Digital Egypt Pioneers Initiative (DEPI) Final Project

Data Exploration: Milestone (1)

Sales Forecasting and Optimization

Group Members:

Basmala Ehab Mohamed Yousry

Mostafa Mahmoud Mohamed Elshahat

Mohab Mohamed Ibrahim Mohamed

Ziad Ahmed Gharieb

SHR2_AIS4_S2

Submission Date:
9 May 2025

Table of Contents

Introduction

This project aims to use past sales data to predict future sales. By doing this, a store can manage stock better, plan promotions, and improve sales. In Milestone 1—*Data Exploration: Milestone (1)*—we focus on collecting data, exploring it, and cleaning it. We look at data quality, find trends and seasons, check how things relate, and fix problems like missing values and outliers. These steps will help us build a good forecasting model later.

Data Overview

- **Dataset:** Daily sales data from 2010 to 2016 for a large store, with over 50,000 records and 24 columns.
- **Important Columns:**
  - *Date* (when the sale happened)
  - *Sales*, *Quantity*, *Discount*, *Profit* (numbers)
  - *Category*, *Sub-Category*, *Region*, *Segment* (labels)
- **Size & Types:** About 51,000 rows. Mix of dates, numbers, and object labels.
- **Variety:** 15+ sub-categories, 23 regions, many products and customers.

Missing Values and Data Quality

- **Missing Data:** Overall, less than 3.35% of values are missing, except in the Postal Code column, where approximately 80% of entries are missing.

- **Fixing Missing Data**:
  - That column wasn't very important, and since most of its values were missing, we dropped it entirely.

- **Duplicates & Format**:
  - There are no duplicate transaction IDs.
  - All dates are formatted as YYYY-MM-DD.
  - Profit values range from negative to positive.

Trend Analysis

- **Overall Trend:** Sales rose from 2010 to mid-2015, then stayed flat in 2016.
- **By Category:**
  - *Technology* grew fastest (~12% per year).
  - *Office Supplies* grew ~5%–7% per year.
- **Top Products**: Phones, Copiers, Chairs, Bookcases, Storage items leading in sales.
- **Charts:**
  - *The monthly sales line plot shows steady growth.*
  - *Bar plot shows the top 10 sub-categories by total sales.*

Seasonality Analysis

- **Weekly Pattern**: Highest sales on Fridays and Sundays; lowest on Mondays.
- **Monthly/Quarterly Pattern:**
  - *Q4 (Oct–Dec) has the biggest peaks (holidays, Black Friday).*
  - *Q2 (Apr–Jun) has moderate boosts (spring sales).*
- **Holiday Effects**: Black Friday boost sales by 30%+. Other holidays vary by region.

Correlation and Influencing Factors

- **Key Correlations**:
  - *Sales* and *shipping cost*: strong positive (high sales → high profit).
  - *Profit* and *Sales*: moderate positive (more discount → more sales).
  - *Discount* and *Profit*: negative (more discount → less profit margin).
- **By Category**:
  - Technology sees a 50% sales lift during promotions.
  - Furniture profit holds up better under discounts.

Outliers and Anomalies

- **Finding Outliers**: Used box plots and IQR on sales, quantity, discount, profit, and shipping cost.
- **Results**: About 14.35% of records are extreme (very large single orders).
- **Handling Outliers**:
  - *Remove records beyond 3× IQR to avoid skewing models.*
  - *Keep mild outliers but mark them for special handling.*
  - *After removal, the dataset has ~ 20,411 fewer records*

Key Findings and Next Steps

- ***Data Quality****: Data is mostly clean and complete.*
- ***Trends & Seasons****: Clear growth trend up to 2015 and strong seasonal patterns.*
- ***Opportunities****: Technology and Office Supplies grow fast; Furniture could use more targeted promotions.*
- ***Modeling Tips****:*
  - *Include holidays\weekends and Black Friday flags in models.*
  - *Create features like day/week/month markers.*

Next, in Milestone 2 we will further clean and preprocess the data, analyze correlations and trends, and create both static charts and interactive dashboards to explore sales patterns.