Data Quality Monitoring System for Real-World Taxi Trips Model

Member: Aashritha Aachi

netID: aa3333

Project Definition:

The reliability of data plays an important role in determining the accuracy and usefulness of data-driven systems. In modern analytics and machine learning pipelines, real-world datasets often contain inconsistencies, missing fields, outliers, or other structural issues that can significantly distort analysis if left unaddressed. This project focuses on creating a data quality monitoring system capable of profiling, cleaning, and evaluating a large real-world dataset derived from NYC Yellow Taxi aggregated trip records for 2024. The system I developed provides a fully automated workflow that ingests the dataset, assesses its quality, performs systemic cleaning operations, detects anomalies using machine learning, generates visualizations that help interpret data quality concerns and stores raw and processed results in a SQLite database. Ultimately, the goal of the project is to construct a unified pipeline that ensures transparency, and trustworthiness in data preparation procedures. By doing so, I address the common but critically important problem of real-world data reliability in analytical environments.

Introduction:

Real-world datasets rarely come in a form suitable for direct analysis, and the NYC taxi trip dataset is no exception. Publicly released datasets often suffer from incomplete records, duplicated entries, type inconsistencies, or anomalous values resulting from system glitches, human error, or irregular real-world behaviors. These imperfections introduce significant sources of noise into statistical models and machine learning systems, reducing their accuracy and dependability. The novelty of my project lies in implementing a complete, reproducible data quality pipeline within a single Jupyter Notebook environment. Rather than treating data quality issues as isolated problems, this project integrates profiling, cleaning, anomaly detection, and structured storage together into a monitoring system.

The approach is significant because data quality is an essential but frequently overlooked part of the data science lifecycle. Researchers and practitioners often focus on model development without critically examining the underlying data. By highlighting and measuring the quality of the dataset before performing deeper analytics, the system demonstrates an effective framework for ensuring trustworthy downstream insights. The results of this project show that even large, seemingly clean datasets contain issues that must be identified and corrected. Through the combination of descriptive statistics, visual analysis, algorithmic anomaly detection, the project contributes a practical template that can be extended to other large-scale datasets where reliability concerts are paramount.

Methodology:

The methodology of this project integrates three major components: a data science profiling and cleaning workflow, a database storage layer, and a machine learning-based anomaly detection model. The entire workflow is executed within the Jupyter Notebook, which allows the system to run end-to-end in a modular yet unified setting.

The data science component begins by loading the NYC aggregated taxi dataset, which includes nearly nine hundred thousand rows of hourly aggregated data. This dataset contains a mixture of numerical variables, such as trip distance and fare amounts, along with categorical attributes identifying boroughs and payment types. The first step in the pipeline is profiling the dataset to understand its basic structure and potential quality issues. I computed missing values, duplicates, and column-level statistics and visualized missingness through a heatmap. Although the datasets appear broadly complete, the heatmap confirms that missing values are virtually nonexistent, which reflects the mature data collection infrastructure used by the NYC taxi operations. Even so, duplicate rows and numerical inconsistencies required further attention.

