M3 Final Project Submission

Group: 10
Project Title: Visualizing Tech Compensation
Video Demo: https://www.youtube.com/watch?v=wZ_ZY8AlCfU

Members	Student Numbers
Benjamin Chang	14554159
Joy Xiao	19317171
Joey Shi	41274648

Overview

Our visualization will be tackling the problem of providing transparency on salary ranges for jobs in tech. People entering the field will not have a full understanding of how much they should be paid, and may be exploited by their lack of knowledge and accepting a lower pay than what is fair. Our visualization addresses this issue by allowing users to see the salary distribution of workers in similar positions. This will allow workers to achieve a better understanding of their monetary value and know when they are being offered a low or high salary for their experience and job title.

Data and Data Preprocessing

Dataset: "Salaries for Jobs in Data Science and Tech"
Link: https://www.kaggle.com/datasets/jackogozaly/data-science-and-stem-salaries
Description:
The dataset consists of tech salary records in North America from January 1, 2019 to September 9, 2020. There are around 62,000 items total in the dataset. Our visualization will consider the following attributes:

| Attribute	Type	Cardinality/Range
company	Categorical	1631
title	Categorical	15
basesalary	Quantitative	[0, 1659870]
yearsofexperience	Quantitative	[0, 69]
yearsatcompany	Quantitative	[0, 69]
location	Categorical	1050

The data was pre-processed in the following ways in a Jupyter notebook:
- The non-American data was filtered out.
- A new column was created representing the American state.
- To remove outliers, we filtered out the top and bottom 5% of the data by salary, years of experience, and years of tenure.
  - We do this because our target audience are new-grads/people just entering the industry.
- We dropped columns that were not relevant to our task abstractions and data visualization.
- We filter for records with one of the top 10 most frequent job titles to avoid clutter in the amount of job roles in our legend

