**Susheela Polepalli**

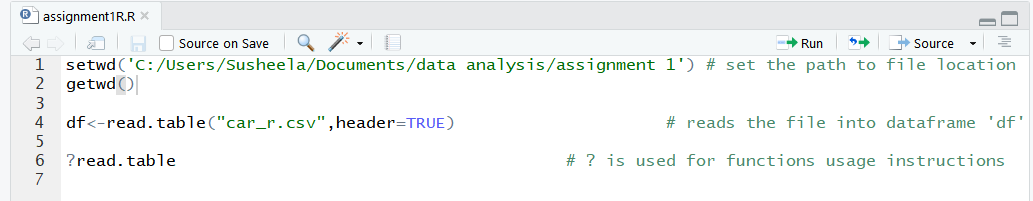
**M10727836**

**BANA7038 Homework 1**

**1. Import the CSV file ‘car\_r.csv’ using the function “read.table()” or “read.csv()”. Where to find the instruction on how to use the functions?**

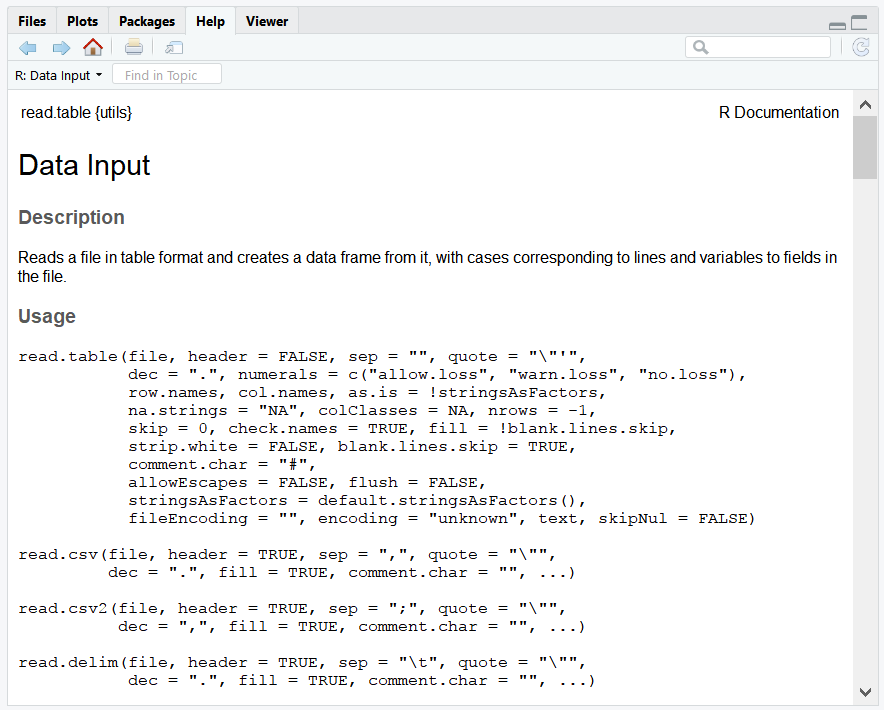
**Solution:**

We can use the ‘read.table()’ function to import csv file “car\_r.csv”



? is used to find the instruction on how to use functions.

Below shows the output of the command ‘?read.table’



**2. How many variables in the data set? What are their names?**

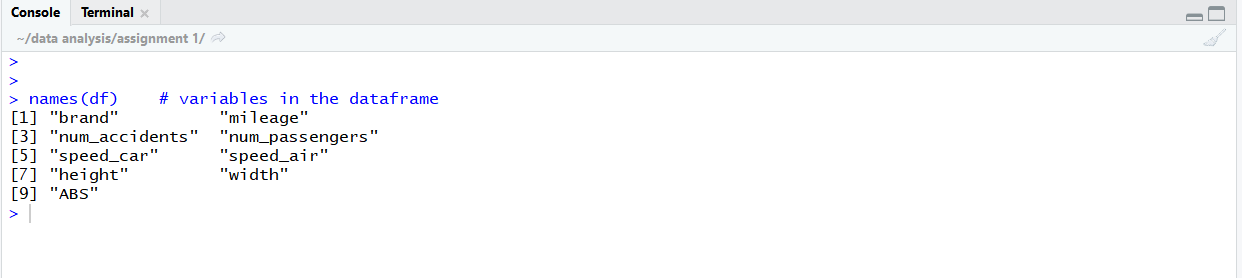
**Solution:**

There are 9 variables in the dataset.

