# Rossmann Drug Store Sales Forecasting

## Project Report

Time Series Analysis & Forecasting using
**VAR, VECM, ARIMA & SARIMAX Models**

**Prepared by: Ankur Parashar**

# Table of Contents

# 1. Problem Statement

Rossmann is a European drug distributor which operates over 3,000 drug stores across seven European countries. Since a lot of drugs come with a short shelf life, that is, they do not have a long expiry date, it becomes imperative for Rossmann to accurately forecast sales at their individual stores. Currently, the forecasting is taken care of by the store managers who are tasked with forecasting daily sales for the next six weeks. With thousands of individual managers predicting sales based on their unique circumstances and intuitions, the accuracy of the forecasts is quite varied. To overcome this problem, the company has requested to forecast Sales data.

## Key Business Challenges

- Short shelf life of pharmaceutical products requiring accurate demand prediction
- Varied forecast accuracy across 3,000+ stores due to subjective predictions
- Multiple influencing factors: promotions, holidays, competition, seasonality
- Need for standardized, data-driven forecasting approach

# 2. Project Goals

## Primary Objectives

1. Build a forecasting model to forecast the daily sales for the next six weeks (42 days)
2. Prepare models specifically for 9 key stores: 1, 3, 8, 9, 13, 25, 29, 31, and 46
3. Determine if sales data is stationary or non-stationary
4. Test for cointegration between Sales and Customers using Johansen test
5. Analyze impact of promotional variables (Promo, Promo2) on sales
6. Report model accuracy using MAPE (Mean Absolute Percentage Error)

# 3. Data Description

The data is provided in two tables: **store** and **train**.

## 3.1 Store Table (Metadata)

*Shape: 1,115 rows × 10 columns*

| Field | Description |
|---|---|
| **Store** | Unique ID for each store |
| **StoreType** | Store model type: a, b, c, d |
| **Assortment** | Assortment level: a=basic, b=extra, c=extended |
| **CompetitionDistance** | Distance to nearest competitor (meters) |
| **Promo2** | Continuous promotion: 0=not participating, 1=participating |
| **PromoInterval** | Months when Promo2 starts (e.g., 'Feb,May,Aug,Nov') |

## 3.2 Train Table (Sales Data)

*Shape: 1,017,209 rows × 9 columns (Full dataset)*

*Filtered for 9 key stores: 8,110 rows*

| Field | Description |
|---|---|
| **Date** | Date of the sales record (2013-01-01 to 2015-07-31) |
| **Sales** | Turnover for the day (TARGET VARIABLE) |
| **Customers** | Number of customers on that day |
| **Promo** | Daily promotion indicator: 0=no promo, 1=promo active |
| **StateHoliday** | Holiday type: 0=none, a=public, b=Easter, c=Christmas |
| **Open** | Store open status: 0=closed, 1=open |

