Background

Our project was for Green Cart Ltd., a growing UK-based e-commerce company focused on eco-friendly household products. The company is preparing for its Q2 performance review, and we were asked you to investigate sales and customer behaviour across regions and product lines. This data-set, after cleaning and wrangling, was left with 3,004 transaction records, combining order, customer, product, and delivery information. The information in the data is structured to provide sales and customers behaviour insight. To determines which product drives the most revenue and in which region, how discount allowed and loyalty programme could contribute to sales. The data also provided insight on how delivery delay contributes to the business logistic challenges and how different group of customers relates or behaves towards different programme provided to improve sales. Each row represents an individual order line, including details such as order ID, customer and product identifiers, quantity, unit and base prices, revenue, discounts, and price bands. It also captures product metadata such as category, launch date, supplier.

Data Cleaning and Future Engineering Summary

As a first rule of thumb in handling raw data, cleaning of the data-set was very crucial before any other actions can be performed with the data-set. To achieve this, different cleaning and future engineering which are listed below was applied so as to make the data ready for EDA.

- The first approach was convert signup_date, order_date and launch_date to DateTime using DateTime query

- Fill missing values in categorical data column where the dtype is object with “Unknown” and filling empty numeric data columns with “0”, using query.

- We dropped columns only in sales and customer data-set, intrapolate and dropped duplicate values in all the data-set

- Standardization of all the the data-set. The essence of this is to have the character in one particular order which “title” was a preferred option for me. This enables the data to be well synchronized

- Filling the NaN values in discount_applied column with “0”, the price column with “mean,” date column with “mode” value and other categorical data columns with “Unknown”

- Quantity column was mapped, and quantity, unit price and discount_applied columns showed false when checked for non-negative value

- New column that were required for data insight was generated. Such columns as revenue (using: quantity, unit_price and discount applied), order_week (using: ISO calender), days_to_order which is the date between order_date and launch_date, is_late column which would help in generating insights for delayed delivery etc.

- Some columns ("order_id", "product_id", "product_name","launch_date", "supplier_code", "email", "gender", "region_y) which I found not necessary for the insight that I want to generate were dropped at this point for the clarity of information sake.

Key Findings and Trends

- Based on product category performance, with an average discount of 0.086, total quantity of 3595, Cleaning generated the highest revenue of 93836.18 pounds and the least on the leather are some unknown items that generated 610.65 pounds as revenue. A closer look at figure 1.0 below, would give a clear understanding.

```
print(category_performance)
```

```
...                Total_Revenue  Total_Quantity  Avg_Discount_Applied
category
Cleaning           93836.177000          3595.0              0.085626
Storage            47037.747500          1733.0              0.080763
Outdoors           40202.297588          1525.0              0.082087
Kitchen            33993.041500          1229.0              0.075558
Personal Care      24965.356500           906.0              0.086184
Unknown              610.656500            22.0              0.150000
```

Figure 1.0

- The line-chart below shows that revenue was at a spike during the early weeks and last weeks of the business calender with the Northern region on the highest at the early weeks and Southern region at the highest at the last weeks. Apart from this spike period mentioned, the regions recorded a horizontal sales output and a downward skewed towards the end of the sales period

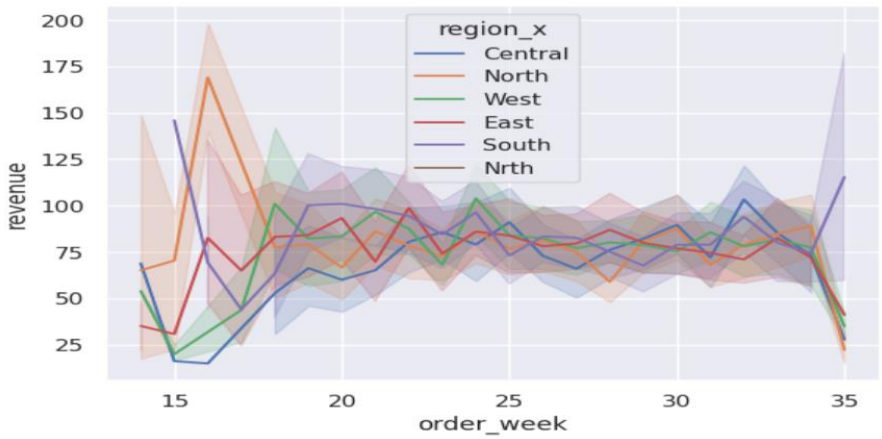

Figure 2.0

Response to Business Questions

1. **Insight on product categories that drives the most revenue, and in which regions**

According to the bar-plot attached as figure 3.0, the cleaning category generated the highest total revenue. This is also supported by the information in the attached table below as figure 4.0 which also portrayed that the northern region generates the highest revenue from this category. To support the output is a one time view on the line-plot with an upward spike in the early weeks