Command used is:

C:\Users\Susheela\AppData\Local\Microsoft\Windows\INetCache\Content.Word\question2.png

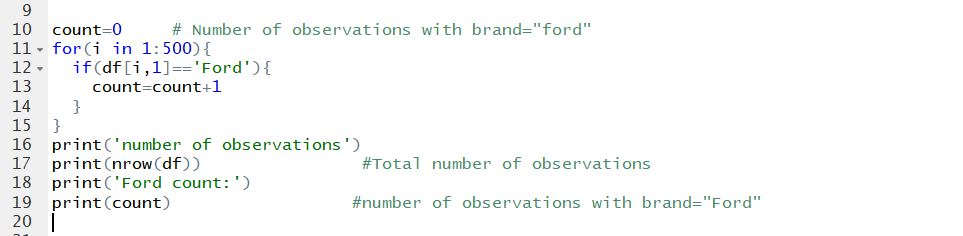
The output console shows the list of variables names in the data set.



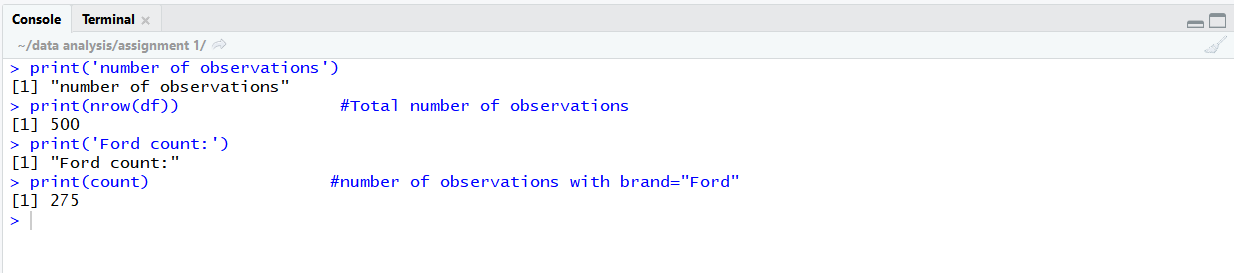
**3. How many observations in total? How many observations for Ford?**

**Solution:**

Code for Number of observations and Ford Observations are:



Output:

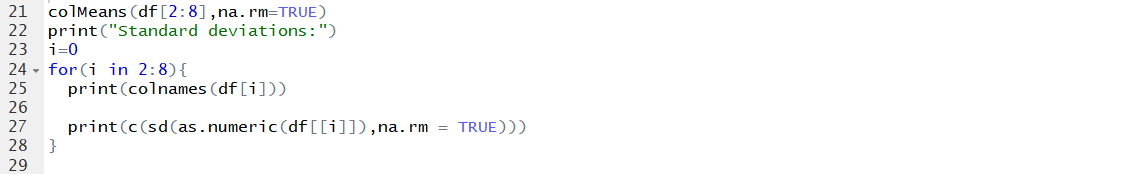


Number of observations are 500 and number of Ford observations are 275.

**4. Calculate the mean for each of the car parameters (measures). Please also report the corresponding standard deviation.**

Solution:

The mean and standard deviation for corresponding car parameters are obtained as follows:



The output is:

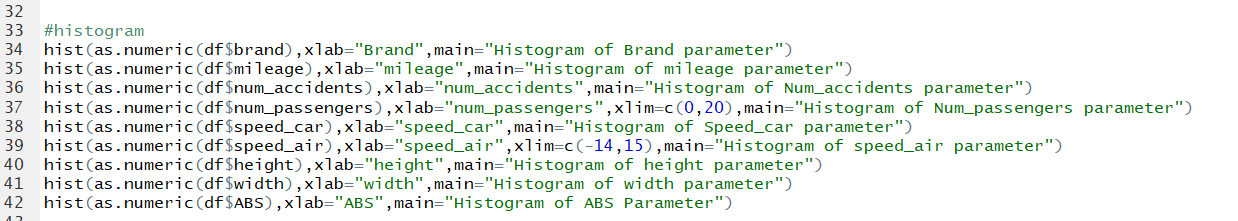


|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Mileage | Num\_accidents | Num\_passengers | Speed\_car | Speed\_air | Height | width |
| Mean | 39564.63 | 2.154 | 6.69 | 50.059 | 0.24 | 5.912 | 6.013 |
| Standard deviation | 10819.68 | 1.423495 | 3.742 | 9.773 | 3.0845 | 1.054882 | 0.47145 |

