## Dec 2, 2023 | [Milestone 3 Meeting](https://www.google.com/calendar/event?eid=MnBwNnI2b285ZnFwbDhoY3BqNGtjbm9vMWUgc204OTY2QG55dS5lZHU)

Attendees: [Zaeem Shahzad](mailto:ms12297@nyu.edu) [Ritin Malhotra](mailto:rm5486@nyu.edu) [Soumen Mohanty](mailto:sm8966@nyu.edu) [Giorgi Kituashvili](mailto:gk2215@nyu.edu)

Notes

* Use statistical analysis for feature selection in MS3

**Action items**

*Milestone 2 Improvements*

1. Visualization and Analysis:

* Review and improve visualizations for clarity and effectiveness.
* Add error bars to average plots.
* Decide on the handling of multiple genres in movies and justify the approach.

The approach will be to create two bar plots:

* One showing one genre per bar
* The other addresses multiple genres as well i.e. Comedy AND Action

2. Notebook Enhancements:

* Improve notebook formatting for readability.
* Document the rationale behind each code segment and decision.
* Ensure each plot has a defined purpose and message.
* Review and update the notebook to include all relevant plots and tests.

3. Statistical Analysis and Feature Selection:

* ~~Clarify the basis of selecting informative features.~~
* ~~Include comprehensive statistical analysis and hypothesis testing.~~
* ~~Consider the age of movies as a potential variable to study.~~

Additions:

* Cramer’s V correlation test for preliminary correlation testing between occupation and genre
* ANOVA testing code block included (though I had it locally, I believe it was accidentally omitted from the Brightspace submission)
* Markdown cells explaining everything
* Outlier detection cell for Milestone 3 (let me know if I can add anything else for milestone 3 further EDA)

4. Additional Research and Analysis:

* Create additional visualizations for key variables (occupations, gender, etc.).
* Explore clustering occupations into categories relevant to the hypothesis.

*Milestone 3*

* Further Exploratory Data Analysis
  + - Heatmap for occupation vs genre
    - Group occupations in dataframe and display counts of genres to look at distribution of votes
    - Correlation Matrix
    - Cluster Occupations and plot the distribution of these clusters
* Feature Engineering based on EDA in Milestone 2 Improvement
* Predictive Model

Idea is to create a Genre Classifier for recommendations based on user demographics i.e. age, occupation etc.

Recommendations for Milestone 3 (from milestone 2 feedback):  
 In Milestone 3, I expect you to justify your choices of features. Why did you select this particular set of variables to study? How are they related to your hypothesis? For example, your hypothesis might be that watching exercise videos leads to a lower chance of getting a heart attack. You need to control for age as an extra feature, as you show that age is highly associated with a person’s chance of getting a heart attack.

A recommendation for milestone 3, maybe cluster occupations into 3 main categories (since this is directly related to your hypothesis)

Reiterating the hypothesis, especially during the presentation, and show how your model relates to it.