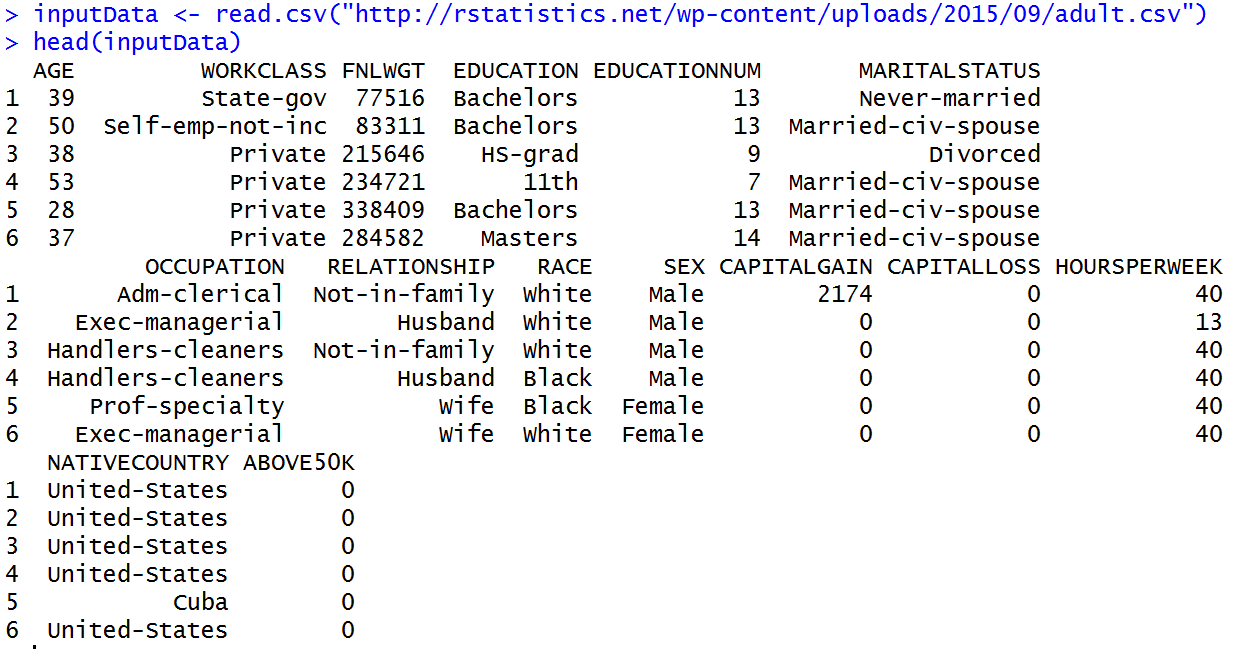
**Module 7: Additional Exercises with Answers**

Predicting Annual Income of Individuals

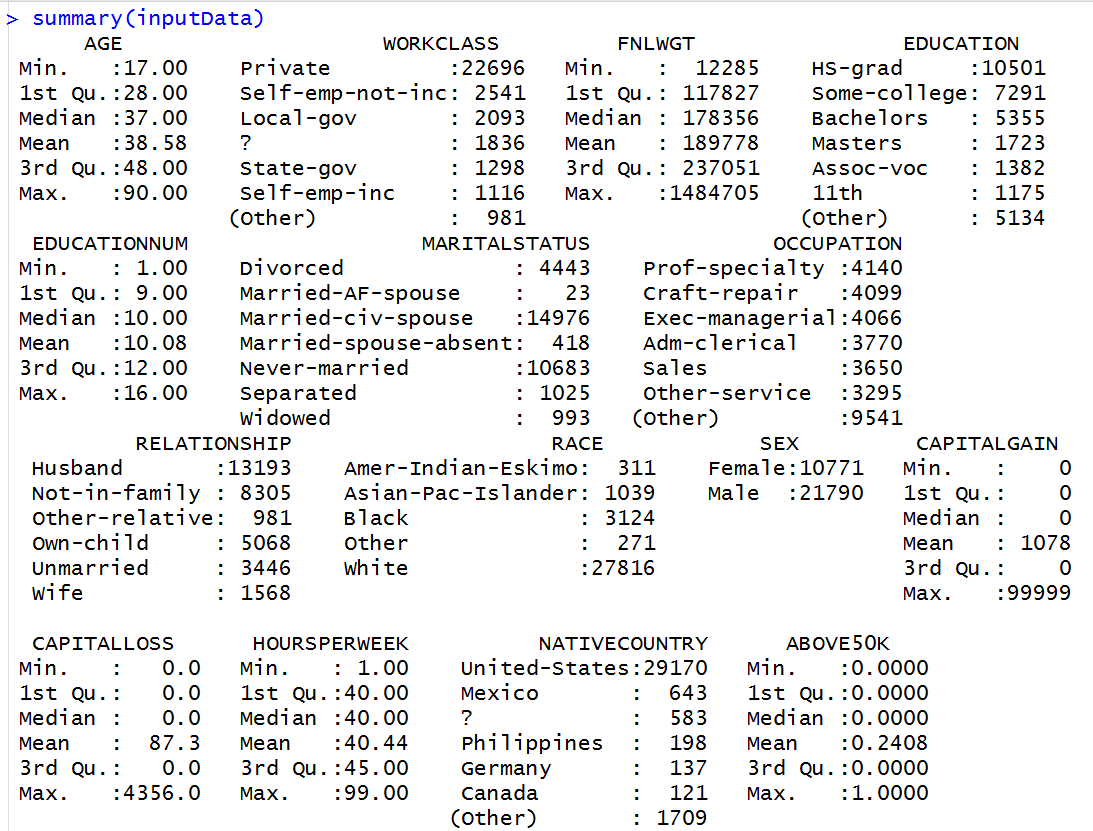
In this example, we will use logistic regression to predict if an individual earns more than $50k in a year or not. First read the ‘Adult’ CSV file using the following command (you need to be connected to the internet and it will take few moments for data to be loaded):

