PhD Candidate Coding Exercise: Repository Insight Analysis

Objective
Analyze a GitHub repository by selecting one predefined research question, collecting and analyzing data, and reflecting critically on the findings — including any challenges faced and any use of AI tools. **Please read the whole document before answering the questions.**

Instructions

1. **GitHub Repository to study**
   You may use tools like PyDriller, GitHub API, or any other method you prefer to extract commit history and metadata from the github project `numpy/numpy`

2. **Pick One Research Question (RQ)**

   You do **not** need to come up with your own RQ. Choose one from the following:

   - RQ1: How does the number of commits change over time (monthly or weekly)?
   - RQ2: Which files are changed most frequently, and what file types dominate the churn?
   - RQ3: How does code churn (lines added/removed) fluctuate over time?

3. **Write Code to Collect and Process Data**

   - Extract commit-level data (e.g., author, timestamp, files, insertions, deletions).
   - Organize the data into a clean structure (e.g., DataFrame or CSV).
   - Save your script in a reproducible format (Jupyter notebook or Python script).

4. **Create at Least 2 Visualizations**

   - Include charts such as line plots, bar plots, or heatmaps to illustrate the trends related to the selected RQ.
   - Ensure that all plots are labeled (axes, title) and include a short caption or description for each.

5. **Answer the Following Reflection Questions**
   Please write your answers in a short document (`README.md` or similar):

   - What difficulties or errors did you face while completing this task?
   - Share the full conversation (link or PDF export) with ChatGPT or any other AI tools that you used to help you.
   - Choose one of your plots: What do you find surprising, confusing, or ambiguous about it? What might explain this?
   - What would be one interesting follow-up question or analysis to pursue based on your findings?

Submission Requirements

Please submit the following:

- Your code as a Jupyter notebook or Python script
- A folder with your plots
- A short summary document answering the reflection questions
- If you used ChatGPT or any AI assistance, include the full session as a transcript or PDF

Evaluation Criteria

Criterion	Excellent	Good	Needs Work
Data Collection	Script is correct, clean, and relevant	Mostly correct with minor issues	Major errors or shallow approach
Visualizations	Clear, well-labeled, insightful	Mostly clear and relevant	Poorly labeled or uninformative
Analytical Reflection	Shows originality, reasoning, and interpretation	Some insight or useful observation	Vague, generic, or surface-level

AI Use Transparency	Fully disclosed and explained	Partial explanation or unclear use	No disclosure or over-reliance
Handling Ambiguity	Thoughtfully interpreted unclear results	Some attempt at explanation	Ignored or guessed poorly