Project Transformation Reflection

The link to the github repo is here.

Original State of the Analysis

The initial analysis consisted of scattered Jupyter notebooks with hardcoded paths, manual data cleaning, and no documentation. Results were hard to reproduce reliably due to missing environment specifications and inconsistent preprocessing steps.

Biggest Challenges in Transformation

Git and version control was the most significant hurdle as a beginner - learning git commands, proper commit practices, and understanding when to commit changes. **Dependency management** presented another challenge - the original notebooks had undocumented implicit dependencies causing "works on my machine" problems.

Most Impactful Improvements for Reproducibility

1. **Automated Pipeline Script**: Converting from manual notebook execution to `run_analysis.py`.
2. **Comprehensive Testing Suite**: Adding data validation and pipeline integrity tests.
3. **Requirements Management**: Pinned versions in `requirements.txt` dealt with the environment inconsistencies.

What I Would Do Differently

In future projects, I would adopt **modular function design** earlier - the current script could benefit from separate modules for preprocessing, visualization, and analysis. Finally, **proper logging** should replace the current print statements for better debugging capabilities.

Implementation Timeline

- **Project Structure Setup**: 0.5 hours
- **Dependency Documentation**: 1 hour
- **Data Pipeline Refactoring**: 3 hours
- **Visualization Standardization**: 2 hours
- **Testing Suite Development**: 2.5 hours
- **Documentation Creation**: 1.5 hours

- **Final Testing & Debugging**: 1 hour