Adult Income Dataset

Arun K. Verma

May 25, 2016

Introduction:

In this project, we are given a large adult income dataset which includes both training dataset and test dataset. Our goal is to predict income based on provided features (or a set of them). For that, I will build prediction model using our training dataset. The training dataset is given in data frames having 32561 rows and 15 columns. Here, we have some information of the dataset:

Load Train Dataset

```
adult <- read.csv("/Users/Arun/Desktop/Income/adult.train")
```

Dimensions of Train Dataset

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

Structure of Train 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 20964
2 45781 159449 ...
##  $ education     : Factor w/ 16 levels " 10th"," 11th",..: 10 10 12 2 10 13
7 12 13 10 ...
##  $ education.num : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ marital.status: 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
...
##  $ capital.gain  : int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ capital.loss  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hours.per.week: int  40 13 40 40 40 40 16 45 50 40 ...
##  $ native.country: Factor w/ 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
...
```

Summary of Train Dataset

we can see summary of our dataset which will provide us a 5 variable summary (mean, median, min, max, quartiles).

```
##       age                    workclass         fnlwgt       
##  Min.   :17.00    Private         :22696   Min.   :  12285  
##  1st Qu.:28.00    Self-emp-not-inc: 2541   1st Qu.: 117827  
##  Median :37.00    Local-gov       : 2093   Median : 178356  
##  Mean   :38.58    ?               : 1836   Mean   : 189778  
##  3rd Qu.:48.00    State-gov       : 1298   3rd Qu.: 237051  
##  Max.   :90.00    Self-emp-inc    : 1116   Max.   :1484705  
##                  (Other)          :  981                    
##          education     education.num                  marital.status 
##   HS-grad     :10501   Min.   : 1.00    Divorced             : 4443  
##   Some-college: 7291   1st Qu.: 9.00    Married-AF-spouse    :   23  
##   Bachelors   : 5355   Median :10.00    Married-civ-spouse   :14976  
##   Masters     : 1723   Mean   :10.08    Married-spouse-absent:  418  
##   Assoc-voc   : 1382   3rd Qu.:12.00    Never-married        :10683  
##   11th        : 1175   Max.   :16.00    Separated            : 1025  
##  (Other)      : 5134                    Widowed              :  993  
##             occupation           relationship  
##   Prof-specialty :4140    Husband       :13193  
##   Craft-repair   :4099    Not-in-family : 8305  
##   Exec-managerial:4066    Other-relative:  981  
##   Adm-clerical   :3770    Own-child     : 5068  
##   Sales          :3650    Unmarried     : 3446  
##   Other-service  :3295    Wife          : 1568  
##  (Other)         :9541                          
##                  race               sex         capital.gain  
##   Amer-Indian-Eskimo:  311    Female:10771   Min.   :    0  
##   Asian-Pac-Islander: 1039    Male  :21790   1st Qu.:    0  
##   Black             : 3124                   Median :    0  
##   Other             :  271                   Mean   : 1078  
##   White             :27816                   3rd Qu.:    0  
##                                              Max.   :99999  
##                                                             
##   capital.loss    hours.per.week         native.country     income     
##  Min.   :   0.0   Min.   : 1.00    United-States:29170    <=50K:24720  
##  1st Qu.:   0.0   1st Qu.:40.00    Mexico       :  643    >50K : 7841  
##  Median :   0.0   Median :40.00    ?            :  583                 
##  Mean   :  87.3   Mean   :40.44    Philippines  :  198                 
##  3rd Qu.:   0.0   3rd Qu.:45.00    Germany      :  137                 
##  Max.   :4356.0   Max.   :99.00    Canada       :  121                 
##                                   (Other)       : 1709                 
```

Exploration and Preparation of Train Dataset

Firstly, we need to pre-process our data, for which I will find all the rows having unknown values and remove them for buildinng our model. The total number of rows are 2399, which have unknown values.

```
##
## FALSE  TRUE
##  2399 30162
```

In our dataset, we have two variables as **fnlwgt** and **education.num**. The fnlwgt represents final sampling wieght and education.num represents the highest level of education. Both variables have continous values, and doesn't play much role in predicting income. So, I will remove both variables.

Now, The dimensions of our Train Dataset is :

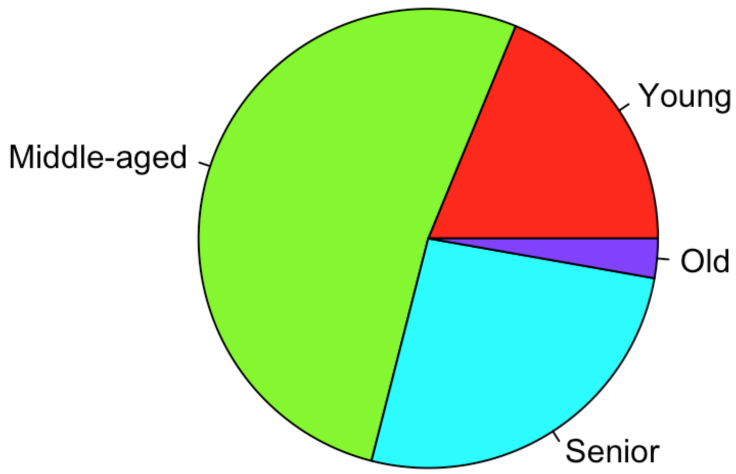
```
## [1] 30162    13
```

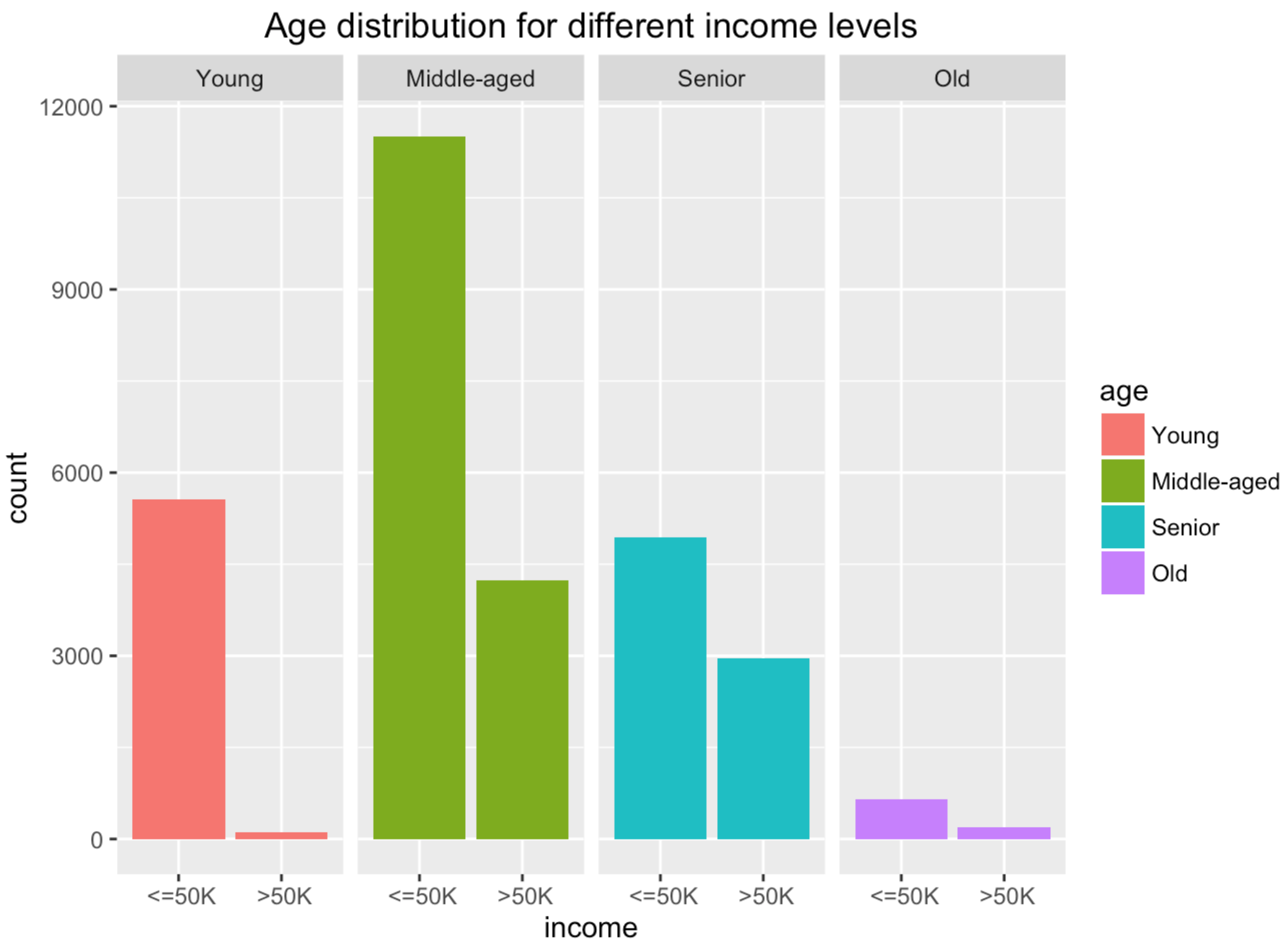
Age

To explore, First I am considering age variable. In our data minimum age is 17 and maximum age is 90. So, I will group them into four as *Young, Middle-age, Senior, Old*. We can visualize them through pie-chart. Also, we can see the age distribution for different income levels. We can see that mostly middle-aged people are working with income less than 50K.

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