**5. Obtain the histogram for each of the car parameters.**

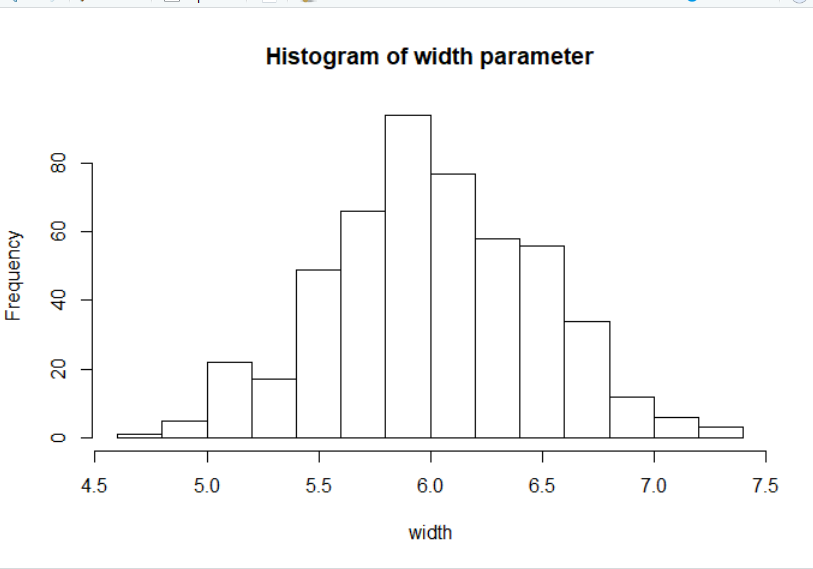
Solution:

