

INFO 5100

PROJECT 3 - INTERACTIVE DATA VISUALIZATION

FINAL REPORT

Team Members

Daisy Liu (hl2445), Eva Kang (yk622), Han Gao (hg359), Alice Chang (cc2353)

Due: Monday, December 13, 2021

PART 1: PROJECT IDEA

1. Motivation & Description

[Big Data](#) is a field that treats ways to handle, analyze, and extract important information from large, complex, hard-to-manage volumes of data that inundate business on a day-to-day basis. With the fast development of technologies, the [importance](#) of big data has been recognized and valued recent years. By analyzing the data, businesses can improve operational efficiency, optimize product development, explore new revenue and growth opportunities, and streamline resource management.

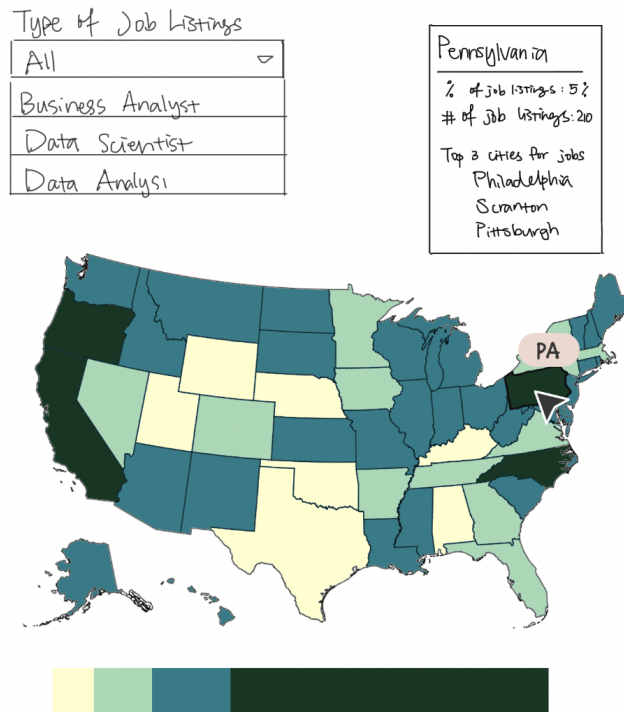
As big data becomes a focus in the business world, more and more data-related job opportunities available. **Business Analyst**, **Data Analyst**, and **Data Scientist**, as three hot positions under the big data path, have significant amount of overlaps in their job responsibilities and required skills. For example, they all work with massive data to solve business problems. Therefore, when it comes to determine a career path, people usually hesitate among these three positions. We observed lots of our friends, who want to pursue a big data path, encountering selection difficulties. They usually wonder *what **differences** are among these three positions, what **industries** they can work in, which position provides more **job opportunities** or **higher salaries**, and which one is more **suitable** for them.*

Therefore, we created this project to help people, who have the same struggles or questions as our friends, gain an **overview of job opportunities** in the big data field. Hopefully, we can help them find some answers to their questions and find their dream positions.

