

Salaries and Rising Police Force Budgets*

An Exploratory Analysis on Expenditures of Toronto Police Force

Meha Grewal

27 April 2022

Abstract

In this paper we will explore the budget of the Toronto Police Service and examine the implications of their fiscal responsibilities. Through the understanding of the central influences of the police force, we can determine how to effectively distribute resources and funding. These findings provide rationale for policy decisions.

1 Introduction

The Toronto Police Service (TPS) are supposed to be the upholders of law and order. They provide a sense of safety to the community and work to prevent and solve criminal cases. The budget for the the TPS is highly significant as it allows them the ability to safely govern the city. To identify the most influential aspects of the TPS budget and thus the parts of policing that they deem as priority, I will analyze the relevant variables and model them. I explore their expenditures and fiscal responsibilities for 2021 and its implications in providing safety for residents of Toronto.

Conducting an exploratory analysis on the budget for TPS would indicate areas that the government prioritizes. Utilizing the dollar amount as the response variable, we can conclude the largest predictors for police force expenditures and its implications. This information can be applied by government officials and policy makers to minimize government spending and allocate finances to areas that most necessitate it.

Beginning with a general overview of the variables provided in the data, I visually identified where majority of the budget is spent on. Through a multiple linear regression model, I was able to determine the most statistically significant variables. This indicates what portions of the police force have the most influence over the budget and the priorities of the force. By analyzing each predictor to understand its importance, I then considered the broader societal implications. The linear regression indicates how finances are allocated and which districts and units relieve the largest portions of the budget. By determining what districts and categories within the police force require the most funding, we can evaluate the future of the force. This analysis provides the determining factors of the fiscal responsibilities for the police and showcases what areas require more resources. Policy decisions influence the acceptable expenditures of the TPS and this paper aims to understand the ramifications of the budget distribution.

The data set provided by the city of Toronto showcases information regarding actual expenditures from the police force in 2021. By examining the relevant summary statistics and plots, I am able to identify key predictors for our model. I then model the data using the amount for the budget as the response variable and conduct a multiple linear regression. This was done to predict the budget and the crucial components. My discussion includes the understanding of our model and how it was assembled. Lastly, I consider the social implications of the model and how the budget has an impact on safety in Toronto.

*Code and data are available at: <https://github.com/meha-g/FinalPaper>.

Table 1: Top ten budget expenditures for Toronto Police Budget

Pillar Name	Unit Name	Feature Category	Amount
Detective Operations	Other	Salaries	55734700
East Field Command	Court Services	Salaries	49314100
West Field Command	Traffic Services	Salaries	33167600
East Field Command	Division 55	Salaries	31265300
Centralized Service Charges	CCC - Employee Benefits	Benefits	27717200
East Field Command	Communication Services	Salaries	27563100
West Field Command	Division 51	Salaries	26530300
West Field Command	Division 14	Salaries	26264800
East Field Command	Division 43	Salaries	25093100
Centralized Service Charges	CCC - Central Paid Duty	Salaries	24667000

2 Data

2.1 Data Source

This data was obtained from the Open Data Toronto portal (Gelfand 2020), which is accessible due to the City of Toronto. It is a breakdown of the approved budget for TPF and actual expenditures. The data was collected in the fiscal year of 2021 and was published by Toronto Police Services. The budget indicates expenses for the operation of TPF approved by City Council and the Toronto Police Services Board. The raw data includes 2,755 observations of 12 variables, and the data was last refreshed on October 20, 2021. An exploratory data analysis was done on the data using **R statistical programming** (R Core Team 2020). The data was cleaned and manipulated for modelling using **tidyverse** (Wickham et al. 2019), **knitr** (Xie 2022), **dplyr** (Wickham et al. 2022), and **modelsummary** (Arel-Bundock 2022).

