Datasheet for Analysis Data 'analysis_data.parquet'*

dataset can be found in data/02-analysis_data

Hanqing Yang

December 3, 2024

This datasheet details the Canadian Municipal Elections Database (CMED), an observational dataset spanning over 150 years of Canadian municipal elections. It compiles candidate-level data on gender, incumbency, election outcomes, and years from diverse sources, including provincial archives and municipal records. This datasheet provides a concise overview of CMED's structure, methodology, and potential applications for electoral research.

Extract of the questions from Gebru et al. (2021).

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 enable analysis of Canadian municipal election outcome. The analysis focusing on study the role of gender and incumbency status in determining electoral outcomes. The dataset supports analysis of long-term trends in female participation and success in municipal elections, as well as the gendered dynamics of incumbency advantages over time. The original dataset, obtained from archival and municipal records, was processed to address missing values and standardize variable types. For example, data cleaning process including remove missing value and duplicate row from raw dataset and recode gender variable from characters to factor type with binary outcome '1' for female and '0' for male. This process allow me to fit logistic regression model in my analysis .

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

*Code and data are available at: [https://github.com/Lorina-Y/Canadian_Municipal_election.git]

- The dataset was primarily created by researchers associated with the Canadian Municipal Elections Project, led by Jack Lucas at the University of Calgary. The project team compiled the dataset from a variety of municipal, provincial, and archival sources to build a comprehensive resource for studying Canadian local government elections. It was lightly adapted by myself, Hanqing Yang, an undergraduate student at the University 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 creation of the dataset was funded by grants associated with the Social Sciences and Humanities Research Council of Canada (SSHRC), specifically under the Canadian Municipal Barometer project.

4. *Any other comments?*

    - The dataset aggregates information from diverse sources, including CivicInfoBC, Données Québec, and historical archives from Toronto and Montreal. It serves as a valuable resource for examining gender equity, incumbency dynamics, and other factors influencing municipal electoral outcomes across Canada.

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.*

    - Each instance in the dataset represents a candidate who participated in a Canadian municipal election. The data includes candidate-level variables such as gender, incumbency status, election year, municipality, and electoral outcome. These variables enable analysis of individual electoral performance as well as broader trends across time and regions.

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

    - The cleaned dataset used for analysis contains 84,028 instances, each representing a unique candidate. And the raw dataset contain 110,800 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).*

- The dataset is not a random sample but rather a comprehensive compilation of available records for municipal elections in Canada from 1862 to 2022. It aims for record every candidates' information and elections feature for every municipal election take place in Canada.

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 consists of both features of candidates and features of a municipal election, These include:
     - **Gender**: Indicates whether the candidate is male or female, with 'M" for male and 'F' for female
     - **Incumbent status**: Whether the candidate held the same office prior to the election. 1 for incumbent and 0 for not incumbent
     - **Election year**: The year of the election.
     - **Municipality**: The name of the municipality where the election occurred.
     - **Electoral outcome**: A binary variable indicating whether the candidate was elected (1) or not (0).
     - **first_name** and **last_name**: name of candidate.
     - **party**: party that a candidate belong to, such as independent, Burnaby Green Party, or Abbotsford First Electors Society and so on.
     - **votes_received** and **total_votes**: number of votes received by the candidate in a election and total votes recevive for this election.
     - **acclaimed** : indicate whether the candidate was acclaimed for a position or not.

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

   - No

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 missing information in variable gender, incumbent and region. It was more prevalent in early election records due to inconsistent archival practices. For example, gender information for earlier candidates was unavailable in some cases and had to be manually inferred. Observations with missing critical information were dropped during data cleaning to maintain dataset integrity.

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.*

- Relationships between instances are implicit rather than explicit. For example, candidates in the same municipality or election year may be related through competition. However, these relationships are not explicitly encoded in the dataset and require additional processing to analyze.

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.*

    - The dataset was split into training (70%) and testing (30%) subsets for model building and validation. The split ensures that model evaluation is performed on unseen data to assess generalizability and prevent overfitting.

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

- No

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 dataset is self-contained and includes all variables necessary for analysis. However, it was compiled from a variety of external resources, such as provincial archives, municipal records, and online repositories. While these sources provided foundational data, the CMED aggregates and cleans this information into a cohesive format.
    - There are no guarantees that external resources used for compiling the dataset will remain available or unchanged over time. For example, some online municipal records may be updated or removed.
    - The CMED dataset serves as an archival version of these records, providing a consolidated and cleaned dataset that reflects the state of the data at the time of its creation.
    - The CMED dataset is made freely available through BorealisData website.

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.*

    - No, the dataset does not contain any confidential information. All data is derived from public records and pertains to municipal election candidates and outcomes,

which are publicly available. No private communications or legally protected data are included.

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.*

    - No.

