Project: Visualization on Technologies

CSE564 - Visualisation Project Proposal

Team: 66

Ranjith Reddy Bommidi

114241300

Manish Reddy Vadala

114190006

Instructor: Klaus Mueller

Department of Computer Science

State University of New York at Stony Brook

April 2021

1 Background

The data that we chose has diverse representation and it's one of the largest data of developer community that stack overflow has collecting over the past decade. Every year stackoverflow conducts survey and collects data all around the world. In the year 2020 stack overflow has collected data from over 65000 people in the developer community. This diverse data contains the information given by people from different parts of the world with different ethnic groups, with different levels of experience , their interests in coding languages, their salaries, level of education and most popular/desired technologies among the developer community and many other things. In total the number of attributes that are present in each dataset is 60.

Dataset: `https://insights.stackoverflow.com/survey`

1.1 Why this Dataset?

We have been in love with technology, since childhood. Every tech enthusiast would like to know the current world stats about the technologies in the industry and some insight over that. Here in our project we would like to provide correlations between our attributes and meaningful insights from the stats. Hoping this might help many upcoming developers including myself.

2 Problem

- Developers in the tech industry need to keep themselves updated on latest technologies so they might have some insights about what technologies are used across different parts of the globe.
- People might not have enough insights about the technologies used in the evolving market and the level of expertise they need, pay-scale they get for the positions they apply.
- People should also know how technology trends have changed over the years, because learning a stable technology might help people to survive at least few years.

3 Approach

3.1 Why do we need to visualise the data?

With the current number of technologies that are in market, we need to get enough insights on technologies and other attributes. So with the help of data visualisation tools in this project we will provide meaningful insights about the latest market trends that helps developers to choose their path and many other relevant things.

3.2 What are we planning to do with the data?

We are Planning to use the datasets of the past 4 years and show how the trends have changed over the years, and show the important trends in each and

every year, if possible we would also like to predict the trend for 2021. More information regarding the dashboard and it's content is mentioned below.

4 What does the Dashboard Contain?

- **Barchart**: First part of our Dashboard is barchart which shows the selected attribute. This shows the stats of many attributes, providing meaningful insight to the user. Selecting the attribute as Technology shows the user which technologies has been used by most of the people in the developer community.
- **WorldMap**: Second part of our Dashboard shows the worldmap. Here we are going to show the difference in countries in terms of attributes. Such as average pay-scale of developers, average age of a developer, average number of hours they put in week, etc from country to country, from year to year.
- **PCP Plot**: Third part of our Dashboard shows the PCP plot. Here we are going to show the Correlation between various features like technologies, pay-scale, level of experience, work satisfaction e.t.c. From this we can deduce some interesting visualizations like which set of people are more satisfied with their work.
- **Scatter Plot**: Last part of our Dashboard shows the Scatter Plot. Here we are going to show the correlation between two attributes and would also like to add a third attribute as a filter. Ex: Technologies vs salaries, Roles vs Salaries, Roles vs Experience, we will also add a third attribute such as country as the filter.

5 What are we getting out of this Visualization?

The Visualizations which we show helps us to know about the developers around the world and many more attributes about them.

- What technologies are majorly used around the world?
  - From the respondents response on their current technologies use, we can deduce the technologies used around the world. This can be shown in Barchart.
- What technologies are most loved around the world?
  - From the respondents response on their desired technologies, we can deduce the technologies they love to learn by majority. This can also be shown in Barchart.
- What set of technologies are getting highest salaries?
  - From the attributes Technologies and their salaries, we can find which set of technologies are getting paid more. This can be shown in ScatterPlot.

- Which set of people are more satisfied with their work?
  - From multiple attributes such as work hours, technologies and work satisfaction, we can find out which set of people are more satisfied with their work. We can show this in the PCP Plot.
- Which roles are the highest paid around the world?
  - From the developer type attribute and the salary we can deduce which roles are highest paid around the world. This can also be shown in ScatterPlot.
- What are the average salaries across the world?
  - From the attributes country and the salaries, we can deduce the average salaries from country to country. This can be shown in the Worldmap.
- How the trends are changing over the years?
  - From the datasets of the past 4 years, we will analyze the trend for major attributes for each and every year.

