Project Report

Preprocessing and Exploratory Data Analysis

Prepared by:

Avantika Roy

Table of Contents

Project Summary

Objective

Identify popular product bundles and predict next-purchase products to help optimize marketing strategies, such as product bundling and targeted promotions. The project aims to increase average order value (AOV) by encouraging customers to buy bundles of frequently purchased products together.

Project Description

This project involves understanding associations among products that customers frequently purchase together, using market basket analysis and supervised modeling to identify patterns and predict a user's next likely purchases. By using association rules to discover product bundling opportunities and sequence-based supervised models to anticipate future purchases, this project provides insights that support effective marketing and personalization strategies.

Data Considered

- Orders Dataset
- Order Products Dataset
- Products Dataset
- Aisles Dataset

Data Description

products.csv

This dataset contains data on product ID, their names, the aisles they are stored in and the department they are located in. The **product_id** column is where numerical values are stored, which are unique to each product; no two products share the same ID.

The **product_name** column stores the names of the products that are available in the store. The **product_id** column corresponds to these item names. This feature by itself does not serve any purpose to the analysis but helps identify the type of product the customers of the store usually buy.

The **aisle_id** column stores unique numerical values for aisles. This also helps identify the aisle in which the items are stored.

department_id also stores unique numerical values that help identify in which department an aisle is present and thus which department an item is present in.

orders.csv

User Behavior:

- **User Loyalty:** The user_id feature can be used to identify repeat customers and analyze their purchasing patterns. Frequent orders indicate a high level of customer loyalty and satisfaction.
- **Order Frequency:** The days_since_prior_order feature provides insights into how often customers make purchases. By analyzing the distribution of this feature, we can identify trends in customer purchasing behavior and identify potential opportunities for targeted marketing campaigns.

Order Patterns:

- **Order Timing:** The order_dow and order_hour_of_day features provide insights into when customers are most likely to place orders. This information can be used to optimize staffing schedules, delivery operations, and marketing campaigns.
- **Order Volume:** The order_number feature indicates the number of orders placed by each user. Analyzing the distribution of order numbers can help identify high-value customers and potential opportunities for upselling and cross-selling.

order_products.csv

Product Information:

- **Product ID:** This unique identifier helps track individual products and their associated attributes.
- **Add-to-Cart Order:** This feature provides information about the order in which products were added to the cart. By analyzing this data, we can gain insights into customer purchasing

behavior and preferences. For example, frequently added products might be popular or essential items.

Customer Behavior:

- **Reordered:** This binary feature indicates whether a product was reordered by a customer. By analyzing reordered products, we can identify popular items, customer preferences, and potential opportunities for targeted marketing campaigns.

aisles.csv

Aisle Identification:

- **Aisle ID:** This numerical feature uniquely identifies each aisle within the store. It serves as a numerical representation of the aisle.
- **Aisle:** This categorical feature provides the textual description of the aisle, such as "Produce," "Dairy," or "Snacks." It offers a more human-readable representation of the aisle.

Preprocessing

products.csv

The dataset contains two features, **product_name** and **product_id**, with a maximum value of 49688. **product_name** can be ignored while **product_id** fulfills the same role in a more computationally efficient manner.

aisle_id has a total of 134 unique values, which means compared to the total number of data, it is fairly less. So, this can be treated as a categorical variable. This indicates that multiple items are stored within the same aisle, making it suitable for categorical grouping.

Similarly, **department_id** serves as a categorized feature. With fewer unique values than **aisle_id**, it reflects a hierarchy where each department encompasses multiple aisles, and, by extension, each aisle houses multiple items, making **department_id** useful for analyzing product distribution across departments.

There are no null values in any of the columns of this dataset, ensuring that the data is complete and thus making the process of preprocessing easier. The dataset also does not contain any duplicate values as well.

orders.csv

The **order_id** feature provides a unique ID for each transaction, even for repeat purchases of the same items.

The **user_id** feature assigns a unique ID to each user, with repeated IDs indicating that users order multiple times.

The order_number feature tracks the total orders per user, reflecting customer engagement and loyalty.

The order_dow and order_hour_of_day features capture the day and hour of each order, helping identify peak times for better staff and delivery planning.

Lastly, the days_since_prior_order feature shows the interval between purchases, which helps in predicting restocking needs and demand cycles.

order_products.csv

The order_id feature contains unique IDs for each order placed by users. The repeated IDs may link to the product_id feature, where repetitions suggest which products are included in each specific order.

The product_id feature is a numerical identifier, unique to each product in the dataset, helping to distinguish individual products.

The add_to_cart_order feature indicates the product's priority in the cart, showing the order in which each item was added by the user.

The reordered feature, a categorical variable, shows whether a product was reordered, offering insights into repeat purchases.

aisles.csv

The aisle_id feature provides unique numerical identifiers for each aisle, though it can be treated as categorical since multiple items can belong to the same aisle. Each value represents a distinct aisle, making it easier to categorize products by location.

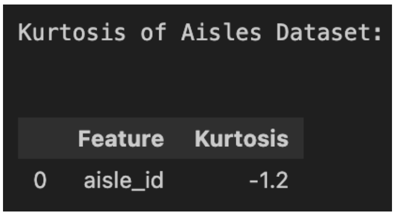
The aisle feature contains the actual names of these aisles, corresponding directly to their aisle_id values. Each aisle has a unique name, which helps in identifying and organizing product locations more intuitively.

Exploratory Data Analysis

Univariate Analysis

Kurtosis

Aisles:



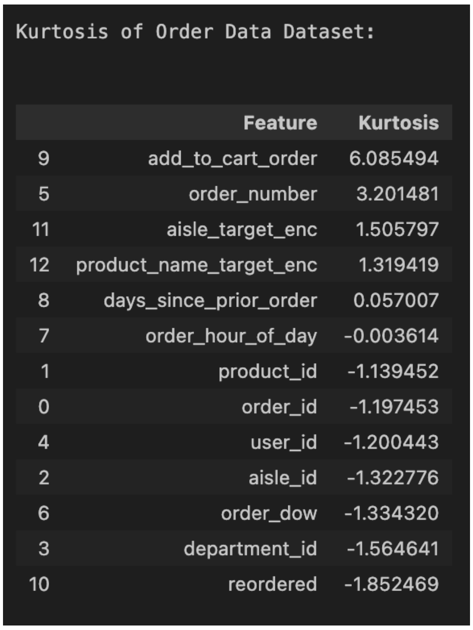
Kurtosis of Aisles Dataset:

	Feature	Kurtosis
0	aisle_id	-1.2

The **aisle_id** has a kurtosis value of -1.2, indicating a platykurtic distribution. This suggests that the data is flatter than a normal distribution, with fewer extreme values and a wider spread of data points. The distribution appears to be more evenly distributed with relatively fewer large deviations from the mean.

Merged:

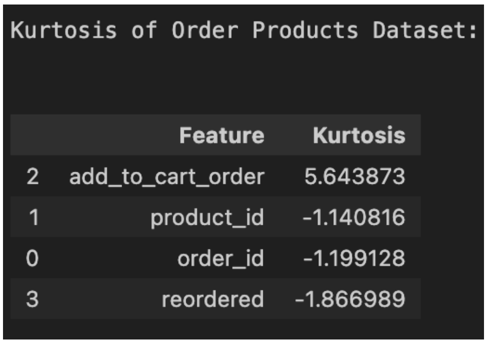
In the merged dataset, features such as **add_to_cart_order** and **order_number** exhibit high kurtosis, with **add_to_cart_order** showing a sharply peaked distribution and **order_number** showing a slightly peaked distribution. Features like **aisle_target_enc** and **product_name_target_enc** have moderate kurtosis, indicating a distribution closer to normal with light tails. The **days_since_prior_order** and **order_hour_of_day** features show very flat distributions, with kurtosis values close to zero. Lastly, **product_id**, **order_id**, **user_id**, **aisle_id**, **order_dow**, and **department_id** all exhibit negative kurtosis, indicating platykurtic distributions that are flatter with fewer extreme values. The **reordered** feature is likely flat, but its kurtosis value is missing, potentially due to it being a binary variable.



