# 

Data science & IA master

# **TFM Academic Memory Summary**

Role-play DS market

Authors:

José Luengo

Nicolás Gaveglio

Marina Sallent

[**TFM Academic Memory Summary 0**](#_4k3mtznuxjth)

[Introduction 3](#_i5xhbrbb3v9n)

[Internal Team Structure and Methodology 4](#_f3bklyjcxvva)

[1. Exploratory Data Analysis (EDA) 5](#_ffczcpnspf6j)

[1.1. Objective 5](#_8pdpozs2dvd)

[1.2. Methodology and initial considerations 5](#_gewohoxrcg0)

[1.3. Actions and steps: 6](#_8uvmb1lz6s4)

[2. Dashboarding & KPI Visualization 8](#_5un0ajt9wuea)

[2.1. Objective 8](#_xh0ja9b6l16o)

[2.2. Methodology and initial considerations 8](#_cu2z4k8bvls3)

[2.3. Actions and steps 10](#_pez2ki7k8j9r)

[3. Product Clustering Analysis 12](#_kcberll4mito)

[3.1. Objective 12](#_dm7kqpf4wre3)

[3.1.1. Clustering for stores 12](#_2wjrt7z0xot1)

[3.2. Methodology and initial considerations 12](#_vcnbyf2sg8mc)

[3.2.1. Product - Seasonality 12](#_6y8h2z3u0k9k)

[3.2.2. Product - Pricing 12](#_ox5nah4hinqp)

[3.2.3. Product - Events 13](#_a4d6pftbqpu4)

[3.2.4. Product - Revenue 13](#_t3cum92vzx1v)

[3.3. Actions and steps 13](#_a6tz3su4s76n)

[3.3.1. Environment Setup & Configuration 13](#_s67dvpc40jax)

[3.3.2. Data Preprocessing Pipeline 14](#_tnf36tm0vask)

[3.3.3. Dimensionality Reduction 14](#_foroil3i0ugw)

[3.3.4.Clustering Model & Optimization 14](#_msqx5s9zs9l8)

[3.3.5. Cluster Analysis and Interpretation 15](#_rr13bj4uip9t)

[3.3.6. Final Assignment 16](#_j9urwqvbn68p)

[4. Sales Forecasting 17](#_1ko4o3b2mgp6)

[4.1. Objective 17](#_pbsv2zxnp8oy)

[4.2. Methodology and initial considerations 17](#_sp52cjib2inp)

[4.2.1. Data Enhancement and Feature Engineering 17](#_j2tlyaebxz9k)

[4.1.2. Core Methodology and Performance Framework 17](#_8qvzn9o2n8gu)

[4.2.2. Algorithm Selection and Modeling Approach 17](#_vzycdmhai4xt)

[4.2.3. Performance Evaluation Framework 18](#_wk4ndd59g8jp)

[4.3. Actions and steps 18](#_osv82zvlsf6b)

[4.3.1. Data Preparation 18](#_fosc0nad6yuu)

[4.3.2. Correlation analysis 18](#_jgrx8m15fkm0)

[4.3.3. Feature creation 19](#_5gqskayblr4f)

[4.3.4. Model evaluation and training 20](#_9fepqg152n3u)

[5. Replenishment & MLOps Proposal applied to pilot case study 22](#_v6sejxdwal0e)

[5.1. Objectives 22](#_mwezi4ko8nki)

[5.2. Methodology and initial considerations 22](#_8y7zdgplpcai)

[5.3. Actions and steps 23](#_kdqxqcub60e4)

[5.3.1. Actual stock flow 23](#_ob3n2my2ckwm)

[5.3.2. Forecasted stock flow 23](#_spg8974snwck)

[5.3.3. Comparison 23](#_b1vcdwhzctzk)

[5.4. MLOPS workflow proposed 24](#_4zsn0en0u3c)

[Conclusions 25](#_b7gbypddkv4v)

# 

## **Introduction**

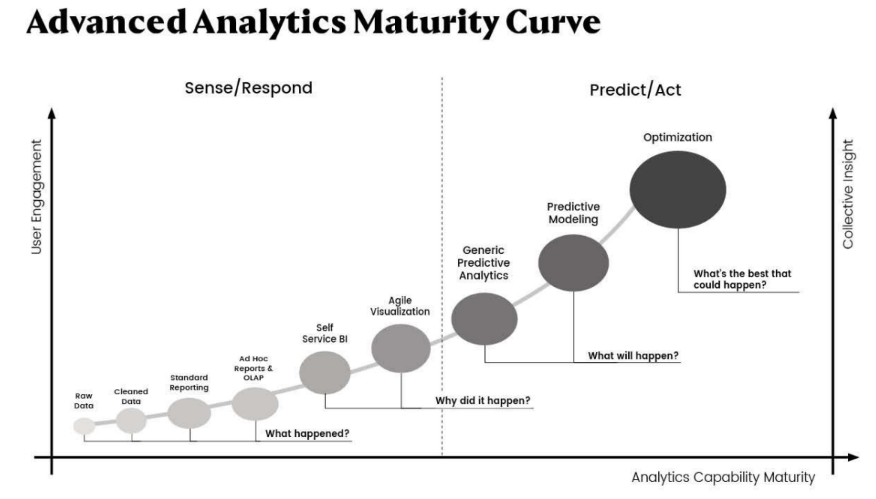
The Final Master Project (TFM) focuses on the end-to-end development of a data-driven strategy for sales analysis and forecasting in a retail context - **DS Market**.

The project aims to address multiple facets of sales data management, including exploratory data analysis (EDA), dashboard visualization, clustering analysis, and sales forecasting with the aim of **transitioning from analog operations into full digital transformation.**

The project was approached as a role-play scenario, with team members assuming the role of *Nicole*, a senior data scientist hired to work under the Finance Director, Paul Rogers, and in collaboration with the new Chief Digital Officer, Michelle Huggins.

The primary objective is to provide actionable insights that enhance business understanding, support strategic marketing and operational decisions, and contribute to improving the supply chain efficiency through predictive analytics.

This simulation was designed to mimic a real-world analytics engagement, complete with ambiguous requests, incomplete data, and cross-functional stakeholder dynamics.



### **Internal Team Structure and Methodology**

To manage complexity and ensure iterative progress, the team applied **Scrum methodology**, dividing the project into multiple sprints and different roles:

* Scrum Master: Ensured the workflow adhered to agile principles, scheduled weekly sprints, and facilitated team coordination.
* Data Engineering Lead: Managed data cleaning, preprocessing, and table joins.
* Data Analytics Lead: Carried out EDA and helped design dashboard visuals in Power BI.
* Machine Learning Lead: Led clustering and sales forecasting modeling.
* Documentation & Reporting: Responsible for formal writing and coordination of deliverables.

Each of the roles were shared among different members to ensure all were aware of different phases and steps involving the project and at the same time being 3 members could not be split into one role specific for each.

