School of Science, Computing and Engineering Technologies

SWINBURNE UNIVERSITY OF TECHNOLOGY

COS30045

Data Visualization

Project and Group Reflection

Student Name: Ta Quang Tung

Student ID: 104222196

Word count: 2233

Introduction

This report presents my reflection on the development of the visualization project for unit COS30045 and what I have learned in the process. The topic for the visualization project this semester was migration, a global phenomenon amplified thanks to greater mobility and ease of travel. The goal of the project is to build an interactive visualization using D3.js to put the issue into perspective. The following sections will detail what I have learned throughout the project in terms of both data visualization principles and data visualization programming. It will also describe my teamwork process and my contributions to the project.

Learning process

Data visualization concepts and principles

Over the semester, I have learned many data visualization concepts and this project has allowed me to put them into practice. One of the most important things I learned was storytelling with data. Data visualizations can be categorized as either exploratory or explanatory. Exploratory visualizations invite viewers to explore the data and reach their own conclusions whereas explanatory visualizations draw the viewer's attention to our findings. For this project, we have decided to design an explanatory visualization to help viewers understand the patterns of international student migration around the world. This distinction between exploratory and explanatory is crucial because it defines the purpose of our project and guides our direction. It requires us to identify what we want to tell from the data and understand our target audience. After doing research, we decided that the story we wanted to tell was where the migrating international students came from and went to, and our main audience was students who were considering which country to migrate to. As our primary audience might not be proficient in reading advanced charts, we agreed that our charts should be kept simple and intuitive.

With the chart criteria established, we had to consider alternative visualizations for the job. To choose the right visualization "idiom", we considered two aspects in order. The first was suitability – we would not want to use a chart type incompatible with our data. We have three types of data to show: (1) rankings of origin/destination countries over time by student numbers, (2) total flow of students from origins to destinations by

country, and (3) change in the number of students from one country to another over time. We considered a line chart and a bump chart for the first type, a bar chart and a sankey diagram for the second type, and a line chart for the third type. The second aspect we considered was effectiveness and ease of use. We wanted to choose charts that could most effectively display the types of data we were aiming for. For this reason, we eliminated the line chart option for the first type of data, as a bump chart is better at showing rankings. We also eliminated the bar chart option for the second type of data, as sankey diagrams, despite being more advanced, are better suited to showing flows from one set of values to another. After finalizing the chart types, we exchanged sketches to share ideas before deciding on the final design.

An important principle I learned during the design process of the project is to distinguish between “design” and “art”. Design is about solving problems whereas art is about making things look nice. Initially, when we started the project, I was too concerned with making the website look nice that I overloaded it with pretty but hollow charts that had little connection to one another. I was thankfully put on the right track during one of the project standups and began to think about how the charts should be linked together to make a more cohesive visualization. This was when my team considered the options of the above section and decided on the combination of a bump chart, a sankey diagram, and a line chart.

Another data visualization concept I learned, and perhaps one central to this unit, is interactivity. Interactivity can make visualizations easier to read by allowing users to zoom in on particular areas of interest, showing details on demand, etc. This is useful when charts have very large data sets and displaying everything on the screen can make them illegible. The visualizations of our project made extensive use of details on demand to declutter the screen. Viewers can hover over a series in the bump chart or a flow in the sankey diagram to reveal additional details only when they need to. Interactivity also allows us to update our existing charts with new data. For example, when the user clicks on a country in the bump chart, the sankey will update to reflect the information of that country.

Data visualization programming

To build the interactive visualizations for the project, we had to use D3.js, a powerful but complex library that has a considerable learning curve. Getting used to D3’s mental

model was the hardest part for me. It utilizes the concept of data joining to correspond each data item to an element in the visualization. Joining data with elements created three objects: the enter selection (containing missing elements), the update selection (containing existing elements joined to data), and the exit selection (containing elements to delete). Being mindful of which selection exists and needs to be dealt with is important to make our interactive charts work. This was a requirement for our sankey diagrams and line charts as we wanted to update these charts every time the user wished to change the data. We had to make sure that elements were correctly reused and that no unnecessary ones were created to avoid bugs in the visualizations. Throughout the project, we also had the chance to explore other features of the library such as animations and event handling to make our charts more interactive. Our charts have animations and events attached to respond to click or hover actions.

Aside from learning the core D3 library, we also had to study example use cases to build more advanced charts. Having to build a bump chart and a sankey diagram, two types of charts not directly supported by the library, meant that we had to research to build them ourselves. For the bump chart, we found an example on Observable that we had to study and extend to accommodate our needs. This required understanding the code and the D3 library at a relatively high level to add or remove features as needed without breaking the program. For the sankey diagram, we found a D3 plugin that simplified the process, but this required learning an additional API. I feel the development of these charts has helped hone my self-studying and exploration skills.

