Predicting Income Level Using the Adult Census Income Data Set

N Ray

2019/12/11

1. INTRODUCTION

The *Adult Census Income* data set contains approximately 32,500 records of the income levels and 14 socio-economic factors, such as race, education, age, etc., of census correspondents in 1994. The purpose of the analysis is to determine whether a group of correspondents earn an annual salary of either less or greater then $50,000.

This report analyses the data set using four different machine learning algorithms. The first is the *Generalised Linear Mode (GLM)* which fits a linear regression model to the data. The second algorithm attempts to improve upon the accuracy of the first by using *K Nearest Neighbours (KNN)*. The third algorithm is *Classification and Regression Trees (CART)* and the fourth is *Random Forests*.

After having applied the data set to the four different algorithms, the final results show that only *Random Forest* improved upon the accuracy of the linear regression algorithm.

The URL for *Adult Census Income* is
https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data

The git-hub directory of the files for this project is
https://git hub.com/yu138538/Adult-Income-Census-Project

2. ANALYSIS

2.1 The *Census Adult Income* Data set

The *Census Adult Income* data set consists of 15 fields and approximately 32,500 records.

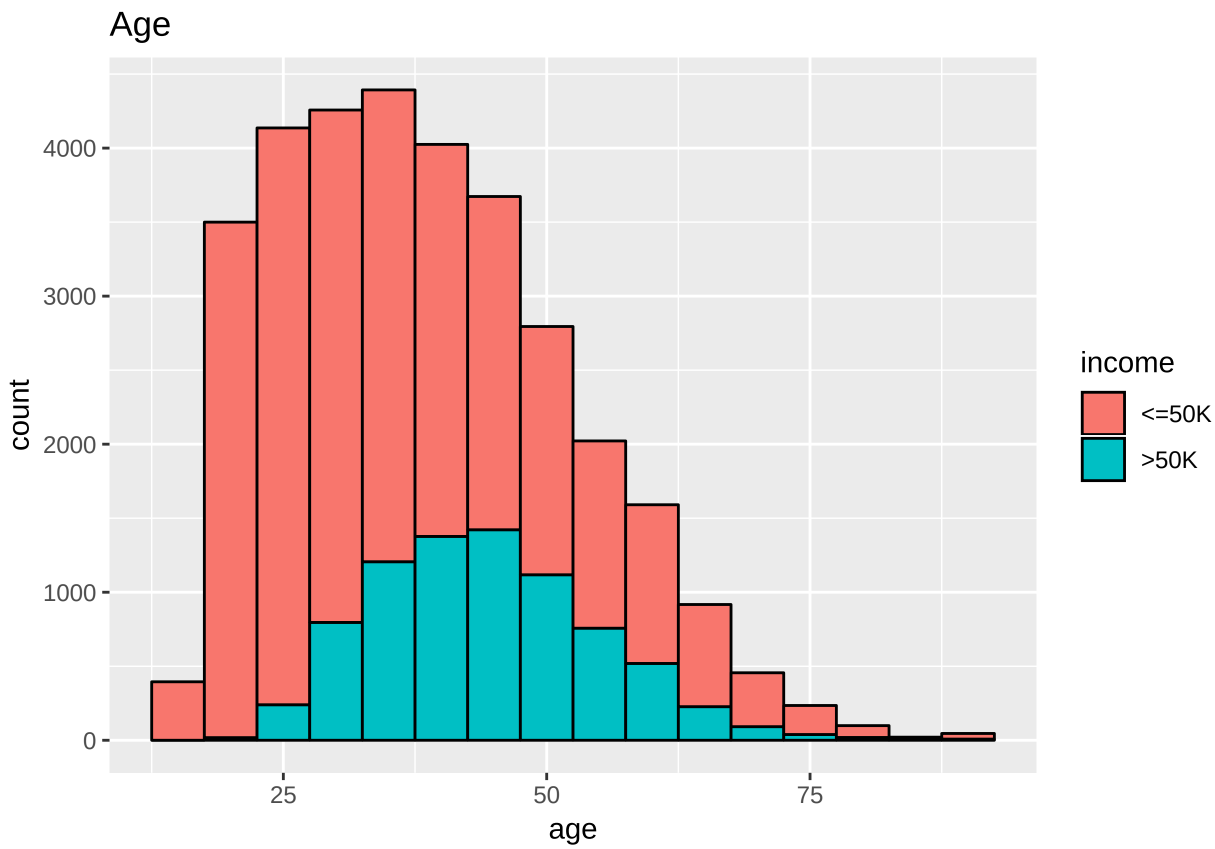
```
## Observations: 32,561
## Variables: 15
## $ age            <int> 39, 50, 38, 53, 28, 37, 49, 52, 31, 42, 37, 30,...
## $ workclass      <fct> State-gov, Self-emp-not-inc, Private, Private, ...
## $ fnlwgt         <int> 77516, 83311, 215646, 234721, 338409, 284582, 1...
## $ education      <fct> Bachelors, Bachelors, HS-grad, 11th, Bachelors,...
## $ education_num  <int> 13, 13, 9, 7, 13, 14, 5, 9, 14, 13, 10, 13, 13,...
## $ marital_status <fct> Never-married, Married-civ-spouse, Divorced, Ma...
## $ occupation     <fct> Adm-clerical, Exec-managerial, Handlers-cleaner...
## $ relationship   <fct> Not-in-family, Husband, Not-in-family, Husband,...
## $ race           <fct> White, White, White, Black, Black, White, Black...
## $ sex            <fct> Male, Male, Male, Male, Female, Female, Female,...
## $ capital_gain   <int> 2174, 0, 0, 0, 0, 0, 0, 0, 14084, 5178, 0, 0, 0...
## $ capital_loss   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ hours_per_week <int> 40, 13, 40, 40, 40, 40, 16, 45, 50, 40, 80, 40,...
## $ native_country <fct> United-States, United-States, United-States, Un...
## $ income         <fct> <=50K, <=50K, <=50K, <=50K, <=50K, <=50K, <=50K...
```

2.1.1 The *Census Adult Income* Summary of Fields

Below is a summary of the fields, or attributes, of the data set .

Age

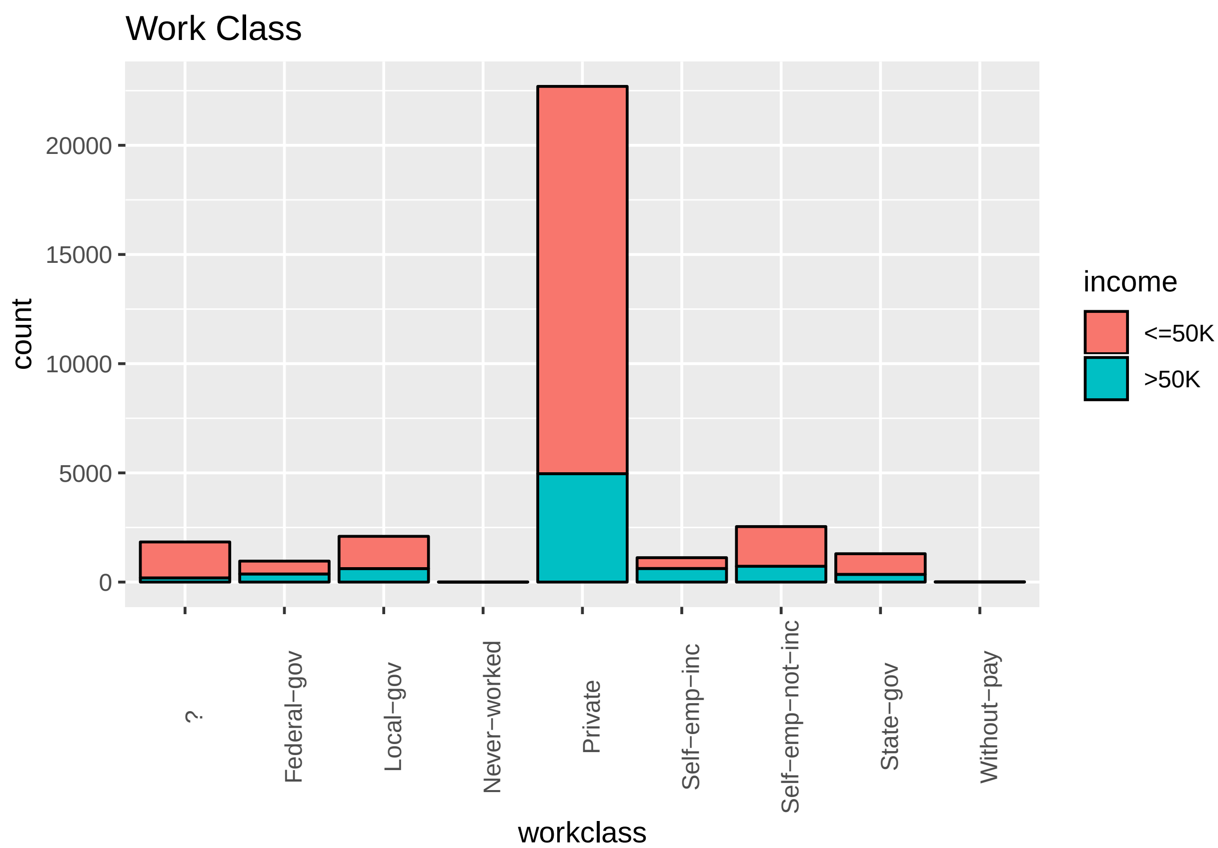
The age of the census correspondents.




```
## # A tibble: 73 x 3
##      age     n      Pct
##    <int> <int>    <dbl>
##  1    36   898 0.027579
##  2    31   888 0.027272
##  3    34   886 0.027210
##  4    23   877 0.026934
##  5    35   876 0.026903
##  6    33   875 0.026873
##  7    28   867 0.026627
##  8    30   861 0.026443
##  9    37   858 0.026351
## 10    25   841 0.025828
## # ... with 63 more rows
```

Workclass

The nature of the employment status or sector of correspondents.



