EDA REPORT

Objective:

Develop a personalized recommendation system that suggests products to users based on their historical purchase behavior and related user and product data. The project aims to enhance user experience by providing tailored recommendations, thereby increasing user engagement and sales.

Project Description:

This project focuses on understanding user purchasing patterns to recommend products they are likely to buy. The dataset contains user purchase history across various products, including when and how often items are reordered. The goal is to use this data to build a recommendation system that leverages both unsupervised and supervised learning methods for optimal performance.

Datasets:

Orders_csv, Products_csv, Order_products_csv, Departments_csv, Aisles_csv

Preprocessing:

1. Merging all the tables together :

Inner join is performed on all datasets and a merged dataset is obtained which contains 32434489 rows × 15 columns.

2. Data cleaning :

Dropped ‘eval_set’ column as it does not capture the values training and test that was present in it before merging. So, it has to be dropped otherwise the model won't be able to capture all the features

3. Missing values treatment:

- ‘days_since_prior_order’ contains a lot of missing values . Since there are too many null values, imputation needs to be performed

- Filled NAN values in the dataset with -1. Then added 1 to each value in the dataset which will convert the NAN values to 0. The overall pattern remains intact in the dataset

EDA:

1. Information about the variables present in the dataset

```
Data columns (total 15 columns):
 #   Column                  Dtype
---  ------                  -----
 0   product_id              int64
 1   product_name            object
 2   aisle_id                int64
 3   department_id           int64
 4   department              object
 5   aisle                   object
 6   order_id                int64
 7   add_to_cart_order       int64
 8   reordered               int64
 9   user_id                 int64
 10  eval_set                object
 11  order_number            int64
 12  order_dow               int64
 13  order_hour_of_day       int64
 14  days_since_prior_order  float64
dtypes: float64(1), int64(10), object(4)
```

Variables description:

Variables in the dataset

- product_id: the id of the products which is an unique identifier
- product_name: the names of the products
- aisle_id: the id of aisle where it is located
- department_id: the id of the department to which the products belong
- department: contains names of various departments available
- aisle: the aisle contains the name descriptions to which the products is located
- order_id: the order id which is the unique identifer for each order
- add_to_cart_order: the position of the product within the customer's cart
- reordered: a binary variable indicating whether the product was previously ordered or not
- user_id: an unique identifier for each user
- eval_set: contains 3 categories prior, training and test
- order_number: the sequence number of the order for the user
- order_dow: the day of the week in which the order was placed
- order_hour_of_day: the hour of the day in which the order was placed
- days_since_prior_order: the number of days since the user's previous order

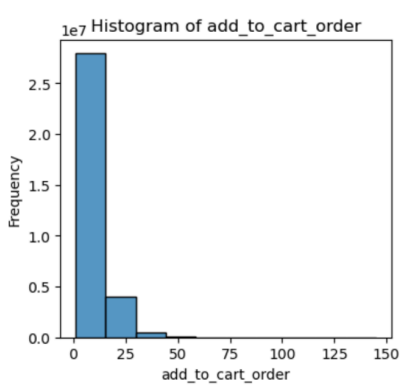
2. Univariate analysis

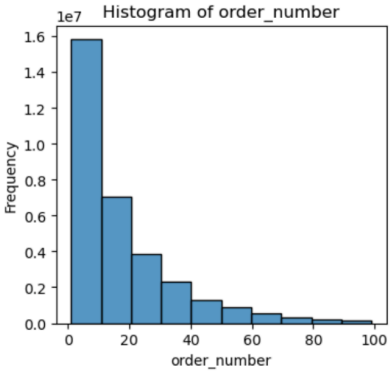
The value counts of all the categorical variables in the dataset states that:

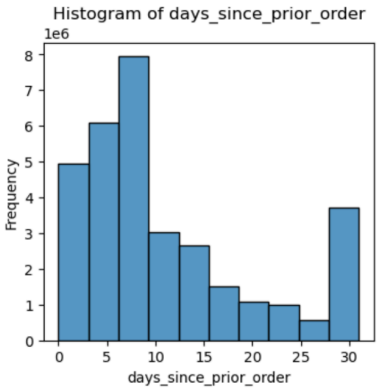
- The product Banana is the most frequently purchased item in the ‘product_name’ category, appearing most often in the dataset. This suggests that bananas are a top choice among customers.
- The Produce department is the most prominent in terms of sales, indicating that fresh produce items dominate the sales in this dataset.
- The Fresh Fruits and Fresh Vegetables aisles are the most frequently represented in the data, reflecting the popularity of fresh items in customer purchases.
- The order_dow variable, which represents the day of the week (ranging from 0 to 6), captures the distribution of orders across the week.
- The reordered column, a binary variable with values 0 and 1, indicates whether a product was reordered.

3. Stratified sampling was performed on the dataset, with the 'days_since_prior_order' column.. A 10% sample of the data was selected, ensuring that the distribution of values in the 'days_since_prior_order' feature was preserved. This approach reduces the risk of bias and ensures that the sample accurately reflects the overall distribution of customer order behavior.

4. Histogram and box plots on numerical columns







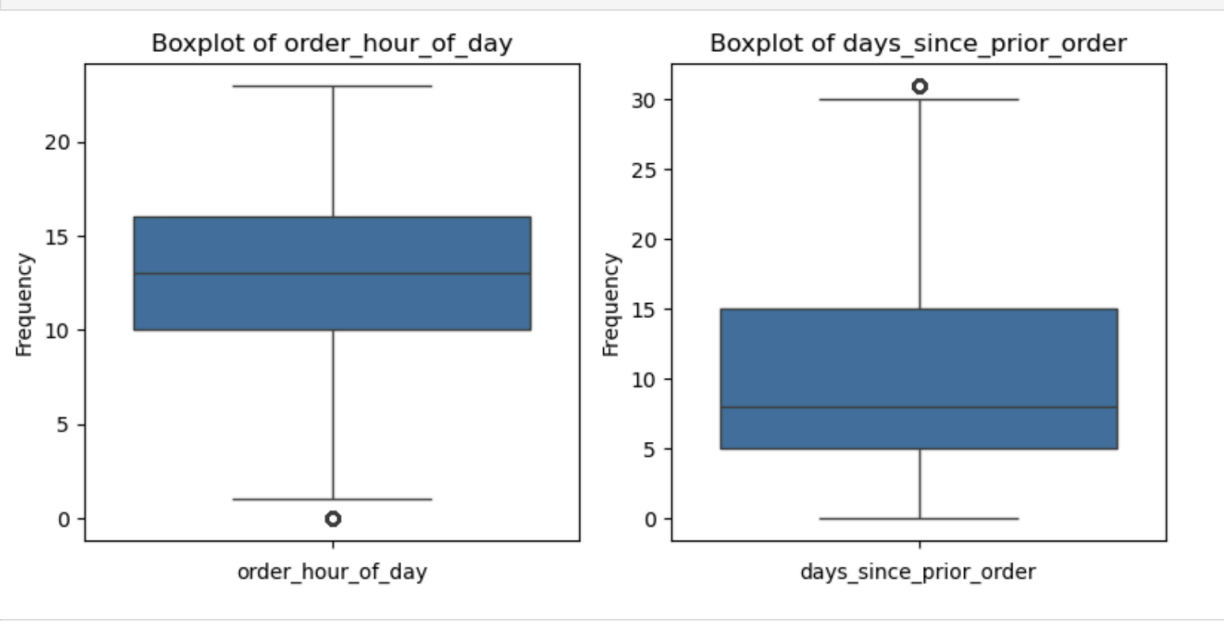
Insights:

- The histogram of ‘add_to_cart_order’ shows that the majority of items are added to cart within the first 25 positions.
- A sharp decline in frequency is seen after the initial positions
- Very few items are added to cart beyond position 50

- The histogram of ‘order_number’ shows a clear exponential decay pattern. The highest frequency is concentrated in the 0-20 range.

- A steady decrease in frequency is noted as the order number increases
- Orders beyond number 80 are rarely observed

- The histogram of ‘days_since_prior_order’ shows a distinct peak around 7-10 days between orders
- The distribution shows a gradual build-up from 0-7 days
- A consistent decline is noted after the 10-day mark
- An interesting uptick is seen at the 30-day mark, suggesting monthly shopping patterns
- The majority of repeat purchases are made within the first 15 days



Boxplot insights:

Order Hour of Day Boxplot:
- The median ordering time is observed at approximately 13
- A relatively wide interquartile range is noted between 10 and 16
- Some outliers are detected in the early morning hours
- The bulk of ordering activity is concentrated during daytime hours

Days Since Prior Order Boxplot:
- The median time between orders is determined to be approximately 8
- An interquartile range spanning from 5 to 15 is observed
- A maximum typical interval of 30 is noted, with some outliers beyond this point
- Outliers are primarily found in the upper range, suggesting some customers have longer gaps between orders
- The compact lower quartile indicates consistent shopping patterns for frequent customers

5. Skewness and kurtosis on numerical columns

column_name	skewness	kurtosis
add_to_cart_order	1.818071	5.643873
order_number	1.756896	3.256605
order_hour_of_day	-0.044083	-0.011658
days_since_prior_order	1.020110	0.005736
order_dow	0.180193	-1.333989

Insights

1. Add to Cart Order:
The skewness of the 'add_to_cart_order feature is 1.82, indicating a right-skewed distribution, meaning that most customers tend to add a smaller number of items to their cart, with a few customers adding significantly more.
The kurtosis value of 5.64 suggests a leptokurtic distribution, meaning that there is a higher concentration of values around the mean, with more extreme values than a normal distribution.

