Discovering Trends From Stack Overflow

Elaine Kim, Changran Hu, Yutong Zhou

Project goals

Stack Overflow is the largest online community for developers to share their knowledge and learn. Programmers can ask questions freely, and the community members can answer them if they can help. As Stack Overflow is a place where many developers have been actively communicating for the previous years, we were confident that we could gain some insights from the data available from Stack Overflow.

We wanted to visualize and inform professional developers and people who are interested to start learning programming languages with our work by keeping them updated with some latest trends that can be observed from Stack Overflow.

Discussion of Related Work

1	Stack Overflow Developer Survey 2015	This survey shows the overall profile of the developers who are using Stack Overflow for the year of 2015.
2	Stack Overflow Developer Survey 2016	This survey shows the overall profile of the developers who are using Stack Overflow for the year of 2016.
3	Stack Overflow Developer Survey 2017	This survey shows the overall profile of the developers who are using Stack Overflow for the year of 2017.
4	Stack Overflow Developer Survey 2018	This survey shows the overall profile of the developers who are using Stack Overflow for the year of 2018.
5	Stack Overflow Developer Survey 2019	This survey shows the overall profile of the developers who are using Stack Overflow for the year of 2019.

6	Stack Overflow Developer Survey 2020	This survey shows the overall profile of the developers who are using Stack Overflow for the year of 2020.
7	Data dump from Stack Exchange	This dump contains all anonymized Stack Overflow data. We use the 2020 data dump for our project.
8	Exploring Isotype Charts: "Our Private Lives"	We were inspired to lay out our gender isotype in multiple men and women figures after looking over this article.
9	Gender Equality Infographic	This infographic shows how there is a big gap in the average salary earnings for men and women. Using the isotype as shown in this work, we employed a similar way to present the demographic information in our work.

Description of Visualization

For the demographic trend, we used isotypes to visualize the gender representation noticeable in the software development field. By having different icons and colors for each gender, we wanted to clearly communicate that the industry is male-dominated, but has been some increase with women developers across the years.

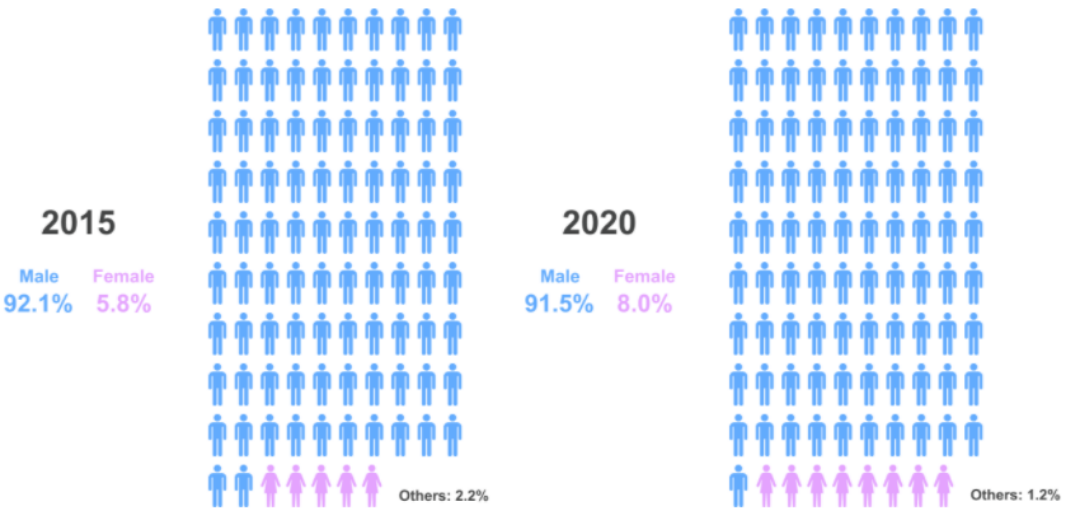
Trends Discovered From Stack Overflow

Introduction

Stack Overflow is the largest online community for developers to share their knowledge and learn. Programmers can ask questions freely, and the community members can answer them if they can help. As Stack Overflow is a place where many developers have been actively communicating for the previous years, we were confident that we could gain some insights from the data available from Stack Overflow. We wanted to visualize and inform professional developers and people who are interested to start learning programming languages with our work by keeping them updated with some latest trends that can be observed from Stack Overflow.

