Data Intake Report

Introduction

The comprehensive datasets utilized for the cab service analysis encompass transactional data, customer demographics, city statistics, and economic indicators. Each dataset contributes valuable insights into the operational and business environment of cab services in the United States. The following is an extensive report on the intake and processing of these datasets.

Datasets Overview

1. **Cab Data (Cab_Data.csv)**:
   - **Source**: Internal company records.
   - **Volume**: 359,392 records.
   - **Variables**:
     - **Transaction ID**: Unique identifier for each transaction.
     - **Date of Travel**: Date when the service was used.
     - **Company**: Cab service provider.
     - **City**: City where the service was provided.
     - **KM Travelled**: Distance covered during the service.
     - **Price Charged**: Fare charged to the customer.
     - **Cost of Trip**: Operational cost incurred for the service.
   - **Data Types**: Mixture of integers, floating-point numbers, and strings.
   - **Purpose**: Used to analyze the profitability and operational aspects of cab services.

2. **City Data (City.csv)**:
   - **Source**: Demographic databases/public records.
   - **Volume**: 20 entries.
   - **Variables**:
     - **City**: Name of the city.
     - **Population**: Number of people residing in the city.
     - **Users**: Number of people using cab services in the city.
   - **Data Types**: Strings, later converted to numerical data for analysis.
   - **Purpose**: Provides demographic context to assess market penetration and potential.

3. **Customer ID Data (Customer_ID.csv)**:
   - **Source**: Customer database.
   - **Volume**: 49,171 records.
   - **Variables**:
     - **Customer ID**: Unique identifier for each customer.
     - **Gender**: Gender of the customer.
     - **Age**: Age of the customer.
     - **Income (USD/Month)**: Monthly income of the customer.
   - **Data Types**: Integers and strings.

- **Purpose**: To understand customer profiles and tailor services accordingly.

4. **Transaction ID Data (Transaction_ID.csv)**:
    - **Source**: Transactional logs.
    - **Volume**: 440,098 records.
    - **Variables**:
        - **Transaction ID**: Corresponding to the **Cab_Data** transactions.
        - **Customer ID**: Linking to customer profiles.
        - **Payment_Mode**: Payment method used for the transaction.
    - **Data Types**: Integers and strings.
    - **Purpose**: To correlate customer data with their transactions and payment preferences.

Data Enrichment and Augmentation

- **Geolocation Data Retrieval**:
    - Using **Nominatim** from the **geopy** library, latitude and longitude coordinates were appended to the **City Data** to enable geospatial analysis on mapping platforms like Folium.
    - The enrichment process allows for a visual representation of service usage across different geographies.

- **Economic Indicators Integration**:
    - **Unemployment Rate (UNRATE)**:
        - Sourced from the Federal Reserve Economic Data (FRED).
        - Time-series data that provide insight into economic conditions that may impact cab service demand.
    - **Gasoline Prices (GASREGCOVW)**:
        - Also from FRED, providing context on operational cost fluctuations due to fuel prices.
    - The economic data were aligned with transactional data based on date to analyze the cab service performance against economic factors.

Data Merging Strategy

- A comprehensive dataset named **combo_data** was created by merging **Cab Data**, **Customer ID Data**, **Transaction ID Data**, enriched **City Data**, and economic indicators (**UNRATE** and **GASREGCOVW**) on the **Date of Travel** attribute.
- This merge strategy facilitated a unified view, linking transactions to customer demographics, operational details, and macroeconomic conditions.
- Additional attributes were computed such as **Profit**, **ProfitPerKM**, and **Usage Percentage** to support in-depth analysis.

Data Quality Assurance

- Rigorous checks for data consistency, type integrity, and null value handling were implemented.
- Conversion of categorical to numerical data was performed where necessary to support quantitative analysis.

- Logarithmic transformation was applied to the **Population** variable to normalize the data and mitigate skewness, enabling a more equitable visual representation of city sizes when plotting on maps.

Analysis Preparedness

- The preprocessing and merging efforts culminated in a rich dataset ready for advanced analysis.
- The dataset is primed to support business intelligence tasks such as profit analysis, customer segmentation, operational optimization, and strategic planning.

Conclusion

The data intake process encompassed careful planning and execution to ensure that the datasets were primed for analysis. The efforts undertaken during this phase are critical to the success of subsequent data exploration, analysis and insights, with an emphasis on integrity and utility to drive strategic business decisions and actionable insights. The meticulous preparation of the datasets ensures that the upcoming analyses are built on a strong foundation of clean, comprehensive, and relevant data, enabling a robust understanding of the complex dynamics within the cab service industry. The integration of geolocation data and economic indicators will allow for multifaceted analyses that can reveal correlations and insights into market dynamics and operational efficiencies.

With the datasets in place, analysts and stakeholders are equipped to delve into questions such as the impact of socioeconomic factors on cab usage, the interplay between economic conditions and service demand, and the geographical distribution of market penetration. By understanding these and other patterns, cab companies can better strategize to serve current markets, expand into new ones, and optimize their operations to navigate the challenges and opportunities presented by the evolving economic landscape.