Cheng Ho Wang 32353391

Project Title: Data Scientists' Salary

Introduction

The aim of this project is to give information through visualization for Master of Data Science student to make better decisions for their career after graduating. Based on the factors that affect data scientists' salaries, students can decide which company is more suitable for them, which country they are going to stay in and what skills they could improve in order to get a better salary. I hope this project can help data science students to decide their future career paths.

The dataset is originating from the Glassdoor website. With the given information, I have a few aspects on how to give precise and concise information to data science students to make their decision. In general, I have investigated the factors in three different aspects which are countries, company, and personal skills. As a result, I could interpret why the factors are affecting data scientists' salary. Hence, data science students can focus on developing their skills and decide based on these factors.

Design

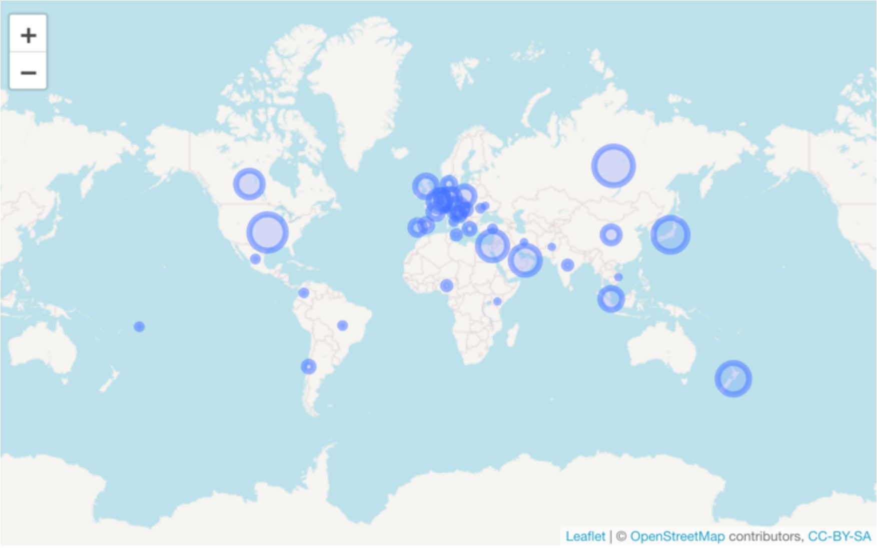
In terms of the design, I would like to go through the design of the layout first. Then, I will dive into the design of each graph.

In the five design sheets, I was trying to figure out which layout is better for my audience. Therefore, I thought of few design layout which maybe easier for audience to understand and more user friendly.

For the first page of the five sheet design, I was trying to summarize the graphs that I have done during the previous project. Then, I was thinking about how to organize the layout of the website and which graphs are suitable for me to do an interactive chart. Since the project is done based on 3 different factors, including company, personal and country. In each factor, I had explored the number of graphs they have and list out their type of graph. Then, I started to do the second to forth page, which are the layout design. For the first layout, I tried to group the charts by factors. Since the company factor includes most of the charts, I placed it at the top of the design. Then, I have got the personal factor and country factor below. This is easier for the user to understand how to read the graphs and the reasoning behind this project. Also, this is the most apparent way to shows the factors to the audience.

For the second layout, I tried to group the type of plot all together. In my project, there are plots including map, box plot, bar charts and scattered plots. Since I know it takes time for reader to process different type of graphs, this is the way for them to look at it without switch their processing mode in their brain. In the third initial design, I used an eye-catching approach. First of all, placed all the interactive charts on the top to attract users' attention. These leads to my final decision which group all the idea together. The final layout follows the eye-catching idea which placed the interactive charts at the top. Then, organized others by the factors. On one hand, it can draw users' attention. On the other hand, the flow of reading the report is logical. The view of the factors is also from a more general global view and dive deeper into company and personal view, which is a more in-depth factor.