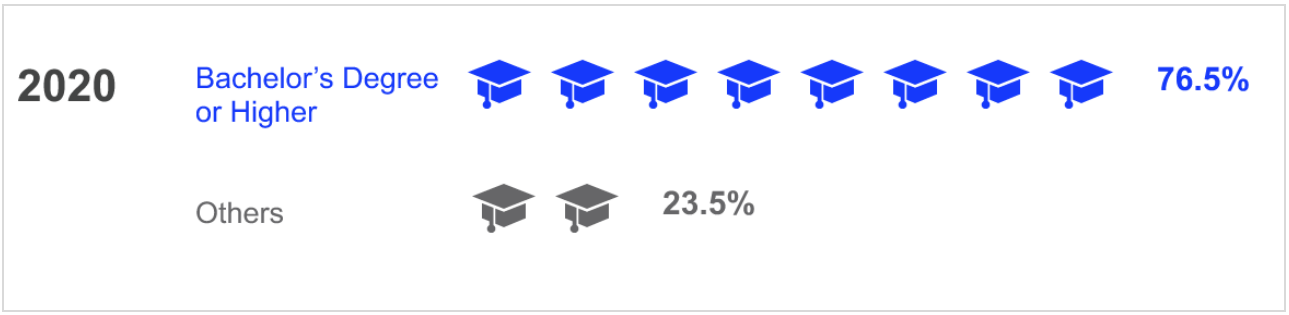
Demographic Trend

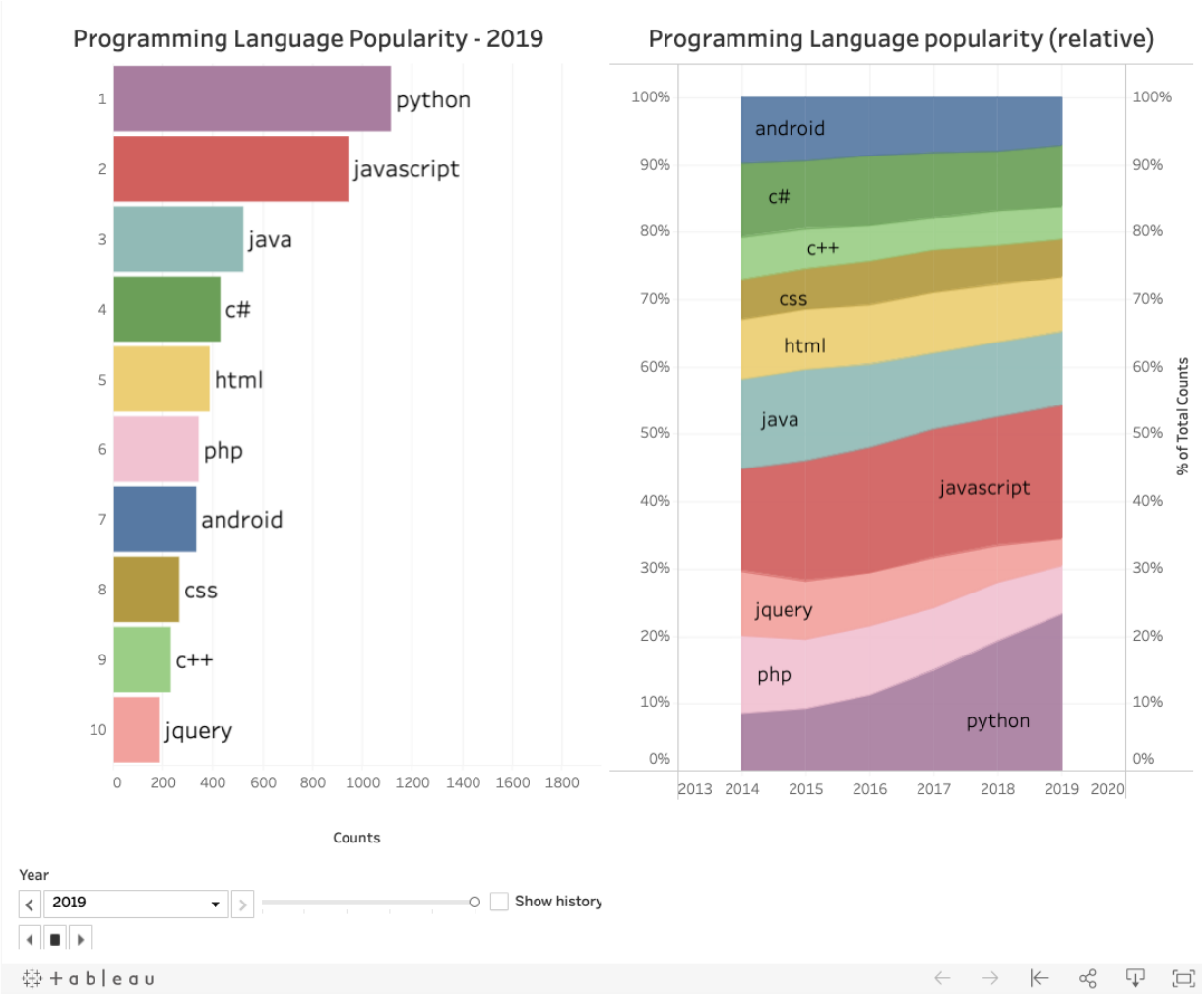
Based on a worldwide survey done by developers who use Stack Overflow over the years, more men are working in software development than women. Although there are still more work to be done to reach appropriate gender representation in the field, women developers have increased if we compare the two years as below.



