- **What problem did you select and why did you select it?**
    - Can we accurately predict a movie's IMDB rating based on various features?
    - We selected this topic because it's very relevant to the movie industry. We anticipate being able to pull out meaningful insight about what predicts a highly rated movie in addition to interesting demographic information about the ratings.
- **What database/dataset will you use? Does it need to be cleaned?**
    - The dataset we will use is the "IMDb movies extensive dataset" on Kaggle. The dataset contains data from IMDb, a website that contains information about movies and actors, that allows users to rate movies. The dataset has information on actors, movies and their ratings, and the users who have rated movies.
    - The dataset contains some missing values and we anticipate there may be some errors in the data or mismatched datasets that we will need to clean.
    - Dataset link
- **What data mining algorithm will you use? Will it be a standard form, or will you have to customize it?**
    - This will be a supervised multiclass classification problem. Some models we can use include: logistic regression, tree based models (random forest or boosting), KNNs.
    - We will narrow down 2-3 of these models and use sklearn's base implementations of these models and adjust hyperparameters during cross validation. We will then choose the final model with the best F-1 Macro score.
        - We will most likely start with a logistic regression as a baseline model and a tree based model to see if we can yield higher performance, however, we will wait on final decisions on choosing specific classification models until we have explored the data more.
- **What packages will you use to implement the model? Why?**
    - For Data preprocessing we'll be using Pandas since our data is in a table format.
    - We also anticipate using Numpy and Scipy for basic statistics
    - For figures we anticipate using Seaborn and Matplotlib
    - For the GUI we'll be using the PyQt5 package we'll be learning later in the semester
    - For the predictions we anticipate using sklearn but we're waiting until we cover this in class and confirm the packages we should be using. Sklearn has many built in functions that make it easy to implement, train, and predict different machine learning pipelines.
- **What reference materials will you use to obtain sufficient background on applying the chosen model to the specific problem that you selected?**
    - Understanding F-1 Scores, Precision, Recall, Confusion Matrix
    - Sklearn:
        - Logistic Regression
        - Random Forest Classifier
        - Gradient Boosting Classifier

        - KNN Classifier
    - Understanding IMDb ratings
- **How will you judge the performance of your results? What metrics will you use?**
    - Using F1 Macro Score (supplement with Confusion Matrix, Precision, and Recall for more granular results and depending on if we want our model to be weighted towards a higher Precision or Recall).
- **Provide a rough schedule for completing the project.**
    - Proposal**:** 11/01
    - Code done + rough GUI done: 11/22
    - Rough Draft of presentation: 11/29
    - Final Presentation and report: 12/06