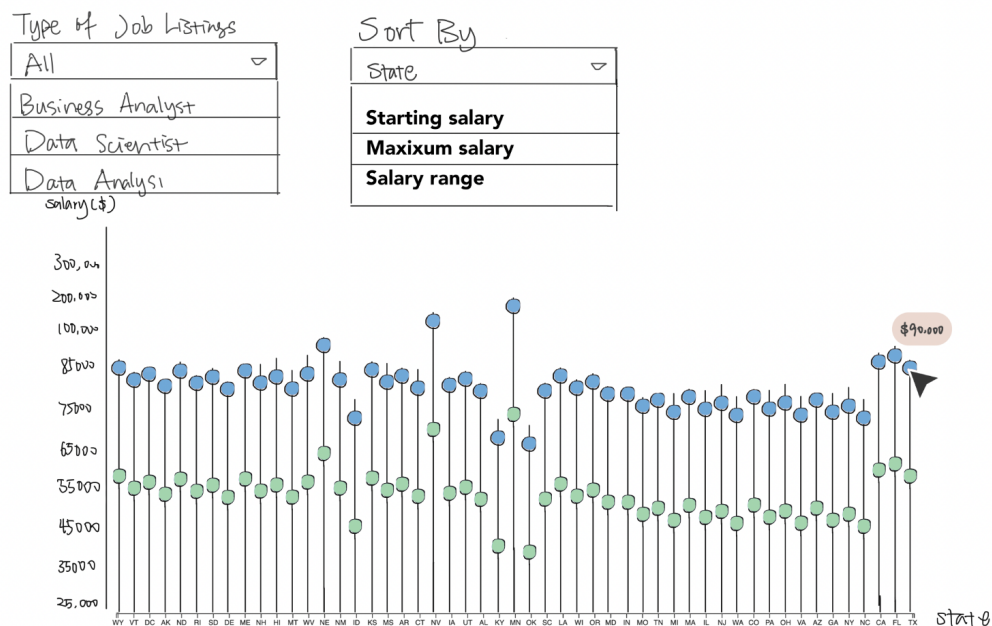
2. Design Brainstorm & Sketch

In our brainstorming session, we considered *what **factors** would affect our **job choices***. We came up with several factors that include self-interest, wage, job location, and company. Therefore, we decided to show **geographic location**, **salary**, and **company-related information** of all job listings in the Business Analyst, Data Analyst, and Data Scientist job positions.

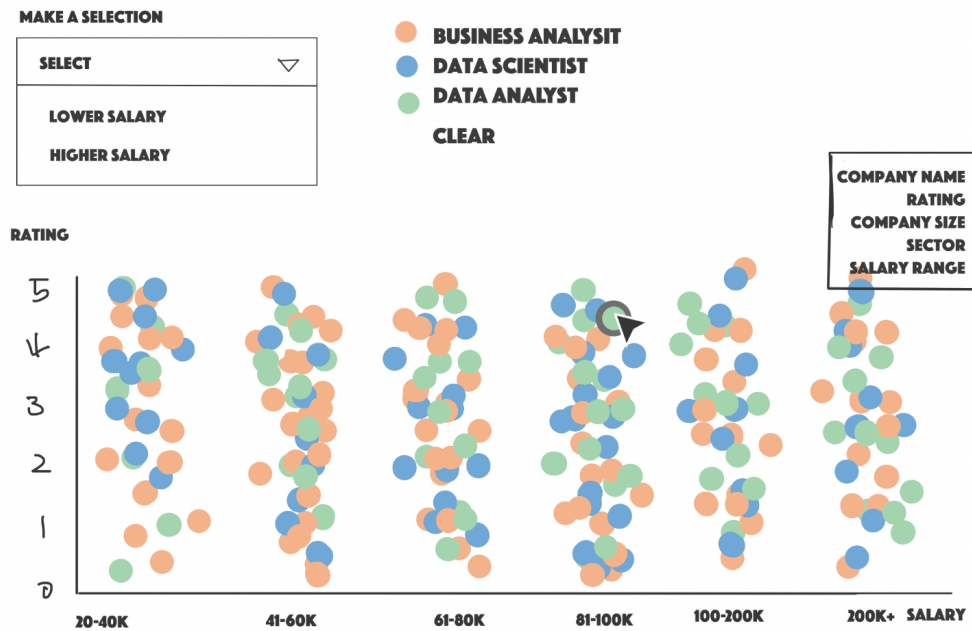
Based on the focus of our project, we created our design sketches which include three visualizations. The first visualization shows the **geographic distribution** of data-related jobs on the U.S. **map**.



The second visualization compares **average salaries** of those jobs among different U.S. states in a **lollipop graph**.



The third visualization shows the **salary distribution** of jobs in giant companies by their **company ratings** in a **scatterplot**.



PART 2: DATA DESCRIPTION

1. Data Source

In this project, we used **five datasets** in total to create our visualization. The primary three datasets are the **job listing data** for three data-related job positions - [business analyst](#), [data analyst](#), and [data scientist](#). These three datasets were originally scrapped from glassdoor and only included job positions in the United States with information such as position title, salary estimation, location, company rating, company size, and more.

The additional two datasets we used are the *TopoJson data file* for the U.S. map and the *U.S. states' names and FIPS codes dataset* from the [INFO5100 FA21 course Github](#). We used those two additional data to create the map visualization.

2. Data Processing

STEP 1: DATA MERGIN

Since the three job listing datasets ([Business Analyst](#), [Data Analyst](#), and [Data Scientist](#)) have the same columns and data structures, we merged the three datasets into one dataset and created a column “position” to indicate the job position with 1 as business analyst, 2 as data analyst, and 3 as data scientist.

