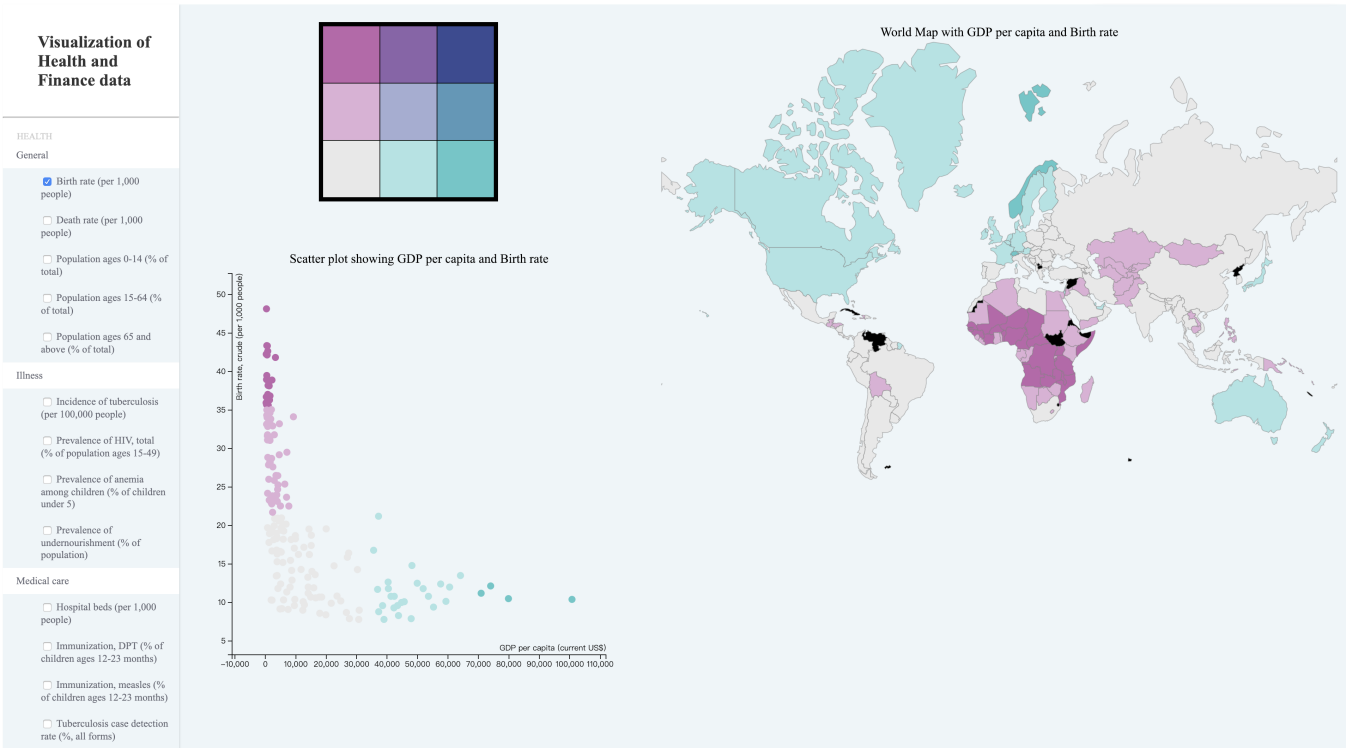
IN4086 Data Visualization InfoVis Project Group 13

Exploring health and finance data

Zheng Liu
4798406
Delft University of Technology
Z.Liu-14@student.tudelft.nl

Tian Tian
4818776
Delft University of Technology
T.Tian-1@student.tudelft.nl

Wenyu Gu
4896246
Delft University of Technology
W.Gu@student.tudelft.nl



1 INTRODUCTION

Health is important and influences all aspects of life. An increasing number of research has been done to examine the relationship between health and other factors, such as environment, health care organizations and health policies. These health-related factors are inseparable from the financial states of a country. They are varied from counties to countries and we believe it is a meaningful to explore the relation between them. Therefore, our project aims to find the relationship between health-related data and financial states of country by using data analysis and visualization techniques. We designed a new visualization web page with the help of D3.js, a JavaScript library for producing dynamic and interactive data visualizations in web browsers[3]. Our work can be divided into three parts: data processing, data visualizing and deliverables preparing. In the project, we designed a selection list, a bivariate color legend, a scatter plot and a choropleth map. We hope this design allows users better explore the health and finance data, and discover some valuable information behind the data.

