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 PAYMENT-TYPE 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;
  - In rows of taxi_zone_lookup, zone 'NV' inconsistent format with other zones. Treated as typo for NA.
- 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.

1.1.4 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 Ehail column is missing. Data must be collected or stop storing the variable.
  - For analysis purposes, the column was dropped as it was not needed.
- The forward flag,

1.1.4.1 Count of Columns with Missing Values

STORE_AND_FWD_FLAG	PASSENGER_COUNT	TRAIL_FEE	PAYMENT_TYPE	CONGESTION_SURCHARGE
32518	32518	83691	32519	32518

1.1.5 Extreme Data Issues

1.1.5.1 Unduplicated Data Summary Report

2 Data Analysis Report