

United States Federal Spending in Georgia

Seth Brice
February 5, 2026

Data Mining M1 Project Proposal

I. DATASET DESCRIPTION

The data set used in this study comes from "USASPENDING.gov", a website that collects and holds records of awards given to each state since 2008 [1]. It breaks down awards into multiple categories, including contracts, contract IDVs, grants, loans, direct payments, and other award types (Figure 1). Each award has several pieces of data giving context to the award, including the recipient name, awarding agency, amount granted, etc.

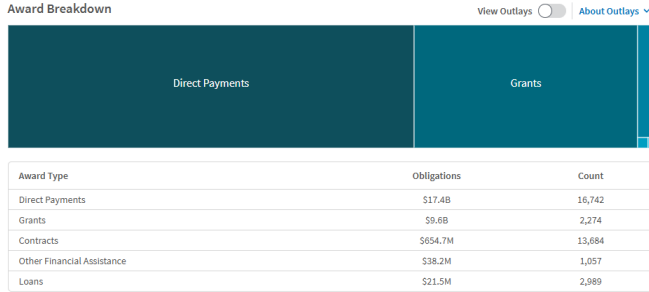


Fig. 1. A graph from "USASPENDING.gov" detailing the sum of all transaction types made to the state of Georgia from fiscal year 2008 to 2026.

This study specifically uses the web page that contains records of all federal awards granted to Georgia. This study only takes into account awards in 2025 totaling over 20,000,000 dollars, as this is the most relevant area for considering current spending. The primary attributes of each data point that will be examined are the recipient name, the awarding agency, the award type, the award subtype, the award description and the obligated amount (Table 1). The obligated value of the award was chosen over the outlay amount due to a significant amount of missing outlays present in the dataset.

While the obligated amount may reduce clarity on where federal money is going, it will still provide context into the interests of the federal government and their spending habits. There may also be a quality issue in the description of each award, as some are more clear than others, while some barely provide a description at all. For this reason, the description is not as considered for analysis as the rest of the data, however it is still provided to give context to each award. After sectioning out the most important portion of time, and carefully considering which pieces of data are most relevant to the study, the data set used contains 1,478 rows of entries, each with 6 columns of data.

transactions	
transaction_id	string
recipient	string
awarding_agency	string
award_type	string
award_subtype	string
description	string
obligations	float

TABLE I
DATA SCHEMA DESCRIBING THE BREAKDOWN OF EACH TRANSACTION AND ITS RELEVANT DATA. THIS STUDY PRIMARILY DEALS IN STRING VALUES, WITH THE EXCEPTION OF OBLIGATIONS, WHICH ARE RECORDED AS A FLOAT VALUE.

II. DISCOVERY QUESTIONS

This data mining project aims to answer the following questions about U.S. federal spending:

- What kind of programs and businesses does the government fund the most within the state of Georgia?
- What businesses or institutions in Georgia benefit the most from federal spending?
- What does the distribution of funds to these companies look like?

In a time where U.S. citizens desire to know more than ever where their tax dollars are being used, answering these questions is an important task. It is vital to provide clarity and rationale for the average tax payer that their money is going to secure areas that benefit and strengthen their community. Unfortunately in the past few years, waves of questionable and fraudulent programs benefiting off of government subsidy have been brought to light, causing restlessness and conflict in America. Examining the state of Georgia should either provide reassurance to its citizens that their taxes are still primarily going to good places, or will expose an underlying problem within the state in need of correction.

III. PLANNED TECHNIQUES

To begin the process, each transaction made within the relevant time period will be recorded within a data matrix created using Python code (Figure 2). Once the data has been

```

transactionMatrix = [
    ["FAB62506C6450", "Lockheed Martin Corp", "Department of Defense", "Contract", "Definitive Contract", "7,261,932,653.67"],
    ["25956A30AP", "GA Department of Community Health", "Department of Health and Human Services", "Grant", "Block Grant", "12,849,962,570.80"],
    ["SLFRP1829", "State of GA", "Department of the Treasury", "Direct Payment", "Direct Payment for Specified Use", "4,853,511,855.92"],
    .
    .
    .
]

```

Fig. 2. Sample data matrix demonstrating what the matrix will look like when it is recorded within the python script. This example demonstrates a recorded contract, a grant, and a direct payment.

consolidated into a table, two primary data-mining techniques will be used to analyze it. Any abnormal transactions would most likely be one time transactions made for specific purposes under specific context, and since this study cares about the most common uses of federal aid within Georgia, it is important to single out these irregular transactions. First, anomaly detection will be employed to further prune any values that are significantly higher or lower than the rest of the set. To do this, the IQR method will be used in context of the total obligations directed to the state of Georgia. Obligations found to be an outlier will not be considered for categorical analysis, but will still be highlighted as a transaction made from the federal state to Georgia.

