# Accessible and Completed Datasets yet Majority are Bronze-Graded and not Updated Regularly with Missing Metadata*

**An analysis of the data quality of datasets available on the Open Data Toronto Portal (As of May 13, 2025)**

Emily Su

May 16, 2025

As one of the central hubs for Toronto-related data, we analyzed the data quality of Open Data Toronto's catalogue. Despite Open Data Toronto's extensive dataset catalogue being accessible and having minimal missing data, 56% of their datasets are graded bronze and bronze-graded datasets are less likely to be updated and have completed metadata fields. These findings can help raise awareness to Open Data Toronto whose datasets play an important role in news reporting and policymaking, and also inform anyone interested in using datasets from Open Data Toronto's catalogue about what goes behind the grade given to datasets.

## Table of contents



*Code and data are available at: https://github.com/moonsdust/data-quality.

# 1 Introduction

In 2024, a story collaboration between the Investigative Journalism Foundation (IJF) and CBC published in the CBC reported that the risk of death and injuries from fires in lower-income Toronto wards was higher compared to higher-income Toronto wards (Penrose 2024). This story as well as other stories published in the news rely on data from Open Data Toronto to bring their stories to life (Penrose 2024). Open Data Toronto serves as a hub for all types of data related to Toronto from crime data to information about shelters across the city (The City of Toronto, n.d.). Open Data Toronto has not only been used in the news but also in civic spaces when establishing city policies (The City of Toronto, n.d.). Given the importance of the hub, it raises concerns about the quality of Open Data Toronto's data catalogue and the following question: What is the quality of the datasets like on the Open Data Toronto portal?

In this paper, we analyzed data provided by Open Data Toronto on the data quality grade of the datasets in their catalogue and their characteristics like "accessibility", "completeness", "freshness", "metadata", and "usability". Farrow analyzed the data quality scores of the different datasets on the Open Data Toronto portal and found that the metadata and freshness scores were poor overall (Farrow 2021). However there has been no analysis done on the characteristics and how they compare for different graded datasets as of 2025. In our findings, all datasets in Open Data Toronto portal are accessible and a majority of them having minimal missing data and are usable. It also showed that 56% of their datasets are graded bronze with bronze-graded datasets being less likely to be updated and have completed metadata fields. These bronze-graded datasets contribute the poor metadata and freshness scores we saw. These findings can inform Open Data Toronto about specific graded datasets that should be given more attention and the specific qualities of them such as metadata completion, which impacts the quality of the data used in the media and in policymaking. These findings can also be informative for users of the portal in order to understand what goes behind the grade of datasets on the portal and what they mean.

For the remainder of the paper, the data section (Section 2) looks at the data used and how it was retrieved alongside the characteristics of the data, the data's limitations, and our variables of interest for our analysis. In the results section (Section 3) we looked at the data more in-depth through graphs. With the discussion section (Section 4), we will provide an overview of what was done in our results, discussing our results and its real-world implications, and indicate areas of improvements for our analysis and directions for future works. Finally, the appendix (Section A) includes acknowledgements and any additional information related to the paper.

# 2 Data

## 2.1 Overview

The dataset used in the paper comes from Open Data Toronto portal titled "Catalogue quality scores" (The City of Toronto 2025). Other datasets like "Toronto Open Data Intake" were considered in the analysis of the paper however, it does not indicate the quality of the datasets that are being requested. This specific dataset looks at the quality of the datasets available from the Open Data Toronto catalogue to inform others how valuable certain datasets are to be used for various situations like reporting on civic issues. The datasets are scored based on characteristics such as its accessibility, completeness, freshness, metadata, and usability, which are then calculated together to give a dataset a grade. This grade is displayed alongside a trophy icon under the details section on a dataset page on Open Data Toronto portal (The City of Toronto 2025).

We used the programming language Python (Van Rossum and Drake Jr 1995), the statistical programming language R (R Core Team 2023), and the following libraries to download, clean, analyze, and test the dataset and the overall paper itself: Requests (Prewitt, Cordasco, and Larson 2011), datetime (Python Software Foundation 2025), Matplotlib (Hunter 2007), numpy (Harris et al. 2020), pandas (The pandas development team 2020), Polars (Vink 2025), Pydantic (Pydantic 2025), seaborn (Waskom 2021), Pointblank (Iannone, Vargas, and Choe 2025), and Pyarrow (Apache 2025).

