Green Taxi:
Competitor Market Analysis

Assessed By Barbara Mikulášová

Report Generated By User
Written By James Wright

Report Generated On
Tuesday 06 June 2023

Date First Published
Wednesday 07 June 2023

Contents

1 Data Collection and Quality Report

1.1 Data Collection

1.1.1 Metadata Quality Issues

Summary of problems and corrections with the metadata on import:

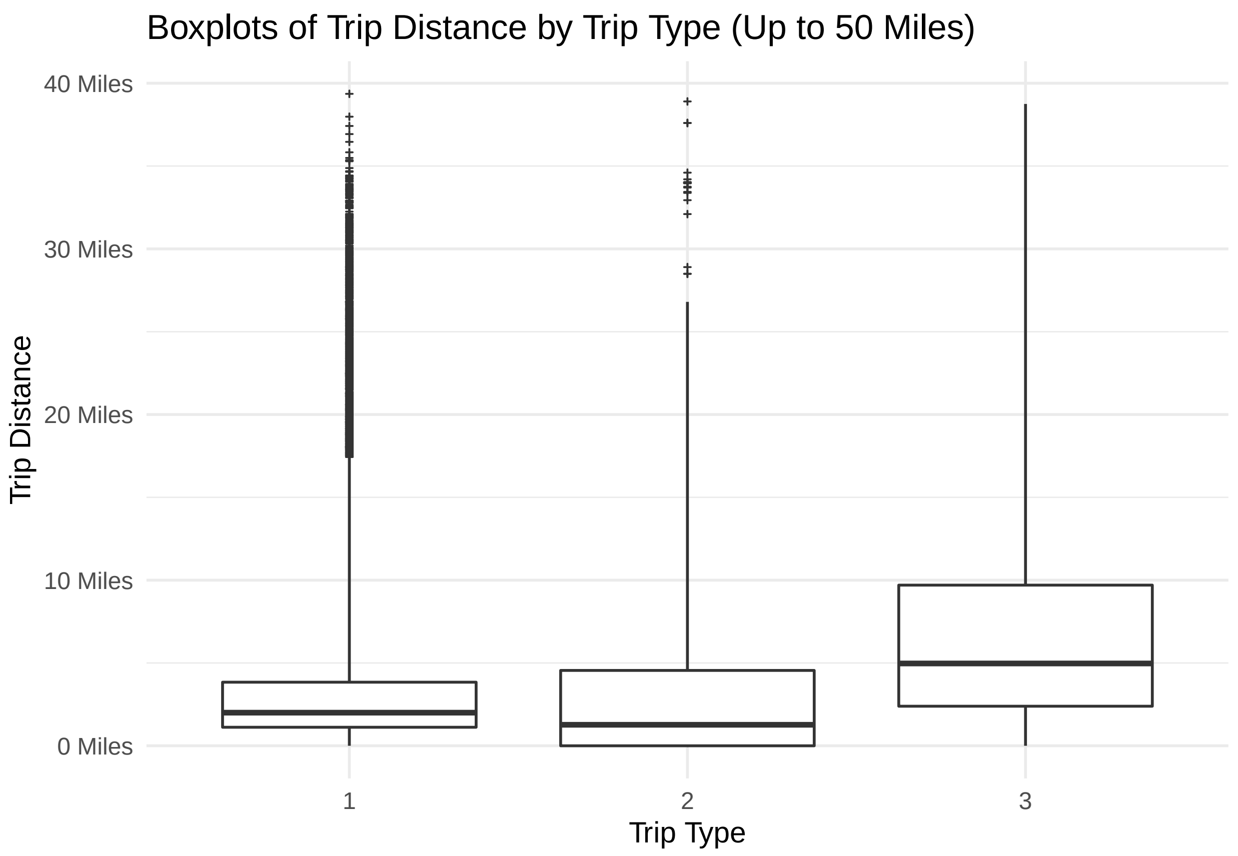
- All variable names should be upper-case by convention for formatting clarity;
- Inconsistent variable name word-separation convention across table variables - variable PAYMENTTYPE in taxi_trips renamed to PAYMENT_TYPE ;
- Inconsistent ID variable name convention and naming - variable LOCATION_CODEID in taxi_time_location renamed to LOCATIONID for a consistent ID format across tables and common name across tables;
- Missing variables - variable LOCATION_DETAILS column in taxi_time_location was missing but is in appendix-added containing NA values for completeness.
- Ill-defined field categories in appendix, 'no charge', 'dispute', 'voided' need definitions.

