Exploratory Data Analysis Report for Store Sales Time Series Forecasting

Contents

1 Introduction

This Exploratory Data Analysis (EDA) report examines the dataset for the store sales time series forecasting project. The dataset contains sales records across multiple stores, product families, and time periods, enriched with external factors such as oil prices, weather conditions, and event information. The objective of this EDA is to uncover patterns, relationships, and insights that can inform feature engineering and model development for accurate sales forecasting.

2 Dataset Overview

The dataset comprises 2,061,903 records and 16 columns, capturing sales and related data from 2013 to 2017. Below is a detailed description of the columns:

- **date**: Date of the sales record (datetime64[ns]).
- **store_nbr**: Store identifier (int64, 1 to 54).
- **family**: Product family (string, e.g., GROCERY, PRODUCE).
- **unit_sales**: Number of units sold (float64, includes negative values for returns).
- **onpromotion**: Indicates if the product was on promotion (int8, 0 or 1).
- **city**: City where the store is located (string).
- **state**: State where the store is located (string).
- **store_type**: Type of store (string, e.g., A, B, C).
- **perishable**: Indicates if the product is perishable (int8, 0 or 1).
- **transactions**: Number of transactions at the store (int64).
- **oil_price**: Daily oil price (float64, relevant due to economic impact).
- **day_type**: Type of day (string, e.g., Work Day, Holiday, Event).
- **Event Scale**: Scale of the event (string, e.g., Non, Local, National).
- **locale_name**: Name of the locale for events (string).
- **temperature**: Daily temperature in řC (float64).
- **precipitation**: Daily precipitation in mm (float64).

2.1 Data Quality

- **Missing Values**: No missing values were found (df.isnull().sum() = 0).
- **Duplicates**: No duplicate records were detected (df.duplicated().sum() = 0).

- **Data Types**: The data types are appropriate, with numerical columns (int64, float64, int8) for quantitative variables and string types for categorical variables. The date column is correctly formatted as datetime64.

2.2 Temporal Coverage

The dataset spans from January 1, 2013, to August 2017, covering:

- **Years**: 2013, 2014, 2015, 2016, 2017.
- **Months**: January to December, with varying coverage per year.
- **Days**: All days of the month are represented, with specific days (e.g., 1830) appearing more frequently in the sample.

This temporal range allows for the analysis of seasonal patterns and long-term trends.

3 Data Exploration

3.1 Summary Statistics

The numerical columns provide insights into the scale and variability of the data:

- **unit_sales**: Mean = 531.81, Min = -12,240, Max = 1,033,988. Negative values indicate returns or adjustments.
- **transactions**: Mean = 116,768, Min = 0, Max = 835,784. High variability suggests differences in store traffic.
- **oil_price**: Mean = 63.62, Min = 26.19, Max = 114.90. Reflects economic fluctuations.
- **temperature**: Mean = 17.26řC, Min = -9.7řC, Max = 33.7řC. Captures diverse weather conditions.
- **precipitation**: Mean = 8.04 mm, Min = 0 mm, Max = 305.1 mm. Indicates varying rainfall levels.

3.2 Categorical Variables

- **day_type**: Includes Work Day, Holiday, Event, Additional, and Transfer. Work Day is the most frequent, with Holidays and Events potentially influencing sales.
- **family**: 33 unique product families, with PRODUCE and GROCERY being common. Perishable items (e.g., PRODUCE) are flagged by the perishable column.
- **store_type**: Types A, B, C, D, E, indicating different store formats or sizes.
- **city** and **state**: Multiple locations, with Guayaquil and Quito being prominent cities.

4 Key Visualizations and Insights

The visualizations generated in the EDA provide critical insights into the relationships between key variables and unit sales, shedding light on potential drivers of sales behavior. This section delves into the two primary visualizationsTemperature vs. Unit Sales for PerGHishable Products and Average Unit Sales by Oil Price Rangeand extrapolates additional insights by considering the dataset's structure and context. These findings are crucial for informing feature engineering, identifying influential factors, and guiding predictive modeling.

4.1 Temperature vs. Unit Sales for Perishable Products

A scatter plot with an ordinary least squares (OLS) trendline was created to examine the relationship between temperature and unit sales for perishable products (perishable == 1), as shown in the notebook using Plotly Express (px.scatter). The plot maps temperature (řC) on the x-axis and unit sales on the y-axis, focusing on perishable items such as PRODUCE, DAIRY, and MEATS, which are sensitive to storage conditions and consumer behavior.

