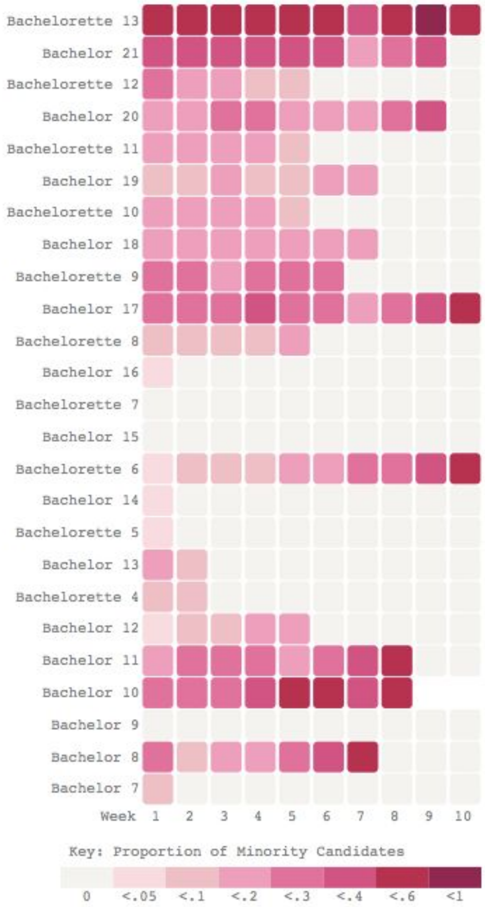
Minority Representation in *The Bachelor* Franchise

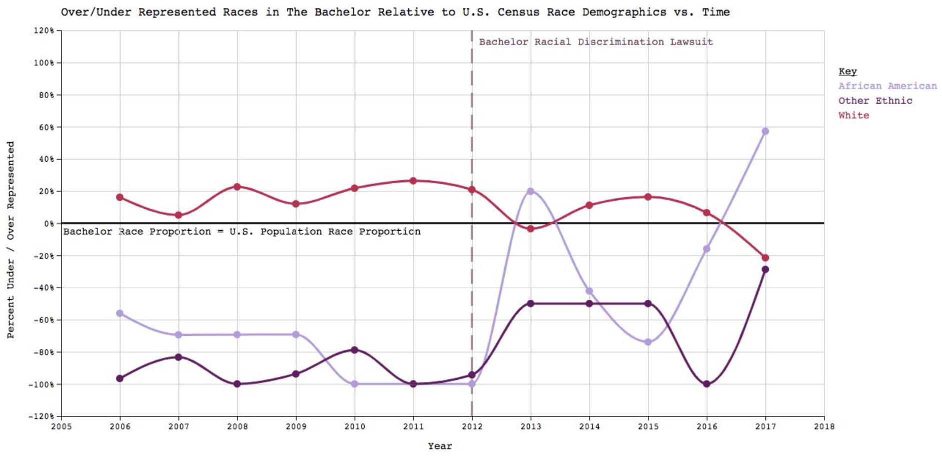
A. GitHub Repository Link: https://github.coecis.cornell.edu/jb2338/3300project1.git

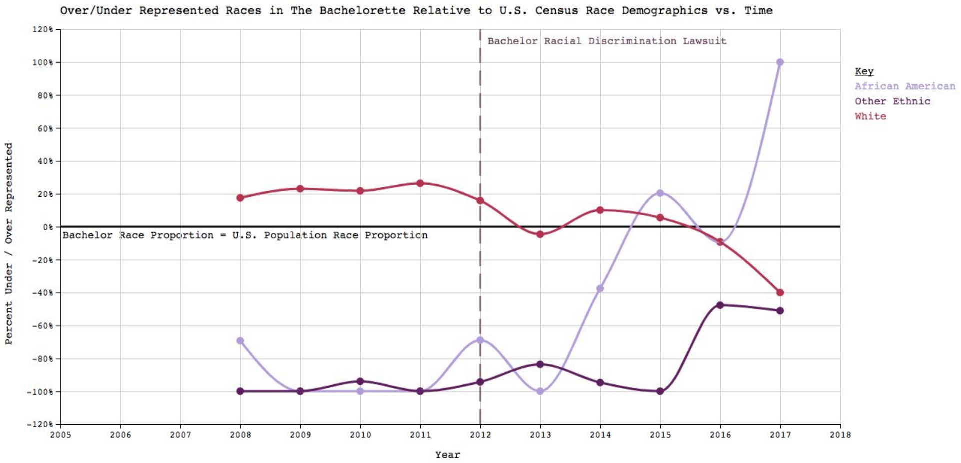
Static Visualizations:

1. *Heat Map*

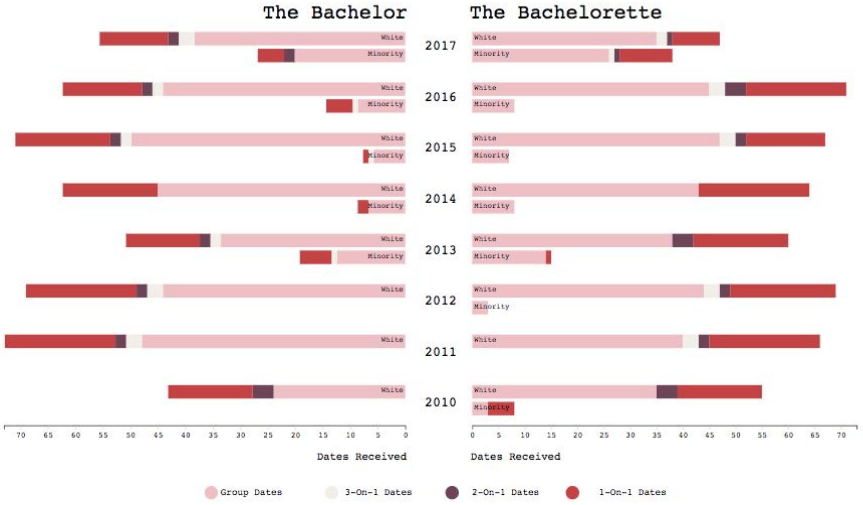


2. *Line Graphs*


Over/Under Represented Races in The Bachelor Relative to U.S. Census Race Demographics vs. Time


Over/Under Represented Races in The Bachelorette Relative to U.S. Census Race Demographics vs. Time

3. *Stacked Bar Graph*



B. Data Description: We sourced our baseline data regarding contestants' information on their time on *The Bachelor* and *The Bachelorette* from the data journalism website FiveThirtyEight. This data set can be found here. Since we knew we wanted to analyze this data in the paradigm of minority representation, we researched online for data sources containing racial information regarding the contestants. However, we were unable to find any comprehensive dataset detailing the race of contestants in the franchise. Thus, we manually inputted the races for all of the seasons included on the original FiveThirtyEight dataset. This information was primarily based on photographic evidence on websites such as bachelor-nation.fandom with additional research via Google, or articles such as "A History of Black Contestants on The Bachelor and The Bachelorette". We also used a supplemental dataset for our second visualization (line graph) in order to compare the proportion of certain races on *The Bachelor* franchise to the proportion of those races in the U.S. population. This dataset was sourced from the U.S. Census Bureau Data for Races and U.S. Census Bureau Data for the Hispanic Population.

The variables within the original FiveThirtyEight Bachelor dataset are Show, Season, Contestant, an elimination variable for each week of the show (1-10), and a date variable for each week of the show (1-10). The show variable simply indicates if the contestant competed on *The Bachelor* or *The Bachelorette.* The season variable indicates the season number the contestant competed on. The contestant variable gives an identifier for each contestant in the format of SeasonNumber_FirstName_LastInitial. The elimination variables for each week are a modified form of dummy coding. If a contestant is eliminated that week, the value for the contestant on that elimination variable is "E", potentially followed by another letter indicating the type of elimination. "EQ" means the contestant quit, "EF" means the contestant was fired by production, "ED" means a date elimination, and "EU" means an unscheduled elimination taking place outside of a date or rose ceremony. If a contestant receives a rose outside of a rose ceremony that elimination week, the value for the contestant on that elimination variable is "R", which is followed by "1" if the rose is a first impression rose. For each date variable, the value for each date variable is either blank, indicating the contestant did not receive a date that week, or "D" followed by the number of women present on the date.

We added two additional variables via manual data entry regarding contestant race. One variable is Race, which indicated if a contestant was African American, Asian, Hispanic, White, two or more races, or other. The value for this variable is simply a string containing the race the contestant was. The other variable we added is a Minority (Yes/No) variable that stated if a contestant was considered a minority. The value for this variable is a string of either "Yes" or "No".

