scientific data

OPEN

DATA DESCRIPTOR

A comprehensive dataset of United States federal procurement from 1979 to 2023

Aymane Omari[1], Nasser Alansari[2], Brian Libgober[3] & Aaron R. Kaufman[1]✉

The United States government spends hundreds of billions of dollars annually obtaining goods and services from commercial bidders, a process known as government procurement. Without transparency, this process is susceptible to manipulation, corruption, and grift. However, the data required to assess the procurement process, including information detailing the amounts and awardees of government contracts, is difficult to obtain or study, though existing scholarship has made important contributions relying on limited samples of this data. We collect, clean, analyze, and make available the entire corpus of federal procurement contracts from 1979 to 2023, nearly 100 million contract actions over 45 years, and introduce an R package for accessing subsets of the data. Overall, these tools hold great promise for studying representation, corruption, and the connections between business and politics in the United States.

Background & Summary

The United States federal government is the world's largest organization with a 2024 annual budget of $6.75 trillion (https://fiscaldata.treasury.gov/) and nearly 3 million employees (https://www.statista.com/statistics/204535). To obtain the goods and services necessary to perform its many functions, the U.S. relies on *government procurement*, a system by which private businesses can bid on contracts to supply everything from office supplies and clothing to rocket engines and satellites. Procurement contracts account for more than 10% of the U.S. annual budget, occupying more than 100,000 government employees[1]. Overseeing this process is a complex web of laws, regulations, legislative committees, executive agencies, and special courts designed to ensure fairness and transparency and prevent fraud and corruption.

On the one hand, a robust suite of legislation implements preferential contract awarding to small, minority-owned, and women-owned businesses. As many as 10% of minority-owned businesses in the U.S. contract with the government[2], potentially strengthening minority entrepreneurs' access to product markets[3], especially as the share of minority employees in the government grows[4]. Procurement regulation can be a powerful force for strengthening democratic representation and the economic success of disadvantaged groups.

However, systems so large and lucrative are vulnerable to grift and corruption[5–7]. As many as 44% of contracts receive only a single bid[8], and bidders or potential bidders may collude[9] to maximize their profits. Politicians also have incentives to corrupt the procurement process: securing grants for constituent firms or copartisan entrepreneurs may be rewarded with votes or campaign contributions[10,11]. Waste, inefficiency, and political manipulation in procurement can reduce trust in government[12], harm taxation morale[13], and may even contribute to democratic backsliding[14].

Studying the impacts of procurement policies on representation and inequality, or corruption and fraud, requires access to reliable and systematic procurement data. Researchers, journalists, and citizens primarily access this data through the Federal Procurement Data System (FPDS, https://www.fpds.gov/). This system provides users with an interface for keyword searches, filtering by variables like the government department issuing the contract, ordering the results, and downloading search results to CSV files. Figure 5 shows the results for a search of the term "boots". The first contract in the search results was posted by the Department of Veterans Affairs for $3,000, and was awarded to "BOOTS PHAR INC", a pharmaceutical company.

[1]New York University Abu Dhabi, Division of Social Science, Abu Dhabi, United Arab Emirates. [2]New York University Abu Dhabi, Center for Research Computing, Abu Dhabi, United Arab Emirates. [3]Northwestern University, Department of Political Science and Pritzker School of Law, Chicago, IL, USA. ✉e-mail: aaronkaufman@nyu.edu

FPDS is a crucial source of rich, granular data for studies in political science, public administration, economics, public policy, and many other areas. A Google Scholar search for "federal procurement data system" shows more than half a million results as of August 2025. Yet crucial hurdles prevent researchers from freely accessing this data. For performing comprehensive analyses on representation, inequality, and corruption, the current interface is certainly insufficient. There is no functionality for downloading bulk data, nor are there tools for calculating aggregate statistics or tabulating trends across time. A wealth of data on government operations is functionally locked behind an outdated website.

In this paper, we overcome this obstacle by introducing two tools for querying and analyzing FPDS data. The first is a comprehensive collection of FPDS data from 1979 to 2024. Although the government FPDS website includes some records dating back to 1957, these are too sparse to be considered complete. This contract-level dataset contains nearly 100 million entries and more than 200 covariates, including timestamps and geolocations; contract data; contracting agency codes; and vendor variables.

As this dataset is large and potentially cumbersome for many researchers, the second tool is a lightweight R package for compiling the results of specific FPDS searches into convenient data frames. First, users search the FPDS website for an arbitrary set of contracts and record the URL. They then pass this URL to the R package, which automatically scrapes and compiles the search results. This R package is available on CRAN and on GitHub (https://github.com/aymane-omari/fpdsScraper). This dataset opens up exciting possibilities for researchers, journalists, and citizens seeking to monitor public spending in the US.

Existing work using FPDS data

Existing literature demonstrates the diverse applications of FPDS data in analyzing various aspects of federal procurement. These studies highlight the importance of accessible and structured data for advancing public administration, policy analysis, and economics research.

For example, a corpus of descriptive work studies procurement practices and trends, focusing on specific agencies, events, or contract types[15] analyzes the utilization of specific procurement programs in the Air Force, while[16] investigates Department of Defense service contract trends from 1990 to 2011.[17] assesses federal contracts awarded for Hurricane Katrina recovery efforts[17], and[18] explores bundled and consolidated contracts at the U.S. Department of the Navy.

A related literature examines the causal effects of contract types across different administrative, economic, or political outcomes. Existing work finds that contract structure and contracting agency capacity can both impact the successful implementation and efficiency of procurement itself[19,20], and that administrative reforms can successfully improve that process[21], but this work crosses many disciplinary boundaries.

