ABSTRACT

This project delves into a comprehensive analysis of sales data, comprising of orders, shipments, customer profiles, and product details. Utilizing Spark SQL functions and data manipulation techniques, the analysis aims to extract actionable insights to aid strategic decision-making and drive business growth.

The primary objectives include understanding market dynamics, assessing customer behavior, evaluating product performance, and scrutinizing logistical efficiencies. Through meticulous data preparation, transformation, and analysis, this project unveils pivotal insights across various business segments.

The findings reveal distinctive market trends, regional variations in customer engagement, seasonal impacts on product profitability, and logistical intricacies affecting shipment efficiency. These insights serve as a foundation for strategic recommendations to enhance marketing strategies, optimize shipment modes, and capitalize on peak demand periods.

INTRODUCTION

In today's dynamic business landscape, leveraging data-driven insights is paramount for strategic decision-making and sustained growth. This project embarks on an in-depth exploration of sales data, aiming to extract meaningful insights from an integrated dataset comprising orders, shipments, customer profiles, and product information. Through the application of advanced analytics and Spark SQL functions, this analysis endeavors to unravel key patterns and trends critical to business success.We aim to achieve the following objectives:

- Market Understanding: The primary goal is to dissect market dynamics, discerning profitable segments, and identifying areas for market expansion or consolidation.

- Customer Behavior Analysis: Understanding customer preferences, regional variations in engagement, and purchase patterns to tailor marketing strategies and enhance customer experiences.

- Product Performance Assessment: Evaluating product-specific sales trends, profitability, and their lifecycle to optimize product offerings.

- Logistics and Shipment Efficiency: Scrutinizing logistical operations to identify inefficiencies, optimize shipment modes, and enhance overall operational effectiveness.

MOTIVATION OF OUR PROJECT

In today's hypercompetitive business environment, the significance of data-driven decision-making cannot be overstated. The motivation behind this project stems from a dire need to harness the power of data to drive informed strategies, enhance operational efficiencies, and seize untapped opportunities within the business landscape.

1. Market Intelligence and Strategy Formulation: Understanding market dynamics, consumer preferences, and emerging trends is pivotal in formulating competitive strategies. This analysis aims to provide valuable market insights critical for strategic planning and market positioning.

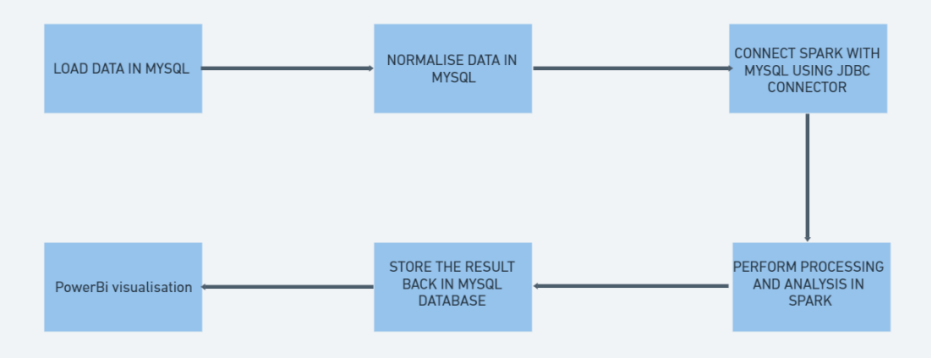
2. Customer-Centric Approaches: In an era where customer experience drives loyalty and retention, understanding customer behavior, preferences, and patterns becomes indispensable. This project seeks to uncover actionable customer insights to personalize offerings and improve overall satisfaction.

3. Operational Optimization: Efficient logistics and streamlined operations are foundational to business success. By scrutinizing shipment modes, assessing logistical efficiencies, and identifying bottlenecks, this analysis aims to optimize operations and reduce costs.

4. Data-Driven Decision-Making: The project's core motivation lies in advocating for a data-centric approach. By leveraging the wealth of information within the datasets, the goal is to empower stakeholders with actionable insights that translate into strategic actions and informed decisions.

In essence, this project is propelled by the ambition to transform raw data into actionable intelligence, fostering a culture where data becomes the guiding force behind every strategic move and operational enhancement within the business.

METHODOLOGY



The foundation of this project lies in the exploration and analysis of intricate sales data, commencing with the initial normalization and structuring of raw information into three core tables: Customers, Products, and Order Shipments. Leveraging the robust capabilities of SQL, these tables were meticulously crafted to establish a structured relational database. The integration of MySQL with Spark using JDBC facilitated seamless data transfer, enabling a bridge between SQL-driven structures and the advanced analytical capabilities of Spark.

