Track B(MVP) Final Report

Richard Appiah
appiah@students.uni-marburg.de

Solomon Mwanga
mwanga@students.uni-marburg.de

Gzouni@students.uni-marburg.de
Salem Gzouni

elkarhan@students.uni-marburg.de
Youssof El Karhani

1. Motivation and Dataset

This dataset comprises structured transactional data pertaining to product orders, with key attributes providing a comprehensive overview of each sales event. Below is a detailed description of the most salient features relevant to the project's analytical objectives:

ORDERNUMBER provides a unique identifier for individual transactions, allowing us to distinguish one sale from another. The **ORDERDATE** precisely records when an order was placed, which is essential for conducting **temporal analysis** and uncovering sales trends over time. We can gauge **product sales volume** by examining **QUANTITYORDERED**, while **PRICEEACH** offers insights into pricing strategies and potential profit margins. The **SALES** figure, derived from the product of unit price and quantity, represents the most direct measure of revenue generated from each order line. **PRODUCTLINE** categorizes the item sold, enabling systematic analysis of how different product categories perform in terms of sales efficacy and customer demand.

The qualitative descriptor **DEALSIZE** indicates the transaction's magnitude, helping us understand the scale of individual orders and segment them accordingly. **COUNTRY** specifies the customer's geographical location, proving invaluable for regional analysis and identifying geographic sales trends. **STATUS** indicates the current state of an order (e.g., completed, pending, or shipped), allowing us to monitor the order lifecycle and assess fulfillment efficiency. Finally, **CUSTOMERNAME** contains the customer's identity; while not always central to large-scale aggregate analysis, it can be beneficial for more granular customer segmentation or personalized marketing efforts.

This dataset is characterized by a high degree of structure and validation, which significantly facilitates its analytical utility. Specifically, the **numerical fields**, encompassing **SALES**, **QUANTITYORDERED**, and **PRICEEACH**, are consistently populated with accurate figures, demonstrating an absence of significant outliers or missing values. The **temporal field**, **ORDERDATE**, maintains a uniform format, enabling straightforward parsing into a datetime object. This uniformity is crucial for conducting sophisticated **time-series analyses**, such as tracking sales evolution over specific periods.

Furthermore, the **qualitative fields**, including **PRODUCTLINE** and **DEALSIZE**, are consistently populated and uniformly formatted, thereby enabling facile segmentation and categorization of data. While it is acknowledged that a limited number of blank entries exist within secondary features such as STATE and ADDRESSLINE2, these omissions do not substantially impede the core analytical objectives, as these particular attributes are not critical for revenue, product or primary customer analysis

This dataset unlocks several significant analytical opportunities. We can analyze **revenue trends** by examining the total sales value over time using the **ORDERDATE**, which helps reveal seasonal patterns, growth trajectories, or declines in sales performance. Furthermore, by evaluating **PRODUCTLINEs** or individual products, we can pinpoint which offerings demonstrate superior performance in terms of sales volume and quantity ordered. The dataset also enables **customer segmentation** based on **COUNTRY** or **DEALSIZE**, providing insights into customer behaviors that can inform the refinement of marketing and sales strategies.

The motivation for choosing this dataset was that it directly addresses real-world business challenges in marketing. AI plays a crucial role in precisely targeting customers, optimizing

advertising expenditure, and elevating conversion rates, which are very critical functions in contemporary business operations. We make working on this interesting because the insights derived from this dataset will profoundly impact MVP's business strategies by enhancing customer engagement, increasing sales, and improving their advertising return on investment. This makes it a highly impactful project, clearly demonstrating AI's capacity to deliver substantial business success.

2. Technical Solution

2.1 Data Loading

The solution includes loading of the dataset. It uses `files.upload()` from `google.colab` to allow users to upload a CSV file. The code then attempts to read the CSV, first with `utf-8` encoding and falling back to `latin1` in case of a `UnicodeDecodeError`, ensuring broad compatibility with different file encodings. Crucially, the 'ORDERDATE' column is immediately converted to datetime objects using `pd.to_datetime()`, preparing it for temporal analysis.