Economists have investigated the relationship between government customer concentration and firm performance[22] and how market conditions impact the choice between fixed-price and cost-reimbursement contracts[23], as well as the overall impact of procurement spending on local economies such as Hawaii's[24].

Public administration scholars have used FPDS data to explore the role of policy consultants[25] and nonprofits[26,27], and to uncover undisclosed subcontractors in public contract records and employment data in federal procurement[25].

And political scientists often study the role of procurement in politics and how politics shapes procurement, finding that procurement contracts are disproportionately allocated to politically important areas[28] or favored firms[29], and how different procurement policies impact contracting with minority-owned small businesses[30,31].

While improvements in open data can significantly advance transparency and accountability in federal procurement processes[32], data availability is also crucial to conducting research on procurement as one of the most important governmental outputs and economic drivers. Convenient and accessible procurement data enables scholars, policymakers, and citizens to engage more deeply with federal procurement data, fostering transparency, accountability, and informed decision-making in public spending.

Methods

FPDS Bulk Download with ATOM. To efficiently collect data from the Federal Procurement Data System (FPDS), we developed a custom Python spider utilizing the Scrapy framework. This tool automates the bulk download of procurement records by interfacing with the FPDS Atom feed.

To manage the large volume of data and adhere to the FPDS Atom feed's limitation of 399,990 records per search query, which will not output results beyond this limit, we implemented date range segmentation. The overall data collection period was divided into weekly intervals, starting on Mondays and ending on Sundays, without crossing calendar months. This segmentation ensured that each query remained within acceptable limits and optimized the scraper's performance by processing manageable chunks of data.

Our scraper constructs query URLs specifying these date ranges using the syntax `SIGNED_DATE:[YYYY/MM/DD,YYYY/MM/DD]`. This approach allows precise filtering of records based on their signing dates and ensures records are retrieved in chronological order.

To prevent overloading the FPDS servers and comply with ethical scraping practices, we incorporated concurrency control into our scraper. Scrapy's concurrency settings were configured to limit the number of simultaneous requests.

The scraper handles pagination automatically for each weekly range request, iterating through all available pages with a fixed page size of 10 records due to the FPDS Atom limitation (max records per request). This ensures comprehensive data collection without manual intervention.

The Atom feed responses are in XML format. We employed a library to parse these responses into Python dictionaries, facilitating easier data manipulation. Special characters and control codes are sanitized to prevent

parsing errors. The parsed data is then serialized into JSON format and compressed using gzip, optimizing storage space and enabling efficient handling of large datasets.

Following the data collection, a post-processing step is implemented to convert the collected JSON data into Parquet format. This conversion is done with monthly partitioning, which allows for efficient querying and analysis of the dataset. The Parquet format, known for its columnar storage, optimizes both storage space and read performance, making it ideal for large-scale data analysis tasks. And while Parquet is not as commonplace as the CSV format, it is equally convenient for loading into common statistical software like R (arrow) or Stata (stata-parquet) through packages that wrap the Apache Arrow C++ library.

By automating the bulk download process with the FPDS Atom feed, our scraper simplifies data retrieval from FPDS. This tool lowers barriers for researchers and analysts by providing a scalable solution for accessing large volumes of procurement data, thereby promoting greater transparency and enabling comprehensive analysis of federal spending patterns.

FPDS Data Extraction with the `fpdsScraper` R Package. We anticipate that for most users, even the reduced CSV dataset will be unnecessary and burdensome for their use cases. To efficiently retrieve and analyze smaller subsets of data from FPDS, we developed the `fpdsScraper` R package. This package streamlines the process of extracting procurement data by automating the collection of search results from the FPDS website into a structured and analyzable format.

The `fpdsScraper` package operates by accepting a search URL generated from the FPDS website specifying query parameters for federal contracts. To generate it, users can perform an advanced search on the FPDS platform, specifying criteria such as date ranges, contracting agencies, vendor names, contract types, and other relevant filters. Once the desired search parameters are set, the FPDS website provides a URL that reflects these criteria.

By passing this URL to the `fpds_scrape()` function within the package, the scraper programmatically accesses the FPDS search results. The function handles pagination automatically, iterating through all available result pages to ensure a complete data collection. It employs web scraping techniques to extract pertinent information from each contract record, such as contract identifiers, award and completion dates, obligated amounts, vendor details, and other covariates specified in the search.

The package is designed to handle medium-sized datasets efficiently. It includes features for error handling and rate limiting to comply with the FPDS website's usage policies and to mitigate issues related to network interruptions or server responses. The scraped data is compiled into a tidy data frame in R, facilitating immediate analysis using standard statistical tools and packages within the R ecosystem.

The `fpdsScraper` package is open-source and available for installation from public repositories. Its accessibility encourages collaboration and enables other researchers to replicate our methodology or adapt the tool for their own data collection needs. By simplifying the data retrieval process from FPDS, the package lowers barriers to entry for researchers, journalists, and policymakers interested in federal procurement data, especially those without massive computing resources, thereby promoting greater transparency and accountability in public spending.

Data Records

In the data we provide, each row corresponds to a *contract action*; each contract may include multiple actions, each of which typically comprises an individual outlay of money. Note that many researchers typically aggregate to the contract level when analyzing procurement data.

We include two versions of our dataset: a complete version in parquet format with 99,057,002 rows and 470 variables, totaling more than 75 GB, and a reduced version in CSV format sized at approximately 25 GB (1.5 GB compressed) and containing only the following variables: unique identifier, Contracting agency, NAICS code, dollar value, ZIP code of the place of performance, congressional district of the place of performance, state of the place of performance, date the contract was signed, the FPDS URL of the contract, and economic indicators for whether the vendor is a small business, is woman-owned, or is minority-owned.