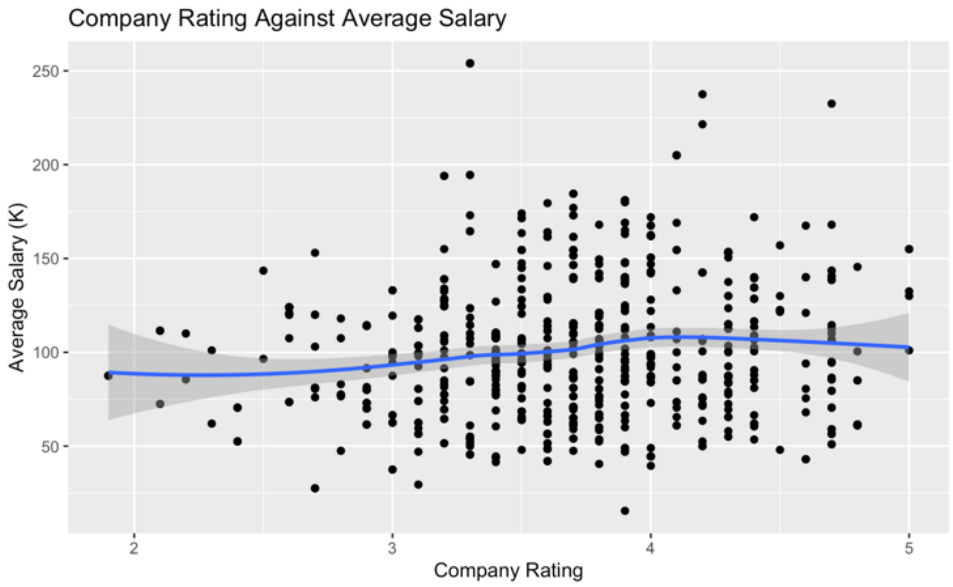
Then, I can start to introduce you each of the graph and the reason behind of choosing the graph. First of all, we can start with the map. I plotted the average salary of each country on the map by using leaflet. Readers can link the salary to country easily. In general view, they can have the idea of location at first glance instead of looking at the table which just stating “US”. The radius of the circle is proportional to the average salary of the data scientists. The bigger the circle, the higher salary of the data scientists. This visual cue give reader a general idea that which country has a relatively higher salary for data scientist. Another good thing is we can also compare the average salary of different country at one glance. After painting a big picture for audiences, we can also add a tooltip for each of the circle. Adding a tooltip on each circle can let readers to get the detail information easily. From general view to in-depth view, we can all summarized in a graph in this case.



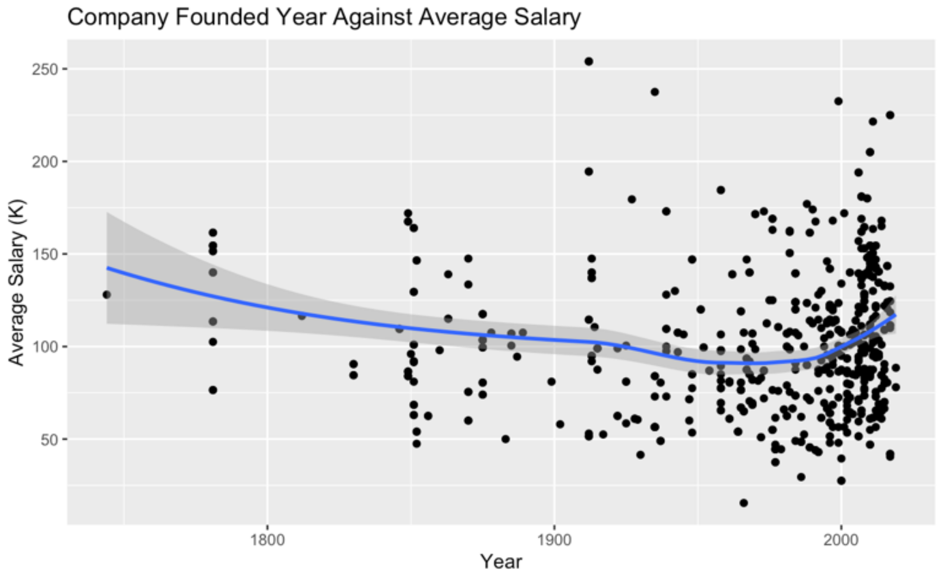
In the following, we can explore different graphs which are related to company’s factors.

First of all, this is the company’s rating versus average salary. I chose to use scattered plot because it fits the data type. In x-axis, it is the company’s rating which is a continuous variable. Therefore, I cannot group them together. The only way to do it is let both x and y axis to be continuous and each plot represent each data. An advantage here is that we can also know how the data disperse. The main point of using scattered plot is for users to spot any cluster point easily. Cluster points represent the majority of the data. So that readers can estimate their expected salary based on those cluster. Assume they are being a part of the majority. Also, we can easily spot the outlier of extremely high or low salary. Then, audiences can also interpret the reason of having those outliers.

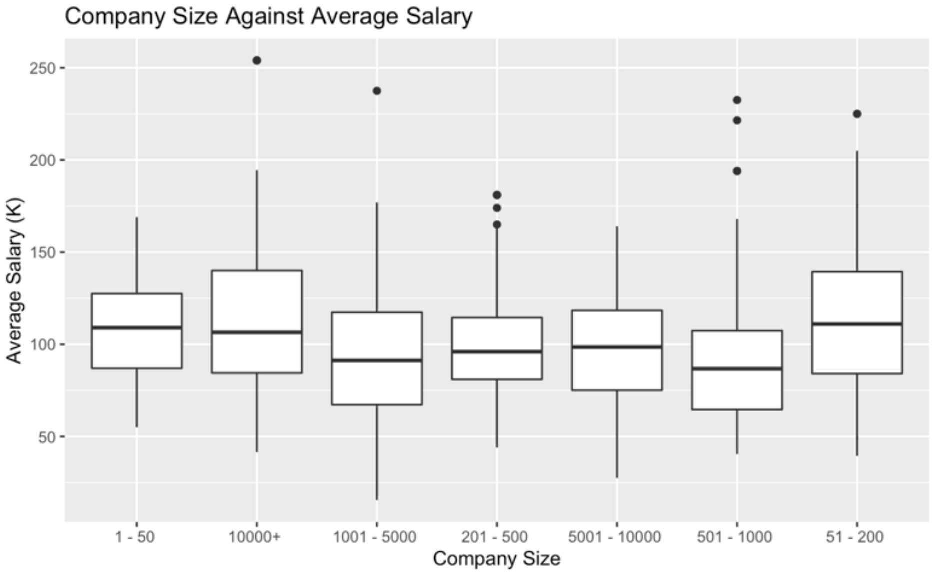
In this case, the data cluster at around rating = 3-4 and average salary = 100K. Also, it is easy for us to spot any outlier. Furthermore, I added a line on the scatter plot which shows a average salary on each company’s rating. From the line graph, we can see the trend in order to find the correlation between rating and salary. Also, we can find is there any peak and trough so that we can summarize the average salary is higher or lower in particular rating of the company.



