Procurements of the Canadian Government, Ballooning Military Awards for Fighter Jets*

Robert Ford

December 3, 2024

This paper gathers data on Canadain Federal Procurement Awards, uses a series of visualizations to analyse the data, builds two linear models to compare general government procurements and military procurements. Over the last three years the Canadian military has tripled their spending in an effort to aquire next generation fighter jets and train pilots, while every other government agency has atrophied. This spending concerns every Canadian citizen and often goes undercovered compared to analysis of the governments general budget.

*Code and data are available at: https://github.com/Ford-Robert . Special thanks to Professor Rohan Alexander for motivating this project, providing key support and insight, and demostrating the crucial skills required in modern statistical analysis. Check out his excellent book: https://tellingstorieswithdata.com/

1 Introduction

In recent years, Canada's federal spending landscape has undergone significant transformations, particularly with a marked increase in Military spending. The government has announced substantial new military spending packages, [REF]. This paper will use data collected by the Investigative Journalism Foundation (Foundation, n.d.) detailing Canada's Federal Procurement Awards (Nass, Allen, and Park 2024) to analyse government procurements over the last four years.

Federal procurements involve the acquisition of goods and services by government departments, Buyers, from companies, Suppliers. These procurements are essential for a wide range of government functions from; acquiring defense equipment, building infrastructure, procuring healthcare supplies, technological services and engineering consultants to name a few.

The primary questions this paper aims to address are: What is the government buying? Who are the primary suppliers? How are spending patterns shifting across different departments and over time?

I collected and processed data on well over 300,000 contracts awarded since January 2020. The dataset includes variables such as contract amounts, durations, award dates, Buyers, and Suppliers. We derived additional metrics to facilitate our analysis, such as Baseline Days (the number of days since January 1, 2020) and Duration Days (the length of each contract), and standardized supplier names to ensure consistency.

Some significant findings presented in this paper include: military procurement spending has surged in the last two years. Where the Department of National Defence now dominates federal procurement spending, with expenditures exceeding $56 billion since 2020. This increase in spending is in part thanks to large contracts, including a $11.2 billion contract to SkyAlyne for fighter pilot training (the largest contract in our dataset) and multi-billion-dollar contracts for fighter jets and naval vessels. While other government departments have experienced a notable decrease in procurement spending. For instance, the Public Health Agency of Canada's spending, which was substantial in 2020 due to the COVID-19 pandemic, has declined significantly almost nothing in subsequent years.

These findings are important for several reasons. They provide critical insights into how government spending priorities are shifting, which has implications for economic policy, industry stakeholders, and public transparency. Understanding the concentration of spending in large military contracts can inform discussions on national defense strategies, budget allocations, and the impact on domestic industries.

The paper is structured as follows: Section 2 details the data collection and processing methodologies, including the sources of procurement data and the steps taken to clean and standardize the dataset. Then I present a series of graphs to demonstrate government procurements from various angles to provide a clear picture of spending over the last four years. Section 3 describes the linear regression models used to assess trends in contract amounts over time and

contract duration. Section 4 presents the results of the analysis, highlighting key trends in military and non-military procurement spending. Section 5 discusses the limitations of the study, including data constraints and methodological considerations, and suggests areas for future research. Finally, Section 5 concludes with a summary of the findings and their implications for policymakers, industry stakeholders, and future studies. Additional information about how the data was extracted using web scraping, how the data was cleaned, and how suppliers names were processed can be found in Section A. Model diagnostics can be found at Section C.

By providing an analysis of federal procurement spending, this paper aims to inform readers on how government consumption of private sector goods and services is changing over time. The insights gained from this study can inform policy decisions, contribute to academic discourse, and enhance transparency regarding government spending practices.

2 Data

Data was analyzed through the R programming software (R Core Team 2023), and `RSelenium` (Harrison 2022) with `rvest` (Wickham 2024) were used to support a web scraping script to download all data. Additional packages such as `tidyverse` (Wickham et al. 2019), `knitr` (Xie 2014), `actuar` (Dutang, Goulet, and Langevin 2022), `arrow` (Richardson et al. 2024), `readr` (Wickham, Hester, and Bryan 2024), `stringr` (Wickham 2022), `broom` (Robinson, Hayes, and Couch 2023) and `dplyr` (Wickham et al. 2023) were used to help clean, simulate, analyze, and test the data. The presentation of this data using tables and graphs was made possible by r-packages; `ggplot2` (Wickham 2016), `lubridate` (Grolemund and Wickham 2011), `kableExtra` (Zhu 2024), and `zoo` (Zeileis and Grothendieck 2005). All colours are taken from `RColorBrewer` (Neuwirth 2022)

A web scraping script was used to collect all Federal Award Data from the IJF (Nass, Allen, and Park 2024). The variables this paper focuses are; Contract, Buyer, Supplier, Amount, Award Date, Start Date, and End Date.

- **Contract**: A few words detailing a brief description of what the contract signed was for.
- **Buyer**: The name of the government department who awarded the contract.
- **Supplier**: The name of the company that fulfilled the contract by agreeing to provide the good or service to the Buyer.
- **Amount**: The value of the contract in Canadian Dollars.
- **Award Date**: The day the Buyer awarded the contract to the supplier.
- **Start Date**: When the contract is due to start.
- **End Date**: When the contract is due to end.

The Figure 1 details the first few rows of the IFJ Database:

Contract	Buyer	Supplier	Amount	Award Date	Start Date	End Date
Institutiona...	Public Servi...	All-Brite El...	1,861,961	2022-11-22	2019-11-27	2023-03-31
Scientific s...	Fisheries an...	SGS AXYS ANA...	77,175	2020-03-04	2019-11-28	2020-03-04
Repair of Sh...	Fisheries an...	SEAMASTERS S...	19,755	2020-01-08	2019-11-29	2020-01-08
Operating Sy...	Fisheries an...	THE MATHWORK...	14,403	2020-02-12	2019-11-30	2020-02-12
Miscellaneou...	Fisheries an...	LEICA MICROS...	23,945	2020-01-13	2019-11-30	2020-01-13

Figure 1: Sample of the Data

Additional Derived Variables, displayed in Figure 2:

- **Baseline Days**: Number of days since 1st January 2020.
- **Duration Days**: End Date subtracted from the Start Date to calculate the duration of the contract.

- **Processed Supplier**: The name of suppliers regularized to allow for aggregation and comparison.

Baseline Days	Duration Days	Processed Supplier
1,056	1,220	allbrite
63	97	sgs axys ana...
7	40	seamasters
42	74	mathworks
12	44	leica micros...

Figure 2: Sample of Derived Data

Processed Supplier is required as many companies have slightly different spellings, such as ["General Dynamics", "GENERAL DYNAMICS", "General Dynamics Mission Systems", "GENERAL DYNAMICS MISSION SYSTEMS INC"]. See Section A for more details on how I processed and regularized the suppliers.

The data was then cleaned. First I began by setting each variable to the correct data type. Then I removed any data point where Amount was negative or zero, because I am not sure what a negative amount means in this context and contracts that are not worth anything are not worth looking at. There were a number of contracts which started and ended on the same day, to simplify later calculation I simply added a day to their End Date, figuring that same day contracts are equivalent to one day contracts. Further data cleaning details are available in Section A. After cleaning I was left with 320,175 contracts spanning January 2020 to November 2024.

2.1 Measurement

The data is recorded and published through various government agencies, who use different accounting standards and procedures. The three sources IJF used to collect the federal procurement data are; Canada Buys, Buy and Sell, and Proactive Disclosures (which makes up the vast majority of the procurement data). However, because the data used in this paper is pulled from the IJF's Federal Awards page this data comes exclusively from the Proactive Disclosures. The Treasury Board of Canada Secretariat is responsible for publishing these Proactive Disclosures under the Access to Information Act [REF].

