**Project 4: Complete the works given below with R script, knit it in PDF format and Turn In here by uploading R script as well as PDF output file with interpretation.**

Part I:

Do the following in RStudio using “airquality” dataset with R script to knit PDF output:

a) Get mean and standard deviation of Wind variable by Month variable using the appropriate “apply” family of function, show both the results in a single table and interpret them carefully

b) Perform goodness-of-fit test on Wind variable by Month variable to check if it follows normal distribution or not

c) Perform goodness-of-fit test on Wind variable by Month variable to check if the variances of mpg are equal or not on am variable categories

d) Discuss which one-way ANOVA must be used to compare “Wind” variable by “Month” variable categories based on the results obtained above

e) Fit the best one-way ANOVA for this data now and interpret the results carefully

f) Fit the most-appropriate post-hoc test if the ANOVA is statistically significant and interpret the result carefully

Part II:

Do the following in RStudio using “USArrests” dataset with R script to knit PDF output:

1. Create an “crime” dataset containing all the variables of USArrests
2. Create correlation matrix plot of the crime data and interpret each scatterplot carefully
3. Randomly split the crime dataset into training and testing data with 70% and 30% cases
4. Fit a multiple linear regression on training data with Murder as dependent variable and all other variables as independent variables and interpret the results carefully using R-squared, RMSE, Regression ANOVA and Regression Coefficients (BLUE?)
5. Check multicollinearity and finalize this model with the appropriate VIF cut-off value
6. Perform residual analysis of this model i.e. LINE tests using suggestive graphs and confirmatory tests and interpret the results carefully
7. Predict the Murder in the testing dataset using the fitted model
8. Report R-square and RMSE of predicted model and interpret them carefully
9. Report R-square and RMSE of predicted model using other cross-validation methods i.e. LOOCV, k-fold and repeated k-fold
10. Which predicted model is the best model? Why?
11. Fit KNN and ANN-MLP regression on the training data and predict on the testing data.
12. Compare R-square and RMSE of KNN and ANN-MLP models with linear regression
13. Which regression model is best? Why?

Part III

Do the following in R Studio using “mtcars” dataset with R script to knit PDF output:

**a)** Divide the mtcars data into train and test datasets with 80:20 random splits

**b)** Fit a supervised logistic regression model classification models on train data with “am” as dependent variable and all other variables as independent variable

**c)** Check multicollinearity of this model and finalize it using appropriate VIF cut-off value for logistic regression

**d)** Get the confusion matrix, sensitivity, specificity of the fitted model and interpret them carefully

**e)** Predict the transmission variable in the test data and interpret the result carefully

**f)** Get the confusion matrix, sensitivity, specificity of the predicted model and interpret them carefully

**g)** Fit a supervised naïve bayes, support vector machine, decision tree, bagging, random forest, tuned random forest model and boosting classification models on training data with “am” as dependent variable and all other variables as independent variables

h) Which supervised classification model is the best predictive model? Why?