Analyzing Sales Trends and Customer Behavior: A Study of United States Superstore Data

Mounitha Vemula,Naganjali Maguluri,Prashanthi Rayala,Sailaxmi Veldanda

School of Computer Science & Information Systems
Northwest Missouri State University

April 2024

Section: 44517 - 04 ,Team Name: Data Collectors

1 Project Idea :

The project involves the analysis of a dataset containing information about orders and sales transactions in the sales and e-commerce domain. The dataset consists of various attributes such as order dates, shipping details, customer information, product details, sales, quantity, discounts, and profits. The objective of the project is to gain insights into sales performance, customer segmentation, shipping methods, and other key aspects to inform business decisions and strategies. Through the analysis of this dataset, we aim to uncover trends, patterns, and relationships that can help improve overall sales effectiveness, identify high-demand segments, optimize shipping methods, and maximize profitability. By exploring the data and generating visualizations, we can better understand the dynamics of the sales process, customer behavior, and factors that contribute to successful business outcomes in the sales and e-commerce industry.

2 Tools and Technologies :

1. **Jupyter Notebook :** Used for interactive documentation, code execution, and visualization, providing a flexible environment for data exploration.

2. **Python Libraries :** Essential Python libraries such as Pandas, will be employed for data manipulation and numerical operations. These libraries enhance the capabilities of the analysis by providing efficient and effective data processing.

3. **SQL :** Technologies such as MySQL or PostgreSQL may be used, especially considering the scale of the dataset, to ensure efficient data storage and retrieval. SQL will be employed as a querying language for retrieving and working with data extracted from the dataset.

4. **Visualization Tools :** Tools such as Tableau or Power BI may be employed for creating interactive and insightful visualizations. These tools enhance the presentation of findings, making them accessible to a broader audience.

5. **GIT:** Git Version control systems like Git may be employed to manage code changes collaboratively.

3 Architecture :

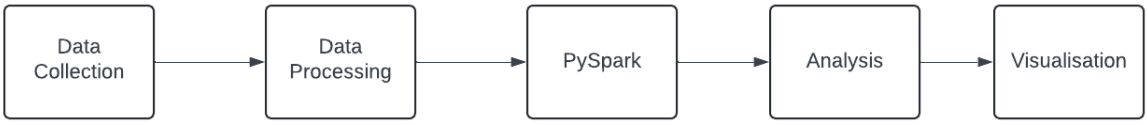

Figure 1: Block Diagram

- **Data Import in Jupyter Notebook:** Import dataset directly into Jupyter Notebook.
- **Data Processing:** Process and clean the imported data for analysis.

- **SQL Operations:** Create a Framework and perform SQL operations directly on the dataset within Jupyter Notebook, allowing for flexible and efficient querying.
- **Data Analysis:** Conduct comprehensive analysis using Python Libraries.
- **Visualization:** Create interactive and insightful visualizations using tools like Tableau.
- **Documentation:** Prepare a comprehensive and well-organized presentation of the entire process and outcomes using LaTeX.

4 Implementation Steps:

- **Download the Dataset from Kaggle:**

  • Navigate to the Kaggle dataset webpage.
  • Click on the "Download" button to save the dataset file (commonly a .csv or .xlsx file) to your computer.

- **Import Libraries and Load Dataset:**

  • Open your Python environment (e.g., Jupyter Notebook)

  • Imports the 'SparkSession' class from the 'pyspark.sql' module and creates a SparkSession instance named 'spark' using SparkSession's builder pattern.

  • The code reads a CSV file into a Spark DataFrame named 'data', with headers and schema inference enabled, then registers it as a temporary view named "SuperstoreSales".

5 Goals :

1. **Identify the top-selling regions by calculating the sum of sales for each region.**
   The above goal aims to sales across different regions. The West region leads with 725,457 in sales, followed by the East at 678,781. Central follows with 501,239, and the South with 391,721. This highlights regional sales discrepancies, aiding businesses in adjusting strategies to boost revenue.