We provide both files in a single Figshare repository (https://doi.org/10.6084/m9.figshare.28057043)[33].

File organisation and formats. The complete corpus, distributed as 589 Parquet shards in the `fpds_data` directory of the FigShare bundle, contains every contract-action record signed between October 1957 and November 2024. Each shard holds the actions for a single calendar month and is named `YYYYMM.parquet` (for example, `195710.parquet` and `202411.parquet`). Collectively, the shards comprise 99,057,002 rows and 470 columns, occupying about 75 GB on disk with Snappy compression. A single row represents one contract action (either a base award or one of its modifications) so contracts with multiple modifications appear in multiple rows. Column names retain the hierarchical paths used in the FPDS Atom XML feed: a dot indicates a nested level and an at-sign marks an attribute (e.g., `content.ID.ContractID.PIID`, `content.PSCCode.@description`, and `content.agencySpecificElements.NASASpecificElements.COTRName`). Fewer than five percent of the variables are agency-specific and therefore sparse.

Row counts vary greatly across shards: early months may contain only a handful of legacy back-fills (for instance, two rows in `195710.parquet`), whereas recent months typically hold between 150,000 and 250,000 actions. For quick starts, a reduced flat CSV (~25 GB) is also provided; it aggregates all months into a single file and retains eleven frequently used fields—contract ID, signing date, agency, NAICS, obligated dollars, place-of-performance ZIP code, state, and indicators for small, minority-, or women-owned vendors.

To assist researchers who prefer to work at the contract (rather than contract action) level, we provide an aggregated derivative of the corpus together with the Python script that creates it. The script,

`aggregate_fpds_contract_level.py`, streams each monthly Parquet shard, groups rows by `content.ID.ContractID.PIID`, sums numeric fields (e.g., `content.DollarsObligated`) and keeps the first non-null value of categorical variables, then writes a single file `fpds_contract_level.parquet` that contains one row per contract. The script is available in the project's public GitHub repository at https://github.com/aaronrkaufman/FPDS_replication/blob/main/Aggregation/aggregate_fpds_contract_level.py, alongside all replication code; users can adjust the aggregation rules by editing the indicated configuration block at the top of the script. The contract level dataset is available through figshare.

Variable dictionary. A companion variable dictionary corresponds to the Parquet shards. For each of the 470 variables, it records the variable name, definition, type, and, where applicable, the list of valid values, as well as a plot showing that variable's availability over time. This document provides essential reference information for interpreting the dataset and selecting columns for analysis.

Data Codebook. To produce the aforementioned codebook, we developed an LLM-based parser to segment the existing codebook into a machine-readable format, which we store as a JSON file. From this parsed codebook, we remove the variables we exclude from our dataset and manually update the variables we produce as combinations of existing variables. Finally, for each variable, we calculate the monthly proportion of erroneous or missing values and plot that proportion, embedding that visualization into the codebook.

Descriptive Statistics of the FPDS Dataset. The data dictionary, as well as all code required to reproduce the analyses and figures, is openly available at https://github.com/aaronrkaufman/FPDS_replication. Figures 1, 3, and 4 can be replicated at https://github.com/aaronrkaufman/FPDS_replication/blob/main/analzye_fpds.R and Fig. 2 at https://github.com/aaronrkaufman/FPDS_replication/blob/main/Figure2.py.

Figure 1 shows the volume of records in our dataset. The top panel shows the log count of contract actions; the bottom panel shows, in log dollars, the total value of contract actions per year. Both show a strong upward trend from roughly 1979 to 2004; the count of contract actions continues to increase after that, while the dollar value plateaus.

One row in our dataset represents a contract action (rather than a contract), and most contracts consist of multiple actions. Figure 2 shows that most contracts have only a handful of associated actions, often fewer than 10, though a small number of large contracts have one hundred or more actions.

Figure 3 disaggregates the top panel of Fig. 1, examining twelve selected agencies' contracting patterns over time. While each agency sees an overall increase in the number of contract actions, the extent of that increase varies considerably. The Drug Enforcement Administration (DEA) and Department of Defense increased greatly, especially post-9/11, while the Environmental Protection Agency (EPA) and Internal Revenue Service (IRS) stayed largely steady from the 1990s to the present day.

As the FPDS dataset includes rich geographical variables, both about the vendor and the place of performance, we can map the number of contract actions performed across the US. Figure 4 shows the log count of contract actions by congressional district from 2012 to 2022, corresponding to the 113th to 117th Congresses. Districts with fewer contract actions are in light yellow; darker districts indicate a higher density of contract actions performed there.

Technical Validation

We conduct two validation tests of this dataset. First, we conduct an audit of random contract actions to ensure that the details of those actions in our dataset match those on the FPDS website, confirming that the data was scraped without error. Next, we compare the descriptive statistics from[28]. We find substantially more contracts than the prior authors do, prompting a more detailed comparison between our contract counts, contract counts from USASpending, and from[28]. We conclude that differing contract totals stem from screens on contracts that are not documented in the replication file, but which could, in principle, very likely be reused on our data due to the similarity of our contract totals with what is available on USASpending.

Manually Auditing Random Contracts. Our first and simplest validation analysis tests whether our scraper collected data from the FPDS website accurately. First, we randomly sample 250 contract actions from our dataset and search for them on the FPDS website by contract ID. We then manually cross-check the contract details to confirm that they match the covariates in the dataset. In all 250 cases, we find 100% fidelity: no variables diverged between our dataset and the online portal in any of the 250 contract actions we hand-checked, confirming that our scraper performed correctly as intended.

