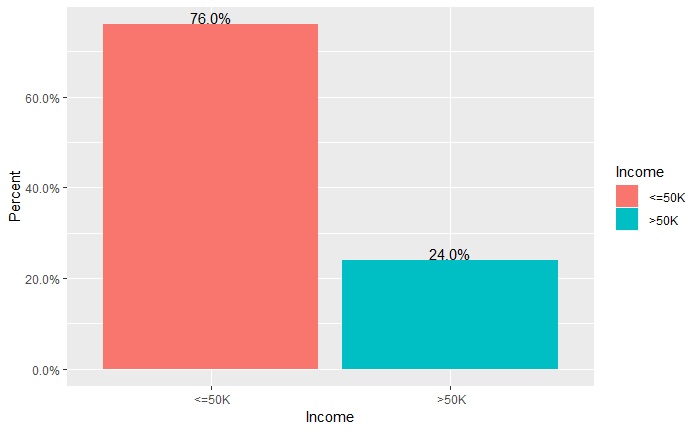
**Part 1: Exploratory**

**Variables:**

**#1** I loaded the data with read.csv, multiple variables from the data are read as character variable. I found that the original data variable names (education-num”, "hours-per-week", "native-country","capital-gain", "capital-loss",) are not eligible for R variable name. So I attempted the changed the variable names (“education\_num", "capital\_gain", "capital\_loss", "hours\_per\_week", "native\_country") and converted them into factor for the analysis.

**#2** Outcome

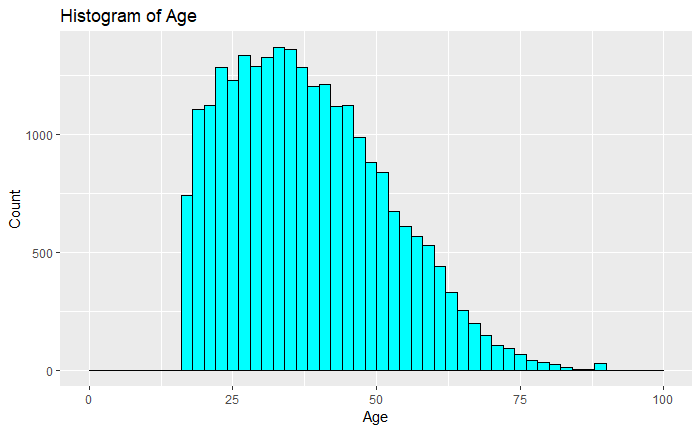
"income": two level factor variable, ‘<50K’ and ‘>=50K’



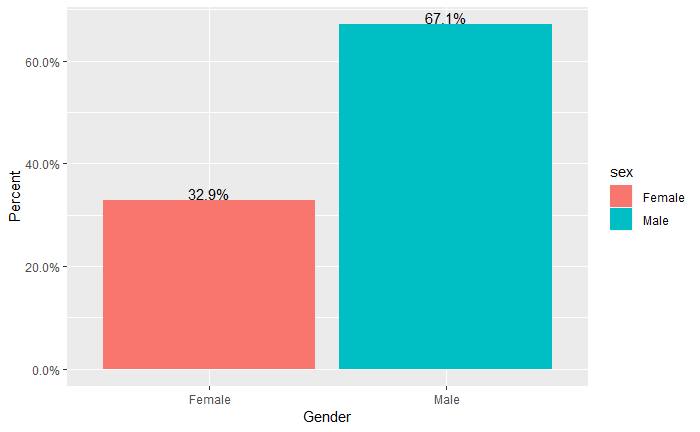
From train data set, about 76% samples form train dataset have an income less or equal to 50K while 24% have an income more than 50K.

**#3** Summary description of numeric predictors:

“**age”:** From train data set, the minimum sample age is 17, and the minimum sample age is 90. The median age is 37, and the mean age is 38.6.



