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

June 11, 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, there was a story by the Investigative Journalism Foundation (IJF) and CBC that said people in lower-income Toronto wards had more risk of dying or being injured in fires than people in higher-income wards (Penrose 2024). This story and other stories in the media used data from Open Data Toronto to help support their stories (Penrose 2024). Open Data Toronto is a place where lots of different kinds of data about Toronto are stored, like crime stats and shelter data (The City of Toronto, n.d.). The site is not just for news but is also used when the city makes decisions or policies (The City of Toronto, n.d.). Because of this, it's important to think about how good the data on the site is and to ask: how good is the data that Open Data Toronto provides?

In this paper, we looked at the data quality grades on the site and some other things like "accessibility," "completeness," "freshness," "metadata," and "usability." Farrow talked about how freshness and metadata were not very good (Farrow 2021). But no one has really looked at what these characteristics are like for the different dataset grades in 2025. We found that all datasets are accessible, and most of them don't have a lot of missing data and are usable. We also saw that 56% of datasets were bronze, and these datasets were less likely to be updated and didn't have good metadata. These bronze datasets are probably the reason that freshness and metadata scores were low. These results might help Open Data Toronto know which datasets to focus on fixing and what parts of them—like metadata—they should work on. It might also help people using the portal understand how datasets are graded and what the grades actually mean.

Next in this paper, the data section (Section 2) will show the data we used, how we got it, what it's like, the limits, and what variables we looked at. Then in the results (Section 3) we'll show some graphs. The discussion (Section 4) will go over what we did, what it means, why it matters, and what we could have done better. At the end, the appendix (Section A) has thank-yous and extra stuff.

2 Data

2.1 Overview

The dataset that we used in this paper comes from the Open Data Toronto portal and it's called "Catalogue quality scores" (The City of Toronto 2025). We also looked at other datasets like "Toronto Open Data Intake" but that one didn't really show anything about the quality of datasets people were asking for. The dataset we used is about how good or bad the datasets on Open Data Toronto are, and it helps people know which ones are more useful for stuff like news stories or other city-related things. The datasets in it are scored based on things like accessibility, completeness, freshness, metadata, and usability. These are all put together somehow to make an overall grade. You can see this grade when you look at a dataset page on the portal—it shows up as a trophy icon (The City of Toronto 2025).

To work with the data and do everything in this paper, we used Python (Van Rossum and Drake Jr 1995) and R (R Core Team 2023) and a bunch of different libraries. Some of them were 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). These helped with downloading, cleaning, analyzing, and testing the data and writing the paper too.

We got the dataset by using the Open Data Toronto API (The City of Toronto 2025) and used Requests (Prewitt, Cordasco, and Larson 2011) to download it as a CSV. After cleaning, the dataset had 39,580 rows, and each row is one of the datasets from the catalogue. There's a table later on (see Table 1) that shows what some of the cleaned data 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 something called the "Data Quality Score" to give datasets a grade like "bronze", "silver", or "gold", and you can see this grade if you go on the page for any dataset on the Open Data Toronto website. To make the "Data Quality Score", they made a group called the Data Quality Working Group. This group is made up of a bunch of different people, like people who use the data and also people who make the data (Open Data Toronto 2023; Carlos Hernandez 2020).

They first looked at a bunch of documents and papers, including academic stuff and white papers, and came up with 15 quality things to use to measure how good a dataset is (Open Data Toronto 2023; Carlos Hernandez 2020). The things they picked to focus on were: Interpretability (which is basically how easy it is to understand the data), Usability (how easy to work with), Metadata (is it described well), Freshness (how new it is), Granularity (how detailed), Completeness (if anything is missing), and Accessibility (how easy it is to access) (Open Data Toronto 2023; Carlos Hernandez 2020).

After that, they gave a survey to the group to rank how important each of the things were, like from 1 to 7, where 1 is the most important one and 7 is the one they care about the least. These rankings were used to come up with how much each thing counts toward the score. Some of the things were taken out or mixed together. For example, granularity was taken out, and interpretability was merged into usability. Then they used some math method called Sum and Reciprocal to decide the final weights and made a table showing the weights and the questions they asked (Open Data Toronto 2023; Carlos Hernandez 2020).

If you want to know more technical stuff about how all the dimensions are actually calculated, you can go to this link: https://github.com/open-data-toronto/framework-data-quality/blob/master/data_quality_score.ipynb (Hernandez 2020). But basically, the score is based on things in the metadata and the actual dataset. If the total score is below 60%, the dataset gets a Bronze grade, and if it's between 60% and 80% it gets Silver, and above 80%

it gets Gold. The Open Data Toronto team uses CKAN's API to collect all the datasets and score them (Open Data Toronto 2023; Carlos Hernandez 2020). These scores get updated every single week (Open Data Toronto 2023).

