Fall 2022
MIS 572/CM 503 Introduction to Big Data Analytics

Group Exercise 1

- Graded out of **100** points. Please typeset your answers, save as an R source code file with title "Your group ID_Exercise_1.R" (e.g. Group01_Exercise_1.R).
- Please submit your code to NSYSU Cyber University before **10/26 11:59pm**. **No late submission**.
- DO NOT use any loops in your answers. Also notice that your code must follow the suggested programming and data analysis styles discussed in the class.

1. **[70 pts]** Please load Credit data in package "ISLR". Enter "?Credit" in RStudio console to check out the data description.

   1.1 Create density plots for continuous variables, and frequency tables along with bar chart for categorical/factor variables to get a sense of the data.

   1.2 Consider doing a series of bivariate analyses on Balance vs. the rest of variables. Specifically, plot your data and perform bivariate statistical tests to understand the relationships among the variables.

   1.3 Please perform normality tests on Balance. Does it seem "normal"? If not, do you think fitting general linear models to predict or explain the outcome is appropriate?

   1.4. Consider fitting linear models with manually selected variables (i.e., multivariate analysis). What is your best model? You may consider those variables with "p < 0.05"

   1.5. Split your Credit dataset into training and testing sets with the following R code.

```
set.seed(1)
train_idx = sample(1:nrow(Credit), 0.7*nrow(Credit))
train_d = Credit[train_idx,]
test_d = Credit[setdiff(1:nrow(Credit), train_idx),]
```

   Build linear models with training set *train_d*. You may consider reusing your selected variables in previous questions.

   1.6 Write a function that computes Root Mean Square Error (RMSE), which is defined as:

$$RMSE=\sqrt{(\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2)}$$

   where y and $\hat{y}$ are actual and predicted values, respectively. Then apply your linear

models to the training and testing sets, *train_d* and *test_d*. What are the RMSEs of your models? What is your best model in terms of accuracy of prediction (with lowest RMSE)?

1.7 Run summary() to get more information about your linear models, and report the variables with p-value < 0.05. Also run any correlation tests and report the variables with high correlations. Do you think the correlation coefficient is a good measurement for variable importance ranking?

2. **[30 pts]** Please load the given LendingClub loan datasets “LoanStats.csv”. Check out the data dictionary if you would like.

2.1 Please remove columns with any NA, and keep those records with *loan_status* in “Fully Paid” and ”Charged Off”. What is the percentage of the “Charged Off” loan?

2.2 Please replace below R code with SQL code that does similar split-apply-combine operations. Suppose “loan” is the name of your R data frame.

```
# Split, by emp_length
sp_loan = split(loan, loan$emp_length)
# Apply, get average loan amounts
result = sapply(sp_loan, function(x) mean(as.numeric(x$loan_amnt)))
# Combine, into a data frame
result = data.frame("Employment_Length" = names(result),
  "Loan_amount_average" = unname(result)); result
```

2.3 Please replace below SQL code with R code that does similar data management tasks. Suppose “loan” is the name of your R data frame.

```
# For those of top (> 5000) loan purposes,
# count the number of loans for different grades
SELECT grade, count() as Grade_Count
FROM loan WHERE purpose IN
        (SELECT purpose FROM loan GROUP BY purpose HAVING count() >=
5000)  GROUP BY grade
```