FIT 5147 Semester 2021

Visualisation Project

External Factors Affecting Average California SAT Scores

Student ID:27181944
Name:Weisheng Jiang
Tutors: Mohit Gupta, Vaibhavi Bhardwaj

Introduction

This report will describe exactly what I want my narrative report to be. It accurately describes the message I wanted my narrative visualization to convey in the last data exploration project and my R shiny. The purpose of this report is to design a new visualization of the project R shiny, which was used to compare different external influences with the average SAT score in California. The purpose of this report is to explain the process of creation, which includes design, implementation and user guide. The audience is those who know the California counties and have a basic understanding of charts.

Background

It is well known that different environments can have an impact on students. Our aim this time is to find the relationship between each external factor and SAT scores of California students through exploration and visualization in the given data.(SAT is the American College Entrance Examination)

Question Explored by Visualization

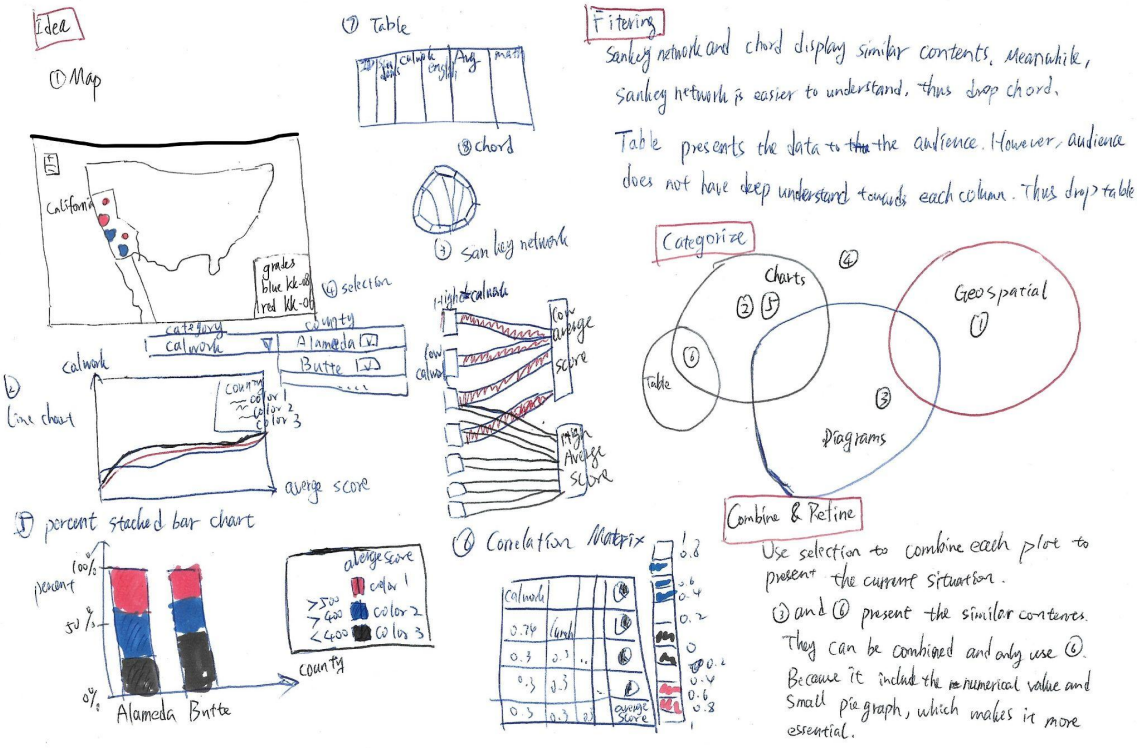
1.Can you display the distribution of scores of different counties in different grades?
2.For variables calworks,lunch,computer,expenditure,income and english, which has higher effects on the scores?

The project has four main parts.
First, the project introduces the design of the visualization, including a description of the visualization design process and a summary of the five design sheets.
Second, the project shows the implementation of R shiny. This section contains the tools of R shiny and their reasons and challenges.
Third, the project illustrates the user guide. It explains how users can use the page and visualize the end result.
Finally, the project summarizes my findings and what I have achieved, as well as the limitations of the product and some of the challenges of this visualization project.