2. Order Number:
The skewness of 'order_number' is 1.76, which also indicates a right-skewed distribution. This suggests that most customers make fewer orders, while a small number of customers place significantly more orders.
The kurtosis value of 3.26 indicates a relatively normal distribution with slightly heavier tails than a perfect normal distribution, meaning there are some customers with either very few or very many orders.

3. Order Hour of Day:
The skewness of 'order_hour_of_day' is -0.04, indicating that the distribution is approximately symmetric.
The kurtosis value of -0.01 suggests a distribution that is very close to normal, with no significant excess of outliers.

4. Days Since Prior Order:
The skewness of 'days_since_prior_order' is 1.02, indicating a right-skewed distribution. This suggests that while most customers tend to place orders relatively soon after their last purchase, a smaller group of customers waits much longer between orders.
The kurtosis value of 0.01 indicates that the distribution is close to normal, with a slight excess of values around the mean and very few extreme outliers.

5. Order Day of Week:

The skewness of 'order_dow' is 0.18, which suggests a slight right skew, meaning there are a few days of the week with more orders than others, but the distribution is fairly even across the week.

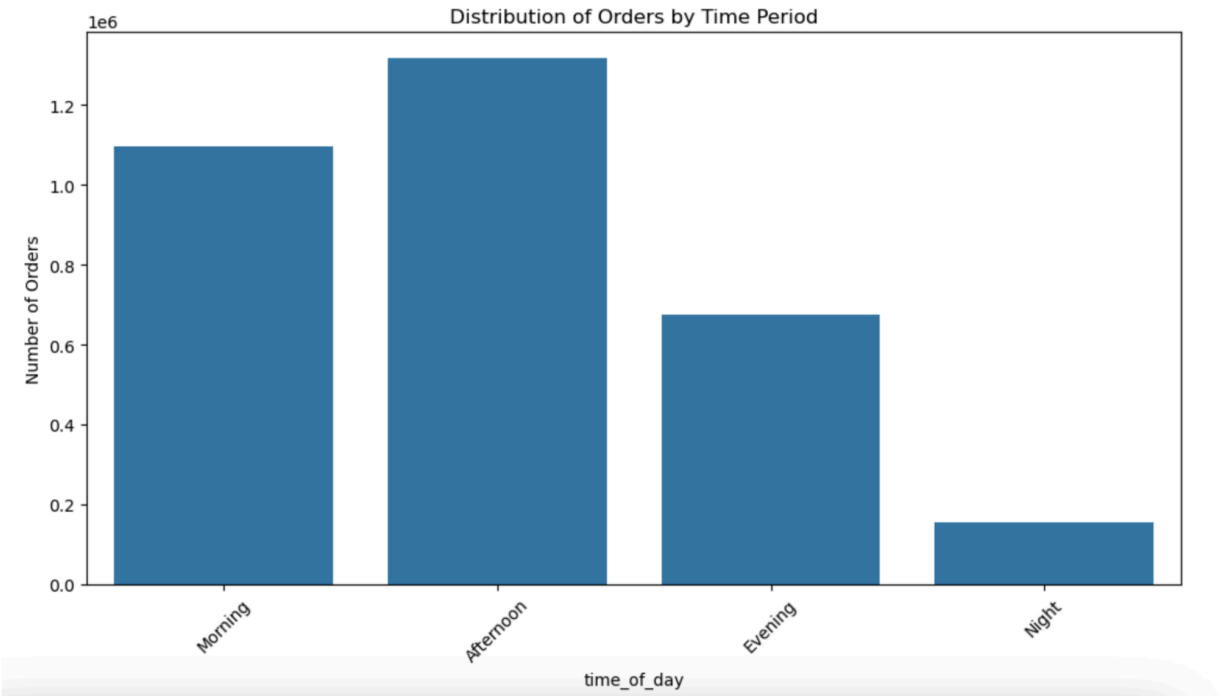
The kurtosis value of -1.33 indicates a platykurtic distribution, meaning the data has fewer extreme values and is flatter than a normal distribution, with most orders distributed relatively evenly across the week.

.

6. Binning and bucketing

For binning and bucketing, the time of day was divided into 4 categories.

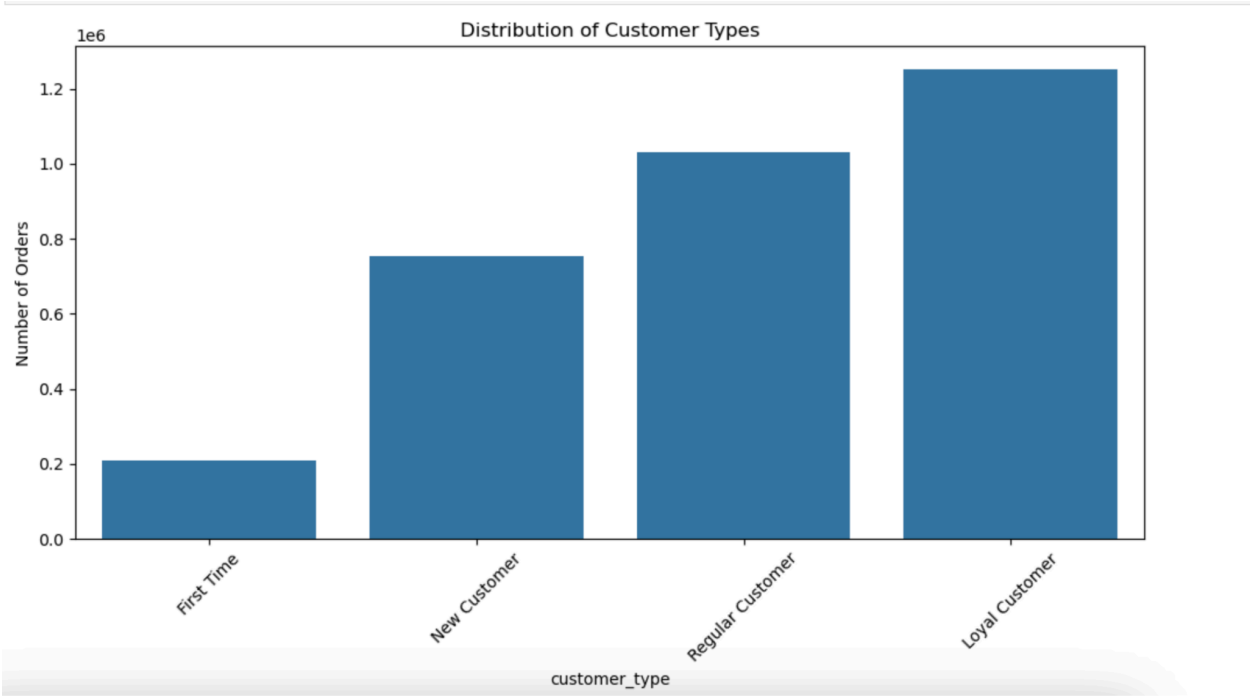
- Morning from 5 to 11 hours
- Afternoon from 12 to 16 hours
- Evening from 17 to 21 hours
- Night from 22 to 5 hours



The distribution shows that the number of orders during the afternoon that is from 12 to 16, is the highest. Meaning customers are more active during the day and are likely to place orders.

For binning and bucketing, the customer types were divided into 4 categories.

- First time whose order number is 1
- New customer whose order number is between 2 to 5
- Regular customers whose order number is between 6 to 15
- Loyal customers whose order number is more than 15

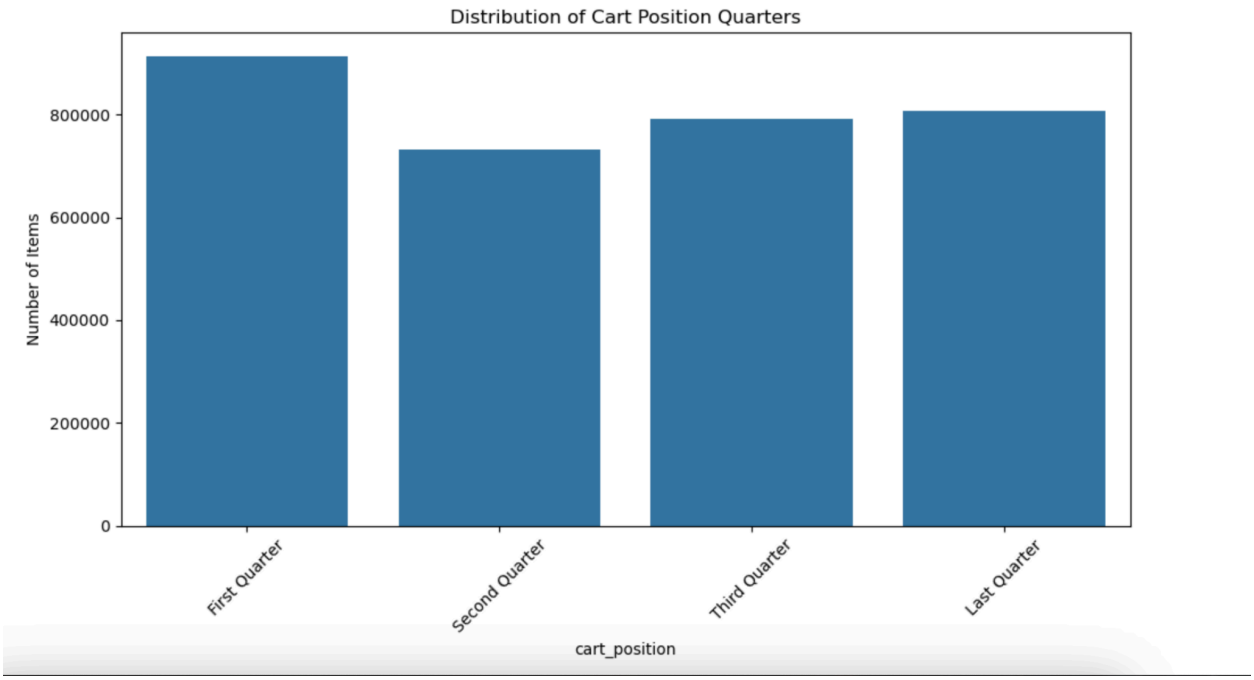


The distribution shows that the number of orders given by loyal customers that is more than 15 are the highest. This indicates that the most active customers are also the most loyal, showing consistent engagement and repeat purchases over time.

