Comprehensive Analysis of User Activity (2020-2022)

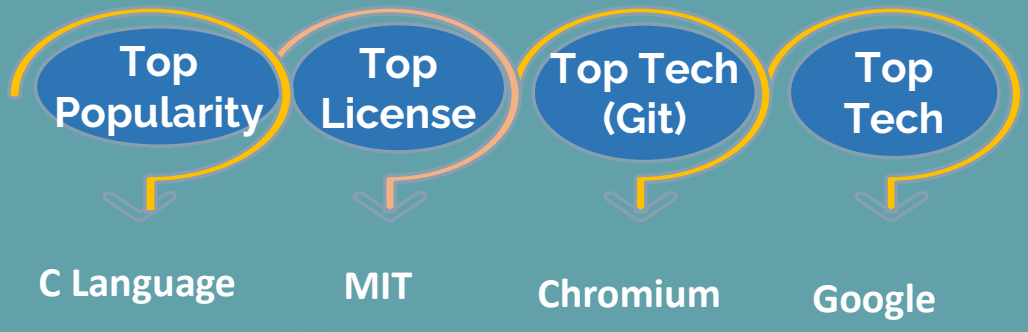
Yu-Chih (Wisdom) Chen

December 8, 2023

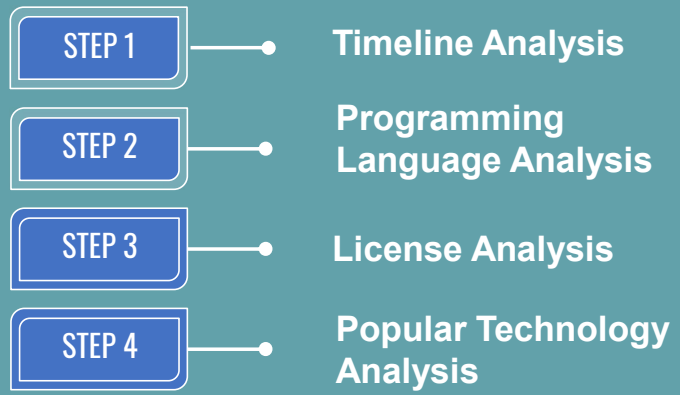
Executive Summary

- This project analyzed how people use GitHub, a popular platform for sharing and collaborating on code. We looked at when people were most active, which programming languages they used most, the types of licenses they preferred for their projects, the most popular technologies, and how often the Big Five tech companies (Microsoft, Alphabet, Amazon, Apple, Meta) were mentioned.

Key Takeaways (2020 – 2022):



Key Milestones:



Methodology

- The project also utilized Google Cloud Platform (GCP) for its robust computing capabilities. GCP provided the necessary infrastructure to handle the large volume of data from GitHub repositories and perform complex computations required for the analysis.
- The findings are based on a comprehensive examination of GitHub user activity, programming language usage, and technology trends between 2020 and 2022. The methodology ensured a rigorous and systematic approach to uncovering insights about the GitHub ecosystem.

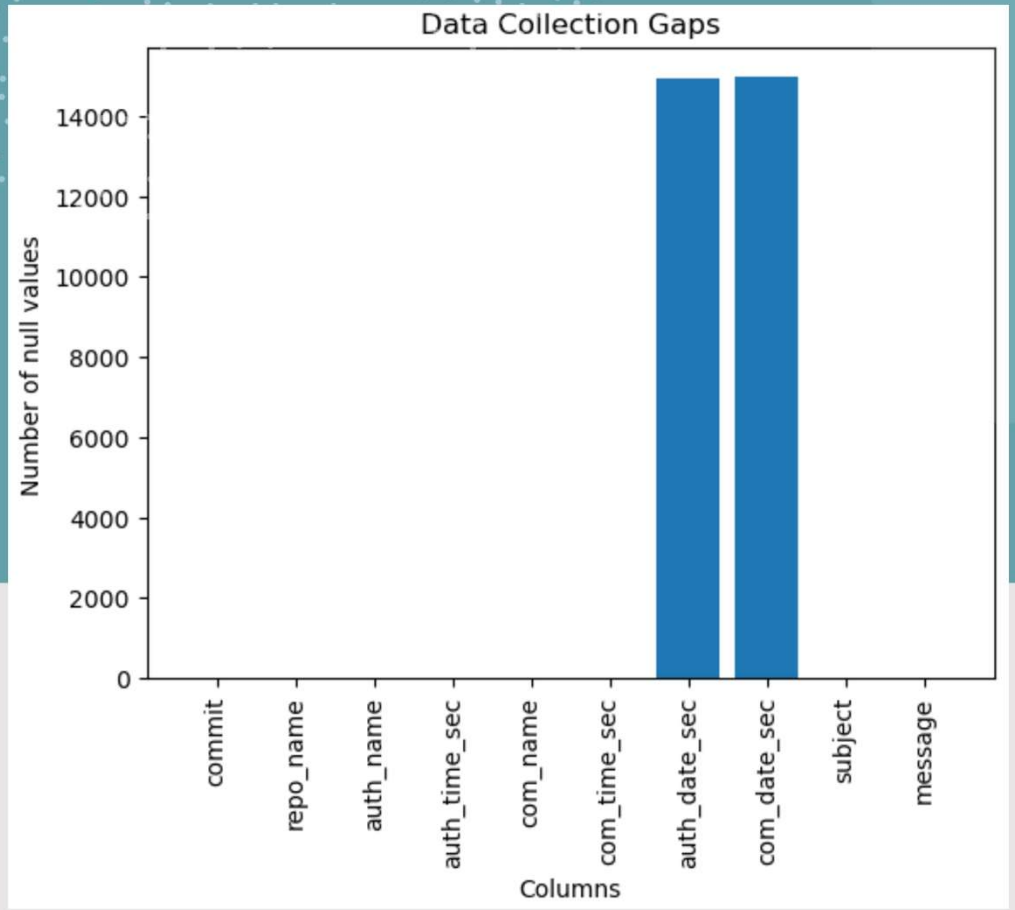
Source Data

- The source data for the project was gathered from a wide array of GitHub repositories. These repositories encompassed a diverse range of topics, including programming, technology, and open-source contributions. The data from these repositories provided a comprehensive view of GitHub user activity, programming language usage, and technology trends between 2020 and 2022. This wide-ranging data set served as the foundation for the project's analysis and findings, offering valuable insights into the GitHub ecosystem

01

Clean up & Filtering

Tweet Clean up & Filtering I

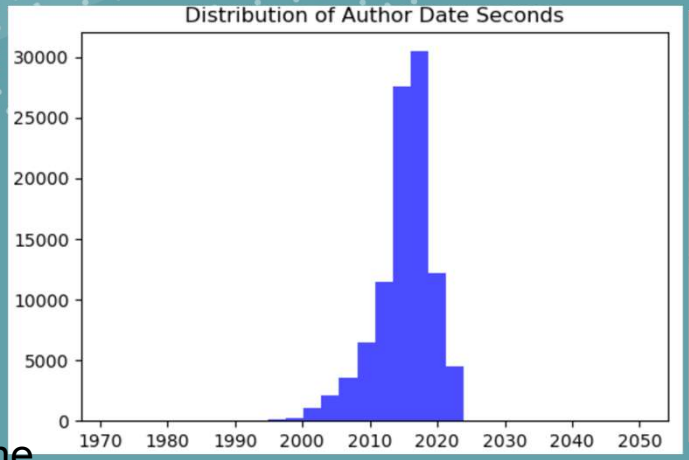


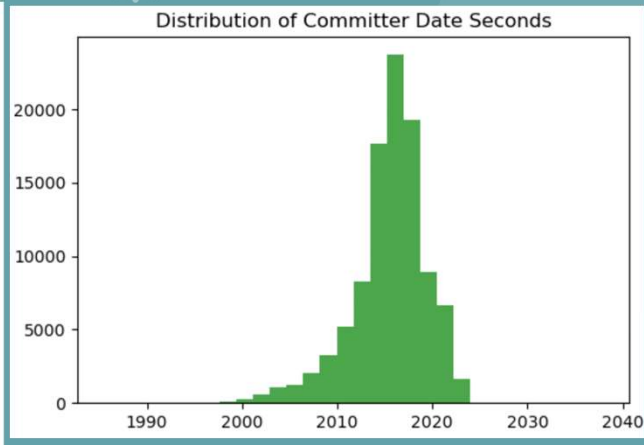
- Two columns with missing values were identified: **author date seconds** and **committer date seconds** (< **1%**).
- These columns represented the same timeline as **author time seconds** and **committer time seconds**.
- To maintain data quality and reliability, the columns with missing values were removed.

```
Percentage of missing values in auth_date_sec: 0.00994%
Percentage of missing values in com_date_sec: 0.00997%
```

- The decision to use complete data and disregard incomplete data simplifies the dataset and avoids potential issues.