6 Rough outline of the dashboard

The following image roughly shows about our dashboard which we are planning present in the final project.

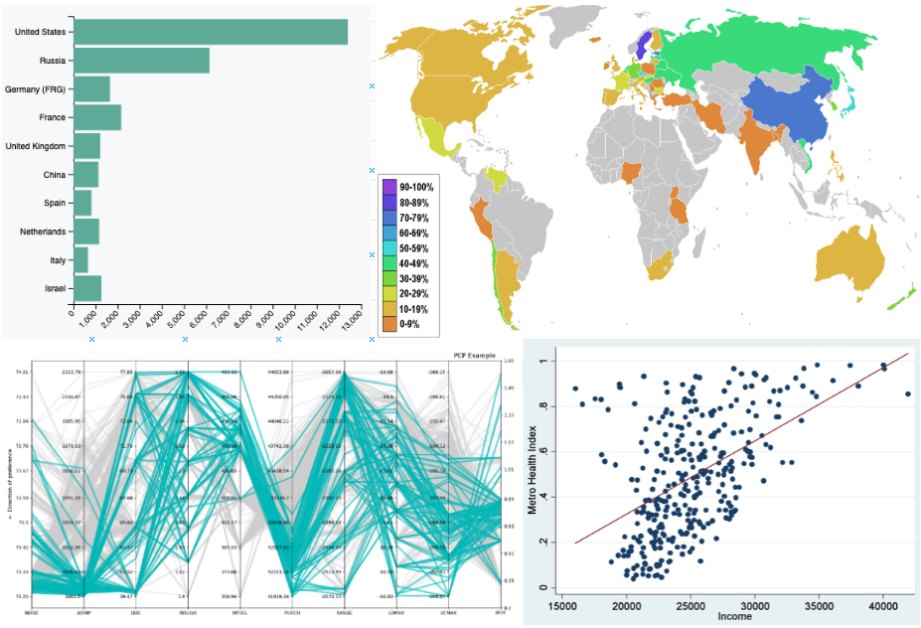

Figure 1: The Researcher and Supervisor

Project: Visualization on Technologies

CSE564 - Visualisation Preliminary Report

Team: 66

Ranjith Reddy Bommidi

114241300

Manish Reddy Vadala

114190006

Instructor: Klaus Mueller

Department of Computer Science

State University of New York at Stony Brook

April 2021

1 Progress

1.1 Back-End Work:

For this project we have used stack overflow developer survey dataset for years 2019,2020. We have merged these two datasets as most of features were similar for these two years compared to previous years. In future, we are planning to use for last 4-5 years stack overflow data. Before attempting to use data we have made effort in cleaning data manually for few columns that contains categorical values. Flask is used as backend server to interact with dashboard. Data that we are handling in server is of 80M currently with dataset size (153344 X 30). For user interaction we need to quickly interact for which data is loaded on start of server. Data is handled as per API calls with the dashboard. Data is handled for all filters that are selected by client.

1.2 Front-End Work:

- ***BarChart***: We have implemented BarChart functionality with a filter option, where we can select the filter such as LanguagesWorkedWith, DatabasesWorkedWith e.t.c. And this shows the top attributes on the left.