The distribution also shows that first-time customers make the fewest orders. Importance must be given on customer retention strategies for loyal customers while also implementing measures to encourage repeat purchases from first-time and new customers.

For binning and bucketing, the cart position quarters were divided into 4 categories.

- First quarter Items added first in the cart (positions 1 to 25% of the total items).
- Second quarter Items added between 26% and 50% of the total cart positions.
- Third quarter Items added between 51% and 75% of the cart positions.
- Last quarter Items added in the final 25% of the cart positions.



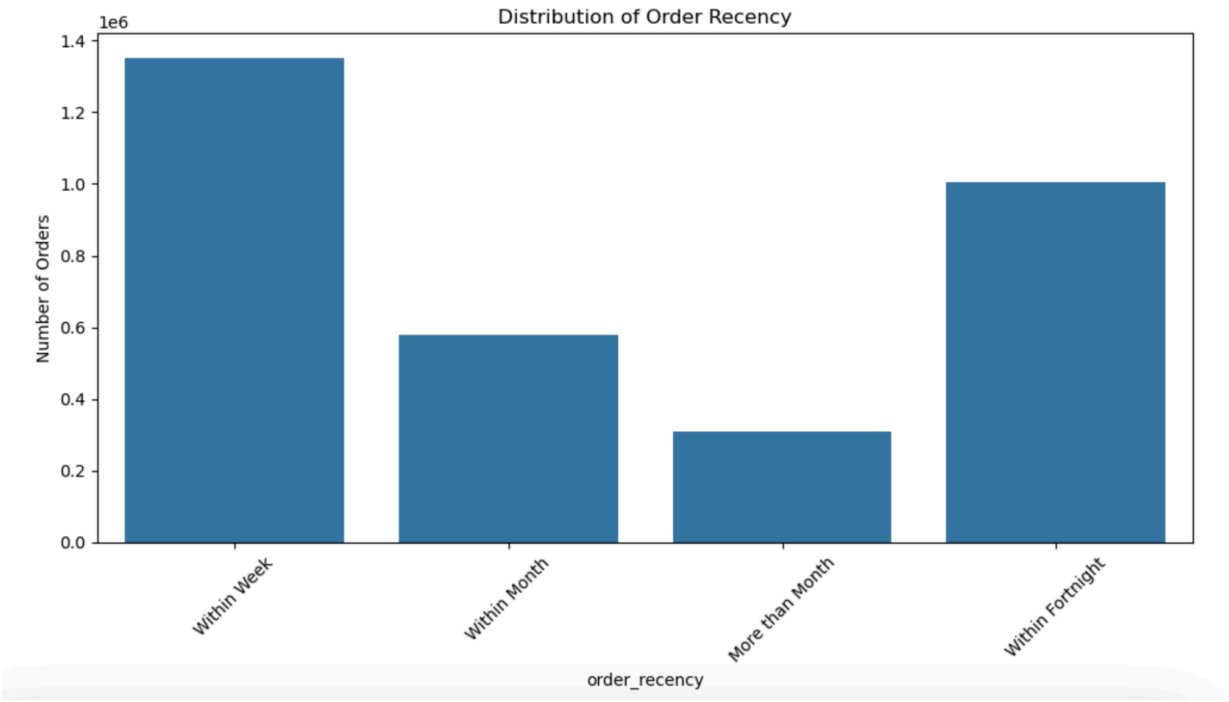
The number of items sold in the first quarter from the add_to_cart_order is the highest. This suggests that customers tend to make decisions about core products early on in their shopping journey.

The first items added to the cart may represent products that are essential or frequently purchased, or they could reflect items customers already know they want.

A decline in the number of items from the first quarter could mean a significant engagement drop.

For binning and bucketing, the order recency was divided into 4 categories.

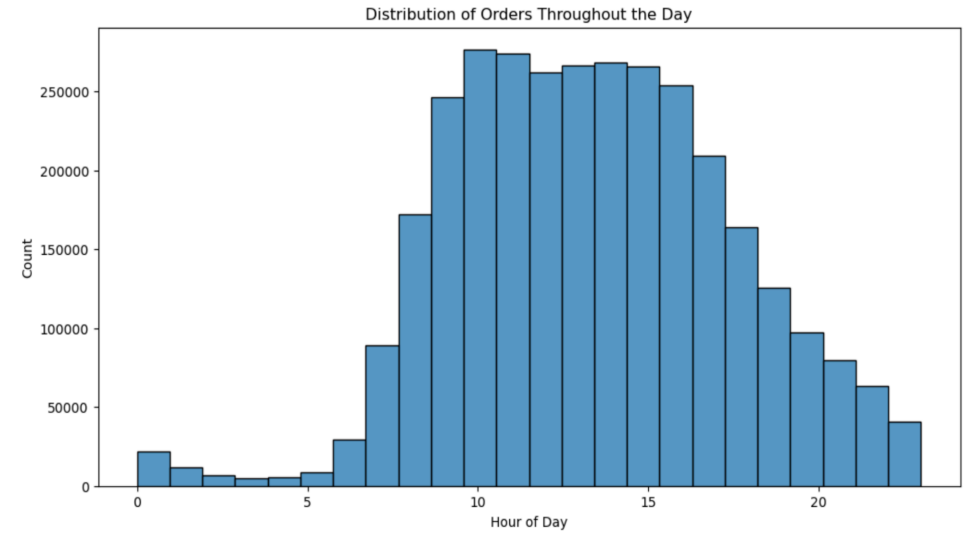
- Within week if days since prior order is less than or equal to 7
- Within fortnight if days since prior order is less than or equal to 14
- Within a month if days since prior order is less than or equal to 30
- More than a month if days since prior order are more than 30



The number of orders based on the days_since_order column. The orders within a week that are less than 7 days are highest. Meaning customers tend to place repeat orders quickly and are highly engaged. The rate of activity starts to slow down after a week

As the time period increases the number of orders drops sharply indicating less engagement and customers are less likely to make a repeat purchase.

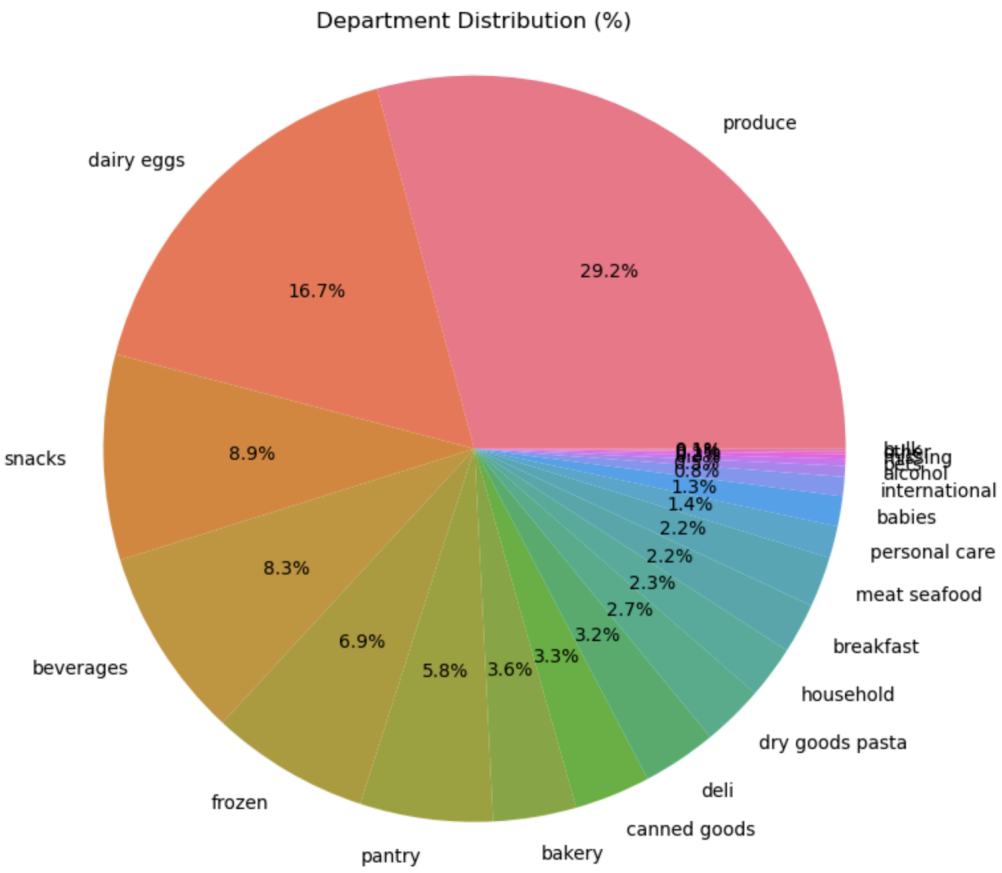
7. Other univariate distributions:



As time increases the order count increases. It reaches its peak at 10 then slowly decreases after 15.