The structure of this report is organized as follows. Section 2 introduces the domain and data selection. Section 3 explains the reason of visualization design. In Section 4 each visualization design is explained in more detail. Section 5 provides some interesting results based on the visualization. Finally, section 6 concludes the project and describes future improvement.

2 DATA SELECTION AND PROCESSING

2.1 Domain Selection

Health is a topic that always worth to research. Different countries have different ways of treating health of their citizens, and even the same country has different understandings of health at different times. Hence, it is necessary for us to explore the development or change in treating health domain between countries. In addition, analysing health data in one dimension is limited for us to better understand. Some topics in this domain are regularly being discussed and provided with accompanying data visualizations. These visualizations though, are displayed in one dimension. We want to provide people with a tool that allows them to explore the data in two dimensions. Hence, we decided to use data of another dimension that would have an impact on health. In terms of a country, finance is its lifeline, as well as the most directly affecting the level of its health care. There are lots of indicators reflect the financial level of the country. Therefore, our domain choice will include health and finance.

2.2 Data Selection and Processing

World Bank [6] is an online database that is open to all users. It provides data for some world development indicators such as health indicators and agricultural indicators. We used health and financial data from this database. The dimension of data we selected is shown in Appendix A.

All the data file downloaded from World Bank is in csv format. Data of different dimensions is stored separately. We noticed that each data file records data in different time ranges. As we valued on the completeness of data for all countries with in the specified time period, we limited the time range from 2000 to 2016.

As we allow users to explore the relationship between two dimensions, data from different files needed to be linked together. We stored all the indicator data in different files in an alphabetical order, and obtained the index added by the user through the function provided by d3.

Because the data we obtained directly from World Bank was raw, it contained some redundancy and something irrelevant to the project. An additional work was to clean the data. The data set included data of both individual countries and some unions as well. As we decided to focus on data for individual countries, data for these unions was not needed. We filtered unions by a standard list of country code in [1] and obtained a data set only for countries. Besides, we unified the data storage format of each file to make it easier to load data in batch. All code related to data processing is stored in a JavaScript file named *data_process.jsp*.

As we worked on data for countries, the geometry data is able to specify locations for countries and is needed for the project. We found a json file containing the country name and its geographic location in pairs of coordinates from [5]. The data for each country is linked to its geographic coordinates by the country code.

3 REASONING OF DESIGN

3.1 What: Data Abstraction

3.1.1 Data Types. In total we used has three kinds of data types, as is depicted in Figure 1. Most data falls in the category of *attributes*, such as the year, birth rate and tax revenue. *Items* also emerge in the data set. For example, a country is an item. In addition, in our project, spatial data, also called position, is needed. The world.json file provides two-dimensional geo-location data for each country. It contains pairs of latitude and longitude which can be useful to specify the location of countries.

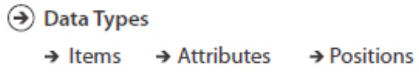


Figure 1: Graph of data types used in the project

3.1.2 Dataset Types. We obtained two kinds of dataset types using three kinds of data discussed above. Figure 2 shows these two kinds of dataset types. All post-processed data can be related with each other using *tables*. The columns in the data could be general health data, illness data, medical care data, finance data, country and year. If data with more than three dimensions are integrated together, a *multidimensional table* is formed. *Geometry* datasets is formed with spatial data.

All our datasets are considered to be static. They were downloaded on the Internet and was processed and filtered before providing to users.

3.1.3 Attributes Types. Countries are considered to as *categorical* variables. All other data are *ordered*. Finance and geometry data are ordered *divergently*. Health data and time are believed as *sequential* values.

