Pavitra Dangati

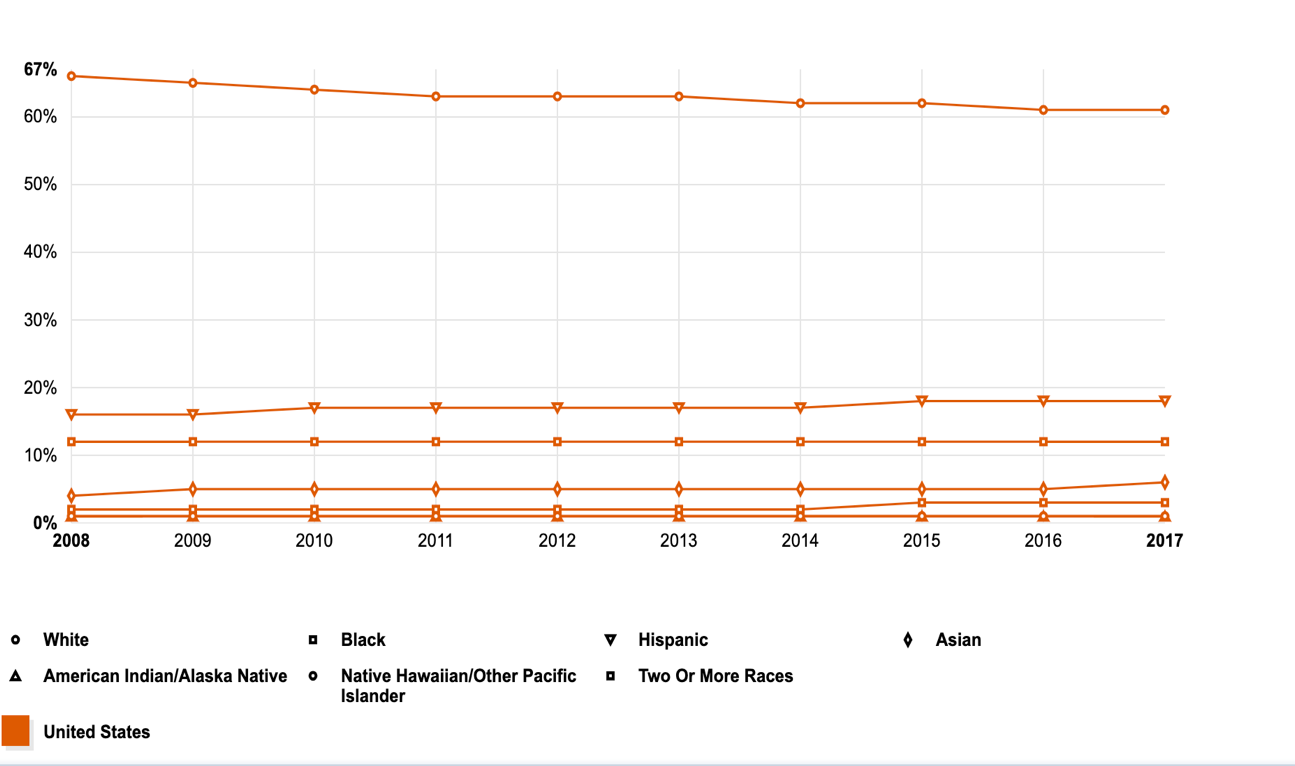
Manan Patel

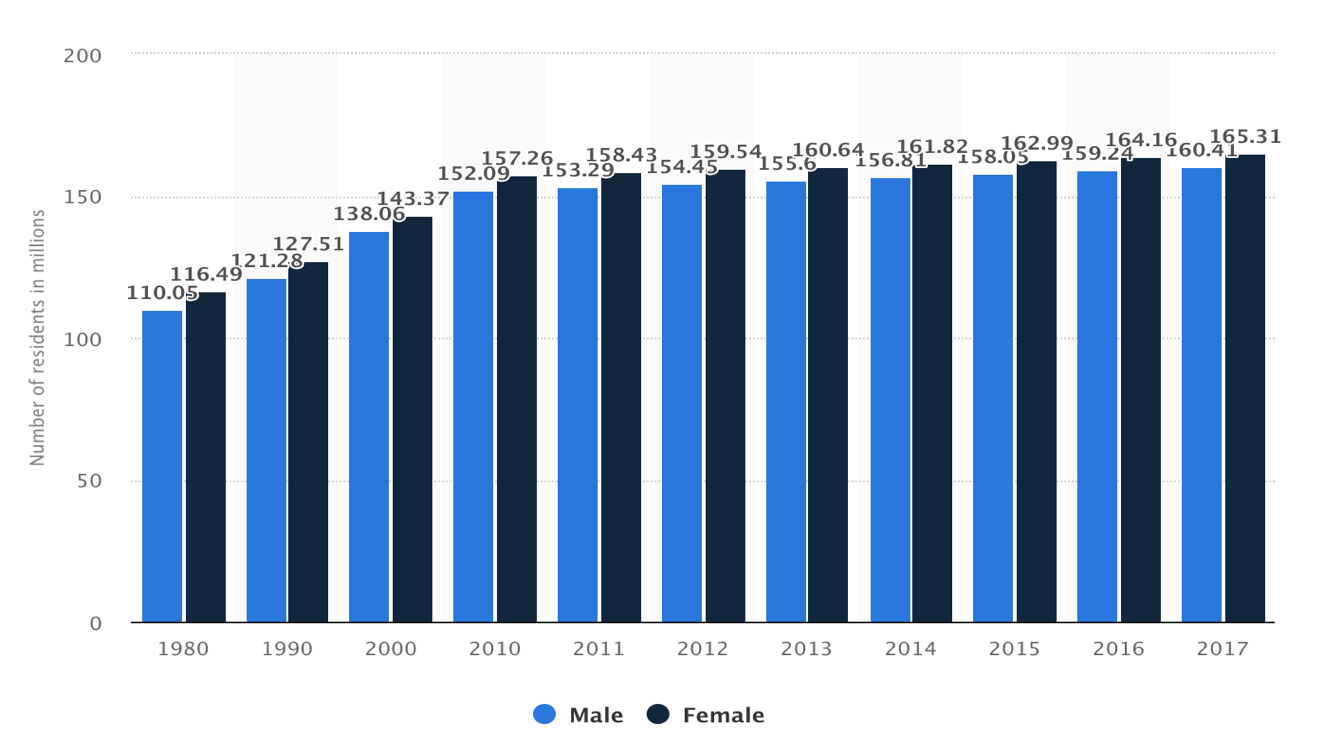
Thomas Lam

**Topic**: Analysis of Gender and Racial Diversity: in Hollywood Movies

**Github Link: https://github.com/thomaslam/hollywood-diversity-dataviz.git**

**Description:**

These are two visualizations depicting the racial and gender diversity in the United States over the past few years taken from the US Demographic Records.



From the two visualizations, you can see that in terms of gender diversity, the female ratio is slightly higher than the male over the past few years and in terms of racial diversity, the minority ratios have been slowly increasing over the years. However, has this growth and opportunities in every field the same for women and the minorities?  
  
Ethnic and gender diversity has been a long-standing issue which is gaining light over recent years in various sectors, especially in Film, Arts and Theatre. . If we consider top 200 movies released in 2016 and cable and digital platform tv shows from 2016-17 in order to which women and people of color are present in front and behind the camera. Although US demographic has become more diverse in recent decades, there’s been various studies which show that minority and female representation in the arts is severely lacking given such demographic trend. For this project, we were looking something for fun visualization on many sources and we found a data set on diversity in Hollywood and we also wanted to look at a subset of the creative outlet, that is, Hollywood movies and perform an analysis of racial and gender makeup of the cast and crew of movies from 1960-2016 timeframe

**Users who would be interested:**

We want to provide a visual aid that helps clearly illustrate this problem to a targeted audience consisting of Hollywood executives, decision makers, directors, scriptwriters as well as other people interested in exploring this social issue.

* Hollywood directors and producers
* Casting directors
* Scriptwriters and authors so they can share the minorities stories to a wider audience
* Film, art and theatre students
* Supporting social causes

**Task Abstraction: What would they want to see?**

Here are some questions that hopefully our visualization design can help answer:

1. How many female and minority actors/actresses are typically represented in movies compared to their white, male counterparts? How did this representation change over time?

