Data Visualisation Project

“Impact of Industry and Educational Qualification on Salary Range in the United States”

November 2023

VINAY MITTAL
Student Id: 33613877
Applied Session Number: 10
Tutor Name: Kadek Satriadi

Table of Contents

Introduction

Many people dream of living a prosperous life in the competitive employment market in the United States, where some people struggle, and others thrive. But in the USA, it's crucial to have a high income to live a good life. That was the motivation behind my exploration project where I analysed the factors that can assist in reaching a desirable salary bracket. To uncover what contributes to achieve a good salary bracket in the USA, we will delve into the findings of the following questions:

Question 1: Does a specific industry type have influence on a particular income range in the USA?
Question 2: Is a higher bracket of income driven by a specific level of education?
Question 3: What is the gender distribution between industries and their evolution over time?

To answer my first question, a key finding suggests which type of industry has a significant effect on how likely you are to acquire a particular salary bracket. Though, it's not enough to just enter the field—what congressional district you work is also of the utmost significance.
One important finding related to the second question highlights how important education levels are in the United States for obtaining a desired income. There is a strong correlation between an individual's having higher level of education and higher income brackets, whereas lower educational attainment is correlated with lower income ranges.
The third question highlights that despite knowing about the most profitable industries, breaking into them can be quite challenging. This is because certain industries are dominated by males, while others are female dominated.

My target **audience** is not limited to aspiring domestic or international students choosing their courses for higher education to get into the right Industry. It also includes working people who are thinking about changing careers in the United States.

Design

Started off by brainstorming with ideas for visualisations using the Five Design Sheet process as shown in *Figure 1.1*.
It's important to show how U.S. individual are distributed within different salaries before diving into the questions. I considered two charts, a line chart and a bar plot, to accomplish this.
The main objective is to show that a significant proportion of the population is in higher pay groups, implying that it is not unrealistic to aim to be in such ranges. I chose bar-plot to visualise this as it is an intelligent choice to abandon the line chart in favour of a bar plot, which is a more useful visual aid for displaying the distribution of categorical data where the categories are distinct salary ranges.

To address the **first question**, where I aim to showcase that an industry type is associated by different income ranges one belongs to, I considered a Sankey diagram. This visualization establishes a distinct connection, highlighting the potential salary brackets one can attain based on their association with a specific industry. I particularly used Sankey

diagram as it is the best visual tool to show many to many relationships between two domains and very easily to comprehend as flow paths are self-explanatory.

In the second visual for this question, a stacked bar plot will be used to illustrate the crucial impact that work location plays in both raising and lowering income ranges while keeping the industry parameter constant. The stacked bar plot's counts are all related to the number of congressional districts offering a particular salary range grouped by different industry type. The following graphic shows how people in different congressional districts are associated with the same industry and different salary range categories.

We'll now analyse which congressional district provides the greatest compensation for a certain industry in the last visual of the question. A great illustration of the best places to work industries to reach a desired income bracket is provided by the U.S. chloropleth map at district level granularity. You may see which congressional district is dominated by a particular industry and the associated pay offerings by moving the pointer over the map. The reason behind selection of the map is very user-friendly and interactive.

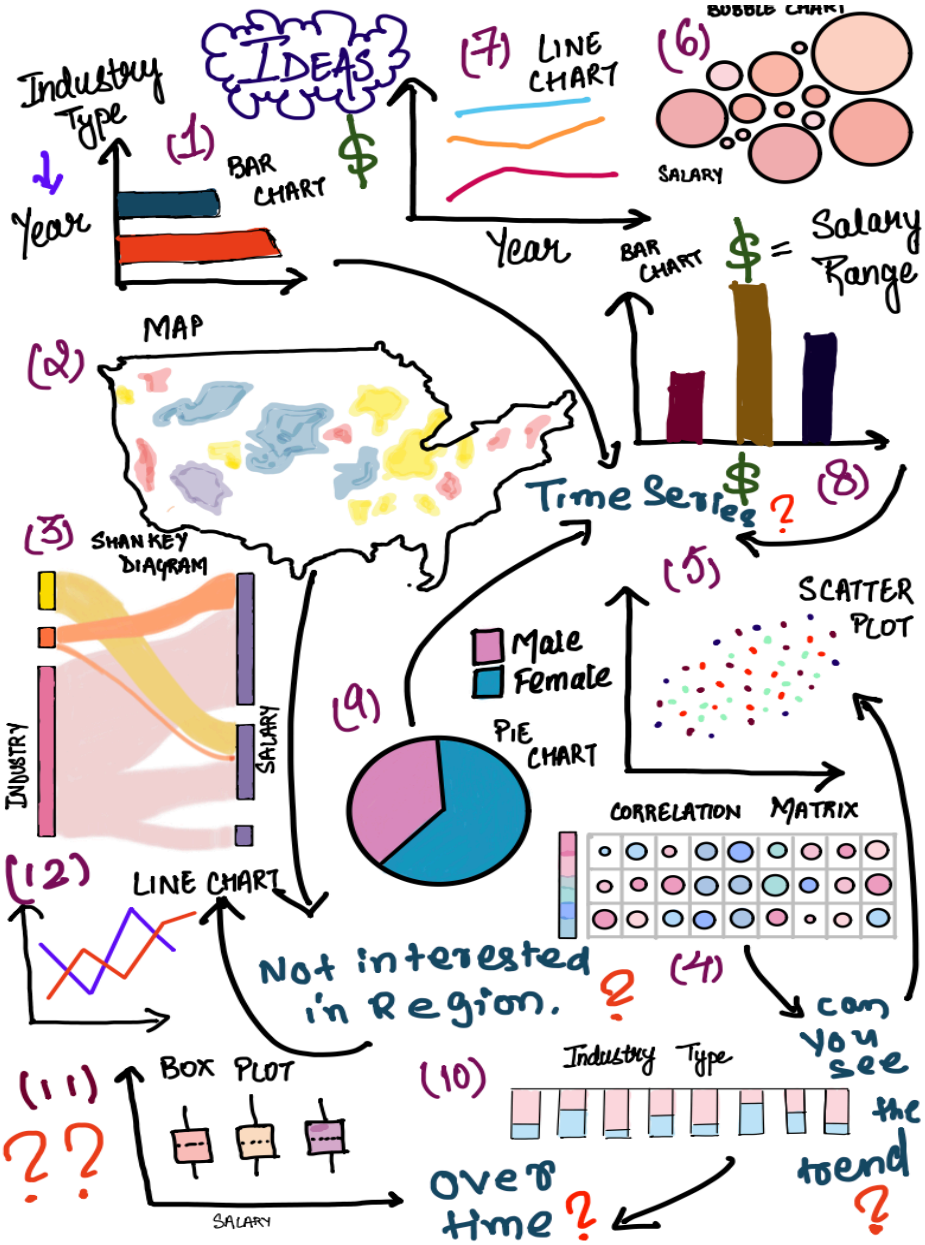


Figure 1.1 Brainstorming

To show the relationship between salary ranges and educational qualification, I choose to create a heat map correlation matrix for the second question. Although colour is sufficient to demonstrate both direction and magnitude, but not that user-friendly, User may find it difficult to grasp the magnitude based just on hue intensity. I have given extra touch by adding circle shape to improve clarity by measuring the magnitude using circle radii. This enhancement makes it easier for users to discern magnitude right away, leading to a more intuitive comprehension of the correlation heatmap. Hence the magnitude and direction of the correlation are shown by the size and colour of the circle; larger circles denote stronger magnitudes, blue for positive correlations, and red for negative correlations. I selected colours like blue and red for the sole reason that our eyes naturally understand them well. Blue means good connections, like between salary and education, and red shows when things aren't connected well or to show a negative impression. The size of the circles in the correlation matrix matters—bigger blue circles mean larger correlation coefficient, and bigger red circles mean larger negative correlation coefficient. This helps people quickly see and understand the relationships between salary ranges and education levels.

Finally, after a high overview of the between different combination of educational qualification and income ranges, the scatter plot serves as an excellent visual for displaying correlations as it allows for a visual representation of individual data points. In the scatter plot, I added a regression line to improve the visual clarity as it easier to human eyes to see the slope directly rather than understanding the trend by points. This line's slope reveals the magnitude of the correlation between income ranges and education levels in addition to its direction. This update provides a precise and straightforward way to visualise the relationship between income levels and education, helping to explain the overall trend and strength of the relationship.

For the third question, I have considered that the best visual aid to demonstrate the gender distribution in the industry would be a pie chart. Its simplicity comes in handy when working with a small set of categories. I decided to use a pie chart to provide a general picture of the subject before getting into the details.
The second visual for presenting the findings involves incorporating an interactive chloropleth map. By clicking on a specific state, a stacked bar plot will dynamically generate, showcasing the gender distribution within the industry for that state. This linked interaction feature adds a dynamic and localized dimension to the presentation, offering a detailed distribution of gender dynamics in different industries at the state level. It wouldn't be realistic to use a pie chart in this as well because of possible clutter given the large number of industries. Alternatively, the stacked bar plot is a suitable option because it can handle multiple categories,
Pie charts, stack bar and chloropleth map plots typically employ the colour palette having pink and blue to symbolise and categorise males and females, respectively. Choosing the traditional colour palette of pink and blue is a simple way since it guarantees general comprehension. In general idea, pink is for girls and blue is for boys. Anyone may easily and quickly understand the gender-related information displayed in the charts with this straightforward colour association. In addition to giving the visuals more aesthetic appeal, the distinct contrast between the two colours helps to effectively communicate demographic data.

Discussed above are the shortlisted charts which I have used in different design sheet to convey my findings. To effectively visualise my findings, I came up with many other charts throughout the brainstorming phase, including bubble charts, box plots, and line charts. Still, I decided to toss these visualisations away. For example, a bubble chart was designed to show the distribution of salaries. It was discarded because the bubbles are round, it can be difficult for the human eye to distinguish minute variations in radius, which can lead to misunderstanding, particularly when comparing income ranges. The audience's ability to appropriately notice and evaluate the tiny differences in income distribution may be hampered by this visual complexity. As a result, I went with different visualisations that do the content justice without adding needless visual complexity. Three nominated designs were chosen to create the charts that were previously discussed. I will now go over each design, emphasising the features or visuals I ultimately chose to keep and those I decided to leave out. This involves discussing the strengths and weaknesses of each design that led to the creation of the final design.