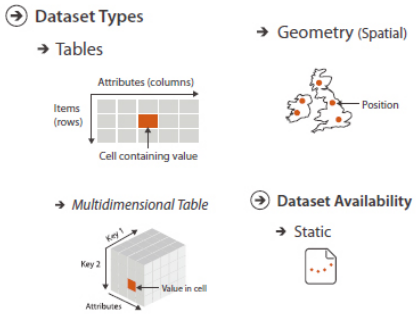


Figure 2: Graph of dataset types used in the project

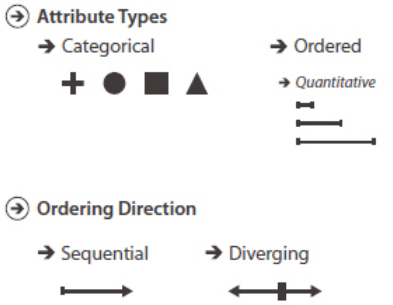


Figure 3: Graph of attribute types used in the project

3.2 Why: Task Abstraction

Health and economy data are collected from all countries all over the world, except some countries' data are missed in some years. The scale of these data is rather vast and unavailable for users to access important information, what they can achieve are only some values in one attribute of a country in a year, which is meaningless and useless. We aim to visualise the data in both scatter plot and map and show the potential inside relations between each attribute.

With our product, users can discover the relations between every two attributes they choose to obtain the inside relation via the scatter plot. For instance, users choose two features and expect to see the relation between them, the visualization result might be negative linear correlated or some other results, in the scatter plot showing that the lower GDP a country has, the higher birth rate it has, then the final and intuitive result users could obtain is that the richer countries are, the lower birth rate they have. As the same time, users can see the visualization results in the map we offer to them. They can identify and compare information based on the geological positions via the diverse colour of map, for example it is easy for them to find which country has higher illness rate as well as to check if the country is a developing country which is poorer compared with most of the others.

3.3 How: Visual encoding

3.3.1 Map. Since the data contains geographic information and indicator information from countries around the world, this geographically based data is best visualized and ranked according to geographic location. Therefore, we decided to use a choropleth map. The map is capable of discovering the spatial pattern and explore the data through the geographical relationship on the map. It also allows users to understand national data more quickly. This is because the data drawn on the map can be understood faster than viewing the data in the list.

Figure 4: Graph of charts using combination of marks and channels.

3.3.2 Scatterplot and Color. Figure 4 shows examples of using marks and channels to represent data. To display bivariate data, both bar chart and scatter plot can be applied. Since we decided to study the relationship between health and finance in worldwide, a bar chart might be inappropriate as it needs huge space to display values in bars. The Scatter plot is a combination of marks and channels and it is able to show two-dimensional data. It can show both vertical position and horizontal position and save space as well. So we decided to use a scatter plot. To make it easier to see difference of values, color could be used. As the example in Figure 4(c), we added an additional categorical data attribute on scatter plot by using visual channel of hue (one aspect of color).

3.3.3 Highlight. In order to highlight particular point in the scatter plot, we used a quantitative attribute encoded with the visual channel of size, as the example in figure 4(d). The change of size is easy to perceive and gives users direct hint of which point he chooses.

For the choropleth map, we also want to highlight the data if a mouse hover over a country or a region. Color can be of great help. The color of a specified country or region becomes different if that country or region is chosen. In this way users could clearly receive a feedback whether the data is chosen or not.

4 VISUALISATION DESIGN

The visualization project is comprised of four parts: a selection bar, a color legend, a scatter plot and a choropleth map. In this section the design of each visualization plot will be explained one by one.

4.1 Selection Bar

To make it easier for users to make choices, we designed a selection bar, as shown in figure 5. The list shows two high-level dimensions of data, health and finance. The health dimension includes three finer dimensions – general, illness and medical care, each of which contains basic attributes. The user can click and select two attributes