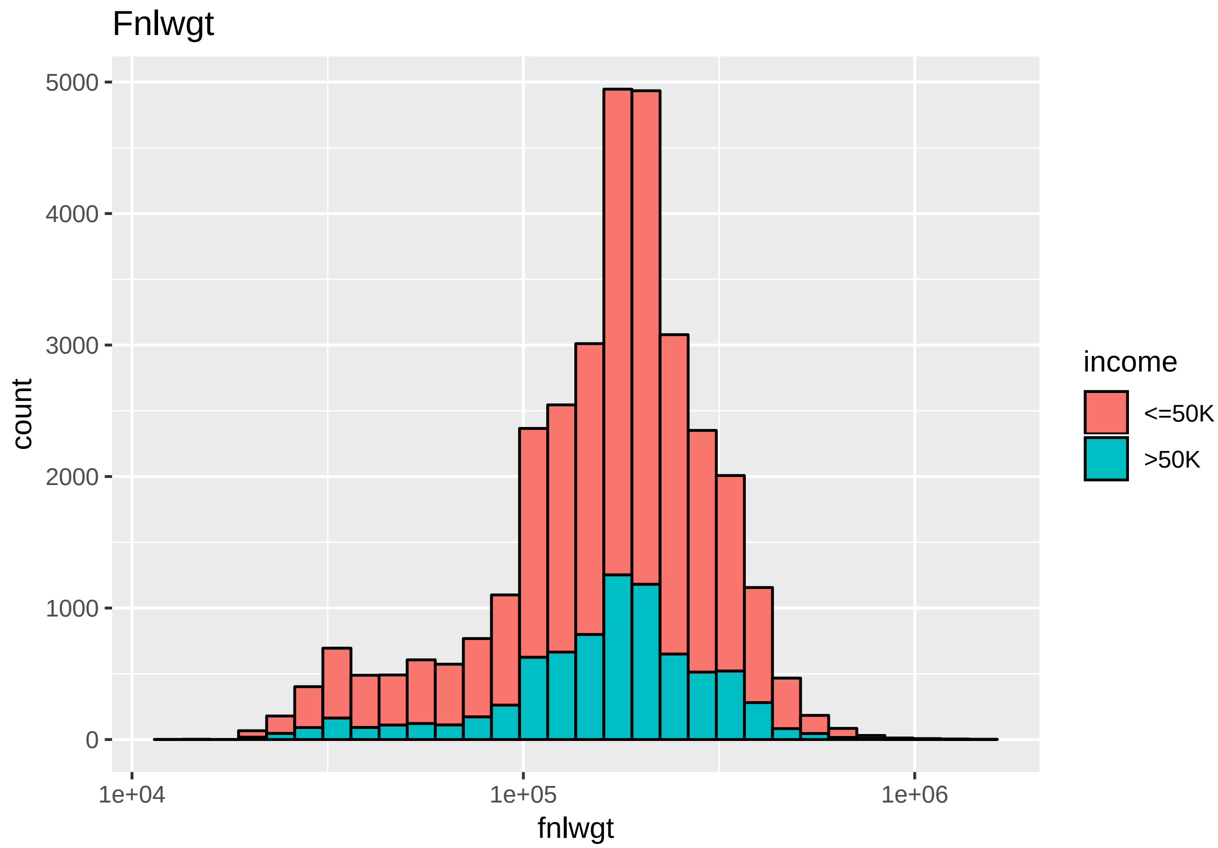

```
## # A tibble: 9 x 3
##   workclass            n       Pct
##   <fct>            <int>     <dbl>
## 1 Private          22696 0.69703
## 2 Self-emp-not-inc  2541 0.078038
## 3 Local-gov         2093 0.064279
## 4 ?                 1836 0.056386
## 5 State-gov         1298 0.039864
## 6 Self-emp-inc      1116 0.034274
## 7 Federal-gov        960 0.029483
## 8 Without-pay         14 0.00042996
## 9 Never-worked         7 0.00021498
```

The data analysis excludes this field as it bares little to no impact on the final results.

Fnlwgt

The number of correspondents grouped by the attributes of each data record.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

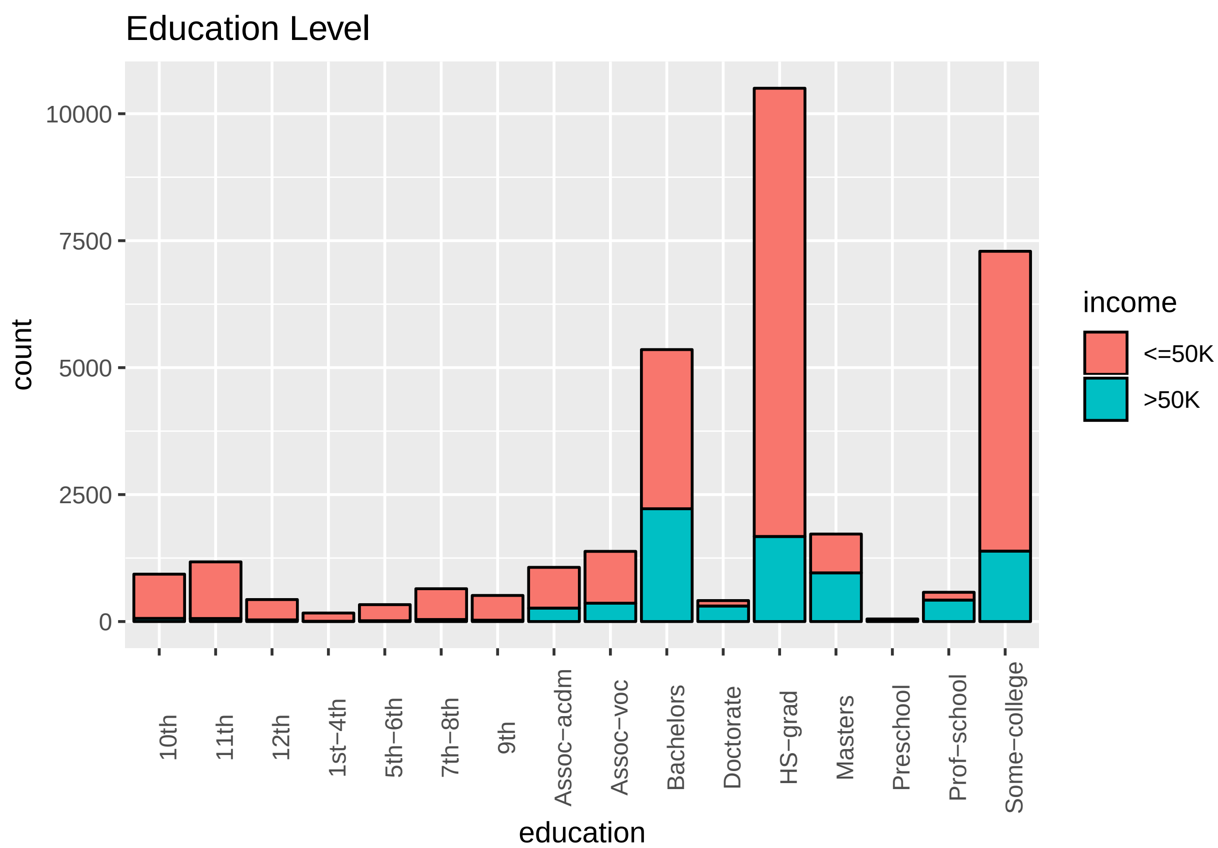



```
## # A tibble: 21,648 x 3
##    fnlwgt     n        Pct
##     <int> <int>      <dbl>
##  1 123011    13 0.00039925
##  2 164190    13 0.00039925
##  3 203488    13 0.00039925
##  4 113364    12 0.00036854
##  5 121124    12 0.00036854
##  6 126675    12 0.00036854
##  7 148995    12 0.00036854
##  8 102308    11 0.00033783
##  9 111483    11 0.00033783
## 10 120131    11 0.00033783
## # ... with 21,638 more rows
```

The data analysis excludes this field as it bares little to no impact on the final results.

Education

The education level of correspondents.

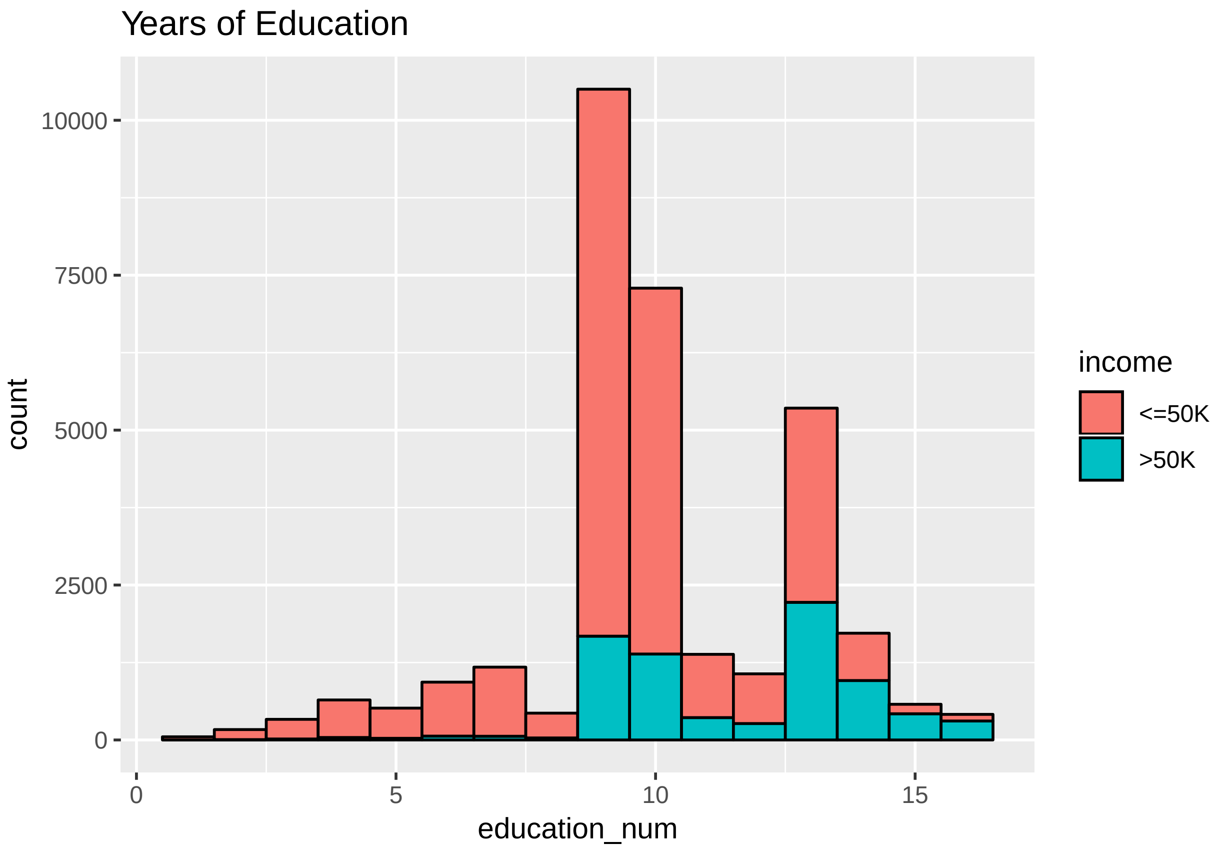



```
## # A tibble: 16 x 3
##    education        n       Pct
##    <fct>        <int>     <dbl>
##  1 HS-grad      10501 0.32250
##  2 Some-college  7291 0.22392
##  3 Bachelors     5355 0.16446
##  4 Masters       1723 0.052916
##  5 Assoc-voc     1382 0.042443
##  6 11th          1175 0.036086
##  7 Assoc-acdm    1067 0.032769
##  8 10th           933 0.028654
##  9 7th-8th        646 0.019840
## 10 Prof-school    576 0.017690
## 11 9th            514 0.015786
## 12 12th           433 0.013298
## 13 Doctorate      413 0.012684
## 14 5th-6th        333 0.010227
## 15 1st-4th        168 0.0051595
## 16 Preschool       51 0.0015663
```

This field is made redundant by the *education_num* field which provides the same information quantitatively. As a result it is excluded from the data analysis.

