Regression Lab

As part of this lab you will use the method of Linear Regression to predict Miles-Per-Gallon performance of a vehicle based on its characteristics. For that you are given a file named **"regression_lab.csv"** that contains a sample dataset of various characteristics of a number of cars. The data-set contains the following attributes:

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

You will need to use that dataset to build "as-accurate-as-possible" predictor. While building the dataset you will need to follow the framework/steps described below. Each of the steps assumes prior knowledge of the main concepts and techniques presented to you during lectures. It is understood that your predictor might not reach "best-possible" accuracy. The aim of this lab is to help student build practical knowledge of the general regression analysis process (as presented during lectures). Adherence to the process and the use of appropriate methods and techniques will constitute the final mark for the lab.

What to Submit: for this assignment the students are required to submit **Python code** that was used to build the model. The code will need to be complemented with a **report** where student will describe the results obtained at each step and their interpretation.

Step No	**Description**
0.	During the lab you are free to use any Python module, though as they are just simply too many, we offer you some advice on what you might need. Those are the modules that have been used in examples presented during lectures. So prior to the lab itself, you should ensure that you have the following installed: • **pandas** - will be used for data import and initial exploration; • **statsmodels** - will be used for Regression Analysis; • **scipy** - will be used for normality and independence assessment.
1.	Using **read_csv** function from **pandas** module, import data from the "**regression_lab.csv**" file. The module returns an object of the type DataFrame. Use its **decribe** and **describeby** methods to gain initial knowledge of the data: • Are there any obvious outliers? • Which numeric variables could come from a normal distribution? • How are the attributes correlated with each other?

- Only consider correlation between two numeric, or two categorical variables. Correlations between numeric and categorical variables lay beyond the scope of this course.
- When establishing correlation between two numeric variables don't forget to check the assumptions if you use Pearson correlation (**Lecture 3**).
- While establishing Normality of a numeric variable follow the steps from **Lecture 2**.
- Use visual assessment to judge on Linear dependency between variables.

2. As part of this lab you are offered to perform a number of steps of what is broadly known as forward selection of stepwise regression building procedure.

   i. Begin with an intercept and just a few other attributes (maybe 1) as the only predictors of the model. (**Lecture 4**)
   ii. Build the current version of the regression model. Ensure it satisfies the regression assumptions (e.g. valid model). Evaluate its predictive capabilities (use Adjusted R-squared). (**Lecture 4**)
   iii. Select another term to be added to the model:
      a. Choose the term from:
         i. One of the original numerical attributes (**Lecture 4**);
         ii. One of the original categorical attributes (**Lecture 5**);
         iii. Cross Product between Numerical and Categorical Variables if the effect of Moderation is suspected (**Lecture 7**);
         iv. Terms of Higher order (**Lecture 5**);
      b. Evaluate potential appearance of the Mediation Effect (**Lecture 6**);
      c. The term offers best improvement of predictive capabilities of the model.
   iv. Updated model performs better than original?
      a. if Yes, update current model, go to step ii;
      b. if No, process concludes.

   The students are not expected to perform the process all the way till completion. Though, at least two iterations of steps ii)-iv) are required to be accomplished.