Tasks

1. Identify distribution of salaries for workers with a particular range of years of experience, time spent at company, and tech role.
2. Compare distributions of salary between 2 different tech roles.
3. Locate highest value/average salary across all companies for workers with a particular range of years of experience, time spent at company, and tech role.
4. Identify geographic distribution of average salaries across different states.
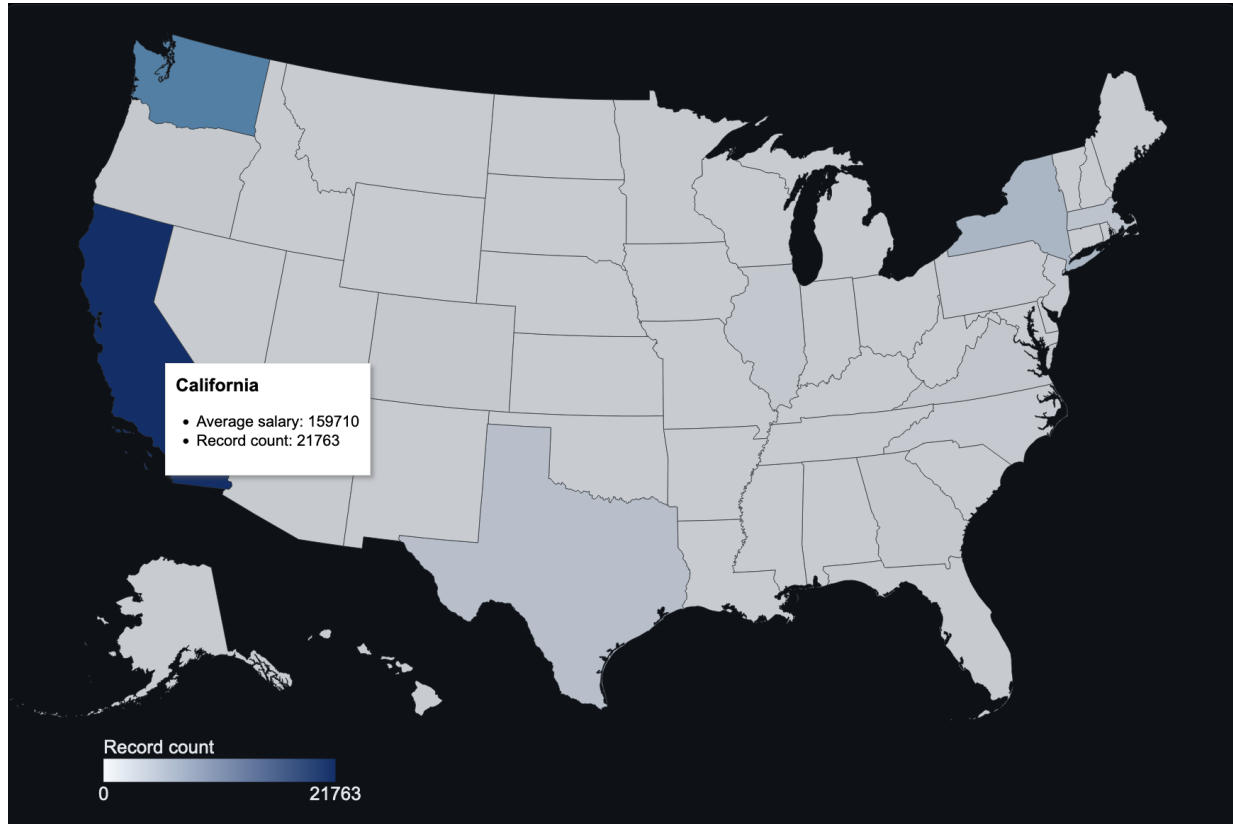
Visualizations
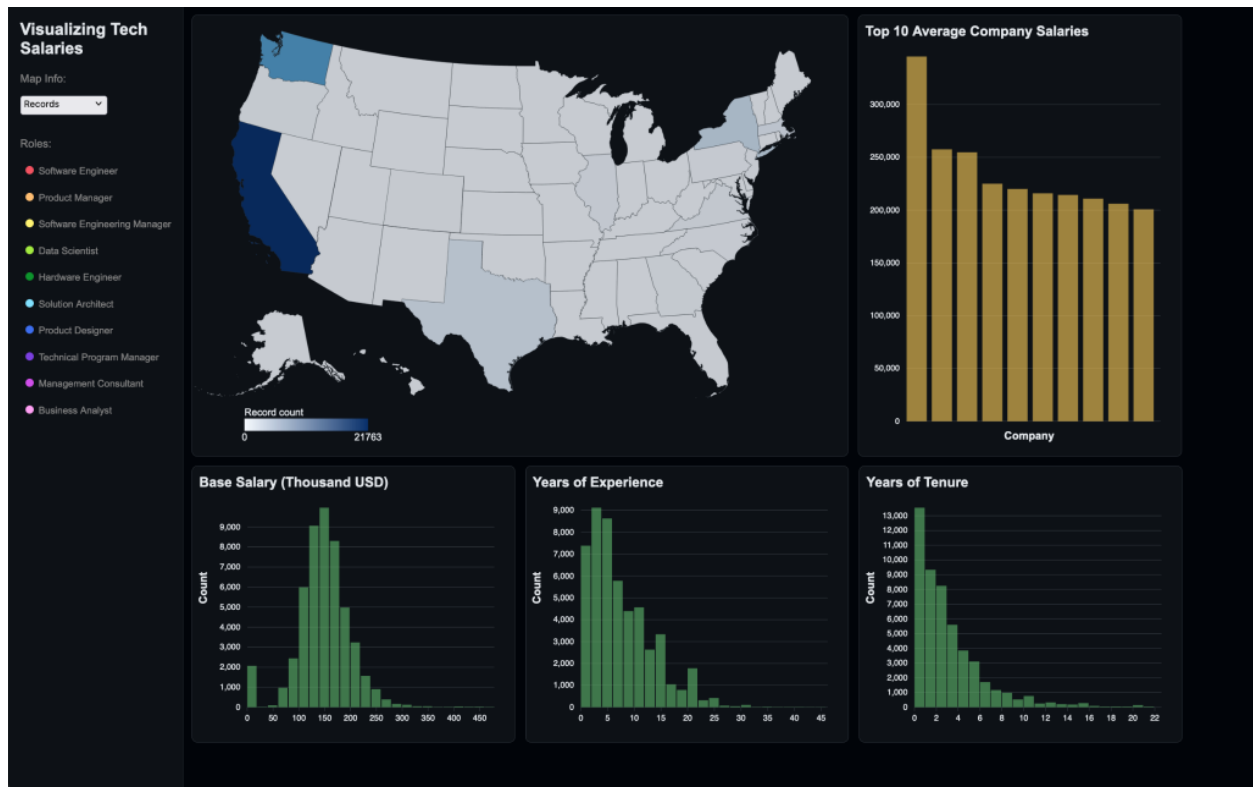
Choropleth map
- Visual encodings
  - Marks: Interconnected areas
  - Channels: Color luminance (number of data points)
- Interactions
  - Hovering over a state will show a tool tip containing
    - State name
    - Average salary value for data points in the state
    - Number of data points for that state
    - Only states with >0 data points are selectable
  - Changing the dropdown selection in "Map Info" will change the data that's being encoded in the map.
  - Clicking on a state will filter the dataset by that state, changing all of the other views to reflect the data records of that state. The borders of the state will also be highlighted when selected.
- Rationale
  - Using an interconnected areas of the US map allows the user to view salary ranges across the United States. Color luminance accentuates where on the map that average salary may be the highest. This visualization ties in with Task 4,