First Design Sheet

In the **first design** as shown in *figure 1.2*, the layout was designed like a sequential guide, where users would click horizontal radio buttons to navigate through the visualization from start to end. However, I found this layout to be less favourable as transitioning to the next visualization using radio buttons disrupted the flow of storytelling and the consistency between the visuals will be broken. I therefore made the decision not to incorporate this navigation in the final design.

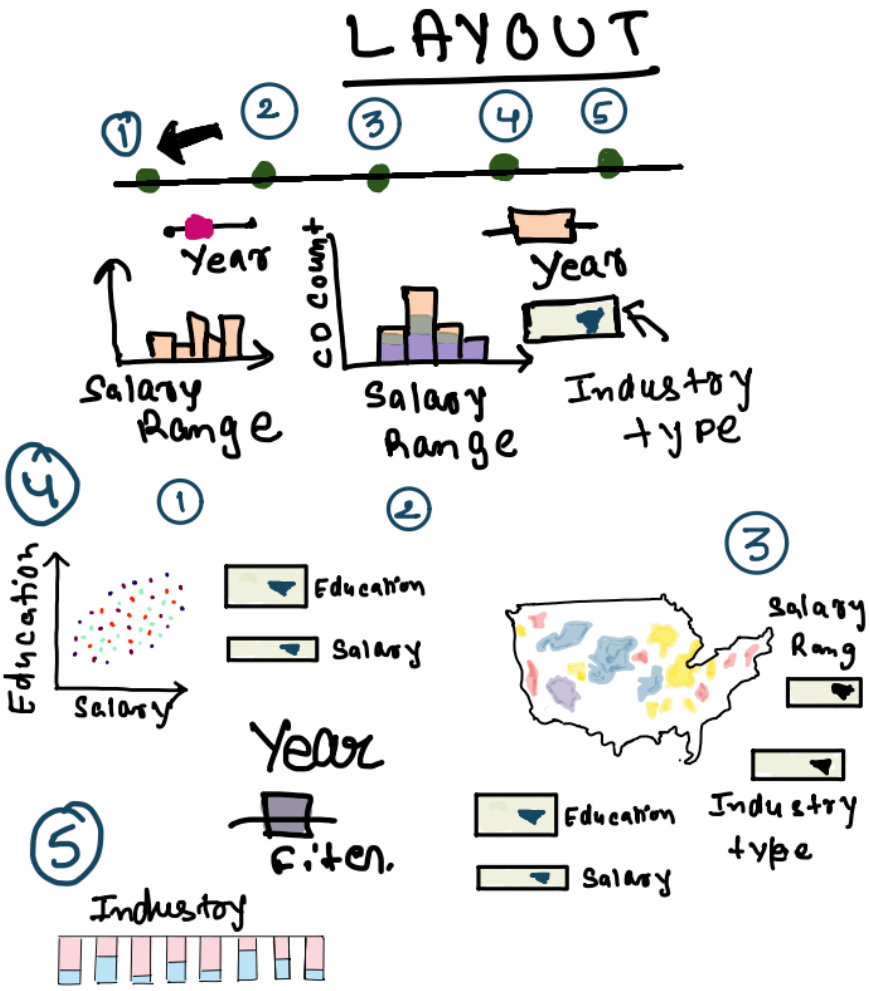


Figure 1.2 Initial Design - 1

However, I favoured the concept of starting my story with a bar plot illustrating the distribution of individuals in the U.S. across different salary ranges. The bar plot, equipped with a year filter, made its way into the final design. Additionally, from this design I also pulled in a stacked bar plot demonstrating how the same industry presents various salary ranges across different congressional districts. Although I felt the design lacked a high-level overview, prompting me to introduce a visualization in my final design sheet that demonstrates how each industry can offer distinct salary ranges. Moreover, I also took an interactive chloropleth map from this design, featuring filters for different salary ranges and industry types. This interactive map allows users to hover over each congressional district, revealing details such as the dominant industry and the corresponding salary offerings. This user-friendly feature simplifies for individuals to visualise locations that offer the highest salaries.

Additionally, I used the scatter plot from this design to show the correlation between different combinations of level of education and income ranges. An interactive filter for education levels and income ranges is included in this scatter plot. However, overview of all possible combination is missing in the design. Furthermore, I saw that the scatter plot by itself did not offer enough quantitative data. Therefore, I thought of adding a regression line to the scatter plot, which made it possible to quantify each pair's magnitude and direction more precisely. The line's slope turns out to be a useful tool for illustrating the direction and intensity of the relationship between education levels and salary ranges.

I think it's a great idea to show the distribution of gender in various industries using a stacked bar plot. Nevertheless, I discovered that adding a year filter obscures the location-based distribution of industries.

Second Design Sheet

In the **second design** as shown in *figure 1.3*, the layout involved navigating through different tabs to progress through the visualization. However, I faced a similar issue as in the first design. This tab navigating layout could disrupt the continuity of the storytelling and break the coherence between visualizations. So, I decided not to go along with this navigation in my final design. Still, there were aspects of this design that I liked. The concept of displaying the distribution of Americans across various income levels really appealed to me. The Sankey diagram, which I included to show how different industry types might offer varying income ranges, was also the design's standout feature.

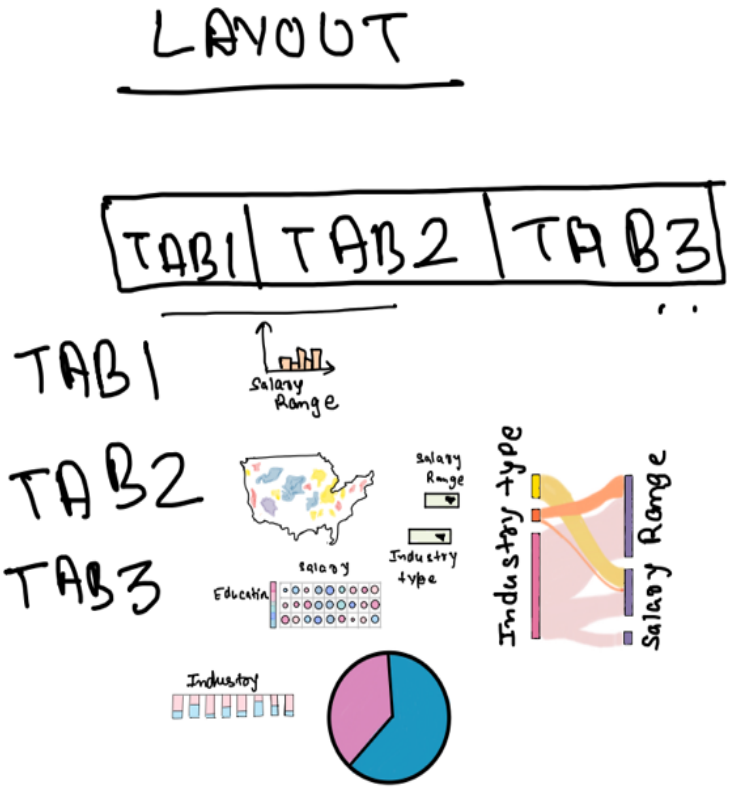


Figure 1.3 Initial Design - 2

Additionally, I incorporated the idea of a correlation matrix from this design, offering a high-level summary of the correlations between various combinations of income and education levels. In addition, I took a pie chart from this design that provides a thorough breakdown of the proportion of men and women among US states. I think it's a great idea to show the distribution of gender in various industries using a stacked bar plot. Nevertheless, I discovered that adding a year filter obscures the location-based distribution of industries.

Third Design Sheet

I picked up the idea of scrolling down as a useful navigation technique from the **third design** as shown in *figure-1.4*, which allowed for a smooth transition across the visualisation. This method adds to the overall coherence and consistency of the visualisations while also preserving the storytelling's continuity. In addition, I decided to illustrate the gender distribution working in different industries unique to each state by utilising a state filter rather than having a year filter only. A dropdown with 52 alternatives could be somewhat complicated. In my final design I have fixed this part to reduce the complexity for user interaction.

