Data Visualisation Project - Comprehensive Analysis of Data Science Salaries: 2020 ~ 2024

Full Name: Yu-Jung Ho (Malone Ho)

Student ID: 33531315

Applied Session Number: 05_Oncampus (Tuesday 12:00 – 14:00)

Teaching Associate: Bruno Mendivez Vasquez

Table of Contents

1. Introduction

This project is trying to deliver the finding from DEP, which is that the location of company and the levels of experience are the 2 most significant factors which can affect the amount of salary. This project will use multiple figures to visualise the relationship between each factor and the amount of salary and explain how these factors affect the amount of salary. The potential audience of this project is the students who are going to graduate from data science related field or some people who are interested to find jobs related to data science. This project allows readers to understand the current average salary for data science related jobs around the world.

2. Design Process

Overall, the genre of this project will be the magazine style, which provide a professional appearance and logically visual path (Clementson, n.d.), and the narrative visualisation style will be hybrid, which represents that half of the website will be author-driven and another will be reader-driven.

The background colour fort this website will be light gradient blue. This colour can make text and images stand out, making the content easy to read and visually appealing (Clayton, 2023). In terms of fonts, this project will use sans serif typeface. To be specific, the title of this project and each section will be Helvetica Neue, and the content will be Helvetica. Helvetica enhances the readability and provides a clear, neutral design. To ensure reader's sight line will be focus on the first part when they just open this website, the font of the three questions will be bold and slightly larger than other content. Otherwise, readers might ignore the introduction and directly dive into other sections with images. Furthermore, the bold style is also applied in other sections to highlight the main findings. To create the symmetry and balance in this website, the structure of the entire website will be two columns and multiple rows. Two columns could effectively shorten the website but will not make the entire website too crowded. Moreover, space will be added between sections to group elements, which can reduce the distraction from other sections.

For the first design sheet, I generated as many graphs as possible, including simple graphs such as line chart and box plot, and complex graphs such as flow map and

Sankey diagram. I filtered out some graphs because some graphs show the same information as others and some graphs contain too little information. For the categorisation, I categorise the remaining graphs to three types, which are map, statistical and flow or relation graphs. These three types are able to conclude all findings from my DEP and present in DVP. To make my website more comprehensive and interactive, I combined map and ridgeline chart together as linked graphs. Additionally, I planned to use animation, filter and tooltips to include more information in graphs.

For the second sheet, I wanted to make my website to present the data from the simple aspect to the comprehensive aspect. Hence, after briefly introducing the project and the three questions, I planned to use simple graphs such as a box and violin plot to generally show the relationship between the amount of salary and the level of experience, and used a filter to allow reader to choose to combine multiple box plots in the first section. Filter allows reader to change to plot for different years. Tooltips can give reader see more information about a particular group. The description for box plot will be provided at the side of the chart. Furthermore, the description will be change according to the year that reader selected. Afterward, I used Sankey Diagram to present more relationships with the salary. This diagram will be animated by years to show the trend over years. For the next section, I planned to present some findings related to employee residence and company location. Thus, I used graphs belonging to the flow type in the first sheet. Reader can use the filter to choose either to show flows via chord or flow map. The default value is chord diagram since I want reader to see the overall trend. If reader want to see more detail about some particular flows, they can change to flow map and see the detailed information. Tooltips are provided in the flow map to present the exact number of each flow. The section will be introducing the relationship between salary and company location via choropleth and linked ridgeline chart. Choropleth map will present the average salary in different countries and ridgeline chart will present the distribution of different experience levels in different countries. The default ridgeline chart will be the global distribution of different experience levels. When reader hovering on a particular country, tooltip will show the detailed information for that country. In addition, if the reader clicks on a particular country, the ridgeline chart changes according to that country. Finally, for the section of the comprehensive

analysis, I decided to use the correlation matrix and decision tree to conclude all findings. The reason why these two graphs were not listed in the first sheet is because this idea was given from the feedback of my DEP and the feedback was released after I completed the first sheet. The correlation matrix will illustrate the relationship between variables in the dataset, and the decision tree will illustrate what are the main factors that could influence the amount of salary based on the machine learning model. In case some people do not know how to analyse the correlation matrix, I will add some animations such as motions to highlight the important information in this matrix. After finishing this design sheet, I felt this website might be too long to read. Readers might have to keep scrolling the website up and down to view each graph. Therefore, I made some improvements in the next sheet.

For the third sheet, I shortened the website by pairing up graphs. Basically, the graphs use is same as previous sheet except the fourth graph. From the viewpoint of website structure, this graph maintains the balance and symmetry of the entire website. From the viewpoint of main findings, this sunburst diagram presents the distribution of different variables such as levels of experience and employment types. Furthermore, this sunburst diagram will be animated to show the trend over years. There is one more difference between this sheet and the previous sheet, which is that I added a filter for choropleth map. Readers can use this filter to choose how many countries are shown in the map. The reason I created this filter is because I think that if all the countries are shown on the map, readers might be overwhelmed by having too much information on one map. After I completed this sheet, I guess the final website might be too crowded at the middle part of the website if reader choose to use flow map to illustrate the finding for the company location and employee residence as there will be two maps in the middle part of the website. Hence, I improved this problem in the next sheet.

For the fourth sheet, I decided to leave more space for the section which includes the chord diagram and flow map, and I moved the sunburst diagram to the top of the website. The reason I moved the sunburst diagram up is because I think this sunburst diagram could provide the general concept about the relationship between salary and other factors for readers. Moreover, I added a line chart which can also provide

general concept about the relationship between salary and other factors for readers. Readers can use filter to select different factor shown in the graph. Tooltips can present more detailed information for particular node on the line. The default type is experience level, as experience level is the most common factor that other people might consider to be the most important factor that can affect the amount of salary.

For the fifth sheet, I decided to remove the sunburst diagram as I used some sample data to create the sunburst diagram and realise that this diagram might be too complicated read. Hence, I decided to just use line chart to present the general concept before readers dive into more detail.

