Data Quality Report - Initial Findings

1. Overview

Introduction to Data Quality Report for TLC Taxi Dataset.

This data quality plan for our dataset aims to ensure that the information we gather, process, and analyse meets the highest standards of quality which can be used for predictive modelling. This report outlines the strategies and methodologies that will be employed to assess, monitor, and enhance the quality of the data related to taxi ride trip data gathered from TLC.

Our dataset includes information on taxi trips around New York city from 01-2023 to 12-2024. It provides us informative features such as pickup times, drop off times, pickup locations, drop off locations, date and time of trips and most importantly the passenger count of each trip.

2. Summary

The parquet files containing monthly trip data were all joined together to create a dataset with 89 million rows of data. In order to be able to clean and view this dataset we must find ways of condensing the dataset while maintaining important information.

We may combine the pickup and drop off locations together in order to improve our predictive accuracy. We are not interested in one over the other as they both tell us important information about the areas. Analysing pickup locations helps understand commuting patterns, popular destinations, and areas of high activity, which can be correlated with events, attractions, or daily routines. Analysing drop-off locations can help identify popular hotspots, areas of high tourist activity, and places with significant economic and social activity. By combining pickup and dropoff locations, we can analyse the overall taxi activity in Manhattan, including both the origins and destinations of trips, equally.

Analysing the null/missing/0 values of our dataset, we can see that only passenger_count contained any null values (around 4%). We decided that the best course of action would be to replace these rows with the median value of passenger_count feature. This way we can still use these rows within our dataset and do not need to remove them.

3.2.Visualise Data

All histograms and boxplots can be found on the appendix. Outliers can evidently be seen in the passenger_count box plot also. However, the points are only seen as ‘outliers’ because

of the huge percentage of data with value 1 and will not skew our data as we will be combining all passenger_counts together by taxi_zone.

4.1. Descriptive Statistics

There are 5 categorical and one continuous feature in the dataset. They can essentially be divided into a few groups.

Time:

- Hour: The hour of the day in 24 hour standard (e.g. 13)
- Day: Day of the month (e.g. 2 - 2nd of the month)
- Week: Day of the week (e.g. Saturday)
- Year_Month: Year and Month of the year (e.g. 01-2022 - January 2022)

Taxi Zone:

Shows the TaxiZoneID number for each zone in which the trips occurred.

Passenger Count:

Shows the number of passengers in the taxi zone within that time frame.

5. Action to take

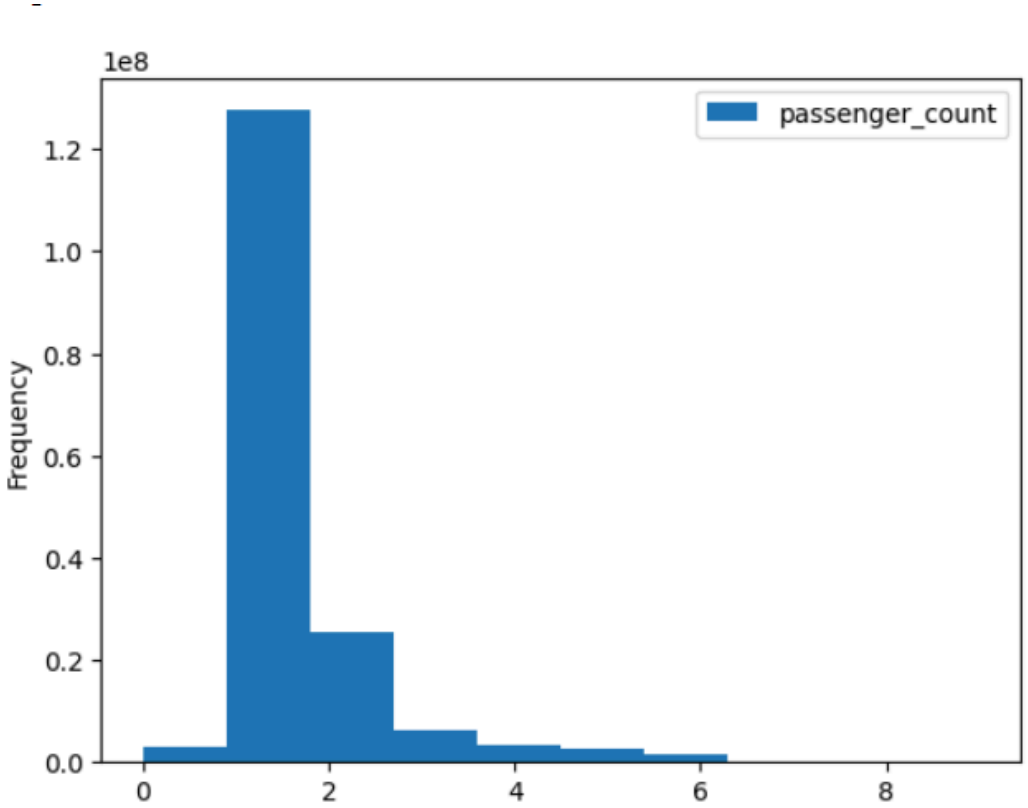
Passenger Count:

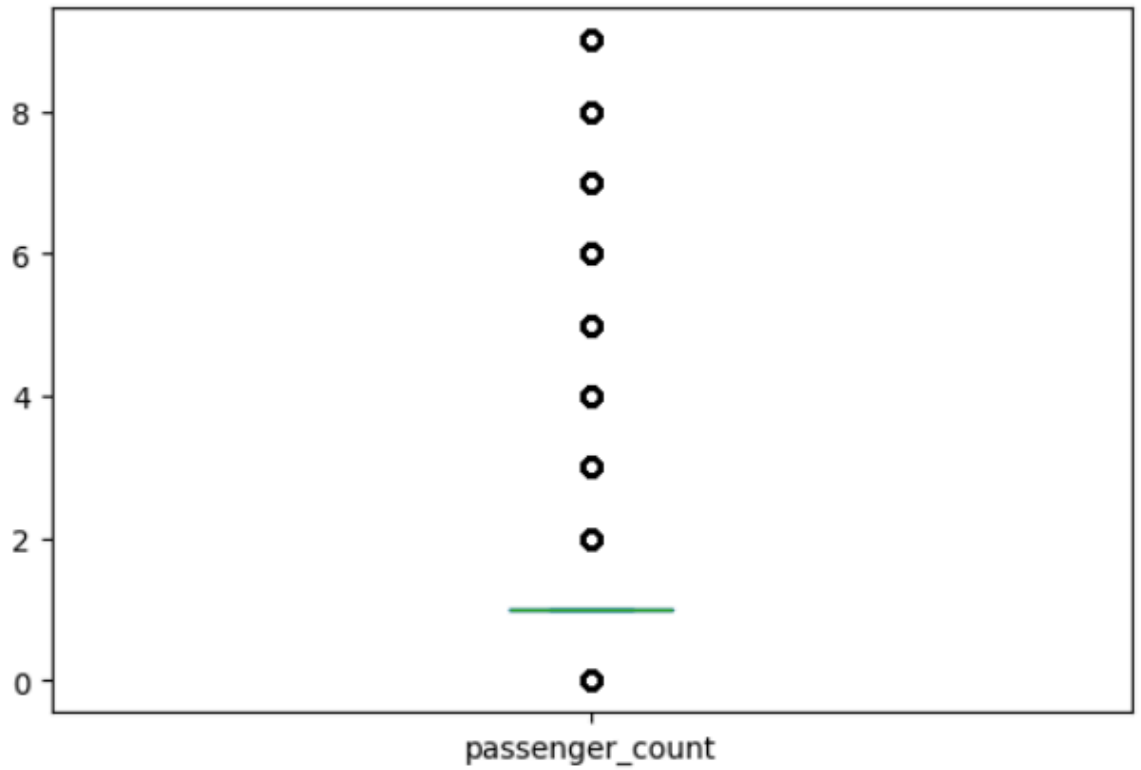
Replace the null values with the median value.

