Accessible and Usable Datasets yet Majority are Bronze-Graded and not Updated Regularly*

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

Emily Su

May 14, 2025

As one of the central hubs for data on the City of Toronto that are used in the media and in policy-making, we conducted analysis on the quality of Open Data Toronto's data catalogue. We found that despite Open Data Toronto's extensive dataset catalogue being accessible and usable, 56% of their datasets are graded bronze and bronze-graded datasets are less likely to be updated the more completed they are. However we also found that metadata fields were less likely to be filled for gold-graded and silver-graded datasets. These findings can help reporters, policy-makers, and anyone interested in using datasets from Open Data Toronto's catalogue make an inform decision of what datasets to choose.

Table of contents

1 Introduction 2

2 Data 2

2.1 Overview . . . 2
2.2 Measurement . . . 3
2.3 Variables of Interest . . . 3

3 Results 3

3.1 Grade of datasets . . . 3
3.2 Accessibility scores of datasets . . . 3
3.3 The relationship between completeness and usability scores of datasets . . . 5
3.4 Completeness versus metadata completion scores of datasets . . . 5

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

3.5 Completeness versus freshness scores of datasets 7

A Appendix 8

A.1 Acknowledgments 8

References 9

1 Introduction

In the analysis of our paper, we looked at the following questions: What is the quality of the datasets on the Open Data Toronto portal? What are the features of different types of datasets on Open Data Toronto?

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 and our variables of interest for our analysis. In the results section (Section 3) we looked at the data more in-depth through graphs. 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 to be used for various situations like 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 that 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 (Nate Prewitt 2011), datetime (Python Software Foundation 2025), Matplotlib (Hunter 2007), numpy (Harris et al. 2020), pandas (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 (Nate Prewitt 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. The cleaned dataset look as follows Table 1:

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

2.2 Measurement

The Information & Technology department at Open Data Toronto collected (Open Data Toronto 2023)

2.3 Variables of Interest

Our variables of interest are "accessibility", "completeness", "freshness", "metadata", "usability", and "grade".

3 Results

3.1 Grade 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".

3.2 Accessibility scores of datasets

As of May 13, 2025, Figure 2 indicates 22294 bronze-graded datasets on the Open Data Toronto portal scored 1 for accessiblity. For gold-graded datasets, it was 9698 datasets and for silver-graded datasets, it was 7588 datasets.

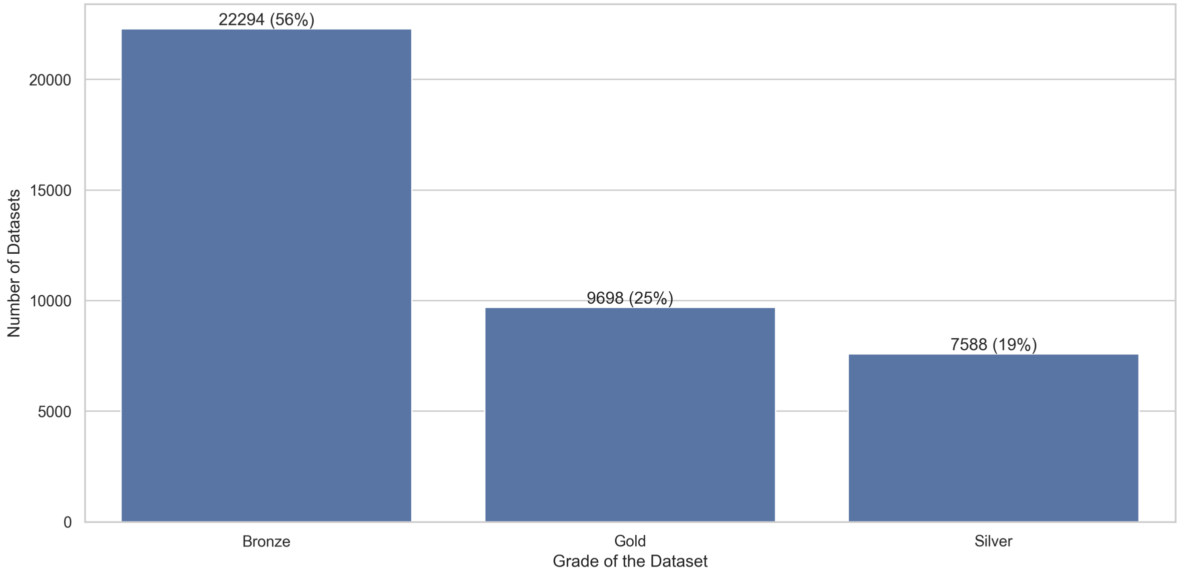

Figure 1: Number of datasets on Open Data Toronto graded bronze, silver, and gold as of May 13, 2025