Comparing with published results. Our next validation test compares our contract universe with that used in a prominent published study of federal procurement data. We show that our dataset contains no fewer observations than those authors find, and that the distributions of variables of those contracts look similar in the aggregate.

We examine[28], which finds that politically responsive federal agencies (but not independent agencies) allocate noncompetitive contracts disproportionately to battleground states, which are important competitive constituencies. They study this problem using approximately 570,000 contracts priced at $150,000 or more between 2003 and 2015 drawn from USASpending.gov, which is an alternative portal for procurement data, available in PostgreSQL format; FPDS provides data to USASpending.gov, FPDS-NG (https://www.fpds.gov/wiki).

These authors consider a total dataset of 2.1 million contracts; with identical dates (2003-2015) and our broader filters, we observe 19 million contracts. Similarly, we find 1.07 million contracts between 2003 and 2015

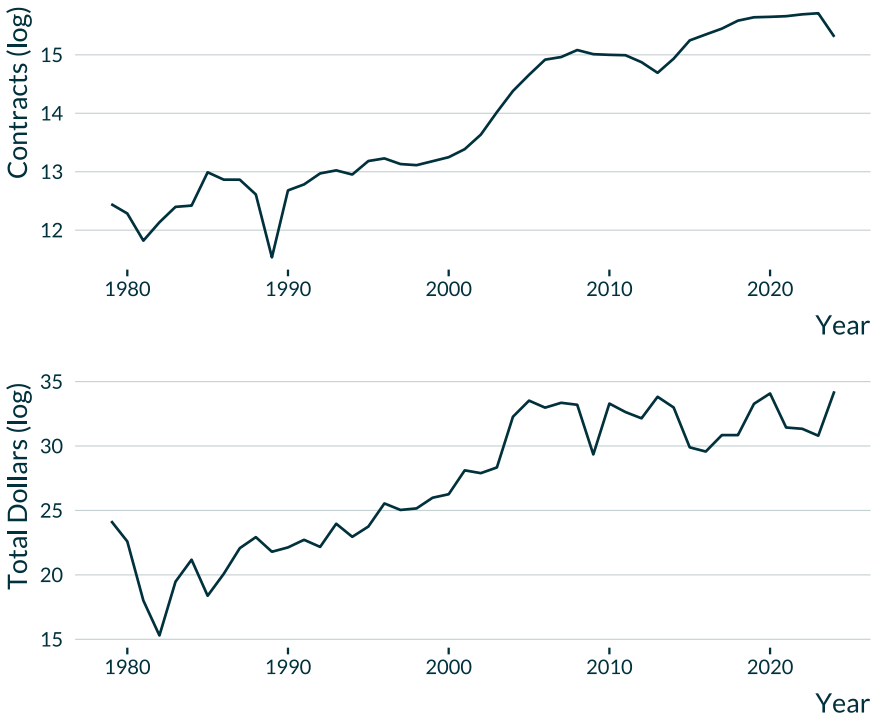

Fig. 1 The log count of contract actions (top) and log total dollar value of those contract actions (bottom) in the FPDS dataset by year.

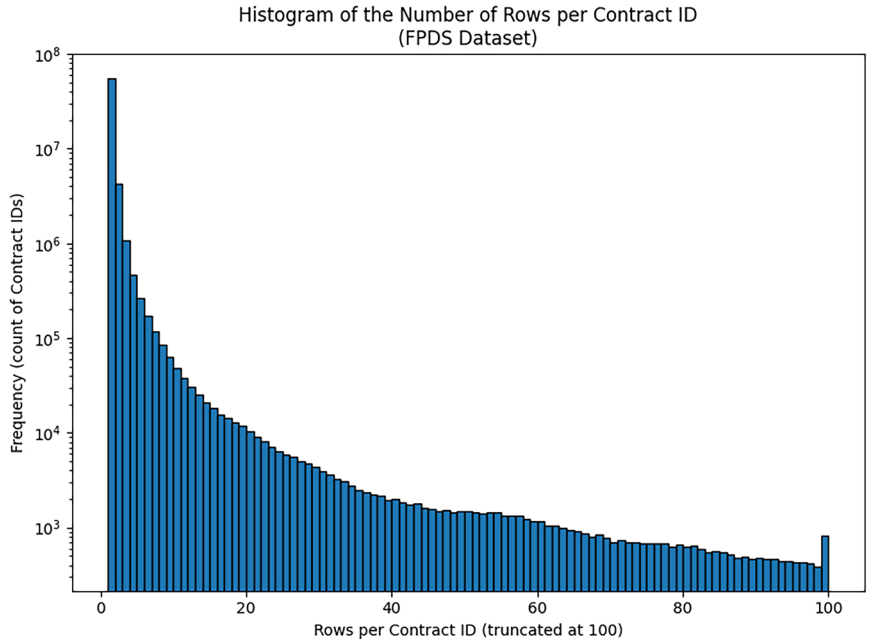

Fig. 2 Most contracts have relatively few associated actions.

worth at least $150,000, nearly twice as many contracts as in[28], with a mean log contract value of 13.647 compared to 13.522 in the original paper. The lack of linking variables in[28] makes a complete assessment of coverage challenging without reverse engineering the linking variables from the replication file.

Comparison with USASpending and contract filters. FPDS-NG (https://www.fpds.gov) and USASpending (https://www.usaspending.gov) both publish the same underlying government feed but apply different processing steps. FPDS-NG keeps every transaction; USASpending collapses all modifications into a single award, drops rows that fail validation against its accounting files, and refreshes on a monthly schedule.