2.2 Analysis

Figure 3 is a break down of total amount awarded by the top 5 biggest spending Buyers. The Other category consists the of the total awards given by the remaining 88 smallest Government

Buyers. It is clear that National Defence dwarfs the procurement budget, spending over $56 Billion on procurements since 2020, making up nearly half of all government procurement.

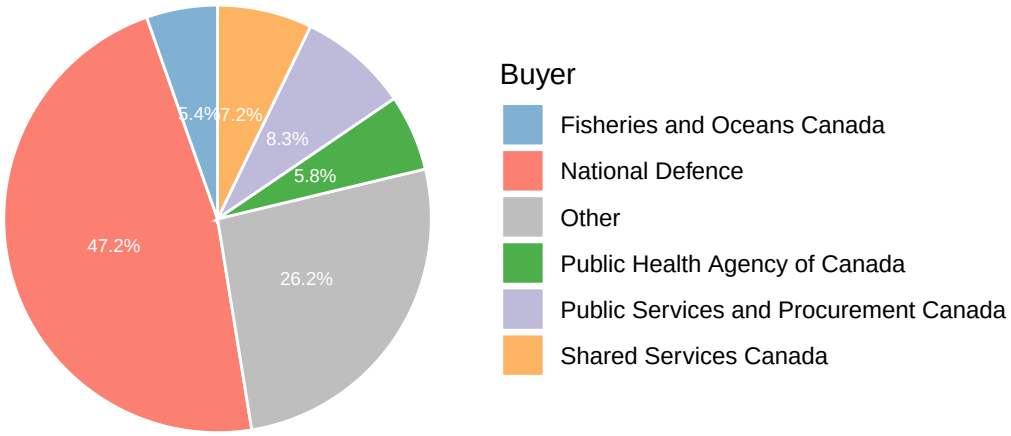

Figure 3: Top 5 Buyers by Total Amount Spent and Others

Figure 4 shows how National Defence spending has taken off in recent years, where is 2021 it had not even crested $5 Billion, but in the last 3 years spending has exploded to $15 Billion a year. From 2021 to 2022, military procurements increased by more than 300%. Though it is important to note that these amounts reflect the amount awarded in any given year, while many awarded contracts last for years. Especially large contracts which can last decades. Meaning the payouts for these lengthy contracts could be amortized over the length of the contract.

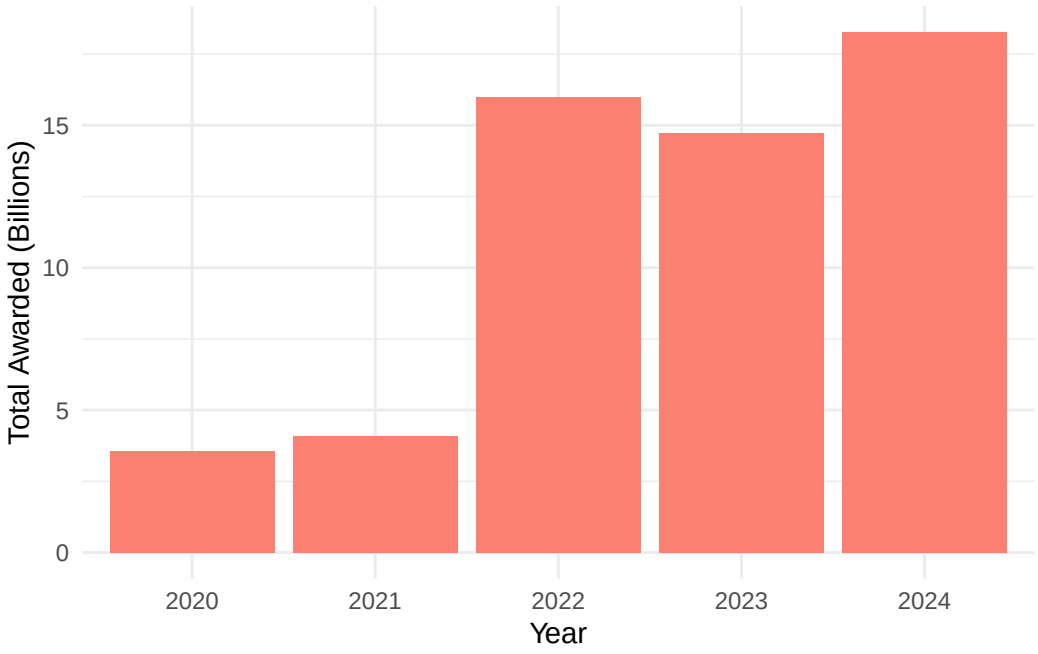

Figure 4: Military Spending Per Year in Billions

Public Services and Procurements Canada (PSPC) is the clearinghouse for the entire federal government. This means they are responsible for managing the procurement process for buyers across the government [REF]. It is important to note that the PSPC sometimes does not publish contracts less than $40,000. Figure 5 shows PSPC spending over time, indicating that general government procurements are taking a hit while military spending balloons.

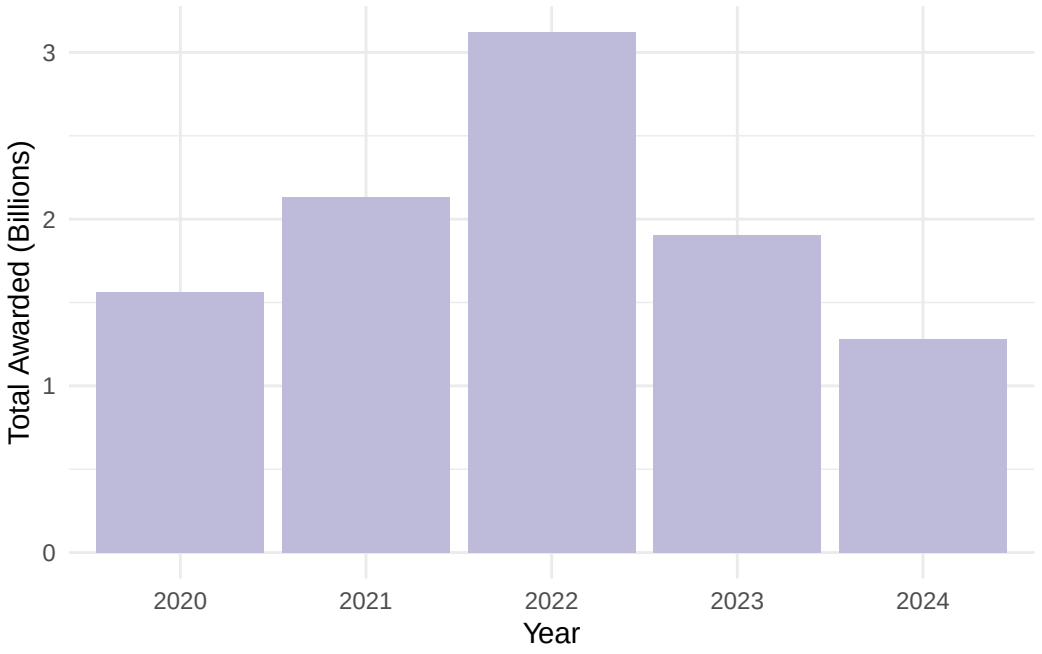

Figure 5: Public Services Spending Per Year in Billions

Figure 6 shows the top five Buyers spending per year since 2020. In the first 2 years military spending held steady and did not dominate the budget by any means. But since then Defence spending exploded not just nominally but compared to the rest of the budget as well. Notably the Public Health Agency of Canada's spending while strong in 2020, decreased to nothing over the following four years. The spending in 2020 is linked to purchases of medical equipment, likely for combating the pandemic. Maybe this over investment has meant that that the Public Health Agency has not needed to invest in equipment in the subsequent years. Though investments across all government agencies, excluding the military, are falling and are lowest in 2024. Note that the 2024 data only goes to November, so more spending may happen in December.