***Communication*** was managed through Slack and Whatsapp, regular weekly calls and meetings.

***Version control*** and collaboration were supported using Google Drive.

***Working practices agreed*** code will be written in english and correspondent memory will be written down and described at the end of each step in order to capture rational decision making.

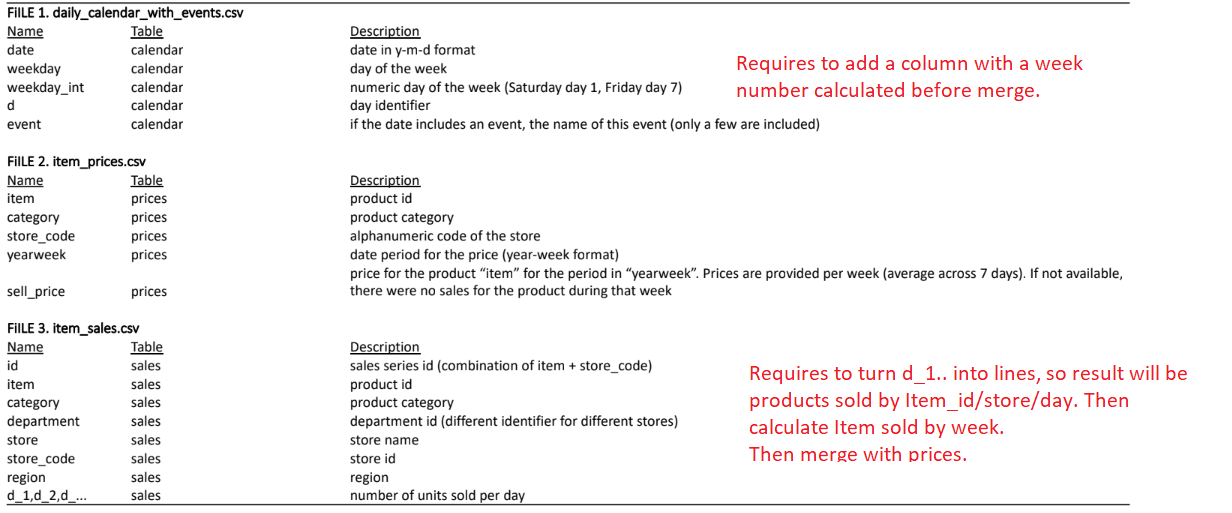
## **1. Exploratory Data Analysis (EDA)**

### **1.1. Objective**

Understand the structure, quality, and key dynamics of the 3 datasets given - Table\_Calendar, Table\_Sales, and Table\_Prices - and merge them in one data frame to be used for further analysis.

### **1.2. Methodology and initial considerations**

It was initially given the following structure as reference:



In red includes initial comments on how to proceed to connect them as initial thoughts.

From this table it can be identified key points to merge them:

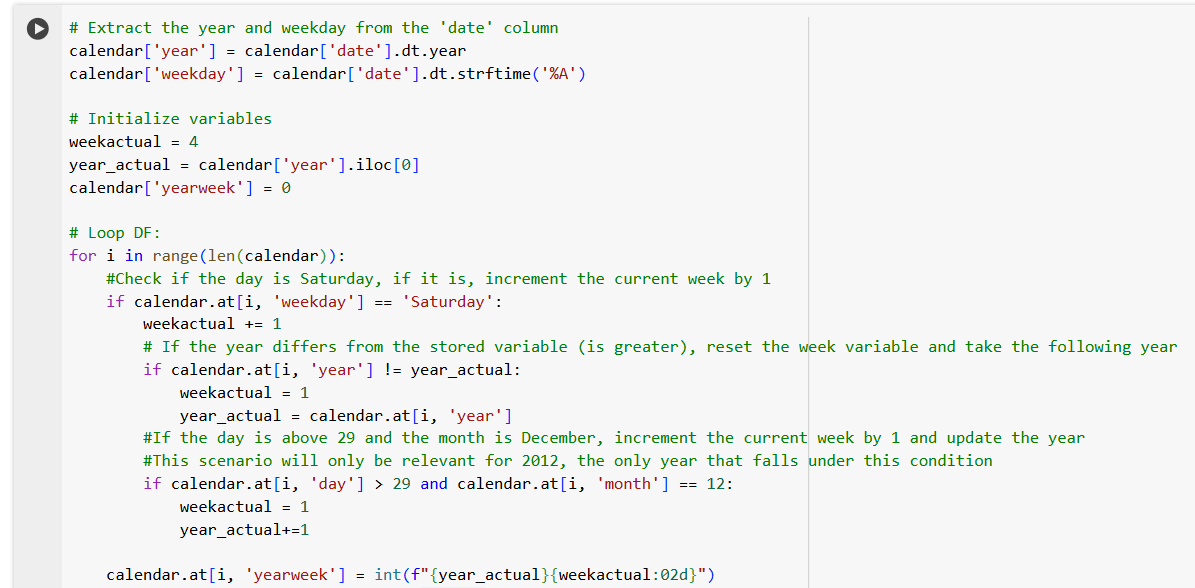
* ID product - unique id product on the database.
* ID store - unique store where items are sold.
* Day of sales - unique single day.

Initial hypothesis to be validated with full dataset:

* ‘yearweek’ metric will be what is used forward to join 3 datasets.
* If yearweek is null, there were no sales that week.

### **1.3. Actions and steps:**

1. Merge multiple data tables with the following steps:
   1. ***Table Calendar:*** create a column with the number of yearweek calculated. Understanding DS market yearweek definition is from Saturday to Friday, following the format as in table\_prices.
   2. ***Table\_Sales:*** melt columns with days into rows. Merge rows by weeks as prediction model will be used by weeks.
      1. Join both previous tables and keep all values as yearweek.
   3. ***Table prices:*** Merge with previous tables.
2. **Definition of yearweek** deepdive - based on a Saturday–Friday logic and transformed daily data to a weekly structure with the following code:



Key of understanding this step has been that only 2012 would be a year containing 53 weeks - and part of initial 2013 data of sales would be included in the previous year one.

This specification was met only in the event that day is avobe ‘29’ and month = ‘December’.

This logic was the main challenge to overcome in this part of the process to make sure data was correctly related and structured.

1. **Event definition** understanding was that will be considered effective as off the whole week, as understood that even though an event might happen only on one specific day, marketing campaigns and sales of specific items happen prior the Day, which should be reflected in the whole ‘yearweek’ period.
2. **Null values** treatment to make sure data was realistic and correctly merged.
3. Crosscheck initial hypothesis that if there are nulls in the price column means that there weren't any sales during that week**.**
4. Crosscheck no rows had sales but no price associated with it (NA) in the price column.

5. **Outliers treatment:**

To understand if all price values associated with each item and store sale were consistent:

* 1. Methodology used - Z-score as graphically or std weren’t accurate enough or viable due to large amounts of data.
  2. If an outlier is detected in price - mode will be applied instead for that product in that specific store.

1. Check if outliers are related to events or specific stores.

6. **Revenue calculation** as sell\_price\*sales for each week/item/store.

