HW 2 Write Up

Mitra Farzami (msf248), Lexi Valachovic (anv35), Clayton Seibel (ccs239)

Data Description

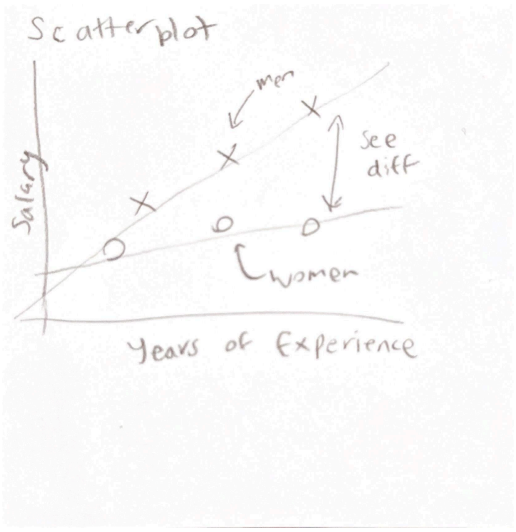
Our team was interested in visualizing the gender wage gap, as it always is a highly talked about social issue. One of the main debates about the wage gap we often hear is what the root cause of the wage gap is – choice of occupation, difference in experience, sexism in the workplace, or other factors. We thought data visualization would be a great tool to explore how different variables affect the average pay difference between genders.

In order to analyze the wage gap in the lens of different factors, we needed to find a dataset that had information on a person's salary as well as their degree, years of experience, age, and job title. We found a Kaggle dataset called Salary Data that collected all of this information from multiple sources including public surveys and job posting sites. It has a total of 5 variables and over 6,000 data points. While we had hoped to find a more reliable source with more entries, such as US census data, we found that this dataset had more of the features that would be relevant to our data visualization ideas.

Interaction Storyboards & Explanation

The gender wage gap is most commonly reported as some sort of number, like the number of cents a woman makes for every dollar a man makes. This simple data point can be a telling one, but without context it can invite many questions about its calculation. Our visualizations aim to explore which factors are associated with inequality in salaries, and how men and women's salaries compare when we adjust for career choice, years of experience, or their age.

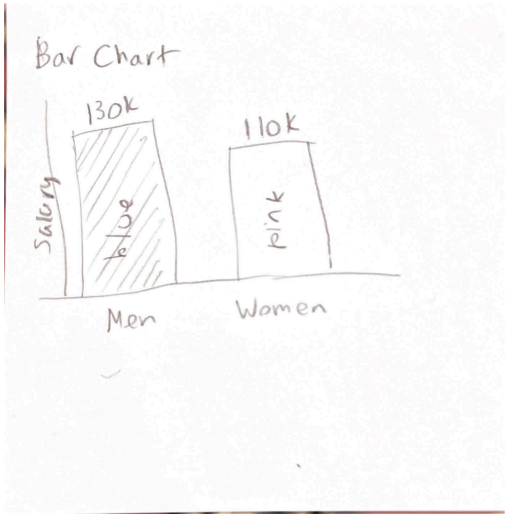
Scatterplots



We knew we wanted to use scatter plots in our data visualization because we wanted to show any patterns that exist when we compare salaries (color coded by men and women*) and our two quantitative variables: age and years of experience. We assumed both would show positive correlations, as older and more experienced people tend to have higher paying jobs, but we were hoping to show whether gender has an effect on this trend.

*We decided to use pink for women and blue for men – which, despite being an outdated cliche, is also a very common cultural gender reference. We learned in INFO3300 how important it can be to match the user's mental model of cultural color cues.