2. Is there a diversity bias towards a select few genres?

3. Do the popular directors who have a bigger platform and opportunities to represent the underrepresented actually have high minority representation in their movies?

**Data:**

The dataset we’re looking at is the IMDB Data set from Kaggle (<https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset>) which contains information on 5000 different movies ranging from 1960-2016 containing attributes such as directors, durations, actors, ratings, genres, etc which is in csv format.

**Data Scraping:**The Kaggle dataset is missing the gender and racial information of the cast and the crew. For this, we implemented data scraping in python using BeautifulSoup module to supplement the data with this information. For each movie in the dataset, the IMDB website link from where we scraped the cast and crew information given on the website. After getting the names, we scraped the Notable Names Database to get the racial and gender information of the people and supplemented by the SexMachine and ethnicolr python modules incase the information wasn’t available on the website.

**Data: Features**

Our data has the following features:

* Movie name
* Director Name
* Actor’s Name
* Genre
* Ratings
* Budget
* Director popularity
* Actors Popularity
* IMDB Link
* Language
* Year
* Count of cast members for gender(Male/Female) and races(Whites/Asians/Hispanics/Blacks)

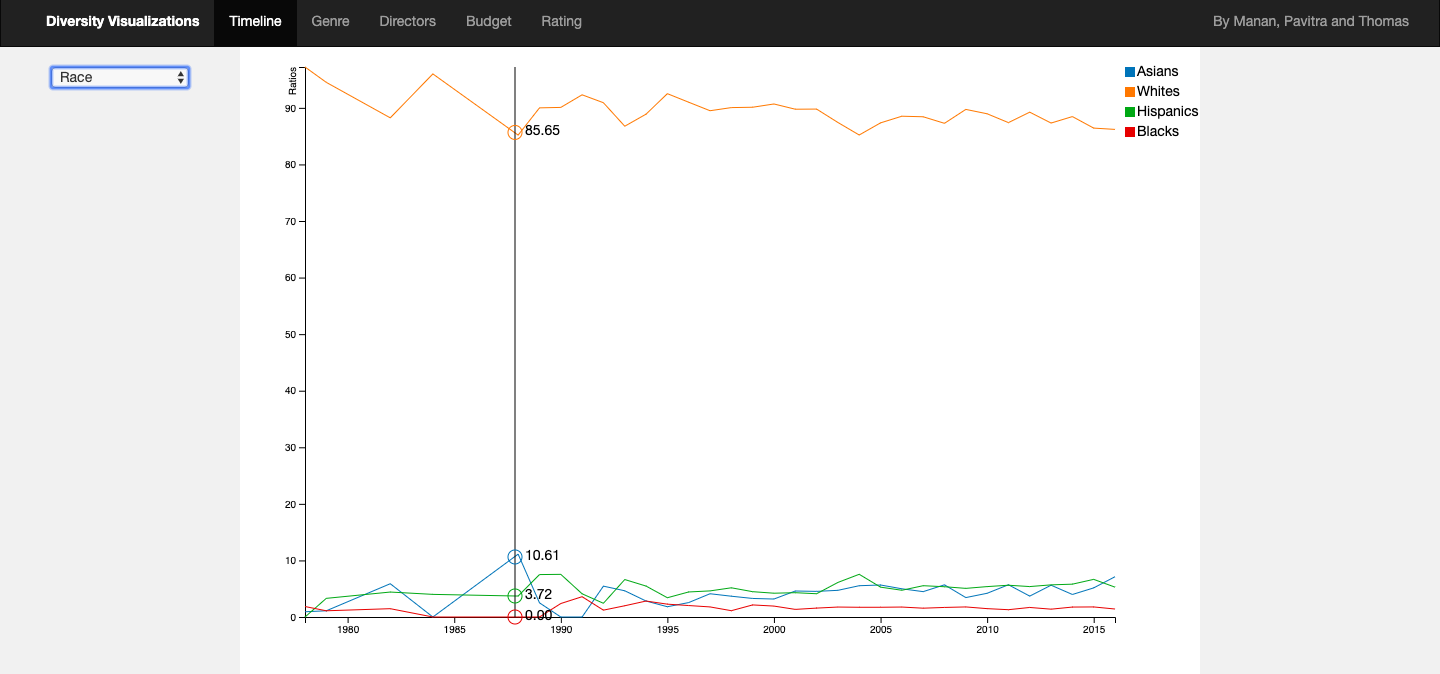
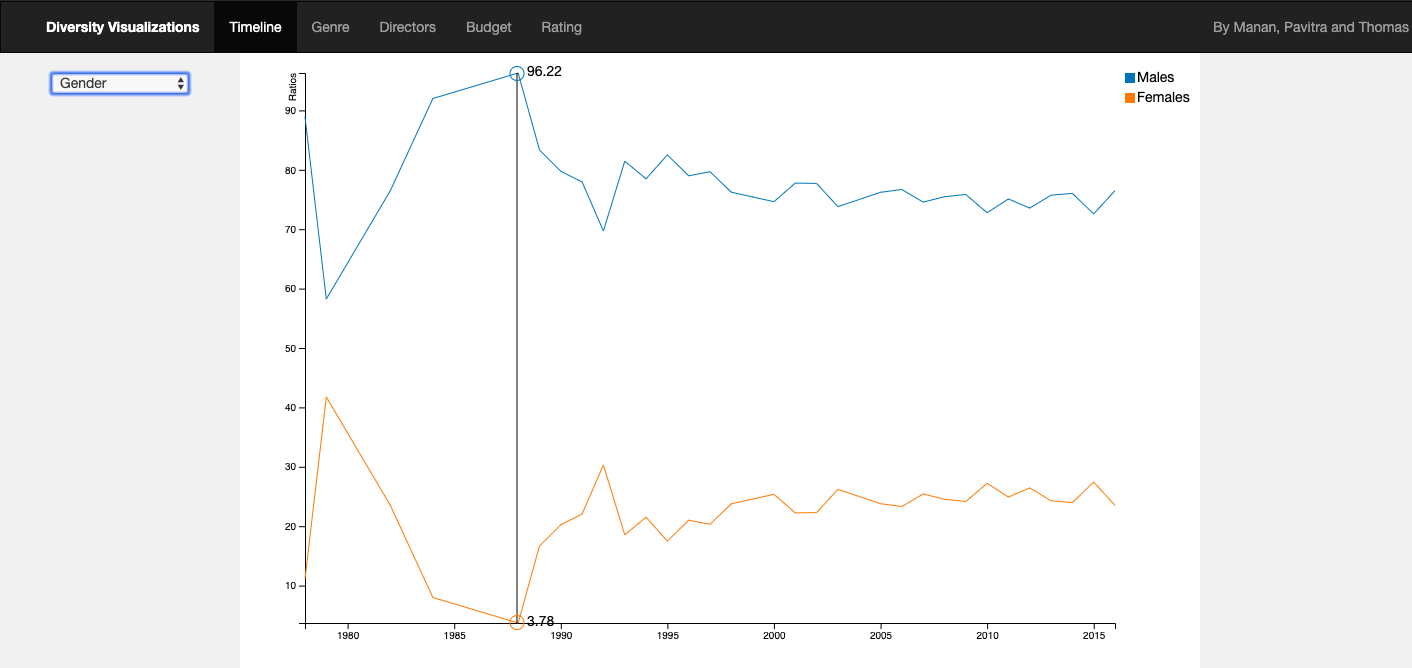
**Technologies used:**