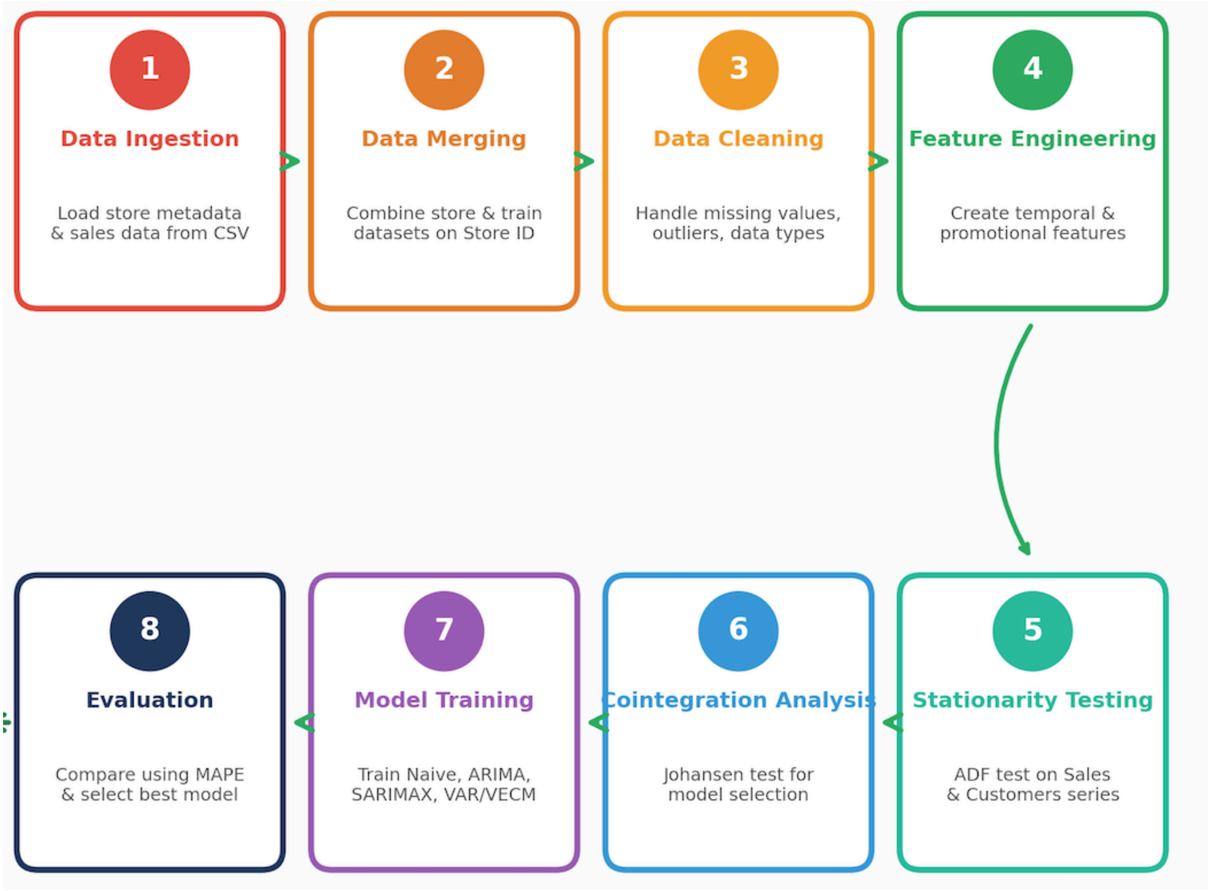
# 4. System Architecture

## 4.1 Technology Stack

| Component | Libraries/Tools |
|---|---|
| **Programming Language** | Python 3.x |
| **Data Processing** | NumPy, Pandas |
| **Visualization** | Matplotlib, Seaborn |
| **Statistical Modeling** | Statsmodels (ARIMA, SARIMAX, VAR, VECM) |
| **Machine Learning** | Scikit-learn (StandardScaler, MinMaxScaler) |
| **Environment** | Google Colab (High Memory Runtime) |

## 4.2 Pipeline Overview

1. **Data Ingestion:** Load store metadata and sales data from CSV files
2. **Data Merging:** Combine store and train datasets on Store ID
3. **Data Cleaning:** Handle missing values, outliers, and data type conversions
4. **Feature Engineering:** Create temporal and promotional features
5. **Stationarity Testing:** ADF test on Sales and Customers series
6. **Cointegration Analysis:** Johansen test to determine model selection
7. **Model Training:** Train multiple models (Naive, ARIMA, SARIMA, SARIMAX, VAR/VECM)
8. **Evaluation:** Compare models using MAPE and select best performer


1
Data Ingestion
Load store metadata & sales data from CSV
2
Data Merging
Combine store & train datasets on Store ID
3
Data Cleaning
Handle missing values, outliers, data types
4
Feature Engineering
Create temporal & promotional features
5
Stationarity Testing
ADF test on Sales & Customers series
6
Cointegration Analysis
Johansen test for model selection
7
Model Training
Train Naive, ARIMA, SARIMAX, VAR/VECM
8
Evaluation
Compare using MAPE & select best model

# 5. Data Processing

## 5.1 Data Filtering

The analysis focuses on 9 key stores selected for their revenue significance and historical importance: **Stores 1, 3, 8, 9, 13, 25, 29, 31, and 46**.

After filtering and removing closed store days (Open=0), the final dataset contains **6,611 records** spanning from January 1, 2013 to July 31, 2015.