Once the data was assimilated into Spark, an extensive phase of preprocessing ensued. This critical stage involved data cleaning, manipulation, and transformation, ensuring data integrity and preparing a refined dataset for in-depth analysis. Spark's powerful functionalities were harnessed to navigate through the complexities of the dataset, preparing it for the subsequent analytical processes.

The crux of the analysis unfolded within Spark's environment, leveraging its capabilities to derive meaningful insights. Key aspects of market trends, customer behaviors, product performance, and logistical efficiencies were meticulously analyzed and dissected. By employing various Spark SQL functions and transformations, the dataset was scrutinized, unveiling profound patterns and trends ingrained within the sales data.

To enhance the interpretability and accessibility of these insights, Power BI was employed to craft visual representations. Visualizations served as a conduit to communicate complex findings in an easily comprehensible format. These visual insights provided stakeholders with a lucid depiction of the data's essence, fostering informed decision-making and strategic planning.

This project represents a holistic journey from data acquisition and normalization to rigorous analysis, encapsulating the synergy between SQL-driven structural efficiency, Spark's analytical prowess, and Power BI's visualization capabilities to distill raw data into actionable insights

DATASET

- Our Dataset is "**Orders and Shipments"** .
- Source : OneGlobe
- The Dataset consist of the data regarding order details , Product details, Customer details and Shipment details.
- Size of the Dataset
    - 30871 rows
    - 24 columns

Attributes

Order ID, Order Item ID,Order Year, Order Month , Order Day, Order Time, Order Quantity, Product Department, Product Category, Product Name, Product ID, Customer ID, Customer Acct Id, Customer Market, Customer Region, Customer Country, Shipment Year, Shipment Month, Shipment Day,Shipment Mode, Shipment Days – Scheduled, Gross Sales, Discount % , Profit

DATABASE STORAGE -MYSQL

MySQL, an open-source relational database management system (RDBMS), serves as the primary repository for storing and managing the extensive datasets utilized in this project.

Normalised Tables

- Products Table: Product Department, Product Category,Product Name, Product ID
- Customers Table: Customer ID, Customer Acct Id, Customer Market, Customer Region, Customer Country
- OrderShipments Table: Order ID, Order Item ID, Order Year, Order Month , Order Day, Order Time, Order Quantity, Shipment Year, Shipment Month, Shipment Day, Shipment Mode, Shipment Days – Scheduled, Gross Sales, Discount % , Profit

MYSQL-SPARK INTEGRATION USING JDBC

1. Objective of Integration

- Data Accessibility: Enable Spark to interact with MySQL databases for reading and writing data.
- Unified Analysis: Utilize Spark's analytical capabilities on MySQL-stored datasets.

2. JDBC Connector

- Usage of JDBC: JDBC acts as a bridge, allowing Spark to connect with MySQL databases.
- Driver Configuration: Integration involves specifying the MySQL JDBC driver within the Spark environment.

3. Reading Data from MySQL into Spark

- DataFrame Creation: Utilize Spark's DataFrame API to read MySQL tables into Spark DataFrames.
- JDBC Connection Properties: Define properties like URL, username, password, and other necessary configurations for establishing the connection.
- Reading Operations: Showcase examples of reading MySQL data into Spark DataFrames using JDBC.

4. Writing Data from Spark to MySQL

- DataFrame to Table: Demonstrate how Spark DataFrames can be written back to MySQL tables.
- Write Modes: Explain different write modes such as overwrite, append, or ignore for data writing operations.
- Parallelism and Optimization: Discuss methods for optimizing write operations to MySQL using Spark parallelism.

IMPLEMENTATION AND RESULTS

- Drop columns OrderYear, OrderMonth and OrderDay after creating OrderDate column from them.
- Drop columns ShipmentYear, ShipmentMonth and ShipmentDay after creating ShipmentDate column from them.

