CS 2704
Term Project

Matthew Kenneth Peterson (#3719754)
https://github.com/MKP157/CS2704-Project

December 1, 2023

1 Introduction

Of all aspects of the culture surrounding computer science, and software development in particular, the concept of open-source software (OSS) and its development is perhaps one of the crowning achievements of the community at large. It is a remarkable tool no matter its shape or size, with applications in education, encouraging greater collaboration and more, all with an emphasis of accessibility. An open-source program is, by definition, open to *anyone* to modify and use, so long as its license permits it[1]. In turn, this means that the majority of OSS is free for anyone to use, regardless of whether or not they intend on editing the source code.

It seems as though many developers who dedicate their time and effort to open-source endeavours do it not only for possible personal gain, but for the love of their community. Participation in OSS projects for most programmers means doing the work on their own time, away from a professional environment. Further, because the majority of open-source projects are made available for free, this means that most who contribute to these projects tend to see little if any monetary gain from their endeavours aside from acts of charity (i.e. donations).

In this way, it is sufficient to say that open-source contribution is much akin to volunteer work. Hard work with nothing expected in return, giving one's time to some sort of cause. But that is simply by definition, and I am curious to see if it holds up in real world practice. Therefore, my goal for this project is to see if there is a possible correlation between areas of the world where it is more common for individuals to participate in bettering their communities, and areas of the world where it is more common for software programmers to contribute towards open-source endeavours. Currently, my hypothesis is that there will be a strong positive correlation between the two.

[1] Red Hat, "What is open source?," October 24, 2019. [Online]. Available: https://www.redhat.com/en/topics/open-source/what-is-open-source

2 The Data

Unfortunately, this is not a heavily surveyed topic. Metadata for these sorts of studies is difficult to gather, especially when necessary APIs are not made available by the major OSS platforms. However, what data there is available on the topic may reveal some interesting results.

For this exploratory project, I will be using the data collected by a group of researchers out of Vienna from 2021, in which they plotted the density of GitHub users who made over 100 contributions to OSS, by country and certain sub-regions. [2] The reason they had done this was to attempt at establishing a positive correlation between a country's gross national income (GNI) and its open-source contribution count. It is remarkable work, as they have gone so far as to account for discrepancies in the number of accounts against the number of actual individual persons who have used the service. They have done this by taking the GitHub archived history of all public commits, and queried its saved email and personal information against the GitHub account lookup API. Considering the tools at their disposal, this may be the best snapshot we have into the location data of the open-source community.

To compare this GitHub data with, I have decided on using data that describes a country's allowance for its citizens to participate in democracy, along with actual participation numbers, to see if there is any correlation between more democratically aligned countries and their likelihood to contribute to OSS. This data will come from the Economist's Intelligence Unit (EIU) Democracy Index for the year 2021, as this data weights each country's democratic performance using multiple factors.[3] The variation of the data set which I have used was cleaned by the organization Our World in Data, and posted to their GitHub account for public use[4]. This includes their free and fair elections score (`elect_freefair_eiu`) and their democratic participation score (`pol_part_eiu`). This data set scores each individual trait on a scale from 0.00 (least democratic) to 10.00 (most democratic).

Democratic performance data, unlike volunteer numbers, is heavily documented. There were many indexes I could have drawn my data from; the only reason I chose the EIU's is because it distinctly separates electoral participation and democratic freedom, which I will use when weighting my results.

[2] J. Wachs, M. Nitecki, W. Schueller, A. Polleres. "The Geography of Open Source Software: Evidence from GitHub." Technological Forecasting and Social Change (2022): https://www.sciencedirect.com/science/article/pii/S0040162522000105

[3] The Economist Intelligence Unit. "Democracy Index 2021: the China challenge." [Online]. Available: https://www.eiu.com/n/campaigns/democracy-index-2021/

[4] Herre, B., "eiu_cleaned.csv," Scripts and datasets on democracy, February 2, 2023. [Online]. Available: https://github.com/owid/notebooks
/blob/main/BastianHerre/democracy/datasets/cleaned/eiu_cleaned.csv

3 Hypothesis

First and foremost, I believe that a country's overall democracy index will quite strongly and positively correlate with its GitHub contributions per capita. However, in order to come up with the best possible results, I have not only compared the base democracy index with the GitHub contributions per capita, but have used three sub-scores from the EIU which they had used to create the combined democracy index. Because this analysis was more of a thought experiment, I thought it may be interesting to take a look at the different things possibly affecting the overall correlation scores, and I have come up with an individual hypothesis for each.

The first metric I used was the "political participation" index, calculated by the EIU and included in the data set under the column `pol_part_eiu`. This index is meant to describe the actual participation by citizens in elections, and I believe this will have the strongest correlation with the data set.

The second metric I used was the "free and fair election" index, found in column `elect_freefair_eiu`. This describes the extent to which a country hosts free and fair elections for its citizens. I believe this one will also correlate pretty strongly with the GitHub data set, because it makes sense that a country where political participation is discouraged to have a less community-oriented culture overall.