Education_num

The number of years of education of the correspondents.

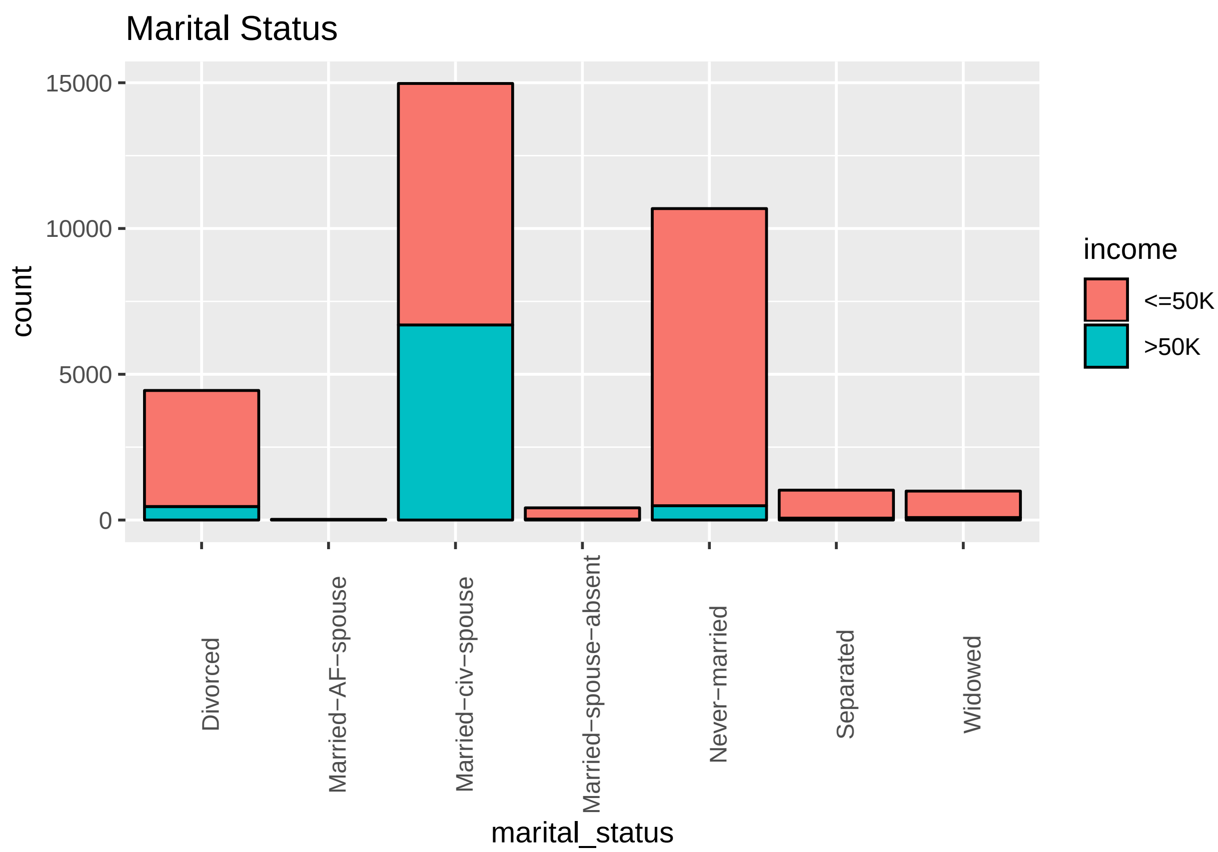



```
## # A tibble: 16 x 3
##    education_num     n       Pct
##            <int> <int>     <dbl>
##  1             9 10501 0.32250
##  2            10  7291 0.22392
##  3            13  5355 0.16446
##  4            14  1723 0.052916
##  5            11  1382 0.042443
##  6             7  1175 0.036086
##  7            12  1067 0.032769
##  8             6   933 0.028654
##  9             4   646 0.019840
## 10            15   576 0.017690
## 11             5   514 0.015786
## 12             8   433 0.013298
## 13            16   413 0.012684
## 14             3   333 0.010227
## 15             2   168 0.0051595
## 16             1    51 0.0015663
```

This fields makes the *education* field redundant.

Marital_status

The marital status of the correspondents.

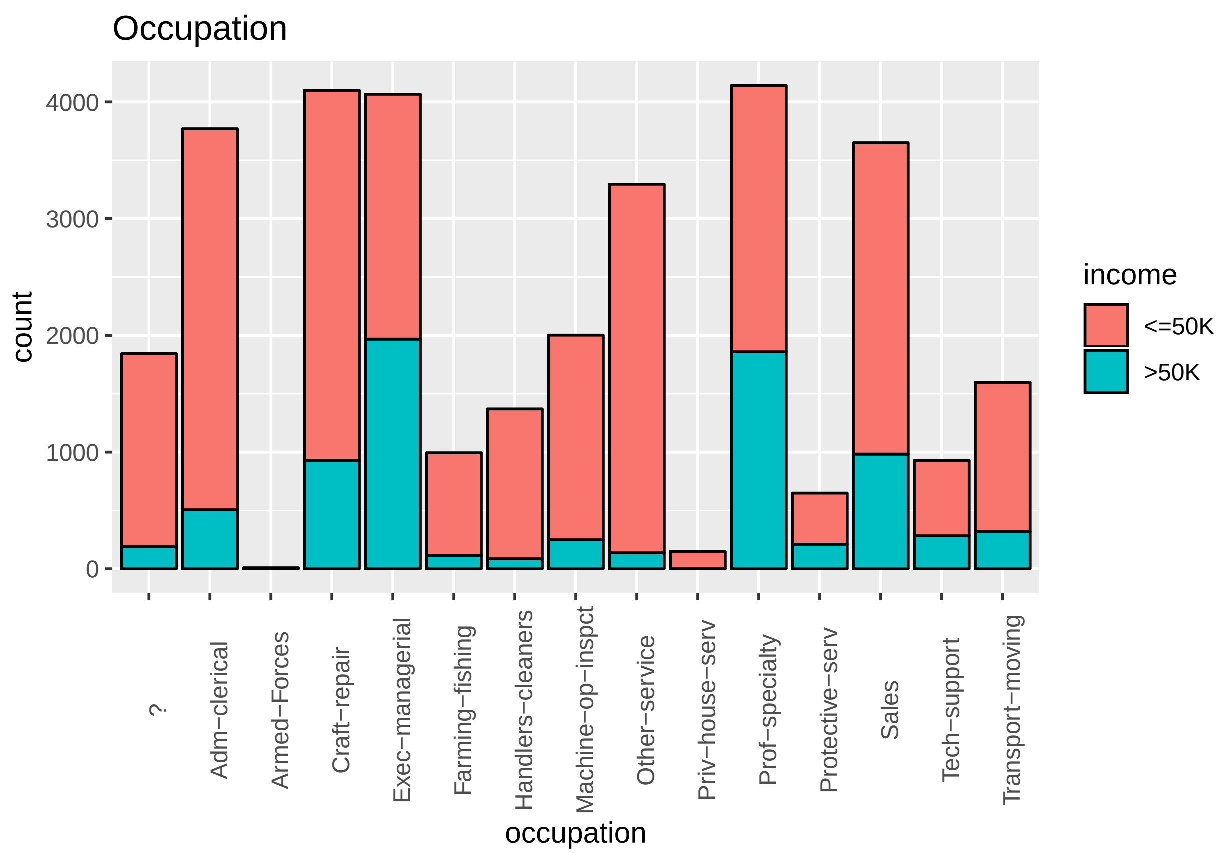



```
## # A tibble: 7 x 3
##   marital_status            n        Pct
##   <fct>                 <int>      <dbl>
## 1 Married-civ-spouse    14976 0.45994
## 2 Never-married         10683 0.32809
## 3 Divorced               4443 0.13645
## 4 Separated              1025 0.031479
## 5 Widowed                 993 0.030497
## 6 Married-spouse-absent   418 0.012837
## 7 Married-AF-spouse        23 0.00070637
```

For the data analysis, this field is grouped into a binary field that assigns a value of *TRUE* for married correspondents.

Occupation

The profession of thew correspondents.