```
##
##         Old       Young      Senior Middle-aged
##         839        5668        7900       15755
```

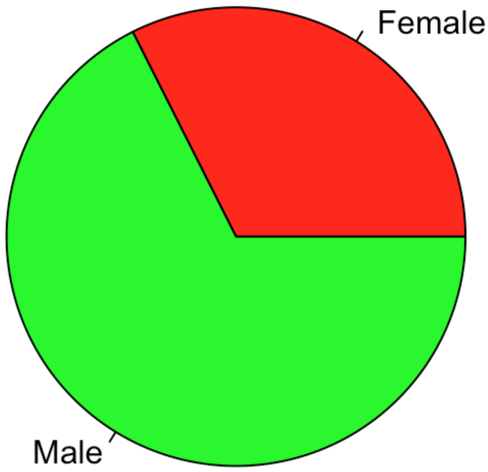

Young
Middle-aged
Old
Senior

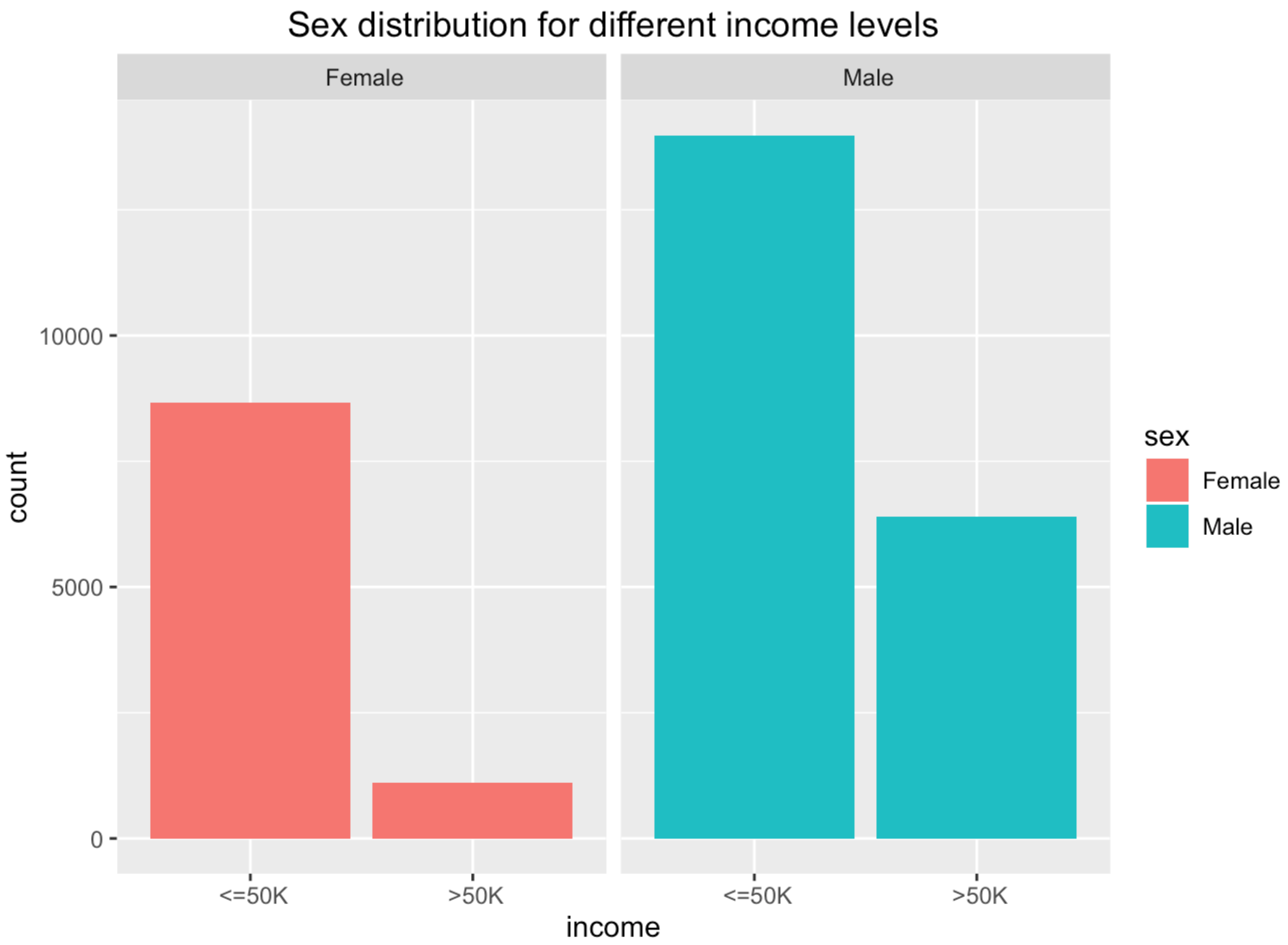

Age distribution for different income levels
Young
Middle-aged
Senior
Old
12000
9000
6000
3000
0
count
<=50K
>50K
income
age
Young
Middle-aged
Senior
Old

Sex

Second, we are considering sex variable to see that how many men and women are in our dataset. We also showed the their distribution for different income levels. We can see that almost 2/3 males are working and most of them have income less than 50K.

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

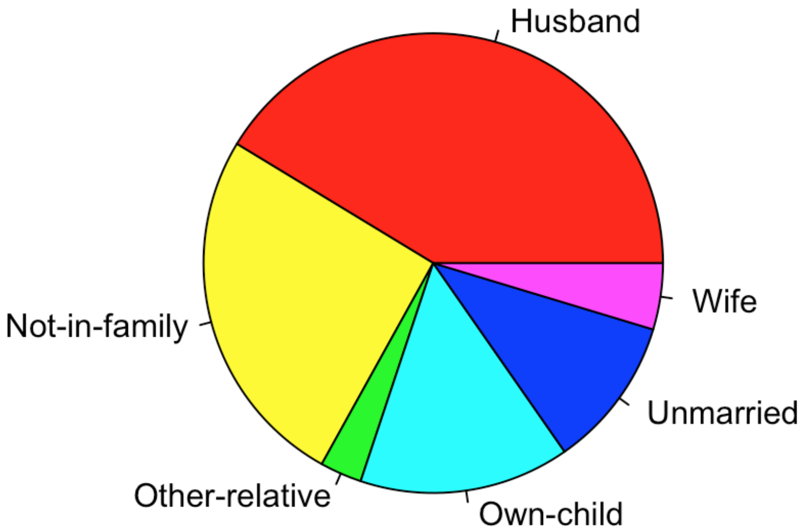


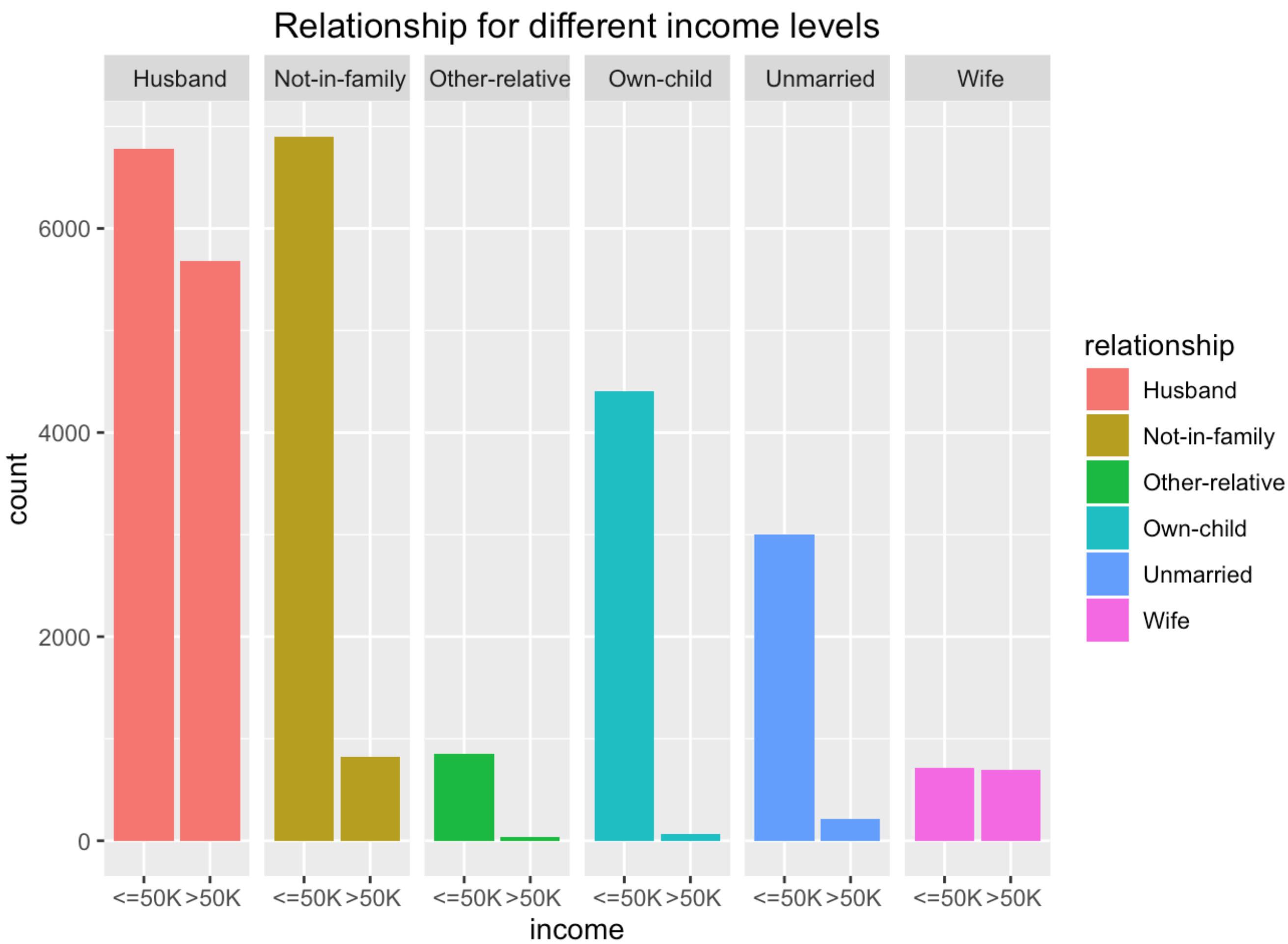


Relationship

Next, we are considering relationship status of people. We can see that moslty husband and those people who are not in family are working having income less than 50K.

