Fall 2016
MIS 413/572 - Introduction to Big Data Analytics

Homework 2

Graded out of 130 points. Due in class on January 9th. Please typeset your homework, save as an R source code file with the title of "your student ID-Homework_2.R" (e.g. B1234567-Homework_2.R), and submit to NSYSU Cyber University. Note that your code must follow the suggested programming and data analysis styles discussed in the class.

1. Please load the dataset “diamonds” in the package *ggplot2* and answer the following questions.

1.1 **[20 pts]** Please write a MapReduce function that calculates average prices for different colors of the diamonds.

1.2 **[20 pts]** Please refer to the page 47 in slide Unit 4. Replace the following SQL with functions in the package *plyrmr.*

sqldf("SELECT cut, avg(carat) FROM diamonds WHERE depth>=65 GROUP BY cut")

1.3 **[15 pts]** Please fit a linear regression model with the *price* as the target/outcome variable and other variables as the predictors. Do any variable transformations or selections if you'd like. Briefly explain the result of your linear model. For instance, what impacts do these predictor(s) have on the target variable?

2. Please load dataset “Hitters” in the package *ISLR*.

2.1 **[20 pts]** Split the dataset into training (90%) and testing data (10%), clean the data and then fit linear models or other regression models you’d like to predict the salary of players.

2.2 [**25 pts]** Apply any feature selection strategies/techniques discussed in the class to your model building process. What is your best model? What variable(s) you believe are most relevant to the outcome (salary of players)?

2.3 **[30 pts]** Write a MapReduce function that compute the 10-fold cross-validation RMSE of your linear model in Question 2.1 or 2.2.