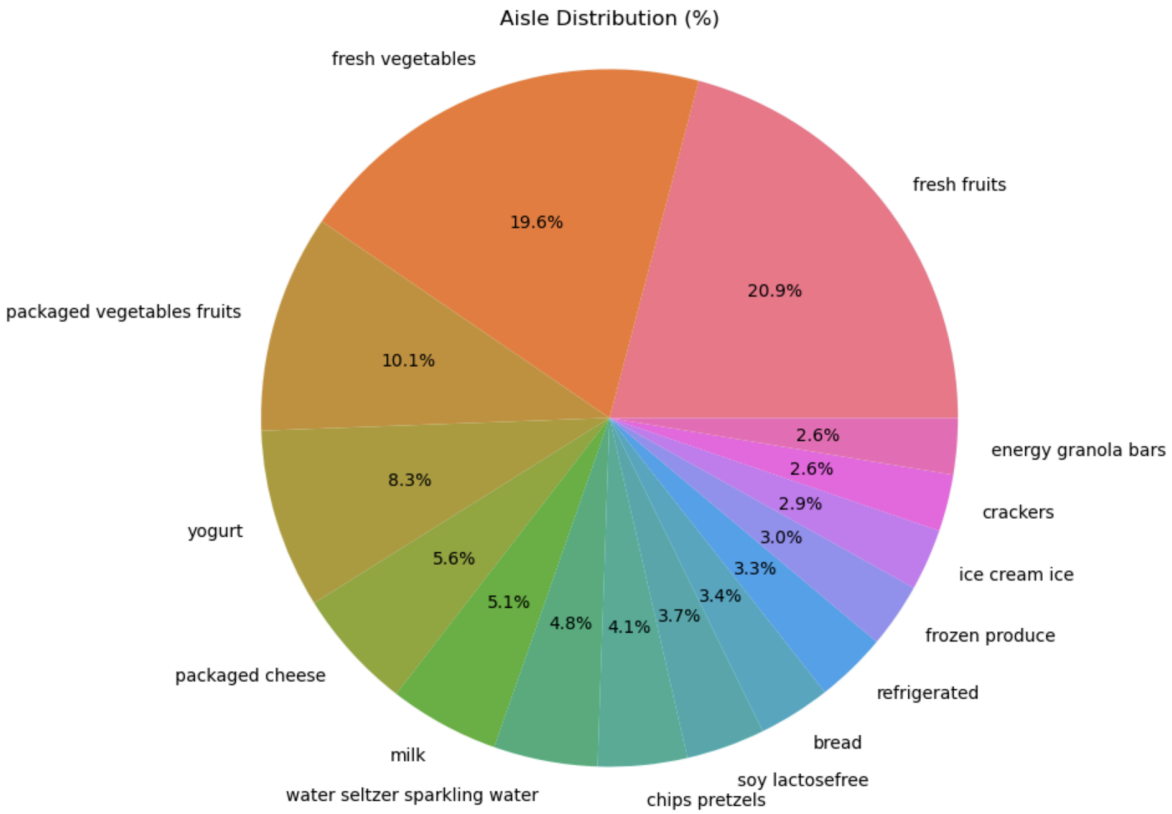
The plot shows a bimodal distribution with two distinct peaks. This indicates that there are two primary time periods during the day when orders are placed most frequently.

A strong emphasis should be placed during the two time periods since there is a demand for products.



A pie chart showcasing the distribution of the department with produce occupying around 30% of the data in the department. The high percentage for ‘Produce’ suggests a focus on healthy eating and fresh ingredients among customers.

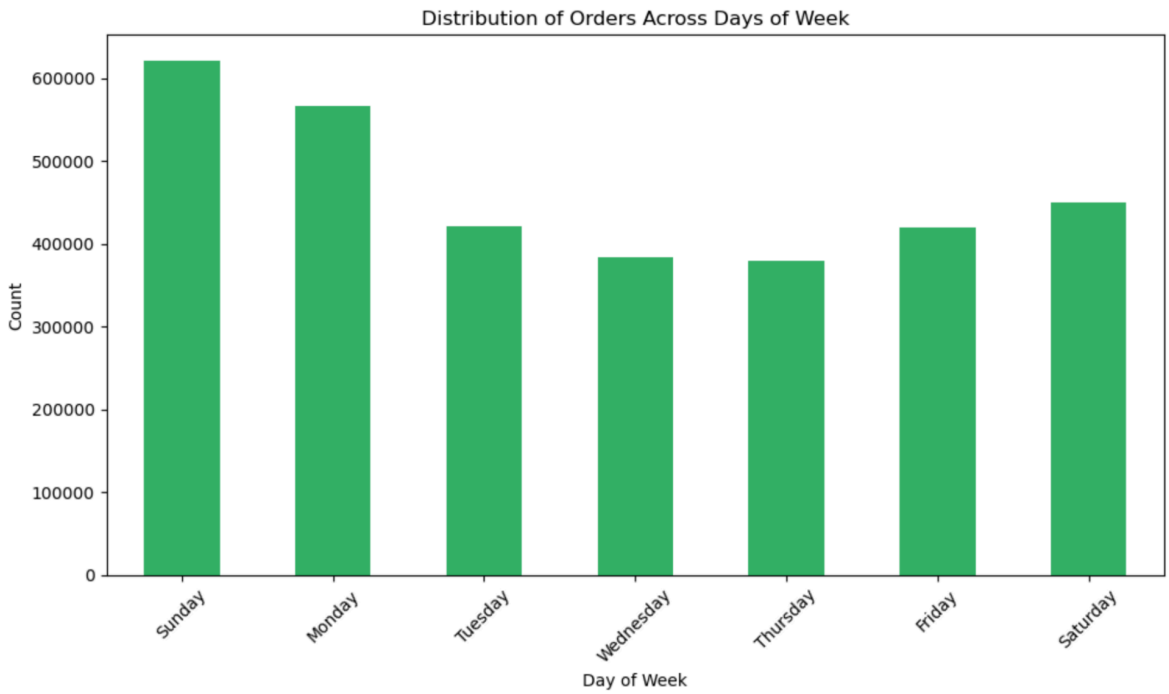
The ‘Dairy Eggs’ department follows closely with 16.7%, highlighting the importance of dairy and egg-based products in customer shopping lists.



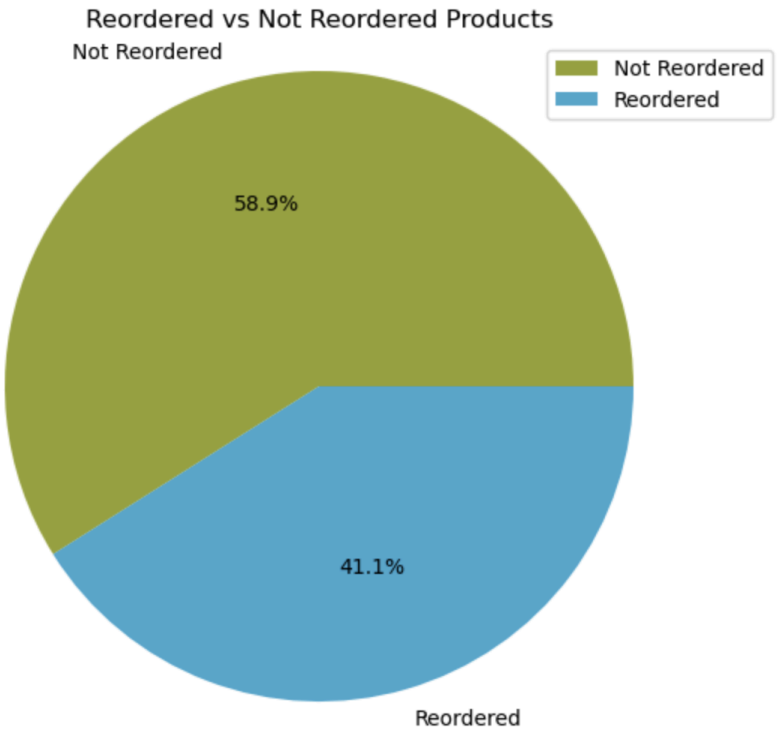
A pie chart showcasing the distribution of aisle in the dataset. Fresh fruits and fresh vegetables are the most popular aisles, accounting for 20.9% and 19.6% of total orders respectively. This indicates a strong preference for fresh produce among online shoppers.