Design

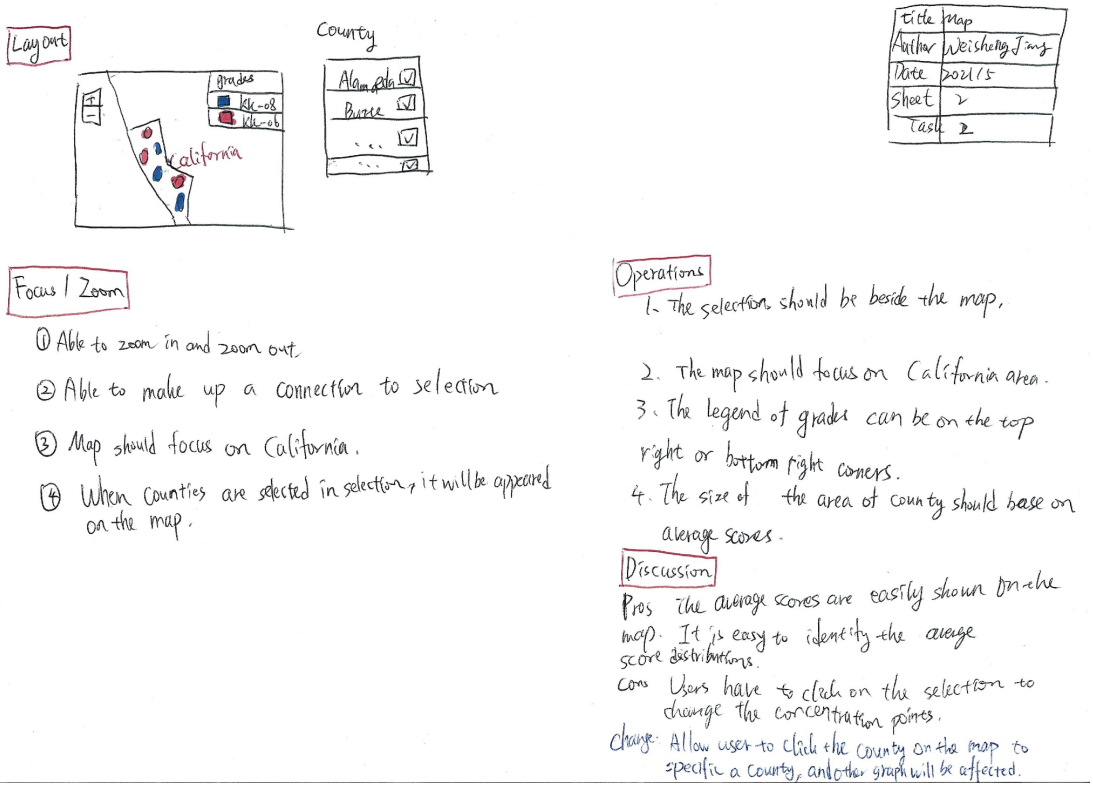
Sheet 1: Brainstorming

The sheet 1 contains the ideas and filter,categorize, combine and refine.
Idea:
To present the data, the idea is to use the table which shows data directly.
To present the distribution of the grades of different counties in California, the good idea is to use the bubble map. Each bubble represents a county, and the size of the bubble depends on the average scaled average score.
To present the relationship between different factors, the idea is to create a plot to do the comparison. It can be a line chart, percentage stacked bar chart, correlation matrix, sankey network, chord, or table.
To connect them together, the idea is to create a selection, e.g.select bar, checkbox.
Filtering:
After comparison, table, sankey network, chord are removed. Because the stack bar, sankey network, chord shows similar contents. Meanwhile, the The bars and selection are more interactive and more visible. Bars can contain multiple counties, but the sankey network and chord will be confusing. Table is unfriendly for the audience.
Categorize:
Chart: bar chart, line chart. Geospatial: Map. Selection: select bar, checkbox.
Combine and refine:
The selection will be used to link different charts and the map together to increase the interactions. The idea of percentage stacked bar chart, sankey network, chord will be combined together.



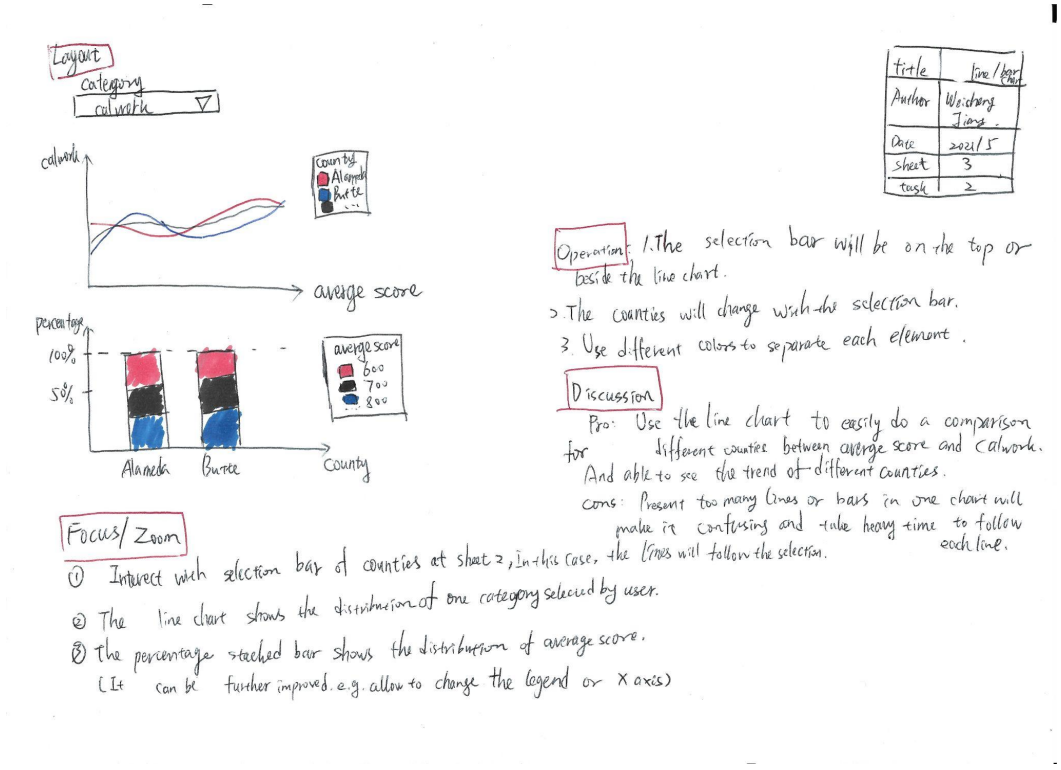
Sheet 2: Design sheet 1

This sheet contains layout, focus/ zoom, operations, discussion.
Layout:
It shows the idea of the map. It is a bubble map of the average score of California. Different colors represent different grades. The bubble size shows the size of the average score. Near the bubble map, there should be a checkbox which allows the user to decide which county/ counties to present.
Focus/Zoom:
Allow users to zoom in or zoom out.
Able to tick the checkbox to decide which county to show.
Map will initially focus on California in the US.
The county checkbox should have effects over the whole pages. Other charts will be rebuilt based on the current counties.
Operations:
The checkbox is near the bubble map, so that users can easily change the county. It can be done by the checkbox or checkbox group in the R shiny.
Map will be generated using the leaflet. It also allows the user to change the zoom.
Map will initially focus on California in the US. When creating the leaflet, initial coordinates should be set up in California.
Discussion:
Pros: the average score is easily shown on the map. It can show the distribution of different average scores of grades.
Cons: unable to identify which county it is. The average scores are close, unable to see differences.
Changes: add up a tip in the bubble map, showing the detailed information of this county. Use the scaled value as the point of average value to make it easier to find differences.