## 5.2 Missing Value Treatment

| Column | Missing % | Treatment |
|---|---|---|
| CompetitionDistance | ~0.3% | Filled with median value |
| CompetitionOpenSince* | ~20.95% | Filled with 0 (no competition) |
| Promo2Since* | ~69.74% | Filled with 0 (not participating) |
| PromoInterval | ~69.74% | Filled with 'None' |

## 5.3 Outlier Removal

Outliers were identified at the 99th percentile for Sales and Customers variables and removed to prevent extreme values from skewing the model predictions.

## 5.4 Standardization

Sales and Customers variables were standardized using StandardScaler before modeling. Per-store scalers were maintained to enable inverse transformation of predictions back to the original scale.

# 6. Exploratory Data Analysis (EDA)

## 6.1 Sales Distribution Analysis

After outlier removal and filtering for open stores only:

- **Mean Sales:** 6,481.96
- **Median Sales:** 5,893.00
- **Std Deviation:** 2,537.48
- **Range:** 0 to 19,080

## 6.2 Customer Distribution Analysis

- **Mean Customers:** 712.16
- **Median Customers:** 615.00
- **Std Deviation:** 349.01
- **Range:** 0 to 2,376

## 6.3 Sales by Store Type

| StoreType | Count | Mean Sales | Std Dev |
|---|---|---|---|
| a | 4,272 | 6,408.47 | 2,653.82 |
| c | 2,155 | 7,156.63 | 3,536.62 |
| d | 2,184 | 6,162.95 | 2,005.84 |

## 6.4 Key EDA Insights

- **Strong positive correlation:** Sales and Customers are highly correlated
- **Promotional impact:** Days with Promo=1 show significantly higher sales
- **Day-of-week patterns:** Clear weekly seasonality with weekends showing distinct behavior
- **Store type variance:** Type 'c' stores show highest mean sales but also highest variability

# 7. Feature Engineering

## 7.1 Temporal Features

The following time-based features were extracted from the Date column:

- **Year:** Year component (2013, 2014, 2015)
- **Month:** Month component (1-12)
- **Day:** Day of month (1-31)
- **WeekOfYear:** ISO week number (1-52)
- **DayOfYear:** Sequential day count (1-365)
- **Weekend:** Binary flag for Saturday/Sunday

## 7.2 Promotional Features

- **Promo2Active:** Binary flag indicating if Promo2 was active on the specific date based on PromoInterval

## 7.3 Standardized Features

- **Sales_Std:** Standardized sales using per-store StandardScaler
- **Customers_Std:** Standardized customer count using per-store StandardScaler

# 8. Model Selection

## 8.1 Stationarity Testing (ADF Test)

The Augmented Dickey-Fuller (ADF) test was performed on both Sales_Std and Customers_Std series for all 9 stores.

**Results:** All series across all stores rejected the null hypothesis at 5% significance level, indicating the standardized series are **stationary**.

## 8.2 Johansen Cointegration Test

The test was performed to determine cointegration between Sales and Customers:

| Store | Optimal Lag | Cointegration Rank |
|---|---|---|
| 1 | 10 | 2 |
| 3 | 9 | 2 |
| 8, 9, 13, 25, 29, 31, 46 | 9-10 | 2 |

**Interpretation:** Rank = 2 indicates full rank, meaning both Sales and Customers are already stationary (I(0)). This supports using VECM with cointegration constraints.

## 8.3 Models Evaluated

1. **Last Value Forecast:** Baseline using last observed value
2. **ARIMA:** Autoregressive Integrated Moving Average with grid search
3. **SARIMA:** Seasonal ARIMA without exogenous variables
4. **SARIMAX:** SARIMA with exogenous variables (Customers, Promo, Promo2Active, Weekend)
5. **VAR/VECM:** Vector models based on cointegration rank

# 9. Model Training

## 9.1 Train-Test Split

- **Training Period:** All data except last 42 days
- **Test Period:** Last 42 days (6 weeks)
- **Forecast Horizon:** 42 days ahead

## 9.2 SARIMAX Configuration

Grid search was performed to find optimal (p, d, q) parameters:

- **p range:** 0 to 2
- **d range:** 0 to 1
- **q range:** 0 to 1
- **Selection Criterion:** AIC (Akaike Information Criterion)
- **Exogenous Variables:** Customers, Promo, Promo2Active, Weekend

## 9.3 VECM Configuration

- **Lag Selection:** Based on AIC with maxlags=10
- **Cointegration Rank:** Determined by Johansen trace test
- **Endogenous Variables:** Sales_Std, Customers_Std

# 10. Model Evaluation

## 10.1 Evaluation Metric: MAPE

Mean Absolute Percentage Error (MAPE) was used as the primary evaluation metric:

$$MAPE = (1/n) \times \Sigma |Actual - Predicted| / |Actual| \times 100\%$$

MAPE was chosen because it provides an intuitive percentage-based error measure that is easy to interpret across different store scales.