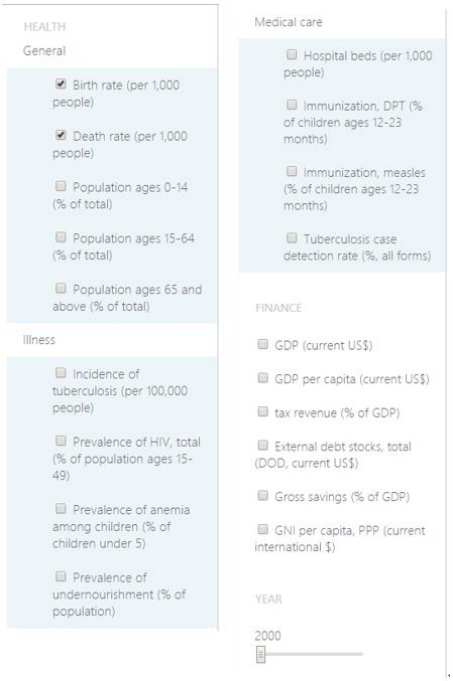


Figure 5: Graph of the selection list

and the year which is from 2000 to 2016 in the list and submit. Selecting this property again will deselect it. After selecting a pair of attributes, the scatter plot, the color legend and the map will update their data visualization accordingly.

4.2 Color Legend

Because of our choice for a multivariate dataset, we designed a bivariate color legend. The univariate graph gives us clues to quantities. The bivariate graph makes it possible to show two variables at once, not only quantities but also a relation between two variables, which are difficult to detect if each variable is plotted separately on its own chart.

Plotting two variables with in one graph is difficult. We decided to use a two-dimensional color legend with sequential colors proposed by Brewer[4]. A sequential data class is logically ordered by number from high to low, and the stepped sequence of categories is represented by sequential lightness steps. So low data values are presented by light colors and high data values are presented by dark colors, which is easy to be captured by users.

Colors were carefully chosen. To make it easier to see the difference of values, we decided to have each variable step through only 3 classes. We used a color scheme commended by [8]. Figure 6 shows the formation of our bivariate color scheme. The first two sub plots used a single sequential hue to represent one variable. The color grid in each hue gradually became darker with the increasing of value. A pair of complementary colors was chosen for these two plots to maximize the difference. We then combined two sequential schemes and obtained a new color palette revealing two variables.

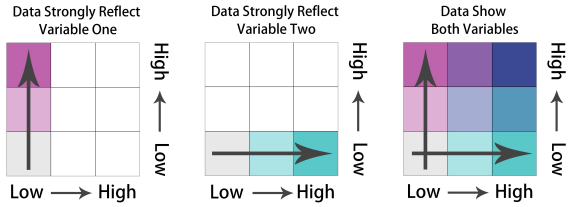


Figure 6: Graph of formation of a two-dimensional color legend.

To show the value range for each grid, we decided to use a tooltip. The tooltip includes the value range for a pair of attributes. Figure 7 shows the final design effect of color legend.

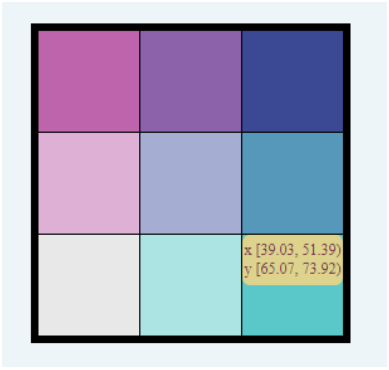


Figure 7: Graph of the design result of the color legend

4.3 Scatter Plot

We decided to use a scatter plot which allows users to see the values of the two attributes chosen in the navigation bar. The domain of each attribute starts from the minimal value and stops at the maximal value in the data set. Each point in scatter plot represents a country. It is color-coded according to the relative position between the coordinate system and the color legend.

Tooltip is also used to give users additional information about the point, including the name of and country and values. In order to give a coherent look, the background color of the tooltip is same with the color of the point representing the same country.

4.4 Choropleth Map

Implementation. In the end we decided to design a map which can illustrate visualised information spatially for users, it could help users to see where is a country located and how it performs within a certain data combination. For achieving the visualised map in D3, we used JQuery [7] and Underscore[2] to assist our work. We applied the Mercator projection to display our map which are easy and familiar for users to recognise. Moreover, because there is no country in Antarctica and thus there is not available data, so we removed Antarctica from our map.