**“Sex”**: From train data set, 67.1% sample of train dataset are male, and female only accounts for 32.9%.



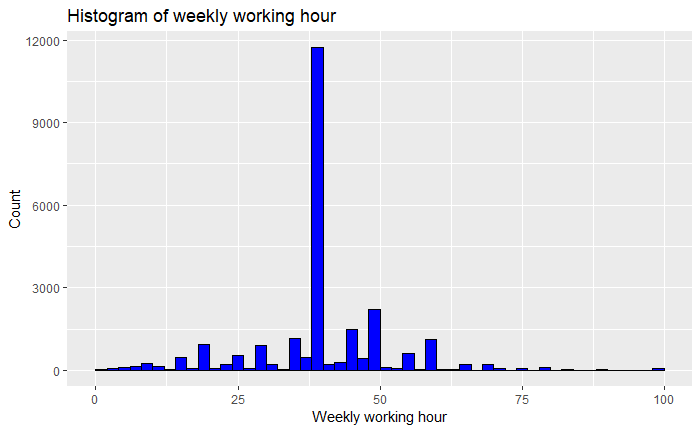
**“education\_num”:** education\_number is based on education factor variables, they are identical.

**“fnlwgt”:** this is weighted number for each respondent.

**"capital\_gain":** The mean capital gain is 1074. Most of people don't have capital gain.

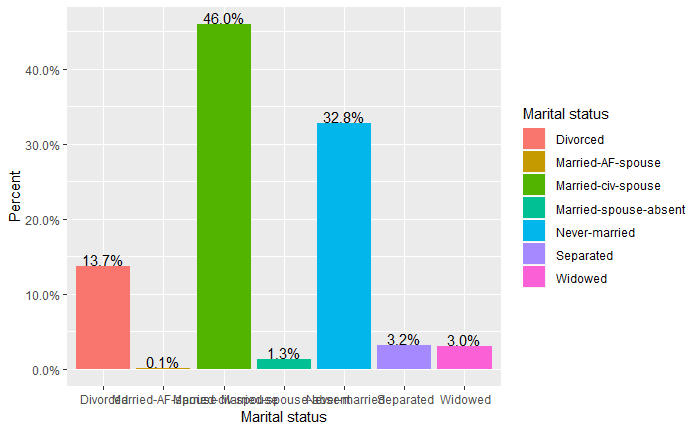
**"capital\_loss":**

**"hours\_per\_week":** The median work hour is 40, and average working hour is 40 hour.



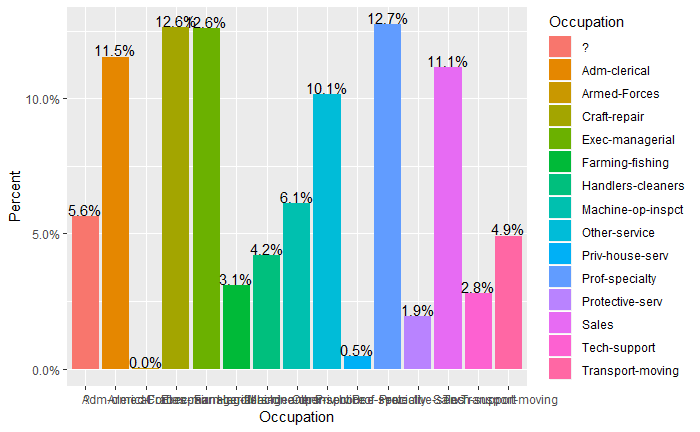
**#4** **Categorical predictors**

**“Marital.status”**

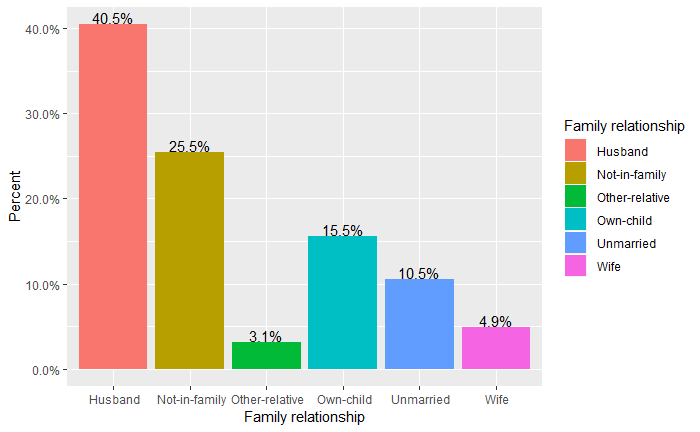


From above table from train data set, we can see 46% people are married couple, 33% are never married, 14% people are divrced.

