# Subway Hourly Ridership Data Quality Report

## Overview

This report will outline the initial data quality findings on the dataset 'MTA\_Subway\_Hourly\_Ridership\_\_Beginning\_February\_2022\_20240522.csv', obtained from <https://data.ny.gov/Transportation/MTA-Subway-Hourly-Ridership-Beginning-February-202/wujg-7c2s/about_data>. It will include an overview of the dataset, and a review of the continuous and categorical features, including histograms and bar charts. On initial review, this dataset is clean and well documented. There is no missing data, and the data appears to be reasonable and logical.

## Summary

This dataset consists of ridership data by station, hour of the day and payment method. There are no missing values in any of the columns. There are no duplicate rows. Distribution of the data is consistent with expectations, and there appear to be few outliers in this dataset. This appears to be a clean dataset, and is excellent data to use for busyness calculation purposes.

## Review Logical Integrity

Test 1: No timestamps are all before Feb 2022 or after the upload date (22nd May).

* 0 instances.

Test 2: No valid timestamp-station\_complex\_id combination has 0 rows in the dataset.

* 8,619,539 / 65,910,871 timestamp – station\_complex\_id combinations are missing.

## Review Continuous Features

### Descriptive Statistics

There are 4 continuous features in this dataset:

* Transit Timestamp
  + This has an even distribution across all months. There are 19,891 unique timestamps in this dataset, which is slightly less than the 19,895 hours that are between the 1st Feb 2022 and 9th May 2024. This means there are 4 hours that has 0 subway ridership data.
* Ridership
  + Mean ridership is 44, with a min of 1 and a max of 14243. From this data it is clear that stations with no ridership in a given hour do not get a line on the table. The outlier of 14243 is expected. This was likely a busy station during a large event.
* Transfers
  + Transfers have a mean of 1.85, a min of 0 and a max of 1242. This is in line with expectations.
* Latitude and longitude
  + Latitude and longitude have a standard distribution. The mean value of 40°43'56.3"N 73°56'06.6"W points directly to the center of New York.

### Histograms

All Histograms can be found in the appendix as a summary sheet. All features show a plausible distribution.

## Review Categorical Features

There are 7 categorical features in the dataset:

* Transit mode
  + This has 3 possible values; subway, railway, and tram. Large majority of trips were taken by subway.
* Station complex id
  + There are 854 unique values. There is a fairly even distribution across all values.
* Station complex
  + There are 428 values, almost exactly half of the station\_complex\_id values. Similar to station complex ID, this has an even distribution across all values. The magnitude is higher than station complex ID as there are multiple IDs per station.
* Borough
  + Data from 5 boroughs. Staten island has the lowest proportion of datapoints, while Brooklyn has the highest.
* Payment Method
  + Two payment methods, metrocard and omny. Metrocard is more frequent than omny.
* Fare Class Category
  + 10 possible values. Metrocard – Full Fare is the most popular, while OMNY – Other is the least popular. This is in line with the distribution of payment methods.
* Georeference
  + There are 976 values. Georeference has the same distribution as station\_complex\_id, which is consistent with expectations. There appear to be more values than station complex ID. This may be due to station complexes moving over time.

## Actions to take

1 action will be taken:

* Missing time – station ID complex combinations
  + Missing time - station ID complex combinations will be added in as new rows with the mode longitude, latitude, station ID etc values, and 0 as ridership and transfer numbers.

# Appendix

## Continuous Features

### Descriptive Statistics



### Histograms

A group of blue bars

Description automatically generated with medium confidence

## Categorical Features

### Descriptive Statistics



### Box Plots

A collage of graphs

Description automatically generated