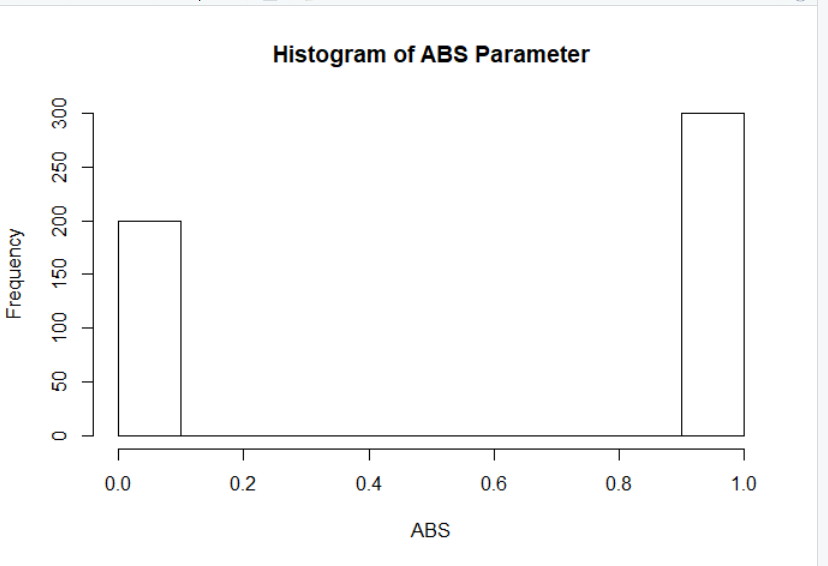
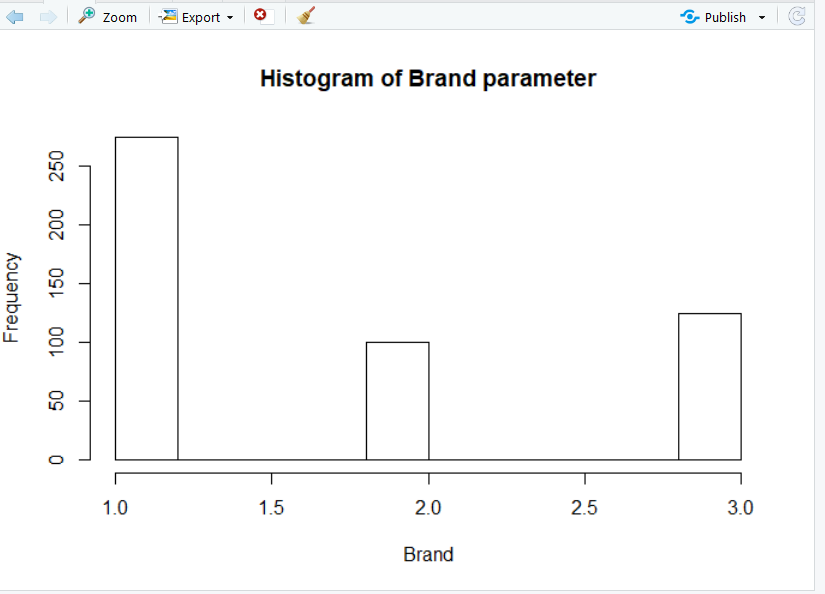
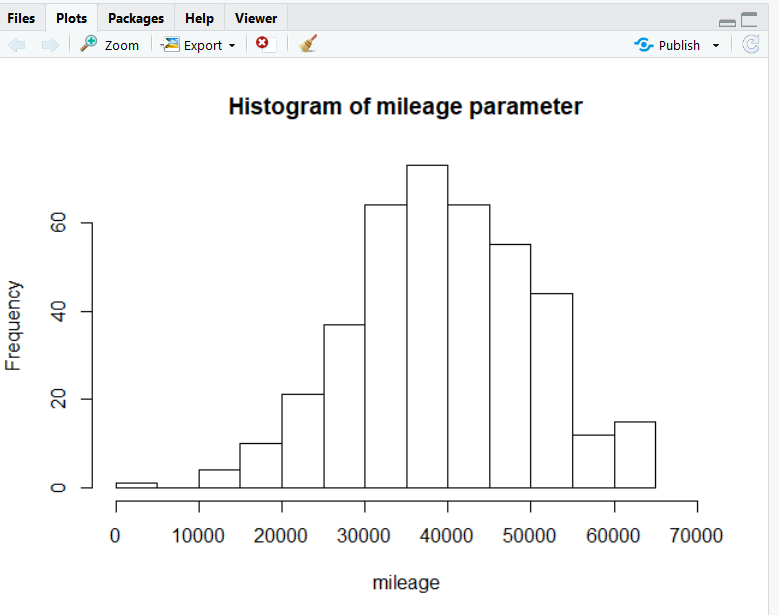
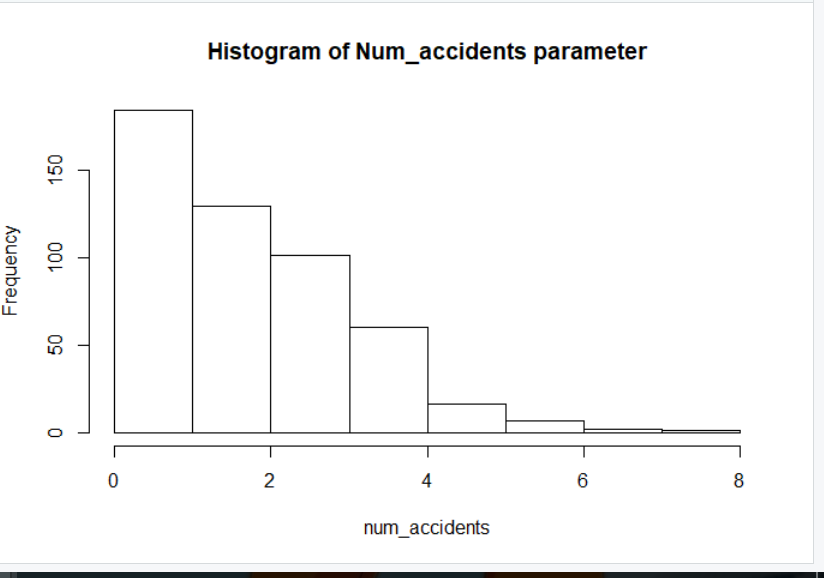
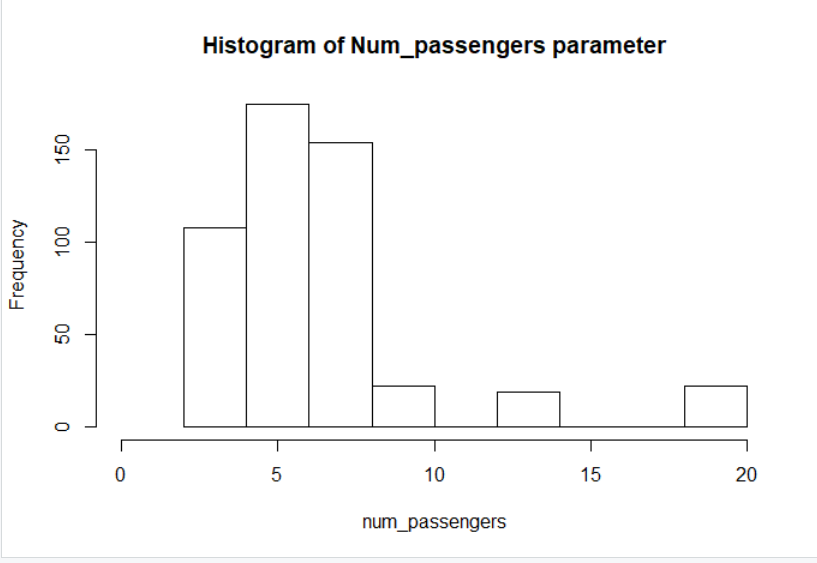