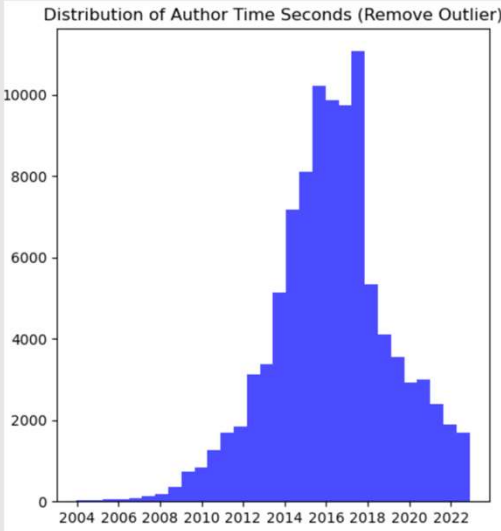
Tweet Clean up & Filtering II

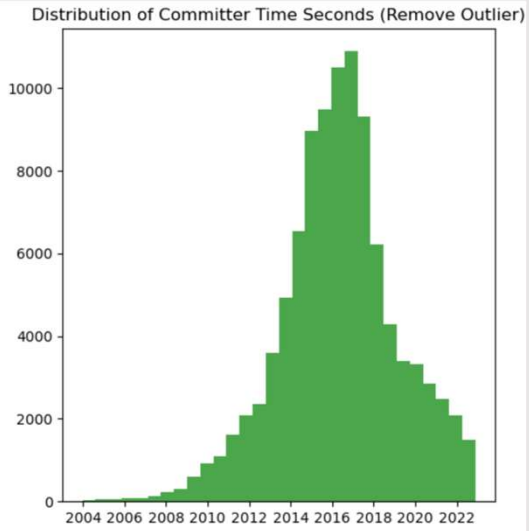




Before cleaning the data: We noticed some **inconsistencies** in our data. For instance, some data points were recorded from years beyond 2023, which doesn't make sense. Additionally, we found data from before 1990, which we considered too **old** to be relevant for our current analysis.

After cleaning the data: The data now appears more consistent and **well-distributed**. The timeline of the data is also more appropriate, focusing on more recent years.



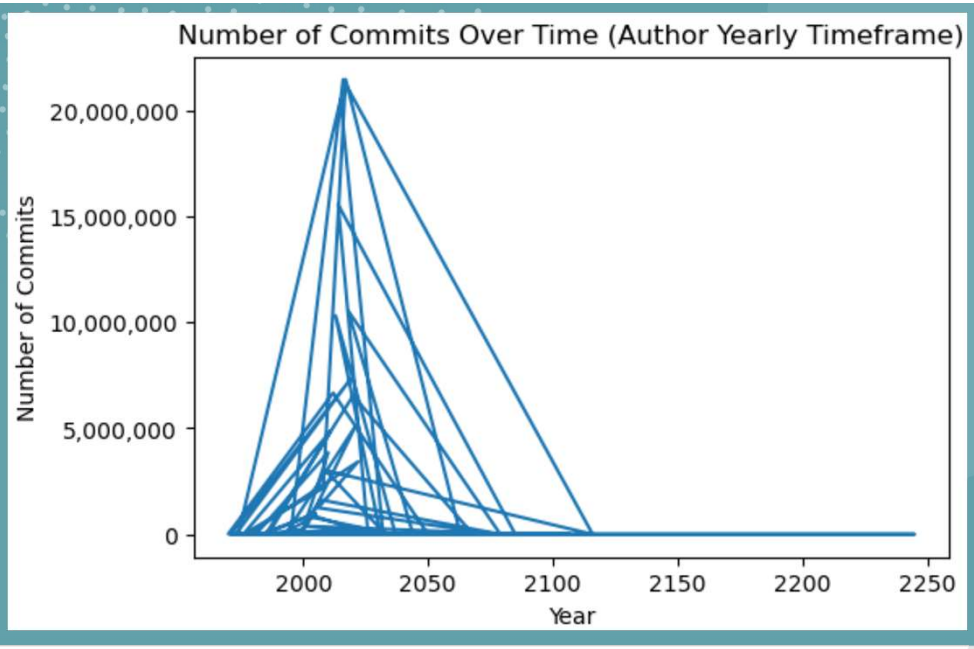


02

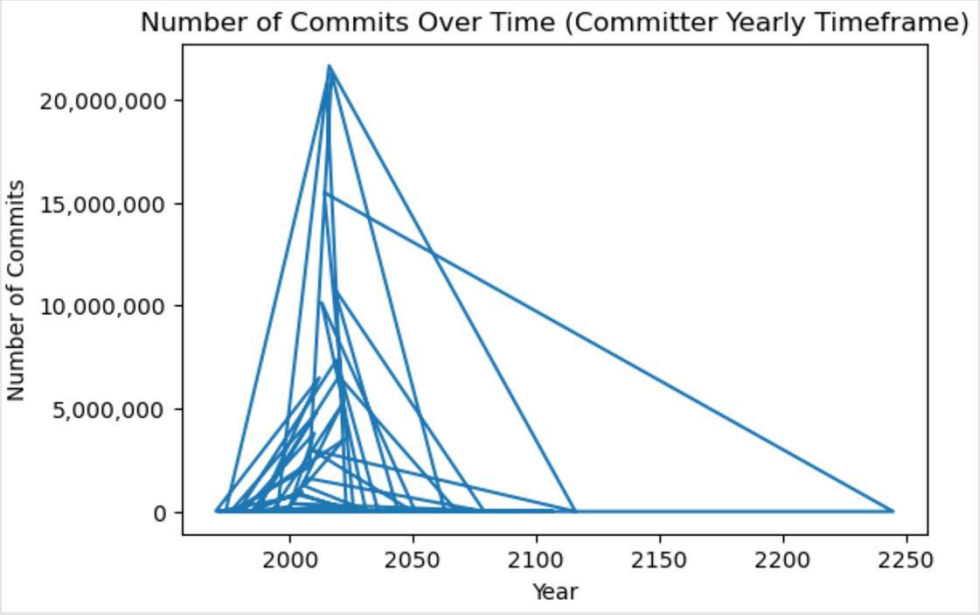
Timeline Analysis & EDA

Timeline Analysis

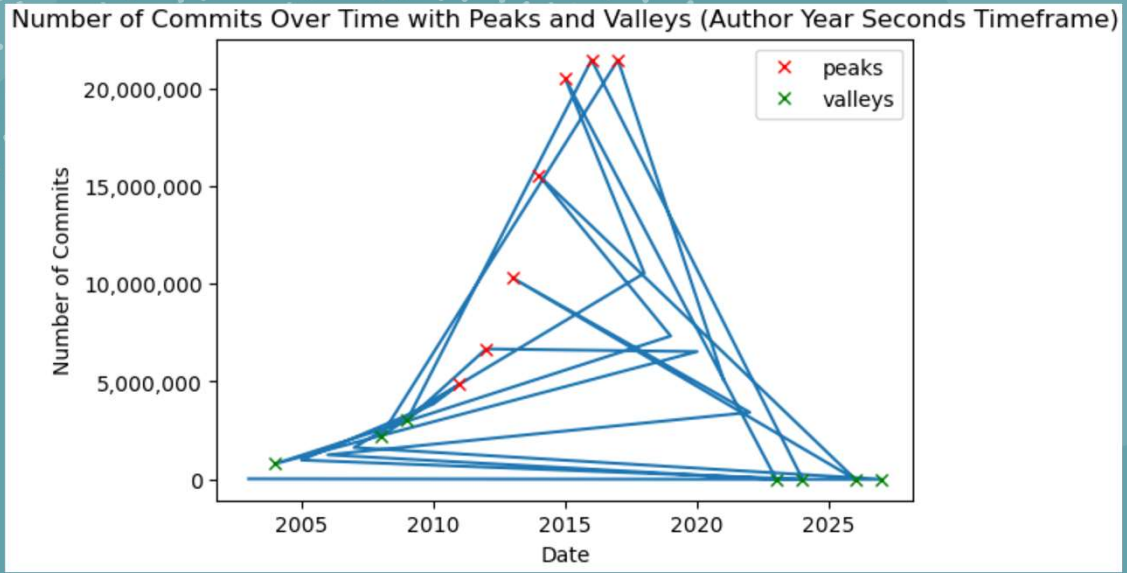
```
The Minimum Author Time:   1970-01-01

The Maximum Author Time:   2245-02-21

The Minimum Committer Time:   1970-01-01

The Maximum Committer Time:   2245-02-21
```



