Section A:

Introduction

We wanted to build a data visualization for a person, student, job seeker who was doing some research on the career opportunities for Data Science related roles in the United State and what the role might entail. We want to provide them with a more complete picture . For example, while salary is important, but it should be placed in context of other considerations to really grasp the overall value of an opportunity. Maybe the user didn't know they valued good weather so much? Maybe that six figure salary isn't so impressive when viewed alongside the cost-of-living index. So, we had to think what might a user find important?

The dataset we chose was from https://www.kaggle.com/andrewmvd/data-scientist-jobs. Which included 3,909 job listings for data scientist roles scraped from Glassdoor. There were 15 variables which include Job Title, Salary Range, Job Description, Rating, Company Location, Headquarters Location, Company Size, Founded, Type of Ownership, Industry, Sector, Revenue, Competitors, and Easy Apply. The dataset included variable types 3 different categorical, ordinal, and discrete.

Section B:

Exploratory Analysis

The data source we analyzed is about the employment of data scientists. Several elements like "index", "salary estimate", "location", "job title" etc. are explored. It is very obvious that we may be able to measure the employment of data scientists in different geographical locations by geographical distribution, for which we can use choropleth to analyze the data. Another charting method cartogram can also provide a good overview of the employment distribution of data scientists. This is a relatively intuitive preliminary idea.

Further analysis and assumptions on the data, we find that "salary estimate" may have a strong correlation with "location", "industry", "job title", etc. For this we can further explore in depth.

After an overview of the data, it is not difficult for us to ask the following questions: For example, do different industries directly determine the range of salary levels? Does the company's rating determine whether the job is easy to apply for? In general, which work industries have the highest employee satisfaction as known as rating? etc.

A large range of companies in the data are geographically located near New York, so we can think behind the hidden data. Can we discover whether different industries have the characteristics of being concentrated in different geographical locations?

In order to enhance the close interaction with the audience, we can use html-based choropleth geographic views. Therefore, when the viewer moves the cursor and wants to further explore the data of the geographic location, the detailed content can be enlarged, expanded, and zoomed in and so on. For salary comparison, we can use a grouped bar chart which has 2 bars, one for low salary and one for high salary.

In the end, we might have a choropleth that shows the distribution of data science jobs in the US. Upon choosing a state, the user is heat map and a bar chart. Heat map is to show the temperature varies in a year, and bar chart is to show the salary estimate, so that they can decide whether or not to look for a job and move to that state. This project is basically to facilitate the job search process. It helps users answer the questions mentioned above and narrow down on their choice based on important factors including salary, weather, location, loss of living.

Section C:
Scope and Design

Our ideas really started to take the shape of a story during Project Milestone 03. We began thinking from a user's perspective: If I were the user, what would I want to know? If I were the user, what information would be useful? We used these questions and considerations to create the mappings found in the flow diagram in Figure A.

Our intention was to provide the user with a more complete picture that reflected more than just a salary when searching/ contemplating a new job/job market. What's the weather like in that state where job *'x'* is offered? What's the cost-of-living? What would my quality of life be like compared to the average American citizen? What industry's offer jobs in each state? How many jobs does each industry offer? We understood that an interactive tool designed to help the user investigate these questions/considerations would be of value.

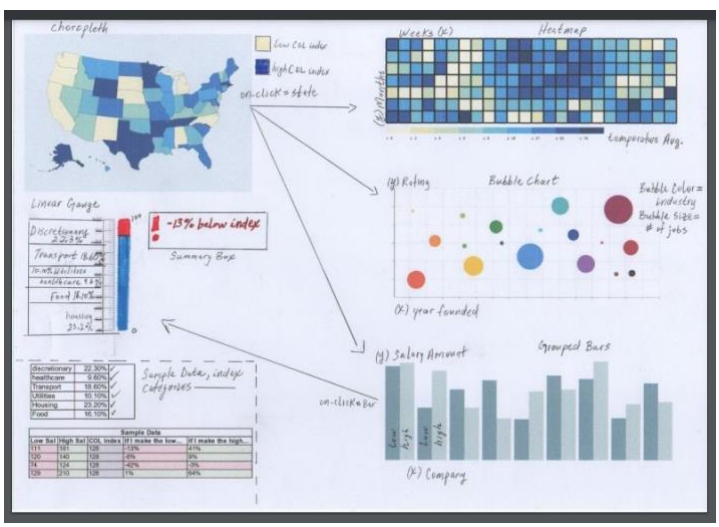
As you will see in the following section (section d), we gave careful consideration for usability and ease of visual interpretation. We achieved this by putting some extra thought into design considerations, such as:

- color scheme selection: ensuring the scheme matched the data.
- color palleted selection; ensuring the color was appropriate for the nature of the data and ensuring the color was appropriate in relation to the other visuals.
- using stroke and opacity when appropriate for visually communicating a 'selectable' item on the screen.
- using a pop-up box to communicate an exact value encoded within an element; allowing the user to easily process the data.

We wanted to display the data in novel ways while maintaining a degree of commonality for implementation of interactivity. We realized we could create consistency amongst the different visuals by using the US state abbreviation to 'connect' the data and visuals — allowing for manipulation and varying levels of examination by the user; following Schneiderman's mantra of overview first –zoom and filter—for details on demand.

While we fell short in many respects at implementing all the planned interactions — the overall scope and plan remained consistent. That is, the choropleth would serve as the foundation and filtering mechanism for the other visuals at the state level. By 'clicking-to-select' a US State in the choropleth, the other graphs would respond by highlighting or filtering data of the selected US State. We also didn't get to fully develop the 'Linear Gauge' which would have provided a summation for the cost-of-living index vs. salary (ascending in accordance with Maslow's Hierarchy of needs). As you will see in the following section (section d), we were able to execute the interaction between the choropleth and heatmap.

FIGURE A:



c) Interactive Visual Analysis Tool:

SEE FIGURE B: VISUALS OVERVIEW

Choropleth…

• Serve as the foundation and filtering mechanism for the web application
• Sequential color scheme chosen to indicate low to high 'Cost of Living Index'.
  - The lighter the color the lower the 'Cost of Living Index'.
  - Similarly, the darker the color the higher the 'Cost of Living Index'.
  - A black-grey-white palette chosen intentionally, as I didn't want to potentially confuse the user [i.e. – we didn't want the user to interpret the choropleth as displaying weather data].

Heatmap…

• Diverging color scheme chosen to indicate cold to hot spectrum of temperature data which I figured would be intuitive for the user.
• Red-Yellow-Blue color palleted chosen deliberately, and matches the data [blue=colder temperatures, red=hot temperatures]
• x-axis indicates the US state abbreviation, y-axis indicates the yyyy-mm

Bubble Chart…

• Categorical color scheme chose to align with the categorical data [number of jobs by industry] — clearly indicates that each 'bubble' is a different industry because it's a different color.
• The size of each 'bubble' corresponds to the number of jobs available.

Grouped Barchart…

• each pair of bars indicates the 'low'/'high' salary for the job title
• Job titles are listed on the x-axis, the y-axis indicates the salary amount.
• Uses two distinct colors to clearly indicate the 'low'/'high' nature of the data.

SEE FIGURE C: MOUSEOVER, CLICK-TO-SELECT, PAN/ZOOM

Choropleth…

• Pan and Zoom functionality included for extendibility; assuming future development iterations would incorporate more detailed data-views [for example, at the city-level].
• Mouse-over highlights the 'State' to clearly indicate selection to the 'user' using a purple boarder and opacity change.
• Mouse-click highlights the 'State' to clearly indicate selection to the 'user' using a purple boarder. Allows for multiple selections.

Heatmap…

• Mouse-over highlights the 'rectangle' to clearly indicate selection to the 'user' using a black boarder and opacity change.
• Mouse-over produces a 'pop-up' box with the temperature value in large font so the user can easily process the value.
  - Following Schneiderman's mantra of overview first –zoom and filter—for details on demand.