where the geographic shapes and color luminance gives a clear look at the distribution of average salary across the US. Interactions with the job legend on the left hand side can give the user a view of the geographic distribution based on certain job titles, and interactions with bar charts can narrow down where top companies are located the most in the US. Interactions with the histograms as well can show the distribution of average salary, job experience, and company tenure for specific states in the US.
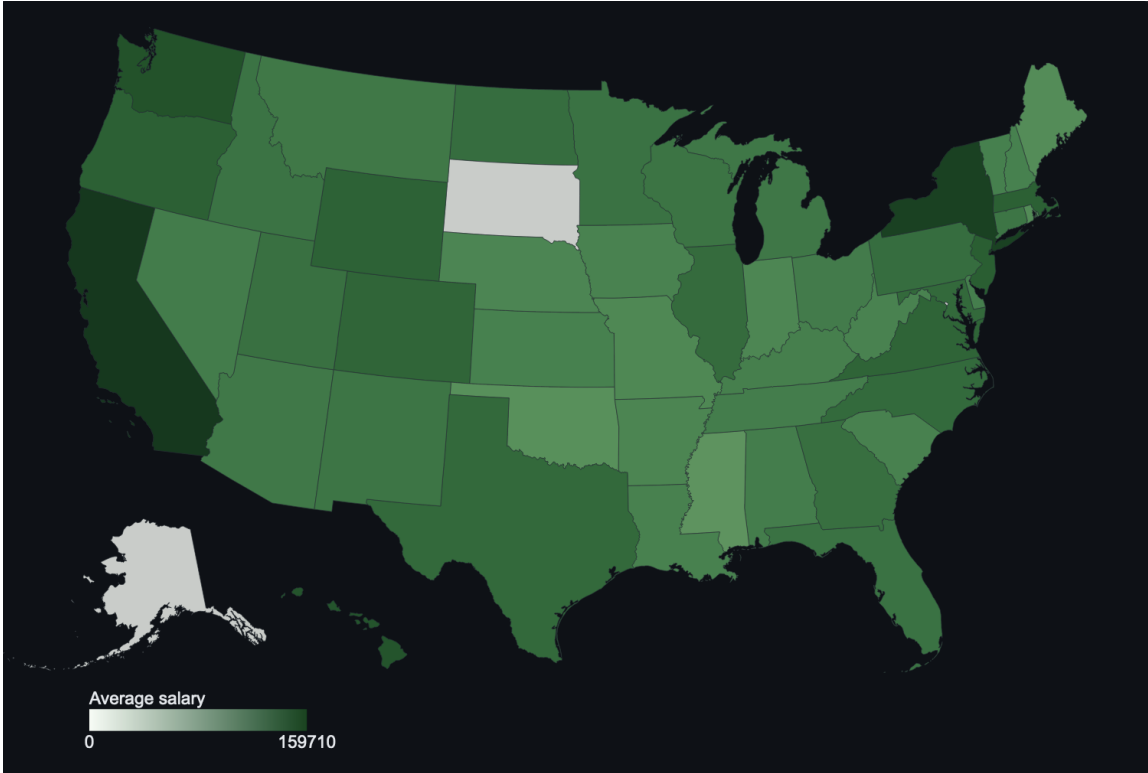
Hovering over a state shows a tooltip

Clicking on a state highlights the state

California
- Average salary: 159710
- Record count: 21763

Record count
0   21763
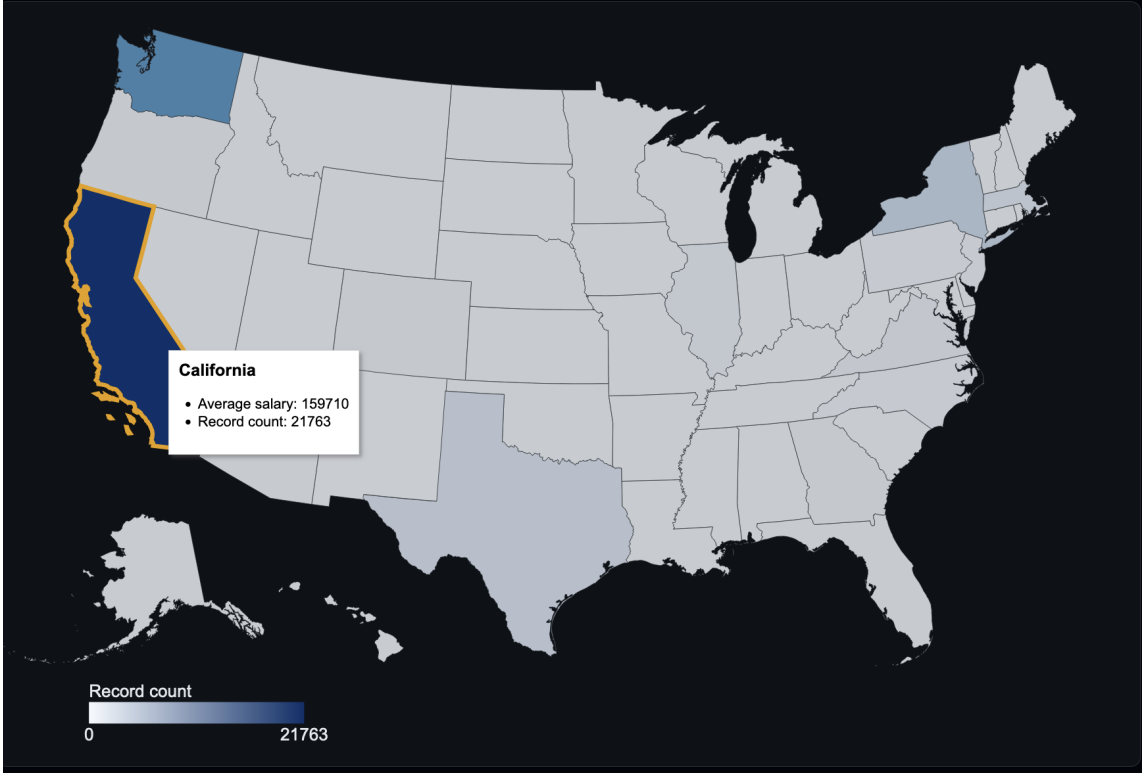
Selecting average salary in the map info dropdown