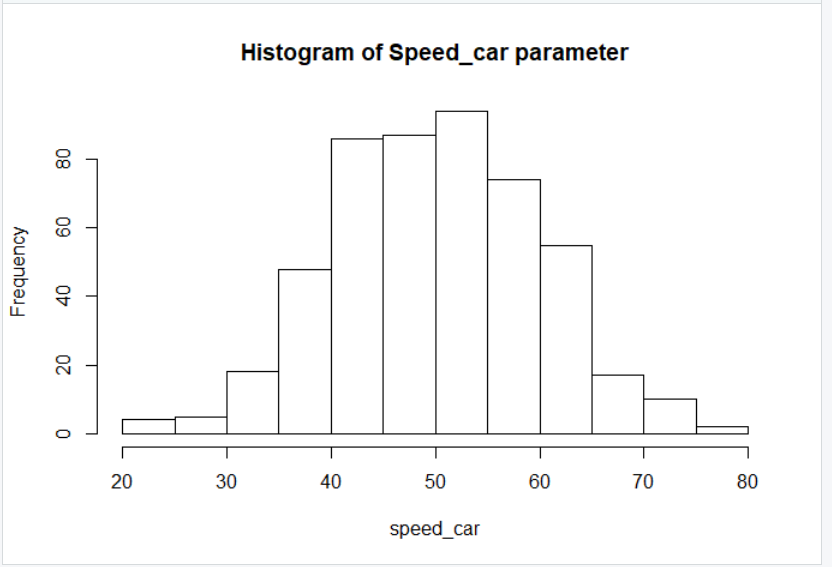
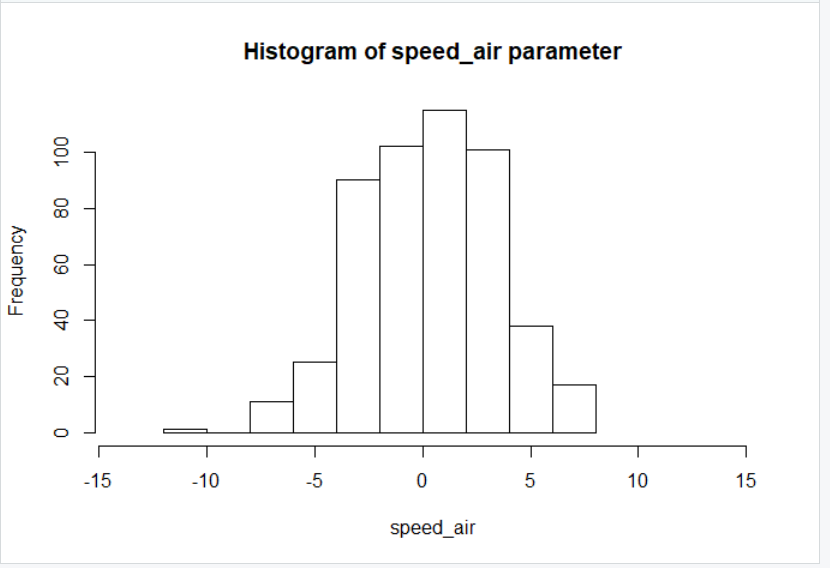
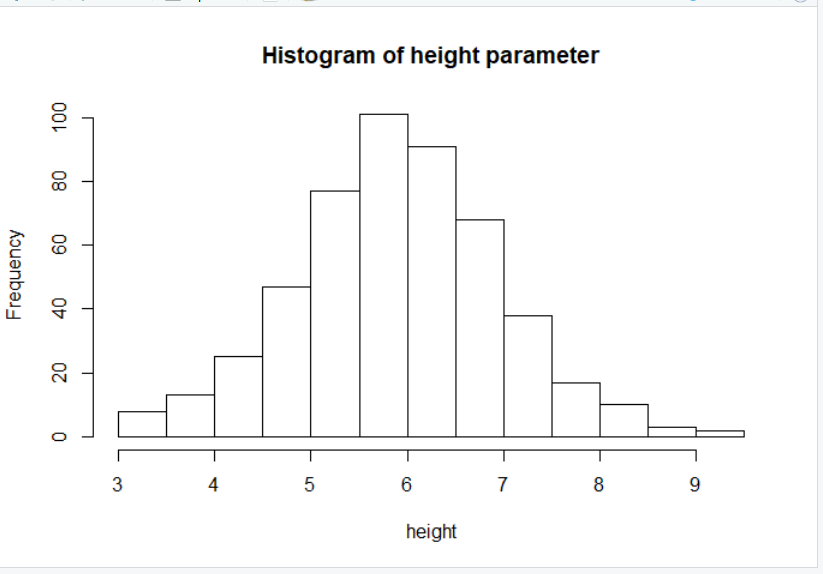
The histogram for each of the car parameters are obtained as follows.