2.2 Variables

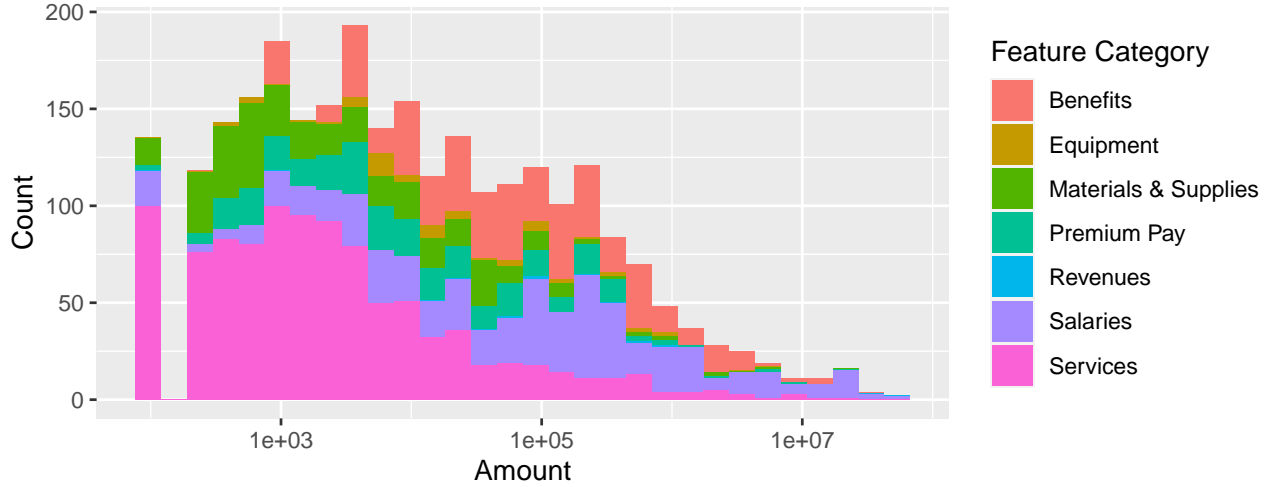
The data collected by the Toronto Police Service includes many variables that explain the budget amount, the organization within TPS, name of the command, pillar, district, and unit. It also includes the feature category which classifies the expenditure, and the cost elements. The dollar amount variable is the most significant sub-section of the data I will be analyzing. The other notable variables will be used as predictors for the budget.

To begin with, I removed the variables for budget type and fiscal year as each data point was the same, all the points in this data set were from the fiscal year 2021 and were approved for the budget. Furthermore, the column for command name and district type were removed as they were similar to pillar name. (Table 1) showcases a preview of the data but only including the pillar name, unit name, feature category, and budget amount. The table showcases the top ten budget expenditures for the TPS in the descending order. Since it is just a brief preview of the data, only the most essential columns are included.

Based off (Table 1), a large portion of budget expenses can be attributed to the salaries of those employed by the police force. Only one data point in the top ten budget expenditures for feature category is associated with benefits while every other expense is due to salaries. Additionally, pillar name includes many points in this table from the East Field Command or the West Field Command, indicating that commandments and particularly those two, require higher budgets.