## **2. Dashboarding & KPI Visualization**

### **2.1. Objective**

Support the marketing and executive teams in understanding product performance across time, regions, and pricing strategies. This step will be **developed in Power BI** as tableau is not available to get with a student licence.

### **2.2. Methodology and initial considerations**

It was required to deliver a dashboard that would be used to keep track of relevant business metrics, so initially it was developed following a list of questions that wanted to resolve with our dashboard outcome:

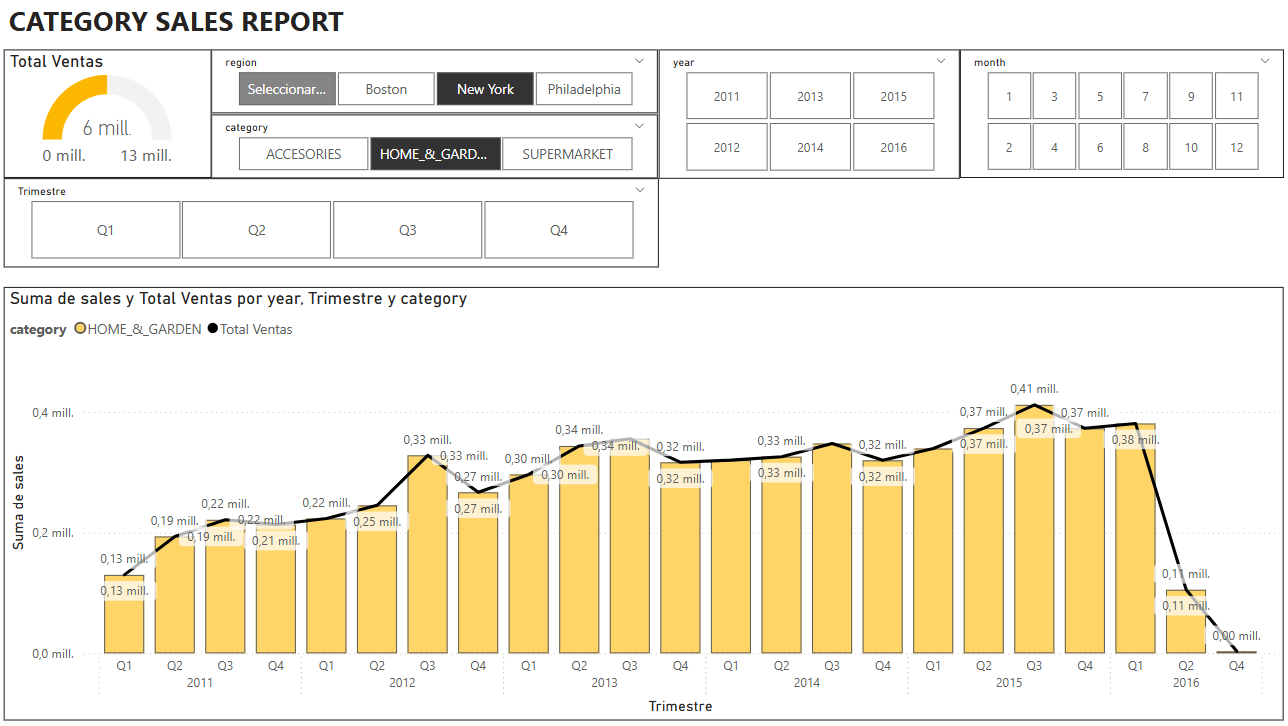
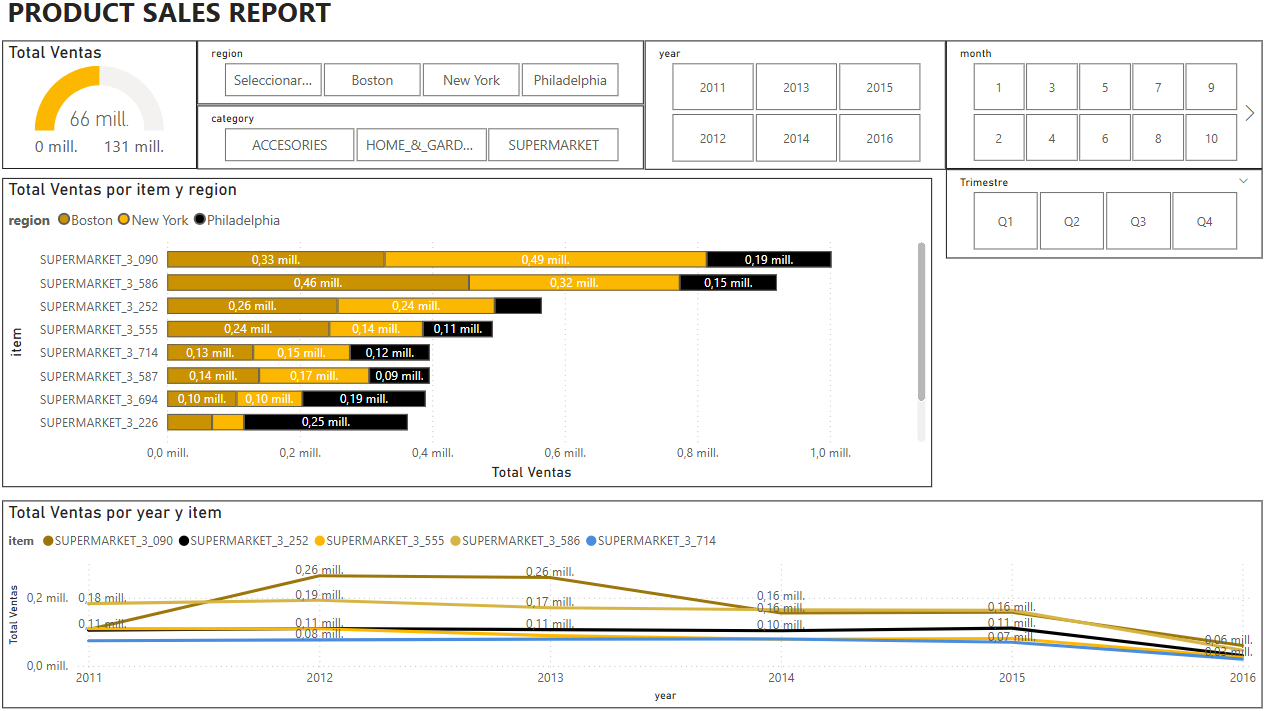
| **Analysis** | **Key Questions** |
| --- | --- |
| Total Sales per Product per Year | What are the total sales of each product by year? How do sales vary annually? |
| Top and Bottom (5–10) Products per Year | What are the best-selling and least-selling products each year? Do the trends remain consistent over time? |
| Total Sales per Product per Month | How much does each product sell monthly? Are there seasonal patterns? |
| Top and Bottom (5–10) Products per Month | What are the best-selling and least-selling products each month? Do they align with annual trends? |
| Total Sales per Product by Category | Which categories have the highest sales volume? Which products lead in each category? |
| Top and Bottom (5–10) Products by Category | Which products stand out within each category? Are there low-performing products in each category? |
| Sales Trends Over Time by Category | How have sales by category changed over time? Are there identifiable growth or decline trends? |
| Total Sales per Product by Store | What are the sales figures for each product in each store? Are there significant differences between stores? |
| Sales Trends Over Time by Store | How have sales in each store changed over time? Are there growth or decline patterns? |
| Top and Bottom (5–10) Products by Store | What are the top and bottom-selling products in each store? Are trends consistent across stores? |
| Total Sales per Product by Event | How do specific events impact product sales? Which products benefit the most? |
| Total Sales per Product per Event per Year | How do product sales vary by event across years? Which products benefit most? |
| Event Influence on Annual Total Sales | What is the contribution of events to annual sales? Which events are most significant? |
| Sales Distribution by Product Volume | How many products account for the majority of sales? Is there a sales concentration in a few products? |
| Sales Distribution by Store by Product Volume | Does each store sell a similar product mix, or are there marked differences? |
| Weekdays with Highest or Lowest Sales | On which days of the week are most products sold? Which days have the lowest activity? |
| Category Sales by Store | How do sales by category vary across different stores? Which categories dominate each store? |
| Compare Same Product (Top and Bottom 5) Across Regions | Does the same product perform similarly across regions? Where does it sell best or worst? |
| Compare Same Product (Top and Bottom 5) Across Stores | Does the same product show significant sales differences across stores? What factors might influence this? |
| Total Revenue by Category | How much revenue does each category generate? Which categories are most profitable? |
| Total Revenue by Category by Product Count | Does the number of products in a category affect revenue? Which categories generate more revenue per item? |
| Revenue by Store | Which stores have the highest and lowest revenue? How does profitability vary among stores? |
| Price Trends Over Time for Top and Bottom 5 Products | How have prices of best- and worst-selling products changed over time? |
| Price Trends Over Time for Top and Bottom 5 Products by Category | Are there specific price trends by category for the best- and worst-selling products? |
| Price Trends Over Time for Top and Bottom 5 Products by Store | How do prices of top and bottom-selling products vary across stores over time? |
| Average Price per Product | What is the average price of each product? Are there products with more stable or volatile prices? |
| Calculate Average Spending (Mean Price × Total Sales) | What is the average customer spending per product? How does it vary between products and categories? |