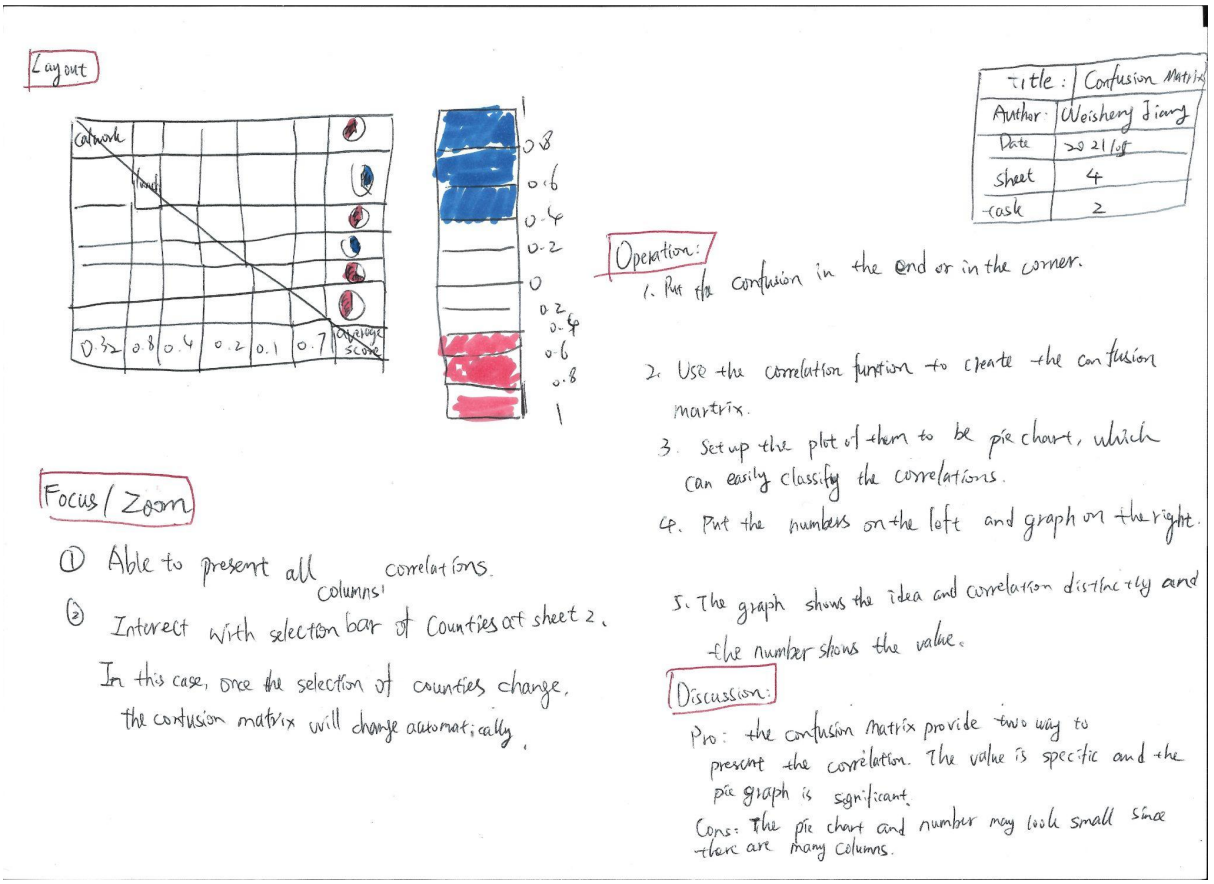
Sheet 3: Design sheet 2

This sheet contains layout, focus/ zoom, operations, discussion.
Layout:
It shows the layout of the line chart, percentage stacked bar chart. Near the line chart and bar chart, there should be a select bar which allows the user to decide which category as the x axis. The line chart will show the relationship between the x-axis and average score. (The idea was forced to change to a bubble chart because the number of counts was too large and there were not enough data samples. As a result, the visualization is ineffective and confusing, far worse than bubble charts.)
Focus/Zoom:
Interactive with the checkbox of counties in the sheet 2.
Create the new select box to choose the x-axis. So it is easier to see the relation to the average score.
The percentage of the bar chart shows the distribution of the level of average score.
The y-axis can be further improved, instead of using the default average score. It can be set to be a select bar of y-axis with default value average score.
Operation:
The select bar will be on the top or near the line chart.
Use different colors to present counties in a line/ bubble chart.
Use color to represent different levels in the bar chart.
Discussion:
Pro: Use bubble charts to easily do the comparison of each point, and capture the trend of the bubble chart. Bar chart allow users to identify the distribution of level of a specific category, especially the average score.
Cons:Bubbles are dispersed. Too many counties makes the line chart messy.
Changes: Use bubble chart, create the trendline, add the y-axis, use plotly.Plotly can greatly improve the interactivity (but can't make an overall line, percentage bar chart directly).



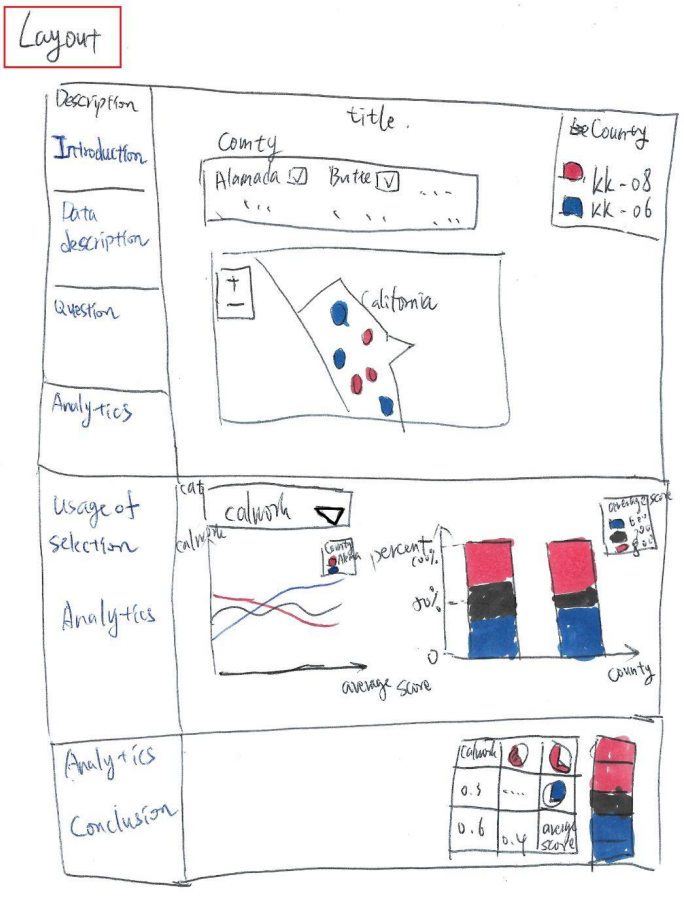
Sheet 4: Design sheet 3

This sheet contains layout, focus/ zoom, operations, discussion.
Layout:
It shows the layout of the correlation matrix. The matrix should be divided into 2 parts. Bottom left to present the correlation in numeric format. The top right shows the correlation in graph format. My idea is to show in the pie chart, it looks more pretty and effective. Near the matrix, there should be a legend which shows the meaning of each color.
Focus/Zoom:
Able to present correlation for all columns.
Interactive with the checkbox of counties in sheet 2, select bars in sheet 3. Once the selection changes, the matrix will be automatically updated.
Operations:
Put the matrix under the bubble chart.
Use the correlation function in R to create the correlation matrix.
Specify the color, graph type as pie chart.
Specify the bottom left to be number, top right to be pie chart.
Discussion:
Pros: The top right graph gives users the idea of correlation. Meanwhile, the bottom left shows the detailed values, which is very detailed and significant.
Cons: The pie chart may look small because it needs to contain many pie charts.



Sheet 5: Final Design sheet

This sheet contains layout, focus/ zoom, details.
Layout:
This layout shows the overview of the visualization. On the left panel, it will show the description of the introduction, questions, and analytics of the graph. On the right panel, it will show different layouts, e.g. map, checkbox, charts, and matrix.
Focus/Zoom:
Different plots will be connected together using the selection, e.g. checkbox, select bar.
Any changes in the county selection will automatically update to the map, charts, and matrix.
Any changes in the x axis, y axis will be updated to the charts and matrix.
Details:
To plot the charts, low level is using ggplot2, high level is using plotly.
To do the map visualization, leaflet is the good choice.
To do the matrix, correlation in R is the good idea.
To set up the frame of the panel, side panel and main panel is the good choice.



title	Sheet 5 layout of whole pages
Author	Weisheng Jiang
Date	2021/05
Sheet	5
task	2

Focus/Zoom

① should have good interaction between the various diagrams. when the user change the county by selection bar or clicking the point inside the map, the other diagrams change.