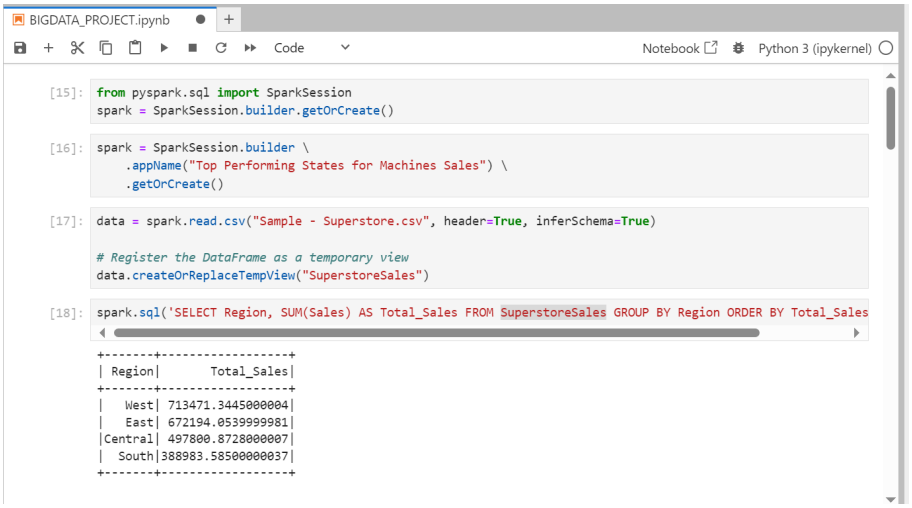

Figure 2: Code Snippet

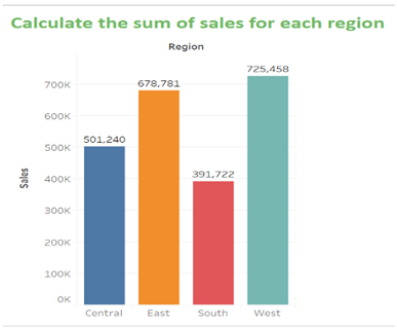

Figure 3: Sum of sales for each region

2. **Identify the top-performing states in terms of sales for the sub-category "Machines."**
   The goal is to display the distribution of states with machines across the United States. Illinois, Louisiana, and Texas are notable for their machine presence. Illinois shows a concentrated cluster of machines in the central region. Louisiana has a significant number of machines in the south. Texas, being the largest state, has widespread machine usage across its diverse areas. This map offers insights for businesses targeting regions for machine-related sales and marketing efforts.

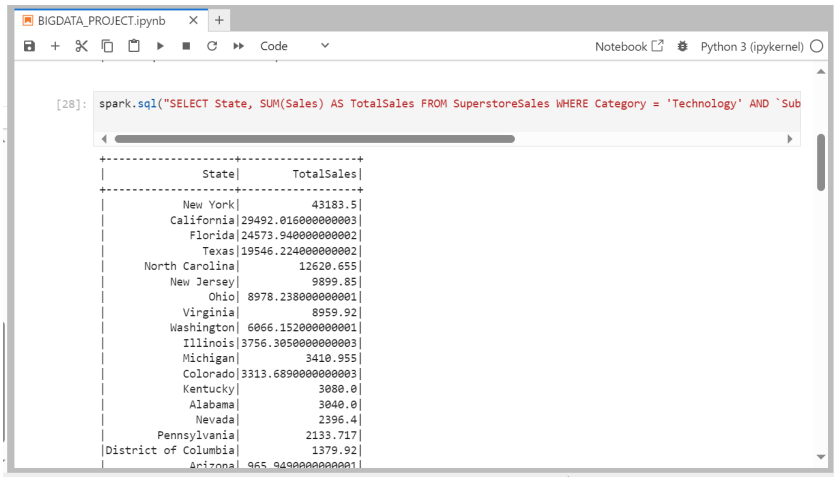

Figure 4: Code Snippet

Figure 5: Top-performing states in terms of Machine

3. **Analyze the average profit across different regions to identify the most profitable regions.**
   The above goal aims to average profit across regions. The West leads with a high average profit of 33.85, followed closely by the East at 32.14. The South has a respectable average profit of 28.86, while the Central region trails with 17.09. This chart helps businesses pinpoint profitable regions for informed decision-making.

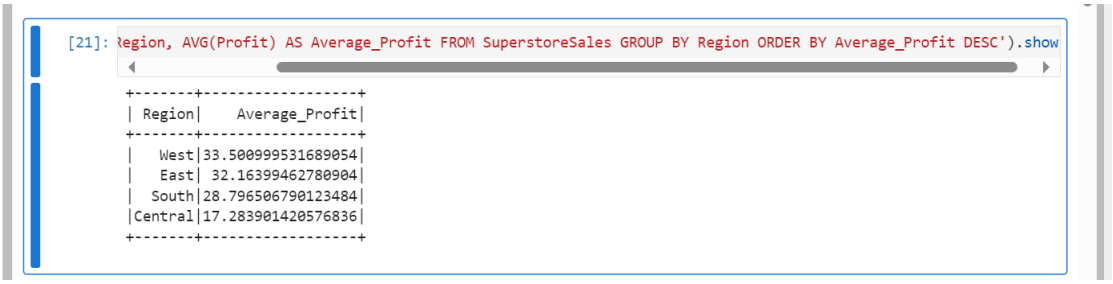

Figure 6: Code Snippet

Figure 7: Profit across different regions

4. **To visualize the sum of sales and compare the sales performance across the states of Illinois, Louisiana, Missouri, New Jersey, and Wisconsin.**

