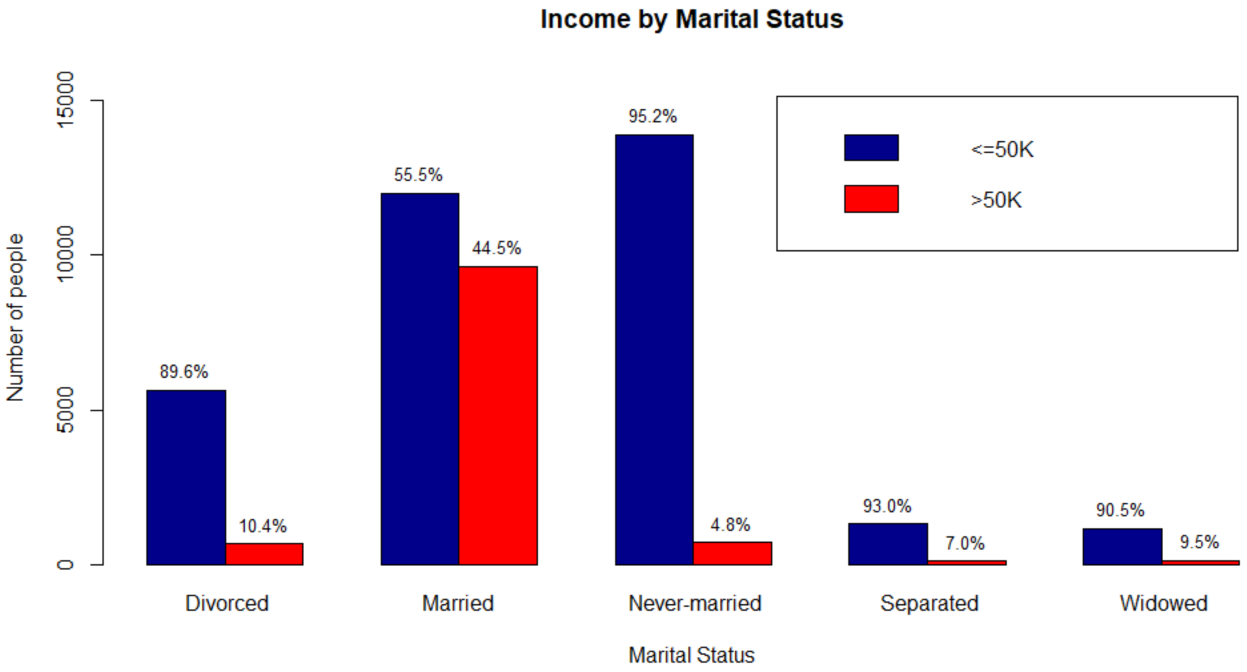
The graph shows that most people in the census data are high school graduates. Also, it can be observed that people who did not graduate high school are the ones with the highest number of people with 50 thousand dollars or less since they consist of 94%, which is the highest among the other groups. Moreover, it can be observed that people who attended professional school are the ones who have annual income higher than 50 thousand dollars since 75.4% of the people from the prof school have an annual income of more than 50 thousand dollars.

<u>“education_num”</u>:



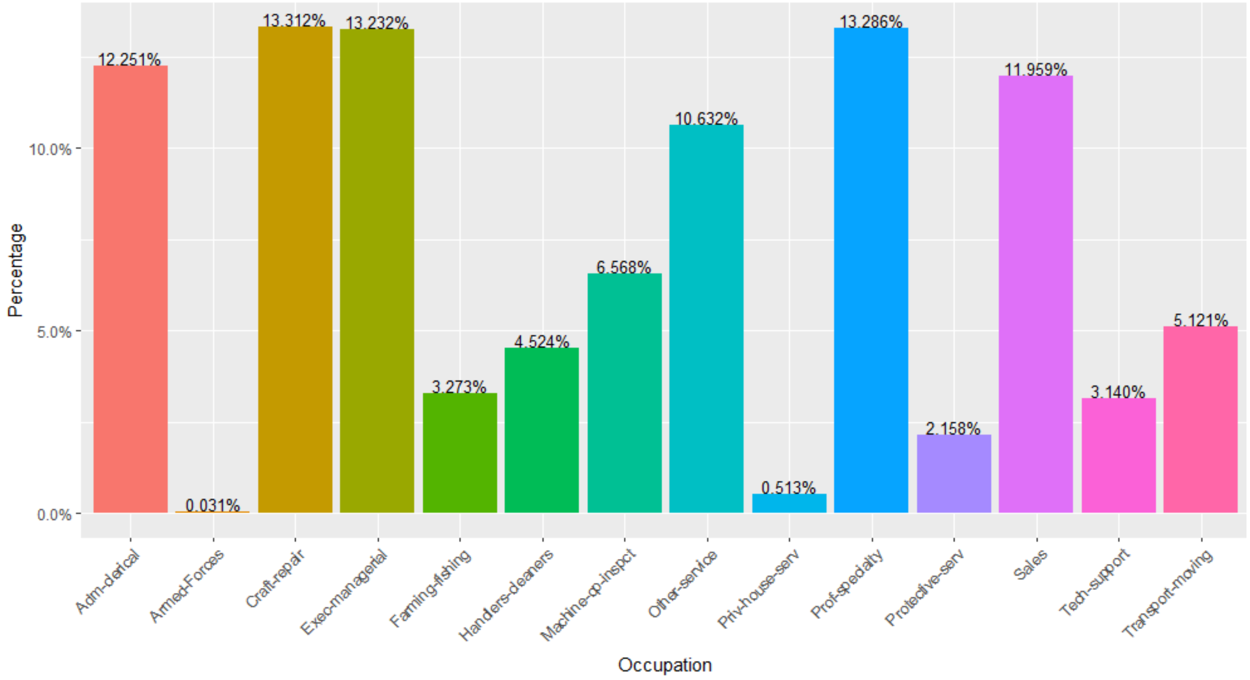
From one of the graphs, it is observed that most of the people who have annual income less than or equal to 50 thousand dollars are the ones who have 9 years of education. From the other graph, it is observed that most of the people who have an annual income of more than 50 thousand dollars are the ones who have 13 years of education.