Operation

① use R shiny/D3 to achieve interactivity.
② use different graph to do the visualisation.
③ use side pannel to set up the left side description.

Details

① plot: ggplot2, leaflect, correlation function
② Interactivity: R shiny/ D3
③ frame: In R shiny: side pannel, main pannel
In D3: sidenav, main by div

Implementation

This section contains a high-level description of the implementation.

Libraries used and reasons (all belong to R package, using in R studio)

library(shiny)
Shiny is a web application framework of R. Shiny allows users to build the interactive web with R coding. It provides the function to define the ui and server and run it in a web frame. It also allows users to combine different plots and descriptions. It provides the side panel in ui and the render function in the server.

library(leaflet)
Leaflet is an open source JavaScript library for map products. It is popular in many languages, e.g. R, python, js. The reason to use it is because it provides a mobile friendly map, and allows users to make up the circle markers and the tips.

library(ggplot2)
Ggplot2 is a system for creating graphics e.g. bar chart, line chart, pie chart. It helps users simply create complex plots from a dataframe. Compared to the built in function in R, it allows users to change many variables and give more options.

library(RColorBrewer)
RColorBrewer is a package to create good looking color palettes. I use it to specify the colors for different counties in charts and colors for different grades in the bubble map. Also for the pie chart in the matrix, it is used to define the color range from negative to positive of the pie chart.

library(dplyr)
The dplyr is a package which provides a set of tools for manipulating the dataset. It provides a set of verbs called to help solve data manipulation. For my work, I use the mutate() and select() functions.

library(corrplot)
Corrplot is a package for a graphical display of a correlation matrix and confidence interval. It is used to generate the correlation matrix.

library(shinydashboard)
Shinydashboard is a package to help users using Shiny to easily generate the dashboard. Instead of using a shiny dashboard, I use the combination of a shiny dashboard and fluid page. Shinydashboard is used to initialize the current environment so that I can use box function from shinyWidgets.

library(shinyWidgets)
ShinyWidgets is a package that allows users to custom input controls and user interface components for Shiny. The box function inside the shinyWidgets is the main reason I import this package.

library(plotly)
Plotly is a package for creating interactive web-based graphs via open source Javascript plotly. It is used for creating bubble charts and percentage stacked bar chart. Also it includes the renderPlotly, which allows users to show plotly graphs in the R shiny pages.

library(broom)
Broom is the package used for converting statistical objects into tidy tibbles. As a result, they are easily combined. I use the augment function from it, to get the fitted data from a linear regression.

User guide

This part introduces the instructions for viewing and exploring the visualization.

Layout 1: Bubble Map

- Side Panel:

The side panel makes a description of introduction, meaning of factors, question explored by visualization, analytics of bubble map and the checkbox of grade and county.

- Main Panel(as figure 1):

Top left corner has the “Selected Counties” box that shows the selected counties as No.2. Meanwhile, all contents are kept in the box, and it enables users to minimize them when they click the top right corner as No.1.

Bottom left is the bubble map made by leaflet. Each bubble represents one grade of a county. It has three functions. First, it allows users to zoom in or out as No.7. Secondly, when the user's mouse is over the bubble, it will show the county name as No.8. Finally, when clicked, it will show the detailed information about the current county as No.3.

Top right is the grades select bar. It allows users to change grades as No.4. It is used to handle the overlapping for bubbles with different grades but the same county. Also it will affect the other layouts’ charts and matrix as a filter.

Bottom right is the county selector. It allows users to choose “All” to get all counties as No.5. Another function is that it allows users to choose specific counties they wish to see in the map and the charts in other layouts.

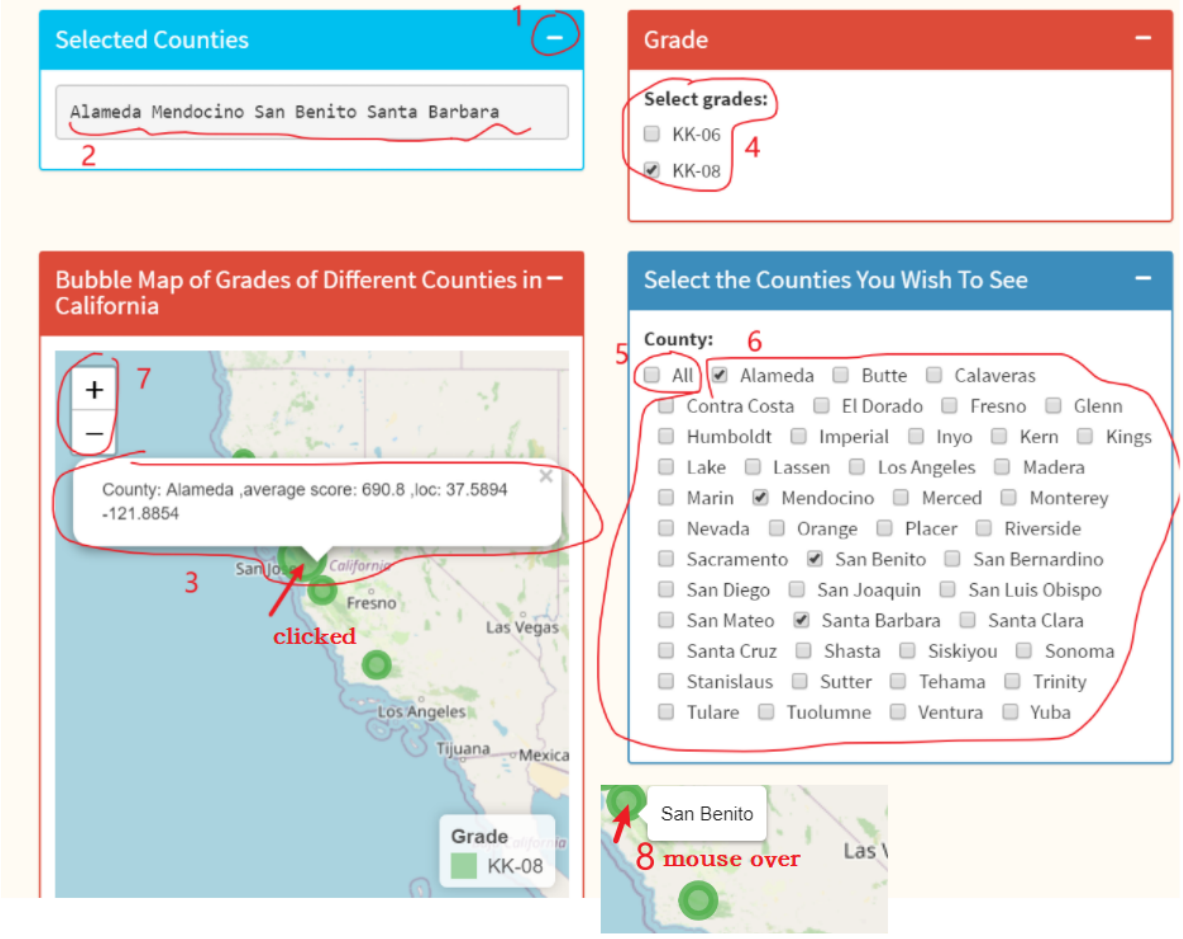


Figure 1: Main Panel of Layout of Bubble Map of Different Counties in California

Layout 2: Bubble Chart and Percentage Stacked Bar Chart

- Side Panel:

The side panel makes a description of select bar of X-axis, select bar of Y-axis, bubble chart of Y-axis on X-axis, e.g.(y: average score on x: calworks), and percentage stacked bar chart of Y-axis, (y: average score).

