

Showcase of d3.js and its possibilities in
infographics using refugee data of the current
Ukraine conflict

Luis Rothenhäusler (20202459)

Brandenburg, June 1, 2022

Abstract - German

Englische Arbeiten brauchen eine Zusammenfassung auf Deutsch. Mal abgesehen davon, dass wenn die Zusammenfassung interessant ist man ohne English eh nicht weiter kommt...

Contents

1	Introduction	3
2	Basics	4
2.1	Data	4
2.1.1	Types of Data	4
2.1.2	Datasets	5
2.1.3	Data Visualization	6
2.2	Diagrams	6
2.2.1	Refugees per country	7
2.2.2	Refugees over time	8
2.3	D3.js	8
2.3.1	What is it?	8
2.3.2	How does it work?	8
3	Implementation	10
3.1	Showcase	10
3.1.1	Integration of each diagram	10
3.1.2	Data Updates	10
3.2	Diagrams	10
3.2.1	Bar Chart	10
4	Conclusion	12
5	Appendix	14

1. Introduction

Describe the background of the thesis, why it is important, what do we want to achieve. Why is this thesis? What do we want to do? What is the status quo? What benefits will result from this thesis? Something about the importance of infographics and comprehensible data.

The postmodern world produces huge amounts of data every second. Analyzing this data leads to better-informed decision-making in every sector. Yet the vast amounts of created data is often hard to comprehend with the human mind. Data visualization is about finding ways to represent this data in visually appealing, yet easily understandable visual representations. Doing this quickly and always up to date can be crucial. There are many tools available to help with the creation of infographics. Some of these tools have a graphical-user-interface, are programming based. This thesis will be a deep dive into the possibilities of one of these tools, the 'd3.js'(D3) library for JavaScript. To show its capabilities and potential D3 will be used to create a showcase containing several different graphics.

2. Basics

Everything necessary to understand the implementation as well as anything which is done beforehand, will come up here

In the following, all concepts, technologies and required backgrounds for understanding this thesis are explained.

2.1 Data

We'll talk about data a bit. Where does it come from? How is it structured? What kind of attributes? What even are attributes?

Since ancient times, humans have recorded data. Recording the ins and outs of available resources was one of the driving factors behind the conceptualization of writing. (TODO: Check sources of the beer brewing video series) With the introduction of computers the amounts of gathered data have grown drastically. Vast amounts of data are gathered across all aspects of life.

2.1.1 Types of Data

Even though data comes from a huge variety of sources and can express a plethora of things, there are only four different types of data. They are split into equal pairs. Two types of categorical data and two types of numerical data. In the following each of the types of data will be explained.

Categorical

What is categorical data? Nominal and ordinal data

Categorical or qualitative data is information collected in groups. It is often of descriptive nature. Whilst the values can be represented in numbers, they do not allow for arithmetic operations. There are two types of categorical data. Nominal and ordinal data.

Nominal data is mostly descriptive in nature. They are independent and have no inherit order. Examples are 'Country', 'Color', 'Brand'.

Ordinal data is mostly similar to nominal data. Yet the data does have some sort of internal order. For example different dates each describe a day, but one day also comes after another.

Numeric

What is numeric data? Continuous and Discrete

Numeric or quantitative data is all data expressed in numbers, where numbers do not represent categories. It allows for arithmetical operations and can be split into Discrete and Continuous data.

Discrete data can only take certain defined values. This usually means whole numbers to represent things that can not be split up further. Like the 'Number of Refugees' or 'Tickets sold'. Discrete data is countable.

Continuous data can be measured. It can have any real number as value. Therefore fractions are possible as well. For example when measuring the temperature, or the length or weight of an object.

2.1.2 Datasets

What are our datasets about? Where do they come from?

Datasets are a collection of several data-points. each of these data-points is made up of different attributes. Each attribute corresponds to a specific data-type.