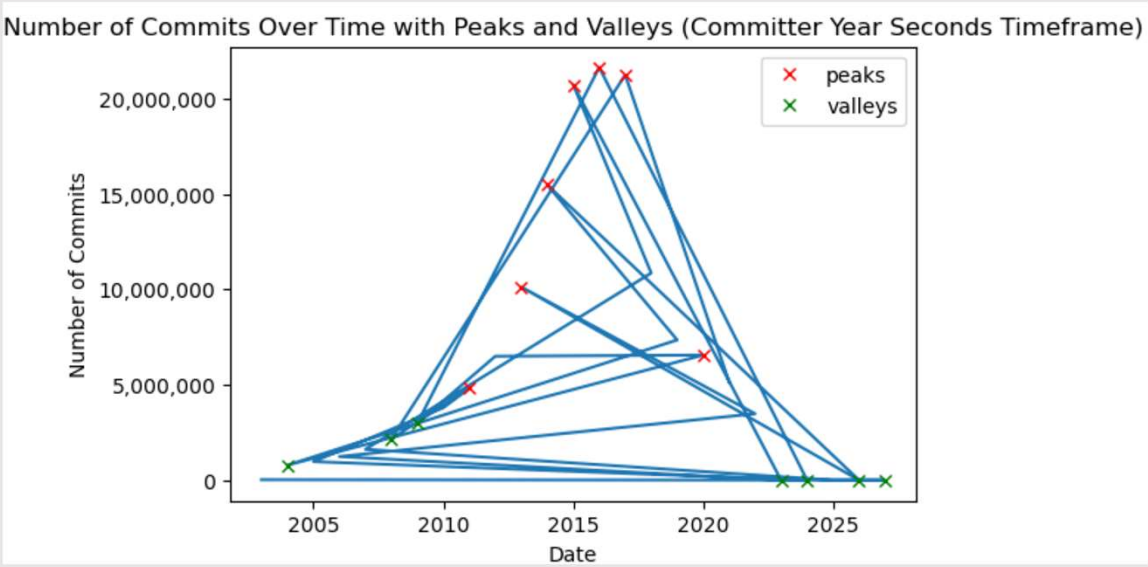
- Timeline covers the period: **01/01/1970 – 02/01/2245**
- We've identified some **anomalies** in our data, with dates extending **beyond the current year (2023)**, which we suspect to be errors. To maintain the integrity of our analysis, we're taking steps to address these outliers
- From our preliminary analysis is that the bulk of the data points, referred to as "commits", started to significantly increase from the beginning of the **21st century**



Timeline Analysis



- Observed **variations** in the number of commits made by authors and committers over time, with distinct periods of **increased** and **decreased** activity
- In **2020**, we noticed a **significant increase** in the frequency of commits made by committers
- The number of commits peaked in 2020 for committers but not for authors, It could be that more **people were contributing to the project in 2020**, or that the project required **more frequent updates**.

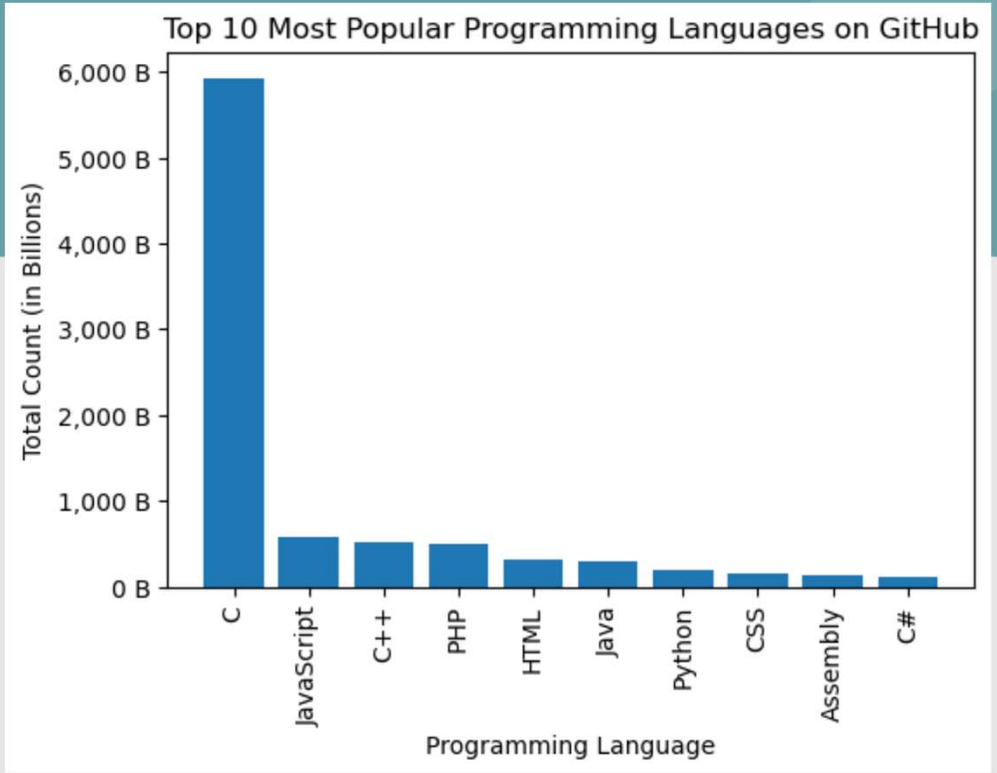


03

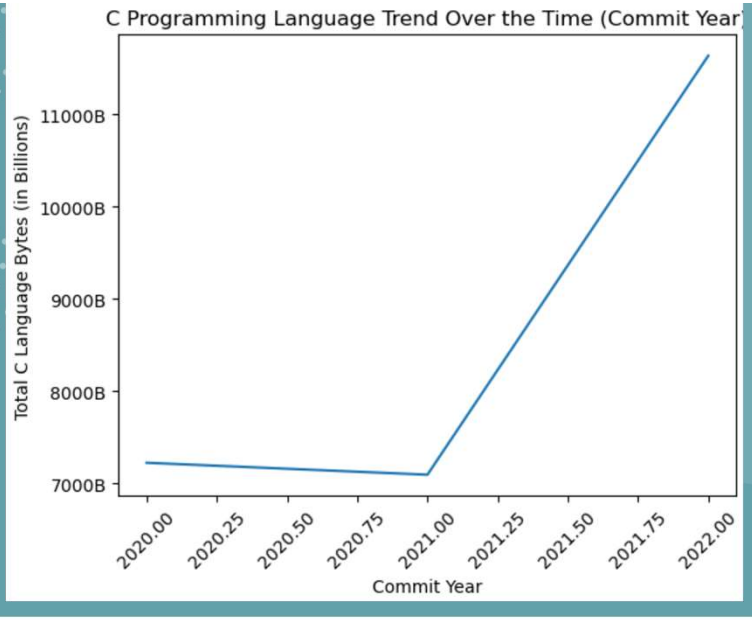
Programming Language, License Analysis & EDA

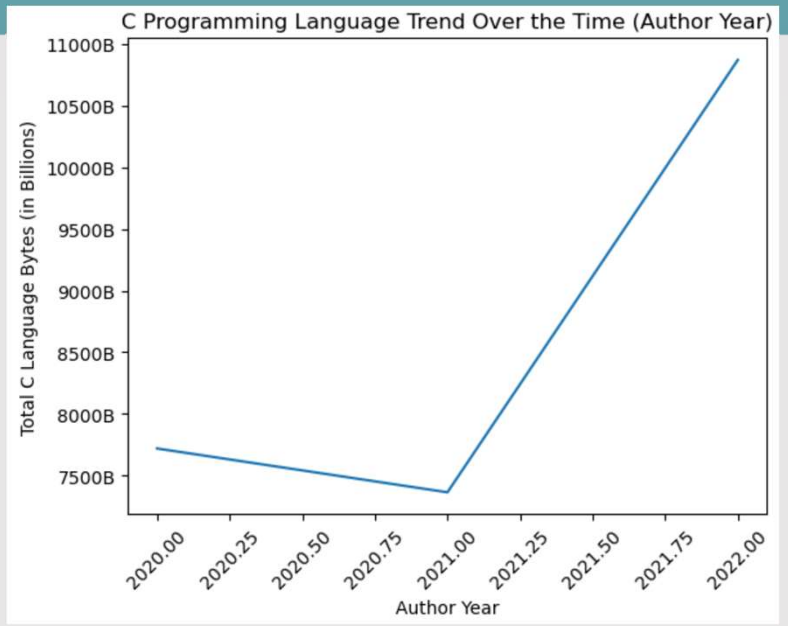
Programming Language

- C, JavaScript, C++, PHP and HTML are the **top 5** popular of programming Languages