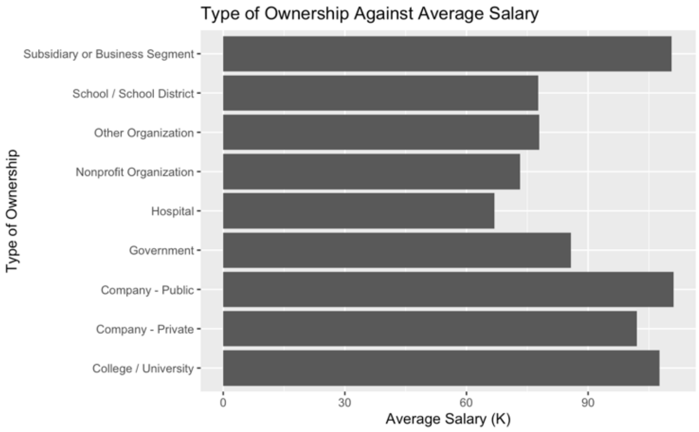
The following graph also follow the same reason of using scatter plot. Year is a discrete variable, but the range of year is too large. It will make the graph so complicated when it is grouped together. Therefore, I used scattered plot to summarize the data. Also, I fit a line into the data and see if there is any trend and correlation between two variable. Here, we can get a relatively obvious decrease at around year of 1930 to 1980.

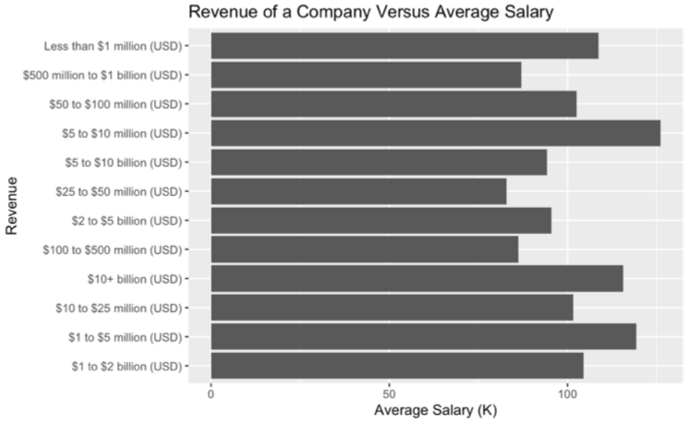


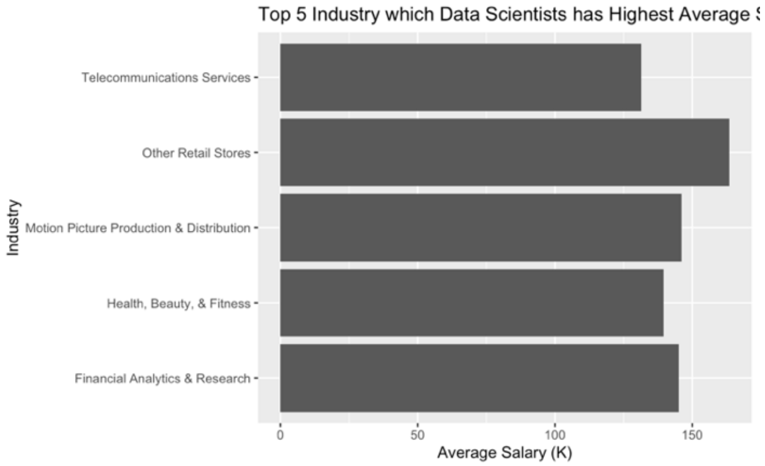
In the following plot, this is the only box plot for my project, and I think it is also an interesting plot. This is a plot for company size versus salary. From the original data, they have already grouped the company size for me. Therefore, the x axis could be using different group. Besides, the reason I chose box plot is that it could show the median, first quartile, and third quartile to us. Even the highest and the lowest point are shown on the graph. Also, the box plots are using the same y axis so that we can compare the statistics easily. The outliers of the data are also shown in the graph as well. We can get a lot of insight from these statistics. And we can compare the statistics between groups in one graph.

