Data Analytics

Global Super Store

Data Analytics

Global Super Store

Fabian Foo
January 2025

Table of Contents

Introduction

The primary objective of this project is to collect, analyze, and visualize business data to derive actionable insights that aid decision-making. This project is completed as part of the RNCP Level 6 certification program, which emphasizes skills in data analytics and machine learning. The project integrates data gathering, cleaning, database design, and machine learning to simulate real-world applications in business intelligence.

Businesses today face unprecedented challenges in a data-driven world, where the ability to extract value from raw information is key to competitive advantage. This project showcases the practical application of advanced data analytics methodologies, demonstrating their impact on operational efficiency, strategic decision-making, and market growth.

Business Use Case

Modern businesses face challenges such as understanding customer needs, optimizing inventory, and improving profitability. By leveraging advanced data analytics, this project focuses on the following use cases:

1. **Revenue Optimization:**
    - By analyzing historical sales data, the project identifies key revenue-generating regions and products, enabling businesses to allocate resources effectively.
2. **Customer Insights:**
    - Through segmentation and purchase pattern analysis, this project provides a clearer understanding of customer behavior, enabling targeted marketing and personalized experiences.
3. **Profitability Analysis:**
    - Highlighting underperforming categories and markets, the project recommends actionable strategies to improve overall profitability.
4. **Scalability:**
    - By building a scalable framework, the project ensures readiness for data growth and seamless integration of advanced machine learning capabilities in the future.

Goal

The overarching goal is to leverage data analytics to drive informed decision-making and operational excellence. Specific objectives include:

- Identifying top-performing regions and products to optimize sales strategies.
- Improving profitability in less lucrative markets by uncovering hidden opportunities.
- Developing a robust data pipeline that ensures accuracy, reliability, and efficiency in analysis.
- Generating predictive insights through machine learning to aid long-term strategic planning.

Plan

To achieve these objectives, a structured four-week plan was implemented:

1. **Week 1:**
    - Define the project scope and objectives.
    - Collect and review datasets, ensuring completeness and relevance.
    - Conduct initial exploratory analysis to identify patterns and trends.
2. **Week 2:**
    - Clean and preprocess the data, addressing inconsistencies and gaps.
    - Perform detailed exploratory data analysis (EDA) to derive insights and inform subsequent steps.
3. **Week 3:**
    - Design a comprehensive database schema and entity-relationship diagram (ERD).
    - Normalize the data to eliminate redundancy and ensure query efficiency.
4. **Week 4:**
    - Develop API endpoints to expose data for analysis and integration.
    - Create dynamic dashboards using Tableau for real-time insights.
    - Implement and evaluate machine learning models to predict sales trends.

Data and Data Sources

Data forms the backbone of this project, sourced from multiple reliable platforms:

- **Primary Dataset:**
    1. **Flat file -** GlobalSuperstore, derived from Tableau demo data, offers a rich dataset covering sales, customer demographics and geography, and product categories.
- **Supplementary Sources:**
    1. **Fixer.io API:**
        - Provides currency exchange rates to standardize monetary values across regions, ensuring uniformity in financial analysis.
    2. **Web Scraping:**
        - Weather data for France, Paris was scraped from timeanddate.com to evaluate its impact on seasonal sales trends.
    3. **AWS BigQuery:**
        - Via S3 and Athena. Facilitates scalable analysis of large datasets, enabling integration of diverse data sources.
    4. **MySQL :**
        - SQL Query for Analysis: To analyze all orders with customer details, the following SQL query was used:

```
-- Retrieve All Orders with Customer Details
SELECT
    o.OrderID,
    o.OrderDate,
    o.ShippingDate,
    c.CustomerName,
    c.Segment,
    cl.City,
    cl.State,
    cl.Country
FROM
    orders o
JOIN
    customers c ON o.CustomerID = c.CustomerID
JOIN
    customerlocations cl ON c.CustomerID = cl.CustomerID
ORDER BY
    o.OrderDate DESC;
```

        - Also used simple SQL queries to verify that the API was working precisely

```
SELECT * FROM customers WHERE CustomerID = 'AA-315';

SELECT * FROM customerlocations where CITY ='New York City';

SELECT * FROM orderdetails where OrderPriority ='High';
```

Data Collection

The data collection process involved consolidating information from various sources into a unified structure for analysis. Key steps included:

1. **Fixer.io Integration:**
    - Currency exchange data was retrieved to harmonize financial figures.
    - API rate limits were managed using batched requests and caching mechanisms.
2. **Web Scraping:**
    - Seasonal weather data was collected to explore potential correlations with sales trends.
    - Error handling and validation ensured reliability and accuracy in the scraped data.
3. **GlobalSuperstore Dataset:**
    - The dataset was imported and validated to ensure it was comprehensive and relevant for analysis.

