- general statement on status/progress, along with any concerns or challenges
- one interesting finding from exploratory analysis (could be a statement, or a data graphic, with a description)
- one insight or output from a model fit
- Analysis plan

Update: 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 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 prep and preprocessing for the portuguese dataset, where we got the data into a more acceptable dataframe format first. For binary attributes- we replaced either possibility 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 column. We checked the correlation matrix. 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: r-squared: 0.16.

Concerns/Challenges: Datasets are small, (especialy mat), but seems like it's big enough to catch a signal. We hope to get a similar performance for mat as portugese.

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 the 'higher' attribute seems most correlated with 'G3' after 'G1' and 'G2'

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: r-squared: 0.16.

Analysis plan:

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