Packaged vegetables and fruits occupy the third spot with 10.1%, followed by yogurt (8.3%). This suggests a significant demand for pre-packaged and convenient food items.

Given the high demand for fresh fruits and vegetables, it is crucial to ensure a consistent supply of high-quality produce as the number of health conscious customers are significantly higher.



This shows the count of orders across each day of week. Sunday being the day with the highest order count. The count decreases a lot after monday and increases again on saturday forming a cycle . Meaning that customers are more likely to order during the weekends.



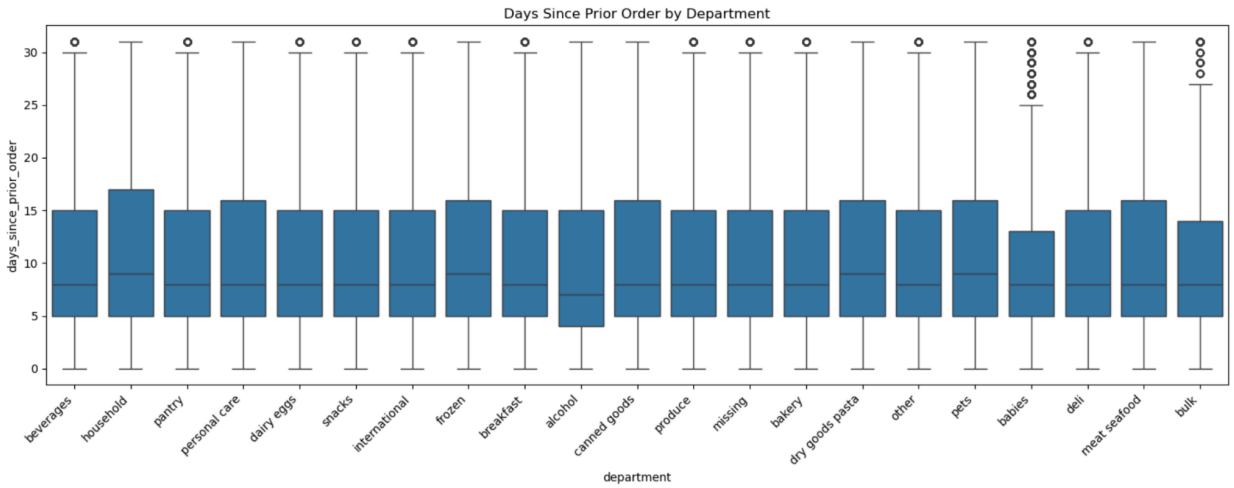
The pie chart showing the % of reordered and ordered. The % of customers who didn't reorder are more. Meaning the majority of products (nearly 59%) are not reordered by customers.

This implies that since the customers may not be properly involved, client retention and repeat business may be areas where the company might improve.

9. Bivariate analysis

reordered	0	1
department		
alcohol	6600	8748
babies	17772	24618
bakery	43856	73809
beverages	93426	175570
breakfast	31134	39830
bulk	1475	1981
canned goods	58082	48708
dairy eggs	178393	362994
deli	41136	63958
dry goods pasta	46656	39997
frozen	102519	121126
household	44314	29535
international	16887	10040
meat seafood	30790	40103
missing	4098	2771
other	2173	1453
pantry	122720	64812
personal care	30232	14457
pets	3922	5834
produce	332233	615688
snacks	123153	165544

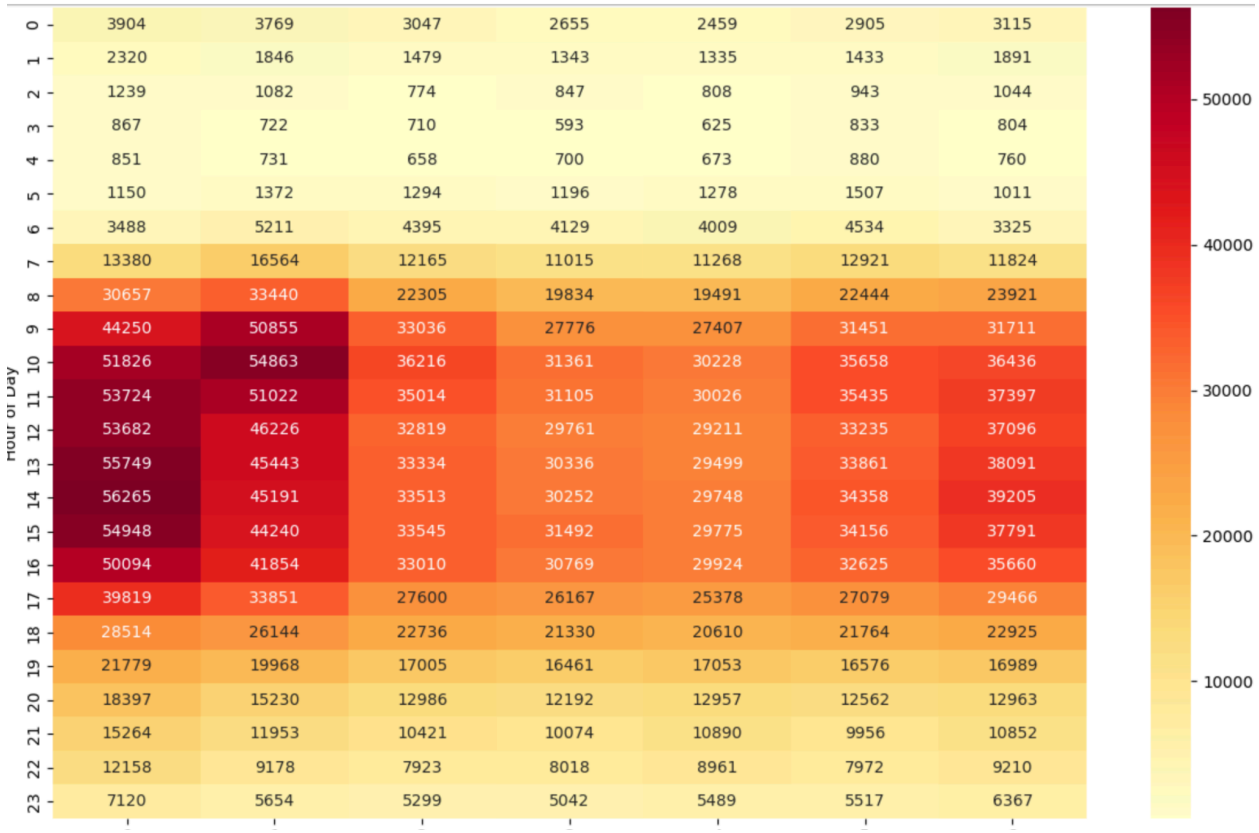
A contingency table between reordered and department showcasing which products from department are frequently reordered and which are not. The most reorders are from the produce category suggesting the demand for fresh items like fruits and vegetables in this category.



The box plot shows that there is significant variation in the typical time between orders across the different departments. Some departments show much shorter intervals between repeat purchases compared to others.

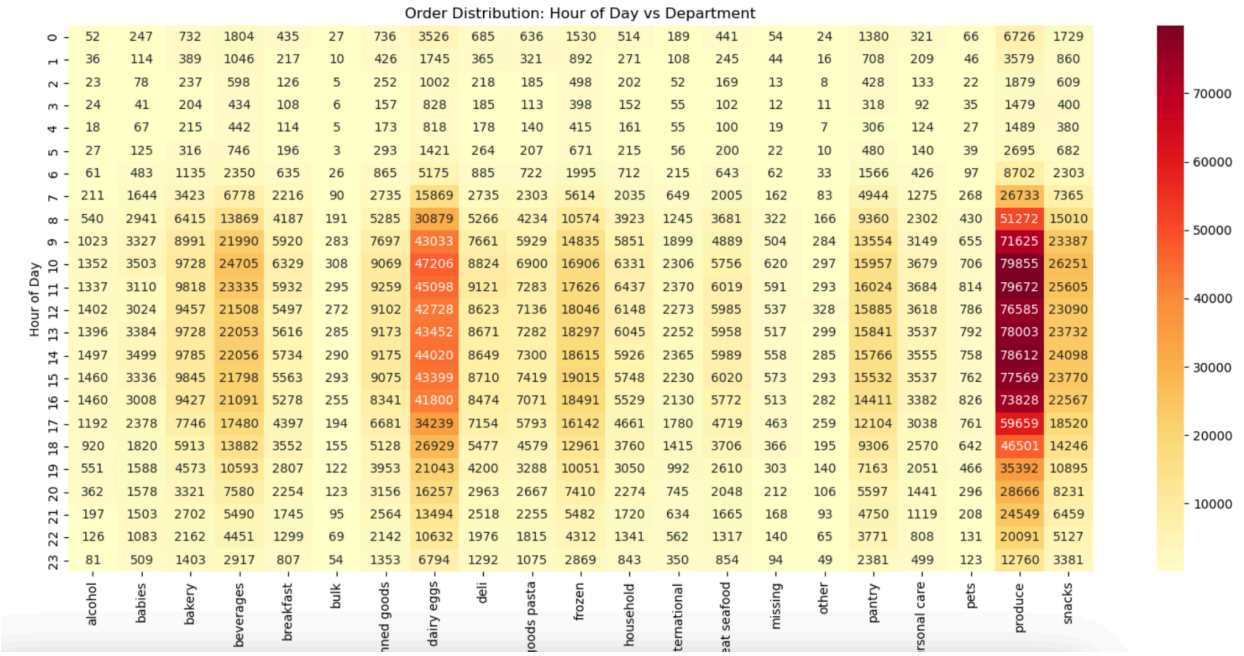
Several departments have notable outliers or extreme values, indicating some customers in those departments have very long gaps between orders compared to the typical customer.

