ASSIGNMENT ONE

1.Using the ncol and nrow commands to find out how many rows and columns are in this dataset.

```
num_col<- ncol(census_data)
num_col
```

15 columns

```
num_rows <- nrow(census_data)
num_rows
```

32561 rows

2. Using the colnames command, list the variables in this dataset.

```
col_names <- colnames(census_data, do.NULL = FALSE)
col_names
```

```
"Age"            "Workclass"      "fnlwgt"         "Education"
"Education.Num"  "Marital.Status" "Occupation"     "Relationship"
"Race"           "Sex"            "Capital.Gain"   "Capital.Loss"
"Hours.Per.Week" "Native.Country" "Income.Level"
```

3. Using the “$” shortcut and the mean and median commands, calculate the mean and median ages of the surveyed people in the dataset.

```
print(mean(census_data$Age))
```

Mean age : 38.58165

```
print(median(census_data$Age))
```

Median age : 37

4. The hist function allows one to create a histogram quite easily. Create a histogram of the ages of the surveyed people in the dataset, and title it appropriately. What do you notice from the distribution of the ages?

```
hist(census_data$Age,main ="Histogram of Age" , xlab = "Age")
```

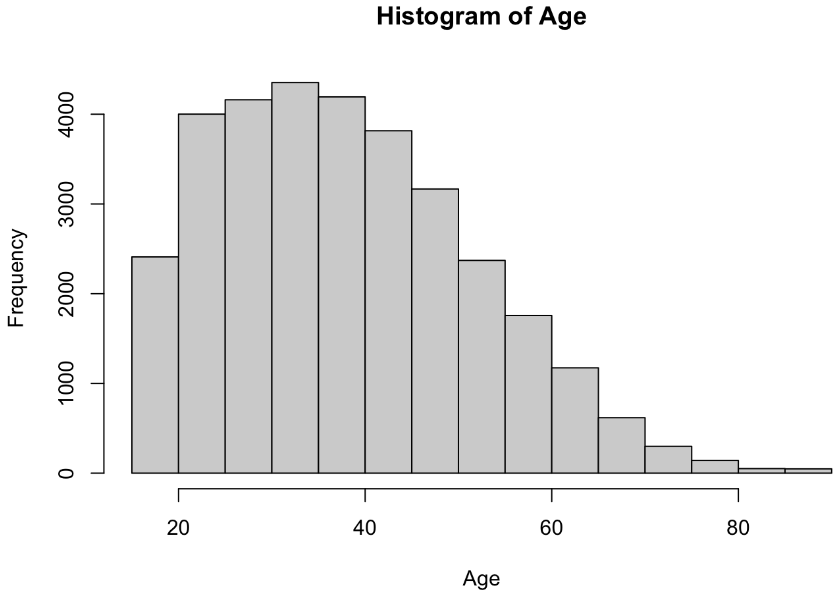
It is right skewed distribution, the mean is larger than the median

5. The Education-Num column in the dataset refers to the number of years of education for a given surveyed person. For this column, using the max, min, mean, and sd functions calculate the maximum, minimum, mean, and standard deviation of this variable.

```
minedu <- min(census_data$Education.Num)
minedu
```

Min education: 1

```
maxedu <- max(census_data$Education.Num)
maxedu
```



Max education: 16

```
sdedu <-sd(census_data$Education.Num)
sdedu
```

Standard Deviation of education :
38.58165
37

5. Finally, the cor function calculates pairwise correlations between all pairs of variables in a dataset. Let us see if any linear relationships exist between one's age, number of years of education and hours of work per week. Do this by first creating a data frame with just those 3 variables using the data.frame command and then calling the cor function on the newly created dataframe. Are there any correlations between the variables? If not, what does this tell you about these variables?

```
newdf <- data.frame(census_data$Age, census_data$Education.Num,
census_data$Hours.Per.Week)
cor(newdf)
```

```
                             census_data.Age     census_data.Education.Num
census_data.Age                1.00000000          0.03652719
census_data.Education.Num      0.03652719          1.00000000
census_data.Hours.Per.Week     0.06875571          0.14812273

                          census_data.Hours.Per.Week

census_data.Age                   0.06875571
census_data.Education.Num         0.14812273
census_data.Hours.Per.Week        1.00000000
```

All three variables have a very weak positive relationship