**Analysis for the project – Q&A’s as per code:**

Summary:

Methods used: Data Preprocessing, Descriptive Statistics, Correlation Analysis, Regression Modeling, Statistical Analysis, Model Evaluation Metrics, Feature Engineering, Visualization, Stepwise Regression

1. Split the dataset into two sets: 'x',  containing all features except 'Alcohol', and 'y' as a separate dataset with only the 'Alcohol' column. What is the mean value of 'Malic\_Acid'? Round the answer to two decimal points.

* 4.46

2. Using the function train\_test\_split from Scikit-Learn, split the data ('x' and 'y') into training and validation sets. Use an 80/20 split and random state = 0. Now, using the training sets only find the correlation between 'Proline'  and 'y'? Round your answer to two decimals..

* 0.67

3. What is the RMSE value of this model? Enter value rounded to two decimal points.

* 109 to 110 inclusive

4. Perform a univariate simple regression model with 'Proline' as predictor. What is the  adjusted value on the training dataset rounding to three decimals?

* 0.448

5. What is the  adjusted value on the validation dataset rounding to three decimals?

* 0.484

6. In view of these results, can you say that you are overfitting your data?

* No

7. Check the assumption of normality of the errors. Do you think that it holds?

* Yes, a histogram shows very little skewness although the data amount is enough. The errors seem to be normally distributed or show very little deviation form a normal distribution, but it could be an effect of the sample size. We need more data to be sure.

8. Perform a t-test to check if the mean error is 0. What is the p-value of this test? Enter it with three decimal numbers.

* 1.00

9. Check the assumption of homoscedasticity. Explain here why it holds or does not hold.

* Homoscedasticity, particularly in the context of regression analysis means, to describe a situation where the variance of the residuals is constant. The assumption of Homoscedasticity states that the spread of errors should be roughly constant.

In the previous plots, especially in the residuals vs fitted values, it is clear that there is no discernable pattern in the spread of errors, which suggests that the assumption for Homoscedasticity holds true. Also, the plots Histogram and QQ plot, for residuals show the residuals are spread normally, meaning that the variance is constant for residuals.

10. Drop all the non significant variables checked above and run another regression model with the remaining ones. Is the resulting model valid?

* Yes, it passes the F-test and all the t-tests.

11. Give the R^2 adj value of this model in the training dataset up to three decimals.

* 0.883

12. Give the R^2 adj value of this model in the validation dataset up to three decimals.

* 0.890

13. In view of these results, can you say that you are overfitting your data?

* No

14. Now, starting with all the variables in , run a stepwise regression. Feel free to use the function shown in class. What is the  value of this model in the training dataset up to three decimals?

* 0.883

15. Give the R^2 adj value of this model in the validation dataset up to three decimals.

* 0.890

16. In view of the results obtained for the three models above (simple linear regression, multiple linear regression, stepwise regression), pick one as the best and justify your answer.

* While all the models displayed good predictive performance(Generalized well), Simple Linear Regression over-simplified the model or maybe underfit the data a little, Multiple Linear Regression complicated the model using non-significant predictors as well and showed high risk of multicollinearity, finally, the Stepwise Regression model is the best choice among the three. Because it is simple(only 5 variables), showed high performance(Highest adj R^2 of 0.890), utilizes only significant predictors(p<0.05) to avoid overfitting which is ideal for both interpretation and application. This model leverages the strengths of both simple and full multiple regression models by selecting only relevant predictors, ensuring efficient and robust predictions.  
    
  The best model to select is the one obtained with MLR, as it gives the best performance with respect to the other two models. This difference is too large to justify any benefit from the sparsity provided by other models.