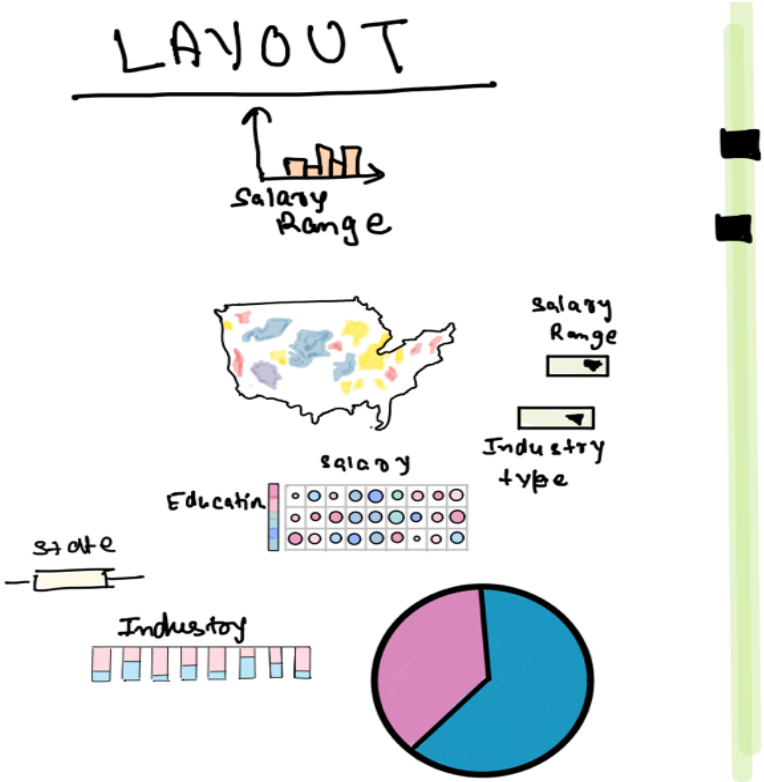


Figure 1.4 Initial Design - 3

Final Design Sheet

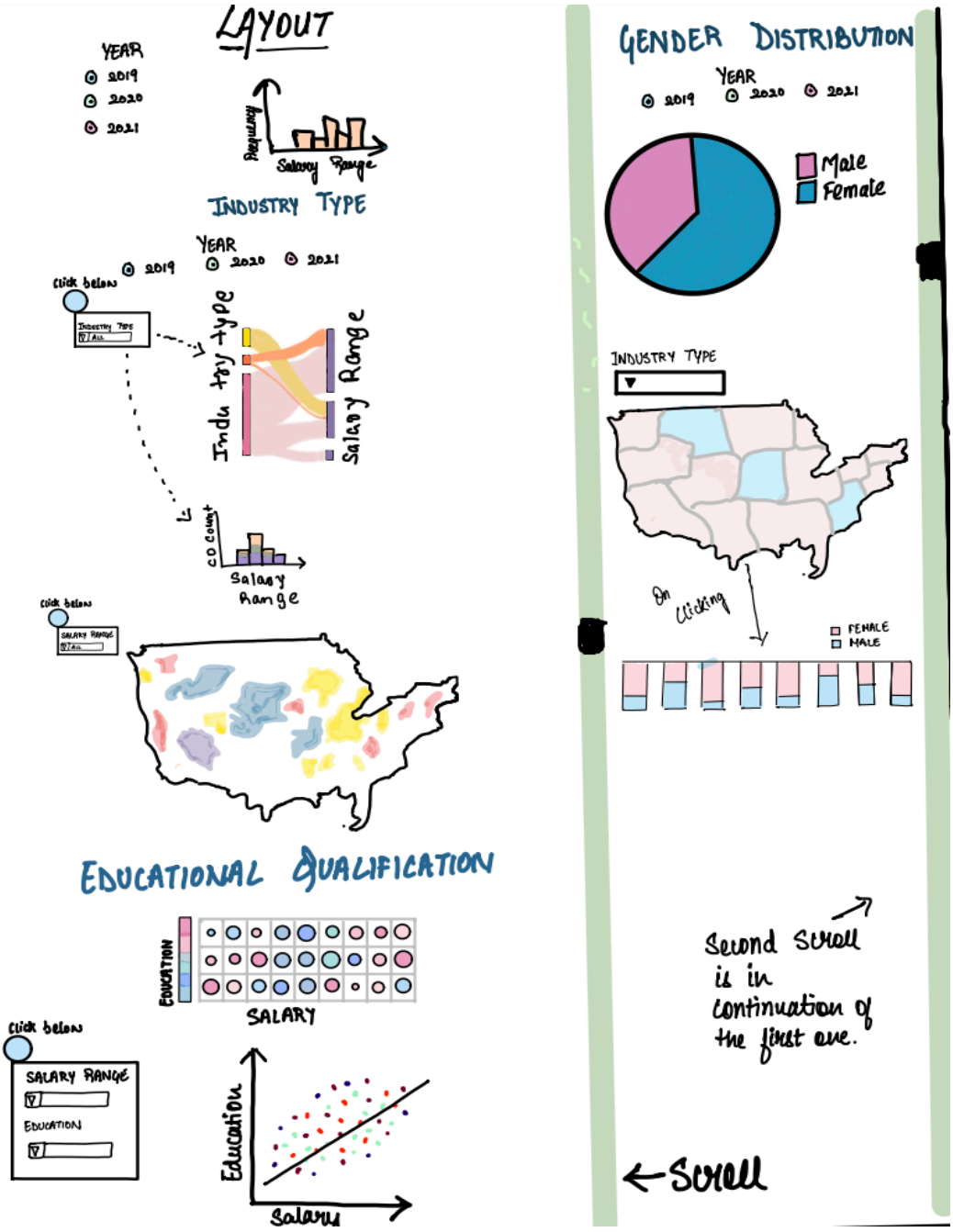


Figure 1.5 Final Design Sheet

I combined the best features from the three designs into the final design as shown in *figure 1.5*. I have implemented scroll down navigation basically a magazine style layout for the final design. Then there are three sub-heading to the narrative visualisation- Industry type, Educational Qualification and Gender Distribution with having individual year filter to refine the data as per selected year. In Industry type section, there is a side drop down icon having multiple choice industry filter which will update sankey diagram and stacked-bar plot. Legends will be updated as per the industry selected. Then there will be congressional district level chloropleth U.S map having a side drop down icon with salary range filter. Selecting salary bracket will update the map and legends. Moreover, hovering over to any of congressional district will share more information like dominant industry and salary range offered by the same industry in that congressional district. In Educational Qualification, there will heat map correlation matrix and a scatter plot with a side drop down icon having salary range and educational qualification filter to plot the scatter plot. Caption of the scatter plot will be changed accordingly as per the selection. In Gender distribution selection, there will be a pie chart in the beginning then I wanted to show the stacked-bar plot of gender distribution for working different industries for each state I decided to utilise a linked interaction chloropleth map instead of a dropdown with 52 filter because having some many features in a drop down is very user unfriendly and complex. Users can see the stacked bar plot displaying the gender distribution in each state's industries by clicking on that state. To enhance interactivity and gain more insights out of this, I introduced an industry dropdown filter on the map. Selecting a specific industry dynamically updates the

map, with states dominated by females turning pink and those dominated by males turning blue. This combination of features ensures a user-friendly.

The sole reason for choosing a blue theme for my final design narrative visualization was intentional, aiming to appeal to an adult target audience rather than children. Moreover, the association of the colour blue with feelings of calmness and tranquillity makes it well-suited for brands specializing in consultation, emotional therapy, or relaxation. This choice aligns with the fact that people, that in general, experience anxiety, uneasiness, and stress when they don't have a desirable income, driving them to explore options like a career change or pursuing higher education to enter the right industry. This narrative visualisation will bring sense of trust and calmness to the user. This decision was driven by my target audience, as described in the introduction.

Implementation

The important libraries imported to build the narrative visualisation are shown in *Figure 2.1*. Plotly was used to provide interactivity to the charts, and Shiny, Shinyjs, and ShinyWidgets were used to add animation and improve the user interface. Furthermore, the highcharter package was essential in helping to plot the U.S. map, and the ggsankey library made it easier to create the Sankey diagram. Robust data manipulation ability was made possible by the combination of tidyverse and dplyr, and corrplot played a crucial role in producing correlation matrix. This comprehensive set of libraries ensures a dynamic and informative narrative visualization.

```
## Library
library(shiny)
library(shinyjs)
library(dplyr)
library(highcharter)
library(stringr)
library(tidyverse)
library(ggsankey)
library(plotly)
library(shinyWidgets)
library(corrplot)
```

Figure 2.1 Imported library

Income Range Bar-Plot

I enhanced the look of the Income Range Bar plot filter by using a more sophisticated design, leveraging prettyradiobutton for an enhanced visual appeal from library shinyWidgets. The addition of subtle animations further elevates the overall layout aesthetics. To enhance interactivity with the bar-plot, Using the Plotly library to improve interactivity, I choose to use renderPlotly function from the library rather than the more common renderPlot. By just hovering over specific bars on the bar plot, this option makes count retrieval simple and straightforward. Finally, I used the ggplot and geom_bar functions in tandem to create and display the bar plot. Reactive function was used to update the data frame with the dynamic input by the user.
Pretty Radio button reference: https://rdrr.io/cran/shinyWidgets/man/prettyRadioButtons.html

Sankey Plot

The Sankey diagram was second most complex plot throughout the visualization process. While there are libraries and functions available for generating Sankey diagrams, I found their appearances to be overly minimalistic and challenging to interpret due to poor labelling and inadequate colour combinations. Additionally, incorporating interactivity using plotly was not feasible with these existing functions. After long search, I discovered ggsankey library which has the geom_sankey function, which integrates with ggplot for Sankey diagram creation. However, I faced some limitations with data flexibility, compelling me to adjust the file for improved compatibility with geom_sankey and I needed to

understand the source code of **make_long** function to transform the data frame accordingly. Furthermore, when introducing plotly to the ggplot, the hover values contained unnecessary information, necessitating datatype conversion to factor before passing it to the aes function. In addition, I have added a side drop down icon having multiple choice industry filter which will update sankey diagram and stacked-bar plot which is our next plot. Legends will be updated as per the industry selected. The shinyWidgets library's dropdown capability was heavily utilised to improve the overall design.

Reference:

Sankey Diagram: https://github.com/davidsjoberg/ggsankey

Drop down: https://dreamrs.github.io/shinyWidgets/reference/dropdown.html

Income Range Stacked Bar-Plot

I have used ggplot in integration with geom_bar to plot the stacked bar-plot. In order to plot it I have to pass stack in position argument as a parameter in the geom_bar function and also passing Industry in the fill variable as a argument in aes function to group the bar into different industries.

Congressional District level U.S.A Map

I worked with a large dataset to get the final data frame for this visualisation, which required wrangling six files, each with a sizable volume (~33k rows x 1k columns). Each file corresponded to a certain year, with three files pertaining to industry data and the remaining three connected to salary data. Since the source data was originally with granularity at the zip code level, there was a great deal of data wrangling and aggregating to the congressional district level. I chose PySpark because as I knew it was the best option for managing massive amounts of data, especially considering how big the dataset was.

Making a map at the congressional district level presented the greatest challenge out of all the visualisations. There was no easy way to automatically plot the U.S. map at the congressional district level displaying with the desired value using a built-in library. Although the HighCharter library came with a built-in function for making maps, but each time the function was used, the map had to be downloaded to the local system. Still, due to project limitations, every file had to be locally stored. However, as per instruction of the project that all files should remain locally present. To resolve this, I downloaded the non-tabular data (JS file) needed by the highchart () function. However, there was a complication while directly passing the JS data into the function directly as the hc_add_series_map function was not able to comprehend the data appropriately, to overcome this challenge.