Analysis using Spark

1)Customer churn identification by region

- Objective:
    - Identify potential customer churn between the years 2021 and 2022 within different regions.
- Analysis Methodology:
    - Joined OrderandShipment with Customers based on CustomerAccountID.
    - Grouped data by CustomerRegion.
    - Calculated the count of unique customers for the years 2021 and 2022.
    - Derived the difference in customer count between the two years.
- Result

```
scala> uniqueCustomersByRegionYear.show()
+---------------+--------------------+--------------------+----------+
| CustomerRegion|UniqueCustomers_2021|UniqueCustomers_2022|Difference|
+---------------+--------------------+--------------------+----------+
| Western Europe|                 911|                 311|      -600|
|Central America|                 960|                 395|      -565|
|  South America|                 489|                 196|      -293|
|Northern Europe|                 353|                 126|      -227|
|Southern Europe|                 339|                 122|      -217|
|      Caribbean|                 268|                 105|      -163|
|   Central Asia|                   7|                  31|        24|
|         Canada|                  17|                  50|        33|
|   Eastern Asia|                 190|                 244|        54|
| Central Africa|                  28|                  82|        54|
|        Oceania|                 243|                 303|        60|
|    East Africa|                  17|                  80|        63|
|Southern Africa|                  20|                  88|        68|
|     South Asia|                 178|                 293|       115|
| Southeast Asia|                 246|                 364|       118|
|   North Africa|                  38|                 162|       124|
| Eastern Europe|                  49|                 201|       152|
|    West Africa|                  65|                 217|       152|
|  South of  USA|                  46|                 229|       183|
|      West Asia|                  82|                 335|       253|
+---------------+--------------------+--------------------+----------+
only showing top 20 rows
```

- Insights:
    - Churn Identification: Regions with a negative difference indicate a potential decrease in customers from 2021 to 2022.
    - High Churn Regions: Regions showing a considerable decline in unique customers might signify higher churn rates.

- Retention Strategies: Highlight regions where customer retention efforts might be needed to mitigate churn.
- Opportunities: Identify regions with an increased customer base, suggesting growth or successful retention strategies.

➢ Potential Actions:

- Retention Strategies: Implement targeted retention campaigns or loyalty programs in regions with significant churn.
- Customer Engagement: Focus on understanding reasons behind churn and enhance customer experience initiatives.
- Market Expansion: Utilize regions displaying customer growth as potential areas for increased marketing or service focus.

2) Customer Purchase Behavior Analysis

➢ Objective:

- Analyze the time gaps between consecutive purchases for individual customers.

➢ Analysis Methodology:

- Joined OrderandShipment with Customers based on CustomerAccountID.
- Defined a window partitioned by CustomerAccountID and ordered by OrderDate.
- Calculated the time gap in days between consecutive purchases (OrderDate and the subsequent NextPurchaseDate).
- Filtered for non-zero time gap days, indicating subsequent purchases.

➢ Result

```
scala> ordersWithTimeGap.show()
+------------------+----------+----------------+-----------+
|  CustomerAccountID| OrderDate|NextPurchaseDate|TimeGapDays|
+------------------+----------+----------------+-----------+
|  100-1-Puerto Rico|2021-08-08|      2022-09-24|        412|
|         1000-1-USA|2022-11-29|      2023-07-06|        219|
|10003-1-Puerto Rico|2021-05-06|      2021-06-19|         44|
|10003-2-Puerto Rico|2022-01-15|      2022-03-08|         52|
|10003-2-Puerto Rico|2022-03-08|      2023-05-15|        433|
|        10007-1-USA|2021-06-14|      2022-08-31|        443|
|        10012-1-USA|2021-04-18|      2022-01-30|        287|
|        10012-1-USA|2022-01-30|      2022-04-21|         81|
|        10012-1-USA|2022-04-21|      2023-03-03|        316|
|        10017-1-USA|2022-05-24|      2023-02-27|        279|
|        10022-1-USA|2021-04-09|      2023-01-26|        657|
|        10022-1-USA|2023-01-26|      2023-05-14|        108|
|        10023-1-USA|2021-02-19|      2021-04-01|         41|
|        10023-1-USA|2021-04-01|      2022-04-15|        379|
|10038-1-Puerto Rico|2021-10-22|      2022-02-11|        112|
| 1004-1-Puerto Rico|2021-07-17|      2022-03-27|        253|
|        10040-1-USA|2022-12-16|      2022-12-20|          4|
|        10040-1-USA|2022-12-20|      2023-08-15|        238|
|10041-1-Puerto Rico|2021-01-30|      2021-03-02|         31|
|        10042-3-USA|2021-01-11|      2021-02-24|         44|
+------------------+----------+----------------+-----------+
only showing top 20 rows
```

- Insights:
    - Purchase Frequency: Understand the frequency of repeat purchases for each customer.
    - Customer Engagement: Identify customers with longer time gaps, potentially indicating decreased engagement or churn risks.
    - Seasonal Trends: Explore patterns in purchase intervals across different customer segments or regions.
    - Opportunities: Target marketing or engagement strategies based on purchase behavior to enhance customer retention.
- Potential Actions:
    - Segmentation Strategies: Segment customers based on purchase intervals for tailored marketing or loyalty programs.
    - Engagement Initiatives: Implement initiatives targeting customers with longer time gaps to re-engage them.