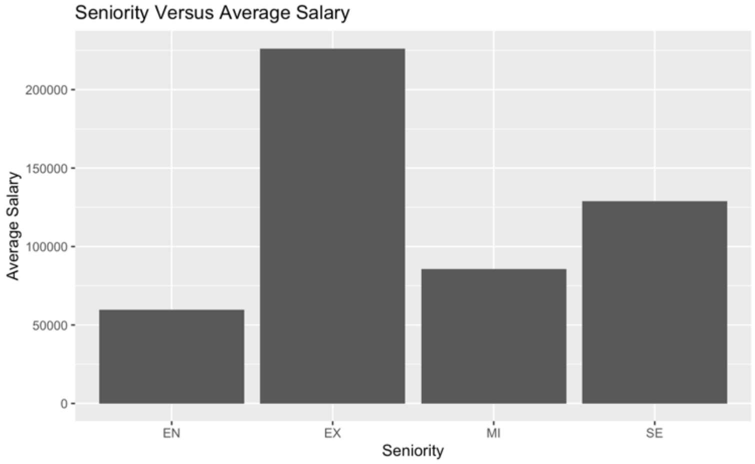


In the following, I chose to use bar chart to summarize the data. This is the easiest way to compare the value between groups. Also, the values on y-axis are already grouped by the data before. Bar chart is good for this kind of categorical variable. The only special thing is that I have placed the value of salary on x-axis and the group on y-axis. It is because the words cluster all together when they are on x-axis. Switching the position of two of them will be clearer for audience to read. This decision applies to these four graphs below.









Implementation

By doing the whole report, I used shiny to develop the web app. For the libraries that plotting the graph, I used ggplot2 for charts and leaflet to draw the map. I have done the data wrangling and plotting graphs on the previous data exploration project. The main task here is to combine everything together in a web app and design the layout for user to read my project.

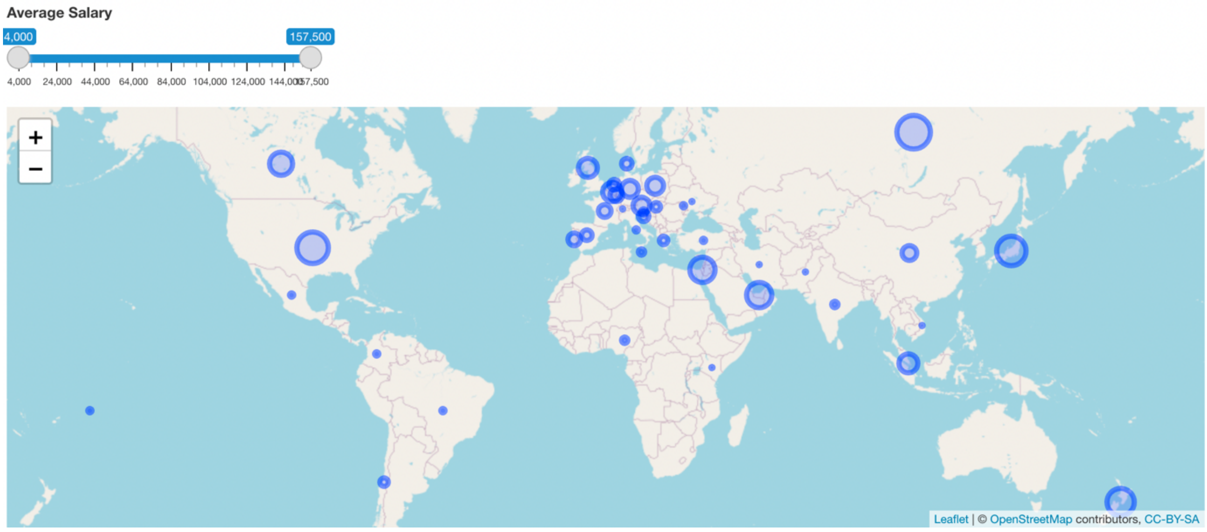
For the general layout, I have placed the title and aim at the top of the web and then a navigation bar below it. Therefore, users can go to different pages to explore the graphs. This is a big difference from what I have done in my design. In the five sheet design, I tried to pack everything into a page and divide them into different part. However, I don’t think it is user friendly when I implement it. The web page will be fully packed with information which is too complicated for user to process the information. Therefore, I found way to get more subpages in a single web. The code is reference to the R documentation. The navigation bar is quite suitable for my project because I grouped the graphs by different factors. The navigation bar can let the audience to spot the factors when it is placed at the top. Besides, the information is organized. In the company’s factors button, I have also added a menu which users can explore different graphs in it. This is also another different from my original design. When I tried to fit all the graphs into a page, the graphs are hard for me to read. Also, I have no idea how the information is organized. Another thing is that the layout is so crowded that is hard for me to add interaction on some of the graphs. That is why I added a menu for user to select different graphs in the company’s factor. In general, this is the whole process for selecting the layout during implementation.

By coding out the web, I chose to use shiny and divide into two document. One is ui and one is server. The ui file contains code for me to manage the layout and different features that the user can see. For example, the slide bar to filter out some of the data. In my project, I am using fixed page for the layout since it is easier for me to organize the page. The extra thing that I have to do is to manage the row and column by hard coding. The special ui feature that I have used is the slid bar for filtering the data. There are two buttons on one slide bar, one is for selecting the minimum value, and one is for selecting the maximum value that user want to see. By default, the website should appear the slide bar with the range of the value which match the range of the data. Therefore, in the coding, I am getting the minimum value and the maximum value of the data in specific column, which is the salary column in addcountry data in this case. Also, I have to set the step of the slidebar which fits with the range of the data.