I examined the source code of the inbuilt function, identified the syntax responsible for transforming the file's structure, and locally applied it to the downloaded JS file. This approach enabled me to prepare the data in a suitable format for the highchart function without the need for repeated external downloads.

I next looked through the JS file to find the key-value pair that Highcharts plots on the map. Upon more investigation, I found that the "from" variable is used to assign the relevant values to the congressional district key, which is designated as "hc-key" in the file. I carried out a join depending on the names of the congressional districts (cd_code) to incorporate this data into my data frame. But the first plot produced a heatmap, which wasn't appropriate for my values that were categorical. I made some adjustments by including hc_colorAxis and passing list of colours to resolve this, manually choosing the colours for the categorical values using stops. After finishing the layout changes, I included hover interaction, which lets users explore various congressional districts and hover over them to get more information.

U.S.A map using HighCharter reference: https://stackoverflow.com/questions/48898192/drilldown-united-states-city-county-map-from-states-using-highcharter

Heatmap Correlation matrix

It is usually simple to obtain the correlation matrix by using the cor function. But in these instances, I found that the population data for income and education needed to be normalised. As variations in the values for the same congressional district were found when populations from several attributes of the data frame were combined so normalising was of utmost significance. To fix this, I first normalised the education and income range data frames separately before joining them according to the relevant congressional district. I plotted the resulting heatmap after using the cor function from the library corrplot.

Although I have previously discussed the colour scheme selection, I discovered that although the heatmap is naturally helpful but not user-friendly, User may find it difficult to grasp the magnitude based just on hue intensity. I added circle shape to improve clarity by measuring the magnitude using circle radii. This enhancement makes it easier for users to discern magnitude right away, leading to a more intuitive comprehension of the correlation heatmap.

Corrplot reference: https://www.rdocumentation.org/packages/corrplot/versions/0.92

Scatter Plot

To incorporate two filter dropdowns—filters for Educational Qualification and Income Ranges—within a single tab, I utilised the shinyWidgets package to construct a dropdown function. I used the ggeom_point function and the ggplot2 library to visualise this data. To improve interactivity, I also included the plotly library.

One significant change from the final design is in the initial method, in which I just plotted points. Later, I added a regression line to the plot to improve the information presented. The lm (linear model) approach and the stat_smooth function was used to accomplish this.

Pie-chart

Integrating renderplotly with ggplot for pie chart is not compatible, I used plot_ly function to make the pie chart.

Reference: https://stackoverflow.com/questions/62021769/plotly-r-pie-chart-how-to-fixate-the-color-assignment-color-per-group

Linked Interaction State Level U.S. Map Populating Stacked Bar-Plot

I used the same method as previously described to create the map. But in the previous map, the data was displayed at the level of congressional districts, where in this plot my intention was to display it at the state level. To accomplish this, I examined the JS file and found the key "postal_code," which is in charge of graphing states. My data granularity was at the congressional district level, which presented a hurdle. I did some data wrangling and combined the dataset to the state level to get around issue. After obtaining the combined data, I went ahead and integrated link interaction. The hc_plotOptions method must be altered to produce an event each time a user clicked on a state.

Next, this event was recorded using the observeEvent function, which extracted the clicked state as input. This data frame for the stack bar plot was then filtered using this input. After the data was filtered, the ggplot function used it as input to create the stack bar plot. The change in the final design which I presented in the class is discussed in the final design section.

U.S.A map using HighCharter reference: https://stackoverflow.com/questions/48898192/drilldown-united-states-city-county-map-from-states-using-highcharter

User Guide

User can filter the data in *Figure 3.1* by selecting a year using the year radio button on the left side of the screen. Hovering over a certain bar can also provide information on U.S individual falling into certain salary range.

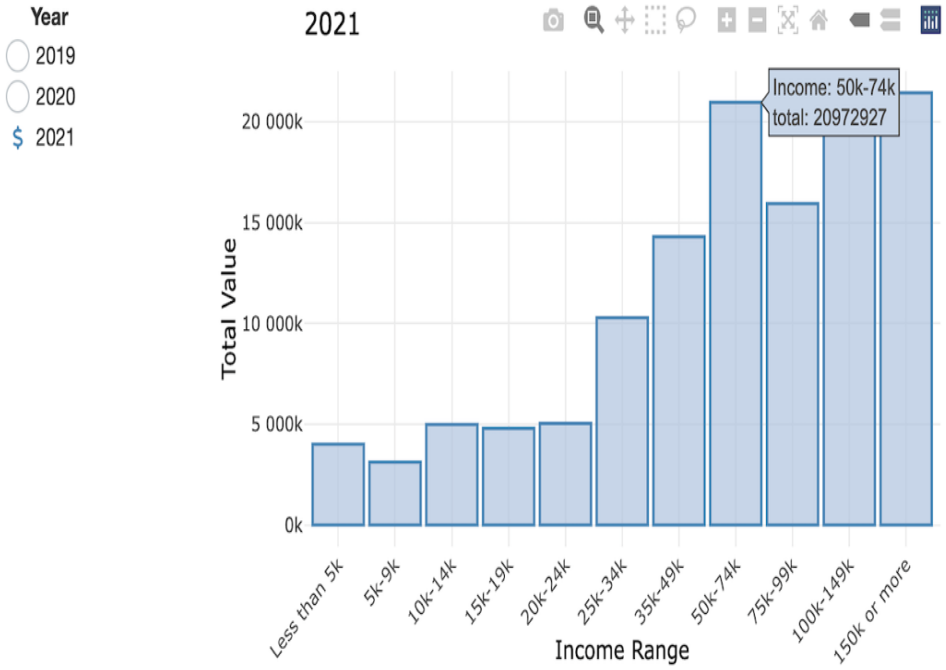


Figure 3.1 Salary Distribution

A year filter is incorporated into the industry type section, as shown in *Figure 3.2*, and it affects the three plots that are included in this section: the Sankey diagram, the bar plot, and the USA chloropleth map. Clicking allows users to change the year, which is set to 2020 default.

This section includes a side drop-down icon with the wording "Click below" above it, as shown in Figure 3.3. This drop-down populates the multiple-choice filter for industry type, which then updates the two plots shown in Figure 3.4.

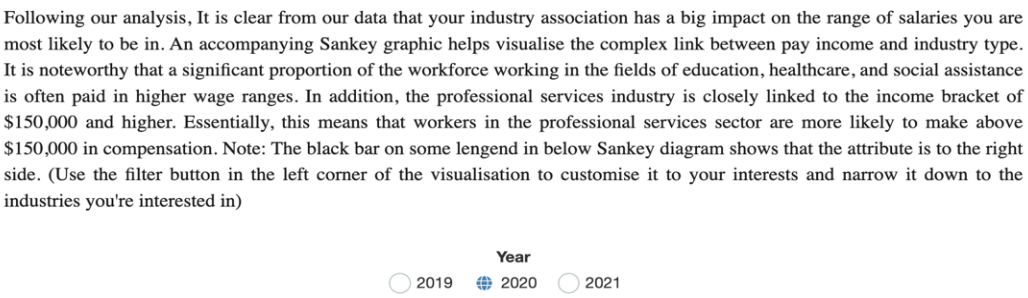


Figure 3.2 Industry Year Filter

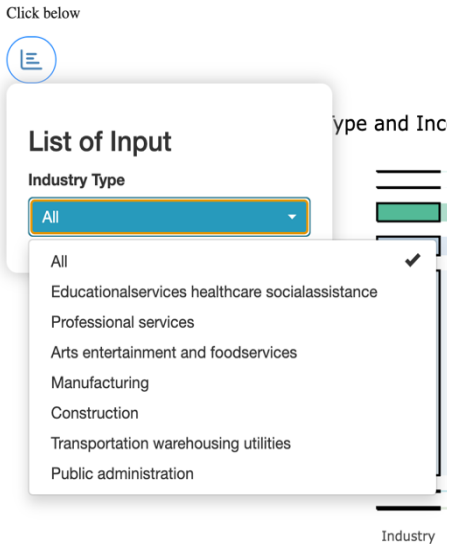


Figure 3.3 Side Drop Down Industry

Users must click on the side drop-down menu and then select their desired industry by navigating to the industry type drop-down menu. Legends will be updated as per the industry type selection.