**"occupation"**: a factor with levels includes dm-clerical’, ‘Armed-Forces’, ‘ Craft-repair’, ‘ Exec-managerial’, ‘ Farming-fishing’, ‘ Handlers-cleaners’, ‘ Machine-op-inspct’, ‘ Other-service’, ‘ Priv-house-serv’, ‘ Prof-specialty’, ‘ Protective-serv ’, ‘Sales’, ‘ Tech-support’, ‘ Transport-moving’.

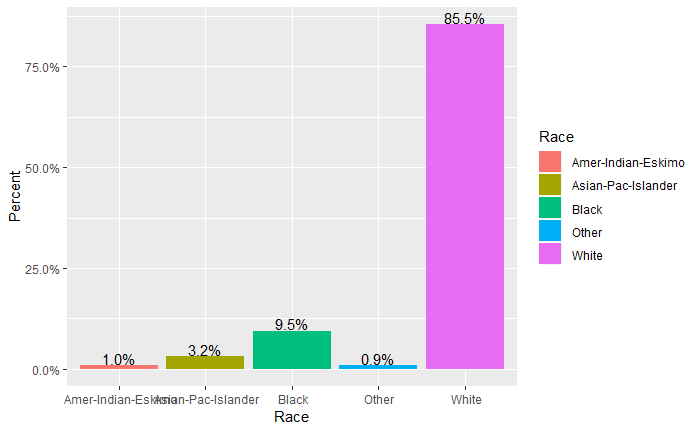


**"relationship":** A factor with Divorced Married-AF-spouse Married-civ-spouse Married-spouse-absent Never-married Separated Widowed.



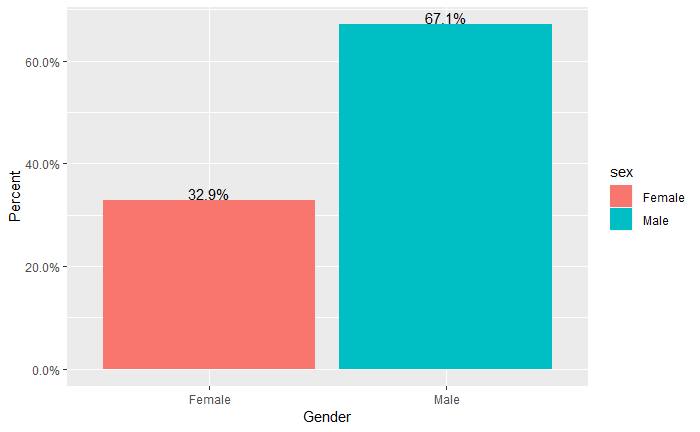
40% family are husband and wife, 25% are not in a family relationship.

**"race":** a factor with levels, include Amer-Indian-Eskimo, Asian-Pac-Islander, Black, Other, White



From above table from train data set, we can see 85% people are white, 9% are black.

**"sex":** a two level factor with male and female.



From train data set, 67.1% sample of train dataset are male, and female only accounts for 32.9%.

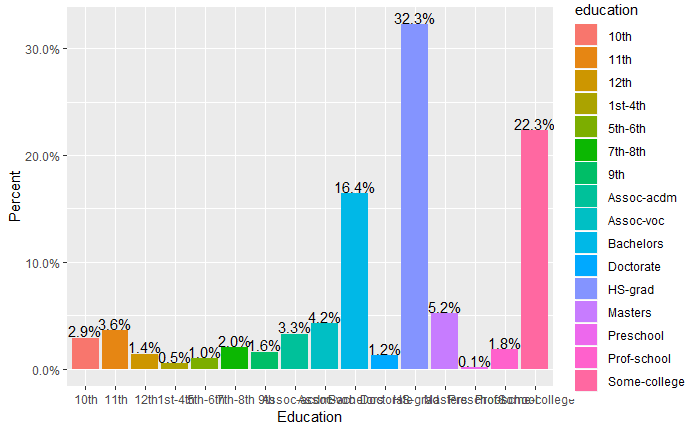
**"native\_country":** a factor with levels include ‘Cambodia’, ‘Canada’, ‘China’, ‘Columbia’, ‘ Cuba Dominican-Republic’, ‘ Ecuador’, ‘ El-Salvador’, ‘ England’, ‘ France ’, ‘Germany’, ‘ Greece’, ‘Guatemala’, ‘ Haiti’, ‘ Holand-Netherlands’, ‘ Honduras’, ‘ Hong’, ‘ Hungary’, ‘ India ’, ‘Iran’, ‘Ireland ’, ‘Jamaica’, ‘ Japan’, ‘ Laos Mexico’, ‘ Nicaragua’, ‘ Outlying-US(Guam-USVI-etc) ’, ‘ Peru Philippines’, ‘’, ‘ Poland’, ‘ Portugal’, ‘ Puerto-Rico’, ‘’, ‘ Scotland’, ‘ South Taiwan’, ‘ Thailand’, ‘ Trinadad&Tobago’, ‘ United-States’, ‘ Vietnam ’, ‘Yugoslavia’, ‘

