Sales & Customer Behaviour Insights

1. Business Scenario

Green Cart Ltd., a growing UK-based e-commerce company focused on eco-friendly household products is preparing for its Q2 performance review. The Data & Insights team's manager asked for investigations on sales and customer behaviour across regions and product lines. The findings will serve upcoming marketing and operational strategies.

They were particularly interested in:

- Cleaning and merging the data;
- Creating new features;
- Analysing patterns and performance;
- Presenting insights using charts.

2. Datasets

Three datasets were provided for this analysis:

- **sales_data.csv** – Contains information about past transactions and has the following columns: *order_id*, *customer_id*, *product_id*, *quantity*, *unit_price*, *order_date*, *delivery_status*, *payment_method*, *region*, and *discount_applied*;

- **product_info.csv** – Contains the basic identifiers for each product and has the columns *product_id*, *product_name*, *category*, *launch_date*, *base_price*, and *supplier_code*;

- **customer_info.csv** – Contains identifiers for each user, such as *customer_id*, *email*, *signup_date*, *gender*, *region*, and *loyalty_tier* (can be 'Bronze', 'Silver', or 'Gold').

3. Data Cleaning Summary

To ensure accurate and relevant analysis, the data must be clean. Data that hasn't been cleaned properly can result in biased or even wrong results. The following data cleaning processes were applied:

- **Data type conversion** – The date columns (*order_date*, *launch_date*, and *signup_date*) type was changed from 'object' to 'datetime'. This ensures flexibility when it comes to analysis operations.

- **Standardisation of text values** – Having too many formats of the same value in the datasets can cause confusion and lead to wrong analysis results. The following values were standardised:
  - 'three' in *quantity* -> '3';
  - 'five' in *quantity* -> '5';
  - '   DELAYED' and 'delyd' in *delivery_status* -> 'Delayed';
  - 'delivered' and 'delrd' in *delivery_status* -> 'Delivered';
  - ' Cancelled' in *delivery_status* -> 'Cancelled';
  - 'credit card' in *payment_method* -> 'Credit Card';
  - 'bank transfr' in *payment_method* -> 'Bank Transfer';
  - 'nrth' in *region* -> 'North';
  - 'male' in *gender* -> 'Male';
  - 'FEMALE' and 'femle' in *gender* -> 'Female';
  - ' gold', 'GOLD', and 'gld' in *loyalty_tier* -> 'Gold';
  - 'bronze' and 'brnze' in *loyalty_tier* -> 'Bronze';
  - 'sllver' in *loyalty_tier* -> 'Silver';

```
Check the unique values in the sales_data dataset

Unique values in the quantity column:  ['3' '5' '1' '2' '4' nan 'three' 'five']
Unique values in the delivery_status column:  ['Delivered' ' DELAYED' 'delivered' ' Cancelled ' 'Delayed' 'delrd'
 'delyd' nan]
Unique values in the payment_method column:  ['PayPal' 'credit card' 'Bank Transfer' 'Credit Card' nan 'bank transfr']
Unique values in the region column:  ['Central' 'North' 'West' 'East' 'South' 'nrth']

Check the unique values in the product_info dataset
Unique values in the category column:  ['Storage' 'Cleaning' 'Kitchen' 'Personal Care' 'Outdoors']

Check the unique values in the customer_info dataset

Unique values in the gender column:  ['Male' 'Female' 'male' 'FEMALE' 'Other' 'femle' nan]
Unique values in the region column:  ['Central' 'West' 'North' 'South' 'East' nan]
Unique values in the loyalty_tier column:  ['Silver' ' gold ' 'GOLD' 'bronze' 'gld' nan 'brnze' 'sllver']
```

```
['3' '5' '1' '2' '4' nan]
['Delivered' 'Delayed' 'Cancelled' nan]
['PayPal' 'Credit Card' 'Bank Transfer' nan]
['Central' 'North' 'West' 'East' 'South']
['Male' 'Female' 'Other' nan]
['Silver' 'Gold' 'Bronze' nan]
```

Figure 1: Unique values in each column, before (up) and after (down) standardisation

- **Numeric columns validation** – The *quantity* column's type was changed from 'object' to 'numeric' to ensure calculations are possible. A check to make sure there aren't any negative values was also made.