In the final design, line chart provides the general concept before readers dive into more detail; box and violin plot provides the distribution of salary in different experience levels; Sankey diagram shows the relationship between experience levels, employment types and the different intervals of salary; chord diagram and flow map can see what is the most popular location for working; choropleth map and ridgeline chart illustrate the average salary and the distribution of different experience levels in different countries; correlation matrix shows the relationship between variables; decision tree provides the main factors which affect the amount of salary from the machine learning model.

3. Implementation

3.1 Technical Implementation

This project has used multiple high-level libraries to produce interactive graphs. The list below is some high-level libraries and the purpose for each library.

- **gganimate & magick & gifski:** Used to produce animation and gif
- **networkD3 & htmlwidgets:** Used to produce Sankey diagram and optimise front-end appearance
- **circlize:** Used to produce chord diagram
- **sf & rnaturalearth & geosphere:** Used to calculate the distance between countries and load the shape for countries
- **ploty & DT:** Used to add the tooltips and optimise the interaction of graphs

Furthermore, this project requires extensive data wrangling to form the require data format to produce different types of graphs. The list below the significant technique some complex graph.

- **Box & Violin Plot:** As the process speed for rendering animation extends the time to render the whole website. I had to store the animation as gif file and load the file to increase the process speed. The parameters for rendering animation should be set properly, as it significantly affects the process speed. Completed code to render animation is provided as the comment in the source code.
- **Sankey Diagram:** Since the required data format to render this diagram is completely different from the original data format. Hence, I had to re-organise the whole dataset and create the files for links and nodes separately. Completed code to render this diagram is provided as the comment in the source code.
- **Chord Diagram:** To render this diagram, I had to produce an adjacent matrix including source locations and destinations from original dataset. Furthermore, original adjacent matrix was too messy to visualise. Hence, I only keep some flows which are important.
- **Flow Map:** To create an accurate flow map, I had to calculate the distance between countries and add the coordinates for each country. Furthermore, the name for some countries is not identical from the library and my dataset. Hence, I manually checked these mismatches and corrected it. For the coordinate of each country, I had to manually insert coordinate for some countries, as the coordinate for those countries are not included in the library. I also changed the data structure to fit the required structure to produce flow map.
- **Choropleth Map & Ridgeline chart (Linked Graph):** For the inconsistency of countries name between library and my dataset, I manually check those mismatches and corrected it. Furthermore, I used shape files to add the polygons for each country. In terms of linked graph, the movement from users should be observed to change the graph in real time, I used observeEvent() function to monitor the movement and change the graph according to the id of the country that user click on.

- **Correlation Matrix:** To highlight a specific row, I added another layer on the specific row, and it should be included in the animation. Completed code to render animation is provided as the comment in the source code.
- **The Difference between Final Design and Implementation**
  The only difference between the final design and the final implementation is that the Sankey diagram becomes static instead of animation. The reason why I remove the animation is because the final animation of this Sankey diagram might mislead readers, and this animation might be too complicated to analyse. Hence, I changed this diagram back to static and add some tooltips to show the detailed information of each flow.

3.2 Interactive Narrative Visualisation Implementation

Since this website is going to introduce the dataset from simple aspect to comprehensive aspect, I firstly introduced the importance of the data science by referring the finding from research paper. Afterward, I explained the purpose of this website and briefly introduce the collected data. Three questions proposed from my DEP are used to be the conclusion of my introduction section.

For those interactive graphs, user guide is provided below each graph to guide readers how to control the graph.

From the simple aspect, I started to show how the amount of salary changes in different levels in different key factors over years. Filter is used to present different chart with different key factors. Tooltips in the line chart provide detailed information for particular time periods and levels. Furthermore, filter and tooltips increase the interaction between the website and readers. The narrative in this section can guide readers to understand these charts deeper by telling readers some subtle differences from these three different charts. Figures 1 is the sample result of this section.

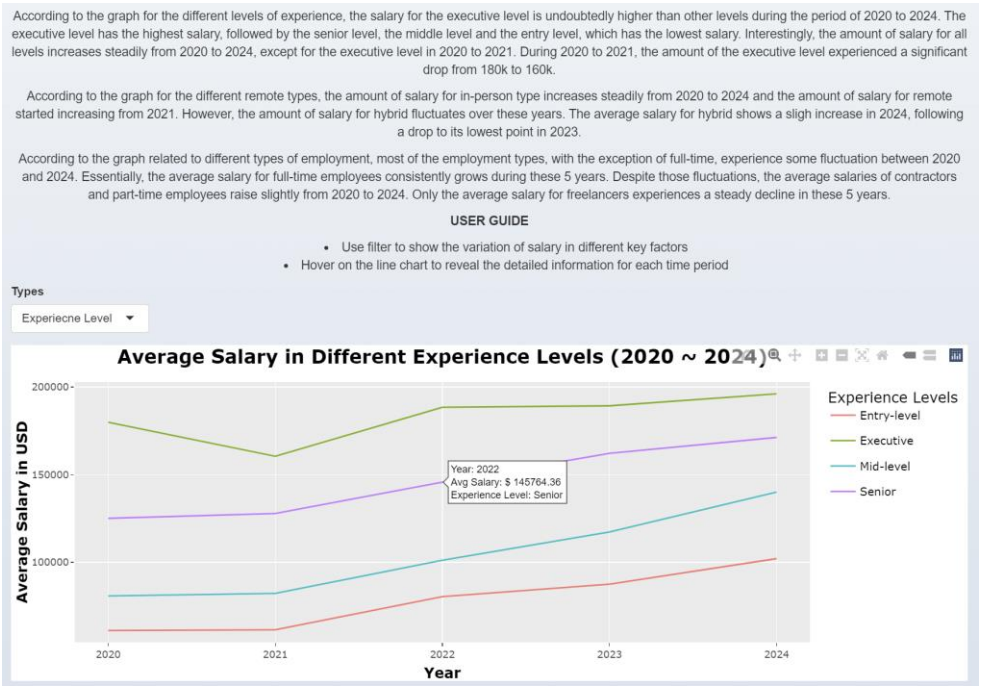


Figure 1. Relationship between salary and experience levels. (2024) [Screenshot]. R Studio. https://posit.co/download/rstudio-desktop/

This website starts going deeper into this dataset after the previous section. The next section explains how the level of experience affects the amount of salary. The reason why I decided to introduce experience level in my second section is that the level of experience is the factor that most people consider it can influence the amount of salary the most. This type of introduction also aligns with the concept of from general to narrow. Additionally, the narrative in this section provides more insights from this animation.