Data Cleaning and Exploratory Data Analysis

Data Cleaning:

- Addressed missing data by:
    - Imputing numeric values with the mean.
    - Replacing categorical values with the mode.
- Outliers were identified and capped using z-score analysis to maintain dataset reliability.
- Duplicate records (10% of the dataset) were removed to enhance accuracy.
- Normalized textual data, ensuring consistency in product names and categories.

Exploratory Data Analysis (EDA): EDA uncovered key insights:

Dataset Overview:

- **Timeframe**: Orders range from January 1, 2011, to December 31, 2014.
- **Total Orders**: 51,290.
- **Categories**: Three main categories: Technology, Office Supplies, and Furniture.
- **Regions and Markets**: Global coverage with segmentation by `Region` and `Market`.

Financial Metrics:

- **Average Sales per Order**: $1,205.
- **Highest Sales**: $135,831 in a single transaction.
- **Profit**: Highly variable, with an average of $106 per order but outliers up to $22,214.

Customer Behavior:

- Most frequent segment: `Consumer` (26,518 orders, ~52% of total).
- Common priority level: `Medium` (29,424 orders).

Product Performance:

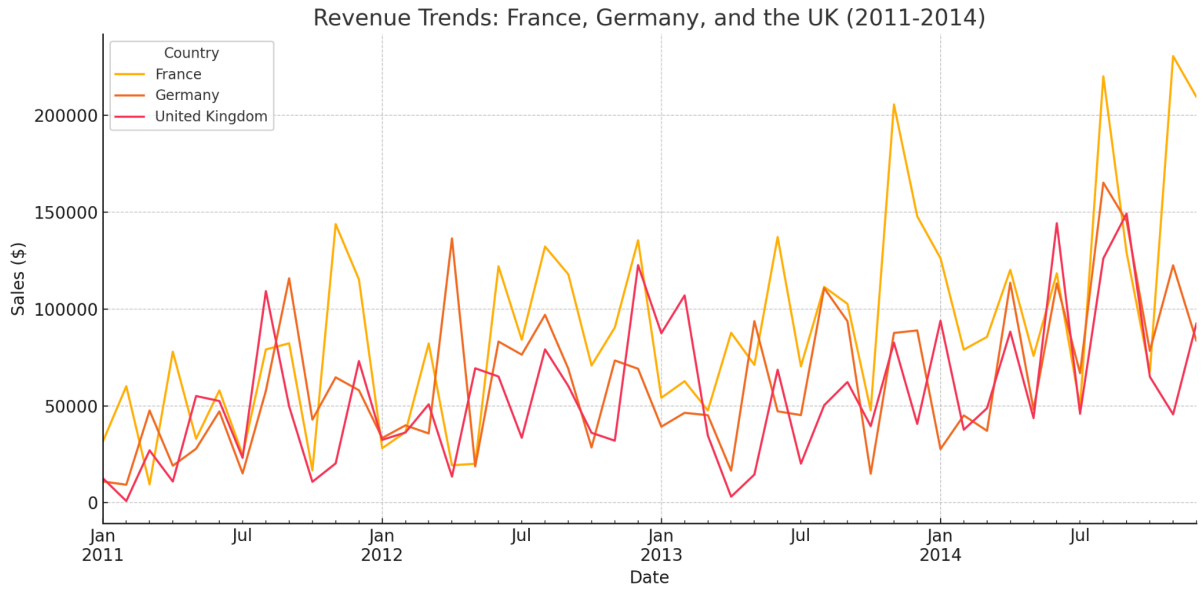
- Top Sub-Category: `Binders` (6,152 orders).
- High product diversity: 3,255 unique product names.

Operational Observations:

- Dominant Shipping Mode: `Standard Class` (60% of orders).
- Postal Code field is highly incomplete, with many missing values (80%).

