Use following file for this practice assignment : "census_income.csv" . Many questions already contain expected output. What you need to do is to write the codes which will result in those outputs

Quick Summary

Take a quick look at the data. Generate quick summary for the numeric variables which looks like this:

```
##                vars     n      mean        sd median   trimmed      mad
## age               1 32561     38.58     13.64     37     37.69    14.83
## fnlwgt            2 32561 189778.37 105549.98 178356 180802.36 88798.84
## education.num     3 32561     10.08      2.57     10     10.19     1.48
## capital.gain      4 32561   1077.65   7385.29      0      0.00     0.00
## capital.loss      5 32561     87.30    402.96      0      0.00     0.00
## hours.per.week    6 32561     40.44     12.35     40     40.55     4.45
##                   min     max   range  skew kurtosis     se
## age                17      90      73  0.56    -0.17   0.08
## fnlwgt          12285 1484705 1472420  1.45     6.22 584.94
## education.num       1      16      15 -0.31     0.62   0.01
## capital.gain        0   99999   99999 11.95   154.77  40.93
## capital.loss        0    4356    4356  4.59    20.37   2.23
## hours.per.week      1      99      98  0.23     2.92   0.07
```

Next, write a for loop to generate summary of categorical variables. [Individual frequency counts]. Output should look like this : [Only two variable outcome is shown, you need to generate it for all the variables]

```
## [1] "Summary for  workclass"
##
##                 ?       Federal-gov         Local-gov      Never-worked
##              1836               960              2093                 7
##           Private      Self-emp-inc  Self-emp-not-inc         State-gov
##             22696              1116              2541              1298
##       Without-pay
##                14
## [1] "Summary for  education"
##
##          10th          11th          12th       1st-4th       5th-6th
##           933          1175           433           168           333
##       7th-8th           9th    Assoc-acdm     Assoc-voc     Bachelors
##           646           514          1067          1382          5355
##     Doctorate       HS-grad       Masters     Preschool   Prof-school
##           413         10501          1723            51           576
##  Some-college
##          7291
```

Similar Categories

You'll study your predictive modelling modules that your data needs to be numeric for applying any predictive modelling technique [few exceptions such as Decision Tress are there]. Categorical variables are converted to dummy variables to deal with this. You make n-1 dummy variables for a categorical variable which takes n

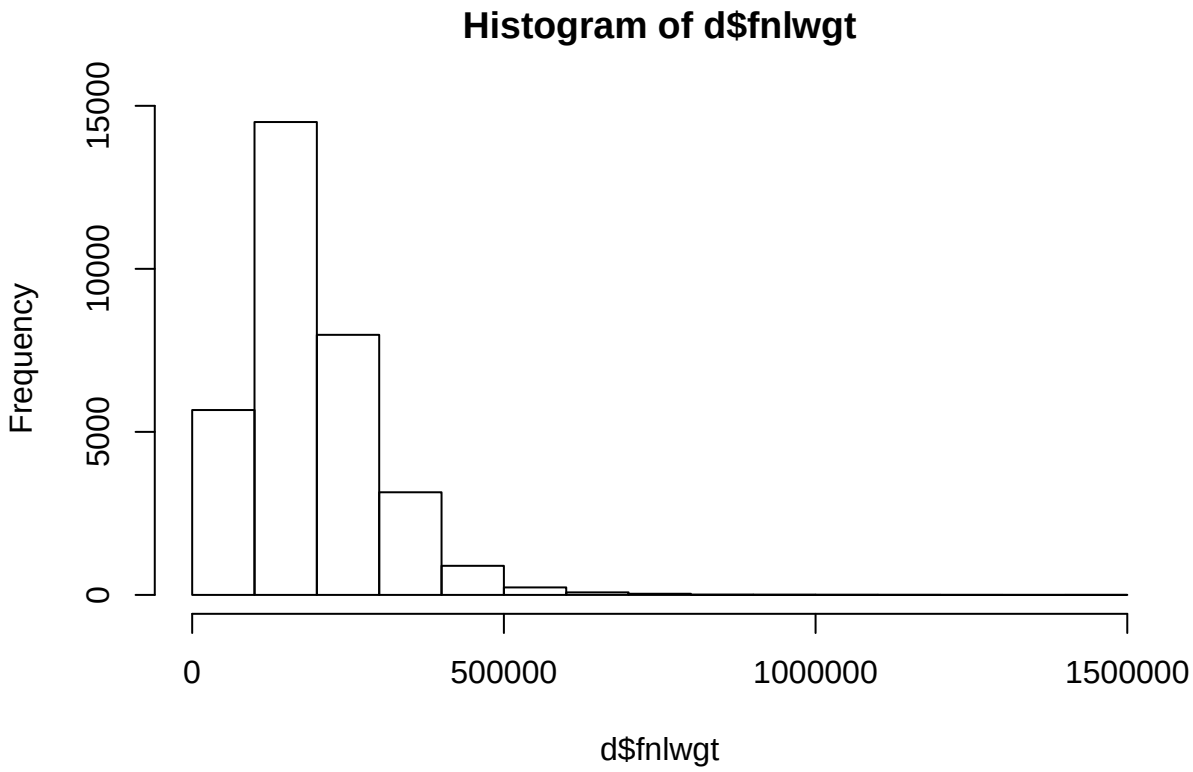
distinct values. Sometimes you can bring down the number of dummy variables that you need to create by finding similar categories for the categorical variables and treating them as one. This discovery is enabled by cross table between categorical variable and target [which is also categorical].