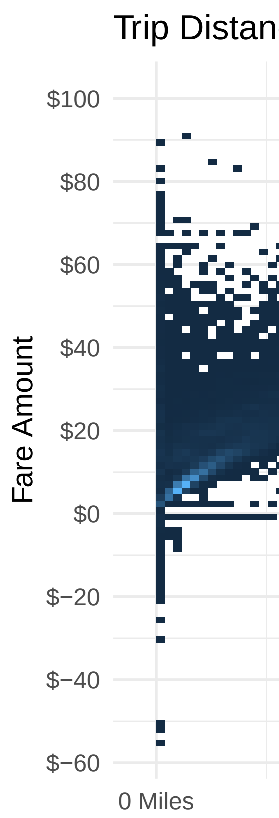
1.1.2 Duplication Issues

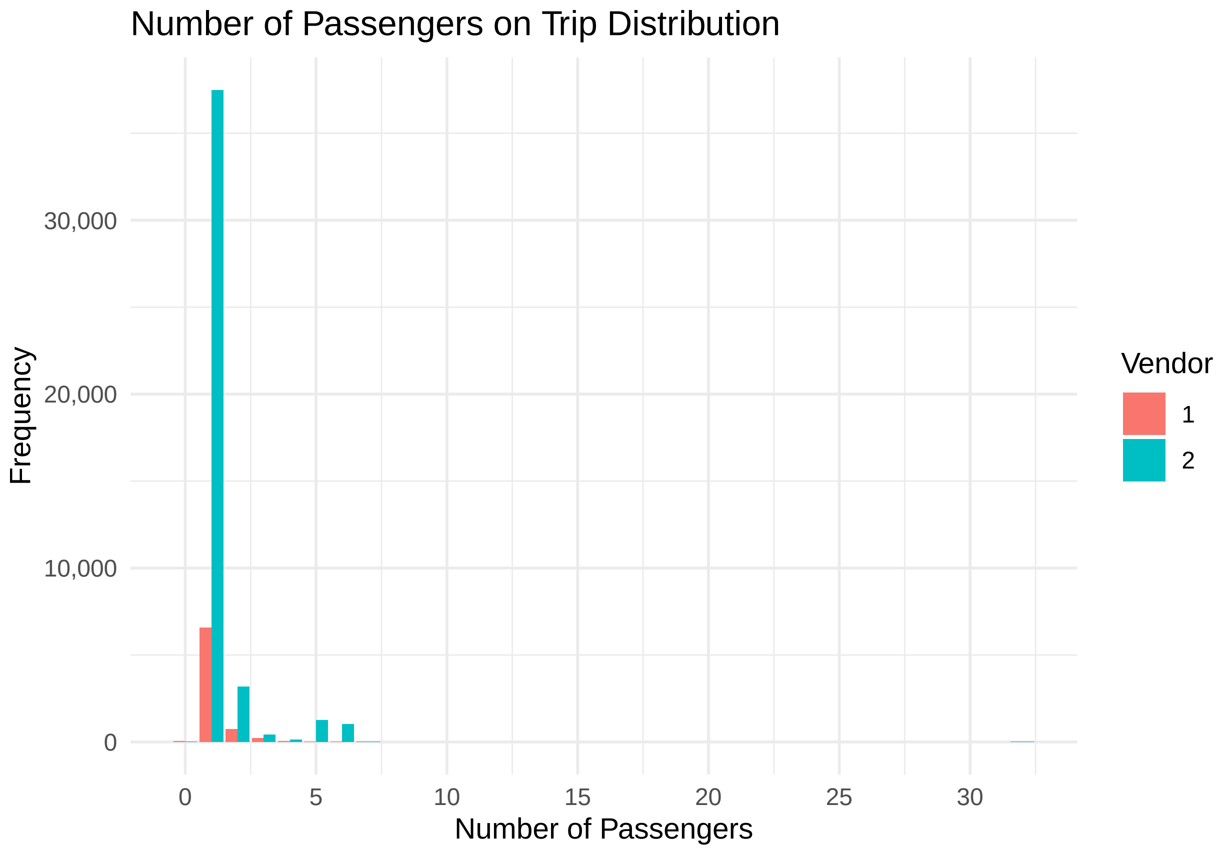
There were 67 duplicates in the taxi_trips data, 0 duplicates in the taxi_zone_lookup data, and 0 duplicates in the taxi_time_location data. The duplicates have been saved as an Excel file in the 'Reports' folder to review which data gets duplicated in the collection process.

