

# Data Science 1 - HW 1

---

**Date:** August 30, 2025

---

## MTA Art Catalogue

### 1) What is the dataset?

#### MTA Permanent Art Catalog

"Through the Permanent Art Program, MTA Arts & Design (formerly Arts for Transit) commissions public art that is seen by millions of city-dwellers as well as national and international visitors who use the MTA's subways and trains. Arts & Design works closely with the architects and engineers at MTA NYC Transit, Long Island Rail Road and Metro-North Railroad to determine the parameters and sites for the artwork that is to be incorporated into each station scheduled for renovation. A diversity of well-established, mid-career and emerging artists contribute to the growing collection of works created in the materials of the system -mosaic, ceramic, tile, bronze, steel and glass. Artists are chosen through a competitive process that uses selection panels comprised of visual arts professionals and community representatives which review and select artists. This data provides the branch or station and the artist and artwork information."

New York State. (n.d.). MTA Permanent Art Catalog (Beginning 1980) [Dataset].

[https://data.ny.gov/Transportation/MTA-Permanent-Art-Catalog-Beginning-1980/4y8j-9pkd/about\\_data](https://data.ny.gov/Transportation/MTA-Permanent-Art-Catalog-Beginning-1980/4y8j-9pkd/about_data)

I selected this dataset because I love NYC, and I love art! Art in public spaces is especially meaningful to me in that it can draw us out of our tiny rectangles filled with metrics and ads, and remind us of the beauty and creativity that exist in our fellow humans. Also, it can really spice up the concrete!

### 2) Outside of the dataset

#### File Structure:

- Single file dataset: [MTA\\_Permanent\\_Art\\_Catalog\\_\\_Beginning\\_1980\\_20250829.csv](#)
- No compression applied
- The license for this dataset is unspecified.
- There was no copyright information provided in the documentation. However, it is stated that Metropolitan Transportation is the owner. There is also a terms of use located [here](#).
- Standalone file with no accompanying documentation files

#### File Format & Size:

- **Format:** CSV (Comma-Separated Values)
- **File size:** 523.4 KB
- **Data dimensions:** 381 rows × 9 columns

### 3) Loading the dataset

I [exported](#) the CSV and moved it into the appropriate folder and used `read_csv` provided by the pandas library.

```
df = pd.read_csv('./MTA_Permanent_Art_Catalog__Beginning_1980_20250829.csv')
```

#### 4) Inside the dataset

This dataset is beautifully documented [here](#) in the "What's in this Dataset?" section. In the table provided, columns map to variables. Descriptions and data types are provided as well.

#### 5) Questions

Questions I think this dataset could at least begin to answer are as follows:

- What lines contain the most art?
- Do lines favor any particular artist or material?

Here is something fun I looked into. I wanted to know if delayed trains had less art. To be clear, I am not suggesting any correlation or causation. I simply wanted to know if those that had to wait had less art to look at. This dataset alone could not answer this question, but I did some internet sleuthing. The good news is this does not seem to be the case!

According to the [nypost](#), the B train was the most delayed in 2024. There are 26 pieces along the B line!

According to [lavocedinewyork](#), the F line was the most delayed in 2023 and it had 13 pieces.

This puts both of these tardy lines in the top 5 in terms of number of installations!