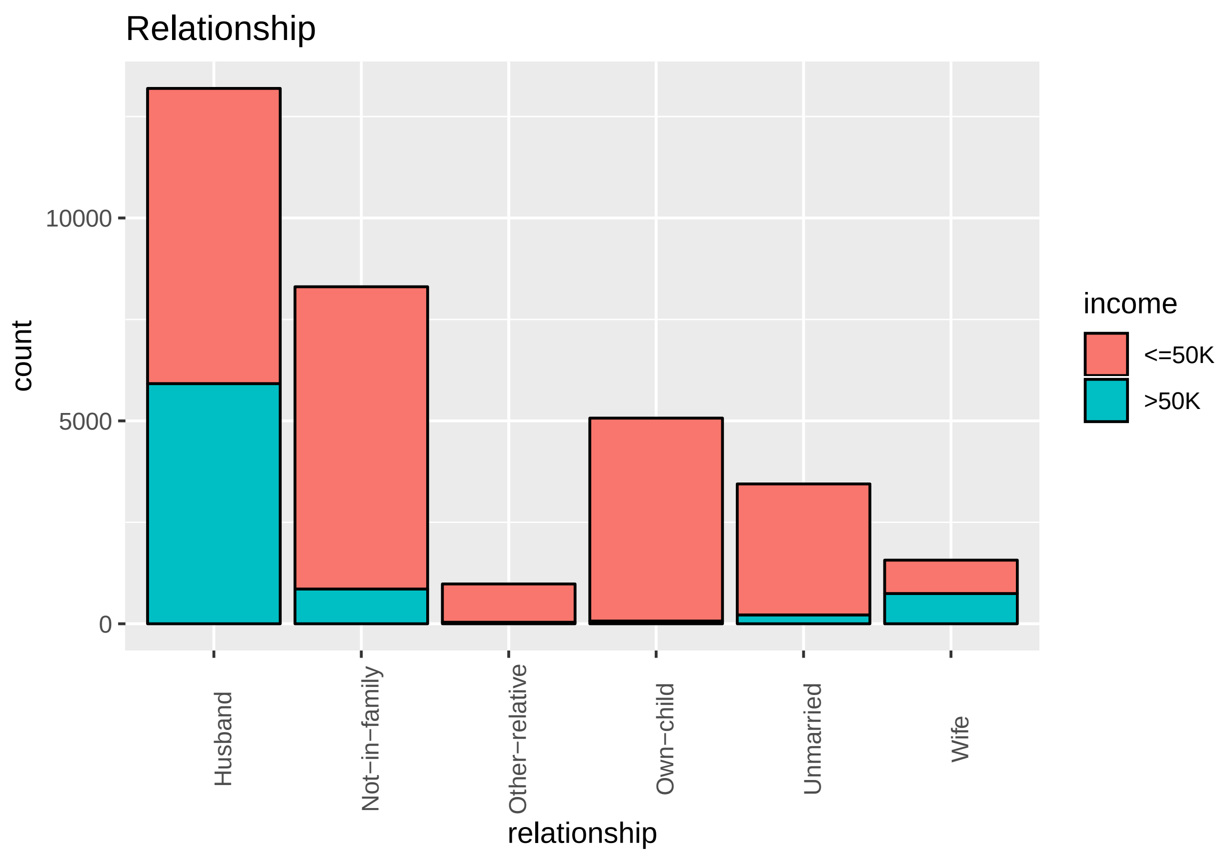



```
## # A tibble: 15 x 3
##    occupation            n        Pct
##    <fct>             <int>      <dbl>
##  1 Prof-specialty     4140 0.12715
##  2 Craft-repair       4099 0.12589
##  3 Exec-managerial    4066 0.12487
##  4 Adm-clerical       3770 0.11578
##  5 Sales              3650 0.11210
##  6 Other-service      3295 0.10119
##  7 Machine-op-inspct  2002 0.061485
##  8 ?                  1843 0.056601
##  9 Transport-moving   1597 0.049046
## 10 Handlers-cleaners  1370 0.042075
## 11 Farming-fishing     994 0.030527
## 12 Tech-support        928 0.028500
## 13 Protective-serv     649 0.019932
## 14 Priv-house-serv     149 0.0045760
## 15 Armed-Forces          9 0.00027640
```

For the data analysis this field is grouped into two categories that separate professional and clerical correspondents from the others.

Relationship

The relationship within a family of the correspondents.

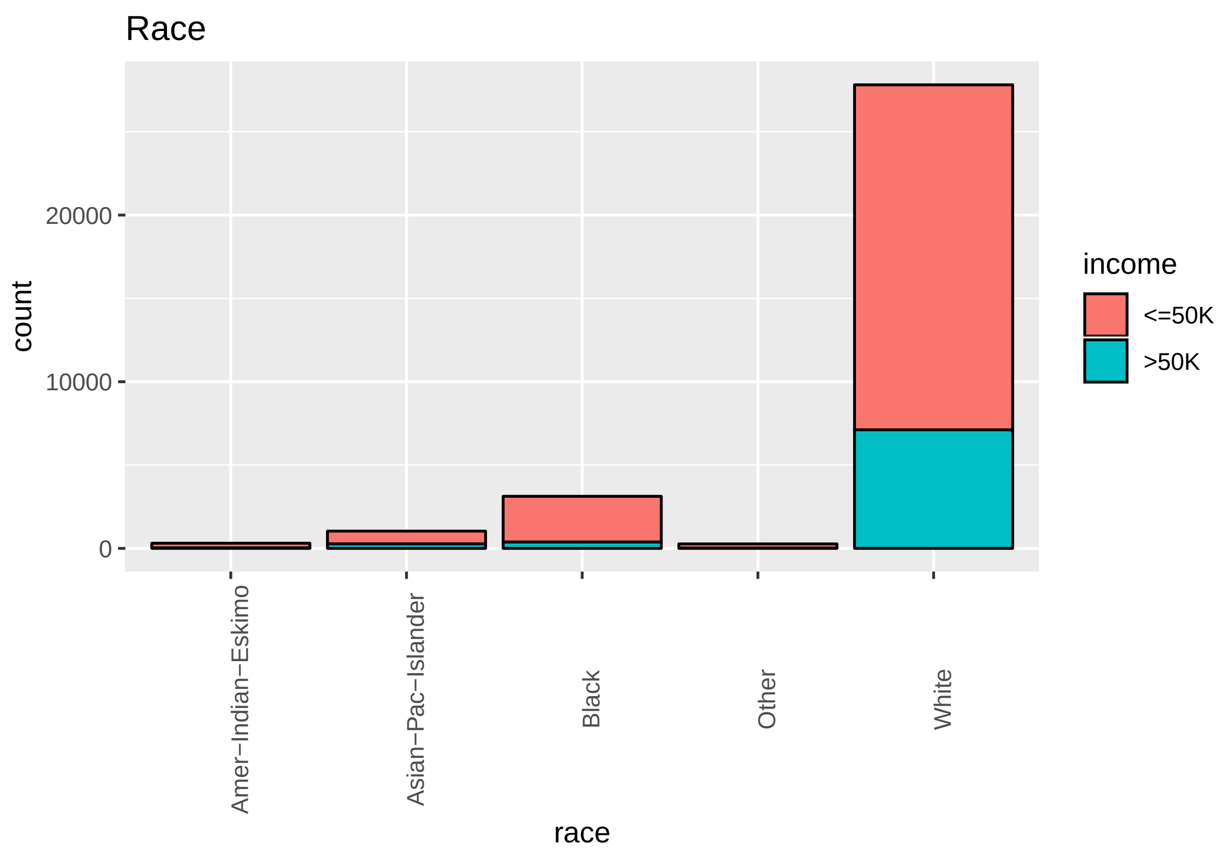



```
## # A tibble: 6 x 3
##   relationship       n      Pct
##   <fct>          <int>    <dbl>
## 1 Husband        13193 0.40518
## 2 Not-in-family   8305 0.25506
## 3 Own-child       5068 0.15565
## 4 Unmarried       3446 0.10583
## 5 Wife            1568 0.048156
## 6 Other-relative   981 0.030128
```

This field is excluded from the data analysis as it has little or no impact. The *sex* and *marital* fields provide more relevant information.

Race

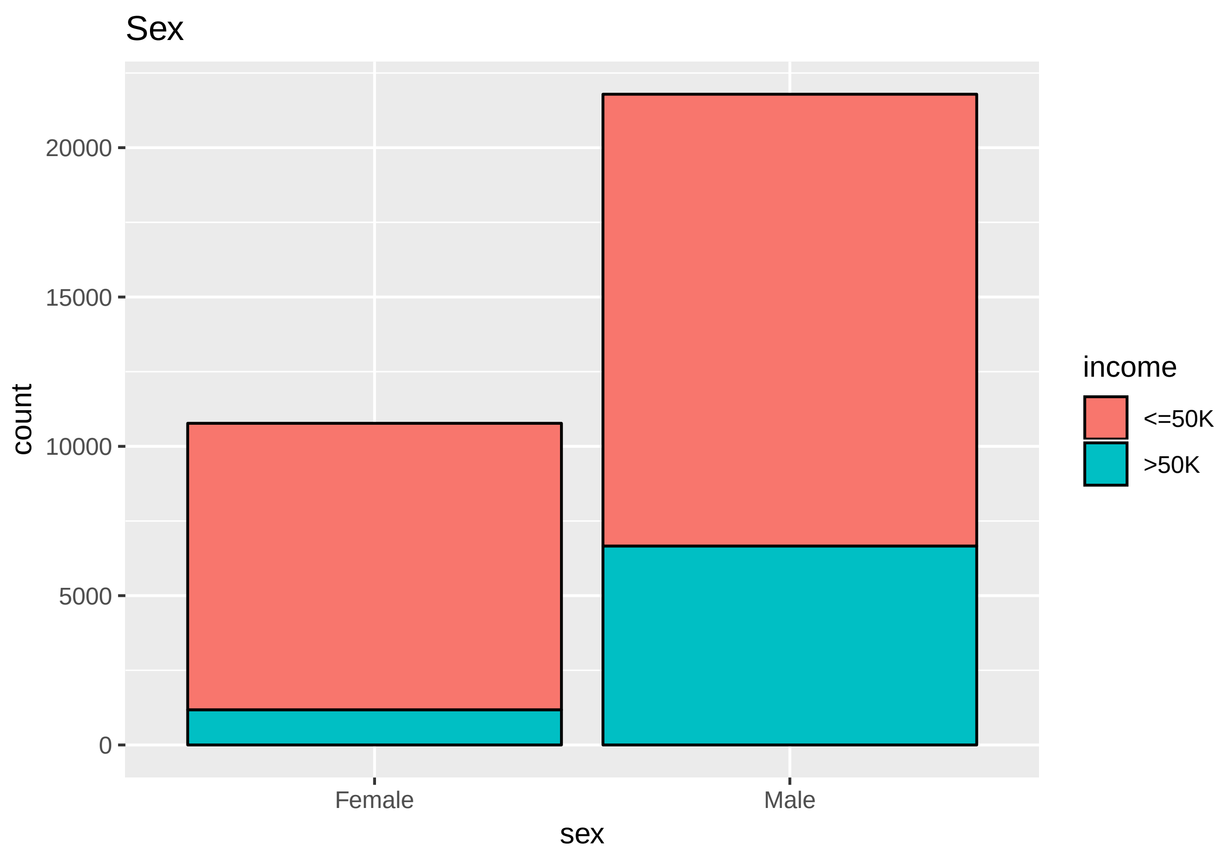
The race of each correspondent.




```
## # A tibble: 5 x 3
##   race                   n       Pct
##   <fct>              <int>     <dbl>
## 1 White              27816 0.85427
## 2 Black               3124 0.095943
## 3 Asian-Pac-Islander  1039 0.031909
## 4 Amer-Indian-Eskimo   311 0.0095513
## 5 Other                271 0.0083228
```

Sex

The sex of each correspondent.