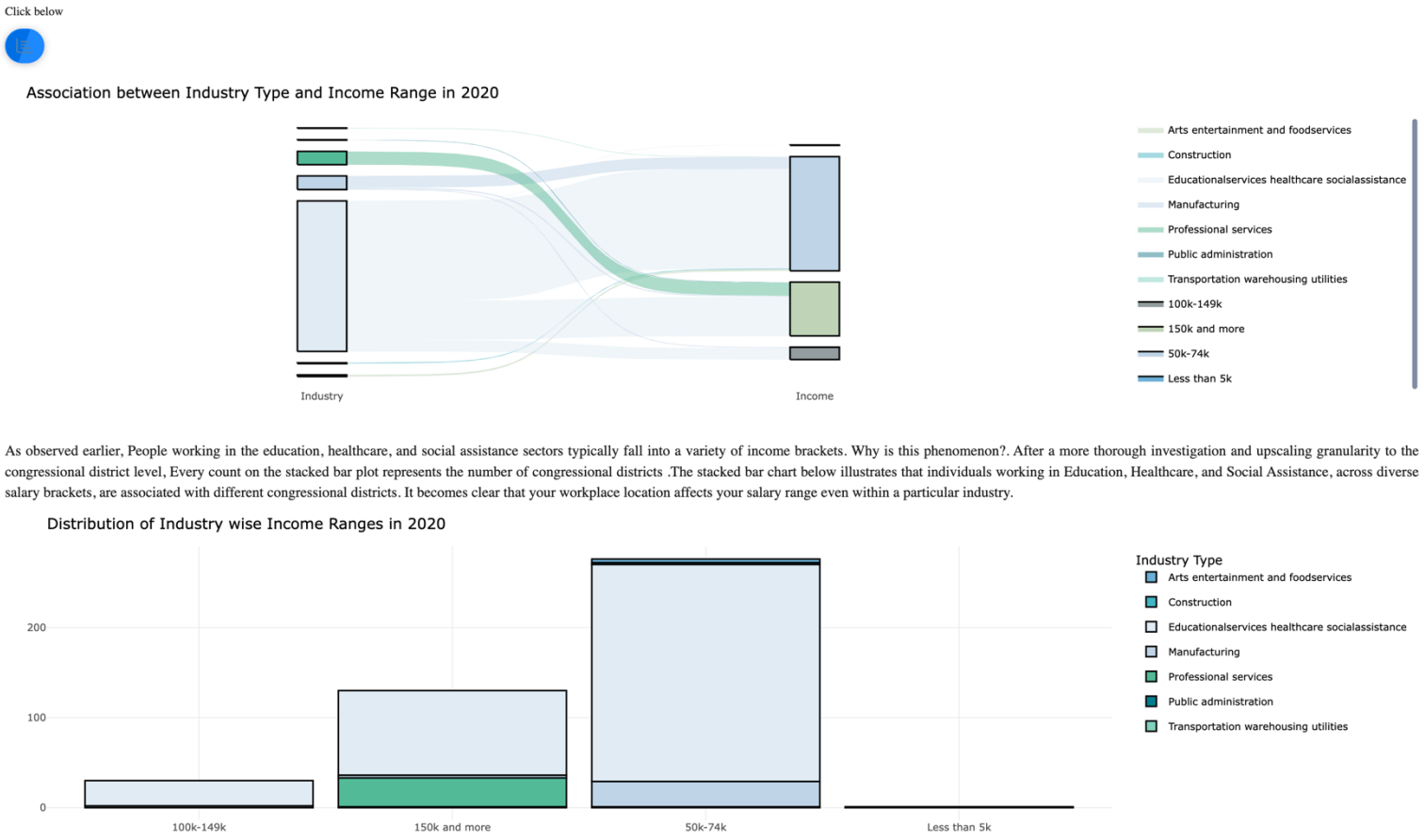


Figure 3.4 Drop down 1 updated Visual.

Hovering over various plots will disclose more information, as shown in Figure 3.5. Information about the node will be shown in the Sankey diagram, and the congressional district count will be shown in the stacked bar plot.

In the same way, if user click the sidebar drop-down button in the US map plot, a filter will show up, as seen in Figure 3.6. This salary range filter is a single choice drop down; thus, it essentially filters the map according to the range that is chosen. Hovering over a congressional state also reveals other information, like the leading industry in the state and the pay scale for that industry, as seen in the graphic below.

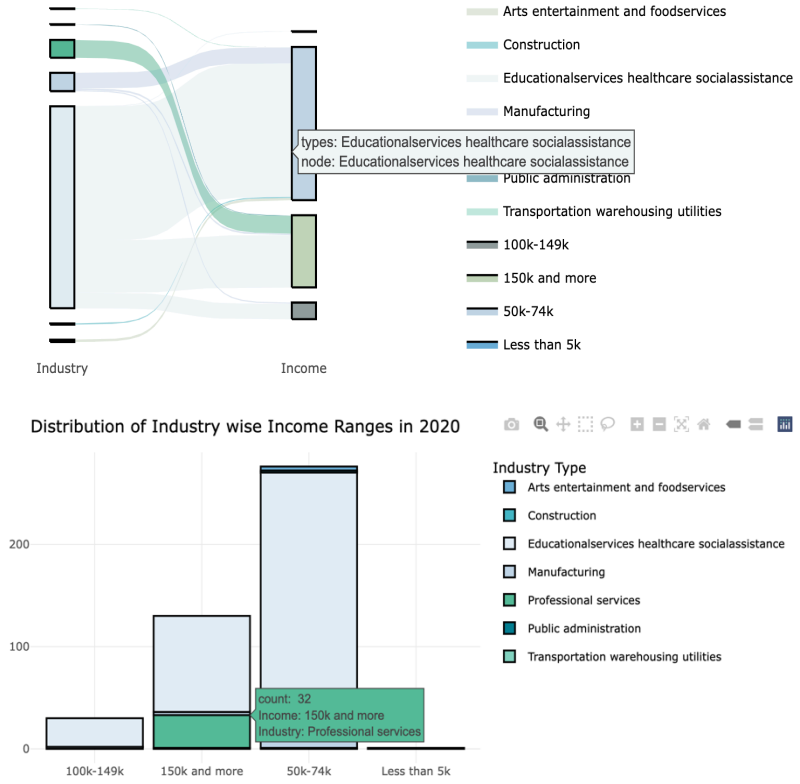


Figure 3.5 Hover feature

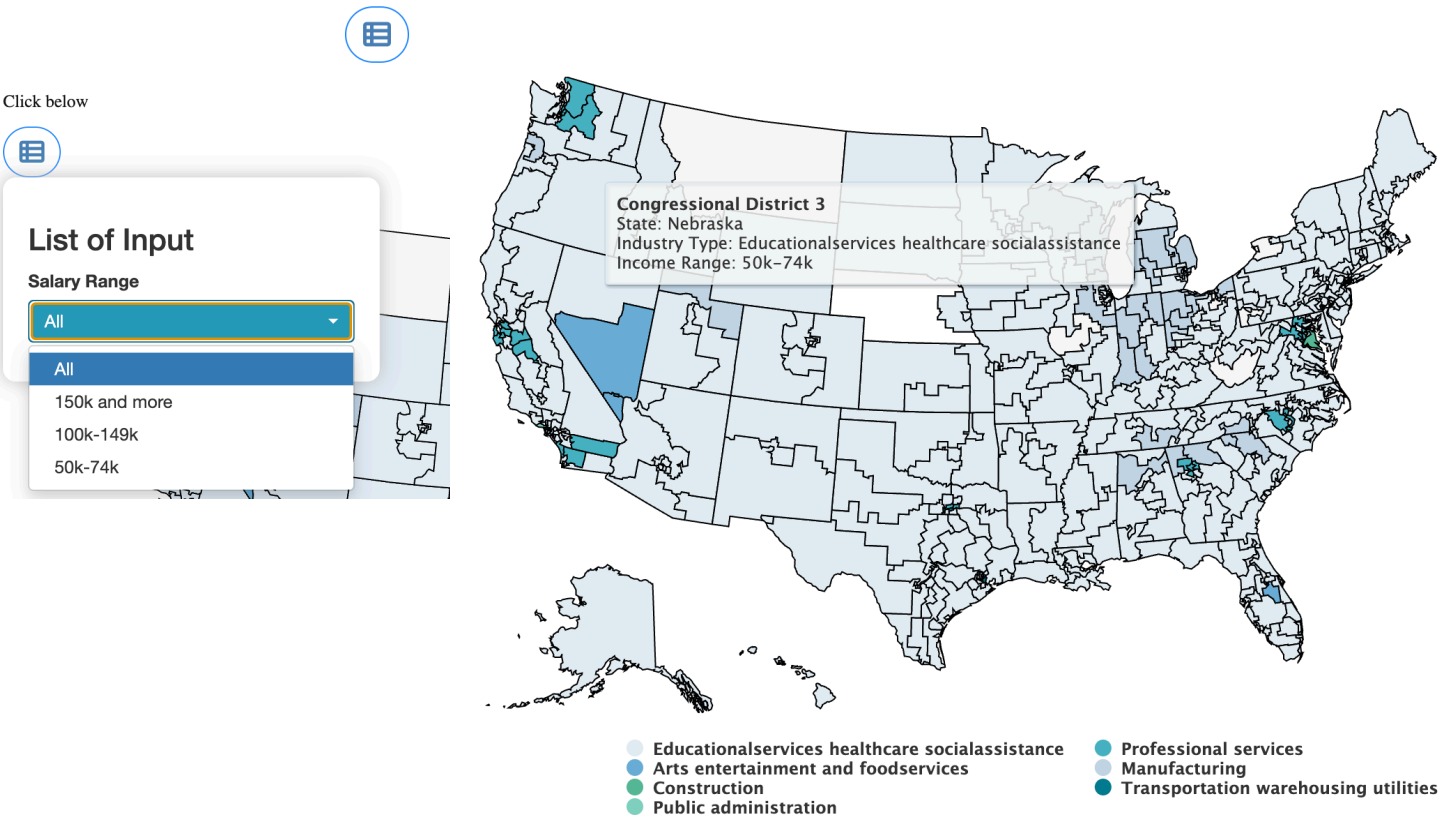


Figure 3.6 Congressional district level U.S. Map

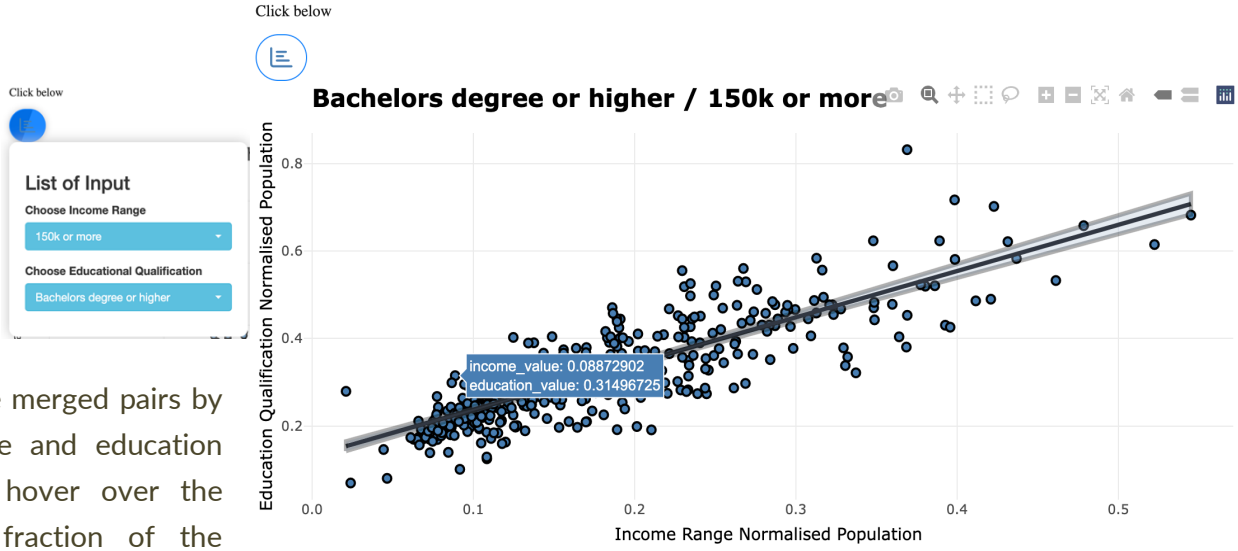


Figure 3.7 Scatter Plot

Comparably, there is a sidebar drop-down with two single-choice drop-down filters, as seen in Figure 3.7. The user can analyse the relationship between the merged pairs by choosing various income and education ranges. Additionally, a hover over the normalised population fraction of the chosen choices will show information. Heading the plot will be changed automatically based on the selected combination.