STEP 2: DATA CLEANING

For the data cleaning, we used Python to finish the process. First, we **deleted** columns that are not useful for our visualization, including “**Job Description**” and “**Headquarter**”. Second, we did **Exploratory Data Analysis (EDA)** to check each column’s validity and evaluate whether the columns were useful for our data visualization. For example, we deleted the column “**Industry**” because it contained nominal data with too many categories, which was not ideal for visualization. We also deleted columns “**Revenue**”, “**Competitor**”, and “**Easy apply**” since these three columns have more than 1/3 null or unknown data.

After confirming the columns to keep, we filtered out all rows containing any **null** or **unknown** values to avoid confusion. We also filtered out positions that are paid by **the hourly rate**, so that we can compare the salary among jobs on the same scale.

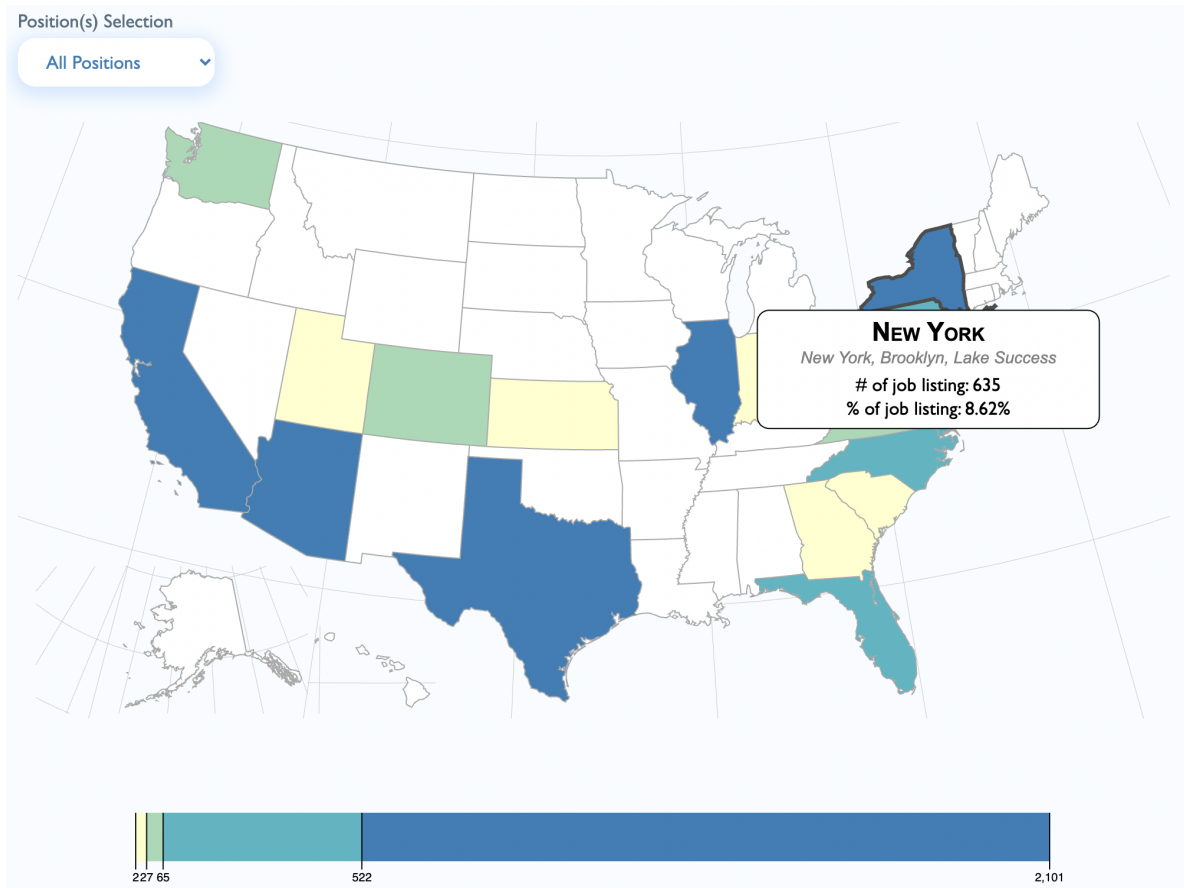
STEP 3: DATA REFORMATTING

The last part of the data processing is reformatting. We reformatted the “**Salary Estimate**” data from string to number and created two columns to record the **lowest** and **highest** number of the salary range. Moreover, we reformatted the **job location** by separating the city and the state of the location into two columns. Finally, we removed the rating at the end of the **company’s name**, for example, from “Asembia\n3.6” to “Asembia”.

