Sir Syed CASE Institute of Technology

Programming For Ai

Project Proposal

Submitted By:

Name: Mohib Ul haq

Roll#: 2530-4005

Submitted To:

Name: Salma Asif

Date: 12/14/2025

Data Science Project Proposal: Cafe Sales Revenue Prediction

1. Project Overview

Title: Data Cleaning and Regression Modeling for Cafe Sales Revenue Prediction

Objective:

The objective of this project is to build a regression-based machine learning model capable of predicting the total amount spent by customers in a cafe. The project strongly emphasizes real-world data preprocessing techniques, as the dataset used is intentionally uncleaned and contains multiple data quality issues.

Target Variable:

Total Spent – a continuous numerical variable, making this project suitable for a **regression** problem.

2. Dataset Description

Dataset Name:

Cafe Sales – Dirty Data for Cleaning Training (Kaggle)

Dataset Size:

- 10,000 rows
- 8 columns
- Total cells: 80,000

Columns:

- Transaction ID
- Item
- Quantity
- Price Per Unit
- Total Spent
- Payment Method
- Location
- Transaction Date

3. Data Quality Issues Identified

The dataset simulates real-world messy data and includes the following problems:

- Missing values across almost all columns
- Inconsistent entries such as “ERROR” and “UNKNOWN”
- Incorrect data types (numerical and date columns stored as strings)
- High missing rates in categorical features such as Location and Payment Method

Approximately 12–13% of the total dataset contains missing or invalid values, making it ideal for practicing data cleaning and preprocessing techniques.

4. Project Methodology

Phase 1: Data Preprocessing

This is the most critical phase of the project.

- Replace invalid entries (ERROR, UNKNOWN) with standard missing values (NaN)
- Convert data types:

- Quantity, Price Per Unit, Total Spent → Numeric
- Transaction Date → Datetime

• Handle missing values:
- Numerical columns: Median imputation
- Categorical columns: Mode or placeholder category

• Feature engineering:
- Extract day, month, and weekday from Transaction Date

• Encode categorical variables using One-Hot Encoding

Phase 2: Exploratory Data Analysis (EDA)

• Analyze distribution of Total Spent
• Study relationships between Item, Location, Payment Method, and sales
• Identify patterns and outliers affecting revenue

Phase 3: Regression Modeling

• Train regression models such as:
- Linear Regression (baseline)
- Random Forest Regressor

• Evaluate models using:
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)

5. Tools and Technologies

• Python
• Pandas – data manipulation and cleaning
• NumPy – numerical operations
• Scikit-learn – regression modeling and evaluation

6. Expected Outcome

The final outcome will be a clean, structured dataset and a trained regression model capable of predicting cafe transaction revenue with reasonable accuracy. The project demonstrates strong understanding of data preprocessing, feature engineering, and regression modeling on real-world uncleaned data.