From the survey, developers who completed at least the equivalent of a bachelor's degree or higher were approximately 76% of respondents in 2020, which has been consistent from previous years.

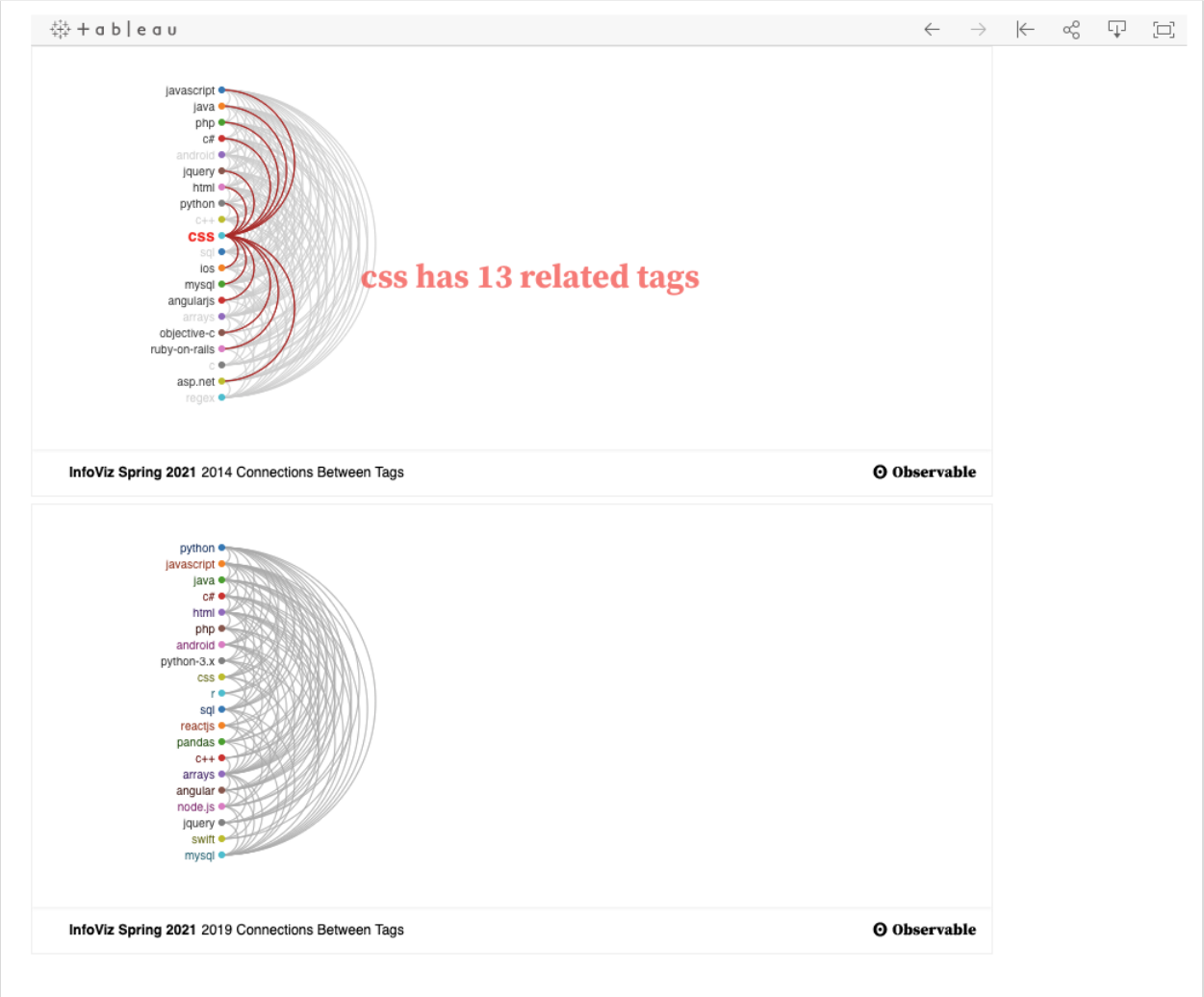
Similarly, we used an isotype to compare the percentage of developers who have at least the equivalent of a bachelor's degree or higher versus those who do not. We labeled the percentage for each group and used different colors to identify them. Only the most recent data (year 2020) was presented since the percentage difference stayed consistent from the previous years.