After the outliers are isolated from the data, classification will be used to sort the data by recipient from highest to lowest frequency. The study aims to determine the highest frequency of data by category, and since there is only one numeric metric the study analyzes, that being the total obligations, classification works best to accomplish these goals. For this, a categorical split decision tree will be created, with three main layers of sorting (Figure 3). The tree will first divide the type of award, which will consist of four branches: contracts, grants, direct payments, and loans. The second layer of branches will represent the awarding agencies responsible for each transaction. Each set of sub-branches will be sorted from left to right by frequency.

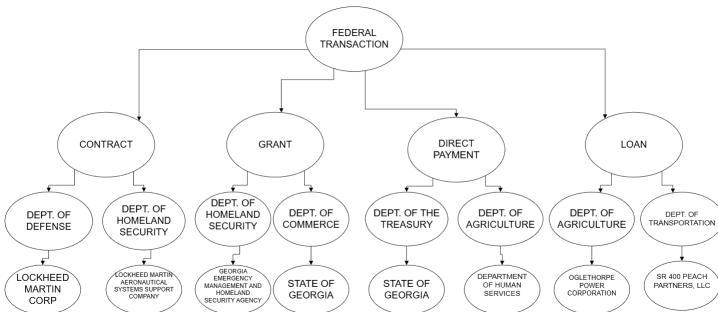


Fig. 3. Sample flowchart representing the classification tree that will be used to determine the frequency of each business which the federal government has made a transaction to within Georgia. The flowchart splits into three layers: award type, awarding agency, and award recipient.

Once the awarding agencies are sorted, the final set of branches will represent the recipient, sorted from left to right by frequency. The total frequency of each agency will be

calculated and used to answer what businesses in Georgia the federal government funds the most. The final step will be to calculate the total funding for each recipient to determine which ones have received the most federal aid within 2025.

IV. TIMELINE

After the proposal is completed, the planned timeline will take into account three major milestone dates: M2 on March 5, M3 on April 2, and M4 on May 3. For M2, the goal is to completely implement the first data mining technique, which being anomaly detection, and to begin to implement the classification tree. To accomplish this, preprocessing and data transformation will need to be complete before the data can be mined. The first half of M2 will be focused on accomplishing these tasks, along with setting up the data matrix within Python. At the same time, the creation of the anomaly detection script will begin. The goal is to complete the initial setup of the data in the first two weeks, then complete and record the anomaly detection in the second two weeks before March 5. After M2 is completed, the next major goal for M3 will be to implement the classification tree and use it on the data matrix. The plan is to dedicate the first half of the milestone to scripting and debugging and to use the second half to run the data matrix through it and have a visualization of the resulting tree ready for interpretation. The last milestone, M4, will be dedicated to finalizing the interpretation and compiling it in the final report, ready to be presented to the wide audience. There will be 2 anticipated challenges throughout this process. The first is the time it will take to transfer the data into a data matrix for M2. There are over 1400 entries, which will be time consuming to accomplish. Despite this, by doing this in tandem with scripting the anomaly detection, and by pacing the creation in the span of two weeks, the matrix creation will be ready in time to be used. Secondly, since the data leans heavily on non integer data pieces, such as names of businesses and federal organizations, the classification tree needs to be able to accurately interpret and class each instance, meaning it has to correctly match multiple string values with one another. There is plenty of room for error here, since it takes more effort and analysis to match strings over integers, and the breadth of data makes this margin of error wider. However, given plenty of time to implement and debug the scripting, the accuracy of the analysis by the program should be close to guarantee.

REFERENCES

- [1] "State Profile: Georgia." *USAspending.gov*, <https://www.usaspending.gov/state/Georgia>. Accessed January 15, 2026.