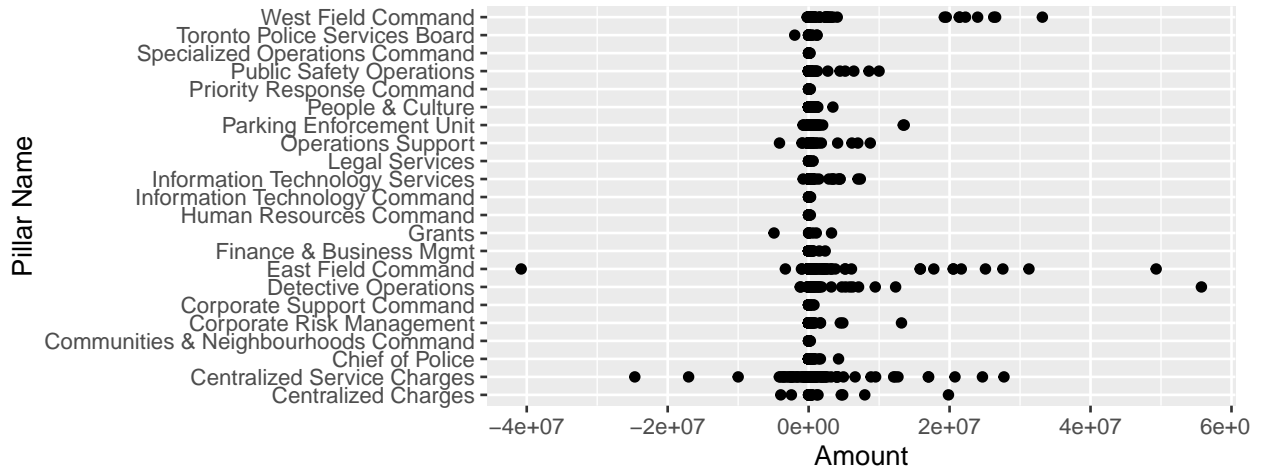
2.3 Features

Figure 1: Distribution of Amount and Feature Category



(Figure 1) displays the distribution of budget expenditures for the TPS, grouped by feature category. The amount of expenses is right skewed and follows a unimodal distribution. Expenses within the salaries category has the largest distribution as it reaches the highest amount, this corresponds with (Table 1) where the highest budgets were for salaries.

Figure 2: Relationship between Amount and Pillar Name







(Figure 2) is a scatter plot of the relationship between budget amount and pillar name which determines where the expense is for. The East Field Command and Detective Operations appear to have outliers, with East Field Command also having a large negative outlier. The largest expenditures derive from West Field Command, East Field Command, Detective Operations, and Centralized Service Charges. This information is useful in determining which sections of the police force require the most or least funding, it can be used for future budget determinations.

2.4 Methodology

To model this data for an exploratory data analysis, the data set was manipulated. I created numerical variables that would allow for a multiple linear regression to be done and statistical significance be tested. I constructed a dummy variable that indicates whether the budget expense was for a command, or other service. Similarly, I built a dummy variable that determined whether the budget was spent on salaries or another category. These variables were created for the model to see whether salaries or commands have a

Table 2: Summary of discrete variables for TPS budget

	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max	
Amount	1043	0	411 215.0	2 923 700.0	−40 790 800.0	6100.0	55 734 700.0	
is_command	2	0	0.4	0.5	0.0	0.0	1.0	
organization_num	3	0	1.1	0.4	1.0	1.0	3.0	
rev_or_sal	2	0	0.1	0.3	0.0	0.0	1.0	

strong impact on the budget. Another variable was made to just assign a numerical value to the organization entity, ranging from 1 to 3.

(Table 2) showcases the summary statistics for the discrete variables in the data set. This includes the amount from the budget, alongside the three new numerical variables constructed for the model.

3 Model

3.1 Multiple Linear Regression

To analyze the budget of the TPF and how different variables impact the expenditures of the police force, I created a model using a multiple linear regression. This was done to identify a relationship between amount and salaries, command, and organization within the police force. This model was appropriate for this data set as we are examining linear relationships for exploratory variables with the predictor variable, which is amount.

$$Y_i = \beta_0 + \beta_1 X_1 + \epsilon_i \quad (1)$$

Equation 1 follows the first model which is just a simple linear regression with amount only being explained by feature category. Equation 2 includes three more variables to explain more variability in the model and transform it into a multiple linear regression

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon_i \quad (2)$$

In this model for Equation 2, Y_i is the budget amount and β_0 is the intercept or the predicted value when all other variables are zero. X_1 is whether the expense is for a command, X_2 is the organizational entity, X_3 is whether the expense is a salary, and X_4 is the feature category.

A multiple linear regression was selected because amount is the variable of interest and this paper is examining what factors impact the budget.

4 Results

(Table 3)

Table 3: Explaining TPF budget based on whether it is a command, organization entity, salary, and feature category