Finally, dashboard was organized by type of analysis that would be conducted, to ensure data visualization was clearer and each tab focused on one specific metric:

* Executive summary - Key points to support business growth
* Detail of product sales
* Detail of store sales

### 

### **2.3. Actions and steps**

* Designed dashboards featuring year-over-year (YoY) and month-over-month (MoM) comparisons.
* Displayed top/bottom product rankings across temporal and regional dimensions.
* Analyzed weekly sales trends, pricing variation, and event-based effects.



## 

## **3. Product Clustering Analysis**

### **3.1. Objective**

Segment products based on sales behavior, price sensitivity, event responsiveness, and seasonality to understand their behaviour and give to marketing team insights on different item clusters that behave closely, so then this information can be used for targeted campaigns.

This step comes after we have understood DS Market actual situation - here it is where it starts to answer ok, but now what? What comes next? what will happen?

#### 3.1.1. Clustering for stores

Store clustering was evaluated but deemed infeasible due to limited store count. Comparative store analytics were used instead inside Power BI analysis.

### **3.2. Methodology and initial considerations**

Prior to start this analysis, a new dataframe had to be build including different features definition of products - mostly based on what would be relevant from business perspective - to allow those features to be defining points on clustering analysis.

features to be created were segmented into larger groups (in line with power BI analysis) that would ensure no questions were left unanswered:

#### 3.2.1. Product - Seasonality

Objective: Identify products with seasonal purchasing patterns (e.g., products with higher demand in winter or summer).

* **Avg. Units Sold per Week**: Measure the average quantity sold weekly to detect fluctuations over time.
* **Sales Frequency**: Analyze the average number of days between repeat purchases to understand product turnover cycles.

#### 3.2.2. Product - Pricing

Objective: Assess price sensitivity among products and group them based on their elasticity to price changes.

* **Above/Below Average Price**: Identify products consistently priced above or below the category/store average.
* **Price Variance During Sales Period**: Evaluate the degree of price fluctuation throughout the product’s sales lifecycle.
* **Price Variation by Store (Same Period)**: Compare pricing for the same product across different stores during the same timeframe to identify inconsistencies or regional pricing strategies.

#### 3.2.3. Product - Events

Objective: Evaluate the impact of promotional events or campaigns on product performance.

* **Top-Selling Products During Events**: Identify which products experience significant sales boosts during events.
* **Event Influence**: Quantify the overall effect of events on sales performance at the product level.
* **Positive/Negative Product Impact**: Detect products that are positively or negatively affected by specific events.

#### 3.2.4. Product - Revenue

Objective: Determine which products contribute most and least to total revenue.

* **Highest / Lowest Revenue Contributors**: Highlight products with the greatest and lowest share of total revenue.

When it comes to which time length should be used as relevant for data analysis it was defined from the beginning that **minimum 1 year should be used as reference** to be able to capture differences between each product.

Finally a full complete year was used from yearweek 201401 to 201553 to have the latest trends captured.

When it comes to the model **k-means** was chosen and would be **validated via elbow curve** analysis.

### **3.3. Actions and steps**

#### 3.3.1. Environment Setup & Configuration

Objective: To prepare the Python environment and its libraries for the data analysis tasks.

This initial phase involves importing all necessary tools.

Key libraries include **pandas** for creating and managing DataFrames, **numpy** for efficient numerical calculations, and **matplotlib/seaborn** for data visualization. Crucially, **scikit-learn** is configured with set\_config(transform\_output="pandas").

#### 

#### 

#### 3.3.2. Data Preprocessing Pipeline

Objective: To clean, transform, and standardize the raw data, making it suitable for the clustering algorithm.

A scikit-learn **Pipeline** is used to chain all preprocessing steps, ensuring a consistent and reproducible workflow. The key actions are:

* **Imputation with KNNImputer:** Missing data by using the values from the 'k' most similar samples (nearest neighbors) to estimate and fill in the missing ones.
* **Scaling with StandardScaler:** Distance-based algorithms like K-Means are sensitive to feature scales. StandardScaler rescales each feature to have a mean of 0 and a standard deviation of 1, giving all features equal weight and ensuring the clustering is based on true patterns, not arbitrary scales.

#### 3.3.3. Dimensionality Reduction

Objective: To reduce the number of features of the dataset while preserving as much important information as possible.

To combat the "curse of dimensionality," **Correlation** is used to reduce model complexity.