The variables within the original U.S. Census Bureau Data for Races are Year, White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian and Other Pacific Islander, and Two or More Races. The year variable indicates the year the census data was collected, and each race variable indicates the number of the people in the U.S. of that race in that year in millions. The variables within the original U.S. Census Bureau Data for the Hispanic Population are Year and Hispanic Population in millions. Similar to the other census dataset, the year variable indicates the year the the census data was collected. The Hispanic Population in millions variable indicates the number of hispanic people in the U.S. in that year recorded by the census.

When inputting our data for race, we discovered that for the first six seasons of *The Bachelor* (seasons prior to 2006), and the first three seasons of *The Bachelorette* (seasons prior to 2008), racial data was often inadequate. Thus, we filtered out data from these seasons.

Data was reformatted in different ways in order to allow for the various visualizations to be made. We explain reformatting and filtering for each visualization below:

1. *Heat Map:* We had to reformat the data through a Jupyter Notebook in order to transform the dataset into a json containing individual data points with the ratio of remaining contestants who were of minority origin for each week of each season. This reformatting can be found in our Git repository as an ipynb file. The final adjusted data can be found in 'teresa.json.'

   First, the seasons were ordered chronologically, as while both *The Bachelor* and *The Bachelorette* were ordered, they were not integrated. We needed the data integrated chronologically to see changes in the heat map over time easily. This is also the reason why this is the only visualization with season numbers rather than years. Since many seasons happen during different times in one year, it makes more sense to have ordered seasons rather years for this visualization in order to see changes over time.

   Following this, the total number of contestants present each week for each season was determined, followed by the number of minority contestants present each week. Finally, these numbers were divided in order to determine the ratio. Our outputted dataset has variables Order, Week, and Ratio. The Order variable indicates the chronological position of the season; 1 indicates the earliest season included, while 25 indicates the most recent season included. The Week variable indicates which week of the indicated season the ratio is for. Finally, the Ratio variable indicates the proportion of minority candidates for the specified week and season. This dataset can be found in teresa.json. Creating a separate dataset allows us to only load in data that will be used in the visualization, allowing for greater efficiency.

2. *Line Graph Comparison:* The U.S. Census data was reformatted into proportions to better suit our purposes. First, in the U.S. Census Bureau Data for Races, we summed the total number of people in the U.S. for each year. Then, for every year, each race was converted to a proportion by dividing the number of people in the race by the total number of people in the U.S. for each year. Similarly, the hispanic population in millions from the U.S. Census Bureau Data for the Hispanic Population dataset is divided by the total number of people in the U.S. for each year. This allows for standardization in terms of data size as the U.S. census number of people is obviously much greater than *The Bachelor* and *The Bachelorette* number of contestants per year. All of these calculations are done in excel, primarily via the SUM function, and simple division. The hispanic proportions are copied over to the file with all the other racial information in proportions, and this is the data that is loaded into our code. The reformatted census data can be seen in our Git repository in census.json.

In terms of formatting and filtering *The Bachelor* data, this is done solely in the code. First, the seasons are mapped to years. This is due to the fact that the census data is in years, as well as the fact that important events regarding the franchise are more easily identified through years. We decided to keep different start years for *The Bachelor* and *The Bachelorette* as the amount of data was already limited, and we wanted to retain as much data as possible. Following this, we made a counter in order to count how many total contestants participated in each year, as well as how many contestants of each race participated in each year. The counts allowed us to then calculate proportions. For each year and each race, the number of contestants in each race was divided by the total number of contestants that year.

After formatting and filtering both *The Bachelor* data and the census data, we created the final dataset to be used in the visualization by calculating the percent difference between The Bachelor race proportions and U.S. Census race proportions. This allows for easy interpretation of percent over representation and under representation. The data for each race described above (Hispanic, Asian, African American, White, two or more, and other) were initially used to plot the line graph, but upon plotting, the graph was overcrowded, with many overlapping lines at -100%. Thus, we decided to group Hispanic, Asian, Other, and two or more into one category (Other Ethnic) by averaging the percent differences to allow for a simpler visualization.

3. *Stacked Bar Graph:* While researching data for this visualization, we discovered that recent date structures were different from date structures prior to 2010. For example, contestants often were invited on more than one date per week prior to 2010, when it is impossible to receive more than one date per week after 2010. Thus, data was selected between the years 2010 and 2017. To produce the stacked bar graph, only relevant attributes from the data set were selected, including the number of dates per contestant. The length of each bar is determined from an additional attribute, their total number of dates, which was added to the object representing each individual. As the dataset we used provided only season numbers, and *The Bachelor* and *The Bachelorette* began at different times, the original data array needed to be sliced in order to select the season corresponding to the years of interest.