Kurtosis of Order Data Dataset:

	Feature	Kurtosis
9	add_to_cart_order	6.085494
5	order_number	3.201481
11	aisle_target_enc	1.505797
12	product_name_target_enc	1.319419
8	days_since_prior_order	0.057007
7	order_hour_of_day	-0.003614
1	product_id	-1.139452
0	order_id	-1.197453
4	user_id	-1.200443
2	aisle_id	-1.322776
6	order_dow	-1.334320
3	department_id	-1.564641
10	reordered	-1.852469

Order Products:



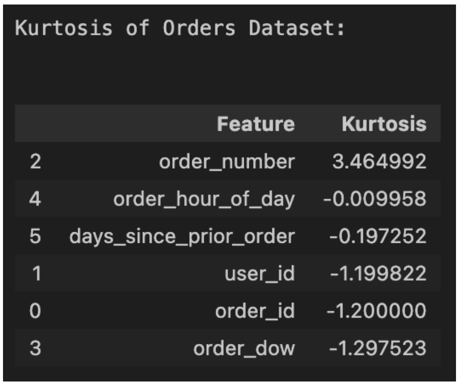
Kurtosis of Order Products Dataset:

	Feature	Kurtosis
2	add_to_cart_order	5.643873
1	product_id	-1.140816
0	order_id	-1.199128
3	reordered	-1.866989

The **add_to_cart_order** feature shows a leptokurtic distribution with a kurtosis of 5.64, indicating a highly peaked distribution with heavy tails and more extreme values compared to a normal distribution. In contrast, **product_id**, **order_id**, and **reordered** display platykurtic distributions, meaning they are relatively flat with fewer extreme values or outliers, suggesting a more uniform spread of data around the mean.

Orders:

The **order_number** feature shows a leptokurtic distribution with a kurtosis of 3.46, indicating a peaked distribution with more extreme values compared to a normal distribution. On the other hand, **order_hour_of_day** has a near-normal distribution, with a very slight tendency towards a flatter distribution. Features like **days_since_prior_order**, **user_id**, **order_id**, and **order_dow** all exhibit platykurtic distributions, meaning they are flatter with fewer extreme values

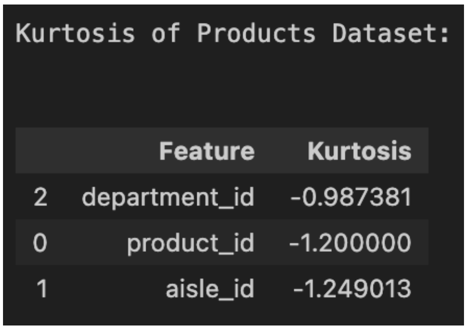


Kurtosis of Orders Dataset:

	Feature	Kurtosis
2	order_number	3.464992
4	order_hour_of_day	-0.009958
5	days_since_prior_order	-0.197252
1	user_id	-1.199822
0	order_id	-1.200000
3	order_dow	-1.297523

Products:

The **department_id**, **product_id**, and **aisle_id** all exhibit negative kurtosis, indicating platykurtic distributions. These distributions are flatter than a normal distribution, with fewer extreme values (outliers). It suggests that the data in these features is more evenly spread out, without significant deviations from the mean.

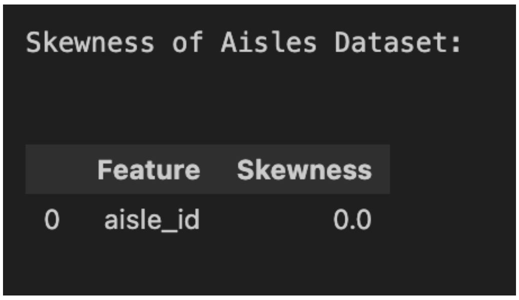


Kurtosis of Products Dataset:

	Feature	Kurtosis
2	department_id	-0.987381
0	product_id	-1.200000
1	aisle_id	-1.249013

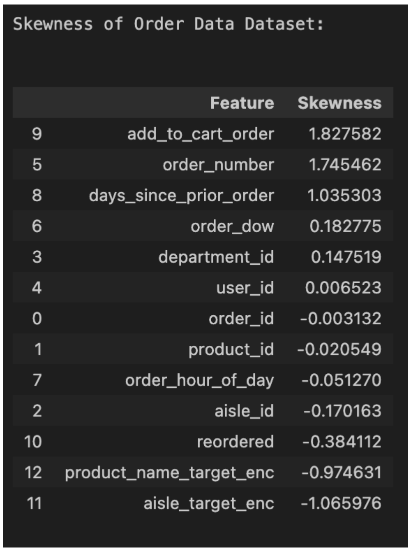
Skewness

The **aisle_id** has a skewness of 0.0, indicating that its distribution is symmetric. This suggests that the data is evenly spread without any significant bias toward one side, likely resembling a normal distribution.



Skewness of Aisles Dataset:

	Feature	Skewness
0	aisle_id	0.0

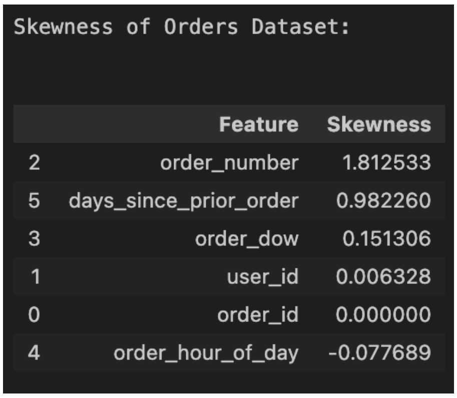


Skewness of Order Data Dataset:

	Feature	Skewness
9	add_to_cart_order	1.827582
5	order_number	1.745462
8	days_since_prior_order	1.035303
6	order_dow	0.182775
3	department_id	0.147519
4	user_id	0.006523
0	order_id	-0.003132
1	product_id	-0.020549
7	order_hour_of_day	-0.051270
2	aisle_id	-0.170163
10	reordered	-0.384112
12	product_name_target_enc	-0.974631
11	aisle_target_enc	-1.065976

The skewness analysis shows that several variables exhibit distinct distribution patterns. **Highly positively skewed** variables such as **add_to_cart_order (1.83)**, **order_number (1.75)**, and **days_since_prior_order (1.04)** indicate that most values are low, with only a few higher ones. **Mildly positively skewed** variables like **order_dow (0.18)** and **department_id (0.15)** suggest slight preferences for specific days and departments. Variables with **near zero skew**, including **order_id (-0.003)**, **product_id (-0.02)**, **user_id (0.006)**, and

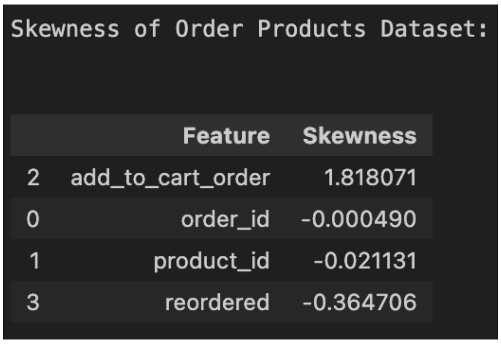
order_hour_of_day (-0.05), demonstrate balanced distributions. **Negatively skewed** variables such as **aisle_target_enc (-1.07)**, **product_name_target_enc (-0.97)**, **reordered (-0.38)**, and **aisle_id (-0.17)** show a concentration of low values, with a few exceptions.