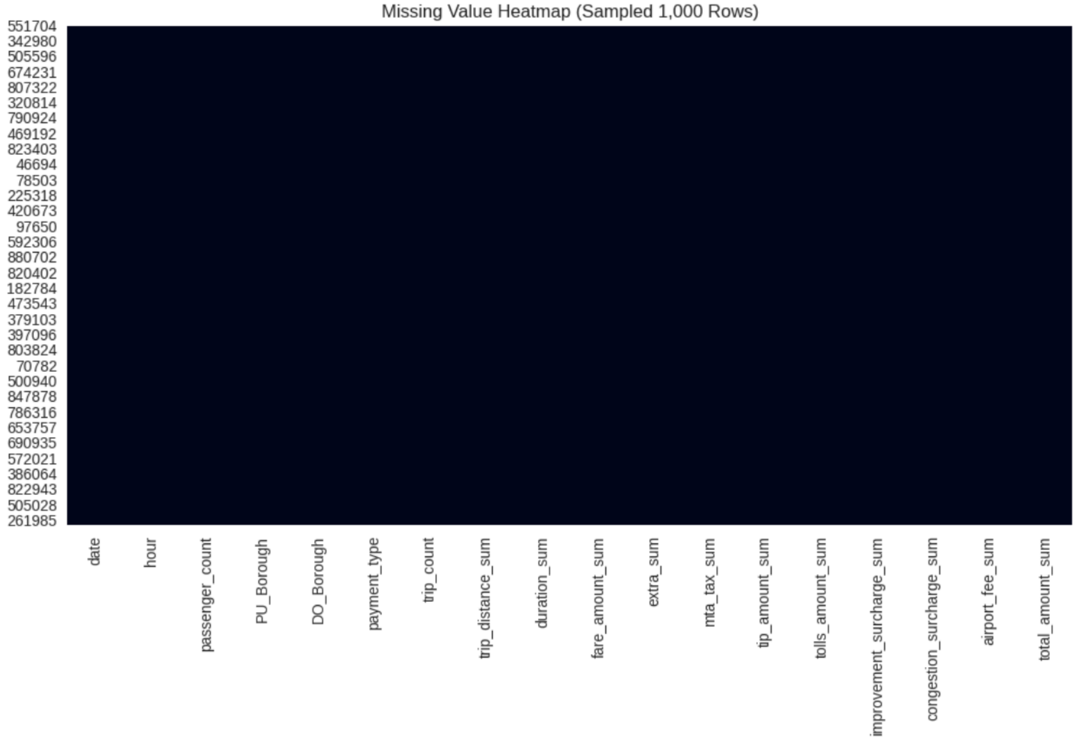
The cleaning phase followed a structured set of transformations. First, all duplicate rows were removed to prevent biased aggregations. Then, I performed type conversions to ensure that numeric fields were properly interpreted by analytical tools. For all numeric columns, missing values if present were imputed using the median of the column, which is an effective method that reduces the influence of genuine but extreme values. To reduce the impact of unusually large or small measurements that could distort later phases of analysis, I applied a three-sigma rule to remove outliers. This approach identifies observations outside three standard deviations from the mean and filters them from the dataset. The combination of these cleaning steps resulted in a dataset that was more consistent, more statistically stable, and more suitable for machine learning.

The second core component of this project is the database infrastructure, which was implemented using SQLite. I constructed a relational schema with six tables: datasets, raw_data, cleaned_data, quality_checks, anomaly_scores, and logs. The schema supports foreign key relationships and preserves a full audit trail fo the entire pipeline. Raw rows were stored as JSON strings, allowing each row to be preserved exactly as it appeared in the dataset. Cleaned data was stored similarly, enabling comparisons between raw and cleaned versions. Quality metrics such as missing value counts and duplicate counts were stored in separate table. Finally, anomaly scores and system log entries document the machine learning results and overall pipeline state. The database therefore functions not only as storage but as a record keeping system for quality assessment.

The machine learning component centers on the Isolation Forest model. This algorithm is well suited to high dimensional numerical datasets such as the NYC taxi dataset. After selecting all numeric columns, I trained the Isolation Forest model with contamination rate of 5%, meaning the model expects approximately five percent of the dataset to consist of anomalous patterns. The model computes an anomaly score for each observation, where lower values indicate unusual behavior. It also assigns a binary label indicating whether an observation is considered an anomaly. I appended these results to the cleaned dataset and stored them into the database for the future. The use of anomaly detection adds a layer to the pipeline by identifying subtle patterns in the taxi data that do not follow general trends. These may include abnormally high trip distances, inconsistent fair totals, or unusual aggregation behaviors that could suggest data entry errors or rare real-world events.