In this thesis, two data-sets are used. They are both from UNHCR[1]. The first data-set contains information about the total number of refugees per country[2]. The second dataset is about the total cumulative total amount of refugees per day[3]. Both data-sets are in the JSON format.

Preprocessing

What is done in preprocessing? Python script which removes all excess / maybe do that in JS as well..?

As both of the used data-sets have a lot of 'unnecessary' information, they are both preprocessed. In both cases the data-sets are read and the valuable information extracted and saved in the csv format. Both of the newly created csv files have two columns and one header row. The resulting

csv of the refugees per country dataset, contains the two columns of country and refugees. The other resulting csv contains a column for the date and one for the cumulative refugees. The data-set about the refugees per country can also easily be converted into using percentages. After adding up the total amount of refugees from each data-point, one can convert the absolute number of refugees into percent.

Data Types

Which data types can be found in our data-sets? Where?

The two chosen data-sets already cover most of the data-types. The refugees per country dataset contains two attributes per data-point. The country is a categorical attribute. The number of refugees is discrete. When converting this data-set into using percentages, the percentage of refugees becomes a continuous attribute. In the refugees per date data-set, the amount of refugees is still a discrete attribute. The date itself is an ordinal attribute though. As one day clearly comes before and after another day.

Choosing data-sets which cover all types of data-types was an important consideration. Different data-types can have different ways of representation, as well as different ways of implementation on the programming side of things.

2.1.3 Data Visualization

What is it? Where is it? Why is it important?

Data visualization is the process of turning data into graphical representations. As the vast amounts of data which are gathered in charts and databases are often hard to comprehend with the human mind and might require a big amounts of space to directly represent, it is often desired to turn these data-sets into a more easily understandable formats. Therefore data is turned into diagrams. We constantly come across the results of data visualization in everyday life. They can be commonly found across all kinds of news sources, but also in reports, information campaigns or as part of user-interfaces in machinery or control systems.

2.2 Diagrams

What diagrams exists? Which are the most common? What possibilities do they offer for encoding data? Which considerations for readability? Why do some diagrams not make as much sense? Which considerations were made for fulfilling the showcase requirements?

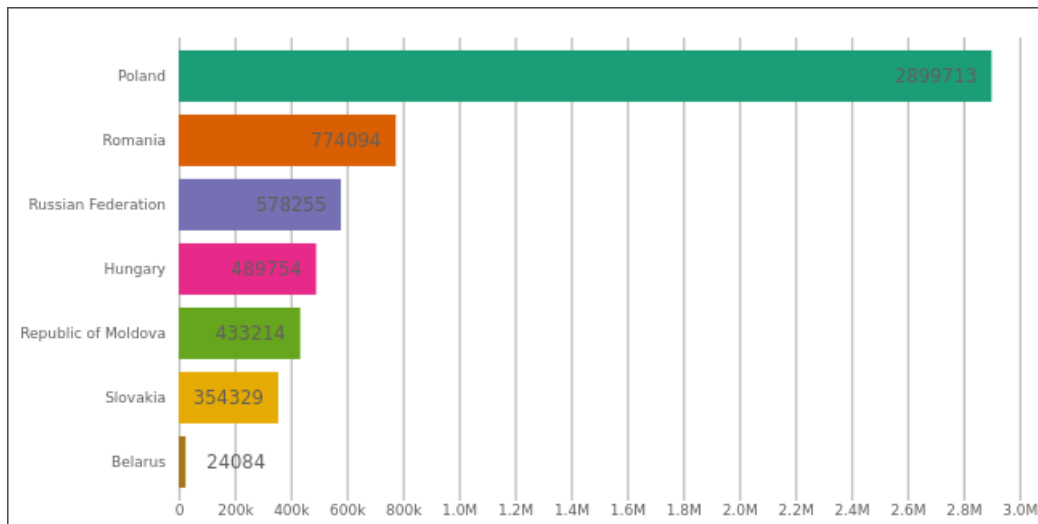


Figure 2.1: This is a bar-chart (TODO: which needs a frame..?)