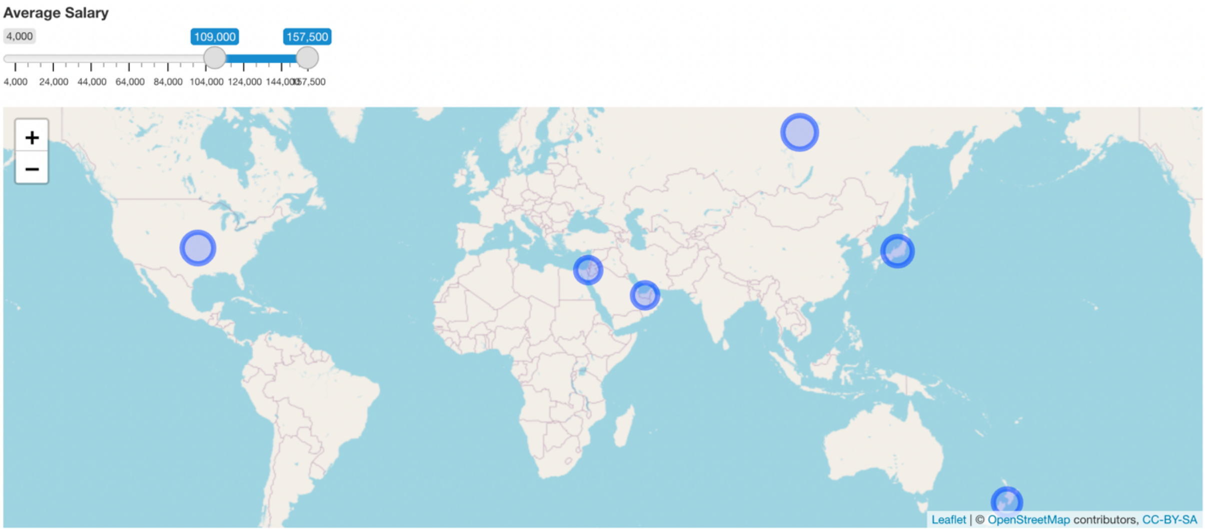
Then, we can have a look at the implementation of the serve file. This is the backend for the website. It contains information of who to plot the graphs and how to process the data from the instruction of the slide bar. First of all, I can explain how the server react when they received the instruction from the slide bar. The slide bar can get the value from user and then they will filter out the data by using a conditional statement, such is getting data which is larger than or smaller than some value. We have to store this into a variable. After that, in the process of generating the plot under renderplot or render leaflet, we have to plug in the variable into the data item so that R can generate the plot based on the filtered data. For generating the plot and the map, I simply used renderplot and render leaflet function. Then, store it into a variable. After that, we can have reference to the ui file and put the variable at the place that we want to place the graph. In general, the server file is not much difference from my original design.

User guide

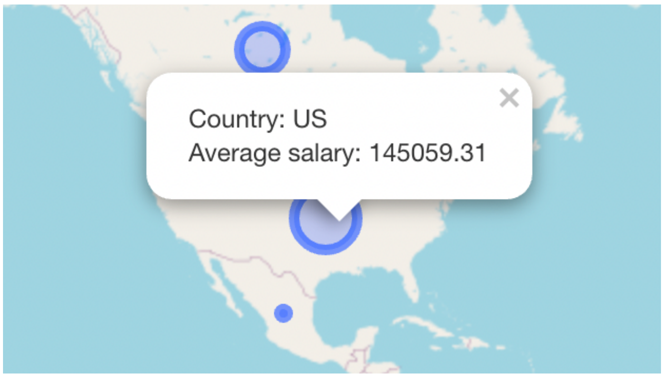
First of all, the home page of the website is showing the country factors which contains a map. The slide bar on the top is used to filter the average salary for each country. By selecting the average salary, user can view the map showing the dots which represents countries that data scientists' average salary within the selected range. The button on the left is selecting the minimum average salary and the button on the right is the maximum average salary. Then, the map should filter out some of the dot on the map. Which looks like this:



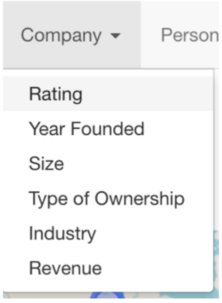
The graph about shows all the data by country, user can make it clear by moving the slide bar which could give us the graph below as a result.