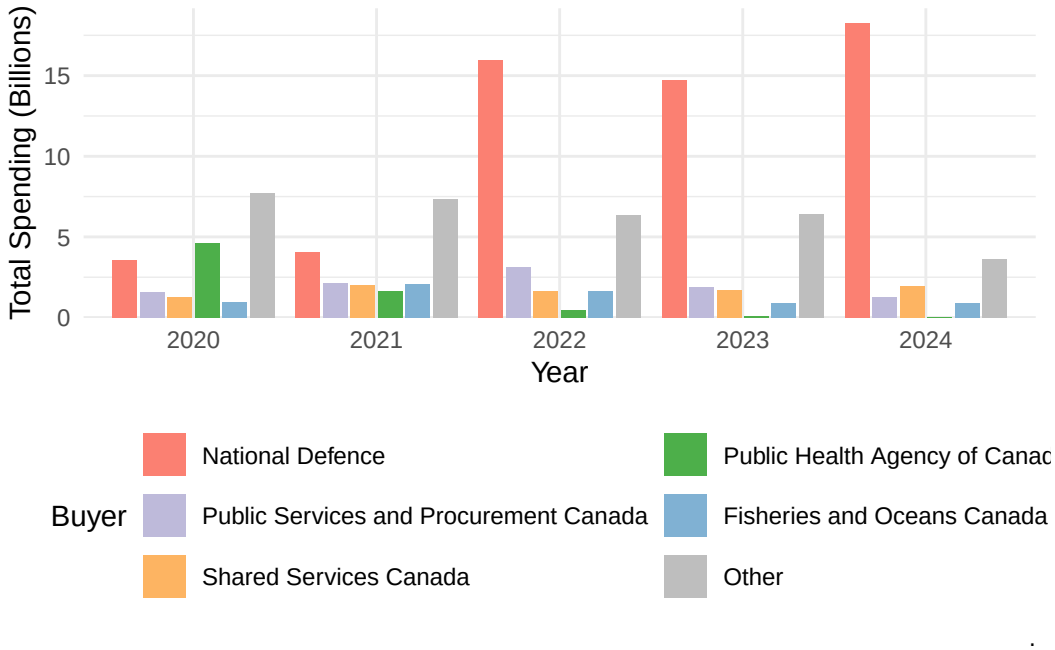

Figure 6: Spending by Top 5 Buyers and Others Per Year

To find out the firms that are benefiting the most through this procurement process we can look to Figure 7. Unsurprisingly, the biggest suppliers are primarily are awarded contracts by the military. Interestingly the top supplier, SkyAlyne, was only awarded one huge $11.2 Billion contract for fighter pilot training services. This is the largest contract in the dataset. Then considering the largest contracts awarded since 2022 to; the F-35 program (Fighter Planes, $4.5 Billion), General Atomics (Airplanes, $1.7 Billion), CAE (Aviation Engineering and Training, $4.4 Billion), Airbus (Airplanes, $3.7 Billion), and Textron Aviation (Aircraft parts, $3.1 Billion) for a grand total of $28.6 Billion. These few large contracts devoted to advanced fighter jets, training pilots, and parts are a significant portion of surge in military spending. Remember military procurement spending since 2022 has been about $15 Billion dollar a year up from $5 Billion in 2021, as per Figure 4. That equates to roughly $30 Billion in extra spending ($10 Billion extra over the $5 Billion year for 3 years), $28.6 Billion of which went to aviation projects.The military also paid Vancouver Shipyards $3.4 Billion in 2024 for Ships and Boat parts.

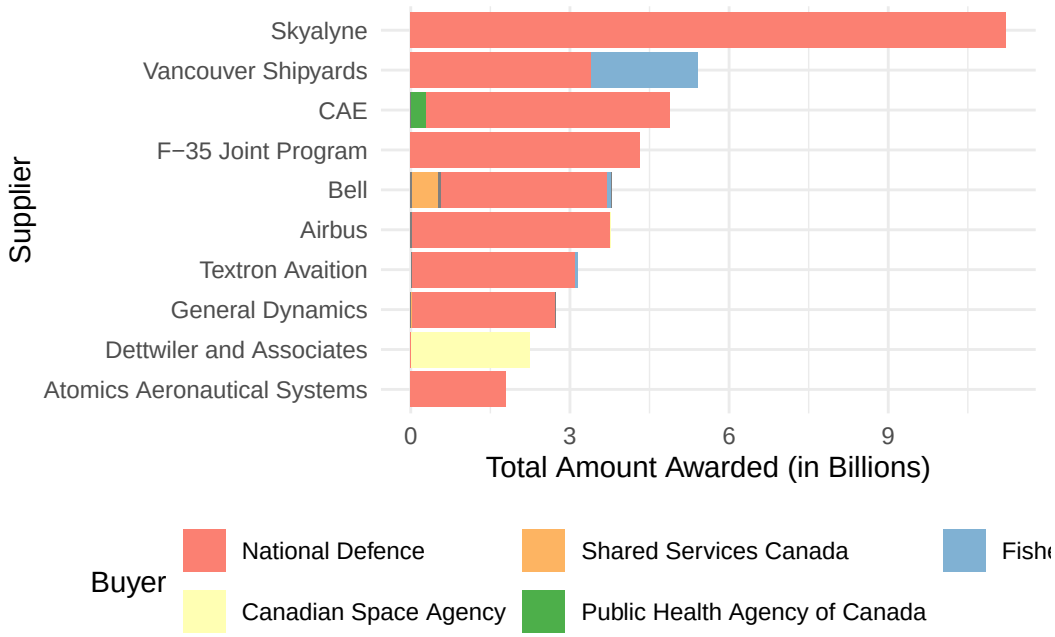

Figure 7: Top 10 Suppliers by Total Amount Awarded

Dettwiler is an engineering firm that provides the Canadian Space Agency with engineering and scientific consulting. Surprisingly 86.3% of the Space Agency's procurement budget goes to Dettwiler.

2.3 Contract Distribution

This graph represents 81.75% of all contracts, indicating that the vast majority of contracts awarded by the government are quite small. Furthermore, it seems that the number of contracts follows a Pareto distribution. There are 102 contracts that are worth $1, this maybe some legal formality.

Though most contracts are small contracts (less than $100,000), the sum of these 246,000 contracts is only worth a measly 7.34 Billion. The SkyAlyne contract itself outweighs all the small contracts.

