See Jane - Research Project Data Normalization	
Contact	Meredith Conroy meredith@seejane.org
Organization	See Jane | The Geena Davis Institute on Gender Media https://seejane.org/
Organization Description	Founded in 2004 by Academy Award Winning Actor Geena Davis, the Institute is the only research-based organization working collaboratively within the entertainment industry to create gender balance, foster inclusion and reduce negative stereotyping in family entertainment media.
Project Type	Data Science
Project Description	*What is the high level goal or purpose for this project* The goal of this project is to normalize data sets that the Institute has related to representation in advertisements. The Institute will use the normalized data to update their database. Generally, the Institute has done analysis as a one off project for different clients but by having normalized data this analysis can easily be done annually or as often as necessary. - Do we just need to normalize the labels (e.g. LGBTQ VS Sexual Orientation)? - IDE preference (Jupyter Notebook/PyCharm/…)?
Data Sets & Sources	*Links to data sets e.g. APIs where data will be sourced from, folders/ files in the google drive that we were given, links to sites to be scraped, etc.* The student teams will be given access to de-identified data sets that need to be normalized.
Suggested Steps	*Steps to complete the project including data collection, data cleaning/ processing steps, and analysis* ● Gather all data sources/datasets (Should be done by clients or us? Waiting for the datasets) ○ Merge the data sources (does this mean clients will give us several data files and we need to merge them into one first?) ○ Normalize the data (just by labels? Other requirements?)

	○ Be able to disaggregate the data by year (Adding a column with regards to year, is enough?) ● Data analysis (does this mean those “questions to be answered”?) ● Stretch Goal: create a dashboard for visualizing the data that has been cleaned and normalized. What kind of dashboards do the clients want? Need an example.
Questions to be answered in Analysis	*Very specific questions that the clients wants answered - maximum of 5* Are there industry differences in representation in advertisements? Is there a change in representation in advertisements over time? Are there representation trends over time? What are the trends?
Ideal Output + Final Deliverable	*What does the client want in-hand at the end of the semester? What format would the client like the final report in (word, ppt)?* The final deliverable for this project will be a CSV (Excel or Google Sheet) containing the normalized data. Potential risks: 1. When can we have the data? 2. Rules for cleaning, merging, and normalizing the data
Additional Information	*Other relevant information including links to previous work if this is a project continuation* For each project, there is a different codebook, which will be shared with the project team. Unfortunately, our variable names and labels weren’t consistent across projects until a few months ago, so the normalization process will also involve streamlining variable names and labels. Need clarification (codebook VS dataset)