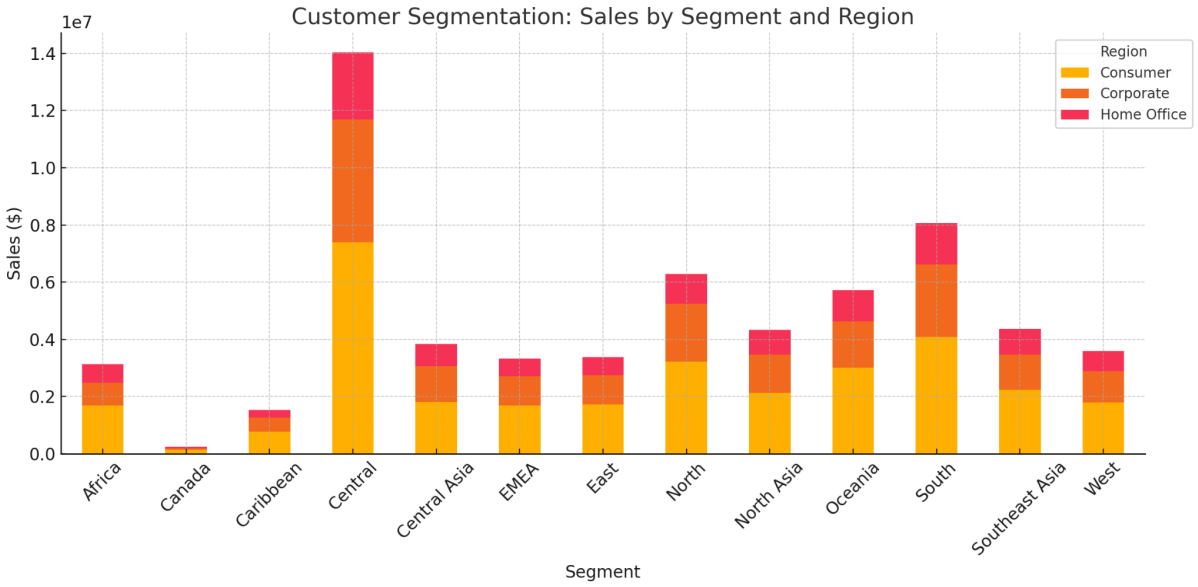
Revenue Trends:

- France, Germany, and the UK were identified as leading revenue contributors.



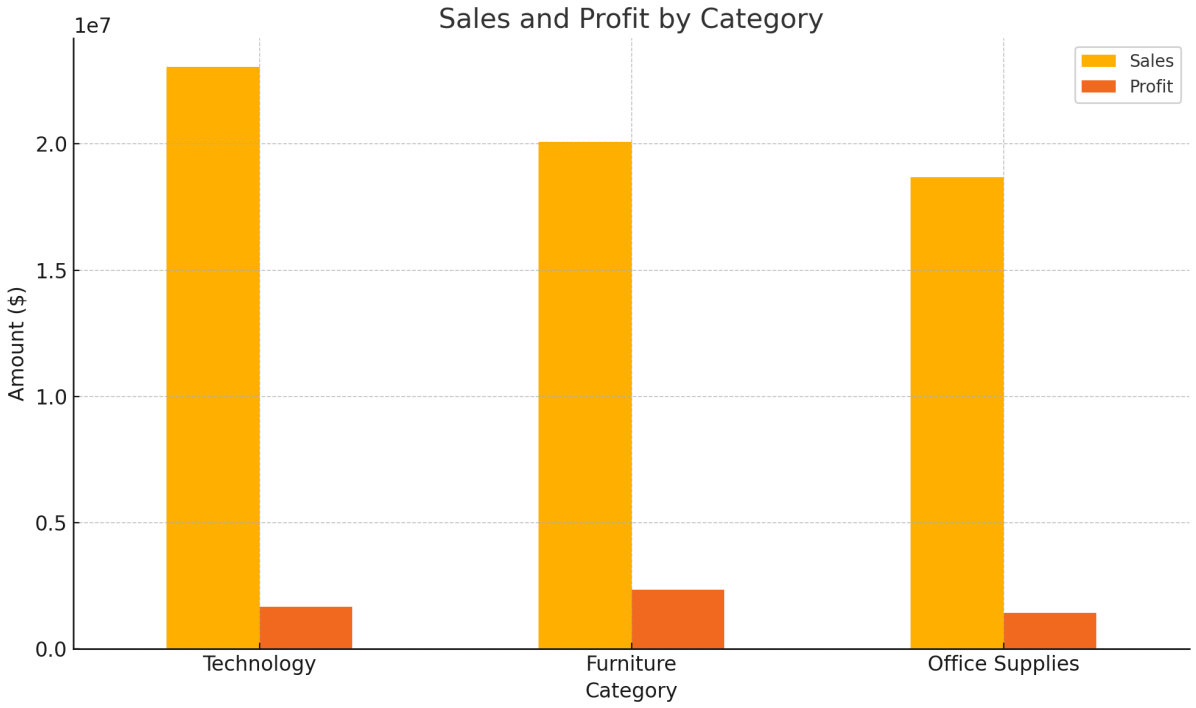
Customer Segmentation:

- Distinct buying patterns were observed across customer segments and regions.



Category Insights:

- Technology dominated sales volume, though profitability remained an area for improvement.