The histograms are

Par(mfrow=c(2,4)



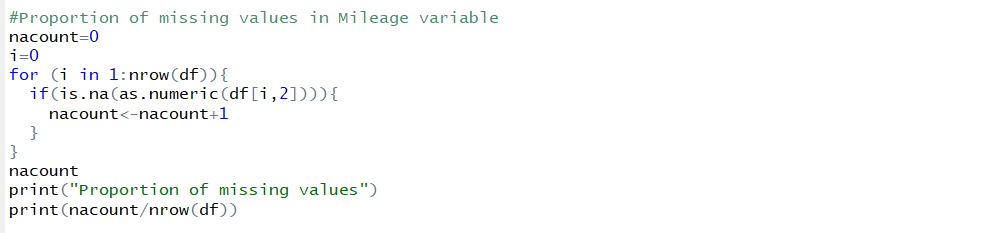


**6. Is there any missing value in the data set? If yes, which variable? What is the proportion of missing values?**

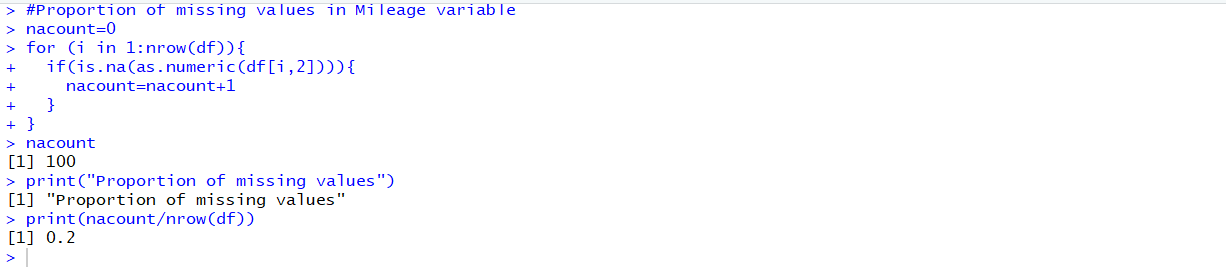
**Solution:**

Yes. There are missing values in the data set for “mileage” variable.

Code:



Output:

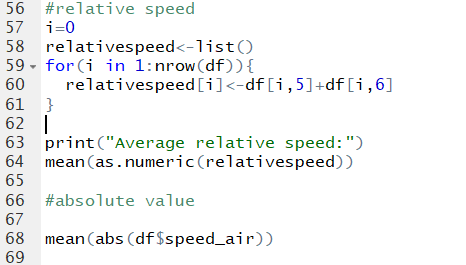


The proportion of missing values in mileage parameter is 0.2.

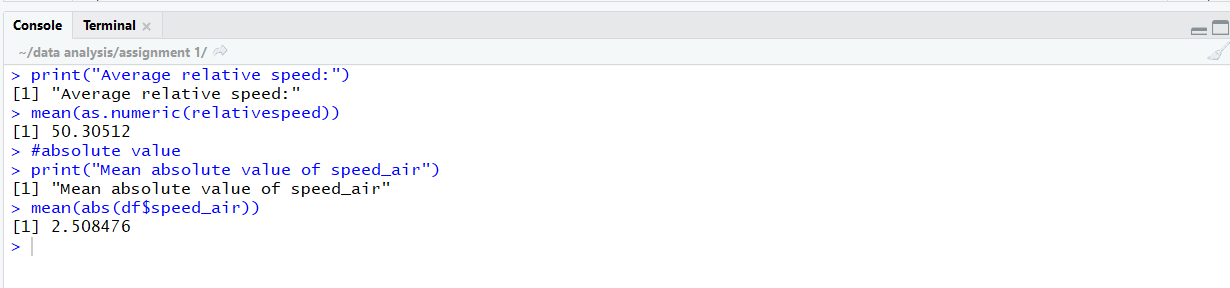
**7. Calculate the relative speed of the car (defined as = speed\_car + speed\_air, where speed\_car is always positive and speed\_air can be positive or negative). What is the average relative speed of the car? Convert speed\_air to absolute value and calculate the average of the absolute value of speed of the air?**

**Solution:**

The code for relative speed of the car and absolute value of speed\_air is ass follows:



Output:



The average relative speed is 50.30512. The mean absolute speed\_air is 2.508476.

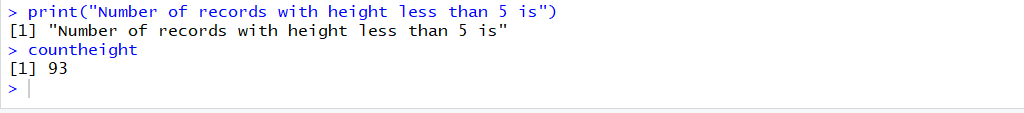
**8. How many cars have mileage less than 40000? How many cars have height less than 5? Please delete those observations (i.e., cars whose mileages are less than 40000 and cars whose heights are less than 5) and delete the observations that contain NAs from the original data set to form a new data set.**

**Solution:**

The following code is used to count the mileage <40000, cars height <5 and to delete the observations from the original dataset to form new data set “updatedDF”.



