Information Visualization Project Report

Matyáš Skalický
Czech technical
university in Prague
skalimat@fit.cvut.cz
ist194904

Ingeborg Sollid
Norwegian University of
Science and Technology
ingeboss@stud.ntnu.no
ist195201

Lenka Obermajerová
Czech technical
university in Prague
obermlen@fit.cvut.cz
ist195110

INTRODUCTION

Our primary motivation for visualizing the data about the Erasmus program is that all of us are currently participating in Erasmus ourselves. Erasmus is widely known among the general population, but it can be hard to understand, where the people (and money) are going. We believed, that by visualizing the data of Erasmus students, we could find a lot of interesting information about our fellow classmates.

The resulting visualization aims to visualize the flow of international students within the ERASMUS program. In the academic year 2013-2014 272,000 students spent time abroad and had a total budget of €580 million (1). It is clear that the Erasmus program touches many people every year.

Our visualisation can also help future ERASMUS students to decide which country is the best for them by showing which countries are popular among students of similar degrees as well as which students decide to visit such countries. It also can be useful for schools, countries and European citizens to visualize how ERASMUS programme is utilized. For example, schools can see popular destination and can try to conclude new agreements with other institutions.

The initial questions we wanted to answer were the following:

- What is the flow of students doing Erasmus between countries?
- Which countries are popular for being an Erasmus destination?
- Is there any difference between bachelor and master’s degrees?
- Does gender effects the choice?
- Does distance from the city of home university to the city of receiving university matter?
- Does the city size of receiving and home university matter?
- How does difference in cost of living in home and target country effect the selection?

During the project, we had redefined the questions, due to scope, and especially course’s criteria on more specific question and available data. The final example questions that is the visualization aims to answer are:

- Which countries do Portuguese bachelor students choose as an Erasmus destination in comparison to masters and PhDs?
- Is it popular for Danish students to go to countries with same or lower cost of living?
- Where do the Erasmus students from Portugal like to go on Erasmus?
- Do Czech females travel smaller distance to their target country than the Czech males?

RELATED WORK

European Commission has some work done on visualizing data from the ERASMUS programme (1), but these are static graphs and the report mainly considers changes from previous years. In addition, the report does not take into account other factors such as cost of living and travel distances.

The initial inspiration for the chloropleth map came from the Medium post by Jules Beley which did a visualization as a part of a tutorial on R data visualization (2). This showed us that there is a potential for a great visualization and that this can be done with such dataset.

THE DATA

The main dataset was obtained from EU Open Data Portal (3). It contains entries for every student abroad for the academic year 2013-2014 with columns containing information about source institution, target institution, gender, degree, and more.

We had the institution and city name for each student. The original idea was to use the city name to obtain information about the place where the student studies and compare those. This has proven to be a difficult task, as the original dataset contained corrupted city names as well as institution names which could not be fixed. This was later solved by joining the original dataset with the EUC dataset (4) (a different dataset that maps the unique institutional ID to a city name which was not corrupted).

After obtaining the city names which were not corrupted, we found out, that the cities were written in different languages (and with obvious hand typing induced errors) which made it impossible to join the city names easily with any geo database such as geonames.org.

Therefore, instead of using city names to obtain the positions, we queried a Google Cloud Geolocation API and got the exact coordinates of each institution this way. From the coordinates, we calculated the distances travelled for each student.

For the use in chloropleth map and sankey diagram, we have precalculated all of the movements (incoming and outgoing student counts for different degrees of each country) to speed the visualization up. Loading the full dataset would be slow and was not needed. This was done by grouping first by the sending and then by the receiving country and then counting the samples. This precalculated dataset is also used in the barchart to sort the bars.

Other than that, barchart uses a dataset with information about cost of living and other price indexes for participating countries. A dataset on cost of living in different cities around the world was obtained from Kaggle (5). Cost of living attribute was calculated per-country as joining based on a city name was previously found to be extremely challenging (duplicate cities, encoding errors, localized names, typing errors).

VISUALIZATION

The layout of our interface is divided into four interconnected views. Each of them takes a different look at Erasmus participants. Each idiom works when no country is selected (Figure 1Figure 2) and also when a country is selected (Figure 4). We can also choose between incoming and outgoing students using buttons on top right of the screen as shown in Figure 2. Country can be selected either by using the dropdown menu on top right part of the page or by selecting the country in one of the different idioms.

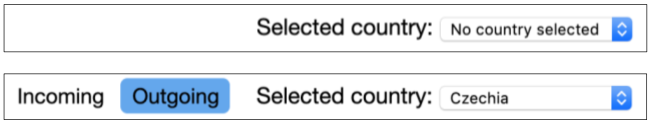

Selected country: No country selected
Incoming
Outgoing
Selected country: Czechia

Figure 2: Controls on the right top part of the visualization. No country is selected (top), Czechia is selected (bottom)

Chloropleth map

In the upper left corner, we have a map of countries participating in the Erasmus programme, which aims to visualize the information about student flows between the countries.

