**Task 1**

**Data Cleaning, Transformation, and Sales Forecasting**

**Task Overview**

The project involves:

1. **Data Cleaning & Transformation** – Standardizing, handling missing values, and merging datasets.
2. **Feature Engineering** – Creating useful features for analysis.
3. **Supervised Learning for Sales Forecasting** – Predicting sales performance using RandomForestRegressor.
4. **Supermarket Performance Classification** – Identifying high-performing supermarkets using RandomForestClassifier.

**Step 1: Data Cleaning & Transformation**

**Steps Taken:**

* **Removed duplicates** to ensure data integrity.
* **Standardized column names** for consistency across datasets.
* **Trimmed whitespace** from string values to avoid formatting issues.
* **Handled missing values** by filling meaningful defaults:
  + Description, Brand, Size → "Unknown"
  + Voucher → 0, CustomerId → "Guest"
  + feature, display → "None"
* **Converted data types** for accuracy:
  + Time of transactions, Day → datetime
  + Amount → float
  + Units → integer
* **Merged datasets** using key columns (code, Supermarket\_No, province).

**Challenges Faced:**

* Handling missing or inconsistent data formats.
* Need for additional feature engineering to capture sales trends.

**Step 2: Feature Engineering**

**New Features Created:**

* **Date-based Features:**
  + Weekday (Day of the week)
  + Month (Month of transaction)
  + IsWeekend (Binary flag: 1 if Saturday/Sunday, 0 otherwise)
* **Basket-level Analysis:**
  + Total Basket Value (Total amount per basket)
  + Average Units Per Basket (Average units sold per basket)
* **Promotion Features:**
  + Promotion Active (Binary flag if promotion is running)
  + Display Impact (Binary flag if promotional display is active)
* **Customer Segmentation:**
  + Classified customers into Low, Medium, and High spenders based on transaction history.

**Step 3: Supervised Learning - Sales Forecasting**

**Problem Definition:**

* **Predict sales amount** based on promotions, store location, and product details.

**Model Choice:**

* **RandomForestRegressor** chosen for its robustness in handling **non-linear relationships** and categorical data.

**Features Used:**

* **Product-related:** code, brand, type
* **Store-related:** Supermarket\_No
* **Time-related:** Month, Weekday
* **Promotion-related:** Promotion Active

**Model Evaluation:**

* **Root Mean Squared Error (RMSE)** used to assess model accuracy.
* Achieved a low **RMSE score**, indicating accurate sales predictions.

**Business Insights:**

* Identified **promotion periods** that positively impacted sales.
* Recognized **high-selling products** to optimize inventory planning.

**Step 4: Supervised Learning - Supermarket Performance Classification**

**Problem Definition:**

* **Classify supermarkets** as **High-Performing (1) or Low-Performing (0)** based on total sales.

**Model Choice:**

* **RandomForestClassifier** chosen for its ability to handle categorical data and rank feature importance.

**Features Used:**

* Month, Weekday, code, Promotion Active, Display Impact

**Model Evaluation:**

* **Accuracy Score** used to assess classification performance.
* Achieved **high accuracy**, confirming that key features influence supermarket performance.

**Business Insights:**

* Identified **top-performing supermarkets** to guide promotional strategies.
* Enabled **targeted marketing efforts** based on key sales drivers.

**Step 5: Business Insights & Impact**

**Key Findings:**

* **Promotion Impact:** Sales significantly increased when promotions were active.
* **Weekend Sales Surge:** More transactions occurred on weekends.
* **High-Performing Stores:** Some supermarkets consistently outperformed others due to **location and promotions**.
* **Customer Segmentation:** High spenders contributed disproportionately to sales.

**Methodologies Used:**

* **Data aggregation and visualization**.
* **Feature importance analysis**.
* **Predictive modeling** to support business decision-making.

**Business Relevance:**

* **Optimized inventory management** by predicting demand.
* **Improved marketing strategies** based on customer spending behavior.
* **Enhanced store operations** through **data-driven decisions**.

**Final Summary**

This documentation provides a structured **end-to-end data analysis pipeline** from **data cleaning and transformation** to **advanced analytics and forecasting** for supermarket sales.