We retrieved the raw dataset by calling the Open Data Toronto API (The City of Toronto 2025) using the Requests library (Prewitt, Cordasco, and Larson 2011) and downloading the file as a CSV. There are 39,580 total observations in the cleaned dataset with each observation being a dataset in the catalogue. Table 1 shows a preview of what the cleaned dataset looks like:

Table 1: Preview of dataset on Open Data Toronto's Catalogue quality scores as of May 13, 2025

| | accessibility | completeness | freshness | metadata | usability | grade |
|---|---|---|---|---|---|---|
| 0 | 1 | 0.69 | 0.5 | 0.84 | 0.86 | Silver |
| 1 | 1 | 1.00 | 0.0 | 0.25 | 0.85 | Bronze |
| 2 | 1 | 0.98 | 1.0 | 0.25 | 0.69 | Bronze |
| 3 | 1 | 0.96 | 1.0 | 0.75 | 0.94 | Gold |
| 4 | 1 | 0.83 | 1.0 | 0.75 | 0.87 | Gold |

Table 2 shows the summary statistics of the cleaned dataset:

Table 2: Summary statistics of dataset on Open Data Toronto's Catalogue quality scores as of May 13, 2025

| | accessibility | completeness | freshness | metadata | usability |
|---|---|---|---|---|---|
| count | 39580.0 | 39580.000000 | 39580.000000 | 39580.000000 | 39580.000000 |
| mean | 1.0 | 0.872202 | 0.555413 | 0.468757 | 0.839316 |
| std | 0.0 | 0.150054 | 0.472661 | 0.292833 | 0.108338 |
| min | 1.0 | 0.150000 | 0.000000 | 0.000000 | -0.130000 |
| 25% | 1.0 | 0.780000 | 0.000000 | 0.250000 | 0.770000 |
| 50% | 1.0 | 0.940000 | 0.750000 | 0.380000 | 0.850000 |
| 75% | 1.0 | 1.000000 | 1.000000 | 0.750000 | 0.920000 |
| max | 1.0 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |

## 2.2 Measurement

Open Data Toronto uses a metric called the "Data Quality Score" in order to give each dataset in their catalogue a grade indicating "bronze", "silver", or "gold", which can be seen through on the webpage for each dataset on Open Data Toronto portal website. In order to create the "Data Quality Score", they assembled the Data Quality Working Group, which consist of a diverse group of people from consumers of datasets to people who create datasets (Open Data Toronto 2023; Carlos Hernandez 2020).

Open Data Toronto first reviewed various literature such as academic papers and industry white papers to compile 15 dimensions used to measure quality (Open Data Toronto 2023; Carlos Hernandez 2020). The dimensions selected were as follows: Interpretability ("How easy it is to understand the data?"), Usability ("How easy is it to work with the data?"), Metadata ("Is the data well described?"), Freshness ("How close to creation time is the data published?"), Granularity ("How atomic is the data?"), Completeness ("How much data is missing?"), and

Accessibility ("Is the data easy to access?") (Open Data Toronto 2023; Carlos Hernandez 2020).

From there the Data Quality Working Group were surveyed and asked to rank the importance of each dimension when it comes to assessing data quality where 1 represents the most important and 7 being the least important. The results from the survey would then help the team determine the weight of each dimension towards the overall data quality score. Some of the dimensions were combined or removed. For example, granularity and interpretability were removed and interpretability was combined with usability. The team then used the ranking weighing method, Sum and Reciprocal, and obtain the following weights for each dimension alongside the type of questions asked for each dimension (Open Data Toronto 2023; Carlos Hernandez 2020):

| Quality Dimension | Weight | Metrics |
|---|---|---|
| **Freshness**<br>Is this dataset up-to-date? | 35% | - Has the data been refreshed on schedule?<br>- Has the data been left unrefreshed for more than 2 years? |
| **Metadata**<br>Is this data well described? | 35% | - Are there metadata missing from the dataset?<br>- Is the contact owner opendata@toronto.ca?<br>- Is the "Learn More" URL a valid URL?<br>- Are data definitions missing? |
| **Accessibility**<br>Is this data easy to access for different kinds of users? | 15% | - Are there any tags on the dataset?<br>- Is the data updated manually or automatically?<br>- Is the data stored as a file, or in the Open Data database? |
| **Usability**<br>How easy is it to work with the data? | 10% | - Do the columns have meaningful names?<br>- Do columns have constant columns? |
| **Completeness**<br>Is there lots of data missing? | 5% | - Does the data consist of more than 50% null values? |