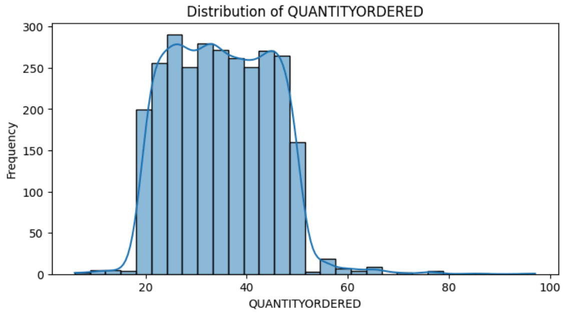
2.2 Initial Exploration and Data Analysis

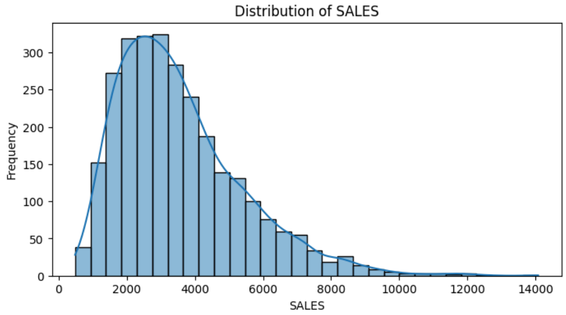
Initial data exploration was carried out, more specifically focusing on generating numeric summaries and value counts for key categorical fields. This step typically involves:

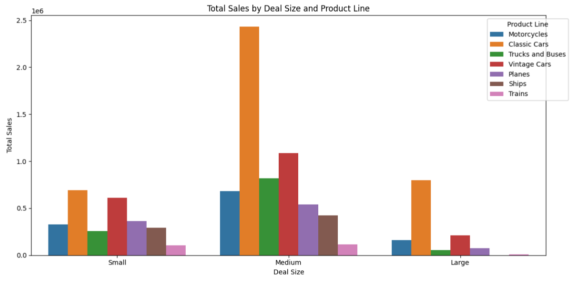
Numeric Summaries: Calculating descriptive statistics (count, mean, standard deviation, min, max, quartiles) for numerical columns such as 'SALES', 'QUANTITYORDERED', and 'PRICEEACH'.

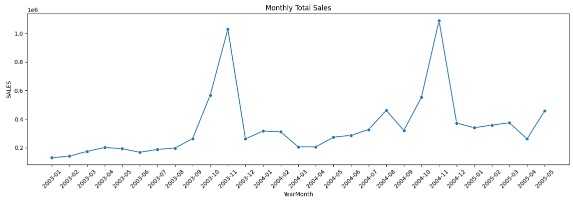
Value Counts: Displaying the frequency of unique values for categorical features like 'PRODUCTLINE', 'COUNTRY', and 'DEALSIZE'. This provides an immediate understanding of the distribution and common categories within these fields.

Below are a few of the different distributions for insight into the data









This initial exploration is fundamental for understanding the dataset's structure, identifying data types, checking for missing values, and gaining preliminary insights into the distributions of key variables before proceeding with more complex analytical tasks.

2.3 Data Preprocessing

Data preprocessing was a crucial step to prepare the raw data for analysis and modeling. The techniques applied include:

- **Column Dropping and Selection**: The process involved strategic column removal to streamline the dataset and eliminate irrelevant or redundant information. Specifically, columns such as 'PHONE', 'POSTALCODE', 'ADDRESSLINE2', and 'TERRITORY' are systematically dropped if they exist within the DataFrame, as they are deemed secondary and not critical for the core analytical objectives. Following this, missing values in the 'STATE' column are addressed by imputing them with 'Unknown'. This ensures completeness for a potentially useful geographical attribute, even if granular state-level data is not always available.

  Further, the solution enriches the dataset by **extracting temporal features** from the 'ORDERDATE' column. The year and month components are derived and added as new, distinct columns ('year' and 'month'). This temporal decomposition is fundamental for time-series analysis, enabling the identification of seasonal trends and long-term patterns

