**Problem Statement**

There is a ML code sample in the VM player which the faculty has shared with you earlier, the data for that program is read from a local data frame in R. The task which the students have to complete are;

1)Save the data frame into a CSV file in local file system

2)Copy the CSV file from local file system to HDFS

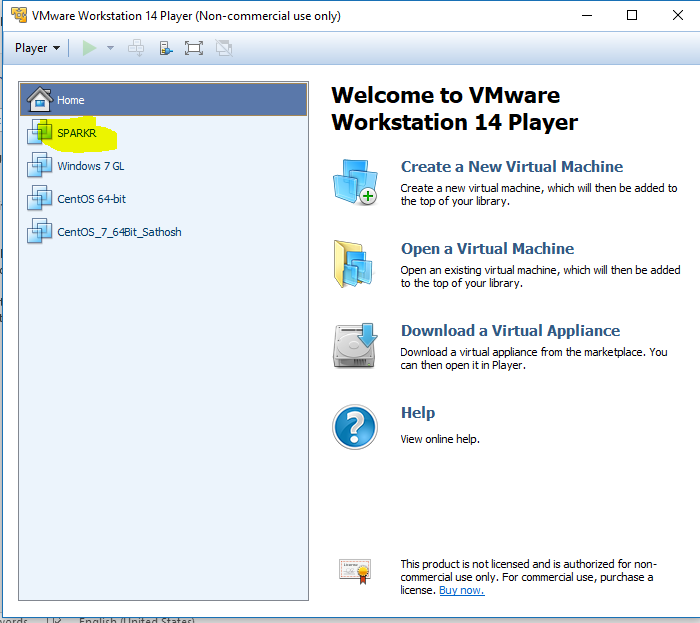
3)Re-run the ML code where the data is now read from HDFS instead of a local data frame in R

You are required to detail down the steps in order to complete above mentioned assignment, print the output and submit the same with their interpretation

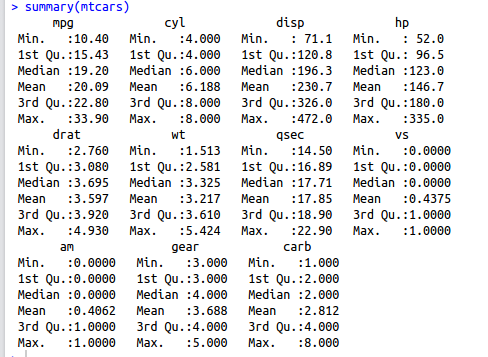
**Solution**

**Step 1**

1. **Open VM that is preloaded with Hadoop and Spark using VMware**



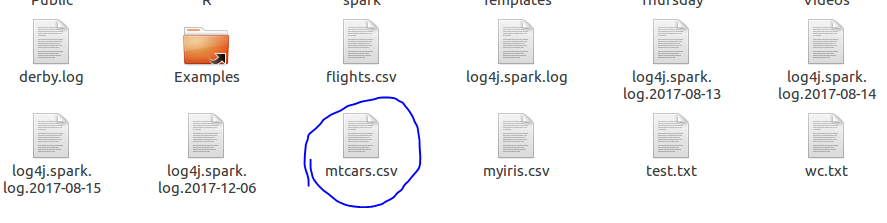
1. **Launch the Spark start scripts after making necessary configuration w.r.t SPARK\_HOME**
2. **Open RStudio inside the VM and Create New Script window.**
3. **Type mtcars and ensure the data is available.**
4. **Type Summary and have high level understanding of data.**



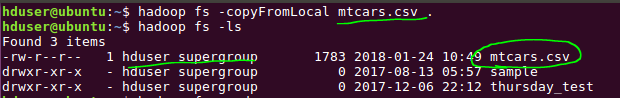
1. **Now save the data which is in format of dataframe to a “CSV” file in following location, “/home/hduser”**
2. **Execute following command**

write.csv(mtcars,file = "mtcars.csv")

1. **You can see the local file is created from the dataframe**



1. **Now we have to move this file to Hadoop**
2. **Run the following command in shell to move this file to HDFS and also check if the file is moved to hdfs.**