<u>"marital_status"</u>:

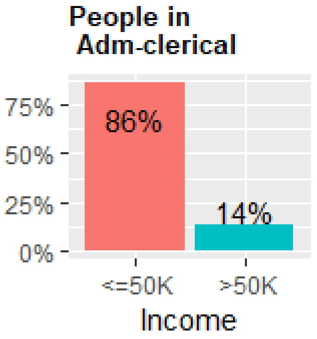


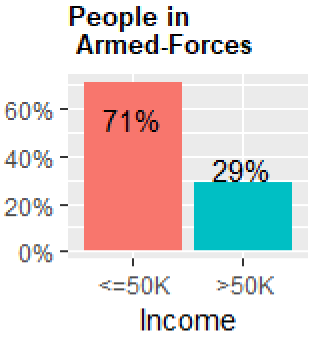
It is observed that the highest number of people from the census data are married. Also, most people who have annual income higher than 50 thousand dollars are married since 44.5% of the married population have annual income higher than 50 thousand dollars, which is the highest among the people with another marital status. Also, most people who have annual income less than or equal to 50 thousand dollars have never married since 95.2% of the never-married population have annual income less than or equal to 50 thousand dollars which is the highest among other marital statuses.

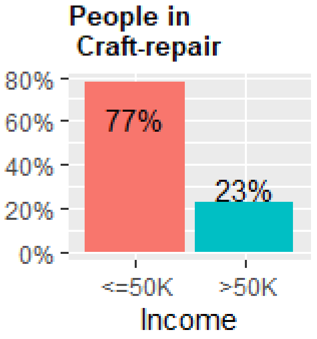
"occupation":

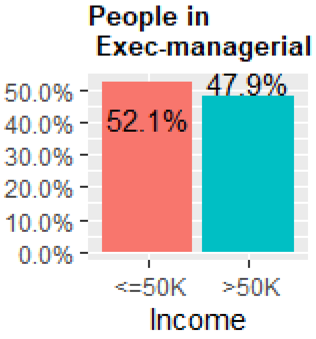


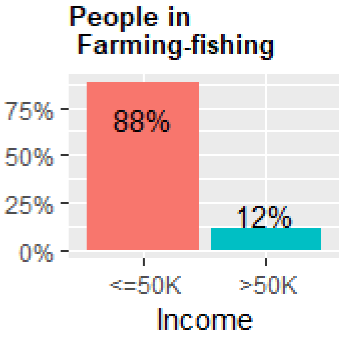
The graph shows the most of the people have an occupation as prof-specialty since the percentage is the highest among the other occupations.