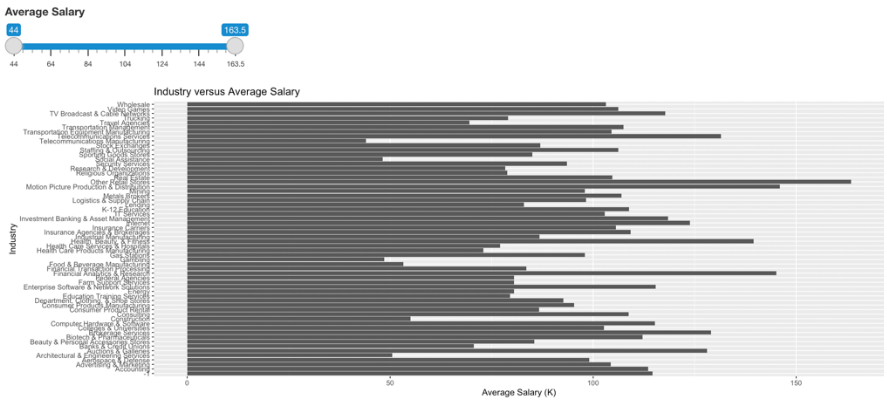
The map can be zoom in and out by clicking the plus and minus button on the top left corner. If the users click on the dot on the map, the tool tip will pop up and showing the country name and their average salary. Then tooltip looks like this:


Country: US
Average salary: 145059.31

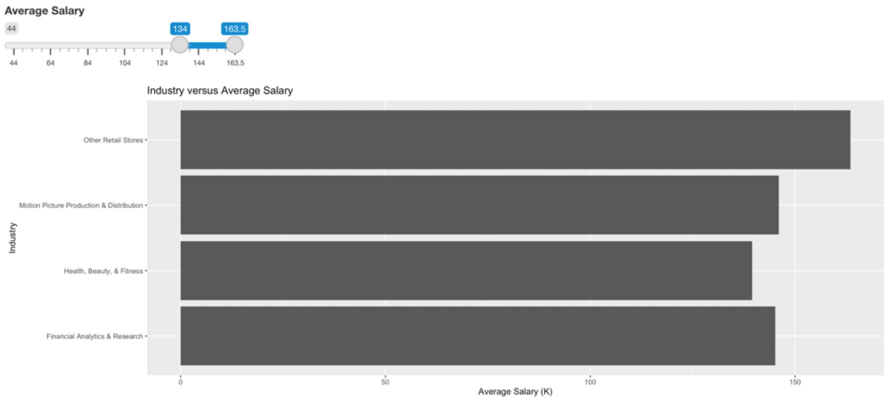
Then, we could start exploring other different factors. For the company factor, users can click on the company button and a menu appears. The menu contains different factors that is in the company aspect. Then, user can click on each of the factors to explore the graphs. Each page shows the graphs of how the factors affect data scientists' salary and the explanation of the graph.



The industry page contains an interactive graph. In default, the graphs show all the industry relative to the data scientists' salary. It is too much information to read, and the bar is crowded. User may use the slide bar on the top to filter out the data. The slide bar can select the range of average salary that user want. The graph will then show the industry that data scientists' salary within the range. The information can tell user that which industry is relatively better for getting their target salary.


Average Salary
44
163.5
44
64
84
104
124
144
163.5
Industry versus Average Salary
Industry
Wholesale
Video Games
TV Broadcast & Cable Networks
Trucking
Travel Agencies
Transportation Management
Transportation Equipment Manufacturing
Telecommunications Services
Telecommunications Manufacturing
Stock Exchanges
Staffing & Outsourcing
Sporting Goods Stores
Social Assistance
Security Services
Research & Development
Religious Organizations
Real Estate
Other Retail Stores
Motion Picture Production & Distribution
Mining
Metals Brokers
Logistics & Supply Chain
Lending
K-12 Education
IT Services
Investment Banking & Asset Management
Internet
Insurance Carriers
Insurance Agencies & Brokerages
Industrial Manufacturing
Health, Beauty, & Fitness
Health Care Services & Hospitals
Health Care Products Manufacturing
Gas Stations
Gambling
Food & Beverage Manufacturing
Financial Transaction Processing
Financial Analytics & Research
Federal Agencies
Farm Support Services
Enterprise Software & Network Solutions
Energy
Education Training Services
Department, Clothing, & Shoe Stores
Consumer Products Manufacturing
Consumer Product Rental
Consulting
Construction
Computer Hardware & Software
Colleges & Universities
Brokerage Services
Biotech & Pharmaceuticals
Beauty & Personal Accessories Stores
Banks & Credit Unions
Auctions & Galleries
Architectural & Engineering Services
Aerospace & Defense
Advertising & Marketing
Accounting
-1
0
50
100
150
Average Salary (K)

The graph above shows all the industry relative to the data scientists' average salary. By filtering the data, we can get the bar chart such as the below image. It shows the top 4 industry which data scientists' have the highest average salary.


Average Salary
44
134
163.5
44
64
84
104
124
144
163.5
Industry versus Average Salary
Industry
Other Retail Stores
Motion Picture Production & Distribution
Health, Beauty, & Fitness
Financial Analytics & Research
0
50
100
150
Average Salary (K)

At last, user can also explore the personal factors by clicking the personal button. Then webpage shows the bar chart for how seniority affects the data scientists' salary.

Conclusion

From the result of the graphs, I have got most of the answer from my purpose of doing this project. In general, I have got the answer of how country, different type of company and personal factors affects data scientists' salary.