- **Duplicate rows handling** – There were two duplicate rows in the *sales_data* dataset and one duplicate row in the *customer_info* dataset. All of them were dropped.

- **Missing values handling** – There were 538 missing values in the *sales_data* dataset and 19 missing values in the *customer_info* dataset. They were handled in the following way:
    - ➢ The rows with missing values for identifiers and columns that are crucial for analysis (*order_id*, *customer_id*, *product_id*, *order_date*, *delivery_status*) were dropped;

- The missing values in the *payment_method* column were filled with ‘Unknown’. This column wasn’t needed for the analysis report, so it was not worth dropping the rows;

- The missing values in the *discount_applied* column were filled with ‘0.0’. This column had the most missing values (517) and it was most probably because a discount simply was not applied.

- Missing values in the categorical columns (*quantity*, *gender*, *region*, *loyalty_tier*) were imputed with mode values.

- Missing values in the numeric column *unit_price* were filled with the median value.

4. Datasets merging

After ensuring the data was properly cleaned, the datasets were merged in order to perform feature engineering and analysis later on.

The *sales_data* and *product_info* datasets were merged on *product_id*, then the result was merged with *customer_info* on *customer_id*. In order to preserve all sales transactions, a left type of join was used.

Since both the *sales_data* and *customer_info* datasets had columns named ‘region’, the merged dataset kept both regions, implicitly assigning them as ‘region_x’ (*sales_data*) and ‘region_y’ (*customer_info*). To avoid confusion, their names were changed to ‘sales_region’ and ‘customer_region’.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2982 entries, 0 to 2981
Data columns (total 20 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   order_id          2982 non-null   object
 1   customer_id       2982 non-null   object
 2   product_id        2982 non-null   object
 3   quantity          2982 non-null   float64
 4   unit_price        2982 non-null   float64
 5   order_date        2982 non-null   datetime64[ns]
 6   delivery_status   2982 non-null   object
 7   payment_method    2982 non-null   object
 8   sales_region      2982 non-null   object
 9   discount_applied  2982 non-null   float64
 10  product_name      2982 non-null   object
 11  category          2982 non-null   object
 12  launch_date       2982 non-null   datetime64[ns]
 13  base_price        2982 non-null   float64
 14  supplier_code     2982 non-null   object
 15  email             2900 non-null   object