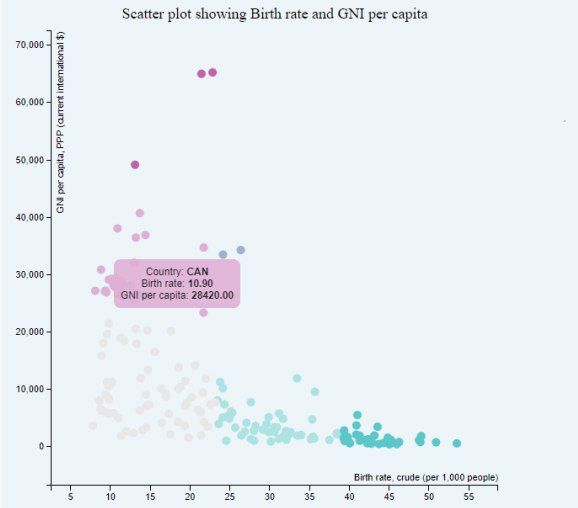


Figure 8: Graph of the final design effect of the scatter plot

Map Initialisation. In the beginning when users do not select two features yet, there will be an initialised empty and static map inserted into the web page, the original color of it is filled by a default one, which does not appear in the color matrix to prevent conflicts. When hovering mouse over the map, the country which mouse is over presently will display its name in a tooltip window and the color will change to another one to highlight it, shown in Figure 9. In this step the only data can be achieved is name of country from *map.json*.

Information Visualisation. All data combinations from users' selection could be visualised in the world map. We decided to utilise the 2-dimensions data to made a color matrix because we have 2-dimensions data as input and a 2D color can show the hierarchical spatial information clearly on the map. After clicking choose button, each country will have its own filling color based on the value it has and the color matrix. When hovering mouse over a certain country, the color of tooltip window will change into the filling color of the country and the filling color of country will change into the same one as the initialisation part, but the contents in tooltip windows would change as well, not only the country name would be displayed, but also the values of two features selected of the certain country would be illustrated, we aimed to make it easy for users to access exact data of two features at the same time they see the hierarchical color in the map, shown in Figure 10. What the tooltip shows when hovering mouse on map is the same content as when doing it on scatter plot. When users select new pairs of features, our map can be updated based on the new two features they select.

Additionally, some data are not collected by the World Bank, in details, some country lost data of some features in some years. In our map, if data of a certain country within a certain combination of two features and a year is missed, then the map will display black and the tooltip window well not be displayed.

Figure 9: Graph of the initial state of world map

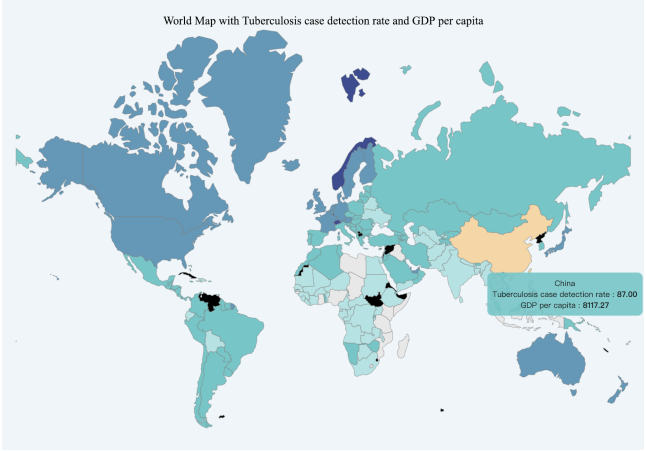


Figure 10: Graph of the world map with visualisation result of GDP per capita (current US$) and Tuberculosis case detection rate (%, all forms)

5 OBSERVATION ANALYSIS

Our data visualisation project offers users 2907 ($\binom{19}{2}\times 17$, there are 19 features, every time only two features can be selected and 17 years in total from 2000 to 2016) visualisation selections totally, users can combine any two features of both health domain and economy domain from the selection bar, a scatter plot and a choropleth map will be illustrated. We chose three kinds of combination, such as *health-finance*, *health-health* and *finance-finance*, and proceeded tests to analyse the relation between some of the features and the effects of our product.

