COMP6234 Data Visualization Coursework Report

YUNHONG LIU
University of Southampton
Southampton, United Kingdom
yl2y17@soton.ac.uk

***Abstract*—A good visualization could help people to communicate with the mass data in an intelligible way. This report provides the detail of data story by demonstrating the justifications for the choices of visualisation, include the chart choices and colour design regarding to techniques and theories. In addition, it outlines the strength, weakness and improvements of the visualisation.**

***Index Terms*—Data visualization, Tableau, Data Story, Author-driven, Reader-driven, Tuftefis Design Principles**

I. Introduction

The visualization of data could improve the ability to interpret information. Finding key factors from vast amounts of data is not easy, but graphs and charts could provide information in seconds. People could quickly get insight that via looking at visualization. However, how to choose an appropriate way to represent data, communicate with the story and to answer the questions is significant. What is more, avoiding the bias and confusion is also necessary. To make it, it is better for developers understand data by exploring and analysing data as much as possible.

In this report, it will provide the overview of the data story. Next, it will demonstrate what kind of tools used in to build visualisation to support the data story. And then is the data set demonstration part, it will provide the data source and explain how it is related to data story after analysing. Then, the report will describe the main part which include how each graph be chosen with its highlight characteristics, strength, weakness and improvement and the theories used to justify the visualisation. Finally, this report will give the conclusion of this coursework.

II. Data Story Overview

This story is based on the current study abroad boom. More and more international students are more or less likely to work in a country where they study abroad, but it is not known how much they can stay.

According to the international students statistic was published by *UKCISA*, about 442,375 Non-UK students were enrolled in higher education inside UK in 2017, second only to the USA. As one of the international students in the UK, I suppose many students would like to know is easy to find a job and then settle in the UK or not? Therefore, this data story would like to lead readers to follow each question with three visualisations and finally find an answer of the main question is "Is a possible route to settle in the UK for Non-EU students by studying in the UK?".

The first question is "Where do the international students come from?, reader could interact with the map and bar chart in the first dashboard by change the slider from 2005 to 2016. Readers could quickly know where the most students come from by looking at the bar chart which provide the top ten countries student migrants in the UK. The second question is "How many Non-EU people work in the UK?. Readers could have a glance on the line chart and find the trend changes between EU people and non-EU people. Thirdly, after reader understand the overall amount of the non-EU workers, they could look at the bar chart they could address this question which is "How many non-EU students could truly settle in after they graduate from UK?. After looking at those three visualizations, readers might have their own answer.

III. Techniques tools summary

The visualizations in this coursewrok produced by Tableau which is the popular programs allowing users to analyse data in a simple way. The data story is shown in a web page which produced by Bracket. It is an open source text editor written in HTML, CSS, and JavaScript.

IV. Dataset summary

In this data story, I used three main datasets are in excel file from the open source website. The datasets will be demonstrated as following:

A. whole year international student enrolled in UK by nationality from 2005 to 2016 data set

This data set was published by HESA institute in 2017, which contains the number of international students enrolled in the UK by different country from 2005 to 2016 academic year in every season. To only display the import factors, the students number of season was merged into student number of year.

B. UK and non-UK people in the labour market data set

The data set was provided by NOS.GOV.DATA website, it contains the number of people who work in the UK from 2012 to 2018 in every season and the region where people come from.

C. Statistics on changes in migrants' visa and leave status: 2016

This data set was published by HOME OFFICE website in 2017. It contains over 10 data tables, in the table MJ03, we could find the relevant data which shows the number of migrants issued a study visa in 2004 to 2011, broken down by their immigration status 5 years later. In this table, it contains the number of students granted settlement, the number of valid leave to remain in the UK and the number of expired leave to remain in the UK. By simple calculate manipulation in excel, the percentage will be shown in the visualization which could better indicate how many students could settle in the UK.

V. Visualisation Summary

In this story, it used four types of visualization. Firstly, it combined a map chart and a horizontal histogram, then it used line chart to show the changes of the trend. Finally, it used a vertical histogram to show the changes of proportional of different part in each year.

To tell a good data story we should balance the Author-driven and Reader-Driven [2].Therefore, all charts are interactive which could allow reader to explore the data.

In addition, based on the Tuftefis Design Principles [1], use colours found in the nature which are red, green and blue. It also to make the visualisations are colour-blind-friendly. The visualisation in this report are designed mostly with red colour with grey colour to assist. Moreover, according to this principles about the graphical integrity [1], all bar charts and line charts show the accurate scale start from 0. The detail of how to design chart is shown as following.