When **no country is selected**, the map shows the ratio of the number of incoming students per one outgoing. This is visualized through a color gradient which is explained in the legend.

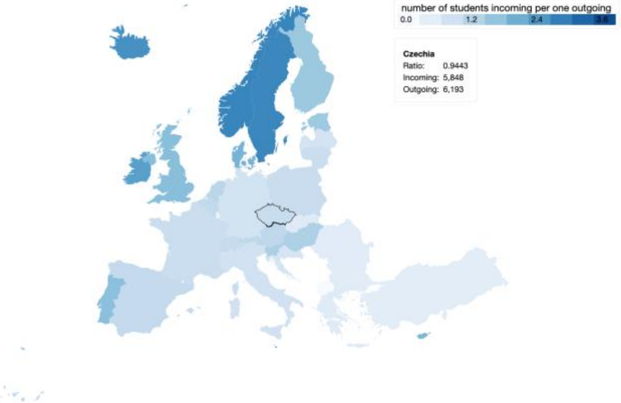


Figure 3: No country is selected, hovered over Czechia

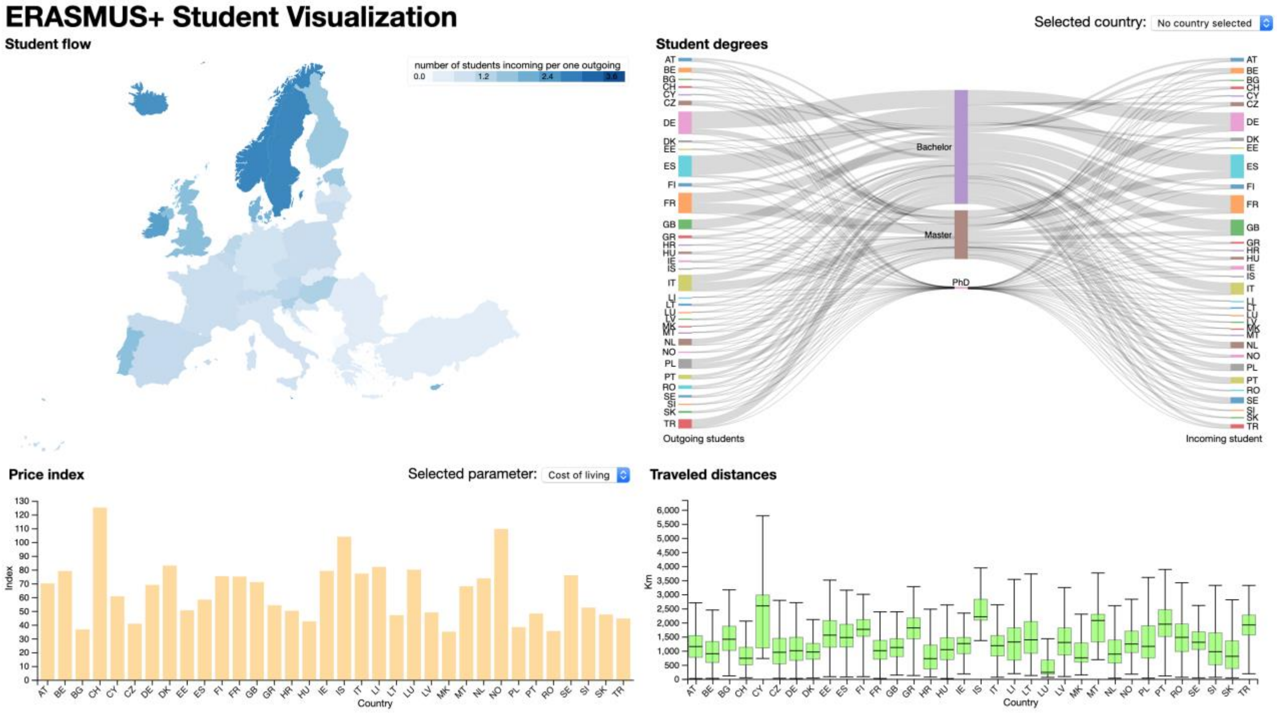


Figure 1: Whole visualization when no country is selected

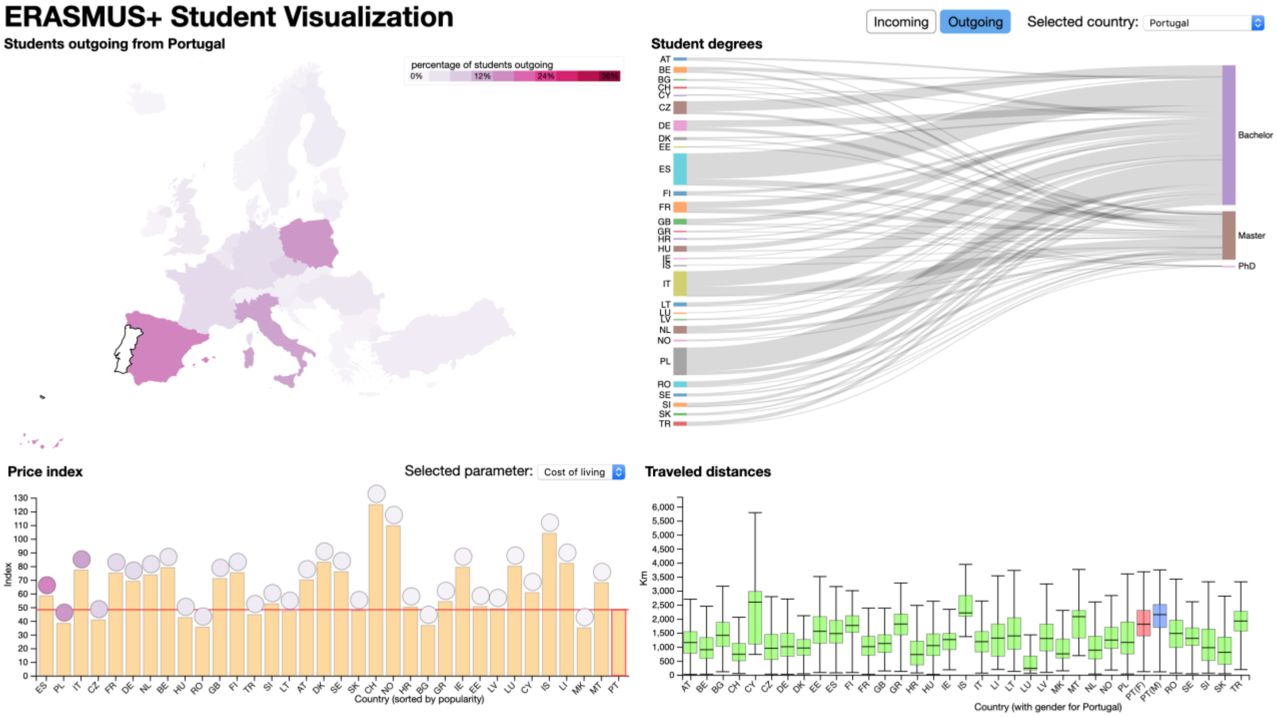


Figure 4: Whole visualization after Portugal was selected

When the user hover over a country, a tooltip is shown and the user can see the exact ratio as well as the number of incoming and outgoing students, as is shown on Figure 3.

When user **selects the country** by clicking on it either on map, or any other idiom, the country gets highlighted and the chloropleth map changes based on the country data and selected direction (incoming/outgoing) – see Figure 5.