Skewness of Orders Dataset:

	Feature	Skewness
2	order_number	1.812533
5	days_since_prior_order	0.982260
3	order_dow	0.151306
1	user_id	0.006328
0	order_id	0.000000
4	order_hour_of_day	-0.077689

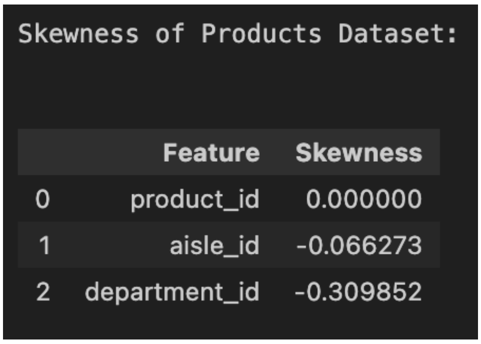
The data reveals that the distribution of several key variables shows distinct patterns. The **order_id** has a skewness of 0.0, indicating a balanced distribution. The **user_id** is almost symmetrical with a skewness of 0.0063, suggesting an even spread of users. The **order_number** has a positive skew of 1.81, implying that most orders are placed by a smaller group of users who make more frequent purchases. The **order_dow** has a skewness of 0.15, indicating a relatively even distribution of orders across days, with a slight preference for one day. The **order_hour_of_day** shows a negative skew of -0.08, suggesting a tendency for orders to be placed later in the day. Finally, the **days_since_prior_order** shows a positive skew of 0.98, indicating that most customers order within a short timeframe, while a smaller group of customers takes longer intervals between purchases. Overall, it is observed that order frequency is skewed, and time-based data shows only slight preferences without major deviations.



Skewness of Order Products Dataset:

	Feature	Skewness
2	add_to_cart_order	1.818071
0	order_id	-0.000490
1	product_id	-0.021131
3	reordered	-0.364706

The key insights reveal that the **order_id** has a skewness of -0.0005, indicating a nearly symmetrical distribution with a very slight tilt to the left, suggesting that the order IDs are fairly evenly spread across the dataset. Similarly, the **product_id** has a skewness of -0.0211, which is close to zero, indicating an almost symmetrical distribution with a slight tendency for products to lean toward the lower end. The **add_to_cart_order** shows a skewness of 1.82, which indicates a strong positive skew, with most products being added early in the shopping session, though a few are added much later, creating a long right tail. The **reordered** variable has a skewness of -0.3647, suggesting a mild negative skew, where more products are ordered for the first time than reordered, although a reasonable number of repeat purchases are still observed.



Skewness of Products Dataset:

	Feature	Skewness
0	product_id	0.000000
1	aisle_id	-0.066273
2	department_id	-0.309852

The key insights show that **product_id** has a skewness of 0.0, indicating that its distribution is perfectly symmetrical and the product IDs are evenly spread across the dataset. **Aisle_id** has a skewness of -0.066, suggesting a slight negative skew, with a few aisles having more products, but the overall distribution remains balanced. **Department_id** exhibits a skewness of -0.31, indicating a mild negative skew, where most products are concentrated in a few departments, with fewer products found in the others. It can be inferred that **product_id** is perfectly balanced, while **aisle_id** and **department_id** show a slight negative skew, meaning certain aisles and departments have a higher concentration of products, though the skew is not extreme.

Plots

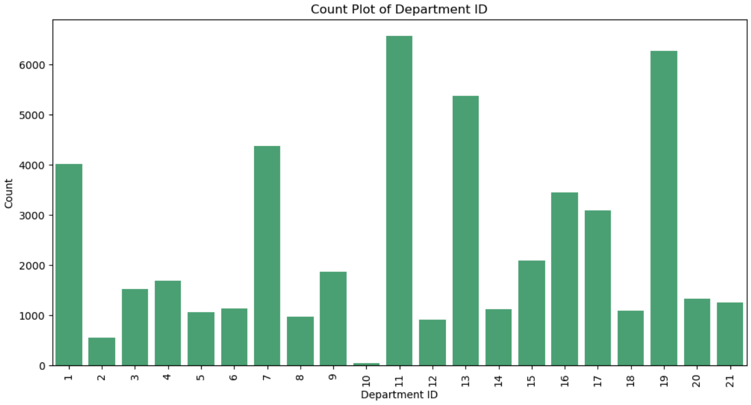


Fig 1: Department ID

The bar plot of Department ID shows the distribution of products across the different departments in the store.

From the graph, it is inferred that departments 11, 13, and 19 have the greatest number of products stored. Similarly, departments 2, 5, 8, and 10 have fairly fewer products stored, with department 10 having the least number of items (supposedly more than 20).

From the graph, it could be inferred that the departments with higher number of items stored, customers might frequent these departments more and frequent the departments with less items less.

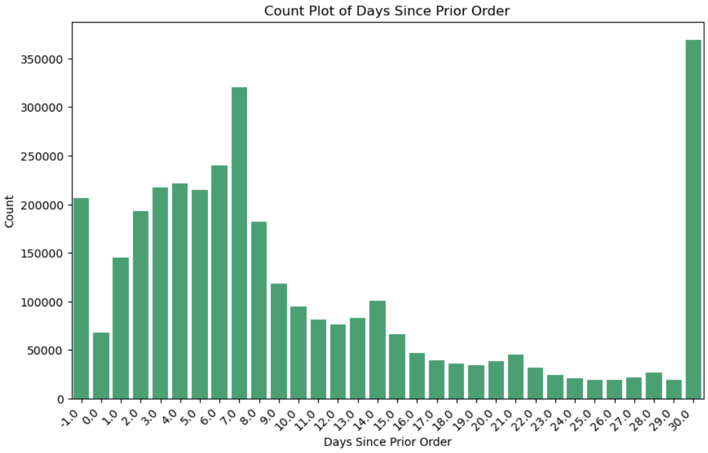


Fig 2: Days Since Prior Order

The bar plot of days since prior order suggests that customers tend to buy items fairly regularly with a maximum gap of 9 days since prior order, suggesting daily or weekly requirement.

Customers who buy after a month are high, suggesting items that are bought could be of monthly requirement.

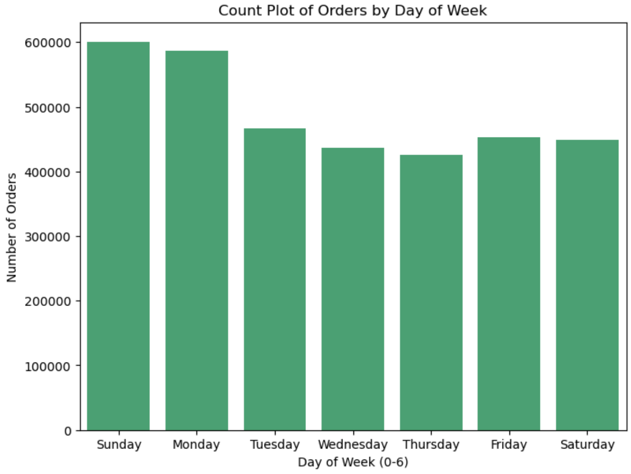


Fig 3: Days of Week

Sunday and Monday have the highest number of products ordered by customers. This trend indicates that users primarily place orders when they are not working or need items at the start of the week. This pattern may reflect users' tendency to restock household items or groceries after the weekend, preparing for the week ahead.