Figure 1: Weight of the selected dimensions for data quality assessment and the metrics for each dimension

For more technical details about how each dimension is calculated for a dataset and the final data quality score, it can be through the following link: https://github.com/open-data-toronto/framework-data-quality/blob/master/data_quality_score.ipynb (Hernandez 2020). However, it is worth noting that each dimension are calculated based on data obtained on the metadata information and the dataset itself. If the data quality score of a dataset has a normalized score of less than 60%, they are given the grade "Bronze", if the normalized score is between 60% and 80%, it is given the grade "Silver", and finally if the normalized score is over 80%, it is given the grade "Gold". The Information & Technology department at Open Data Toronto collected datasets on their portal using the CKAN Datastore API in order to give each of them a data quality score and grade (Open Data Toronto 2023; Carlos Hernandez 2020). The score for each dataset in the portal is recalculated every week by the team (Open Data Toronto 2023).

### 2.3 Variables of Interest

Our variables of interest that we used in our analysis are the following: "accessibility", "completeness", "freshness", "metadata", "usability", and "grade". "Accessibility" is a score from 0 to 1 that indicates the degree that the dataset can be access or not through the Open Data Toronto API, keywords or tags, and automated data pipelines with 1 being that it can not be accessed with the various methods noted and 0 if not at all. "Completeness" is a score from 0 to 1 that indicates how much of the data is missing with 1 being that there is no missing data and 0 being that all data fields are empty or missing. The "freshness" variable is a score from 0 to 1 that indicates how up-to-date the data where a shorter time duration between the recent refresh date and time and the previous refresh date and time before that gives a higher score. The "metadata" variable indicates how complete the following metadata fields are for a dataset from a scale of 0 to 1: Description, Limitations, Topics, Contact Email. The more metadata fields completed, the higher the metadata score. The "usability" variable is a score from 0 to 1 indicating how easy it would be to use the dataset and this is determine by the proportion of meaningful column names or in other words column names with English words in it. However a limitation of the "usability" variables is that it does not consider datasets that are in different languages other than English.

## 3 Results

### 3.1 Grade and accessibility of datasets

As of May 13, 2025, Figure 2 shows that 56% of datasets on the Open Data Toronto portal had a grade of "bronze". Following this, 25% of datasets are graded "gold" and finally 19% of datasets are graded "silver". This means half of the datasets on the Open Data Toronto portal are ranked "bronze". However since all the datasets have an accessibility score of 1, which indicates they are accessible, and the mean accessibility score is 1 by Table 2 as well, it indicates all datasets on the Open Data Toronto portal can be accessed directly using methods like an API, tags or keywords, or automated data pipelines accessing Open Data Toronto's catalogue.

### 3.2 The relationship between completeness and usability scores of datasets

As seen in Figure 3, for all grades, there's a slight positive relationship between the completeness of a dataset on the Open Data Toronto portal and its usability. However, this relationship is more apparent with the datasets that are graded bronze. This means as the completeness score increases, the usability score of the dataset increases. We can also see most of the scores for bronze-graded datasets are more spread out along the completeness score axis compared to gold-graded and silver-graded datasets. This indicates that more bronze-grade datasets