   For this graph, the individual ethnicities were not taken into account. Rather, the Minority (Yes/No) attribute was used to separate the minority contestant data from the larger contestant pool. A new data array was created which held the attributes of interest for this graph. For a particular year, the objects in the data array are stored with *The Bachelorette* data first and *The Bachelor* data second. After this point, only the new data array was referenced in code.

C. Design Rationale: The general design for all of our graphs was based off of the aesthetic of *The Bachelor* franchise. Since *The Bachelor* is often associated with romance, we chose a color palette with different shades of pink, purple, red, and white. Furthermore, we chose courier as the font for labels in our plots as it has a classic connotation. Our design rationale for each visualization can be found below.

1. *Heat Map:* Marks include rounded squares. Visual channels include the aligned x and y positions of the squares, saturation of the color in the square, as well as text. The proportion of the number

of minority candidates remaining was mapped to the saturation of the squares at the designated position for each season and each week. The legend for the color mapping is located at the bottom of the visualization. The seasons are arranged chronologically such that the most recent season is at the top. Weeks are arranged in chronological order from left to right.

The boundaries for the ratio color cutoffs were determined based on the distribution of the data. It was important that the boundaries were not evenly spaced throughout the domain since most of the data points were less than 0.15, which would result in a relatively mono-saturated heatmap. While even numerical cutoffs would be more intuitive for users, boundaries based on the data distribution allows for greater differences in saturation on the heat map. We chose to value differences in saturation over evenly spaced cutoffs as saturation differences are a key visual channel in heat maps. Furthermore, since *The Bachelor* season 10 ended after week 8, we needed to decide whether to put 0 saturation colored squares there, or leave it blank. Since the season ended, we decided it would be a better visualization decision to leave those remaining spaces blank with no square, as putting 0 saturation colored squares there could imply that the show continued, and all ethnic contestants were eliminated. We make a note of this before the visualization. We decided to not put it on the visualization itself due to cluttering, and those curious about the missing squares could look in the description.

2. *Line Graph:* Marks include circles and lines. Visual channels include the aligned x and y positions of the lines and circles, hues of the lines and circles, as well as text. The percent difference in proportions data for each year were mapped to visualizations via circles positioned appropriately on the x axis and y axis. Points for each race were then connected by lines in order to elucidate trends over time. Colors were used as a visual channel to separate categorical data for each race. Each race was associated with a different color, allowing for easy interpretation. As mentioned in the data description section above, we decided to combine overlapping races that were similar to one another in order to de-clutter the graph. While the data loses specificity via this decision, it allows for clearer visualization.

   In order to allow for easy interpretation of the over vs. under representation, a thick black line at the 0% difference mark was included marking the positioning of the y axis where the race proportion in *The Bachelor* was equivalent to the race proportion in the U.S. census. This was labeled with nearby text to allow users to be able to easily understand the purpose of this line. Furthermore, the year in which The Bachelor faced a racial discrimination lawsuit is indicated via a brown dotted line. This was similarly labeled with text in order to allow for easy user interpretation. While text labels on the graph adds clutter, we concluded that the instantaneous information given by text labels outweighs the negative aspects of the extra clutter. Since we were already using a key to map colors to races, an additional key to explain these marker lines would increase confusion.

   When brainstorming visualizations for over and under represented races on *The Bachelor*, a bar graph was considered, as this is a common way to map over and under representation (Appendix 1). However, bar graphs are better representations for over and under representations at a static point in time. Since we wanted to see the trend over time, we decided that a line graph was the optimal visualization. This benefit of being able to visualize representation over time is not without drawbacks--the area aspect of the bar graphs allows users to easily map over-representation to the rectangles above the zero cutoff line, and under-representation to those below. To combat this, we added the black equal representation line described above. We additionally considered adding a bar for each race at each point in time, but upon implementing this visualization, it was evident that the graph was extremely cluttered, and it was difficult to discern percent differences changes for each race over time. Thus, the simple line graphs over time were determined as the best way to visualize this dataset.

To further elucidate over representation and under representation, we considered adding arrows on the right y-axis labeled with over and under representation respectively. However, this design cluttered the graph, and user feedback indicated confusion about which axis to refer to. Another approach we attempted to combat the issue of over/under-representation interpretation was shading the chart area above the zero cutoff line as a white-pink, and the chart area below the zero cutoff line as a white-blue. However, this once again added confusion, due to the potential misinterpretation that the shaded chart areas themselves were representative of over/under representation. Thus, we concluded that the black zero cutoff line was the most appropriate visual channel to convey over and under representation.