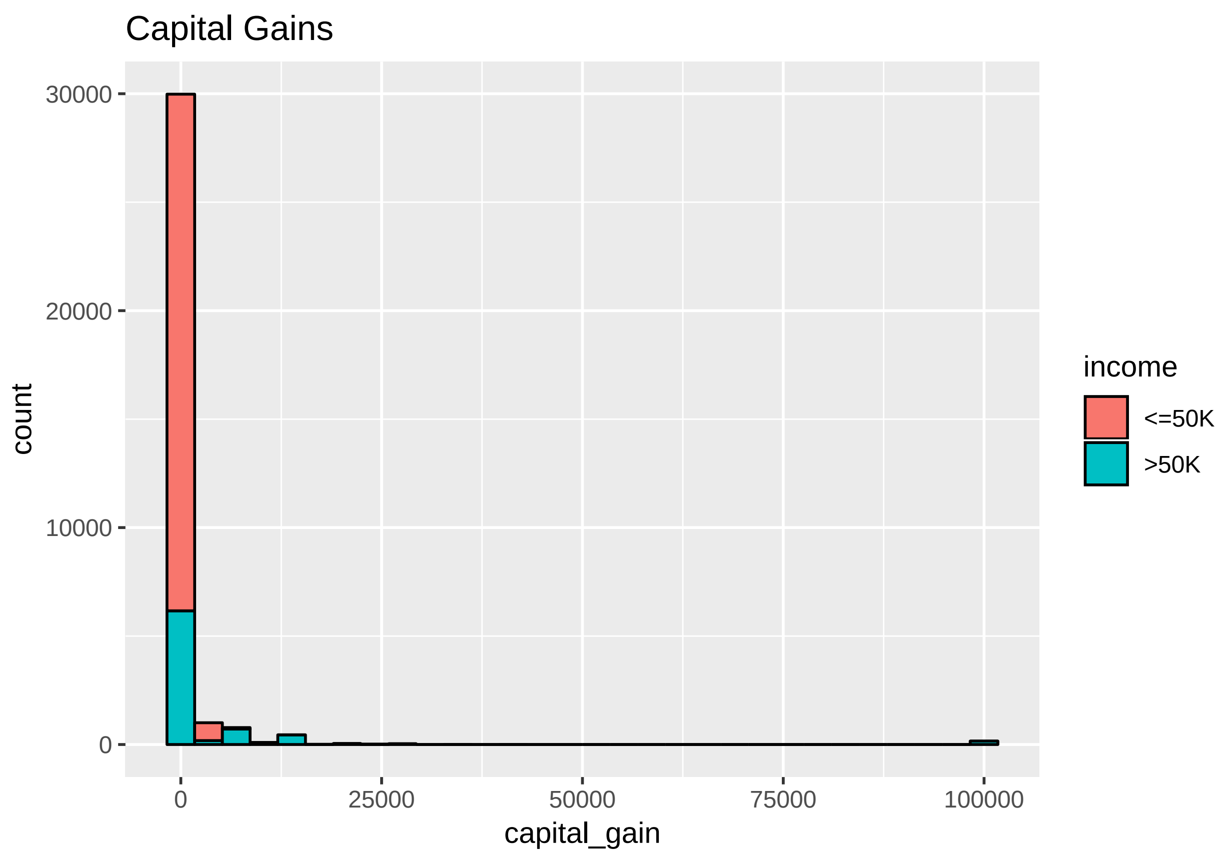



```
## # A tibble: 2 x 3
##   sex         n     Pct
##   <fct>   <int>   <dbl>
## 1 Male    21790 0.66921
## 2 Female 10771 0.33079
```

Capital_gain

The amount of capital gained by the correspondents.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

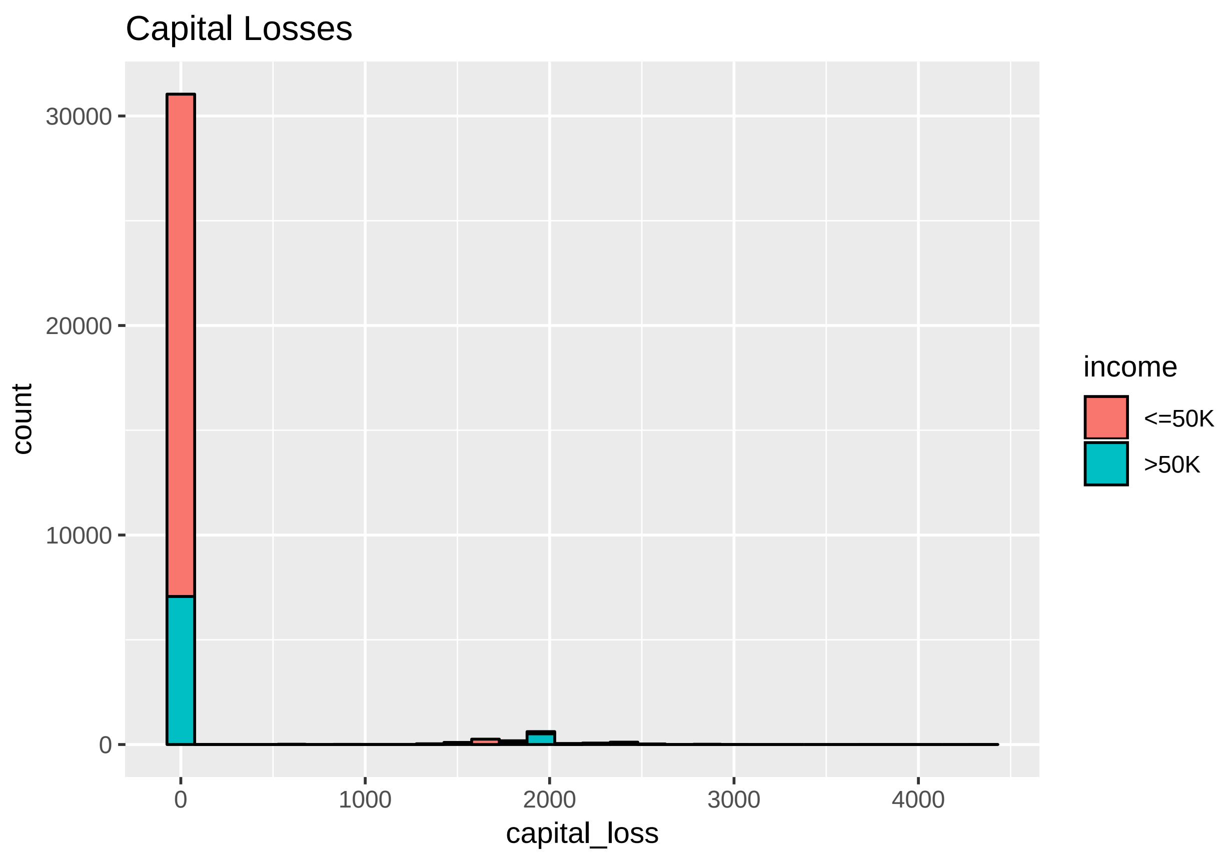



```
## # A tibble: 119 x 3
##    capital_gain     n       Pct
##           <int> <int>     <dbl>
##  1            0 29849 0.91671
##  2        15024   347 0.010657
##  3         7688   284 0.0087221
##  4         7298   246 0.0075551
##  5        99999   159 0.0048831
##  6         3103    97 0.0029790
##  7         5178    97 0.0029790
##  8         4386    70 0.0021498
##  9         5013    69 0.0021191
## 10         8614    55 0.0016891
## # ... with 109 more rows
```

Capital_loss

The amount of capital lost by the correspondents.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

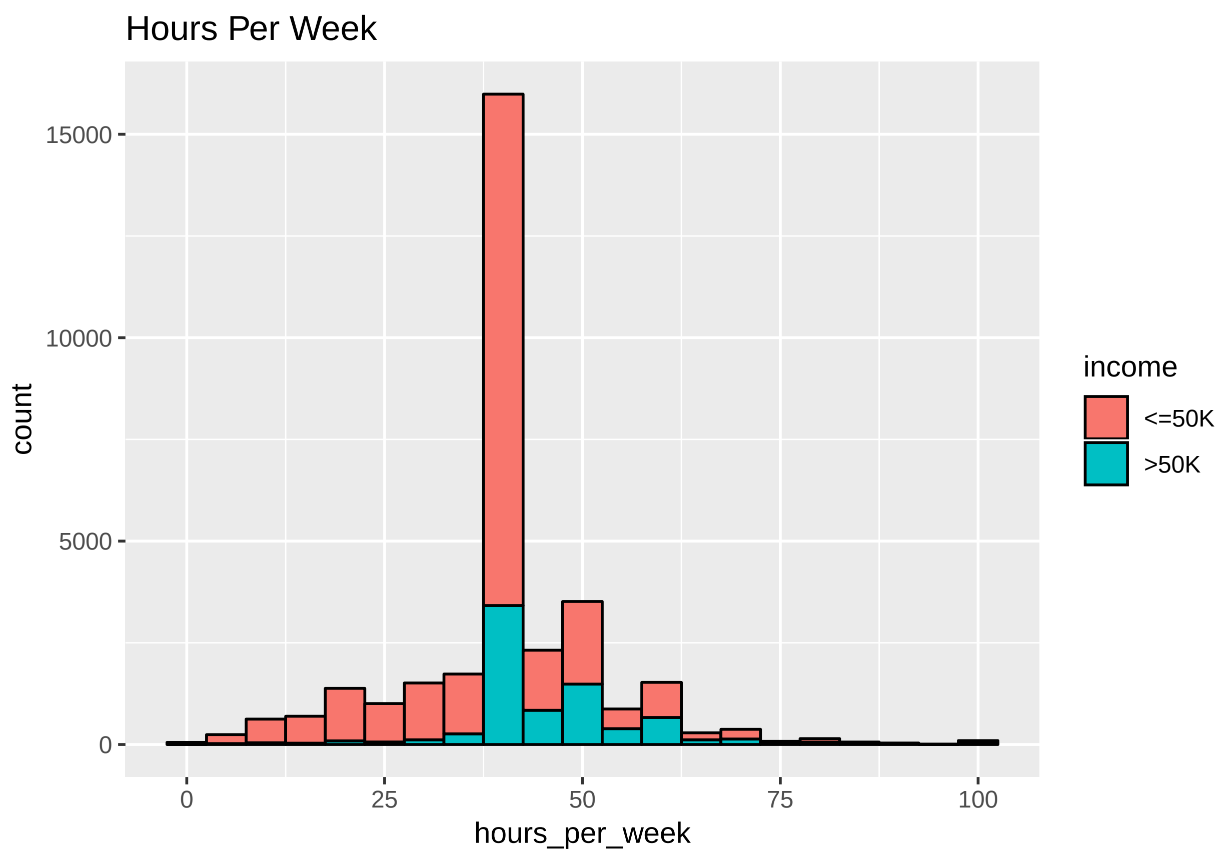



```
## # A tibble: 92 x 3
##    capital_loss     n       Pct
##           <int> <int>     <dbl>
##  1            0 31042 0.95335
##  2         1902   202 0.0062037
##  3         1977   168 0.0051595
##  4         1887   159 0.0048831
##  5         1485    51 0.0015663
##  6         1848    51 0.0015663
##  7         2415    49 0.0015049
##  8         1602    47 0.0014434
##  9         1740    42 0.0012899
## 10         1590    40 0.0012285
## # ... with 82 more rows
```