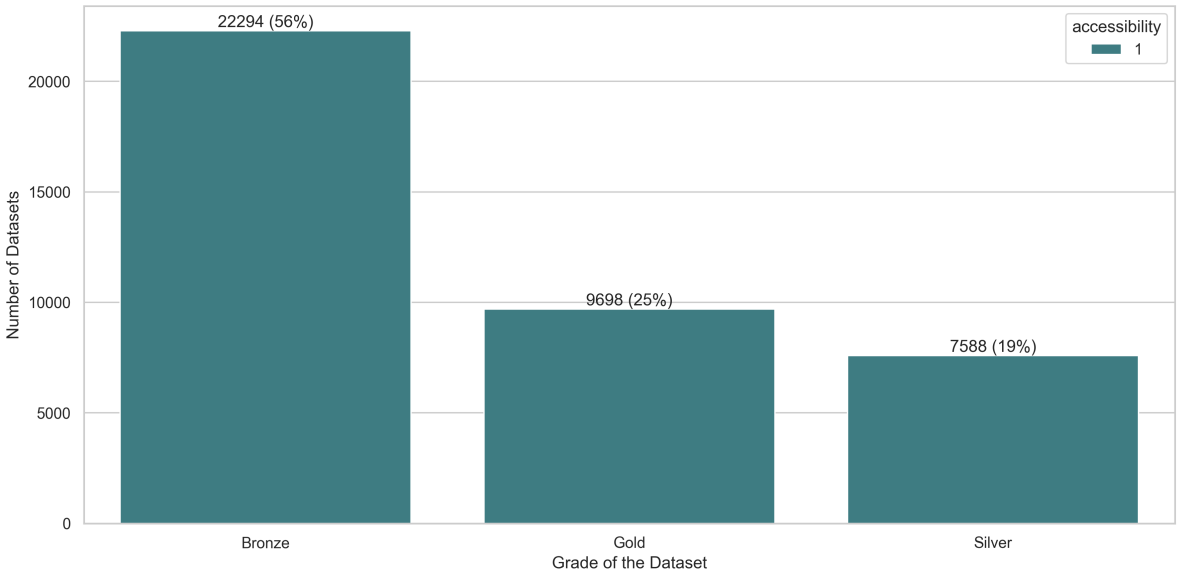


Figure 2: Number of datasets and their accessibility on Open Data Toronto graded bronze, silver, and gold as of May 13, 2025

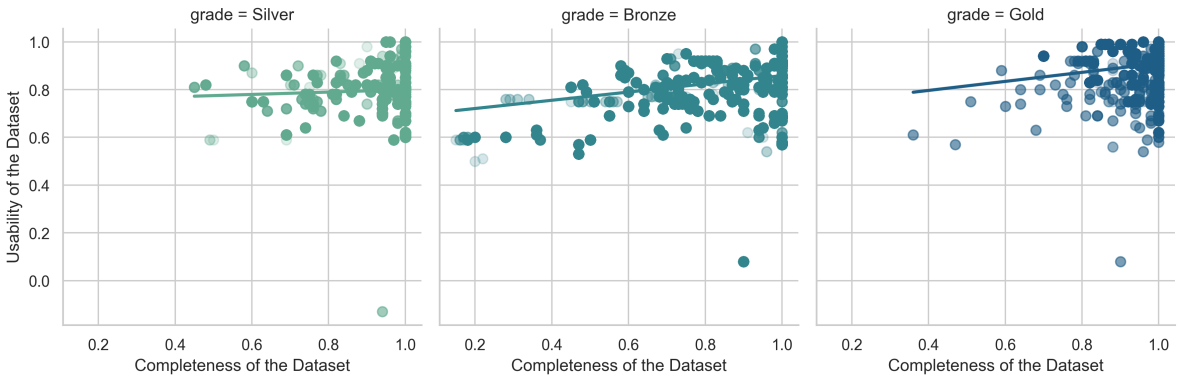


Figure 3: The relationship between completeness scores and usability scores of Open Data Toronto's datasets across different grades as of May 13, 2025

contain more missing data than the other graded datasets. Despite this, Figure 4 shows that the completeness score of datasets on Open Data Toronto skews left with their peaks being above 0.6 (60%), this indicates that across all grades, the datasets have minimal missing data. Table 2 also indicates that the mean values for completeness and usability scores across all datasets are 0.87 (87%) and 0.84 (84%), respectively.