C Programming Language

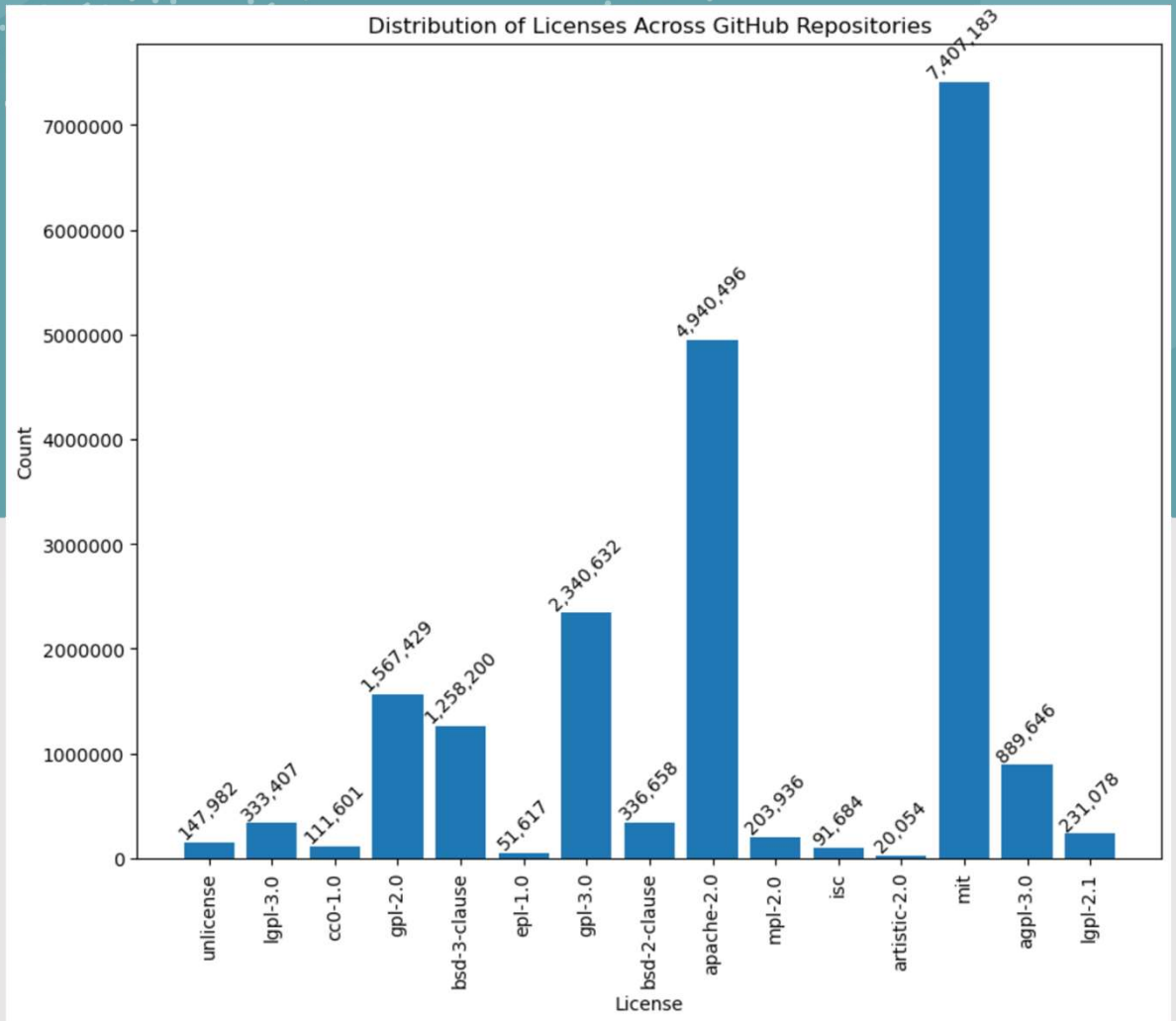




- In 2021, we noticed a **significant increase** in the use of the C programming language by **both** authors (those who originally wrote the code) and committers (those who last applied the changes to the project)
- In 2022, the C programming language became **more popular** among committers compared to authors. This means that more people were applying **changes to the code written in C than those who were originally writing it**

License Analysis I

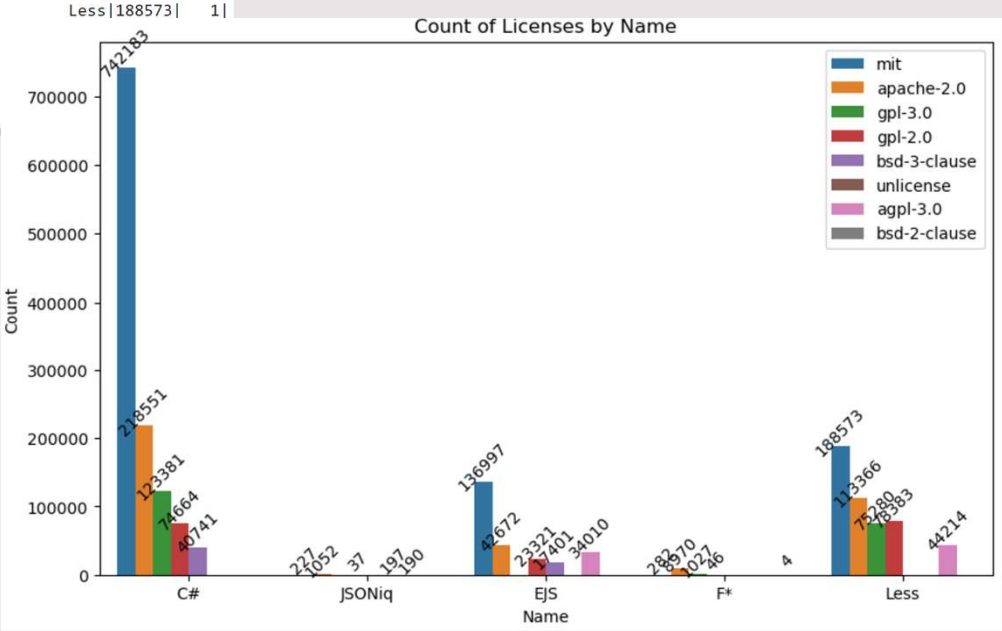
- Mit, Apache, Gpl-3.0, Gpl-2.0 and bsd-3-clause are the **top 5** frequency License across GitHub Repositories between 2020 and 2022



License Analysis II

- **Top 5** of programming languages associated with the license.
- **No specific programming language** is more likely to be **one** associated with the MIT license.
- In the data visualization, the **MIT** license is the most common license across different programming languages.
- The MIT license's popularity is due to its **flexibility**, allowing developers to freely use, modify, and distribute the code, making it a common choice across various programming languages

license	language_name	count	rank
mit	C#	742183	1
apache-2.0	C#	218551	2
gpl-3.0	C#	123381	3
gpl-2.0	C#	74664	4
bsd-3-clause	C#	40741	5
gpl-2.0	Cairo	290	1
mit	EJS	136997	1
apache-2.0	EJS	42672	2
agpl-3.0	EJS	34010	3
gpl-2.0	EJS	23321	4
bsd-3-clause	EJS	17401	5
apache-2.0	F*	8970	1
gpl-3.0	F*	1027	2
mit	F*	282	3
gpl-2.0	F*	46	4
bsd-2-clause	F*	4	5
gpl-2.0	Gleam	290	1
mit	Gleam	242	2
gpl-2.0	Grace	32	1
gpl-3.0	Grace	9	2
apache-2.0	JSONiq	1052	1
mit	JSONiq	227	2
unlicense	JSONiq	197	3
agpl-3.0	JSONiq	190	4
gpl-2.0	JSONiq	37	5
mit	Less	188573	1
apache-2.0			
gpl-2.0			
gpl-3.0			
agpl-3.0			



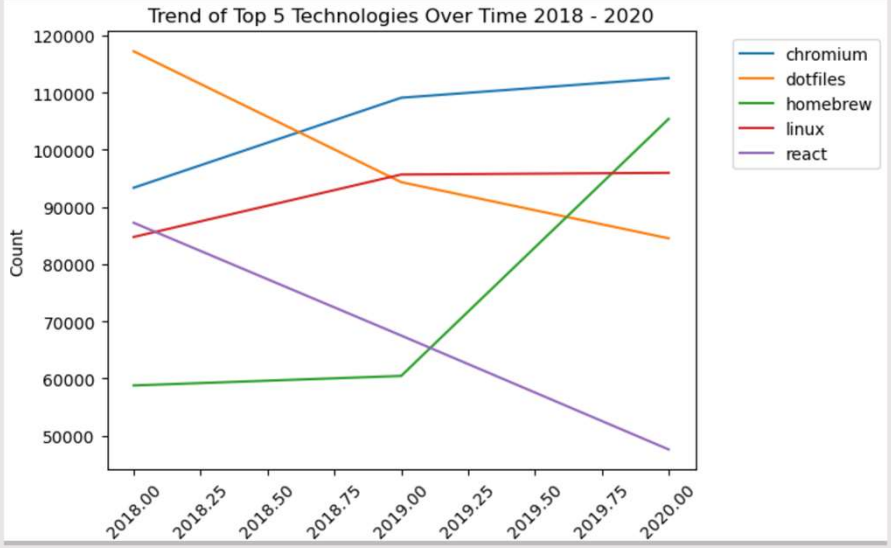
04

Popular technology, Repositories Analysis & EDA

Popular Technology 2018 - 2020

technology	count
chromium	315043
dotfiles	296110
linux	276430
homebrew	224655
react	202359