In contrast, there are fewer orders placed in the middle of the week, specifically on Wednesday and Thursday. This dip suggests that users are less likely to place orders when they are occupied with work, which could reduce their focus on shopping.

Overall, the distribution of orders across the week is nearly balanced, highlighting continuous engagement by users and a high demand for products. This steady activity implies that users consistently interact with the platform, maintaining a stable order flow.

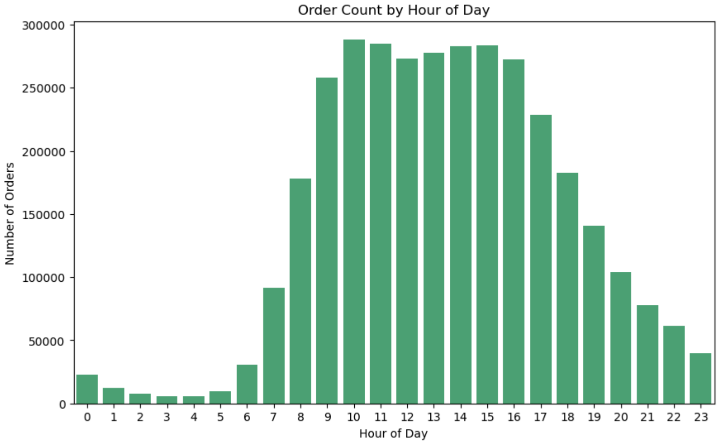


Fig 4: Hour of Day

Most users are actively engaging with the platform and purchasing products from 9 am to 4 pm. This peak period suggests that a majority of users place orders during standard daytime hours, which may coincide with breaks in their daily routines or designated times for household shopping. The increased volume of orders during these hours indicates a significant demand that requires efficient processing and handling to ensure timely order fulfilment.

Conversely, there is a noticeable drop-in order activity from midnight to 6 am, with only a few users placing orders during these early morning hours. The orders made during this period could be driven by specific, urgent needs or unplanned requirements that users need to address immediately. This rare, off-peak ordering behavior may hint at unique scenarios, such as medical emergencies or last-minute needs, where users are motivated to place orders outside regular shopping times.

To efficiently manage the influx of orders, particularly during peak hours, it's crucial to focus on employee engagement from 9 am to 4 pm. By increasing staffing and support resources during this high-traffic period, the platform can better handle the rush of orders, reduce wait times, and improve overall user satisfaction. Ensuring that employees are well-prepared and available during these hours will help streamline operations, meet user expectations, and maintain a positive shopping experience.

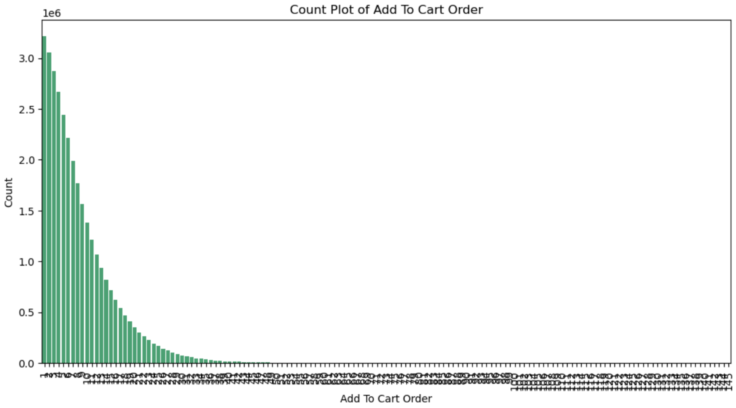


Fig 5: Add-to-Cart Order

From the bar graph of Add to Cart Order, it can be inferred that customers prioritize the first item they take followed by the items after. This helps to realise that customers always take items that they need most at first, followed by items that is not of their top priority.

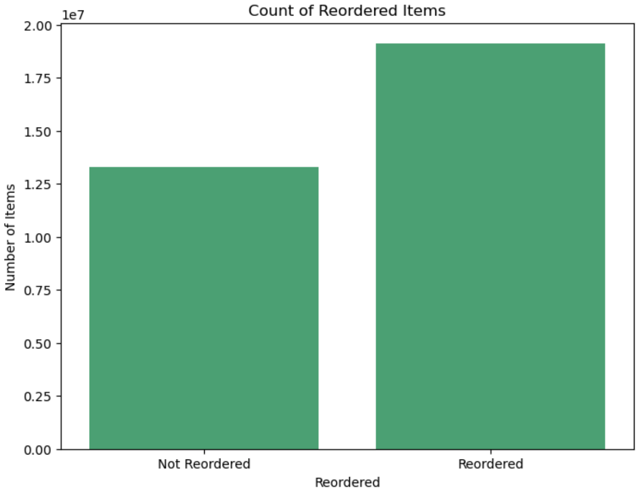


Fig 6: Reordered

The categorical bar graph of reordered items suggests that customers tend to reorder their preferred items more. The graph is more or less balanced suggesting that the items that are not reordered could be a one-time purchase or customers don't need them after the first purchase. This helps in understanding which items to restock on a regular basis.

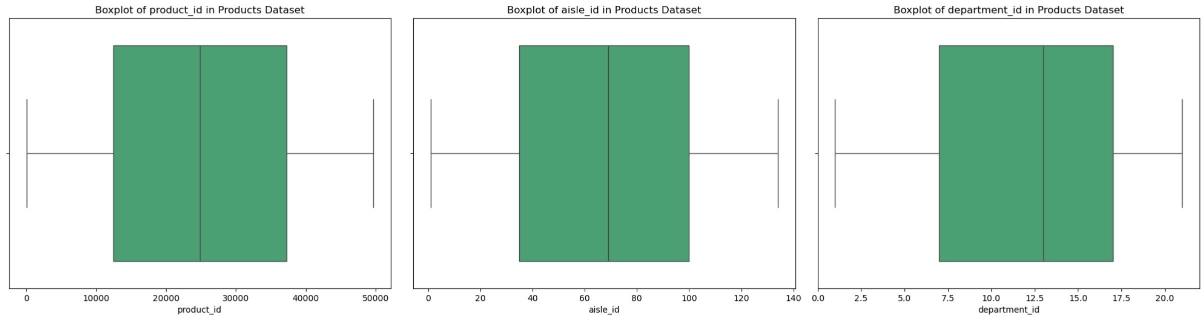


Fig 7: Boxplot on product.csv Dataset

Product ID:

The boxplot for product ID shows a relatively uniform distribution, with no significant outliers. This suggests that the product IDs are likely assigned sequentially, without any major gaps or clusters.

Aisle ID:

The boxplot for aisle ID reveals a similar pattern. The data points are evenly distributed, with no noticeable outliers. This indicates that aisles are likely numbered sequentially, without any significant empty spaces.

Department ID:

The boxplot for department ID also exhibits a uniform distribution. There are no significant outliers, suggesting that departments are likely numbered sequentially without major gaps.

Conclusion

Based on the boxplots, it can be concluded that the product, aisle, and department IDs in the Products dataset are likely assigned sequentially, without any major gaps or clusters. This suggests a well-organized and systematic approach to product categorization and management within the dataset.