- **Observation**: The scatter plot reveals a dispersed distribution of unit sales across a wide range of temperatures (from -9.7řC to 33.7řC). The OLS trendline is relatively flat, indicating a weak linear correlation between temperature and unit sales for perishable products. There are no pronounced clusters or patterns suggesting a strong dependency, though outliers with high unit sales appear at moderate temperatures (1525řC).
- **Insight**: The weak correlation suggests that temperature alone is not a dominant driver of sales for perishable products. This is surprising, as one might expect higher temperatures to reduce demand for perishables due to spoilage concerns or lower temperatures to increase demand for certain items (e.g., dairy in colder weather). Several factors could explain this:
  - **Regional Variation**: The dataset includes stores across multiple cities (e.g., Guayaquil, Quito) with diverse climates. Aggregating temperature data across regions may mask location-specific effects. For instance, coastal Guayaquil experiences warmer, more stable temperatures, while Quito, at a higher altitude, has cooler and more variable weather. Segmenting the analysis by city or state could reveal localized temperature effects.
  - **Seasonal Confounding**: Temperature is likely correlated with seasonality (e.g., warmer months in summer). The lack of a clear trend may indicate that seasonal purchasing patterns (e.g., holidays, back-to-school periods) overshadow temperature effects. Adding time-based features like month or day_type to the analysis could clarify this.
  - **Consumer Behavior**: Perishable product sales may be more influenced by factors like promotions (onpromotion), store traffic (transactions), or day types (e.g., Holidays). For example, PRODUCE sales might spike during festive seasons regardless of temperature.
- **Implications**: While temperature may not be a primary predictor, it could still contribute to sales in specific contexts. For instance, extreme temperatures (below 0řC or above

30řC) might affect store visits or logistics, indirectly impacting sales. Future analysis should:

  - Explore interactions between temperature and other variables, such as precipitation or day_type, using multivariate models.
  - Create temperature bins (e.g., cold, moderate, hot) to test non-linear effects.
  - Investigate specific perishable product families (e.g., DAIRY vs. PRODUCE) to identify category-specific trends.

4.2 Average Unit Sales by Oil Price Range

The second visualization is a bar plot showing the average unit sales across five oil price ranges (Low, Medium Low, Medium, Medium High, High), created by binning the oil_price column and grouping by these bins (pd.cut and df.groupby). The plot, generated using Plotly Express (px.bar), displays oil price ranges on the x-axis and average unit sales on the y-axis, providing insight into the economic impact of oil prices on consumer purchasing behavior.

- **Observation**: The bar plot shows that average unit sales vary slightly across oil price ranges, with values hovering around 500600 units. The Medium and Medium High ranges (approximately 5080 USD per barrel, based on the dataset's oil price range of 26.19114.90) exhibit marginally higher average sales, while the Low and High ranges show slightly lower sales. However, the differences are not dramatic, and there is no clear monotonic trend (e.g., increasing or decreasing sales with rising oil prices).
- **Insight**: The subtle variation in sales across oil price ranges suggests that oil prices, as an economic indicator, have a limited direct impact on store sales in this dataset. This finding can be interpreted in the context of the dataset's geographic and economic setting (Ecuador, where oil is a significant economic factor):
  - **Economic Context**: Ecuador's economy is oil-dependent, and fluctuations in oil prices can affect consumer purchasing power, transportation costs, and retail prices. The lack of a strong trend may indicate that these effects are diluted across the diverse product families and store types in the dataset. For example, essential goods (e.g., GROCERY, CLEANING) may be less sensitive to oil price changes than discretionary items (e.g., HOME APPLIANCES).
  - **Lagged Effects**: Oil price changes may influence sales with a delay, as economic impacts (e.g., changes in disposable income) take time to manifest. The current analysis uses daily oil prices, but aggregating prices over weeks or months could reveal stronger relationships.
  - **Confounding Factors**: Other variables, such as onpromotion, day_type, or store_type, may have a stronger influence on sales, overshadowing oil price effects. For instance, promotional campaigns or holidays could drive sales spikes regardless of oil price levels.