After completing the data processing, we exported the cleaned data into a single dataset, named “*final_dataset.csv*”.

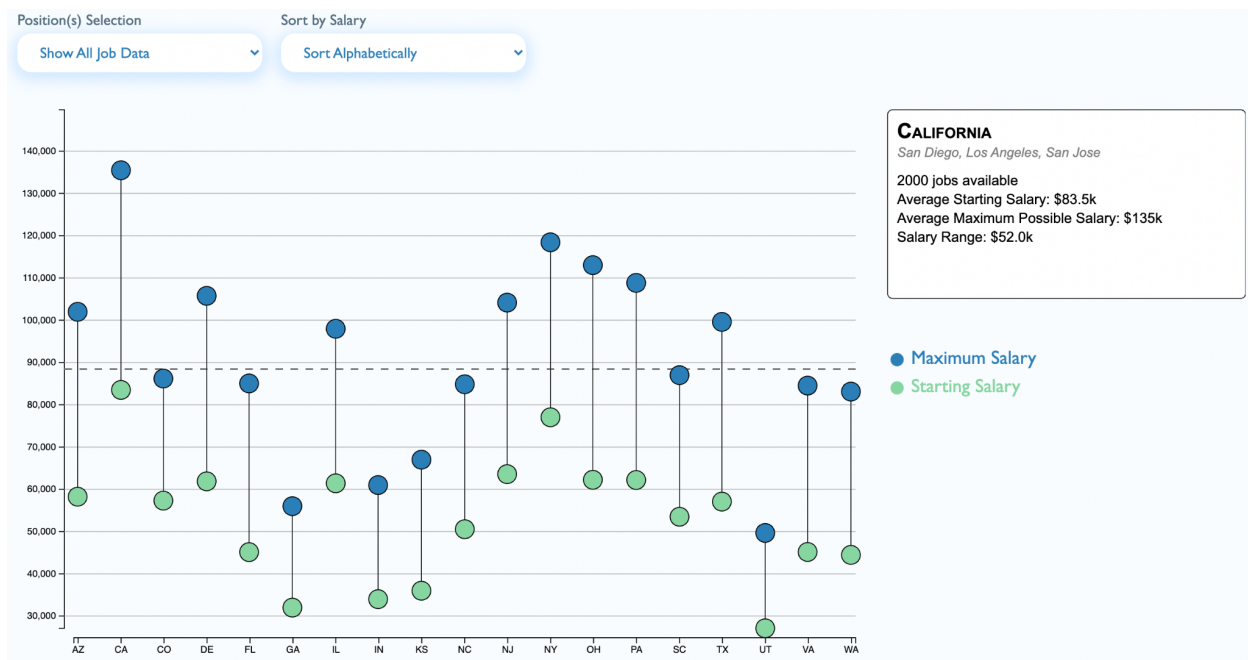
PART 3: VISUAL DESIGN RATIONALE

Graph 1: Geographic Distribution of Big Data Jobs in the U.S.



We plotted the geographical distribution of job listings in the first graph using a map. The visual **marks** are the **boundary lines** and **graticules** on the map of the U.S. The **channel** is the **color hue** of the states, which represents the amount of job postings, or the density of data points, available in a particular state. Considering, our data points are not evenly distributed, we used *d3.scaleQuantile()* function to build our color scale. We built a color palette, ["#ffffcc", "#a1dab4", "#41b6c4", "#2c7fb8"], for an intuitive reading, where darker hues signify higher density of data-related jobs in that area.

