

Mining Software Repositories Assignment 2 (MSR2)

Submit Assignment

Due Saturday by 10am **Points** 15 **Submitting** a file upload **File Types** pdf and ipynb

This is the second of two mining repositories assignments that will teach you how to investigate data within real-world project repositories. In this assignment, we will focus on using quantitative analysis and graphical data representations to explore and present data-driven results.

This is an individual assignment, and each student is expected to develop their own methods for answering the questions. This type of analysis can be accomplished in multiple ways, so please try to develop an intuition for your chosen methodology.

Problems:

FILE: [MSR2.ipynb](#)

The Jupyter notebook file included above should be your starting point for this assignment. All questions outlined below are also included in the file. The assignment has only one part, but the questions are expected to take more time than on the first MSR assignment:

Part 1: bitcoin/bitcoin repo

For this assignment, we will investigate the [bitcoin/bitcoin](https://github.com/bitcoin/bitcoin) [_ \(https://github.com/bitcoin/bitcoin\)](https://github.com/bitcoin/bitcoin) project.

- Question 1: Using a barchart plot to illustrate your answer, show which day of the week receives the most new issues.
- Question 2: Using the issues that you collected in Question 1, determine the descriptive statistics for:
 - word counts in the issue text
 - number of comments
 - word counts of comments
- Question 3: Now create three plots to illustrate the descriptive statistics you found in Question 2.
- Question 4: Does the number of comments or the word count in comments impact the duration that the issue remained open?

What to hand in:

Report (filename: *[onid-username]-msr2.pdf*)

- Detail your methodology and results for each question, and include citations for any references/resources used.
- Write in an academic tone similar to the research papers assigned in the course. Be precise and write only as much as needed to convey your results.
- Any format is allowed, but if possible follow the [ACM](https://www.acm.org/publications/proceedings-template) [_ \(https://www.acm.org/publications/proceedings-template\)](https://www.acm.org/publications/proceedings-template) or [IEEE](https://www.ieee.org/conferences/publishing/templates.html) [_ \(https://www.ieee.org/conferences/publishing/templates.html\)](https://www.ieee.org/conferences/publishing/templates.html) formats since it will provide you with practice.
- If you include plots in your report, be sure to include captions that provide an adequate description of the data being visualized.
- Do not simply include code to demonstrate your methods (those implementation details belong in the notebook file).

Jupyter Notebook (filename: *[onid-username]-msr2.ipynb*)

- Include all data extraction and analysis code used to answer each question.
- Provide code for any plots used to answer each question.
- Scrub any personal or confidential information (e.g. authentication tokens) from the file.
- Do not hard-code data into your notebook, but instead include your commands for interacting with the GitHub API.

Some Rubric (1)