**Introduction**

We are students in the Data Science and AI track at DEPI (Digital Egypt Pioneers Initiative), which is a national program in Egypt that provides advanced training in technology and data-related fields.

As part of our graduation project, we wanted to work on real-world data to gain practical experience. That’s why we chose the Rossmann Store Sales dataset, since it provides a rich and realistic business case for applying data science and AI techniques.

**About us**

We are students in the IBM Data Science track at DEPI (Digital Egypt Pioneers Initiative).

**Team Members**

* **Yousif Salah Muhammad**
* **Asmaa Ali Abdelaaty**
* **Abdelrahman Osama**
* **Esraa Taha Abdelhasseb**
* **Aml Abdullah shehata**
* **Hanan Hamdy Mostafa**

**Project goal**

To predict daily sales for Rossmann stores and produce a reproducible pipeline demonstrating data cleaning, feature engineering, model selection, evaluation, and a deployment-ready forecast solution. The project focuses on both predictive performance and actionable business insights (promotions, seasonality, competition).

**About Rossmann Store Sales**

Rossmann is one of the largest drugstore chains in Europe, headquartered in Germany. It operates thousands of stores across several countries, selling products such as health, beauty, household, and personal care items.

The Rossmann Store Sales dataset was originally released as part of a Kaggle competition. It contains real-world historical sales data from Rossmann stores. The dataset includes information such as:

* Sales figures (daily sales for each store).
* Store details (store type, assortment type, whether it runs promotions, etc.).
* Promotions and marketing campaigns (Promo, Promo2, and their intervals).
* Competition data (distance to the nearest competitor, when competitors opened nearby).
* Dates and holidays (to capture seasonality, weekends, and special events).

This dataset is widely used because it provides a realistic business case: predicting future sales for retail stores. It’s useful for learning and applying:

* Time series forecasting
* Regression models
* Feature engineering
* Optimization for inventory and business decisions

**Quick dataset summary (concise)**

* **train.csv**: daily historical records including Sales, Customers, Open, Promo, StateHoliday, SchoolHoliday, Date, Store.
* **test.csv**: same but without Sales (what we predict).
* **store.csv**: static store info: StoreType, Assortment, CompetitionDistance, CompetitionOpenSinceMonth/Year, Promo2 fields.

Key modelling target: Sales (predict daily sales for each store-day). On closed days Open=0 sales should be 0.

**Kaggle**

[Rossmann Store Sales | Kaggle](https://www.kaggle.com/c/rossmann-store-sales/data)

**Dive into Data**

It’s historical sales data for 1,115 Rossmann stores. The task is to forecast the "Sales" column for the test set.

**Files**

* train.csv - historical data including Sales
* test.csv - historical data excluding Sales

**Data fields**

**Train.csv**

historical data including Sales

**Store: a unique Id for each store**

**DayOfWeek:** 1 → Monday … 7 → Sunday

**Date:** YYYY-MM-DD

**Sales:** the turnover for any given day (this is what you are predicting)

**Customers:** the number of customers on a given day

**Open:**

* **0 = closed → sales = Zero**
* **1 = open**

**Promo**

* **0 → no**
* **1→Yes**
* **indicates whether a store is running a promo on that day**

**StateHoliday:**

**All schools are closed (public holidays , weekends)**

* **a = public holiday**
* **b = Easter holiday,**
* **c = Christmas**
* **0 = None**

**SchoolHoliday: indicates if the (Store, Date) was affected by the closure of public schools**

* **0→No**
* **1→Yes**

**store.csv**

supplemental information about the stores

**Store:** a unique Id for each store

**StoreType**: differentiates between 4 different store models: a, b, c, d

**Assortment:**

* a = basic
* b = extra
* c = extended

**CompetitionDistance:** distance in meters to the nearest competitor store

**CompetitionOpenSinceMonth:** approximate month of the time the nearest competitor was opened

**CompetitionOpenSinceYear:** approximate Year of the time the nearest competitor was opened

**Promo2:** Promo2 is a continuing and consecutive promotion for some stores:

* 0 = store is not participating,
* 1 = store is participating

**Promo2SinceWeek:** describes the week when the store started participating in Promo2

**Promo2SinceYear:** describes the year and calendar week when the store started participating in Promo2 • DurationPromo2 = CurrentYear - Promo2SinceYear

**PromoInterval:** describes the consecutive intervals Promo2 is started, naming the months the promotion is started anew. E.g. "Feb,May,Aug,Nov" means each round starts in February, May, August, November of any given year for that store

The first step in our project was to merge the different datasets into a single dataframe to facilitate analysis and modeling.

* We merged **train.csv** and **store.csv** on the Store column.
* We applied the same merge to **test.csv** and **store.csv**.  
  This allowed us to enrich each record with additional store information (e.g., store type, assortment, competition, and promotion details).

A screen shot of a computer code

AI-generated content may be incorrect.A screen shot of a black background

AI-generated content may be incorrect.



**Exploratory Data Analysis (EDA)**

is a important step in data analysis which focuses on understanding patterns, trends and relationships

**Step 1: Understand the Problem and Data**