The third and final metric I used was the democratic culture index. That is, the extent to which a country's citizens prefer democratic institutions over non-democratic ones. This is found under the column `dem_culture_eiu`. This hypothesis is trickier for me to formulate, as there are many democratic countries where their respective citizens have a particular distaste for democracy. Mexico, for example, is a fairly democratic country with low approval for democracy within my dataset. So while I still believe this will correlate positively with the GitHub data, I cannot be certain how strong this correlation will be.

4 Descriptive Analytics

In order to compare the data, I plotted each index set against the GitHub data on a scatter plot, using Seaborn[5]. Seaborn uses Matplotlib's[6] `pyplot` functionality in order to create its graphs, and also allows for regression lines to be plotted without needing to calculate them first. It also easily allows for polynomial-degree regression lines for better fitting. After initial testing, I went with order-2, as it was a better fit. Considering Seaborn was easy to use for these purposes, I went with it for my analysis.

To calculate a necessary p-value, I used Scipy's statistics module's implementation of linear regression, as it would calculate for me the necessary metrics.

[5] Waskom, M., Seaborn: Statistical Data Visualization Library for Python. [Online]. Available: https://seaborn.pydata.org/.

[6] Hunter, J. D., "Matplotlib: A 2D graphics environment," [Online]. Available: https://matplotlib.org/.

It is also what Seaborn uses to calculate its own regression plots, and so using Scipy ensured the results to match. After plotting each index and its regression lines, as well as outputting their p-values, here were my results:

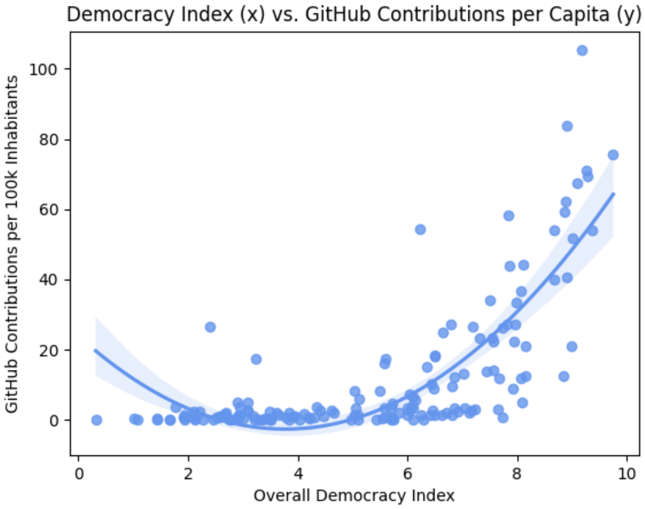

Figure 1: Democracy Index vs. GitHub Contributions per Capita

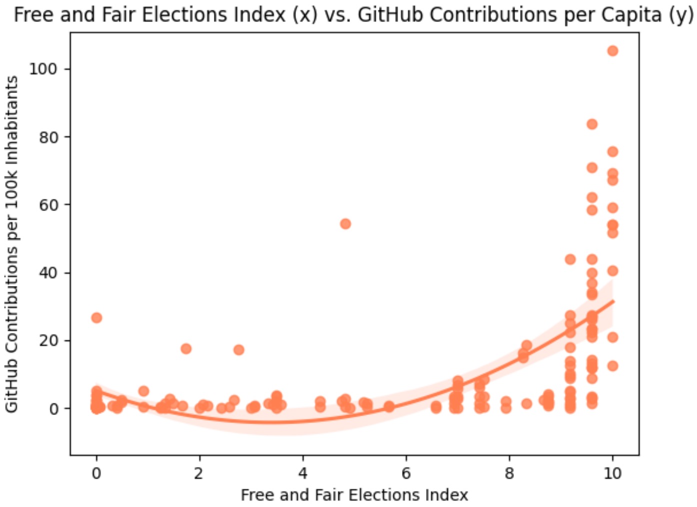

Figure 2: Free and Fair Elections Index vs. GitHub Contributions per Capita

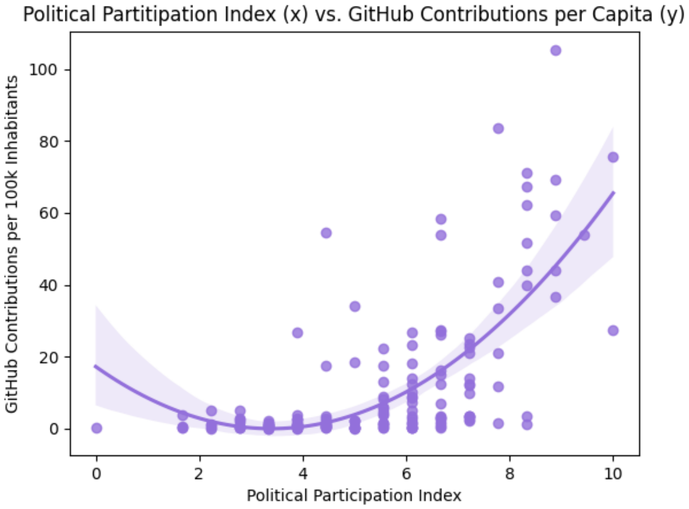

Figure 3: Political Participation Index vs. GitHub Contributions per Capita

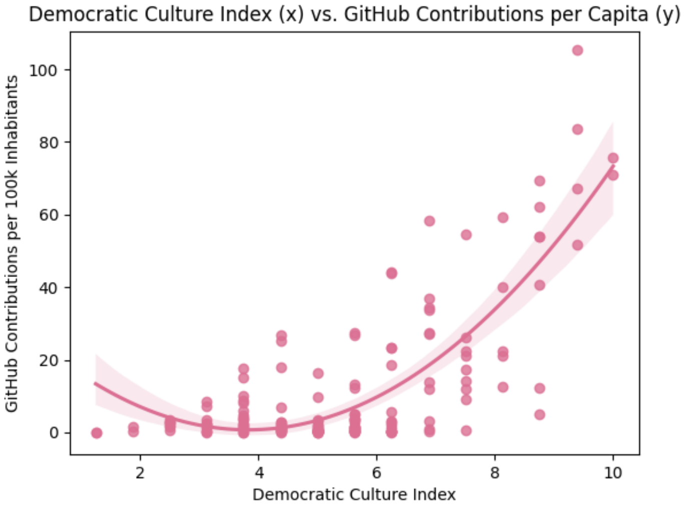

Figure 4: Democratic Culture Index vs. GitHub Contributions per Capita

Index	Democracy	Free and Fair Elections	Political Partic.	Democratic Culture
R^2 Value	$7.66 \times 10^{-}21$	$2.05 \times 10^{-}11$	$3.79 \times 10^{-}17$	$2.57 \times 10^{-}22$

Table 1: Calculated p-values for each data set.

The p-values seen in 1 are remarkably impressive, and indicate a strong statistical significance found when computing the correlation between the data sets.

5 Predictive Analytics

In order to test the data sets' strength as a prediction model, I had to reuse and reshuffle my data over many variations and take the average R^2 score of them all. I accomplished this by using Scikit-learn's[7]`utils.shuffle` in order to randomly reorder my aggregated data efficiently with each iteration, then splitting it in half with Pandas' sample function. I then would retrain and retest a new Scikit-learn `LinearRegression` model with each shuffle, because Scikit-learn's `LinearRegression` has a user friendly scoring function to easily test predictive performance. Over 10,000 iterations (my provided source code says only 1000 so as to not explode your machine), I was able to produce the average prediction scores as seen in Table 2.

Index	Democracy	Free and Fair Elections	Political Partic.	Democratic Culture
R^2 Value	0.3824	0.2079	0.3125	0.3958

Table 2: R^2 results for predictive performance, averaged over 10,000 iterations.

The R^2 scores convey a strong predictive ability found within each variation of my data. While these are not as good as the regression scores found by the initial researchers who collected the GitHub data for comparison against countries' GNIs (their R^2 was 0.65), I am still quite happy with the results. Interestingly enough, it seems that my hypothesis for the free and fair elections index and the democratic culture index were reversed. It seems that the democratic culture index was, in fact, the most strongly correlated index against the GitHub contributions per capita. This is confirmed when comparing their previously calculated p-values, where the democratic culture index's score is magnitudes smaller than the free and fair elections index.

[7] Pedregosa, F. et al., "Scikit-learn: Machine Learning in Python," [Online]. Available: https://scikit-learn.org/stable/.

6 Discussion

This analysis has been an interesting exercise in data analytics. In one go, we have been able to determine a strong, statistically significant correlation between two otherwise independent data sets, using multiple forms of available regression models using Python. I am very proud of the work I have been able to put into this project, especially considering my weak statistical background. That being said, there are a lot of fundamental flaws with the way I have approached this analysis, as I have been rather naive in its execution.

First of all, I do not believe the data I used lends any credence towards any real-world implications. There are likely many more fundamental factors that come into play than just a country's overall participation in democracy which may contribute towards its overall likelihood to participate in open source software. To name a few, I believe access to technology, access to education in software engineering, GDP per capita, computing culture, and availability of software engineering employment would all heavily affect my results.

Second of all, I cannot be sure how reliable this data is. While the original researchers mentioned earlier claim to have pinned down each accounts' geographical location through a variety of sampled metadata, including the locations of posts from associated Twitter (now X.com) accounts and randomly polled GitHub contributions from each account, this does not account for possible interference by Virtual Private Network (VPN) usage, which grows ever more popular by the day in order to maintain privacy online. There are more reasons why I think this research is inconclusive, however these two are the strongest counter arguments posed to me.

7 Further Research

For the above listed reasons, I believe further research is required before making any concrete conclusions. There are many ways I would thus improve my approach going forward:

- Use more data from other years, as I only used data from the year 2021
- Introduce other indicators, such as human development indices and education rates
- Account for Virtual Private Network (VPN) usage
- Account for other OSS platforms and their users, such as GitLab
- Account for companies which explicitly contribute to open-source (i.e. IBM's Red Hat), and analyze their contribution towards the numbers
- Use a singular (maybe self-produced) linear regression package for Python with all necessary features for the test, in order to reduce variability in results