Departments with shorter median and interquartile range values (e.g. "beverages", "household") appear to have more active and consistent repeat purchasing behavior compared to departments with longer intervals (e.g. "beauty", "pets").



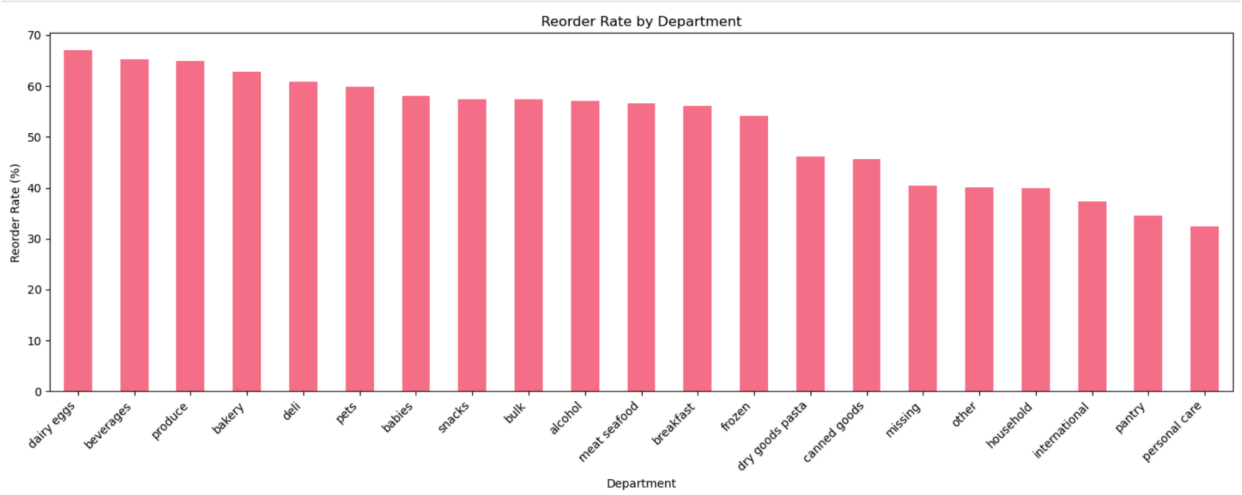
A heatmap between order hour and day of week showcasing the time period where the orders were maximum. Sunday and Monday were the days when customers were most active. On Sunday, from 10 to 16 and Monday from 9 to 11 the orders were maximum. This indicates that customers prefer to do their shopping during weekends.

Weekdays on the other hand have a more spread out distribution throughout the day with slight peaks during morning and afternoon.



A heatmap between department and hours of day showcasing from which department and hour of day the order was the highest. Items from produce and from 9 to 16 are the highest followed by dairy eggs. Meaning customers ordering fresh products during that time are higher.

Snacks and beverages also have a peak demand from 9 to 16 could be as a complement to after lunch.
Delivery optimization, scheduling more staff during peak hours and implementing target marketing might further enhance the customer satisfaction and engagement leading to more sales.

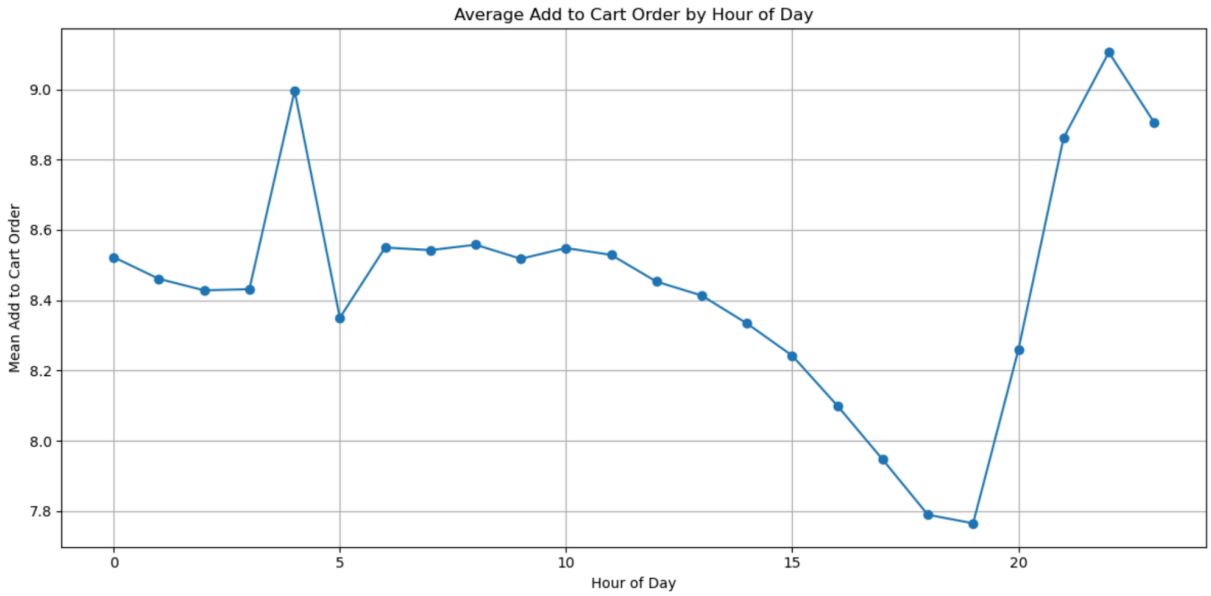


A barplot showcasing the reordered rate by department.

High Reorder Departments:

- Dairy Eggs: The high reorder rate for dairy and eggs suggests that these are essential items frequently purchased.
- Beverages**:** The high reorder rate for beverages indicates a strong demand for drinks, especially for daily consumption.
- Produce**:** The high reorder rate for produce reflects the frequent purchase of fresh fruits and vegetables.

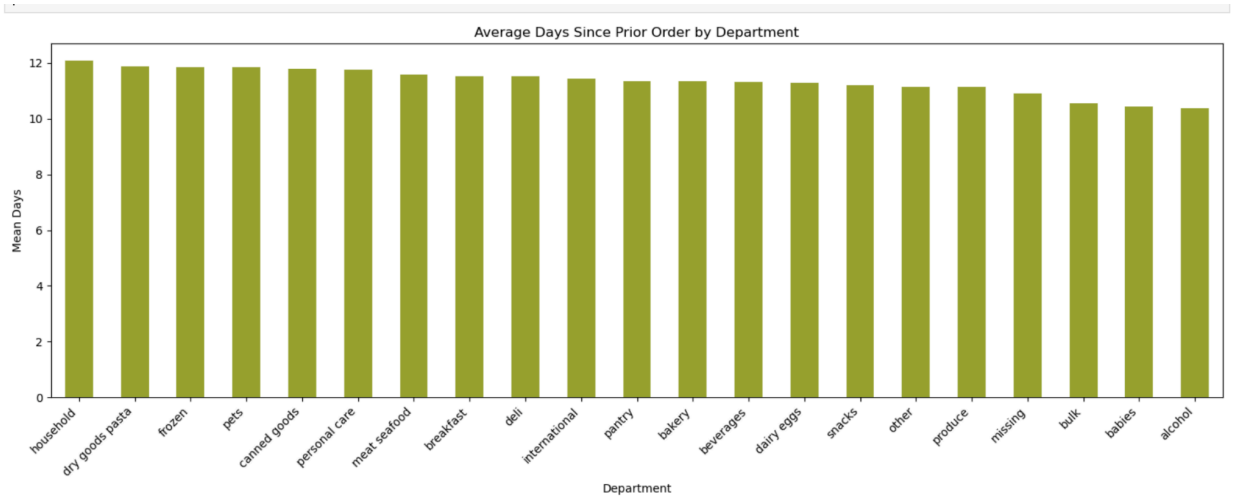
Targeted promotions and personalized recommendations can help to encourage repeated purchases and boost sales.



A line plot of the mean of add_to_cart and hour of day showcasing the average add to cart order by hour of day. The graph has a peak before reaching 5 and then decreases. It goes up again before reaching 20.

This pattern suggests that the business experiences two primary shopping periods during the day - one in the morning and another in the afternoon.

Understanding these trends can help the business optimize their operations, marketing, and product availability to better serve their customers throughout the day.



The bar chart reveals a relatively consistent pattern in the average days since the prior order across most departments.

