**Assignment – 2**

**-Sai Sree Mithra Sripathi**

**Steps for cleaning the data file:**

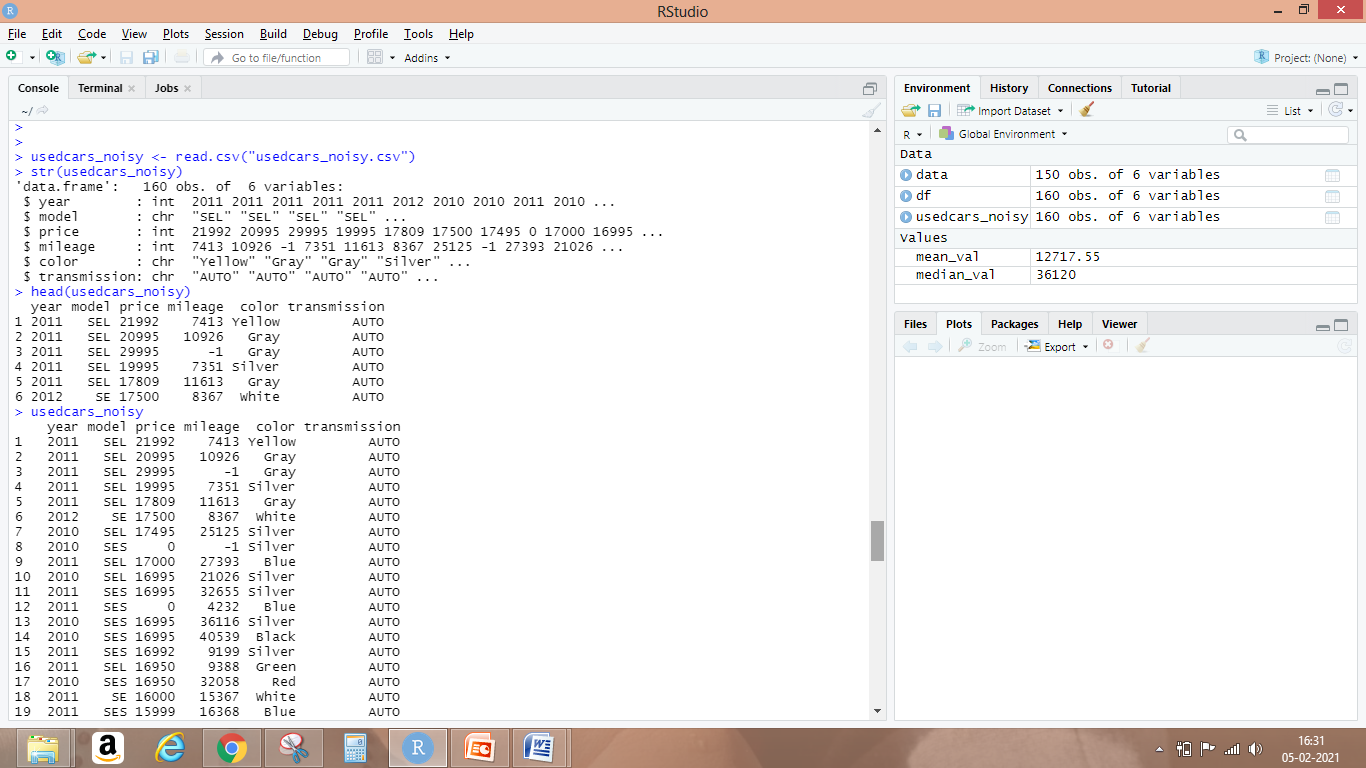
**To complete this Assignment:**

**Download the “usedcars\_noisy.csv” file linked to Week 2 module to your local machine. This file contains the data in “usedcars.csv” file that you used in Week 1, in addition to some noisy data. The noisy data are in the form of instances (i.e., rows/examples) in which either the “price” attribute/feature has the value of “0”, the “mileage” attribute/feature has the value of “-1”, or both.**

**Using RStudio, do the following:**

**Load “usedcars\_noisy.csv” dataset and store it into a local R variable named “usedcars\_noisy” using “read.csv” R command.**

**Display the content of the “usedcars\_noisy” variable by typing the variable name in RStudio. In the output, note the noisy data with “0” price and “-1” mileage.**



usedcars\_noisy <- read.csv("usedcars\_noisy.csv") --- to read the csv into usedcars\_noisy variable

str(usedcars\_noisy) -> lists the columns in the data file

head(usedcars\_noisy) -> lists the first 6 rows in the csv file

usedcars\_noisy -> displays all records in the csv file

This displayed 160 rows of used cars data. Among them there were entries with price value 0 and mileage value -1.