```
## 
## Other-relative           Wife      Unmarried      Own-child 
##            889           1406           3212           4466 
##  Not-in-family        Husband 
##           7726          12463
```


Husband
Wife
Not-in-family
Unmarried
Other-relative
Own-child

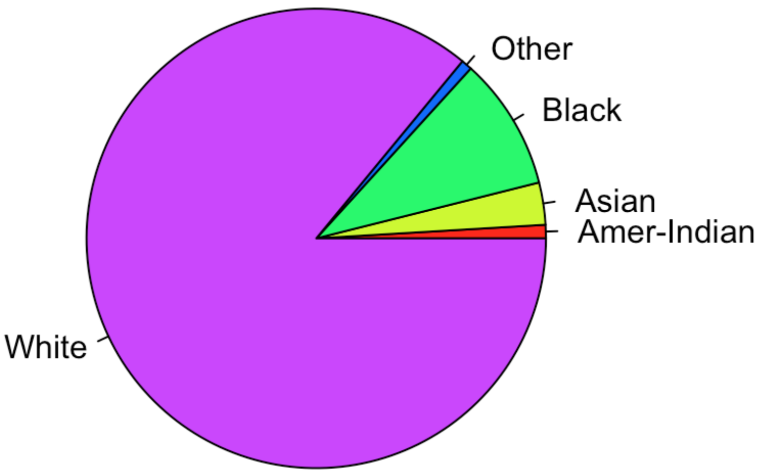

Relationship for different income levels
Husband
Not-in-family
Other-relative
Own-child
Unmarried
Wife
6000
4000
2000
0
count
<=50K >50K
income
relationship
Husband
Not-in-family
Other-relative
Own-child
Unmarried
Wife

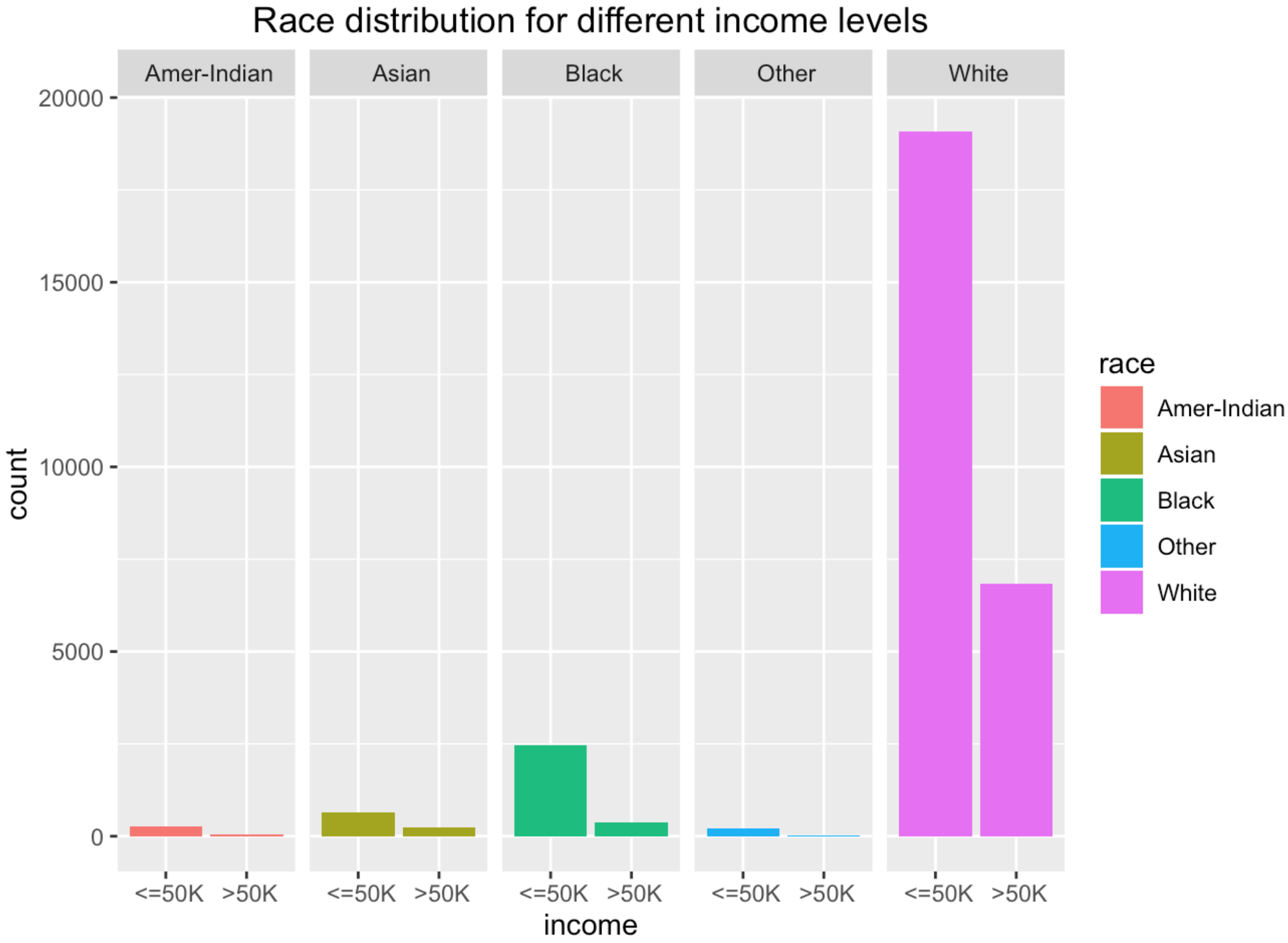
Race

Next, we are also looking race variable. Almost 80% white people are working, also we can that most of the white people have income less than 50K.