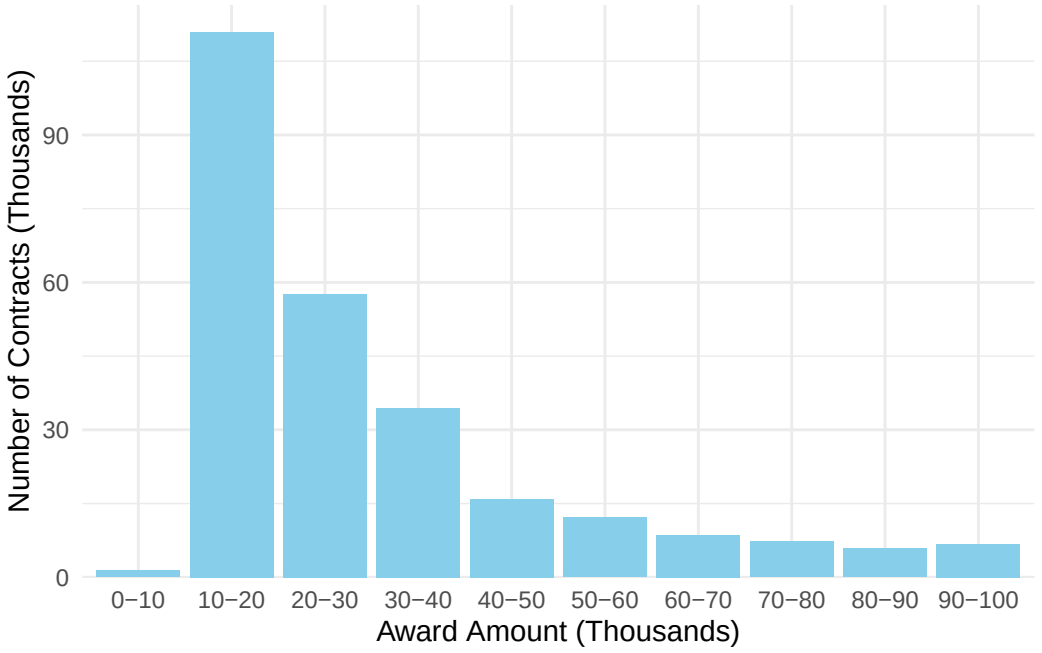

Figure 8: Distribution of Contracts that are less than $100,000

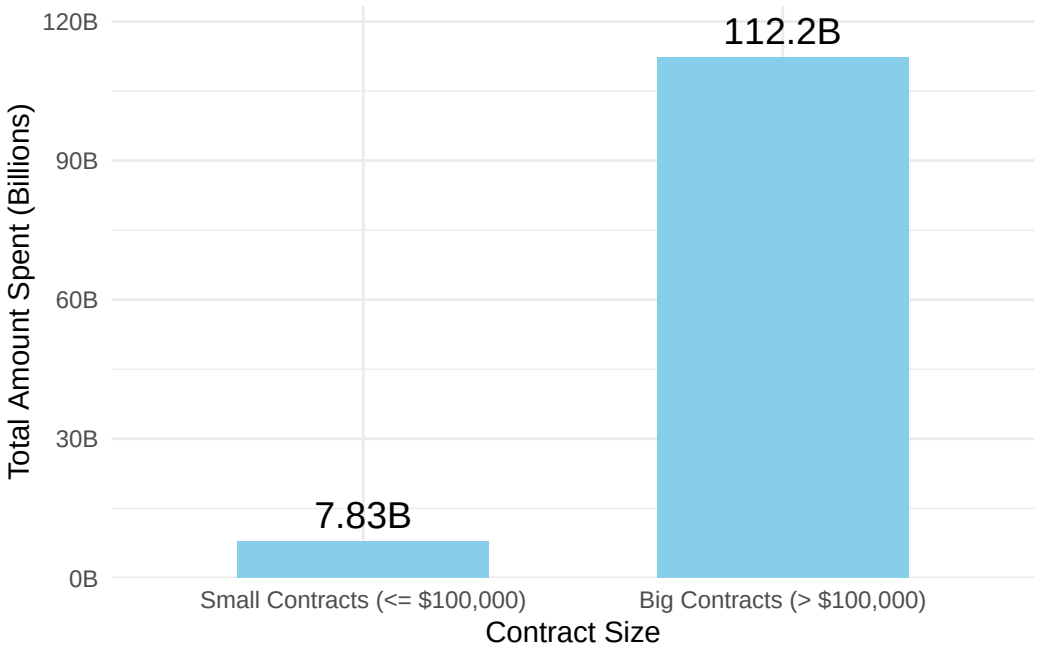

Figure 9: Total Amount Spent on Contracts

3 Model

This paper will use a simple linear regression to better understand the how contract award amounts are affected over time compared to their lengths. To do this the response variable will be Amount, and the two dependent variables will be the number of day since January 1st 2020 (start of the data set) and the length of the contract. The coefficients of these variables will indicate the relative importance of when the contract was issues compared to the length of contracts on determining the size of contracts. I will create two models, one to measure the coefficients for all the contracts excluding military contracts and a model for solely military contracts. This way we can compare to see if military contracts are getting larger because they are relatively longer or because recently they have purchased more expensive equipment.

3.1 Model set-up

All model construction and diagnostics were conduction using R (R Core Team 2023), and the following packages:

- Tidyverse: For data manipulation and visualization (Wickham et al. 2019)
- Dplyr: For data manipulation and transformation (Wickham et al. 2023)
- Car: For advanced regression modeling tools and diagnostics (Fox and Weisberg 2019)
- Caret: For classification and regression training, including model tuning and validation (Kuhn and Max 2008)

Model Formula:

$$\text{Amount}_i = \beta_0 + \beta_1 \times \text{Base Line Days}_i + \beta_2 \times \text{Duration Days}_i + \epsilon_i$$

Variables:

- Amount: The Amount awarded to a contract
- Baseline Days: The number of days since 1st of January 2020
- Duration Days: The number of days that the contract spans

The point of Baseline Days is to measure when the contracts take place, so when Baseline days is zero then it is one of the earliest contracts, as it gets larger that describes more recent contracts.

3.2 Limitations

One significant limitation is the assumption of a linear relationship between the predictors and the contract amount. In reality, the relationship may be more complex or nonlinear, especially over an extended period where economic factors and policy changes could introduce fluctuations not captured by a simple linear trend.

Another limitation is the potential omission of relevant variables that influence contract amounts. Factors such as inflation, changes in government procurement policies, economic cycles, or industry-specific trends can significantly affect contract values. The exclusion of these variables may lead to omitted variable bias, where the estimated effects of the included predictors are distorted because they are capturing the influence of missing factors.

Additionally, the presence of outliers or extreme values, particularly in financial data like contract amounts, can disproportionately influence the model estimates and lead to misleading conclusions.

This model treats all contracts as a homogeneous group. But considering that contracts both for National Defence and general government awards vary widely in scope and cost that would suggest a linear model is not appropriate. A more segmented model may be appropriate. This may be difficult as dividing the data in to groups based on contract size would not be sufficient as there is still a wide variety of contracts that are of similar size. Possibly using the Contract descriptions to divide the data, but this would require more advanced Natural Language processing techniques and in many cases the descriptions are hardly descriptive.

Moreover, the model assumes that the effect of each predictor is constant over time. If the influence of contract duration or time on contract amounts changes due to evolving market conditions or government strategies, the model's fixed coefficients would not reflect these dynamic relationships.

3.3 Diagnostics

The model diagnostics highlight the limitations of these two linear models, see (**diagnostics?**) for relevant diagnostic plots. The Residuals Vs. Fitted plots indicate large clumps in the first quadrant. This indicates a significant bias in the predictions for both models. The QQ-Normal plots indicate that the majority of the residuals follow a normal distribution. However, there is a significant tail off in the end likely due to a large number of outliers which conclusively violates the normality assumption. The Scale-Location plot for the General Model indicates a funnel shape near the origin, indicates the assumption of homoscedasticity is violated. While the National Defence Model has a clump there is no clear pattern so the assumption of homoscedasticity maybe intact. For all predictors across both models their VIF values are very close to 1, indicating no multicollinearity. Both Cook's distance plots show there is a significant number of outliers in both models.

4 Results

The model general contracts indicates that the contract amount tends to decrease over time. Because Baseline days is slightly negative. This value suggests that every day that passes on average the value of non-military contracts is going down by $61.62 a day. This decline could reflect shifts in government spending priorities away from non-military departments. Additionally, as expect Duration Days is positive of the contract amount . This means that longer contracts are more expensive for the government.

In contrast, the model for military contracts reveals that the contract amount increases over time. The positive coefficient for Baseline days suggests an upward trend in military contract values since 2020. In this case the model suggests that the cost of military contracts is increasing by $852.50 every day. This is trend is due to the huge increases in military procurement we observed in Figure 4.

The impact of contract duration on the contract amount is substantially larger in military contracts compared to non-military ones. Where military contracts cost $38,146 more per day of the length of the contract. This is again due largely by the $11 Billion SkyAlyne contract, and the other large long contracts for training and fighter jets.