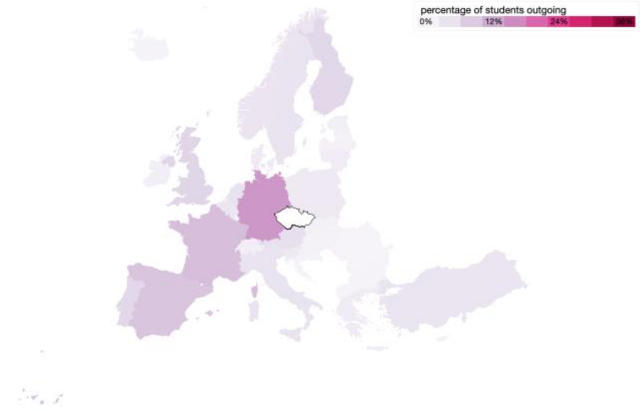


Figure 5: Czechia was selected

After the country is selected, hovering over a different country than selected, shows the connecting lines between institutions located in these countries as can be seen on Figure 6. This is animated and the direction of the animation shows the currently selected student direction. Tooltip now shows the information about the number of students outgoing from Czech Republic to Spain.

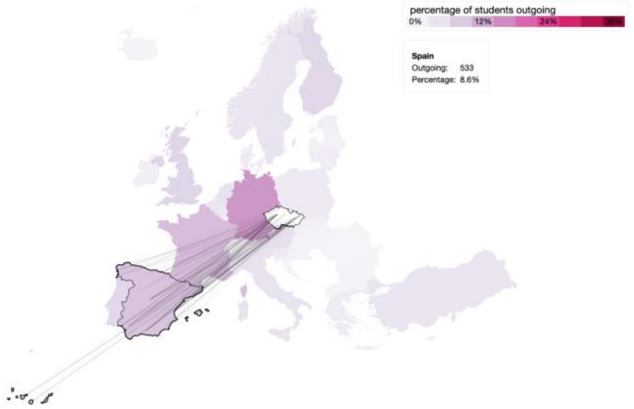


Figure 6: Hovered over Spain

Sankey diagram

In the upper right corner, we have a Sankey diagram. This idiom shows the at what degree level students go abroad and to which country. When **no country is selected**, it shows the degree type of incoming and outgoing students from the whole Erasmus program.