- Top 5 most common technologies used (**Chromium, Dotfiles, Linux, Homebrew, React**), with Chromium being the most popular
- **Chromium's** popularity surged in the **second quarter of 2018**, establishing its dominance.
- **Dotfiles** and **React** experienced a notable decline in usage at the beginning of 2018.
- **Homebrew** saw a **significant increase** in usage compared to Linux and Dotfiles starting in the **first quarter of 2019**
- Reasons behind the **decrease** in **Dotfiles** and **React** usage at the beginning of **2018** and the **increase** in **Homebrew** usage over Linux and Dotfiles (It could be due to changes in technology preferences)

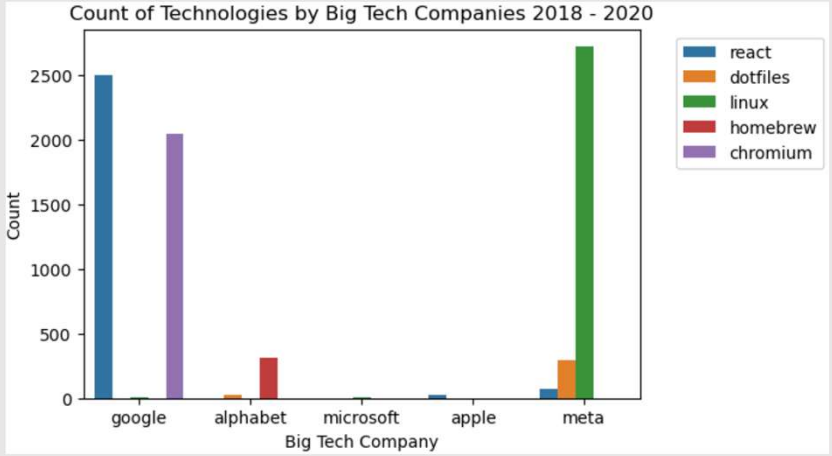


Big Tech Analysis I 2018 - 2020

big_tech	technology	count
google	react	2502
alphabet	dotfiles	29
apple	react	29
microsoft	linux	3
google	linux	3
meta	homebrew	2
meta	linux	2721
meta	react	70
alphabet	homebrew	313
meta	dotfiles	297
google	chromium	2048

- We identified the **top 5** technologies used between 2018 and 2020: **Chromium, Dotfiles, Linux, Homebrew, React**
- **Linux** is connected to major companies like **Microsoft, Google, and Meta**.
- **Meta** has a **stronger connection** to various technologies, such as **Linux, Dotfiles**, and **React**, compared to other Big Tech companies like Apple and Microsoft.

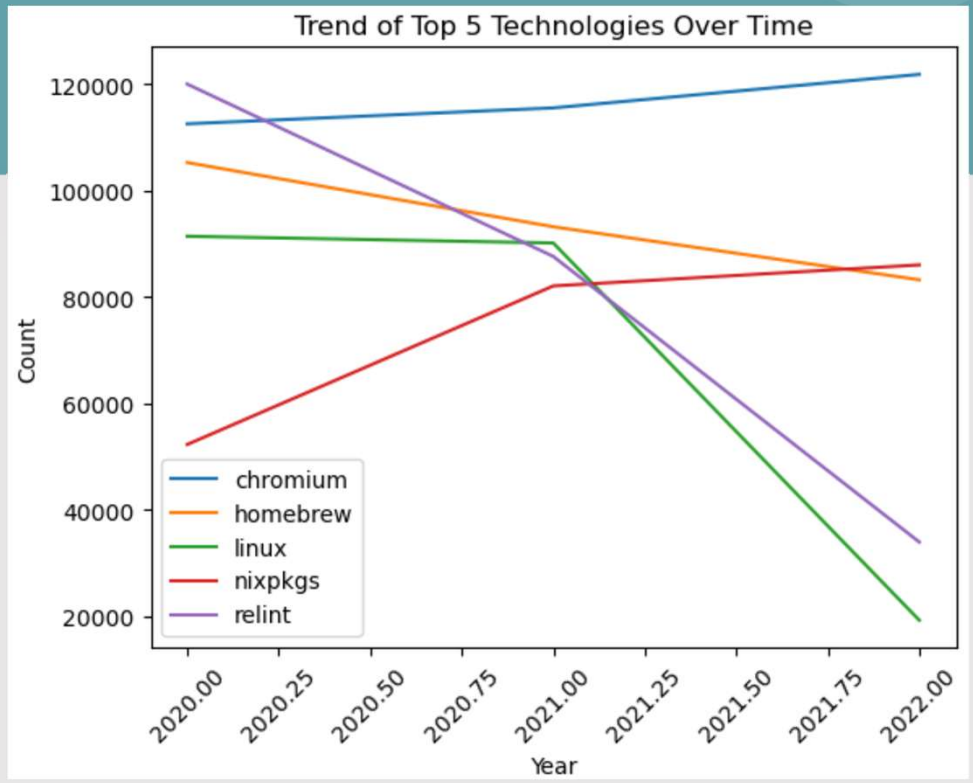
1) In **2020**, Meta added over **13,000 employees**, a **30% increase**, marking the biggest year of hiring in the company's history
2) This growth was fueled by the **increased importance of internet applications** during the widespread COVID-19 lockdowns, which supercharged business for many tech companies



Popular Technology 2020 - 2022

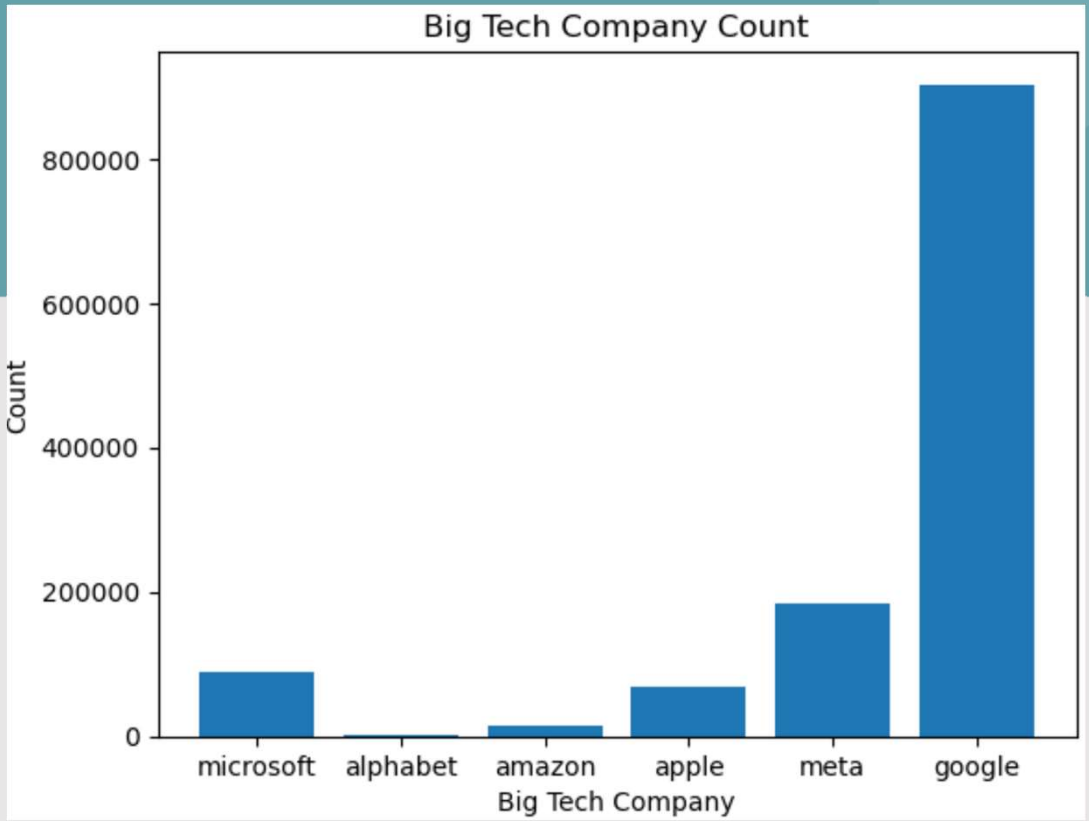
technology	count
chromium	349925
homebrew	281725
relint	241616
nixpkgs	220430
linux	200867

- Top 5 most frequent technologies (**Chromium, Homebrew, Relint, Nixpkgs, Linux**), **Chromium** is the most popular technology
- **Linux** and **Relint** saw a **significant decrease** towards the end of 2020
- Chromium's popularity can be attributed to its dominance in the web browser market, with a **market share of around 62.92%** worldwide in 2023 →This popularity could have led to increased usage and contributions on Github
- The **decrease** in frequency count for **Linux** and **Relint** could be due to various factors, such as changes in user preferences, the emergence of competing technologies