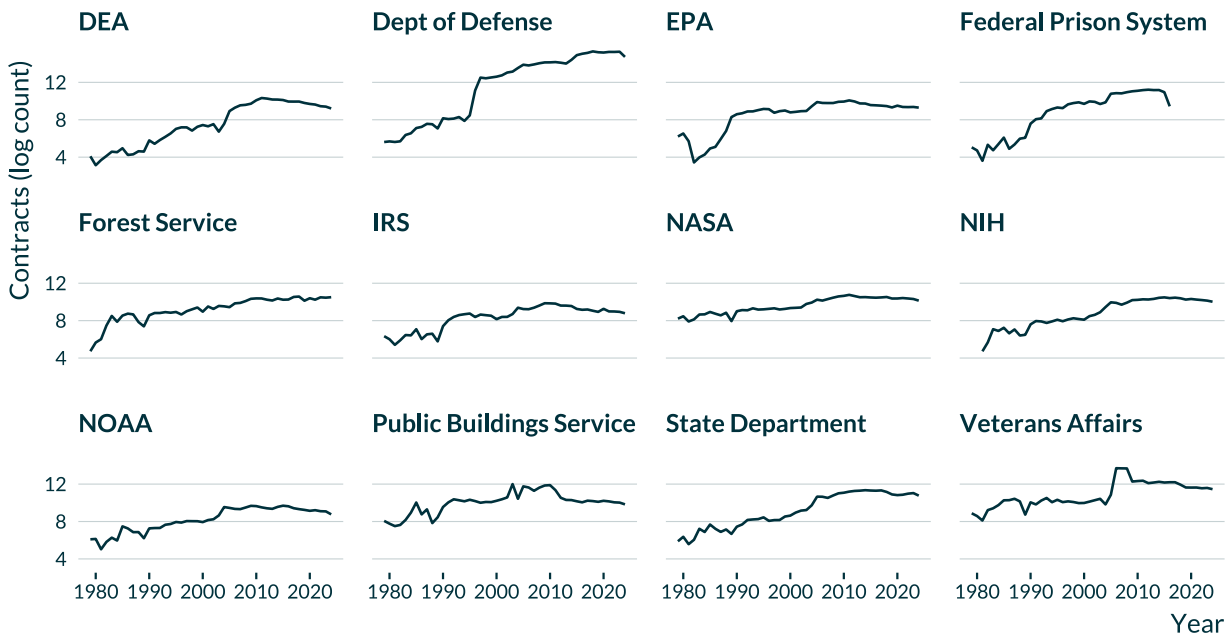

Fig. 3 Log count of contracts by year for selected government departments and agencies, 1979 to 2024.

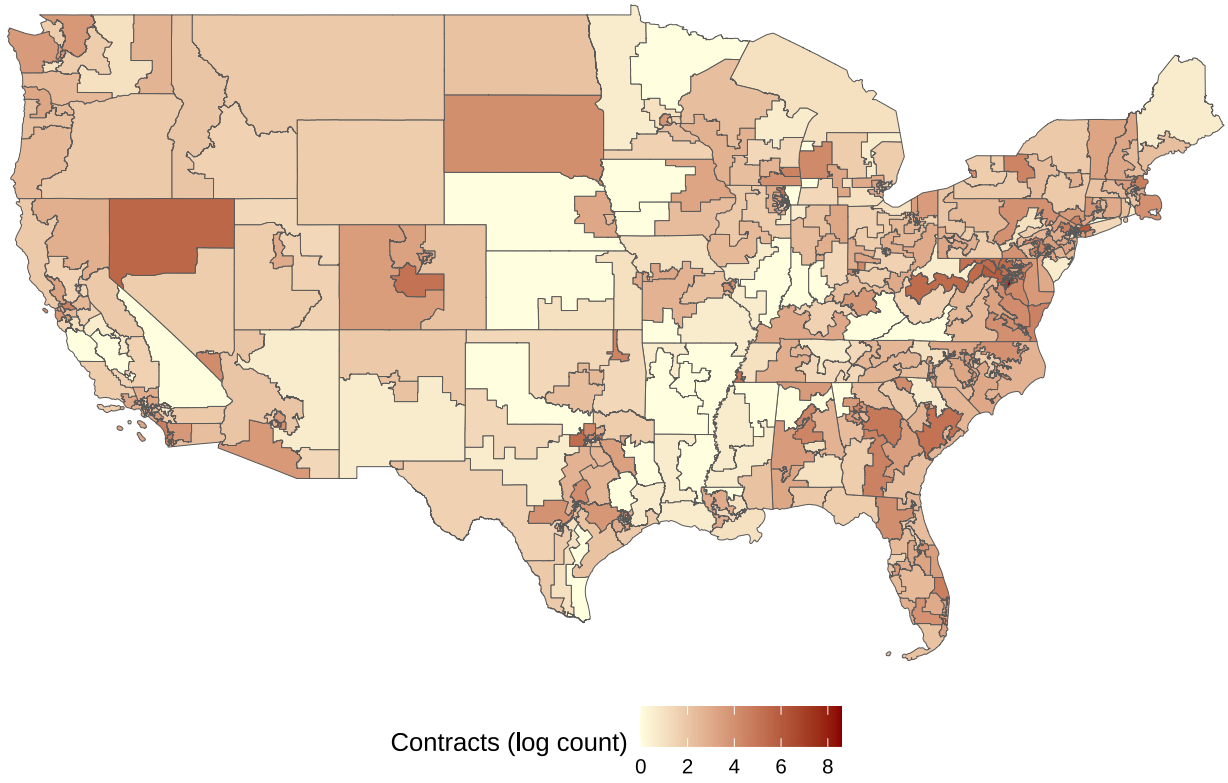

Fig. 4 Log count of contract actions by congressional district, 2012 to 2022.