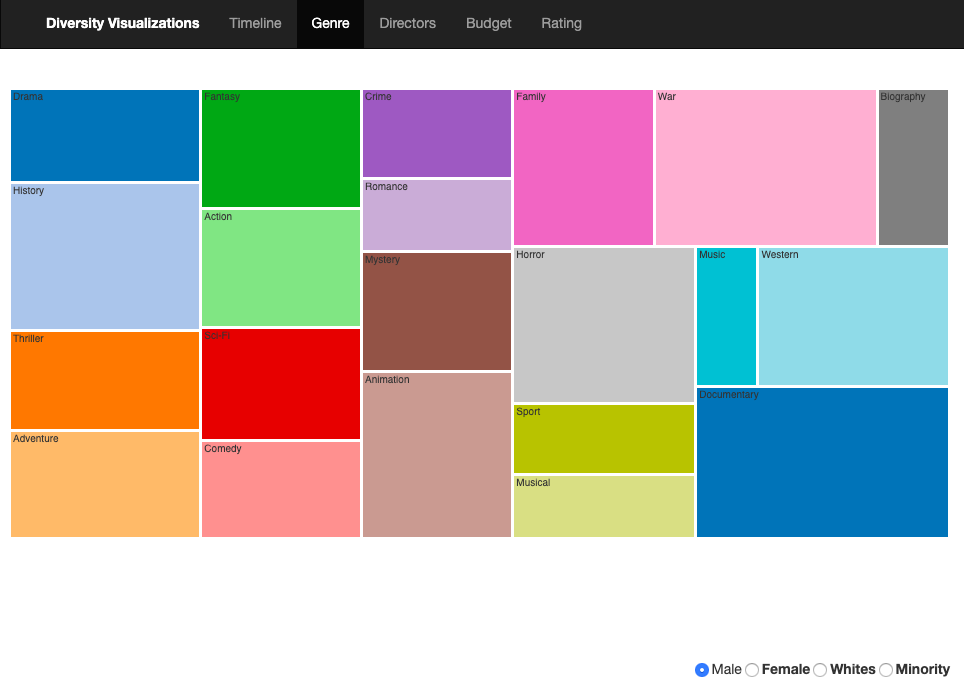
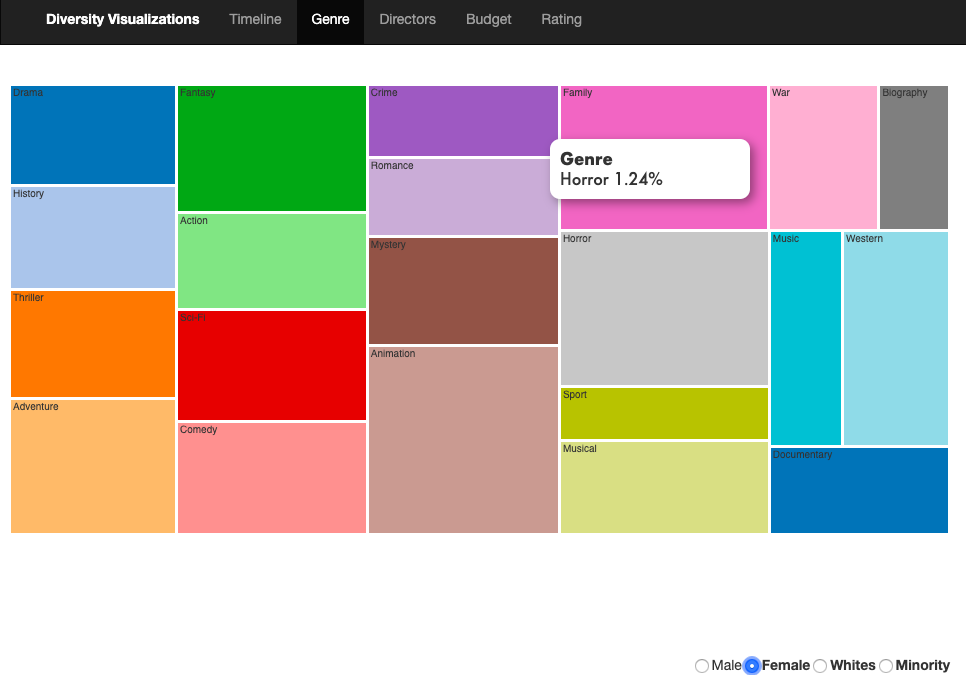
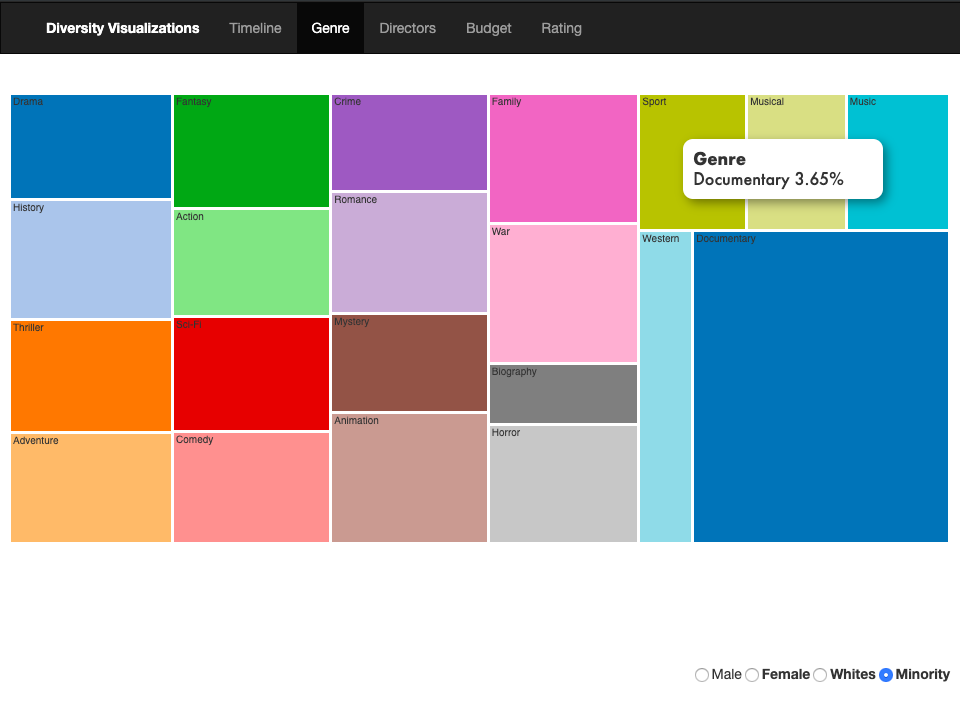
* Python
* Beautiful Soup, Sex Machine, Ethnicolr
* D3
* Javascript
* HTML/CSS
* Bootstrap