3. *Stacked Bar Graph:* Marks include rectangles with constant vertical height. The visual channels used include x-aligned position (horizontal length of bar), y-aligned position (year), and color of the bar segment. The most important consideration was in how to represent the proportion of dates accurately. The length of each bar represents the total number of dates received by minority or other contestants on both shows. Bars were scaled linearly according to this total metric. A linear scale was used for both *The Bachelor* and *The Bachelorette* data for ease of comparison. Each segment of the bar is meant to show the proportion of a certain type of date received by a certain group of contestants.

   By having two bars for each show and season, the distinction between both the number of dates received and the proportion of types of dates received for different groups of contestants is more clear. The spacing between bars corresponding to the same year is smaller than the spacing between bars of different seasons for ease of viewing. The colors were selected to represent the categorical data for each date type. Adjacent colors were selected to be distinct enough from one another while also keeping with the theme of the show. In addition, the order of the bars was selected such that less sought after date types are shown closer to the year labels and highly sought after date types are shown at the end of each bar. Finally, labels were added to each bar to make clear which bar corresponded to minority contestants and which corresponded to other contestants.

   A potential alternate design of our stacked bar graph was to have minority bars on one side, with Caucasian bars on the other side, instead of the current configuration with *The Bachelor* bars on one side, and *The Bachelorette* bars on the other side. However, this would result in most of the space on the minority side of the graph to be unused. If we scaled the different sides differently, the comparison of dates between Caucasian contestants and minority contestants would be difficult to comprehend visually. We also considered making the colors within each bar representing the types of different dates between minorities and Caucasians differently. However, having too many colors on the graph makes mapping the colors more complicated for the users. Thus, we decided to simply use labels.

D. Key Takeaways (The Story): We chose to produce three visualizations to tell the story of minority representation on *The Bachelor* and *The Bachelorette*. All of our visualizations attempt to expose different aspects regarding ethnic representation on the franchise.

1. *Heat Map:* This visualization shows several key insights. Preceding Bachelor Season 17, a class action racial discrimination lawsuit was filed against the Bachelor for under-representing minorities, and the heat map is noticeably darker above this point. This visualization also shows that Week 5 for the Bachelorette and Week 7 for the Bachelor have been popular "cutoff" weeks for minority candidates. Viewers of the show have speculated about a token minority theory- that the lead will keep minority contestants until a certain time frame of Week 5-7, and these decisions are likely dictated by producers. As evidenced by the trend towards darker shades as you travel up the map, minority representation has increased on the show. Further, in multiple seasons, a minority candidate has been proposed to, such as in Season 6 of the Bachelorette when Roberto Martinez, the only minority candidate of the season, proposed to Ali Fedotowsky- hence the ever darkening shades as the user progresses through the row of Bachelorette Season 6 data. In addition, the lead for *The Bachelorette* Season 13 was the first African American lead, which likely explains the darker heat map for that specific season.

2. *Line Graph:* Overall, for both The Bachelor and The Bachelorette, it is evident that the Caucasian race is over-represented in the shows in comparison to the overall U.S. population demographics. On the other hand, African Americans and other ethnic groups are severely under-represented in these shows. However, over time, it is evident that minority and ethnic representation on the shows has improved. The degree of under-representation of these minorities is worse from 2006-2012. As mentioned above, in 2012, Season 17, *The Bachelor* franchise faced a racial discrimination lawsuit. While this lawsuit was dismissed, it is evident that it had an impact on the franchise, with minority representation increasing dramatically, especially for The Bachelor. Additionally, in 2017, Rachel Lindsay became the first black lead on The Bachelorette. This explains the large over-representation of African American men in 2017. Overall, our line graphs show the improvement in adequate racial representation over time; however, it must be noted that these improvements are not without external factors, such as the lawsuit in 2012.

3. *Stacked Bar Graph:* It is clear that across 2010-2017, the number of dates received by Caucasian contestants outnumbers the number of dates received by minority contestants on both *The Bachelor* and *The Bachelorette*. However, the major takeaway from this visualization is that the proportion of date types given to minority versus other contestants is often very different. When looking at *The Bachelorette* from at least 2010-2016, Caucasian contestants are given a higher

proportion of sought after dates, like one-on-ones, relative to the total number of dates received. In 2017, when the first African-American Bachelorette was cast, this discrepancy in proportion of sought after dates decreased. Looking at *The Bachelor*, it is easier to see a slight overall trend in a reduction in discrepancy between the proportion of sought after dates received by minorities versus Caucasian contestants. Overall, one conclusion to be drawn from this graph is that minorities are, on average, not considered for higher value dates in the same way that Caucasian contestants are.