Output:



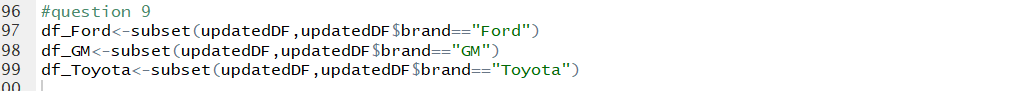


Number of records with mileage less than 40000 is 310.Number of records with height less than 5 is 93.The new dataset is named as UpdatedDF which has 158 observations and 9 variables.

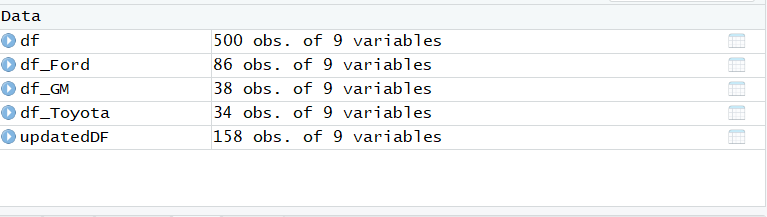
**9. Divide the new data set (as obtained in Step 8) into three subsets: Ford, GM and Toyota.**

**Solution:**

The code for dividing new data set “updatedDF” into three subsets is:



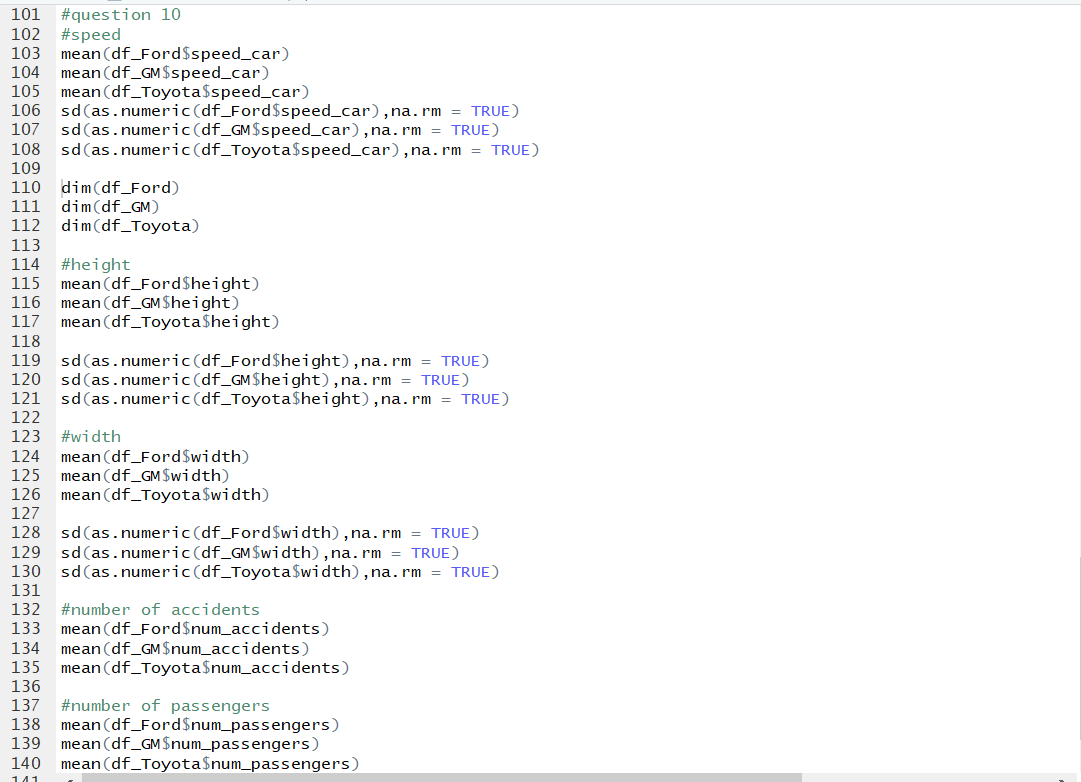
The dimensions of the subsets are as follows:



**10. Using the new data set (as obtained in Step 8), is there any difference between these three brands (in terms of speed, height, width)? You can compare their means, variances.**