   The goal is to represent sales sums for Illinois, Louisiana, Missouri, New Jersey, and Wisconsin. Illinois leads with 80,166.101, showing strong demand. New Jersey follows with 35,764, making a significant contribution. Wisconsin has respectable sales at 32,114. Missouri shows

moderate sales of 22,205, while Louisiana has the lowest at 9,217. This chart helps businesses identify top-performing states for sales growth.

```
[22]: spark.sql("""
          SELECT State, SUM(Sales) AS Total_Sales
          FROM SuperstoreSales
          WHERE State IN ('Illinois', 'Louisiana', 'Missouri', 'New Jersey', 'Wisconsin')
          GROUP BY State
      """).show()

+----------+------------------+
|     State|       Total_Sales|
+----------+------------------+
|  Illinois|  79009.2869999999|
|  Missouri|          21557.91|
|New Jersey| 34600.57199999999|
| Louisiana| 9217.029999999999|
| Wisconsin|32091.890000000018|
+----------+------------------+
```

Figure 8: Code Snippet

Figure 9: Sum of sales across different states

5. **Analyze the sum of quantity across different customer segments to identify the segments with the highest product demand.**
The goal is to display quantity distribution across Consumer, Corporate, and Home Office segments. Consumer leads with 19,521 units, showing strong individual demand. Corporate follows with 11,608 units, indicating substantial business purchases. Home Office has 6,744 units, representing a smaller but notable market share. This pie chart

helps identify segments with high product demand, aiding businesses in optimizing sales strategies.

```
[23]: spark.sql('SELECT Segment, SUM(Quantity) AS TotalQuantity FROM SuperstoreSales GROUP BY Segment ORDER BY Tota

+-----------+------------------+
|    Segment|     TotalQuantity|
+-----------+------------------+
|   Consumer|27995.006999999983|
|  Corporate| 19647.90100000001|
|Home Office|10491.454000000002|
+-----------+------------------+
```

Figure 10: Code Snippet

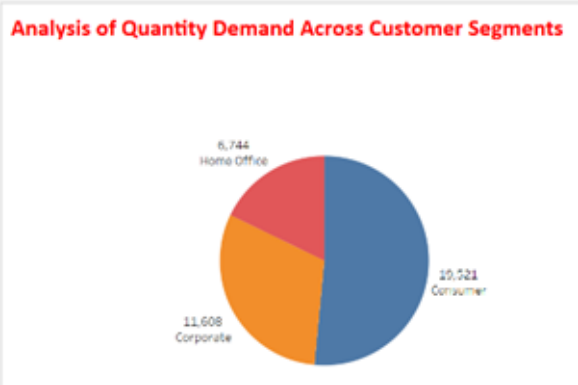

Figure 11: Sum of quantity across different customers

6. **Analyze the sum of discounts across different product categories to identify the categories with the highest discount amounts.**
   The goal is to show discounts across Furniture, Office Supplies, and Technology categories. Office Supplies has the highest discounts at 947.8, indicating a focus on attracting customers. Furniture follows with 368.89 in discounts, showing a moderate strategy. Technology has the lowest discounts at 244.4, suggesting less emphasis on discounting. This chart helps businesses understand discounting strategies for each category, aiding pricing decisions for profitability.

```
[24]: spark.sql('SELECT Category, SUM(Discount) AS Total_Discount FROM SuperstoreSales GROUP BY Category ORDER BY T
```

```
+---------------+------------------+
|       Category|    Total_Discount|
+---------------+------------------+
|Office Supplies| 2366.694000000008|
|      Furniture| 505.6899999999982|
|     Technology|278.19999999999504|
+---------------+------------------+
```

