Enhancing E-Commerce Efficiency using Advanced SQL and Python

Team: Saurav Shakti Borah, Bhavan Sekar, Kavyasri Jadala, Liana Simopoulos, Amrutha Gabbita, Anurati Kulkarni

I.Background

The comprehensive analysis that was performed for The STEM Store, an e-commerce platform specializing in educational tools focused on Science, Technology, Engineering, and Mathematics (STEM), is aimed at addressing the operational challenges that have surfaced as the company expands into new cities. This initiative is crucial for maintaining and enhancing customer satisfaction and operational efficiency in the face of increasing competition within the e-commerce sector for educational tools. The analysis is multifaceted, incorporating several key components designed to refine the operational dynamics of the company.

The core of the analysis begins with generation of synthetic data (7 tables) using Python with the company's original data as a reference. Rigorous data preprocessing was performed to ensure the cleanliness and structured organization of the data for analysis. A Recency, Frequency, Monetary (RFM) analysis, performed using SQL aims at segmenting customers based on their purchasing behavior to identify and enhance the satisfaction of priority customers. This involves extracting a sizable sample of records from the database, calculating, and updating key metrics for each customer, and ultimately, segmenting customers effectively based on a composite RFM score. This segmentation enables targeted marketing strategies and personalized customer engagement, focusing on the most valuable customer segments.

Simultaneously, the analysis extends into optimizing vendor allocation within the logistics framework, a critical step towards improving delivery efficiency and cost-effectiveness. This process encompasses extensive data preparation, manipulation through temporary tables, and the application of sophisticated SQL procedures for vendor selection based on multiple criteria. The objective is to enhance operational efficiency by carefully selecting vendors who offer the best balance of cost, performance, and reliability.

Driver allocation is another pivotal aspect of the analysis, aimed at ensuring efficient and punctual delivery services. By introducing performance scores for drivers, adjusting vehicle capacities, and establishing a dynamic view for real-time order assignment adaptation, the process is meticulously designed to allocate orders to the most suitable drivers, thereby maintaining the integrity and punctuality of the delivery service.

Furthermore, the analysis includes a cost savings aspect, focusing on the financial impact of delayed orders and developing strategies to mitigate these delays. By analyzing delay patterns across geographical areas, calculating financial losses, and visualizing the impact of delays through

advanced analytics tools like Tableau, the analysis uncovers areas for logistical improvements and operational enhancements.

The strategic implementation of these analyses involves upgrading the database system with integrated SKU Prices and implementing dynamic delivery fees, tailored to the specific demands of each city. The expected outcome of this extensive analysis and strategic implementation is a significant boost in customer satisfaction, especially among identified priority customers, and a reduction in the Cost per Delivered Order (CPDO). By focusing on data-driven vendor allocation and optimizing driver utilization, The STEM Store aims to streamline its operations, thereby not only reducing operational costs but also strengthening its position in the competitive market, ensuring continued customer loyalty and operational excellence.

II.Data

The database with synthetic Python generated data described below constitutes a sophisticated framework aimed at monitoring and optimizing an e-commerce delivery system, focusing on the intricacies of order processing, product oversight, customer engagement, delivery mechanics, and vendor collaboration. The structure of this database contains several integral tables, each with a specific function within the broader operational schema.

The **OFD (Out For Delivery) Table** contains orders currently in the delivery phase. It encapsulates comprehensive details such as the dates of order placement and promised delivery, any instances of delay, the actual date of delivery commencement, modes of payment, and specifics of the ordered products. Furthermore, it ties each order to customer and delivery personnel information, thereby outlining the delivery logistics from the distribution hub to the destination.

The **Products** table is a catalog the entirety of the product inventory, detailing SKU codes, product descriptions, classifications, and potentially, pricing or stock levels. This table is essential for inventory management and facilitates an understanding of the available product range.

The **Orders** table serves as the cornerstone of the order management system, documenting each transaction with comprehensive details such as order numbers, dates, customer identifications, payment specifics, and associated product SKUs. This connects customer purchases with the inventory and fulfillment operations.

The **Vendor Rate Card** table is designed to manage information regarding vendors, including their pricing strategies, service tiers, and the categories of products they supply. This is vital for cost control and in making informed vendor selection decisions.