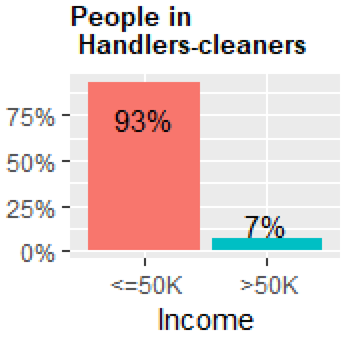

People in
Adm-clerical
75%
50%
25%
0%
86%
14%
<=50K
>50K
Income

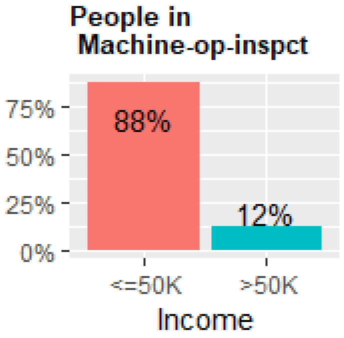

People in
Armed-Forces
60%
40%
20%
0%
71%
29%
<=50K
>50K
Income

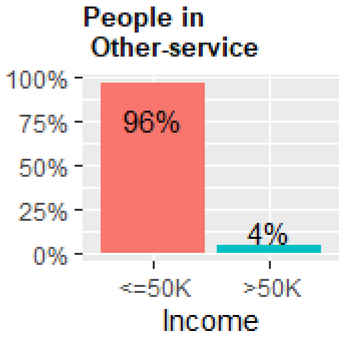

People in
Craft-repair
80%
60%
40%
20%
0%
77%
23%
<=50K
>50K
Income

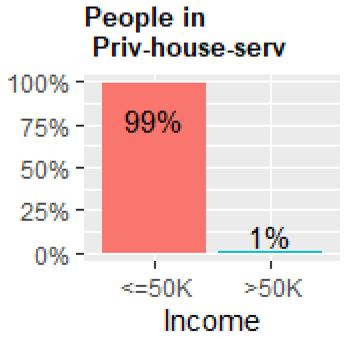

People in
Exec-managerial
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
52.1%
47.9%
<=50K
>50K
Income

