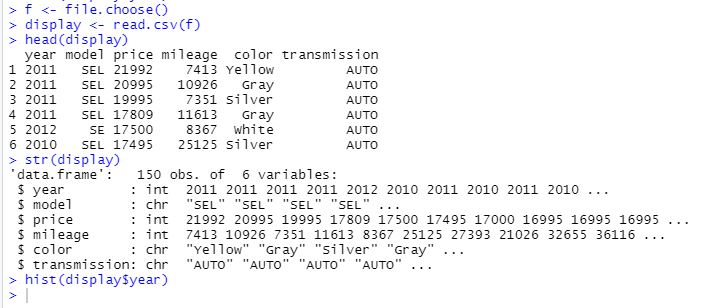
**Assignment 1**

**-Sai Sree Mithra Sripathi(@01414266)**

**Exercise :**

Loaded the ‘usedcars.csv file’ and run the following –



How you used the “str” and “hist” commands in the R Programming exercise along with the outcome of the “str” command and the plot generated by the “hist” command.

**Str :**

Str is a compact way to display the structure of an R object.

This allows you to use str as a diagnostic function and an alternative to summary. Str will output the information on one line for each basic structure. Str is best for displaying contents of lists. The goals is to get an output for any R object.

**Hist :**

Histogram can be created using the hist() function in R programming language.

First the data from usedcars.csv is read into a variable display using read.csv . The following is displayed when str is run:

'data.frame': 150 obs. of 6 variables:

$ year : int 2011 2011 2011 2011 2012 2010 2011 2010 2011 2010 ...

$ model : chr "SEL" "SEL" "SEL" "SEL" ...

$ price : int 21992 20995 19995 17809 17500 17495 17000 16995 16995 16995 ...

$ mileage : int 7413 10926 7351 11613 8367 25125 27393 21026 32655 36116 ...

$ color : chr "Yellow" "Gray" "Silver" "Gray" ...

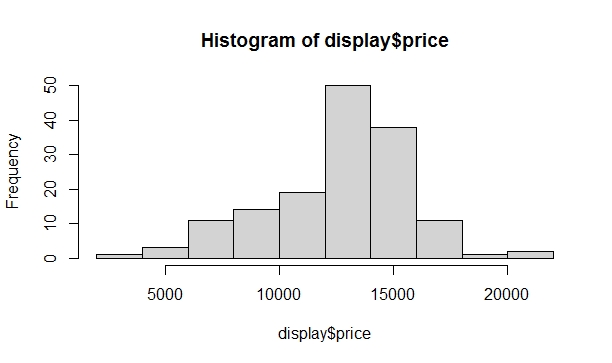
$ transmission: chr "AUTO" "AUTO" "AUTO" "AUTO" ...

From this we can say there are 6 variables or columns in the display variable.

To obtain a histogram we use hist() function, to generate a histogram for the year variable we can do it with – display$year.

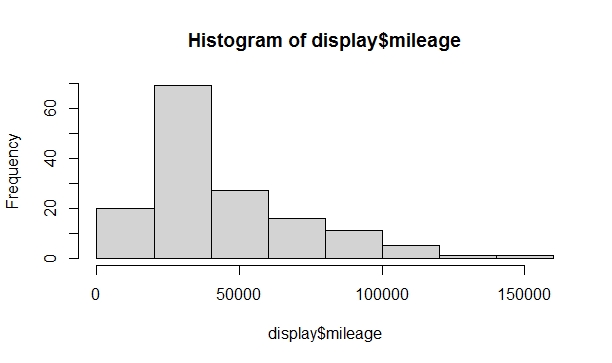
The input to histogram must be numeric. If we try a histogram for the price column :

hist(display$price)



From this histogram we can interpret that most prices range between 10000-15000, going for a peak at 13000-14000 . <5000 and >20000 prices have very low frequencies, close to 0.

Similary if we run - hist(display$mileage)



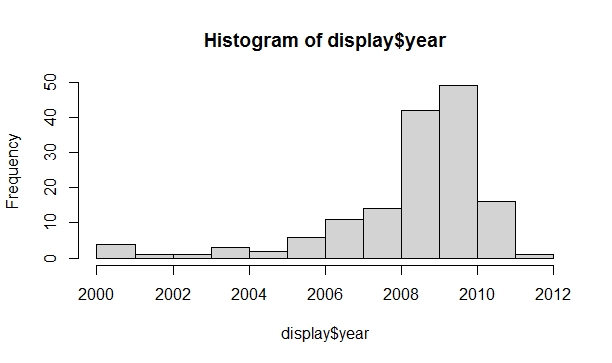
Mileage of 30000-50000 has peak frequency. Too high mileages of >100000 are almost near to 0.

From these interpretations the sales team and management can know where the maximum demand is and improve their business.

**Answer the following questions:**

**What is your interpretation of the histogram data/plot?**

The histogram for year in usedcars.csv :



The x axis contains the provided field display$year and y-axis contains the frequency of occurrences of the x axis values.

The x-axis is divided as two year time frames ranging from the least year in the display$year values to the highest year. Here the x axis is divided for 2 units. We can change the labels and division in x axis to display every year.

hist(display$year, xaxt='n')

axis(side=1, at=seq(2000,2012, 1), labels=seq(2000,2012,1))

The frequency displayed on y axis can also be modified as per the requirement.

Some data sets have a distinct shape. Here are three shapes that stand out:

Symmetric. A histogram is symmetric if you cut it down the middle and the left-hand and right-hand sides resemble mirror images of each other

Skewed right. A skewed right histogram looks like a lopsided mound, with a tail going off to the right

Skewed left. If a histogram is skewed left, it looks like a lopsided mound with a tail going off to the left.

**What interesting insights you can draw from the plot?**

From the histogram for display$year , the peak value in year is in 2009 with a max frequency of 50. In 2008 it was 40. The least frequencies close to 0 were recorded in 2001, 2002, 2004 and 2012.

The display$year histogram is a skewed left histogram and display$price and display$mileage are skewed right histograms.

Based on the maximum frequency years the manufacturers can draw estimations for future sales.

**References:**

<https://www.dummies.com/education/math/statistics/how-to-interpret-the-shape-of-statistical-data-in-a-histogram/>

<https://www.datacamp.com/community/tutorials/make-histogram-basic-r>

https://www.datamentor.io/r-programming/histogram/