Term Project Summary

The data provides insights on the factors that affect the first year grade point average (GPA) of college students. The EDA reveals the summary statistics and correlation coefficients of the six columns: sex, sat_v, sat_m, sat_sum, hs_gpa, and fy_gpa. The mean value of sex is 1.48, indicating that there are more females in the dataset, while the mean values of sat_v, sat_m, and sat_sum are 48.93, 54.43, and 103.36, respectively. These values suggest that the students' overall SAT scores are moderate. The mean value of hs_gpa is 3.20, and the mean value of fy_gpa is 2.47, which is lower than hs_gpa, indicating a drop in performance in the first year.

The correlation matrix shows that the highest correlation with fy_gpa is hs_gpa (0.535207), followed by sat_sum (0.454933), sat_v (0.393295), and sat_m (0.384160). These values suggest that the high school GPA is the most significant predictor of the first year GPA, followed by SAT scores. The OLS regression analysis confirms that all four predictors: sex, sat_v, sat_m, and sat_sum, are statistically significant in predicting fy_gpa, along with hs_gpa.

However, there are some limitations to this analysis. One significant limitation is the absence of other variables such as socioeconomic status, major, and study habits that could impact the first year GPA. These variables could potentially confound the relationships found in this analysis. Additionally, there are no insights into the students' personal backgrounds or circumstances that could explain the lower first year GPA.

Moreover, there is a possibility of incorrect assumptions. For instance, the assumption of linearity may not hold, and there could be a nonlinear relationship between the predictors and the first year GPA. There is also a chance that the data is not normally distributed, and the summary statistics could be affected by outliers or extreme values.

The challenges faced in this analysis include a lack of understanding of the context of the data, such as the selection criteria for the students in the dataset, the timing of the data collection, and the availability of data on other potential factors that could affect the first year GPA.

In conclusion, while this analysis provides insights into the factors that impact the first year GPA, there are some limitations to the analysis that should be addressed. To better understand the predictors of first year GPA, future studies should incorporate a broader range of variables and consider nonlinearity and non-normality assumptions.