Firstly, take an assumption that *the richer the country is, the lower birth rate it has.* We selected features **Birth rate (per 1,000 people)** and **GDP (current US$)** in year 2016 to do the test, Figure 11 shows the relationship between these two features, it generally shows that with the increase of GDP, birth rate decreases. It can be more obvious to draw this conclusion in map. We can see that most

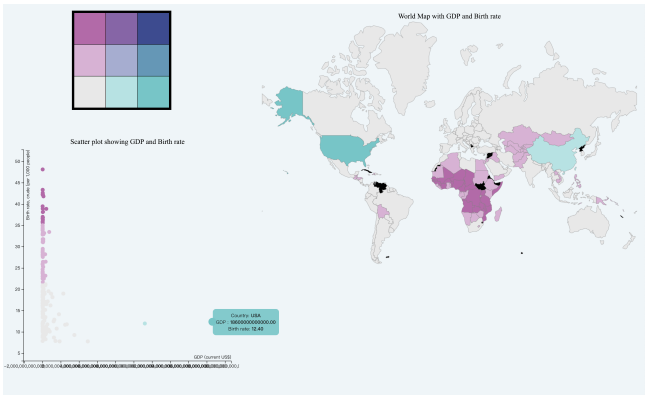


Figure 11: Graph of visualisation result of feature GDP (current US$) and Birth rate (per 1,000 people)

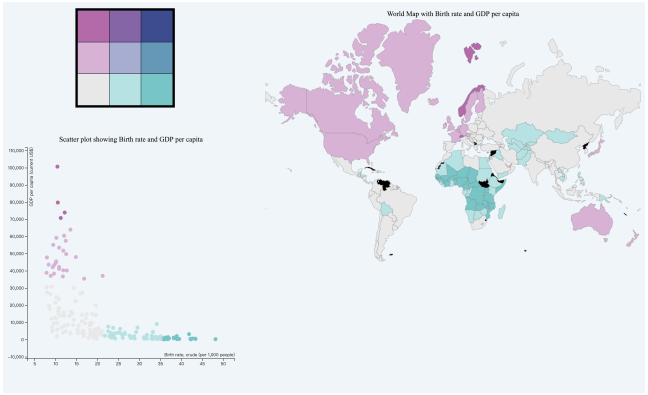


Figure 12: Graph of visualisation result of feature GDP per capita (current US$) and Birth rate (per 1,000 people)

countries in Africa and west Asia have a high birth rate. And at the same time, they also have a low GDP. But in Europe and North America, the situation is just opposite.

However, it is obvious that USA and China are two outliers whose economy scale are rather large and their scatters in the plot are far away from all the other countries. So, we used another features combination **Birth rate (per 1,000 people)** and **GDP per capita (current US$)** in 2016 to do the observation one more time, Figure 12 shows our assumption perfectly. The average value solves the problem that developed/poor, developing countries data sometimes are way larger/lower than the others. The latter combination behaves better and clearer than the former one to illustrate the assumption. Also for the map, the color among different countries in the map clearly proved our assumption. Because of the result of USA and China in that scatter plot, it made that all the scatters of the other countries were squeezed into a limited space. The visualisation method is easy to detect outliers, but meanwhile, the outliers cause the squeezing space problem above sometimes.

For health-health test, some other general knowledge such as the higher **Birth rate (per 1,000 people)** a country has, the larger **Population ages 0-14 (% of total)** it has could be easily approved

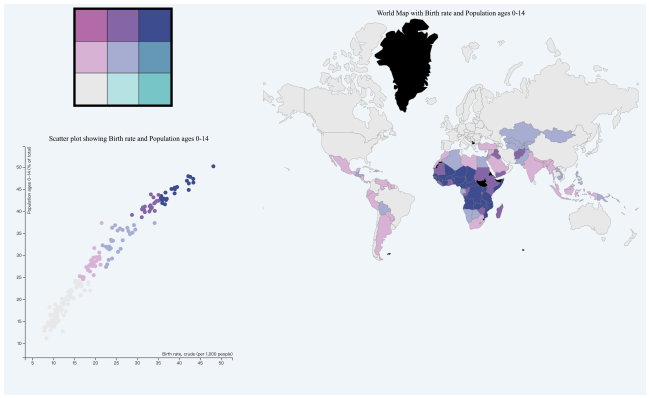


Figure 13: Graph of visualisation result of feature Birth rate (per 1,000 people) and Population ages 0-14 (% of total)

by our visualisation product, shown in Figure 13, these two features have high linear positive correlation and our assumption was proved.