	Model 1	Model 2
(Intercept)	501 332.340 (130 936.505)	461 880.660 (198 199.561)
Feature_CategoryEquipment	−427 859.927 (395 060.589)	−401 584.043 (385 700.313)
Feature_CategoryMaterials & Supplies	−432 398.504 (203 686.856)	−437 204.835 (198 495.342)
Feature_CategoryPremium Pay	−329 792.614 (211 517.743)	−309 342.804 (206 948.526)
Feature_CategoryRevenues	−3 094 183.218 (398 133.409)	−3 163 292.946 (389 221.838)
Feature_CategorySalaries	1 066 261.593 (178 764.085)	−357 846.198 (209 512.692)
Feature_CategoryServices	−387 739.147 (158 777.477)	−388 355.574 (154 753.092)
is_command		118 590.829 (112 987.133)
organization_num		−4663.404 (123 446.891)
rev_or_sal		2 882 793.173 (235 822.807)
Num.Obs.	2751	2751
R2	0.059	0.108
R2 Adj.	0.057	0.105
AIC	89 569.3	89 428.8
BIC	89 616.7	89 494.0
Log.Lik.	−44 776.648	−44 703.417
F	28.881	36.939
RMSE	2 838 635.78	2 765 580.51

Figure 3: Feature Category Budget and Is Command

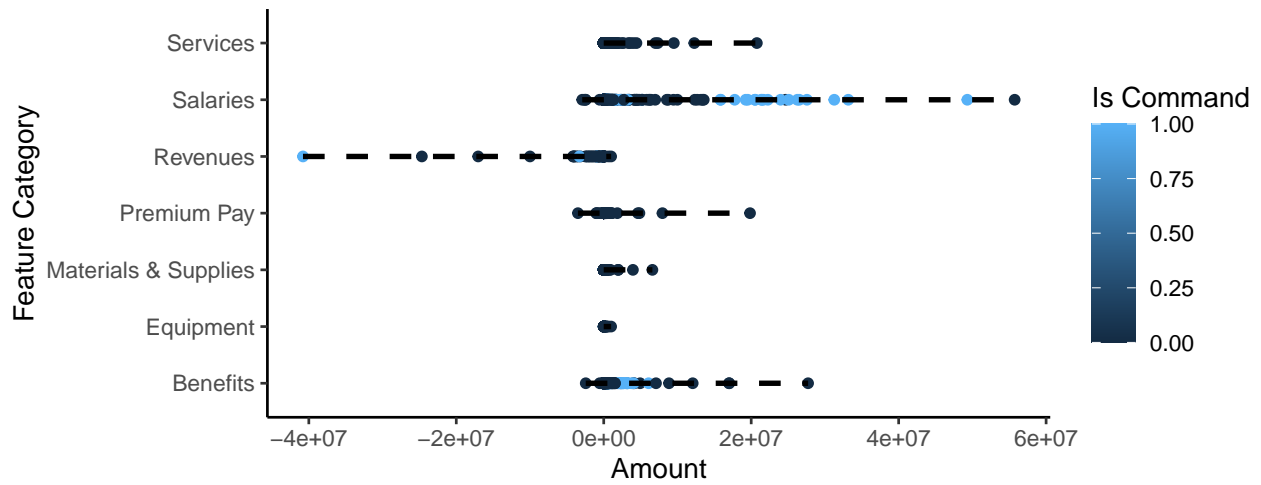


Figure 4: Salary or Not Budget and Organization Entity

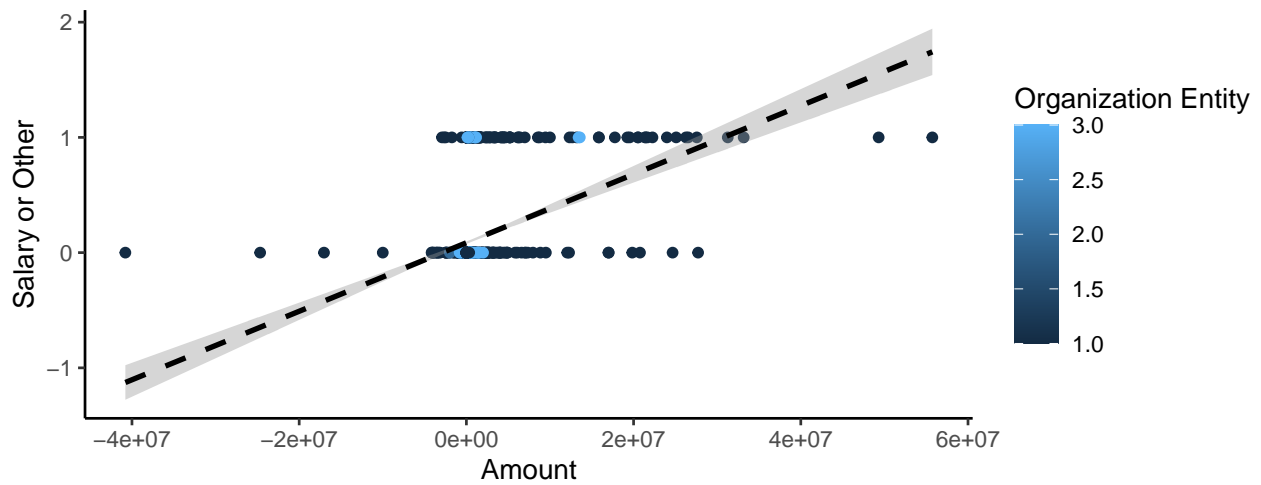
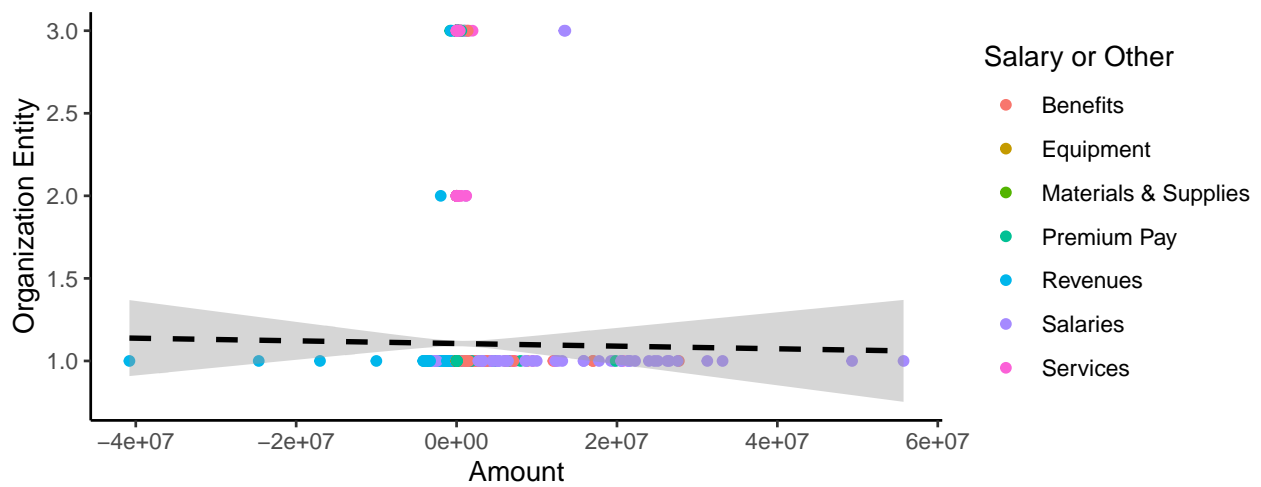


Figure 5: Organization Entity Budget and Feature Category



5 Discussion

5.1 Findings

If my paper were 10 pages, then should be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Implications

5.3 Limitations and next steps

Weaknesses and next steps should also be included.

Appendix

.1 Datasheet

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
 - The dataset was created to record the Toronto Police Budget for the 2021 fiscal year.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
 - The dataset was published by the Toronto Police Services on behalf of the City of Toronto.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
 - The data was funded by the City of Toronto as it is a governmental responsibility.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
 - The instances in the dataset represent expenditures for the police force. There is one type of instance, the expense on the budget and its corresponding descriptions.
2. *How many instances are there in total (of each type, if appropriate)?*
 - There is one instance in total.
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
 - The dataset contains all possible instances for the budget, it is the observational data for the Toronto Police Service.
4. *What data does each instance consist of? “Raw” data (for example, unprocessed text or images) or features? In either case, please provide a description.*
 - each instance contains raw data that describes the expense on the budget.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
 - There is a label associated with the instance, there is a unique ID variable for each data point.
6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*
 - There were four missing variables within the instance for the budget, it was unavailable and not accessible.
7. *Are relationships between individual instances made explicit (for example, users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit.*
 - There are no relationships between individual instances.
8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*
 - There are no recommended data splits.
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
 - TBD
10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any*

restrictions associated with them, as well as links or other access points, as appropriate.

