DOWNLOAD THE DATASETS

Full dataset:
https://drive.google.com/file/d/1fdpvAp3u7hksi0q-E4eJhAFPJogkur3z/view?usp=sharing
Sample dataset:
https://drive.google.com/file/d/1EzbSfRFda8rYjluxjlD-uBF-rjvMabUp/view?usp=sharing

Tech report

Q: How many data points are there in total? How many are there in each group you care about (e.g. if you are dividing your data into positive/negative examples, are they split evenly)?
In our Stack Overflow data there are 17.7 million rows, in “Most Popular Programming Languages Since 2004” there are 247 rows, and in the Github Programming Languages data there are 3375 rows. The main groups we care about are the years; for the latter two datasets the years are split approximately evenly, but for the Stack Overflow data the number of entries per year is slightly more random. Getting a meaningful exact count for each year (and other groupings, like count by language) will require significantly more preprocessing, but we anticipate that overall counts will on average increase over time as Stack Overflow usage increases.

Q: Do you think this is enough data to perform your analysis later on?
Yes. With millions of rows of unique data points before the join, this should give us a robust and plentiful source of data.

Q: What are the identifying attributes?
In the first dataset, the id field uniquely identifies Stack Overflow questions. In the second dataset, the Date field uniquely identifies each row. In the third dataset, the name of the programming language along with the year and quarter uniquely identify each row. So the tuple (id, Data, name, year, quarter) uniquely identifies each row in the joined dataset.

Q: Where is the data from?
The data is from three Kaggle datasets. The first pulls from Stack Overflow data that is updated quarterly. The second pulls from PYPL, the PopularitY of Programming Language Index, which measures the frequency of language tutorial Google searches. The third originates from Github data on programming language popularity from 2011 to 2021.

Q: How did you collect your data?
We downloaded it from Kaggle.

Q: Is the source reputable?
Yes: Stack Overflow and Github are known to be reputable, and since the data for the second dataset comes from analyzing Google searches using a straightforward methodology, this is also reputable.

Q: How did you generate the sample? Is it comparably small or large? Is it representative or is it likely to exhibit some kind of sampling bias?
Our sample was taken by joining the tables in the same dataset, then randomly sampling 1500 rows.

Q: Are there any other considerations you took into account when collecting your data? This is open-ended based on your data; feel free to leave this blank. (Example: If it's user data, is it public/are they consenting to have their data used? Is the data potentially skewed in any direction?)
User-specific data is pulled from user-generated content on Stack Overflow, but since it is public and users understood they were posting to the Internet, using it shouldn't pose an ethical issue.

Q: How clean is the data? Does this data contain what you need in order to complete the project you proposed to do? (Each team will have to go about answering this question differently but use the following questions as a guide. Graphs and tables are highly encouraged if they allow you to answer these questions more succinctly.)
The data is fairly clean, though there are some blank values, which we replaced with “N/A.”

Q: How did you check for the cleanliness of your data? What was your threshold reference?
We scanned the sample. We had no reference, because no important fields were empty or unclean.

Q: Did you implement any mechanism to clean your data? If so, what did you do?
Yes, we replaced empty/blank values with the “N/A” string.

Q: Are there missing values? Do these occur in fields that are important for your project's goals?
Yes, but they only occur in unimportant fields. We changed them to “N/A”.

Q: Are there duplicates? Do these occur in fields that are important for your project's goals?
There could theoretically be duplicates in some fields (e.g. two posts with the exact same text), though this is unlikely. However, each row in each table has at least one field with a unique value (e.g. an ID), so there will never be duplicate rows.

Q: How is the data distributed? Is it uniform or skewed? Are there outliers? What are the min/max values? (focus on the fields that are most relevant to your project goals)
Details on range of values of the fields can be found in our data spec.

Q: Are there any data type issues (e.g. words in fields that were supposed to be numeric)? Where are these coming from? (E.g. a bug in your scraper? User input?) How will you fix them?
No

Q: Do you need to throw any data away? What data? Why? Any reason this might affect the analyses you are able to run or the conclusions you are able to draw?
Only data that is irrelevant to our analysis. More details in the data spec.

Q: Summarize any challenges or observations you have made since collecting your data. Then, discuss your next steps and how your data collection has impacted the type of analysis you will perform. (approximately 3-5 sentences)
One challenge will be joining the data in an efficient and meaningful given the size of the Stack Overflow dataset. Our next step here will be to identify the precise questions we would like to answer and only join the data on relevant conditions, filtering rows that can are irrelevant to the question at hand (e.g., a time period outside the scope of the period we are examining). The upside of this is that the collection process has given us an abundance of data to work with, so we have a rich trove of raw material to analyze, and well-structured analyses are more likely to be statistically significant by the LLN. We can thus perform more specific, refined analyses with reasonable confidence.