Checkpoint V – Group 28

Carlos Sequeira
Nº 87638
Instituto Superior Técnico,
Taguspark

Cassandro Martinho
Nº 87642
Instituto Superior Técnico,
Taguspark

Rui Nóbrega
Nº 87703
Instituto Superior Técnico,
Taguspark

ABSTRACT

Nowadays information is all around the web but most of the times is hard to find and is spread across different locations. Our visualization was designed to solve all those issues by displaying information about the demographics of Stack Overflow users. The data that our solution was built on is a dataset that contains the answers to a survey which users from Stack Overflow answered, covering everything from developers' favorite technologies to their job preferences. To built our solution, we developed the project with D3 Version 6 and it makes use of idioms like a choropleth map, a radar chart, a sunburst and finally an idiom composed of an image of a computer with the stack overflow symbol, this picture will be measuring our selected attribute in the configuration panel by how much of the image is colored. To answer all questions someone may have. the visualition can be controlled and adjusteb by a set of filter options.

Author Keywords

D3; Choropleth map; Radar Chart; SunBurst; Stack Overflow; JavaScript; CSS; HTML; Visualization;

INTRODUCTION

"Visualization gives you answers to questions you didn't know you had." – Ben Schneiderman

As the years go by, technology is rapidly evolving so, working in Computer Science became a very popular career choice. In order to get a better insight on the demographics of such area, we decided to turn to the organization that best describes the programmers/developers of the world.

Every year, Stack Overflow conducts a survey that addresses everything from favorite technologies to their job preferences. As mentioned by Stack Overflow itself, such survey is the largest and most comprehensive survey of people who code around the world. By using the dataset from the survey, we get to understand that the audience is as wide as one could think ranging from students with almost no experience to professionals with years and years of work experience.

In our visualization, we will be able to understand thing like where the users are located, which programming languages they work with, their job satisfaction, they salary etc. We can then cross reference these data with other data like gender, ethnicity, age, years of experience or how many work hours per week.

Nowadays, the internet contains the answer to everything we questions but things might not be that simple. Most of the time when we need to analyse information from different countries we have to look for data in several website masking the whole process annoying and hard. The solution proposed in this paper is a visualization capable of answering a whole set of questions in one place. With our visualization, someone who is struggling to decide where to work after college, for example, will easily get all of his/her questions answered. By knowing all the data like salaries or the amount of work hours per week from the Stack Overflow users, a decision can be taken accordingly.

With all being said, our solution intends to answer questions like:

1. What country has the most users on Stack Overflow with more than 10 years of code?
2. How many full-time students use Stack Overflow with a bachelor?
3. Average work week hours per man per country?
4. What's the most popular coding language for people younger than 25 years old and have at least 1 year of coding?
5. Average salary in the USA for an Asian American?
6. Education difference between employed and unemployed people?

RELATED WORK

In order to accomplish any of the visualizations we used the help of many websites being these the most important ones:

- https://www.d3-graph-gallery.com/
- https://d3js.org/
- https://observablehq.com/@d3/gallery
- https://www.d3-graph-gallery.com/donut.html
- https://observablehq.com/@d3/sunburst
- http://bl.ocks.org/Kuerzibe/338052519b1d270b9cd003e0fbfb712e

The last three links were especially important for the development of the radar chart and sunburst as they were bigger and required more knowledge of the d3 library, especially the second last time

THE DATA

In this section, we present the core of our visualization, the data itself. The dataset that we based our work on is called "Stack Overflow Developer Survey Results 2019" and can be found on kaggle [1].

Initially, the dataset was composed by nearly 88863 instances (users that took the survey) and a total of 85 attributes describing all the questions that each person answered regarding everything from developers' favorite technologies to their job preferences.

Data Processing

When it was time to analyse the data we had in order to set everything up before starting the implementation of the code, we realised that there were major flaws with the data that needed either some cleaning or some fixing.

Firstly, the amount of information portrayed in the csv file (later converted to a json file) was excessive as it was impossible for us to display all the information in the visualisation in an intuitive and pleasant way so, to solve such issue, we decided to reduce the size of the attributes at hand, pruning the 85 attributes ending up with just 16. To end up with such a reduction on the size of attributes, when pruning the process was not to just delete the unwanted columns but we also merged , for example, the transgender column into the gender column enabling us to save space by not having the transgender column. Meanwhile, it is represented as an option in the Gender column.