Graph 2: Average Salary Distribution of Big Data Jobs by State



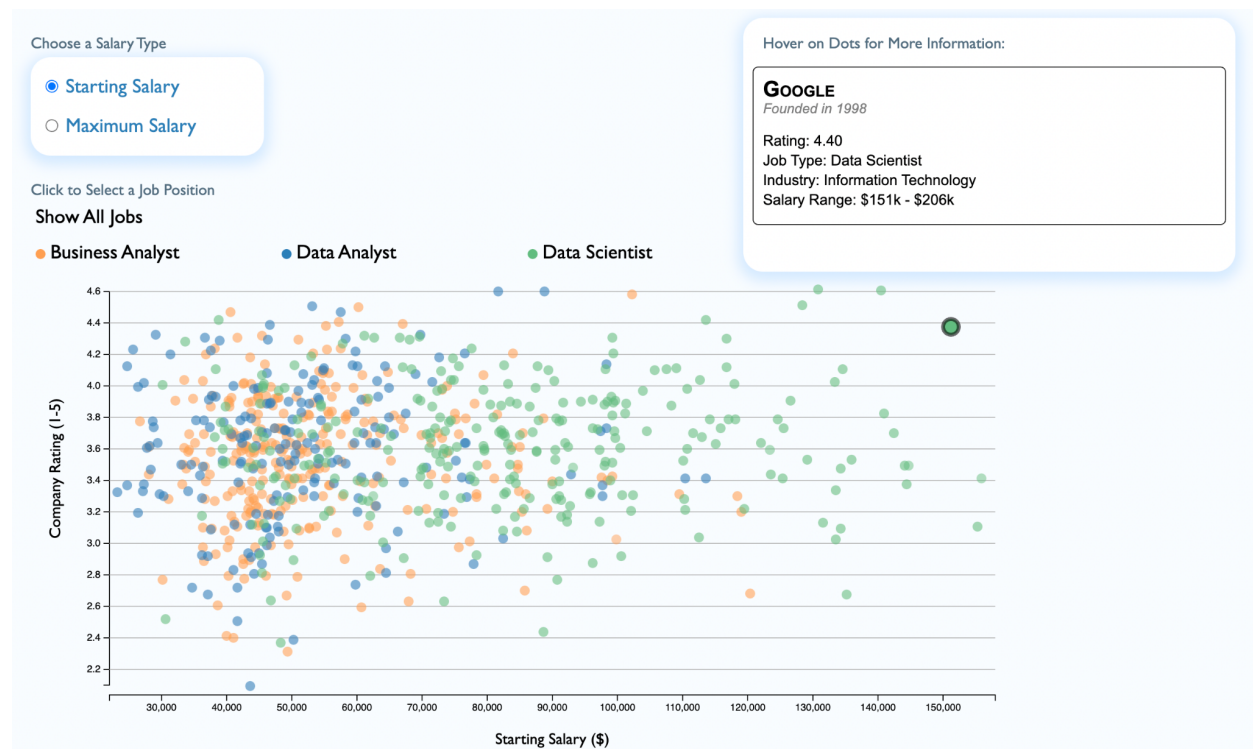
In the second part of the visualization, we plotted average salary ranges of data-related jobs in different states in the U.S. We used a lollipop chart to allow comparison of lower salary bounds, higher salary bounds, and average salary ranges across states. The **x-axis** of the chart shows the **states** where data-related jobs are available, and the **y-axis** plots the **average salaries** in dollars for these states.

The visual **marks** of the chart are the **circles** and **lines** that connect the circles, which make up the lollipops. Each circle represents the average salary of a state, and each line connecting two circles represents the salary range within a state. We used **vertical** and **horizontal aligned positions** as **channels**, allowing comparison of average salaries across different states. **Length of the connecting line** is also a visual channel, which allows comparison of salary ranges across states. Another channel in this chart is **color hue** with blue representing the higher bounds of average salaries and green representing the lower bounds of average salaries.

We included a dotted horizontal line across this plot, representing the national average salary for the three positions we visualize in the chart. This allows the user to compare salaries of different states in relation to the national average. To achieve an internal consistency, the color usage in this plot is the same as the first plot: [“`rgb(135, 214, 161)`”, “`rgb(44, 127, 184)`”].

Graph 3: Giant Company Rating vs Job Salary in Big Data Field

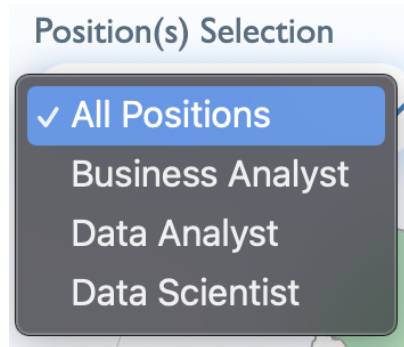
Considering lots of people prefer to start their career paths in famous, giant companies, we created the third plot focusing on companies with **more than 10,000 employees**. In the third part of the visualization, we plotted all data-related jobs at those giant companies in a scatterplot. The **x-axis** of the scatterplot plots the **average salaries** in dollars, while the **y-axis** plots the **ratings of companies** on a scale of 1 to 5.