Average salary
0   159710

Sorted company bar chart
- Visual encodings
    - Marks: Line
    - Channels:
        - Horizontal Position (company)
        - Vertical Position (number of data points)
- Interactions
    - Hovering over a bar will show a tool tip containing
        - Company name
        - Average salary for the company
        - Number of data points for the company
    - Clicking on a bar will filter the other views by the company selected
        - Can filter on top of other filters
        - Can select multiple companies in the barchart
        - However, this selection will be cleared once another filter is selected, since the new filtered data may not contain the selected companies anymore
- Rationale
    - The horizontal position of a bar chart is able to visualize the differences between each of the top 10 companies. Having it sorted from highest to lowest also immediately determines which company is the top company for average salary and can compare bar heights of subsequent companies on how much lower their average salary is. This sorted bar chart accommodates Task 3, being able to see the highest average salaries across different tech companies, and interactions with the histograms and the job title legend will allow the user to view the top companies specific to the filtered selection of various job titles, salary ranges, years of experience and company tenure by the user.

Selecting the "Netflix" company bar

Top 10 Average Company Salaries

Netflix
- Average salary: 343500
- Number of records: 134
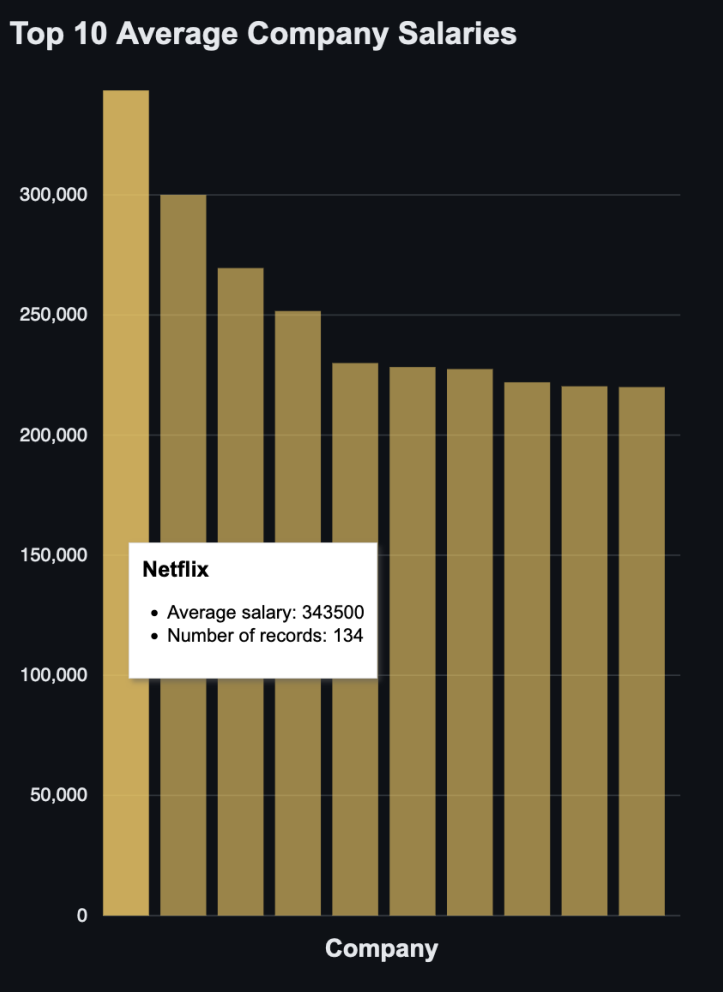
Company

