1. Introduction

This documentation assesses the quality of the provided dataset and identifies issues affecting its reliability, accuracy, and completeness. It outlines the findings of the assessment and highlights the importance of data quality for decision-making.

2. Data Assessment Methodology

The data assessment was based on the document specification provided (gs://data_eng_test/data_dictionary_trip_records_yellow.pdf), which served as a guideline for evaluating the data quality. The document specification outlined the expected characteristics, structure, and content of the dataset. This ensured a systematic and comprehensive evaluation of the data against the specified criteria, facilitating a thorough assessment of its quality.

3. Data Quality Assessment

The assessment revealed several data quality issues:

1. tip_amount field should be of the float data type. However, a thorough examination of the source data exposed a mix of string and float values
2. passenger_count field should be of the integer data type. However, a thorough examination of the source data exposed a mix of float and integer values
3. payment_type field should be of the integer data type. However, a thorough examination of the source data exposed a mix of float and integer values
4. rate_code_id field should be of the integer data type. However, a thorough examination of the source data exposed a mix of string and integer values
5. mta_tax field should be of the float data type. However, a thorough examination of the source data exposed a mix of string and float values
6. dropoff_longitude, dropoff_latitude, pickup_longitude, pickup_latitude fields. It was observed that certain latitude values exceeded the expected range of -90 to 90 degrees, indicating potential data entry errors or inaccuracies.
7. The data quality assessment revealed a discrepancy in field delimiters across the dataset.The majority of files follow the standard practice of using tab delimiters for separating fields. However, upon closer inspection, it was identified that a subset of files within the dataset deviates from this standard and employs the pipe symbol (|) as the delimiter instead.

4. Data Issues

Several data issues were identified during the assessment:

1. Mixed Data Types: The dataset contains fields with inconsistent data types, such as numeric, string, or other unexpected formats. This inconsistency hinders accurate data analysis and processing.
2. Incorrect Latitude Values: Some latitude values within the dataset are outside the valid range of -90 to 90 degrees. This discrepancy indicates potential data entry errors or inaccuracies in recording geographic locations.
3. Different Delimiters: While most files in the dataset use tab delimiters for field separation, a subset of files deviates from this standard and adopts different delimiters, such as the pipe symbol (|). This variation poses challenges in data integration and consistent processing.