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.

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 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.

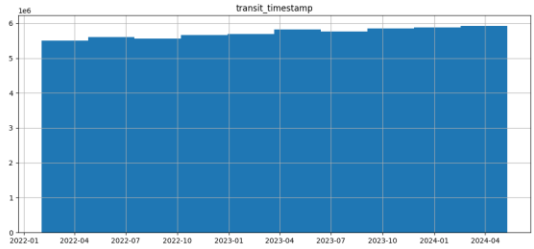
Appendix

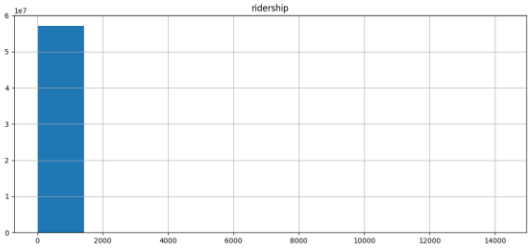
Continuous Features

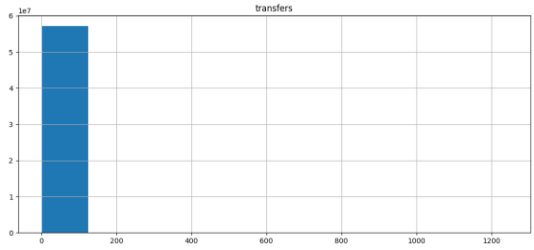
Descriptive Statistics

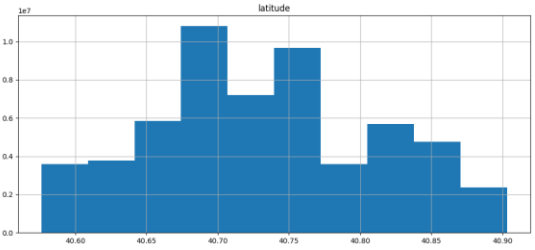
Feature	count	mean	min	25%	50%	75%	max	std	%missing	card
transit_tir	57291332	23:39.4	01/02/2022 00:00	02/09/2022 19:00	31/03/2023 02:00	21/10/2023 14:00	09/05/2024 23:00		0	19891
ridership	57291332	44.194172	1	4	12	35	14243	142.05360	0	7096
transfers	57291332	1.8580036	0	0	0	1	1242	11.523552	0	1032
latitude	57291332	40.732297	40.576126	40.677315	40.72433090209961	40.79164123535156	40.903126	0.078356	0	926
longitude	57291332	-73.9352	-74.07484	-73.98123169	-73.94747925	-73.89948	-73.7554	0.056157	0	927

