Summary of Methodology and Results

The project was executed in three phases, successfully addressing significant data quality issues and leading to a high-performance predictive model.

Phase 1: Data Preprocessing and Feature Engineering

This phase was the most critical due to the extreme "dirtiness" of the input data, which contained missing values, inconsistent text entries ("ERROR," "UNKNOWN"), and incorrect data types.

Key Preprocessing Actions:

- **Data Cleaning:** Invalid text entries were systematically converted to standard missing values (NaN).
- **Type Conversion:** Numerical columns (Quantity, Price Per Unit, Total Spent) were converted from strings to the appropriate numeric type, and Transaction Date was converted to datetime.
- **Imputation Strategy:** Missing numerical values were imputed using the **median**, while missing categorical values (Item, Location, Payment Method) were assigned a **'Missing' placeholder category**.
- **Feature Engineering:** Time-based features (Month, Day, Weekday) were extracted from the Transaction Date to capture temporal sales patterns.

The successful cleaning resulted in a dense, structured dataset ready for modeling.

Phase 2: Exploratory Data Analysis (EDA)

EDA confirmed the distribution of Total Spent and provided initial insights into customer spending behavior:

- The Total Spent distribution was observed to be right-skewed, typical of sales data, indicating that most transactions are small but a few are significantly larger.
- Initial analysis of categorical features showed that the type of Item sold has a strong correlation with the mean Total Spent, a factor that was encoded and used effectively in the subsequent regression models.

Phase 3: Regression Modeling and Evaluation

Two regression models were trained to predict Total Spent: Linear Regression (as a baseline) and Random Forest Regressor.

Model	**Root Mean Squared Error (RMSE)**	**Mean Absolute Error (MAE)**
Linear Regression (Baseline)	2.1716	1.5175
Random Forest Regressor	1.1401	0.3042

Conclusion

The project successfully achieved its core objective of predicting cafe sales revenue from complex, uncleaned data.

The **Random Forest Regressor** significantly outperformed the Linear Regression baseline, demonstrating superior ability to capture non-linear relationships within the features. Its low **Mean Absolute Error (MAE) of 0.3042** signifies that, on average, the model's prediction for the total amount spent on a transaction is within **0.30USD** of the actual amount.

This final model is a valuable asset, proving that robust data preprocessing is the foundation for accurate machine learning predictions in real-world scenarios. The project validates a strong proficiency in the end-to-end data science pipeline, from cleaning and transformation to advanced model building and rigorous evaluation.