Next section adds one more factor, employment type, to analyse the relationship between each other. The relationships are presented by Sankey diagram. This diagram can not only see the majority of the flows but also observe the flows that cannot be noticed easily. Moreover, this diagram includes tooltips to show the exact number of each flow. The narrative in this section provides the information how so observe this diagram and the insights of this diagram.

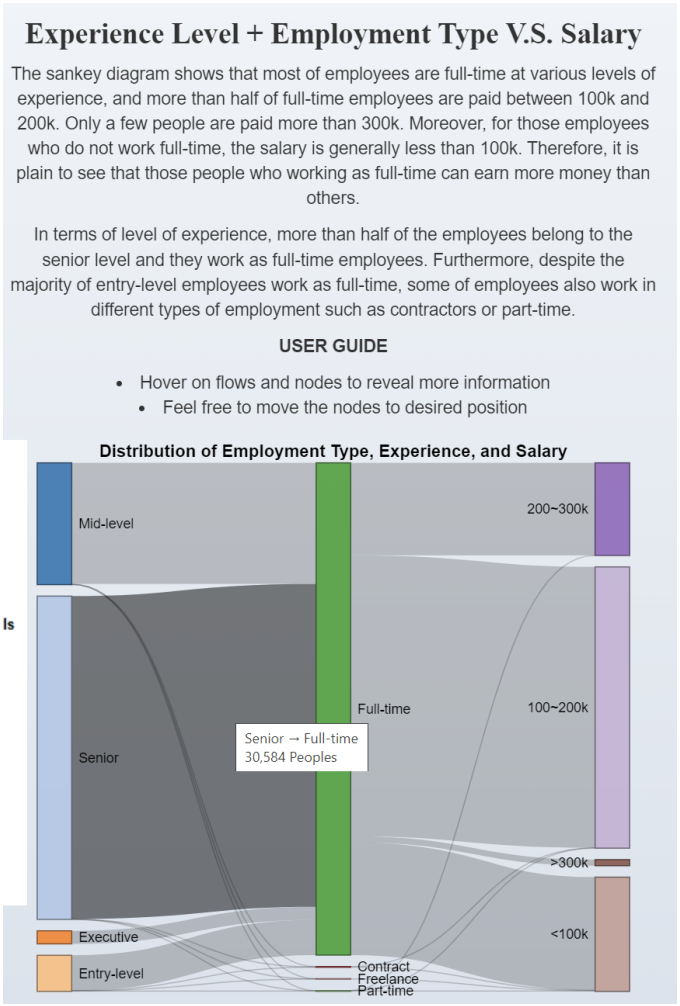


Figure 2. Sample Result for Sankey Diagram. (2024) [Screenshot]. R Studio.

https://posit.co/download/rstudio-desktop/

Afterward, this website starts analysing the geographic location. This section firstly uses chord diagram to present the general flows from residences to company locations, as chord diagram can be observed easier than flow map. The narrative in this section supports readers to understand the graphs and conclude the findings in this section. The figure 3 is the sample result if readers select the flow map in this section and filter some countries out.

Figure 3. Sample Result for Flow Map. (2024) [Screenshot]. R Studio.

https://posit.co/download/rstudio-desktop/

After the analysis of residence and company location, next section is going to analyse the relationship between company location and salary. The section not only uses choropleth map to show the average salary in different countries but also uses ridgeline chart to show the distribution of employees in different experience levels in different countries. Ridgeline chart explain the reason why the salaries in some countries are relatively high or low. The default ridgeline chart is the global distribution of employees in different experience levels. Readers can click on the specific country to see the distribution in that country. The filter above the map allows readers to select how many countries are shown in the map if readers think the current map is too crowded. The legend in the map represents the overall amount of salary across the whole world. The higher percentage and deeper colour, the higher amount of salary in that country. As this section might include too many information, the narrative in this section has summarised some vital findings for readers. The figure 4 is the sample result for choropleth map and linked ridgeline chart.

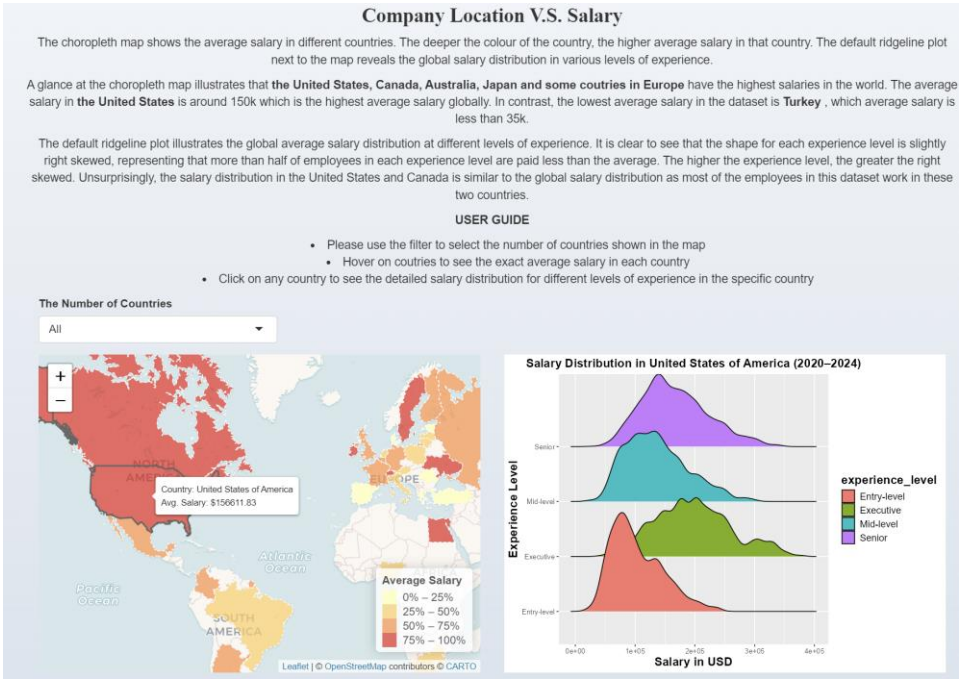


Figure 4. Sample Result for Choropleth Map and Ridgeline Chart. (2024) [Screenshot]. R Studio. https://posit.co/download/rstudio-desktop/

For the comprehensive analysis section, correlation matrix and decision tree are used to explain the overall findings. Since not everyone understands how to read the correlation matrix and the decision tree, I made the correlation matrix as an animation and highlighted the row which is the most important one to support readers to understand the matrix. The narrative in this section not only tells readers how to read the matrix and decision tree but also summarises the findings from these two graphs. Figure 5 is the result of this section.