Grouped Barchart…

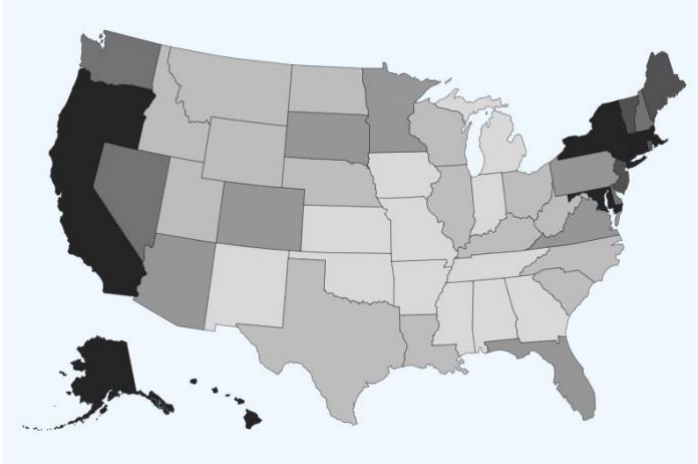
- Mouse-over highlights the ‘rectangle’ to clearly indicate selection to the ‘user’ using opacity change.

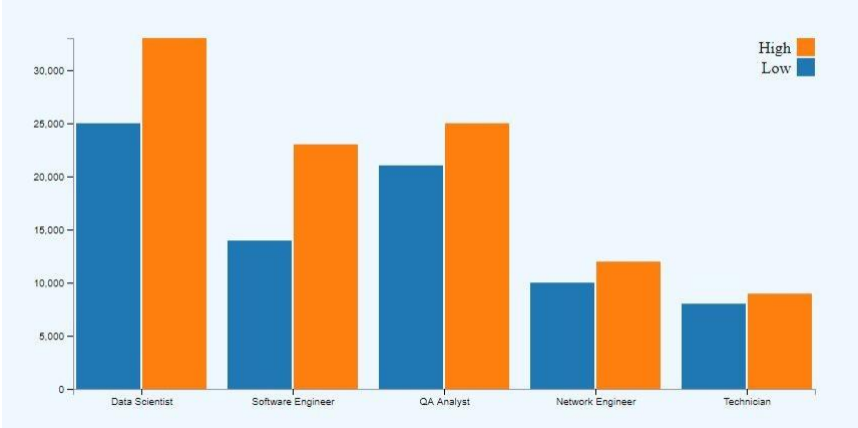
SEE FIGURE D: INTERACTIVITY

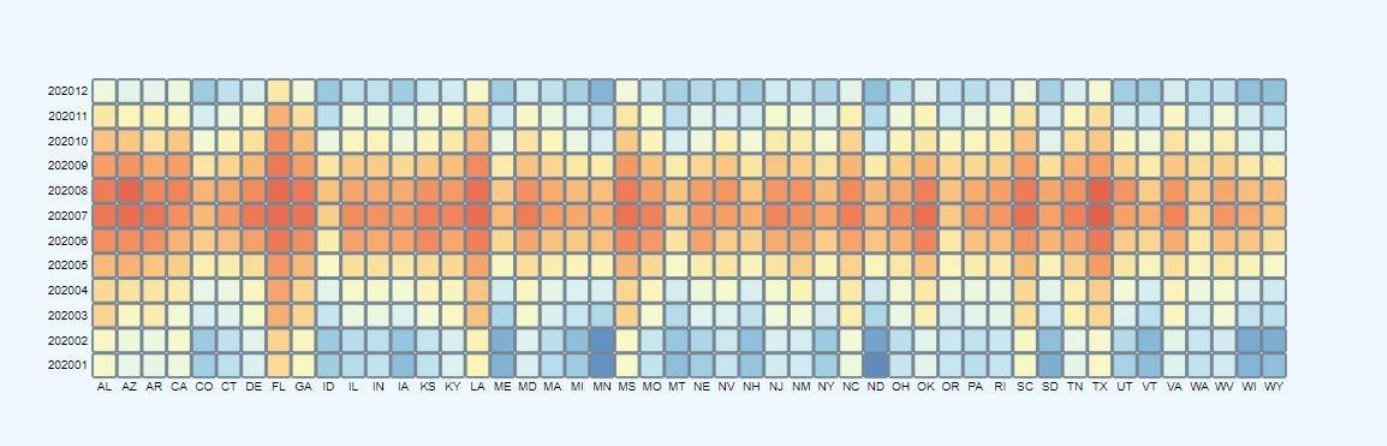
Choropleth and Heatmap…

- On-click within the choropleth map highlights the ‘State’ in purple and triggers a d3.dispatch() event which highlights the selected State’s ‘cells’ in the heatmap.
- The d3.dispatch() event is ‘undone’ by clicking the State again which unselects the choropleth State and corresponding heatmap cell. This is managed both visually and with a ‘Last-In,First-Out’ [LIFO] structure.
- The choropleth and heatmap connect through a common ‘key’ in the data [US State Abbreviation]. The code can be found under the d3.dispatch() call in index.html.
- Encoded in ‘purple for a few reasons: 1) it has no intuitive connection with weather data, 2) to give it a ‘highlighter’ feel like you might expect when reviewing a document by hand, and 3) the contrast is decent against both color palettes.