```
##
##                Other  Amer-Indian-Eskimo  Asian-Pac-Islander
##                  231                 286                 895
##                Black               White
##                 2817               25933
```

```
##
##         Other  Amer-Indian         Asian         Black         White
##           231          286           895          2817         25933
```



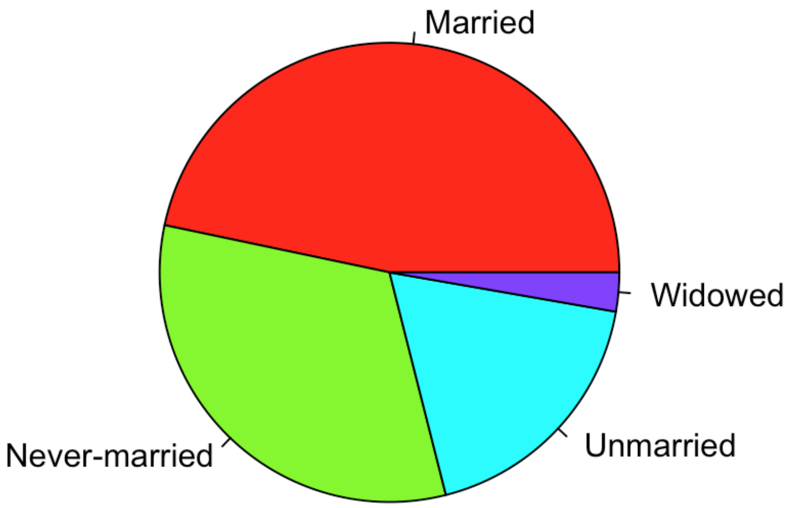


Marital Status

Here again, I will group them like *Married-AF-spouse* and *Married-civ-spouse* are **Married** people. And, *Married-spouse-absent*, *Separated* and *Divorced* are **Unmarried** people. Now, we can say for most of unmarried and never-married people have income less than 50K. While for married people, almost half of them have income greater than 50K and other half have income less than 50K.

```
## 
##     Married-AF-spouse Married-spouse-absent               Widowed 
##                    21                   370                   827 
##             Separated              Divorced         Never-married 
##                   939                  4214                  9726 
##    Married-civ-spouse 
##                 14065
```

```
## 
##       Widowed     Unmarried Never-married       Married 
##           827          5523          9726         14086
```


Married
Widowed
Unmarried
Never-married

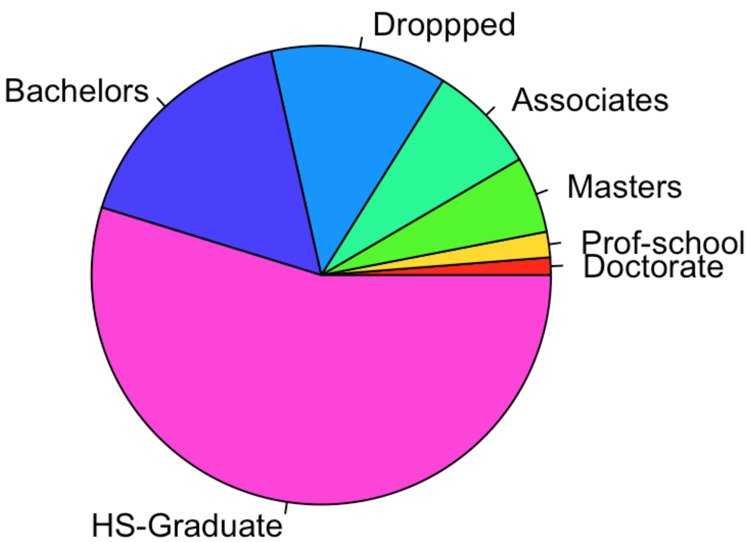

Marital relation distribution for different income levels
Married
Never-married
Unmarried
Widowed
count
7500
5000
2500
0
<=50K
>50K
income
marital.status
Married
Never-married
Unmarried
Widowed

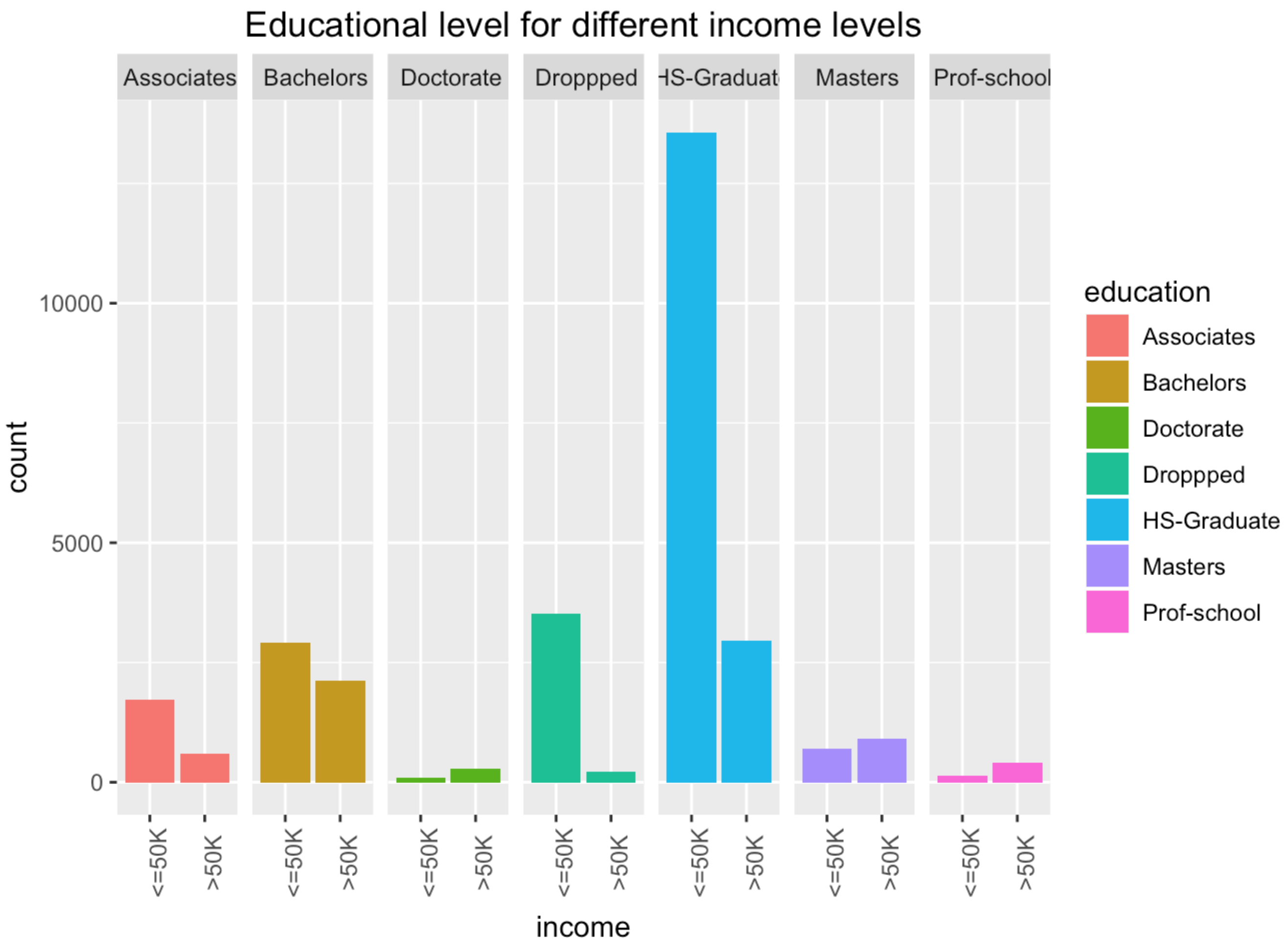
Education

Similar to previous section, I will group them like people of educational standard as *5th-6th, 7th-8th, 9th, 10th, 11th, 12th* are **Dropped**. While *Assoc-acdm, Assoc-voc* are **Associates** and *HS-grad, Some-college* are **HS-Graduate**. Here, we can see that most of the HS-Graduate and Dropped people are working with income less than 50K. While for Bachelor people, almost half of them have income greater than 50K and other half have income less than 50K.