The charts that we designed would not be possible without the data set, which was obtained from the data portal of the UNESCO Institute for Statistics (UIS). The relevant data sets are *"Outbound internationally mobile students by host region"* and *"Inbound internationally mobile students by country of origin"*. I was the principal data processor in this project as I am more skilled at Excel and writing custom scripts to transform the data. This experience has taught me the influence of size and file format on the performance of the visualizations. As our our sankey diagrams and line charts update their data frequently, we have to make sure our data set is optimal. This means removing all unused data items and preparing a separate data file for each chart instead of using one large, shared file to minimize the need for sorting and filtering. Our data sets are either CSV or JSON files, the former being used for tabular data that needs some sorting and filtering and the latter being used for data that requires quick look-up.

```csv
origin,destination,year,students
Belarus,Portugal,2017,16
Belarus,Portugal,2018,31
Belarus,Portugal,2019,22
Belarus,Portugal,2020,22
Belarus,Portugal,2021,24
Hungary,Luxembourg,2017,19
Hungary,Luxembourg,2018,14
Hungary,Luxembourg,2019,13
Hungary,Luxembourg,2020,12
Hungary,Luxembourg,2021,14
Croatia,Poland,2017,17
Croatia,Poland,2018,25
Croatia,Poland,2019,25
Croatia,Poland,2020,33
Croatia,Poland,2021,36
Belarus,Georgia,2017,3
Belarus,Georgia,2018,2
Belarus,Georgia,2019,2
Belarus,Georgia,2020,5
```

Fig. 1: Data set on international student migration from one country to another, raw CSV (approx. 50,000 rows).

```json
{
    "Poland": {
        "Ukraine": {
            "2017": 34692,
            "2018": 26864,
            "2019": 26938,
            "2020": 27068,
            "2021": 30903
        },
        "Belarus": {
            "2017": 5002,
            "2018": 5258,
            "2019": 6025,
            "2020": 7472,
            "2021": 8994
        },
        "India": {
            "2017": 2084,
            "2018": 2497,
            "2019": 2153,
            "2020": 2960,
            "2021": 2665
        },
```

Fig. 2: Data set transformed into JSON. One pair of orign-destination can be quickly looked up.

Team reflection

Nguyen Quang Huy was my teammate in this project. We collaborated mainly through Messenger (for communication), Confluence (for sharing documents), and GitHub (for programming). Despite scheduling conflicts (especially near the end of the semester), we tried our best to keep up communication and stay updated on each other's work.

In this project, we followed a structured design process, first starting with identifying the domain. After doing research, we decided that we would focus on the migration patterns of international tertiary students and that our main audience would be students who are considering which country to migrate to. With this audience in mind, we had to make sure the charts we designed were simple and intuitive to use. Establishing the focus and audience early on was very helpful as it limited the scope of our project and helped us save time. It also helped us to focus on one specific area instead of scratching the surface of a broader domain.

The next step was to gather the data for our topic. With this topic, we sought to answer three questions:

(1) What are the top origin and destination countries of international students?

(2) What are the biggest destinations and origins of the countries found in (1)?

(3) How have the numbers changed for each pair of origin-destination over the years?

The data to address these questions was collected from the UNESCO Institute for Statistics (UIS), which has abundant data on international student migration between 2017 and 2021. For the sake of data integrity and consistency, we decided to not take data from other sources to patch up missing data items in the UNESCO data set. My teammate and I both participated in cleaning and transforming the data set, although I did the bulk of the work since I was more proficient in Excel and writing data transformation scripts. Depending on the complexity of the dataset, I would use either Excel or a custom JavaScript script to read and transform the data. Although I did a larger share of the work, I made sure to incorporate feedback from my teammate throughout the process. Looking back, this communication helped me design a cleaner and more optimal data set for our visualizations.

After gathering our data, we worked on designing the visualizations together. Both of us were equally involved in this process. We exchanged sketches to share ideas and made sure we were on the same page. This communication was crucial as without it, our work would not be as cohesive. We included three charts in the final design: a bump chart, a sankey diagram, and a line chart. We also discussed how the user interactions would take place for each chart. This helped link our charts together to produce a logical connection among the data presented.

We then moved on to programming the visualization. I did more programming than Huy since I was more proficient in JavaScript. Huy worked mainly on the bump chart, for which he found an example that we studied to accommodate our use case. He got the basic chart working and I extended it to support additional interactions, such as clicking to update other charts. I also worked on the sankey diagram and line chart. My efforts allowed the charts to work together smoothly to produce a cohesive visualization.

To compensate for not doing as much programming, Huy worked more on the process book than I did. He helped describe in detail the visualization design and programming processes in the report while I only wrote the data collection section. Despite doing different areas of work, we made sure to stay in constant contact to get updates on each other's progress and to get cross-feedback. Thanks to this, I could ensure that my programming output matched our initial design, while Huy could ensure that the report closely aligned with the design and programming processes. While programming, I also discussed the code frequently with Huy to get his feedback and to explain to him how the code worked, allowing him to stay involved without having to code directly. After completing our website, both of us set out to collect feedback from users to validate our work. We jointly designed a form and asked our friends and acquaintances to test the website. We then gathered the results and discussed them to make any final improvements before submission.

Overall, I think that we collaborated quite effectively on this project. We were able to create a supportive environment where both of us could confidently exchange ideas. We tried our best to involve one another in every stage of the project to stay on the same page and receive feedback. Despite our conflicting schedules, we tried not to miss any deadlines and meetings. That said, our process was not without flaws. Near the end of the semester, because we both faced a lot of deadlines from other units, one time we decided to work independently for a few days and combine our work afterward.

We essentially worked in a vacuum during that period, and the resulting work from each of us (my programming and his report writing) was different from what the other person was expecting. Realizing that this lack of communication was making our work less cohesive, we decided to adopt the strategy of meeting each other for 20 minutes every day to show our progress and communicate issues. This approach worked rather well as we produced fairly consistent work and very few major changes had to be made.