Executive Summary-

The goal of this project is to perform an in-depth analysis of data related to orders, operations, payments, customers, sellers, and products from a prominent eCommerce company in Brazil. This analysis aims to improve customer satisfaction, boost sales, and eliminate operational inefficiencies. The insights gained such as bottlenecks in the order approval process, top performing regions, customer lifetime values, top products and categories by season, etc., provide a deeper understanding of the company's customer base, operational dynamics, and product offerings. These insights will enable the company to develop data-driven business strategies and make informed decisions that will ultimately enhance overall business performance.

Introduction-

In the rapidly evolving eCommerce landscape of Brazil, staying competitive requires continuous improvement and adaptation. For this prominent eCommerce firm, the drive to enhance operational efficiency and customer satisfaction is paramount. This project was initiated to dissect various facets of the company's operations, from order processing and delivery systems to customer engagement strategies. By scrutinizing these areas through a data-driven lens, the company aims to identify bottlenecks and opportunities that could significantly impact its performance and customer perception.

The purpose of this comprehensive analysis is twofold:

1. **Operational Optimization**: To evaluate the efficiency of existing processes, particularly focusing on order fulfillment and delivery systems. This aspect of the analysis seeks to uncover inefficiencies in the supply chain and logistical operations that could be streamlined for faster and more reliable service delivery.
2. **Customer Experience Enhancement**: To deepen understanding of customer behaviors and preferences, enabling more personalized and engaging customer interactions. This involves segmenting the customer base, analyzing purchase patterns, and assessing the effectiveness of marketing and payment options.

By addressing these objectives, the analysis aims to furnish actionable insights that will not only refine operational tactics but also bolster customer relationships. The ultimate goal is to foster a more responsive, efficient, and customer-centric business model that aligns with the dynamic demands of Brazil's eCommerce sector.

Data Cleaning-

During the data cleaning phase, several key steps were taken to prepare the datasets for analysis:

Missing Value Detection and Treatment:

- **Products, Orders, and Payments Datasets**: These datasets contained missing values. In the products dataset, all columns except for product_id had missing values. Initially, I assessed the proportion of missing values in each variable. If a variable contained a disproportionately high number of missing values relative to the total observations, it was considered for removal due to insufficient data.
- **Products Dataset**: An observation with all values missing was identified and removed. For numeric product attributes like height, weight, length, and width, mean imputation was used, while mode imputation was applied to other categorical product variables to maintain consistency.
- **Orders Dataset**: For missing values in order_approved_at, order_delivered_carrier_date, and order_delivered_customer_date, I imputed these using calculated averages from observed timelines, such as the average approval time and the average time taken for carriers to pick up an order.
- **Payments Dataset**: This dataset didn't show missing values explicitly. However, entries under payment_type labeled as "not_defined" and with a payment_value of zero were considered invalid and subsequently removed.

Duplicate Detection:

- Duplicates were identified across datasets by checking for repeated entries in columns that serve as unique identifiers, such as order_id, order_item_id, product_id, and seller_id.
- The payments dataset was found to contain duplicates, characterized by entries with identical order_id, payment_type, payment_installments, and payment_value. These were presumed to be errors in the system where payments were processed multiple times accidentally.

