FIT 5147 Visualisation Project: The trends of Data Science Jobs in Australia and Other Countries in 2021

Napoj Thanomkul

Student ID:32338589

Tutorial 08

Tutor: Tam Vo

Introduction

The purpose of this narrative visualisation of the exploration project is to identify the differences in the data science job trends between Australia and the other countries (United States and United Kingdom). From the former statement, these differences can be identified by observing the three following visualisations:

1. Bubble/Choropleth Map of each countries with number of job posts by each states/cities along with the bar chart of top three states/cities posting data science job with different type of jobs. These visualisations will provide the different trends of Data Science job postings between each country.
2. The multiple regression plot of salary prediction by two predictors ("skill count" and "seniority") which identify the factor which predict the salary
3. The Bar Chart of top10 technical skills requiring for the data science job between these countries. These visualisations show the differences in the trends of data science skills requiring for data science job in each country.

Based on the following visualisations project initial information, the main audience of this project is anyone who interested in Data Science career to somebody who plan to study a data science to get hired by the company in these three countries.

Design

In this section, the design of this visualisation project will be explained by the 5-sheet design as the following:

Sheet 1.

In the first sheet, it will be covered on the initial ideas for visualisation design which will answer the questions of the project starting by the following process:

1.Brainstorming/Idea:

In this sub-section, it will provide all the visualisations that might be related to the topics. Regarding earlier statement, these are all the ideas (See Appendix for the visualisation idea):

1. Line graph
2. Bar chart with differences in color
3. Heat map
4. Bubble/Choropleth map
5. Scatter plot
6. Contour map
7. Pie Chart
8. Flow Map
9. Box-plot

Once the ideas are prepared, the next process is filtering and categorising only a necessary visualisation for the project. Here is the filtering/categorizing result:

1. The line and scatter plot in a 2D plots are chosen for statistical analysis part for salary prediction because it shows a trend in a prediction. Whereas all the other remaining statistical plots(box-plot, and pie chart) are filtering out of the design. The reason for the removal of the other two plot because they are difficult for interpretation of the analysis insight. In addition, the graph will be presented in 2D because it capable for audience to observe without involving a hard work on human visualisation system (Satriadi,2021a).
2. The bubble/choropleth map is chosen instead of other map visualisations. The reason for the following decision is that bubble/choropleth map can easily represent a sufficient color hue for convenience in differentiate the information based on the color on the map. For example, when there is a 5 colour hue intensity on the plot which including red,green,blue and yellow, the audience who observe the visualisation can differentiated which colour represented high occurrence and vice versa (Satriadi,2021a)..
3. Lastly, the bar chart which different in color is chosen for showing a frequency. The reason it is plot with different color because it can show that each information can be stand out of one another which will help audience to differentiate the information. Also, the bar chart is suitable for showing the frequency type information (Satriadi,2021b)..

After the filtering/categorizing is proceed, all of the chosen visualisations will be combined and refined for a visualisation design in the next sheets for the project to answering the following questions:

1. What is the different in the trend of data science jobs posting in each country?

2. What are the factors predicting the estimated salary of the data science jobs in each country?

3. What is the trend of relevant skills require for a specific data science job in each country?

Sheet 2.

On the following section, it will focus on the first initial design coming up from the first sheet as the following:

Layout:

The page will be provided with a drop-down selection of the countries. Once the county on the drop-down bar is selected, the visualisations include the bar charts of top-10 skills for each country, the choropleth map of the number of job posts by each states/cities for each country and the multiple regression plot of salary. All of these visualisations is static.

Focus:

The focus of this visualisation is the interactive features of drop-down bar. It will show that the audience can select the information based on their interaction with the drop-down bar which will show the information by each country.

Operation:

All the interactive visualisations are worked by the drop-down bar. The bar contains all the countries list within it. Once one of the lists is selected, all the visualisations of a certain country will appear in front of the audience.

Discussion:

Pros: It allows an audience to look at all the visualisation in one page, while the audience can also select the information by the country they want to see.

Cons: The audience might confuse which visualisations to look at first sight, and the visualisation is static so it might not stand out when answering the project questions.

Sheet 3.

In this section, it will focus on the second initial design of the 5-design sheet. Here is the third sheet information:

Layout:

This design will be focus mainly on the salary regression plot which is the only static visualisation, while the other visualisations will appear based on the drop-down bar which is the only interactive part of this design.

Focus:

The focus of this visualisation is similar to the previous design. However, instead of the list of countries, it will be the list of visualisation instead that will appeared next to the salary plot.

Operation:

Only the plot of the salary prediction is static without any interactive feature, but all the visualisations which inside the selected drop-down bar is interactive (when the mouse is move on the visualisation it will show the value in a form of hovertext).

Discussion:

Pros: This visualisation will be mainly focused on the salary part of the project, which will help the audience to focus on this part of the project. While the other visualisation is an addition information, and it will not make it hard for user to lose their focus on looking at the visualisations.

Cons: It only focus on the salary part of the project only, which is not good for audience who does not focus only on the job salary.

Sheet 4.

In the following section, it will focus on the third design of the 5-sheet design as the following:

Layout:

The third design will be the visualisations separated by tabs on the left sidebar panel on the left side of the user-interface. The visualisation will be displayed on the user-interface main page along with some text of the finding from data exploration project. In addition, some visualisations will have an addition of drop-down bar with list countries to change the data for the visualisation

Focus:

This design will be mainly focus on the two features: the sidebar tab and drop-down bar on some visualisation. The tab will separate each visualisations while the drop_down bar will change the information of the visualisation based on each country data.