In a global view, we could also compare the data scientists' average salaries in different countries. Those companies in developed countries such as the USA, Canada, and Russia could pay more for data scientists. If the student is ambitious to get a higher-level job, the success rate in these countries is higher.

Data scientists can get more pay and a higher potential to get a high-level position while they are working in the large companies working in the business or financial industry. In terms of companies' revenue, companies earning $5M to $10M pay the most for the data scientists which is a surprising result. It means companies earning more are not those who pay the most for their employees. On a personal level, data science students can keep working on improving their skills by getting more experience. Since the seniority can boost their salary.

Throughout the visualization implementation process, I have tried to summarize the result that I have found in the data exploration project and turn it into a visual report. The most challenging and interesting thing is that I have to consider a lot of factors during the design. For example, I have to consider the logical flow of the website and is it user friendly to the audience. I also need to consider the target audience of my project so that I can put information and adjust the complexity of the website in order to give the most information to the users.

The coding part of the project in shiny is a bit challenging at first, but once I understand how I can organize with using the navigation bar. The whole implementation is not that difficult. The only thing that I have to aware of is the variable between ui and server file have to match. Also, the number of brackets should be match. Since some of the syntax error will cause nothing appears in shiny. It is time consuming to find of the error if I made some careless mistake.

In the aspect of organizing the data, I am quite satisfy with what I have achieved. I could summarize and plot the data for most of the factors. Still I did not analyse all the factors that I have planned to do. The reason is I do not get any data relative to it and some of the data is a NA in the dataset that I am doing. In the future, I may have to do research on getting more dataset with relative information before doing the wrangling and data analysis.

In order to improve the result, the thing that I can do is to better organize the data. Since there are some of the axis such as the company's revenue, the x axis is not ranked by the order. Therefore, it may be a bit not user friendly. The thing I have to do is to remove the words and symbol of the column and then turn it into a number data type. Then, I may sort it in order. This may makes the order of the bar charts looks better. I think the web that I made is pretty good which has shown a lot of information. However, I am not good at design so that it is not that beautiful. In the future, I may get familiar with D3 which include more interactive features and beautiful design. The information that I have included in this project is great but the web will be more attractive for general audience if the design and graphs look better. Another thing is that I can think about how to make the charts more interactive before drawing it. In this project, I have chose to use the most traditional graphs for presenting the statistics. However, there are also other different ways to represent it. Also, adding some animation to it can also be more eye catching.

Bibliography

Countries.csv. https://developers.google.com/public-data/docs/canonical/countries_csv

Elena, Z., Tony, A., Robert, v. L. (2008) *Overview of Interactive Visualisation.* DOI: 10.1007/978-1-84800-269-2_1

Nikhil,B. (2021) *Data scientist salary*. Kaggle. https://www.kaggle.com/nikhilbhathi/data-scientist-salary-us-glassdoor?select=data_cleaned_2021.csv

RDocumentation. Ggplot2. https://www.rdocumentation.org/packages/ggplot2/versions/3.3.5

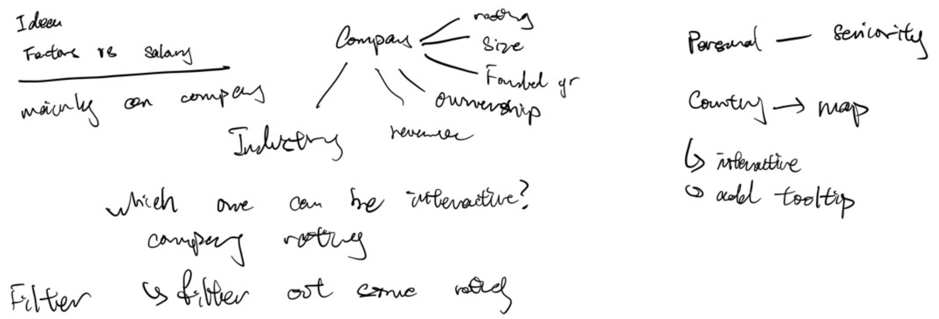
R Shiny. https://shiny.rstudio.com/gallery/navbar-example.html

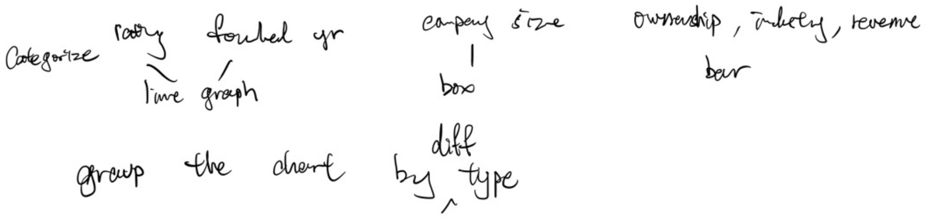
Saurabh, S. (2021) *Data Science Jobs Salaries Dataset*. Kaggle. https://www.kaggle.com/datasets/saurabhshahane/data-science-jobs-salaries/discussion