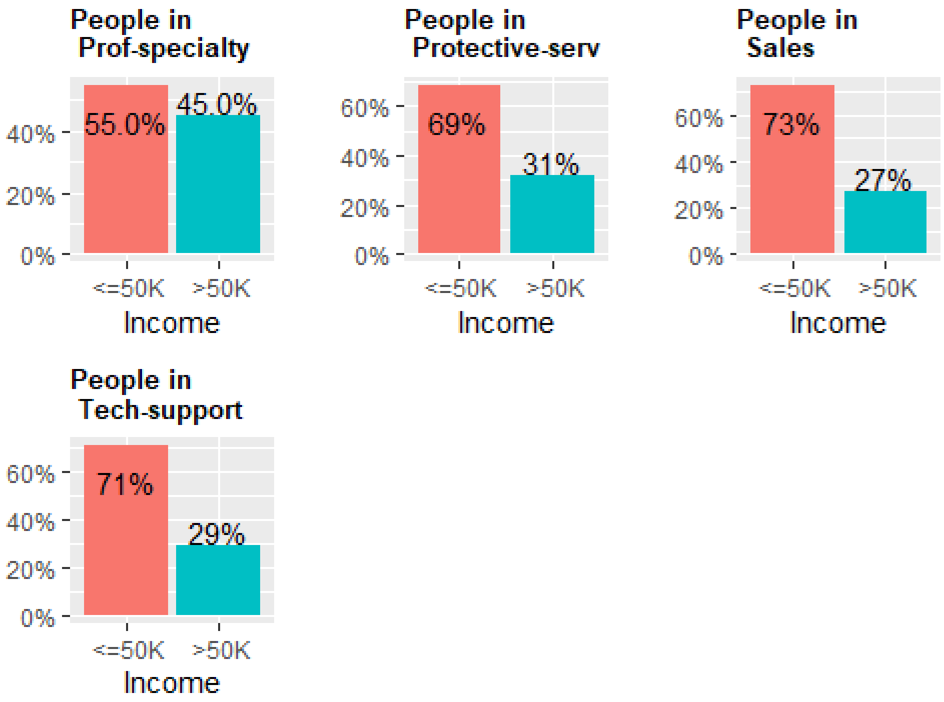

People in
Farming-fishing
75%
50%
25%
0%
88%
12%
<=50K
>50K
Income


People in
Handlers-cleaners
75%
50%
25%
0%
93%
7%
<=50K
>50K
Income


People in
Machine-op-inspct
75%
50%
25%
0%
88%
12%
<=50K
>50K
Income

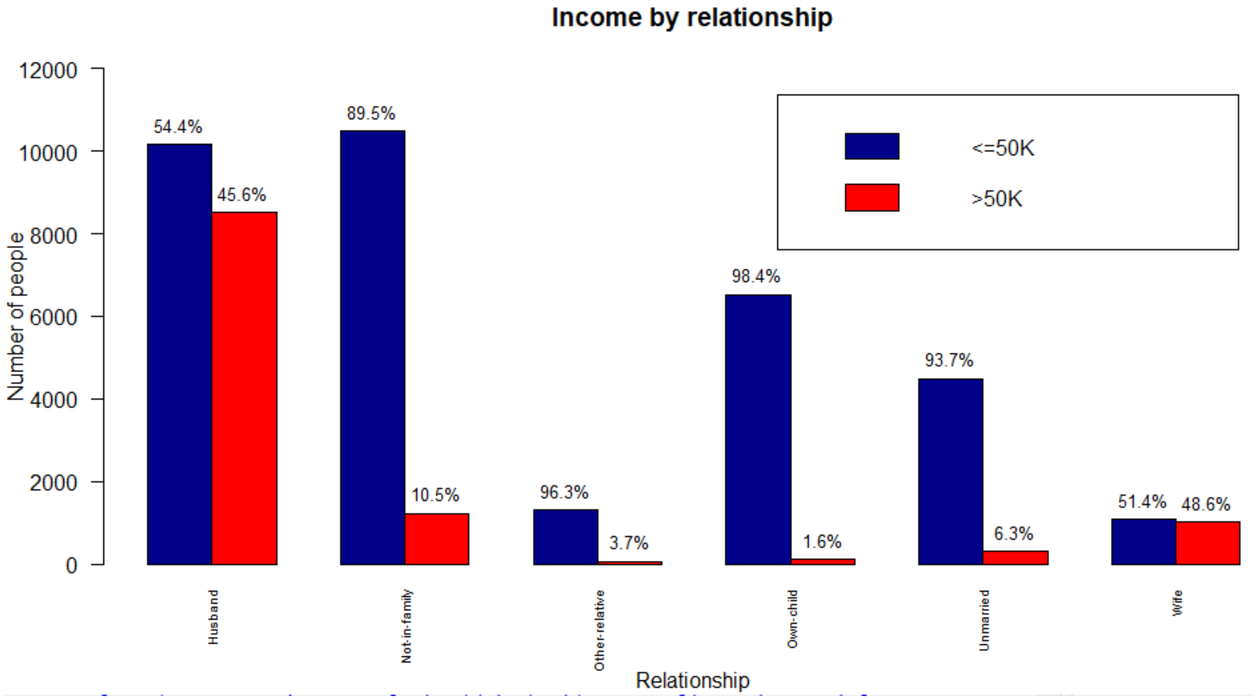

People in
Other-service
100%
75%
50%
25%
0%
96%
4%
<=50K
>50K
Income


People in
Priv-house-serv
100%
75%
50%
25%
0%
99%
1%
<=50K
>50K
Income



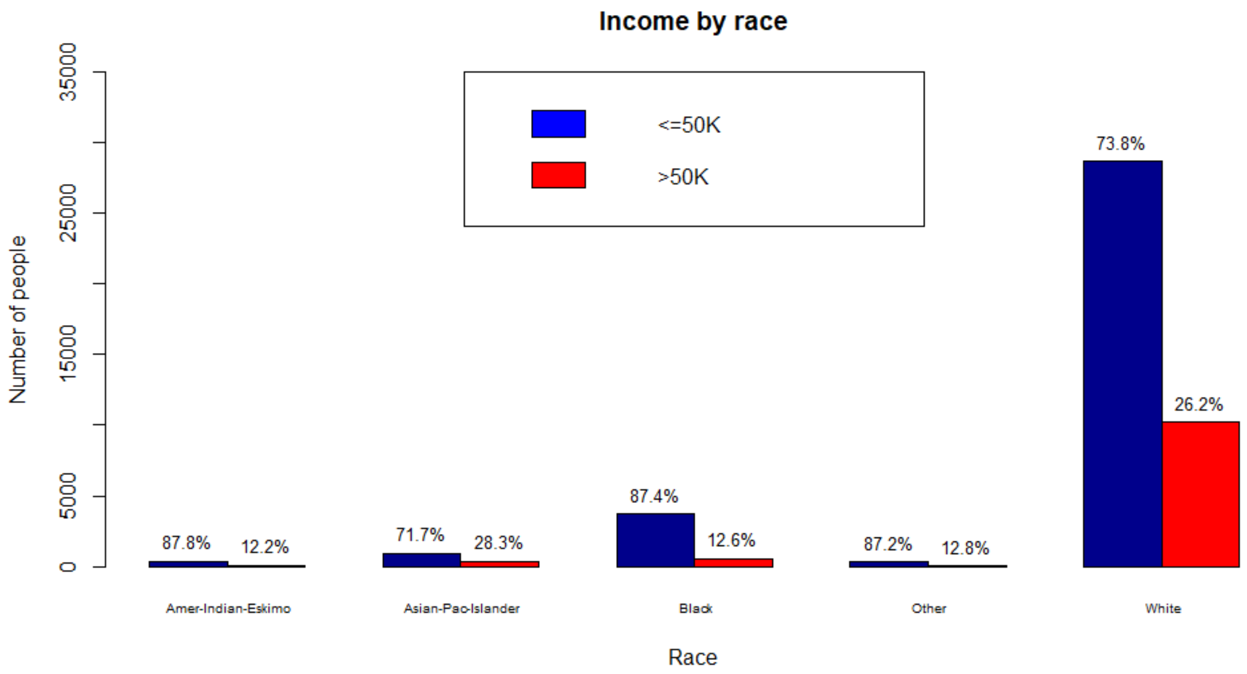
These graphs show that most of the people who have annual income less than or equal to 50 thousand dollars have an occupation as “priv-house-ser” since 99% of the population who has this occupation have annual income less than or equal to 50 thousand dollars which is the highest percentage among other occupations. Also, it shows that most people who have an annual income of more than 50 thousand dollars have an occupation as “prof-specialty” since 45% of the population who has this occupation have an annual income of more than 50 thousand dollars which is the highest percentage among other occupations.