For this modeling principal components were selected and **excluded correlation of above 95%.**

#### 3.3.4.Clustering Model & Optimization

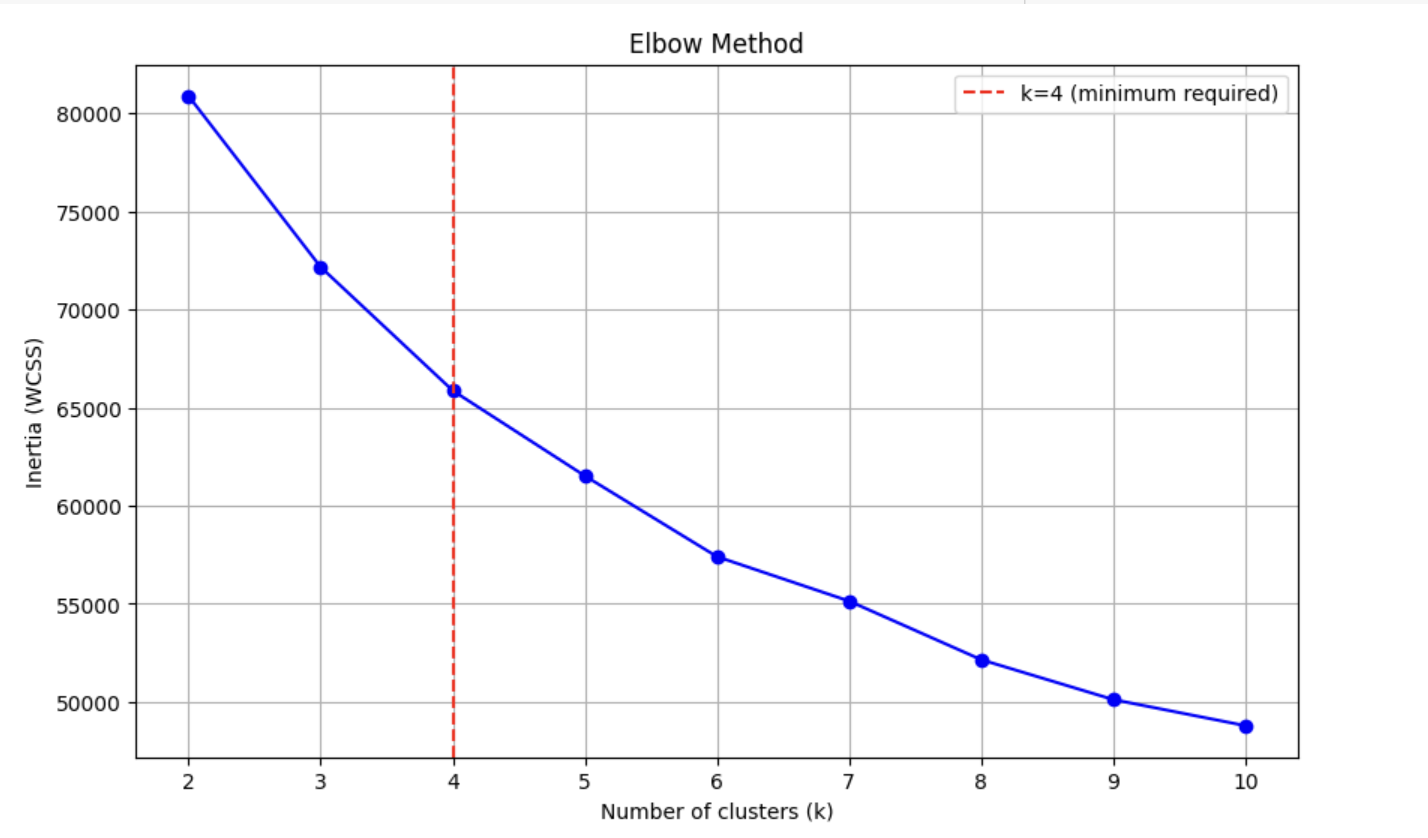
Objective: To group the preprocessed data into distinct clusters and scientifically determine the best number of clusters to use.

The core clustering task is performed using the **K-Means** algorithm.

The key challenge faced in this step was choosing the number of clusters, k. Main metric learned during the course to determine k has been the elbow curve, but sometimes can be quite tricky to identify graphically.

To solve this, an optimization loop had been runned, training the model with different k values.and **silhouette\_score** had been calculated for each iteration

Finally, the best iteration has been chosen as the most suitable as the number of k (the one with the highest value).



#### 3.3.5. Cluster Analysis and Interpretation

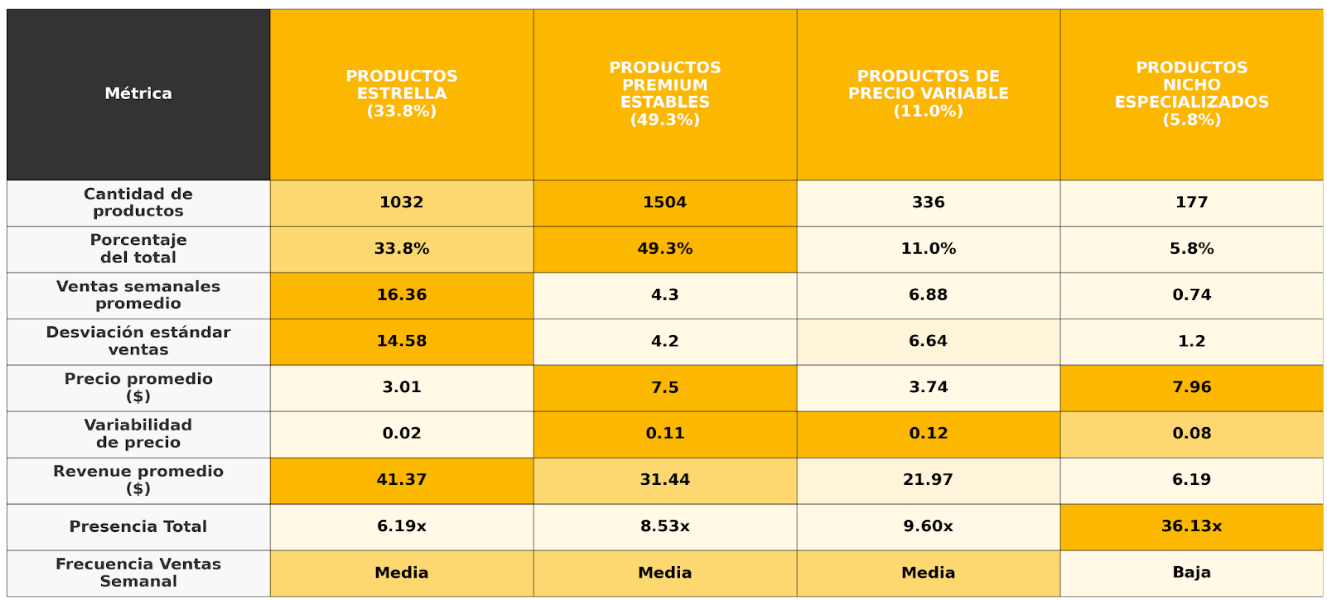
Objective: To translate the numerical cluster assignments into meaningful business insights by understanding the defining characteristics of each segment.