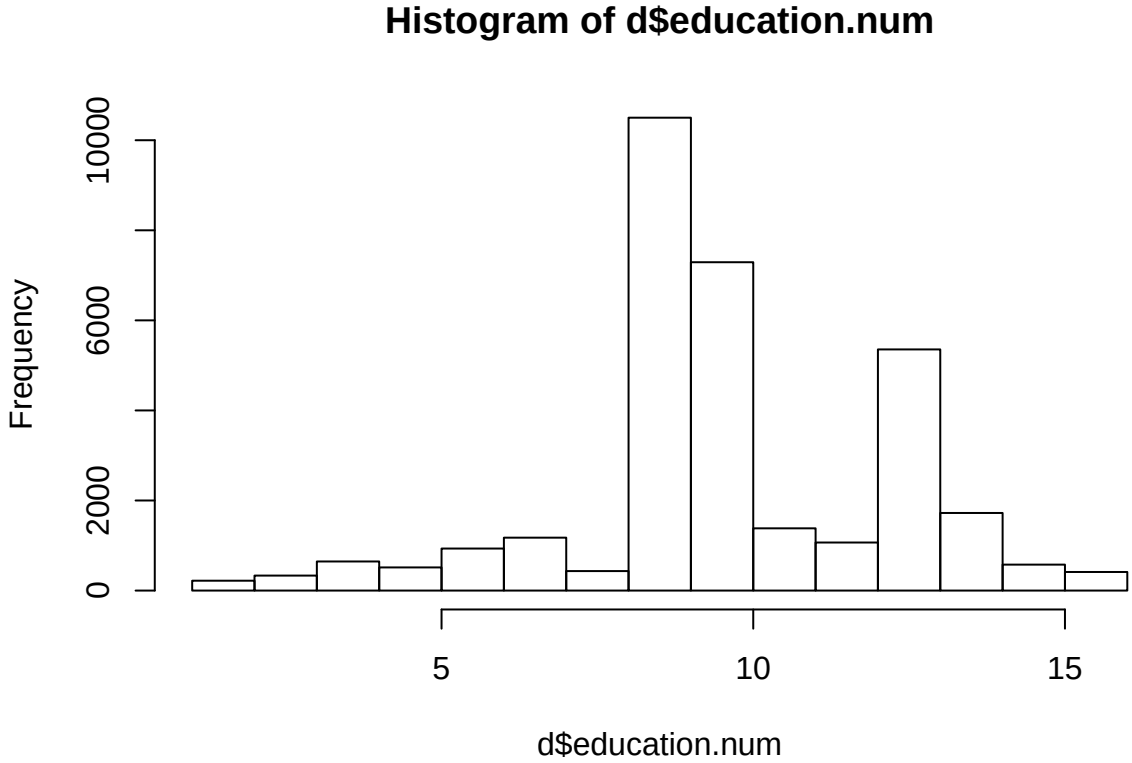
For this purpose prepare a cross table between variable education & Y. This needs to be a percentage cross table where row percentages should add up to 1. output should look like this:

```
##
##                  <=50K  >50K
##    10th           0.93  0.07
##    11th           0.95  0.05
##    12th           0.92  0.08
##    1st-4th        0.96  0.04
##    5th-6th        0.95  0.05
##    7th-8th        0.94  0.06
##    9th            0.95  0.05
##    Assoc-acdm     0.75  0.25
##    Assoc-voc      0.74  0.26
##    Bachelors      0.59  0.41
##    Doctorate      0.26  0.74
##    HS-grad        0.84  0.16
##    Masters        0.44  0.56
##    Preschool      1.00  0.00
##    Prof-school    0.27  0.73
##    Some-college   0.81  0.19
```

Finding Outliers

Plot histogram for variables fnlwgt and education.num.





As you can see that these are skewed distributions of values and if you were looking for outliers; a simple $\mu \pm 3 * \sigma$ limits will not work. Find q1,q2 and IQR values for these variables and use following limits to report number of outliers according to each variable : [q1 - 1.5*IQR, q3 + 1.5*IQR].

HINT : Use function "quantile" to find q1 and q3 which are nothing but 25 and 75 percentiles of the data.

Your Results should be as follows

```
## [1] "Outlier Limits For fnlwgt are :"

## [1] -61009 415887

## [1] "Number of outliers according to these limits for fnlwgt:"

## [1] 0

## [1] "Outlier Limits for education.num are :"

## [1]  4.5 16.5

## [1] "Number of outliers according to these limits for education.num:"

## [1] 0
```

Also see what would be the result if you go by $\mu \pm \ 2 * \sigma$ limits.