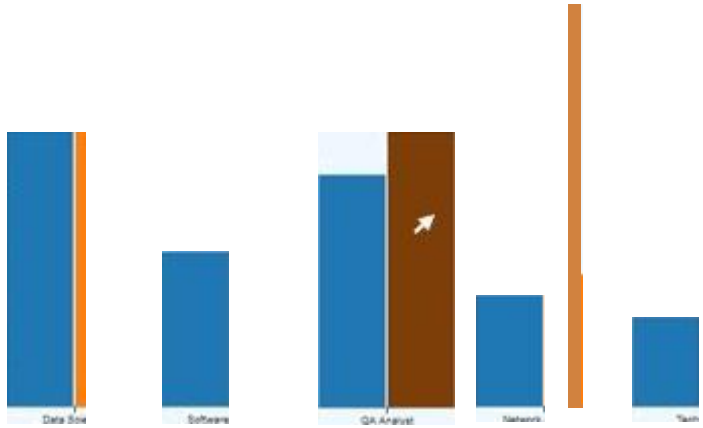
Figure b: Visuals Overview





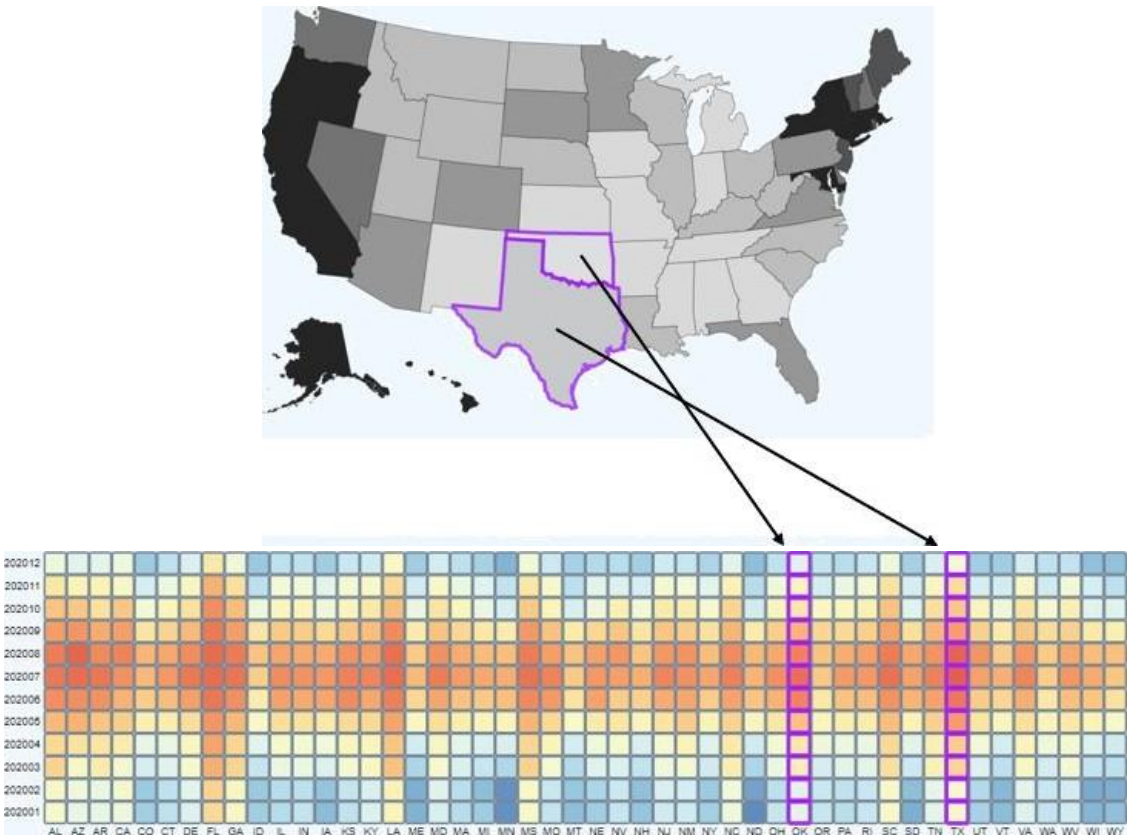


Fiqure c Mouseover



Temperature: 75.6

Figure d: Interactivity



Section E:

In Choropleth, the color scheme from light to dark shows the increase of the cost-of-living index. The lighter the color, the lower the cost-of-living index. The darker the color, the higher the cost-of-living index. Correspondingly, the cost-of-living index is higher on the east and west coasts, and the cost-of-living index is lower in the central region. On the other hand, from the observation of our data, there are more job opportunities on the east and west coasts, which is undoubtedly a double-edged sword. These factors are undoubtedly extremely important reference information for job seekers.

In the heat map that interacts with Choropleth, we can explore the temperature changes in different states in different months of 2020. Generally speaking, job seekers may be more reluctant to relocate in too hot or too cold weather. On the contrary, in the non-extreme weathers, we predict that this will be the days when job seekers will be more active in their desire to start a new life.

In the bubble chart, we show the average annual net income of data science jobs in different industries. Surprisingly, the data science corresponding positions in the real estate industry are the

most highly paid, not in the information technology industry as our assumptions. Due to the special location of New York City, the education industry is also the top three optional industries for the same job salary. The biotechnology pharmaceutical industry is a new generation of giant upstarts. Therefore, data scientists who want to gain a foothold in New York City can refer to the above information.

In the bar chart, looking at various data science positions, we can tell that the positions with the highest average salary are senior data scientist, and the positions with the lowest average salary are machine learning engineer and data engineer. Relatively speaking, the position with the largest increase in the pay gap is senior data scientist. This conclusion proves that the data science field is still an industry with considerable income, and at the same time it can also help job seekers to make further career choices. Accumulating work experience may be the most popular thing on the job market.