A. Map chart

Figure 1 shows the number of international student study in the UK by countries from 2005 to 2016. The visualization based on the location of each country automatically generates the geographic graph. It is used colour as a visual cue and sub-continent are filed based on the number of enrolled students and time series scale. It used two contrast colours, which is grey and red. The higher values indicate higher saturation, the lower values represent the lower saturation. The different values in the middle are transition from grey to red in different shade of those two colours.

1) Strength: Utilizing the map form to reflect the data with many different countries is a good choice. It is more intuitive than general table or chart, and it is more effective to communicate with data. In addition, it could be more interesting for reader to explore data. This map aims to answer the question 'Where do most of international students come from. It could be seen in this figure that the most international students which shown as the reddest come from India, China,

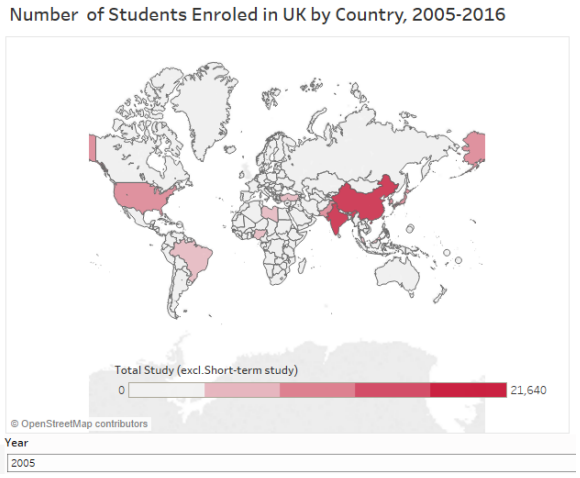

Fig. 1. Number of non-UK students enrolled in UK by country, 2005-2016

Pakistan and America. The number and year in the Tableau detail could make users explore data and see the change of each country easily.

2) Weakness: Some countries with small land areas cannot be easily exploited. In addition, the names of each country are shown on the map is make the visualisation seems messy

3) Improvement: Allow reader to zoom and drag the mapthus small countries also could be exploited. To deal with the messy problem, it could do not display the country names. However, this may not be convenient for those people with poor geography.

B. Horizontal histogram

Figure 2 shows the top then countries student migrants in the UK from 2005 to 2016 ranked by the number of international students who come to the UK.

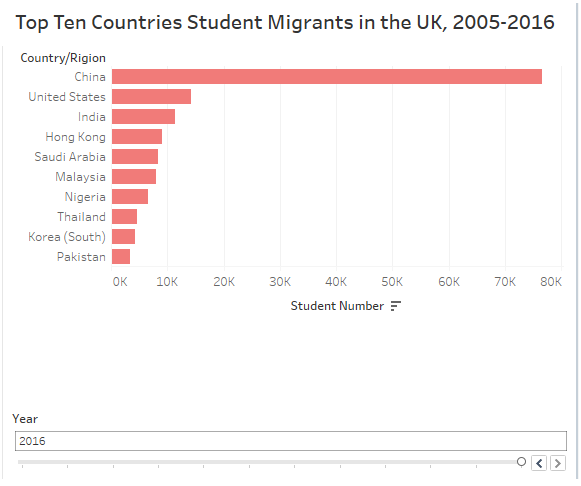

Fig. 2. Top ten countries student migrants in the UK, 2005-2016

1) Strength: The horizontal axis of the figure shows the number of international students and descending sort by the number. This bar chart is designed for helping users to get

the insight from the map quickly. Reader could see the most student migrants in the UK from China, the US and India. To avoid user may read the specific number from the surface of the bar hardly, this visualisation adds the number into the detail. When mouse move on the bar, there are relevant year and student number in the tool text. In addition, by using the sliderreaders could notice the change of the numbers based on the length of bar.

2) Weakness: It could be notice that except China, the number of international students from other countries are close to each other. Readers might not notice the comparison between them at one glance. It also might mislead the reader that is the percentage is represented by decimal. what is more, the notation of the percentage is not readable and clear.

3) Improvement: It is better use another colour, such as grey to fill the last seven countries that could the top three countries. In addition, it should display the number of percentage with % mark. And change the notation of percentage into 'the percentage of the total international student number'.

C. Line Chart

This simple line chart represent the nubmer of EU worker in the UK verse to the nubmer of non-EU worker from 2012 to 2017. It use the time series with four seasons in every year. The horitical axis shows the nubmer of people, the vertical axis as a year scale.