The number of students is visualized by the width of the link between the nodes. I.e. as in Figure 7, when hovering over Italy, we can see the flow of outgoing students from Italy, and the amount at each degree. By reading the tooltip the user can get more detailed information.

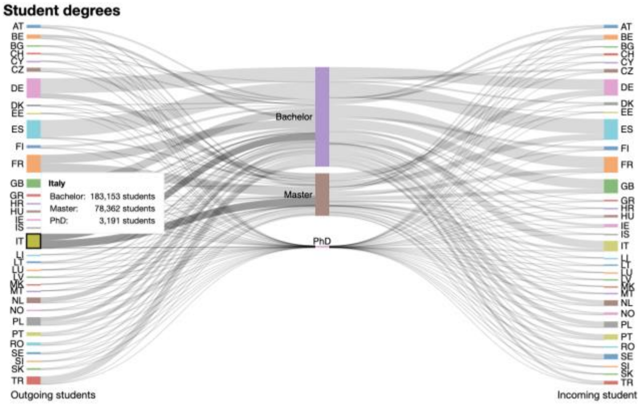


Figure 7: Sankey diagram when no country is selected

When a **country is selected**, you can either see the degree flow of incoming or outgoing students to the selected country. Figure 8 is showing outgoing students from Italy. When hovering over a link a tooltip shows the exact amount of students.

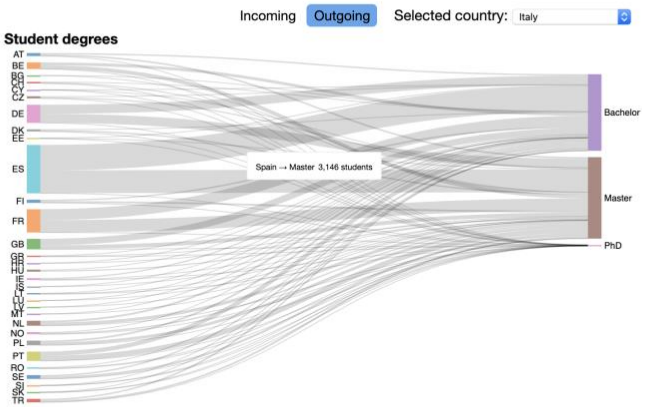


Figure 8: Sankey diagram for outgoing student from Italy

Bar chart

In the lower left we have a bar chart visualizing information about country's different price indexes. User can select cost of living, rental index or price of the beer (Figure 9) to be shown using a dropdown menu situated in right top corner of this section (shown on Figure 10). When **no country is selected**, bars are sorted alphabetically. When the user hovers over a specific bar, a tooltip with information about all three price indexes is shown (Figure 11).

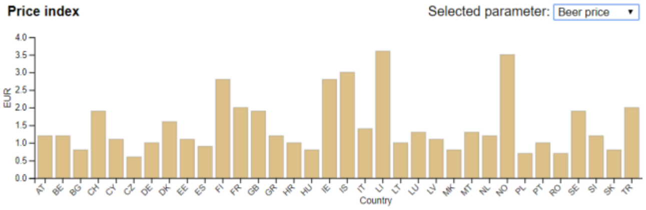


Figure 9: Bar chart showing beer price (no country selected)

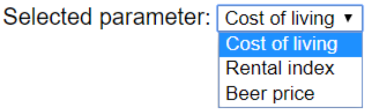


Figure 10: Dropdown menu for selecting the price index

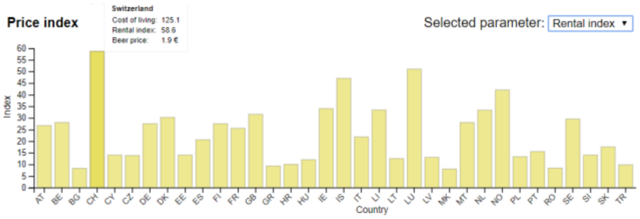


Figure 11: Bar chart – rental index (no country selected)

When the **country is selected**, it becomes a reference to all the other countries. The countries (on the x axis) become sorted by their popularity. User can easily see the difference to the selected country thanks to the red horizontal line, which shows the reference value of the selected country. The bubbles over bars show the popularity of the country with same colours as in the chloropleth map. Bar chart with selected country is shown on Figure 12.

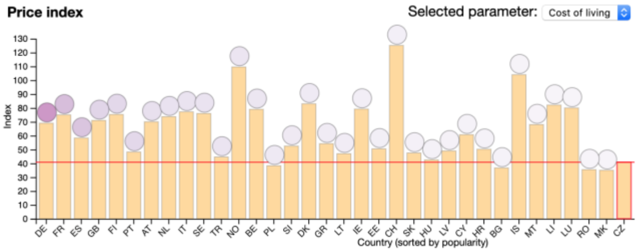


Figure 12: Bar chart showing cost of living when Czech Republic and outgoing direction are selected

Boxplot

Boxplot aims to visualize the distance that the student from a given country had travelled. Once again, when **no country is selected**, it shows information about all of the students traveling from the country (Figure 13).