The gender distribution section has a year filter (see Figure 3.8), much like the industry section does. All the charts under gender distribution section will be modified by this filter according to the chosen year. Furthermore, if user mouse over the pie chart that shows the gender distribution of working U.S individuals, User will see the percentage as well as the overall count.

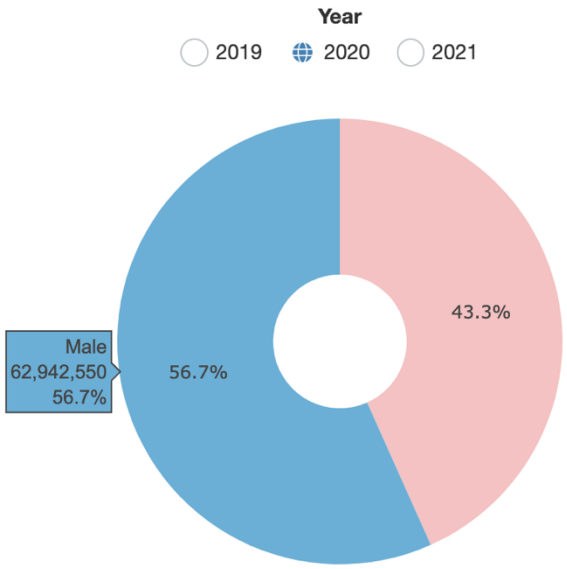


Figure 3.8 Gender Distribution pie-chart

Users can choose an industry in Figure 3.9 and see which gender is more prevalent in that industry in each state.

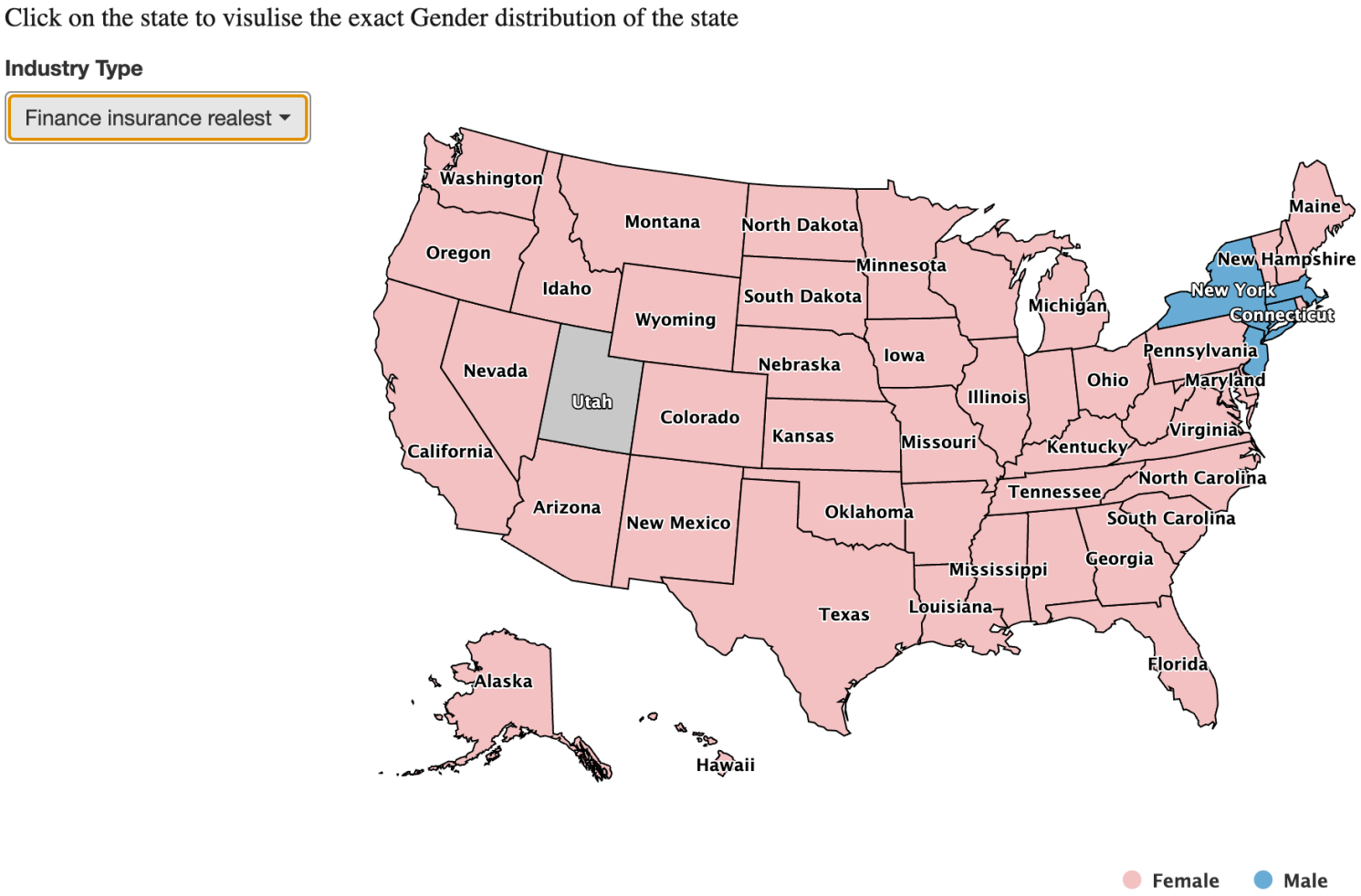


Figure 3.9 Gender Distribution U.S. state map

I've added a link interaction as an extra feature. When a state is clicked on the map, a stacked bar plot displaying the gender distribution of working U.S. individuals in different industries will appear for that state. This feature is like the example shown in figure 3.10, where choosing Utah causes the matching stack bar plot to appear.

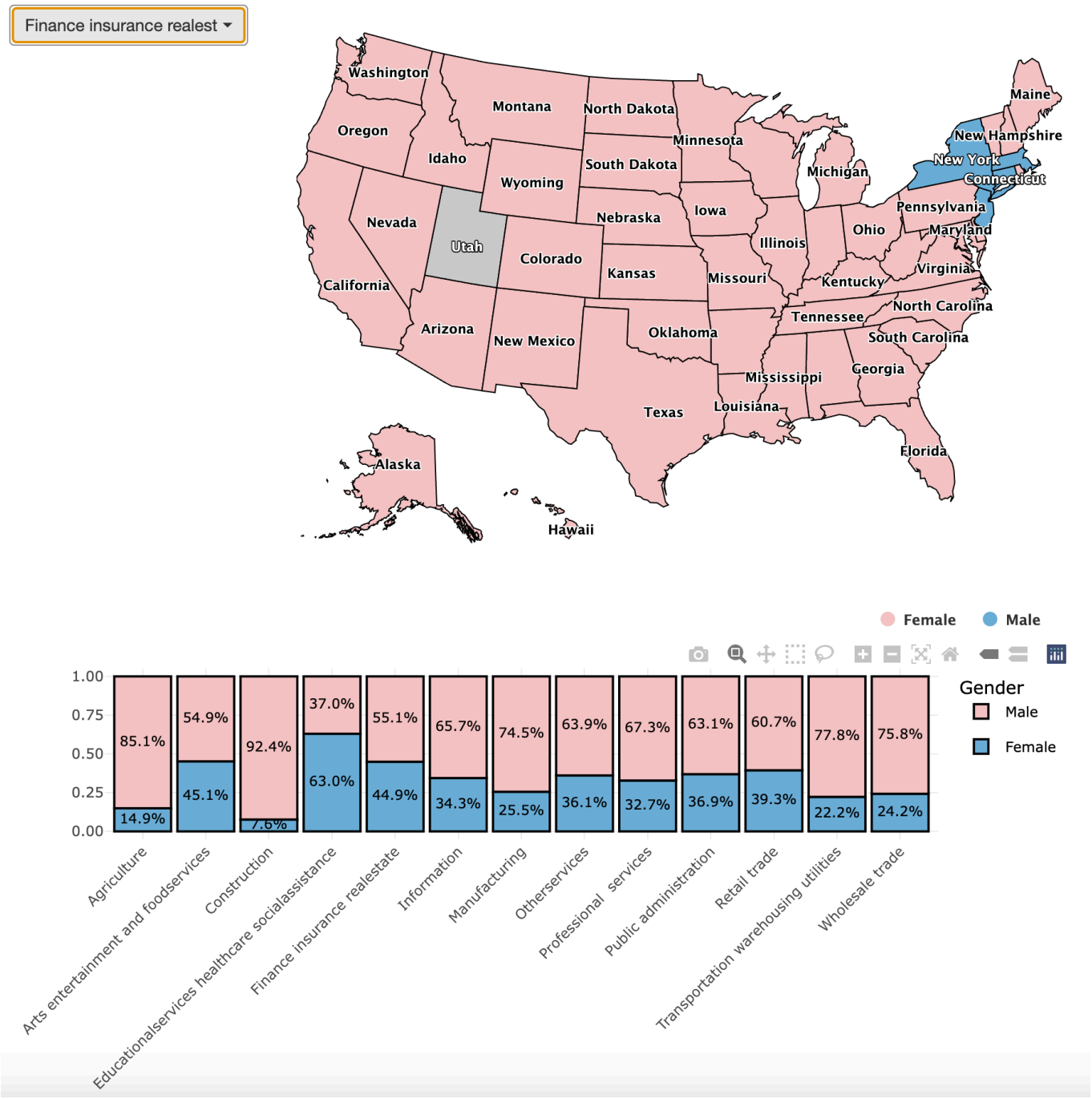


Figure 3.10 Link Interaction

Conclusion

We can state with confidence, based on the detailed visualisations, that choosing the appropriate industry and obtaining better education have a substantial association with reaching greater pay range. But it's not just about breaking into a field; it's also important to work at the best possible location. The difficulty is not just getting into the field but also overcoming gender domination, which might make it harder to break into it. Therefore, to guarantee a greater income,

it is necessary to emphasise the location's equal significance at the same time, Gender also becomes an important consideration for certain businesses.

