In the name of God

Project 3- Data science and HPC

- Write your codes in a notebook file so we can see codes and results together.
- If you want to write your codes in a jl file, take a couple of photos of your terminal and attach them in a zip file

1) A CSV file namely "Salary Data.csv" is taken from a public survey in Tehran which shows the years of experience of each person and the amount of their salaries. Run an OLS regression, report the results and interpret the coefficients.

2) In "50_Startups.csv", a researcher has collected the dataset of 50 startups from three states in the US to check the relationships between the amount of expense each startup spent and the profits they made.

a) Without considering the location, regress profits on each type of expense, report the coefficients and R-Squared and interpret the results. Which type of expenses is a better explanation of profits in the US?
b) We want to know whether adding another type of expense is able to explain the profits better or not. In order to do that, First, regress profit on R&D expense, and next, add another expense, and so on. For each regression, interpret the coefficients, report the R-Squared, and write your conclusion.
c) By grouping firms for each city, regress the profits on the marketing expenses and interpret the results.

3) The dataset "Social_Network_Ads.csv" is collected from a marketing campaign. In which people saw the social media ad and some of them have purchased the product. The CSV file contains three columns namely Age, EstimatedSalary of each person and Purchased with 0 and 1 values where 1 indicates he/she has bought the product and 0 indicates no purchase has happened.

a) For this dataset, run a Logistic regression and interpret the results.
b) Consider 20%, 50%, and 80% of the data, respectively, and predict three forward points. Using the confusion matrix analyze the performance of each step.

4) As a junior analyst in JP Morgan, you are asked to predict the Gold price. To do that, using MarketData package (or other alternatives), collect the historical price of the Gold, Crude oil, and the Federal funds rate for the months between 2015-01-01 & 2022-05-01.

Considering the dates until 2022-02-01, run the following regression models: first regress Gold on the first lag of Gold, next add Crude oil, then add Federal funds rate.

(Note that the price levels are normally non-stationary, and it may depend on many factors. Hence, we use differencing or calculate the return of the price)

1) In each regression model, estimate the coefficients and predict the Gold price for three forward days.
2) For each model, plot the actual and predicted value of the Gold and report the errors. Suggest which model has lower errors for our prediction and which one has higher errors.