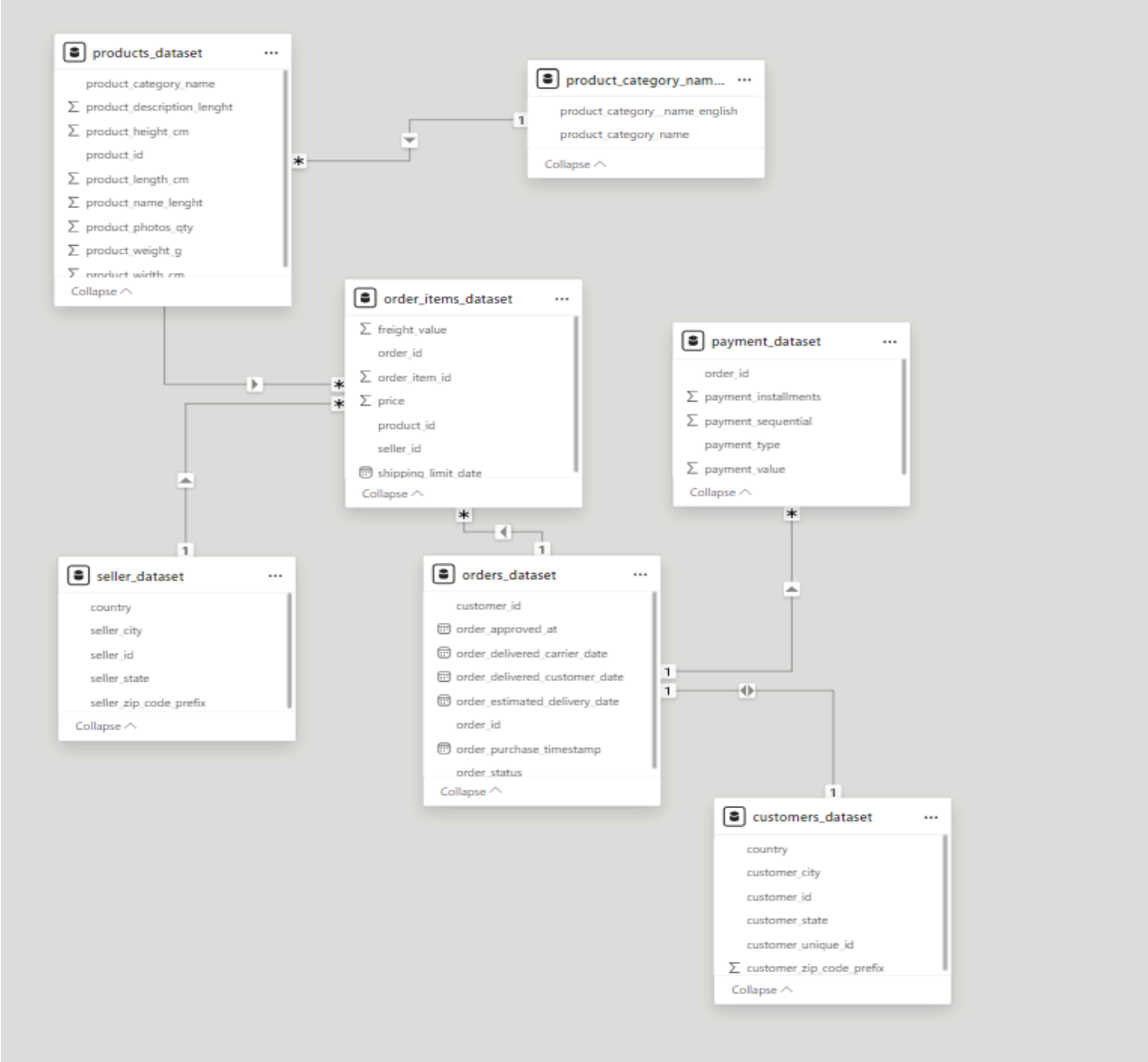
Data Consistency and Standardization:

- **Zip Code Standardization**: Zip codes were standardized to a five-digit format to address inconsistencies where some were recorded as four digits due to integer formatting.
- **City Names in Sellers and Customers Datasets**: Spaces between city names were standardized to underscores (e.g., "rio de janeiro" became "rio_de_janeiro") to ensure uniformity across datasets, as underscores were used elsewhere to represent spaces.
- Additional steps included stripping whitespace from string entries and converting text to lowercase across all datasets to ensure consistency and facilitate accurate data handling.

Data Integration-

Given the disparate datasets, integration was essential to create a unified data frame that would enable comprehensive insights and detailed analysis. The integration relied on key identifiers including customer_id, order_id, product_id, seller_id, and product_category_name. My objective was to compile a dataset centered around customer information to facilitate a customer-centric analysis.

Here is an overview of the data integration process:



Customer and Order Data Integration:

- A left join was performed between the customer dataset (as the left dataset) and the order dataset using customer_id. This step associated all relevant order information with individual customers, setting the stage for analyzing order patterns.

Incorporating Order Items:

- The combined customer-orders dataset was then joined with the order_items dataset using a left join, with the customer-orders dataset as the left. This integration brought in detailed order-related information.

Product Category Translation:

- A left join was used to merge the products and translations datasets, with product_category_name serving as the key. This was crucial for translating product categories from Portuguese to English, enhancing the clarity and usability of product-related data.

Adding Product Information:

- The product details were then merged into the customer-orders dataset using product_id in a left join. This action populated the dataset with comprehensive product information for each order.

Integrating Payment Information:

- Payment data was next integrated using order_id in a left join. This step enriched the dataset with payment details for each order, linking financial transactions to their respective orders.

Seller Information Integration:

- Finally, seller information was incorporated using a left join with seller_id. This enriched each product entry within an order with corresponding seller details.

Outlier Analysis

After assembling the final combined dataframe, an outlier analysis was conducted. This analysis focused on product attributes, sales, and cost data, employing the Interquartile Range Approach to identify potential anomalies. The analysis concluded that there were no significant outliers present in the dataframe.

Data Analysis & Insights-

Customer Segmentation:

- **Methodology:** Utilized Recency, Frequency, Monetary (RFM) segmentation to categorize customers into 'High Value', 'Medium Value', and 'Low Value' groups based on their interactions with the eCommerce platform.
- **Metrics Used:** Recency was measured using the `order_purchase_timestamp`, frequency was gauged by the count of transactions per `customer_id`, and monetary value was assessed using the `price` variable.
- **Findings:**
    - Revenue contributions, geographic distribution, and transaction completion times were analyzed for each segment.
    - Coastal regions predominantly housed Medium and High Value customers, whereas inland regions had a higher presence of Low Value customers.
    - Insights indicate a need for operational optimization in coastal regions to maintain retention of high lifetime value customers.

