Data Analysis:

1. The data collection part has been tricky because the GitHub API has not been giving consistent results.
   * 1. Last Friday, when I had collected the data for the 13k repository form, the first look, everything seemed fine, but then I later realized there were many false negatives.
     2. So, I tried using spark to parallelize the process because the program can take hours to run.
        1. But this resulted in more false negatives as GitHub API takes time to fetch the data, in which case you need to retry the API
        2. I went ahead with collecting the data for 8k results out of the 13k using a single thread mechanism.
2. Analysis of running the program on 8k repositories
   1. There are a total of 1869 GitHub repositories that have data.
   2. There are 512 repositories that no longer exist or have been removed from GitHub.
   3. There are a total of 6204 repositories that don't have dependency data. From the initial analysis, I have concluded that they don't have dependencies because GitHub could not find any dependency files for that project.
3. Analysis of running the program on the remaining 5k repositories
   1. 1143 of the total repositories have data
   2. 4457 repositories currently don't have files that allow GitHub to extract dependency data
   3. 500 of the repositories with an error message or timeout