Bar charts

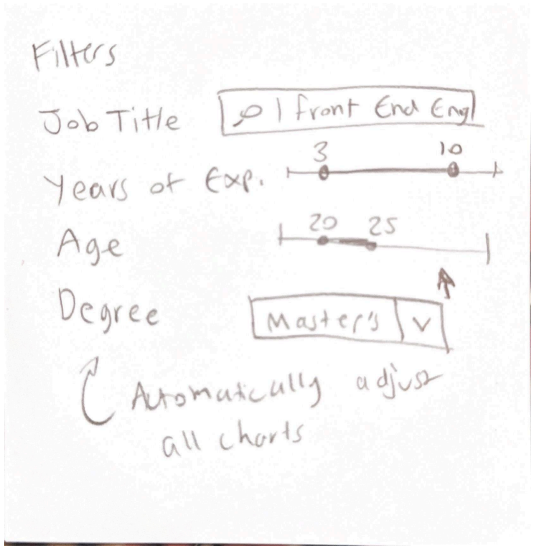


We also wanted to show a very simple but telling graphic: a bar chart with average salary per gender. While a two bar bar chart is a simple choice, we thought the visual channel of length would illustrate the difference in average salary the user may find in

an intuitive and striking manner (whereas other visual channels are less obvious). In the bar chart, we wanted to include a clear numerical value label in instances where the average salaries may be close.

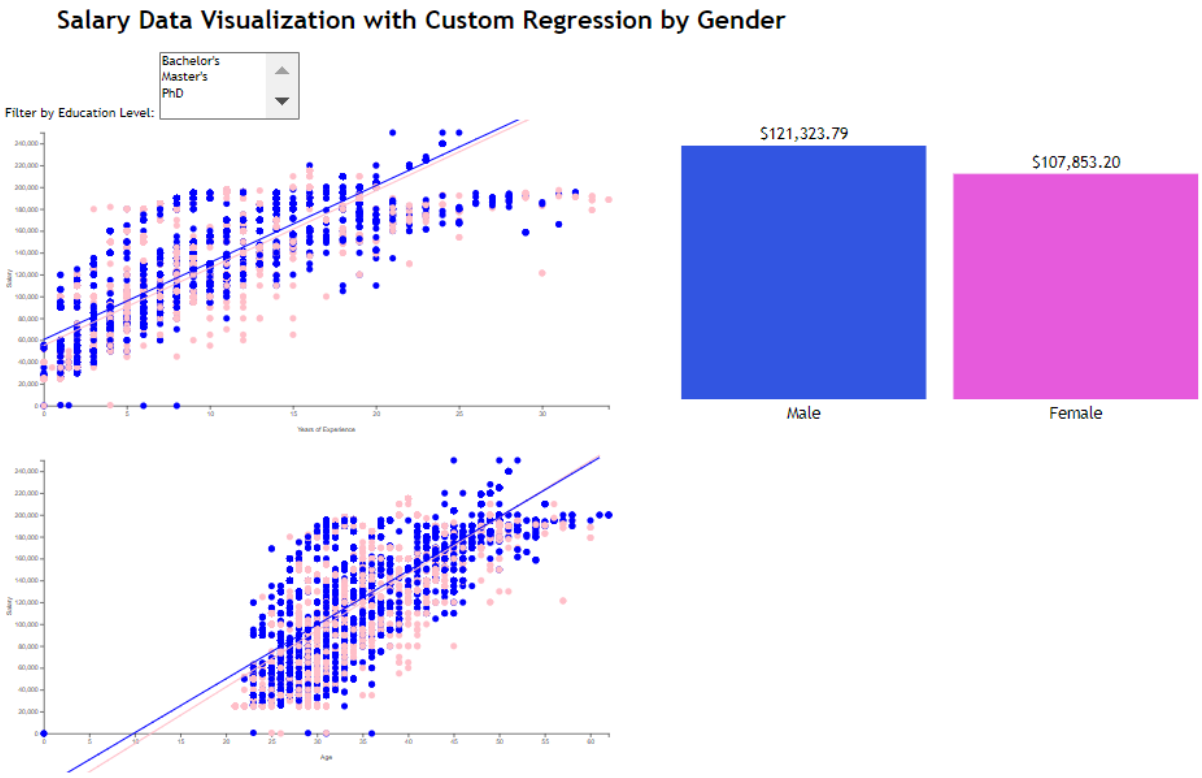
Both the scatter plot and barchart combined make a really effective visualization as the scatterplot shows the distribution of salaries, showing that women can be higher than men in a lot of individual cases, but the bar chart narrows this down and shows the average per gender, which highlights the overall status of where we are at with the wage gap. But including the scatterplot with women represented at higher salaries can give the user motivation and insights that women have the ability to earn more than men.