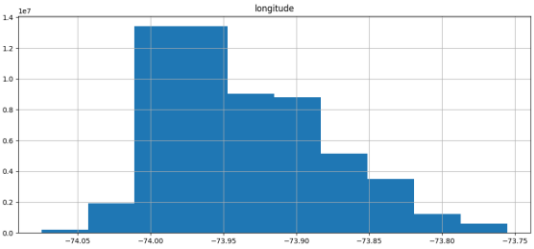
Histograms









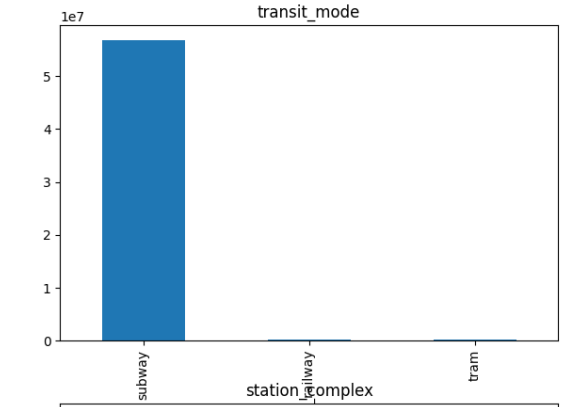


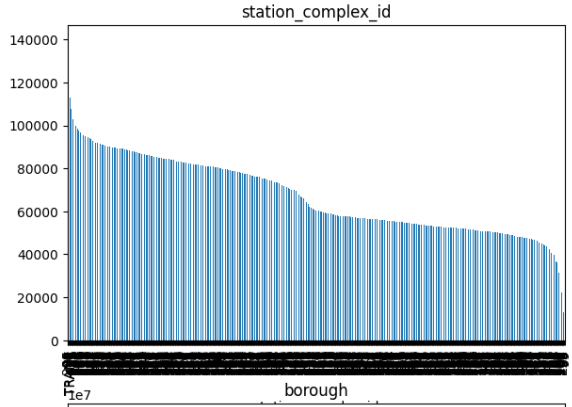
Categorical Features

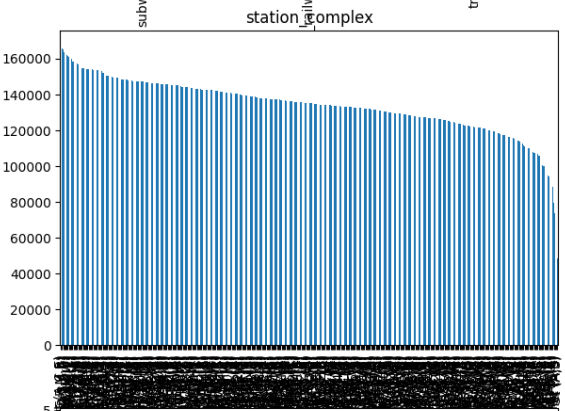
Descriptive Statistics

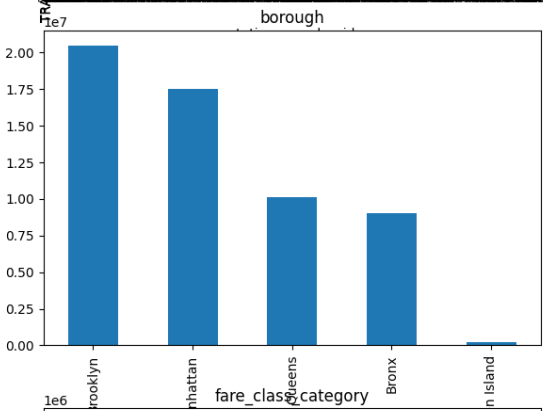
Feature	count	unique	top	freq	second	second_freq	%missing	card
transit_mode	57291332	3	subway	56851724	staten_island_railway	222855	0	3
station_complex_id	57291332	854	225	139715	98	138916	0	854
station_complex	57291332	428	Times Sq-42 St (N,Q,R,W,S,1,2,3,7)/42 St (A,C,E)	167286	74-Broadway (7)/Jackson Hts-Roosevelt Av (E,F,M,R)	165697	0	428
borough	57291332	5	Brooklyn	20454130	Manhattan	17492836	0	5
payment_method	57291332	2	metrocard	47880503	omny	9410829	0	2
fare_class_category	57291332	10	Metrocard - Full Fare	7950433	OMNY - Full Fare	7920107	0	10
Georeference	57291332	976	POINT (-73.83406066894531 40.668235778808594)	142299	POINT (-73.97684478759766 40.75177764892578)	116171	0	976

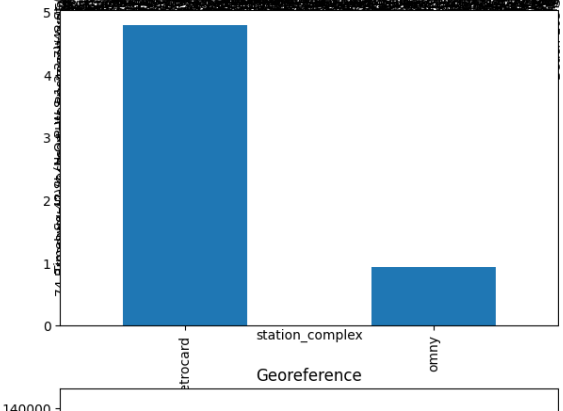
Box Plots

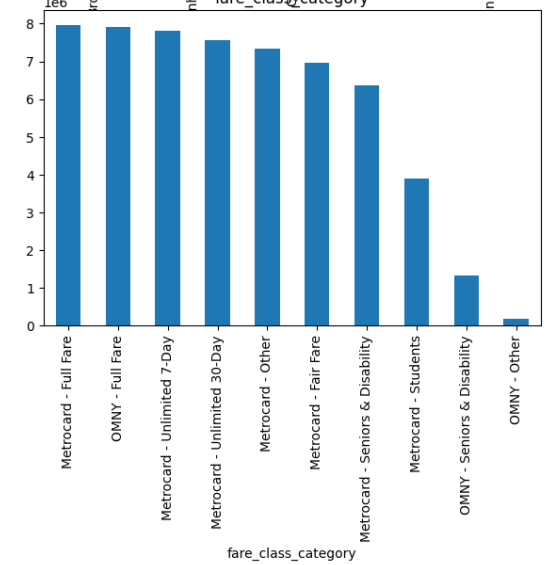

transit_mode
1e7
subway
railway
tram

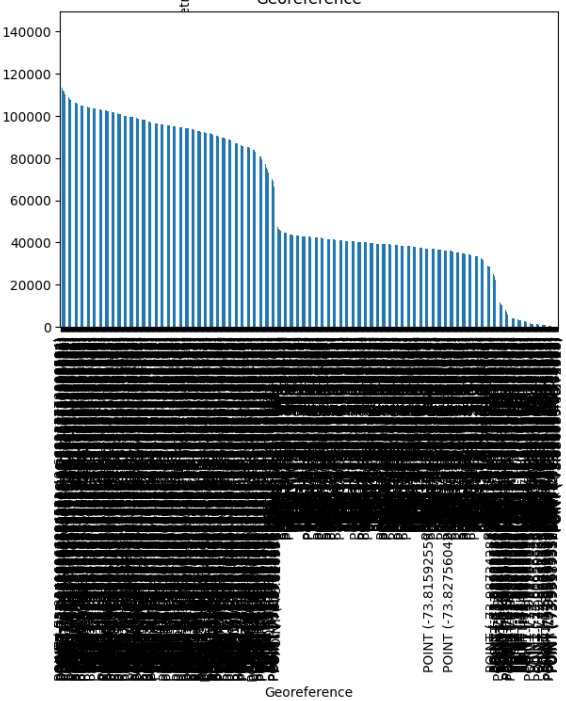

station_complex_id


station_complex


borough
1e7
Bronx


station_complex
omny


fare_class_category
1e6
Metrocard - Full Fare
OMNY - Full Fare
Metrocard - Unlimited 7-Day
Metrocard - Unlimited 30-Day
Metrocard - Other
Metrocard - Fair Fare
Metrocard - Seniors & Disability
Metrocard - Students
OMNY - Seniors & Disability
OMNY - Other
fare_class_category


Georeference
POINT (-73.81592558
POINT (-73.82756048
Georeference