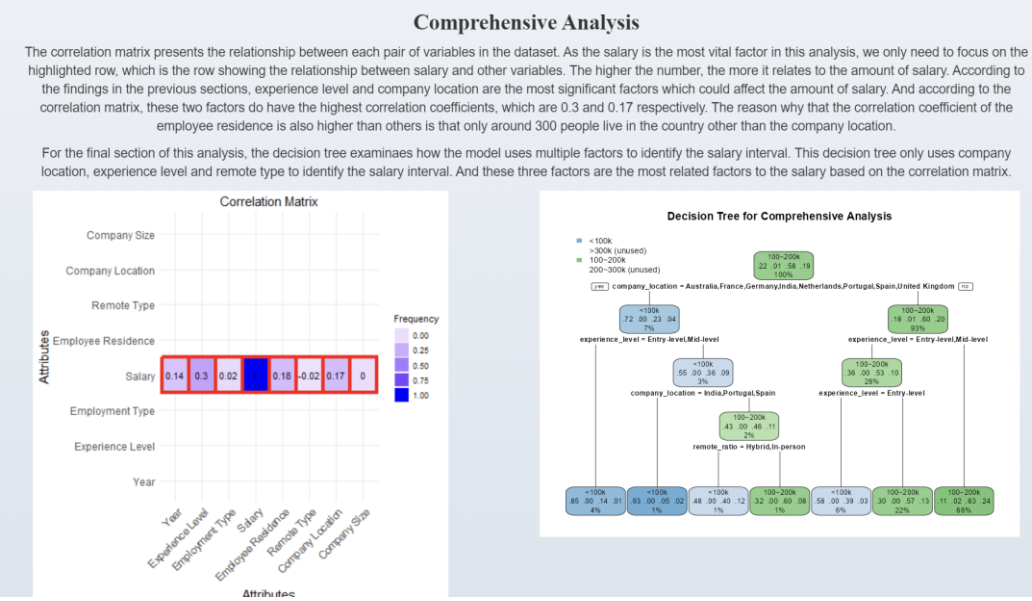


Figure 5. Sample Result for Correlation Matrix and Decision Tree. (2024) [Screenshot]. R Studio. https://posit.co/download/rstudio-desktop/

The conclusion is used to summarise the findings from the graphs above and answers the three questions mentioned in the introduction section. The data source section explains the information about the collected dataset.

3.3 Using the Implementation

- **Introduction & Data Source**

  The word with blue colour means that click on those words can lead readers to reference website.

- **Line Chart**

  The filter above the line chart can let reader choose different factors shown in the line chart. Hovering on the line chart can also present the tooltip for detailed information.

- **Sankey Diagram**

  The nodes in this diagram can be moved up and down. Hovering on the nodes and flows can show the tooltip for detailed information.

- **Chord Diagram & Flow Map**

  The filter next to the graph section can let reader to select either use chord diagram or flow map to present the information in this section. There is also a filter contains in the flow map. This filter allows readers to filter the flows from

some origin countries. For instance, if readers uncheck India, any flows depart from India will be invisible. Readers can also hover on each flow to show the exact number of migration shift.

- **Choropleth Map & Ridgeline Chart**

  The filter above the map can let readers to select how many countries should be shown in the map. Hovering on countries reveals the details information by tooltip. Moreover, click the country can also change the ridgeline chart next to the map. The ridgeline shows the distribution of salary in different experience levels.

4. Conclusion

In conclusion, after the analysis of this project, I found out that the location of company and levels of experience do affect the amount of salary significantly. If an employee works in the United States or Canada can generally get a higher salary than other employees working in other countries. In terms of levels of experience, even an employee working at the entry level for a very long time, the amount of salary might still lower than other employees who work at mid-level or senior level.

After completing this project, I have learnt how typography, typeface and colour potential affect the user experience. Furthermore, different types of charts with the same data can also deliver different information to users. Technically, I have learnt how to explore the insights of data and use Tableau, R and D3 to create a meaningful graph.