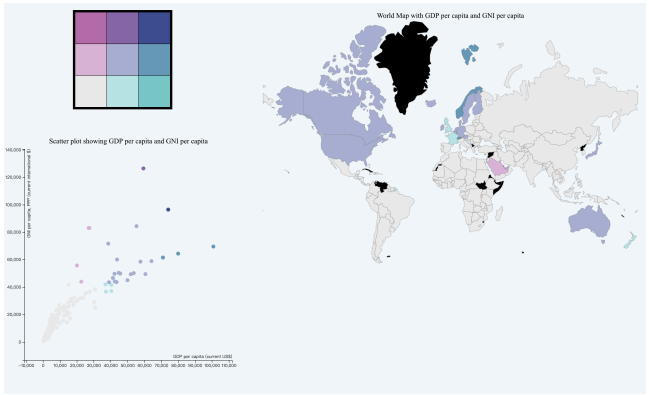


Figure 14: Graph of visualisation result of feature GDP per capita (current US$) and GNI per capita, PPP (current international $)

In the end, we did an finance-finance test and selected **GDP per capita (current US$)** and **GNI per capita, PPP (current international $)** as our features. The test result in Figure 14 shows that if a country is rich, people in the country always earn more money than others whose are not.

For all those tests above, the choropleth map illustrates the geographical distribution of different feature values. It is obvious to explain which parts in the world are rich and developed and which parts have poor behaviour in the medical and health fields.

In this testing progress, we achieved plenty of good results, either they fit our assumption or show some characteristics what we did not expect. However, some combinations of features do not have objectively related regulation. Because of the missing data in some years and in some countries, it will cause the related country blocks are filled into black color, such as Prevalence of HIV, total (% of population ages 15-49).

6 CONCLUSION

In this information visualization project, we created an exploratory view of health and financial data. The data visualization application provides an easy way to explore data. Special attention is on data processing, including selecting features and coupling data sets. In order to visualize the processed data, we designed a navigation bar, a scatter plot, a color legend and a map. By combining related graphs, we allow users to explore data in a simple and direct way, looking for relationships and trends between data.

6.1 Future Perspectives

Due to time limitations, we came up with many ideas but did not achieve them all. There is still something could be improved in future. First, a link could be created between the scatter plot and the map. It will help users perceive the bivariate data and geographical location in the map at same time. It will be better when hovering over the map, at the same time, in the scatter plot, the corresponding point will be enlarged. Vice versa, when hovering over the scatter plot, the corresponding country on the map will be highlighted. Second, we could also design a line chart or an animated scatter plot to show the trend of how one dimensional data changes with time. We believed it is worthy to compare the trend of data for different countries and some interesting result might be found. These are the improvement we hope to make in the future.

REFERENCES

[1] Central Intelligence Agency. 2009. *The CIA World Factbook 2010.* Skyhorse Publishing Inc.
[2] Jeremy Ashkenas. 2009. Underscore.js. https://underscorejs.org/.
[3] Mike Bostock. 2016. D3.js - Data-Driven Documents. https://d3js.org Accessed by 04-11-2018.
[4] Cynthia Brewer. 1994. Color Use Guidelines for Mapping and Visualization. http://www.personal.psu.edu/cab38/ColorSch/Schemes.html Accessed by 08-12-2018.
[5] Natural Earth. 2018. Free vector and raster map data at 1:10m, 1:50m, and 1:110m scales. http://www.naturalearthdata.com/downloads/
[6] World Bank Group. 2018. World Bank Open Data. https://data.worldbank.org/ Accessed by 10-11-2018.
[7] Inc. jQuery Foundation. 2005. jQuery. https://jquery.com/.
[8] Joshua Stevens. 2015. Bivariate Choropleth Maps: A How-to Guide. http://www.joshuastevens.net/cartography/make-a-bivariate-choropleth-map/ Accessed by 08-12-2018.

