Assignment 5

CS215: Data Analysis and Interpretation

Shaik Awez Mehtab, 23B1080
Vaibhav Singh, 23B1068
Satyam Sinoliya, 23B0958

Solutions

Solution 1
Data Icebreaker

Categorize columns

Categorical Columns

- `pickup_community_area` and `dropoff_community_area`: These columns represent predefined geographic zones. They don't have inherent numeric meaning, so we classify them as categorical.
- `trip_start_month` and `trip_start_day`: While represented as numbers, these actually denote categories of months and days. Treating them as categorical data makes it easier to encode and interpret as discrete categories.
- `payment_type` Payment methods, such as "Credit Card" or "Cash," are categorical.
- `company`: The taxi company that provided the service is a categorical variable. It doesn't have inherent numeric meaning, so it's best treated as categorical.

Numerical Columns

- `fare`: This column represents the fare amount, a continuous numerical variable that can be analyzed directly (e.g., for averages, totals).
- `trip_start_hour`: Although it ranges from 0 to 23, this column represents the hour of the trip and is naturally numerical. Treating it as numerical allows calculations, like mean trip hours, if relevant.
- `trip_miles`: This column indicates the distance traveled in miles, a continuous numerical measurement suited for calculations such as averaging distances.
- `pickup_latitude`, `pickup_longitude`, `dropoff_latitude`, and `dropoff_longitude`: These columns represent geographic coordinates, which are inherently numerical. They're suitable for operations like distance calculations between pickup and dropoff points.
- `trip_seconds`: Trip duration is best suited as a numerical column, as it allows for calculations of averages or total trip time.
- `tips`: This represents the tip amount in currency, a continuous numerical variable, so it fits into numerical analysis.

Mixed Columns

- `trip_start_timestamp`: This column stores timestamps, which don't fit cleanly into categorical or numerical types without conversion. Converting it into datetime format allows for time-based analysis.
- `pickup_census_tract` and `dropoff_census_tract`: These columns represent census tracts, a geographic categorization. They're typically large integer codes that may include null values or non-numeric identifiers (like ZIP codes). Because these codes don't imply numeric relationships, treating them as a complex type is useful unless they're simplified