- TBD
- 11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*
 - TBD
- 12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
 - TBD
- 13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*
 - TBD
- 14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*
 - TBD
- 15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*
 - TBD

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*
 - The data was directly observable raw data in a csv file. All the data was validated by the City of Toronto.
2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*
 - The data was collected directly by the Toronto Police Service.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
 - The data is not a sample.
4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*
 - Employees from the Toronto Police Service were involved in the data collection process, they were compensated with governmental salaries.
5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*
 - The data was collected over the 2021 fiscal year, it does match instance.
6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*
 - The expenses were reviewed by the Toronto Police Services Board and the City Council.
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
 - The data was collected via third party from the Open Data Toronto portal.
8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to,*

or otherwise reproduce, the exact language of the notification itself.

- The individuals at Open Data were not notified about the data collection.
9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*
 - The individuals did consent to the use of the data, the purpose of the dataset was to provide open access through the Open Data Toronto portal since it is governmental information. The consent was provided in their license information, <https://open.toronto.ca/open-data-license/>
 10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*
 - Consent cannot be revoked since it was developed through public consultation with the government.
 11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*
 - There has not been an analysis of the potential impact of the dataset.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
 - Missing values were removed and unnecessary columns were removed.
2. *Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.*
 - The raw data is saved on the Open Data Toronto portal: <https://open.toronto.ca/dataset/toronto-police-budget/>
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
 - R statistical programming was used to process and clean the data (R Core Team 2020).

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
 - The dataset was used to examine the variables within, used for tables, plots, and summary statistics. The data was also manipulated and used for a multiple linear regression model.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
 - There is no repository to all papers or systems that use the dataset.
3. *What (other) tasks could the dataset be used for?*
 - The dataset can be used for policy planning regarding government budgets. Additionally, it can be used to analyze safety in the city of Toronto.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
 - The dataset is a breakdown of the approved budget for Toronto Police Service, risks involving financials for different classifications within the police force may arise.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
 - The dataset should not be used for topics not involving the expenditures of the police force, as that is the only suitable utilization of this data.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution,*

organization) on behalf of which the dataset was created? If so, please provide a description.

- It is possible the dataset may be distributed within the municipal government, but to my knowledge there has been no distribution besides its availability on Open Data Toronto.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
 - The dataset is available on Open Data Toronto, the unique identifier for getting the package is 668434ee-9541-40a8-adb6-0ad805fcc9b6.
 3. *When will the dataset be distributed?*
 - The dataset was last refreshed in October 20, 2021.
 4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
 - The dataset is distributed under an Open Government License, a link: <https://open.toronto.ca/open-data-license/>
 5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
 - There have been no imposed restrictions.
 6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
 - Regulatory restrictions from the City of Toronto must apply.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
 - The dataset is maintained by the Toronto Police Services
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
 - The Toronto Police Service can be contacted by email: Analytics.Innovation@torontopolice.on.ca
3. *Is there an erratum? If so, please provide a link or other access point.*
 - There is no erratum.
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
 - It is refreshed annually by the Toronto Police Services.
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
 - The dataset does not relate to people or any individuals.
6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*
 - The 2020 fiscal year budget will continue to be maintained by being refreshed annually on the Open Data Toronto portal.
7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*
 - If others want to augment the dataset they can use the “For Developers” section to copy and paste the code in their respective coding language. Their contributions will not be verified. There is no process of communicating to dataset consumers.

References

- Arel-Bundock, Vincent. 2022. *Modelsummary: Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready*. <https://CRAN.R-project.org/package=modelsummary>.
- Gelfand, Sharla. 2020. *Opendatatoronto: Access the City of Toronto Open Data Portal*. <https://CRAN.R-project.org/package=opendatatoronto>.
- R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Golemund, 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, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation*. <https://CRAN.R-project.org/package=dplyr>.
- Xie, Yihui. 2022. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. <https://yihui.org/knitr/>.