Individual Report – Thang Tran

At first, we wanted to go with Twitter data. It contains customer's reviews about their flight experience with certain airlines. It looks very interesting, but we realize that visualizing text data is not an easy task. Furthermore, we would like to inform classmates of the potential in data science field which is why we switched to using data science job data for the final project. For someone who is looking to get a job in data science, in addition to salary, they may need to know if they can cope with the weather in that city, the amount of money they make is enough to sustain their lifestyle so they can live comfortably, etc.

In the first milestone, I was responsible for presenting our group's plan and conveying the feedback to everyone. Although the data looks promising, I received a few suggestions from Professor Brown to improve the quality of our project. We were suggested to add cost of living so that user can make the decision whether or not the salary paid justifies the cost of living. Access to education, entertainment was also mentioned, but we were not able to find the appropriate data.

In milestone 2, I did research on which graphs are most suitable for our data, and possibilities of making them interact with one another. One of the suggestions that I made was kept for the final project which I myself implemented. It is the bar chart. We would like to show the users the highest and lowest salary of a certain job title.

d3 is used to make the bar chart. The first part is to simplify the data. Although the job title is the same, each company pays their employees differently. Therefore, we decided to take the average of the lowest and highest pay of each job title in each state. Luckily, we did not run into an issue of outlier where one person gets paid extremely high resulting in distortion in the average. We chose the top 5 common job titles. The reason is because these jobs are easier to look for while searching for jobs, the job description is often clear so prospect applicants know what to expect, and its data (pay, benefits, etc.) can be easily found on the internet.

