**Assignment 3 (10% CA):**

**Due at 11:59pm through Moodle on the 11th of December**

**Provide all the code (as a separate .py file) and report results and plots from python and interpretation of answers in a document.**

Load in the data “mtcars” from the csv file that is available on moodle. Make sure to read the README attached with the data on moodle.

1. **Relationships [50 marks]**
2. Briefly describe the dataset, dimensions and what type of variables there are.
3. Investigate if any of the continuous numerical variables have a linear relationship by producing scatterplots, interpret these plots.
4. What is the response variable and why? What is the research question of interest in this dataset?
5. Investigate if any of the categorical /discrete variables seem to have a relationship with the response variable (selected in part b) using boxplots. Interpret these plots.
6. Using part d), choose one categorical /discrete variable with at least 3 categories to test to see if there is any difference between the means of the response variable (selected in part c) using an one-way ANOVA test. Explain your choice and hypothesis, check the assumptions and interpret the results if appropriate. Conclude your findings.
7. Create a boxplot to see if there is any difference between the means of the response variable (selected in part c) across the two variables “vs” and “am”. Also, create a plot to see if there is an interaction effect for these two variables with the response variable. Interpret these plots. Test to see if any of these factors and/or interactions have a significant relationship with the response variable (selected in part c) using an two-way ANOVA test.
8. **Multiple Linear Regression [50 marks]**
9. Fit a suitable multiple linear regression model, based on your answers from question 1, explain your choice. Comment on whether the assumptions are satisfied, interpretation of the results and the fit of the model.
10. Are there any variables you would like to remove/add from/to the model and why? Re-run the multiple linear regression model. (You can do this more than once, in a background stepwise elimination if you think appropriate, to find the most suitable model). Compare the fit of this model to the model in part a. Perform an F-test to compare the two models, stating your hypothesis and the conclusion of this test.
11. Conclude your overall results. Use your preferred model to predict the fitted values for your response variable and calculate the residual term if you are given the following data:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| mpg | cyl | disp | hp | drat | wt | qsec | vs | am | gear | carb |
| 27 | 4 | 143 | 89 | 3.85 | 2.95 | 27.8 | 1 | 1 | 4 | 2 |