Table 1 tracks contract counts for fiscal years 2003-2015 after progressively stricter screening rules. Columns list every contract and, in the right-hand column, only those with obligations of at least $150,000. To interpret the resulting counts, note three important facts about procurement contracts. The Procurement Instrument Identifier (PIID) is the serial number assigned by the issuing contracting office when an award is first created; every subsequent modification or delivery order inherits that same root PIID. An Indefinite-Delivery Vehicle (IDV) (often called an "umbrella" contract) is a master agreement that fixes a contract's ceiling price and scope but leaves the actual spending to later task or delivery-order awards, so analysts must decide whether to count

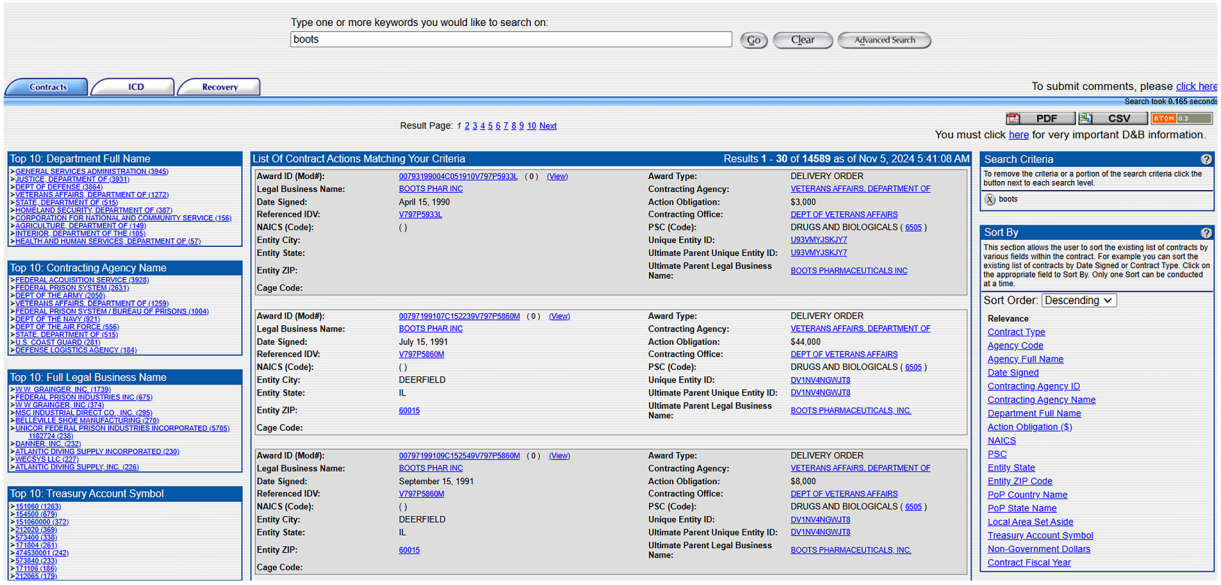

Fig. 5 A search for procurement contracts mentioning the term "boots" on the Federal Procurement Data System website, recorded November 2024.

Counting rule	All contracts	≥ $150k
Raw totals		
FPDS-NG, unique `PIID`	19,029,518	1,066,659
USASpending, unique `PIID`	22,225,530	1,058,521
Published filters		
Keep procurement records only	21,684,657	1,045,751
Retain base awards only (`modification_number` = 0)	21,593,039	967,475
Additional IDV filter		
Drop task or delivery orders for IDVs (`parent_award_id` is null)	10,353,439	441,404

Table 1. Effect of common preprocessing filters on contract counts, FY 2003-2015.

the umbrella, its orders, or both. Finally, the Federal Acquisition Regulation (FAR) is the government-wide rulebook that sets competition thresholds and dictates which data elements must be reported to FPDS; restricting analysis to FAR-covered awards is a common strategy for maximizing data completeness.

Before any bespoke filters, the two sources differ by fewer than 8,200 high-value contracts, or less than one percent. Applying the same published filters (procurement only, base awards only) preserves this near-parity. The large discrepancy between the two sources appears only when task and delivery orders under IDV umbrellas are removed, a step that is implicit but not always documented in many political economy studies. Adding the IDV screen brings the USASpending total close to the 390,000 high-value contracts reported by Dahlstrom *et al.*, confirming that differences arise from undocumented preprocessing rather than different data availability gaps.

Table 1 and the analyses above demonstrate the range of contract counts that arise once different quality screens are applied. Although we cannot match the filters used by Dahlstrom *et al.*, we show that reasonable choices produce totals both above and below theirs. Deciding which screens are appropriate will depend on the specific research task and is left to future users of the data.

Usage Notes

This manuscript introduces the most comprehensive dataset on federal procurement to date. We provide two interfaces for this data: a bulk download option available through figshare, and an R-based webscraper for downloading smaller subsets of data. We believe this dataset holds enormous promise for scholars of public administration and political economy, journalists and practitioners, and taxpayers interested in government transparency. We conclude with some advice for accessing and manipulating the dataset, and we note several points of caution.