For the graphic display, we opted for a grouped bar chart where each job has 2 bars, one is for lowest salary and one for the highest salary sitting next to each other for easy comparison. The one of most challenging parts is to process the input data in a specific way so that when we call d3 methods, 2 bars are drawn next to each other, and they are filled with 2 distinguishable colors. The simplest, although not the most optimal solution is to hardcode the data into JavaScript file and process from there rather than relying on d3 loading methods.

Another problem is drawing legend. The bars could be too high and cover legend. In order to avoid that, the right-most job title on the bar chart can't have the highest salary comparing the remaining 4. Bar's height is scaled according to salary. As for interaction, upon hovering the mouse on any bar, that bar is highlighted so that user know what they have chosen. This is achieved by using mouseover, mouse out event, and it's fairly simply to implement.

What I learned about data visualization during the development is even though JavaScript gives developers a lot of freedom, it comes with the cost of messy code and hard to follow. Another key takeaway is d3 is version dependent. One piece of code can work on version 3, but it does not necessarily mean it can work on version 4.

Data visualization is powerful way to extract valuable information from a dataset and convey it to the users. However, how to make all the data we extract interact with one another is easier said than done. Because of the lack of experience in using JavaScript and d3, I was not able to get the bar chart to interact with the choropleth Kyle implemented. I chose to visualize the data of the state of New York in my bar chart because that is essentially the core of the bubble chart. Furthermore, if time permit, I would like to add a linear gauge so that user can see the cost of living in each state.

Individual Report - Vi Nguyen

My contribution involved researching and presenting potential datasets and ideas of the final project to the group and also picking out our group name. I suggested a twitter dataset that had flight information. However, we ended up selecting data science salary because we found it all relevant to the group members. I also examined this potential . I further examined and discussed the benefits and drawbacks of the data science salary data set. For example, it was clear that it would require regular expression to clean it since the data is not in ready to use format. I also discuss the challenges of using a self-report dataset as the salary may be inflated rather than accurate. I attended an after class check in point and informed our professor of our progress and direction and reported back to the group. My main contribution involved cleaning and transforming the dataset. I took the requirement of the project and team members and was tasked to clean and transform this dataset in a usable form using a Jupiter Notebook.

There were a few things I did to clean the dataset. I also had to revise the dataset cleaning later during implementation when one team member found the problem with mapping the numerous data with the data visualization. I applied a regular expression to remove all unnecessary characters from data. I also did feature engineering in order to transform unusable variables and columns into usable data. I research how to normalize salary data. This took several attempts, as a team member also suggested a way to normalized salary but it didn't not seem accurate so I researched one that would be best. I also did some analysis on data to suggest to team members where we should go with a project such as picking out the top 5 data science job titles rather than use them all.

Next as we were expanding the scope of the visualization, I found a cost-of-living dataset and merged it into our dataset to meet the requirements of the Project. This involved splitting location into city and state in order to merge a second dataset. I also filmed and participated in presenting the final project to the class. Originally, I advocated for us to all record together but due to a big time zone difference the group believed it would be best for the team members to record separately.

I also did some ad-hoc data analysis, when a team member requested a more succinct and smaller dataset with just salary by top 5 jobs with salary and state because of data mapping issues. I did some extra data cleaning in order to deliver him that abridged version of the data for his task. I also participated in picking out which data visualization would make sense for our project. I also format the final report.

<u>What I learn</u>

Details are incredibly important, and an effective data visualization takes into many details and consideration. When I first started this course, I was only interested in nice looking visualization. I thought sunburst is a very nice data visualization and should be implemented. However, after taking this class, I realize that your data visualization has a purpose, and you take into consideration other considerations like for our data it may not make sense to pick a data visualization just because it looks nice and does not communicate the story we wanted. I also learn about using D3 and Javascript. The best ways to store data and I never believed it before but using D3 and Javascript, I can actually see how hard coding certain variables and data would be bad and inefficient.

Individual Report – Ximan Liu

Overview of the data that has been cleaned for the first time, variables include job title, location, company size, industry, salary level etc. attracted my attention. What they have in common is that they all belong to the data science jobs. The information of data science positions is arranged and distributed according to different geographic locations. The data that attracts me the most are geographic location, salary, and cost of living, because these three have a certain correlation, perhaps these three factors can directly affect the employment choice of job seekers.