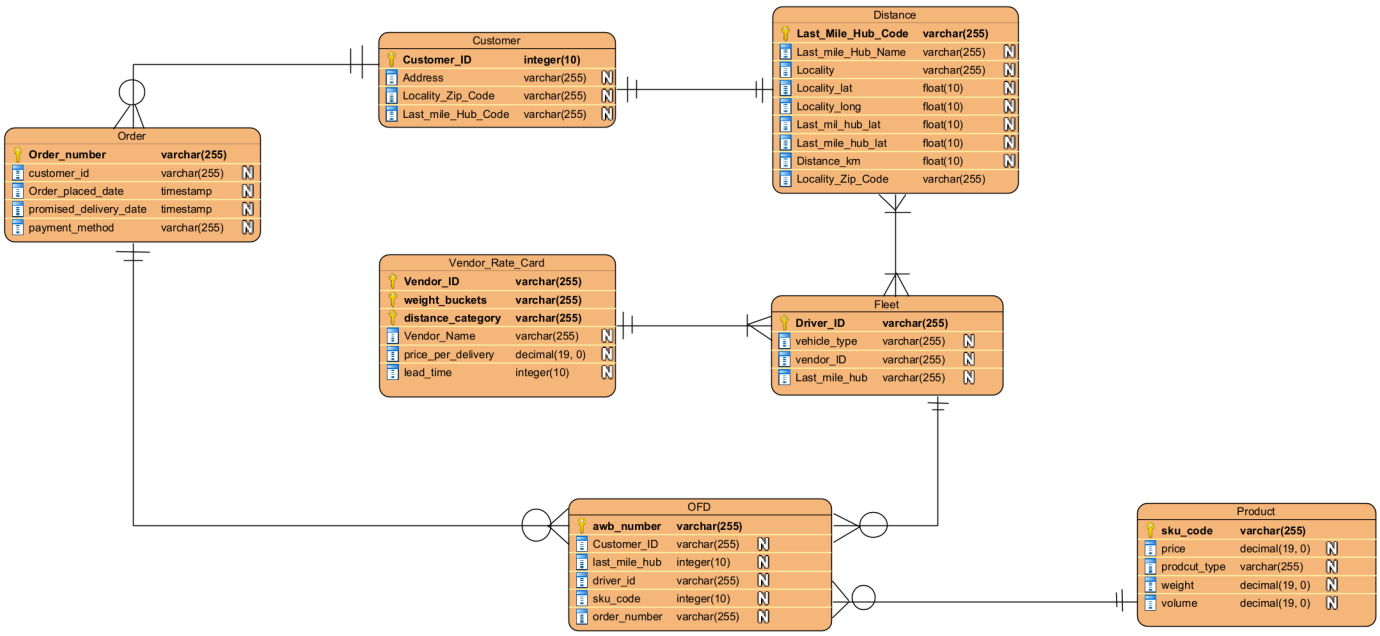
The table **New Localities Distance** likely contains data related to geographical logistics, such as delivery zones, the distances between distribution hubs, or customer locales. This supports logistical planning, enabling route optimization and the computation of delivery costs.

The **Customer** table is tailored to house profiles for each customer, featuring identification numbers, contact data, addresses, and possibly insights into order history or preferences. It's instrumental in fostering customer relationships and executing targeted marketing strategies.

Lastly, the **Fleet** Table details the delivery infrastructure, listing driver IDs, vehicle information, hub assignments, and metrics on delivery performance. This is crucial for the efficient management of the delivery fleet and enhancing delivery services.

Collectively, these tables create a comprehensive database designed to house the operational, analytical, and strategic requisites of an e-commerce delivery enterprise, facilitating the seamless transition of orders from placement through to delivery, the strategic management of inventory, engagement with vendors for procurement purposes, optimization of delivery routes, and the cultivation of customer relations.

III.Relational Data Model



IV. Analysis

A. RFM Analysis

The provided code executes a comprehensive RFM (Recency, Frequency, Monetary) analysis on customer data extracted from an e-commerce delivery system database. Initially, a random sample of 120,000 records is drawn from the 'ofd_final_new' table and stored in a temporary table named 'ofd_temp'. This step likely aims to reduce computational load while ensuring a representative subset of order data for subsequent analysis.

Following the sampling, the analysis proceeds by calculating three key metrics for each customer: recency, frequency, and monetary value. Recency is determined by identifying the last order date for each customer, frequency is measured by counting the number of orders placed, and monetary value is computed as the total amount spent by each customer. These metrics are then ranked accordingly, providing insights into customer behavior and purchasing patterns.

Subsequently, the customer table is updated with the computed recency, frequency, and monetary ranks. A new column named 'RFM_Metric' is added to the table, representing the composite RFM score for each customer. This score is derived by multiplying the individual recency, frequency, and monetary ranks together. Finally, the RFM metric is normalized to a scale between 0 and 1 for consistency, facilitating comparison across customers. The top 100 zipcodes with the lowest RFM metric are taken and used for the subsequent analysis described below. Ultimately, this RFM analysis enables the segmentation of customers based on their buying habits, providing valuable insights for targeted marketing strategies and customer relationship management within the e-commerce delivery system.