Table 1: Model Results Summary

Model	Term	Estimate	Std. Error	t-Statistic	p-Value
Excluding Military Contracts	Intercept	-39457.6159	24751.90295	-1.594124	0.1109095
Excluding Military Contracts	Baseline Days	-61.6224	23.80006	-2.589170	0.0096214
Excluding Military Contracts	Duration Days	1341.7462	32.82717	40.873040	0.0000000
Military Contracts Only	Intercept	-5485218.7975	402365.90658	-13.632414	0.0000000
Military Contracts Only	Baseline Days	852.4927	393.51927	2.166330	0.0302892
Military Contracts Only	Duration Days	38145.9891	677.11532	56.336030	0.0000000

Figure 10: Model Results Summary

5 Discussion

5.1 What Was Done

This study conducted an in-depth analysis of Canadian federal procurement data, with a focus on military contracts. The data was collected and processed to ensure accuracy and relevance. Using web scraping techniques, we gathered information from the Institute for Government's (IJF) Federal Award Data. The primary variables included contract descriptions, buyer and supplier names, contract amounts, and key dates such as award, start, and end dates.

To facilitate the analysis, we derived additional variables: Baseline Days, representing the number of days since January 1st, 2020, and Duration Days, calculated by subtracting the start date from the end date of each contract. These transformations were necessary because date variables are not directly suitable for linear modeling. We also standardized supplier names through a Processed Supplier variable to ensure consistency, given the variations in how supplier names were recorded.

Data cleaning was a critical step. Contracts with negative or zero amounts were excluded due to their ambiguous nature. Contracts that started and ended on the same day were adjusted by extending the end date by one day to simplify duration calculations.

For the modeling phase, we employed a simple linear regression to explore how contract amounts are influenced by time and contract duration. Two separate models were constructed: one encompassing all contracts excluding military contracts, and another focusing solely on military contracts. This bifurcation allowed us to compare trends between military and non-military spending.

5.2 State of Canadian Procurements

The analysis revealed significant trends in Canada's federal procurement spending:

Surge in Military Spending: Military spending has dramatically increased over the past couple of years. National Defence emerged as the dominant spender, allocating over $56 billion since 2020. While military spending was under $5 billion in 2021, it escalated to approximately $15 billion annually in the following years. This spike is attributed to substantial contracts for advanced military equipment and services, such as the $11.2 billion contract awarded to SkyAlyne for fighter pilot training—the largest in the dataset. Other notable contracts include multi-billion-dollar deals with Lockheed Martin, General Atomics, CAE, and Airbus for fighter planes and aviation services.

Decline in Other Departments' Spending: In contrast, other government departments have experienced significant reductions in procurement spending. For instance, the Public Health Agency of Canada's spending, which was robust in 2020 due to pandemic-related expenditures, diminished to negligible levels in the subsequent years.

Prevalence of Small Contracts: The majority of government contracts are relatively small. Approximately 81.75% of all contracts are valued under $100,000, with the most common contract size ranging between $10,000 and $20,000. This pattern indicates a governmental preference for smaller contracts, possibly for regulatory reasons. There may be a rule that if a contract is below $20,000 then it requires less scrutiny to be passed, in effect encouraging department to award more smaller contracts.

5.3 Weaknesses

While the study provides valuable insights, several limitations must be acknowledged:

Supplier Name Analysis: The methodology for standardizing supplier names, requires more rigorous techniques. The method used ensured that the top 10 suppliers are accurately aggregated. However, some contracts that these suppliers may have been missed. But beyond the top 10, the names of suppliers not well parsed out, with problems like variations of "incorporated" and that sort of thing, dominating the rest of the suppliers list.

Model Limitations: The linear regression model demonstrated weaknesses, primarily due to its simplicity and the limitations of the dataset. The assumption of a linear relationship between the predictors (Baseline Days and Duration Days) and contract amounts may not capture the complexities of procurement spending, which can be influenced by many things such as; policy changes, economic conditions, and sector-specific dynamics.

Methodological Constraints: The model did not account for external variables such as inflation, exchange rates, or changes in government procurement policies, which could significantly influence contract amounts. Additionally, treating all contracts as a homogeneous group overlooks the diversity in contract types and scopes.

5.4 Next Steps

To address these weaknesses and build upon the findings, the following steps are recommended:

Enhance Data Analysis Techniques: Supplier Name Processing: Implement advanced text processing techniques, such as fuzzy matching, to accurately standardize supplier names and improve data reliability. ADD MORE

Contract Description Analysis: Utilize text mining and content analysis methods on contract descriptions to uncover patterns in procurement objectives and priorities. ADD MORE

Expand Data Collection: Comprehensive Spending Data: Include other forms of government spending, such as in-house expenditures on salaries and operational costs, as this would contextualize the procurement budget in the larger scheme of federal spending overall. How much

of the total federal budget goes into procurements? Are procurements more cost effective than in house spending?

Investigate Specific Trends: Public Health Agency Spending: Examine the factors contributing to the decline in the Public Health Agency of Canada's procurement spending.

Proliferation of Small Contracts: Analyze the rationale behind the high frequency of small contracts, exploring whether this is a strategic decision to support smaller businesses or a reflection of procurement policy regulations.

Refine Modeling Approaches: Advanced Statistical Models: Explore more complex modeling techniques, such as nonlinear regression, time series analysis, or machine learning models, to capture the intricate relationships between variables.

Incorporate Additional Variables: Include economic indicators, policy changes, and other relevant factors to improve the explanatory power of the models.

By undertaking these steps, future research can provide a more comprehensive and nuanced understanding of Canadian federal procurement spending. This further analysis could inform policy decisions, promote transparency, and ensure that government resources are allocated effectively and efficiently.

A Data Details

A.1 Web Scraping

These are step by step instructions on how to Web Scrape the IFJ procurement data page, because as of November 2024 there is no easy/automated way of downloading the 330 pages of data.

Note: It is important that docker, java, python and Firefox are installed on your machine as RSelenium requires these packages. To download Selenium and create a docker image, open the docker application and search for the Selenium package for the browser you wish to use, in this case Firefox.

1. Run a compatible Selenium docker image. You must make sure that the version of Selenium you are using is compatible with Firefox or whichever browser you choose. You must launch docker from your terminal whose directory matches that of the r file's path to 01-download_data.R . You can check if the docker image has successfully been launched using the docker application's nice GUI. It is also important in this step to define the path of the local directory that the data should be saved into, this is shown as USER_PATH in the script.

Terminal Command:

user:canadian_gov_procurement USER_PATH$ docker run -d -p 4444:4444 -p 5900:5900 -v /USER_PATH/canadian_gov_procurement/data/raw_data:/home/seluser/Downloads selenium/standalone-firefox:3.141.59

2. It is possible that the tags used to find the buttons on the web-site may change over time. So visit the IJF procurement data website and right click and inspect the download and next page buttons. This code uses the aria labels to locate these buttons, but there are many ways to identify them using html and css elements.

3. Now the 01-down-load-data.R file can be executed. Be sure to modify USER_PATH in both to download_dir and host_down_load_dir to match the path of you docker and personal computer path.

4. The for-loop simply runs "click download -> click next page" process 329 times, however IJF may add more pages so modify total_pages to scrape the number of pages IJF is offering. In the event of an error mid-way through scraping take note of which page failed to download (page that you are on is printed to the console) and simply set the start of the for-loop. After restarting the loop be sure to verify that the next downloaded page is the correct one, otherwise you will have to restart the process from step 1. While it may be tempting to reduce the Sys.sleep times to download faster (11 seconds delay per loop and 329 loops that's about 1 hour!), I found reducing the time cause errors as the pages were slow to load and thus ruining the process.

5. Done!

A.2 Data Cleaning

I began by converting Amount into a numeric type and removing all the $ signs. Then I converted all the date columns, Award Date, Start Date, End Date into date types. There was one date whose start year was was in 2204, this was clearly a typo so I set it to 2024. There were a significant number of contracts whose Start Date was the same as their End Date. This complicated some Per Day calculations I did, so I simply added one day to the End Dates of those contracts. I believe this hardly impacted the dataset because a same day contract and a one day contract are roughly the same in my view. I then filtered out any contracts that were awarded or started before 1st of January 2020. There were an number of contracts which had ridiculous start dates (i.e. 2156) so I removed any contracts beyond 31st of December 2099, which may still be to late. Finally I removed any duplicate rows, I believe that the IJF database has a significant number of duplicate rows. It is also possible that rows were duplicated in the web scraping process as I noticed some strange behavior on IJF's website. If you go to the last page take note of the top companies there, and then the hit Previous page button, then the Next page button you will notice that its a difference set of companies. In any case I removed any duplicate rows. Finally I processed the supplier names.

