To: Taylor Corbett

From: Doug Hummel-Price

Date: October 1st, 2019

Re: Collecting and Cleaning NASA Contract Expenditures for Fiscal Years 2005 to 2019

Executive Summary

This memo discusses the collection, cleaning, and wrangling of data from NASA regarding the amounts and locations of all its contracts for the fiscal years 2005 through 2019. This document should provide all the necessary information to be obtain and similarly clean the data for your own analyses.

Source

I obtained the data sets from the NASA Procurement Data View (“NPDV”).[[1]](#footnote-1) By selecting a single fiscal year, leaving all other options at their most general level (usually "All"), and submitting the query, a user can download an .xls file containing NASA data from that fiscal year. I did this for each of the 15 Fiscal Years available, FY05 - FY19. NB: it takes anywhere from 5 to 30 seconds for the page to populate correctly. Wait until the number of contracts is shown before downloading the Excel file; otherwise, the dataset will be missing significant numbers of contracts. Further, the webpage includes several caveats that delineate the specific types of data included.

License

The NPDV site contained the following language:

Individuals using archive data agree by virtue of downloading this information that they will not misrepresent the data provided. Misrepresentation of the data includes both intentional and accidental omission of rules, processes, exclusions or other guidance that is used to generate legitimate information. A few examples of misrepresenting data include creating a report that purports to show socio-economic information but which contains none of the requisite Small Business Administration rules and exclusions or identifying a vendor as being small or large without benefit of the business process and rules regarding size determination, vendor representations, vendor certifications, etc.

I interpret that language, combined with the fact that the data is budgetary federal data that is by definition for public use, to mean that the data can be used freely for any non-fraudulent purpose.[[2]](#footnote-2)

Collection Methods

The source website said the following regarding the collection mechanism:

Government agencies are responsible for collecting and reporting data on federal procurements through the Federal Procurement Data System–Next Generation (FPDS-NG).

Contracting Officers (COs) must submit complete reports on all contract actions, as required by the Federal Acquisition Regulation (FAR).

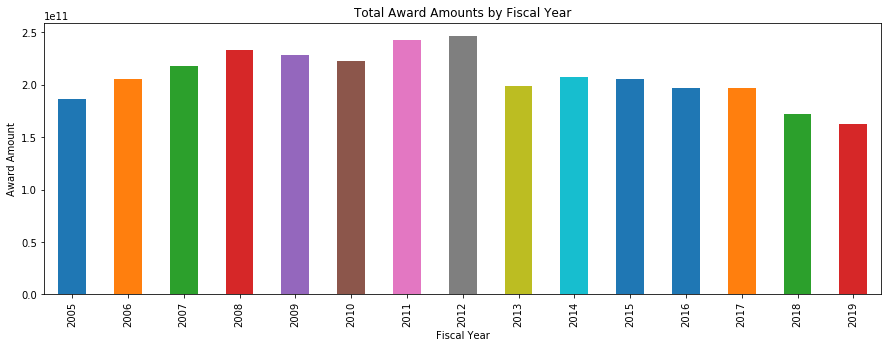
Bias/Sampling

There is no reason to believe that the data suffer from any particularly obvious biases or sampling issues. Given the collection mechanism, there likely is a small but potentially significant amount of error in the data from misreporting due to human error.[[3]](#footnote-3)

Data Cleaning Method and Identified Issues

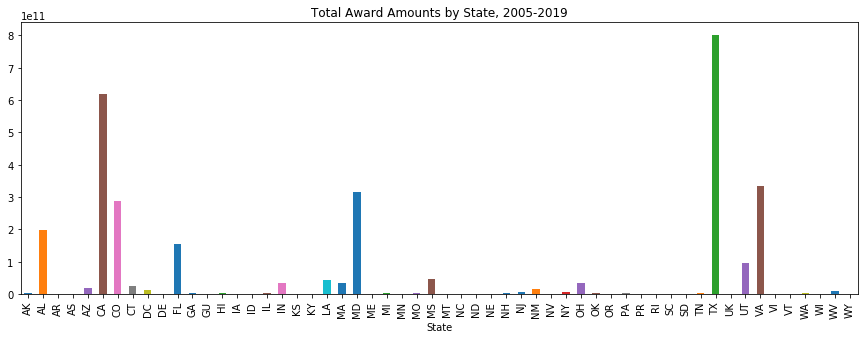
The original data files were not easily read into Python. I solved this by manually opening each file in Excel and resaving it explicitly as a .csv file, which fixed the problem. I cycled through each csv, dropping all unneeded variables, adding a fiscal year variable, and then saving the remaining data to a master data frame that will include all 15 Fiscal Years. Minor cleaning and wrangling included shortening variable names, dropping rows with missing values, converting the award strings into floats, and creating a State-District variable.

