Lab 2: Regression Analysis.

During this lab you will gain practical knowledge of the regression analysis techniques that have been presented during Lectures 5-11. To refresh the material, please consult the notes available for the module on Moodle.

You are provided with a supplementary file, **"BSCY4_Lab_2.csv"**. The file contains a subset of **"Auto MPG Data Set"** available from the UCI Machine Learning Repository[1]. As part of this you are required to build a predictor for the **MPG** field that records fuel efficiency of various motor vehicles. Alongside mpg, the following information is recorded:

1. cylinders: categorical
2. displacement: numeric
3. horsepower: numeric
4. weight: numeric
5. acceleration: numeric
6. model year: categorical
7. origin: categorical
8. car name: string (unique for each instance)

In this lab, you will need to use **pandas** module of Python to import contents of both of the files, cleanse the data and merge it into a single data frame. Note that some of the formatting and otherwise errors have been introduced into the data. Use the techniques presented to you in class to correct those errors. Upon completion of this lab you will need to submit python code that help identify problems in data from both files as well as the code that corrects errors.

Please, ***comment your code sufficiently*** to avoid possible misunderstandings during the marking process. Additionally in your comments ***you will need to provide sufficient justification to the steps you take building the regression model.***

[1] https://archive.ics.uci.edu/ml/datasets/auto+mpg

Data Exploration "BSCY4.csv" (4 points)

1. Import data from "BSCY4.csv".
2. Assess normality of MPG values. Do the numbers appear to come from a normal distribution? If not, can a transformation be applied so that its result is normal?[2] **(1 point)**
3. Perform the same assessment (i.e. as Step 2) for other numerical fields included in the data set. **(2 points)**
4. Which of numerical fields satisfy the assumptions of regression analysis? **(1 point)**

Regression Analysis (6 points)

5. Build an initial regression model that incorporates only one numerical predictor. Ensure the model satisfies all of the regression assumptions. **(1 point)**
6. Introduce at least 1 other numerical predictor. Ensure the extended model remains valid (i.e. all of the regression assumptions are satisfied). **(1 point)**
7. What can you say about the extended regression model? Is there mediation effect present? **(1 point)**
8. Introduce a categorical variable into the model. Are all of the categories significant? **(1 point)**
9. Is there a potential mediation effect governed by the categorical variable? If yes, how should the model be updated? **(2 points)**

[2] Consider logarithm as one of the possible transformations (i.e. available in numpy).