Because more than 90% of correspondents have capital gains or losses of zero, these field are grouped into binary fields where values greater than zero are assigned as *TRUE*.

Hours_per_week

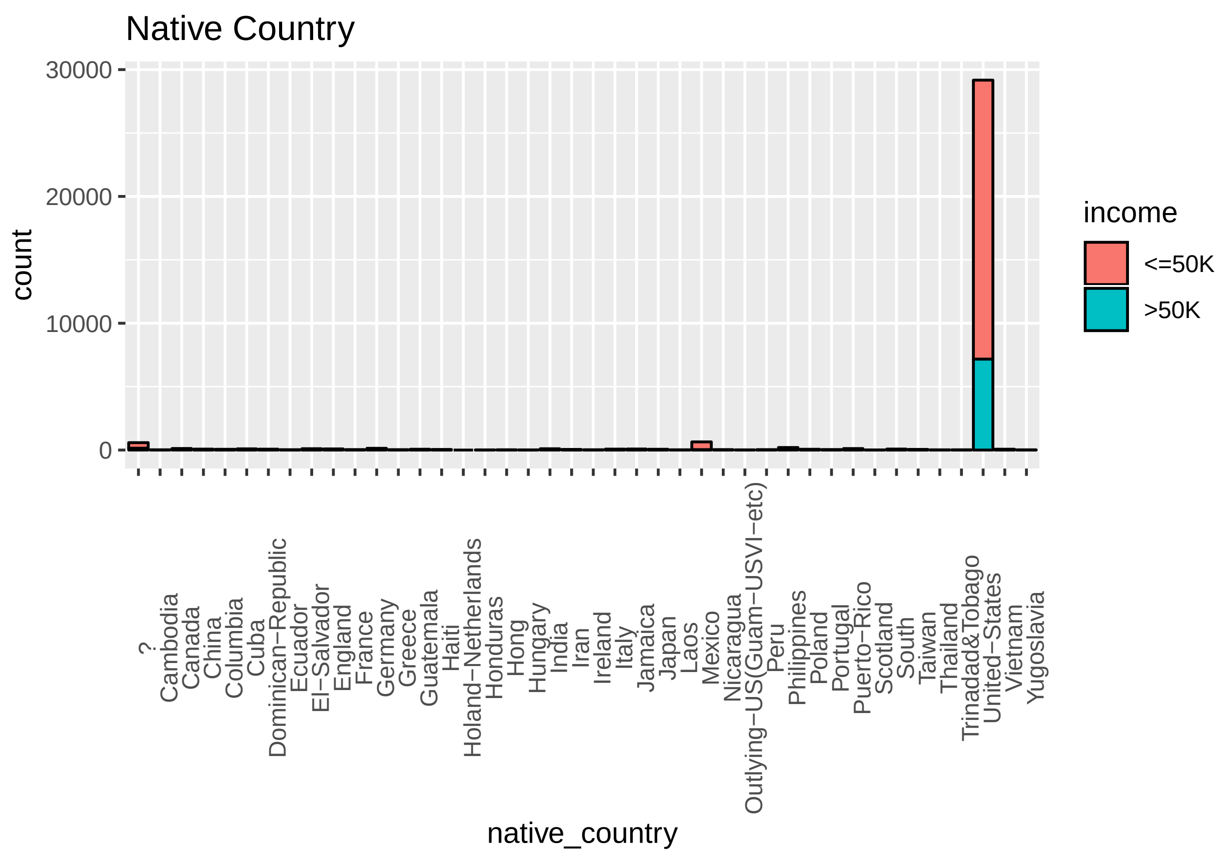
The number of hours worked per week.




```
## # A tibble: 94 x 3
##    hours_per_week     n      Pct
##             <int> <int>    <dbl>
##  1             40 15217 0.46734
##  2             50  2819 0.086576
##  3             45  1824 0.056018
##  4             60  1475 0.045300
##  5             35  1297 0.039833
##  6             20  1224 0.037591
##  7             30  1149 0.035288
##  8             55   694 0.021314
##  9             25   674 0.020700
## 10             48   517 0.015878
## # ... with 84 more rows
```

Native Country

The country of birth of the correspondents.

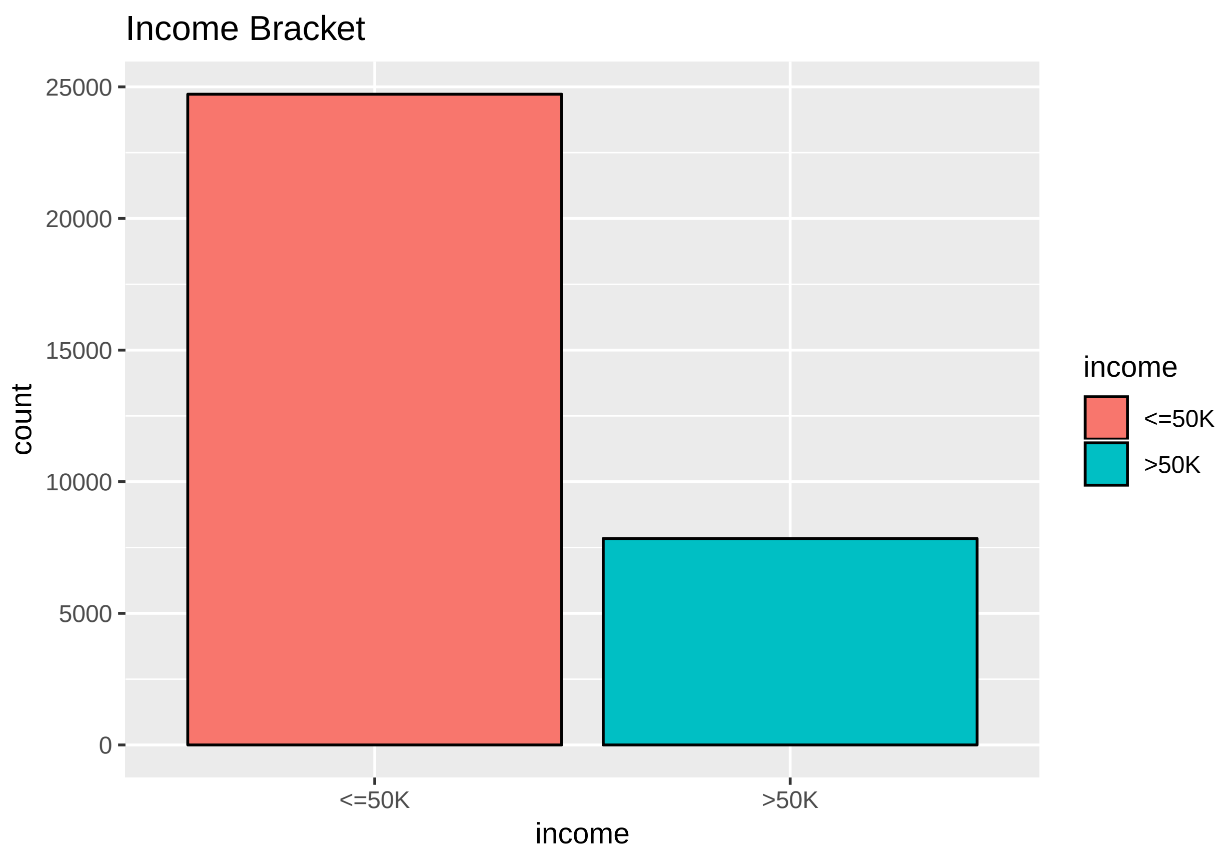



```
## # A tibble: 42 x 3
##    native_country     n       Pct
##    <fct>          <int>     <dbl>
##  1 United-States  29170 0.89586
##  2 Mexico           643 0.019748
##  3 ?                583 0.017905
##  4 Philippines      198 0.0060809
##  5 Germany          137 0.0042075
##  6 Canada           121 0.0037161
##  7 Puerto-Rico      114 0.0035011
##  8 El-Salvador      106 0.0032554
##  9 India            100 0.0030712
## 10 Cuba              95 0.0029176
## # ... with 32 more rows
```

Because the high proportion of correspondents born in America, this field is grouped into a binary field assigning "American" as *TRUE*.

Income

The classification of correspondents into those who earn less or greater than $50,000 annually.




```
## # A tibble: 2 x 3
##   income     n     Pct
##   <fct>  <int>   <dbl>
## 1 <=50K  24720 0.75919
## 2 >50K    7841 0.24081
```

In the data analysis, records classified as greater than $50 thousand (>50K) are assigned a value of 1 and the rest (<=50K) are assigned 0. In the Regression Tree analysis the categorical values are used in addition to the numerically assigned ones.