“workclass”: Factor with 9 levels includes: Federal-gov , Local-gov, Never-worked, Private, Self-emp-inc, Self-emp-not-inc, State-gov, Without-pay.

89.6% people form train data set is originally come from United states.

**“education”** : ordered factor with 16 levels. I converted it into an ordered factors;

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



From the summary table, the top three levels are: 32.3% of sample have high school education, 22.3% have some college education, and 16.4% have a bachelor degree.

**The exploratory analysis of response variable and priditors**

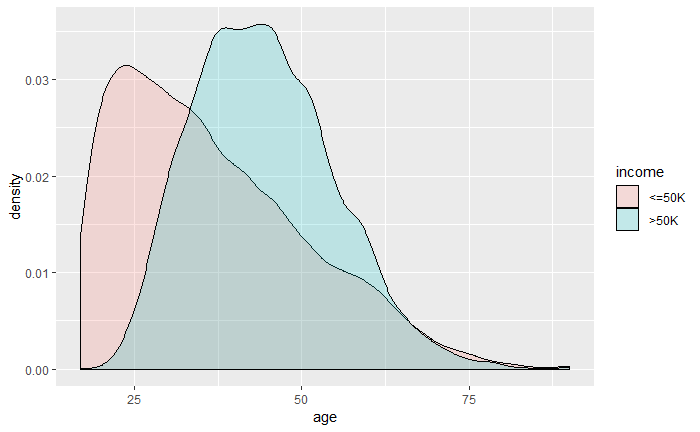
**The work class and income gap**

Based on above test, at the 0.05 significance level, there is strong indication that “workclass” and “income” are correlated. Almost 70% peope working in fedaral govement make more than 50K. 34% peope working in private entity make more than 50K.

**The age and income gap**

At the 0.05 significance level, T test significantly different.

We see that the majority of people earning more than 50K a year are between 33 and 55 years old, whereas the greater number of people who earn less than 50K a year are between 18 and 45. The mean age of people who make more than 50K are 44 year old, majority of are from 25-60 year old. The mean age of people who make less or equal to 50K are 37 year old, large proportion of them are under 25 year old.



**The work hour per week and income**

At the 0.05 significance level, T test significantly different.

The mean weekly work hour of people who make more than 50K are 46 hours. The mean weekly work hour of people who make less or equal to 50K are 39 hours, large proportion of both groups work 35~40 hour per week.

**The capital gain/loss and income**

The mean capital gain of people who make more than 50K are 4013. The meancapital gain of people who make less or equal to 50K are 146, much less than the higher income class.

The mean capital loss of people who make more than 50K are 200. The mean capital loss of people who make less or equal to 50K are 53. This suggests that people who earn high income have higher capital gain and loss than those of low income. So the high income have certain capital risk, but the average capital gain (4013) is mchc larger than average capital loss (200) the higher income class.

**The native\_region and income**

At the 0.05 significance level, Chi-squared test significantly different.

89.6% people form train data set is originally come from United states.

**The occupation and income**

At the 0.05 significance level, Chi-squared test significantly different.

**The marital status and income**

At the 0.05 significance level, Chi-squared test significantly different.