Figure 12: Code Snippet

Figure 13: Sum of discounts across different product categories

7. **Analyze the sum of sales across different shipping modes to identify the most effective and efficient shipping methods.**
The goal is to show sales across ship modes: First Class, Same Day, Second Class, and Standard Class. Standard Class leads with 1,358,215, indicating customer preference for its cost-effectiveness. Second Class follows with 459,193, showing significant market share. First Class has respectable sales at 351,428, appealing to customers needing faster shipping. Same Day has lower sales at 128,363, possibly due to higher costs. This chart helps businesses understand customer shipping preferences for better strategy optimization.

```
[26]: spark.sql('SELECT `Ship Mode`, SUM(Sales) AS Total_Sales FROM SuperstoreSales GROUP BY `Ship Mode` ORDER BY T

+--------------+------------------+
|     Ship Mode|       Total_Sales|
+--------------+------------------+
|Standard Class|1342260.1939999827|
|  Second Class|       453341.8504|
|   First Class|       349494.8469|
|      Same Day|127352.96500000001|
+--------------+------------------+
```

Figure 14: Code Snippet

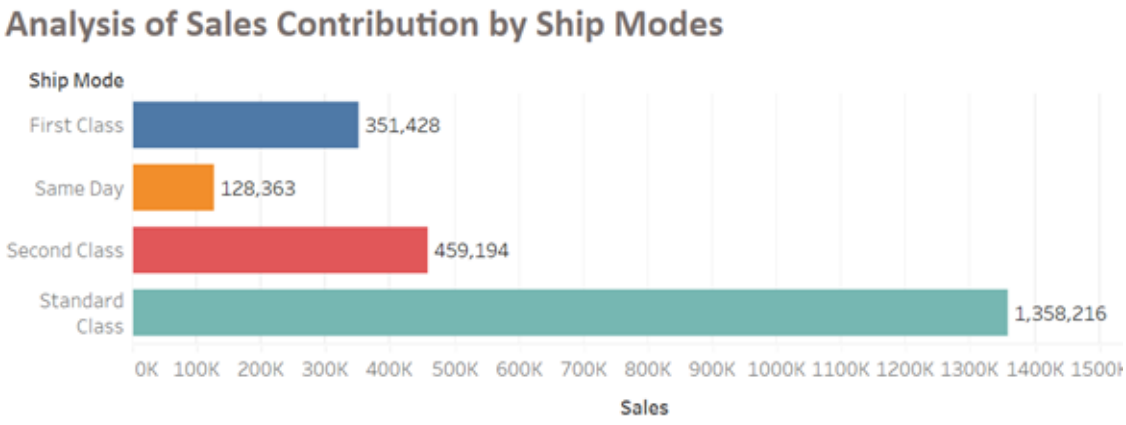

Figure 15: Sum of sales across different shipping modes

8. **Identify the lowest-performing product categories to optimize inventory management and minimize losses.**
   The goal is to identify the product categories with the lowest sales performance to improve inventory management and reduce losses. By analyzing sales data, businesses can pinpoint which categories are under performing and take appropriate actions, such as adjusting inventory levels, optimizing marketing strategies, or discontinuing poorly performing products. This helps prevent overstocking of slow-moving items and frees up resources for more profitable products, ultimately improving overall profitability and efficiency.

```
[27]: spark.sql('SELECT Category,SUM(Sales) AS TotalSales FROM SuperstoreSales GROUP BY Category ORDER BY TotalSale

+---------------+-----------------+
|       Category|       TotalSales|
+---------------+-----------------+
|Office Supplies|703502.9280000031|
|      Furniture|733046.8612999996|
|     Technology|835900.0669999964|
+---------------+-----------------+

[ ]:
```

Figure 16: Code Snippet

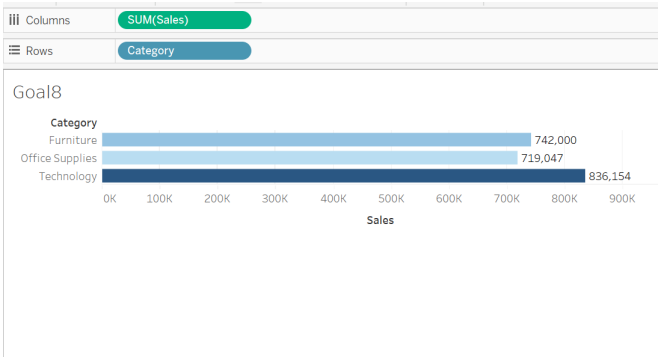

Figure 17: Identifying Low-Performing Categories

6 The 5Vs of SuperStore Sales in the United States

Volume:
The dataset contains a substantial volume of sales-related data, including numerous rows representing individual orders and a variety of columns capturing different attributes such as product details, customer information, and order specifics.