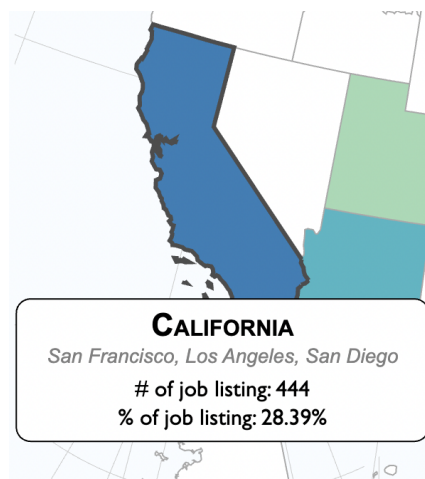
The visual **marks** in this plot include the colored **circles**, each representing a job listing. To increase visibility of the marks, we **reduced overlapping** of the circles using a **vertical** and **horizontal jitter**, and lowered the **opacity** to **0.5**. The visual **channels** include the **horizontal** and **vertical aligned positions** of the circles. Horizontal position allows the comparison of job salaries, whereas vertical position allows the comparison of ratings of companies at which jobs are offered. Because we decrease the opacity of the circles, **density** is another visual channel in this plot, with higher density signifying more jobs within a specific company rating and salary combination. An additional visual channel is the **color hue**, which maps to the three different data-related job types, orange [`rgb(255, 160, 82)`] to business analysts, blue [`rgb(44, 127, 184)`] to data analysts, and green [`rgb(98, 189, 128)`] to data scientists. The use of color hue in the scatterplot allows comparison of salaries across different job types.

PART 4: INTERACTIVE DESIGN RATIONALE

Graph 1: Geographic Distribution of Big Data Jobs in the U.S.



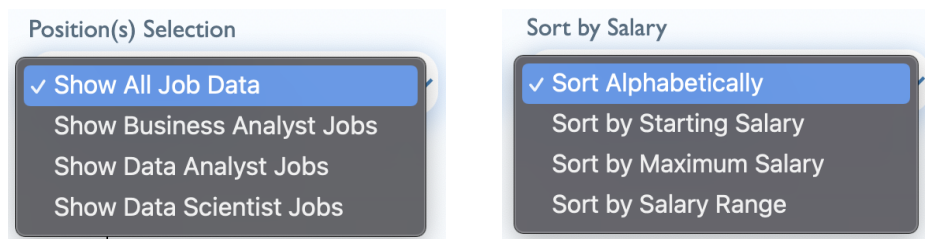
Interaction affordances we included in the first part of the visualization include **filtering** with a **dropdown menu** and **hovering**. By default, the map shows all data-related jobs across different states in the U.S.; however, a user may be interested in a specific job type, for example, one may be looking for data scientist positions and where in the U.S. they can find most data scientist opportunities. Thus, we added a filter that allows the user to choose to only view jobs of a specific type, instead of all data-related positions. We used a dropdown menu to do this, with the current selection visible in the box. The use of a dropdown conserves screen space, allowing the user's attention to focus on the map itself, and puts **constraints on input**, such that the user can only choose to view all jobs or one of the three valid job types available in our dataset.



The map visualization also affords hovering, which allows the user to view more details of a specific state they are interested in. We made this discoverable by **thickening the outline** of a state that is hovered over, signifying that it has been selected, and adding a textbox containing relevant information in the top center of the map. To facilitate reading, the textbox is positioned

near the hovered state. The textbox displays the **name of the state** selected, the **number of jobs** of the currently selected job type (from the dropdown menu) in that state, the **percentage of jobs** in that state out of all jobs in the U.S., and the **top 3 cities** where most jobs are found in that state. We incorporated this hovering feature so that the user can more closely inspect job opportunities in each state of interest.

