## Project Update

### Group Members: Pallabi Chandra, Luwei Wang, Sunny Shah, Matias Gutierrez

**General Progress:** So, we've decided to go ahead with the student performance dataset for the subjects of Portuguese and Mathematics, where we're trying to predict the final grade ('G3') for the student. So far, we've done EDA for both sets of data and obtained statistical summaries.There were no missing values or outliers for any feature. We started off with the data preparation and preprocessing for the Portuguese dataset, where we got the data into a more acceptable data frame format first (original data had a list of entries with values for attributes separated by ';' as a separator instead of a traditional dataframe csv). For binary attributes- we replaced either possibility (represented by strings) with 0 and 1. We used the pd.dummies function for categorical features and finally added these columns to the dataframe and dropped the original categorical columns. We checked the correlation matrix between all attributes. We finally used Linear regression to fit the dataset after doing a 80/20 split. If 'G1','G2' are used as features, performance on the test set is very good: r-squared=0.85, whereas if we don't use 'G1' and 'G2', r-squared value is= 0.16.

**Concerns/Challenges:** Datasets are small, (especially mathematics with only 395 rows), but seems like it's big enough to catch a signal. We hope to get a similar performance for mathematics as we did for portuguese.

**Interesting fact from EDA:** There seem to be no missing values, erroneous values or significant outliers in the dataset. It is seen that most of the binary features are in the form of strings so they need to be preprocessed. Another important fact is that the 'higher' attribute seems most correlated with 'G3' after 'G1' and 'G2'. Yet another interesting observation is that the attributes of 'Medu' and 'Fedu' are correlated to a decent degree which means that multicollinearity for the attributes needs more investigation.

**Model fit insight:** A simple model like Linear regression does very well on the test set if we use 'G1' and 'G2' (strongly correlated to 'G3'). However, if we don't use it, then we don't get good performance showing us that we may want to use more complex models/more feature engineering for this case. If 'G1', 'G2' are used as features, performance on the test set is very good: r-squared=0.85, whereas if we don't use them, r-squared value is= 0.16.

**Analysis plan:**

1. Repeat the same exercise for the mathematics dataset.
2. Study the attribute correlation a bit more to either do feature selection/weighted feature importance/multicollinearity tackling and feature engineering to see if performance improves.
3. Study different models and their performance for the same problem- decision tree regressor, MLP, neural network etc for 'G3' prediction.