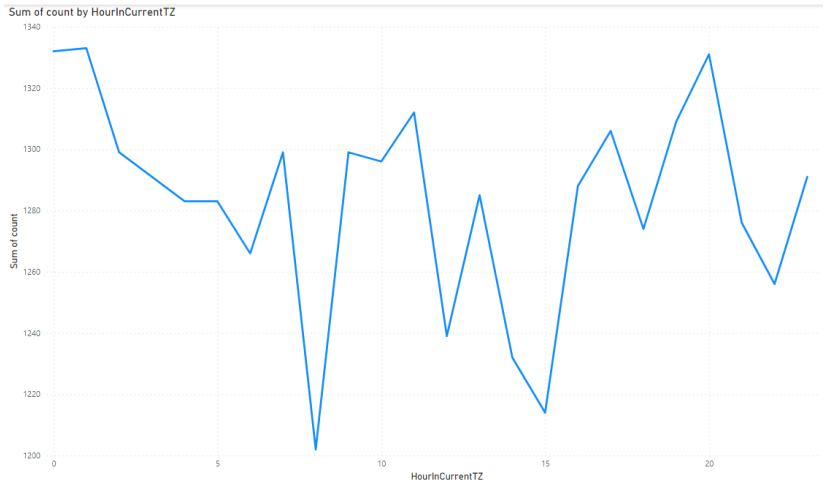
3)Peak Hours of Order Placements Analysis

- Objective:
    - Identify the peak hours during which most orders are placed in the current time zone.
- Analysis Methodology:
    - Created HourInCurrentTZ column by extracting the hour component from the OrderTime.
    - Grouped the data by HourInCurrentTZ.
    - Aggregated the count of OrderQuantity to determine the number of orders for each hour.
    - Ordered the results in descending order to identify the peak hours.

- Result

HourInCurrentTZ	count
1	1333
0	1332
20	1331
11	1312
19	1309
17	1306
9	1299
7	1299
2	1299
10	1296
3	1291
23	1291
16	1288
13	1285
5	1283
4	1283
21	1276
18	1274
6	1266
22	1256
12	1239
14	1232
15	1214
8	1202

24 rows in set (0.00 sec)



- Insights:
  - Order Density: Determine the hours with the highest order density, indicating peak activity periods.
  - Operational Efficiency: Understand resource allocation needs during peak hours for efficient order processing.
  - Customer Behavior: Identify preferred ordering times, aiding in targeted promotional or service initiatives.
- Potential Actions:
  - Resource Allocation: Allocate staff or resources according to peak order hours for timely order processing.
  - Marketing Strategies: Schedule promotional campaigns or offers during high-order density hours for maximum reach.
  - Optimized Operations: Streamline operations during peak hours to ensure efficient order fulfillment.

4) Profit Analysis by Year and Product Category

- Objective:
  - Analyze the total profit generated by different product categories across different years.
- Analysis Methodology:
  - Joined OrderandShipment with Products based on ProductID.

- Calculated the total profit for each ProductCategory grouped by Year.
- Categorized the total profit into different tiers (High, Medium, Low, Very Low) based on predefined profit thresholds.

➢ Result

```
scala> categorizedProfitsByYearAndCategory.show()
+----+--------------------+-----------+---------------+
|Year|     ProductCategory|TotalProfit| ProfitCategory|
+----+--------------------+-----------+---------------+
|2021|         Accessories|    7765.00|    High Profit|
|2021|                Baby|     112.00|Very Low Profit|
|2021| Baseball & Softball|    4168.00|  Medium Profit|
|2021|               Books|      30.00|Very Low Profit|
|2021|        Boxing & MMA|    1814.00|     Low Profit|
|2021|                 CDs|      36.00|Very Low Profit|
|2021|             Cameras|    3066.00|  Medium Profit|
|2021|    Camping & Hiking|  132150.00|    High Profit|
|2021|    Cardio Equipment|  192205.00|    High Profit|
|2021|              Cleats|  217413.00|    High Profit|
|2021|Consumer Electronics|     123.00|Very Low Profit|
|2021|              Crafts|    1617.00|     Low Profit|
|2021|                DVDs|     316.00|Very Low Profit|
|2021|         Electronics|   20047.00|    High Profit|
|2021|             Fishing|  207600.00|    High Profit|
|2021| Fitness Accessories|    1386.00|     Low Profit|
|2021|              Garden|     258.00|Very Low Profit|
|2021|      Girls' Apparel|    6882.00|    High Profit|
|2021|        Golf Apparel|     877.00|Very Low Profit|
|2021|          Golf Balls|    3760.00|  Medium Profit|
+----+--------------------+-----------+---------------+
only showing top 20 rows
```