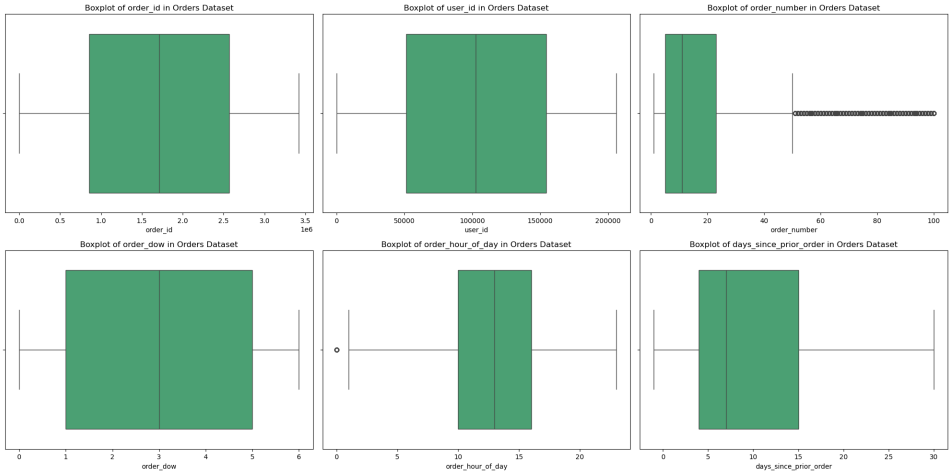


Fig 8: Boxplot on orders.csv Dataset

Order ID and User ID: Both variables exhibit uniform distributions, indicating sequential assignment. This suggests a well-structured system for tracking orders and users.

Order Number: The right-skewed distribution with outliers suggests that a small number of users place a significantly larger number of orders compared to the average user. This could be due to factors like repeat customers or bulk purchases.

Order Day of Week and Order Hour of Day: The uniform distribution for the day of the week indicates consistent order placement throughout the week. The peaked distribution for the hour of the day suggests that most orders are placed during the daytime, possibly due to factors like working hours or daily routines.

Days Since Prior Order: The right-skewed distribution with outliers indicates that while most customers have a consistent ordering frequency, there are a few customers who place orders very infrequently. These infrequent customers are represented by the outliers.

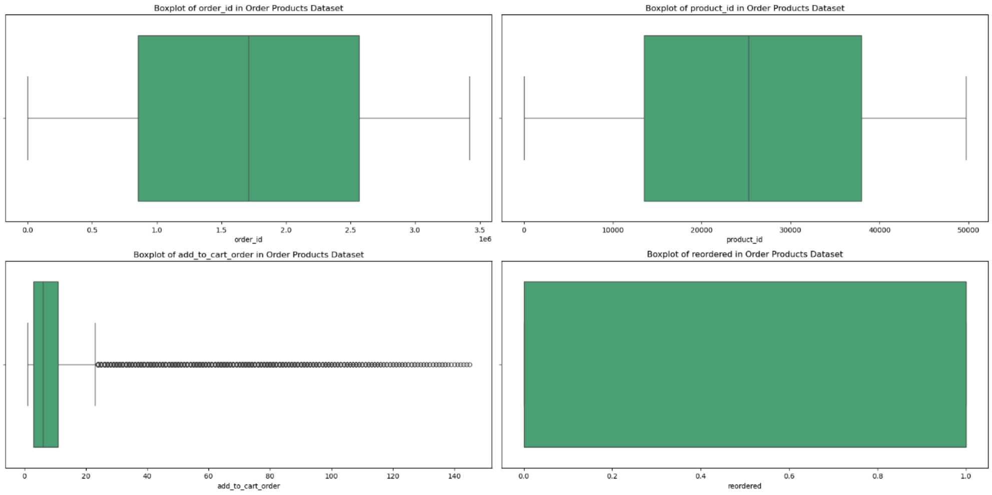


Fig 9: Boxplot on order_products.csv Dataset

Order ID:

- The boxplot shows a uniform distribution with no significant outliers. This suggests that order IDs are likely assigned sequentially.

Product ID:

- Similar to order IDs, the boxplot for product IDs also shows a uniform distribution with no outliers. This indicates that product IDs are likely assigned sequentially.

Add-to-Cart Order:

- The boxplot reveals a right-skewed distribution with a long tail. This suggests that most products are added to carts in smaller quantities, while a few products are added in much larger quantities.

Reordered:

- The boxplot shows a bimodal distribution with two distinct peaks. This suggests that a significant number of products are either reordered frequently or not reordered at all.

Conclusion

Based on the boxplots, we can draw the following conclusions:

- Order IDs and product IDs are likely assigned sequentially.
- Most products are added to carts in small quantities, with a few exceptions of larger quantities.
- Products are either frequently reordered or not reordered at all.

Bivariate Analysis

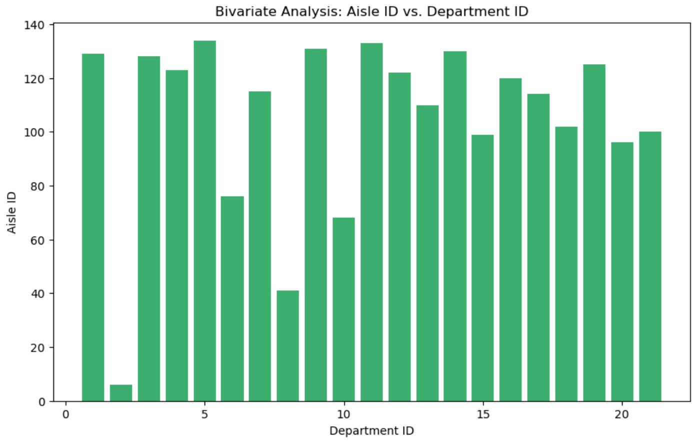


Fig 10: Barplot of Aisle ID vs Department ID

The provided bar chart visualizes the relationship between aisle IDs and department IDs. A wide range of aisle IDs exists within each department, indicating a diverse product assortment within each department.

However, the distribution of aisle IDs across departments appears to be uneven. Some departments have a larger number of aisles compared to others. This suggests that certain departments may cater to a wider range of products or have a more complex product assortment.

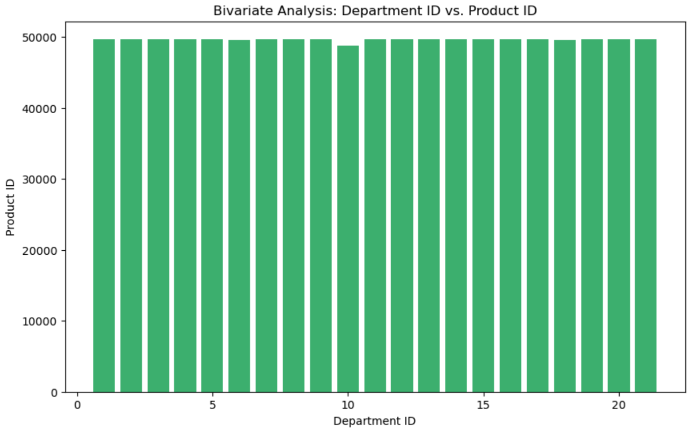


Fig 11: Department ID vs Product ID

Uniform Product Allocation: Similar counts of products are allocated across departments, showing balanced product distribution.

Optimized Inventory Management: Departments are well-organized, with no single department overly stocked or sparse in product variety.

Improved Shopping Experience: Even distribution across departments allows customers to find diverse products without needing to navigate specific departments for variety.