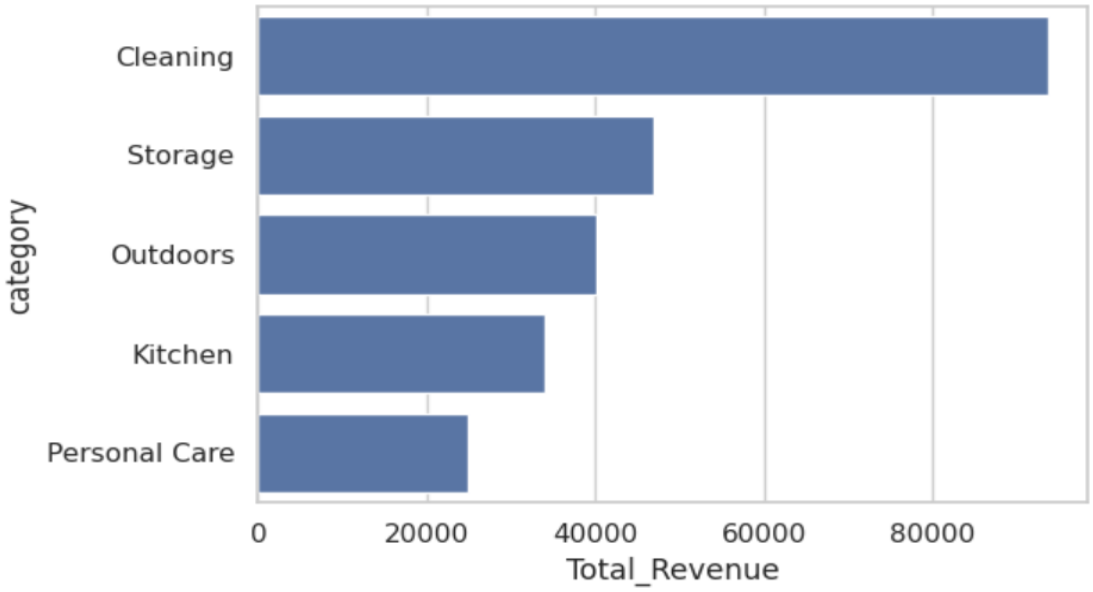

Figure 3.0

category	Total_Revenue	Total_Quantity	Avg_Discount_Applied	Region
Cleaning	93836.177000	3595.0	0.085626	North
Storage	47037.747500	1733.0	0.080763	Central
Outdoors	40202.297588	1525.0	0.082087	West
Kitchen	33993.041500	1229.0	0.075558	South
Personal Care	24965.356500	906.0	0.086184	West
Unknown	610.656500	22.0	0.150000	West

Figure 4.0

2. How discounts lead to more items sold

Based on the available information on the category performance, We were able to deduce that that discount applied doesn't lead to increase in sales. Looking at the above figure 4.0, the average discount applied to all category would produce a horizontal line if plotted on a chart but sale and revenue would produce a linear shaped graph, which shows that there is no correlation between the two variables. This can be clearly seen in the correlation plot below in figure 5.0

Figure 5.0

3. Loyalty Tier that Generates the Most Value

Gold-tier customers according to figure 6.0 accounts for the largest share of revenue, contributing approximately 57% of the total, making them the most valuable customer segment. Silver and Bronze tiers contribute comparable portions, at around 22% and 20% respectively, indicating steady but lower overall value than Gold customers.

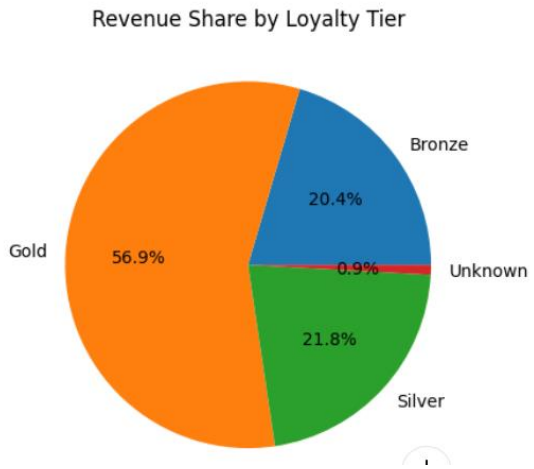

Figure 6.0

4. Are certain regions struggling with delivery delay?

According to the available the data as can be seen in figure 7.0, the West is said to performs slightly better but still has a high delay rate (≈37%). Hence, it can not be said that the West is struggling with delivery delay if the marginal rate is to be calculated.

region_x	price_band	Total_Orders	Avg_Days_to_Order	Late_Delivery_Rate
North	Medium	205	201.965854	0.443902
East	Medium	239	219.451883	0.422594
South	High	297	208.619529	0.420875
East	Low	96	200.770833	0.416667
	High	267	212.003745	0.408240
Central	High	265	202.845283	0.392453
	Medium	225	206.648889	0.391111
West	Low	77	211.883117	0.389610
North	Low	107	210.028037	0.383178
West	High	270	212.377778	0.381481
Central	Low	115	216.400000	0.373913
North	High	293	205.358362	0.361775
South	Medium	201	210.771144	0.358209
West	Medium	248	209.568548	0.350806
South	Low	98	216.448980	0.336735
Nrth	Medium	1	257.000000	0.000000
	Low	0	NaN	NaN
	High	0	NaN	NaN

Figure 7.0

5. Do customer signup patterns influence purchasing activity?

Customer signup timing clearly impacts purchasing behavior. Earlier signup cohorts are significantly more valuable, indicating strong lifetime value effects.This is also explained in the line-plot in figure 2.0 where most of the regions experienced an upward skew from the beginning and sudden downward spike towards the end of the business period. This means the early cohort contributed more to the revenue generated considering the

signup time. It was also x-rayed in figure 6.0 that different tiers of customers contribute differently to the total revenue of the company.

6. Recommendations

- Sequel to the high revenue recorded in the Northern region in the cleaning category, the company prioritize inventory availability and marketing spend for Cleaning products in the Northern region and make effort to introduce region-specific campaigns. The reason is for the company to take advantage of proven demands in different regions

- Loyalty programme has introduce more revenue than the discount programme. The company should shift more towards the loyalty driven incentives, reduce general discounting and redirect value into loyalty benefits.