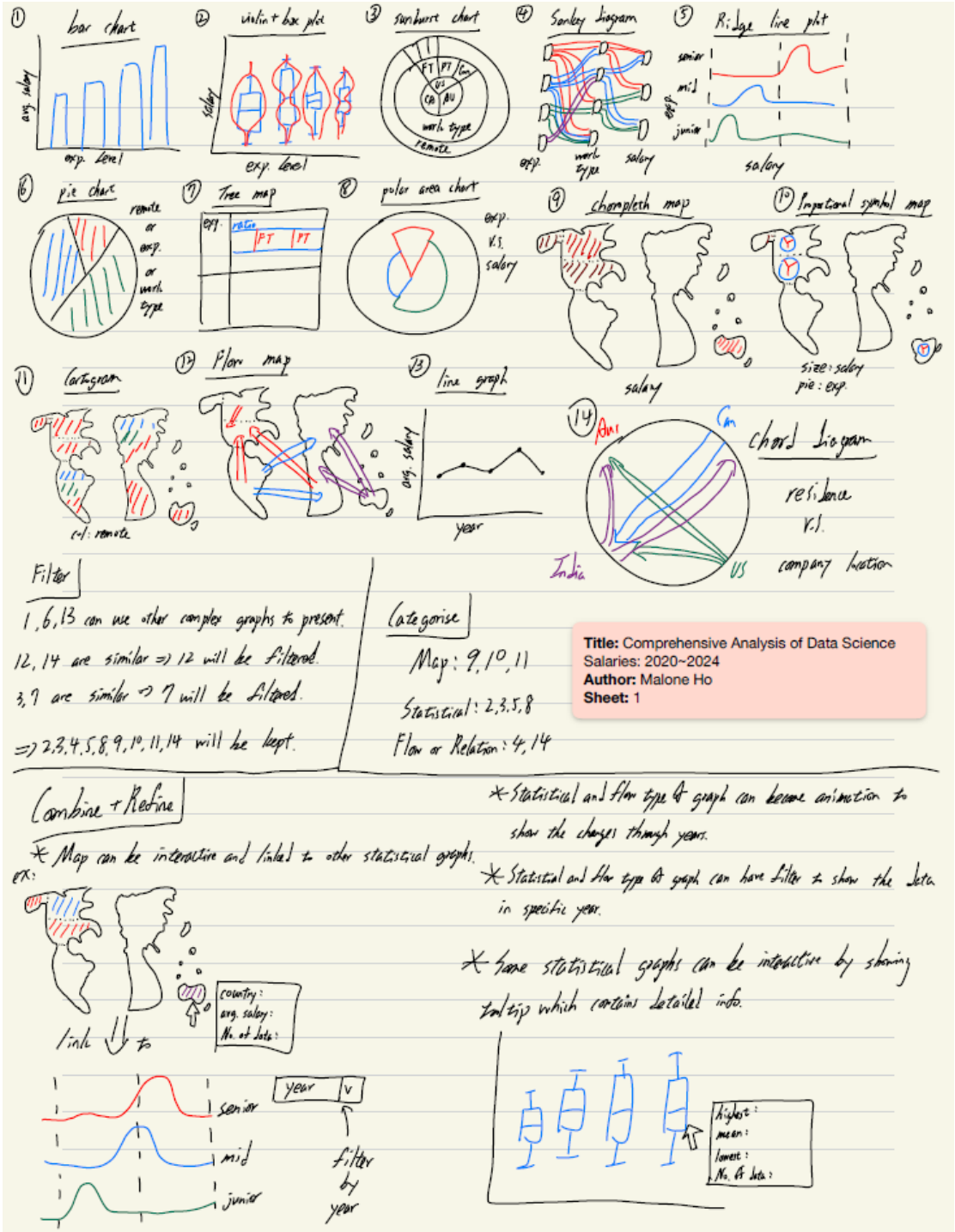
5. Reference

Clementson, J. (n.d.). *Magazine Page Layout: What You Need to Know.* Azura. https://azuramagazine.com/articles/magazine-page-layout-what-you-need-to-know

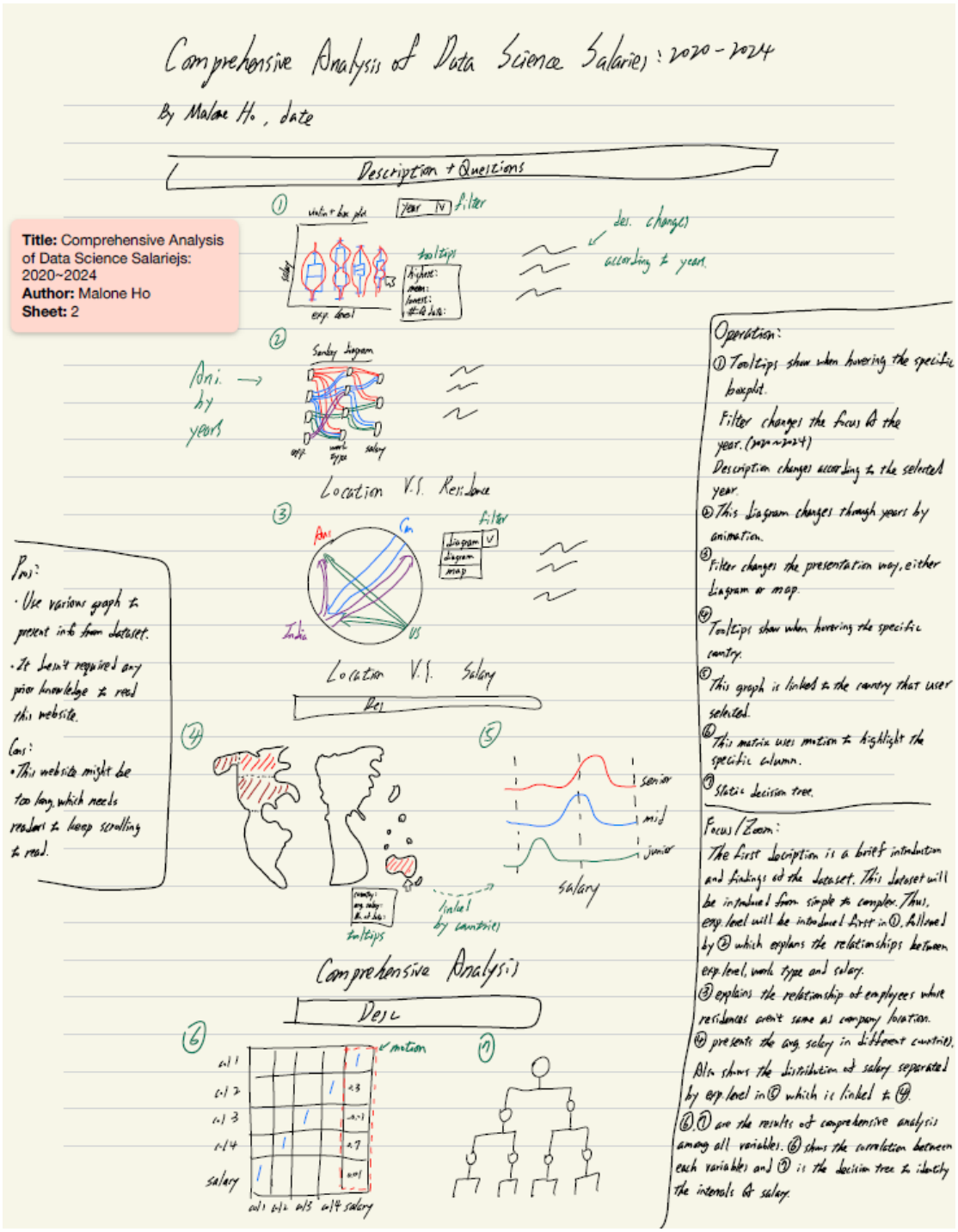
Clayton, T. (2023). Best Color Combinations For Readability. Rigorous. https://rigorousthemes.com/blog/best-color-combinations-for-readability/

