README file for "The Role of Case Management in Misdemeanor Prosecution"

By Lindsay Graef and Aurélie Ouss

This readme file contains instructions for how to replicate the results of our paper. The data that we use for this project are protected. The materials we provide here are our attempt to balance the need to make research transparent and replicable against the need to protect privacy and abide by our data provider.

Data Availability Statement

This replication deposit contains all of the code used to create our main dataset, run our analyses, and construct our tables and figures. However, this project uses protected data that cannot be shared publicly.

The data that we use for this project were accessed through a Memorandum of Understanding (MOU) with the Philadelphia District Attorney's Office (DAO), as part of a long-term collaborative research partnership. These data include internal human resources records and sensitive information on criminal cases, including for people whose case was eventually dismissed, expunged, or sealed. To protect privacy and abide by our data-use agreements, we are sharing only the code used in our analysis.

Our main analysis dataset will be kept on file at the DAO. We note that this dataset was produced in 2024, pulling data from several different administrative databases. Future data pulls may vary, due to expungements and changes in data systems. Additionally, one of our analyses makes use of L2 Voter files, which were accessed separately through Penn Libraries. We include the code that was used to merge and analyze these voter data, but we cannot include variables on political preferences in our main replication dataset, due to data use restrictions.

Requests to access these replication data can be made to the Philadelphia District Attorney's Transparency Analytics (DATA) Lab. For more details on the DATA Lab and to request data access, complete the research request form located here: https://data.philadao.com/Research

The DAO will preserve these records for at least 5 years, and we will assist with any request for reproducibility checks.

Project Folders

The following diagram shows the levels of the project tree structure at a high level. In this setup, we ran the analyses on a secure DAO server. Before running the code, it is necessary to update the paths for the relevant file structure on your computer.

dev/project_folder/
- PCM_README.pdf
- code/
- exports/

srv/project_folder/
- data/
    - benchmark_res_hearing_team/

    - benchmark_res_robustness/

R and Python Scripts
The following scripts are included in this replication packet.

- RCM_Data_Build.R : creates our main analysis dataset
- RCM_Benchmarking.R : runs benchmarking analyses
- RCM_Tables_and_Figures.R : creates all tables and Figures A1 through A3
- RCM_Figures.ipynb : creates Figures 3 through 5 and Figures A4 and A5
- RCM_Misd_Caseflow : gets numbers for Figure 1

Software and Computational Requirements
This section describes the software and computational requirements for running the relevant analysis code, as well as the relevant packages that must be installed.

- R (version 4.4.0)
    - **tidyverse** 2.0.0
    - **tidytable** 0.11.1
    - **tidyr** 1.3.1
    - **stringr** 1.5.1
    - **lubridate** 1.9.3
    - **dplyr** 1.1.4
    - **conflicted** 1.2.0
    - **daotools** 0.1.72
    - **daocore** 0.8.6
    - **arrow** 17.0.0.1
    - **modeest** 2.4.0
    - **NIBRSFunction** 0.0.0.9000
    - **data.table** 1.16.2
    - **odbc** 1.5.0
    - **dbplyr** 2.5.0
    - **DBI** 1.2.3
    - **readxl** 1.4.3
    - **readtext** 0.91
    - **fastDummies** 1.7.4
    - **janitor** 2.2.0
    - **modelsummary** 1.4.3
    - **fixest** 0.12.1
    - **raster** 3.6-30
    - **fastDR** 1.0-1
    - **survey** 4.4-2
    - **gbm3** 3.0
    - **kableExtra** 1.4.0
    - **ggplot2** 3.5.1
    - **patchwork** 1.3.0
    - **extrafont** 0.19

- Python / Jupyter Notebooks (version 7.3.2)
    - **pandas** 2.2.3
    - **matplotlib** 3.10.0

- **seaborn** 0.13.2
- **statsmodels** 0.14.4