## **1. 🗂️ Introduction & Project Overview**

* **Business Context** Analysis of multi-store, multi-department grocery retail sales from 2010 to 2012.
* **Objectives**
  1. Detect sales anomalies
  2. Segment stores and departments
  3. Discover cross-selling patterns
  4. Build demand forecasting models
  5. Validate marketing campaign effectiveness

## **2. Data Preprocessing & Feature Engineering**

* **Raw Data Description** Features: Store, Dept, Date, Weekly\_Sales, IsHoliday, CPI, Temperature, Fuel\_Price, MarkDown1–5, Unemployment, Size.
* **Type Conversions** Dates parsed; categorical conversion for Store and Dept.
* **Missing Data Handling** Median imputation for Markdown fields; filled lag features and percent changes.
* **New Feature Creation**
  + Sales transformations: normalized, Square root transformation, rolling, EMA
  + Lagged sales and percent change for CPI, Temperature, Fuel Price
  + Aggregation by Store and Dept

## **3. Sales Anomaly Detection**

* **Methodology**
  + Rolling average + threshold
  + Exponential Weighted Moving Average (EWMA)
* **Key Finding** Over 90% of anomalies occurred on non-holidays, suggesting unexpected spikes, potential data errors, or local events.

## **4. Segmentation Analysis**

### **4.1 Store Segmentation**

* **PCA** dropped features to 3 components (capturing ~X% variance)
* **K-Means clustering** evaluated via elbow & silhouette → optimal K=4
* **Cluster Profiles**:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cluster** | **Avg. Sales** | **CPI** | **Store Size** | **Interpretation** |
| 0 | 10,334 | 182 | 137,619 sq ft | Large-format premium |
| 1 | 10,750 | 164 | 126,416 sq ft | Mid-sized, high-volume |
| 2 | 10,275 | 172 | 132,208 sq ft | Budget retailers |
| 3 | 10,572 | 160 | 121,491 sq ft | Compact value stores |

### **4.2 Department Segmentation**

* Similar PCA + K-Means methodology on dept-level aggregated features
* Clusters ranged from high-sales to low-sales departments
* Visualization and summary tables help identify performance tiers

## **5. Market Basket Analysis**

* **Transaction format**: Dept lists per Date
* **Apriori rules** (support ≥1.1%, confidence ≥80%, lift ≥3):  
  + Several large-item-sets → Dept 29 & 30 show strong co-occurrence (lift 8–9)
  + Dept 37 → group of 20+ other depts, lift ≈77
  + Dept 35 → same group as above, lift ≈9
* **Business Implications**:  
  + Layout related departments adjacently
  + Offer bundle promotions
  + Enhance cross-store loyalty & pricing strategies

## **6. Demand Forecasting**

### **6.1 Time-Series Preparation**

* Aggregated weekly time series (2010–2012)
* Seasonal decomposition: strong seasonal + trend component
* Correlation: CPI ↔ Fuel Price (+), Unemployment ↔ both (-); no direct link to sales found
* Holiday weeks inconsistent in driving sales—higher variances but not always higher sales

### **6.2 Short-Term (SARIMA) Forecasting**

* SARIMA(1,1,1)(1,1,1)[52] selected via auto\_arima
* Evaluation (no exogenous variables):  
  + MAE ≈ 10,768; RMSE ≈ 12,259
* With external factors (CPI%, Temp%, Fuel%):  
  + Error distributions compared (plots included); slight improvement observed

### **6.3 Long-Term (Random Forest) Forecasting**

* RF trained on 20+ features (lags, percent changes, external factors)
* Baseline MAE ≈ 1,693, RMSE ≈ 2,935 (potential overfitting)
* Optimized tuning → MAE ≈ 8,943; RMSE ≈ 9,232
* Feature importance highlighted CPI%, Temp%, and Fuel changes

### **6.4 Holt-Winters Method**

* Seasonal + trend + damped + Box–Cox adjustments applied
* Forecast 34 weeks: high errors (MAE ≈ 3.2M, RMSE ≈ 4.25M)
* Tuned model forecasted 2013 with future 56 weeks—plotted actual vs. predicted

## **7. Causal Inference: Marketing Campaign**

* Focus: Dept 4 during Q3 2012
* Synthetic increase (+45%) in sales during campaign period
* CausalImpact report:  
  + +14,573 average weekly uplift (~30%)
  + Cumulative effect ≈ +247,735
  + 95% CI = [27%, 33%], p ≈ 0 → highly significant

## **8. Strategic Recommendations**

### **Inventory & Pricing**

* Use CPI & Weather to optimize stock levels and promotions
* Dynamic pricing in high CPI/season demand periods

### **Marketing & Layout**

* Leverage cross-selling: bundle co-occurring departments
* Campaign proven effective → scale to similar stores

### **Operations**

* Supply chain timing based on fuel price volatility
* Adjust store size and assortment by segment profile

## **9. Challenges & Limitations**

* Missing values and frequency inconsistencies
* Model instability (e.g., Holt-Winters convergence)
* Overfitting risk in RF; need cross-validation
* CausalImpact requires careful date alignment

## **10. Conclusion**

A data-driven pipeline uncovered:

* Hidden anomalies
* Meaningful segmentation
* Strong cross-department associations
* Robust forecasting boosted by external features
* Valid marketing campaign impact  
   These insights can power targeted strategies in inventory, marketing, and operations.