“relationship”:



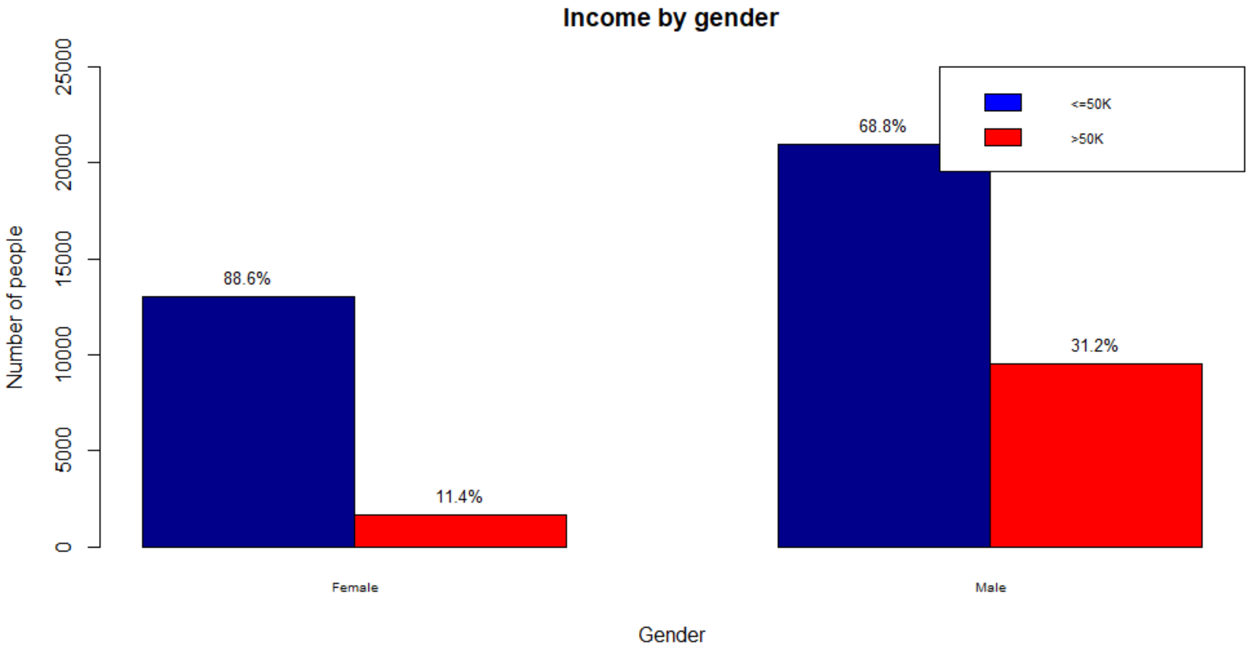
This graph shows that most of the people who have annual income less than or equal to 50 thousand dollars belong from “own-child” since 98.4% of the population from “own-child” have annual income less than or equal to 50 thousand dollars is the highest percentage among other relationships. Also, it is observed that most of the people who have an annual income of more than 50 thousand dollars are wives since 48.6% of the population from this category have an annual income of more than 50 thousand dollars which is the highest among other relationships.

<u>“race”</u>:



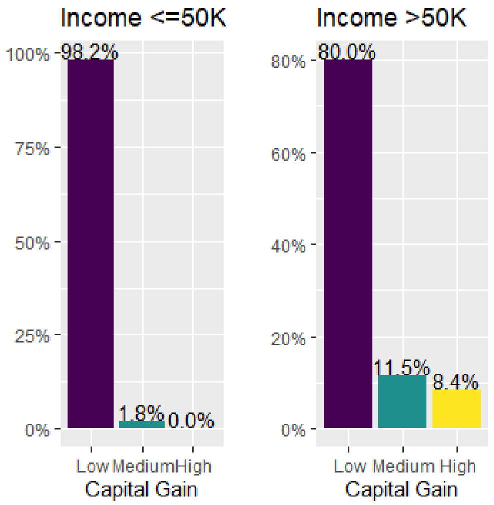
This graph shows that most of the participants in the census data are white. Also, it is observed that people who have annual income less than or equal to 50 thousand dollars belong from "black" as 87.4% of the population from this category have annual income less than or equal to 50 thousand dollars which is the highest percentage among other races. Also, it is observed that most people who have an annual income of more than 50 thousand belong to the "Asian-Pac-Islander" race since 28.3% of the population from this race have an annual income greater than 50 thousand dollars which is the highest percentage among other races.