**inputData <- read.csv("http://rstatistics.net/wp-content/uploads/2015/09/adult.csv")**

Let’s have a look at the first 6 records and also a summary of the dataset



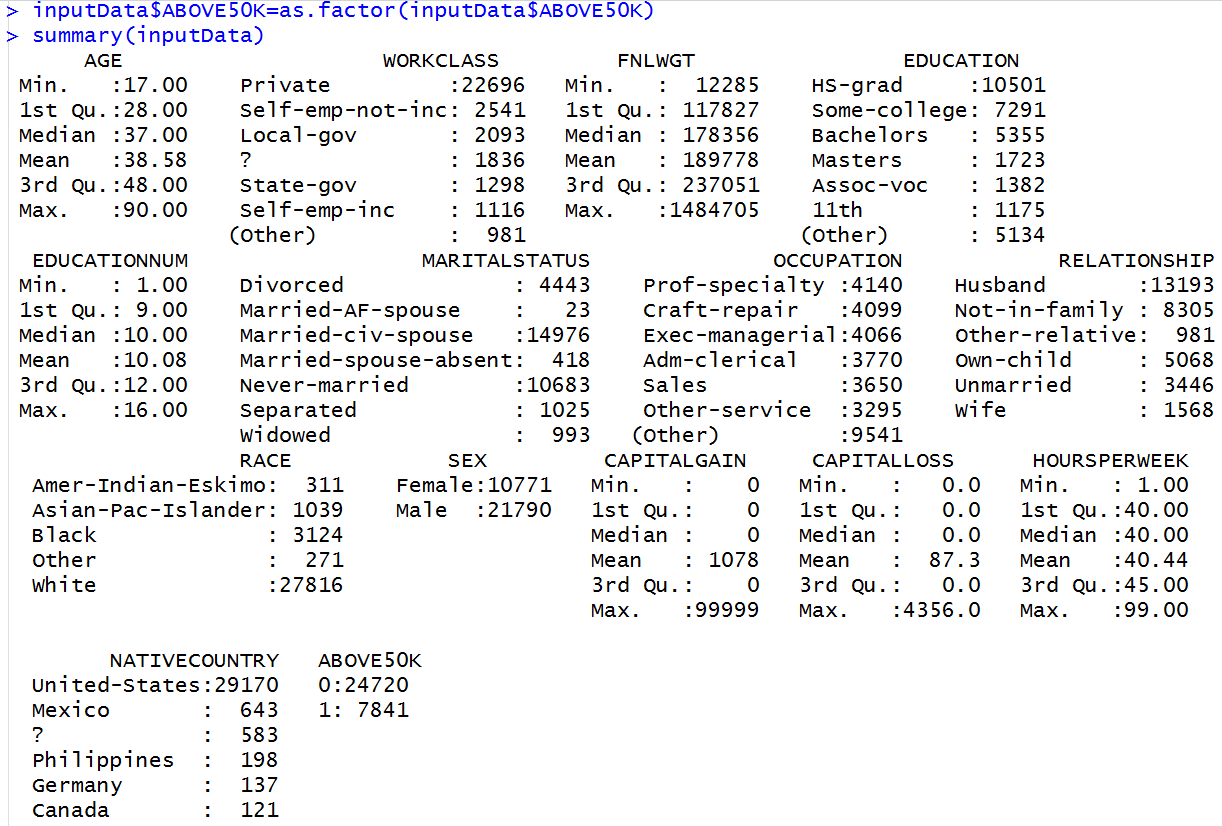
And



The variable ‘ABOVE50K’ is the variable that we are trying to predict. Currently, the variable is coded as a numeric variable that takes 0 and 1s. To use logistic regression, that is a classification method, we need to convert this variable to a factor (i.e. categorical variable):

inputData$ABOVE50K=as.factor(inputData$ABOVE50K)

let’s look at the summary again:



**Questions**

**Q1. Build a decision trees model based on AGE, EDUCATIONNUM (Number of years of Education), SEX and HOURSPERWEEK (number of hours worked per week) to predict ABOVE50K.**

**Q2. How do we judge the statistical significance and the importance of variables in decision tree models?**

**Q3. How does the probability of earning above $50k changes with these variables?**

**Q4. James and Hannah are two individuals. Given the information below, what is the probability that each of them is earning more than $50k a year?**

**Hannah: Female, 25 Year, 16 Years of Education, Working part time 25 hours a week   
James: Male, 31 Year, 16 Years of Education, Working 44 hours a week**

**Q5. What is the accuracy of this model in terms of Area Under Curve (AUC) of ROC?**