Report Outline

I. Introduction

1. What are we doing?
2. Why are we doing it?
3. What's in the report?

II. Collecting Source From GitHub

1. Introduce the websites involved
    1. What is GitHub?
    2. What is Ghtorrent?
    3. What is Big Query?
2. Introduce what we are collecting from these sited
    1. metadata from Ghtorrent w/ Big Query
    2. GitHub project data
    3. GitHub commit data
3. Discuss pros/cons of collecting source from GitHub
4. Give an overview of the process of collecting data from GitHub
    1. getting metadata
    2. collecting project files
    3. collecting commit files
5. Give a technical description of each process
    1. Getting metadata
        1. how are we doing it
        2. what does the data look like
        3. technical issues not included in pros and cons
    2. Getting project files

1. how are we doing it
2. what does the data look like
3. technical issues not included in pros and cons

3. Getting commit changes
    1. how are we doing it
    2. what does the data look like
    3. technical issues not included in pros and cons

6. Final Thoughts

III. Collecting Source From Codeforces

1. Introduce the website
2. Introduce what we are collecting
3. Discuss pros/cons of collecting from Codeforces
4. Give an overview of the process
    1. Collecting and transforming user data
    2. Collecting submissions using selenium
    3. Collecting submissions using links
5. Give the technical details for each process
    1. Collecting and transforming user data
        1. how are we doing it
        2. what does the data look like
        3. technical issues not included in pros and cons
    2. Collecting submissions with selenium
        1. how are we doing it
        2. what does the data look like
        3. technical issues not included in pros and cons

3. Collecting submissions with links
    1. how are we doing it
    2. what does the data look like
    3. technical issues not included in pros and cons
6. Final Thoughts

IV. Collecting Gender Information

1. Introduction to how we are labelling the data
    1. Collecting gender for first names
2. Information on available sources for this
    1. The APIs and websites available
3. The pros/cons of the sources
    1. how accurate are they
    2. what do they cost
4. Which source we chose and why
5. Explanation of the process
    1. Overview
    2. Technical details
    3. What this data looks like
    4. extra technical issues
6. Final Thoughts

V. Conclusions

1. Final thoughts
2. Discuss the strengths and weaknesses of our methods
3. Discuss future extensions and improvements