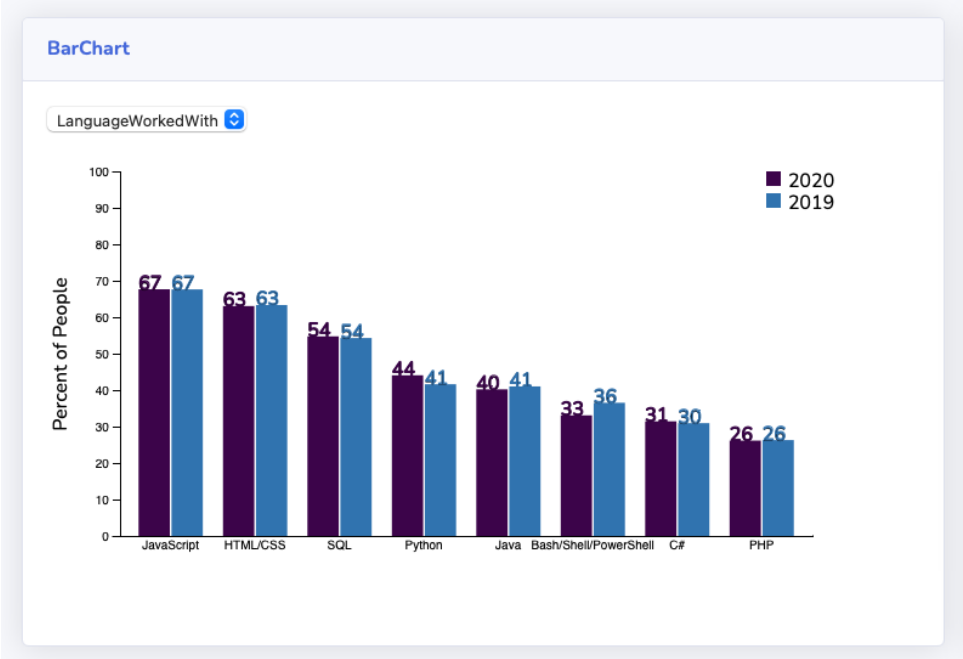

Figure 1: Above BarChart shows the percent of people who work on these languages for the years 2020 & 2019

Observations: We have Observed that the JavaScript is the most used Language across the world in the year 2019 & 2020. We can also see that Python demand is also increased from the year 2019 to 2020. Look-

ing at database technologies, Redis remains the most used, followed by PostgreSQL and Elasticsearch.

- ***WorldMap***: We have also implemented the world map, with on mouse-hover functionality and here we can see the world stats on average global compensations for developers, Average Work hours of Developers with respect to country. Attaching a screen shot below.

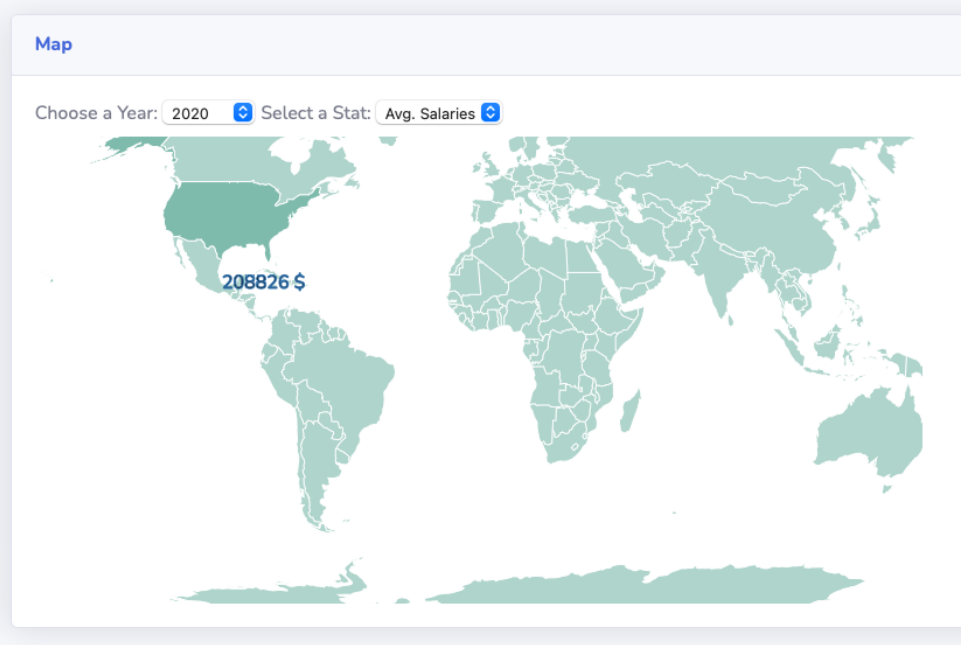

Figure 2: Above World map shows the desired attribute of a Country that is hovered on for selected 2020 year

  - Observations: On Mouse-Hover on a country we can find the desired attribute of that Country.
  - We have observed the Average compensation for United States is the highest among all the countries.
  - From the data observation the Average Age of developers in United States is highest among other countries.
  - We have also observed that Average Years of coding experience among the developers is highest for Australia.