Big Tech Analysis I
2020 - 2022

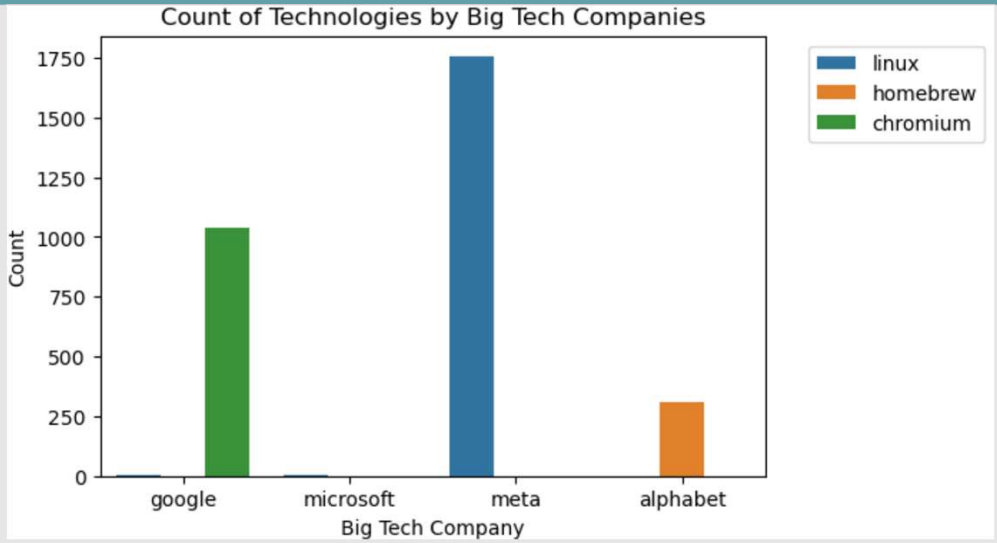
- Big Five tech companies: **Microsoft, Alphabet (Google), Amazon, Apple, Meta**
- "**Google**" is the most frequently mentioned Big Five tech companies in GitHub repositories
- **Google** became a subsidiary of Alphabet in **2015**. (That's the reason there is no "Alphabet" between 2020 and 2022)
- **Google** is more commonly mentioned because it is the most **well-known** and **widely** used subsidiary of Alphabet.



Big Tech Analysis II 2020 - 2022

- We identified the **top 5** technologies used between 2020 and 2022: **Chromium, Homebrew, Relint, Nixpkgs, and Linux**
- **Linux** is connected to major companies like **Microsoft, Google, and Meta**.
- **Homebrew** → Alphabet, which is the parent company of Google.
- **Chromium** is associated with Google
- Many technologies are connected to **Google**, which is why Google is the most frequently mentioned Big Five tech company in GitHub repositories.
- Chromium, Homebrew, and Linux are all open source.
    1) **Chromium** is an open-source browser → **Google**
    2) **Homebrew** is a free and open-source software package → **macOS & Linux**
    3) **Linux** is the best-known and most-used open source operating system

big_tech	technology	count
microsoft	linux	2
google	linux	6
meta	linux	1753
alphabet	homebrew	310
google	chromium	1038



Big Tech Analysis III 2020 - 2022

technology	data science
pythonvscode	11

technology	machine learning
azure	22

technology	artificial intelligence
translateproject	2

technology	neural network
fast	15

technology	deep learning
hanhandatascience...	12

- A significant number of Data Science and AI projects leverage Microsoft Azure technology.
  1) Comprehensive Analytics Tools → It provides a comprehensive portfolio of algorithms and analytics resources, making it a versatile platform for various machine learning problems
  2) Scalability and Flexibility → It offers the ability to scale resources up or down as needed, allowing organizations to pay only for the resources they use
  3) Democratizing AI → It provides user-friendly interfaces and tools, making AI accessible to both seasoned data scientists and non-technical roles

- **Data Science & AI Projects**: Data Science, Machine Learning, Artificial Intelligence, Neural Network, Deep Learning
  1) Data Science: Python VSCode
  2) Machine Learning: Azure
  3) Artificial Intelligence: Translate Project
  4) Neural Network: Fast
  5) Deep Learning: Hanhan data Science

Technological Breakthroughs Driving Adoption 2018 – 2022

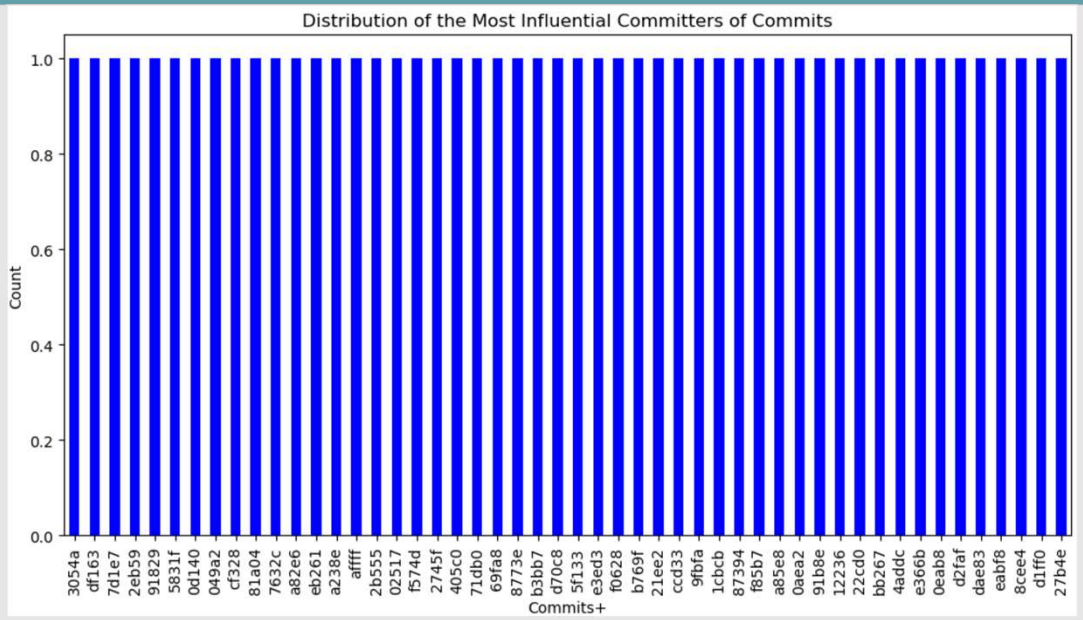
1) **Open-Source Movement** → The open-source movement has been a significant driver of technology adoption
   - **Chromium, Homebrew, and Linux** are all open-source, which means they are freely available for developers to use, modify, and distribute
   - This has led to a surge in their adoption and contribution on platforms like Github
2) **Internet Dominance** → Dominance of internet applications, particularly during the COVID-19 pandemic, has fueled the growth of many tech companies and the adoption of their technologies
   - **Chromium's** popularity can be attributed to its dominance in the web browser market
3) **Company Influence** → Influence of major tech companies like Google, Microsoft, and Meta has also driven the adoption of certain technologies
   - **Linux** is connected to these major companies, and Google is associated with Chromium

Commit Analysis 2020 - 2022

```
The total volume of commits for the committer with the most commits is 6795066.
```

com_name	count	commit
GitHub	6795066	dda77
GitHub	6795066	3be39
GitHub	6795066	fb021
GitHub	6795066	27339
GitHub	6795066	e294e
GitHub	6795066	d88af
GitHub	6795066	521de
GitHub	6795066	076af
GitHub	6795066	66e8e
GitHub	6795066	db6b1

- Top Contributors on GitHub
- Highest number of commits by a single user: 6,795,066
- Most influential committers on GitHub are individual users or organizations who contribute to various projects. For example, according to a list of most active GitHub users, visionmedia (TJ Holowaychuk) had the highest number of contributions during a specific period



