Fall 2023

MIS 413_572/CM 503 Introduction to Big Data Analytics

Group Exercise 1

- You can use **either R or Python** to perform your work (choose one).
- Graded out of **100** points. Please typeset your exercise answers, **save answers to two parts (Part 1 and Part 2) as two R/ipynb source code files with title "*<Your_Group_ID>*_Exercise1_*<Qx>*.R/.ipynb"** (e.g., Group1_Exercise1_Q1.R/Group1_Exercise1_Q1.ipynb).
- Please submit your code to NSYSU Cyber University before 10/11 11:59am. **Note that there is a 25% deduction for each day of late submission.**
- DO NOT use any loops in your answers, and your code must follow the suggested programming and data analysis styles discussed in the class. **Also note that 5 points will be deducted from each part (Part 1 and Part 2) if you do not add comments explaining your code in that part.**

1. Please load the given Car datasets("Car_Merge_A.csv", "Car_Merge_B.csv" and "Car_Concat.csv"). Consider the following data analytics questions.

car_ID - Unique id of each observation (Integer)
drivewheel - Type of drive wheel (Categorical)
wheelbase - Wheelbase of car (Numeric)
enginesize - Size of car (Numeric)
boreratio - Boreratio of car (Numeric)
stroke - Stroke or volume inside the engine (Numeric)
compressionratio - Compression ratio of car (Numeric)
horsepower - Horsepower (Numeric)
peakrpm - Car peak rpm (Numeric)
mpg - Mileage (Numeric)
PRICE - Price of car (Numeric)

1.1. **[10 pts]** Perform an inner join on the "Car_Merge_A'' and "Car_Merge_B" datasets using "car_ID" as the primary key. Then, concatenate the merged dataset with the "Car_Concat" dataset along the row axis. Convert any character columns to factor in R and categorical in Pythons. Note that the remaining 'car_ID' is not one of our variables, delete it or convert it into the index.

1.2. **[10 pts]** For continuous variables, create density plots to understand the distribution of the data. For categorical/factor variables, generate frequency tables and bar charts to summarize the counts of each category.

1.3. **[10 pts]** Consider doing a series of bivariate analyses on "PRICE vs. the rest of variables". Specifically, plot your data and perform bivariate statistical tests to understand the relationships among the variables.

1.4. **[10 pts]** Please perform normality tests on PRICE. Does it seem "normal"? If not, do you think fitting general linear models to predict or explain the outcome is appropriate?

1.5. **[10 pts]** 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".

1.6. **[5 pts]** Split the Car dataset into 70% training and 30% testing sets using random seed “20230929”. Using the training set, build multiple linear regression models with the predictor variables selected in previous analyses.

1.7. **[5 pts]** Computes Root Mean Square Error (RMSE), which is defined as:

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

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)?

1.8. **[10 pts]** 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. Please load the given Titanic dataset ("Titanic.csv"). Consider the following data analytics questions.

PassengerID - The ID of each passenger
Survived - Survival (0 = No; 1 = Yes)
Pclass - Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
Name - Name
Sex - Sex
Age - Age
Sibsp - Number of Siblings/Spouses Aboard
Parch - Number of Parents/Children Aboard
Ticket - Ticket Number
Fare - Passenger Fare
Embarked - Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

2.1. **[5 pts]** Please convert the categorical variables into R factors or Python Pandas Categoricals .

2.2. **[10 pts]** Calculate the number of NA values in each column with any “apply”functions in R or Python, and remove those records (rows) with NA values.

2.3. **[5 pts]** Calculate the average fare (‘Fare’) among the different classes (‘Pclass’). Please sort the average fare in ascending order and show the result.

2.4. **[10 pts]** Calculate the correlation matrix between numeric variables. Which two features have the strongest positive correlation? Which two have the strongest negative correlation? Please explain your answer.