Sales Trends Over Time:

- **Observations:**
    - Sales peak during the tourist season (March-August) and dip towards September.
    - Sales increased significantly in 2017 compared to 2016 but saw a slowdown in 2018, indicating a need for enhanced marketing strategies to attract more customers.

Popular Product Categories:

- **Analysis Techniques:** Employed bar charts to analyze the number of orders and revenue by product category across different regions and time periods.
- **Key Insights:**
    - 'Bed, Bath & Table' was most popular by order count, while 'Health & Beauty' generated the highest revenue.
    - Notable seasonal trends included increased popularity of 'Watches & Gifts' in April and 'Computers' in September.
    - Recommended phasing out the 'Security & Services' category due to poor performance.

Top Performing Regions:

- **Geospatial Analysis:** Examined sales data across various Brazilian cities and states.

- **Findings:** Coastal areas, especially São Paulo, are top performers in terms of revenue and order volume. In contrast, inland and border-near regions showed the lowest performance metrics.

Descriptive Statistics:

- **Metrics Calculated:** Total revenue, shipping costs, customer count, and manufacturer/seller count on the platform.
- **Utility:** Provides a quick overview of key performance indicators for the firm.

Preferred Payment Modes:

- **Analysis:** Investigated the popularity and revenue generation of different payment methods including credit cards, debit cards, vouchers, and boleto.
- **Findings:** Credit cards dominate transactions, often processed in a single installment, suggesting potential areas for enhancing payment system usability and customer convenience.

Operational Delays:

- **Focus:** Evaluated logistical efficiency from order purchase to delivery, including order cancellation and non-delivery rates.
- **Critical Findings:**
    - São Paulo exhibited significant operational delays and higher cancellation/non-delivery rates compared to other states, pointing to major logistical challenges.
    - The average time taken to approve orders was approximately 10.5 hours, highlighting potential for process acceleration.

Recommendations-

1. Optimize Operations in Coastal Areas

- **Action Plan:** Prioritize operational efficiency in coastal regions, especially in São Paulo, where high-value customers are concentrated. Consider implementing faster processing and delivery systems to improve customer retention.
- **Implementation:** Invest in additional logistical support such as new distribution centers or partnerships with local couriers to expedite shipping times. Evaluate the potential for using advanced technology like automation and AI to streamline order processing.

2. Enhance Marketing Efforts During Off-Peak Seasons

- **Action Plan:** Develop targeted marketing campaigns to boost sales during the low season (post-August). Focus on promoting products that appeal to non-tourist demographics or launch new product lines that cater to seasonal needs.
- **Implementation:** Utilize data analytics to identify products with year-round appeal and increase advertising spend on digital platforms during slower months to maintain revenue flow.

3. Expand Product Offerings in High Revenue Categories

- **Action Plan:** Capitalize on the popularity of 'Health & Beauty' products which generate the highest revenue. Explore expanding this category and introducing new, innovative products.
- **Implementation:** Conduct market research to identify trending items within the health and beauty sector and forge partnerships with popular brands to enhance the product lineup.

4. Phase Out Low-Performance Categories

- **Action Plan:** Gradually reduce or eliminate product categories such as 'Security & Services' which show poor performance in terms of orders and revenue.
- **Implementation:** Analyze the inventory turnover rates and profit margins of these categories, and plan a phased withdrawal. Redirect resources to more profitable categories.

5. Improve Payment System Flexibility

- **Action Plan:** Since credit cards dominate transactions, enhance the payment system to offer more flexibility, such as multiple installment payment options or rewards for using preferred payment methods.
- **Implementation:** Work with payment processors to offer customizable payment plans and promotional offers that encourage larger purchases or frequent shopping.

6. Address Logistical Bottlenecks in São Paulo

- **Action Plan:** Specifically target the logistical issues in São Paulo that lead to high cancellation and non-delivery rates.
- **Implementation:** Investigate the root causes of these delays through detailed audits and customer feedback. Enhance training for local teams, upgrade warehouse management systems, and possibly restructure team deployments across the region to manage demand more efficiently.

7. Utilize Seasonal Trends to Inform Inventory and Marketing

- **Action Plan:** Leverage seasonal trends observed, such as the popularity of 'Watches & Gifts' during tourist seasons, to adjust inventory levels and marketing strategies.
- **Implementation:** Develop a dynamic inventory management system that adjusts stock levels based on predictive analytics from sales trends data. Create marketing campaigns that align with these trends to maximize exposure and sales potential during peak months.

8. Improve Customer Data Utilization

- **Action Plan:** Deepen customer relationship management by utilizing RFM and other segmentation techniques to personalize marketing and sales strategies.
- **Implementation:** Integrate a CRM system that leverages machine learning to predict customer behavior and preferences, enabling more targeted communications and promotions.