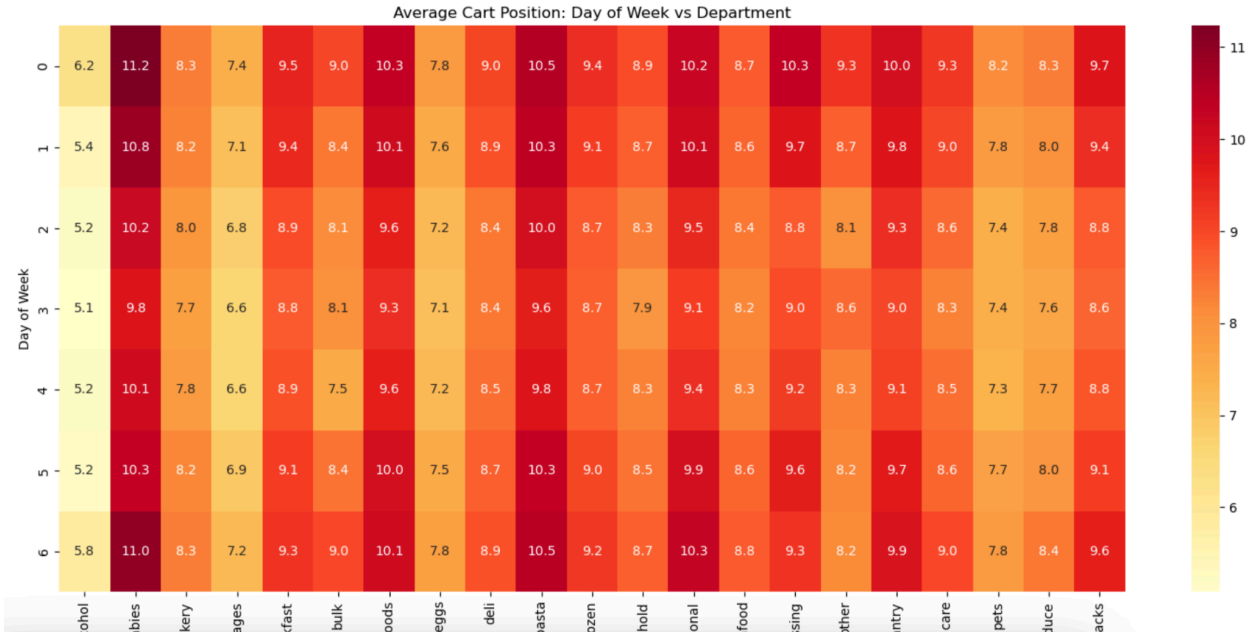
Short Reorder Cycles:

- Household**:** The short average days since the prior order for household items suggest frequent replenishment of essential items.
- Dry goods pasta**:** Similar to household items, dry goods and pasta are frequently reordered.

Longer Reorder Cycles:

- Alcohol**:** The longer average days since the prior order for alcohol might be due to less frequent consumption or specific purchasing occasions.
- Babies**:** The longer reorder cycle for baby products could be attributed to the specific needs and developmental stages of infants.

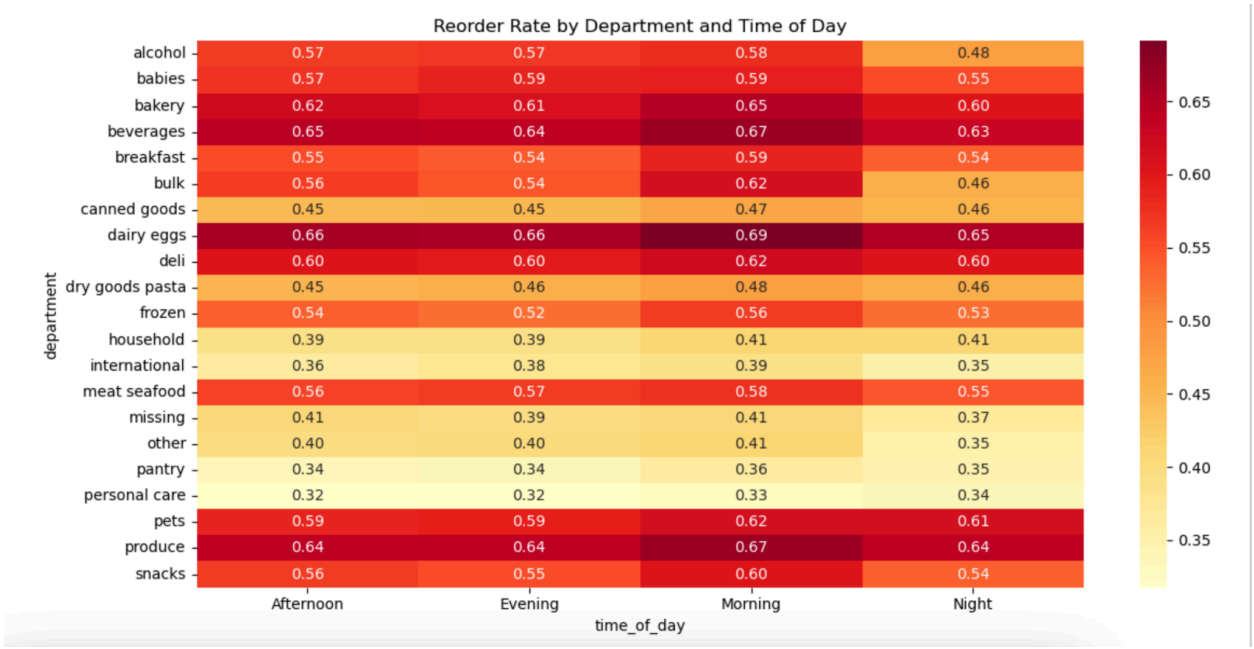
Targeted promotions or discounts for departments with longer reorder cycles might help to stimulate demand and shorten the time between purchases.



A heatmap between days of week and department showcasing how many items from the department were added and at which time of the day.

‘Produce’ and ‘Bakery’ tend to be placed higher in the cart, suggesting that they are frequently added to the cart early in the shopping process.

Analyzing cart position data by customer segment might help to identify preferences and tailor product recommendations.

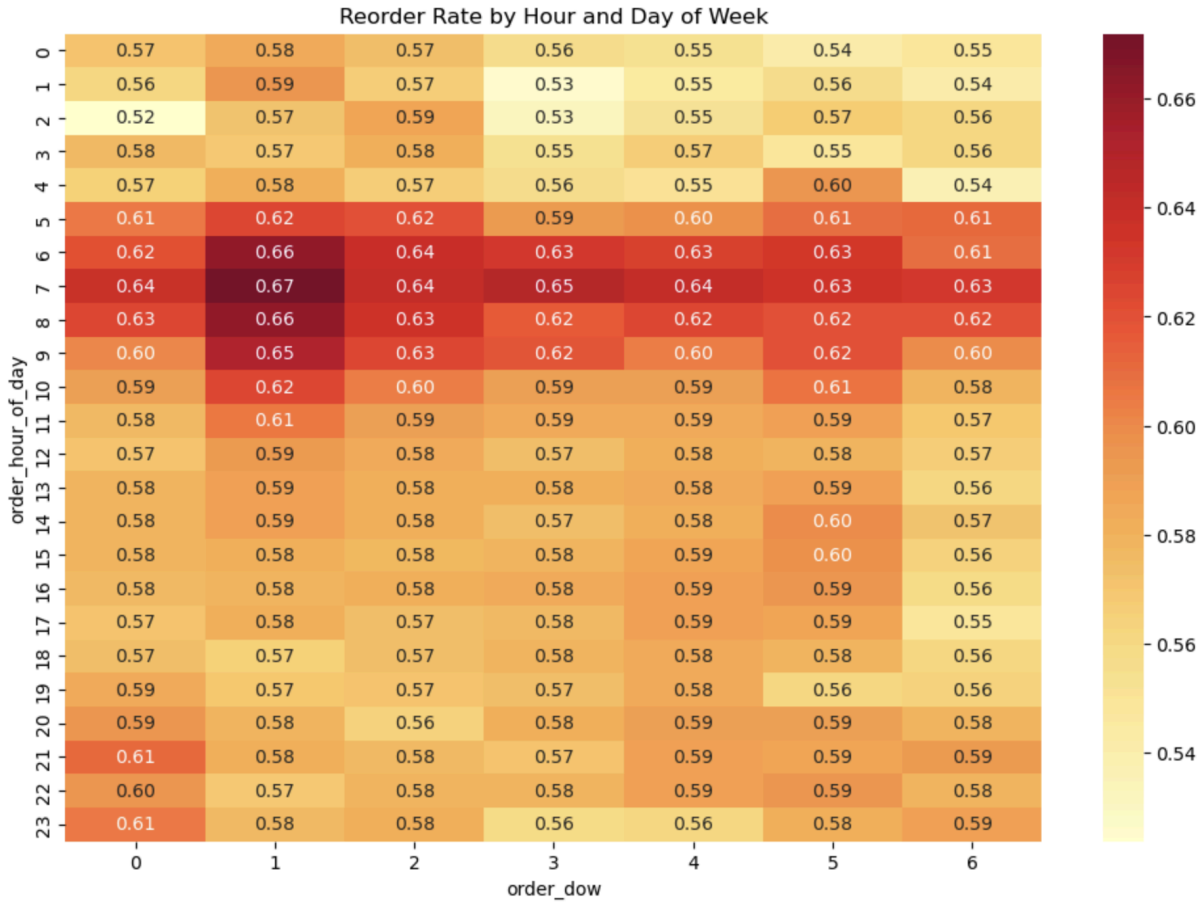


A heatmap between department and time_of_day showcasing the time period when the customers reordered the most and from which department. The ‘Afternoon’ period has slightly higher reorder rates compared to other times.