Diagrams exist in many different shapes and sizes. There are many factors which should be taken into consideration when deciding which diagram to choose. There are usually many ways to represent the desired data. Some of the most common diagrams include bar-charts, pie-charts and scatter-plots (TODO: Find source (1el))

All diagrams use a combination of marks and channels to encode data. Marks usually depend on the type of data. They are usually single points, lines or areas in a diagram. Channels depend on the singular data-points and influence the marks. The most common channels are position, color and size. It is important to note, that not all channels work with all data-types.

To understand this more easily, we can look at diagram 2.2. This diagram uses lines as marks. The y-position, as well as the hue, are both used to encode the same categorical attribute. The area/size encodes a discrete attribute.

Different combinations of marks and channels can influence the readability and the correctness of how the represented data is comprehended [4] [5].

2.2.1 Refugees per country

Which ones did I choose? Why? Which data attributes do they encode?

Bar chart

What is it? Why is it here?

Pie chart

What is it? Why is it here?

Tree map

What is it? Why is it here?

Sankey

What is it? Why is it here?

2.2.2 Refugees over time

Which ones did I choose? Why? Which data attributes do they encode?

Area graph

What is it? Why is it here?

Circle graph

What is it? Why is it here?

2.3 D3.js

This is all about d3. What is it? Where does it come from? What is it used for? Who uses it? Why should it be used? How does it work? Enter, update and exit pattern. Something about the modular structure of D3 as well. Might be worth mentioning "observables" as well.

2.3.1 What is it?

General definition. What is it, where does it come from, what is it for?

2.3.2 How does it work?

General functioning of D3.

General Update Pattern

What is it? What can it do? Describe data joins and dom element links.

Modules

The way D3 is split up into modules, the core package and what kind of extensions are there.

3. Implementation

How are the chosen diagrams implemented? Which D3 modules have been used? How was the implementation done?

3.1 Showcase

How is the showcase structured? How can you get there? Why does it exist? Who might benefit? How can you reuse a part the interesting parts?

3.1.1 Integration of each diagram

How is each diagram integrated? How can you access them? Where can you grab them standalone?

3.1.2 Data Updates

How can you simulate data changes? Why is this useful?

3.2 Diagrams

describe all the diagrams and why they are special and what makes them tick

3.2.1 Bar Chart

How does it work? Which d3 features does it use? how do they work?

Pie chart

How does it work? Which d3 features does it use? how do they work?

Tree map

How does it work? Which d3 features does it use? how do they work?

Sankey

How does it work? Which d3 features does it use? how do they work?

Area graph

How does it work? Which d3 features does it use? how do they work?

Circle graph

How does it work? Which d3 features does it use? how do they work?

4. Conclusion

How well did it work? Was it worth the effort? What could be improved?

Bibliography

- [1] UNHCR, “Operational data portal,” accessed:12.05.2022. [Online]. Available: <https://data2.unhcr.org/en/situations/ukraine>
- [2] —, “Refugees per country,” accessed:12.05.2022. [Online]. Available: https://data.unhcr.org/population/get/sublocation?widget_id=312121&sv_id=54&population_01-01
- [3] —, “Refugees per day,” accessed:12.05.2022. [Online]. Available: https://data.unhcr.org/population/get/timeseries?widget_id=312123&sv_id=54&population_01-01
- [4] J. Heer and M. Bostock, “Crowdsourcing graphical perception: using mechanical turk to assess visualization design,” in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2010, pp. 203–212.
- [5] J. Mackinlay, “Automating the design of graphical presentations of relational information,” *Acm Transactions On Graphics (Tog)*, vol. 5, no. 2, pp. 110–141, 1986.

5. Appendix

I guess this should contain all the source code. Maybe there are ways to import it automatically too?