```
##
##     Preschool        1st-4th        5th-6th      Doctorate           12th
##            45            151            288            375            377
##           9th    Prof-school        7th-8th           10th     Assoc-acdm
##           455            542            557            820           1008
##          11th      Assoc-voc        Masters      Bachelors   Some-college
##          1048           1307           1627           5044           6678
##       HS-grad
##          9840
```

```
##
##    Doctorate  Prof-school      Masters   Associates     Droppped
##          375          542         1627         2315         3741
##    Bachelors  HS-Graduate
##         5044        16518
```



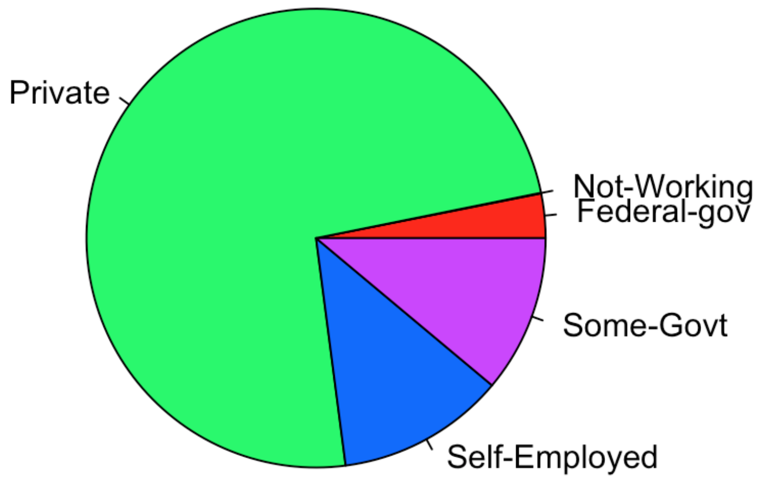


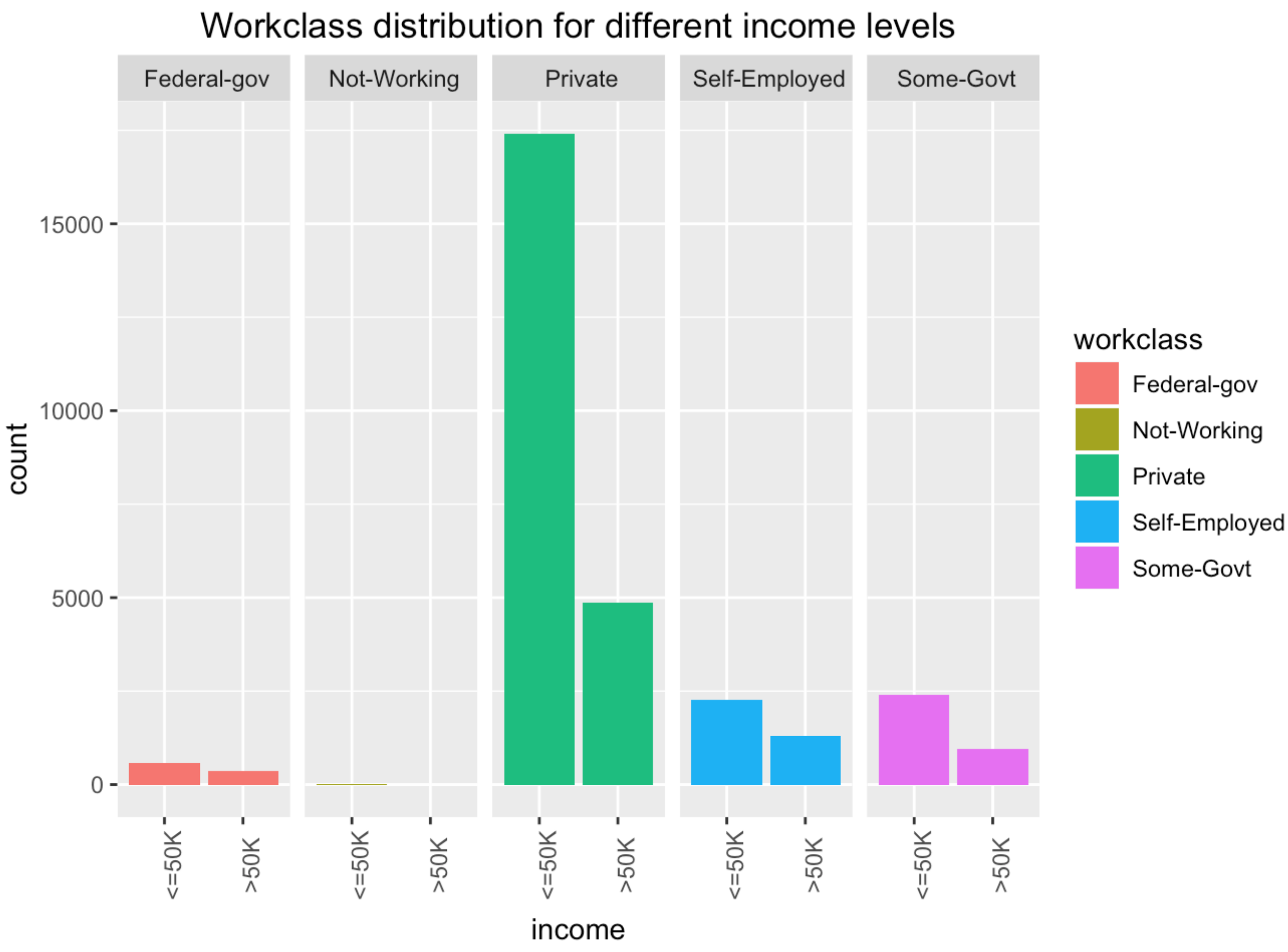
Workclass

Again, we'll group people based on their workclass like people having *Local-gov, State-gov* are **Some-Govt** people. Similarly, *Self-emp-inc, Self-emp-not-inc* are **Self-Employed** people, and *Without-pay, Never-worked* are **Not-Working** people. Now, we can see that most of the people have private job. Also, the distribution graph shows that many of them are working with income less than 50K.

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