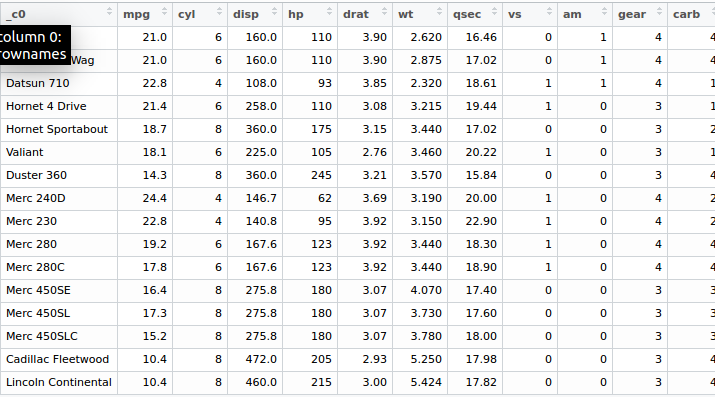
1. **Going back to “R”. Import “sparklyr” library and initiate spark connection (sc).**



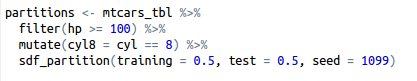
1. **Read mtcars data from HDFS using spark read ad store the results in a dataframe.**



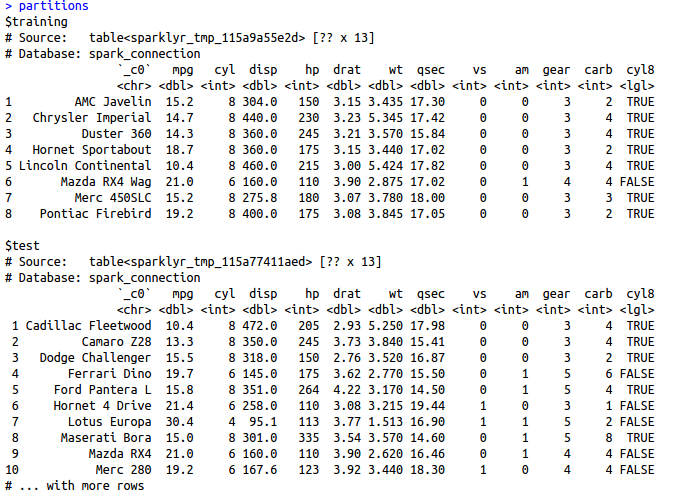
1. **Sample outputs**



1. **Create partition for training and test**



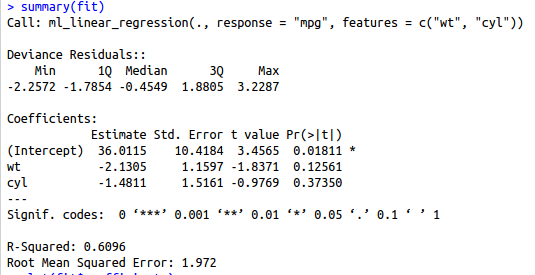
**Output**



1. **Running the Liner Regression**



1. **Results summary**



**RCode**

library(sparklyr)

sc <- spark\_connect(master="local[\*]")

summary(mtcars)

write.csv(mtcars,file = "mtcars.csv")

mtcars\_tbl <- spark\_read\_csv(sc, "mtcarstblread", "hdfs://localhost:9000/user/hduser/mtcars.csv", header = TRUE, columns = NULL,

infer\_schema = TRUE, delimiter = ",",charset = "UTF-8", null\_value = NULL, options = list(),

repartition = 0, memory = TRUE, overwrite = TRUE)

partitions <- mtcars\_tbl %>%

filter(hp >= 100) %>%

mutate(cyl8 = cyl == 8) %>%

sdf\_partition(training = 0.5, test = 0.5, seed = 1099)

partitions

fit <- partitions$training %>%

ml\_linear\_regression(response = "mpg", features = c("wt", "cyl"))

summary(fit)

plot(fit$coefficients)

sc <-spark\_disconnect\_all()