## 10.2 Model Comparison Results

| Store | Naive | ARIMA | SARIMA | SARIMAX | VECM | Best Model |
|---|---|---|---|---|---|---|
| 1 | 16.30% | 15.86% | 15.86% | **3.96%** | 13.29% | **SARIMAX** |
| 3 | 23.07% | 23.97% | 23.95% | **9.10%** | 20.10% | **SARIMAX** |
| 8 | 26.60% | 28.01% | 28.01% | **5.66%** | 20.59% | **SARIMAX** |
| 9 | 15.93% | 17.35% | 17.26% | **4.53%** | 13.98% | **SARIMAX** |
| 13 | 28.46% | 29.83% | 29.83% | **5.89%** | 26.60% | **SARIMAX** |
| 25 | 16.74% | 18.36% | 18.36% | **5.90%** | 16.37% | **SARIMAX** |
| 29 | 18.94% | 20.01% | 20.01% | **4.88%** | 14.15% | **SARIMAX** |
| 31 | 15.24% | 15.62% | 15.63% | **4.30%** | 13.58% | **SARIMAX** |
| 46 | 19.64% | 20.25% | 20.25% | **6.12%** | 18.37% | **SARIMAX** |

# 11. Results & Interpretation

## 11.1 Best Model Summary

| Store | Best Model | MAPE |
|---|---|---|
| 1 | SARIMAX | **3.96%** |
| 3 | SARIMAX | **9.10%** |
| 8 | SARIMAX | **5.66%** |
| 9 | SARIMAX | **4.53%** |
| 13 | SARIMAX | **5.89%** |
| 25 | SARIMAX | **5.90%** |
| 29 | SARIMAX | **4.88%** |
| 31 | SARIMAX | **4.30%** |
| 46 | SARIMAX | **6.12%** |

## 11.2 Key Findings

1. **SARIMAX consistently outperforms all other models** across all 9 stores, with MAPE ranging from 3.96% to 9.10%.
2. **Exogenous variables significantly improve forecast accuracy.** The inclusion of Customers, Promo, Promo2Active, and Weekend reduces MAPE by 10-20 percentage points compared to univariate models.
3. **Sales and Customers series are stationary** (after standardization) based on ADF tests, with all p-values < 0.05.
4. **Johansen test indicates full rank (r=2),** suggesting both variables are I(0) stationary, supporting the use of VECM with cointegration constraints.
5. **Customer count is the strongest predictor** of sales, demonstrating the critical importance of foot traffic in retail sales forecasting.

## 11.3 Business Recommendations

1. **Deploy SARIMAX models** for all key stores to replace manager-based forecasting.
2. **Integrate customer traffic forecasting** to enable more accurate sales predictions.
3. **Use promotional calendar data** as a key input for improved forecast accuracy.
4. **Retrain models quarterly** to capture evolving customer behavior and market conditions.
5. **Expand to all 3,000+ stores** using the validated SARIMAX framework.

# 12. Appendix: Notebook Extracts

## 12.1 Key Libraries Used

```
numpy, pandas, matplotlib, seaborn, statsmodels, sklearn

ARIMA, SARIMAX, VAR, VECM, coint_johansen, adfuller
```

## 12.2 ADF Test Function

```
def adf_test(series, name):

    result = adfuller(series.dropna(), autolag='AIC')
    return {'test_stat': result[0], 'p-value': result[1]}
```

## 12.3 MAPE Calculation

```
def mape(y_true, y_pred):

    mask = y_true != 0
    return np.mean(np.abs((y_true[mask] - y_pred[mask])
                          / y_true[mask])) * 100
```

## 12.4 Model Selection Logic

```
# Best model selection based on minimum MAPE
best_model_info = min(models_info, key=lambda x: x['mape'])
```

*— End of Report —*