- **Implications**: While oil prices may not be a primary driver, they could still contribute to sales forecasting models as an exogenous variable, especially in combination with other economic or temporal factors. Recommended next steps include:
  - Analyze oil price effects by product family or store_type to identify specific categories or stores most affected by oil price fluctuations.
  - Incorporate lagged oil price features (e.g., 7-day or 30-day moving averages) to capture delayed economic impacts.
  - Use regression models to quantify the contribution of oil prices relative to other factors like promotions or events.

4.3 Additional Visualization-Derived Insights

Beyond the two explicit visualizations, the EDA's data exploration (e.g., sampling day_type, precipitation, and examining temporal distributions) suggests additional visualization opportunities that could yield further insights:

- **Day Type and Sales**: The day_type column includes categories like Work Day, Holiday, Event, Additional, and Transfer. Sampling revealed that Work Day is the most common, but Holidays and Events appear less frequently. A box plot or violin plot of unit sales by day_type could reveal whether special days drive significant sales increases. For example, Holidays might boost sales for BEVERAGES or GROCERY due to celebrations, while Events (e.g., festivals) could increase traffic in specific cities.
- **Insight Potential**: Visualizing sales distributions by day_type could quantify the impact of special days, informing inventory planning and marketing strategies. For instance, stores could stock more perishable goods before Holidays to meet demand.
- **Precipitation and Transactions**: The precipitation column shows high variability (0305.1 mm). A scatter plot of precipitation vs. transactions could test whether heavy rainfall reduces store visits, as consumers may avoid shopping in bad weather.
- **Insight Potential**: If precipitation significantly reduces transactions, stores in high-precipitation areas (e.g., coastal cities like Guayaquil) might benefit from online delivery services or weather-targeted promotions to maintain sales.
- **Promotions by Product Family**: Only 3.7% of records involve promotions (onpromotion mean = 0.037). A bar plot comparing average unit sales for promoted vs. non-promoted items by family could highlight which categories (e.g., CLEANING, BEVERAGES) respond most to discounts.
- **Insight Potential**: Identifying promotion-sensitive product families could guide marketing strategies, optimizing promotional budgets for maximum sales impact.

4.4 Synthesis of Visualization Insights

The visualizations and exploratory analyses suggest that while temperature and oil prices have limited direct effects on unit sales, other factorssuch as promotions, day types, and potentially precipitationmay play significant roles. The weak temperature-sales relationship for perishable products highlights the need for segmented analysis (e.g., by city or product family), while the subtle oil price effects suggest incorporating lagged or aggregated price features. Additional visualizations targeting day types, precipitation, and promotions could further refine these insights, providing actionable recommendations for inventory, marketing, and forecasting.

5 Additional Insights

- **Promotions**: The onpromotion column indicates whether products were on promotion. Only 3.7% of records (mean = 0.037) involve promotions, suggesting a potential for analyzing promotion effectiveness.
- **Event Impact**: The day_type and Event Scale columns highlight special days (e.g., Holidays, Events). These could drive sales spikes, particularly for specific product families like GROCERY or BEVERAGES.
- **Geographic Variation**: Sales and transactions vary by city and store_type, indicating potential regional preferences or store-specific strategies.
- **Weather Effects**: Precipitation shows high variability (0 to 305.1 mm). Heavy rainfall could reduce store visits, impacting transactions and sales.

6 Next Steps

Based on the EDA, the following steps are recommended:

1. **Feature Engineering**:
   - Create time-based features (e.g., day of week, month, year) to capture seasonality.
   - Aggregate sales by store, family, or city to analyze trends at different granularities.
   - Encode categorical variables (e.g., family, day_type) for modeling.
2. **Advanced Analysis**:
   - Investigate the impact of promotions on specific product families.
   - Analyze sales during Holidays and Events to quantify their effect.
   - Explore interactions between weather (temperature, precipitation) and geographic location.
3. **Modeling**:
   - Use time series models (e.g., ARIMA, Prophet) or machine learning models (e.g., XGBoost, LSTM) to forecast sales.
   - Incorporate external factors (oil price, weather, events) as exogenous variables.

7 Conclusion

This EDA provides a comprehensive understanding of the store sales dataset, highlighting key patterns and relationships. While temperature and oil prices show limited direct impact on sales, factors like promotions, day types, and geographic differences are likely to play significant roles. The absence of missing values and duplicates ensures a robust dataset for further analysis. The insights gained will guide feature engineering and modeling efforts to develop accurate sales forecasts.