"sex":



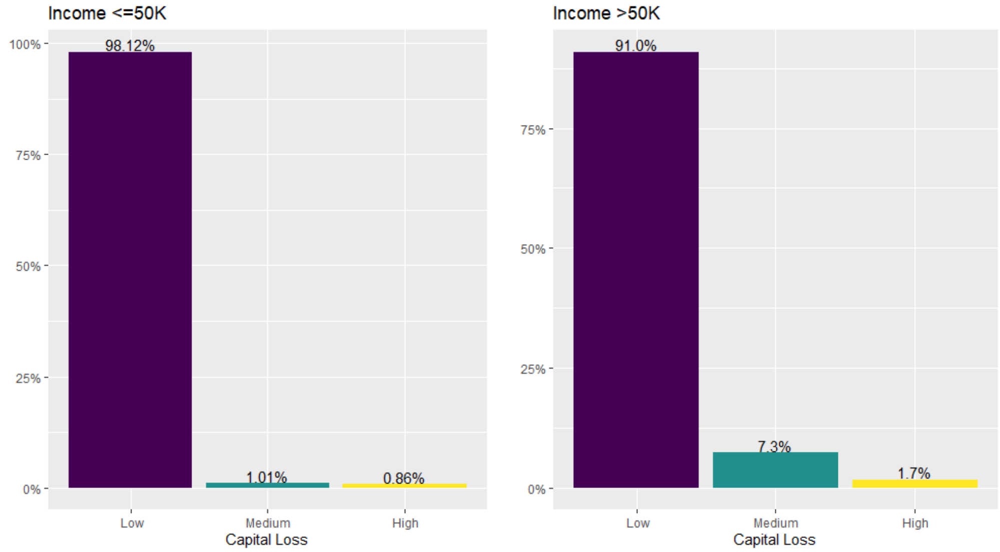
This graph shows that most of the participants in the census data are male. Also, it is observed that people who have annual income less than or equal to 50 thousand dollars belong are female as 88.6% of the population from this sex have annual income less than or equal to 50 thousand dollars which is the highest percentage among other genders. Also, it is observed that most people who have an annual income of more than 50 thousand are males since 31.2% of the population from this sex have an annual income greater than 50 thousand dollars which is the highest percentage among other genders.

“cap_gain”:



The first graph shows that 98.2% of the people who have annual income less than or equal to 50 thousand dollars have low capital gain, and rest of the 1.8% of these people have medium capital gain.The second graph shows that 80% of the people who have annual income more than 50 thousand dollars have low capital gain, 11.5% of these people have medium capital gain, and 8.4% of these individuals have high capital gain.

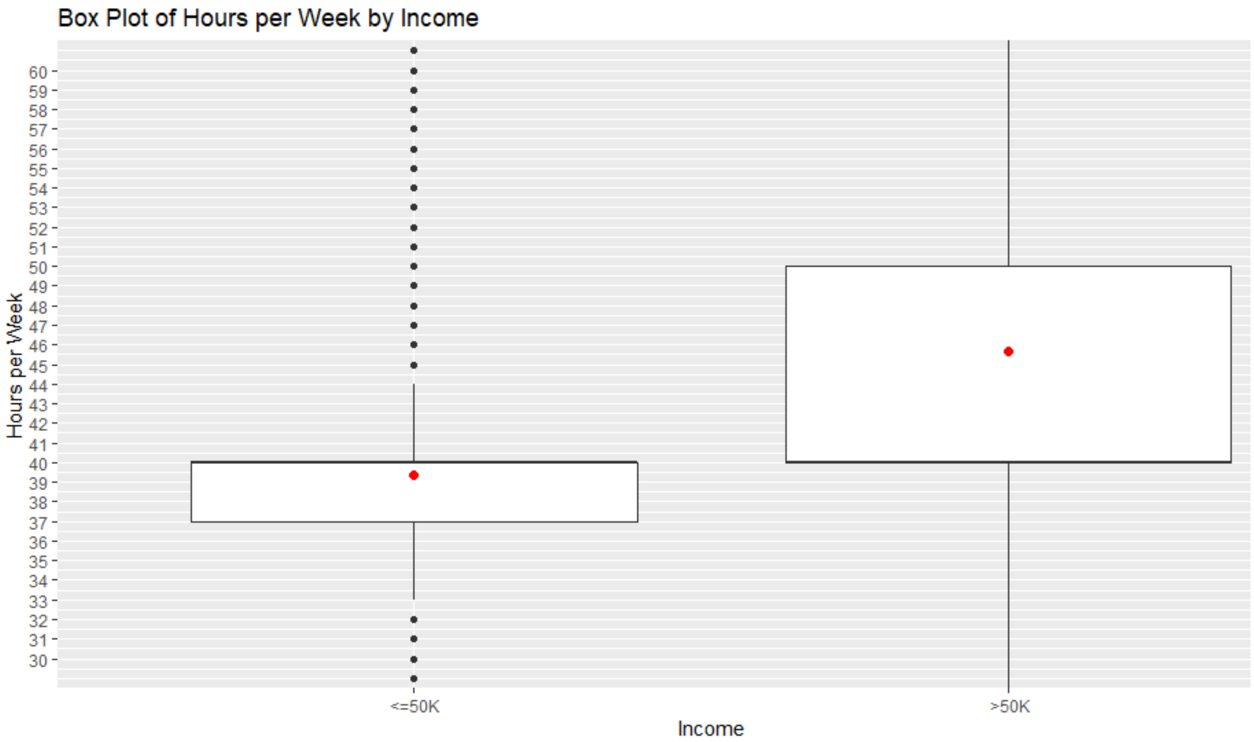
“cap_loss”:



The first graph shows the 98.12% of the people who have annual income less than or equal to 50 thousand dollars have low capital loss, 1.01% of these individuals have medium capital loss, and 0.86% of these people have high capital loss.
The second graph shows that 91% of the people who have annual income more than 50 thousand dollars have low capital loss, 7.3% of these individuals have medium capital loss, and 1.7% of the individuals have high capital loss.

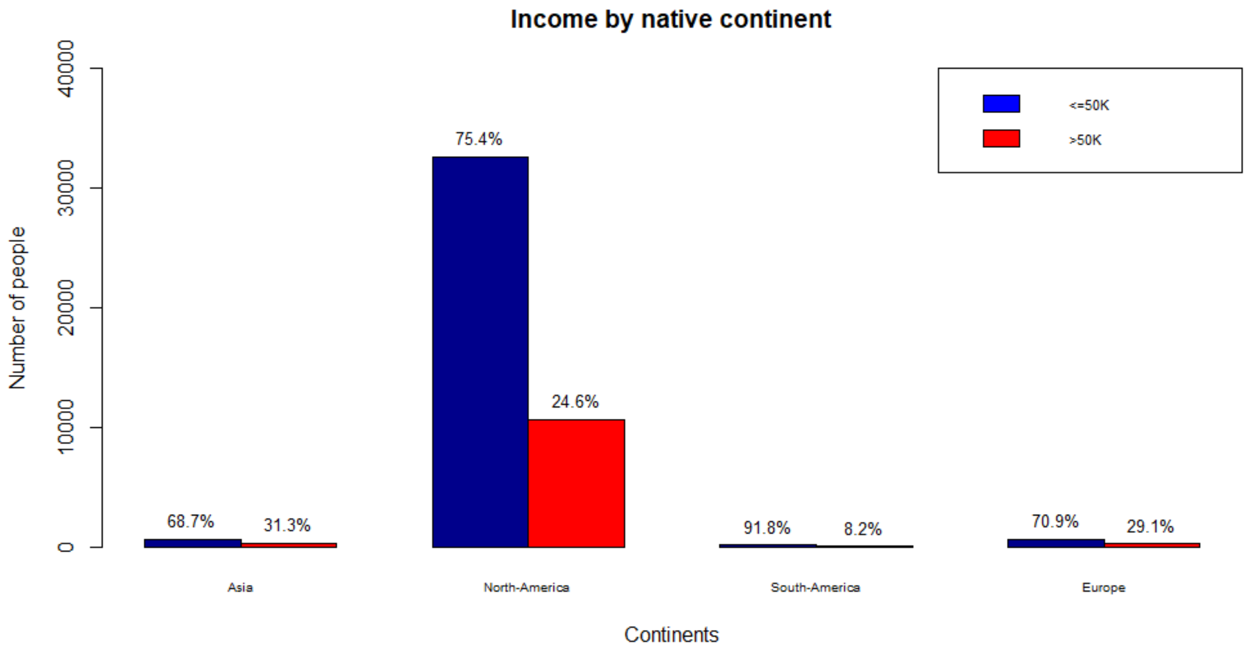
Overall, looking at the graphs of capital gain and capital loss, it can be inferred that most of the people do not have neither capital gain nor any capital loss.

“hours_per_week”:



From the boxplot is observed that the mean value is approximately 39 hours per week for annual income less than or equal to 50 thousand dollars. 46 hours per week is the mean value for an annual income of more than 50 thousand dollars.

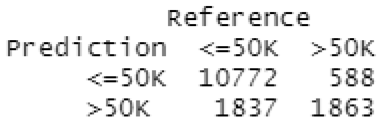
"native_country":



This graph shows that most of the participants in the census data are from North America. Also, it is observed that people who have annual income less than or equal to 50 thousand dollars are people from South America as 91.8% of the population from this continent have annual income less than or equal to 50 thousand dollars which is the highest percentage among other continents. Also, it is observed that most people who have an annual income of more than 50 thousand are people from Asia since 31.3% of the population from this continent have an annual income greater than 50 thousand dollars which is the highest percentage among other regions.