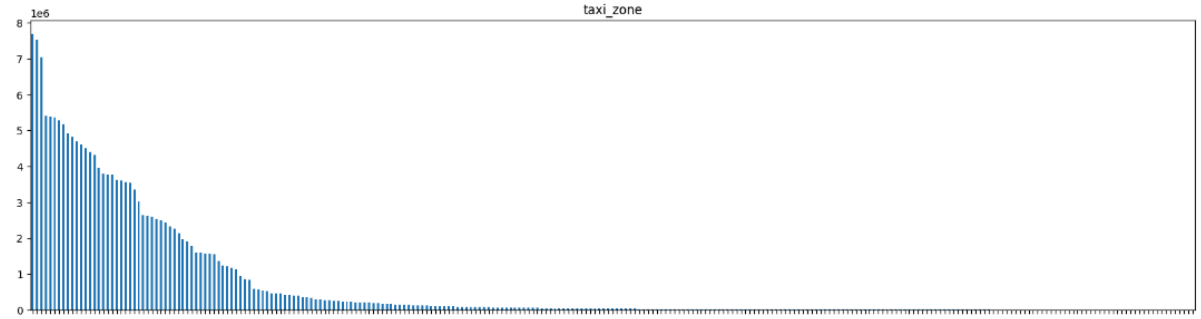
Combine Features:

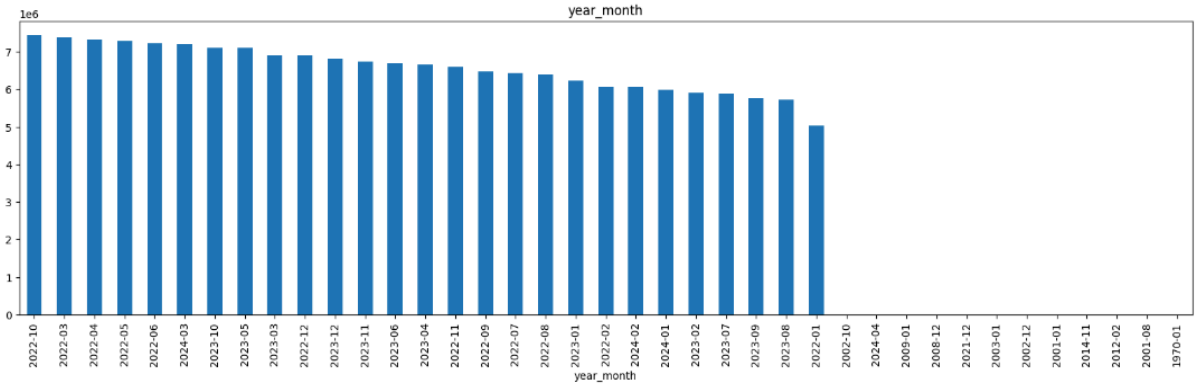
Combine the existing features from the raw dataset to condense the data into useful features in which we can extract useful information which may help our predictions.

