PA 446 Final Project Guidelines

Fall 2025

Overview

The final exam for PA 446 is a project where you will demonstrate your ability to apply the technical and conceptual skills learned throughout the semester to a real-world civic data problem. You will identify a civic issue, acquire and analyze relevant data, build reproducible outputs, and communicate findings to a civic or policymaking audience.

This project should showcase both your technical mastery in R and your ability to think critically about equity, fairness, and the public value of data.

Step 1: Choose a Civic Problem (Context)

- Select a civic or public administration issue such as:
  - Housing code violations
  - 311 service requests (potholes, sanitation, noise)
  - Budget or expenditure transparency
  - Transit access and equity
  - Environmental justice or health disparities
- Define a clear research question (e.g., *Are 311 pothole requests reported evenly across neighborhoods? Can housing violations be predicted by property characteristics?*).

Step 2: Data Acquisition

- Obtain at least one dataset using:
  - Web scraping (e.g., scraping a city agency's posted records), OR
  - API access (e.g., Census, Chicago Data Portal, Data.gov).
- You may combine multiple datasets (e.g., join Census demographic data with 311 data).
- Document your acquisition method.

Step 3: Data Wrangling & Quality Checks

- Clean and structure the data using R (dplyr, tidyr, janitor).
- Evaluate data quality and fairness:
  - Missing values, representativeness across neighborhoods, and potential biases.
  - Visualize missingness (naniar) or compare to population demographics.
- Document decisions you make during cleaning.

Step 4: Reproducible Reporting

- Build a Quarto report that integrates narrative, code, and visualization.

- Required components:
    - At least 3 visualizations (charts, maps, or dashboards).
    - An interactive component (dashboard element, HTML widget, or parameterized report).
    - Clear explanations of methods and findings.

Step 5: Advanced Analysis

Apply at least one advanced method from the course:

- Machine Learning (Intro): classification or clustering of civic data.
- NLP: analyzing public comments, transcripts, or civic documents.
- Network Analysis: mapping civic or nonprofit-government relationships.

Step 6: Workflow & Automation

- Use GitHub for version control (your repo must include all scripts, reports, and documentation).
- Demonstrate automation:
    - Example: parameterized Quarto reports for multiple neighborhoods.
    - Example: an R script that can be re-run to regenerate all outputs.

Step 7: Ethics & Fairness Reflection

- Include a section in your report discussing:
    - Biases or limitations in your dataset.
    - How these biases might affect decision-making.
    - How would you responsibly present findings to policymakers?

Deliverables (Due Dec 10 – Final Exam Date)

1. **GitHub Repository**
    - Must include:
        - Clean folder structure (data/, scripts/, reports/).
        - R scripts for scraping/API access and cleaning.
        - Quarto files for reporting.
        - README.md explaining the project, datasets, and instructions for replication.
2. **Quarto Report/Dashboard**
    - A polished, reproducible HTML or PDF report OR an interactive Quarto website/dashboard.
    - Should integrate narrative, analysis, and visualization.
    - Includes at least one advanced analysis (ML/NLP/Networks).
3. **Reflection Memo (2–3 pages)**

- Written for a non-technical civic audience (e.g., city officials, nonprofit leaders).
- Explain:
  - Civic problem addressed.
  - Steps you took to reach the findings. Show code and results
  - Key findings.
  - Equity and fairness concerns.
  - Policy implications.

Grading Rubric (100 points)

- **Problem Framing & Context (10 pts)** - Clear, relevant civic issue defined.
- **Data Acquisition & Cleaning (15 pts)** - Effective use of scraping/API, tidy data.
- **Project Structure & Reproducibility (15 pts)** - GitHub repo, documentation, reproducibility.
- **Visualization & Communication (15 pts)** - Clarity, interactivity, storytelling.
- **Advanced Analysis (20 pts)** – Correct application and interpretation of ML, NLP, or network analysis.
- **Workflow & Automation (10 pts)** - Parameterization, reproducibility, Git integration.
- **Ethics & Fairness (10 pts)** - Critical reflection on bias, transparency, civic impact.
- **Clarity & Professionalism (5 pts)** - Report polish, memo readability.

Tips for Success

- Start early - scraping and cleaning take more time than expected.
- Choose a dataset that genuinely interests you.
- Think about the audience - could a city agency use your work tomorrow?
- Don't aim for complexity - aim for clarity, reproducibility, and civic impact.

Example Final Project Prompts

1. Predict Housing Code Violations with Open Data + Census API

Civic Problem: Housing code violations often cluster in certain neighborhoods, but city inspectors have limited resources. Can data help predict where violations are most likely?

Data Sources:

- Scrape Chicago housing/building violations from the City of Chicago data portal.
- Pull neighborhood demographic & housing data from the Census API.

Tasks:

- Clean & merge violations with Census demographics (e.g., poverty, housing age).
- Use classification (ML) to predict whether a property will have a violation.
- Create an interactive dashboard showing predicted high-risk areas.
- Reflect on fairness: what are the risks of using predictive models in housing enforcement?

2. Mapping & Clustering 311 Complaints for City Services

Civic Problem: 311 requests (potholes, graffiti, sanitation, streetlights) provide insights into service needs - but how can the city identify patterns to allocate resources better?

Data Sources:

- Pull 311 service request data from the City of Chicago data portal API.
- Optional: join with neighborhood boundary shapefiles (GIS data).

Tasks:

- Wrangle and clean 311 complaint data (filter by type, location, date).
- Use clustering (ML) to group neighborhoods by service complaint patterns.
- Build an interactive map dashboard of complaint clusters.
- Discuss bias: are some communities under-reporting requests due to lack of awareness or trust in city services?

3. Analyzing City Council Meeting Transcripts with NLP

Civic Problem: City council meetings contain rich policy debates, but the volume of transcripts makes them hard to digest for the public.

Data Sources:

- Download or scrape Chicago City Council transcripts (many cities post them online).
- Supplement with FOIA or policy document text datasets.

Tasks:

- Use NLP to:
    - Identify the most discussed policy issues over time (topic modeling).
    - Extract key phrases or sentiment by speaker.
    - Build a word network of how issues are connected.
- Present results in a Quarto report or interactive dashboard (e.g., word embeddings visualization).
- Reflect: does NLP miss nuance or context (e.g., sarcasm, tone) that matters for policymaking?