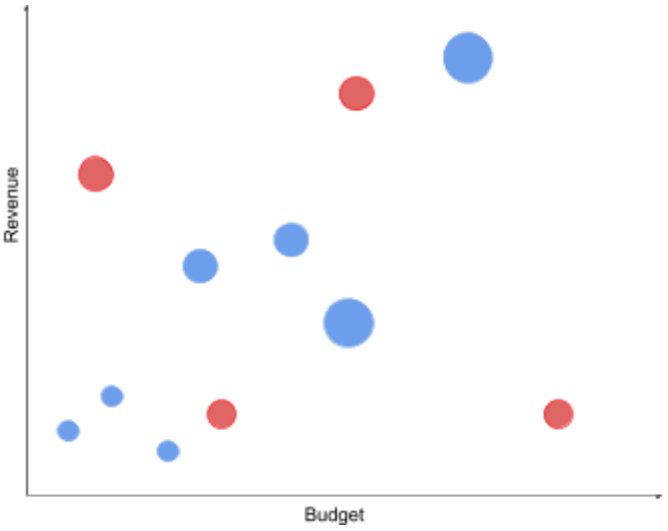
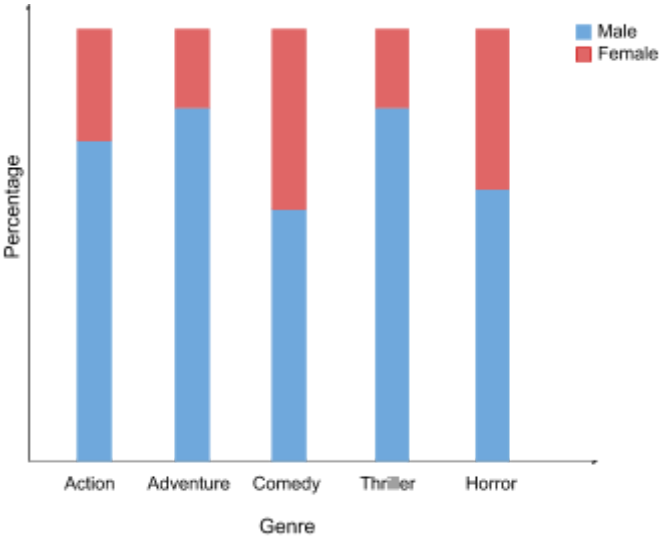
In our project, we want to develop a visualization that will enable us to explore the gender representation in Hollywood throughout the years. To do this, we will use the TMDB 5000 movies dataset that we described in Milestone 1. We would like to develop several different visualizations that will explore different dimensions of the dataset. All visualizations will allow the user to filter the movies included in the visualization using a slider where the user can select a range of years. An example of such a slider can be found at https://codepen.io/jameswilson/pen/LEjgem. In the following, we will sketch and describe the visualizations that we plan to use.

1. **Scatter plot**



In this visualization the user will be able to see the movies in the selected time range. The x-axis will represent the movie's budget and the y-axis will represent the movie's revenue. The size of the point will represent the movie rating and the color will represent the gender of the main actor. In this way, we will be able to analyze several different dimensions at the same time. We will be able to see how the gender representation changes over time and whether there is a correlation between the gender of the male actor and the budget, revenue and success of the movie. Apart from this core functionality, we plan to add some extra features like being able to click on a movie and see more details like the distribution of male and female actors or crew. Furthermore, we might add a functionality to choose additional criteria for coloring the nodes. Instead of coloring according to the main actor, we might color the nodes according to the gender of the director or some other behind-the-scenes roles. Additionally, we might add a functionality to color the nodes using a gradient which represents the overall percentage of female cast/crew in the movie.

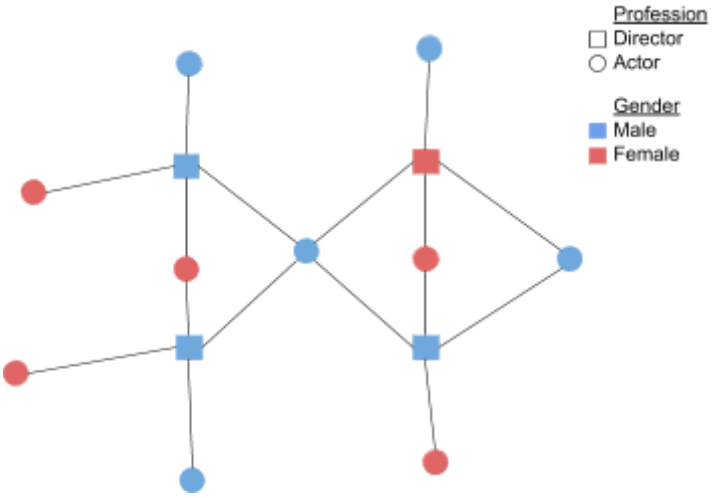
2. **Stacked bar plot**



In this plot, we will allow the user to choose a time period and the categorical feature to explore. The categories of the chosen categorical feature will be shown on the x-axis. There will be one stacked bar for each category representing the percentage of movies in that category with a leading male and female role. The categorical features we want to

explore include the genre, the production company, the country and the original language. This will enable us to explore the representation of female actors in different genres, countries and production companies. If the time permits us, we might add a functionality that enables the user to choose another role in the movie to explore like the director or some other behind-the-scenes role or explore the overall percentage of female character or crew in the movie.

3. Graph visualization



In this visualization, we will let the user choose a time period and we will present the most popular directors and actors using a graph representation. The directors will be presented with squares and the actors will be presented with circles, so that they can be easily differentiated. The color of the node will represent the gender of the person. There will be a link between a director and an actor if they collaborated in some movie. In this way, we will be able to see if some directors have a tendency to collaborate with male or female actors and whether the gender of the director or the time period have any influence in this. An extra feature that we might add is the ability to click a node in a graph and look at the distribution of female actors/crew members that a given director/actor has collaborated with.

4. Word Cloud

In this visualization, we will let the user choose a time period, and we will show two different word clouds, one from movies in which the main actor is male, and one from movies in which the main actor is female. The word clouds will be generated using the column that contains the overview for each movie. This will enable us to analyze whether there is thematic difference between movies starring male and female actors, and whether the themes change with time. An additional feature that we might add is the ability to choose another role for creating the word clouds instead of the main actor, like the director or other crew member.

Our initial idea is to create one dashboard with all four visualizations in it. At the bottom of the dashboard we will have the time range selector and the selected range will be applied to all four visualizations. This dashboard with the core functionalities for each visualization will be our minimal viable product. We will use Python with Pandas for preprocessing the dataset and D3.js for creating the visualizations. We expect that we will need all the D3.js lectures from before the Easter break for designing and implementing our visualizations. Additionally, we will need the lecture on Graphs for our graph visualization and the lecture on Text visualization for our word cloud visualization.