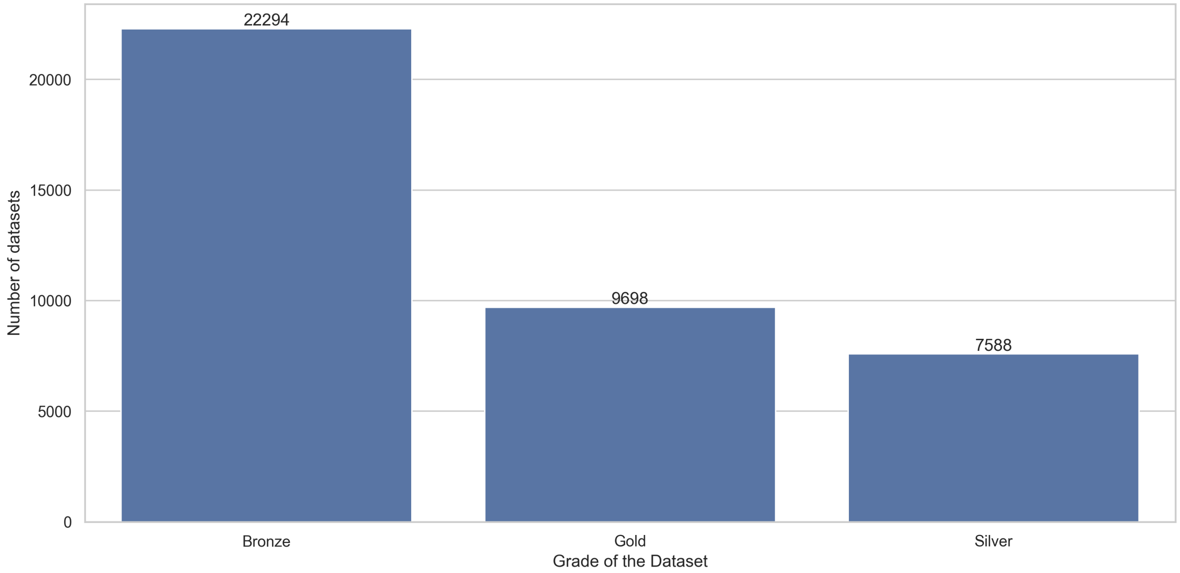

Figure 2: Number of datasets with accessiblity score of 1 on Open Data Toronto as of May 13, 2025