E. Team Contributions: We divided up the work such that each group member focused primarily on producing one visualization. As a team, we discussed these visualization ideas together beforehand and brainstormed the visualizations together. After running these ideas by our TA mentor, we each created our assigned visualization. We met up again after creating our visualizations, giving each other feedback and points of improvement regarding each visualization. We then worked together to create our final website and this rationale document. We spent around 5 hours per person working and polishing the final website with all of the visualizations integrated and this report.

For the milestones, each group member researched potential datasets, developed one or two visualization ideas once the dataset we wished to use was selected, and manually input additional data values to assist in our analysis- approximately 4 hours total for each of us. Please refer back to our milestone PDFs for specific tasks. We also met before milestones were due to discuss in person and assign tasks.

Teresa worked on producing the heat map and the text description for this visualization. The data analysis for this visualization took approximately 2 hours. Designing, creating, and editing this visualization took approximately 4 hours.

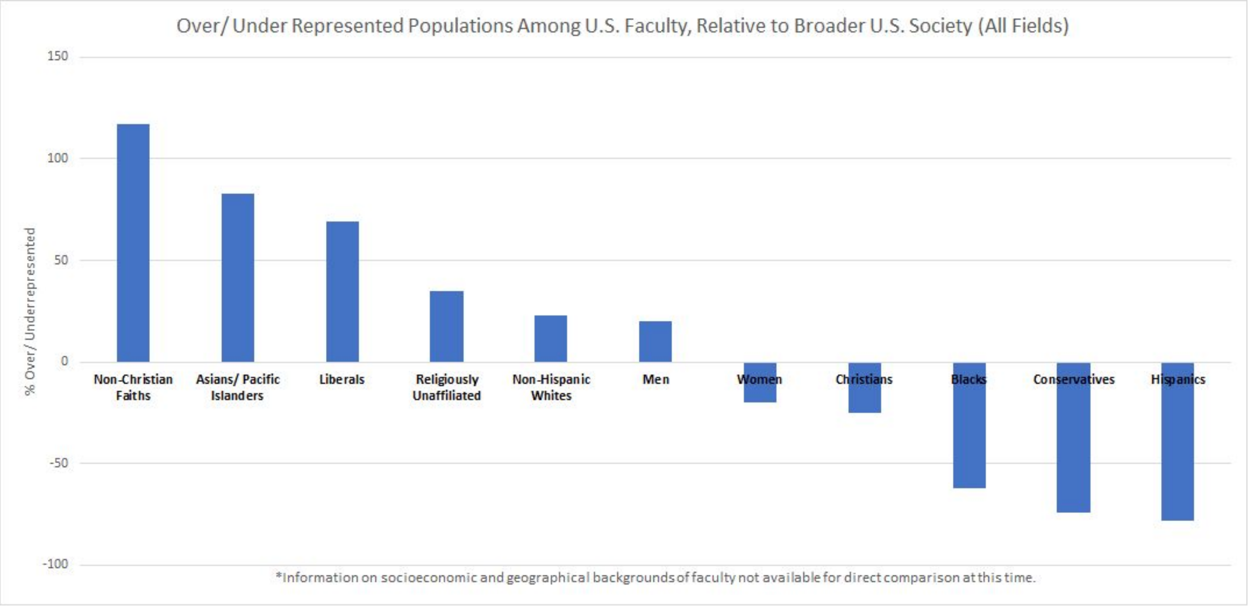
Joyce worked on producing the line graphs and the text description for this visualization. Cleaning, filtering, loading, and calculating the data took around 1.5 hours. In addition, creating the visualization took approximately 4 hours.

Sruthi worked on producing the stacked bar chart and the text description for this visualization. This visualization took around three hours to create. An additional 1.5 hours was required to clean up the visualization and add text descriptions.

The most time consuming parts of our project as a whole was coding the visualizations, and the initial challenge of finding and creating an appropriate dataset.

Appendix

Figure 1.



Source:
https://heterodoxacademy.org/ideological-underrepresentation-compared-to-race-gender-sexuality