- **customer-level aggregation** was performed to derive new, insightful features. This involves grouping the data by 'CUSTOMERNAME' and then applying transformation functions to calculate aggregate metrics for each customer. Specifically, `total_rev_per_customer` is computed by summing the 'SALES' for each customer, providing a cumulative measure of their revenue contribution. Concurrently, `order_count_per_customer` is determined by counting the number of unique 'ORDERNUMBER' entries per customer, indicating their purchasing frequency. These aggregate features enrich the dataset with higher-level insights into customer behavior.
- **Ordinal encoding** is applied to the 'DEALSIZE' categorical variable. Given that 'DEALSIZE' inherently possesses an ordered relationship (Small < Medium < Large), it is mapped to a numerical scale (1, 2, 3 respectively). This transformation converts a qualitative ordinal variable into a quantitative one, preserving the intrinsic order and making it amenable to numerical algorithms.
- **One-hot encoding** is utilized for nominal categorical features, specifically 'PRODUCTLINE', 'COUNTRY', and 'STATUS'. This technique converts each category into a new binary (0 or 1) column, where a '1' indicates the presence of that category and a '0' indicates its absence. The `drop_first=True` parameter is strategically employed during this process. This practice mitigates the issue of multicollinearity, which can arise when a perfect linear relationship exists between predictor variables, thereby ensuring the independence of the newly created dummy variables and improving the stability of subsequent statistical or machine learning models

2.4 Customer segmentation (agglomerative clustering)

With the RFM features appropriately scaled and prepared, **Hierarchical Clustering**, the **Agglomerative Clustering** method, was employed to identify distinct customer segments. This approach builds a hierarchy of clusters by progressively merging individual data points or existing clusters.

A predefined number of `k=3` clusters was chosen, instructing the algorithm to partition the customer base into three primary groups. The `ward` linkage method was selected for the clustering process.

This method is particularly effective as it aims to minimize the variance within each merged cluster, tending to produce compact and well-defined groups. Upon execution, the algorithm assigned a unique cluster label (0, 1, or 2 in this case) to each customer based on their RFM characteristics, effectively grouping them into their respective segments.

To quantitatively assess the quality of the resulting clusters, the **Silhouette Score** was calculated. This metric provides a measure of how similar an object is to its own cluster compared to other clusters. A higher Silhouette Score indicates that the clusters are well-separated and internally cohesive, confirming the effectiveness of the segmentation.

The **Gaussian Mixture Models (GMM) clustering** was also applied to the scaled RFM data. GMM, which assumes data points within clusters are probabilistically generated from Gaussian distributions, provided an alternative perspective on customer groupings, particularly effective for identifying clusters with varying shapes or densities. The profiles of these GMM clusters were then examined by calculating the average RFM values for each segment. To visually represent these identified customer segments, **Principal Component Analysis (PCA)** was performed, reducing the high-dimensional RFM data into two principal components (PC1 and PC2). These components were then used to create a 2D scatter plot, clearly illustrating the spatial distribution and separation of customer segments. Finally, the cluster assignments from both the hierarchical and GMM models were meticulously merged back into the original sales dataset, enabling comprehensive analysis of transaction details within the context of their assigned customer segments

2.5 Time Series Forecasting

The Prophet model was used in predicting future sales trends, particularly in accounting for seasonal fluctuations and utilizing historical sales data. The process, which involved preparing monthly aggregated sales data, fitting the Prophet model, and generating forecasts, was proved to be a reliable method for projecting future sales performance..

By incorporating the Prophet model's capabilities to model seasonal trends, the forecasts generated offered a sound basis for anticipating future demand and optimizing operational decisions. Specifically, the identification of months with the highest forecasted sales can inform targeted marketing campaigns inventory planning, ensuring that resources are allocated effectively in anticipation of peak sales periods**.**

2.6 Predict modelling pipeline

the Apriori algorithm was implemented to effectively uncovered meaningful associations between products in the sales data. By transforming the data into a suitable format, applying frequent itemset mining, and generating association rules based on lift, this approach was able to identify the most significant product pairings. which provided more insightsabout enhancing customer experiences through personalized recommendations, and driving targeted campaign.

3. Insights

Trend: The sales forecast shows a clear cyclical or seasonal pattern with peaks and troughs over time. There are notable high peaks around mid-2004 and late 2005.

Volatility: The forecast indicates significant fluctuations in sales, suggesting that demand for MVP's products might be influenced by external factors or specific periods.

Uncertainty: The shaded area around the forecast line represents the confidence interval, indicating the range within which actual sales are expected to fall. This band widens towards the end of the forecast, implying greater uncertainty further into the future.

Specific Forecasted Sales: The sales forecast graph provides specific monthly sales predictions, with the highest predicted sales occurring in August 2005 ($1,203,896) and December 2005 ($1,217,906), aligning with the visual peaks.

