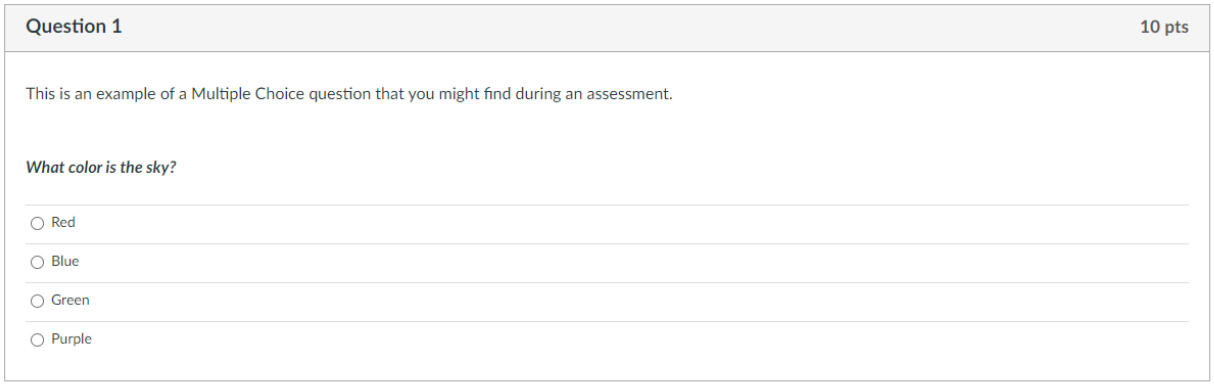
BU510.650 – Data Analytics
Assignment # 6

Please submit two documents: Your answers to each part of every question in .pdf or .doc format, and your R script, in .R format. In your document with answers, please do not respond with R output only. While it is okay to include R output in that document, please make sure you spell out the response to the question asked and also include all the plots. Please submit your assignment through Blackboard and name your files using the convention LastName_FirstName_AssignmentNumber. For example, Yazdi_Mohammad_6.pdf and Yazdi_Mohammad_6.R.

RPNow Practice Quiz

Review the RPNow Instructions and complete the RPNow Practice Quiz located in Module 6. The quiz is designed to verify your ability to use RPNow and consists of the question below for assessment.



Conceptual Part

For this conceptual section, there is no requirement to use any R coding.

Question 1: Suppose we have loan data, which includes the borrowed amount (in 000s of dollars, denoted by X_1), borrower's annual income (in 000s of dollars, denoted by X_2), whether or not the borrower is a student (X_3, which is 1 if the borrower is a student and 0 otherwise), and whether or not the borrower defaulted (Y, which is 1 if the borrower defaulted and 0 otherwise). We want to predict the probability that a borrower will default. After running the logistic regression, we obtain the coefficients:

$$\widehat{\beta_0} = -1, \widehat{\beta_1} = 0.05, \widehat{\beta_2} = -0.01, \widehat{\beta_3} = -0.5$$

(a) Suppose that Shengqi, who is a student, borrowed \$5000 ($X_1 = 5$) and has an annual income of \$25,000 ($X_2 = 25$). Estimate the probability that he will default.

(b) Determine the odds that Shengqi will default.

Programming Part

1. In this question, you will use logistic regression to predict whether a passenger will survive or not.
   To begin your work on this question, first read the data from the file `"TitanicforLogReg.csv"` to a data frame named `Titanic`. (Note: Please review the data before proceeding. You will notice that it has five columns: `Survived, Gender, Child, Fare, Class`, and three of them – `Gender, Fare, Class` – are categorical variables that R will convert to 0-1 columns when you run logistic regression.)
   Next, split the data into training data and test data, using random selection. Include half of the records in the training data and the rest in the test data. Remember to include `set.seed(1)` before the random selection in your code, so we all end up making the same split.

(a) What is the proportion of passengers who survived in the training data, and the proportion of passengers who survived in the test data? (R Hint: Check TitanicExploration.R to see how we did this when we worked on the original Titanic data set.)

(b) Run logistic regression on the training data, with `Survived` as the response variable and `Gender, Child, Fare, Class` as predictor variables. Display a summary of the results. Examine the output: Which predictors are statistically significant? Which predictors are not statistically significant?

(c) Based on part (b), remove the predictors that are not statistically significant, and run logistic regression again on the training data. Display a summary of the results. Examine the output: are all remaining predictors statistically significant?

(d) Using your regression results from part (c), predict the probability of survival for each passenger in the test data. Using these probabilities, assign each passenger in the test data a final prediction of 1 (will survive) or 0 (will not survive). When making this final prediction, adopt the following rule: If the passenger's probability of survival is greater than 0.5, then we predict the passenger will survive. If the passenger's survival probability is less than 0.5, then we predict the passenger will not survive.

(e) Compute the accuracy of the predictions you made for the test data: What is the percentage of passengers for whom your prediction was accurate?