A.3 Proccessing Supplier

To begin I made every suppliers name lowercase. Then I divided each word in a suppliers name and stored them in a list in the column processed_supplier. Then I made a list of 'banned' words, which are word that I will remove from each suppliers name. These banned words are thing like: incorporated, llc, the, service, basically any useless non-identifying words. But sometimes a company name is two banned words so they are left with and empty list in processed_supplier. Most notably this happened with Vancouver Shipyards, to fix this whenever the program found a list with vancouver and shipyards it would concatenate these words into one string. Then I loop through the the lists in processed_suppliers and removed any banned words. Finally I created a dictionary that would go through each processed_suppliers and pick out the individual words and set that as a key. Then I would loop through an if a key was present in a processed_suppliers list it would add the Amount of that contract to the value which would be a running sum for the suppliers with that key.

I would then print out the top ten keys, at first many of these top ten were not descriptive and could not be related to a specific company. Then I would add this key to the banned words list and re-run the program. Eventually I had banned enough non-descriptive words that the top ten keys were aggregating only one supplier each and could be related to a specific company.

Though this is the draw back of this method. If I wanted to get the top 15 or top 100 companies it would be very difficult to root out all of the confounding words. I also had to be vigilant

when added a word to the banned words list because I could rule out an entire company, like what happened with Vancouver Shipyards. Then I had to create a work around that allowed be to remove the words "Vancouver" and "Shipyards", while being able to keep "Vancouver Shipyards". Essentially the problem with this method is that it is matching companies up if the share a single word in common, and that if fine for most companies (after getting rid of the banned words), but some company names are made up of generic words like Vancouver Shipyards.

This problem would only get worse the further down the list you go and at some point it may be impossible to continue with this method. To make matters worse when adding new data you may already be banning all the words in their names and would have no idea. Because at least when you are looking to add a banned word you can check to see if that word is the last word for a lot of companies, but you won't get that opportunity for novel companies. In short I am sure the top 10 companies that I am displaying truly are the top 10, however using this method I cannot say what the top 11 or more companies are and if new data is introduced then all bets are off.

To create a more rigorous supplier name processing method I would completely delete this method and start again. Some good candidates that I have looked into for better supplier name parsing would be: Fuzzy Matching and String Distance Algorithms, Natural Language Processing (Tokenization, Named Entity Recognition (NER)). It would be nice to apply these algorithms to the contract descriptions as well, in that case you could gain further insights into what the government is buying.

B Contract Surveys

C Model Diagnostics

Predictor	General Model	National Defence Model
Baseline Days	1.000395	1.000054
Duration Days	1.000395	1.000054