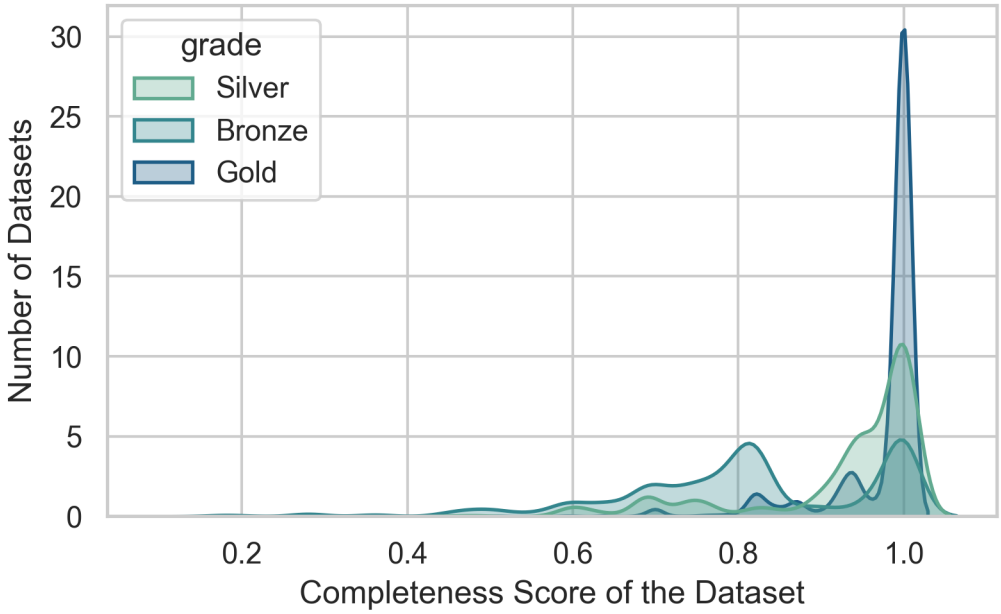


Figure 4: The distribution of completeness scores of Open Data Toronto's datasets across different grades as of May 13, 2025

### 3.3 Metadata completeness scores of datasets

Figure 5 shows that the metadata score for all datasets of Open Data Toronto's datasets has a multimodal distribution. However, the distribution of gold-graded datasets skew left overall. This means that most of the gold-graded datasets have metadata that is almost or is completed filled on the Open Data Toronto portal. On the other hand, the distribution of bronze-graded datasets overall skew right with its largest peak being below a metadata score of 0.5 or 50%. This indicates that the metadata fields for bronze-graded datasets are not sufficiently field or yet not filled out on the Open Data Toronto portal. Table 2 indicates that the mean metadata completeness score is 0.47 (47%) for all datasets on the portal. This means that the average metadata completeness score is below 50% for all datasets on the portal.

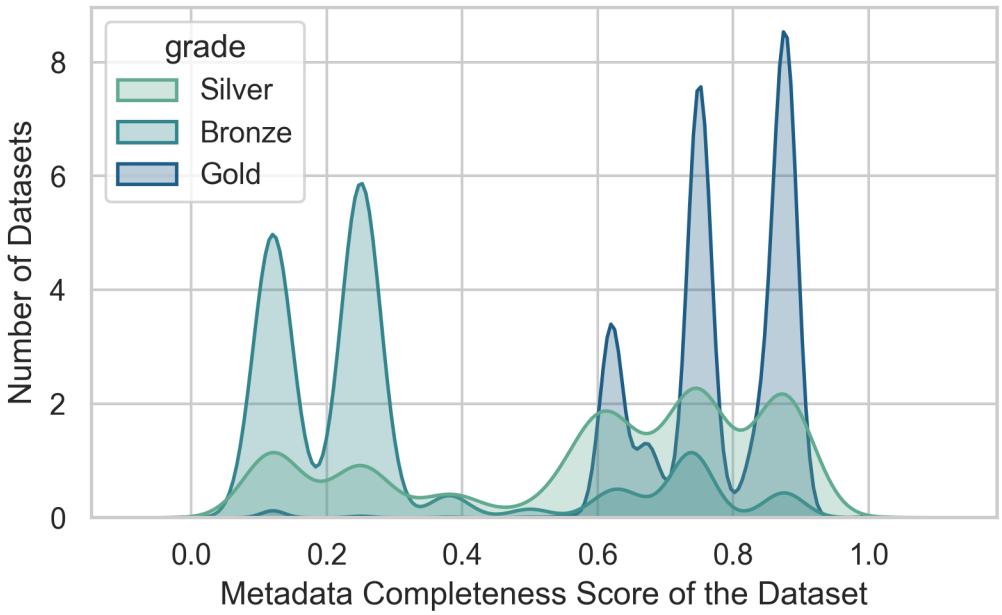


Figure 5: The distribution of metadata completeness scores of Open Data Toronto's datasets across different grades as of May 13, 2025