➢ Insights:

- Yearly Performance: Identify profit trends for different product categories over the years.
- Category Contribution: Determine which product categories contribute the most to the overall profit.
- Profit Distribution: Visualize the distribution of profits across different profit tiers within each category.

➢ Potential Actions:

- Resource Allocation: Allocate resources or marketing efforts to categories yielding higher profits.
- Product Strategies: Identify categories with low profitability for potential improvements or strategic decisions.
- Financial Planning: Plan budgets or forecasts based on the categorized profit trends.

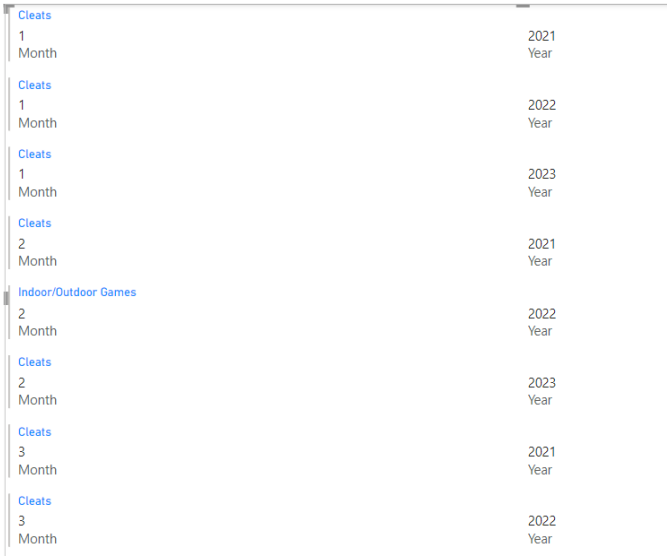
5)Highest Selling Category by Month and Year

➢ Objective:

- Determine the highest-selling product category for each month in different years.

- Analysis Methodology:
  - Joined OrderandShipment with Products based on ProductID.
  - Grouped the data by Year, Month, and ProductCategory, aggregating the total order quantity.
  - Utilized window functions to rank categories by total quantity sold within each month-year combination.
  - Filtered for the top-ranked category for each month and year.
- Results

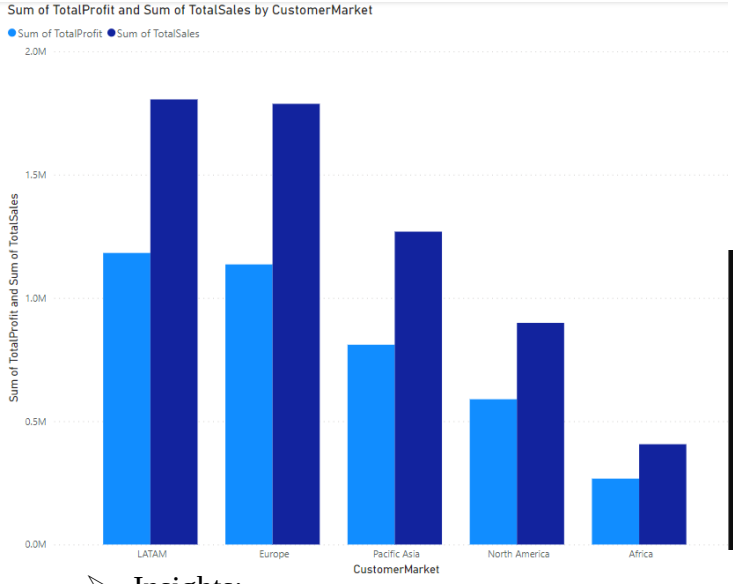
```
scala> rankedCategories.show()
+----+-----+-------------------+-------------+
|Year|Month|    ProductCategory|TotalQuantity|
+----+-----+-------------------+-------------+
|2021|    1|             Cleats|          456|
|2021|    2|             Cleats|          378|
|2021|    3|             Cleats|          442|
|2021|    4|             Cleats|          403|
|2021|    5|             Cleats|          330|
|2021|    6|    Women's Apparel|          373|
|2021|    7|             Cleats|          398|
|2021|    8|             Cleats|          374|
|2021|    9|Indoor/Outdoor Games|          349|
|2021|   10|    Women's Apparel|          333|
|2021|   11|             Cleats|          392|
|2021|   12|Indoor/Outdoor Games|          346|
|2022|    1|             Cleats|          378|
|2022|    2|Indoor/Outdoor Games|          314|
|2022|    3|             Cleats|          487|
|2022|    4|             Cleats|          400|
|2022|    5|             Cleats|          493|
|2022|    6|             Cleats|          423|
|2022|    7|             Cleats|          387|
|2022|    8|             Cleats|          478|
+----+-----+-------------------+-------------+
only showing top 20 rows
```