Secondly, to clean the dataset, we started by handling the values of the column "Salary in USD" in a way that if there were any values set to unemployed in the column "Employment", we assigned the value 0 in “Salary in USD” since it would not make sense to have a salary. Furthermore, entries with odd values for people who answered "employed" in the "Employment" attribute, were considered as missing values, for instance, an employed person claiming to only earning 50 dollars a year or working more than 120 hours a week. Depending on a combination of values, some attributes entries were predicted with the use of the KNN classifier, thus, identifying a possible and coherent value for that entry based on neighbors with similar values. In some cases, some instances had a lot of missing values and it would not make sense to predict values for such a high number of entry so, those instances were deleted from the dataset. After, propertly, cleaning the data, the dataset was reduced to 74380 instances expressing a reduction of 17% of our data. Such value of reduction is not to be considered as a bad aproach since the data deleted made no sense and predicting that much values would affect the integrity of the data.

Thirdly, to the 16 attributes, a new one was added, making it 17 attributes in total. This last attribute is named "Dollars per hour" and it was calculated by dividing "Salary in USD" by "WorkWeekHrs". Such measure is of the utmost importance as it allows the user to see the pay based on work rate.

Last but not least, two datasets were added to the project. The first one is called "world.json" and it contains the name of the countries and its coordinates. The purpose of such json is to aid in creation of the choropleth map, later explained in 4.1. The second dataset created is called "json_map.json" and it was obtained by calculating the average numbers of total users (and percentage), age, salary, dollars per hour, working hours per week and years of coding experience for each country mentioned in the main dataset. Since the main dataset could not be parsed into smaller datasets, we face some scalability issues when sometimes, based on the requirements of the user using the visualization, the response time can be higher than desired and other times it can be faster.

VISUALIZATION

The purpose of this section is to talk about the visualization developed and how it works. Furthermore, it is also displayed the process of decision that was taken in order to choose the idioms we chose. Finally, we intend to show the real potential of our solution by executing a specific case where someone could use the created dashboard.

Overall Description

As previously mentioned, our solution intends to display the demographics of the Stack Overflow users. An overview of the solution, here presented, can be found in Figure 1 and consists in a filter section and four idioms to show the data.

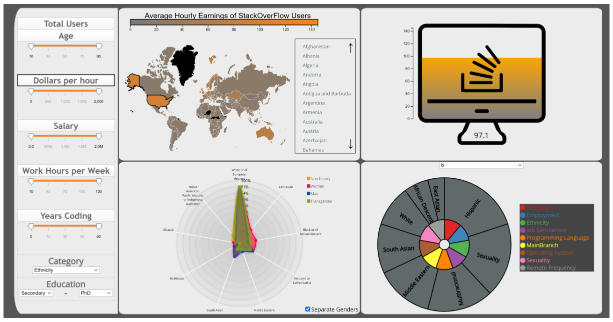


Figure 1. Overview of the entire visualization

The idioms presented in our dashboard are:

1. A choropleth map that can be found in the top left side of our visualization.
2. A Radar chart that is located at the bottom left side of the dashboard.
3. A Sunburst at the bottom right of the visualization.
4. A color based new idiom, invented by us.

As seen in figure 1, there is also a filter options where there are 6 main buttons, that if clicked perform a reset in the dashboard and perform the following:

1. "Total Users" that if clicked, it will alter the visualization in a way that the choropleth map will display information about the total users of Stack Overflow and the percentage that the users from a specific country represent in the entire Stack Overflow community.
2. "Age" will alter the map to display the average age of the users from a selected country.

3. "Dollars per hour" makes the map show the derived measure in which it displays the hourly earnings around the world.

4. "Salary" alters the map to show the average salary of the users of answered the survey.

5. "Work Hours per week" changes the world map to display the amount of work hours the average user is obligated to.

6. "Years of coding" displays the how many years of coding experience each subject has on average.

Furthermore, the filter options also have sliders in which someone using our dashboard can apply filters to the information displayed in the idioms. Such sliders were designed to reduce the population and adapt the data to the options the users of our solutions desires. Lastly, there are 3 dropdown filters, 1 for a category at choice and 2 other to represent the minimal and maximum level of education degree that is intended to look for, for instance, someone might only desire to see information for people who have at least a bachelor´s degree and less than a PhD.

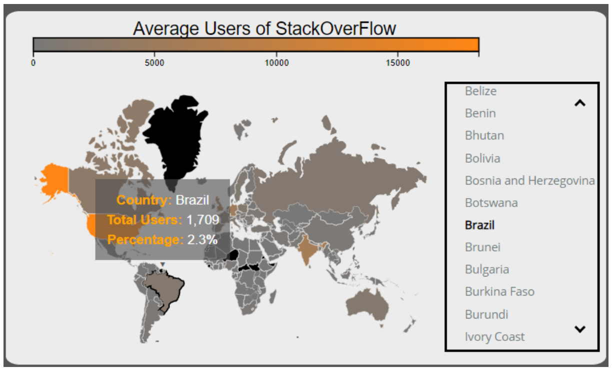


Figure 2. Choropleth map

The first idiom to be described is the choropleth map as shown in figure 2. The map displays different colors representing the attribute value of the selected visualization attribute. On the right side we have a scroll-bar with all the possible countries, in order to facilitate the search of a country. When clicking on a country on the list, that specific country is also highlighted and selected on the map. In the event of a "mouse over" a country on the list, that specific country is also highlighted on the map but it is not selected as when clicking and the highlight disappears when the mouse leaves the name of the country. We can also select countries by clicking on the map and each time a country is selected on either the map or the list, the information displayed in all idioms is related to the selected countries. When there is no highlighted countries all the countries will display color depending on the chosen visualization attribute and all the other idioms will display information as if all countries were selected.