```
## 
##   Not-Working   Federal-gov     Some-Govt Self-Employed       Private 
##            14           943          3346          3573         22286
```


Private
Not-Working
Federal-gov
Some-Govt
Self-Employed

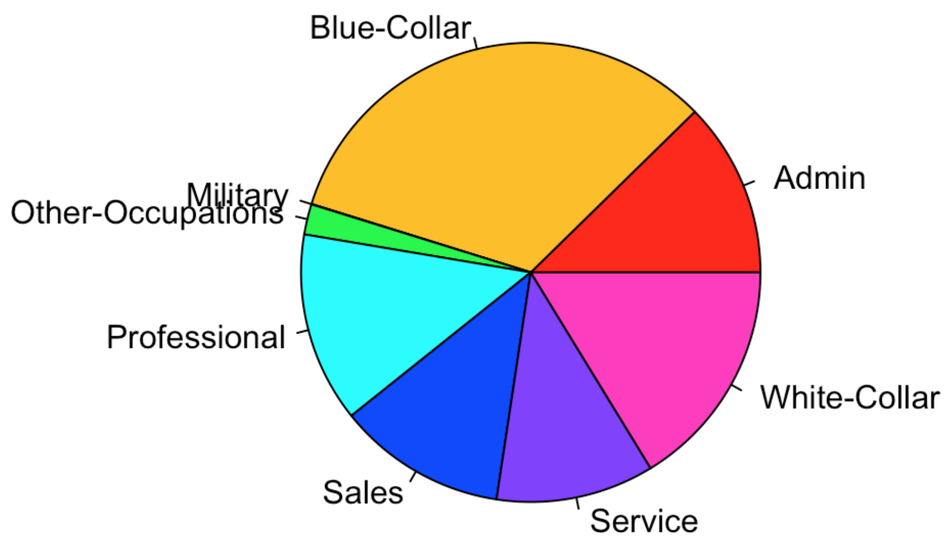

Workclass distribution for different income levels
Federal-gov
Not-Working
Private
Self-Employed
Some-Govt
count
15000
10000
5000
0
<=50K
>50K
income
workclass
Federal-gov
Not-Working
Private
Self-Employed
Some-Govt

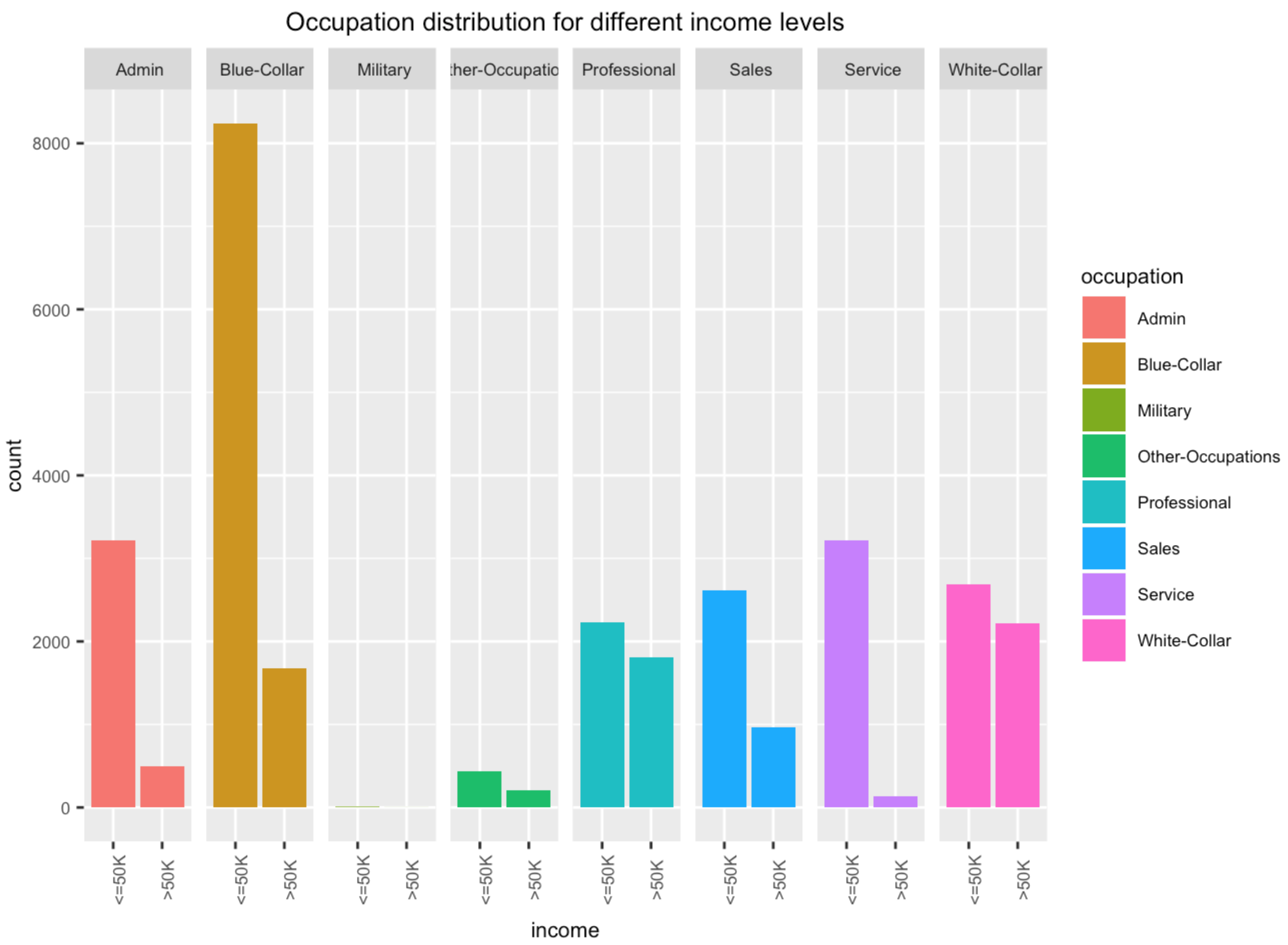
Occupation

Here, I am grouping people based on their occupation. People having occupation as *Craft-repair, Farming-fishing, Handlers-cleaners, Machine-op-inspct, Transport-moving* can be considered as **Blue-Collar** people. Similarly, we can group *Exec-managerial, Tech-support* people as **white collar**, and *Other-service, Priv-house-serv* people as **Service**. Mostly, Blue-collar people are working approximate 30%. From income distribution graph we can see that most of the people having job as Blue-collar, Admin and Service are working with income less than 50K.

```
## 
##                 ?      Armed-Forces   Priv-house-serv 
##                 0                 9               143 
##   Protective-serv      Tech-support   Farming-fishing 
##               644               912               989 
##  Handlers-cleaners Transport-moving Machine-op-inspct 
##              1350              1572              1966 
##     Other-service             Sales      Adm-clerical 
##              3212              3584              3721 
##   Exec-managerial      Craft-repair    Prof-specialty 
##              3992              4030              4038
```