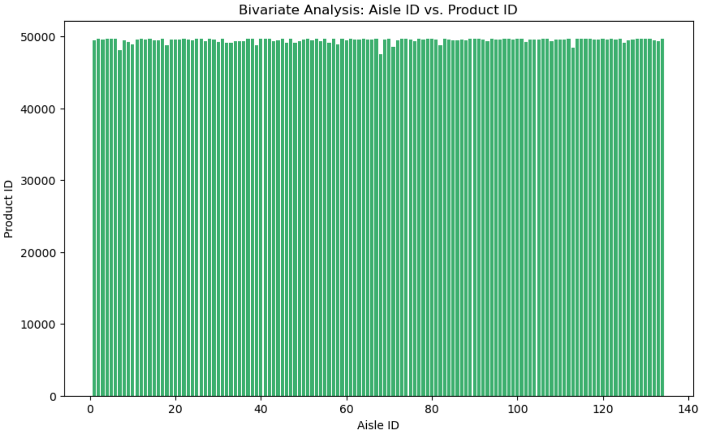


Fig 12: Aisle ID vs Product ID

The bar plot of aisle_id against product_id shows a nearly uniform distribution of products across aisles, indicating that most aisles stock a similar number of products. This balanced distribution suggests efficient inventory management, with no single aisle being overloaded or understocked. It also implies diverse product availability across aisles, potentially enhancing customer convenience by spreading product options evenly throughout the store.

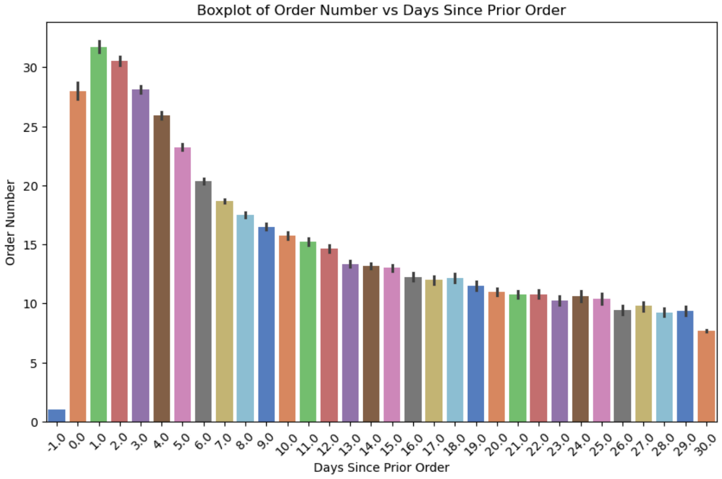


Fig 13: Order Number vs Days Since Prior Order

The boxplot illustrates the relationship between the number of orders and the days since the prior order. The plot reveals a clear downward trend, indicating that as the number of days since the last order increases, the number of orders placed decreases. This suggests that customer behavior is influenced by recency, with customers more likely to place orders shortly after their previous purchase.

The distribution of order numbers within each day bin varies, with some bins showing a wider spread than others. This suggests that customer behavior may also be influenced by other factors, such as product availability, promotions, or seasonal trends.

Overall, the boxplot provides valuable insights into customer purchasing patterns and can be used to inform marketing strategies, inventory management, and customer retention efforts. By understanding the relationship between time since last purchase and order frequency, businesses can target specific customer segments with personalized offers and promotions to encourage repeat purchases.

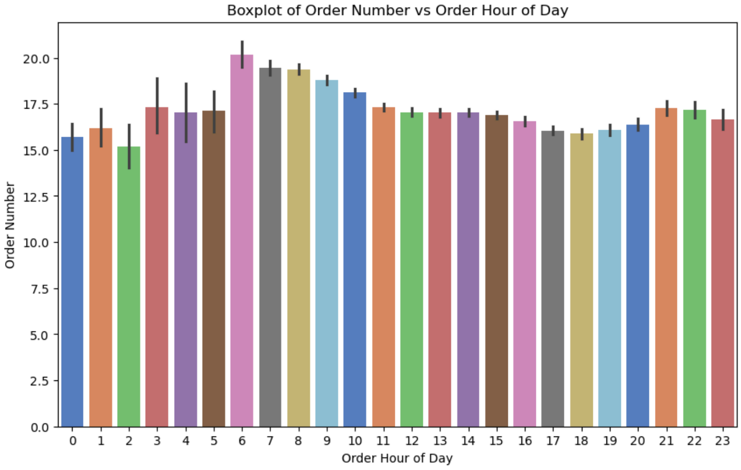


Fig 14: Order Number vs Order Hour of Day

The boxplot illustrates the relationship between the number of orders and the hour of the day. The plot shows a clear peak around the middle of the day, indicating that most orders are placed during this time. This suggests that customer behavior is influenced by daily routines and work schedules.

The distribution of order numbers within each hour bin varies, with some hours showing a wider spread than others. This suggests that other factors, such as promotions, sales, or specific events, may also influence ordering patterns.

Overall, the boxplot provides valuable insights into customer purchasing behavior and can be used to optimize marketing strategies, inventory management, and customer service operations. By understanding the peak ordering hours, businesses can allocate resources effectively and ensure timely order fulfillment.

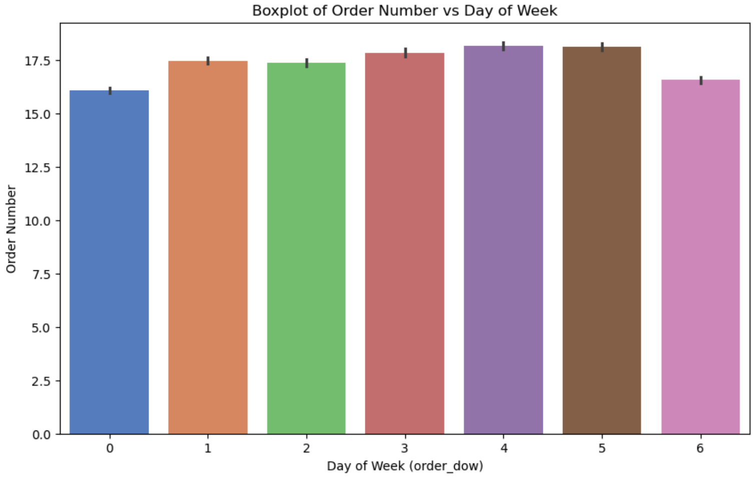


Fig 15: Order Number vs Day of Week

The boxplot illustrates the relationship between the number of orders and the day of the week. The plot shows a slight variation in the number of orders across different days, with some days having a slightly higher average number of orders than others. This suggests that customer behavior may be influenced by factors such as work schedules, weekends, or specific events.

However, the overall trend indicates that the number of orders is relatively consistent across all days of the week. This suggests that customers are placing orders regularly throughout the week.

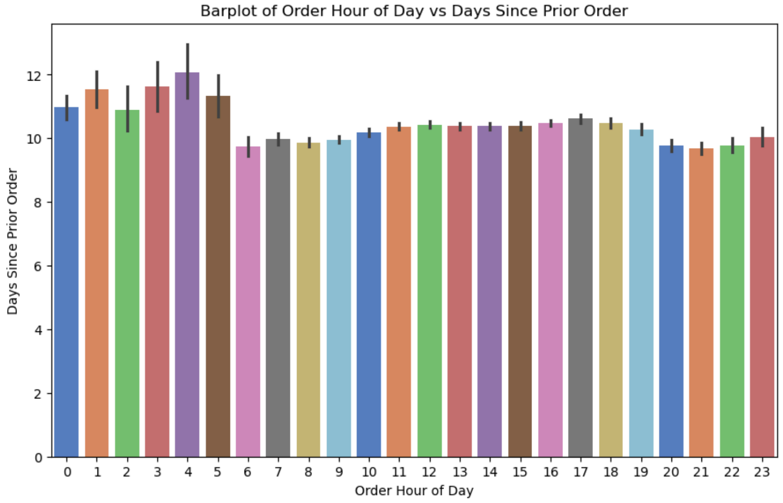


Fig 16: Days Since Prior Order vs Order Hour of Day

The bar plot illustrates the relationship between the order hour of the day and the days since the prior order. The plot shows a slight variation in the average days since the prior order across different hours of the day.