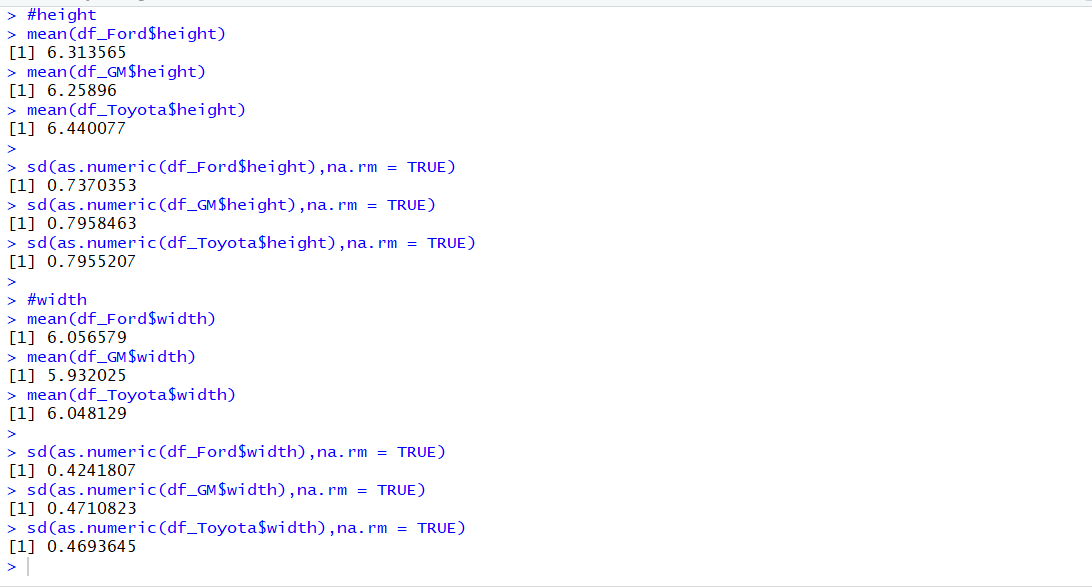
**Solution:**

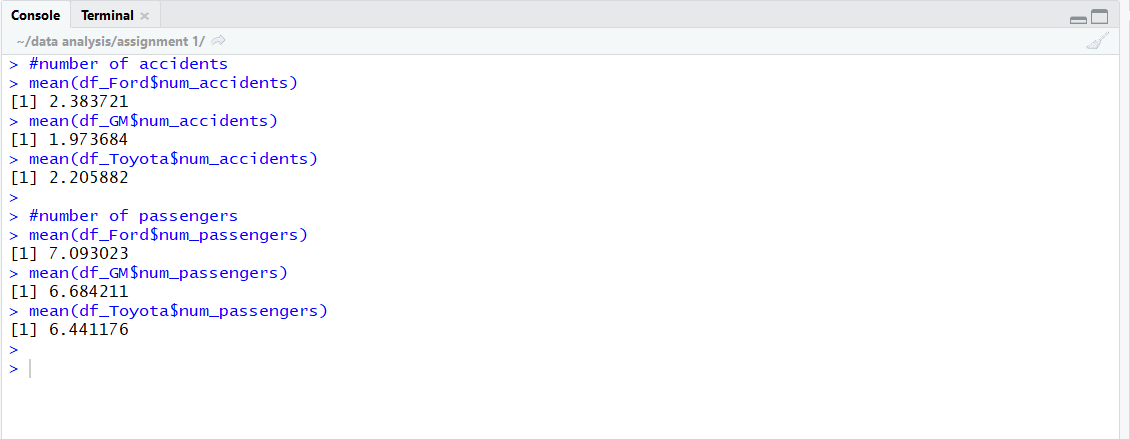
The means and standard deviations of speed, height and width are computed ans the means of other car parameters such as number of passengers, number of accidents and number of cars are also computed as follows.



The outputs are as follows:







**Summary:**

|  |  |  |  |
| --- | --- | --- | --- |
| Parameter | Ford | Toyota | GM |
| Mean speed  Mean Standard deviation | 51.43112  9.666181 | 52.25184  9.880731 | 52.23822  8.778464 |
| Number of cars | 86 | 34 | 38 |
| Height | 6.313565 | 6.440077 | 6.25896 |
| Width | 6.056579 | 6.048129 | 5.932025 |
| Number of accidents | 2.383721 | 2.205882 | 1.973684 |
| Number of passengers | 7.093023 | 6.4411 | 6.6842 |

When comparing the speeds, Toyota is slightly higher than GM followed by Ford.

Number of cars, Ford is having the highest number followed by GM and Toyota.

Height: Toyota > Ford > GM.

Width: Ford > Toyota >GM.

Number of accidents: Ford > Toyota > GM.

Number of passengers: Ford > GM > Toyota.

Key points:

Par(mfrow=c())

Apply(d,1,function(x){sum(is.na(x)})

Look for functions and apply in R.

Which () function

Head()

**ddply**() package

ggplot2() package