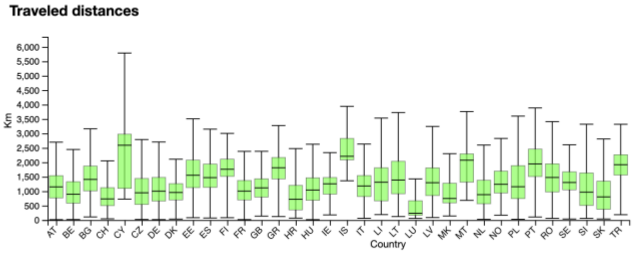
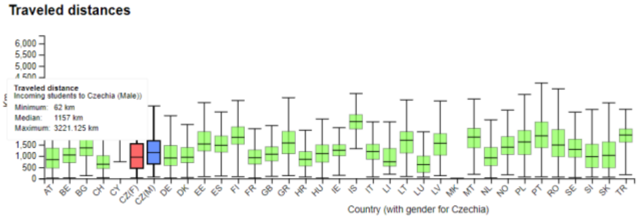


Figure 13: Boxplot (no country selected)

When user **selects a country**, the boxplot for selected country subdivides to specifically shows how far away did male travel compared to the female.

To get more accurate information, the user needs to hover over the box of the country he wants to know more and tooltip with information about minimum, maximum and median travelled distance is show as seen on Figure 14.



Traveled distance
Incoming students to Czechia (Male))
Minimum: 62 km
Median: 1157 km
Maximum: 3221.125 km

Figure 14: Boxplot (CZ is selected, hover over male)

Rationale

We built our visualization on top of the data containing records of students participating in the Erasmus programme. We thought that creating the map would be very interesting as it is a natural way to show geo information.

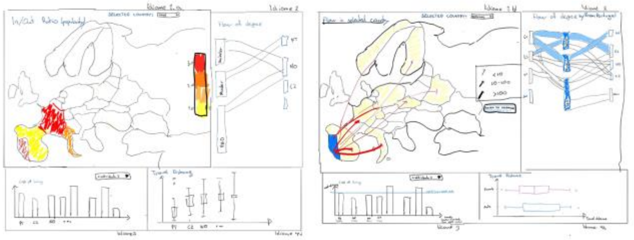
Sankey diagram felt like a good way to visualize the different degrees of the students coming because it creates different groups from the same set of students.

We were really interested from which background the students come, so we decided to use the barchart to explain the data about the country's different financial metrics.

The boxplot is great to visualize data which varies a lot in mean as well as its variance as in case of the travel distance.

Initial sketch

The initial sketch (Figure 15) differs from the first prototype mainly in different orientation of the boxplot which shows the travelled distance of the students when a country is selected. Although making the boxplot horizontal would utilize the space better, we could not easily animate the graph between different states.



We also originally thought that we should visualize the student flows on the map using country-level arrows.

Figure 15: Initial sketch – without and with country selected

First prototype

The Figure 16 demonstrated our first prototype. The boxplot is not implemented. The first prototype also showed all of the incoming and outgoing lines when a country was selected which was optically disturbing (section *Development of the chloropleth map* explains this further). The first prototype lacked integration between all of the idioms, contained several errors and also tooltips, which turned out to be extremely useful for the user.

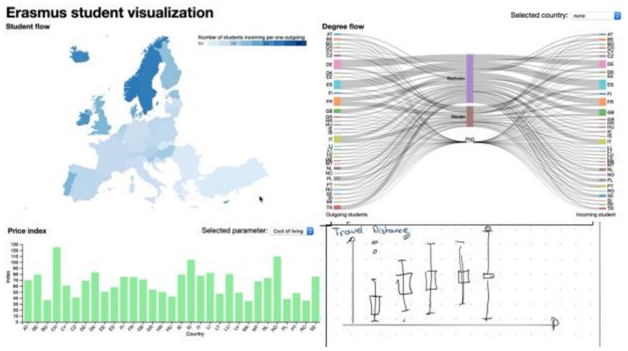


Figure 16: First prototype

Chloropleth map development

The original idea was to visualize the flow of the students between the countries using arrows, where the arrow width also shows the number of students incoming or outgoing from the country (see Figure 17).