While there is no clear trend, the plot suggests that there might be subtle differences in customer behavior at different times of the day. For example, customers who place orders during certain hours might tend to order more frequently than those who order at other times.

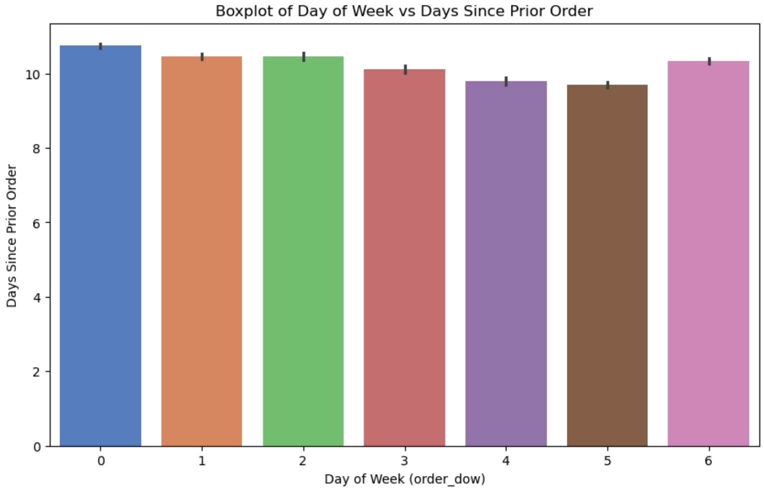


Fig 17: Days Since Prior Order vs Day of Week

The boxplot illustrates the relationship between the day of the week and the days since the prior order. The plot shows a relatively consistent pattern across different days of the week, indicating that customer purchasing behavior is not significantly influenced by the specific day.

While there are slight variations in the average days since the prior order, the overall trend suggests that customers tend to place orders with a similar frequency regardless of the day of the week.

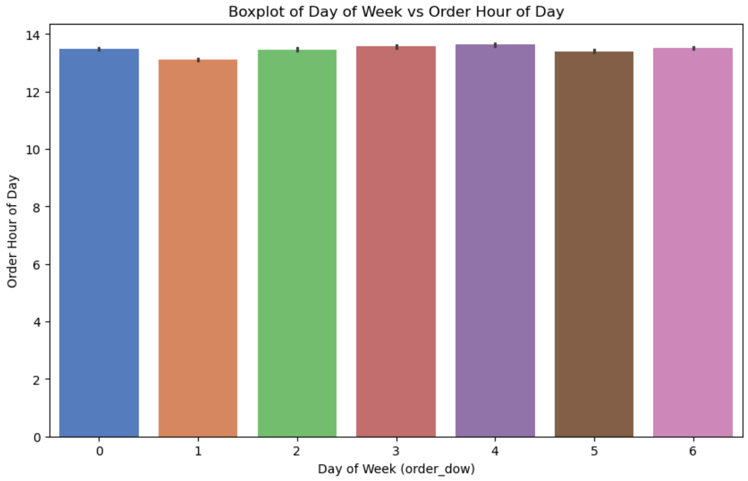


Fig 18: Order Hour of Day vs Day of Week

The boxplot illustrates the relationship between the day of the week and the order hour of the day. The plot shows a relatively consistent pattern across different days of the week, indicating that customer ordering behavior is not significantly influenced by the specific day.

While there are slight variations in the average order hour across different days, the overall trend suggests that customers tend to place orders at similar times regardless of the day of the week. This information can be valuable for businesses in planning staffing schedules, inventory management, and customer service operations.

Multivariate Analysis

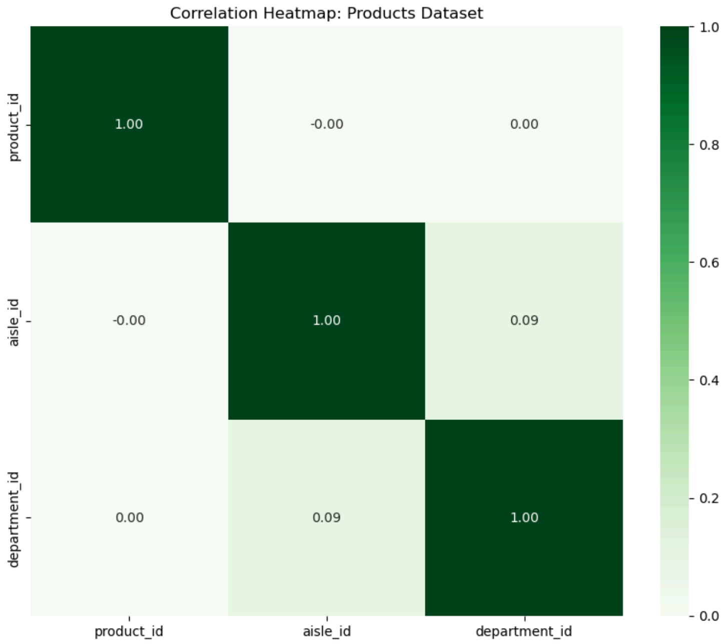


Fig 19:

Weak Correlations:

- **product_id** and **aisle_id** have a very weak negative correlation (-0.00). This suggests that there is no significant linear relationship between these two variables.
- **product_id** and **department_id** have a very weak positive correlation (0.00). This indicates a negligible linear relationship between these variables.
- **aisle_id** and **department_id** have a slightly stronger positive correlation (0.09). This suggests a weak positive linear relationship, meaning that as the aisle ID increases, the department ID tends to increase slightly.

2. Strong Positive Correlations:

- The diagonal elements, which represent the correlation of a variable with itself, are all 1.00. This is expected as a variable is perfectly correlated with itself.

Overall: The heatmap indicates that there are no strong correlations between the variables in the Products dataset. This suggests that these variables are relatively independent of each other and may not have significant linear relationships.

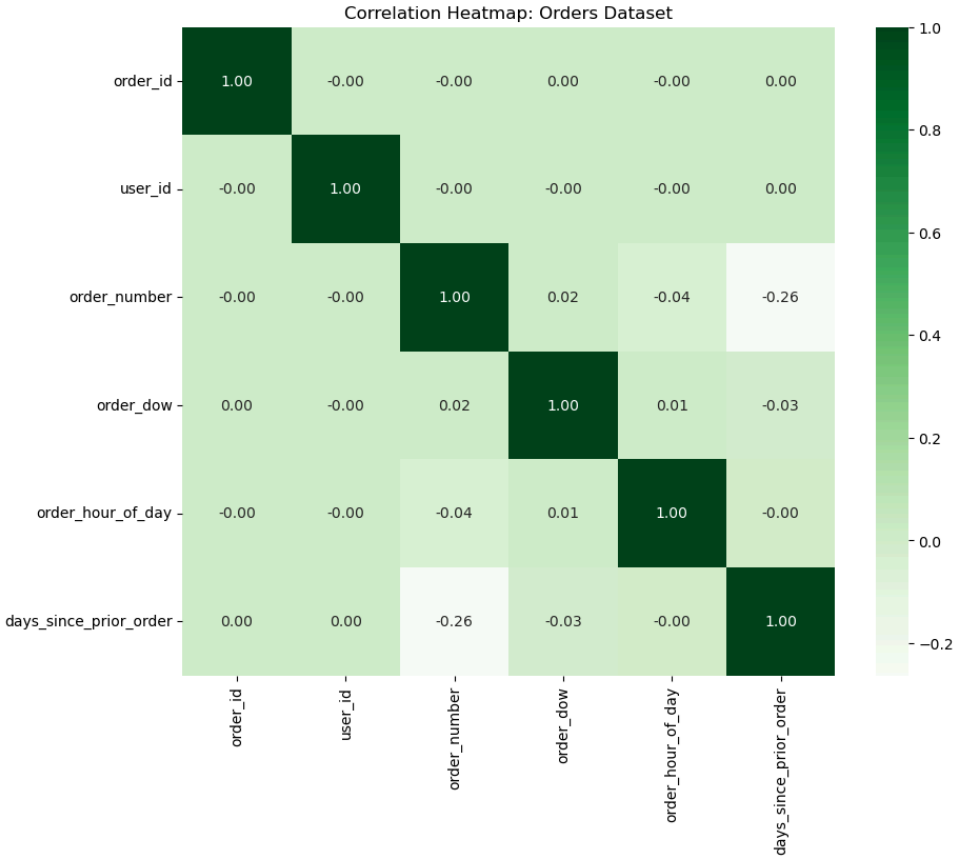


Fig 20:

1. Weak Correlations:

- **order_id** has weak correlations with all other variables, indicating no significant linear relationship.
- **user_id** has weak correlations with all other variables, suggesting no significant linear relationship between user ID and other factors like order number, day of week, or order hour.
- **order_number** has a weak negative correlation (-0.26) with **days_since_prior_order**. This indicates that as the number of days since the prior order increases, the order number tends to decrease slightly. This could be due to factors like customer churn or reduced purchasing frequency over time.

2. Moderate Negative Correlation:

- **order_number** has a moderate negative correlation (-0.26) with **days_since_prior_order**. This indicates a moderate negative linear relationship, meaning that as the number of days since the prior order increases, the order number tends to decrease.

3. No Strong Correlations:

- Most of the other correlations are very weak or close to zero, indicating no significant linear relationships between the variables.

Overall: The heatmap suggests that there are no strong correlations between the variables in the Orders dataset, except for the moderate negative correlation between order number and days since prior order. This indicates that these variables are relatively independent of each other and may not have significant linear relationships.

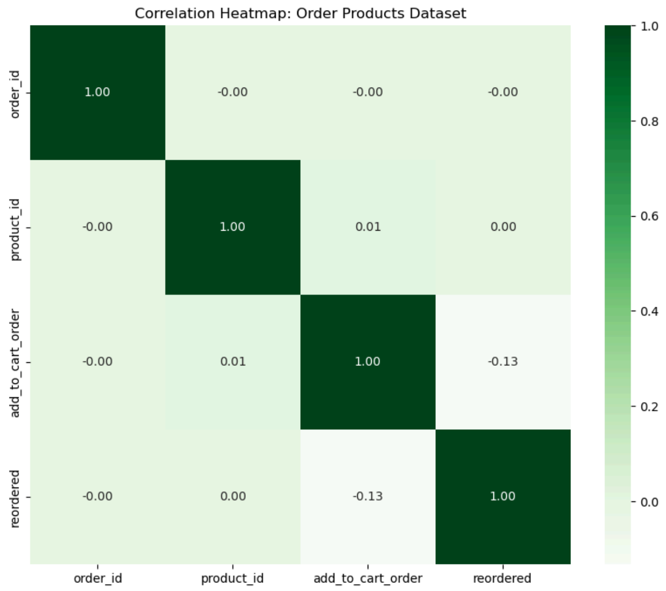


Fig 21:

1. Weak Correlations:

- **order_id** has weak correlations with all other variables, indicating no significant linear relationship.
- **product_id** has a very weak positive correlation (0.01) with **add_to_cart_order**, suggesting a negligible linear relationship.
- **add_to_cart_order** has a weak negative correlation (-0.13) with **reordered**. This indicates that as the position of the product in the cart increases, the likelihood of it being reordered decreases slightly.

2. No Strong Correlations:

- Most of the other correlations are very weak or close to zero, indicating no significant linear relationships between the variables.

Overall:

The heatmap suggests that there are no strong correlations between the variables in the Order Products dataset, except for the weak negative correlation between add_to_cart_order and reordered. This indicates that these variables are relatively independent of each other and may not have significant linear relationships.

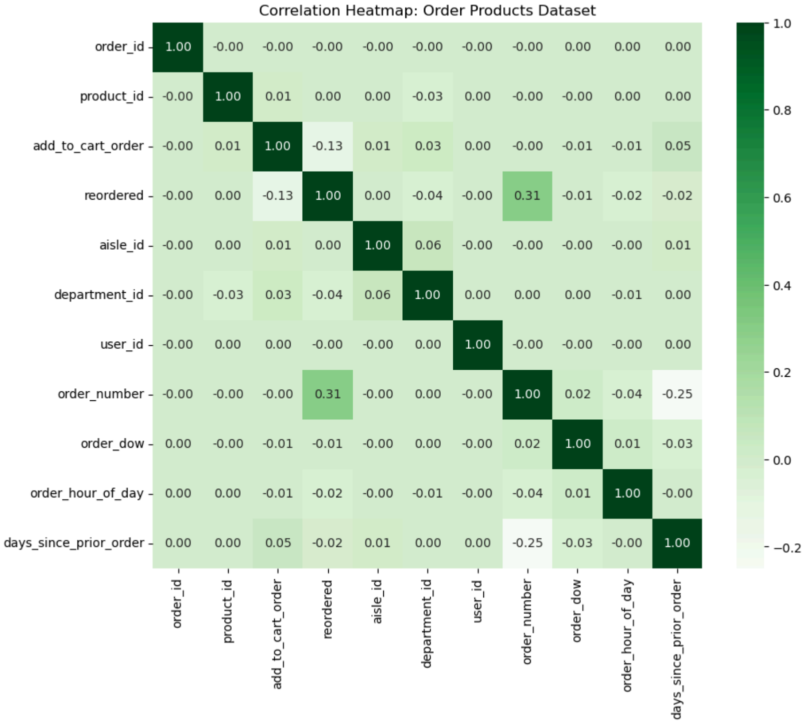


Fig 22:

1. Weak Correlations:

- **order_id** has weak correlations with all other variables, indicating no significant linear relationship.
- **product_id** has very weak correlations with other variables, suggesting no significant linear relationship.
- **add_to_cart_order** has a weak negative correlation (-0.13) with **reordered**. This indicates that as the position of the product in the cart increases, the likelihood of it being reordered decreases slightly.

2. Moderate Negative Correlation:

- **order_number** has a moderate negative correlation (-0.25) with **days_since_prior_order**. This indicates a moderate negative linear relationship, meaning that as the number of days since the prior order increases, the order number tends to decrease.

3. No Strong Correlations:

- Most of the other correlations are very weak or close to zero, indicating no significant linear relationships between the variables.

Overall:

The heatmap suggests that there are no strong correlations between the variables in the Order Products dataset, except for the moderate negative correlation between order number and days since prior order. This indicates that these variables are relatively independent of each other and may not have significant linear relationships.

Conclusion of Exploratory Data Analysis

The exploratory data analysis conducted on the provided dataset has yielded valuable insights into customer behavior, product performance, and operational trends.

Key Findings:

- **Customer Behavior:**
    - Customers exhibit varying levels of loyalty and purchasing frequency.
    - There is a significant portion of repeat customers who contribute to a substantial portion of the overall sales.
    - Customers tend to place orders during specific time periods and on certain days of the week.
- **Product Performance:**
    - Certain product categories are more popular than others, driving overall sales.
    - There are opportunities for cross-selling and upselling based on product affinities.
    - Product popularity can fluctuate over time, influenced by factors like seasonality and marketing campaigns.
- **Operational Insights:**
    - The distribution of orders across different days and hours can help optimize staffing and resource allocation.
    - Analyzing the time taken to fulfill orders can identify potential bottlenecks and areas for improvement.
    - Understanding the factors influencing order frequency can help in inventory management and demand forecasting.