2.3 Variables of Interest

The variables we looked at in our analysis were: "accessibility", "completeness", "freshness", "metadata", "usability", and "grade". "Accessibility" is a number from 0 to 1 that tells how easy it is to get the dataset using the API, keywords, or other stuff like that. If the score is 1, then it's good, if it's 0, then it's not. "Completeness" is also 0 to 1 and tells how much data is missing. A 1 means nothing is missing, and a 0 means all of it is missing. "Freshness" is how new the data is, and it checks the time between updates—shorter time = higher score. "Metadata" looks at how many of the metadata fields are filled in, like the description, limitations, contact email, and topics. The more fields filled out, the better. "Usability" is also from 0 to 1 and tells if the dataset is easy to use by checking how many column names are actual English words. One issue is that this doesn't work well if the dataset is in a different language, which kind of messes up the score.

3 Results

3.1 Grade and accessibility of datasets

As of May 13, 2025, Figure 1 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 2, 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 contain more missing data than the other graded datasets. Despite this, Figure 3 shows that

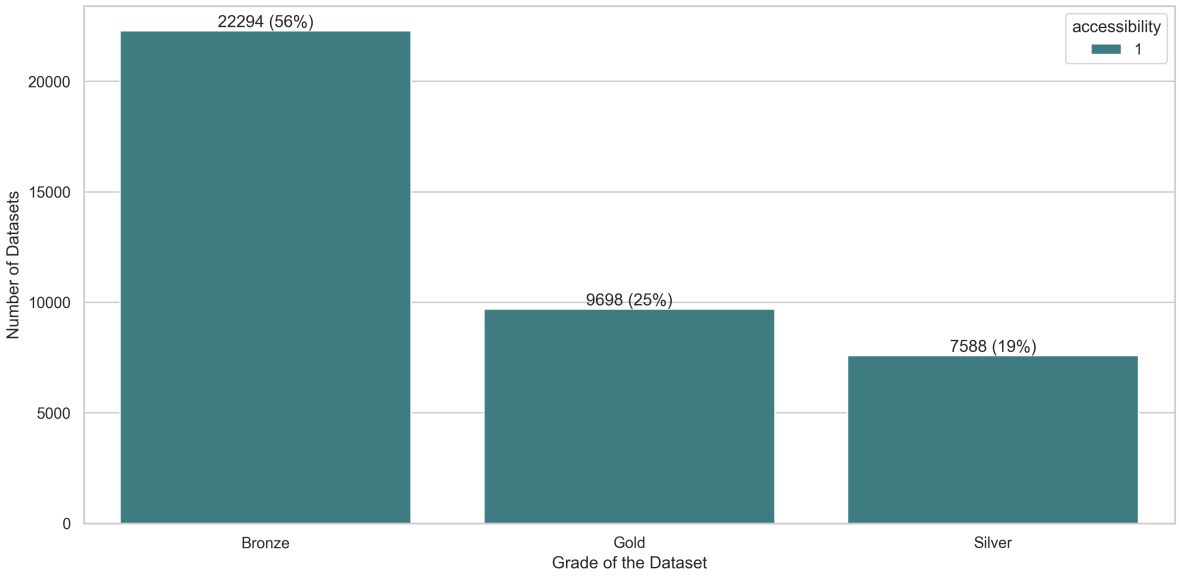

Figure 1: Number of datasets and their accessibility on Open Data Toronto graded bronze, silver, and gold as of May 13, 2025

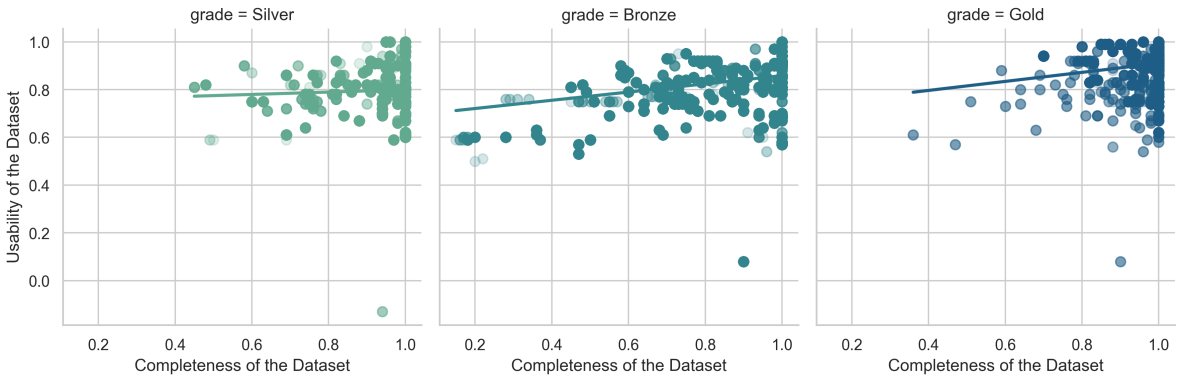

Figure 2: The relationship between completeness scores and usability scores of Open Data Toronto's datasets across different grades as of May 13, 2025

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 3: 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 4 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.