Accessing and Manipulating FPDS Data. From more than 1,000 original variables on the FPDS website, we reduce our dataset to 470 variables by reducing redundancy and converting sets of columns into single columns. This allows the dataset to be stored in a relatively compact parquet format. Parquet files are most accessible using the `arrow` library in `R` or the `dask` library in Python, and can be read and analyzed very quickly despite their massive size; for a primer on accessing parquet files in R, see https://r4ds.hadley.nz/arrow. For simplicity, we

suggest that researchers first read the parquet files, identify the columns or subsets of rows they are interested in, and then export that subset to a traditional data storage method.

Cautionary Notes. Despite our best efforts to validate our dataset of procurement contracts, we cannot account for errors in the underlying data. To the contrary, we have significant evidence that these data are flawed. For example, a February 2013 Department of Defense memorandum notes "the difficulties associated with gathering detailed information on contracts [...]. These data difficulties are very serious and range from no reporting of contracts in some jurisdictions, high dollar figure cut-offs for reporting in many, time lags in others, as well as serious variations in reporting across departments and governments, variations in such practices across time, secrecy provisions regarding contracts, and the general inability to identify contractees from publicly available contract information, as well as difficulties encountered separating out policy-related versus administrative or management consulting activities and contracts, among others." We briefly discuss such threats below.

First, while our dataset contains every record available on FPDS, we cannot directly verify that this contains every government contract: some contracts may be redacted from the website for security or by accident, and as many records are entered by hand, they may be susceptible to misinput error. Note that we remove certain columns, especially those containing dates, that have many such errors. For example, some columns contain dates indicating years greater than 3000 or less than 1950. We also discard records dated before 1979 since we have suspiciously few observations.[34] indicates that these data are most reliable after 2001, so researchers interested in precision may choose to begin their analyses in that year.

Second, while we have verified that we have accurately collected data as it exists on the U.S. government's official portal, we cannot account for errors in the original data we collect. We note several categories of such errors. Some contracts, for example, have obligations listed as zero or even negative dollar values. These are clear errors. However, the error rate in these verifiable fields suggests that other variables may have incorrect entries that are more difficult to discern. As such, we suggest that our dataset is most useful when considering broad aggregations rather than small subsets or individual contracts.

Existing work using FPDS data corroborates these concerns and goes through great efforts to mitigate them.[17] find "that not all contracting actions have been entered into FPDS yet. Some contracting officials may not have access to the necessary computer systems or may not have time to submit information to FPDS."[23] note that "there is no systematic way to monitor how contract managers actually input the data; as a result, many records are incomplete," and[15] argue that "data entry mistakes by the agencies' contracting personnel and timeliness of reporting have historically undermined the system's accuracy."

Further, users should be aware of several systematic error mechanisms that we have documented in the data dictionary codebook while auditing the corpus. First, many variables exhibit pronounced time-varying missingness. For example, the 'systemEquipmentCode' field is essentially blank until the late 1980s, after which coverage improves unevenly, while 'countryOfOrigin' shows prolonged plateaus of 50–70% missingness in the early 2000s before stabilising after 2012. Analyses that pool early and late periods without adjustment will therefore attribute spurious secular shifts to what is in fact differential data capture; we recommend restricting inference to after 2001 for such fields or, at a minimum, including year-specific missing-value indicators. Second, several categorical variables "re-set" after statutory or administrative reforms, producing discontinuities that can masquerade as substantive change. Notable break points follow the FAR rewrites in 2004 and 2010 (affecting extentCompeted, typeOfSetAside, and related flags) and the April 2022 replacement of DUNS with UEI vendor identifiers. Where longitudinal consistency is essential, researchers should collapse pre- and post-reform codes into harmonised super-categories and, for vendor-level work, employ the DUNS-UEI cross-walk we provide. Third, the corpus contains non-trivial frequencies of implausible entries—zero or negative obligations, signing dates in the 1800s or 3000s, contract actions recorded months before their parent contracts are issued, and so forth. These almost certainly arise from manual data-entry mistakes and electronic transfer errors. We trimmed patently impossible dates (<1979 or $>$ current year $+1$) and flagged monetary outliers above the 99.9th percentile, yet users analysing micro-level outcomes should still winsorise extreme values or conduct sensitivity checks that exclude them.

Code availability

All code used to clean and analyze this dataset, as well as to produce the figures in this article, is available on GitHub without restriction at https://github.com/aaronrkaufman/FPDS_replication. The FPDS Scraper is available on GitHub at https://github.com/aymane-omari/fpdsScraper.

Received: 26 December 2024; Accepted: 28 July 2025;
Published online: 06 August 2025

References

1. Thai, K. V. & Drabkin, D. A. Us federal government procurement: Structure, process and current issues. In *Public Procurement*, 117–131 (Routledge, 2012).
2. Bates, T. & Williams, D. Do preferential procurement programs benefit minority business? *The American Economic Review* **86**, 294–297 (1996).
3. Shelton, L. M. & Minniti, M. Enhancing product market access: Minority entrepreneurship, status leveraging, and preferential procurement programs. *Small Business Economics* **50**, 481–498 (2018).
4. Smith, C. R. & Fernandez, S. Equity in federal contracting: Examining the link between minority representation and federal procurement decisions. *Public Administration Review* **70**, 87–96 (2010).
5. Celentani, M. & Ganuza, J.-J. Corruption and competition in procurement. *European Economic Review* **46**, 1273–1303 (2002).
6. Burguet, R. & Che, Y.-K. Competitive procurement with corruption. *RAND Journal of Economics* 50–68 (2004).