 16  signup_date       2900 non-null   datetime64[ns]
 17  gender            2900 non-null   object
 18  customer_region   2900 non-null   object
 19  loyalty_tier      2900 non-null   object
```

Figure 2: The merged dataset's columns and structure

5. Feature Engineering

The following new columns were created:

- *revenue* = quantity × unit_price × (1 - discount_applied)
- *order_week* = ISO week from order_date
- *price_band* = Categorise unit price as Low (<£15), Medium (£15–30), High (>£30)
- *days_to_order* = Days between launch_date and order_date
- *email_domain* = Extract domain from email
- *is_late* = True if delivery_status is ‘Delayed’

	revenue	order_week	price_band	days_to_order	email_domain	is_late
0	117.750	27	High	275	mills-logan.com	False
1	94.600	27	Medium	169	morgan.com	True
2	25.228	27	Medium	103	walters-smith.com	False
3	26.208	27	High	356	gmail.com	False
4	38.096	27	High	136	hotmail.com	True

Figure 3: The new columns after feature engineering

6. Key Findings & Trends

Cleaning products generated the highest revenue overall, with a significant lead over all other categories. This trend was consistent across regions, making *Cleaning* the top-performing category in both volume and value.

	category	revenue	quantity	discount_applied
0	Cleaning	93276.3445	3575.0	0.085684
1	Kitchen	33933.6760	1226.0	0.075622
2	Outdoors	40116.0260	1519.0	0.082016
3	Personal Care	24892.2765	900.0	0.087043
4	Storage	46762.3675	1726.0	0.080783

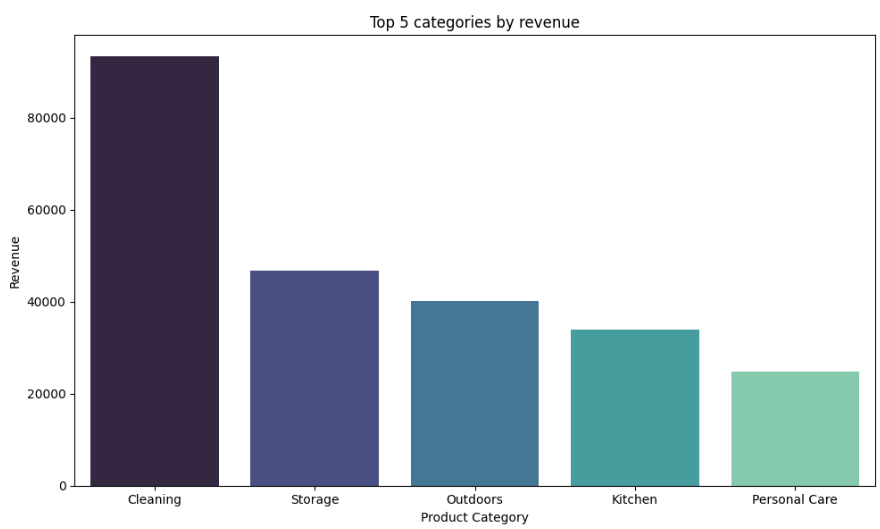

Figure 4: Revenue generated by each category

Discounts had little to no effect on quantity sold, indicating that promotional pricing did not significantly influence purchasing behavior. Most customers purchased regardless of discount levels, and high-discount orders did not correspond to higher quantities.

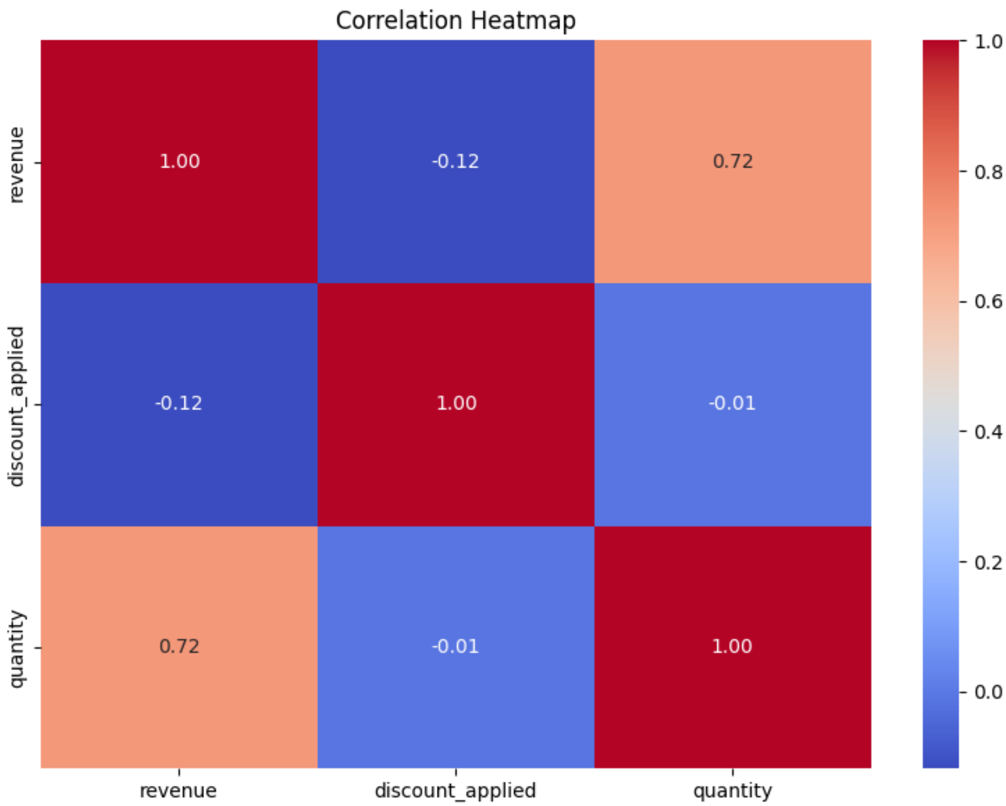

Figure 5: Heatmap of correlation between revenue, discount, and quantity

Gold loyalty tier customers placed the most orders across all regions, contributing the most to total revenue. This highlights the value of retaining and rewarding high-tier loyal customers to drive consistent sales.

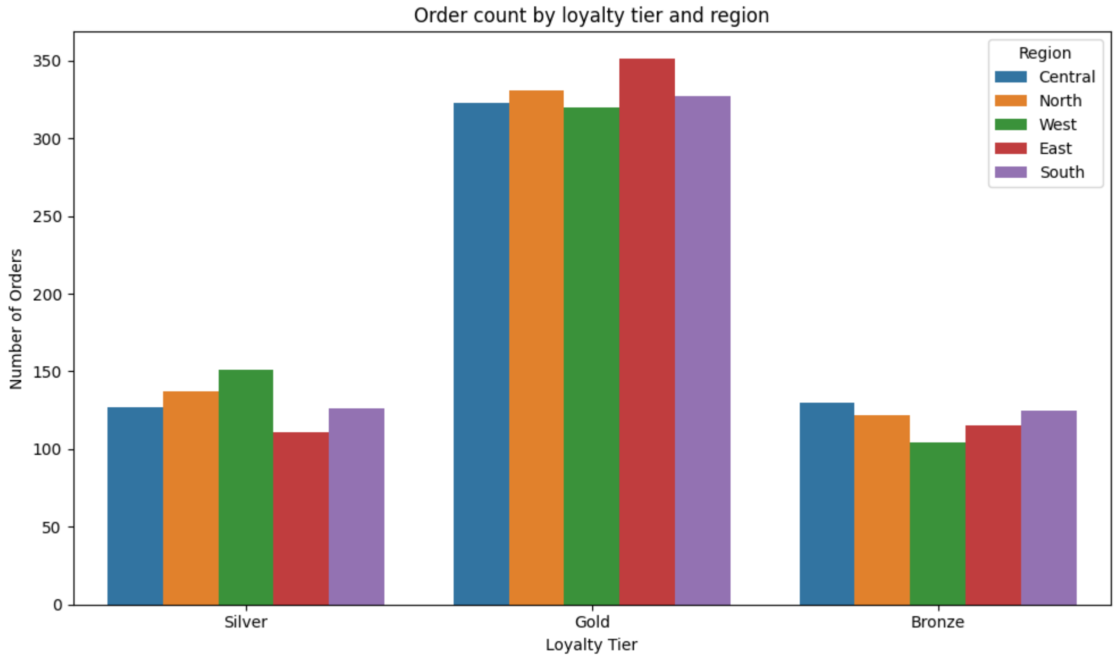

Figure 6: Orders by loyalty tier

There are 29 underperforming products (low quantity, high discount, delayed deliveries).

	product_id	product_name	category
0	P0016	Cleaning Product 53	Cleaning
1	P0002	Cleaning Product 82	Cleaning
2	P0019	Kitchen Product 42	Kitchen
3	P0009	Outdoors Product 13	Outdoors
4	P0029	Cleaning Product 69	Cleaning
5	P0026	Storage Product 50	Storage
6	P0006	Cleaning Product 16	Cleaning
7	P0017	Personal Care Product 11	Personal Care
