IE 360 Statistical Forecasting and Time Series
Homework 5, due February 18th, 2021

Instructions: Please solve the following exercises using R (http://www.r-project.org/) or Python (https://www.python.org/). You are expected to use GitHub Classroom and present your work as an html file (i.e. web page) on your progress journals. There are alternative ways to generate an html page for you work:

- A Jupyter Notebook including your codes and comments. This works for R and Python, to enable using R scripts in notebooks, please check:
  - https://docs.anaconda.com/anaconda/navigator/tutorials/r-lang/
  - https://medium.com/@kyleake/how-to-install-r-in-jupyter-with-irkernel-in-3-steps-917519326e41

  Things are little easier if you install Anaconda (https://www.anaconda.com/). Please export your work to an html file. Please provide your *. ipynb file in your repository and a link to this file in your html report will help us a lot.

- A Markdown html document. This can be created using RMarkdown for R and Python-Markdown for Python

Note that html pages are just to describe how you approach to the exercises in the homework. They should include your codes. You are also required to provide your R/Python codes separately in the repository so that anybody can run it with minimal change in the code. This can be presented as the script file itself or your notebook file (the one with *.ipynb file extension).

The last and the most important thing to mention is that academic integrity is expected! Do not share your code (except the one in your progress journals). You are always free to discuss about tasks but your work must be implemented by yourself. As a fundamental principle for any educational institution, academic integrity is highly valued and seriously regarded at Boğaziçi University.

Stepwise Regression in Forecasting Tasks
The file "sales.txt" contains a sample data to forecast the SALES (per month) of a person. Using the following variables, we are trying to forecast if an applicant will be a good salesperson or not.

APT: Selling aptitude test score
AGE: Age (in years)
ANX: Anxiety test score
EXP: Experience (in years)
GPA: High school GPA

Here, all of these variables may not be needed to forecast sales of a person, so you need to implement stepwise regression to reach a sensible final model.
a) Calculate the correlation matrix of all 6 variables and look at all scatter plots between the variables. Which variables do you think are needed to forecast sales values?
b) Implement stepwise regression by following the steps below and obtain a final regression model.
c) Apply appropriate type of stepwise regression using "step" function and compare the final model with the model obtained in (b).
d) Clearly write down your estimates for the intercept, coefficient(s) for the variables and residual variance.
e) Based on your final model, test if high school GPA of a person has an influence on sales value (Use $\alpha = 0.1$). State H0, H1 and the p-value of the test explicitly.

Stepwise regression tasks:
Step 1: Choose the variable having the highest absolute correlation value. Construct an initial simple linear regression model using this variable and the response.

Step 2: Out of the variables that are not in the model, build a new model by adding one variable into your current model. Use the command anova(currentmodel,newmodel) to test the significance of this new variable with an F-test. Do this for all variables which are not in the current model. Choose the variable that corresponds to largest F-statistic (smallest p-value) and update your current model by adding this variable.

Step 3: Once a new variable is added into your current model, build a reduced model by removing one of the variables which was already in your current model (except the last one added in the previous step). Use the command anova(currentmodel,reducedmodel) to test the significance of the removed variable with an F-test. If the p-value of this test is larger than a sensible significance level (if F-statistic is small then critical F value), then update your current equation by removing this variable. Otherwise, do not touch that variable. Do this for all variables in your current model, except the last variable added in the second step.

Step 4: Repeat steps 2 and 3 until all possible additions are nonsignificant and all possible deletions are significant. (For this question, do not focus on the model assumptions.)