Part 3: Predictive Analysis

Decision tree:

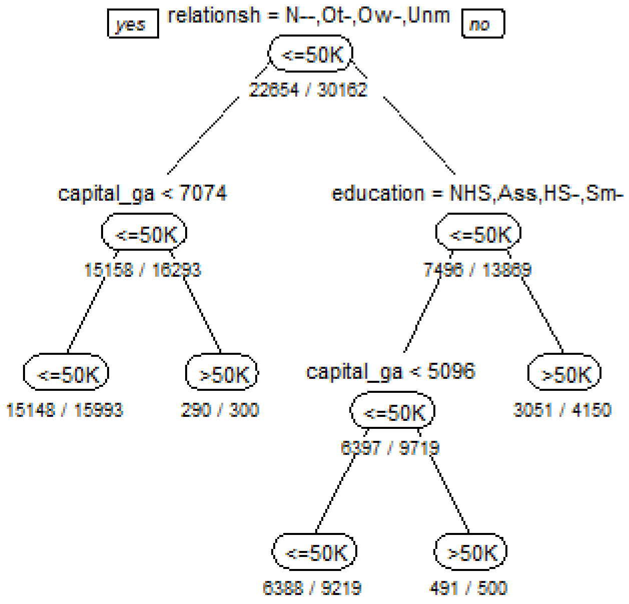


```
             Reference
Prediction  <=50K  >50K
     <=50K  10772   588
     >50K    1837  1863
```

Total: 10772+588+1837+1863 = 15060

Accuracy: (10772+1863)/15060 = 83.9%

Not Accurate: (1837+588)/15060 = 16.10%



The number of leaves in the decision tree: 5

Left part of the decision tree:

1. Relationship: “not-in family”, “other-relative”, “own-child”, and “married”
   Capital gain: < 7074
   Income: <= 50k
   Number of records: 15148
2. Relationship: “not-in family”, “other-relative”, “own-child”, and “married”
   Capital gain: >= 7074
   Income: >50k
   Number of records: 290

The right part of the decision tree:

1. Relationship: “husband” and “wife”
   Education: “Bachelors”, “Doctorate”, “Masters”, “Prof-school”

   Income: >50k
   Number of records: 3051
2. Relationship: “husband” and “wife”
   Education: “Not HS-grad”, “HS-grad”, “Some-college”, and “Associates”
   Capital gain: <5096
   Income: <=50k
   Number of records: 6388
3. Relationship: “husband” and “wife”
   Capital gain: >= 5096
   Income: >50k
   Number of records: 491

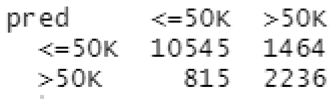
<u>Logistic Regression</u>

Backward:
income ~ age + workclass + education + education_num + marital_status +
  occupation + relationship + race + sex + capital_gain + capital_loss +
  hours_per_week + native_country + cap_gain + cap_loss

Forwards:
income ~ relationship + education + capital_gain + occupation +
  cap_loss + hours_per_week + age + sex + marital_status +
  workclass + education_num + capital_loss + race + native_country +
  cap_gain

Stepwise:
income ~ relationship + education + capital_gain + occupation +
  cap_loss + hours_per_week + age + sex + marital_status +
  workclass + education_num + capital_loss + race + native_country +
  cap_gain

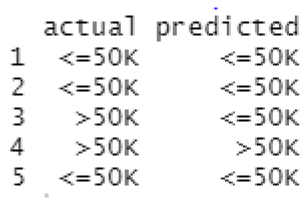
As it can be seen that no variables were eliminated with backward, forwards, and stepwise algorithms.



```
pred     <=50K  >50K
  <=50K  10545  1464
  >50K     815  2236
```

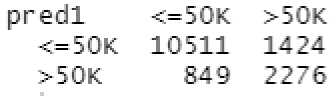
Total: 10545+1464+815+2236 = 15060
Accuracy: (10545+2236)/15060 = 84.87%
Misclassification: (1464+815)/15060 = 15.13%

First 5 records:



```
  actual predicted
1  <=50K     <=50K
2  <=50K     <=50K
3   >50K     <=50K
4   >50K      >50K
5  <=50K     <=50K
```

<u>Neural Network</u>



```
pred1    <=50K  >50K
  <=50K  10511  1424
  >50K     849  2276
```

Total: 10440+1387+920+2313 = 15060
Accuracy: (10440+2313)/15060 = 84.9%
Not accurate: (797+1416)/15060 = 15.1%

Conclusion/Findings

It was found that all of the variables provided in the dataset are strongly related to the response variable, income, except for the variable ‘fnlwgt’, which was not used in the analysis. Some of the findings were found for people who have a higher chance of fitting into a higher income bracket. For example, people with a higher chance of having an annual income of more than 50 thousand dollars had an age between 35 and 50, they were self-employed, attended professional school, had 13 years of education, were married, had an occupation as ‘prof-specialty’, were “Asian-Pac-Islander”, were male, worked for an average of 46 hours per week, and were from Asia.

From the perspective of predictive models, the neural network is the best as it has the highest accuracy rate with 84.9%, then comes the logistic regression with 84.87%, and then comes the decision tree with an 83.9% accuracy.

References

https://archive.ics.uci.edu/ml/datasets/adult