Furthermore, the map still has one more functionality. If someone who is using the map is not sure if it desires to select a country or even where a country is, the user can, as seen in figure 2, put the mouse on top of a country and a tip with information regarding that country will appear. Lastly, the countries that are black colored are not clickable or selectable through either the map or list and the tip functionality does not work since the survey, previously mentioned, did not contain data regarding those specific countries.

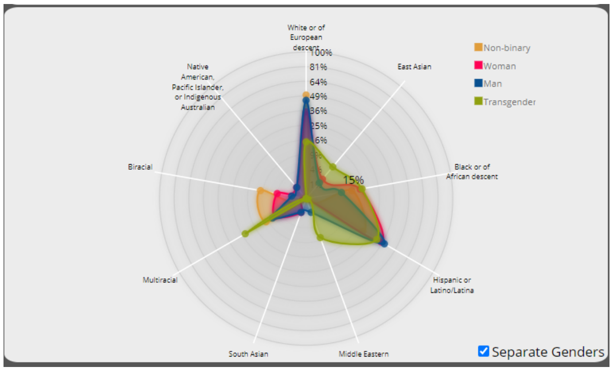


Figure 3. Radar Chart

The second idiom to be presented is the Radar Chart as seen in figure 3. This idiom displays information regarding the countries in the map or the data for all countries if no specific country is selected. The data that it shows is the category selected in the "Category" dropdown filter in the filter options on the left of the dashboard. If the user does not select a category, the category, as predefined, is selected as "Ethnicity", however, if the user selects a category then it will be shown in the radar chart. The radar chart displays the percentage of each entry of the category selected and separates each entry in genders so each color represents a gender and the legend is on the right of the radar chart but if the user does not want to see data by gender then it can deselect the option "Separate genders" and the data will be displayed by just a color. The scale used to draw each polygon was the square root due to the fact that by using such scale, shorter values distance from the the middle and can be seen easily.

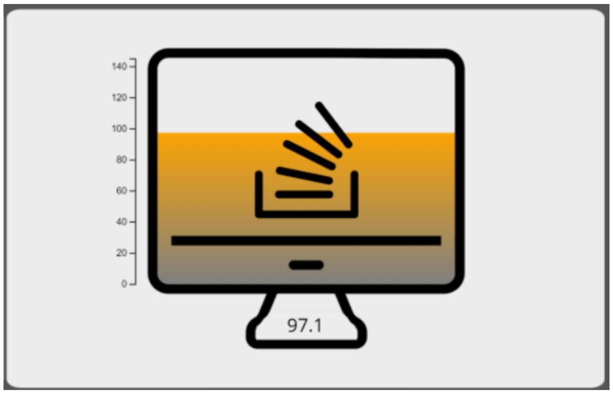


Figure 4. New Idiom invented by us

The third idiom is the one invented by us and an overview of it can be found in figure 4. This complex visualization is composed by an image of a computer with the stack overflow symbol, this picture will be measuring our selected attribute in the configuration panel by how much of the image is colored, starting from bottom to top like the scale on the left indicates.

At the bottom of the computer, in its base, it is indicated the value of the attribute that we are measuring for the selected countries or for all countries if none are selected.

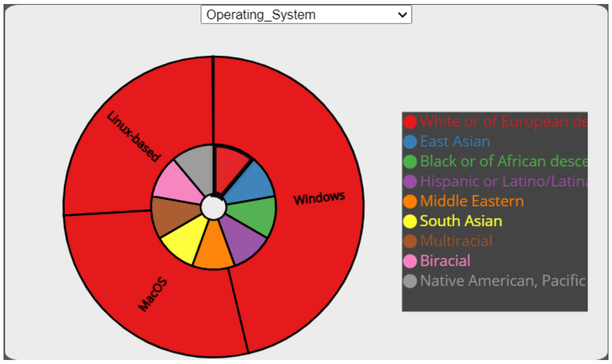


Figure 5. Sunburst invented by us

The fourth idiom is a sunburst that enables to cross information from our database, the inner circle is clickable and will restrict all the displayed data on the outer layer to be instances of that subcategory (the inner one). The first and inner category is the one defined in the settings tab and the second category (outer ring) can be changed by the drop down menu on top of the visualization. The inner ring is separated by color and we can find a meaning to those on the right where there is a name linked to each color, the outer ring however is initially displayed as grey because it is not linked to any inner subcategory, as we click on a inner subsections the outer ring adapts to the new restriction in the data and changes its colour to reflect that. With this visualization we can visualize the number of users that cross these category's where the first category behaves like a restriction and we can see the different distributions of the second category in our data set. Additionally, all the selected countries in the choropleth map and all the settings in the settings bar apply to this visualization and we believe that this provides a big sense of freedom to the user where he can explore the data set freely.

