## Nov 11, 2023 | [Applied Data Science](https://www.google.com/calendar/event?eid=XzY4cGo4ZHBrNjBvajJjaGo3MTBrOGMxZzY4bWoyZzM5ZHBxNzRvYmpjbGlpc3Nya2VsaTZhcmprYWRpbW9waWpjbHA3Y3FiM2NrXzIwMjMxMTE0VDEwMTAwMFogcm01NDg2QG55dS5lZHU) | Milestone 2 (Visualizations) Meeting

Visualization ideas:

1. Map that shows average rating by state (using the zip-code variable). Possible relation to our hypothesis: We can then possibly look at the popular occupations and / or politics of each state to see if they have an effect on movie ratings in a given state
2. Visualization that shows average ratings for movies released across decades / 5 years. So, “1970-1975 movies had average rating of x while movies released in 1985-1990 had average rating of y”
3. How K-12 (and maybe college) students rate older movies (pre 1985) compared to newer movies (post 1985)

Ideas for (iii) Statistical analysis:

For the statistical analysis part of your project using the MovieLens 1M dataset, here are some suggestions to explore relationships between different variables and outcomes:

1. \*\*Computing and Plotting Correlations\*\*:

- Use a heatmap to visualize the correlation matrix of your dataset. This is particularly useful for identifying potential predictors for a predictive model. *Good idea we can do that for some of our variables*

- Focus on variables like user age, movie ratings, number of ratings, and perhaps timestamp data (if available), to see if there are any strong associations.

2. \*\*Analysis of Categorical Variables\*\*:

- Compute and visualize the average movie rating for each category within a categorical variable, such as movie genre, user occupation, or age group.

- A bar chart or a grouped bar chart can be effective to compare these averages.

3. \*\*ANOVA (Analysis of Variance)\*\*:

- Use ANOVA to analyze statistical differences among the means of different groups. For instance, you could compare the average ratings across different genres or user age groups.

- This analysis is valuable in identifying if the differences in means are statistically significant.

4. \*\*Pair Plots\*\*:

- Utilize pair plots to understand bivariate relationships between different variables. This is a great way to visually inspect potential relationships in a multi-dimensional dataset.

- Focus on pairs of variables that are likely to be related based on your understanding of the dataset, such as user age and movie preferences, or movie ratings and number of ratings.

5. \*\*Discussion and Interpretation\*\*:

- In your notebook and presentation, include a discussion section where you interpret the results of your statistical analysis.

- Justify why you chose certain tests and what the outcomes suggest about the relationships between variables in your dataset.

6. \*\*Further Exploration\*\*:

- Depending on the results of your initial analysis, you may identify areas for further exploration. For example, if certain genres have significantly higher ratings, you might explore what specific characteristics of these genres contribute to their popularity.

Describing and visualizing feature variables: Once you have identified the relevant features, we

would expect you to describe each feature/variable. This means:

1. Computing basic statistics for each informative feature variable (eg. min, max, range, median,

mean, standard deviation, and count -- if categorical, etc.) - *using describe / summary functions maybe?*

2. Plotting histograms or box plots of the most informative continuous variables to describe and

understand their distribution.

3. Plotting variables across time (if available). Average rating of movie/genre across time??

4. Plotting variables after stratifying/grouping your data wherever necessary, and visualizing each

stratum/group separately or in comparison with each other (such as male vs. female, old vs.

young, etc)

Statistical Analysis: Finally, once you have described your data and relevant and informative features, we would like to understand the relationship between different variables, and the chosen outcome measure. This would require you to conduct some statistical analysis. Possible examples include:

1. Computing and plotting correlations between different variables of interest. This is especially

useful if your hypothesis necessitates creation of a predictive model. In such a case, correlations

can help indicate possible associations between variables for further exploration.

2. In case of a categorical variable, computing the average outcome of each element.

3. Using ANOVA if analyzing statistical differences among the means of two or more groups.

4. Using pair plots to understand relationships between different variables.