Group G Work Split ups:

1. Get everything set up - Jonny and Cihan
   1. Download Data
   2. Preprocess - find NANs missing values
2. Exploratory Data Analysis - Jared and Manny
   1. Find some relationship between predictors
   2. Dimension reduction to words
   3. Make some Plots
3. Break into groups - Each take one Method
   1. Predict Rating from plot description - Did not have time
      1. Make a heatmap of ratings. Try to cluster
   2. Trends of genres over time - Did not have time
   3. Predict Genre from plot - Did not have time
   4. **Predict if spoiler from review** - Main Goal -Jared, Cihan, Manny, Jonny
      1. Compare the reviews to movie
      2. Approaches to try:
         1. Deep network
         2. RNN
         3. SVM and kernel method - Jared
      3. Compute precision and recall (and other ways to compare models?) - Jonny
   5. Cluster Reviews - Did not have time
      1. Maybe see if there is some connection between reviews clusters and ratings
4. Need to:
   1. Read JSON to pandas - Jared
      1. df = pd.read\_json('**data/simple.json**')
   2. Text reading - Cihan
      1. NLTK
5. If extra time: Can we find another data set with reviews and predict on it - Did not have time
   1. Maybe RNN with each word as a full vector then pass in sequentially
   2. Can use some form of lower dimension embedding to reduce data size

**Progress: Friday evening**

* As a group, we made shared documents and Colab notebooks. We prepared a plan and imported the data set.

**Progress: Saturday morning**

* Cihan began preprocessing the data. This stage allows each of us to use the data on our own ends for different purposes
* Manny worked on editing the dates of reviews so they are in a standard, usable format
* Jared worked on plotting the data and finding trends between different attributes
* Johnny worked on standardizing the duration column and on the document

There were a few separate stages to preprocessing, so we all worked on it at different parts. Given a review as a single string, we had to split it into separate words, remove stopwords (unnecessary words), . This stage took a lot of computational time

**Progress: Saturday Afternoon**

* Many memory issues, data set was too large to work with in Colab
* Jared and Jonny got the data set up and exported to .pkl files to save time rerunning the first preprocessing stages each time
* Jonny set up a random selection of data to reduce from ~500,000 to 50,000. Still had to reduce to 10,000 to get more stability
* Cihan finished NLP part and we were able to turn sentences into 1-hot encoded vectors
* Jared, Cihan, and Manny begin testing out various models and set ups
* Jonny begins working on presentation

**Progress: Sunday Morning**

* Clean some last things up, bring all code back together, final touches on presentation, etc

**One Sentence Summary:** Used Natural Language Processing to convert reviews/descriptions of movies to a 1-hot encoded vector, and using that built several models to predict if a review contained a spoiler for the movie.