DS 303 Assignment 1

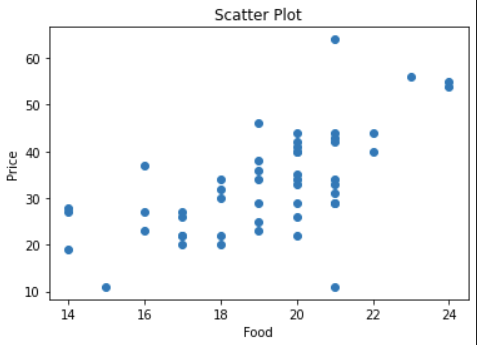
Name: Rohan Rajesh Kalbag Roll no: 20D170033

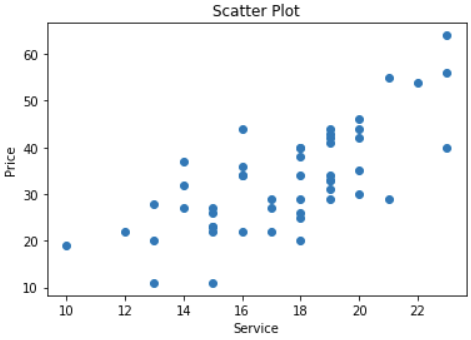
Honour Code:

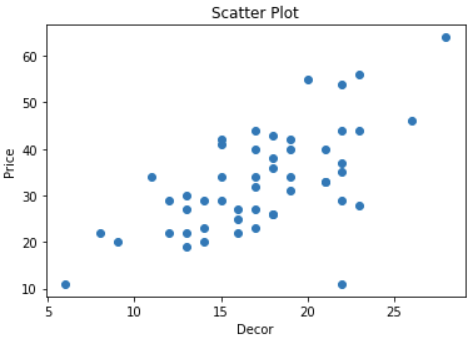
I hereby declare that I have not indulged in any academic malpractices such as plagiarism and would like to assure you that, any code written in this is written by me alone. I would like to add that any resources/threads found on online sites will be mentioned in the document that was used for reference.

The dependence of the price of a meal on the food, service and decor were studied using scatter plots that were made using matplotlib.pyplot library of python.

a) The obtained plots were







We observe that there is no exact relationship between the Food, Service, Decor and the Price but we can infer that there is a nearly linear relationship between each of the input variables and the output variable. We notice that the price increases as the decor, service, food variables increase.

We can expect that after performing the multivariable linear regression we can expect the coefficients to be positive indicating nearly positive linear relationship with each of the input variables.

The scikit-learn LinearRegression model was used to perform the multivariate linear regression and after fitting 50 randomly sampled values from restaurant.csv we obtain the linear model as

where the coefficients that were obtained were

b) To find the 95% confidence interval we find the standard deviation with degrees of freedom as n-2 for the training dataset as

Thus using the properties of gaussian distribution, we get the 95% confidence interval for the predicted as

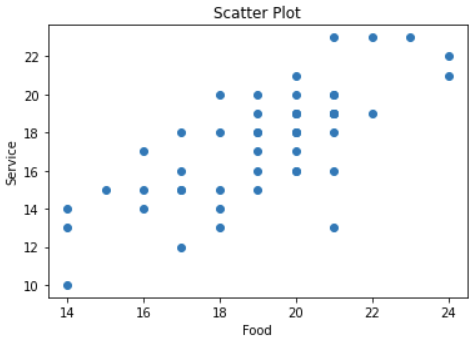
Thus for input as we get the confidence interval as

Reference used: <https://datascience.stackexchange.com/a/41989>

c) We get the coefficient estimate for food from the model as which is the change in the mean response for price per unit increase in food when all other predictors are held constant

Reference used: [What is a regression coefficient? - Minitab Express](https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/supporting-topics/regression-models/what-is-a-regression-coefficient/#:~:text=Coefficients%20are%20the%20numbers%20by,in%20an%20equation%20are%20multiplied.&text=Each%20coefficient%20estimates%20the%20change,other%20predictors%20are%20held%20constant).

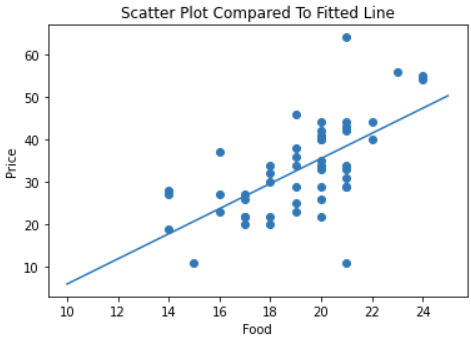
d) The scatter plot for food vs service is as shown below



On regressing the price with the food variable training another scikit-learn model “model2” using the same training data sampled before from restaurant.csv, we obtain the linear model as

where the coefficients that were obtained were

, the modelled line was plotted along with the scatter plot is shown below



We observe that the coefficient for the food is 2.947 in this case whereas it was 1.949 in the first case. The discrepancy is because of the correlation between independent variables in the multivariate regression. From the plot for food and service we see that food and service are positively correlated, i.e if the food variable is high service variable is high. Thus coefficient of food decreases in the multivariate regression because of its high correlation with service.

Reference used: [Linear vs. Multiple Regression: What's the Difference?](https://www.investopedia.com/ask/answers/060315/what-difference-between-linear-regression-and-multiple-regression.asp)

e) Since in (d) we saw that food and service have high positive correlation, I would expect that the model wouldn’t give a very accurate prediction since the food is high and the service is low contrasting to their positive correlating behavior.

The price prediction obtained when a multiple regression was performed for the said inputs

was