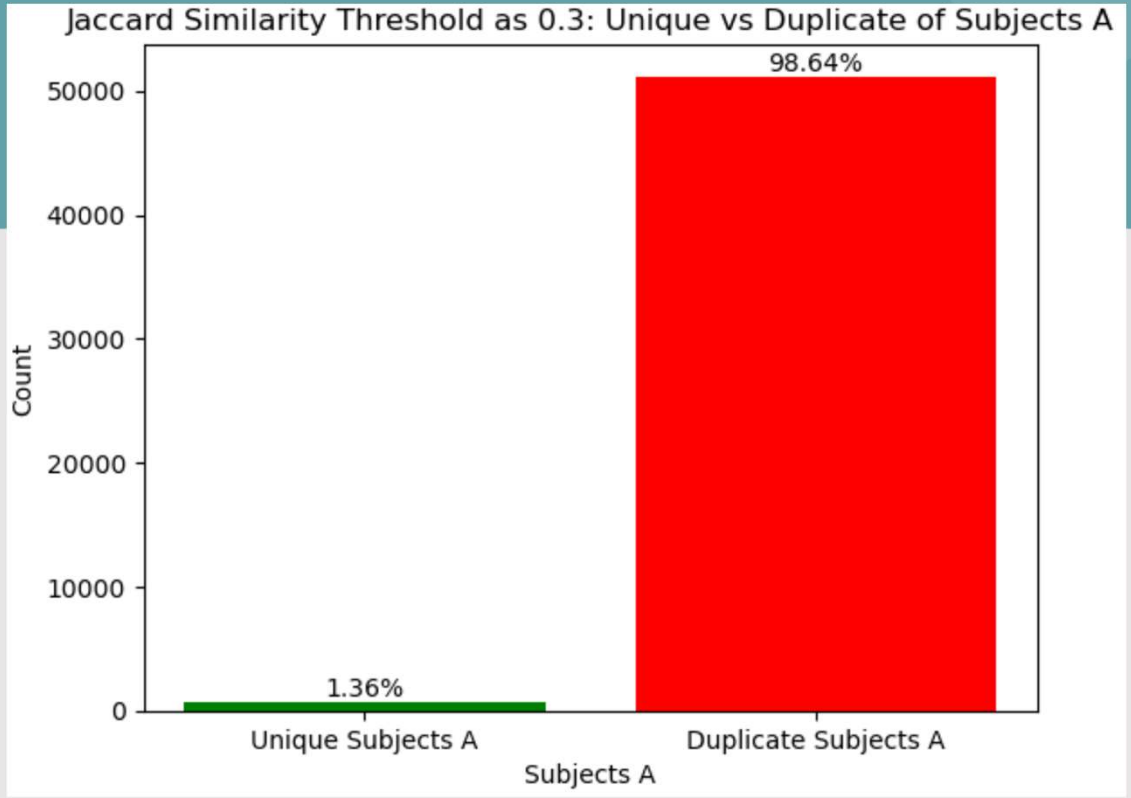
05

Subject & Message Analysis

Subject Analysis 2020 - 2022

```
Number of duplicate subject A with Jaccard Similarity Threshold = 0.3:  51169

Number of unique subject A with Jaccard Similarity Threshold = 0.3:  703
```

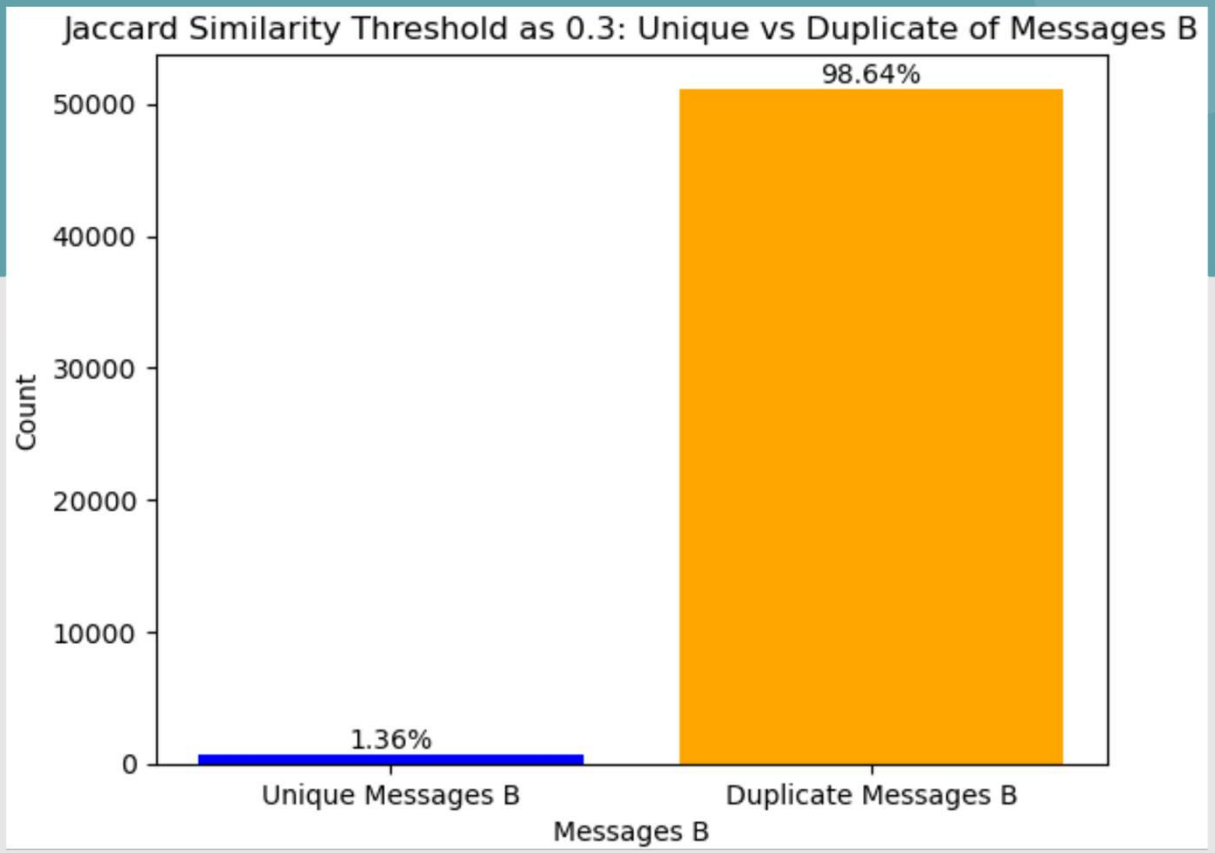
- Similarity measure used: 30% shared content
- In the Subject content, 98.64% are repeated, while only 1.36% are unique
- The high percentage of duplicates in GitHub subject contents could be due to several factors. One possible reason is that there is a lot of duplicate content on GitHub, as many users might reuse code or text from other repositories or projects



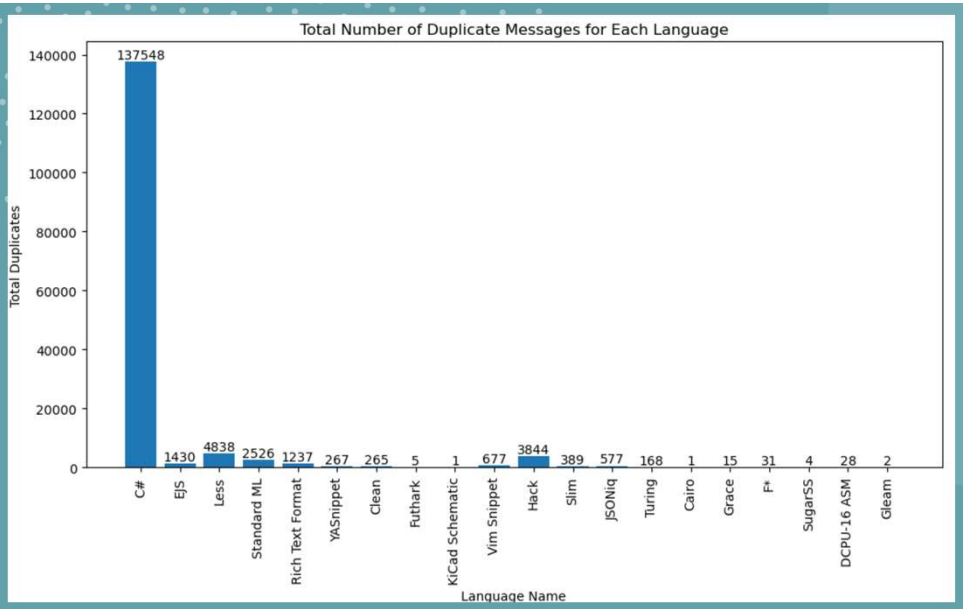
Message Analysis 2020 - 2022

```
Number of duplicate message B with Jaccard Similarity Threshold = 0.3:  51167

Number of unique message B with Jaccard Similarity Threshold = 0.3:  705
```