- Main Panel (figure 2, 3, 4):

Top left side has a select bar for X-axis as No.1. It allows users to drop the bar and choose their preferred X-axis. It is used to control the X-axis for the bubble chart.
Top left side has a select bar for Y-axis as No.2. It allows users to drop the bar and choose their preferred Y-axis. The default Y-axis is average score. It is used to control the Y-axis for the bubble chart and percentage stacked bar chart.
Bottom left is the bubble chart. It is used to show the relationship of different factors towards average score. It has many functions. Firstly, when the mouse is over the bubble, it will show the information of the current county in format of "(x value, y value) county" as No.3. Secondly, it allows users to use the drop down bar to see different counties as No.4. Thirdly, the bubbles is based on the county and grade checkbox in the previous layout.
Bottom right is the percentage stacked bar chart it shows the percentage of average score (can be changed into other factors). One of the good functions is that when the mouse is over the bar chart, it will show the information in format of "county Amount percentage". When there are many counties in the chart, it helps users to identify each bar.
Meanwhile,the bubble chart and percentage stacked bar chart have many **common functions** as shown in **figure 3,4**. Firstly, the plotly allows users to move inside the plot using "pan" as No.7. Secondly, it allows highlighting elements (e.g. bubble, bar) as No.6. It makes comparison easier. Another function is that plotly allows users to compare data on hover,which allows users to compare y values with the same X value as No.8 and No.11. Fourthly, it allows the user to toggle spike lines, which shows the horizontal and vertical line of the current bar. It is useful for doing horizontal comparison. Final function is filtering, the label in the legend is like a switch, it allows the user to filter the legend by simply clicking the legend label. Click to hide and click to show again. It makes comparison between two specific categories easy and fast to implement.