Demonstrate the Potential

After developing all the idioms we can now answer the questions described in the first checkpoint. Below there are some demonstrations on how to proceed to answer these questions.

What Country as the most users on Stack Overflow?

First we need to select the button "Total Users" on the configurations section.

Figure 6. Total Users Button

Then in the second step we just need to look at the map and verify which country on the map has the brightest orange color.

We can easily identify USA as the country with the brightest color ant therefore it is in deed the country with the most users on Stack Overflow. If we don't recognize the country we also do a mouse over it and a tool tip will show us the country name and its total users.

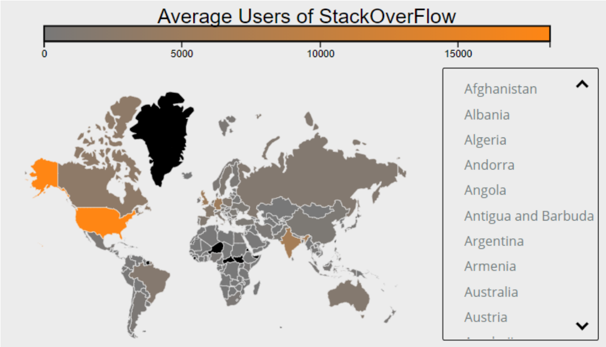
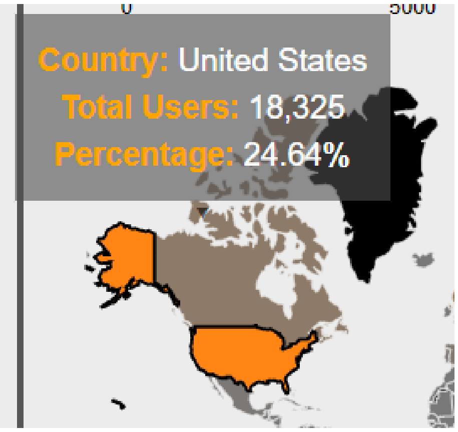


Figure 7. Map Idiom with Total Users



Country: United States
Total Users: 18,325
Percentage: 24.64%

Figure 8. Map Tool Tip of USA

On average, how many Work Week Hours does the population of each Country have?

First we need to select the button "Work Hours per Week" on the configuration section.

Figure 9. Work Hours per Week Button

Then we can use our tool tip do identify the averages of the Work Week Hours for each country. Let's use Portugal as an Example:

As we can see in the image above Portugal as an average of 41 hours of work per week.

Most popular coding language for people with at least two years of coding?

To answer this question we first need to filter out people with less than two years of codi

IMPLEMENTATION DETAILS

CONCLUSION AND FUTURE WORK

Sunburst

When it comes to the sunburst visualization there is a major improvements to be made, this is due to the fact that the displayed names are the names directly used in the csv used

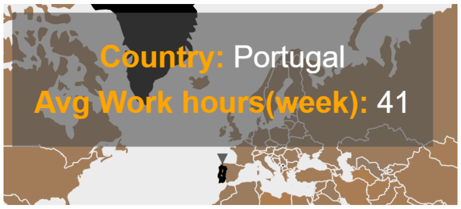


Figure 10. Portugal Tool Tip

for this vis and therefore some of them are too long to be easily displayed anywhere. It would be useful to map the current category and subcategory names into shorter names that are easily displayed when we lack the space like in this case.

Radar Chart

In our radar chart visualization there are two improvements that we would like to see implemented in the future, the first is related to the positions of the names displayed as we feel like they are not easily read, the other issue is due to the fact that currently it is had to select the smaller ares displayed in the radar chart because they overlap with bigger ones that cover the same area and therefore make it harded to select the smaller areas displayed, we wish to further improve the vis by making smaller areas get priority.

Conclusion

To conclude, our visualizations offers much freedom to the user, as he is able to explore and discover much of the data involved in the users of the stack overflow website, there are some points that we would like to work in the future but as it is it provides a good exploration of our dataset.

REFERENCES

[1] 2019. Stack Overflow Developer Survey Results 2019 | Kaggle. `https://www.kaggle.com/mchirico/stack-overflow-developer-survey-results-2019?select=survey_results_public.csv`. (2019).