AI Assistant Testing & Validation Checklist

Use this checklist to guide your testing and validation. The purpose is not to show that everything works. There will be bugs or weaknesses that you can't fix completely this semester and some of these checks may not be implemented in your app. The goal is to document what you tested and what you learned about your app. It's important to keep a log of the issues and any fixes you implemented. Also note if there's anything on the checklist that your app doesn't do (yet). This will serve as a log of things you might want to do in the future to make your app more safe, accurate and reliable and to inform users that the app is still under development and may not work well in all situations.

Functional Testing

- ☐ App runs without crashing
- ☐ Dataset upload works correctly
- ☐ User questions trigger appropriate analysis
- ☐ Generated code runs and returns a result
- ☐ Plots and tables are displayed as expected

Input Validation

- ☐ App shows a clear error if a bad or empty file is uploaded
- ☐ App handles invalid or confusing questions gracefully
- ☐ App doesn't crash with missing or messy data columns

Code Accuracy Check

- ☐ LLM-generated code runs in a separate script or notebook
- ☐ Code uses correct columns and logic
- ☐ Output matches expectations (e.g., correct means, valid plots)
- ☐ Errors in generated code are detected and explained or caught

Output Invariance

- ☐ Asking the same question twice returns consistent results
- ☐ Small changes in question wording don't change answers dramatically
- ☐ If results vary, explain why you think your app might not be working reliably

Usability Testing

- ☐ A classmate or friend (outside your team) was able to use the app
- ☐ You noted confusing steps or unclear messages
- ☐ You collected at least one suggestion for improvement
- ☐ Optional: You asked for a quick usability rating (1–5 stars)

Scenario / Edge Case Testing

- ☐ You tested your app on at least 1 different dataset
- ☐ You tried a mix of numeric and categorical questions
- ☐ App still worked with datasets with missing or oddly named columns

Testing Summary Report

- ☐ You created a short summary of what you tested
- ☐ You noted what worked well and what still needs work
- ☐ You listed at least one improvement you would make with more time