This by far, has been the hardest part of the project as data understanding it is quite complex in this case.

The actions performed are:

1. **Descriptive Statistical Analysis by Cluster:** The dataset is grouped by the new cluster column. For each cluster, key descriptive statistics (e.g., mean, median, standard deviation) are calculated for every original feature. This reveals the quantitative profile of a typical member of each cluster.
2. **Identification of Defining Characteristics:** By comparing the statistical profiles, we can identify what makes each cluster unique. This step answers critical business questions: "Which cluster contains the highest-priced products?", "Is there a segment with significantly higher sales volume?", or "What defines our smallest niche segment?".
3. **Comparative Visualization:** To make interpretation intuitive, visualizations are created to compare the clusters such as box plots or bar charts.
4. **Creation of Business "Personas":** This is the final synthesis where quantitative data is translated into qualitative archetypes. Each cluster is given a descriptive name or "persona" that summarizes its core identity, making the results accessible to all stakeholders.

Clusters achieved have been defined as follows:



#### 3.3.6. Final Assignment

Objective: To apply the final cluster labels to the dataset and save the results for analysis.

This label is added as a new column, **cluster**, to the original DataFrame so will be able to use this figure for further analysis.

## **4. Sales Forecasting**

### **4.1. Objective**

Develop a predictive modeling framework for store-product level weekly sales, leveraging advanced time-series methodologies to replace conventional heuristic forecasting approaches capable of delivering precise 28-day rolling - in our model segmented by weeks as 7 days each.

### **4.2. Methodology and initial considerations**

#### 4.2.1. Data Enhancement and Feature Engineering

The model development process utilized a pre-existing dataframe that required substantial enhancement to improve predictive accuracy.

#### 4.2.2. Core Methodology and Performance Framework

The methodology prioritizes the identification of optimal model-entity combinations through systematic performance assessment, ensuring deployment readiness and superior predictive capabilities compared to existing approaches.

* Comparative evaluation of multiple forecasting algorithms with emphasis on gradient boosting methodologies.
* Systematic performance assessment using industry-standard metrics (RMSE, MAPE, bias analysis).
* Identification and selection of optimal model-entity combinations for deployment.
* Demonstration of superior predictive performance over existing heuristic forecasting methods.

#### 4.2.3. Algorithm Selection and Modeling Approach

A critical consideration in the modeling approach was the dataset's **non-conventional temporal structure,** which employed a "yearweek" metric rather than standard datetime formatting.

This unique data structure, combined with the complex non-linear relationships inherent in sales demand patterns, necessitated careful algorithm selection to ensure optimal predictive performance.

Understanding these restrictions, **XGBoost** was finally used.

#### 4.2.4. Performance Evaluation Framework

Model performance evaluation was conducted using a comprehensive multi-metric approach to ensure thorough assessment of predictive capability and reliability across different entities and time periods.

* **RMSE (Root Mean Squared Error)** as the primary metric for measuring average prediction error magnitude and penalizing large deviations.
* **Cross-validation scores** across different time periods to assess model stability and consistency

### **4.3. Actions and steps**

#### 4.3.1. Data Preparation

This critical initial phase establishes the analytical framework and prepares raw data for modeling.

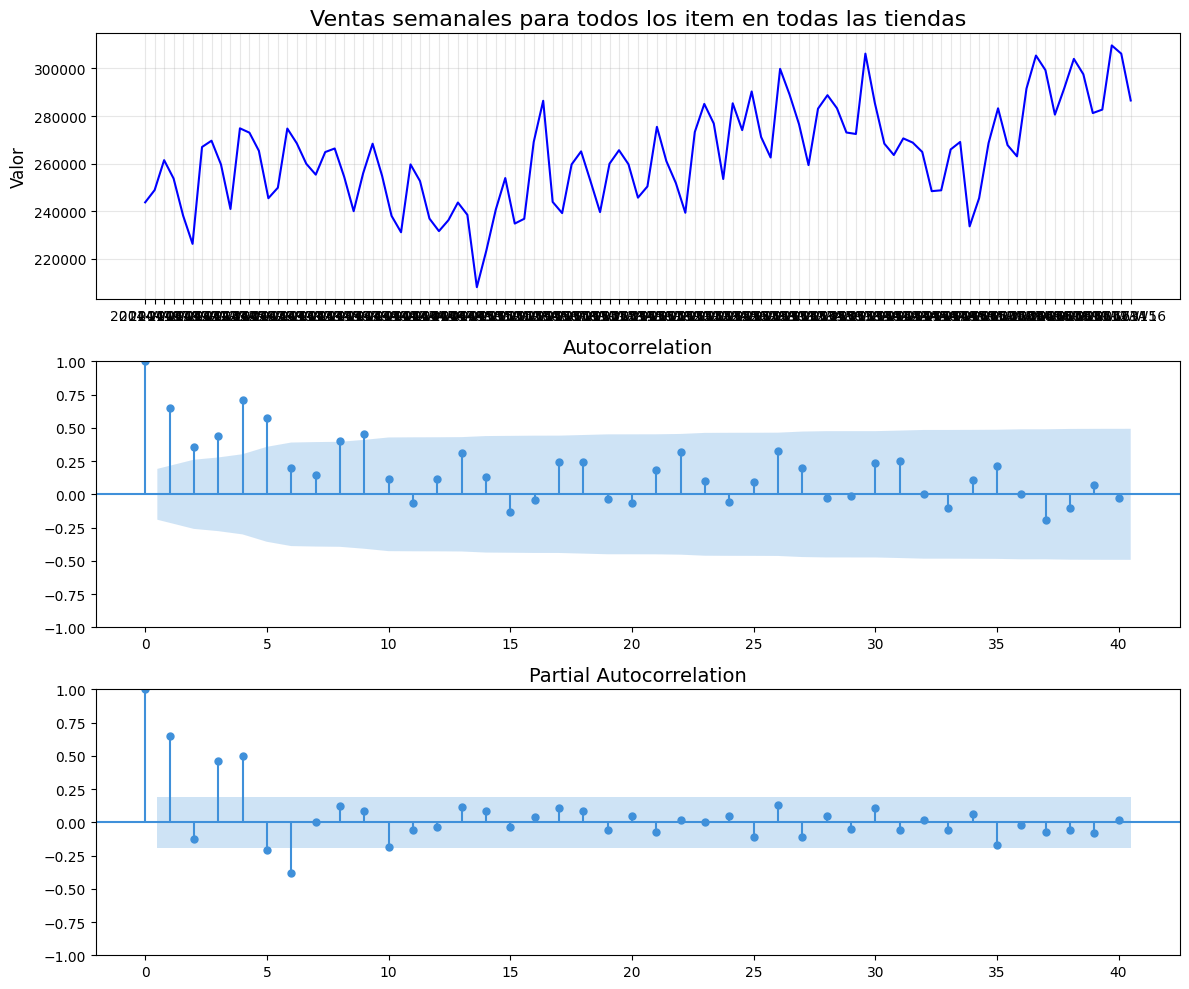
Missing data was **systematically imputed**, using logic tailored to the nature of each feature:

* Time series features: forward-filled within logical groups (e.g., item/store).
* Event flags: defaulted conservatively (e.g., 0 for no event).
* Market trends and price features: filled with median or trend-based methods.