Results:

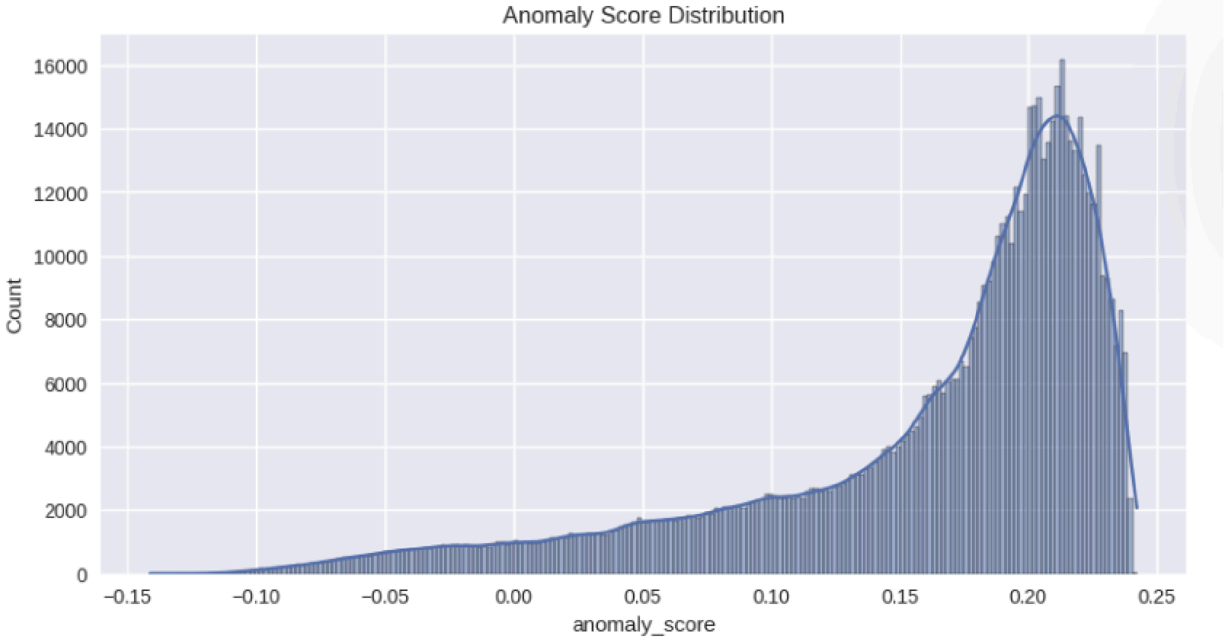
The results of the data quality monitoring system show that the NYC 2024 aggregated taxi dataset is structurally good but still shows meaningful patterns that are uncovered through profiling, cleaning and anomaly detection. Initial missing-value analysis revealed that the dataset contains no missing entries across any column. To confirm this visually, I generated a missing value heatmap using a random sample of one thousand rows. The heatmap appears as a uniformly solid block, indicating complete coverage and no observable missingness in the sampled data.



This visualization supports the numerical findings and highlights the reliability of the dataset's underlying report process.

Despite having no missing values, the dataset did contain duplicate rows, which were removed during cleaning. Numerical fields were standardized through type conversion, and outliers were further reduced using the three-sigma rule, allowing the cleaned dataset to reflect more stable central tendencies. These cleaning steps helped ensure that the subsequent anomaly detection stage was not influenced by extreme but unrepresentative values.

The Isolation Forest model revealed deeper structure in the data by assigning anomaly scores to each observation. The distribution of anomaly scores exhibits a strong right-skewed shape, with most observations clustered within a narrow high-frequency range, while a smaller number extends into the left tail representing increasingly atypical patterns.



These lower-scoring observations are flagged as anomalies and may correspond to unusual aggregation behavior, irregular trip durations or distances, or unique fare structures. The visualization of the anomaly score distribution confirms that, although the dataset is largely consistent, it contains statistically distinct edge cases that need closer inspection than this.

Overall, the results show that the data quality monitoring system successfully identified structural consistency in the dataset while also uncovering meaningful irregularities through cleaning and machine learning.

Contributions:

I completed the project by myself.

References:

https://www.kaggle.com/datasets/mohamedsalamh/nyc-yellow-taxi-trips-2024-aggregated-dataset - Dataset used from kaggle