Datasheet for ‘Operating Budget Program Summary By Expenditure Category’[*]

Murrumbidgee Paper

ZeWei Zhou

March 12, 2024

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.* This data set identifies the approved and recommended annual operating budget summary by expenditure category in each program or division starting from 2011. A new budgetary file is published annually in this data set. The data set needs more variables in order to proceed more professional statistical analysis since the established models shows very low value in R squared.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - Winnie Chen 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.*
    - City of Toronto.
4. *Any other comments?*
    - Maybe provides more options will be reasonable.

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.*

[*] Code and data supporting this analysis is avaliable at:https://github.com/ zhouJoe2033/MurrumbidgeePaper/blob/main”

- The instances comprises the dataset represent documents. There are no multiple types of instances. All the data file has been stored properly in a certain format and has been updated yearly.

2. *How many instances are there in total (of each type, if appropriate)?*

   - From year 2011 to the year of 2023, ther are 12 instances.

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).*

   - It possibly contains all instances since the data source is from the government, because if that's the open source that the government is willing to reveal then that is pretty much all the instances to the public to know therefore it can be considered to the public as all the instances unless public becomes smarter.

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 contain a .xlsx file that can be considered as raw data since it requires work to modify the data from text to usable numeric values to process statistical analysis.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

   - There is no label or target associated with each instance.

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 is no information missing from individual instances.

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.*

   - The relationhips between individual instances does not make explicit.

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 is no recommended data splits.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

   - There is no errors, source of noise, or redundancies in the data set.

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.*

    - The data set if self-contained since there is owner who is from the government.

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.*

    - The datas et does not contain data that can be considered confidential.

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.*

    - The data set deos not contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety.

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.*

    - The data set does not identify any sub-populations.

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.*

    - It is not possible to identify any sub-populations.

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.*

    - The data set does not contain data that might be considered sensitive in any way.

16. *Any other comments?*

- The data set needs more variables.

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 is associated with each instance by having the same variables.

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?*

    - Possibly collects from the certain department by emailing them for the data source. Author did not mention.

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 set 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)?*

    - Only the author who is from the city.

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.*

    - From the year of 2011 to the year of 2023.

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.*

    - There is no ethical review process conducted.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

    - I collect the data from the source's website directly.

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 individual might notify the data collection by having a one increment in the numebr of downloads.

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.*

    - Yes, since the data is open source.https://ckan0.cf.opendata.inter.prod-toronto.ca/dataset/budget-operating-budget-program-summary-by-expenditure-category

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).*

    - The consenting individuals does not provide with a mechanism to revoke their consent.

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.*

    - The analysis of the potential impact of the dataset and its use on data subjects has not been conducted.

12. *Any other comments?*

    - Need more variables.

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.*

    - The data is clean, since all the text value has been well organized, however, the amount variables does not seem legit.

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 was saved in addition to the preprocessed/cleaned/labeled data. https://ckan0.cf.opendata.inter.prod-toronto.ca/dataset/2c90a5d3-5598-4c02-abf2-169456c8f1f1/resource/d55a2458-f116-456e-a3be-4a0d867fa190/download/approved-operating-budget-summary-2016.xlsx

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

   - There is no software that was used to preprocess/clean/label the data avaliable.

4. *Any other comments?*

   - Need more variables.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

   - The data set has not been used for any tasks already.

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 that links to any or all papers or systems that use the data set.

3. *What (other) tasks could the dataset be used for?*

   - Big data analysis.

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?*

   - There is no a=other things about the composition of the data set or the way it was collected and preprocessed/cleaned/labeled that might impact future uses.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

   - There re no tasks for which the data set should not be used.

6. *Any other comments?*

   - Need more variables.

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.*

    - The data set will not be distributed to third parties outside of the entity.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

    - There should be some hidden code written in the excel file that declates the author's integrity.

3. *When will the dataset be distributed?*

    - I don't think it will be distributed.

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 data set will not be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU).

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 is no third parties imposed IP-based or other restrictions on the data associated with the instances.

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.*

    - There is no export controls or other regulatory restrictions apply to the dataset or to individual instances.

7. *Any other comments?*

    - Need more variables.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

    - City of Toronto

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

- winnie.chen@toronto.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)?*

   - There is no labeling errors so far.

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 data set does not relate to people.

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.*

   - All versions of the data set has been listed.

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.*

   - There is no mechanism for them to do so.

8. *Any other comments?*

   - Need more variables.