#### 4.3.2. Correlation analysis

Understanding the underlying time series characteristics is essential before model development. This step was especially complicated as products and time series to be analyzed in this case were multiple, so it required multiple analysis to understand the trend.

Below one of the most useful examples to understand it - sales aggregate by week, regardless of the year.:



As a conclusion, it can be easily observed this series is NOT stationary (p-value: 0.240465 > 0.05), which will require transformations before predictive modeling.

#### 4.3.3. Feature creation

The process involves integrating various business-relevant signals from multiple sources while rigorously preventing **data leakage** based on previous analysis.

1. Multi-Dimensional Feature Enrichment: Sales and pricing history were added from different angles:
   * Product-Level: Recent sales trends (last 1, 7, 14 weeks).
   * Category-Level: Aggregated behavior by product type.
   * Store-Level: Weekly sales dynamics by store.
   * Store-Category Mix: Sales trends for categories within stores.
   * Price Signals: Recent pricing trends at the item level.
   * Market Signals: Overall market behavior to add macro trends.
2. Event and Seasonality Integration  
   * Included encoded features for holidays and special events.
   * Captured seasonal patterns using trigonometric transformations (sin/cos of time).
3. Contextual & Behavioral Factors
   * Integrated metrics like product counts per store and sales performance during events.
   * Captured category-level event responsiveness and typical item stocking behavior.

#### 4.3.4. Model evaluation and training

Prior using the model, the following criteria has been established to part data into different parts:

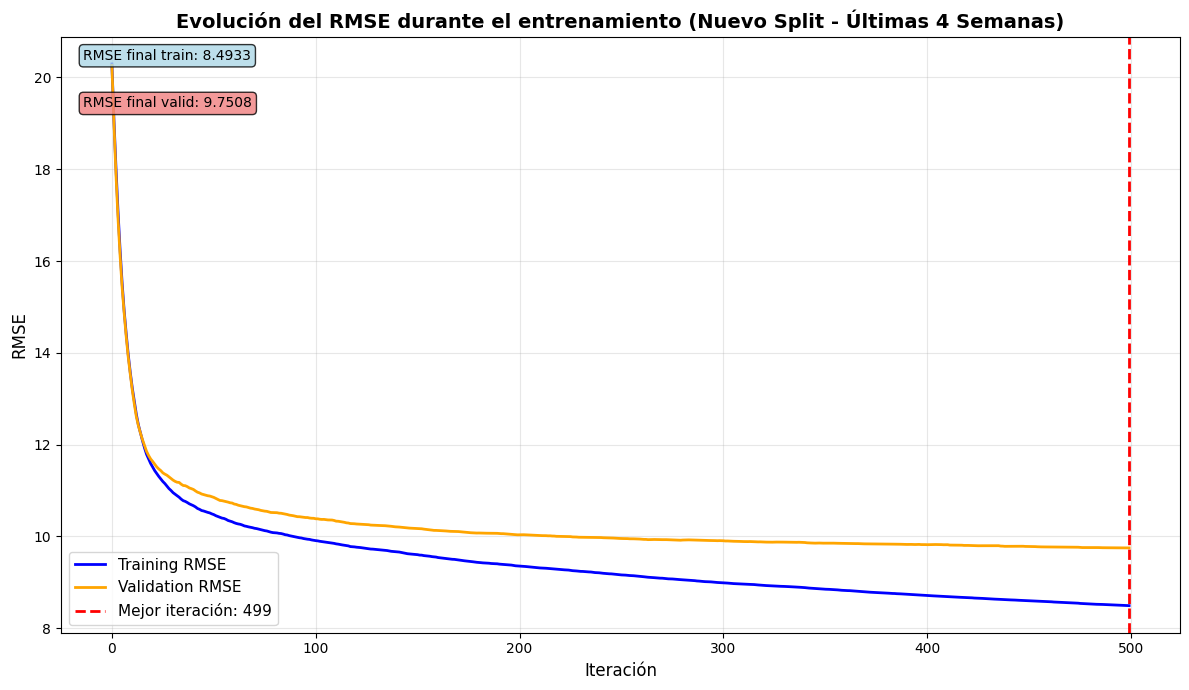
* The last **4 weeks** in the dataset are designated as the **TEST** set.
* The **4 weeks prior** to those are designated as the **VALIDATION** set.
* All previous data becomes the **TRAINING** set.

Based on this, the following results were achieved for the whole dataset. It can be observed that there is a slight overfitting, which results in a good output from the model.

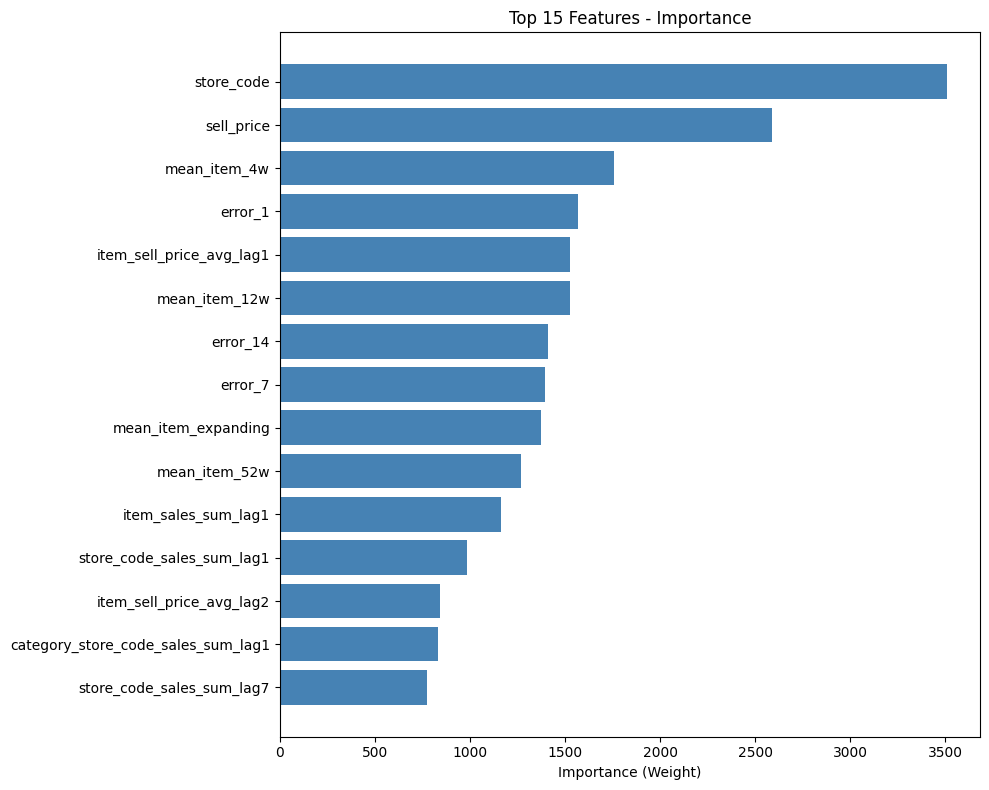
Gap Train → Test: 1.4736

Gap Valid → Test: 0.2161

Overfitting: Moderated (gap < 2.0)