I used d3 to create a bubble chart and represent the average net income of data science jobs in different industries in New York City. The reason why I choose bubble chart because it is more intuitive to show the salary influence of different industries and can directly compare the differences between different industries. In view of the fact that our group has already appeared bar chart to show other data research, so the application of bubble chart here is more novel.

As a graphic display, the most important step is to sort out the data that may be used. First, I estimated the average salary level of each position based on salary estimate lower range and salary estimate upper range, and then offset the cost of living in NYC for a year (approximately $33k). The calculated result, as known as average net income of different industries, is prepared as the basic data of the bubble chart. There is no need for too many interactive displays since we only need to combine and connect bubble chart with the choropleth map.

In order to better design the bubble chart, I choose different colors to correspond to different industries. As we know, the radius of the bubble determines the size of each bubble, and the size of the bubble represents the difference in the size of the average net income.

In the development process, I think cleaning the data is the most challenging step, because the average net income needs to be calculated according to the formula to achieve the next step.

For the visual design of our group, we chose choropleth map, heatmap, and bar chart, etc. It is necessary to analyze various factors because it may determine the candidate's choice of data science positions. My bubble chart is display and exploration based on zooming in the certain part on choropleth map.

The conclusion from my visualization is that the industry with the highest average net income for data science jobs in NYC is the real estate industry, not the technology industry or the financial industry as we imagined. Given the location the Big Apple, the education industry is also considerable. In addition to traditional manufacturing, the biotechnology pharmaceutical industry is also a high-income upstart industry. These characteristics are related to the unique geographical location of the city itself.

If there is more time, I will add one more variable, such as rating. I will consider both the average net income and rating as common influencing factors. For job seekers in data science positions, it is also important to refer to the company’s evaluation and satisfaction when making industry and company selection.

Individual Report – Kyle Arick Kassen

The following document contains information related to my responsibilities and contributions to the group project alongside a short reflection of my experience during the development of the project. I included my ***programming*** and ***design*** contributions, explanations, and considerations.

Planning:

- Consolidated group ideas, created a flow-diagram mapping of the interactions + 'story', and presented these ideas in PM03
- Sourced and cleaned weather data: https://www.ncdc.noaa.gov/cag/statewide/mapping/110/tavg/202005/1/value
- Sourced and cleaned ***supplementary*** cost-of-living Index dataset: https://meric.mo.gov/data/cost-living-data-series

Combining Code:

- Combined all group members code into one organized application.
- Transformed all group member code into JavaScript files enclosed in an outer wrapper, wrapped in a function, and returned a new object.
- Coded the index.html file and created new objects for each group members visualization.
- Coded/Created the d3.dispatch() and set it up to allow the other group members to easily connect their code for interaction with the choropleth.
- Reworked the source data so that all group members could connect their visualization to the choropleth through a common 'key' (splitting off 'US State abbreviation' into a separate column).

Server—Code and Design:

- Coded/Created the server.py program based on Lab 5 to ensure robustness of the web application and avoid errors in retrieving data.
- Allowed the opportunity for extendibility of the web application [for example, adding some Machine Learning functionality].

Choropleth—Code and Design:

- Coded/Created/Designed
- Sequential color scheme chosen to indicate low to high 'Cost of Living Index'.
    - The lighter the color the lower the 'Cost of Living Index'
    - Similarly, the darker the color the higher the 'Cost of Living Index'
    - Black-grey-white palette chosen intentionally, as I didn't want to potentially confuse the user [i.e. – I didn't want the user to interpret the choropleth as displaying weather data].
- Pan and Zoom functionality included for extendibility; assuming future development iterations would incorporate more detailed data-views [for example, at the city-level].
- Mouse-over highlights the 'State' to clearly indicate selection to the 'user' using a purple boarder and opacity change.
- Mouse-click highlights the 'State' to clearly indicate selection to the 'user' using a purple boarder. Allows for multiple selections.