Operation:

All the visualisation is an interactive visualisation, which can be click on or move the cursor to the visualisation to see the insight of the visualisation. Additionally, the text that come along with the visualisation will change based on the drop-down bar accordingly. Also, the visualisations are separated by tabs and all the visualisation will be separated by this order:

1. First tab: Bubble/Choropleth map with bar chart of the type of job posted by each country, and a drop-down bar to select the data based on the selected country which will also providing insight information.
2. Second tab: The regression plot of the salary prediction with some insight information, similar with the second design
3. Third tab: The dropdown which select the visualisation along with the bar chart of top 10 skills requiring for data science job by each country.

Discussion:

Pros: The user will get all the information in an instant, and they can choose which information they want to look by the tab. Also, all the visualisation is interactive so the audience may look at the specific insight on the visualisation.

Cons: It may take a while to run the visualisation.

Sheet 5.

On this sheet, it is a final design based on the previous sheet designs. It will show the layout and design as the following details:

Layout and focus:

The design is based on the fourth 5-sheet design sheet. The focus of the following design wil be focused on the interactive part of the visualisation which are the plot,drop-down bar, and the side bar tab

Operation:

Combination of all design which are separated by the sidebar tab of the dashboard form, each tab which sort in order based on the question will provide all the necessary visualisation for answering the project questions. In additional, some of the features of the visualisation is reduced or modified such as the drop-down bar which show the selected visualisation along with the top 10 skills bar chart will be change to the drop-down with the list of country to change the data of the bar chart instead.

Detail:

This visualisation will be implement by using the R shiny with a several packages such as plotly(make the interactive graph), dashboard(make a dashboard), ggplot(make a static visualisation) and a shapefile related package(getting coordinate of some place for each country).

Based on this, the addition of the shapefile will be required for making a map visualisation (Jumble,2020).

Lastly, it is recommended to look through the top tab to the last tab to get a full insight of the information.

Implementation

In this part, it will provide the process to implement the final design to the real implementation as the following process:

Initial step:

Based on the exploration project, the three datasets will need to be pre-process by using python (the filename for the cleaning part is called “Clean data.py”) which is already done in the exploration project (PlayingNumbers,2020). After that, the cleaned dataset will used for creating the visualisation by using the combination of plotly() and ggplot() libraries to make an interactive plot. Then, the map will be

created by using plot_geo() for information based on the United States dataset, while the other dataset using the addition of shapefile or the maps package to find a city/states coordinate for country which is not United States(Jumble,2020). All the map which are involved with shapefile and map packages are a choropleth map of Australia and bubble map of UK.

Shiny and Shinydashboard: (The template for create the UI and Server for dashboard derived from https://rstudio.github.io/shinydashboard/)

The shiny library such as shiny and dashboard will be used for creating a dashboard environment and rendering all the visualisation on the page. On the sidebar, it will provide the tab to separate each visualisation as the following:

The tab is created by using dashboardSidebar(sidebarMenu(menuItem)) to add tab and putting a tabItem() inside the dashboardBody() to display the content on the body of the dashboard. All of the following method will be put inside the ui code. Then all of the ui will be displayed by the output command in the server function. Lastly, it will be executed by using shinyApp()

Now each tab will be implemented by the following:

The first tab("job trends"): the used of textOutput() for execute text and splitlayout(plotlyOutput()) will be used for create two visualisation output together side-by-side (MLavoie,2015) , and selectInput () for creating a drop-down bar which contain list of countries. After the code in ui part is created, the visualisation will be display once the command observeEvent() is put inside the server with an addition of 'input$...' command on plot and 'output$...' command on the plot to the observeEvent() with an ifelse function to show the visualisation and text based on the country list after selection (Amrrs,2018).

The second tab("salary prediction"): the command in the ui will be similar to the first tab, but it do not include selectInput this time because it will show only the interactive plot of two regression plot with some information. For server function, it do not use the observeEvent() because it do not need any condition selection to show the visualisation. It only need output$ command on the plot only to show the visualisation.

The last tab('skill trends'): the ui and server function is the same as the first tab. The only difference in this tab creation is that it used the plot of top 10 skills not the top of job posts.

User guide

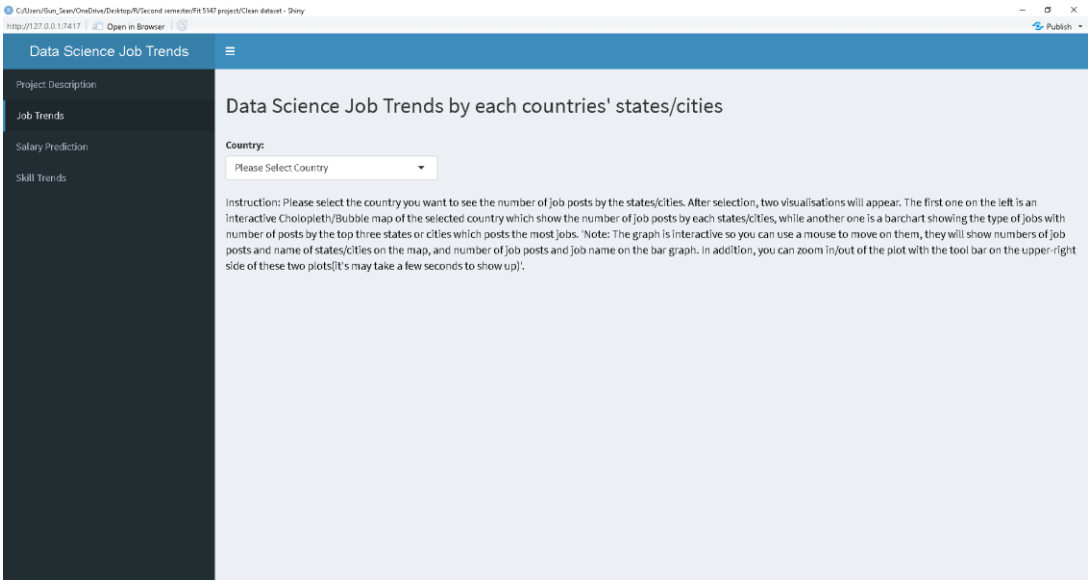
In this section, it will provide on the instruction on how to access the file to the using instruction as the following:

Initial run:

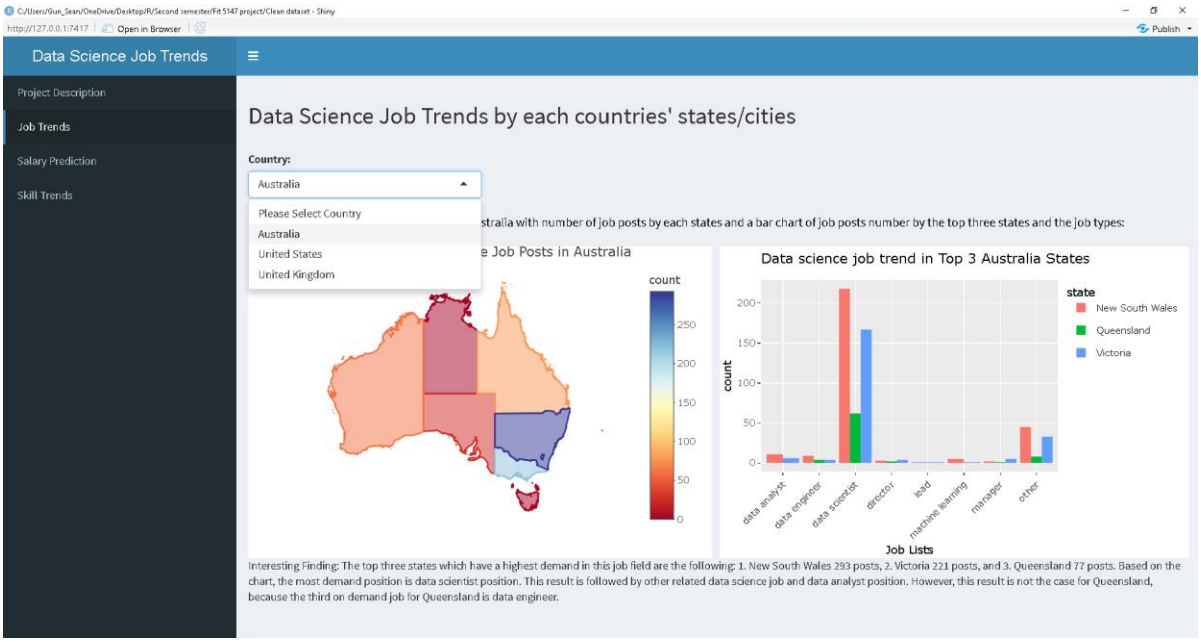
Before running the visualisation, all the r script including ‘Aus_proj.R’, ‘Uk_proj.R’, and ‘Usa_proj.R’ will need to be executed to get all the visualisation before putting all of the visualisation to the R shiny dashboard app. Also, please make sure that all the cleaned csv.file(aus_clean.csv, uk_clean.csv, and us_clean.csv is on the working directory). Once all the necessary plot are obtained, the shiny app code within the ‘proj_app.R’ will be run(I have created a separate folder for the data visualisation project “Visualisation Project” folder. All the other file is all the file involved with both exploration and visualisation project). Here is the instruction on how to use each tab:

Tab1:

The first tab will show up with only a drop down-bar of the country. Once the country is selected it may take a second for visualisation to show up.



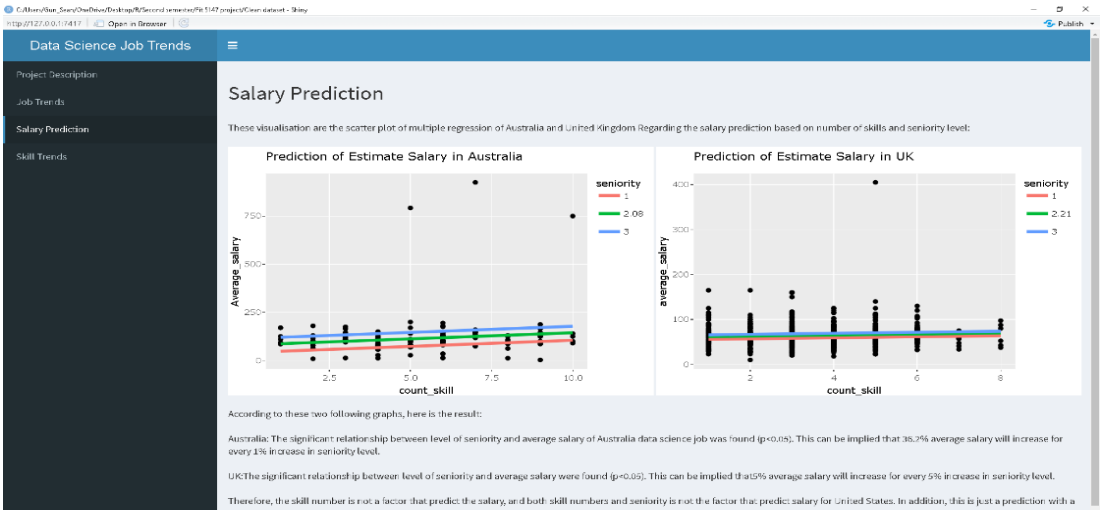
The show up result of the visualisation will provide two visualisation of interactive map and bar plot of the job type post by each country states like below:



When the audience move the mouse cursor to the visualisation the hovertext of visualisation will show up. Also, the audience can zoom in/zoom out by hold the click on the visualisation and drag the part they want for zooming up. In addition, when the audience do not want to see information of one of the state/cities they can click on the legend and the plot represented by that legend will be removed. Lastly, the visualisation can be reset to the first one by click on the tooltip on the visualisation after the zoom in.

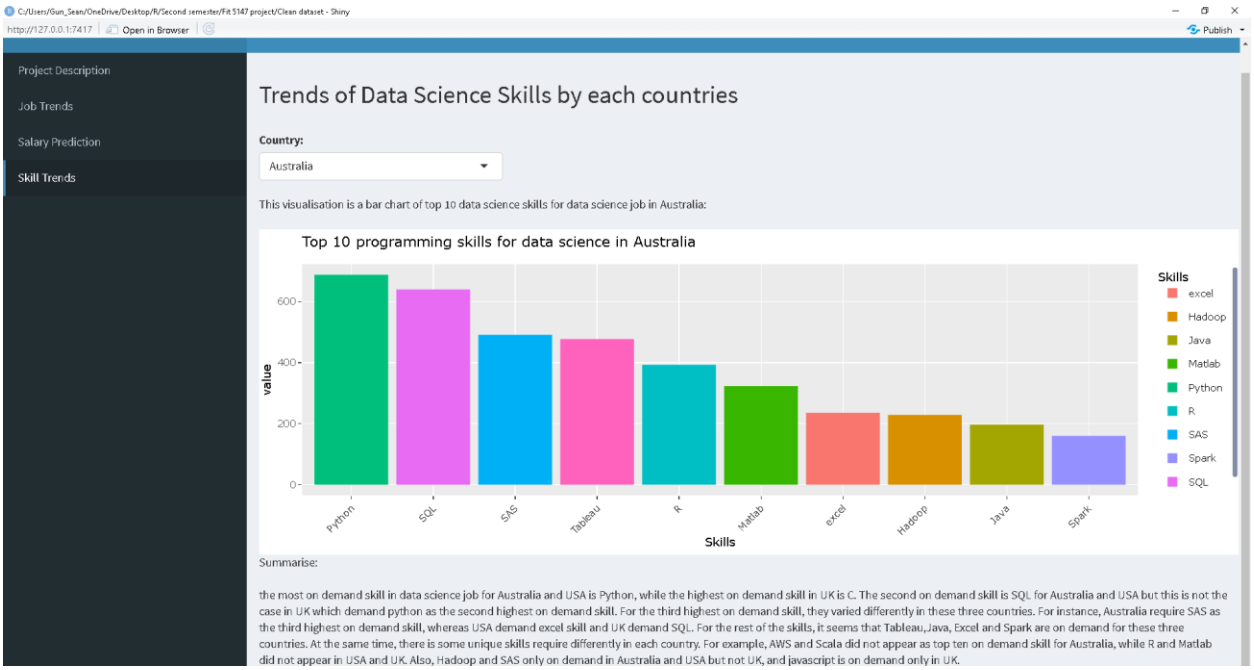
Tab2:

The second tab does not have anything to focus much for the instruction because there is not any transition between graph. The only thing that will be focused on these visualisations are the graph and the text which summarise the graph result based on the data exploration project. Also, similar to the first tab clicking on the legend will remove some unwanted information.



Tab3:

The last tab is similar to the first tab, instead this time there is only one visualisation. Also, all the usage will be the same as the first tab the only different is the text will show only the summary of the exploration result. In addition, the function which click on the legend and the unwanted information is sort out also applied for this visualisation as well.



Conclusion

Based on the result of the visualisation project, it provided me with all the answer to the question. It is expected that the states which posts many jobs for data science position is the place that people want to visited. The only unexpected result found is that UK want more other data science related job than the data science roles. Due to these differences in the job demand, it may affect the skill on demand by the countries. Also, the result of the salary prediction plotting is not as linear as expected although the significant between the predictors and dependent variable is found (Bevans 2020). This salary plot might be displayed as the follow due to the error or the lack of sufficient number of predictors. I have learnt that the interactive plot makes the visualisation look more attractive. Once, the interactive plot used with the color match with the visualisation it will make more impact on the narrative visualisation. Lastly,if I have to do another visualisation project this time, I will try to find any other data analysis or visualisation technique so I can gain more insight for the following dataset.

