# Multinomial Regression (Module -10)

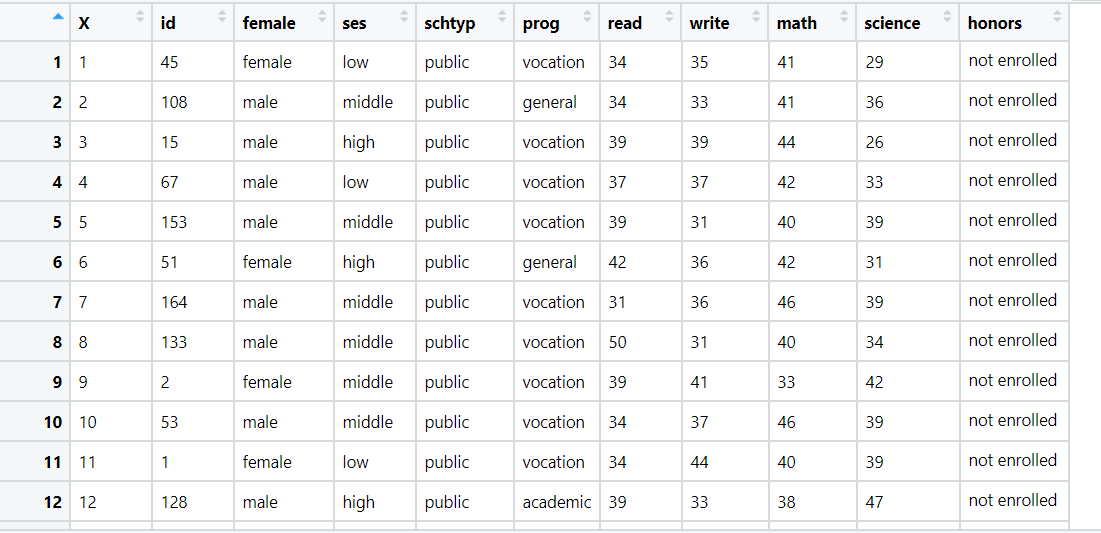
1. You need to predict the type of program a student is in based on other attributes --> mdata.csv

**prog**: is a categorical variable indicating what type of program a student is in: “General” (1), “Academic” (2), or “Vocational” (3)

**Ses**: is a categorical variable indicating someone’s socioeconomic class: “Low” (1), “Middle” (2), and “High” (3)

**read, write, math, science** are their scores on different tests

**honors**: Whether they have enrolled or not



**Ans:**

**Analysis of input and output variables:**

Output Variable (y) = Type of program (prog) which the student takes the most

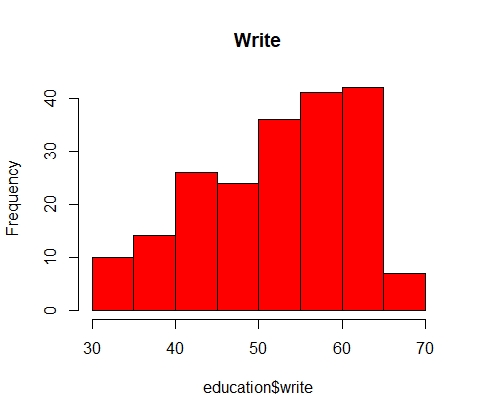
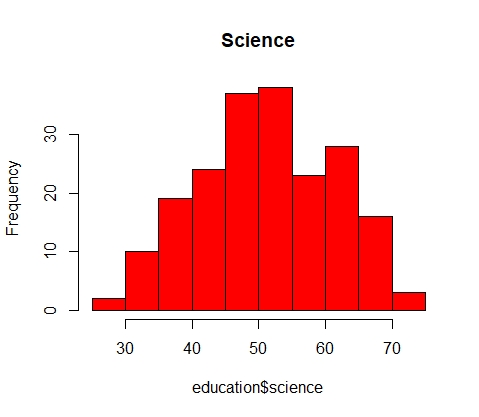
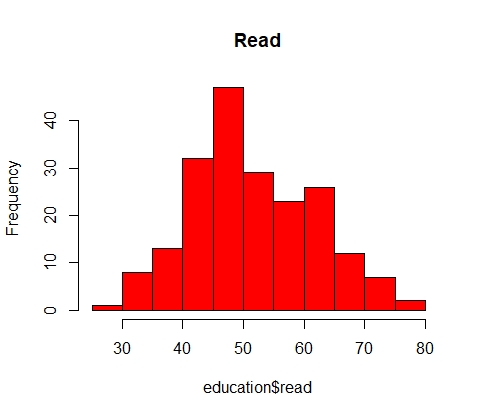
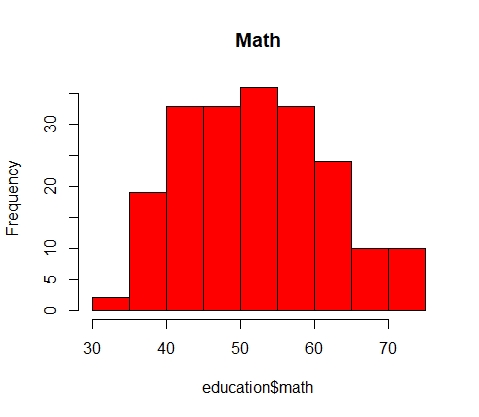
Input Variables = x, id, female, ses, schtyp, read, write, math, science, honors

**Data Preprocessing:**

* As there are no NA values, need to proceed for next step for removing unwanted columns, renaming the columns and reordering the variable columns
* Need to factorize the variables and create dummy variables in R and by using Label Encoder in Python.