Filters



The most important part about both of these visualizations is the ability to filter them by the education level. By standardizing the data to be all people of a certain degree, or experience level, users can contemplate where salary inequality stems from. We ourselves weren't sure what results these filters would return, but we were confident that we would see a substantial salary difference between genders. We contemplated adding more filters for variables such as career choice, years of experience, and age, but decided that this many moving parts may overwhelm the user and make it harder to track differences among salaries.

Final Data Visualization



Our visualization shows a scatterplot of salary versus years of experience and right below we have one showing salary versus age. To increase the effectiveness of these regressions, we included a bar chart to the right of average salaries of the filtered group to show the exact change. All charts update with our education level filter.

Issues & Trade Offs

Complexity vs Functionality Increasing the functionality, such as adding more filters for job title, age, etc., can make the visualization more powerful but also more complex for users to understand and interact with. Therefore, we decided to have a filter for education level and our visualization still accounts for years of experience and age within it. Education level is a main concern for many. Most people feel as though they won't be as successful without higher education, so our visualization serves to allow them to see the differences between degree levels.

How strong of the correlation do the regression lines represent R^2 measures the strength of the relationship between explanatory variables of education and the dependent variable of salary. With this information better insights about the strength of explanatory relationships can be made, but we decided to keep these numbers off of our visualization and utilize a bar graph instead. Since we have 3 graphs at once, it is a

lot of information for the user to track, we decided that 4 more numbers floating around would cause less comprehension of the visualization. Our bar graph gives 2 concrete numbers that the user can directly compare, serving as a better measure of strength when comparing salaries.

Development Process

Originally, the ambition was to develop a 'calculator' feature within our interactive data visualization, aimed at predicting whether a given salary data point was a man or woman based on different variables such as years of experience, education level, and job title. This predictive functionality was intended to offer users a deeper, more analytical interaction with the dataset, using machine learning/ statistical modeling techniques to make gender predictions based on the correlations observed in the data.

However, as development progressed, it became clear that integrating a fully functional predictive calculator was much more ambitious than initially anticipated. Challenges included the complexity of designing an accurate prediction model, and trying to integrate such a model into a website. One specific issue we had when trying to integrate a model into our app.py file was that we had to install packages into our environment like scikit learn and even numpy/pandas that we thought may cause issues for graders/users of the published site. Despite stepping back from this initial idea, the underlying concept of providing predictive insights remains our core message we are trying to convey. So, we plan to provide the user with filtering tools and statistical analyses that highlight the salary disparities between genders.

Work Breakdown

Clayton set up the GitHub repo and spent time coding the scatterplot and regression line for our salary versus age and salary versus years of experience charts, both of which were color coded by gender. This also involved coding filtering functionality by degree type. Mitra found the dataset we used on Kaggle and spent time coding the bar chart visualization, which included filtering functionality for age, years of experience, degree, and job title. Lexi spent time on both types of chart, working on combining the filtering functionalities into one streamlined filtering UI and on formatting the final web page itself. We all had equal workloads and ended up helping each other with assigned tasks along the way, which helped us all be in the loop.

Sources:

https://stackoverflow.com/questions/6195335/linear-regression-in-javascript
https://www.kaggle.com/datasets/mohithsairamreddy/salary-data?resource=download