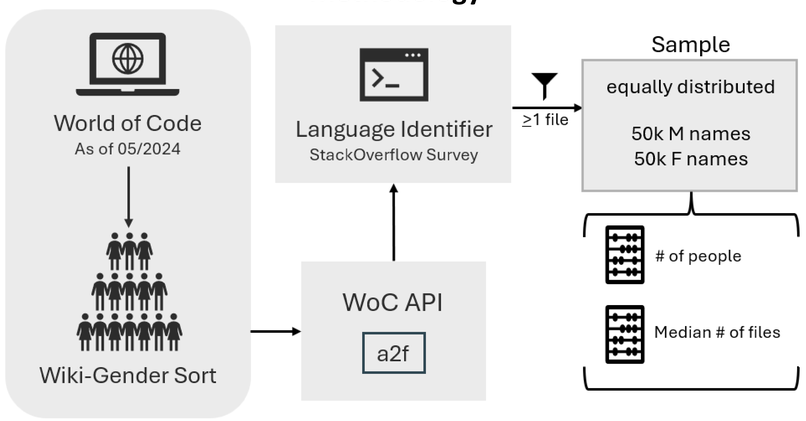
Impact of Gender on OSS File Contributions

Leilani Torres (The College of Wooster), **Heather Guarnera** (The College of Wooster),
Michael Collard (The University of Akron), **Amber Garcia** (The College of Wooster)

Abstract

We examine how gender impacts the use of specific programming languages, as analyzed across a stratified sample of 100k unique open software (OSS) developers from the World of Code (WoC) archive. We use Wiki-Gendersort to identify 50,000 masculine (M) and 50,000 feminine (F) developer names. The top fifteen programming languages according to the 2024 Stack Overflow Developer survey are considered. For each developer, we count the number of files that are edited in each programming language and compute the median across gender categories. Men and women tend to edit the same number of files among most programming languages, with the exception of developers using C#, C, Go, and Rust, which had more edits among men.

Methodology



Extensions	Language	F	M
js | js.map	JavaScript	8	9
html | htm	HTML	3	3
css | scss | less | sass	CSS	5	5
py | py3 | pyo |pyx | pyw | whl | pyd	Python	5	5
sql | sqllite | sqllite3 | mysql	SQL	2	2
ts | tsx	TypeScript	17	17
sh | zsh | bash | profile | bashrc | zshrc	Bash/Shell	2	2
java | iml | jar | dpj | xrb | aidl | mf | classpath	Java	9	9
cs | cspro	C#	17	19
c | h	C	6	7
php | twig	PHP	12	12
ps1	PowerShell	2	2
go	Go	4	5
rs | rlib | rst	Rust	4	5

Table 1: Programming languages with their file extensions and median number of files edited per gender category.

Programmers using TypeScript edited the most files (17) and those using PowerShell edited the fewest (2).

Research Questions

This study seeks to challenge the stereotype that men contribute more to OSS. We examine the top fifteen programming languages based on the 2024 Stack Overflow Developer Survey to identify any gender disparities within contributions according to programming language

Specific research questions are:
RQ1) How does gender influence the number of files edited within a given programming language?
RQ2) What are the characteristics of languages that share gender similarities?

Data Cleaning

- We began with 51.4 million author names from WoC.
- After data cleaning, there were 28.4 million authors. We removed:
  - Duplicate names - based on author alias and email address pairs - (15.8 million)
  - Blank names (0),
  - Names with over 100 characters (85),
  - Names containing more than 10% non-letter characters (7.1 million).
- Randomly select a stratified sample of 100,000 developers.
- We gathered file names through the WoC API call a2f then the filename extension was used to determine the programming language it was written in.

Results

- There are only a few gender differences in participation when using equal representation (i.e., using a stratified data set which consists of half men and half women, rather than a random sample of the OSS population which is heavily male-dominated).
- **RQ1**: Out of the top 15 programming languages, there are only minor gender differences in median file edits.
  - 1 file median difference for JavaScript, C, Go, & Rust
  - 2 file median difference for C#
- **RQ2**: The only languages with more women involved are web development languages HTML and CSS, which include 37.9% and 32.1% of all women, respectively; in comparison, there were 35.6% and 30.4% of all men who were HTML or CSS programmers, respectively

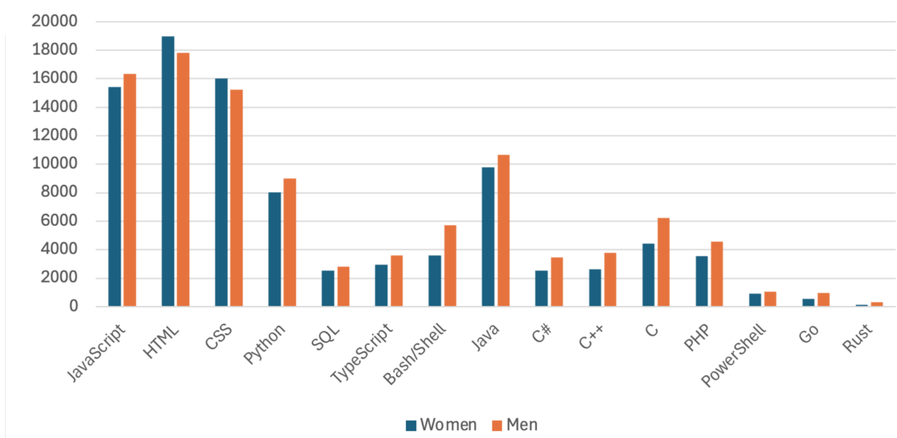


Figure 1: The number of individuals involved in the top 15 programming languages. HTML and CSS have more women involvement than men.

Limitations

- Exclusion of other gender identities.
- Usernames may not reflect one's gender identity.
- Wiki-Gendersort is highly accurate, though there are challenges with non-English alphabet names

Future Works

- Consider more languages, including JavaScript dialects.
- Evaluate the content of files (e.g., types of edits made).
- Compare beginner developers and those with more experience.

References

[1] N Berube, M Sainte-Marie, and V Lariviere. 2020. Wiki-Gendersort: Automated gender detection using first names in Wikipedia. *OSF* (2020).
[2] Y Ma, T Dey, C Bogart, S Amreen, M Valiev, A Tutko, D Kennard, R Zaretzki, and A Mockus. 2021. World of code: Enabling a research workflow for mining and analyzing the universe of open-source VCS data. *Empirical Software Engineering*.
[3] StackOverflow. 2024 Developer Survey.

Acknowledgments: This work was partially supported by the Sophomore Research Program at The College of Wooster. I would like to thank my advisor, Heather Guarnera, for their support in every way.