**How to run:**

**Data Preprocessing**: Run the preprocess.py followed by merge\_csvs.py  
**Visualizations:** A python server needs to be started in the Html folder in the terminal.   
For python 2: python -m SimpleHTTPServer 8000   
For python 3: Python3 -m http.server 8000   
These commands start a server on port 8000. To open the index page, open localhost:8000 on the browser.

**Task 1:** **How did the gender and racial diversity change over the years?**

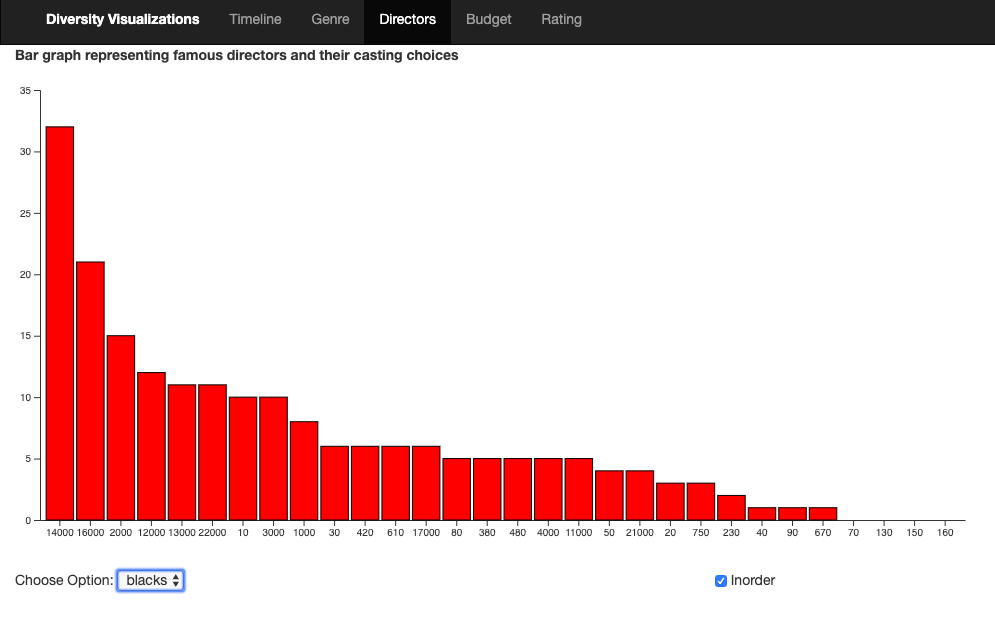