- Insights:
  - Seasonal Trends: Identify product categories that perform exceptionally well in specific months or seasons.
  - Yearly Highlights: Highlight categories that consistently rank as the highest-selling across different years.
  - Inventory Management: Aid inventory planning by understanding monthly variations in popular product categories.
- Potential Actions:
  - Marketing Strategies: Tailor marketing efforts or campaigns to capitalize on seasonal category trends.
  - Inventory Optimization: Ensure adequate stock for high-selling categories during peak months.
  - Product Development: Consider expanding or optimizing offerings within consistently top-selling categories.

6)Market Analysis: Total Sales and Profit by Customer Market

- Objective:
  - Analyze the total sales and profit across different customer markets.
- Analysis Methodology:
  - Joined OrderandShipment with Customers based on CustomerAccountID.
  - Grouped the data by CustomerMarket.
  - Calculated the total sales and total profit for each market.
- Results




```
mysql> SELECT * FROM MARKETANALYSIS ;
+----------------+------------+-------------+
| CustomerMarket | TotalSales | TotalProfit |
+----------------+------------+-------------+
| Europe         | 1787294.00 |  1135975.00 |
| Africa         |  406692.00 |   267161.00 |
| Pacific Asia   | 1268895.00 |   810212.00 |
| LATAM          | 1805157.00 |  1182464.00 |
| North America  |  898853.00 |   589099.00 |
+----------------+------------+-------------+
5 rows in set (0.00 sec)
```

- Insights:
  - Market Performance: Identify markets contributing the most to total sales and profit.
  - Profitability: Analyze markets with higher sales but comparatively lower profits for potential optimization.
  - Market Segmentation: Understand the distribution of sales and profits across different customer markets.
- Potential Actions:
  - Market Focus: Allocate resources or marketing strategies to high-performing markets for continued growth.
  - Profit Maximization: Investigate markets with high sales and low profits for potential pricing or cost optimizations.
  - Market Expansion: Explore opportunities in underperforming markets for strategic expansion initiatives.

7) Top Shipping Country Analysis by Day of the Week

- Objective:
  - Identify the top shipping country for each day of the week based on shipment counts.
- Analysis Methodology:
  - Joined OrderandShipment with Customers based on CustomerAccountID.
  - Extracted the day of the week from the OrderDate.
  - Grouped the data by OrderDayOfWeek and CustomerCountry.
  - Aggregated shipment counts for each country per day of the week and determined the top shipping country.
- Results

```
scala> topshippingcountryByDayOfWeek.show()
+--------------+------------------+------------+
|OrderDayOfWeek|TopShippingCountry|MaxShipments|
+--------------+------------------+------------+
|             1|               USA|         666|
|             2|               USA|         569|
|             3|               USA|         598|
|             4|               USA|         712|
|             5|               USA|         649|
|             6|               USA|         551|
|             7|               USA|         646|
+--------------+------------------+------------+
```

- Insights:
  - Shipping Patterns: Understand shipping preferences or trends for different countries on specific days.
  - Geographical Focus: Identify key markets or countries driving shipments across various days.
  - Operational Insights: Provide insights into logistical or operational needs concerning shipping across weekdays.
- Potential Actions:
  - Logistics Planning: Optimize logistics or shipping operations for high-demand days in specific countries.
  - Market Targeting: Tailor marketing or promotional strategies based on shipping trends in different countries.
  - Customer Experience: Enhance customer experience by optimizing shipping services or options based on shipping trends.