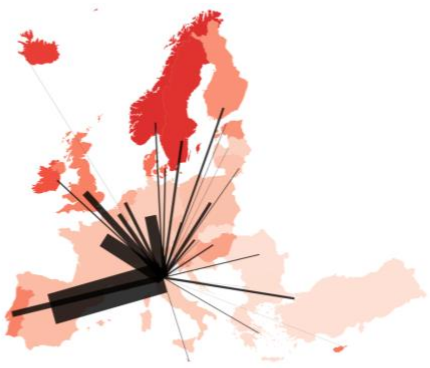


Figure 17: Country–level lines

However, this information is already expressed in the chloropleth map when a country is selected. So the user is not gaining any new information. Also, it would be challenging to correctly implement the arrows as some of the countries have a very large flow which would make some arrows very wide. Also, using the arrows on a country level may mislead the user to the idea that the students are going just from one city of the country since the arrow starts from and ends in one point.

Since we had obtained the exact coordinates, we were able to visualize the exact places from and to where students go. The result of doing so was promising, however since there was a very large number of lines, the map got too cluttered and the information was not visible at all. We tried to solve this by adjusting the transparency, but the visualization was still not readable, as can be seen on Figure 18.

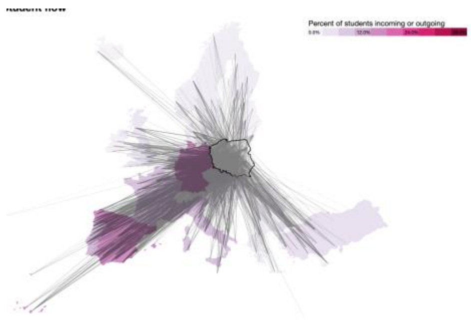


Figure 18: Individual lines for each student

The final solution we came up with was to show the connections only when a country is first selected and the user hovers over another country. This solution provides the user with interesting information without the drawback of being too distracting (Figure 19).

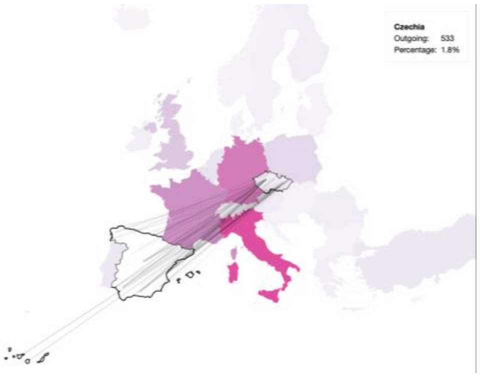


Figure 19: Lines are shown only when country is selected and the user hover over another country

Other ideas and visualizations

Interesting visualization which would show the numbers of students travelling between countries could be done by visualizing the matrix containing student counts in rows as well as columns (Figure 20). This way, we can see easily the biggest flows between countries. This was done just as a proof of concept using the matplotlib library. We did not use this visualization as this data can be figured out using the chloropleth map.

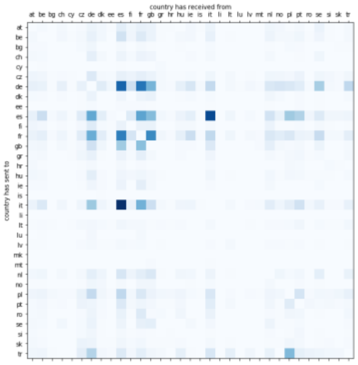


Figure 20: row – receiving, column – sending

Demonstrate the Potential

In this chapter, we describe step by step how to answer question asked in Introduction. A country can be selected in many different ways, here we show just one example in the first question.

Which countries do Portuguese bachelor students choose as an Erasmus destination in comparison to masters and PhDs?

Portuguese bachelor students go mostly to Spain, Poland and Italy. In comparison with that, Portuguese master students go mostly to Italy, then to Spain and then to Poland. There are no Portuguese PhDs students going to Poland. For description on how to get the answer, see Figure 21 and Figure 22.

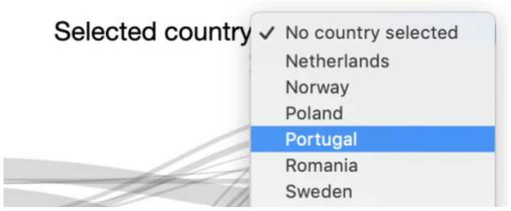


Figure 21: Select Portugal

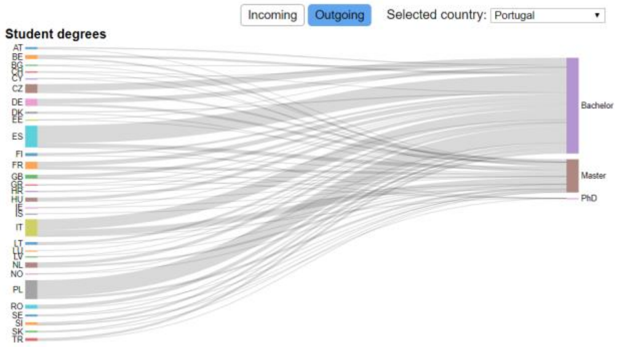


Figure 22: Look on the Sankey diagram

Is it popular for Danish students to go to countries with same or lower cost of living?