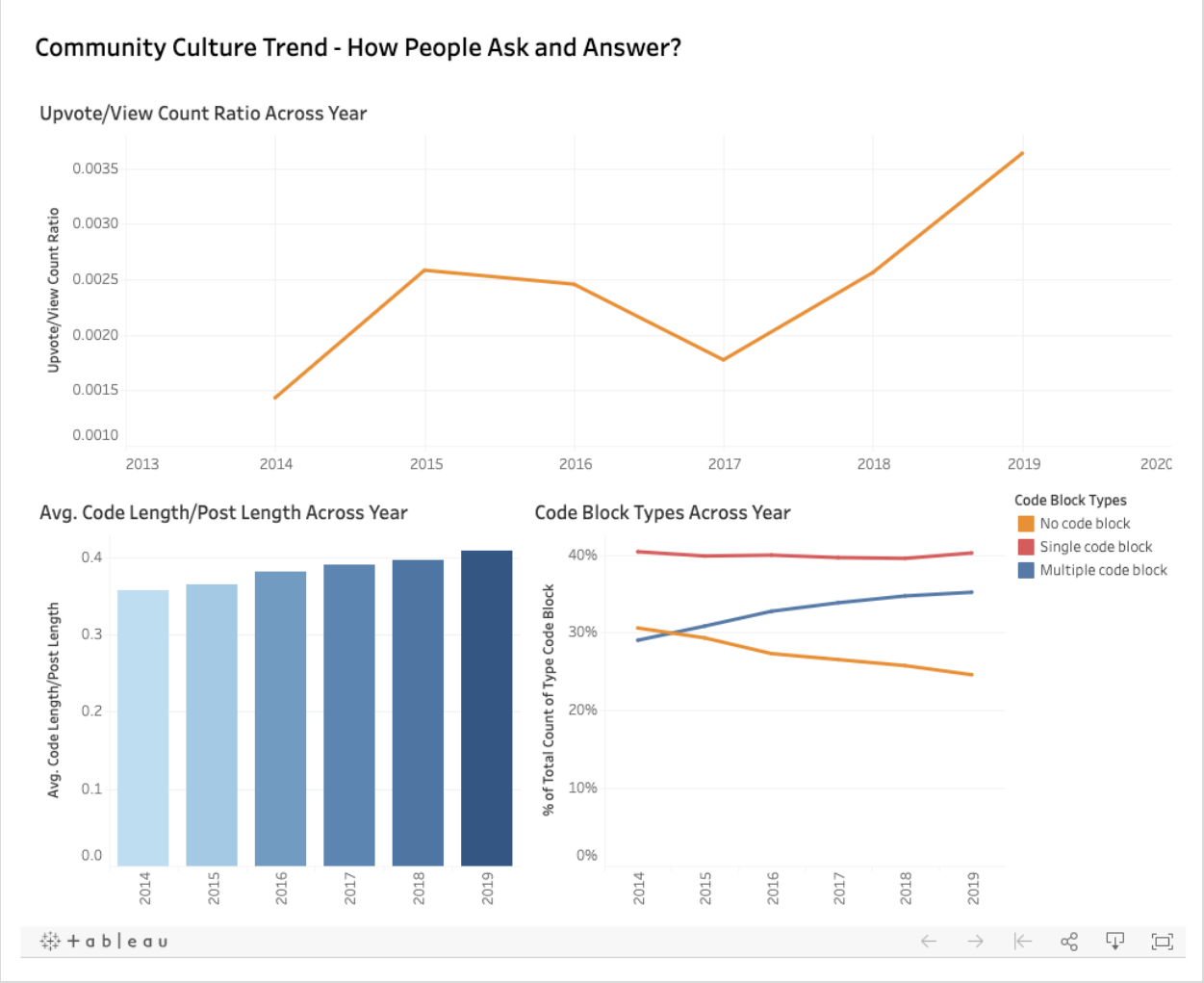


As for the programming trend, we used a bar chart to compare the changes to show the popularity changes in programming languages from 2014 to 2019. The bar chart is interactional, and the year can be easily selected using the drop down menu below the graph. Next to the bar chart, we added a stacked area chart to compare the relative changes that could be observed in the popularity of each programming language across the years. Using arc diagrams, we visualized the connections between different tags used in Stack Overflow. We labeled how many

