Cleanup_handling the outliers

the df_main dataframe in the Jupyter Notebook.

The cleaning process focused on handling missing values, removing irrelevant columns, and addressing outliers, specifically negative values.

Column Drops Initial Column Drops: Columns col1, col2, and col3 were dropped if they existed in the dataframe. This was likely done because these columns were deemed irrelevant for the analysis. If both airport_fee and Airport_fee columns existed, they were dropped after being combined into a new column called combined_airport_fee.

If only one of these columns existed, it was renamed to combined_airport_fee and the original column was dropped. Dropping Columns with Monetary Parameters: Columns mta_tax, extra, and tolls_amount were dropped.

The reasons for dropping these columns are not explicitly stated in the code but might be due to redundancy or irrelevance for the analysis. Handling Negative Values Negative Values in

Numerical Columns: The code identified numerical columns (integer and float types) with negative values and stored them in a list called negative_value_columns.

It then focused on specific columns like fare_amount, total_amount, and trip_distance, addressing negative values in these columns. For fare_amount and total_amount, negative values were converted to their absolute values using np.abs(). For trip_distance, rows with negative values were removed using filtering: df_main = df_main[df_main['trip_distance'] >= 0]. The improvement_surcharge column was handled by replacing negative values with 0.30.

Handling Outliers: Entries where trip_distance was close to 0 and fare_amount was above 300 were likely dropped, suggesting removal of data points with unrealistic values. Similarly, entries where both trip_distance and fare_amount were 0 but pickup and dropoff locations were different were also removed. Outliers with trip_distance greater than 250 miles were likely dropped, assuming these to be data entry errors. Entries with payment_type as 0 were removed as there was no payment type 0 defined in the data dictionary. Additional Cleanups The store_and_fwd_flag column was filled with 'N' for missing values. The passenger_count column was cleaned by replacing 'N' with 0. The RatecodeID column was cleaned by replacing 'N' with 1.0 and removing values equal to 99.0, likely due to data quality issues. The congestion_surcharge column had its missing values replaced with 2.5.