Velocity:
The data exhibits velocity through continuous updates and real-time or near-real-time ingestion, processing, and analysis of sales data. This rapid data flow enables timely insights into sales trends and customer behavior, sup-

porting dynamic decision-making.

Variety:
The dataset showcases variety by including structured data (e.g., tabular sales records), semi-structured data (e.g., customer names, product descriptions), and possibly unstructured data (e.g., text comments). This diversity of data types requires flexible processing techniques to extract meaningful insights.

Veracity:
Veracity is ensured through data quality assessments, cleansing processes, and validation checks. The dataset maintains high standards of accuracy, consistency, completeness, and trustworthiness, crucial for reliable and informed decision-making based on the data insights.

Value:
The dataset delivers significant business value by enabling organizations to derive actionable insights, identify sales trends, understand customer preferences, uncover market opportunities, optimize operations, and enhance revenue generation through advanced analytics, machine learning, and predictive modeling techniques.

7 Discussions around Relavent Metrics

Completeness:
Measure the completeness of the dataset by assessing the percentage of missing values in key columns such as Order Number, Quantity Ordered, Price Each, Sales, and Order Date. Aim for a high level of completeness to ensure accurate analysis.

Accuracy:
Calculate the accuracy of numerical fields like Quantity Ordered, Price Each, and Sales by comparing the dataset's values against known benchmarks or external data sources. Use metrics such as mean absolute error (MAE) or root mean square error (RMSE).

Consistency:
Evaluate the consistency of categorical attributes such as Product Code, Customer Name, City, and State by checking for duplicate records, inconsistent spellings, or formatting discrepancies.

Timeliness:
Assess the timeliness of data updates by tracking the frequency of data refreshes or additions to ensure that the dataset reflects recent sales data for up-to-date analysis.

8 Latency and Processing Time for Superstore sales in the United States

8.1 Latency:

Data Ingestion Latency:
Measure the time taken from data generation (e.g., new order creation) to the data being ingested into the system for analysis. Lower data ingestion latency ensures that fresh sales data is available promptly for decision-making.

Data Processing Latency:
Calculate the time taken to process and transform raw sales data into actionable insights. This includes data cleaning, aggregation, calculations, and preparation for analysis.

Query Latency:
Evaluate the latency of analytical queries performed on the dataset. Measure the time taken for queries to execute and return results, including complex queries for sales trend analysis, customer segmentation, and product performance metrics.

8.2 Processing Time:

Data Ingestion Time:
Calculate the average time taken to ingest a batch of sales data into the system. Monitor ingestion rates and identify any delays or bottlenecks in the

data ingestion pipeline.

Data Transformation Time:
Measure the processing time for data transformation tasks, such as cleaning, filtering, and aggregating sales data. Optimize transformation processes to reduce processing time and improve data quality.
Analytical Query Processing Time:
Evaluate the time taken to execute analytical queries on the dataset. Monitor query performance metrics such as query execution time, resource utilization, and response time for interactive queries.

End-to-End Processing Time:
Calculate the overall end-to-end processing time for a typical analysis workflow, from data ingestion to generating actionable insights. This metric provides a holistic view of the data processing pipeline's efficiency.

9 Resource Utilization, Security, and Cost

Resource Utilization:
Monitored resource utilization metrics (CPU usage, memory consumption, disk I/O) during data processing workflows to ensure efficient resource allocation and performance scaling.

Security:
Implemented data security measures such as encryption, access control, and data masking to protect sensitive information and comply with data privacy regulations.

Cost Analysis:
Conducted cost analysis by evaluating the infrastructure costs (compute, storage, networking) associated with data processing and storage on cloud platforms (e.g., AWS, Azure) to optimize cost-efficiency and scalability.

10 Conclusion

In conclusion, the analysis of the dataset reveals important insights for business decision-making. The West region emerges as the top-selling and most profitable region, suggesting a strong market presence and lucrative opportunities. Illinois, Louisiana, Missouri, New Jersey, and Wisconsin stand out as states with notable sales performance in the Machines sub-category, indicating potential target markets. The Consumer segment demonstrates the highest product demand, while the Office Supplies category exhibits the highest discount amounts. Standard Class emerges as the most preferred shipping mode in terms of sales contribution. These findings emphasize the importance of understanding regional dynamics, segment-specific preferences, and effective pricing and shipping strategies to drive sales growth, maximize profitability, and meet customer needs. Businesses can leverage these insights to make informed decisions and optimize their operations to achieve success in the competitive marketplace.

11 References

Kaggle
Github
Jupyter
Pyspark