related tags each tag has and enabled it to interactively work upon hovering over. By comparing the tag connections for 2014 and 2019, we wanted to show how each tag has either gained or lost its relative popularity between the two years.



Finally, we used a bar graph and two line graphs to visualize the changes in trend that could be observed in the Stack Overflow community culture. The upvote versus view count ratio across 2014 to 2019, and the code block types across years were presented in line graphs, and the average code length versus post length across years was made in a bar chart. We initially used a skinny bar graph to visualize the code block types across years, but changed into a line graph to improve its readability.



Data Used to Accomplish Goals

We took reference of the demographics and educational level information from Stack Overflow Developer Surveys published between 2015 and 2020. Other than that, the majority of the data we used for this project was taken from a data dump from Stack Exchange, preprocessed at our end to accomplish our goals accordingly.

Tools Used to Accomplish Goals

We used different tools to serve our goals appropriately. For instance, we made our isotype in the Demographic Trend with Figma, the bar graphs in the

Programming Trend and Community Culture Trend with Tableau, and the arc diagrams under the Programming Trend with D3. We also used a Github page to present our work publicly.

Results

From our usability test, we were able to identify how effective our visualization was in terms of successfully communicating the information that our audience should be able to learn from our work. Based on the feedback from our participants, we made some iterations for further improvements.

For each visualization, we first asked our participants to think aloud about what they see and how they feel immediately when they saw the visualization. By doing so, we managed to get more understanding of whether the overall readability of our visualizations worked and whether they were able to convey the message in a clear and concise way.

Next we asked our participants some questions that required them to read and understand data values and identify trends, relationships and extrema from the visualizations. For open ended questions, we individually rated the answer right or wrong based on our understanding.

We also used likert questions at the end of the session to evaluate how they liked the overall visualization and how much they thought they had learned.

We noticed that all of our participants were able to answer the questions right, although it took them some time to understand the context due to the lack of contextual description. The average rating for the overall impression and learning were around 5 out of 7, which showed some area of improvement for future iterations.

Demographic Trend	Average Accuracy
How much increase do you see in women developers?	100%
What can you tell from the comparison image of developers' education level?	100%

Programming Trend	Average Accuracy
What is the most popular language in 2017?	100%

Can you describe how the relative popularity changed for Python from 2014 to 2019?	100%
What are some tags that “Python” is connected to?	100%

Community Culture Trend	Average Accuracy
Can you explain any trend that you notice about the number of code block types from 2014-2019?	100%
Can you describe what you can see on the line graph on the bottom left?	100%
Can you describe what you can see on the bar graph on the bottom right?	100%

Impression and Learning	Average Rating
On a scale from 1 (not at all) to 7 (very much), how much did you enjoy reading this?	5
On a scale from 1 (not at all) to 7 (very much), how much did you feel like you learned?	5.33

From the feedback from our participants, we made the following changes to our final visualizations:

- Labeled the year next to the isotype to clearly indicate which year it is referring to for educational level visualization.
- For the programming trend, we used the top 10 languages instead of 20 to reduce cognitive overload.
- The arc diagrams show connections between tags were adjusted to use a fixed stroke width for all arcs, and we also reduced the number of languages being shown to increase readability.

We were unable to conduct a second round of usability testing due to the time constraints, but we hope our iteration can improve the overall experience for our audiences who are interested to learn more from our work.

Links

- Link to our prototype page used for usability testing
- Link to Observable
    - 2014 Connections Between Tags
    - 2019 Connections Between Tags

- Link to Programming Trend Visualization
- Link to Community Culture Trend Visualization
- Link to GitHub

Distribution of Work

	Elaine Kim	Changran Hu	Yutong Zhou
Data Preprocessing	10%	60%	30%
Demographic Trends	80%	10%	10%
Programming Trends	10%	60%	30%
Community Culture Trends	10%	10%	80%
Usability Testing	50%	25%	25%