Salary, company tenure, years of experience histograms
- Visual encodings
    - Marks: Line
    - Channels:
        - Horizontal Position (encodes salary, company tenure, years of experience in respective histograms)
        - Vertical Position (encodes number of data points)
        - Color hue (encodes the job role(s))
- Interactions
    - The brush filter can filter for data points of a specified range of salary, company tenure, years of experience when used in the respective histograms.
        - This filter is applied to all of the other views as well, so that the information of that specific selected range can be displayed on the choropleth map and the top companies for those ranges.
    - Job roles legend
        - By default every view will visualize the entire dataset
        - If 1 role is selected, every view will filter for datapoints for the selected role
        - If 2 roles are selected, every view will filter for datapoints for the 2 selected roles and the histogram views will have overlapping bars.
- Rationale
    - The horizontal position of the histograms for each view will show the differences of ranges in average salary, years of experience, and company tenure, allowing the user to easily identify the distribution of these attributes, tying into Task 1. By using the job title legend in the sidebar, the user can filter for specific jobs and view the differences of the distribution of salary of the selected jobs. When jobs are selected, the histograms will have overlapping bars, corresponding to the color encoding of the job titles. It is easy to then compare the heights of each pair of overlapping bars, which demonstrates the differences in the distributions, linked to Task 2
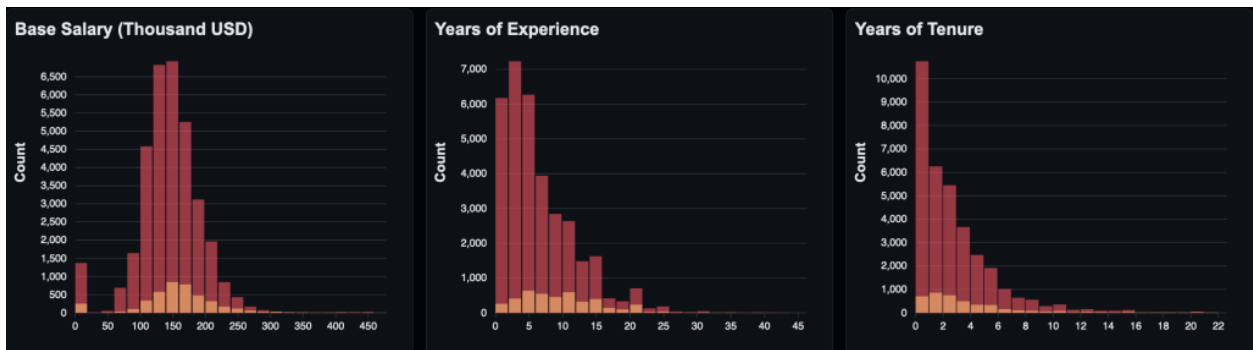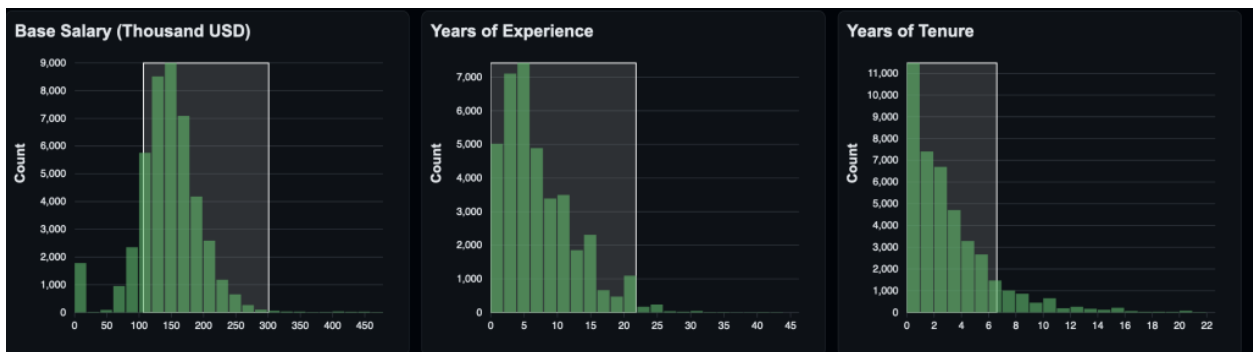
Initial views
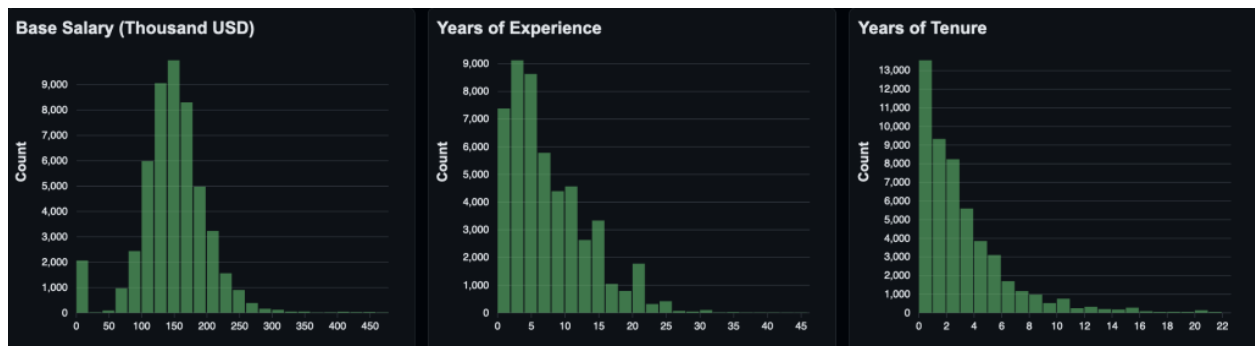
Selecting "Software Engineer" from the roles legend
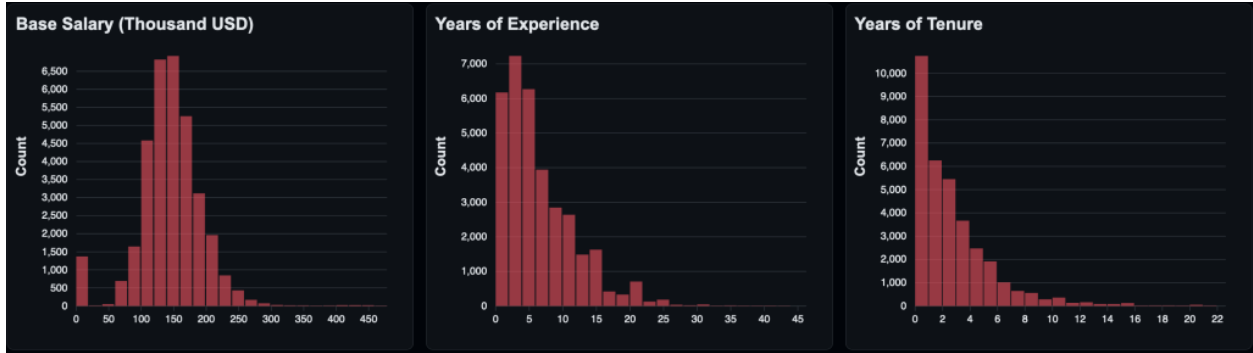
Selecting "Software Engineer" and "Product Manager"