B. Vendor Allocation

The provided SQL script details a sophisticated procedure for vendor allocation within a logistics or delivery framework, emphasizing the systematic handling and analysis of diverse datasets. Initially, the script engages in data preparation by extracting information from several tables, including customer, distance, fleet, orders, products, and vendor rate cards. This foundational step is essential for amassing the requisite data for further analysis and decision-making processes involved in vendor allocation.

Subsequently, the script progresses to create and manipulate a series of temporary tables, such as ofd_temp, fleet, vendor_card_final, and ofd_temp_1. These tables are instrumental for various analytical purposes, including data randomization (ofd_temp), data transformation for enhanced analysis (vendor_card_final with delineated weight ranges and vendor ratings), and data preparation tailored to the final allocation strategy (ofd_temp_1 with calculated lead times and volume assessments).

At the heart of the script lies the vendor allocation logic, utilizing intricate SQL constructs like common table expressions (CTEs), window functions (e.g., DENSE_RANK), and conditional logic through CASE statements. This segment meticulously allocates vendors by carefully considering an array of criteria such as weight buckets, distance categories, and promised lead times. The allocation mechanism is adept at differentiating orders based on delay statuses—specifically segregating those without delays from those necessitating company or vendor reallocations—thereby applying distinct logic to optimize vendor selection based on cost-efficiency, lead times, and vendor performance metrics.

The establishment of the ofd_vendor_alloc_final table aggregates the outcomes of the vendor allocation process across all orders. This table categorizes orders into those unaffected by delays, those allocated by the company due to delays, and those requiring vendor reallocation, ultimately showcasing the allocation results. This sophisticated SQL script exemplifies a methodical approach to tackling the complexities of vendor allocation, striving for enhanced operational efficiency and performance in the logistics sector.

C. Driver Allocation

The driver allocation process involves a series of queries being run in a database. The initial query is adding a new field to the fleet table, where a performance score is assigned to each driver. These scores are not arbitrary; they are generated through a formula to ensure fairness, with each driver receiving a score between 4 and 10.

Subsequently, another query updates the database to reflect adjustments in the vehicle capacities, scaling down the original capacity to a third. This alteration could be a strategic decision to optimize load management and operational efficiency.

Following these updates, a new query rolls out, effectively creating a temporary table that records all orders. This includes a cumulative tally of order volumes, which is essential for monitoring the flow of goods. The query also logs the origin of each order and the preceding hub in the delivery chain, providing critical data for route planning and logistical coordination.

To streamline the allocation process, a query establishes a view within the database. This view acts as a live chart, presenting a snapshot of which drivers are assigned to which orders, based on their performance ratings and vehicle capacities. This dynamic overview is designed to adapt to real-time changes, ensuring that the allocation of orders remains as efficient as possible.

The concluding queries update the central allocation table, which is the definitive record of driver-order pairings. In the event of a delay, the system intervenes with a new query to reassign drivers, thus ensuring that the most suitable driver is tasked with the delivery. This reassignment is crucial to maintaining the integrity and punctuality of the delivery service.

Overall, these queries are the orchestrated commands that ensure the database not only remains current but also optimizes the delivery system's performance, assigning the best drivers and adjusting to changes and challenges in real-time.

D. Cost Savings Analysis

The code provided is part of a comprehensive analysis for a company aimed at understanding the financial impact of delayed orders and finding ways to improve operations.

The analysis starts by assuming that when an order is delayed, there's an average financial loss equal to 5% of the product's price. This figure is based on common market standards. From there, the code sets up a new framework for analysis, which is like a new lens to view the data. It looks at each geographical area and tallies the number of orders that were delayed, calculates what the delay might have cost the company on average, and compares the initial and final delivery costs to determine how much was saved on each order after addressing the delays.

Next, the code pulls all this information together to calculate the average costs and savings across all areas. This gives a bird's-eye view of how much delays are costing the company and how much is being saved by improving delivery times.

For a more detailed look at delays, Tableau was used to create a graph that shows the percentage of total spending that's going toward delayed orders for each area, and how long these delays are on average. Although not active in the code, there's a plan to pinpoint the products or categories that are most often delayed. This would help to identify specific items that may be problematic.

Finally, the code proposes looking more closely at the last stages of delivery or at specific hubs that have higher numbers of delays or longer wait times. This could uncover issues with how goods are routed or if certain hubs are over capacity.

The next graph created in tableau digs into the types of products that are getting delayed. It counts the number of delays and calculates the average length of delay for each product type, highlighting which types of products are most susceptible to delays.