8)Top Country by Hourly Average Profit

- Objective:
  - Identify the country that generates the highest average profit for each hour of the day.
- Analysis Methodology:
  - Derived the OrderHour from the OrderTime.
  - Grouped the data by OrderHour and CustomerCountry.
  - Calculated the average profit per hour for each country.
  - Utilized window functions to rank countries based on their average profits for each hour.
- Results

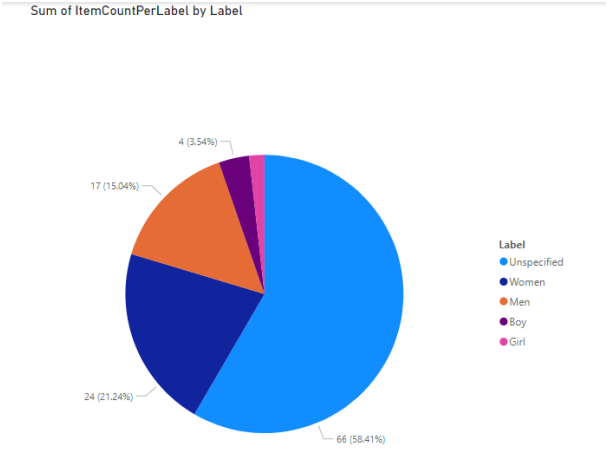
```
scala> topCountryByHour.show()
+---------+--------------------+----------+
|OrderHour|     CustomerCountry| AvgProfit|
+---------+--------------------+----------+
|        0|               Sudan|245.000000|
|        1|          Costa Rica|245.000000|
|        2|             Bolivia|245.000000|
|        3|                Togo|245.000000|
|        4|            Portugal|245.000000|
|        5|           Lithuania|245.000000|
|        6|            Thailand|245.000000|
|        7|               Egypt|206.166667|
|        8|           Hong Kong|258.000000|
|        9|               Qatar|245.000000|
|       10|Bosnia and Herzeg...|245.000000|
|       11|            Zimbabwe|245.000000|
|       12|              Gambia|200.000000|
|       13|            Thailand|178.750000|
|       14|              Angola|245.000000|
|       15|              Norway|245.000000|
|       16|          Uzbekistan|245.000000|
|       17|             Finland|245.000000|
|       18|               Syria|200.000000|
|       19|           Hong Kong|258.000000|
+---------+--------------------+----------+
only showing top 20 rows
```

- Insights:
  - Hourly Profit Trends: Understand which countries drive higher profits during specific hours.
  - Geographical Profitability: Identify countries with varying profitability trends throughout the day.
  - Operational Insights: Offer insights into potential time-specific market strategies or operational considerations.
- Potential Actions:
  - Hourly Strategies: Tailor marketing efforts or product promotions for specific countries during high-profit hours.

- Regional Optimization: Adjust resource allocation or operational strategies based on hourly geographical profitability.
- Customer Engagement: Improve customer engagement strategies during peak-profit hours for specific countries.

9) Product Categorization and Item Count

- Objective:
  - Categorize products based on their names into specified labels.
  - Count the number of items falling under each labeled category.
- Analysis Methodology:
  - Joined OrderandShipment with Products based on ProductID.
  - Created labels (Men, Women, Girl, Boy, or Unspecified) based on product name patterns.
  - Grouped the data by the labels created and counted the items per label.
- Result



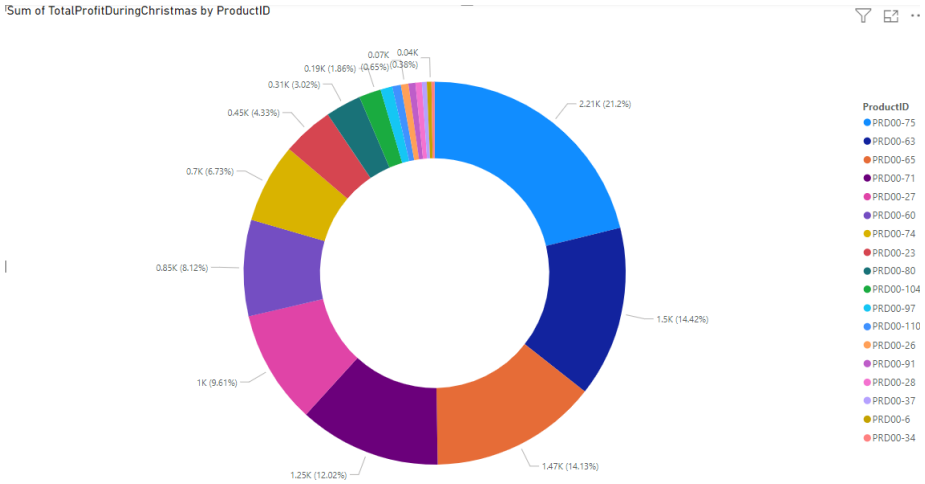

```
scala> itemCountPerLabel.show()
+-----------+-----------------+
|      Label|ItemCountPerLabel|
+-----------+-----------------+
|        Boy|                4|
|Unspecified|               66|
|        Men|               17|
|       Girl|                2|
|      Women|               24|
+-----------+-----------------+
```