VIF Table

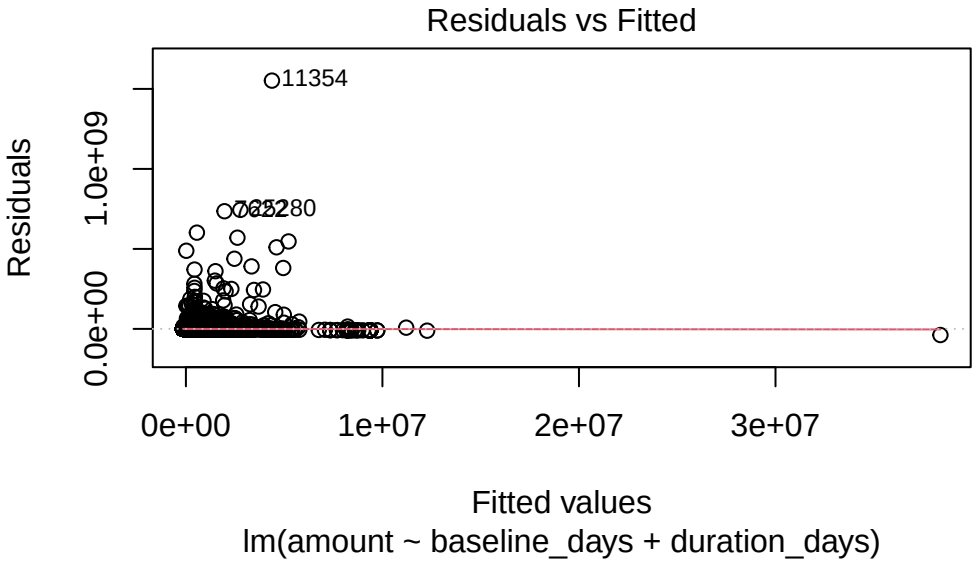

Figure 11: Residual Vs Fitted for General Model

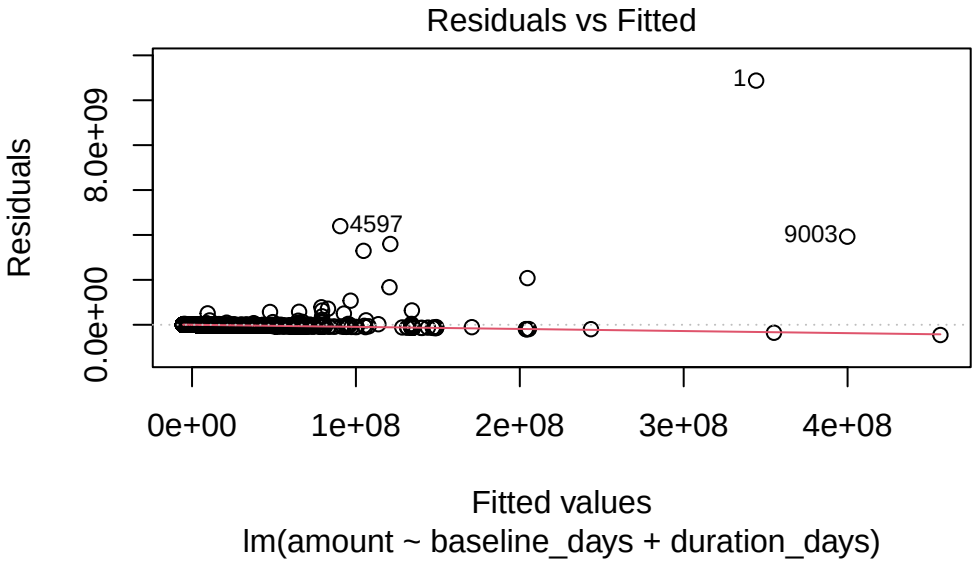

Figure 12: Residual Vs Fitted for Military Model

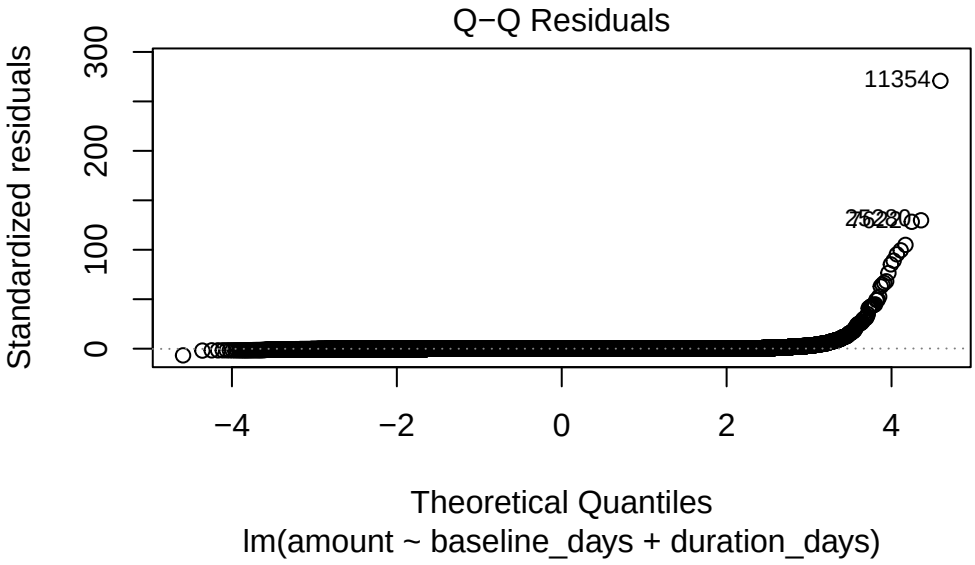

Figure 13: QQ-Normal General Model

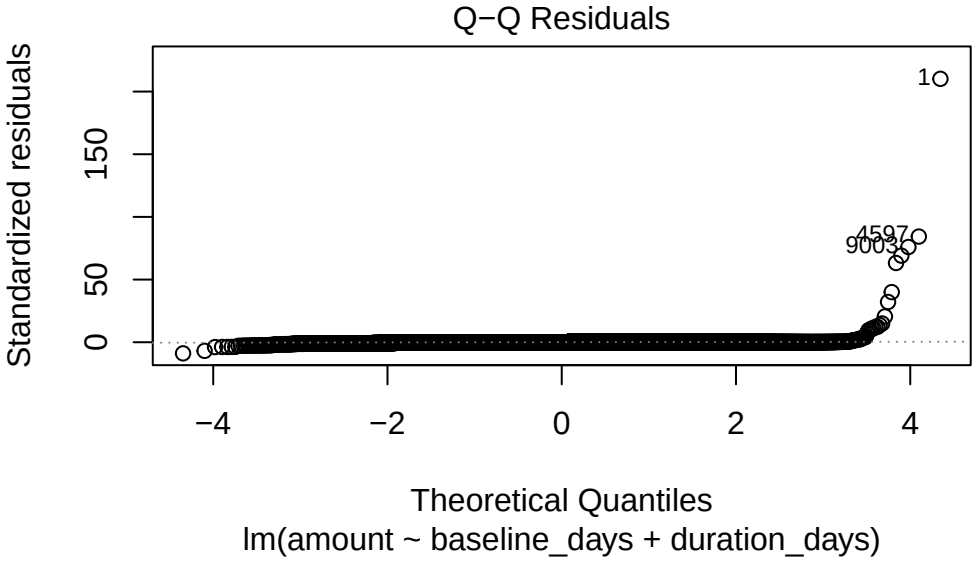

Figure 14: QQ-Normal National Defence Model

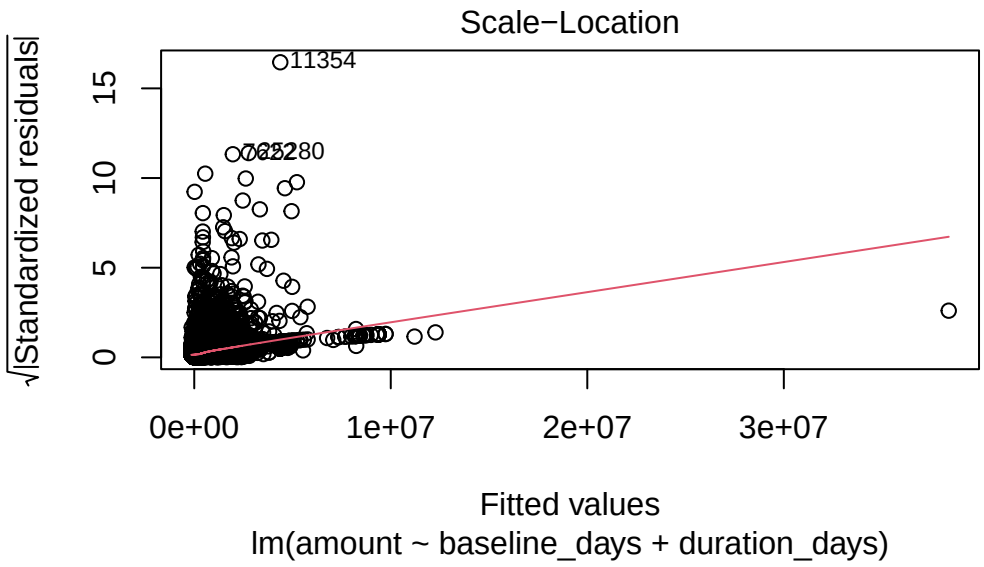

Figure 15: Scale-Location Plot for General model

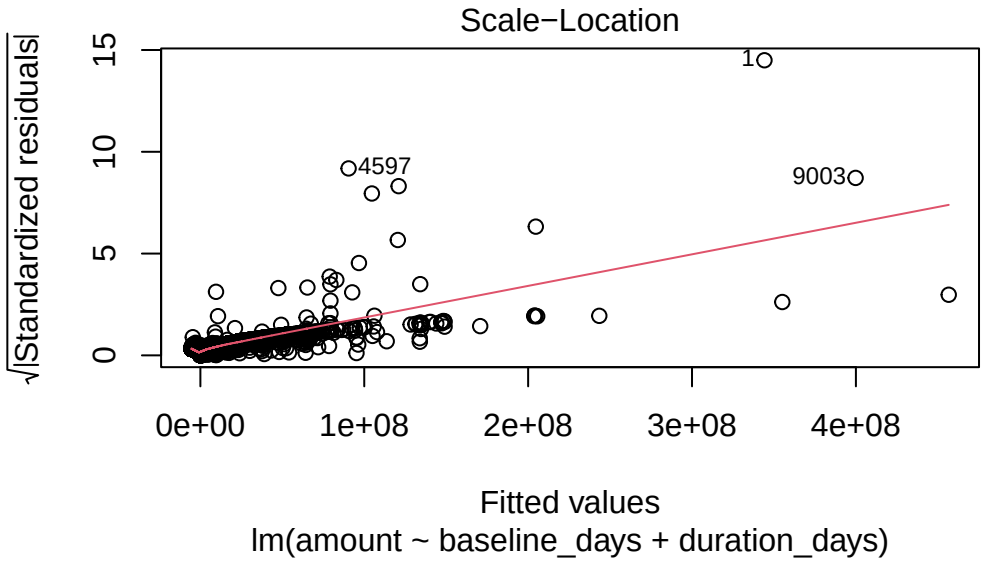

Figure 16: Scale-Location Plot for National Defence model

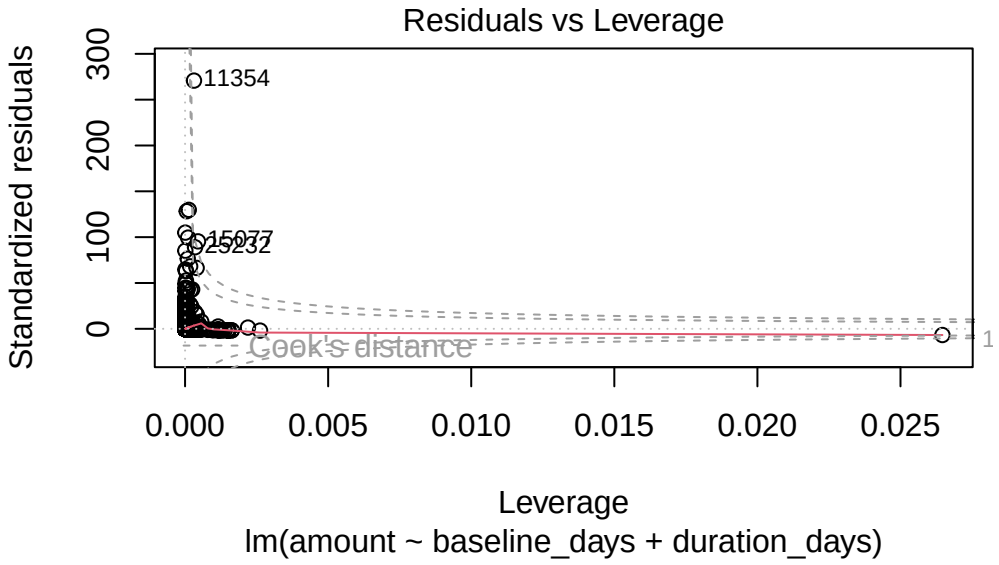

Figure 17: Residuals vs Leverage for Full model

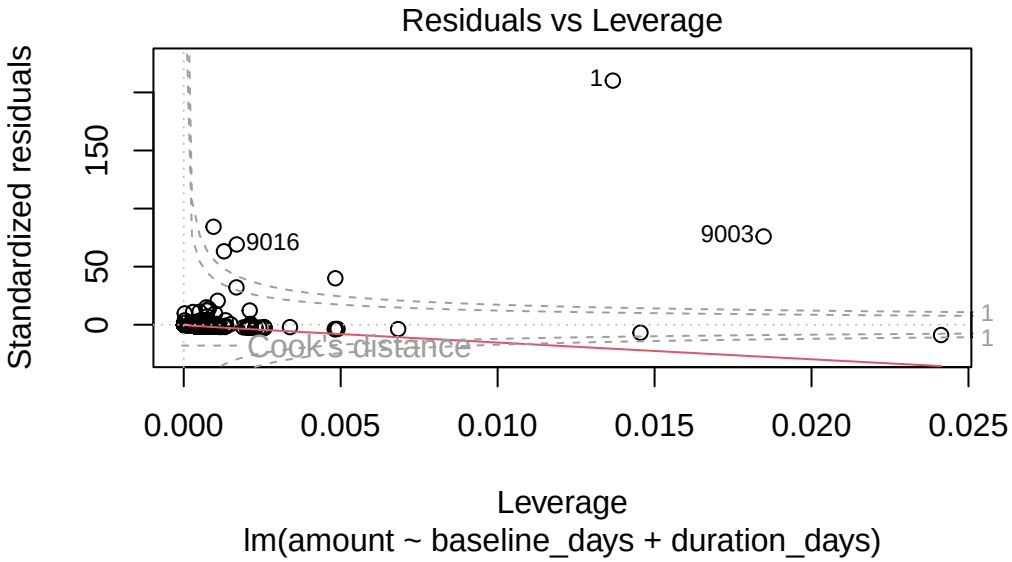

Figure 18: Residuals vs Leverage for National Defence model

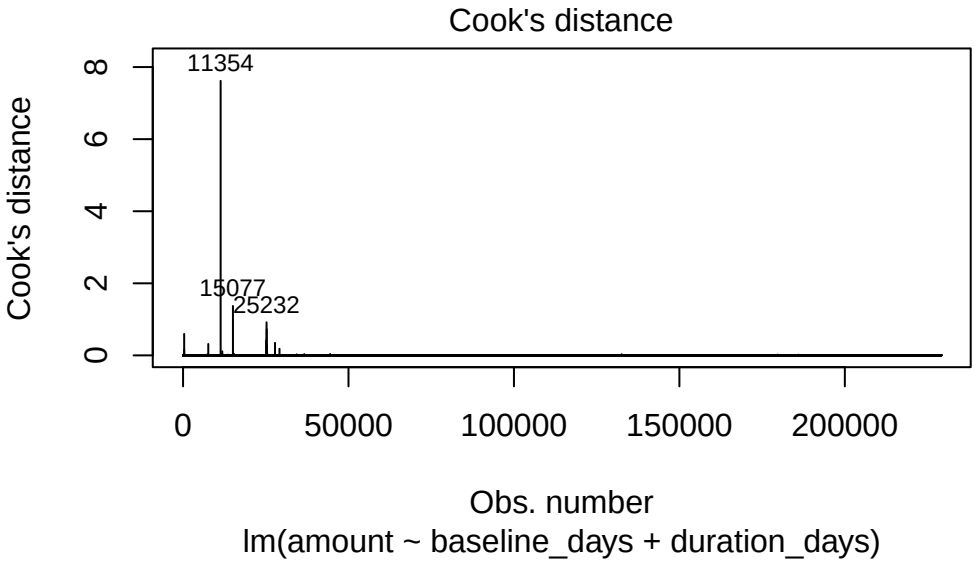

Figure 19: Cook's Distance for full model

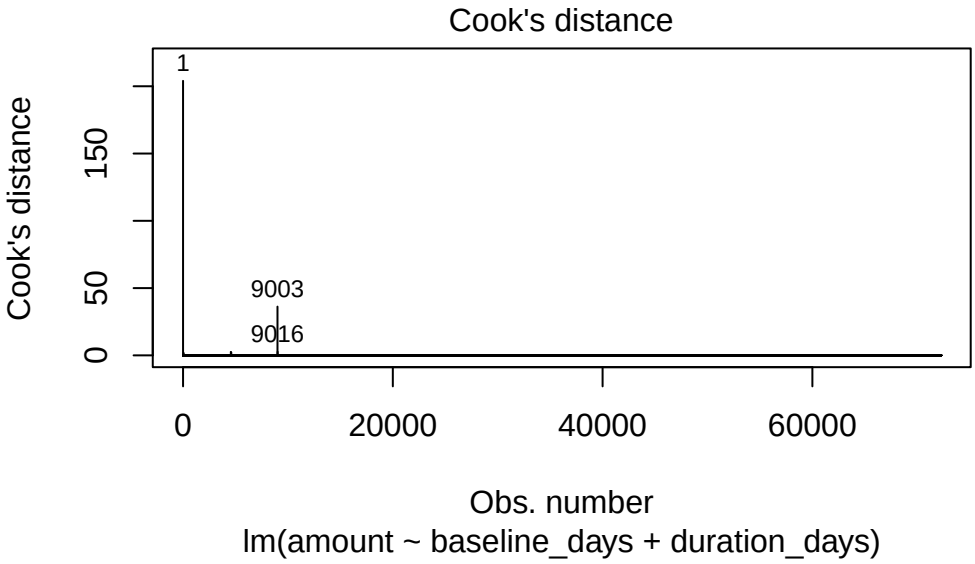

Figure 20: Cook's Distance for National Defence model

References

Dutang, Christophe, Vincent Goulet, and Nicholas Langevin. 2022. "Feller-Pareto and Related Distributions: Numerical Implementation and Actuarial Applications." *Journal of Statistical Software* 103 (6): 1–22. https://doi.org/10.18637/jss.v103.i06.