Bibliography and Code References (some citation is on the code script/comment)

Amrrs. (2018, February 15). RShiny Generating Dropdown Menu for Plotly Charts - using subelements. https://stackoverflow.com/questions/48805897/rshiny-generating-dropdown-menu-for-plotly-charts-using-subelements

Bevans, R. (2020, February 20). *An introduction to multiple linear regression.* https://www.scribbr.com/statistics/multiple-linear-regression/

Bevans, R. (2020, February 25). *A step-by-step guide to linear regression in R*. https://www.scribbr.com/statistics/linear-regression-in-r/

Bubble map with ggplot2. (n.d.). https://www.r-graph-gallery.com/330-bubble-map-with-ggplot2.html

Choropleth Maps in R (n.d.). https://plotly.com/r/choropleth-maps/

Clarke, M. (2021, February 8). *UK Data Science Jobs dataset*. https://www.kaggle.com/devario/uk-data-sciencejobs-dataset?select=deduped-jobs.csv

Divibisan (2018, August 7). *How to remove all whitespace from a string?* https://stackoverflow.com/questions/5992082/how-to-remove-all-whitespace-from-a-string

Interactive web-based data visualization with R, plotly, and shiny. (n.d.) https://plotly-r.com/maps.html

Jaap (2014, September 4). *Reorder bars in geom_bar ggplot2 by value*. https://stackoverflow.com/questions/25664007/reorder-bars-in-geom-bar-ggplot2-by-value

Jdharrison. (2014, June 5). *Change the color and font of text in Shiny App*. https://stackoverflow.com/questions/24049159/change-the-color-and-font-of-text-in-shiny-app

Jumble. (2020, September 27). *Plotting a chloropleth map of Australian states* https://stackoverflow.com/questions/64087391/plotting-a-chloropleth-map-of-australian-states

Lathiya, K. (2021, September 8). *grepl in R: How to Use grepl() Function in R* https://r-lang.com/grepl-in-r/

MLavoie (2015, December 21). *How can put multiple plots side-by-side in shiny r?* https://stackoverflow.com/questions/34384907/how-can-put-multiple-plots-side-by-side-in-shiny-r/34392254

Nomilk. (2021, July 24). *Data Science Job Listings - Australia - 2019-2021*. https://www.kaggle.com/nomilk/datascience-job-listings-australia-20192020?select=listings2019_2021.csv

PlayingNumbers. (2020, April 13). *Data Science Salary Estimator* [only the data cleaning part] . https://github.com/PlayingNumbers/ds_salary_proj/blob/master/data_cleaning.py

Rahman, R. (2021, March 9). *Data Science Job Posting on Glassdoor*. https://www.kaggle.com/rashikrahmanpritom/data-science-job-posting-onglassdoor?select=Cleaned_DS_Jobs.csv

R Maps: Beautiful Interactive Choropleth & Scatter Maps with Plotly. (2020, October 1). [Video]. YouTube. https://www.youtube.com/watch?v=RrtqBYLf404

Satriadi,K. (2021a, April 27). *The human visual system.* https://lms.monash.edu/mod/book/view.php?id=8900106&chapterid=959465

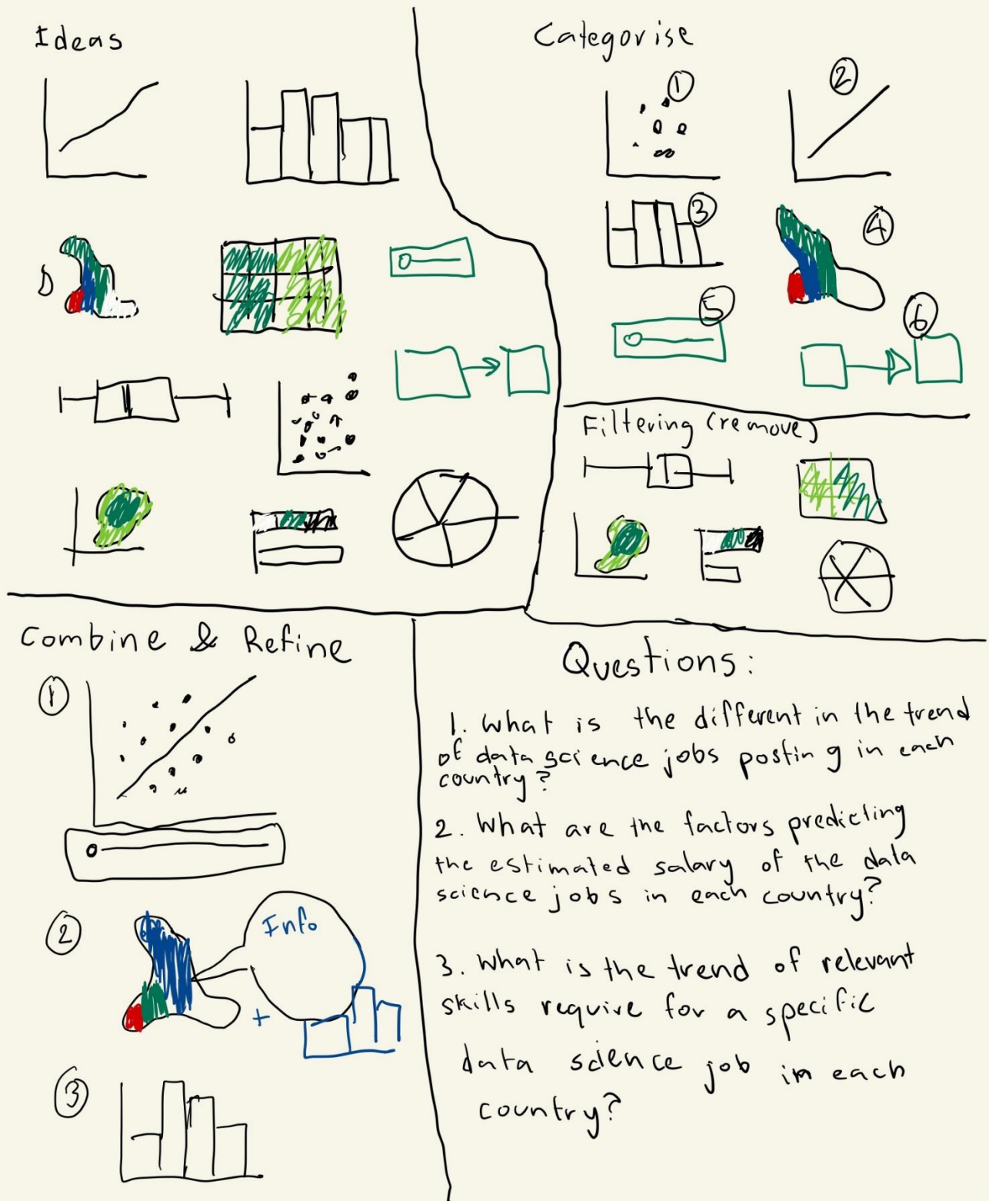
Satriadi,K. (2021b, April 27). *Visual Communication* . https://lms.monash.edu/mod/book/view.php?id=8900106&chapterid=959465