Top 5 countries where Danish students go have same or lower cost of living can be seen in the Figure 23 as the 5 left-most columns in the barchart.

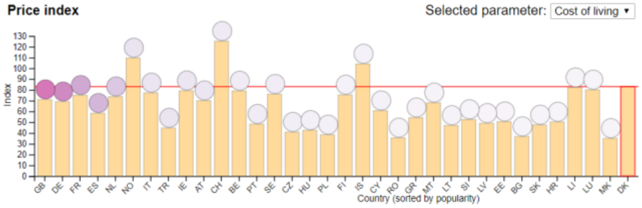


Figure 23: Select Denmark and look on the bar chart to compare the bars, which are now sorted based on the popularity of the country for Danish students, with red line

Where do the Erasmus student from Portugal like to go on Erasmus?

Top 3 countries for students from Portugal are Spain, Poland and Italy (as shown on Figure 24 and Figure 25).

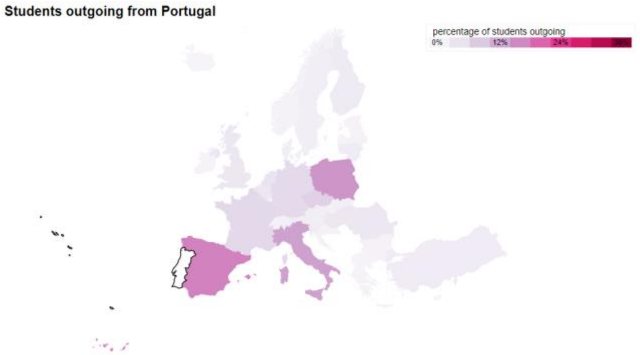

Students outgoing from Portugal
percentage of students outgoing
0%
12%
24%
36%

Figure 24: Select Portugal and Look on Chloropleth map, Color says how popular every country is.

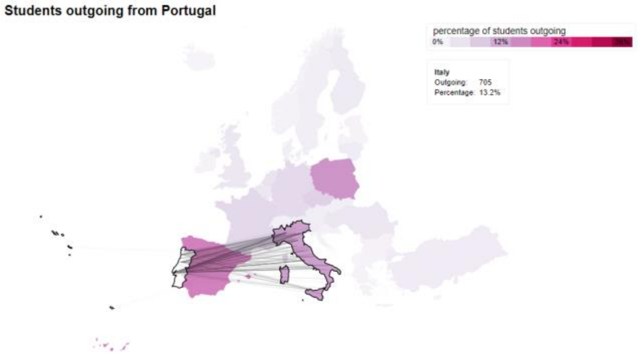

Students outgoing from Portugal
percentage of students outgoing
0%
12%
24%
36%
Italy
Outgoing: 705
Percentage: 13.2%

Figure 25: For seeing more accurate information for a specific country, hover over the country – tooltip says exact number and the links shows universities, where the students travel

Do Czech females travel smaller distance to their target country than the Czech males?

Czech females and males travel similar distances, but Czech males travel little bit further. How to answer this question is shown on Figure 26 and Figure 27.

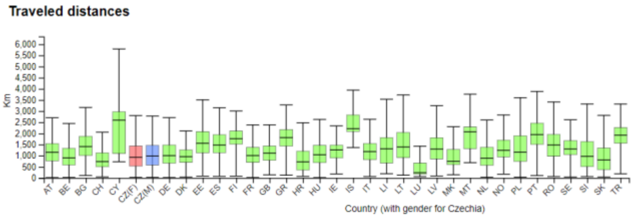


Figure 26: Select Czechia and look at the boxplot. Two boxplots for Czechia compare Czech females and Czech males.

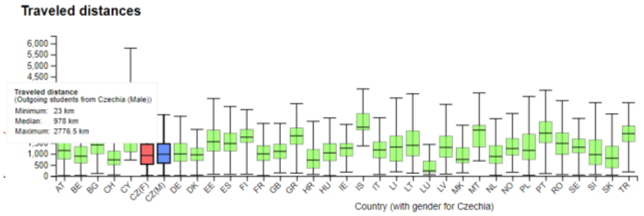


Figure 27: For seeing more accurate information for a country, hover over the box – tooltip says exact number

Surprising Information that we found using our visualization

Thanks to our visualisation, we discovered several interesting facts and trends:

- There is almost no exchange of students between northern countries.
- Portuguese students prefer Spain, but Spanish students prefer the Italy (and Italians prefer Spain).
- Macedonia has no incoming students. Not even one Macedonian goes to Greece.
- Girls usually travel slightly further than boys.
- Every country has some student travelling really near (except for islands). For example, Slovaks have minimum distance of 30 km. There were students only 16km from their home university on border of Austria and Hungary.
- Sweden and Norway have many people coming in compared to going out although they are countries with high cost of living.
- People in Greece almost never go on Erasmus during their master's degree.
- Most of the people that come to Britain are from very comparable price index countries.