## Using the brush filter on each histogram

Usage Scenarios

Aspiring tech workers can use this visualization to identify which company would provide them the best compensation. The visualization could also help young workers understand an expected range of salaries and prevent them from lowballing their salaries.
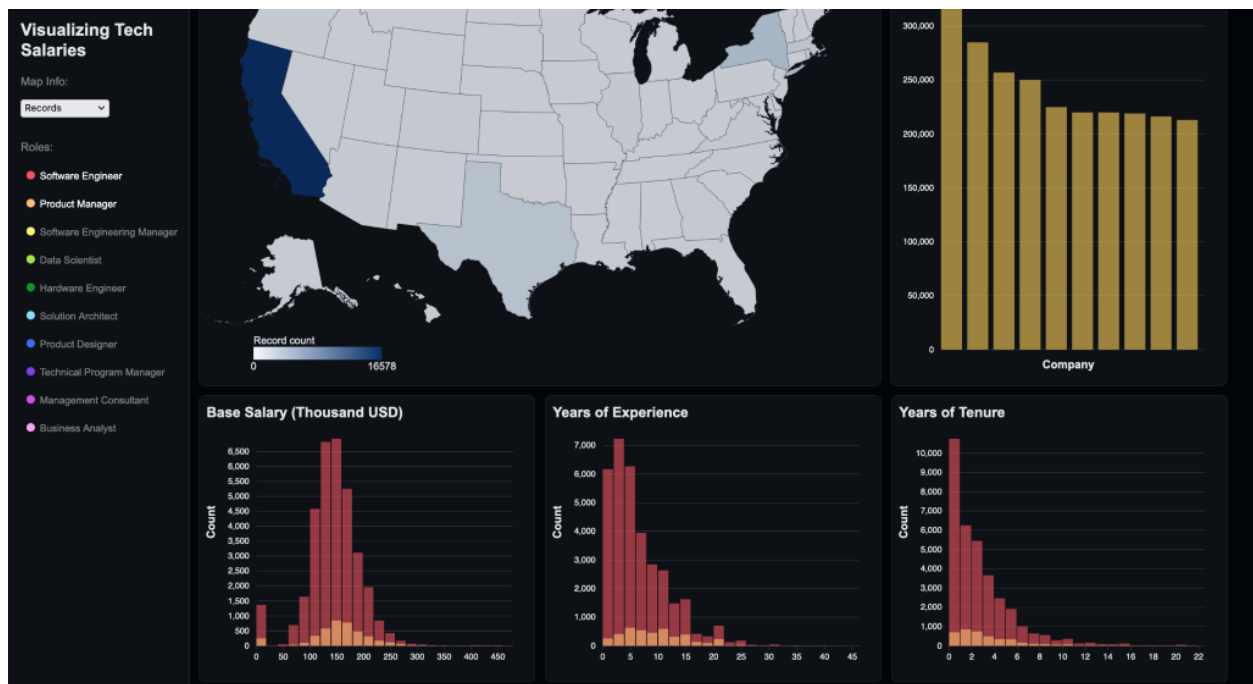
Example Scenario 1