Shinydashboard (n.d.). https://rstudio.github.io/shinydashboard/

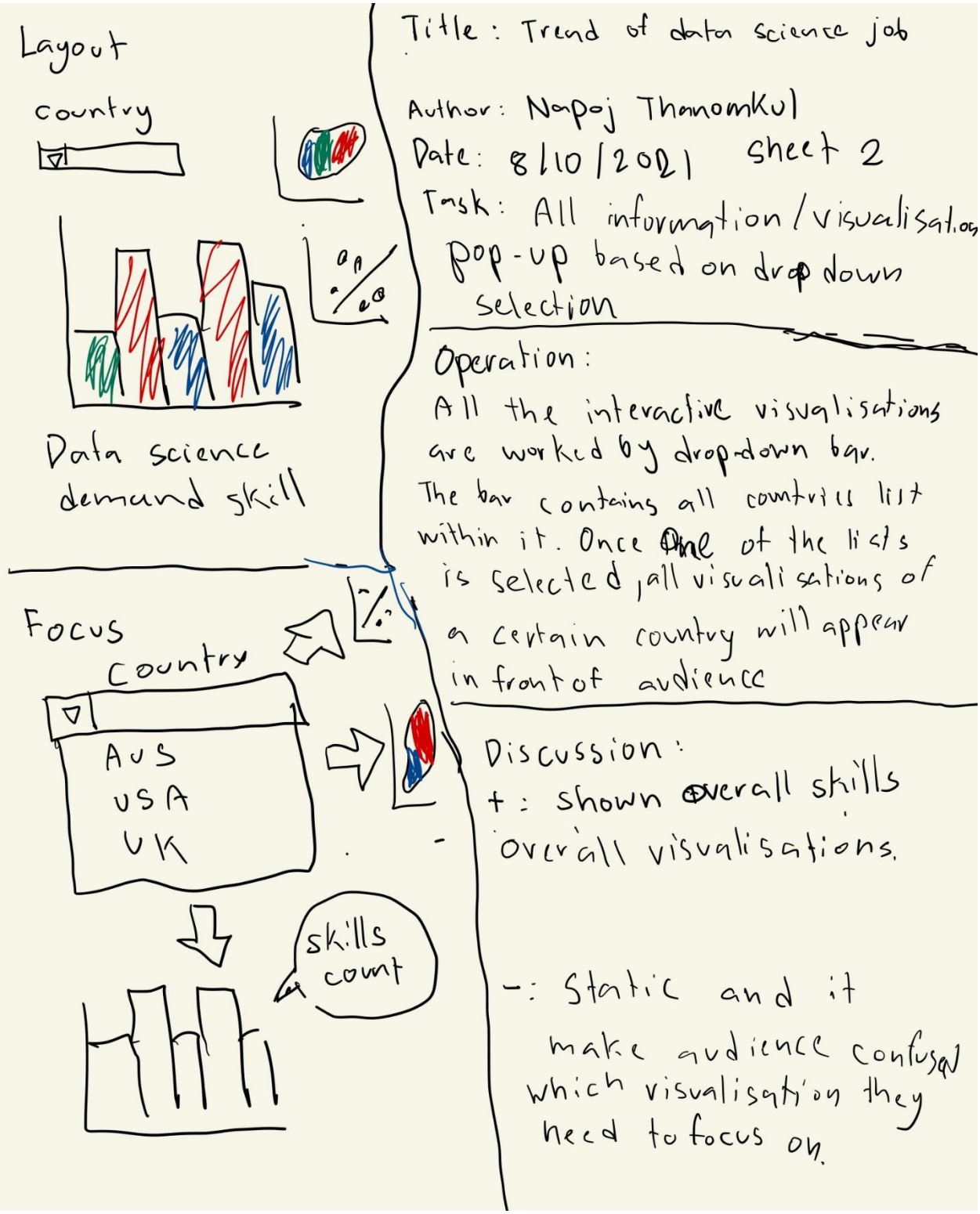
Talat.(2016,January 18). *Count number of rows by group using dplyr*. https://stackoverflow.com/questions/22767893/count-number-of-rows-by-group-using-dplyr