2.1.2 Excluded Fields

As previously mentioned some fields are excluded from the analysis due to their redundancy. The *Education_num* field quantitatively provides the same information as *Education*.

```
## # A tibble: 16 x 3
## # Groups:   education_num [16]
##    education_num education        n
##            <int> <fct>        <int>
##  1             1 Preschool       51
##  2             2 1st-4th        168
##  3             3 5th-6th        333
##  4             4 7th-8th        646
##  5             5 9th            514
##  6             6 10th           933
##  7             7 11th          1175
##  8             8 12th           433
```

```
## 9              9 HS-grad      10501
## 10            10 Some-college  7291
## 11            11 Assoc-voc     1382
## 12            12 Assoc-acdm    1067
## 13            13 Bachelors     5355
## 14            14 Masters       1723
## 15            15 Prof-school    576
## 16            16 Doctorate      413
```

Sex and *marital_status* render *relationship* redundant.

```
## # A tibble: 54 x 4
## # Groups:   marital_status, relationship [29]
##    marital_status    relationship   sex        n
##    <fct>             <fct>          <fct>  <int>
##  1 Divorced          Not-in-family  Female  1177
##  2 Divorced          Not-in-family  Male    1227
##  3 Divorced          Other-relative Female    65
##  4 Divorced          Other-relative Male      45
##  5 Divorced          Own-child      Female   151
##  6 Divorced          Own-child      Male     177
##  7 Divorced          Unmarried      Female  1279
##  8 Divorced          Unmarried      Male     322
##  9 Married-AF-spouse Husband        Male       9
## 10 Married-AF-spouse Other-relative Female     1
## # ... with 44 more rows
```

2.1.3 Modified data set for analysis

The original data set has been modified for analysis purposes. The training and testing data sets were generated using the following modified data set:

```
## 'data.frame':    32561 obs. of  12 variables:
##  $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ education_num : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...
##  $ sex           : num  1 1 1 1 0 0 0 1 0 1 ...
##  $ capital_gain  : logi  TRUE FALSE FALSE FALSE FALSE FALSE ...
##  $ capital_loss  : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ hours_per_week: int  40 13 40 40 40 40 16 45 50 40 ...
##  $ income        : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 1 2 2 2 ...
##  $ married       : logi  FALSE TRUE FALSE TRUE TRUE TRUE ...
##  $ prof_clerical : logi  TRUE TRUE FALSE FALSE TRUE TRUE ...
##  $ American      : logi  TRUE TRUE TRUE TRUE FALSE TRUE ...
##  $ income_class  : num  0 0 0 0 0 0 0 1 1 1 ...
```

The modified data set drops the *education, fnlgwt, relationship,* and *native_country* fields for reasons explained in the previous section.

The binary *married* field replaces the categorical *marital_status* field.

Prof_Clerical replaces *Occupation* , and *American* replaces *native_country*, in each case grouping categorical fields into binary fields.

capital_gains and *capital_loss* were both converted to binary fields. And *income_class* converts the values *income* into ones for (>50k) and zeroes (<=50K).

Below is a summary of the modified model:

```
##       age        education_num                   race
##  Min.   :17.00   Min.   : 1.00   Amer-Indian-Eskimo:  311
##  1st Qu.:28.00   1st Qu.: 9.00   Asian-Pac-Islander: 1039
##  Median :37.00   Median :10.00   Black             : 3124
##  Mean   :38.58   Mean   :10.08   Other             :  271
##  3rd Qu.:48.00   3rd Qu.:12.00   White             :27816
##  Max.   :90.00   Max.   :16.00
##       sex         capital_gain    capital_loss    hours_per_week
##  Min.   :0.0000   Mode :logical   Mode :logical   Min.   : 1.00
##  1st Qu.:0.0000   FALSE:29849     FALSE:31042     1st Qu.:40.00
##  Median :1.0000   TRUE :2712      TRUE :1519      Median :40.00
##  Mean   :0.6692                                   Mean   :40.44
##  3rd Qu.:1.0000                                   3rd Qu.:45.00
##  Max.   :1.0000                                   Max.   :99.00
##    income       married        prof_clerical    American
##  <=50K:24720   Mode :logical   Mode :logical   Mode :logical
##  >50K : 7841   FALSE:17144     FALSE:15358     FALSE:3391
##                TRUE :15417     TRUE :17203     TRUE :29170
##
##
##
##   income_class
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2408
##  3rd Qu.:0.0000
##  Max.   :1.0000
```

For analysis, the training data set uses 80% of the the original analysis data set, and the test set uses the remaining 20%. The *income class* distribution for the is as follows:

```
##
##     0     1
## 19776  6272
```

2.2 GENERALISED LINEAR REGRESSION MODEL (GLM)

The first analysis fits training data to a *Generalised Linear Regression Model*.

A summary of the model is provided below:

```
##
## Call:
## glm(formula = income_class ~ ., family = binomial("logit"), data = .)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8941  -0.5470  -0.2399  -0.0672   3.5383
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -9.797222   0.280349 -34.947  < 2e-16 ***
## age                     0.026360   0.001550  17.002  < 2e-16 ***
## education_num           0.297553   0.009044  32.899  < 2e-16 ***
## raceAsian-Pac-Islander  0.426782   0.257553   1.657  0.09751 .
```

```
## raceBlack               0.268499   0.242006   1.109  0.26723
## raceOther              -0.372311   0.372653  -0.999  0.31775
## raceWhite               0.462547   0.231184   2.001  0.04542 *
## sex                     0.342300   0.051743   6.615 3.71e-11 ***
## capital_gainTRUE        1.704636   0.060206  28.313  < 2e-16 ***
## capital_lossTRUE        1.167992   0.077556  15.060  < 2e-16 ***
## hours_per_week          0.031551   0.001635  19.291  < 2e-16 ***
## marriedTRUE             2.265470   0.048918  46.312  < 2e-16 ***
## prof_clericalTRUE       0.806610   0.043390  18.590  < 2e-16 ***
## AmericanTRUE            0.229513   0.073317   3.130  0.00175 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28756  on 26047  degrees of freedom
## Residual deviance: 18083  on 26034  degrees of freedom
## AIC: 18111
##
## Number of Fisher Scoring iterations: 6
```

The coefficients of the variables:

```
##
## Call:  glm(formula = income_class ~ ., family = binomial("logit"), data = .)
##
## Coefficients:
##           (Intercept)                    age          education_num
##              -9.79722                0.02636                0.29755
## raceAsian-Pac-Islander              raceBlack              raceOther
##               0.42678                0.26850               -0.37231
##             raceWhite                    sex       capital_gainTRUE
##               0.46255                0.34230                1.70464
##      capital_lossTRUE         hours_per_week            marriedTRUE
##               1.16799                0.03155                2.26547
##     prof_clericalTRUE           AmericanTRUE
##               0.80661                0.22951
##
## Degrees of Freedom: 26047 Total (i.e. Null);  26034 Residual
## Null Deviance:       28760
## Residual Deviance: 18080     AIC: 18110
```

The GLM model yields the following results:

```
##
##       0    1
##   0 4553  391
##   1  696  873
```

```
## [1] 6513
```

```
##    fit
##              0          1
##   0 0.69906341 0.06003378
##   1 0.10686320 0.13403961
```

```
## # A tibble: 1 x 2
```

```
##   Method                      Accuracy
##   <chr>                          <dbl>
## 1 Generalised Linear Model  0.83310
```

GLM yields an accuracy of approximately 0.833103.

2.3 K NEAREST NEIGHBOURS (KNN)

In an effort to improve upon the results of the GLM model, the analysis data set is now analysed using *K Nearest Neighbours*. This algorithm classifies or estimates data based on the similarity of other data points.

The optimal tuning parameter for the model:

```
##   k
## 3 9
```

A summary of the KNN model:

```
## 9-nearest neighbor model
## Training set outcome distribution:
##
## <=50K  >50K
## 19776  6272
```

The KNN model yields the following results:

```
##
## knn_pred <=50K >50K
##    <=50K  4568  785
##    >50K    376  784
```

```
## [1] 6513
```

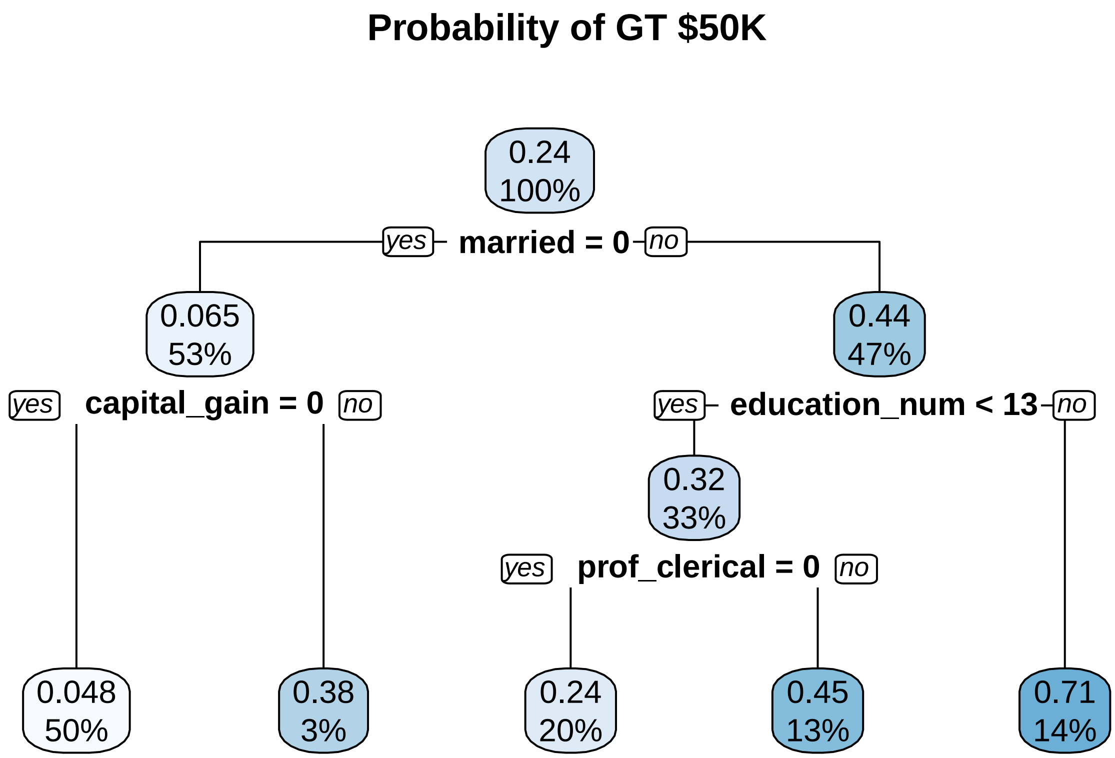
```
##
## knn_pred      <=50K       >50K
##    <=50K 0.70136650 0.12052817
##    >50K  0.05773069 0.12037464
```

```
## # A tibble: 1 x 2
##   Method                Accuracy
##   <chr>                    <dbl>
## 1 K Nearest Neighbours  0.82174
```

The accuracy of 0.8217411 is slightly less than that yielded by the linear regression model.