6. Appendix

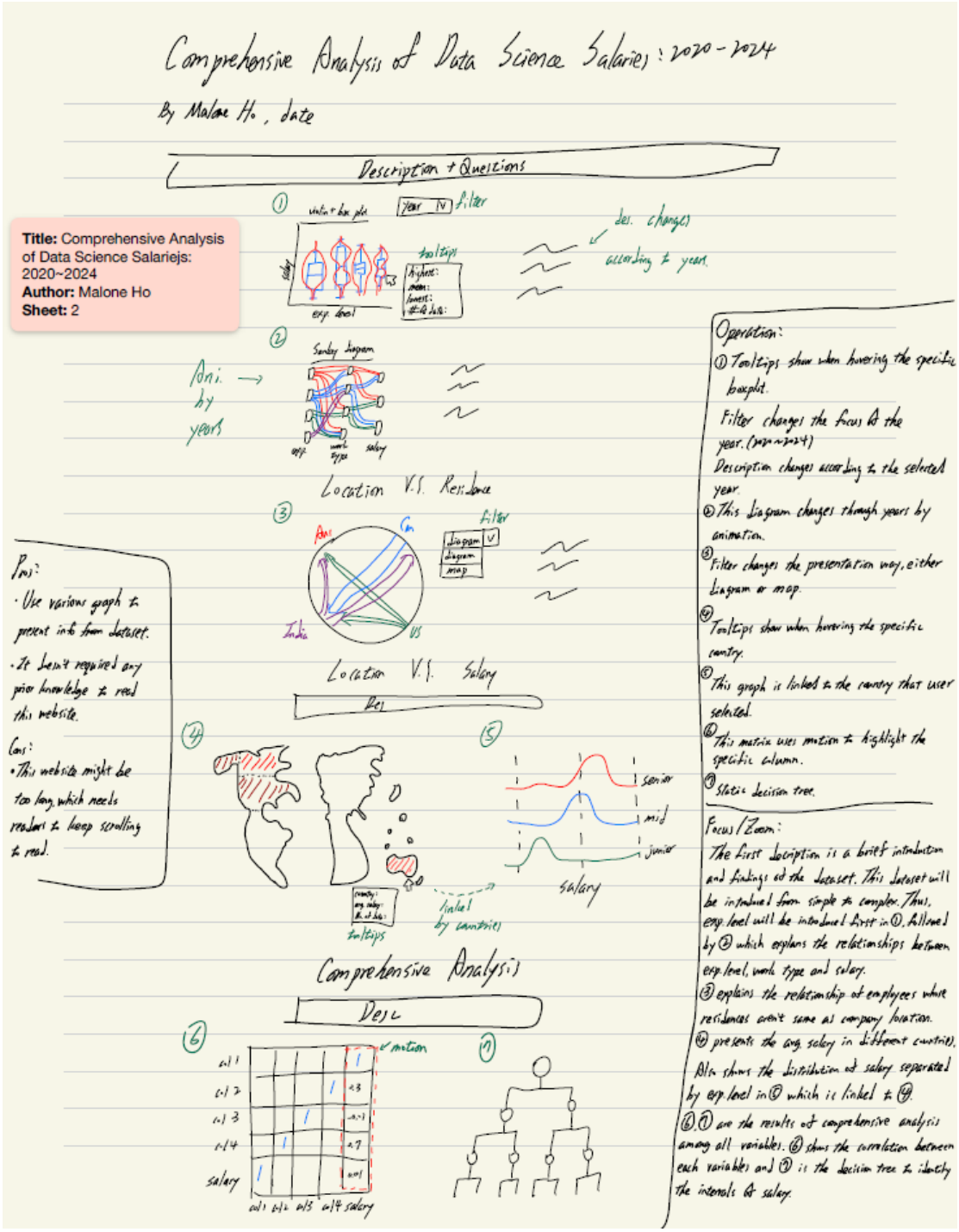
- **Sheet 1**



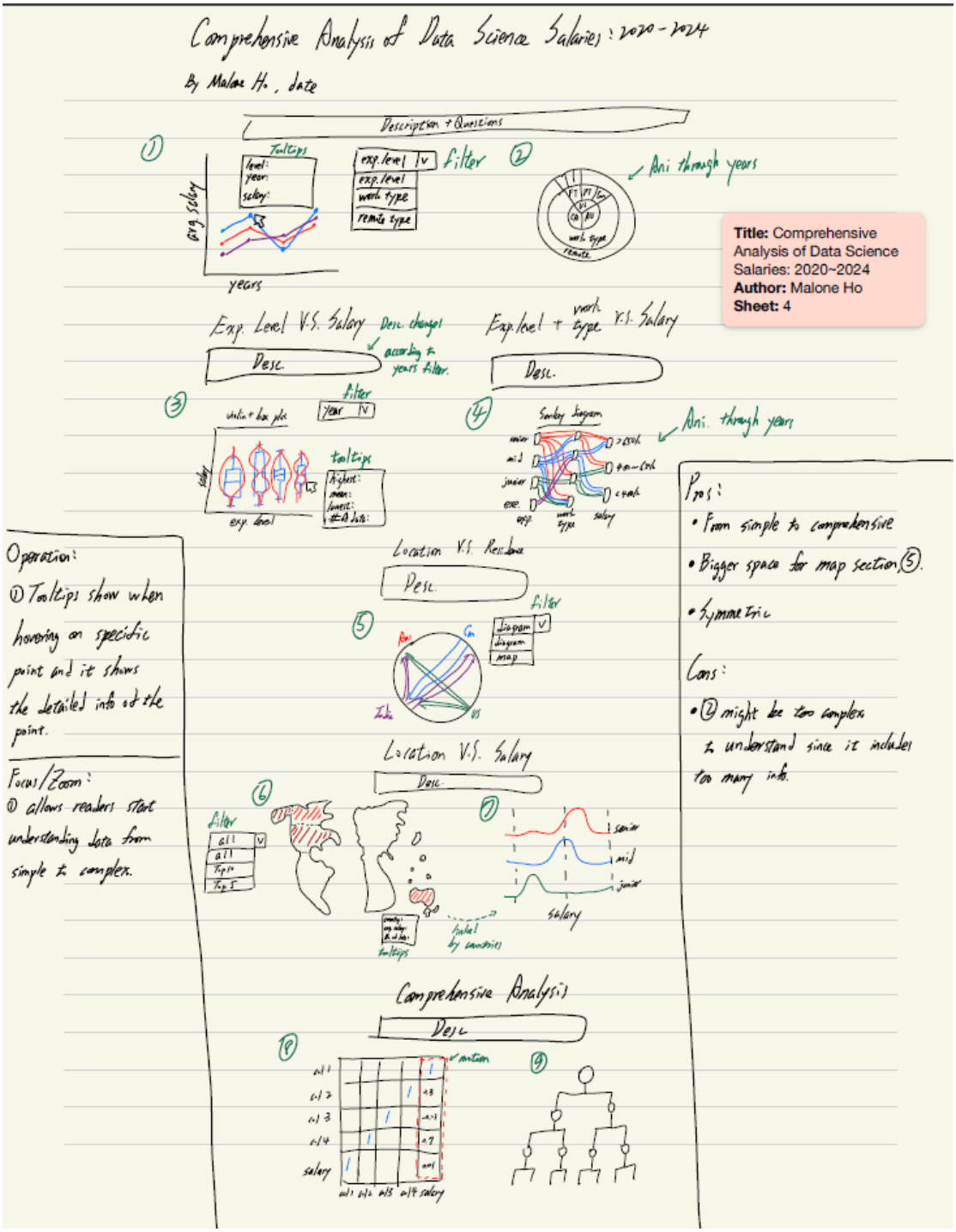
- **Sheet 2**


Comprehensive Analysis of Data Science Salaries: 2020 - 2024
By Malone Ho, date
Description + Questions
Title: Comprehensive Analysis of Data Science Salariejs: 2020~2024
Author: Malone Ho
Sheet: 2
Sankey diagram
Location V.S. Residence
Location V.S. Salary
Comprehensive Analysis
Pros:
· Use various graph to present info from dataset.
· It doesn't required any prior knowledge to read this website.
Cons:
· This website might be too long, which needs readers to keep scrolling to read.
Operation:
① Tooltips show when hovering the specific boxplot.
Filter changes the focus of the year. (2020~2024)
Description changes according to the selected year.
② This diagram changes through years by animation.
③ Filter changes the presentation way, either diagram or map.
④ Tooltips show when hovering the specific country.
⑤ This graph is linked to the country that user selected.
⑥ This matrix uses motion to highlight the specific column.
⑦ Static decision tree.
Focus/Zoom:
The first description is a brief introduction and findings of the dataset. This dataset will be introduced from simple to complex. Thus, exp level will be introduced first in ①, followed by ② which explains the relationships between exp level, work type and salary.
③ explains the relationship of employees whose residences aren't same as company location.
④ presents the avg salary in different countries.
⑤ shows the distribution of salary separated by exp level in ② which is linked to ④.
⑥ ⑦ are the results of comprehensive analysis among all variables. ⑥ shows the correlation between each variables and ⑦ is the decision tree to identify the intervals of salary.