I looked at the following diagnostic plots to identify any issues. Award amount is the only quantitative variable, so scatterplots are not possible. The Year variable is purposefully kept categorical to avoid inadvertently applying continuous regressions or ML techniques. The Total Amounts by Fiscal Year bar graph shows that total contract awards peaked in 2012 and have been declining since. This does not immediately strike me as a problem, though deeper domain knowledge is required to be certain.



Note: I also looked at this as a line chart, but the year labels disappeared from the x-axis. As the purpose of this plot is diagnostic, I opted for a bar chart that explicitly labeled the years.

I looked at the following plot, award amounts grouped by state to identify any issues that may make themselves known:



The plot shows significant disproportionate spending in Texas, California, Virginia, Maryland, Alabama, Florida, and Colorado. Given the locations of NASA facilities across the country, these numbers make sense.

EDA did not identify any significant issues, although several data points appeared to have negative award amounts. I’ll need to do additional research to determine whether these are errant or indicative of paying pack overpayments or something similar. Regardless, these data points will likely be dropped for final analysis. The max award amount is also suspiciously large at $27 billion- this is possible though highly unlikely since NASA’s entire budget tends to be less than that.

Descriptive Statistics

**The Descriptive Statistics for Four Categorical Variables**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Contractor** | **NASA Center** | **State** | **State-District** |
| **count** | 406559 | 406559 | 406559 | 406559 |
| **unique** | 31846 | 13 | 56 | 513 |
| **top** | Caltech | NSSC | CA | MD-05 |
| **freq** | 12623 | 135177 | 76689 | 16756 |

**The Descriptive Statistics for**

**Contract Awards in Thousands of Dollars**

|  |  |
| --- | --- |
|  | **Award** |
| **mean** | $7,682 |
| **std** | $232,905 |
| **min** | -$33 |
| **25%** | $16 |
| **50%** | $90 |
| **75%** | $404 |
| **max** | $27,378,658 |

Data Dictionary

**Award** – The total contract amount award for the given contract. The number is provided in US Dollars.

**Fiscal Year**- The US Federal Government Fiscal Year encompasses October 1st through September 30th. E.g., Fiscal Year ’19 is from October 1st, 2018 through September 30, 2019. Possible values are 2005-2019, inclusive.

**Contractor-** The name of the business awarded the contract, with the city and state in brackets at the end.

**NASA Center-** The NASA location that is associated with the contract award.

Possible Values:

'KSC - Kennedy Space Flight Center',

'AFRC - Armstrong Flight Research Center',

'LaRC - Langley Research Center',

'GRC - Glenn Research Center',

'GSFC - Goddard Space Flight Center',

'JSC - Johnson Space Flight Center',

'MSFC - Marshall Space Flight Center',

'ARC - Ames Research Center', 'SSC - Stennis Space Center',

'HQ - Headquarters Acquisition Division',

'NSSC - NASA Shared Services Center',

'NMO - NASA Mgmt Office at JPL',

'APL - NASA Management Office at APL'

**State**- The state in which the contractor’s business address is located. There are 56 possible states because this includes US Territories like Puerto Rico and Guam.

**State-District**- The state plus the congressional district in which the contractor’s business address is located.

1. <https://prod.nais.nasa.gov/cgibin/npdv/adhoc.cgi>; The [Federal Open Data Policy](https://project-open-data.cio.gov/policy-memo/#c-ensure-information-stewardship-through-the-use-of-open-licenses)states: “Agencies must apply open licenses, in consultation with the best practices found in Project Open Data, to information as it is collected or created so that if data are made public there are no restrictions on copying, publishing, distributing, transmitting, adapting, or otherwise using the information for non-commercial or for commercial purposes.” [↑](#footnote-ref-1)
2. <https://project-open-data.cio.gov/open-licenses/> [↑](#footnote-ref-2)
3. The [OPEN Government Data Act](https://www.datacoalition.org/policy-issues/open-data/open-government-data-act/), passed in January 2019 should help reduce the amount of human error due to automation and machine-readability requirements from the Act. This won’t affect these data sets, but future fiscal years might be affected. [↑](#footnote-ref-3)