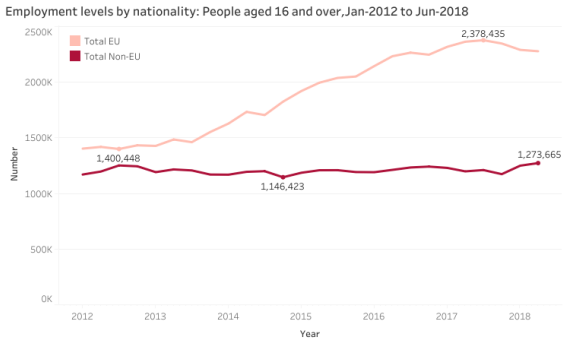

Fig. 3. Line chart with employment levels by different region

1) Strength: Choosing line chart is a straightforward way to compare the change over the same period for the two groups and to see the trend. This chart is designed for answering the question "How many non-EU people work in the UK. Readers could see the number of EU people who work in the UK were increase rapidly after 2012, and then are most as the two times as the non-EU people in 2018. Conversely, the mount of non-EU people was fluctuate stably. In additional, reader could see the specific data when they explore the visualisation. What is more, the visualisation show the highest and lowest value on follow the each line, users could get a insight from it at one glance.

2) Weakness: According to the Richard Gregory theory "Top-down processing theory seeing, eyes, light, brain in 1970, color differences between content and background may help focus our attention on the content itself" [3]. But in this visualization, the colour comparison is not obvious, the two colors are somewhat close. Besides, the data is not easy to read from the scale.

3) Improvement: It could use the red and grey to make more obvious contrast. To answer the question the visulization should be answered, it could highlight the number of people from non-EU countries, then mark it as red and the EU workers as gray. Add specific data to the detail to make data readable.

D. Vertical histogram

This figure shows the number of migrant issued a study visa in 2004 to 2011, broken down by their immigration status 5 years later. The x-axis is the year scale and the y-axis is the number of students. The percentage of three parts which is granted settlement, valid leave to remain in the UK and expired leave to remain in th UK respectively are shown on the correspond bar except the percentage less than 5%.

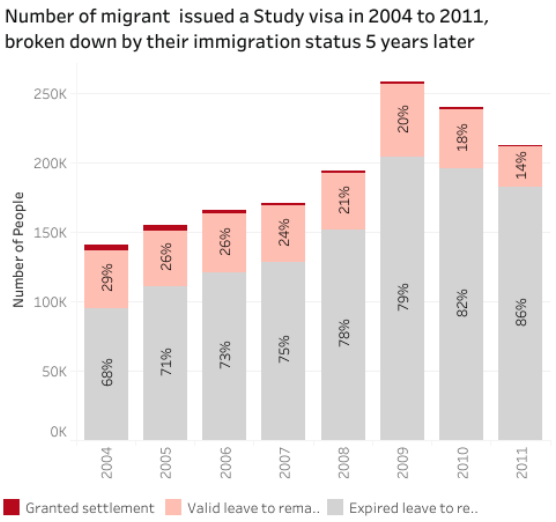

Fig. 4. The number of migrant issued a study visa in 2004 to 2011, broken down by their immigration status 5 years later

1) Strength: This chart is designed for leading reader to answering the question "How many non-EU students could truly settle in after they graduate from UK. Users could based on the color to distinguish the percentage of the each type of immigration statues. It could see the trend the percentage of valid leave to remain in the UK was decreasing over the period. By further explore the data, users could know the rate of settlement has declined steadily. What is more, the visualization uses the contrast color to make user distinguish the type of status better.

2) Weakness: It is hard to read the number of people at a glace.

3) Improvement: Add the specific number of people into the detail.

VI. Conclusion

This report firstly provided the overview of the data story. Secondly it demonstrated the dataset. Thirdly, the report described the main part which include how each graph be chosen with its highlight characteristics, strength, weakness and improvement and the theories used to justify the visualisation. In every visualization, the story gives a question to lead reader to go to next question and finally to find their own answer of the main question.

References

[1] Edward R. Tufte. 2001. Principles of Information Display for Visualization Practitioners.

[2] Shneiderman, B., 1996, September. The eyes have it: A task by data type taxonomy for information visualizations. In VisualLanguages, 1996. Proceedings., IEEE Symposium on (pp. 336-343). IEEE.

[3] Gregory, R. L. (1970). Princeton science library. Eye and brain: The psychology of seeing NJ: Princeton University Press.