Document

I initially started with data exploration and creating the necessary plots to visualize the data effectively. I used boxplots or bar graphs, depending on whether the data was numeric or categorical. Then, I cycled between feature selection and creation, model creation and training, and inference.

During feature selection and creation, I struggled to find four good features for many reasons. I used various methods to select features, such as analyzing the correlation matrix and SelectKBest. However, the first roadblock I faced was the issue of G2 influencing the prediction so much more than G1 even though they were both similarly correlated to G3. In fact, for every increase of G2 by 1, the prediction for G3 increased by a factor of close to 3.4. This factor was below 0.5 for G1. Every other feature besides the number of failures did not have a significant influence on the prediction for G3. I fixed this issue by averaging G1 and G2.

Next, I faced the issue of positively correlated features negatively affecting the model predictions. I tried selecting and creating new features hundreds of times. Additionally, I tried modifying the model by adding L2 (Ridge) regularization to penalize larger terms. I stayed in the feature selection process for hours until I decided to add the features *failures* and *higher* (higher_no after One-Hot Encoding with pd.get_dummies).

I ended up selecting avg_grade and failures_plus_higher_no as my two created features to train the linear regression model. I did not use L2 regularization and opted for the standard linear regression provided by Scikit-Learn instead.

The main insight I found from this model is that grades matter. Therefore, it makes sense to implement learning strategies and try to understand the material to perform well on assignments and tests. Additionally, one must have the desire to pursue higher education because this results in more effort, which helps boost grades.