In essence, this code is a structured approach to quantify the cost of delays in a business and to spotlight areas where logistical tweaks could lead to significant improvements and cost savings.

E. Capacity Reallocation

The analysis builds on the previous analysis on the dataset, focusing on the intricacies of driver allocations and the effectiveness of order deliveries. The initial query leverages a Common Table Expression (CTE) named `DriverOrders` to prepare data for a deeper analysis. This preparation

involves selecting information from the `ofd_final_alloc` table, which likely details orders and their corresponding driver allocations. It calculates two distinct ranks: one (`rn_orig`) for the original assignment of orders to drivers and another (`rn_new`) for any subsequent reassignments to new drivers. This ranking occurs within partitions that group the data by the date the order was placed, the driver ID, and the order number, all sorted by the order placement date. The objective here is to uniquely identify and rank each order based on its allocation to ensure accuracy in counting.

Following this setup, the main selection from the `DriverOrders` CTE counts the distinct orders delivered by both the original and new drivers for each day and driver. It accomplishes this by counting order numbers uniquely identified by their respective rank being 1, indicating the first occurrence of an order within its group. This method avoids the potential pitfall of double-counting in scenarios where multiple records per order exist. The outcome of this query offers insights into the daily workload managed by each driver, distinguishing between the orders initially assigned to them and those acquired through reassignment.

The subsequent query assesses the day-to-day changes in the number of active delivery drivers, contrasting the original allocations with any adjustments made. This assessment is conducted through two additional CTEs: `RankedDrivers` and `DriverCounts`. The `RankedDrivers` CTE assigns a dense rank to both the original and new driver IDs for each day, ensuring every driver receives a unique rank based on their ID for that day, irrespective of the number of orders they handle. The `DriverCounts` CTE then aggregates this ranked information to determine the maximum rank for each day, which correlates to the total count of drivers engaged in deliveries. This aggregation yields the daily tally of drivers before and after reassignments, reflecting operational decisions such as route reallocations or workforce adjustments in response to demand fluctuations.

In essence, these queries collectively offer a view into the delivery operations of the e-commerce system, highlighting the dynamics of order allocation among drivers and the operational shifts in driver engagement. By analyzing the efficiency of the delivery process and the adaptability of the driver workforce, these insights are invaluable for enhancing logistical operations, aiming to optimize delivery schedules, minimize costs, and elevate customer satisfaction through improved delivery services.

V.Conclusion

The analysis presented encompasses a comprehensive examination of various mechanisms of an e-commerce delivery system's operations, leveraging data to optimize delivery mechanisms, enhance customer segmentation, improve vendor and driver allocation, and analyze the financial implications of delivery delays. Through the execution of a detailed RFM analysis, the study effectively segments customers based on their interaction with the service, enabling targeted marketing strategies and personalized customer engagement. The analysis starts by sampling a

significant subset of data to manage computational demands while ensuring the representativeness of the dataset for accurate RFM metric computation.

The vendor allocation process, detailed through sophisticated SQL scripting, underscores the complexity of managing logistics within an e-commerce framework. By extracting and preparing data from multiple sources, the analysis facilitates a strategic approach to vendor selection, emphasizing efficiency and cost-effectiveness. This methodical allocation, rooted in criteria such as weight, distance, and delivery timelines, showcases a nuanced understanding of operational logistics aimed at optimizing delivery outcomes.

Driver allocation is determined through a series of queries aimed at assigning and reassessing drivers based on a performance scoring system. This process not only ensures the efficient distribution of orders but also adapts to real-time logistical challenges, highlighting the system's flexibility and responsiveness. By monitoring and adjusting vehicle capacities and driver assignments, the analysis contributes to a more streamlined and effective delivery process.

The cost savings analysis provides a critical evaluation of the financial impacts of delivery delays, offering insights into how operational improvements can mitigate these costs. By analyzing delays across geographical areas and correlating them with financial outcomes, the study identifies opportunities for logistical adjustments that can lead to significant savings and operational enhancements.

Furthermore, the capacity reallocation analysis, through the manipulation of data within the `DriverOrders` and `RankedDrivers` CTEs, offers a granular look at the day-to-day operations of delivery drivers. This analysis not only highlights the adaptability of the workforce to changing demands but also provides a framework for operational decision-making based on data-driven insights.

In conclusion, the comprehensive analysis conducted provides a multi-dimensional view of the operational challenges and opportunities within an e-commerce delivery system. By leveraging data to inform strategic decisions, the study outlines a path toward operational excellence, characterized by optimized delivery routes, improved customer segmentation, efficient resource allocation, and enhanced financial performance. Through meticulous data analysis and strategic implementation, the project sets a foundation for continued innovation and improvement in the competitive landscape of e-commerce logistics.