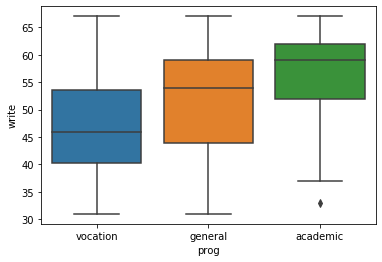
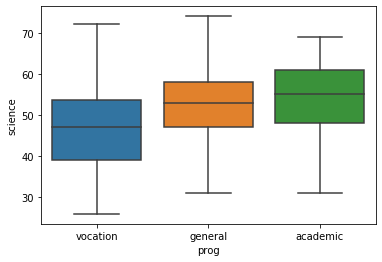
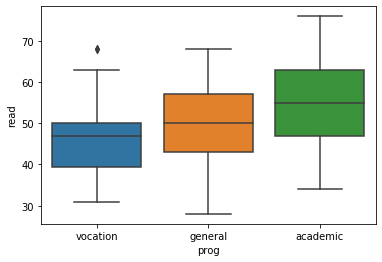
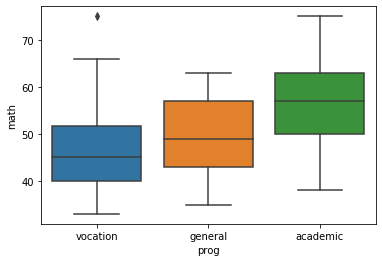
**Graphical Representation:**

**Histogram:**

****

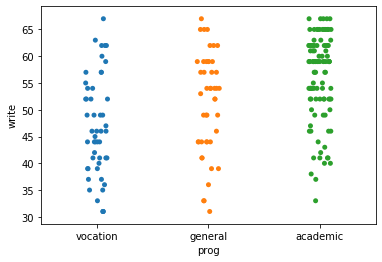
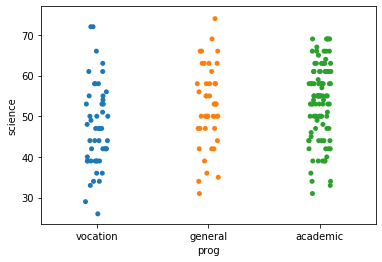
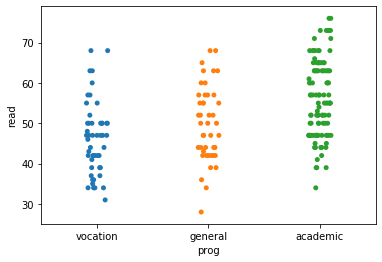
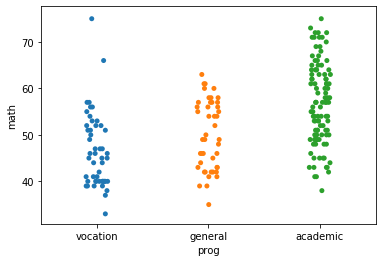
* From the above histogram the individual variables are distributed normally

**Box Plot:**

****

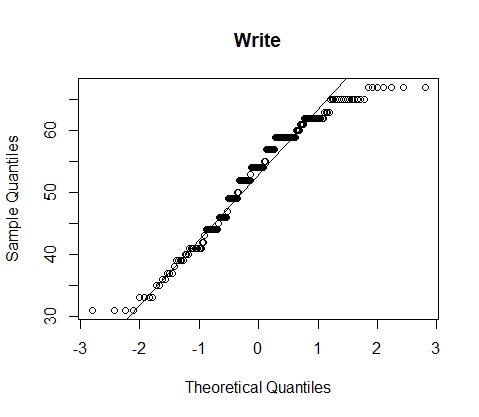
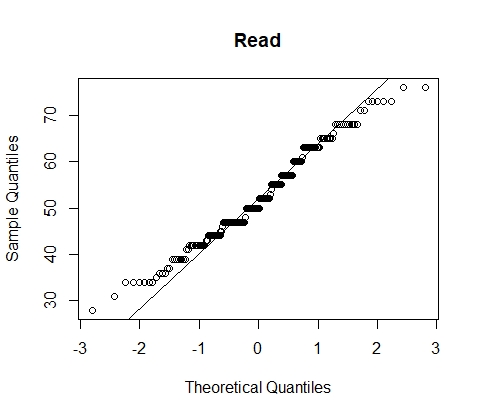
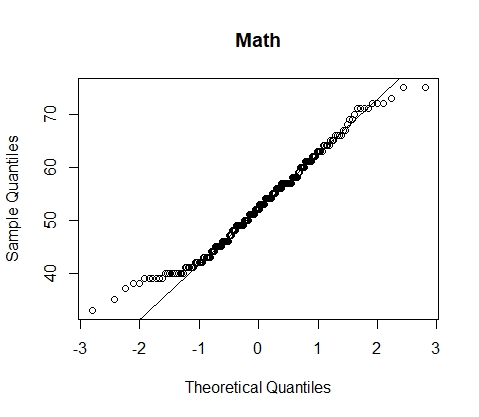
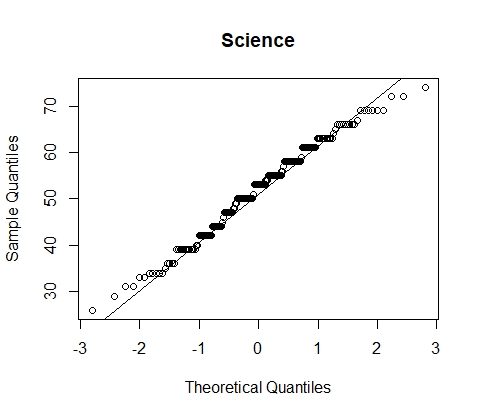
* From the above graphical representation there are outlier present in vacation – read and academic-write

