**Business Analytics**

**Assignment II**

**Instructions: Please answer all questions from 1 to 3. You should use R to solve the questions and include the screen shots in your submission. The Golden questions are optional and carries additional marks. This means that you will not lose marks if you do not answer that question. Your solution should include the calculation steps and your conclusion that is clearly expressed. Please use the link provided on the Blackboard, under the assessment section, to upload your submissions. Late submissions, up to two days, are subject to 30% penalty. Submissions made more than two days after the deadline will not be graded.**

**\*\*\*\*\***

1. **Run the following code in R-studio to create two variables X and Y.**

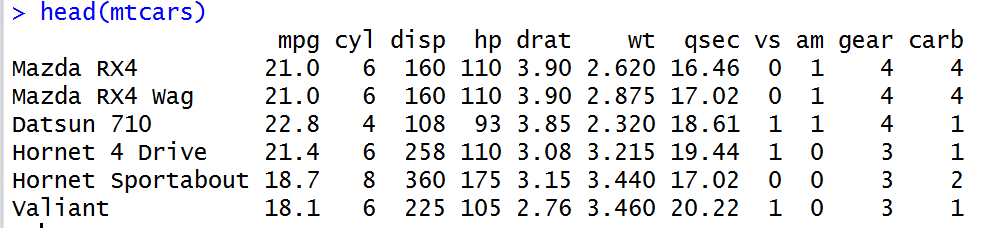
set.seed(2017)

X=runif(100)\*10

Y=X\*4+3.45

Y=rnorm(100)\*0.29\*Y+Y

1. **Plot Y against X. Include a screenshot of the plot in your submission. Using the File menu you can save the graph as a picture on your computer. Based on the plot do you think we can fit a linear model to explain Y based on X? (5 Marks)**
2. **Construct a simple linear model of Y based on X. Write the equation that explains Y based on X. What is the accuracy of this model? (5 Marks)**
3. **How the Coefficient of Determination, R2, of the model above is related to the correlation coefficient of X and Y? (5 marks)**
4. **Investigate the appropriateness of using linear regression for this case (10 Marks). You may also find the story** [**here**](http://blog.minitab.com/blog/statistics-and-quality-data-analysis/violations-of-the-assumptions-for-linear-regression-the-trial-of-lionel-loosefit-day-1) **relevant.   
   More useful hints: #residual analysis, #pattern of residuals, #normality of residuals.**
5. **We will use the ‘mtcars’ dataset for this question. The dataset is already included in your R distribution. The dataset shows some of the characteristics of different cars. The following shows few samples (i.e. the first 6 rows) of the dataset. The description of the dataset can be found** [**here**](https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html)**.**

****

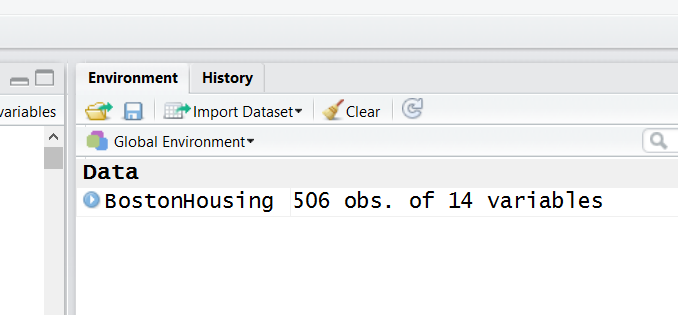
1. **James wants to buy a car. He and his friend, Chris, have different opinions about the Horse Power (hp) of cars. James think the weight of a car (wt) can be used to estimate the Horse Power of the car while Chris thinks the fuel consumption expressed in Mile Per Gallon (mpg), is a better estimator of the (hp). Who do you think is right? Construct simple linear models using mtcars data to answer the question. (10 marks)**
2. **Build a model that uses the number of cylinders (cyl) and the mile per gallon (mpg) values of a car to predict the car Horse Power (hp).**
3. **Using this model, what is the estimated Horse Power of a car with 4 calendar and mpg of 22? (5 mark)**
4. **Construct an 85% confidence interval of your answer in the above question. Hint: use the predict function (5 mark)**
5. **For this question, we are going to use BostonHousing dataset. The dataset is in ‘mlbench’ package, so we first need to install the package, call the library and the load the dataset using the following commands**

install.packages('mlbench')

library(mlbench)

data(BostonHousing)

**You should have a dataframe with the name of BostonHousing in your Global environment now.**



**The dataset contains information about houses in different parts of Boston. Details of the dataset is explained** [**here**](https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/Boston.html)**. Note the dataset is old, hence low house prices!**

1. **Build a model to estimate the median value of owner-occupied homes (medv)based on the following variables: crime crate (crim), proportion of residential land zoned for lots over 25,000 sq.ft (zn), the local pupil-teacher ratio (ptratio) and weather the whether the tract bounds Chas River(chas). Is this an accurate model? (Hint check R2 ) (5 marks)**
2. **Use the estimated coefficient to answer these questions?**
3. **Imagine two houses that are identical in all aspects but one bounds the Chas River and the other does not. Which one is more expensive and by how much? (5 marks)**
4. **Imagine two houses that are identical in all aspects but in the neighborhood of one of them the pupil-teacher ratio is 15 and in the other one is 18. Which one is more expensive and by how much? (Golden Question: 10 extra marks if you answer)**
5. **Which of the variables are statistically important (i.e. related to the house price)? Hint: use the p-values of the coefficients to answer.(5 mark)**
6. **Use the anova analysis and determine the order of importance of these four variables.(5 marks)**