Graph 2: Average Salary Distribution of Big Data Jobs by State



Similarly to the first part, the lollipop chart includes interactive elements of **filtering** and **sorting** with **dropdown menus** and **hovering**. Keeping consistent with the first plot, we also incorporated a filter on job types with a dropdown menu, with the default display of all jobs, so users can select to view only the job types they are interested in. A user may be interested in seeing the lollipops of average salaries in different order, for instance, one might care more about starting salaries, or the average minimum amount of wages they can get, wanting to see such data in the order from highest to lowest average starting salary. Therefore, we added a sorting feature using a dropdown menu next to the job type dropdown. This allows the user to **sort the data** points based on **what is important to them**: starting salaries, maximum salaries, or salary ranges. All of these are ordered from **highest to lowest**, since a user would be more interested in higher salaries than lower.

OHIO
Columbus, Dublin, Westerville
154 jobs available
Average Starting Salary: \$62.2k
Average Maximum Possible Salary: \$113k
Salary Range: \$50.8k

Also similarly to the first plot, this chart affords hovering over lollipops. Users, curious about a specific state, can hover over a lollipop to view more details about jobs and average salaries in a particular state. When they hover over a lollipop, a textbox appears to the right of the chart, displaying relevant information such as the **name of the state**, **top 3 cities** where jobs are found in that state, as well as important information about salary in that state, including **average starting salary**, **average maximum salary**, and **salary range**. This allows the user to examine the specific salary numbers within a state they are interested in.

Graph 3: Giant Company Rating vs Job Salary in Big Data Field

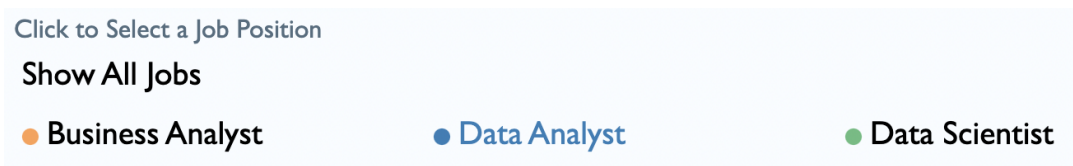


Choose a Salary Type

☒ Starting Salary

☐ Maximum Salary

For the scatterplot, we added interactive elements of **radio buttons**, **interactive color legends**, and **hovering**. We selected interactive radio buttons, one for starting salary, and one for maximum salary, because different users may be interested in seeing either in the scatterplot. The user's selection of either changes the **x-axis label**; for example, if the user chooses to view starting salaries by clicking on the first radio button, the x-axis label shows "Starting Salary (\$)," responding to the user's selection.



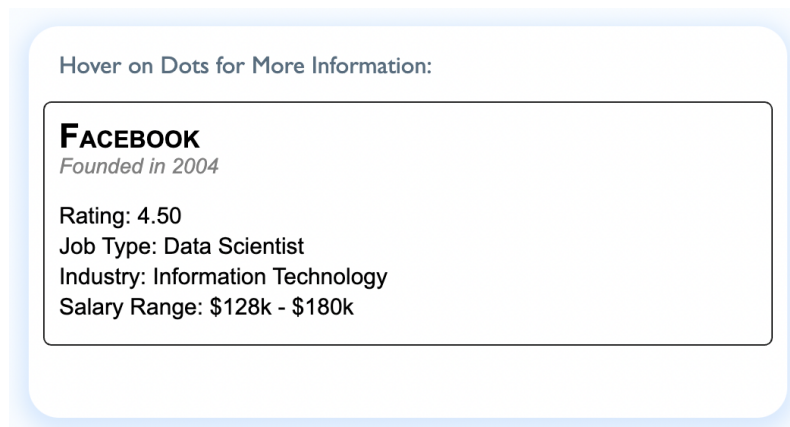
Click to Select a Job Position

[Show All Jobs](#)

☒ Business Analyst ☒ Data Analyst ☒ Data Scientist

Like the first and second plots, we wanted the user to be able to view only the job types that are relevant or interesting to them by choosing a job type filter. However, instead of doing this with a dropdown menu like the previous plots, we used an **interactive legend** in this scatterplot. Since we used color hue as a visual channel that represents the job types of each posting, we needed a legend to make the color mapping clear to users. Therefore, instead of adding another dropdown menu that would include similar information, we added interactivity to the legend so it **affords clicking**. When the user clicks a specific job type in the legend, only the postings within that job

type are filtered in, while other job types are filtered out. We made this interaction discoverable by adding a **short instruction** above the legend and **changing the color** of each job type blue when the user hovers over it, signifying the affordance of clicking when the user mouses over each job type in the legend.