**Strip Plot:**

****

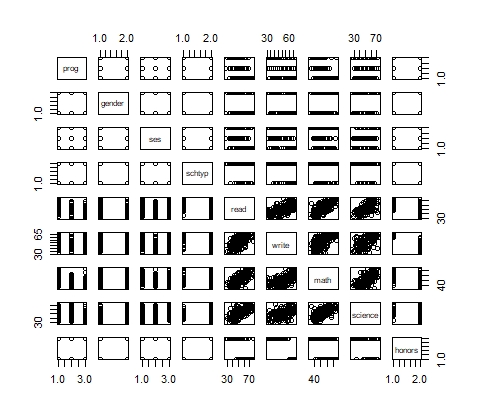
* The above graphs represent the spread of the data at different points.

**QQ - Plot:**

****

* From the above qq-plot it clearly shows that there exists a strong correlation

**Scatter Plot:**

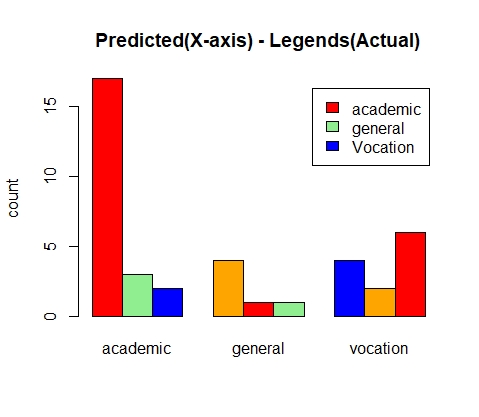
****

* From the above scatter plot we can say that there exist strong correlation between read, write, math and science.
* The summary output has a block of coefficients and another block of standard errors. Each block has one row of values corresponding to one model equation. In the block of coefficients, we see that the first row is being compared to prog = “general” to our baseline prog = “academic” and the second row to prog = “vocation” to our baseline prog = “academic”.

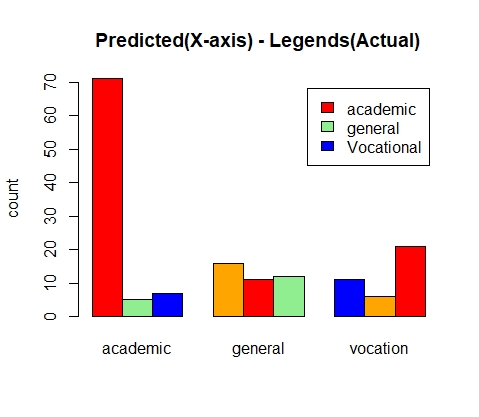
**Building Model:**

* From the give data the data is divided into train and test as 80% and 20%
* From the summary of the model the residual value = 250.0223 and AIC value = 286.0223.
* The accuracy rate for test data = 0.60 and train data = 0.64375, which shows the model is the **Right Fit.**

**Test Prediction:**

****

**Train Prediction:**

****

**Hints:**

1. Business Problem
   1. Objective
   2. Constraints (if any)
2. Data Pre-processing

2.1 Data cleaning, Feature Engineering, EDA etc.

1. Model Building
   1. Partition the dataset
   2. Model(s) - Reasons to choose any algorithm
   3. Model(s) Improvement steps
   4. Model Evaluation
   5. Python and R codes
2. Deployment

4.1 Deploy solutions using R shiny and Python Flask.

1. Result Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.

**Note:**

1. For each assignment the solution should be submitted in the format
2. Research and Perform all possible steps for improving the model(s) accuracy.

Ex: Transformations, Feature Engineering, Hyper Parameter tuning, Outlier treatment, etc.

1. All the codes (executable programs) are running without errors
2. Documentation of the module should be submitted along with R & Python codes, elaborating on every step mentioned here.