John is in his final year of his computer science degree, a soon-to-be new grad. He is in the job pipeline with multiple companies and is expecting offers from several of them. He would like to know the expected salary ranges for someone similar to him so that he may gain some negotiating leverage. This can be done by filtering the data by particular job positions by clicking on a job role in the left panel and filtering by years of experience and company tenure by using the brush filter on their respective histograms. Then, the salary histogram will give the user a visualization of the distribution of salaries for workers similar to themself.
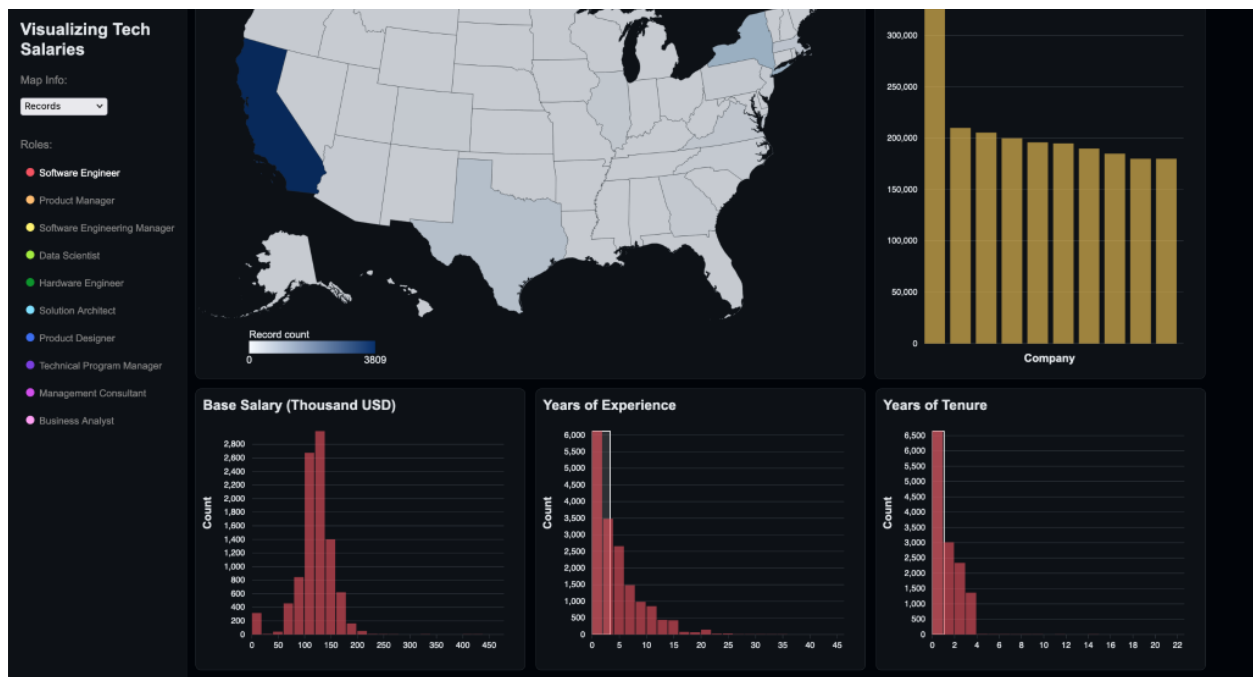
Example Scenario 2

Jane is a freshman and is interested in pursuing a career in tech. The problem is however that she doesn't know what kind of educational stream or discipline she would like to go down. Many of her friends and classmates are intent on pursuing software engineering, but perhaps she would like to use her degree to become a product manager, or go down a more math and statistics heavy degree in order to become a data scientist. By clicking on "Software Engineer" and "Product Manager" titles in the list of roles on the left panel, she will be able to see the salary distribution histogram of software engineers overlapped with the salary distribution histogram of product managers. This can help her get an understanding of what she can expect in terms of salary for particular job roles upon graduation and perhaps down the road in career development.

Example Scenario 3

Johnny is an aspiring tech worker. They are planning on moving to a different state and are curious to know how the average salary in that state compares to their home state (e.g. New York vs California). Our choropleth map is interactive in that upon hover, it will display an average salary for the state. In addition it will also allow users to select specific states and filter the data in the other views according to the selected location.
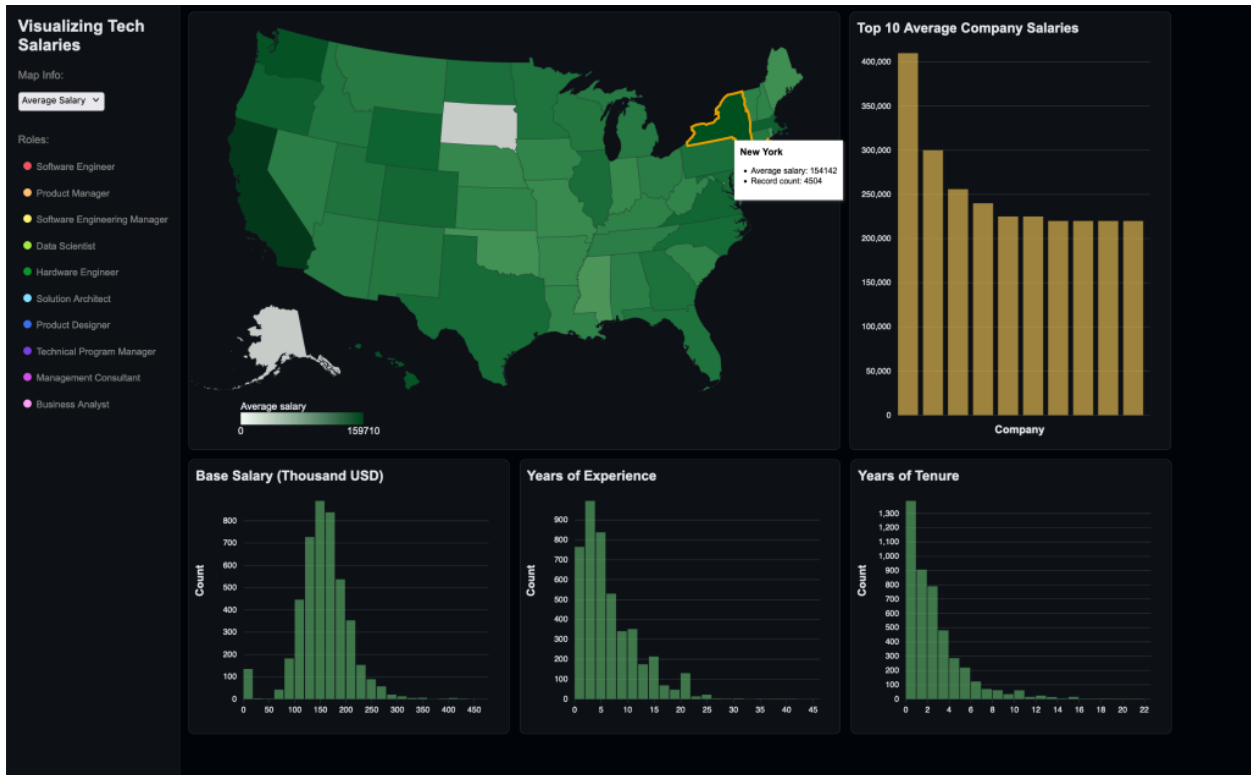
Reflection