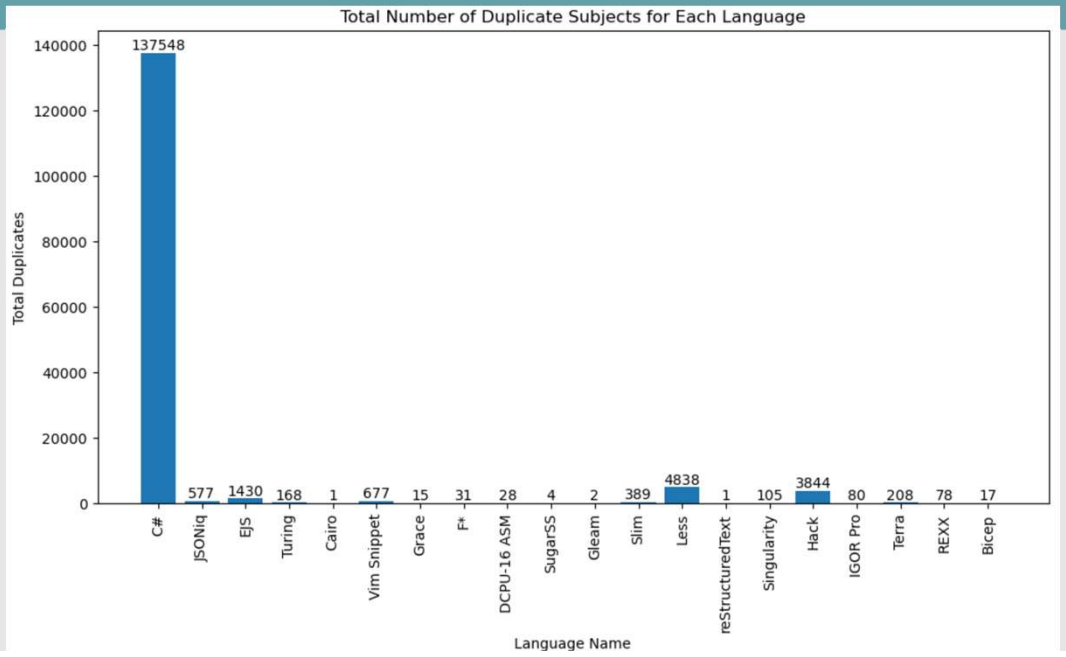
- Similarity measure used: 30% shared content
- In the Message content, 98.64% are repeated, while only 1.36% are unique
- The high percentage of duplicates in GitHub message contents could be due to several factors. One possible reason is that there is a lot of duplicate content on GitHub, as many users might reuse code or text from other repositories or projects



Subject & Message of Language Analysis I 2020 - 2022

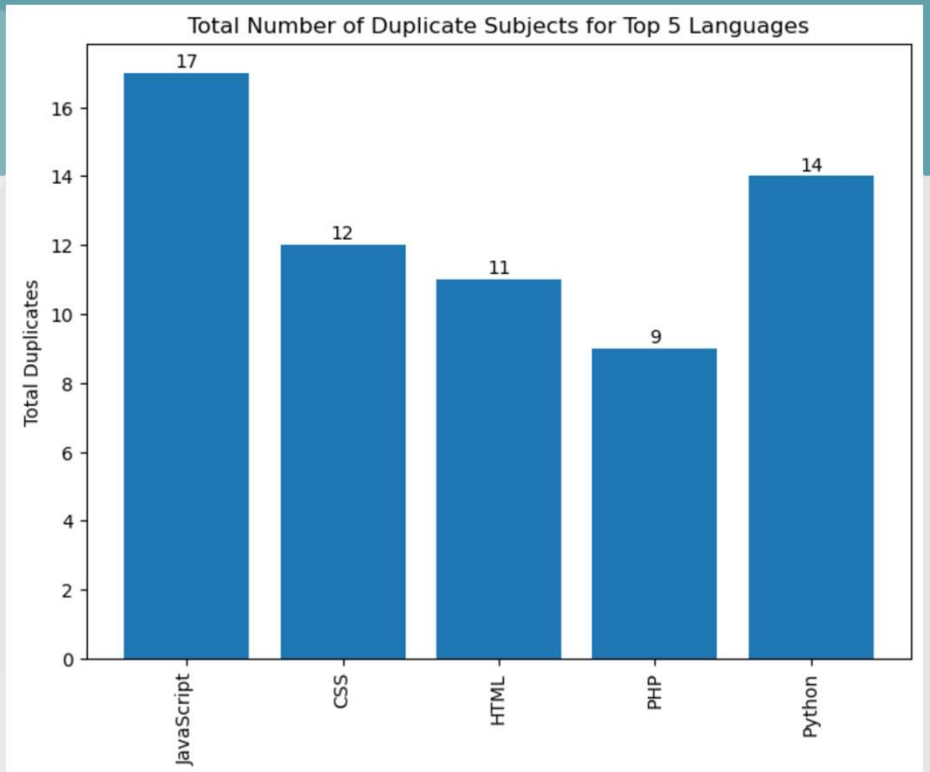


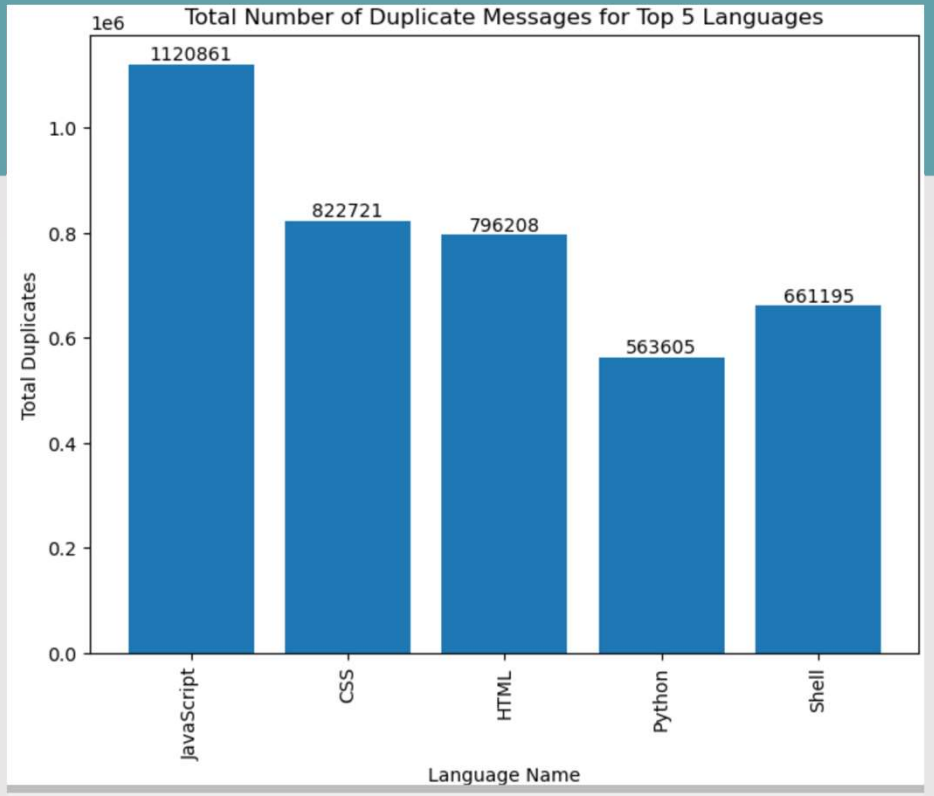
- We found that the programming language "C#" was most frequently duplicated in both the subject and message content.
- The second most frequently duplicated programming language in both the subject and message content was "Less".
- This means that "C#" and "Less" are the two most commonly used languages in the projects we analyzed on GitHub.



Subject & Message of Language Analysis II 2020 - 2022

- We found that JavaScript is the most commonly used programming language.
- The second most popular language varies: for some users, it's Python, while for others, it's CSS.





Recommendations

- **Understand the Influence of Major Tech Companies**: The influence of major tech companies like Google, Microsoft, and Meta is evident in the GitHub ecosystem. Understanding this influence can help in predicting future trends and making strategic decisions. For example, the popularity of technologies associated with these companies, such as Chromium and Linux, indicates their significant role in shaping the tech landscape.
- **Embrace AI Tools for Productivity**: AI tools, particularly generative AI, have been shown to significantly increase developer productivity. They can help complete coding tasks faster, optimize existing code, and document code functionality more efficiently. Developers and organizations should consider integrating these tools into their workflow to maximize productivity and stay competitive.
- **Upskill to Work with AI**: As AI automates lower-level tasks, the demand for developers with skills to work with AI is increasing. Developers should consider upskilling in areas such as machine learning and AI implementation to stay relevant in the evolving tech landscape.

Conclusions

- AI tools can help developers work more efficiently by automating repetitive tasks. This allows developers to focus on more complex and creative work, which can lead to faster development cycles and better software. However, AI is not likely to completely replace human developers.
- Developers bring important skills to the table, like critical thinking and the ability to collaborate effectively, that AI can't replicate. AI is changing the role of developers, making them more like architects and problem solvers.
- In the case of data scientists, AI can automate many of their tasks, but the best results often come from a combination of AI and human input. AI can help data scientists create many different models and simulations to find the best solution.
- In conclusion, while AI can make developers and data scientists more productive, it's not expected to replace them. Instead, AI is a tool that can help these professionals work more efficiently.

Reference

- https://gs.statcounter.com/browser-market-share
- https://www.cnbc.com/2023/01/18/apple-had-slower-headcount-growth-than-tech-peers-no-layoffs-yet.html
- https://www.pnas.org/doi/10.1073/pnas.2219396120
- https://gist.github.com/paulmillr/2657075/a314557294406724 67ada20ac10452d74a871e54