Finally, consistent with the other parts of the visualization, each data point in this scatterplot affords hovering. When a circle is hovered over, it **expands in size**, **increases in opacity**, and **outlines in dark grey**, providing feedback to the user that it has been selected. A textbox appears to the right of the plot, showing the company at which the job is posted, its **founding year**, **rating**, **industry sector**, as well as **salary ranges**. We designed this to allow the user to inspect and learn more about a specific job posting on the scatterplot. For example, a user curious about the data point located at the upper right corner, which represents a job that is located at a high rating company and pays well, can hover over that circle to examine relevant information about that particular job.

PART 5: THE STORY

Story 1: *Where Can People Find More Job Opportunities?*

This visualization informs the viewers about the geographic and salary distribution of three types of data-related jobs in the U.S., business analysts, data analysts, and data scientists. In the map, we see that most data-related jobs are found in **big** and **populated states** like California, Texas, Illinois, and New York. **California** and **Texas** combined contain **more than 50%** of all data-related jobs in the U.S. It might be surprising to some users that there are also many data positions in Arizona as well. This trend in geographical distribution is similar in all three types of positions, which suggests that regardless of the specific type of position a viewer is looking for, whether they are interested in business analyst positions or data scientists positions, they are likely to find more opportunities in states like California and Texas.

Story 2: *Which State Offers Higher Salary?*

In the lollipop chart, we see salary distributions of data-related positions across different states in the U.S. Among all data-related jobs, we see that **California** and **New York States** offer the **highest starting salaries** and **maximum salaries**, compared to all other states in this chart. California also has the highest salary range. With an average starting salary of around \$83k and an average maximum salary of around \$135k, **California offers most compensation and room for salary increase** within its large salary range for data-related job positions. On the other hand, **Indiana, Georgia** and **Utah** are consistently on the right of the chart, suggesting that they offer the lowest starting and maximum salaries compared to other states in this plot.

For **business analysts**, the trend is similar in that the highest starting and maximum salaries are in California and New York states. For **data analysts**, the highest starting salaries occur in Illinois, which also has the smallest salary range, meaning that the data analyst job postings in Illinois, in this dataset, offer high compensation but little room for salary growth. However, California continues to offer high starting and maximum salaries with a wide range for data analysts. Finally, for **data scientists**, the highest starting and maximum salaries occur in Delaware, which lie high above the national average. However, this may be attributed to the small number of jobs available in Delaware. Similarly to other job types, California also offers higher salaries for data scientists compared to other states. This lollipop chart informs the viewer

of how much compensation they can expect from data-related job opportunities in different states, and from these observations, viewers looking for higher salaries may be interested in states like California for better compensation.

Story 3: *Which Position Offers Higer Salary?*

In the final scatterplot, viewers can directly compare salaries of the three job types. With the company ratings, we see that there is **no direct correlation** between the **ratings of companies** and the **salaries** they offer. We see in the scatterplot that giant companies with data-related positions clutter at a rating of around 3 to 4. For starting salaries, these jobs clutter at around \$40 to \$50k, and for maximum salaries, these jobs clutter at around \$80k. However, we also see that both starting and maximum salaries for data scientists tend to have a wider range, meaning that **data scientist jobs can offer higher salaries** than data analysts and business analysts.

PART 6: TEAM CONTRIBUTION

We had meetings twice a week to keep track of project's progress, make decisions for project direction, and discuss problems we each faced while coding or designing the visualization.

1. Team Roles & Responsibilities Assigned

Eva Kang was the visualization designer working on the visual designs and CSS styling. **Daisy Liu** was the coder, working on the data cleaning using python and coding for the first map graph. **Han Gao** was also the coder, working on the coding for the second lollipop graph and the third scatterplot graph. **Alice Chang** was the coder for the third scatterplot and writer for the project write up. All team members helped others when they had problems and contributed to the report writing.

2. Time Distribution

In total, our team spent about **21 hours** in this project.

- Data Collecting & Merging: 1 hours
- Data Processing: 3 hours
- Visualization Coding: 12 hours
- Visualization Styling: 2 hours
- Report Writing: 3 hours