- ***ScatterPlot*** : In our earlier report we planned to present scatter plot with filters X vs Y for features such as Experience vs Salaries, years of experience vs salaries etc. Now we have improvised the scatterplot to bubble plot showing 4 attributes showing size of bubble and colour of bubble as extra 2 features. For example X-axis: Represent Experience, Y-axis: Represent Salaries, Color: Shows the Languages they work on, Size: Shows the No.of People.

Figure 3: ScatterPlot shows the bubble with number of people having same years of software experience working on language having average compensation on y-axis

Observations: After plotting the compensation of people on y-axis and experience on x-axis, we can see some positive correlation between the compensation and experience. People who have more experience likely to have more salaries in the tech industry. Also the majority of the cluster are in the experience range of 1-10 with compensation from 50k-120k USD.

2 DashBoard as of now

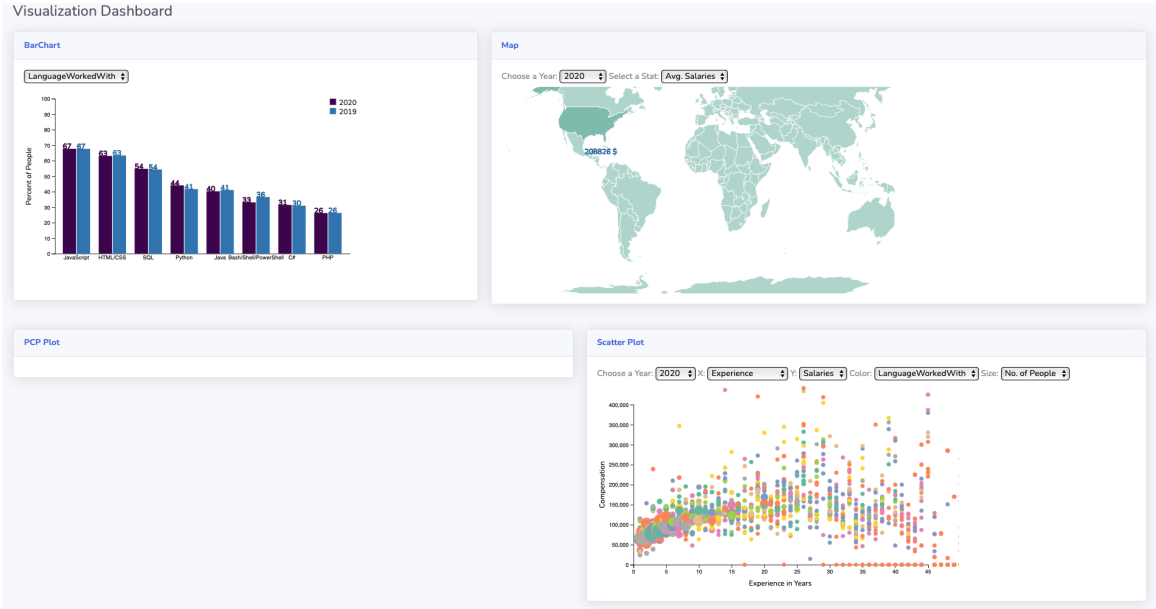

Figure 4: Above DashBoard shows the BarChart, Map & ScatterPlot as of now

3 Tasks to be completed

- We have to yet implement PCP plot.
- We are Planning to add even more interesting stats for the World Map.
- We are also Planning to include 2017 & 2018 data to the entire visualization dashboard.
- We are also planning to remove outliers from the data, as we can see clear outliers in the scatterplot. After removing we can analyze the plot even better.
- We are also planning to apply Machine learning techniques on interesting features from past years data and would like to predict the stats for 2021.
- We are also planning to use some colors on the map to visually represent the desired attribute strength.