Main features that delivered what we considered optimal result had been relevant as follows:



## **5. Replenishment & MLOps Proposal applied to pilot case study**

### **5.1. Objectives**

On this section two main objectives were defined:

1. Apply sales forecasting to store replenishment processes, minimizing overstock and stockouts.
2. Develop a pilot case study on how, when and where would start to be tested in a replenishment case with MLOPS insights.
   1. Demonstrate the business impact of predictive analytics on sales and inventory management.

### **5.2. Methodology and initial considerations**

Initial key point of this exercise was to fully understand how stock process worked and which had been historical stock flow. Main goal was to iIntegrate logic to align product arrival with predicted demand

Weekly prediction output used for next week’s supply, **changing store replenishment definition** from Monday to Sunday to Saturday to Friday, to be able to cover better weekend demand - as it has been historically higher.

Designed system for weekly batch execution (not real-time API) was selected as only needed 1 execution a week.

Pilot case was proposed to be developed with the most challenging environment as off to fully demonstrate the efficiency of this new approach to stock supply - therefore, most sold product in the biggest store, as smaller was easier to manage and control as fluctuation of demand was more stable along the time.

Use case with item Supermarked\_3\_090 in NCY\_4

### **5.3. Actions and steps**

#### 5.3.1. Actual stock flow

Understanding actual workflow - by making a few assumptions - was key to be able to use it as a baseline to compare it with outputs from forecasting models and spot main differences and quantify improvements.

Assumptions in stock:

* DS market actual forecasting method uses average last year sales as of how to stock stores.
  + if actual store falls below 50% of expected sales, a midweek restock is assumed but, given that a lot of products have been sold, store will be restored with an addition of 20% to ensure demand is covered.
* Supermarket items would expiry quite fast - 2 weeks were used for this exercise.

Assumptions in costs:

* Average cost of supermarket items represents 36,65% of item\_price.
* transport cost represents 5% of item\_price per article, all included inside average cost.
* if a product has expired (aka not sold during 2 weeks) would represent a loss of 63,35% from item\_price.

#### 5.3.2. Forecasted stock flow

Same assumptions were added but demand of sales per week substituted by forecasting output.

#### 5.3.3. Comparison

It was demonstrated with these analyses that iInitial model assumed larger badges of supply but would allow larger quantities of product to expire on shelfs, whereas the forecasted approach would represent more transport but wouldn't be incurred in expiry loss.

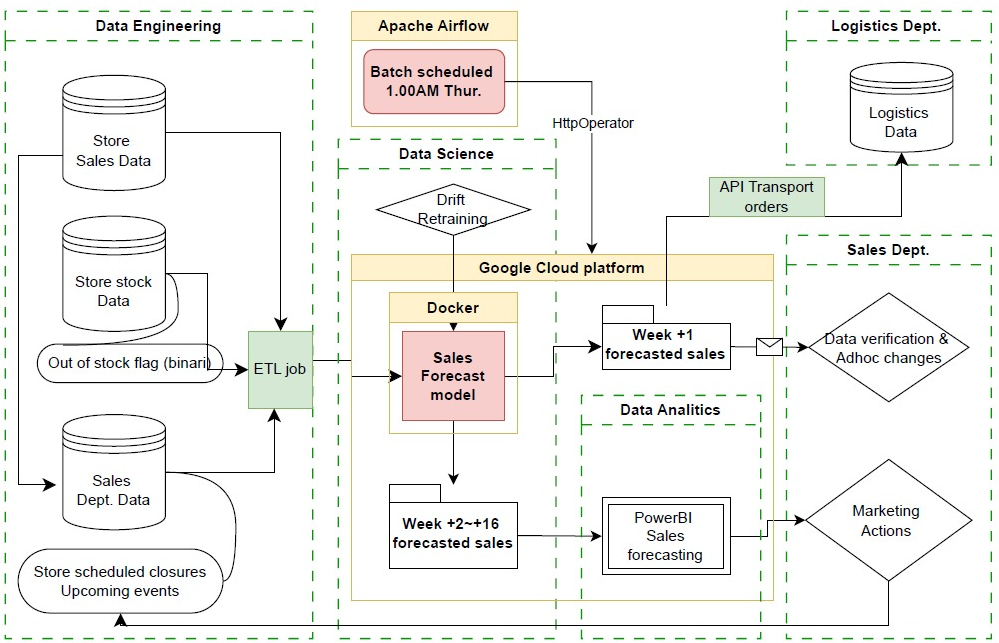
Overall, it is cost wise more economic to do different transports a week assuming forecasted.

Given an item picked to do a study case, it would represent the initial model vs. forecasted one 27% more benefits than previously. It needs to be addressed that not all forecasted items do have the same RMSE accuracy, therefore the result might vary.

### **5.4. MLOPS workflow proposed**

Justification for different options included:

* Apache airflow as it can be easily integrated with Google cloud platform
* Google Cloud base server is a better option for a retail business than AWS.



## **Conclusions**

The TFM project successfully addressed complex retail data analysis challenges through a systematic approach encompassing comprehensive data exploration, advanced visualization techniques, predictive modeling, and seamless operational integration. The initiative demonstrated the practical application of data science methodologies in solving real-world business problems within the retail sector.

By leveraging agile development methodologies combined with sophisticated analytical techniques, the project delivered substantial value beyond traditional analytical outputs. The developed framework provides not only meaningful business insights but also actionable decision-making tools that directly support strategic planning and operational efficiency. The implemented models and interactive dashboards are designed with adaptability and scalability at their core, ensuring seamless integration into existing business systems and processes.

The strategic implementation of collaborative teamwork principles, reproducible modeling frameworks, and dynamic interactive dashboards effectively simulated the high-impact contributions expected from data science professionals in rapidly evolving retail environments. This comprehensive approach validates the methodology's effectiveness for future data science initiatives and establishes a solid foundation for continued analytical advancement.

The DSMarket project exemplified a complete end-to-end data science pipeline, serving as a valuable proof-of-concept that rigorously tested analytical reasoning capabilities and technical proficiency in Python-based solutions. This foundational experience provides essential insights for subsequent project development and methodology refinement.

Ultimately, this project has empowered key stakeholders at DSMarket with access to actionable insights while establishing a robust analytical infrastructure that supports ongoing data science initiatives and drives continued business transformation through data-driven decision-making.