8	P0020	Cleaning Product 40	Cleaning
9	P0021	Kitchen Product 70	Kitchen
10	P0007	Personal Care Product 64	Personal Care
11	P0027	Outdoors Product 55	Outdoors
12	P0013	Cleaning Product 94	Cleaning
13	P0014	Outdoors Product 91	Outdoors
14	P0005	Personal Care Product 1	Personal Care
15	P0008	Storage Product 47	Storage
16	P0015	Storage Product 10	Storage
17	P0024	Storage Product 87	Storage
18	P0028	Outdoors Product 53	Outdoors
19	P0011	Kitchen Product 53	Kitchen
20	P0004	Kitchen Product 82	Kitchen
21	P0023	Outdoors Product 32	Outdoors
22	P0003	Cleaning Product 85	Cleaning
23	P0018	Storage Product 37	Storage
24	P0010	Cleaning Product 70	Cleaning
25	P0030	Cleaning Product 72	Cleaning
26	P0012	Cleaning Product 29	Cleaning
27	P0022	Cleaning Product 86	Cleaning
28	P0025	Cleaning Product 84	Cleaning
29	P0001	Storage Product 39	Storage

Figure 7: Underperforming products

7. Business Questions & Answers

Question 1: Which product categories drive the most revenue, and in which regions?

Answer: The categories that generate the most revenue are *Cleaning*, *Storage*, and *Outdoors* (Figure 4). The *Cleaning* and *Storage* categories perform well in all regions, while the Outdoors category performs best in the South region.

sales_region	category	
East	Cleaning	19874.3030
South	Cleaning	18979.5760
North	Cleaning	18751.7035
West	Cleaning	17952.9345
Central	Cleaning	17717.8275
West	Storage	10014.4465
East	Storage	9454.3105
South	Storage	9345.3590
	Outdoors	9327.4165
Central	Storage	9233.2050
North	Storage	8715.0465
Central	Outdoors	7931.8460
West	Outdoors	7875.1030
North	Outdoors	7806.3265
South	Kitchen	7761.4045
East	Outdoors	7175.3340
	Kitchen	6695.8370
West	Kitchen	6674.0010
Central	Kitchen	6518.9710
North	Kitchen	6283.4625
Central	Personal Care	5616.1620
West	Personal Care	5213.3370
North	Personal Care	5207.5505
East	Personal Care	4616.1995
South	Personal Care	4239.0275

Figure 8: Categories performance by region

Question 2: Do discounts lead to more items sold?

Answer: Discounts do not lead to more items sold (Figure 5). Either customers are not very price-sensitive, or discounts are not strategically targeted.

Question 3: Which loyalty tier generates the most value?

Answer: The loyalty tier that generates the most value is *Gold* (Figure 6). *Gold tier* customers consistently placed the most orders across all regions and also contribute to the most revenue, making it the most engaged and valuable customer segment.

Question 4: Are certain regions struggling with delivery delays?

Answer: Yes, there are regions struggling with delivery delays. The East and North regions had some of the highest late delivery rates.

	sales_region	price_band	late_delivery_rate