3.3 The relationship between completeness and usability scores of datasets

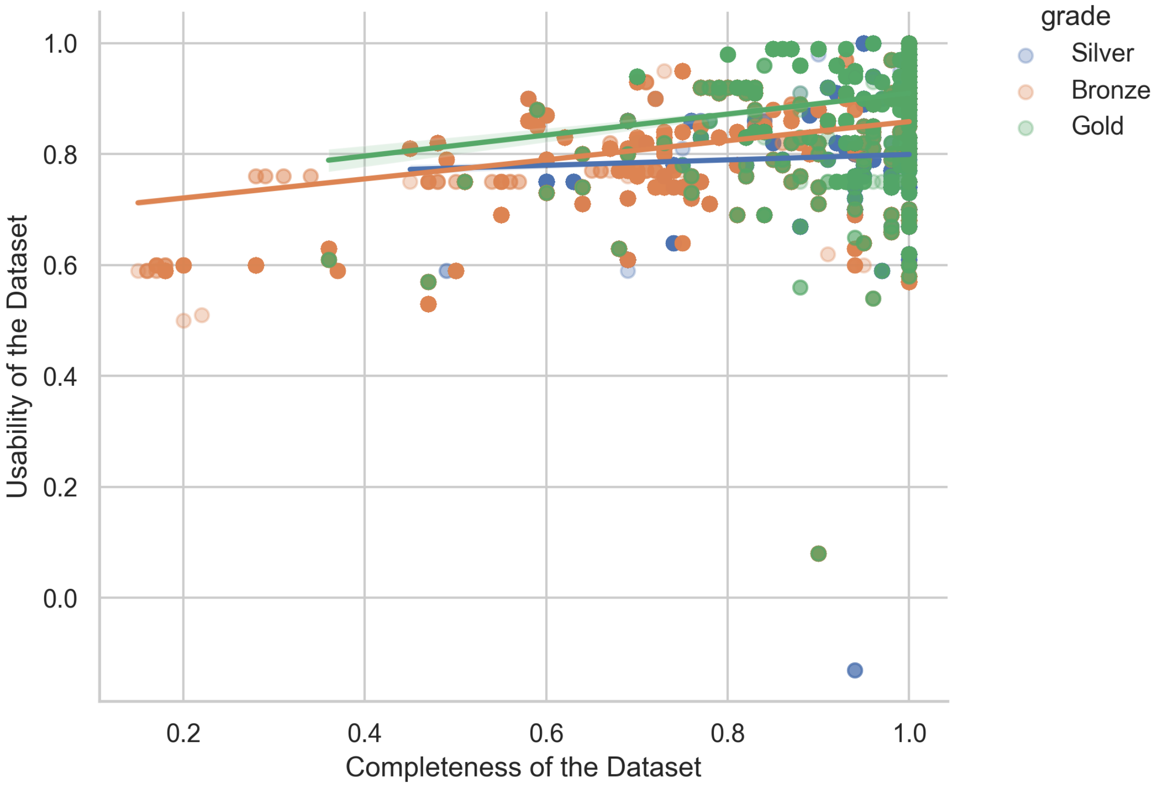

Figure 3: The relationship between completeness scores and usability scores of datasets on Open Data Toronto across different grades as of May 13, 2025

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.

3.4 Completeness versus metadata completion scores of datasets

In term of the relationship between the completeness score and the metadata completion score of datasets on the Open Data Toronto portal in Figure 4, there is a slight positive relationship for bronze-graded datasets while there is a negative relationship for silver-graded and gold-graded datasets. This means that for bronze-graded datasets, as the completeness score increasing, the metadata score increases. However for gold-graded and silver-graded datasets, as the completeness score increases, the metadata score decreases.

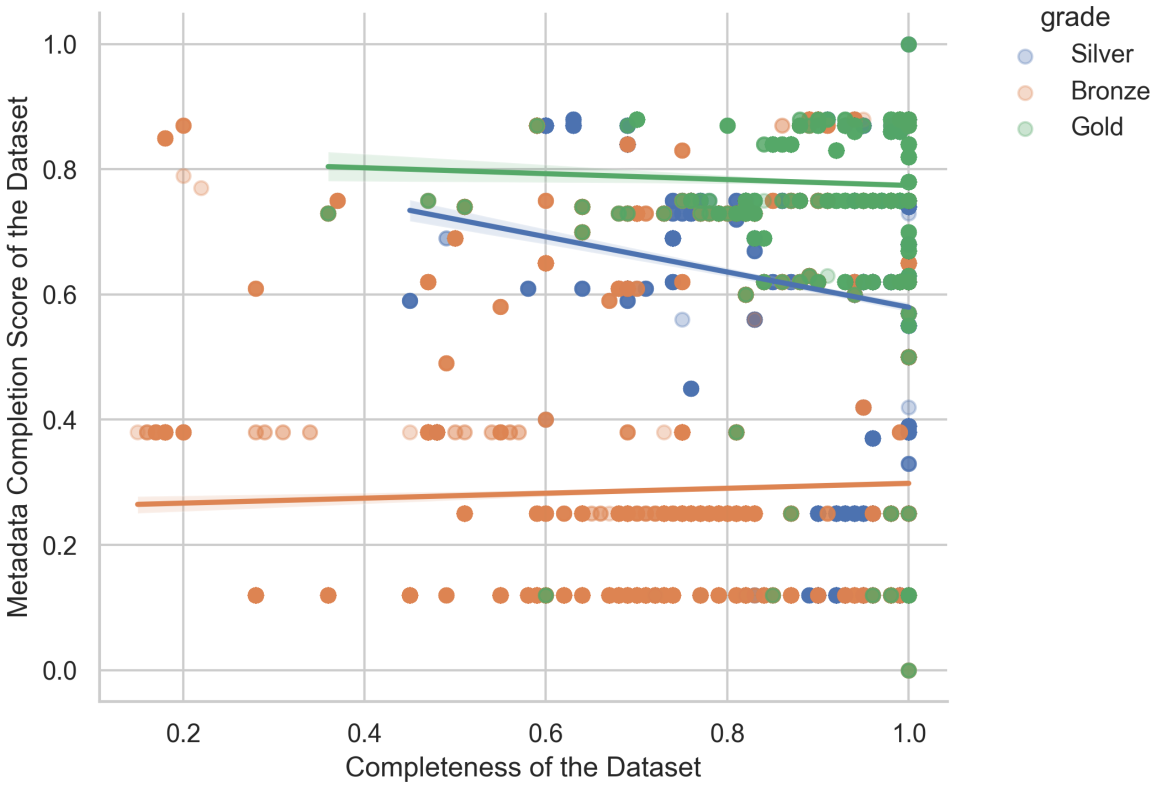

Figure 4: The relationship between completeness scores and metadata completion scores of datasets on Open Data Toronto across different grades as of May 13, 2025