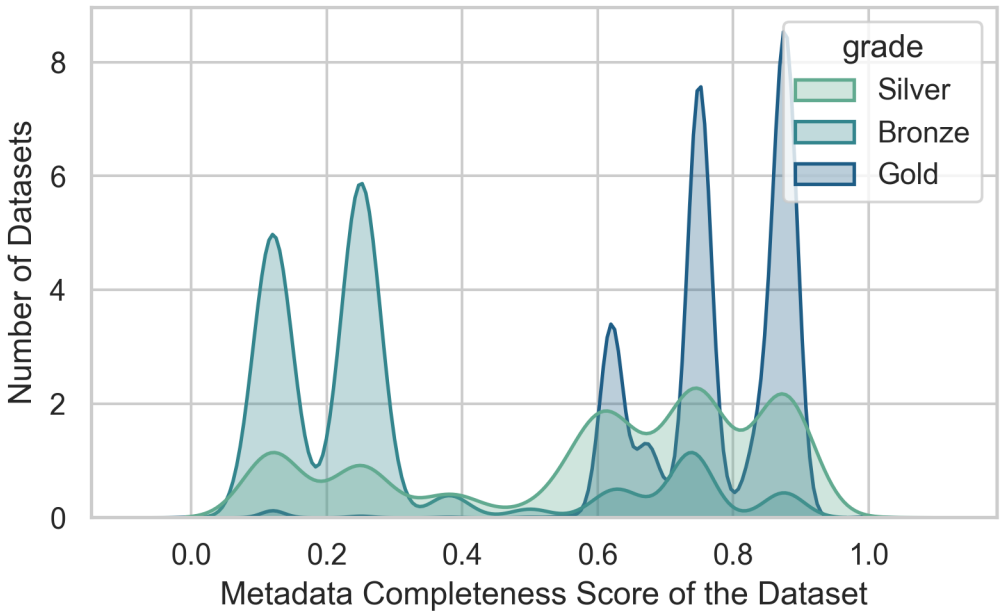

Figure 4: The distribution of metadata completeness scores of Open Data Toronto's datasets across different grades as of May 13, 2025

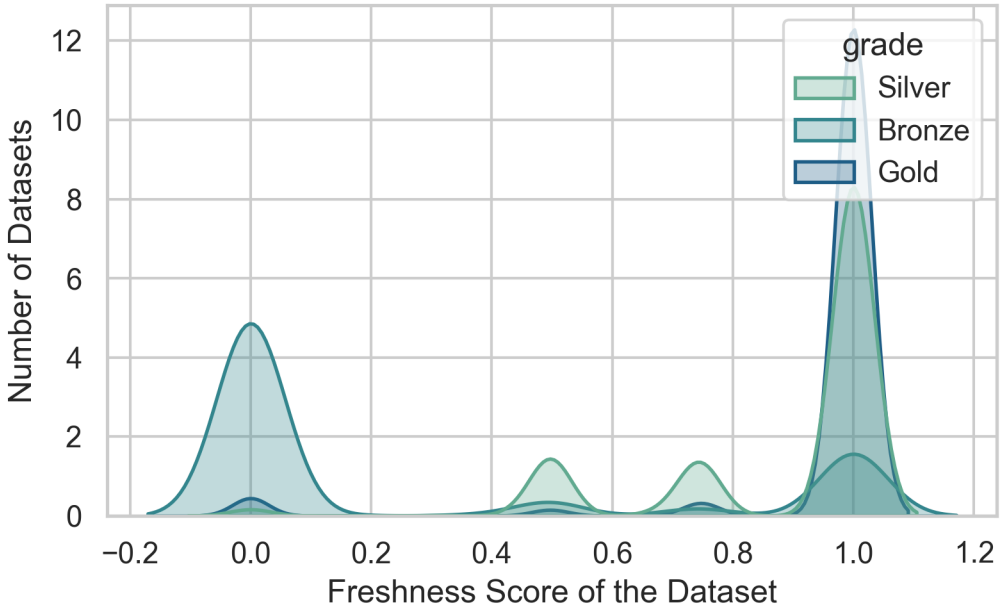

Figure 5: The distribution of freshness scores of Open Data Toronto's datasets across different grades as of May 13, 2025

3.4 Freshness scores of datasets

As of May 13, 2025, Figure 5 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 1 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 4 and Figure 5 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.