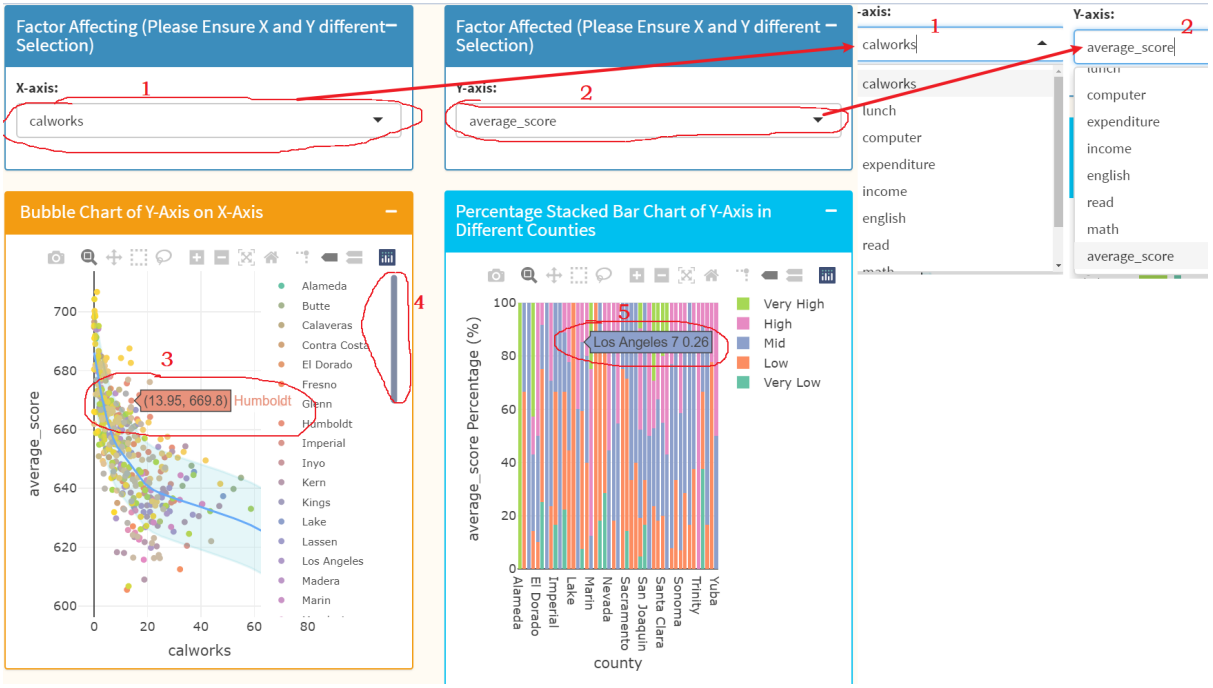


Figure 2: Layout of Bubble Chart and Percentage Stacked Bar Chart

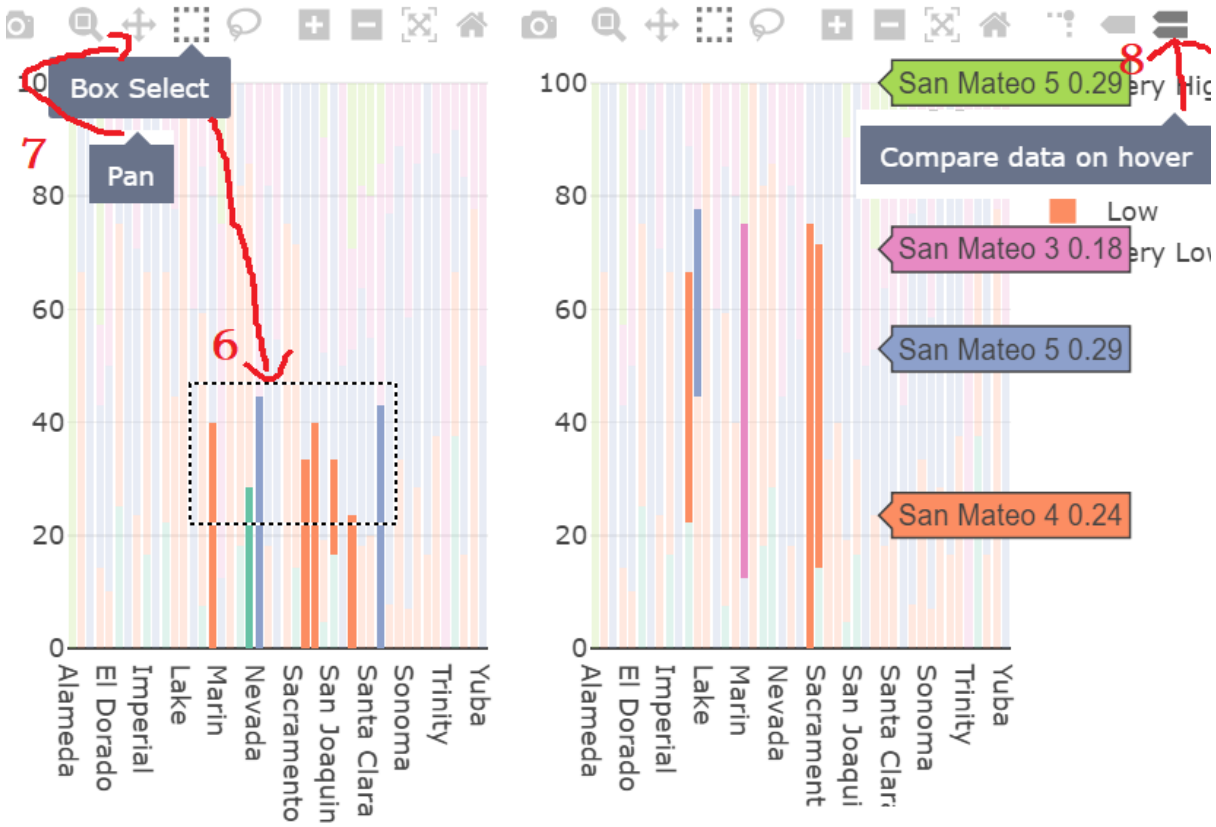


Figure 3: Common functions based on plotly

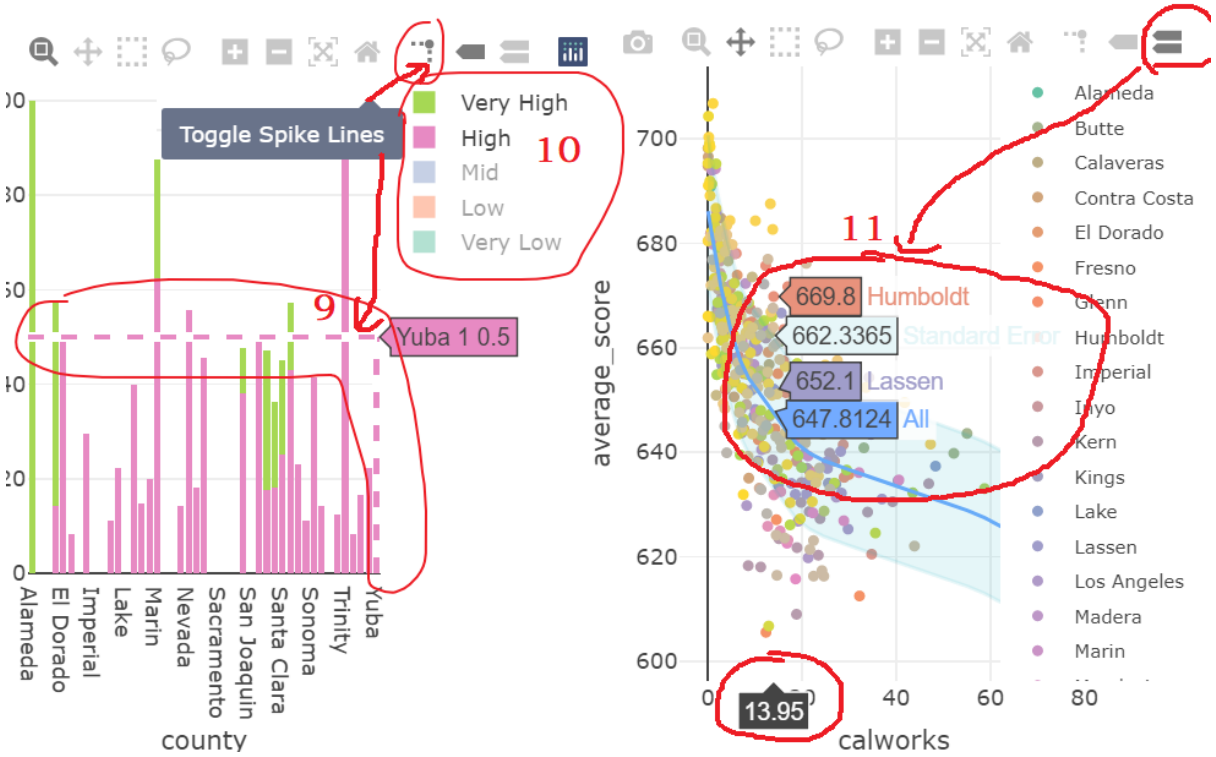


Figure 4: Common functions based on plotly

Layout 3: Correlation Matrix of Different features

- Side Panel(Image of it inside the Appendix within the full image):

The side panel makes a description of correlation matrix and explanation of correlation of average score.

- Main Panel(figure 5):

Left side is the correlation matrix of different features. It shows the correlation between different features. Our focus is the last column as No. 1and last row as No.2. The last column shows the pie chart of correlation between average score and other features. The last row shows them in numbers. Although the correlation matrix is fixed, the value can be changed. It will be controlled by the county checkbox and grade checkbox in layout 1. The value will automatically update and generate a new correlation matrix. So that it can be used for specific counties and grades.

The right side is the conclusion for this exploration, I use it to answer the explored questions and share my ideas to the audience.