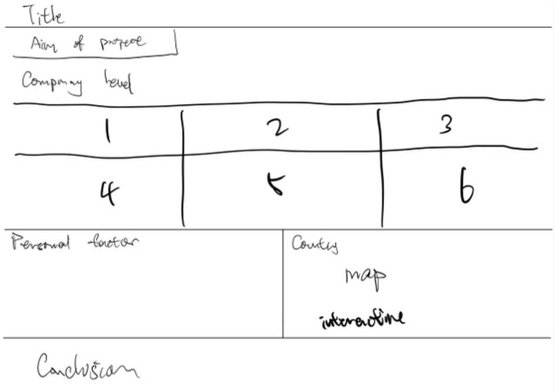
Situ, W., Zheng, X.Y., Daneshmand, M. (2017) Predicting the Probability and Salary to Get Data Science Job in Top Companies*. 2017 Industrial and Systems Engineering Research Conference.*

https://www.proquest.com/openview/dca60d29b2cc1067fe5dd5c6cf8ed168/1?pq-origsite=gscholar&cbl=51908

Appendix







Operation

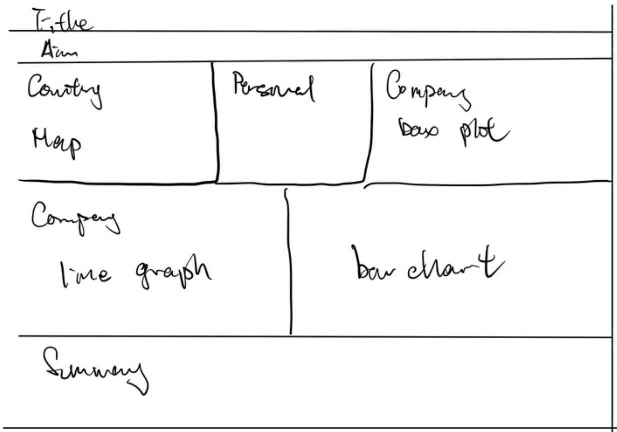
This is the whole layout for the website which divide the factors affecting Data scientist salary into 3 aspect. Then we can have a conclusion space at the bottom, So far the map is interactive.

Focus

Categorize into 3 general factors. and then packed all the graph into each section

Discussion

Organize like this can be more clearer for audience to spot the factors.



Operation

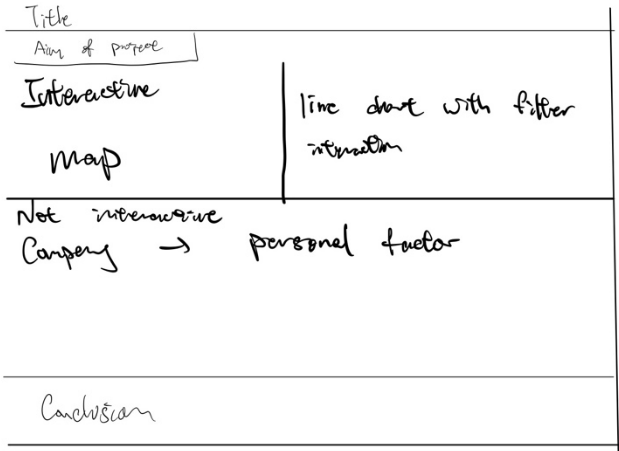
This time we also categorize the layout into 3 general factor. but this time I put the county Map first. This may give readers a general feeling at the first glance. Then, I have categorise the type of chart in company ~~factor~~

Focus

This design focus on categorising different graph into same area

Discussion

This may makes audience easier to process the graph. Their mode in their brain do not have to change suddenly.



Operation

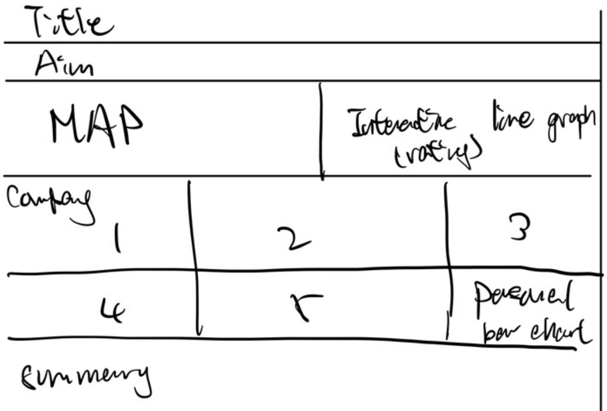
This is the whole layout for the website which divide into interactive and non-interactive graph. Two to Three interactive graph are placed on the top which catch the eyes of the audience.

Focus

Attract the audience attention by placing interesting graph on top of the design.

Discussion

This design is attractive to audience by having interaction but it is a bit messy for the information that I have to deliver. Therefore, I have to be more organize for the graph



operation

The map is interactive which may show the tooltip when I click on the circle.

The line graph should also be interactive which may filter out some unused data.

Focus

The layout summarize my idea which placed the interactive charts on top. The charts at the bottom are non-interactive and grouped by the factors.

Detail

At the start of the graph, we placed some interactive graphs for eye catching purpose. Then, the grouping by factors should be easier for user to process the data. It starts with global aspect to personal factors.