PART-1 REGRESSION ANALYSIS MODEL EQUATION AND DATA VALIDATION

Take a dataset for regression analysis with a minimum of 5 predictors . Describe the dataset (predictors, response, number of data points). Perform exploratory data analysis (with scatterplot matrix and heatmaps) to find out how the variables are related. Please divide the data for training and testing. Use stepwise regression to fit a model and find out the most significant variables in the model.

How well did the model fit the test data? Describe the model equation, root mean squared error (RMSE) and comment if the assumptions are met. If the assumptions are not met, how would you improve the model? Please describe in words about the relationship between the significant predictors and the response.

Please share the notebook (the code and the screenshots of the visualizations) to show this work with the assignment.

The dataset taken from https://Kaggle.com. This dataset is created for prediction of Graduate Admissions from an Indian perspective. This dataset is created for prediction of Graduate Admissions from an Indian perspective. The dataset contains several parameters which are considered important during the application for Masters Programs.

The parameters included are :

- GRE Scores (out of 340)
- TOEFL Scores (out of 120)
- University Rating (out of 5)
- Statement of Purpose and Letter of Recommendation Strength (out of 5)
- Undergraduate GPA (out of 10)
- Chance of Admit (ranging from 0 to 1)

Predictors:

1. GRE Score
2. TOEFL Score
3. CGPA
4. University Ranking
5. SOP
6. LOR

Response Variable: Chance of Admission

No. of data points = 400

Model Equation:

Chance of admission = -1.4139+(0.0023*GRE Score)+(0.0028*TOEFL Score)+(0.0061*University Rating)- (0.0020*SOP)+(0.0227*LOR)+(0.119*CGPA)

All the assumptions are successfully met. The most significant variable is CGPA.

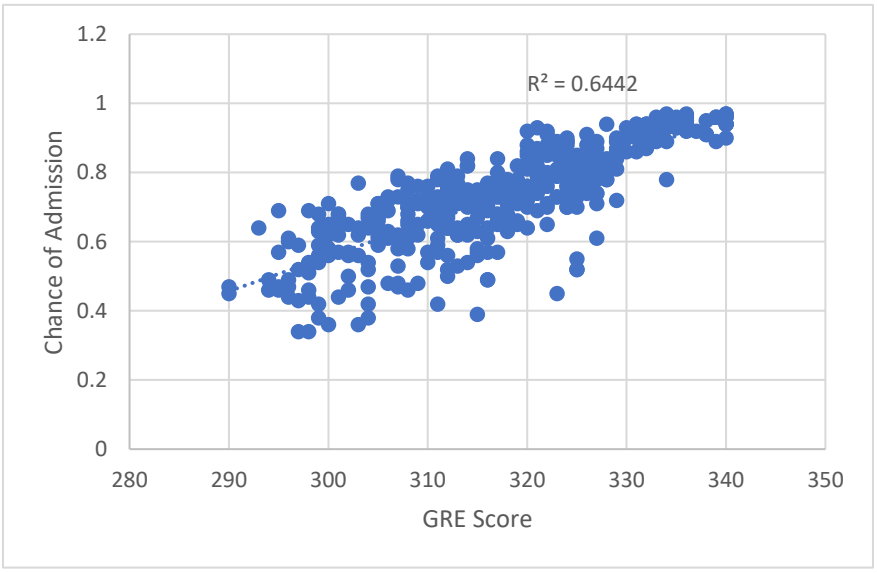
VALIDATION IN EXCEL (GRE vs Chance of Admission)

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.803
R Square	0.644
Adjusted R Square	0.643
Standard Error	0.085
Observations	400.000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1.000	5.227	5.227	720.554	0.000
Residual	398.000	2.887	0.007		
Total	399.000	8.115			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-2.436	0.118	-20.677	0.00	-2.668	-2.204	-2.668	-2.204
X Variable 1	0.010	0.000	26.843	0.00	0.009	0.011	0.009	0.011



The R squared value for GRE Score obtained through Python code and Excel are same. Hence, the model is validated.

SUMMARY OUTPUT (All 6 variables)_								
Regression Statistics								
Multiple R	0.8937							
R Square	0.7987							
Adjusted R Square	0.7956							
Standard Error	0.0645							
Observations	400.0000							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	6.0000	6.4813	1.0802	259.9042	0.0000			
Residual	393.0000	1.6334	0.0042					
Total	399.0000	8.1146						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.4139	0.1154	-12.2470	0.0000	-1.6408	-1.1869	-1.6408	-1.1869
X Variable 1	0.0023	0.0006	3.9385	0.0001	0.0011	0.0034	0.0011	0.0034
X Variable 2	0.0028	0.0011	2.5034	0.0127	0.0006	0.0049	0.0006	0.0049
X Variable 3	0.0061	0.0048	1.2576	0.2093	-0.0034	0.0155	-0.0034	0.0155
X Variable 4	-0.0020	0.0056	-0.3500	0.7265	-0.0130	0.0091	-0.0130	0.0091
X Variable 5	0.0227	0.0056	4.0626	0.0001	0.0117	0.0338	0.0117	0.0338
X Variable 6	0.1199	0.0123	9.7089	0.0000	0.0956	0.1441	0.0956	0.1441