High Reorder Departments:

- Produce: The consistently high reorder rate for produce indicates that customers frequently purchase fresh fruits and vegetables.
- Bakery**:** The high reorder rate for bakery items suggests a strong demand for fresh bread, pastries, and other baked goods.
- Dairy Eggs**:** High reorder rates for dairy and eggs reflect their essential role in daily diets.

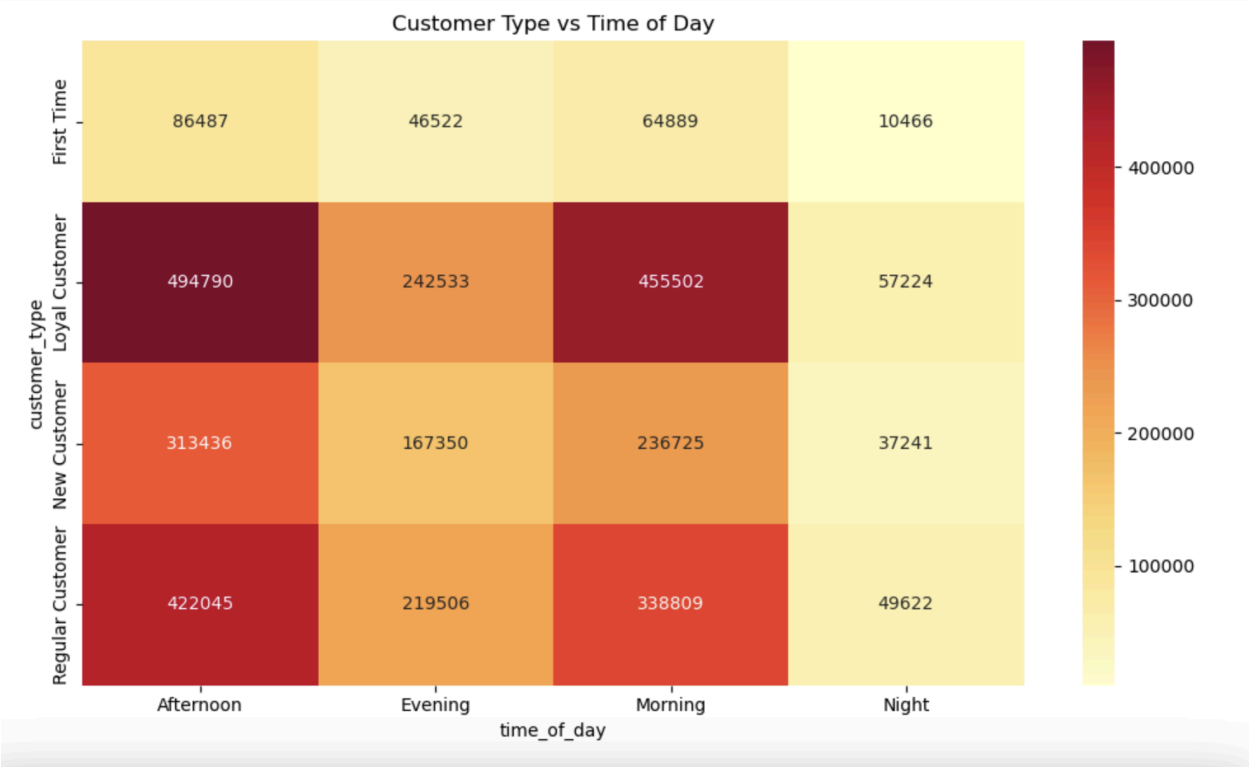
Segmenting customers based on their reorder behavior might be a good way to tailor marketing campaigns and product recommendations to further increase sales.



A heatmap between order day of week and order hour of day showcasing that on which day of the week and hour of day the reorder rate was high.

The reorder rates are higher during afternoon and on weekends indicating weekly shopping.

Targeted promotions or discounts during off peak hours might see a demand and thus a spike in reorder rates.

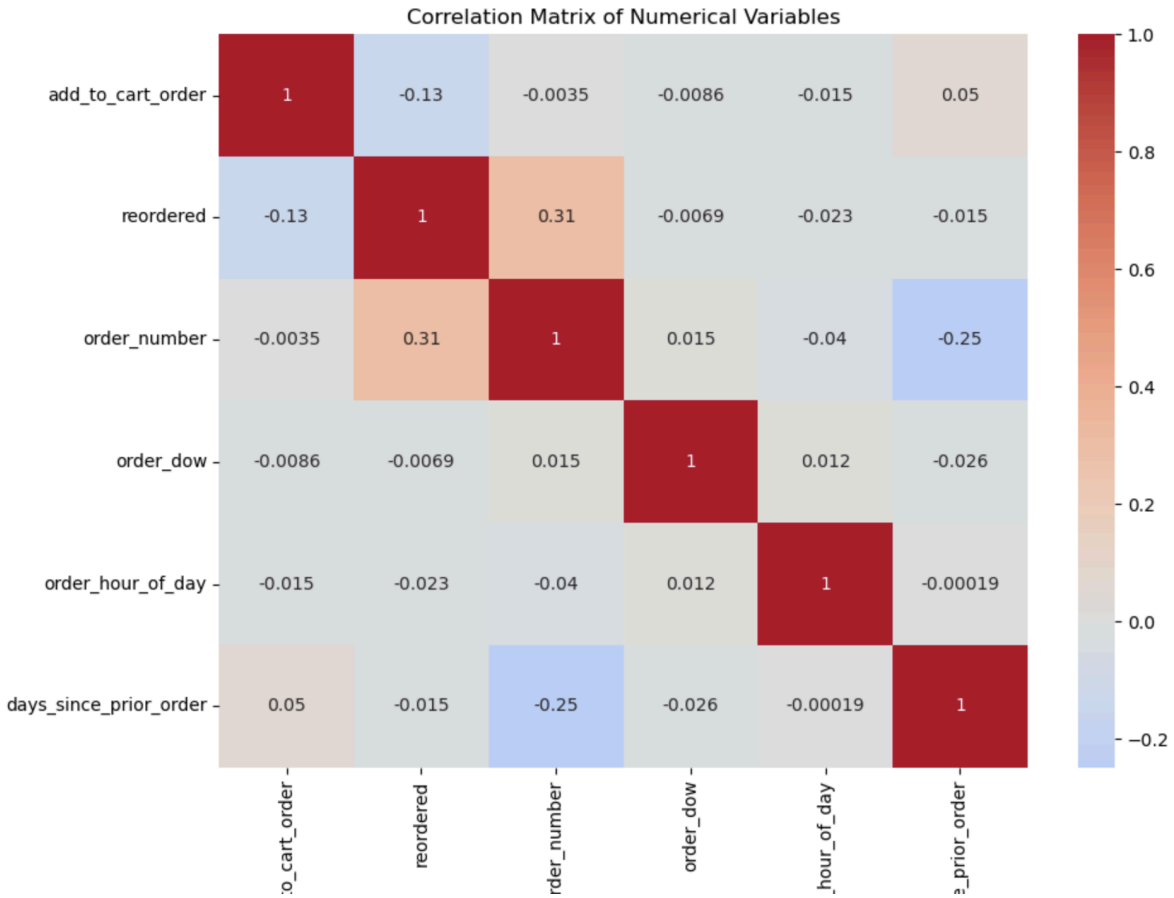


A heatmap between ‘customer_type’ and ‘time_of_day’ showcasing that loyal customers (who are those whose order_number are more than 15) are more active during afternoon and morning. Time from 5 to 11 is morning and from 12 to 16 is afternoon.

We can hence conclude loyal customers consistently generate the highest number of orders across all time periods. This indicates that customer loyalty is a key driver of sales volume.

First time customers' volume is very low compared to other customers. This highlights the importance of customer retention strategies to convert first-time customers into repeat customers.

10. Multivariate analysis



- The heatmap of the numerical columns showcasing absence of high correlation between the groups. ‘Order number’ and ‘reordered 'have the highest positive correlation with 0.31 score indicating that there is a direct relationship between the two groups.

- ‘Order number’ and ‘days since prior order’ are negatively correlated with -0.25 score indicating that there is an indirect relationship between the two groups.

11. Statistical tests

```
ANOVA Test: Days Since Prior Order across Departments:
Statistic: 165.0116
P-value: 0.0000
Significant at 0.05 level: True

Kruskal-Wallis H-test: Days Since Prior Order across Departments:
Statistic: 2926.5895
P-value: 0.0000
Significant at 0.05 level: True

Chi-square test results for Department vs Reordered:
Chi-square statistic: 127520.84
p-value: 0.0000000000
Significant at 0.05 level: True
```

The tests ANOVA & Kruskal-Wallis Results:

- Both tests show significant differences in **"**Days Since Prior Order**"** across departments.

Chi-square Test:

- The Chi-square test shows that department is significantly related to reordering behavior.

After performing Post hoc test analysis:

Significant Differences:

- Alcohol vs Bakery: The mean difference is 0.9661, with a p-value of 0.000, indicating a significant difference.
- Alcohol vs Canned Goods: The mean difference is 1.4183, with a p-value of 0.000, indicating a significant difference.
- Babies vs Household: The mean difference is 1.6566, with a p-value of 0.000, showing a significant difference.
- Bakery vs Produce: The mean difference is -0.2129, with a p-value of 0.0, indicating a significant difference.