Retail Store Sales Analysis using Azure Databricks

Author: Ashwin Kumar

Qualification: MBA in Data Analytics

Tools: Azure Databricks, Python, Pandas, NumPy, Matplotlib, Seaborn

This document presents a complete end-to-end Retail Store Sales Analysis project implemented using Azure Databricks. The project focuses on data creation, preprocessing, exploratory data analysis, visualization, and business insights.

1. Introduction

Retail analytics plays a crucial role in understanding customer behavior, product performance, and store efficiency. With the increasing volume of transactional data, cloud-based analytics platforms like Azure Databricks provide scalable, efficient, and collaborative environments for data analysis. This project demonstrates how Azure Databricks can be used to analyze retail sales data and derive actionable business insights.

The objective of this project is to perform descriptive and exploratory analytics on retail sales data and present insights using visualizations that support business decision-making.

2. Azure Databricks Overview

Azure Databricks is a cloud-based big data analytics platform built on Apache Spark and optimized for Microsoft Azure. It enables fast, collaborative data engineering, data science, and analytics workflows.

Key features used in this project include: - Interactive notebooks - Python-based analytics - Scalable compute clusters - Integrated visualization support

3. Dataset Description

The dataset used in this project is a synthetically generated retail transaction dataset. It contains 100 records representing individual sales transactions across multiple stores and products.

Key attributes include Transaction ID, Customer ID, Product, Quantity, Price, Store, Date, and derived fields such as Total Sales, Day of Week, and Month.

4. Data Processing and Feature Engineering

Data preprocessing was performed using Pandas within Azure Databricks. The following steps were applied: - Creation of synthetic data - Calculation of Total Sales - Extraction of Day of Week and Month from Date - Validation of data types and consistency

Feature engineering helps in uncovering hidden patterns and enables better analytical insights.

5. Exploratory Data Analysis

Exploratory Data Analysis was conducted to understand the distribution and relationships within the data. Various statistical summaries and group-based aggregations were used.

Key EDA activities include: - Product-wise sales analysis - Store-wise revenue comparison - Customer spending behavior - Time-based sales trends

6. Data Visualization

Data visualization is a critical component of analytics. In this project, multiple visualization techniques were used: - Bar charts for product and store comparison - Line charts for sales trends - Pie and donut charts for contribution analysis - Box and violin plots for distribution analysis - Bubble charts to represent multi-dimensional relationships

These visualizations help stakeholders quickly interpret complex data patterns.

7. Business Insights

The analysis produced several valuable business insights: - Identification of top-performing products - Recognition of high-revenue stores - Detection of peak sales periods - Understanding customer purchasing patterns

These insights can be used for inventory planning, pricing strategies, and targeted marketing campaigns.

8. Conclusion and Future Scope

This project successfully demonstrates how Azure Databricks can be used for end-to-end retail data analysis. The combination of cloud scalability, Python analytics, and visual storytelling makes it a powerful solution.

Future enhancements may include: - Integration with real-time data sources - Predictive analytics and forecasting - Machine learning-based customer segmentation - Dashboard integration using Power BI