A ATTRIBUTE LIST

- Health data
  - General
    - Birth rate, crude (per 1,000 people)
    - Death rate, crude (per 1,000 people)
    - Population ages 0-14 (% of total)
    - Population ages 15-64 (% of total)
    - Population ages 65 and above (% of total)
  - Illness
    - Incidence of tuberculosis (per 100,000 people)
    - Prevalence of HIV, total (% of population ages 15-49)
    - Prevalence of anemia among children (% of children under 5)
    - Prevalence of undernourishment (% of population)
  - Medical care
    - Hospital beds (per 1,000 people)
    - Immunization, DPT (% of children ages 12-23 months)
    - Immunization, measles (% of children ages 12-23 months)
    - Tuberculosis case detection rate (%, all forms)
- Finance data
  - GDP (current US$)
  - GDP per capita (current US$)
  - tax revenue (% of GDP)
  - External debt stocks, total (DOD, current US$)
  - Gross savings (% of GDP)
  - GNI per capita, PPP (current international $)

B INDIVIDUAL REFLECTION

This project is a team effort for around five weeks. In the first week we met two times and did brainstorm to come up with ideas. In following weeks we spent most of time in coding. We separated the work based on our design sketch and everyone took a part. We met at least once a week. In the meeting we shared our progress and discussed the problem. In the last week we worked mostly on writing the report and making a screencast.

B.1 Reflection from Wenyu Gu

I improved my JavaScript skills and learned to use D3 through the project. My work mainly consists of three parts: preparation, coding and writing report.

For the preparation, I helped select the topic and set up the working environment on git.

For coding, I designed a color legend and a scatter plot. As I did not have much experience with JavaScript before, learning D3 was a kind of bittersweet process. I followed the tutorial from practical sessions and drew the plot step by step. I started by drawing axes, and then added circles, and then colored the circles, and finally improved the design. My teammates Zheng and Tian gave me some useful suggestions on how to improve the plot. While working with D3 and JavaScript, I met a bunch of problems. I got help from some online resources such as Stack Overflow and D3 official tutorials. Two teaching assistants also gave me great help, not only for technical issues, but also for suggestions of design and the priority of work of whole project.

My last contribution went to writing part of report.

B.2 Reflection from Zheng Liu

It is my first time to use JavaScript and html, in the beginning it was bit difficult for me but I kept practice and got familiar with the programming. My part consists four part: preparation, coding, writing report and screencast.

In the preparation period, I discussed the topics with my teammates. In coding stage, my part is to design the map. I started from the course laboratory work and searched the syntax online to get familiar about d3 and JavaScript. The step of my work is to prepare the map data as .json file, then insert them into web page as empty and static map, then I gradually inserted the mouse-over and mouse-out function, added tooltip into the function, displayed data in the tooltips, fill colours in each country after selecting data and in the end I cooperated with my team members to assemble all components together, in this period I got lots of support from my teammates for getting colour parameter.

In the end I was in charge of a part of report writing and making screencast video.

B.3 Reflection from Tian Tian

I have learned a lot from this project. From the initial topic selection to project design, it is a new challenge for me. Through this project, I have improved my ability to program web pages, and I also learned how to use d3 for web design.

In this project, I take charge of the selection and processing of data and the implementation of the selection bar in the web page. In the process of implementation, I have encountered many difficulties. In the original form of data storage, I chose to store it as an array. However, when linking with the part of the other team members, the data format did not meet their requirements, so I modified it and changed it to json format for transmission. In the final integration section, in order to make the page look neat and tidy, we have re-edited the page layout to meet the requirements of each section. I am very grateful to my teacher and two teacher assistants who helped me a lot.I am also grateful to my teammates Wenyu Gu and Zheng Liu, they gave me a lot of advice when I faced problems.

In addition, in this group cooperation project, I also learned how to communicate with teammates, how to adapt to each other and progress together. I believe that after this cooperation, our next project can be more successful.