**Clean up the data in “usedcars\_noisy” by:**

**Replacing the “0” price values with the mean price of all cars in the dataset. Replacing an old value with a new value can be done using the “replace” R command.**

Calculated the mean of prices using mean function.

Syntax:

Generic function for the (trimmed) arithmetic mean.

## Default S3 method:

**mean**(x, trim = 0, na.rm = FALSE, ...) where

x is an R object and na.rm is a logical value indicating whether NA values should be stripped before the computation proceeds.

For the given usedcars\_noisy mean for price is calculated as:

mean\_val <- mean (usedcars\_noisy$price, 0, FALSE)

mean\_val

[1] 12717.55

Replacing this mean value obtained at where ever price value is 0. For this we use the replace function.

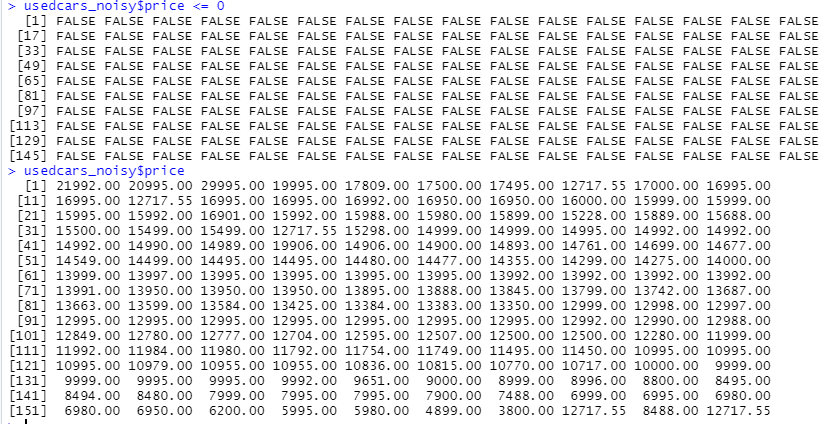
**Replace:** replaces the values in x with indices given in list by those given in values.

Syntax: replace(x, list, values) where

X is vector, list is indexed vector and values are replacement values.

For given usedcars\_noisy we apply replace as:

usedcars\_noisy$price <- replace (usedcars\_noisy$price, usedcars\_noisy$price <= 0, mean\_val)



**Replacing the “-1” mileage values with the median mileage of all cars in the dataset.**

**Calculating median of mileages:**

**Median Value :** Compute the sample median.

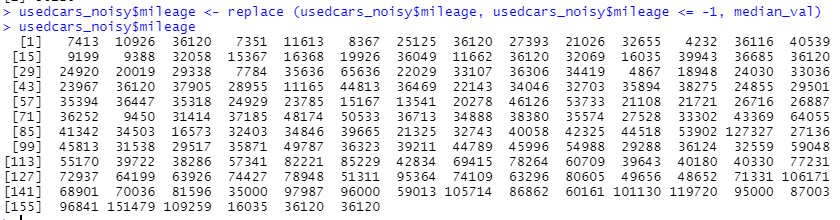
median(x, na.rm = FALSE, ...), where x is an R object and na.rm is a logical value indicating whether NA values should be stripped before the computation proceeds.

Calculate median value using - median\_val <- median (usedcars\_noisy$mileage, FALSE)



Replace the median value with values having -1 using replace function –

usedcars\_noisy$mileage <- replace (usedcars\_noisy$mileage, usedcars\_noisy$mileage <= -1, median\_val)



**Save the cleaned up data in a file named “usedcars\_clean.csv”.**

To save the cleaned values to csv, load it to a data frame first and write it to a csv

Syntax:

write.csv(Your DataFrame,"Path where you'd like to export the DataFrame\\File Name.csv", row.names = FALSE)



The cleaned data is now saved in usedcars\_clean.csv

**References:**

<https://www.rdocumentation.org/>

https://datatofish.com/export-dataframe-to-csv-in-r/