- Insights:
  - Product Labeling: Understand how products align with predefined categories or labels.
  - Category Itemization: Determine the distribution of items across different labeled categories.
  - Inventory Overview: Offer insights into the composition of products based on label categories.

- Potential Actions:
    - Marketing Strategies: Tailor marketing approaches for specific product categories based on their labeling.
    - Inventory Management: Optimize inventory based on the distribution of items across labeled categories.
    - Customer Targeting: Develop targeted campaigns or strategies focusing on specific product categories.

10) Profit Analysis on Christmas Day (Day 359)

- Objective:
    - Analyze the profits generated by products on Christmas Day (Day 359).
- Analysis Methodology:
    - Joined OrderandShipment with Products based on ProductID.
    - Filtered the data for orders specifically on Day 359 (presumably Christmas Day).
    - Grouped the data by ProductID, ProductName, and YearOfOrder.
    - Calculated the total profit generated by each product on Christmas Day.
- Result



```
scala> profitByProductDuringChristmas.show()
+---------+--------------------+-----------+--------------------------+
|ProductID|         ProductName|YearOfOrder|TotalProfitDuringChristmas|
+---------+--------------------+-----------+--------------------------+
| PRD00-80|         Smart watch|       2023|                    314.00|
| PRD00-28|Fighting video games|       2023|                     57.00|
|  PRD00-6|  Adult dog supplies|       2023|                     40.00|
| PRD00-75|Perfect Fitness P...|       2022|                   2058.00|
| PRD00-65|Nike Men's Free 5...|       2022|                   1470.00|
| PRD00-63|Nike Men's Dri-FI...|       2022|                   1125.00|
| PRD00-71|O'Brien Men's Neo...|       2022|                   1125.00|
| PRD00-27|Field & Stream Sp...|       2022|                   1000.00|
| PRD00-60|Nike Men's CJ Eli...|       2022|                    845.00|
| PRD00-74|Pelican Sunstream...|       2022|                    600.00|
| PRD00-23|Diamondback Women...|       2022|                    450.00|
|PRD00-104|Under Armour Girl...|       2022|                    194.00|
| PRD00-97|Titleist Pro V1x ...|       2022|                    104.00|
|PRD00-110|Under Armour Wome...|       2022|                     75.00|
| PRD00-26|ENO Atlas Hammock...|       2022|                     68.00|
| PRD00-91|Team Golf Tenness...|       2022|                     60.00|
| PRD00-37|Glove It Women's ...|       2022|                     43.00|
| PRD00-34|Glove It Imperial...|       2022|                     28.00|
| PRD00-63|Nike Men's Dri-FI...|       2021|                    375.00|
| PRD00-75|Perfect Fitness P...|       2021|                    147.00|
+---------+--------------------+-----------+--------------------------+
only showing top 20 rows
```

- Insights:
    - Christmas Profitability: Understand which products performed best in terms of profitability specifically on Christmas Day.
    - Seasonal Product Performance: Identify products with higher profitability during festive seasons for potential seasonal marketing strategies.
    - Product-Specific Seasonal Trends: Analyze if certain products consistently perform better on Christmas Day across different years.
- Potential Actions:
    - Seasonal Campaigns: Focus marketing efforts on products that consistently perform well on Christmas Day.
    - Product Optimization: Optimize inventory or promotions for products that drive higher profits during festive seasons.
    - Customer Engagement: Tailor customer engagement strategies or offers around high-performing products during holidays.

CONCLUSION

The comprehensive analysis conducted on the integrated dataset comprising orders, shipments, customers, and products has yielded valuable insights across various dimensions of the business operations. Through a series of data transformations and analytical queries using Apache Spark, we have gained substantial insights that can drive strategic decisions and operational optimizations.

The integration of Spark SQL with MySQL facilitated seamless data processing and storage, enabling efficient analysis and retrieval of insights. Leveraging this integration, we conducted extensive data preprocessing, normalization, and diverse analytical queries, culminating in actionable insights for various business functions.