Cafe Sales Data Analysis, Prediction, and Interactive Power BI Dashboard

---

Project Description

This project demonstrates a complete **end-to-end data analytics and machine learning workflow** using a real-world cafe sales dataset. The dataset was initially sourced from Kaggle in a raw and unstructured format (dirty_cafe_sales.csv) and underwent extensive cleaning, transformation, and analysis to produce actionable insights and a professional Power BI dashboard.

The work was completed in three main phases:

1. **Data Cleaning & Preprocessing (Python)**
2. **Exploratory Data Analysis (EDA) & Machine Learning (Python)**
3. **Interactive Dashboard Development (Power BI)**

Phase 1 – Data Cleaning & Preprocessing

- **Dataset Source:** Kaggle (dirty_cafe_sales.csv)
- **Data Size:** 10,000 rows × 8 columns
- **Issues in Raw Data:**
    - Missing values in Item, Quantity, Price Per Unit, Total Spent, Payment Method, Location, and Transaction Date.
    - Inconsistent formatting for categorical values (e.g., “UNKNOWN”, “In-store”, “In-Store”).
    - Erroneous numeric entries such as "ERROR" in Price Per Unit.
    - Duplicate transaction dates preventing proper data modeling.
- **Cleaning Steps Implemented:**

1. Removed leading/trailing spaces from text fields.
2. Standardized categorical values to a uniform format.
3. Converted numeric columns (Quantity, Price Per Unit, Total Spent) to float and handled non-numeric entries.

4. Parsed Transaction Date as datetime.
5. Handled missing values using imputation or removal based on business logic.
6. Removed duplicate records.
7. Final dataset saved as cleaned_cafe_sales.csv for further analysis.

Phase 2 – Exploratory Data Analysis (EDA) & Machine Learning

- **EDA Highlights:**
  - Distribution of sales by item category (Coffee, Cake, Cookie, etc.).
  - Payment method usage and preferences.
  - Location-based sales distribution (In-Store, Takeaway).
  - Sales trends across months and days.
- **Machine Learning Model:**
  - **Goal:** Predict Total Spent based on quantity, price per unit, item type, payment method, location, and transaction date features.
  - **Approach:**
    - Feature engineering: Extracted Year, Month, Day from Transaction Date; one-hot encoded categorical variables.
    - Train-test split: 80/20 ratio.
    - Model: Random Forest Regressor with hyperparameter tuning using GridSearchCV.
  - **Best Model Parameters:**
    - n_estimators=300, max_depth=10, min_samples_split=10, min_samples_leaf=2
  - **Performance Metrics:**
    - R^2 on test set: **0.9196**
    - MAE: **0.65**
    - RMSE: **1.66**
  - Feature importance analysis identified Quantity and Price Per Unit as the most significant predictors.

- Predictions exported as predicted_cafe_sales.csv.

Phase 3 – Power BI Dashboard Development

- **Data Source in Power BI:** cleaned_cafe_sales.csv
- **Data Model:**
  - Created a Date Table using DAX:

DAX

CopyEdit

```
DateTable =
  CALENDAR(
    MIN('Cleaned_cafe_sales'[Transaction Date]),
    MAX('Cleaned_cafe_sales'[Transaction Date])
  )
```

  - Established relationships between Date Table and Sales Table (many-to-one where possible).
- **Visualizations Included:**
  - KPI Cards: Total Sales, Average Price, Predicted Sales, Sales Gap.
  - Donut Charts: Sales by Item, Payment Method.
  - Bar/Column Charts: Sales by Location, Monthly Trends.
  - Line Chart: Sales over Time.
  - Slicers: Item, Payment Method, Location, Date.
- **Design Settings:**
  - **Canvas Size:** Custom, optimized for single-page layout.
  - **Color Theme:** Primary color **#4A90E2** with lighter shades for category differentiation.
  - **Layout:** All visuals aligned and proportioned for professional presentation.

Key Skills Demonstrated

- Data Cleaning & Wrangling (Pandas, NumPy)
- Exploratory Data Analysis & Visualization (Matplotlib, Seaborn)
- Feature Engineering & Machine Learning (scikit-learn, Random Forest)
- Business Intelligence Dashboard Design (Power BI, DAX)
- Data Storytelling & Presentation

Files in Repository

- dirty_cafe_sales.csv – Raw dataset from Kaggle.
- cleaned_cafe_sales.csv – Cleaned dataset after preprocessing.
- predicted_cafe_sales.csv – Dataset with ML predictions.
- cafe_sales_eda_ml.ipynb – Jupyter Notebook containing the complete Python workflow from cleaning to prediction.
- Cafe_Sales_Dashboard.pbix – Power BI file with the final dashboard.
- README.md – Project documentation.

How to Use

1. Clone this repository.
2. Open cafe_sales_eda_ml.ipynb to review the Python workflow.
3. Open Cafe_Sales_Dashboard.pbix in Power BI Desktop to interact with the dashboard.
4. Replace the dataset paths if necessary to match your environment.

Outcome

This project delivers a **fully cleaned dataset**, an **accurate predictive model**, and a **modern, interactive dashboard** ready for business decision-making, demonstrating the entire analytics lifecycle from raw data to insights.