It is noteworthy that a significant segment of the labour force working in social services, healthcare, and education typically earns higher salaries. In the same way, the $150,000 and higher salary range is closely related to the professional services sector. This suggests that people who deliver professional services have a higher likelihood of making more than $150,000 in pay. Education, healthcare, and social assistance associated with a range of salary brackets, in contrast to professional services.
Looking more closely reveals that the place of employment has a significant impact on the range of salaries. For instance, we can evaluate various congressional districts by focusing on same industry which social assistance, healthcare, and education. In this industry, the level 2 congressional district in Oregon pays between $50,000 and $74,000, whereas the level 4 congressional district in California pays more than $150,000.

Also, we need to understand, educational qualification plays a significant role too. There is evidence of a correlation between education level and income type in the correlation matrix. As one's education increases, especially after having a bachelor's degree or more, the correlation with higher income groups grows This suggests that there is a positive correlation between income and education level. For instance, there is an almost perfect positive correlation between the income range of $150,000 or more and the term "bachelor's degree or higher" Furthermore, obtaining a "bachelor's degree or higher" is negatively correlated with falling into a lower income category, indicating that individuals are less likely to do so after completing their higher education. Conversely, there is a clear positive association between less education and lower income levels. For example, there is a substantial correlation between earning between $15,000 and $74,999 and having a high school diploma. Furthermore, there is a marginally correlation between those earning between $5,000 and $35,000 and those without a high school education.

As previously mentioned, finding the industry, and having a good educational qualification that offers the best salaries is only the first step. A closer look at the data reveals an intriguing feature: the distribution of genders across different industry. The map was illustrating a clear distinction, the information sector tends towards male domination whereas industries like educational services, healthcare, and social assistance are controlled by women. This finding is significant because it implies that there may be distinct opportunities when you join an industry where the gender distribution is like your own. The pie chart below shows the gender distribution of U.S. working individuals.

Limitation and Further Improvements

One significant limitation of the project was the dependence on Cramer's V, as explained in the exploratory project, to comprehend the correlation between salaries and major industries. The requirement to concentrate on a single industry per congressional district led to this restriction. Having access to a dataset with real industry-specific compensation numbers would have allowed for more thorough analysis.
In terms of further improvements, the idea is to use the correlation matrix and link interaction for future upgrades. Rather than having to choose their level of education and income range, users could just click on any correlation matrix cell to populate a scatter plot. As of right now, no interactive library that works with corrplot has been found, however more research is being done to find possible solutions.

Bibliography

References

1. *Fiverr - Freelance services marketplace for businesses*. (n.d.). Your Access To This Website Has Been Blocked. https://www.fiverr.com/resources/guides/graphic-design/color-meanings#5-blue

2. *prettyRadioButtons: Pretty radio Buttons Input Control in shinyWidgets: Custom Inputs Widgets for Shiny*. (n.d.). Rdrr.io. Retrieved November 4, 2023, from https://rdrr.io/cran/shinyWidgets/man/prettyRadioButtons.html

3. Sjoberg, D. (2023, November 2). *ggsankey*. GitHub. https://github.com/davidsjoberg/ggsankey

4. *Dropdown — dropdown*. (n.d.). Dreamrs.github.io. Retrieved November 4, 2023, from https://dreamrs.github.io/shinyWidgets/reference/dropdown.html

5. *Drilldown United states city/county map from states using highcharter*. (n.d.). Stack Overflow. https://stackoverflow.com/questions/48898192/drilldown-united-states-city-county-map-from-states-using-highcharter

6. *corrplot package - RDocumentation*. (n.d.). Www.rdocumentation.org. https://www.rdocumentation.org/packages/corrplot/versions/0.92

7. *Plotly (R) - Pie chart: How to fixate the color assignment color per group?* (n.d.). Stack Overflow. Retrieved November 4, 2023, from https://stackoverflow.com/questions/62021769/plotly-r-pie-chart-how-to-fixate-the-color-assignment-color-per-group

Data Sources

A. Industry Population Data - Tabular Data

Industry file 1: rows 33120 X columns 542, Year: 2019

Industry file 2: rows 33120 X columns 542, Year: 2020

Industry file 3: rows 33120 X columns 542, Year: 2021

Source: https://data.census.gov/table?q=industry&g=010XX00US$8600000&kd=ACSST5Y2021.S2404

B. Finance Population Data - Tabular Data

Finance file 1: rows 33120 X columns 1106, Year: 2019

Finance file 2: rows 33120 X columns 1106, Year: 2020

Finance file 3: rows 33120 X columns 1106, Year: 2021

Source: https://data.census.gov/table?q=Financial+Characteris<cs&g=010XX00US$8600000&<d=ACSST5Y202 1.S2503

C. Education Population Data - Tabular Data

Education file 1: rows 33120 X columns 1106, Year: 2020

Education file 2: rows 33120 X columns 1106, Year: 2021

Source: https://data.census.gov/table?q=Educakonal+Aeainment&g=010XX00US$8600000&kd=ACSST5Y2021.S1501

D. Zip code - Tabular Data

Zip code file 1: rows 41924 X columns 4

Zip code file 2: rows 42735 X columns 14

Source file 1: https://github.com/OpenSourceActivismTech/us-zipcodes-congress/blob/master/zccd.csv

Source file 2: https://www.unitedstateszipcodes.org/zip-code-database/

E. U.S. MAP Data - Non - Tabular Data (JS File)

Source: https://code.highcharts.com/mapdata/countries/us/us-all.js

Appendix

COMBINE AND REFINE

a) Chart 1, 2 and 3 can be used to in the same sequence to pull the information.

b) 4 and 5 chart can be used together.

c) 9 and 10 can be used together to pull some insight.

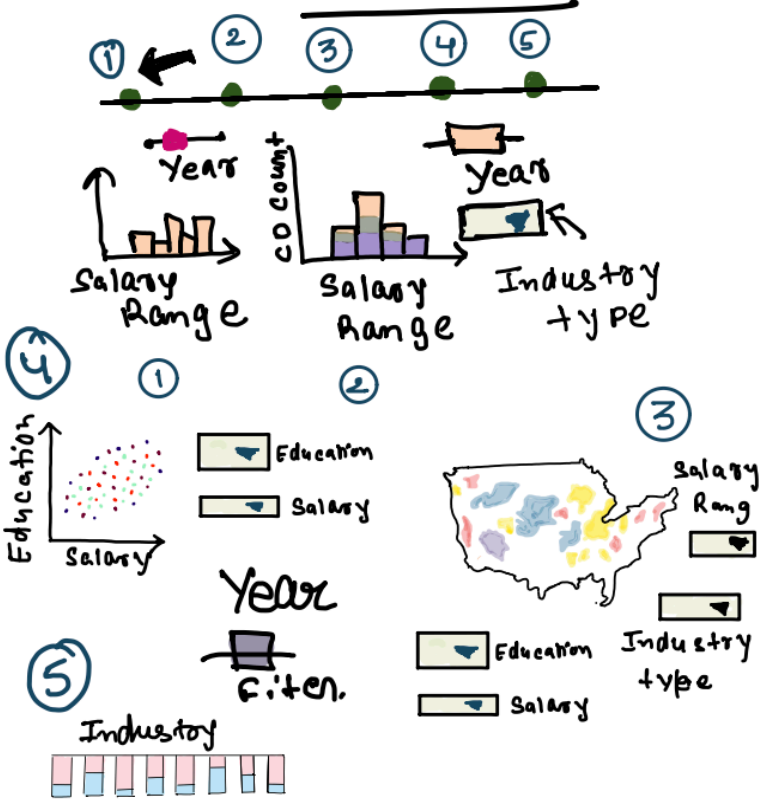
QUESTIONS.

Q1> Does a specific industry type have influence on a particular income range in the USA? (1), (2), (3)

Q2> Is a higher bracket of income driven by a specific level of education? (4), (5)

Q3) What is the gender distribution between industries and their evolution over time? (9), (10)

LAYOUT



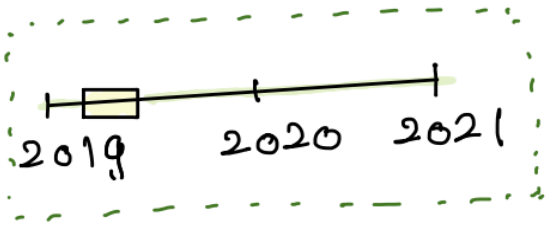
TITLE: IMPACT OF Industry and Education on Salary
AUTHOR: VINAY MITTAL
DATE: 8|OCT|2023
SHEET: 2
TASK: INITIAL DESIGN

Operation.

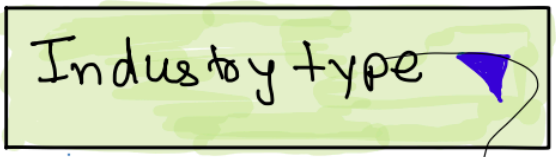
(1) It has main bar at the top which helps the the user to navigative. through different visualisation.