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.*

    - Yes, the dataset identifies sub-populations based on gender (Male and Female) and incumbency status (Incumbent and Non-Incumbent). These sub-populations are explicitly coded as variables and are central to the analysis of electoral success. Gender distribution shows approximately 24.2% female candidates and 75.8% male candidates. Incumbency distribution indicates that approximately 34.09% of candidates are incumbents.

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.*

    - Yes, individual candidates can be directly identified by their first and last names, which are included in the dataset.

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.*

    - No

16. *Any other comments?*

    - No

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 acquired primarily through direct observation and manual curation. Raw data from sources like provincial archives, municipal records, and online repositories were compiled into the CMED. Where data was missing (e.g., candidate gender in earlier years), manual coding based on archival photographs and text descriptions was conducted to fill gaps. While no explicit validation is documented for these manual processes, the use of established archival sources adds credibility to the dataset.

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?*

    - Data collection involved both manual curation and digitization. For early records, historical election posters and written descriptions were manually reviewed to code missing values such as gender. For modern records, digital resources like CivicInfoBC and Données Québec were used to compile structured data. Validation procedures include cross-referencing multiple sources to ensure consistency where available.

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 dataset represents a complete aggregation of municipal election data for Canada, rather than 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)?*

    - The data was collected by researchers and archivists associated with Canadian universities and institutions responsible for compiling the CMED, for example the University of Calgary. Compensation details for contributors are not specified.

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 dataset itself spans elections from 1867 to 2022. Data collection for CMED occurred over several years recently, the dataset was compiled since 2020 but represents historical data collected over a long period, aligning with the timeframes of the municipal elections it covers (1867–2022).

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.*

    - No

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 obtained the data entirely from third-party source, from borealis website (Merrill et al. 2020).

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.*

    - As the data pertains to public election records, individual notification was not required. All information was already publicly available at the time of 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.*

    - No, since consent was not required.

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).*

    - No.

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.*

    - No.

12. *Any other comments?*

    - No.

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.*

    - Yes, preprocessing steps included the removal of rows with missing values, ensuring categorical variables (e.g., gender, elected, and incumbent) were correctly encoded as factors, and eliminating duplicate entries.

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.*

    - Yes, both raw data and analysis data were saved to allow verification of result and support future researches. The raw data is stored in data/01-raw_data.

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

    - Yes, R was used.

4. *Any other comments?*

    - No

Uses

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

    - Yes, the dataset has been used in other studies to evaluate Canadian municipal election outcome. Such as analyze geographic disparities in election outcomes.

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.*

    - No

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

    - The dataset can be used for study voter turnout trends,investigate party affiliations' influence on elections, or model the effects of socioeconomic factors on municipal election dynamics.

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?*

    - Yes, the dataset's reliance on manual gender coding and historical archival records introduces potential biases. Users should note these limitations when conducting analyses, particularly if aiming to generalize findings across different contexts or regions.

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

- The dataset is not suitable for making causal inferences due to its observational nature. It also lacks sufficient detail for studies involving individual voter behavior or campaign-specific dynamics (e.g., advertisement effects).

6. *Any other comments?*

   - No

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.*

   - Yes, the dataset is publicly available for third-party use through the Borealis Data Repository, allowing researchers and practitioners to access and analyze the data.

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 distributed via the Borealis platform, with a Digital Object Identifier (DOI): 10.5683/SP2/4MZJPQ, ensuring persistent and reliable access for future use.

3. *When will the dataset be distributed?*

   - The dataset is already available for distribution as of its publication on the Borealis repository in 2020.

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.*

   - Yes, the dataset is distributed under the terms specified by Borealis, which typically align with Creative Commons licensing. Users should consult the repository page for specific terms of use and restrictions, if any.

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.*

   - No IP-based restrictions from third parties are indicated for the dataset.

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.*

   - No.

7. *Any other comments?*
    - No.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - The dataset is hosted and maintained by the Borealis Data Repository, which is affiliated with the Canadian academic and research community.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - The dataset can be accessed and inquiries can be directed through the Borealis Data Repository platform. Specific contact information for dataset managers is not provided but users can use the repository's support system.
3. *Is there an erratum? If so, please provide a link or other access point.*
    - No.
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 unclear if regular updates will be made. If updates occur, such as corrections or additions of new records, these would be communicated via the Borealis repository, potentially in the dataset's metadata or accompanying documentation.
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 directly involve personal data from individuals. Instead, it focuses on publicly available information about election candidates. Therefore, no specific retention limits apply.
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 repository typically retains historical versions of datasets for reproducibility purposes. Older versions are likely accessible through versioning systems provided by Borealis.

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.*

    - Contributions or extensions to the dataset are not currently supported through a formal mechanism.

8. *Any other comments?*

    - No.

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

Merrill, Reed, Jack Lucas, Kelly Blidook, Sandra Breux, Laura Conrad, Gabriel Eidelman, Royce Koop, et al. 2020. "Canadian Municipal Elections Database." Borealis. https://doi.org/10.5683/SP2/4MZJPQ.