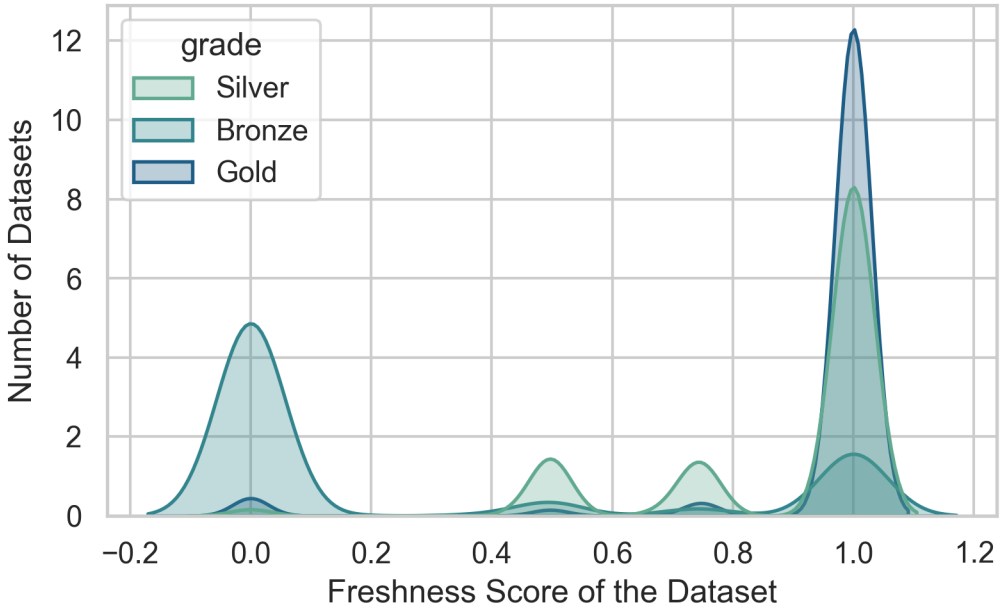


Figure 6: The distribution of freshness scores of Open Data Toronto's datasets across different grades as of May 13, 2025

### 3.4 Freshness scores of datasets

As of May 13, 2025, Figure 6 indicates that for gold-graded and silver-graded datasets, their distributions skews left and that the highest peaks of their distributions are around a freshness score of 1.0 or 100%. This indicates that the datasets that are gold-graded and silver-graded are frequently updated. However with bronze-graded datasets, its distribution skews right with its highest peak being around a freshness score 0.0 or 0%. This indicates that the datasets are not updated frequently or at all. Table 2 also shows that the mean freshness score is 0.56 (56%) across all datasets.

# 4 Discussion

In Section 3, we looked at the data quality of 39,580 datasets on the Open Data Toronto as of May 13, 2025 and the different characteristics of the datasets. We found that our analysis was consistent with what Farrow (2021) found where the metadata and freshness scores of datasets overall was poor but also we found that the low scores were contributed from the bronze-graded datasets.

## 4.1 Majority of Datasets are graded "Bronze"

We saw with that with Figure 2 that 56% of datasets in the Open Data Toronto portal are graded "Bronze". This indicates that 56% of datasets had a data quality score of less than 60%. This also raises concerns regarding the quality of the datasets used in news report for example as well as bring awareness of the quality of datasets currently on the portal. Fortunately based on what Farrow (2021) found in comparison to our results in 2025, there has a decrease in bronze-graded datasets since 2021 from 78% to 56%.