(2) Has various Dropdown and slider filters.

Focus / ZOOM



YEAR SLIDER



Similarly for education and salary Range.

Discussion

+ Detail presentation.

— Lack high level overview between education and salary range. , also for male and Female Distribution.

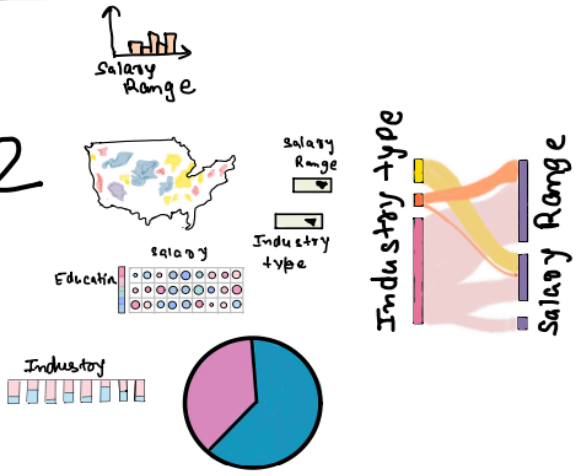
LAYOUT

TAB1 | TAB2 | TAB3

TAB1

TAB2

TAB3



TITLE: IMPACT OF Industry and Education on Salary
AUTHOR: VINAY MITTAL
DATE: 8|OCT|2023
SHEET: 3
TASK: INITIAL DESIGN

Operation.

* Navigation Between visualisation. through Buttons.
* Has various filter Button.

Focus/ZOOM

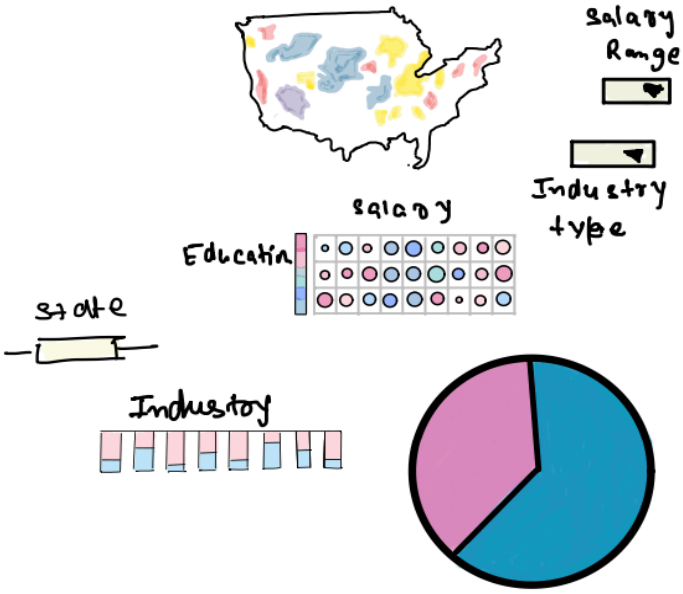
Industry type

Similarly for education and salary Range.

Discussion

+ Has snankey. Plot
− Lacks time series
− lacks coherence and story telling. Looks like a dashboard.

LAYOUT



TITLE: IMPACT OF Industry and Education on Salary
AUTHOR: VINAY MITTAL
DATE: 8|OCT|2023
SHEET: 4
TASK: INITIAL DESIGN

OPERATION

* TO navigate through different visualisations.
* map visualisation has two filter
→ Drop down filter
* Salary Range
* Industry type.

Focus | Zoom.

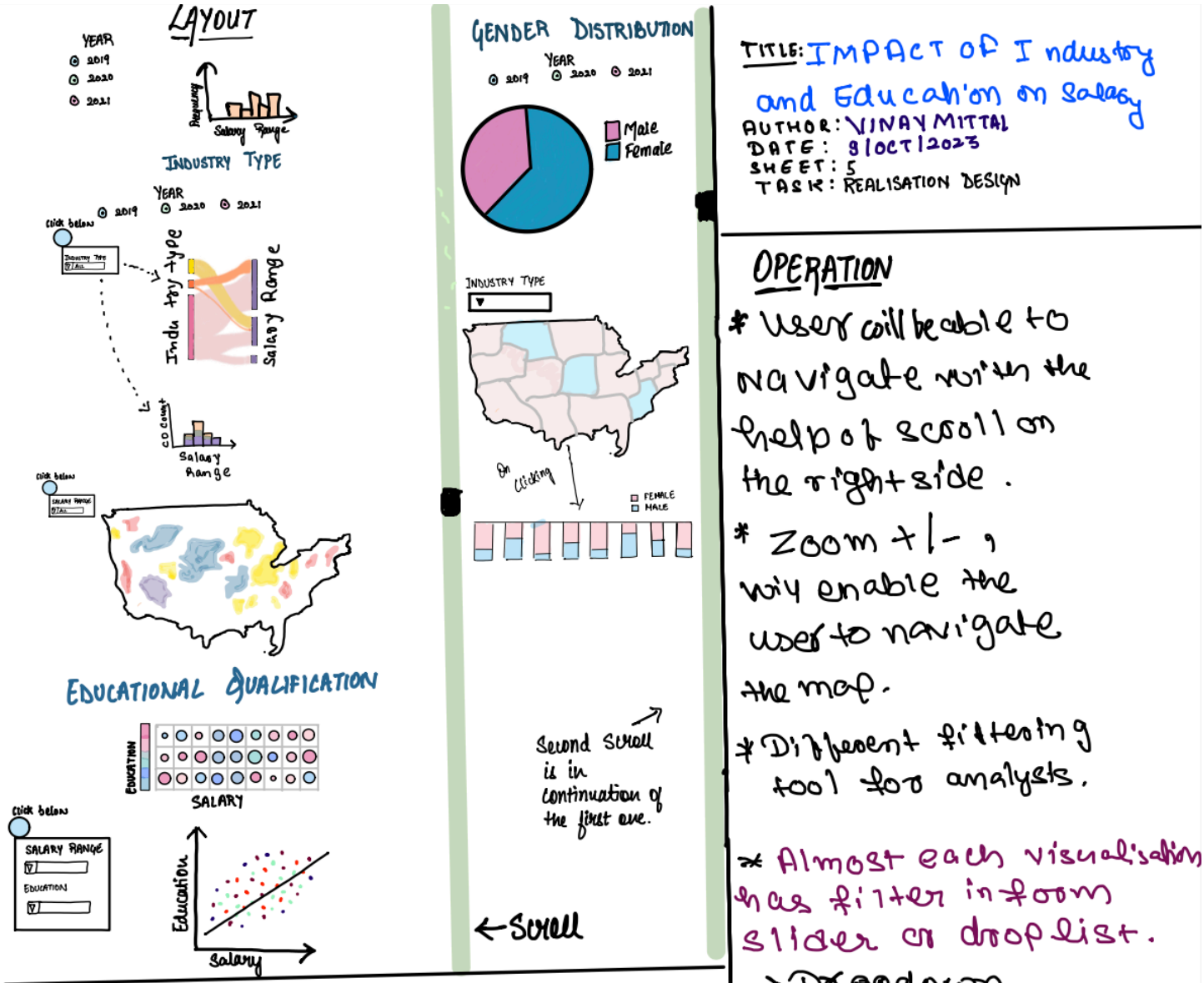
Industry type

Similarly for Salary Range.

Discussion

\+ Easy to navigate and follow coherence

— Correlation matrix gives a top view; but less detailed Information



TITLE: IMPACT OF Industry and Education on Salary

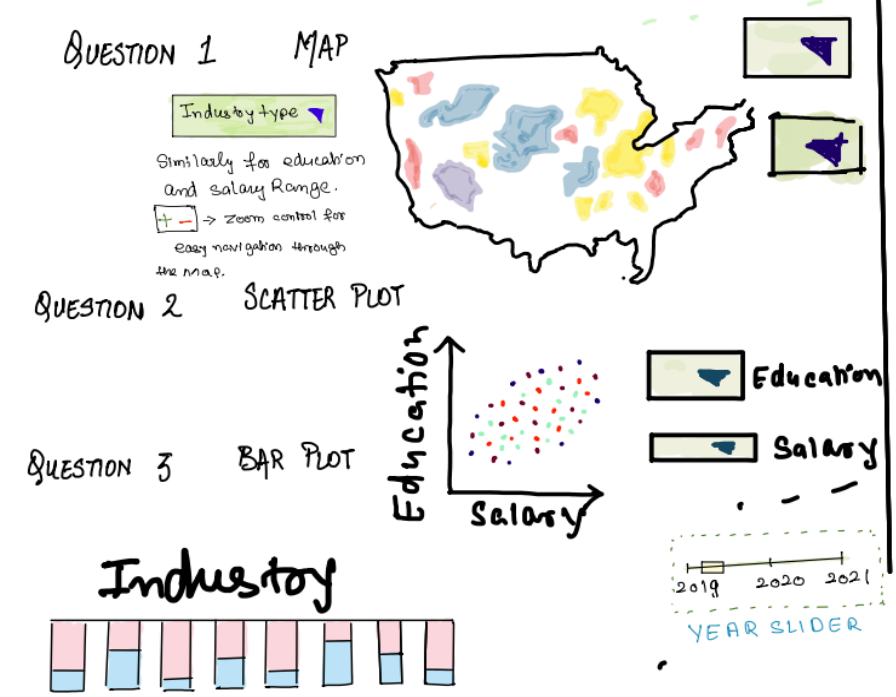
AUTHOR: VINAY MITTAL
DATE: 9 OCT 2023
SHEET: 5
TASK: REALISATION DESIGN

OPERATION

* User will be able to navigate with the help of scroll on the right side.
* Zoom +/- , will enable the user to navigate the map.
* Different filtering tool for analysts.

* Almost each visualisation has filter in form slider or droplist.
→ Dropdown
→ Slider

FOCUS / ZOOM



DISCUSSION

\+ Very easy navigation and have a coherance, each visualization connecting with a story.

— Excessively detail oriented.