1.1.3 Unexpected Data Value Issues

- Unexpected values;
  - There are negative and positive fare amounts, this might make sense as they could be refunds, but for clarity data should be collected or flags assigned to point out where this comes from.
  - In rows of taxi_zone_lookup, zone 'NV' inconsistent format with other zones. Treated as typo for NA.
  - Number of passengers has values up to 30. The typical capacity of a taxi is only going to be 6 people, so clearly data is being collected on the wrong kinds of vehicles.
- Unexpected categories;
  - Payment type 7 is not defined in the appendix but exists in the data.
    - For analysis purposes, these were redefined as NA as all corresponded to Vendor 3 which has mostly unknown data.
  - Rate code 8 is not defined in the appendix but exists in the data. Leaving as it is, as all correspond to vendor 3 who might have their own rate code like London cabs.
  - There is a category unknown for payment type - just leave these as NA for convention.
- Extreme values
  - There are distances travelled as large as 270,000 miles - data only really needs to be collected up to around 40 miles as this is the maximum distance journey you can expect a New York taxi to travel, and this is reflected by the scatter density of outliers for the pink taxi and uber categories once filtered to 40 miles.
    - We restrict the data to up to 40 miles for analysis.
  - Overall, there are outlier fairs of over $150 for values of around 0 Miles, while also unexpected fairs of $50 or lower for distances over 10,000 miles - which is clearly wrong intuitively. There must be mistakes in how the data for fare amounts and distances are collected, or the wrong data is being collected.
  - Most fairs in the 40 mile range are between -$100 and $100, with a few outliers of $480 and around $150.
    - We restrict the data to fares between -$100 and $100 for analysis.
- Value distribution
  - There is a disproportionate amount of Other and Uber data vs Green taxis, this may reflect the proportion of the true market share by each vendor, but this must be checked.







1.1.4 Erroneous Value Issues

- Total amount does not include the congestion charge, this needs to be corrected or highlighted in the appendix.
- For analysis purposes, we create a new variable to also include the congestion charge.

1.1.5 Missing Data Issues

- Inconsistency with naming convention for missing values - e.g. Unknown, N/A, NA, NV.
  - All not available data should follow default naming convention NA.
- The entire E-Hail fee column is missing. Data must be collected or stop storing the variable.
  - For analysis purposes, the column was dropped as it was not needed.
- Other vendors have entirely missing columns of values for STORE_AND_FWD_FLAG, RATECODEID, PASSENGER_COUNT, PAYMENT_COUNT, CONGESTION_SURCHARGE. This data needs to be collected or other data collected that enables you to infer these values.

1.1.5.1 Count of Columns with Missing Values

STORE_AND_FWD_FLAG	PASSENGER_COUNT	EHAIL_FEE	PAYMENT_TYPE	CONGESTION_SURCHARGE
32518	32518	83691	32519	32518

1.1.6 Extreme Data Issues

1.1.6.1 Unduplicated Data Summary Report

2 Data Analysis Report