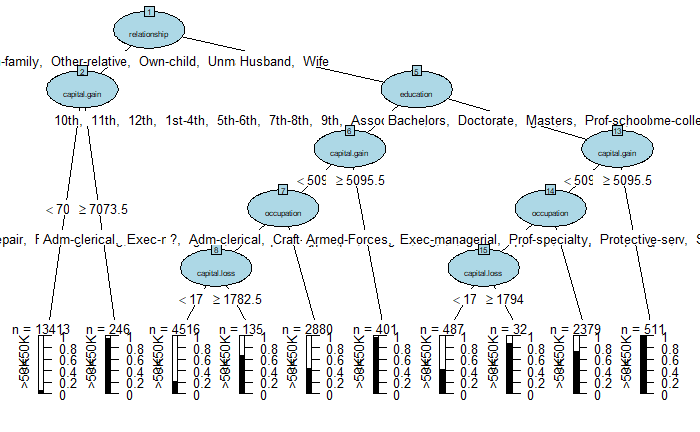
**The relationship and income**

At the 0.05 significance level, Chi-squared test significantly different.

**Part II Prediction models**

**I CART model report**

1. Summary: I converted education as ordered factors. I fitted a CART model with 59 nodes and did the prune with bestCP value 0.0001. And I got 29 nodes tree after the prune. The error rate is (721+2995)/(18281+3003)=0.17. See the prune tree plot as below:



1. The CART model suggests that relationship is the most important factor contributing to binary outcome income. The other predictors such as capital gain, age, education, occupation, capital gain and loss, contribute to income from CART model.
2. I summarized four branches which pridict higher income:

A, node 1-2: the splitting variables are relationship, capital gain and occupation. There is very higher prediction for 246 people who have relationship ( not in a family, other relative, has child), capital gain more than 7073 and occupation as craft-repair will make more than 50K.

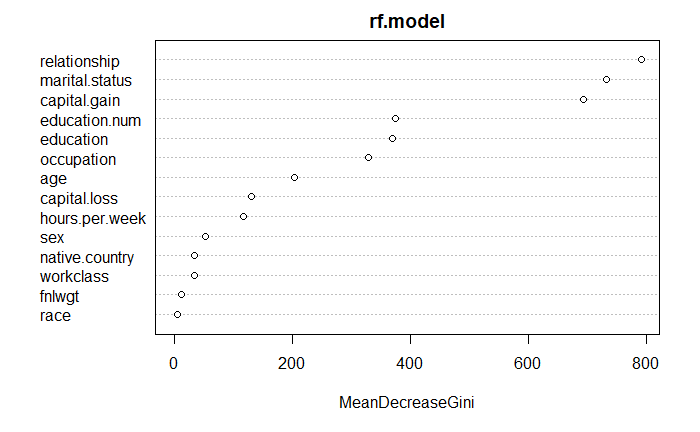
B, node 1-5-13: the splitting variables are relationship, education and capitial gain. There is very higher prediction for 511 people who have relationship (husband and wife), education (association above) and capital gain more than 5059 will make more than 50K.

C, node 1-5-13-14: the splitting variables are relationship, education and capital gain and occupation. There is very higher prediction for 2379 people who have relationship (husband and wife), education (association above), capital gain more than 5059 and occupation with professional specialty will make more than 50K.

D, node 1-5-6: the splitting variables are relationship, education and capital gain. There is very higher prediction for 401 people who have a relationship (husband and wife), education (association above), capital gain more than 5059, will make more than 50K.

**II Random Forest model report**

I fitted a 2000 tree random forest model. The random forest model suggests that relationship is the most important factor contributing to binary outcome income. Then marital status, capital gain are very important variables. Next, education, occupation, age, capital loss and weekly work hours also contribute to the model. But race, workclass, native country and gender contribute less to the model.



**III Generalized linear model:**

1. fitted a generalized linear model. But I found the workclass variable is a trouble and so I removed this variable and also reload the data, and treat education as unordered factors.

2. my model didn’t include education.num as it is identical to education. Therefore, my glm model have 12 predictors.

3. I checked the correlation between predictors, the correlation between relationship and marital\_status is 0.59, Correlation between relationship and sex is 0.42. These are collearilzed variales. These are potential confounders. See below plot.

4. The fitted glm model has an AIC as 15891.8 and BIC as 16631.32, has 31 parameters

5. After the cross validation with test data, the sensitivity of this model to predict the income is 60.34% and the specificity is 93.03%.