Appendix

First sheet

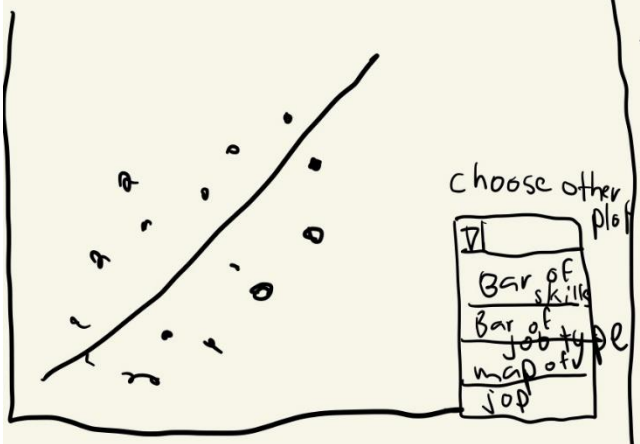


Second sheet



Third sheet

Layout

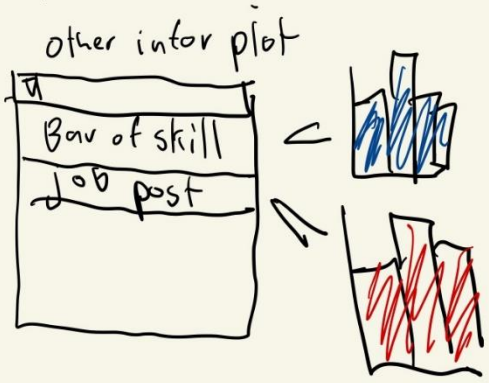


Title: Trend of Data Science Job
Author: Napoj Thanomkul
Date 8/10/2021 sheet 3
Task: Identify factors for Salary prediction and other plot

operation:

Only the plot of the salary prediction is static without any interactive feature, but all the visualisations which inside the selected drop-down bar is interactive (when the mouse is move on the visualisation, it will show the value in a form of hover(x,y))

Focus



show list of visualisation.

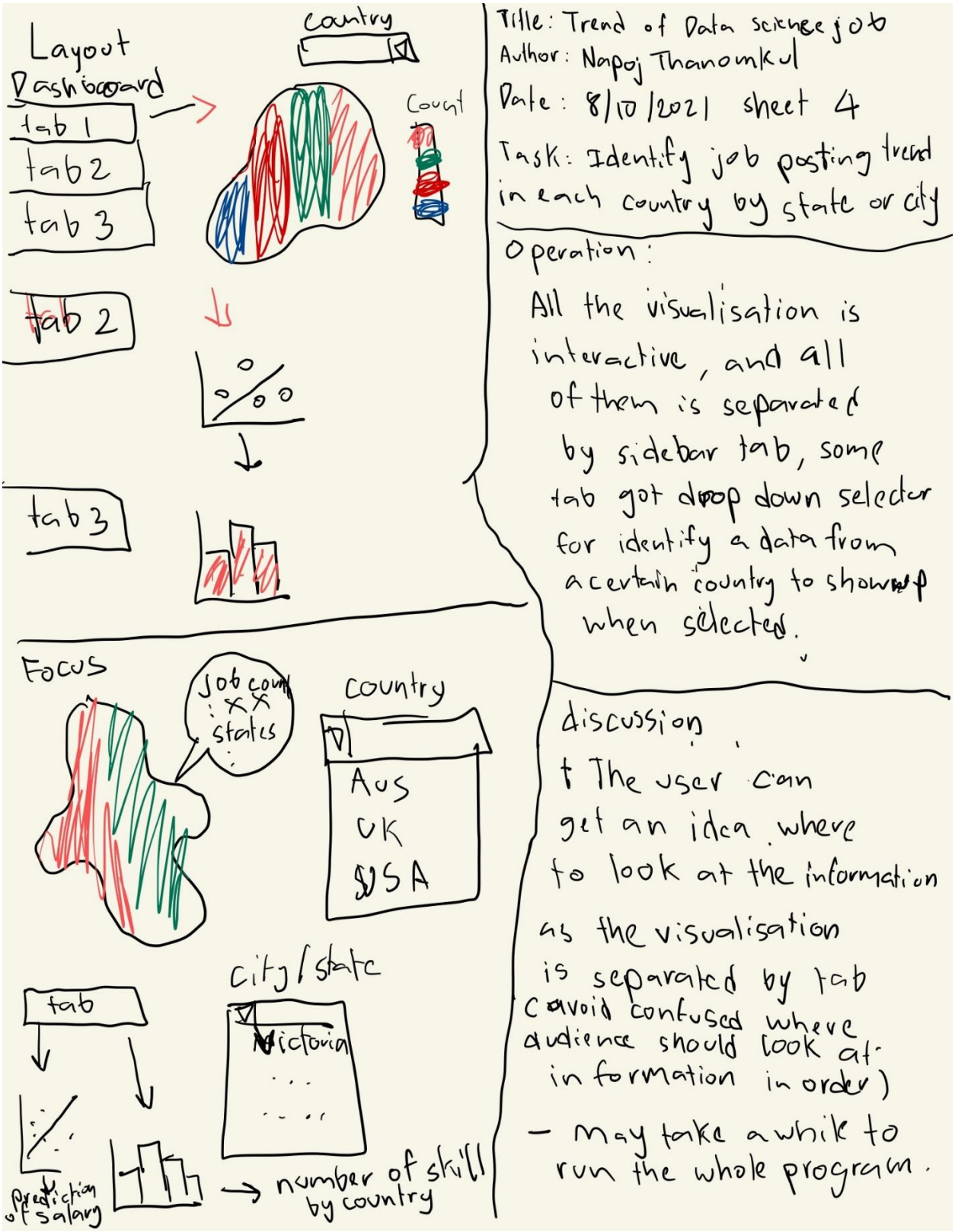
Discussion:

+: Easily defy the prediction trend of salary

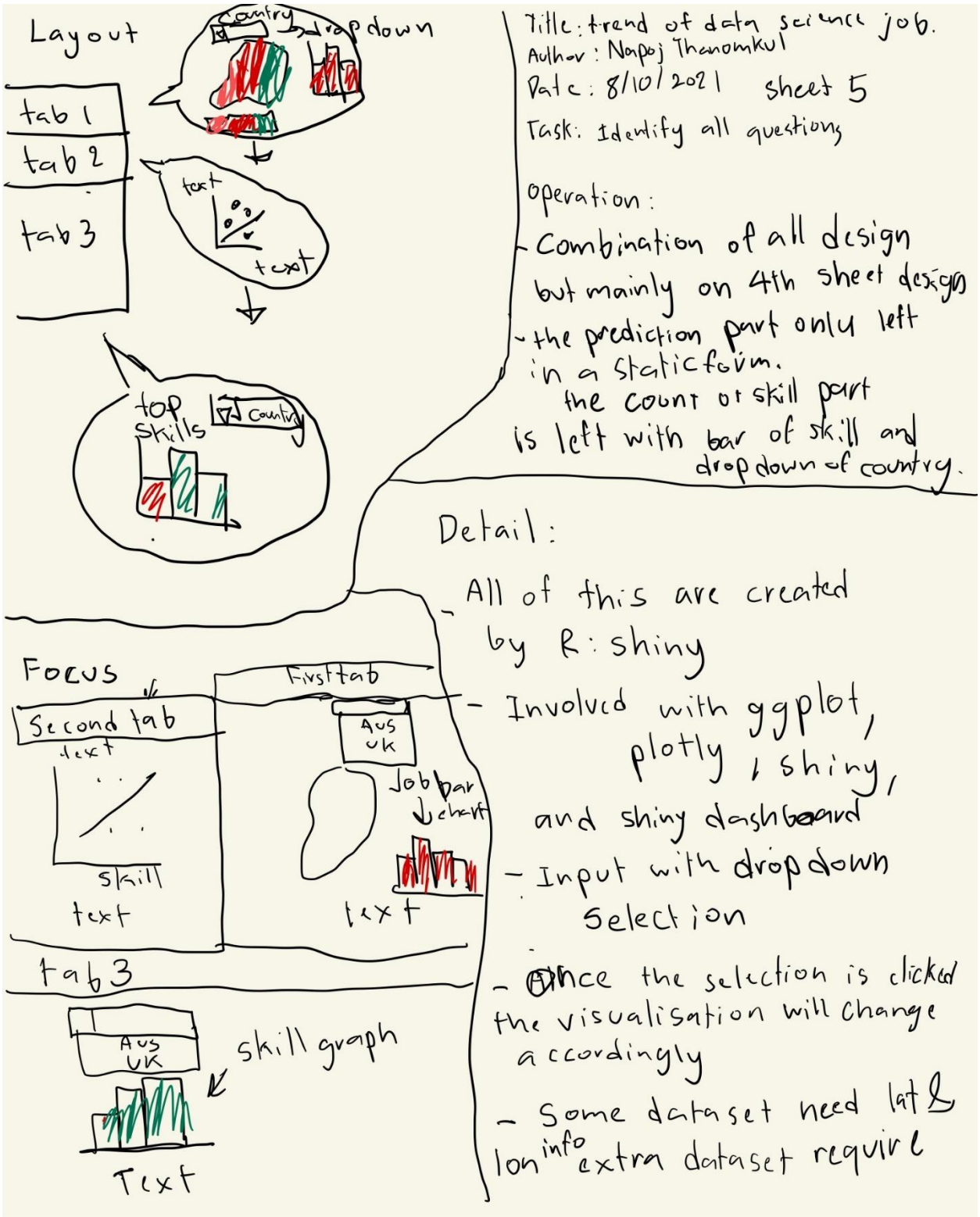
+: Give overall review

-: Only attract the audience who only interested in salary part only.

Fourth sheet:



Fifth Sheet


Layout
country
drop down
tab 1
tab 2
tab 3
text
text
top skills
country
Focus
Second tab
text
skill
text
Firsttab
Aus UK
Job bar chart
text
tab 3
Aus UK
skill graph
Text
Title: trend of data science job.
Author: Napoj Thanomkul
Date: 8/10/2021 sheet 5
Task: Identify all questions
operation:
- Combination of all design but mainly on 4th sheet design
- the prediction part only left in a static form.
the count of skill part is left with bar of skill and drop down of country.
Detail:
- All of this are created by R: shiny
- Involved with ggplot, plotly, shiny, and shiny dashboard
- Input with drop down selection
- Once the selection is clicked the visualisation will change accordingly
- Some dataset need lat & lon info extra dataset require