Foundation, Investigative Journalism. n.d. "Investigative Journalism Foundation." https://theijf.org.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Harrison, John. 2022. *RSelenium: R Bindings for 'Selenium WebDriver'*. https://CRAN.R-project.org/package=RSelenium.

Kuhn, and Max. 2008. "Building Predictive Models in r Using the Caret Package." *Journal of Statistical Software* 28 (5): 1–26. https://doi.org/10.18637/jss.v028.i05.

Nass, Daniel, Martin Allen, and Sam Park. 2024. "Investigative Journalism Foundation Procurement." https://theijf.org/procurement.

Neuwirth, Erich. 2022. *RColorBrewer: ColorBrewer Palettes*. https://CRAN.R-project.org/package=RColorBrewer.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Robinson, David, Alex Hayes, and Simon Couch. 2023. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2022. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

———. 2024. *Rvest: Easily Harvest (Scrape) Web Pages*. https://CRAN.R-project.org/package=rvest.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2014. "knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich

Leisch, and Roger D. Peng. Chapman; Hall/CRC.

Zeileis, Achim, and Gabor Grothendieck. 2005. "Zoo: S3 Infrastructure for Regular and Irregular Time Series." *Journal of Statistical Software* 14 (6): 1–27. https://doi.org/10.18637/jss.v014.i06.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.