2.4 CLASSIFICATION AND REGRESSION TREES (CART)

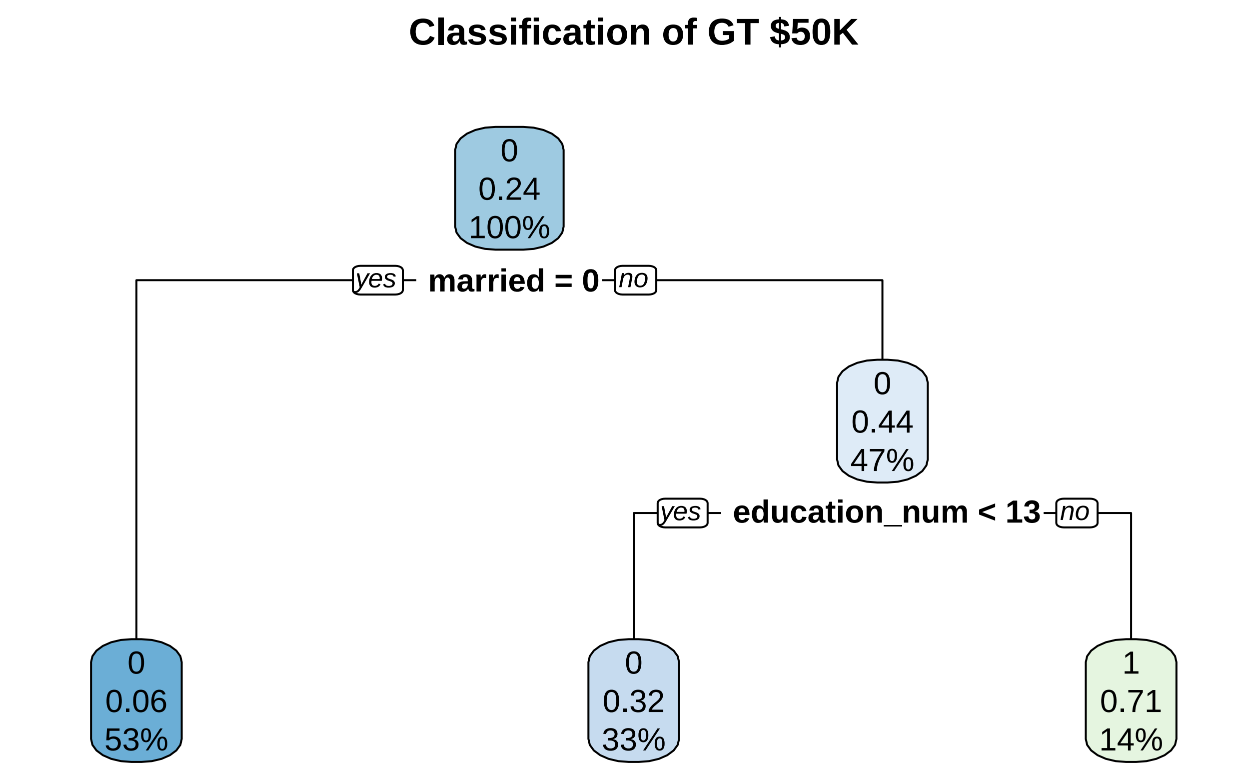
The data set is now analysed using *Classification and Regression Trees (CART).* This method recursively partitions the data set and fits regression models to each data subset.



An output of the rules of the regression tree shown above:

```
##   income_class                                                                                        cove
##          0.048 when married is 0                        & capital_gain is 0                             50
##          0.237 when married is 1 & education_num <  13                       & prof_clerical is 0       20
##          0.382 when married is 0                        & capital_gain is 1                              3
##          0.453 when married is 1 & education_num <  13                       & prof_clerical is 1       13
##          0.712 when married is 1 & education_num >= 13                                                 14
```

This first regression groups data into floating point decimal values which measure the probability of earning above $50,000 annually. For a binary classification, Values of greater than 0.5 would be rounded to 1.0 and thus classified as ">50K". This form of the algorithm is a *Regression Tree*. But because of the rounding involved in fitting the training data, this functions as a *Classification Tree*.



An output of the rules of the regression tree shown above:

```
##  income_class                                                 cover
##          0.06 when married is 0                                 53%
##          0.32 when married is 1 & education_num <  13           33%
##          0.71 when married is 1 & education_num >= 13           14%
```

This regression tree classifies data into discrete values of 0 and 1, which represent the categorical classes of “<=50K” and “>50K” respectively. This modification of the algorithm is a *Classification Tree*.

The CART model yields the following results:

```
##
## rt_predc    0    1
##        0 4663  923
##        1  281  646
```

```
## [1] 6513
```

```
##
## rt_predc          0          1
##        0 0.71595271 0.14171657
##        1 0.04314448 0.09918624
```

```
## # A tibble: 1 x 2
##   Method                                      Accuracy
##   <chr>                                          <dbl>
## 1 Classification and Regression Trees (CART)   0.81514
```

The accuracy of 0.815139 is slightly less than that yielded by the linear regression model.

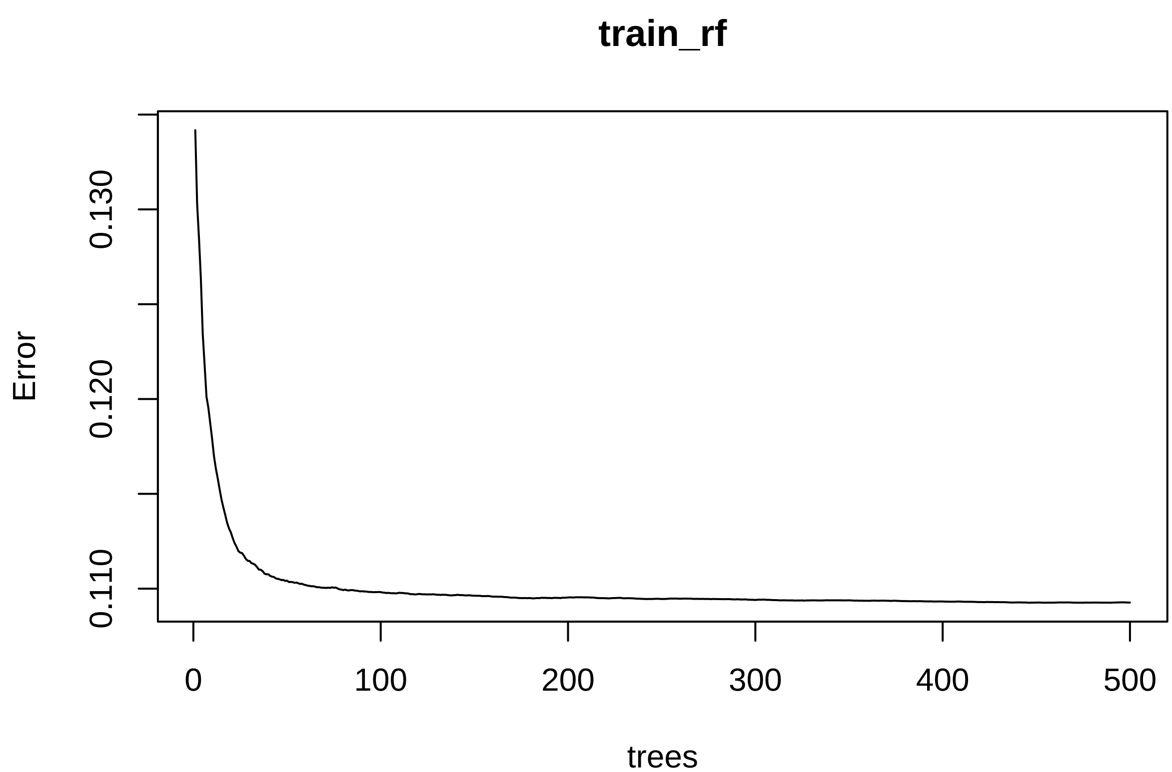
2.5 RANDOM FORESTS

The data set is now analysed using the *Random Forest* algorithm. This algorithm fits data by aggregating the results of a large number of individual regression trees.

A summary of the Random Forest model:

```
## Warning in randomForest.default(m, y, ...): The response has five or fewer
## unique values. Are you sure you want to do regression?
```

```
##
## Call:
##  randomForest(formula = income_class ~ ., data = train_set4)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##           Mean of squared residuals: 0.1092682
##                     % Var explained: 40.23
```



The above plot demonstrates how the model error diminishes with the number of trees used to analyse the data.

The Random Forest model yields the following results:

```
##
## rf_pred    0    1
##       0 4587  681
##       1  357  888
```

```
## [1] 6513
```

```
##
## rf_pred          0          1
##       0 0.70428374 0.10456011
##       1 0.05481345 0.13634270
```

```
## # A tibble: 1 x 2
##   Method         Accuracy
##   <chr>             <dbl>
## 1 Random Forests  0.84063
```

The accuracy of 0.8406264 is a slight improvement upon the accuracy of linear regression model.

3. RESULTS

The final results yielded the following for the four different machine learning algorithms applied to the data set:

```
## # A tibble: 4 x 2
##   Method                                      Accuracy
##   <chr>                                          <dbl>
## 1 Generalised Linear Model (GLM)               0.83310
## 2 K Nearest Neighbours (KNN)                   0.82174
## 3 Classification and Regression Trees (CART)  0.81514
## 4 Random Forests                               0.84063
```

Only the *Random Forest* algorithm improved upon the accuracy of the *Generalised Linear Model*.

4. CONCLUSION

As demonstrated, the *Random Forest* algorithm produces the most accurate estimate for the income of census correspondents. It is also the only one that yields greater accuracy than the *Genralised Linear Model (GLM)* algorithm. The accuracy of this model could have been further increased by experimenting with different combinations of the socio-economic factors that determine income. Socio-economic factors could have also been further revised to group attributes into binary values or a smaller number of super-grouped categories. The analysis could have also been used to fit the *Adult Census Income Test Set* (https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.test) in order to determine how well the analysis models perform against a completely different data set.