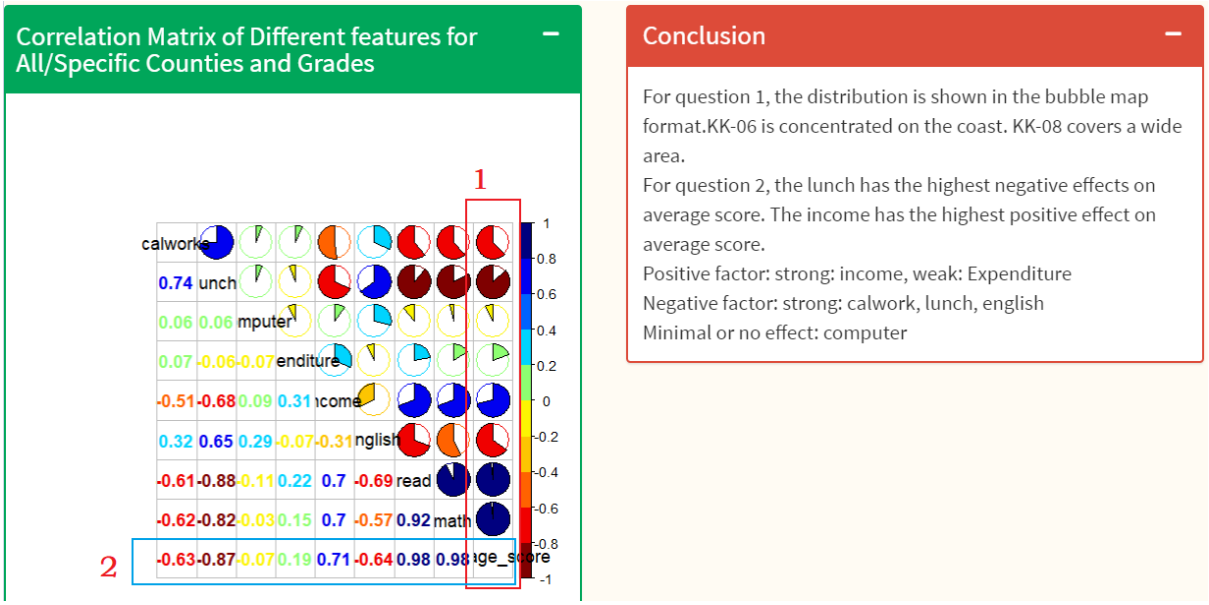


Figure 5: Layout of Correlation Matrix of Different features

Conclusion

Finding:
First of all, answering the questions.
For question 1, the distribution is shown in the bubble map format.KK-06 is concentrated on the coast. KK-08 covers a wide area in California.
For question 2, lunch has the highest negative effects on average score. The income has the highest positive effect on average score.
Positive factor: strong: income, weak: Expenditure
Negative factor: strong: calwork, lunch, english
Minimal or no effect: computer
Personal findings:
When they or their families have higher incomes, they can focus on their schoolwork with few distractions by the external environment, which has a positive impact on students.
When students have financial difficulties, they worry about lunch, living (calworks), and these can negatively affect them and make scores slip.
The level of mastery of English will also affect the score. The more English learners there are, the lower the average score of this county. I believe this is due to the fact that english learners do not fully understand the teacher's content and that this is not conducive to academic communication among classmates.

Achieved:
I create 3 layouts. The main panels of 3 layouts include the bubble map using leaflet, two checkboxes, one selected input print, two select bars for x and y, one bubble chart, one percentage stacked bar chart generated by plotly, one correlation matrix, and a conclusion box.
The side panel includes the description of introduction, factors, explored questions, and the description for bubble map, checkboxes, select bars, bubble map, percentage stacked bar chart, correlation matrix.
I put all of them into a frame of a combination of fluid page and shiny dashboard. I use the layout and panel from the fluid page and box from the shiny dashboard.

Reflection/learnt from this project:

- Hindsight

At the beginning, I am trying to build in html using D3. It takes me many days and the only output I can get is the leaflet bubble map. The hindsight of it is that we should start with something we know if it is hard. Afterwards, I go back to R shiny, which makes my processing faster.
Another is to try to draw out what you really expected, e.g. five sheets. It is really helpful, especially sheet 1 and sheet 5. Sheet 1 gives me the sketch /appearance of the plots. And sheet 2 to 4 gives the focus and improvement I can make based on the basic sketch. The sheet 5 gives the sketch of the frame. So when I am confused about the appearance, it provides constructive advice.

- Difficulty

At the beginning, I tried to plot the bubble chart and percentage stacked bar chart in the ggplot. It is easy to do that. After that, I want to improve the interaction, so I go to plotly.

Plotly is harder compared to ggplot2. It does not provide enough options as ggplot2. The description of it in the plotly website is in low detail. It gives sample codes with few meaningful and detailed explanations. The plotly is difficult for me to achieve the expected result. Plotly does not provide the overall trendline, and options to do the percentage stacked bar. In this part it takes a long time to handle it.
Another difficulty is the combination of fluidpage and shiny dashboard. Since I want all selections to be used in different plots to increase interactions, I want them inside one page rather than multiple pages. At the same time, I want to use the box function inside the shiny dashboard, so that I can present two plots in one line. At the end I handle it. As a result, users are able to minimize every box in side panels and main panels they do not want. This greatly improves the interactivity.

Overall, I learn more of R shiny by facing these challenges.

Bibliography

DataCamp, useShinydashboard: Use 'shinydashboard' in 'shiny'
https://www.rdocumentation.org/packages/shinyWidgets/versions/0.4.4/topics/useShinydashboard

Excel2R, 2018 Mar, R stacked 100% bar chart
https://www.youtube.com/watch?v=7d7ftyLExPA&t=3s

Plotly, Legends in R
https://plotly.com/r/legend/

Salim-B,2016 Nov, GitHub, Add legend 'auto' x|y and/or 'container' (x|y)ref
https://github.com/plotly/plotly.js/issues/1199

Tuomastik, 2017 Sep, StackOverflow,Change legend size in plotly chart
https://stackoverflow.com/questions/37245004/change-legend-size-in-plotly-chart

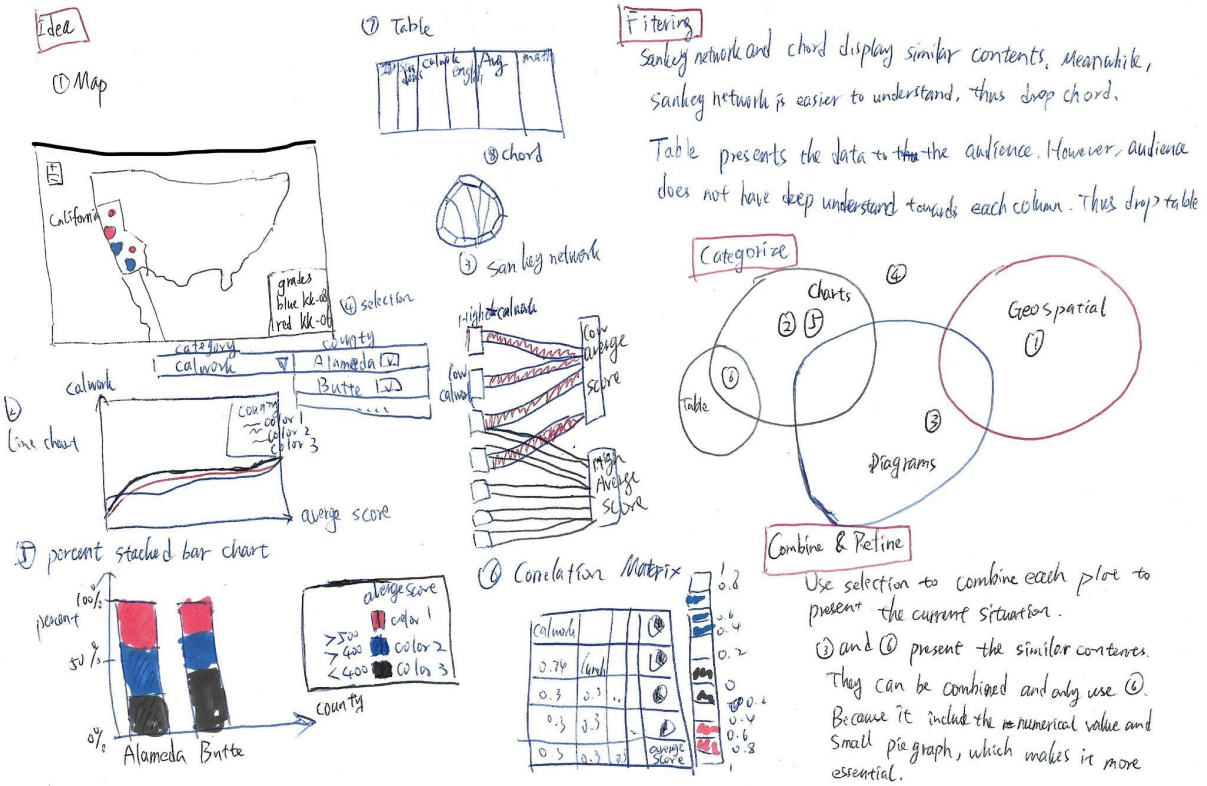
Shiny, Application layout guide
https://shiny.rstudio.com/articles/layout-guide.html