3.5 Completeness versus freshness scores of datasets

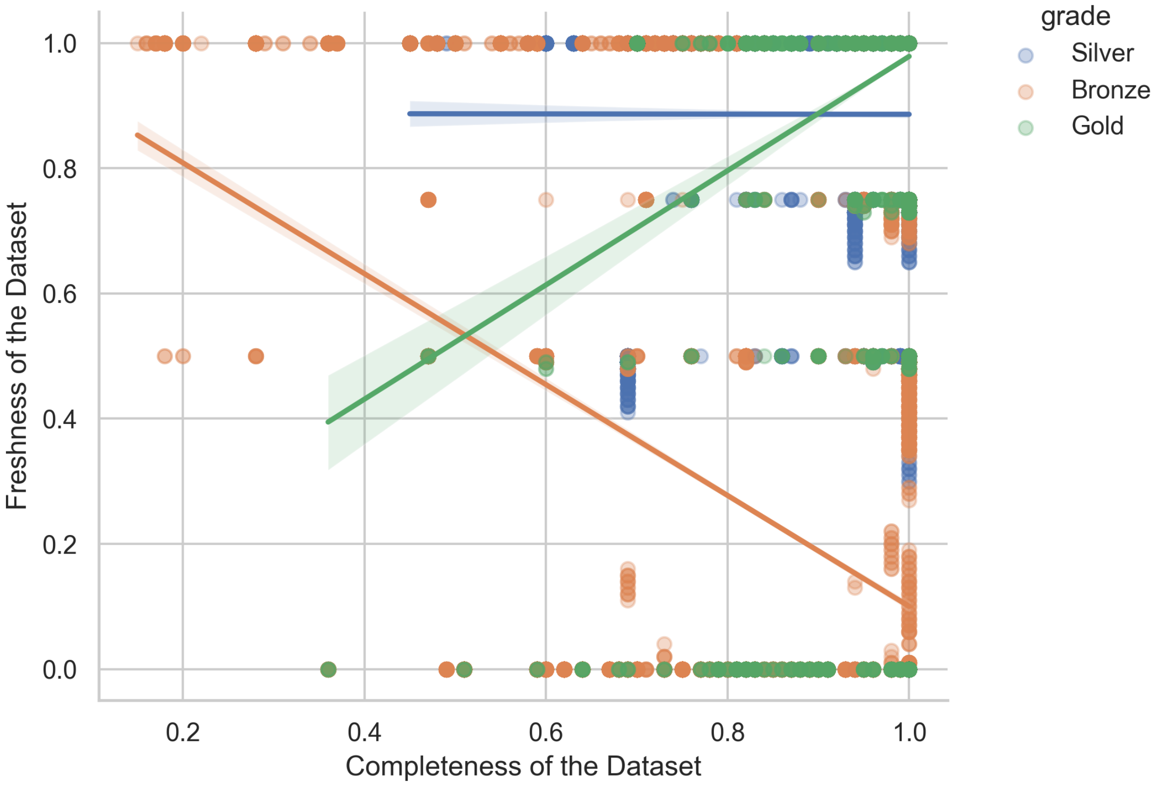

Figure 5: The relationship between completeness scores and freshness scores of datasets on Open Data Toronto across different grades as of May 13, 2025

As of May 13, 2025, when we look at the relationship between completeness score and freshness score in Figure 5, there is a negative relationship for bronze-graded datasets, positive relationship for gold-graded datasets, and there is almost no relationship between the two for silver-graded datasets. This means that for bronze-graded datasets as the completeness score increases, the freshness score decreases while for gold-graded datasets as the completeness score increases, the freshness score increases. However with silver-graded datasets, the freshness score doesn't change as the completion score increases.

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.

References

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

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

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.

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/.

Nate Prewitt, or Seth Michael Larson, Ian Cordasco. 2011. *Requests*. https://CRAN.R-project.org/package=opendatatoronto.

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/.

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/.

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

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

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.