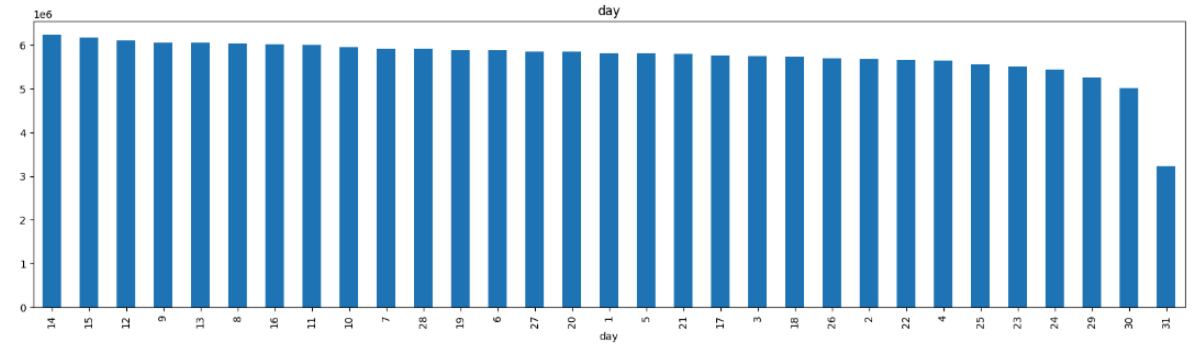
6. Appendix

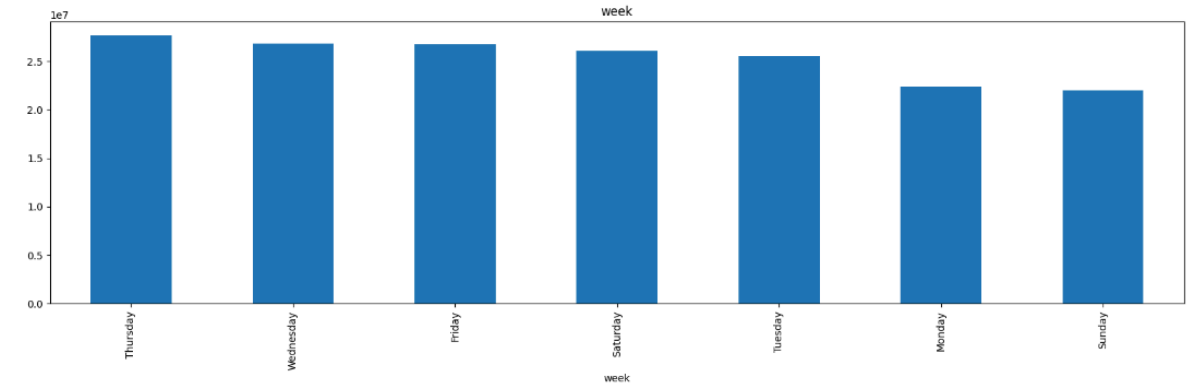

1e8
passenger_count
Frequency
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8

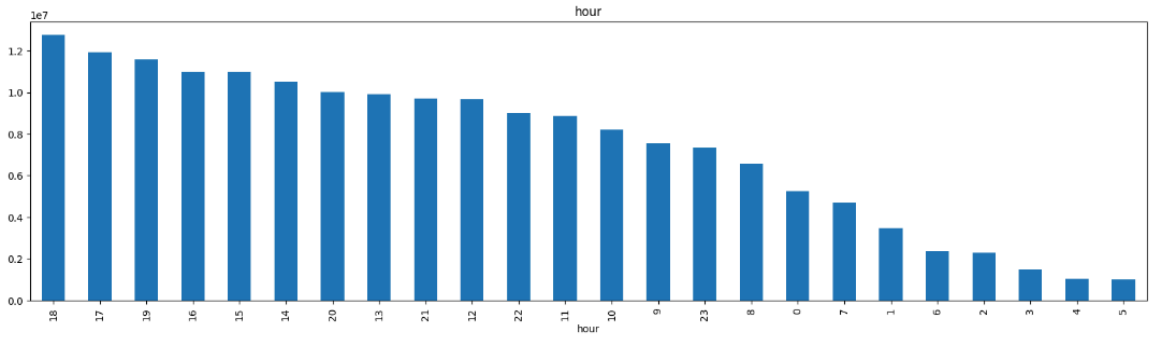

8
6
4
2
0
passenger_count


taxi_zone
1e6
8
7
6
5
4
3
2
1
0


year_month
1e6
7
6
5
4
3
2
1
0
2022-10
2022-03
2022-04
2022-05
2022-06
2024-03
2023-10
2023-05
2023-03
2022-12
2023-12
2023-11
2023-06
2023-04
2022-11
2022-09
2022-07
2022-08
2023-01
2022-02
2024-02
2024-01
2023-02
2023-07
2023-09
2023-08
2022-01
2002-10
2024-04
2009-01
2008-12
2021-12
2003-01
2002-12
2001-01
2014-11
2012-02
2001-08
1970-01
year_month


day
1e6
6
5
4
3
2
1
0
14
15
12
9
13
8
16
11
10
7
28
19
6
27
20
1
5
21
17
3
18
26
2
22
4
25
23
24
29
30
31
day


week
1e7
2.5
2.0
1.5
1.0
0.5
0.0
Thursday
Wednesday
Friday
Saturday
Tuesday
Monday
Sunday
week


hour
1e7
1.2
1.0
0.8
0.6
0.4
0.2
0.0
18
17
19
16
15
14
20
13
21
12
22
11
10
9
23
8
0
7
1
6
2
3
4
5
hour