```
## 
##          Military Other-Occupations           Service 
##                 9               644              3355 
##             Sales             Admin      Professional 
##              3584              3721              4038 
##      White-Collar       Blue-Collar 
##              4904              9907
```


Blue-Collar
Admin
Military
Other-Occupations
Professional
White-Collar
Sales
Service

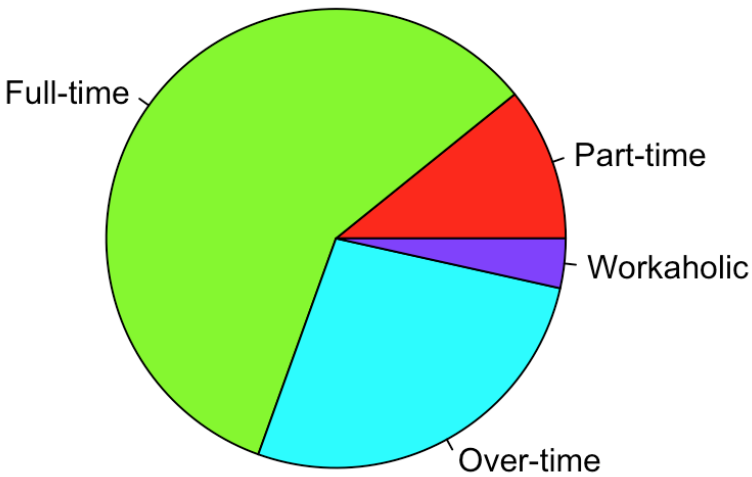

Occupation distribution for different income levels
Admin
Blue-Collar
Military
ther-Occupatio
Professional
Sales
Service
White-Collar
count
8000
6000
4000
2000
0
<=50K
>50K
income
occupation
Admin
Blue-Collar
Military
Other-Occupations
Professional
Sales
Service
White-Collar

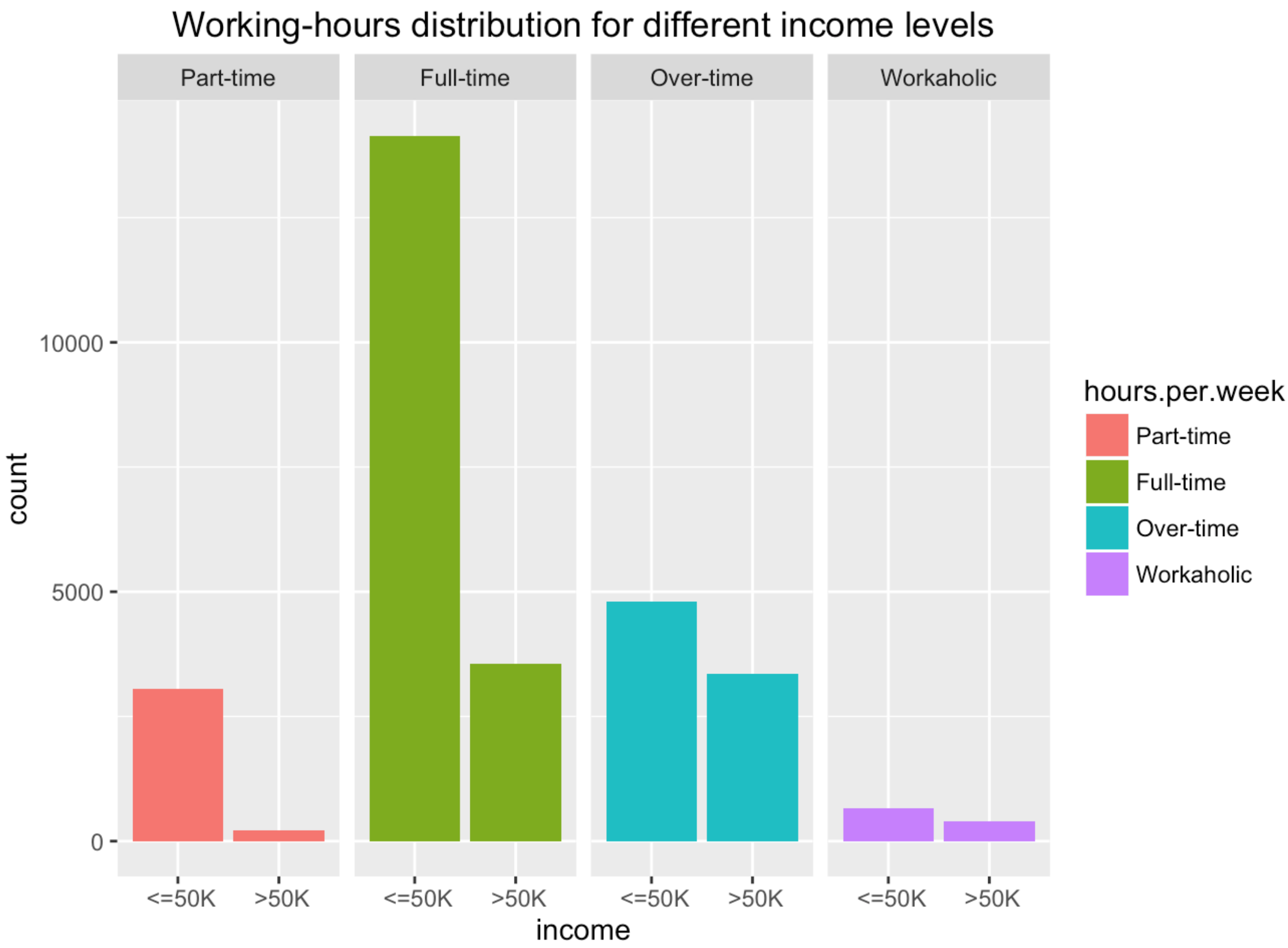
Hours-per-Week

For this variable, we can see that maximum working-hour is 99 and minimum is 1 hour. I will group people into **Part-time, Full-time, Over-time** and **Workaholic**. Most of the people has Full-time job with income less than 50K. Likewise, most of part-time people are working with income less than 50K.

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

```
##
## Workaholic  Part-time  Over-time  Full-time
##       1052       3261       8145      17704
```





Dealing with the Test dataset

I will also follow the same steps mentioned above with the test data for building our prediction models. Now, we can see the dimesions of test dataset.

```
##
## FALSE  TRUE
##  1221 15060
```

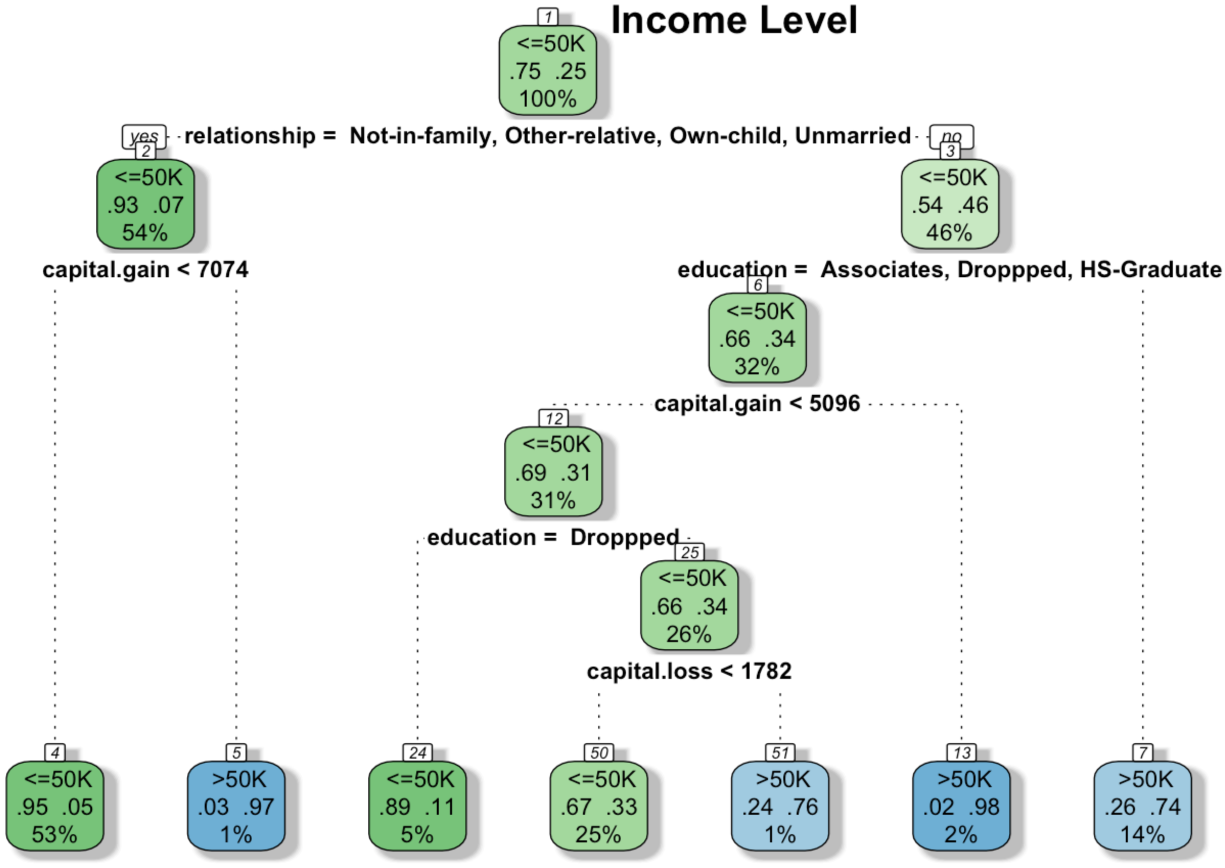
Building Prediction Models

To predict income for test dataset, I will use two prediction Models as:
A. Decision-Tree Prediction Model
B. Naive-Bayes Model

Decision-Tree Prediction Model

In this model, first we need to make a tree based on which I will predict the income for test dataset. I will consider variables age, workclass, education, capital.gain, capital.loss, relationship, sex, race, hours.per.week as important factor for income prediction. Here, I am using rpart library which uses a feature selection methodology. It selects some predictors to build the decision-tree.