Database Type Selection

Relational databases were chosen as the optimal solution due to their:

1. **Structured Data Handling:**
    - Well-suited for managing tabular data with defined relationships.
2. **Querying Capabilities:**
    - Enable complex SQL queries for advanced analysis.
3. **Data Integrity:**
    - Ensures consistency and avoids redundancy through normalization.

The database design adhered to normalization principles, reducing storage redundancy and improving query performance.

Entities and ERD

The database schema was designed to support the project's analytical objectives. Key entities and their attributes include:

1. **Customers:**
    - Attributes: CustomerID, CustomerName, Segment, Region, Country.
    - Relationship: Linked to Orders (one-to-many).
2. **Orders:**
    - Attributes: OrderID, OrderDate, ShipDate, ShipMode, CustomerID, Profit.
    - Relationship: Linked to Customers (many-to-one) and Sales (one-to-many).
3. **Products:**
    - Attributes: ProductID, ProductName, Category, Sub-Category.
    - Relationship: Linked to Sales (one-to-many).
4. **Sales:**
    - Attributes: SalesID, OrderID, ProductID, Quantity, Discount, Sales.
    - Relationships: Linked to Orders (many-to-one) and Products (many-to-one).

API

The API provides seamless access to the processed data:

1. **Framework:** Flask.
2. **Endpoints:**
    - `/customers`: Retrieve customer information.
    - `/orders`: Fetch order details.
    - `/sales`: Access sales records.

3. **Security Measures:**
    - Token-based authentication protects sensitive data.
    - Input validation ensures integrity and prevents vulnerabilities.

Model Evaluation and Insights

Model Performance Summary

The Linear Regression model was developed to predict sales based on key features from the dataset, such as `Quantity`, `Discount`, `Profit`, `Shipping Cost`, `Category`, and `Sub-Category`. The model achieved the following performance metrics:

- Mean Squared Error (MSE): 5,769,928.63
- R-squared (R^2): 0.57

These metrics indicate that the model explains approximately 57% of the variance in sales, suggesting it captures some key relationships between the input features and sales. However, the high MSE shows that the predictions still have a significant deviation from actual sales values.

Interpretation of Results

- MSE Analysis: The MSE value reflects the average squared error in the model's predictions. While it is expected for the MSE to be high due to the scale of the sales data, this metric emphasizes the need for improvement in predictive accuracy.

- R^2 Analysis: An R^2 value of 0.57 suggests that the model captures a moderate amount of variability in the sales data. It also implies that 43% of the variability in sales remains unexplained, possibly due to missing features, randomness, or non-linear relationships not captured by the linear model.

Model Limitations

1. Feature Set:
    - The current feature set may not include all important predictors of sales. For instance, customer-specific information or external market conditions could add valuable context.
    - Interactions between variables, such as how discounts interact with product categories, are not explicitly modeled.
2. Model Choice:
    - Linear Regression assumes a linear relationship between predictors and the target variable. This assumption might not hold true for all features, limiting the model's ability to capture complex patterns in the data.

Recommendations for Improvement

1. Feature Engineering:

- Include additional features, such as customer demographics, seasonality, or regional trends.
- Transform existing features to capture interactions (e.g., Category × Discount) or non-linear relationships (e.g., log-transform skewed variables).

2. Advanced Models:
    - Explore more sophisticated algorithms like Random Forests, Gradient Boosting, or Neural Networks, which can handle non-linear relationships and interactions between variables more effectively.
3. Data Enrichment:
    - Integrate external data sources, such as macroeconomic indicators or competitor pricing, to provide additional context for sales predictions.
4. Evaluation Metrics:
    - Use additional metrics like Mean Absolute Error (MAE) or Median Absolute Error to complement the MSE and provide a more robust evaluation of model performance.

Conclusion

This Linear Regression model serves as a strong starting point for understanding the drivers of sales within the dataset. While the model achieves moderate explanatory power with an R^2 of 0.57, there is significant potential for improvement. Future iterations can focus on enriching the dataset, engineering new features, and experimenting with more advanced machine learning algorithms to enhance predictive accuracy.

The insights gained from this model can guide decision-making processes, particularly in understanding how factors like discounts, product categories, and shipping costs impact sales. However, further refinements are recommended to achieve more actionable and precise predictions.

GDPR

This project adheres to GDPR guidelines to ensure ethical data handling:

1. **Anonymization:**
    - Personal data was anonymized to protect customer identities.
2. **Access Controls:**
    - Role-based permissions restricted data access to authorized personnel.
3. **Consent Mechanisms:**
    - Explicit agreements were established for data collection and usage, ensuring compliance.

By aligning with GDPR, the project ensures that data privacy and security remain a top priority while facilitating impactful analysis.