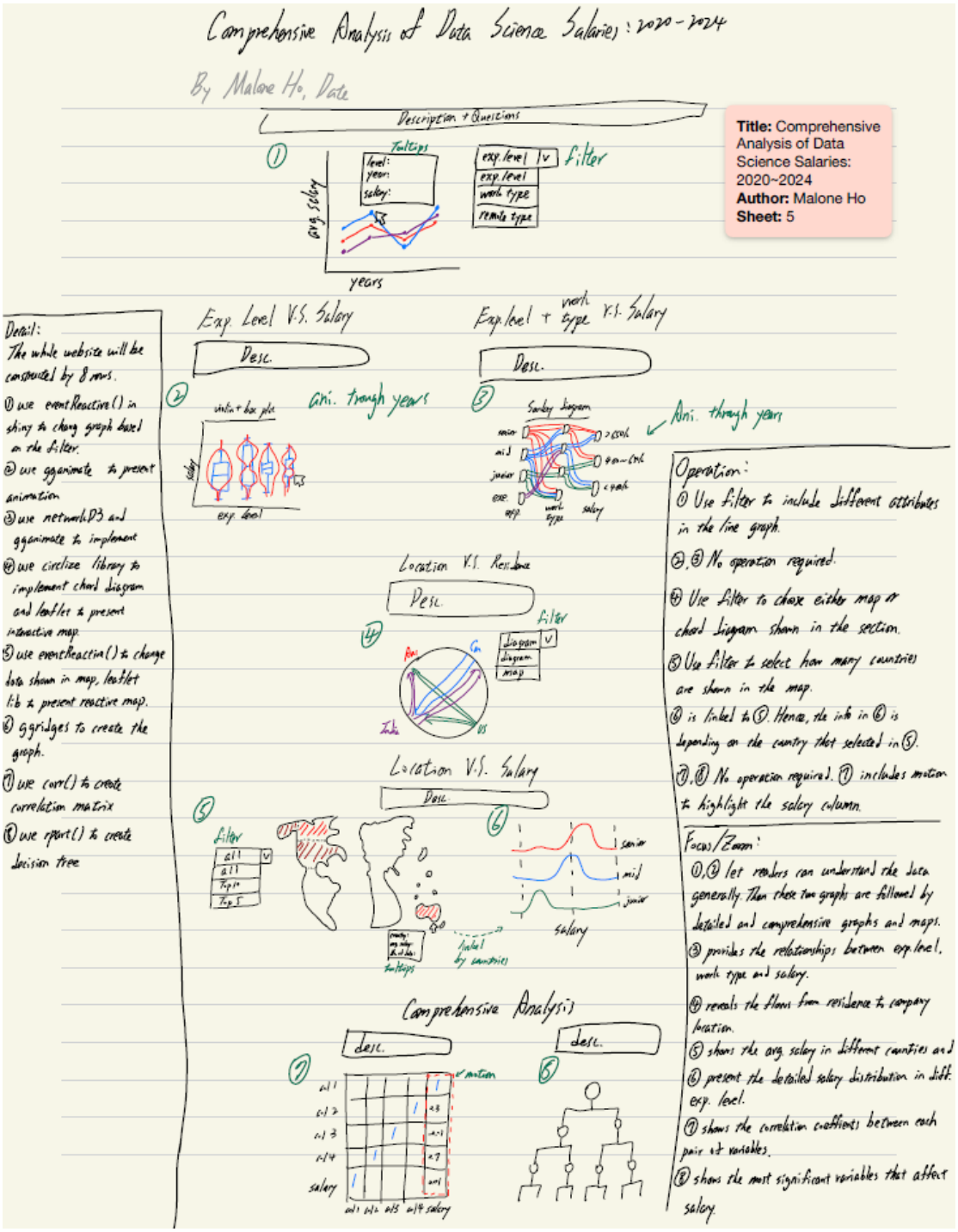
- **Sheet 3**


Comprehensive Analysis of Data Science Salaries: 2020 - 2024
By Malone Ho, date
Description + Questions
Title: Comprehensive Analysis of Data Science Salariejs: 2020~2024
Author: Malone Ho
Sheet: 2
Location V.S. Residence
Location V.S. Salary
Comprehensive Analysis
Pros:
· Use various graph to present info from dataset.
· It doesn't required any prior knowledge to read this website.
Cons:
· This website might be too long, which needs readers to keep scrolling to read.
Operation:
① Tooltips show when hovering the specific boxplot.
Filter changes the focus of the year. (2020~2024)
Description changes according to the selected year.
② This diagram changes through years by animation.
③ Filter changes the presentation way, either diagram or map.
④ Tooltips show when hovering the specific country.
⑤ This graph is linked to the country that user selected.
⑥ This matrix uses motion to highlight the specific column.
⑦ Static decision tree.
Focus / Zoom:
The first description is a brief introduction and findings of the dataset. This dataset will be introduced from simple to complex. Thus, exp level will be introduced first in ①, followed by ② which explains the relationships between exp level, work type and salary.
③ explains the relationship of employees whose residences aren't same as company location.
④ presents the avg salary in different countries.
⑤ shows the distribution of salary separated by exp level in ④ which is linked to ④.
⑥ ⑦ are the results of comprehensive analysis among all variables. ⑥ shows the correlation between each variables and ⑦ is the decision tree to identify the intervals of salary.

- **Sheet 4**


Comprehensive Analysis of Data Science Salaries: 2020-2024
By Malone Ho, date
Description + Questions
Tooltips
filter
Ani through years
Exp. Level V.S. Salary
Desc. changes according to years filter.
Exp level + work type V.S. Salary
Desc.
filter
Sankey diagram
Ani. through years
tooltips
Location V.S. Residence
Desc.
filter
Location V.S. Salary
Desc.
filter
tooltips
Comprehensive Analysis
Desc
salary
Operation:
① Tooltips show when hovering on specific point and it shows the detailed info of the point.
Focus/Zoom:
① allows readers start understanding data from simple to complex.
Pros:
• From simple to comprehensive
• Bigger space for map section ⑤.
• Symmetric
Cons:
• ⑫ might be too complex to understand since it includes too many info.
Title: Comprehensive Analysis of Data Science Salaries: 2020~2024
Author: Malone Ho
Sheet: 4

- **Sheet 5**


Comprehensive Analysis of Data Science Salaries: 2020 - 2024
By Malone Ho, Date
Description + Questions
Tooltips
level:
year:
salary:
avg. salary
years
exp level
exp level
work type
remote type
filter
Title: Comprehensive Analysis of Data Science Salaries: 2020~2024
Author: Malone Ho
Sheet: 5
Detail:
The while website will be constructed by 8 parts.
① use eventReactive() in shiny to chang graph based on the filter.
② use gganimate to present animation
③ use networkD3 and gganimate to implement
④ use circlize library to implement chord diagram and leaflet to present interactive map
⑤ use eventReactive() to change data shown in map, leaflet lib to present reactive map.
⑥ ggridges to create the graph.
⑦ use cor() to create correlation matrix
⑧ use rpart() to create decision tree
Exp Level V.S. Salary
Desc.
violin + box plot
salary
exp level
Gini. trough years
Exp level + work type V.S. Salary
Desc.
Sankey diagram
senior
mid
junior
exp
work type
salary
Gini. through years
Location V.S. Residence
Desc.
filter
diagram
diagram
map
Location V.S. Salary
Desc.
filter
all
all
Top 10
Top 5
tooltips
label by country
senior
mid
junior
salary
Comprehensive Analysis
desc.
desc.
motion
all 1
all 2
all 3
all 4
salary
all1 all2 all3 all4 salary
Operation:
① Use filter to include different attributes in the line graph.
②, ③ No operation required.
④ Use filter to chose either map or chord diagram shown in the section.
⑤ Use filter to select how many countries are shown in the map.
⑥ is linked to ⑤. Hence, the info in ⑥ is depending on the country that selected in ⑤.
⑦, ⑧ No operation required. ⑦ includes motion to highlight the salary column.
Focus/Zoom:
①, ② let readers can understand the data generally. Then these two graphs are followed by detailed and comprehensive graphs and maps.
③ provides the relationships between exp level, work type and salary.
④ reveals the flows from residence to company location.
⑤ shows the avg salary in different countries and
⑥ present the detailed salary distribution in diff. exp. level.
⑦ shows the correlation coefficient between each pair of variables.
⑧ shows the most significant variables that affect salary