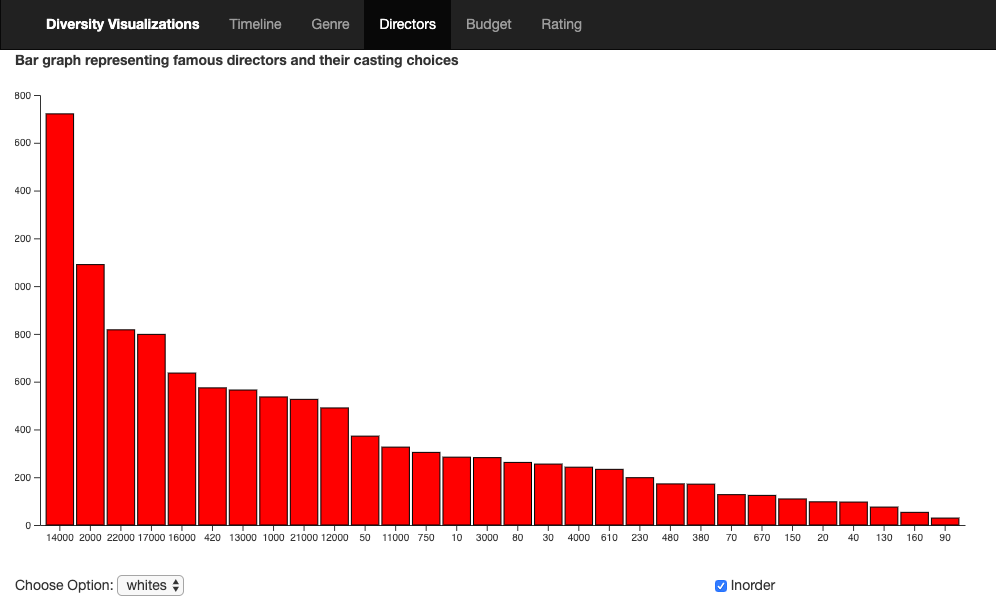
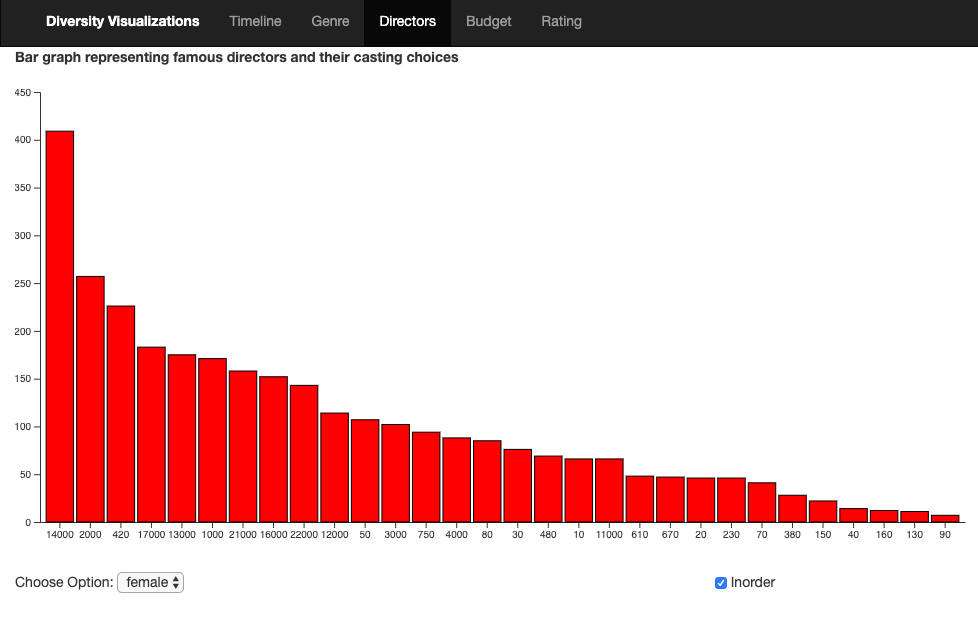
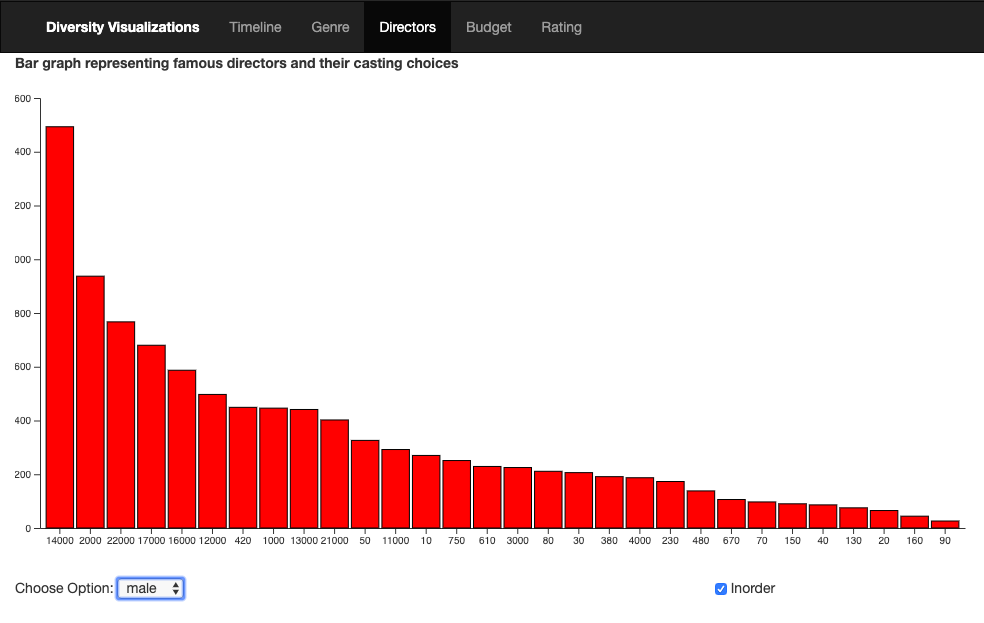
For this task, our proposed solution is a line chart with interaction to showcase the values when hovering over the points.   
Marks: Line  
Channels used: X axis Position for timeline, Y axis position for the percentages, color to show different categories such as Male/Female, Whites/Asians/Blacks/Hispanics.  
  
