Term Exam

Please submit your commented Jupyter Notebook named as **FirstName_LastName.ipynb.**

1. In the MASS package, you'll find the data frame cats, which provides data on sex, body weight (in kilograms), and heart weight (in grams) for 144 household cats (see Venables and Ripley, 2002, for further details); you can read the documentation with a call to ?cats. Load the MASS package with a call to library("MASS") and access the object directly by entering cats at the console prompt.

   a) Plot heart weight on the vertical axis and body weight on the horizontal axis, using different colors or point characters to distinguish between male and female cats. Annotate your plot with a legend and appropriate axis labels.
   b) Fit a least-squares multiple linear regression model using heart weight as the response variable and the other two variables as predictors and view a model summary.
   c) Write down the equation for the fitted model and interpret the estimated regression coefficients for body weight and sex. Are both statistically significant? What does this say about the relationship between the response and predictors?
   d) Report and interpret the coefficient of determination and the outcome of the omnibus F -test.
   e) Tilman's cat, Sigma, is a 3.4 kg female. Use your model to estimate her mean heart weight and provide a 95% prediction interval.
   f) Use predict to superimpose continuous lines based on the fitted linear model on your plot from (a), one for male cats and one for female. What do you notice? Does this reflect the statistical significance (or lack thereof) of the parameter estimates?

2. A certain spare part is manufactured, X is the number of parts produced in a month, and Y is the needed man-hours. The following table contains data on the recent 10 month.

Month i	X(i)	Y(i)
1	30	73
2	20	50
3	60	128
4	80	170
5	40	87
6	50	108
7	60	135
8	30	69
9	70	148
10	60	132

   (a) Draw a scatter plot with horizontal axis X and vertical axis Y
   (b) Suppose that the manager of the facility suggested the following approximate relation between Yh and X is the equation:

   $$Yh(i) = 9.5 + 2.1\ X(i)$$

   Compute Yh(i) for the given ten X(i)'s
   (c) Run a regression model Y= a + bX + error, where a = the intercept and b = the slope.
   (d) Compute the estimates Yhat(i) = a + bX(i); where a and b are obtained from the model of part (c) for the given ten X(i)'s.
   (e) Compare the manager model and part (c) model. Which model is better? And why?

3. Consider airquality data set with Y=Ozone as the dependent variable and X=Solar.R as the independent variable
    a. Show the scatter plot of Y ~ X
    b. Design the following 4 models:
        i. $Y1 = a + bX$
        ii. $Y2 = a + bX + cX^2$
        iii. $Y3 = e^{a + bX}$
        iv. $Y4 = e^{a + bX + cX*X}$
    c. Evaluate these 4 models by computing R_sq and MSE in each case.
    d. What are the best the worst model? Please explain your answer.

Note: The airquality dataset is built-in R so there is nothing to install or prepare, it is already there as an R object.