IMPLEMENTATION DETAILS

Although we have tried to use the Pentaho Data Integration tool to prepare our dataset, it seemed unstable and was not performing well. Since we had experience with Python library Pandas, we have generally preferred Pandas for data preparation.

Chloropleth map

Chloropleth map is drawn using a JSON containing the country shapes which are parsed and decoded with the help of the TopoJSON library. To draw the map and lines, we use the Mercator projection.

The connecting lines are straight which allow us to easily animate the student's direction of travel. We also tried implementing the curved lines which respect the projection, however the curvature was almost neglectable and we could not easily animate the direction of travel.

Sankey Diagram

In order to make the Sankey diagram we used a d3-sankey, that is an extension of D3. This required the data to be on a certain format. Simply explained, each node had to be in a list with indexes in place where we wanted to position them. In addition, each flow link data had to be on the format {source: x, target: y, value: z}. This was done in Jupyter notebook with Python and Pandas. When this data format was obtained a function computed the positions of the node, as well as the node and link objects. This extension facilitated the implementation of this more complicated graph, but we still had to craft the visualization our self.

Bar chart

Bar chart simply visualizes the selected measure by the user. We have utilized the provided example from the laboratory as well as example from D3.js Graph Gallery (6) a starting point when creating this visualization.

Box plot

Box plot is the only visualization which utilizes a large dataset which contains the travelled distances of each student. An example from D3.js Graph Gallery (6) was used as a starting point for this idiom.

Interconnection and linking mechanism

The whole visualization works in two modes: without, or with a country selected. The user can select the country either by using a dropdown menu at the top of the screen or by clicking at the state in any of the idioms. The student direction is selected by using two buttons which are hidden when no country is selected.

The whole visualization is interconnected by using 3 global variables: *selectedCountry, highlightedCountry* and *studentDirection*. These variables are changed through 4 different d3 events: *stateOnMouseOver, stateOnMouseOut, stateSelectedEvent* and *studentDirectionEvent*. These can be invoked from any of the idioms with the state code as their parameter. The global variables are updated based on these events.

CONCLUSION AND FUTURE WORK

During our work on this visualization we learned, that we are capable of using the D3.js library to create stunning visualizations on top of the given data.

Our primary interest was to visualize the flow of Erasmus students, but during the project, the questions were changed to more concrete ones which also utilized the other idioms. Part of them are now better specified but part of them had to be changed or deleted. These questions were especially about cities (their sizes etc..). The reason for their deletion were problems with the merging data based on city names.

We think that it is disappointing, that the data is available only until the year 2014 as we would love to visualize more recent data. On this purpose, we have contacted the European Open Data Portal with a question about this but did not get the answer so far.

The data from years prior to 2014 is available, so it would be possible to create a visualization, that takes in mid the changes over time. For this project, this was decided as out of scope, but it would for sure provide some exciting insights.

For visualizing the travelled distances by boxplot, we use a fairly large dataset, which is processed in the browser. It therefore takes some time to load the boxplot. For future work (and better online performance and scaling) it should be possible to precompute this data for each country prior to downloading it.

It could be also interesting to utilize data about more variables that can influence the student's decision (such as ranking of the university – we were not able to find an unpaid dataset for this topic).

Our current solution works well on high-resolution laptop screens for which it was developed, but it is not responsive which makes it useless on mobile and old devices. This is an improvement which also could be done with more time.

REFERENCES

1. **Comission, European.** Erasmus – Facts, Figures & Trends. *The European Union support for student and staff exchanges and university cooperation in 2013-14.* [Online] https://ec.europa.eu/assets/eac/education/library/statistics/erasmus-plus-facts-figures_en.pdf.

2. **Beley, Jules.** Making a map with EU data on R: Erasmus exchanges by country. *Medium.* [Online] https://medium.com/@jules.beley/making-a-map-with-eu-data-on-r-erasmus-exchanges-by-country-3f5734dcd4ff.

3. **Portal, EU Open Data.** *Raw data of Erasmus staff mobility for training in 2013-14.* [Online] https://data.europa.eu/euodp/en/data/dataset/erasmus-mobility-statistics-2013-14/resource/22ba16a1-d3aa-4e91-accc-af8e1f1a6584.

4. **EU Open Data Portal.** List of all institutions participating in the programme for the academic year 2013-2014. [Online] https://data.europa.eu/euodp/en/data/dataset/erasmus-mobility-statistics-2013-14/resource/9ad29b2b-d63f-4660-ba21-772925d57362.

5. **Tran, Andy.** *Kaggle.* [Online] https://www.kaggle.com/andytran11996/cost-of-living/version/3.

6. **Holtz, Yan.** D3 Graph Gallery. *Boxplot.* [Online] https://www.d3-graph-gallery.com/boxplot.html.

7. —. D3 Graph Gallery. *Barchart.* [Online] https://www.d3-graph-gallery.com/barplot.html.