### M3 Group Assignment 1 – Data Analysis

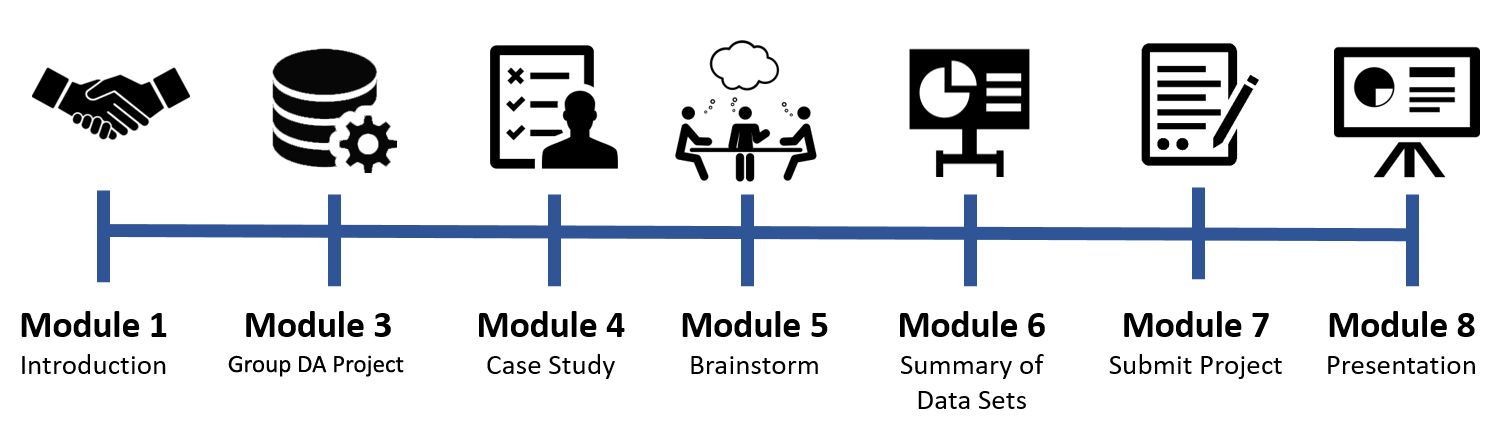
#### BU.510.650

#### Data Analytics

#### Fall 2018

#### Arnab Bisi

## M3 Group Assignment 1 Overview



In this assignment you will complete your first data analysis project. This is a group assignment. It is important for you to schedule a meeting with your group over Zoom or any other group meeting space of your choice. You will work together to complete the following problem set using the “auto.csv” data file, and will have to solve the problems. There are a total of 10 problems, worth 5 points each.

It’s important for everyone to participate equally in the group and the Professor will be monitoring the collaboration to ensure that all members are fully engaged. To ensure that this is, in fact, the case, it is recommended that you meet as a group and discuss the problem set, then divide the problems among the members. Each group member completes their problems and then you reconvene and discuss the results and findings and collaborate on your final submission. If you find this is not ideal for your group, feel free to determine a process amenable to all group members.

You will then discuss your analysis and conclusions in M3 Discussion – Assignment 1 due at the end of the module. You have to post your responses to the questions before you can see peer posts.

Prepare two files:

1. The first file is this document. Type your answers to each problem below in this word file. Name your file ***GroupName*.M3A1.docx**

**Note**: You can also submit an Excel file, or as a scanned handwritten file. If using Excel, each problem should be in a separate tab that is named. If your instructor finds your handwritten submission illegible, he or she may ask you to resubmit it as a typed submission. To submit handwritten work, see “[Instructions for Scanning](https://blackboard.jhu.edu/bbcswebdav/pid-3631299-dt-content-rid-16382471_2/xid-16382471_2)” document posted on blackboard.

1. The second file is your R script. Name the script ***GroupName*.M3A1.R**. Your Professor or TA will use this file if we need more information about your answers or for verification. Both files should be submitted in the Assignment in blackboard.

## Grading

* The following problems count as 5 points each. Total points 10 X 5 = 50. Problems are graded against the standards in the rubric located in the Syllabus for this course.

## Section 1: Questions 1-5

Download the Auto data set, from the course Blackboard page to answer questions 1 - 5.

1. Which of the predictors are quantitative, and which are qualitative? [Hint: Use the codes: Auto <- read.csv(“Auto.csv”, na.strings=”?”), summary(Auto)]
2. What is the range and median of each quantitative predictor? [Hint: Use the codes: require(psych), describe(Auto)]
3. What is the mean and standard deviation of each quantitative predictor?
4. Remove the 25th through 115th observations. What is the range, median, mean, and standard deviation of each predictor in the subset of the data that remains?
5. Suppose that we wish to predict gas mileage (mpg) on the basis of other variables. Using the full data set which variables do you believe will be useful in predicting mpg? Explain your answer using plots and correlation coefficients of the data.

## Section 2: Questions 6-8

GPA’s (Grade Point Averages) for 16 graduating MBA students, and their GMAT scores taken before entering the MBA program are given below. Use this data to respond to questions 6-8.

|  |  |
| --- | --- |
| x=GMAT | y=GPA |
| 560 | 3.20 |
| 540 | 3.44 |
| 520 | 3.70 |
| 580 | 3.10 |
| 520 | 3.00 |
| 620 | 4.00 |
| 660 | 3.38 |
| 630 | 3.83 |
| 550 | 2.67 |
| 550 | 2.75 |
| 600 | 2.33 |
| 537 | 3.75 |
| 610 | 3.85 |
| 570 | 3.30 |
| 590 | 3.50 |
| 650 | 3.65 |

1. Create a linear regression model that uses GMAT scores as a predictor of GPA. Obtain and interpret the coefficient of determination *R2.*
2. Calculate the fitted value for the fourth student on the list (GMAT = 580).
3. Test whether GMAT is an important variable using a significance level of 0.05.

## Section 3: Questions 9-10

Use the rnorm() function to create a vector of 150 observations drawn from a N(0,1) distribution (call this vector x), and another vector of 150 observations drawn from a N(0, 0.2) distribution (call this vector Error). Use these to create a vector *y* according to the model *Y = -1.5 + 0.8 X + Error.* Consider this data to answer questions 9 and 10.

1. Fit a least squares linear model to predict y using x. How do  and  compare to the actual values of  and  ?
2. What are the 95% confidence intervals for  and  and what is the prediction interval for a case with x = 1?