Heatmap—Code and Design:

- Coded/Created/Designed
- Diverging color scheme chosen to indicate cold to hot spectrum of temperature data which I

figured would be intuitive for the user.
- Red-Yellow-Blue color palleted chosen deliberately, and matches the data [blue=colder temperatures, red=hot temperatures]
- Mouse-over highlights the 'rectangle' to clearly indicate selection to the 'user' using a black boarder and opacity change.
- Mouse-over produces a 'pop-up' box with the temperature value in large font so the user can easily process the value.
- Following Schneiderman's mantra of overview first –zoom and filter—for details on demand.

Interactivity—Code and Design:

- Coded/Created/Designed the interaction between choropleth and heatmap.
- On-click within the choropleth map highlights the 'State' in purple and triggers a d3.dispatch() event which highlights the selected State's 'cells' in the heatmap.
- The d3.dispatch() event is 'undone' by clicking the State again which unselects the choropleth State and corresponding heatmap cell. This is managed both visually and with a 'Last-In,First-Out' [LIFO] structure.
- The choropleth and heatmap connect through a common 'key' in the data [US State Abbreviation]. The code can be found under the d3.dispatch() call in index.html.
- I chose 'purple for a few reasons: 1) it has no intuitive connection with weather data, 2) to give it a 'highlighter' feel like you might expect when reviewing a document by hand, and 3) the contrast is decent against both color palettes.

Group Report:

- wrote section *c) Scope and design and put together figures a,b,c,d*
- wrote section *d) Interactive Visual Analysis Tool*

Reflections:

Main takeaways for what I learned about data visualization during the project…

1) Taking the time to get the color scheme and palette to align with your data and visual goes a long way towards building an attractive, intuitive, and user-friendly visualization.
2) User friendly — 'intuitive' functionality can be challenging to accomplish programmatically. For example, the LIFO structure for managing the 'undo' interactivity between the choropleth and heatmap seems perfectly reasonable to a programmer. However, the more user-friendly (and more challenging solution) is to make the 'undo' interaction more abstract [probably using (***this***) instead of the LIFO structure] so the user can unselect states in any ordering. If a I had more time, this is something I would have strived to implement.
3) Connecting the data between separate visualizations using a common key. Web programming can get awfully messy and accessing the objects and data can be challenging. That said, once you figure out how to get them connected, the overall quality of the experience, visualization, and application increases tremendously making the payoff worth the effort.
4) I need to continue to learn, practice, and strengthen my web programming skills (html, CSS, JavaScript). Often times it seemed like what's considered the more advanced programming concepts came instinctively (experience with C/C++, Java, Pythoness) while the 'easier' stuff was super difficult for me. For example, I could never figure out how to make a legend the way I wanted for the choropleth. I even tried to build one with just a 'high' value and 'low' value, and I was successful; however, this caused my 'on-click' functionality to stop working. For example, I could never figure out how to properly format the date on the y-axis of the heatmap (I could get the formatting but not

'put' the correct data into the formatting).

Sources:

Color palettes: https://colorbrewer2.org/

Server.py: Professor Eli T. Brown | CSC 468, Week 7 Live Lecture on Flask Servers using Python

Weather Data: https://www.ncdc.noaa.gov/cag/statewide/mapping/110/tavg/202005/1/value

Cost of Living Index: https://meric.mo.gov/data/cost-living-data-series

Choropleth: Interactive Data Visualization for the Web, 2nd Edition, Scott Murray, Chapters 14

Json Data: [https://github.com/scotthmurray/d3-book/tree/master/chapter_14] Scott Murray: "This file consists of GeoJSON data generated by Mike Bostock from U.S. Census data, and is in the public domain. This file used to be included in earlier versions of the D3 repository and included this summary on its origins:"…"These are derived from the cartographic boundary files from the 2000 U.S. Census. Then, MapShaper was used to simplify the geometry, and ogr2ogr to convert the shapefiles to GeoJSON. Some additional work was done to preserve the FIPS codes, which are dropped from the shapefiles by MapShaper."

Heatmap: https://www.d3-graph-gallery.com/graph/heatmap_style.html

Instructions for running the application:

#1) open terminal or command prompt under the directory containing the project file #2) $set FLASK_APP= server.py
#3) $python -m flask run
#4) type in browser: http://localhost:5000/index.html