  
**Insights:**  
Based on Race: The percentage represented in the movies is lower than actual percentages seen in page 1. However, the trends are similar. The Minority representation increases over the years, similar to the actual growth trend seen previously and the white representation reduces over the years.  
Based on Gender: According to the US Demographic trends, females are slightly more than males. However, over the years in hollywood, the female representation has been always lower than 50%. There has been a growth from 1995 onwards. The graph shows a peak at 1960 however due to the number of movies considered for that year are very low and the average is higher.

**Task 2**: **Is there a diversity bias towards a select few genres?**For this problem, we considered genres individually with each category to see how the genre distribution is.  
Marks: Area   
Channels: Color used to make transition easier between different categories. **  
**

**Insights:**

Male representation: Documentary and war have the highest split with Musical being the least.  
Female representation: Horror, Thriller, Musical, History being the highest with biography, sport being the least.  
Whites: Documentary, Family being the highest and the musical being the lowest.

Minorities: Biography, Sport and Musical are the least.  
  
The Sport and Biography genres barely have any underrepresented categories. Documentaries usually are biased towards Male/White representation.

**Task 3: Do the popular directors who have a bigger platform and opportunities to represent the underrepresented actually have high minority representation in their movies?**For this problem, we chose a bar chart to compare the values easily.  
Marks: Line  
Channels: X axis represents the director’s popularity in terms of facebok likes.  
 Y axis represents the cast ratios of a category. **  
  
Insights:**   
The directors who are in the medium range of popularity represented more of the minorities.

**Design Ideas:**