0	Central	Low	0.383929
1	Central	Medium	0.391111
2	Central	High	0.396947
3	East	Low	0.410526
4	East	Medium	0.428571
5	East	High	0.411321
6	North	Low	0.386792
7	North	Medium	0.436275
8	North	High	0.361775
9	South	Low	0.336735
10	South	Medium	0.358209
11	South	High	0.420339
12	West	Low	0.389610
13	West	Medium	0.349794
14	West	High	0.380597

Figure 9: Delivery rate by region

Question 5: Do customer signup patterns influence purchasing activity?

Answer: Yes. The customers who signed up in Q2, ordered within 14 days, and received a discount made purchases with decent revenue. Early engagement after signup is possible and may be influenced by discounts.

	delivery_status	payment_method	sales_region	discount_applied
32	Delayed	Credit Card	East	0.2
155	Delivered	Credit Card	Central	0.2
809	Delivered	Credit Card	East	0.2
1431	Cancelled	PayPal	South	0.2
1543	Delayed	PayPal	North	0.2
1874	Delayed	Credit Card	Central	0.2
2498	Delivered	PayPal	West	0.2

	customer_region	loyalty_tier	revenue	order_week	price_band \
32	West	Gold	15.328	27	Low
155	West	Bronze	87.840	27	Medium
809	West	Bronze	89.120	27	Medium
1431	East	Bronze	20.592	27	Low
1543	North	Gold	68.944	27	High
1874	South	Bronze	36.512	27	High
2498	East	Bronze	42.640	27	Medium

Figure 10: Customers who signed up in Q2 and placed an order within the first 14 days

8. Recommendations

Based on the previous findings, the following suggestions would help increase the service's popularity and improve the overall customer experience:

Investigate logistics partners, warehouse capacity, or staffing issues in the East and North regions to reduce delays, especially for Medium priced products, where late delivery rate is ~42% (Figure 9). Improved shipping reliability could have a positive impact on the ordering rate and increase revenue.

Reevaluate discount strategies, as higher discounts (>15%) did not lead to higher quantities sold (Figure 5). Focus on optimising pricing and design promotions.

Invest in Gold-tier loyalty programs, since the Gold customers consistently placed the most orders and brought the highest revenue (Figure 6). Consider reviewing the other tiers programs too, as there was a steep difference between the ordering rate and revenue across the loyalty tiers.

Standardise and consolidate region labels. The columns should have different names if they refer to different things. Choosing one authoritative region column for reporting is also a great idea for fixing this problem. If the columns refer to the same thing, they should have consistent entries.

Google Colab link:

https://colab.research.google.com/drive/1RgoDI15ScDVkdg_ZkVY81uEEWeoWdx6Z?usp=sharing