Since our initial proposal, we have decided to visualize tech salaries rather than compensation. This is because total compensation includes additional benefits, such as bonuses and stock options, which does not reflect the actual take-home pay of an individual. Thus, we choose to visualize the base salaries of workers instead so that it is easier to identify where pay disparities exist. The design of our visualization has also changed. We now focus more on comparing salaries between different roles and including views that help answer the question of whether a particular worker is paid similarly to others in his position (e.g., example scenario 1).

A major change from our initial proposal in M1 was changing the scatterplot to a choropleth map in order to visualize distribution of salaries based on geographical location in the US. We also increased the number of views to have small multiples of histograms to more easily analyze distributions of salary, years of experience and company tenure. In addition, we added a barchart view to look at the top companies based on annual salary so that the user has information about what companies may have the best salary or company retention. While our initial proposal was realistic and technically possible, we actually realized we wanted to push the technical details even more than what we at first designed. These changes we have made since M1 and followed through in M2 allowed us to further visualize the information that is given by the dataset we've chosen and use D3 to a fuller potential.

In M3, we implemented the interaction goals we set in M2 and were able to accomplish most of the interactions and filtering between views that we set out to do. During our work for M3, we realized that some interactions were not fully feasible or practical, such as continuously layering filters based on company since certain selections may not have the same top average salaried companies. For this company filter for example, we decided to reset the company filter whenever another filter was applied so that users would end up filtering by a company that was no longer visible. Our technical approaches in coding the visualization have remained consistent.

If we were to make the project again from scratch, we would like to spend more time determining the behaviour of our visualization on various edge cases to have a better understanding of the user experience before writing any code.

Project Management & Team Assessment

Notes:
- *After accounting the changes to our visualization made after our initial proposal, some previous tasks were removed and tasks were reassigned.*
- *If there are more than 1 person assigned, the time estimates are per person.*

Task	Time Estimate (hours)	Actual Time (hours)	Target Date	Actual Date	Assignment
EDA & Data Wrangling	3	3	Mar 20	Mar 29	Ben, Joey
Add docs + detailed specs for static views	3	3	Mar 20	Mar 27	Joey
Elaboration on code organization	2	2	Mar 21	Mar 21	Joey, Ben, Joy
Convert project to typescript and bootstrap with webpack	8	4	Mar 24	Mar 23	Joey
Implement static histogram	12	4	Mar 26	Mar 29	~~Joy~~ Ben
Implement static bar chart and histograms for experience and tenure	8	6	Mar 26	Mar 30	~~Ben~~ Joy
Implement static choropleth map	8	4	Mar 26	Mar 29	Joey
Refine views, CSS styling according to specs	12	2	Mar 28	Mar 30	~~Joey, Joy, Ben~~ Joy

Task					
M2 writeup	8	8	Mar 31	Mar 31	Joey, Ben
Implement interactive filter for histogram (salary range)	10	4	Apr 2	April 8	~~Ben~~ Joey
Implement interactive filter for barchart (company)	10	6	Apr 2	Apr 8	Joy
Implement interactive filter for choropleth map (state)	10	5	Apr 2	April 8	Joey
Link barchart filter -> histograms	4	4	~~Apr 6~~ Apr 8	April 11	Ben, Joy
Link histogram filters -> barchart	4	4	Apr 8	April 11	Ben, Joy
Link barchart filter -> choropleth map	4	5	Apr 8	Apr 10	Joy
Link histogram filters -> choropleth map	4	4	Apr 8	April 11	Ben
Link choropleth map filters -> bar chart	4	4	Apr 8	April 11	Joy
Link choropleth map filters -> histograms	4	4	Apr 8	April 11	Ben, Joey
Add choropleth legend	4	6	Apr 12	April 12	Joey
Final touches/refactoring Presentation Prep	8	10	Apr 12	April 12	Ben, Joey, Joy
M3 Writeup	8	4	Apr 13	April 13	Ben, Joey, Joy

Totals time estimates:
- Ben: 57
- Joey: 72
- Joy: 64
- Total worktime: 193

Total actual hours:
- Ben: 47
- Joey: 57
- Joy: 47
- Total actual worktime: 161

Credits

- Our choropleth map was inspired by "Visualizing Gun Violence in America" from the hall of fame. Major changes were made to strip out all unused code.
    - https://www.students.cs.ubc.ca/~cs-436v/22Jan/fame/projects/project_g06/index.html
- We built upon the histogram example in the d3 gallery. Major additions were made to refactor the code into a class following the init, update, render pattern and make it reusable for all our 3 histograms.
    - https://d3-graph-gallery.com/graph/histogram_basic.html