Extremely Strong Co-Purchasing Patterns: The data reveals almost perfectly correlated purchasing behavior among specific product categories. When certain combinations of products, predominantly Planes and Vintage Cars, are purchased (as antecedents), specific Vintage Cars, Planes, and Ships are *always* purchased alongside them (100% confidence).om these categories are purchased.

Niche, Yet Potentially High-Value, Transactions: While the support values are relatively low (~2.28%), meaning these specific exact combinations occur in a small percentage of overall transactions, the perfect confidence and very high lift indicate that *when* these specific purchasing events happen, they follow a very strict and predictable pattern. This points to a distinct segment of customers whose buying behavior is highly structured for these particular product groupings, potentially representing high-value, comprehensive orders rather than frequent, small ones

4. Recommendation

Based on these insights, here are some recommendations for MVP:

4.1 Prioritize High-LTV Customers for Exclusive Campaigns:

Focus: Direct a significant portion of the targeted ad budget towards "Euro Shopping Channel," "Mini Gifts Distributors Ltd," and "La Rochelle Gifts." These are the most valuable assets.

Strategy: Develop highly personalized and exclusive campaigns for these top-tier customers. This could include Dedicated account managers.

- Early access to new products or special promotions.
- Volume discounts or loyalty programs tailored to their purchasing patterns.

4.2 Leverage Sales Forecasts for Campaign Timing and Inventory:

- **Timing:** Plan ad campaigns to coincide with predicted sales peaks (before August and December based on the provided forecast). This can amplify the impact of their marketing efforts.

4.3 Segment Customers Beyond Top 10 LTV for Tiered Campaigns:

- **Mid-Tier Customers:** Identify the next tier of high-LTV customers (e.g., beyond the top 10) and design campaigns to nurture them, aiming to increase their LTV. This could involve targeted promotions or special offers that encourage repeat purchases.

5. Reflection

This sales data analysis project has been a rewarding learning journey, and upon reflection, I realize how the final project evolved from the original plan. While the initial vision was to encode the relevant feature data pass it to the clustering algorithms , the project grew to incorporate more advanced methodologies and a deeper exploration of the dataset than initially anticipated.

1. **Expanded Data Handling and Preprocessing:**
   Initially, the plan was to handle the data using pandas for cleaning and basic transformations. However, as the complexity of the dataset became more apparent, we had to implement more advanced preprocessing steps. This included working with different file encodings, converting date formats (ORDERDATE), and applying various

scaling methods like StandardScaler, RobustScaler, and MinMaxScaler to prepare the data for machine learning algorithms.

2. **Advanced Exploratory Data Analysis (EDA) & Visualization:**
   While the original plan involved using basic visualizations, the project evolved to include a more comprehensive EDA using seaborn and matplotlib.pyplot. This allowed me to visualize sales distributions, product popularity, and regional performance, providing deeper insights than initially planned. This shift emphasized the value of visual exploration before diving into more formal modeling.

3. **Incorporating Clustering Techniques:**
   The plan initially focused on using K-Means clustering, but eventually expanded this to include AgglomerativeClustering and GaussianMixture models. The project grew to include methods for evaluating the effectiveness of these clustering techniques using silhouette_score. This was a significant shift from the original focus on a single clustering model.

4. **Dimensionality Reduction:**
   The use of Principal Component Analysis (PCA) was not originally planned, but it became a necessary step as we sought to manage high-dimensional data and improve model performance. The addition of PCA helped with both visualization and the performance of machine learning models, especially when dealing with large datasets.

5. **Time Series Forecasting:**
   Initially, the project was meant to provide an overview of sales data, but extended the scope to include predictive analytics using Prophet. This addition allowed for the forecasting of future sales trends, which added significant value in terms of business intelligence and future planning.

6. **Association Rule Mining:**
   Initially not considered, association rule mining became a crucial part of the project as we explored product relationships using the mlxtend library and the Apriori algorithm. This shift allowed us to discover hidden patterns in the data, useful for cross-selling and recommendation systems.

In conclusion, the final project expanded significantly from the original plan, incorporating more advanced techniques and broader insights into the sales data. This allowed for a deeper understanding of the dataset and the development of more sophisticated business intelligence tools. The journey from initial exploration to actionable insights through machine learning models highlighted the importance of flexibility in project planning, allowing for continuous adaptation and improvement.