```
## n= 30162
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 30162 7508  <=50K (0.75107751 0.24892249)
##    2) relationship= Not-in-family, Other-relative, Own-child, Unmarried 1629
3 1135  <=50K (0.93033818 0.06966182)
##      4) capital.gain< 7073.5 15993  845  <=50K (0.94716438 0.05283562) *
##      5) capital.gain>=7073.5 300   10  >50K (0.03333333 0.96666667) *
##    3) relationship= Husband, Wife 13869 6373  <=50K (0.54048598 0.45951402)
##      6) education= Associates, Droppped, HS-Graduate 9719 3322  <=50K (0.658
19529 0.34180471)
##       12) capital.gain< 5095.5 9219 2831  <=50K (0.69291680 0.30708320)
##         24) education= Droppped 1442  153  <=50K (0.89389736 0.10610264) *
##         25) education= Associates, HS-Graduate 7777 2678  <=50K (0.65565128
0.34434872)
##           50) capital.loss< 1782.5 7455 2434  <=50K (0.67350771 0.32649229)
*
##           51) capital.loss>=1782.5 322   78  >50K (0.24223602 0.75776398) *
##       13) capital.gain>=5095.5 500    9  >50K (0.01800000 0.98200000) *
##      7) education= Bachelors, Doctorate, Masters, Prof-school 4150 1099  >50
K (0.26481928 0.73518072) *
```



Rattle 2016-May-24 01:24:36 Arun

In the above decision tree we can see the probability and percentage of income distribution. For example, tree shows that people having relations as Not-in-family, Other-relative, Own-child, or Unmarried are 55%, while other married people are 46%. If we consider those 55% people, out of those, people with capital gain less than 7074, has income <=50K. Similary, if we consider 46% married people, we again put some rules on them like capital.gain and capital.loss. For income prediction, if we have probability 0.5 at least, then it will predict income less than 50K otherwise it will predict income greater than 50K. Now, we compare this predicted income with the given income for test data, and see how many times it predict correct income. So that we can see the accuracy of our model.

```
## [1] "The Decision tree model predicted the income of adult.test dataset with
84.5 % of accuracy."
```

Naive-Bayes Model

The Naive-Bayes model classify entities based on *conditional probability* concept. Or in other words, it will find the probability of something being happen, based on something else has already happened. In the adult dataset we are going to predeict income by making decision rules (Bayes rules) which seems most probable. I am classifying income as less/equal to 50K and greater than 50K based on provided features (age, education, relationship, marital.status, capital.gain etc). Naive-Bayes model has an bayesian eqaution to calculate posterior probability for these class. For exapmple income of a person is less than 50K if he is young HS-graduate with private job. Similarly, this model contains probability of other bayes rules. We can also see the naive-bayes grouping tables and levels as follows:

```
## $age
##          var
## grouping      Young Middle-aged    Senior        Old
##    <=50K 0.24529884   0.5081663 0.2178423 0.02869250
##    >50K  0.01478423   0.5651305 0.3949121 0.02517315
##
## $workclass
##          var
## grouping  Federal-gov  Not-Working   Private  Self-Employed  Some-Govt
##    <=50K   0.02551426 0.0006179924 0.7685177     0.09971749  0.1056326
##    >50K    0.04861481 0.0000000000 0.6494406     0.17501332  0.1269313
##
## $education
##          var
## grouping  Associates  Bachelors  Doctorate   Droppped  HS-Graduate
##    <=50K  0.07570407  0.1288073 0.00419352 0.15520438    0.5987905
##    >50K   0.07991476  0.2831646 0.03729355 0.02996803    0.3933138
##          var
## grouping   Masters  Prof-school
##    <=50K 0.0312969  0.006003355
##    >50K  0.1222696  0.054075653
##
## $marital.status
##          var
## grouping   Married  Never-married  Unmarried    Widowed
##    <=50K 0.3388806     0.40858127  0.2195639 0.03297431
##    >50K  0.8536228     0.06259989  0.0731220 0.01065530
##
## $occupation
##          var
## grouping      Admin  Blue-Collar     Military  Other-Occupations
##    <=50K 0.14227068    0.3636002 0.0003531385         0.01915776
##    >50K  0.06632925    0.2224294 0.0001331913         0.02797017
##          var
## grouping  Professional     Sales    Service  White-Collar
##    <=50K    0.09830494 0.1153880 0.14222654     0.1186987
##    >50K     0.24120938 0.1291955 0.01771444     0.2950186
##
## $relationship
##          var
## grouping   Husband  Not-in-family  Other-relative   Own-child  Unmarried
##    <=50K 0.2994615      0.3047144     0.037697537 0.194314470 0.13238280
##    >50K  0.7563932      0.1096164     0.004661694 0.008524241 0.02836974
##          var
## grouping       Wife
##    <=50K 0.03142933
##    >50K  0.09243474
##
## $race
##          var
## grouping  Amer-Indian      Asian     Black       Other     White
##    <=50K  0.011123863 0.02856008 0.1081928 0.009269886 0.8428534
##    >50K   0.004528503 0.03303143 0.0487480 0.002797017 0.9108950
```

```
##
## $sex
##         var
## grouping    Female      Male
##    <=50K 0.3827139 0.6172861
##    >50K  0.1481087 0.8518913
##
## $capital.gain
##              [,1]       [,2]
##  <=50K  148.8938   936.3923
##  >50K  3937.6798 14386.0600
##
## $capital.loss
##             [,1]     [,2]
##  <=50K  53.4480 310.2703
##  >50K  193.7507 592.8256
##
## $hours.per.week
##         var
## grouping  Part-time Full-time Over-time Workaholic
##    <=50K 0.13472234 0.6244372 0.2117065 0.02913393
##    >50K  0.02783697 0.4738945 0.4460575 0.05221097
##
## $native.country
##         var
## grouping  ?     Cambodia      Canada        China     Columbia         Cuba
##    <=50K  0 0.0004855655 0.003134104 0.002118831 0.0023836850 0.002957535
##    >50K   0 0.0009323388 0.004794885 0.002663825 0.0002663825 0.003329782
##         var
## grouping  Dominican-Republic      Ecuador  El-Salvador     England
##    <=50K         0.0028692505 0.0010152732  0.004016951 0.002471970
##    >50K          0.0002663825 0.0005327651  0.001198721 0.003995738
##         var
## grouping       France     Germany       Greece    Guatemala        Haiti
##    <=50K 0.0006621347 0.003707954 0.0009269886 0.0026485389 0.0016774080
##    >50K  0.0015982952 0.005860416 0.0010655301 0.0003995738 0.0005327651
##         var
## grouping  Holand-Netherlands     Honduras         Hong      Hungary
##    <=50K         4.414231e-05 0.0004855655 0.0005738501 0.0004414231
##    >50K          0.000000e+00 0.0001331913 0.0007991476 0.0003995738
##         var
## grouping       India         Iran       Ireland       Italy     Jamaica
##    <=50K 0.002648539 0.001059416 0.0008387040 0.001942262 0.003089962
##    >50K  0.005327651 0.002397443 0.0006659563 0.003196590 0.001331913
##         var
## grouping       Japan         Laos       Mexico    Nicaragua
##    <=50K 0.001589123 0.0006621347 0.025470116 0.0013684118
##    >50K  0.003063399 0.0002663825 0.004395312 0.0002663825
##         var
## grouping  Outlying-US(Guam-USVI-etc)         Peru  Philippines      Poland
##    <=50K                0.0006179924 0.0012359848  0.005650216 0.001986404
##    >50K                 0.0000000000 0.0002663825  0.007991476 0.001465104
##         var
## grouping      Portugal  Puerto-Rico     Scotland        South      Taiwan
```

```
##     <=50K 0.0013242694  0.004281805 0.0003972808 0.002516112 0.001015273
##     >50K  0.0005327651  0.001598295 0.0002663825 0.001864678 0.002530634
##          var
## grouping     Thailand  Trinadad&Tobago  United-States      Vietnam
##     <=50K 0.0006179924     0.0007062770      0.9053147 0.0026043966
##     >50K  0.0003995738     0.0002663825      0.9316729 0.0006659563
##          var
## grouping   Yugoslavia
##     <=50K 0.0004414231
##     >50K  0.0007991476
```

```
## grouping
##     <=50K      >50K
## 0.7510775 0.2489225
```

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

Firstly, I predicted the income for test dataset and maintained into a seperate column (pred_income). Then I compared the predicted income with the given income, to check where my model predicted income correctly.

```
## [1] "The Decision tree model predicted the income of adult.test dataset with
78.85 % of accuracy."
```

Conclusion

I worked on adult income dataset to build prediction models. I realized that variables as age, education, marital-status, workclass, capital-gain are good factor to predict income. So, I did some manipulation with my dataset by making some groups inside some variables. I build two models as Decision-Tree model and Naive-Bayes Model. The income predicted from Decision-Tree model is much accurate (85 %) than the Naive-Bayes model (79 %).