Shinydashboard,Get started
https://rstudio.github.io/shinydashboard/get_started.html

Shinydashboard, Appearance
https://rstudio.github.io/shinydashboard/appearance.html

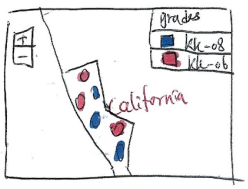
Appendix

5 design sheet
Sheet 1:



Sheet 2:

Layout



County

Title	Map
Author	Weisheng Jiang
Date	2021/5
Sheet	2
Task	2

Focus / Zoom

① Able to zoom in and zoom out.

② Able to make up a connection to selection

③ Map should focus on California.

④ When counties are selected in selection, it will be appeared on the map.

Operations

1. The selection should be beside the map.
2. The map should focus on California area.
3. The legend of grades can be on the top right or bottom right corners.
4. The size of the area of county should base on average scores.

Discussion

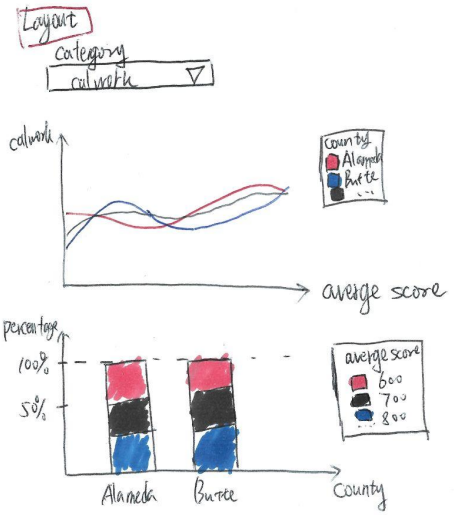
Pros The average scores are easily shown on the map. It is easy to identify the average score distributions.

Cons Users have to click on the selection to change the concentration points.

Change: Allow user to click the county on the map to specific a county, another graph will be affected.

Sheet 3:

Layout



Title	line / bar chart
Author	Weisheng Jiang
Date	2021/5
Sheet	3
task	2

Operations: 1. The selection bar will be on the top or beside the line chart.

2. The counties will change with the selection bar.

3. Use different colors to separate each element.

Discussion

Pro: Use the line chart to easily do a comparison for different counties between average score and calwork. And able to see the trend of different counties.

Cons: Present too many lines or bars in one chart will make it confusing and take heavy time to follow each line.

Focus/ Zoom

① Interact with selection bar of counties at sheet 2, in this case, the lines will follow the selection.

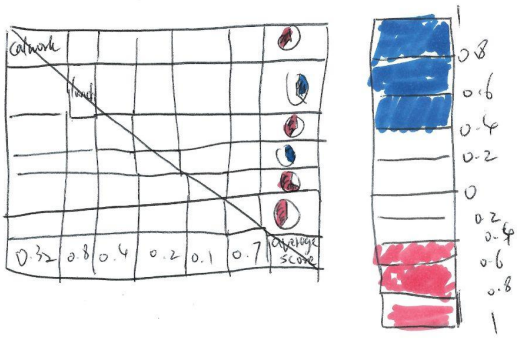
② The line chart shows the distribution of one category selected by user.

③ The percentage stacked bar shows the distribution of average score.

(It can be further improved. e.g. allow to change the legend or X axis)

Sheet 4:

Layout



Title:	Confusion Matrix
Author:	Weisheng Jiang
Date	2021/05
Sheet	4
Task	2

Operation:

1. Put the confusion in the end or in the corner.
2. Use the correlation function to create the confusion matrix.
3. Set up the plot of them to be pie chart, which can easily classify the correlations.
4. Put the numbers on the left and graph on the right.
5. The graph shows the idea and correlation distinctly and the number shows the value.

Focus / Zoom

① Able to present all columns' correlations.

② Interact with selection bar of Counties at sheet 2.
In this case, once the selection of counties change, the confusion matrix will change automatically.

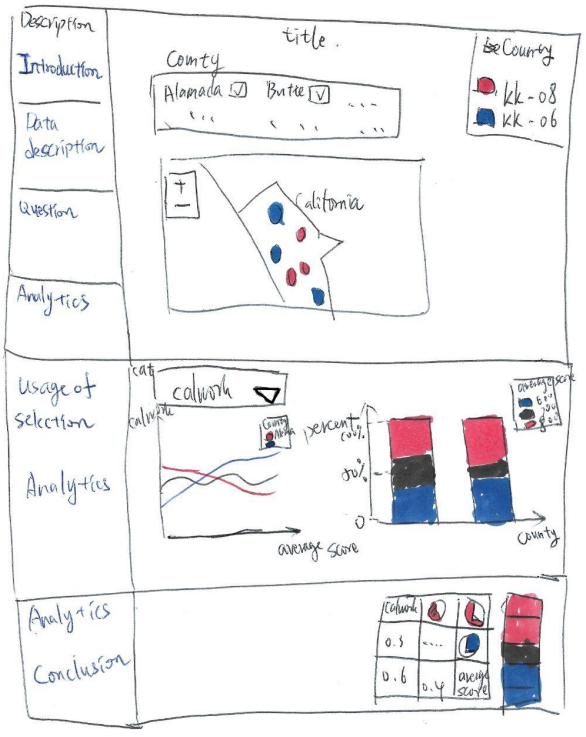
Discussion:

Pro: the confusion matrix provide two way to present the correlation. The value is specific and the pie graph is significant.

Cons: The pie chart and number may look small since there are many columns.

Sheet 5:

Layout



title	Sheet 5 layout of whole pages
Author	Weisheng Jiang
Date	2021/05
Sheet	5
Task	2

Focus/Zoom

① should have good interaction between the various diagrams. when the user change the county by selection bar or clicking the point inside the map, the other diagrams change.

Operation

① use R shiny/D3 to achieve interactivity.
② use different graph to do the visualisation.
③ use side pannel to set up the left side description.

Details

① plot: ggplot2, leaflect, correlation function
② Interactivity: R shiny / D3
③ frame: In R shiny: side pannel, main pannel
In D3: sidenav, main by div

Overview of R shiny