## 4.2 Bronze-graded datasets are less likely to update and have missing metadata

Our results from Figure 5 and Figure 6 shows that bronze-graded datasets have low metadata and freshness scores close to 0 or 0%. This indicates that the bronze-graded datasets contributes to the low metadata and freshness score seen of Open Data Toronto's entire data catalogue. As noted by IBM, Metadata plays an important role in "data governance and data management" (Badman and Kosinski 2024). This means that the lack of metadata for a dataset decreases the experience for users and organization of using the datasets leading to potentially consequences due to issues such as the lack of information of the dataset's limitations and the lack of information about the dataset author's and their contact information.

## 4.3 Areas of improvement

As mentioned in Section 2, datasets have a higher usability score if their column names contains more English words. However, this criteria does not consider datasets that could be still useful but are not in English. Another limitation of our analysis is that the weight of the different dimensions in the data that goes towards the grade of a dataset is subjective in nature since the weighing is based on survey data that had people rank the perceived importance of each dimension.

### 4.4 Next steps

Results from our analysis can be used help the Open Data Toronto team figure out the qualities of bronze-graded datasets that leads to their low scores and also be insight for users of the datasets from Open Data Toronto about what goes behind the grade of the datasets. Future works regarding Open Data Toronto's catalogue can look into the data quality score of the datasets from different divisions.

# A Appendix

## A.1 Acknowledgments

We would like to thank Alexander (2023) for providing assistance with the code used to produce the graphs in this paper. We would also like to thank the team at the IJF for their feedback.

# References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Apache. 2025. *Apache Arrow*. https://github.com/apache/arrow.

Badman, Annie, and Matthew Kosinski. 2024. *What Is Metadata?* https://www.ibm.com/think/topics/metadata.

Carlos Hernandez. 2020. *Towards a Data Quality Score in Open Data (Part 2)*. https://medium.com/open-data-toronto/towards-a-data-quality-score-in-open-data-part-2-3f193eb9e21d.

Farrow, Amy. 2021. *Open Data Quality Is Poor but Slowly Improving*. https://tellingstorieswithdata.com/inputs/pdfs/paper_one-2021-Amy_Farrow.pdf.

Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, et al. 2020. "Array Programming with NumPy." *Nature* 585 (7825): 357–62. https://doi.org/10.1038/s41586-020-2649-2.

Hernandez, Carlos. 2020. *Open Data Toronto: Data Quality Score (DQS)*. https://github.com/open-data-toronto/framework-data-quality/blob/master/data_quality_score.ipynb.

Hunter, J. D. 2007. "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering* 9 (3): 90–95. https://doi.org/10.1109/MCSE.2007.55.

Iannone, Richard, Mauricio Vargas, and June Choe. 2025. *Pointblank: Data Validation and Organization of Metadata for Local and Remote Tables*. https://github.com/posit-dev/pointblank/.

Open Data Toronto. 2023. *Towards an Updated Data Quality Score in Open Data*. https://open.toronto.ca/towards-an-updated-data-quality-score-in-open-data/.

Penrose, Carly. 2024. *Deadly Fires: Risk of Death, Injury Highest in Toronto's Poor Neighbourhoods*. https://www.cbc.ca/news/canada/toronto/fatal-fires-lower-income-1.7177356.

Prewitt, Nate, Ian Cordasco, and Seth Michael Larson. 2011. *Requests*. https://requests.readthedocs.io/en/latest/.

Pydantic. 2025. "Pydantic/Pydantic: Pydantic." https://github.com/pydantic/pydantic.

Python Software Foundation. 2025. *Datetime*.

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

The City of Toronto. 2025. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://open.toronto.ca/.

———. n.d. *City of Toronto Open Data*. https://open.toronto.ca/about/.

The pandas development team. 2020. "Pandas-Dev/Pandas: Pandas." Zenodo. https://doi.org/10.5281/zenodo.3509134.

Van Rossum, Guido, and Fred L Drake Jr. 1995. *Python Tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.

Vink, Ritchie. 2025. "Pola-Rs/Polars: Polars." https://github.com/pola-rs/polars.

Waskom, Michael L. 2021. "Seaborn: Statistical Data Visualization." *Journal of Open Source Software* 6 (60): 3021. https://doi.org/10.21105/joss.03021.