7. Auriol, E. Corruption in procurement and public purchase. *International Journal of Industrial Organization* **24**, 867–885 (2006).
8. Kang, K. & Miller, R. A. Winning by default: Why is there so little competition in government procurement? *The Review of Economic Studies* **89**, 1495–1556 (2022).
9. Ferwerda, J., Deleanu, I. & Unger, B. Corruption in public procurement: finding the right indicators. *European journal on criminal policy and research* **23**, 245–267 (2017).
10. Dávid-Barrett, E. & Fazekas, M. Grand corruption and government change: an analysis of partisan favoritism in public procurement. *European Journal on Criminal Policy and Research* **26**, 411–430 (2020).
11. Artes, J., Kaufman, A. R., Richter, B. K. & Timmons, J. F. Are firms gerrymandered? *American Political Science Review* 1–21 (2022).
12. Uslaner, E. M. Political trust, corruption, and inequality. In *Handbook on political trust*, 302–315 (Edward Elgar Publishing, 2017).
13. Boly, A., Konte, M. & Shimeles, A. Corruption perception and attitude towards taxation in africa. *Journal of African Economies* **30**, i140–i157 (2021).
14. Haggard, S. & Kaufman, R.*Backsliding: Democratic regress in the contemporary world* (Cambridge University Press, 2021).
15. Fleharty, M. J. & Sharkey, J. J. Too good to be used: Analyzing utilization of the test program for certain commercial items in the air force https://api.semanticscholar.org/CorpusID:110242729 (2014).
16. Berteau, D. J. *et al*. An analysis of department of defense services contract trends, 1990–2011 https://api.semanticscholar.org/CorpusID:166828275 (2012).
17. Halchin, L. E. Hurricane katrina recovery: Contracts awarded by the federal government https://api.semanticscholar.org/CorpusID:150550000 (2005).
18. Kidalov, M. V. Examining reasons for, and impact of, bundled and consolidated contracts at the u.s. department of the navy https://api.semanticscholar.org/CorpusID:167024983 (2012).
19. Kim, Y. W. & Brown, T. L. The importance of contract design. *Public Administration Review* **72**, 687–696 (2012).
20. Spagnolo, G., Decarolis, F., Iossa, E., Mollisi, V. & Giuffrida, L. M. Buyer quality and procurement outcomes: Explorative evidence from the us https://api.semanticscholar.org/CorpusID:168970725 (2016).
21. Hunter, A., Sanders, G., McCormick, P., Ellman, J. & Riley, M. Measuring the success of acquisition reform by major dod components https://api.semanticscholar.org/CorpusID:55720116 (2015).
22. Falcone, E. C., Fugate, B. S. & Waller, M. A. Growing, learning, and connecting: Deciphering the complex relationship between government customer concentration and firm performance. *Journal of Supply Chain Management* https://api.semanticscholar.org/CorpusID:268869552 (2024).
23. Kim, Y. W., Roberts, A. N. & Brown, T. L. Impact of product characteristics and market conditions on contract type: Use of fixed-price versus cost-reimbursement contracts in the u.s. department of defense. *Public Performance & Management Review* **39**, 783 – 813 (2016).
24. Hosek, J., Litovitz, A. & Resnick, A. C. How much does military spending add to hawaii's economy? https://api.semanticscholar.org/CorpusID:150847504 (2011).
25. Howlett, M., Brouillette, C., Coleman, J. C. & Skorzus, R. C. Policy consulting in the usa: New evidence from the federal procurement data system - next generation. *Political Institutions: Bureaucracies & Public Administration eJournal* https://api.semanticscholar.org/CorpusID:157862689 (2016).
26. Thornton, J. & Lecy, J. D. Good enough for government work: When should government hire nonprofits. *PSN: Public Administration (Institutions) (Topic)* https://api.semanticscholar.org/CorpusID:167338498 (2015).
27. Thornton, J. & Lecy, J. D. Good enough for government work? an incomplete contracts approach to the use of nonprofits in u.s. federal procurement. *Nonprofit Policy Forum***10** https://api.semanticscholar.org/CorpusID:208141267 (2019).
28. Dahlström, C., Fazekas, M. & Lewis, D. E. Partisan procurement: Contracting with the united states federal government, 2003–2015. *American Journal of Political Science* **65**, 652–669 (2021).
29. Fazekas, M., Ferrali, R. & Wachs, J. Agency independence, campaign contributions, and favoritism in us federal government contracting. *Journal of Public Administration Research and Theory* **33**, 262–278 (2023).
30. Snider, K. F., Kidalov, M. V. & Rendon, R. G. Diversity governance by convenience? federal contracting for minority-owned small businesses. *Public Administration Quarterly* **37**, 393 (2013).
31. Arnav, A. M. Sustainable society 2020: The case of ethnic preferences. *Educational Research* **11**, 1–3 (2020).
32. Jaipaul, C. Update required: How improvements in open data can promote transparency, accountability, and collaboration in federal procurement. *Public Contract Law Journal* **53**, 433 (2024).
33. Omari, A., Alansari, N., Libgober, B. & Kaufman, A. R. A comprehensive dataset of us federal procurement (1979–2023) https://doi.org/10.6084/m9.figshare.28057043 (2025).
34. Potter, R. A. Buying evidence? policy research as a presidential commodity. *Journal of Politics* (2024).

Acknowledgements

The authors acknowledge Muataz Al Barwani and NYU Abu Dhabi's Center for Research Computing for their advice, encouragement, and assistance. This research was supported by NYU Abu Dhabi's Center for Interdisciplinary Data Science and Artificial Intelligence. All remaining errors are our own.

Author contributions

A.K. conceived the project, performed the analysis, and produced visualizations. N.A. built the data collection pipeline, collected the dataset, developed the data repository, and wrote the data collection section. A.O. conducted initial data collection, wrote the R package, and performed analysis. All authors wrote, reviewed, and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.R.K.