Fall 2022
MIS 572 Introduction to Big Data Analytics

Homework 1

- Graded out of **110** points. Please typeset your homework, save as an R source code file with title "your student ID_Homework_1.R" (e.g. B024020001_Homework_1.R).
- Please submit your code to NSYSU Cyber University before **10/30 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. Please load the given dataset “kbopitchingdata.csv” and answer the following data management questions.

   1.1 **[5 pts]** Please remove the observations any NAs.

   1.2 **[5 pts]** Please convert the "year" into multiple groups, "<2005", "2006-2010", "2011-2015", "2016-2020", "2021-2025" and assign to a new variable called “year_interval”

   1.3 **[5 pts]** Calculate the average ERA by year_interval for those with wins larger than 60

   1.4 **[5 pts]** Calculate the total wins for each team and sort by the total wins in ascending order. Please do it in SQL.

   1.5 **[10 pts]** Create a density plot of the distribution of the average_age. Does it seem “normal”? Justify your answer with any statistical methods if you would like.

   1.6 **[10 pts]** Is the average_age associated with the ERA? Use any statistical methods to justify your answers with a brief description.

2. Please load the given dataset “vaccine_2022_TW.csv” for the news about vaccine. Please answer the following questions.

   2.1 **[10 pts]** Please remove observations with any NAs from the dataset and calculate the number of news for each domain. Which domain has the most news?

   2.2 **[5 pts]** Please calculate the number of news per day for the most news domain.

   2.3 **[10 pts]** Please print a line chart with the most news domains in the way of moving average. Please round off the number to integer and label date for every week on the x-axis.

3. Please load the given dataset “HW_car_purchasing_final.csv” and answer the following questions.

   3.1 **[10 pts]** Consider a series of bivariate analyses on "car purchase amount" vs. the rest variables. Specifically, plot your data and perform bivariate statistical tests to understand the relationships among the variables. Are “annual Salary” and “net worth” associated with “car purchase amount”? Use any statistical methods to justify your answers. Also notice that you may consider any data transformation on the “car purchase amount” that helps understand the associations or better predict the “car purchase amount”.

   3.2 **[10 pts]** Please split the dataset into training set (70%) and testing set (30%) with set.seed(1). Then rescale continuous variables into the values ranging from 0 to 1 without centralizing. (Hint: you may use the “caret” package, and you should rescale the training set and testing set with the same feature transformation plan.)

   3.3 **[5 pts]** Write an R function that computes Mean Absolute Error (MAE), which is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

   3.4 **[5 pts]** Build a general linear model with the rescaled training set, then report the training and testing MAEs (round up to the fourth decimal digits).

   3.5 **[5 pts]** Remove the predictors with higher p-values(> 0.05), then build a new general linear model. Does the new model have lower errors in terms of training and testing MAE?

   3.6 **[10 pts]** Again, we would like another new model that considers all the two-way interactions without removing any predictors. Please report the training and testing MAEs. Does the new model have lower errors in terms of training and testing MAE? Can this complex model with more parameters improve the prediction? Which model would you recommend to be used in real-world production and why?