Project Coversheet

Full Name	Kirana Dhinakar Raj
Email	kiranadhinakar@gmail.com
Contact Number	+49 179 5665425
Date of Submission	7th Sept 2025
Project Week	Week 2

Project Guidelines and Rules

1. Submission Format

- **Document Style**:
    - Use a clean, readable font such as *Arial* or *Times New Roman*, size 12.
    - Set line spacing to **1.5** for readability.
- **File Naming**:
    - Use the following naming format:
      Week X – [Project Title] – [Your Full Name Used During Registration]
      Example: Week 1 – Customer Sign-Up Behaviour – Mark Robb
- **File Types**:
    - Submit your report as a **PDF**.
    - If your project includes code or analysis, attach the **.ipynb notebook** as well.

2. Writing Requirements

- Use formal, professional language.
- Structure your content using headings, bullet points, or numbered lists.

3. Content Expectations

- Answer **all** parts of each question or task.
- Reference tools, frameworks, or ideas covered in the programme and case studies.
- Support your points with practical or real-world examples where relevant.

- Go beyond surface-level responses. Analyse problems, evaluate solutions, and demonstrate depth of understanding.

4. Academic Integrity & Referencing

- All submissions must be your own. Plagiarism is strictly prohibited.
- If you refer to any external materials (e.g., articles, studies, books), cite them using a consistent referencing style such as APA or MLA.
- Include a references section at the end where necessary.

5. Evaluation Criteria

Your work will be evaluated on the following:

- Clarity: Are your answers well-organised and easy to understand?
- Completeness: Have you answered all parts of the task?
- Creativity: Have you demonstrated original thinking and thoughtful examples?
- Application: Have you effectively used programme concepts and tools?
- Professionalism: Is your presentation, language, and formatting appropriate?

6. Deadlines and Extensions

- Submit your work by the stated deadline.
- If you are unable to meet a deadline due to genuine circumstances (e.g., illness or emergency), request an extension **before the deadline** by emailing: **support@uptrail.co.uk**
  Include your full name, week number, and reason for extension.

7. Technical Support

- If you face technical issues with submission or file access, contact our support team promptly at **support@uptrail.co.uk.**

8. Completion and Certification

- Certificate of Completion will be awarded to participants who submit at least two projects.
- Certificate of Excellence will be awarded to those who:
  - Submit all four weekly projects, and
  - Meet the required standard and quality in each.
- If any project does not meet expectations, you may be asked to revise and resubmit it before receiving your certificate.

YOU CAN START YOUR PROJECT FROM HERE

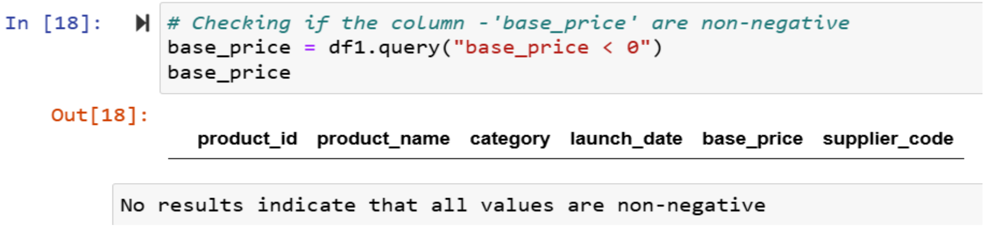
SALES AND CUSTOMER BEHAVIOUR INSIGHTS - GREEN CARD LTD.

1. INTRODUCTION:

This week's project focused on the role of data wrangling in preparing datasets for analysis and visual exploration. Three datasets were used: sales_data, product_info, and customer_info, which were first imported separately as a .csv and converted into a dataframe. In the later steps, they were merged into a single consolidated dataset. The task involved identifying missing values, standardising inconsistent entries, handling duplicates, and performing feature engineering. By transforming the data from a .csv format to a dataframe, the task was to explore customer behaviour, product performance, and business insights.

2. DATA CLEANING SUMMARY:

For the sales_data dataset, missing values in critical columns like 'payment_method' and 'unit_price' were handled, while text inconsistencies such as "delrd" or "delyd" were standardised to "Delayed." Similarly, product_info was cleaned to ensure base prices were valid and launch dates were converted to datetime format for consistency. Also, in both datasets, checking was done ensuring values such as 'unit_price', 'discount_applied' and 'base_price' were all non-negative. The customer_info dataset required standardisation of gender and loyalty tier categories, removal of duplicates, and dropping rows with missing region data. These steps ensured that all three datasets were free from common quality issues like typos, null values, and mismatched formats.

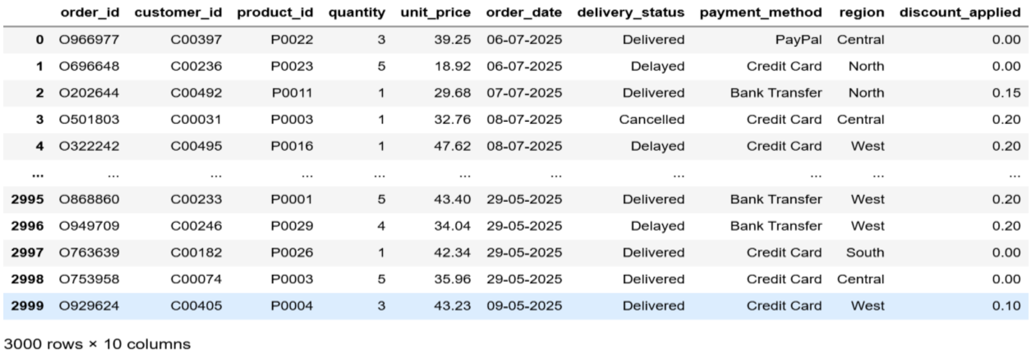


```
In [18]:  # Checking if the column -'base_price' are non-negative
           base_price = df1.query("base_price < 0")
           base_price

Out[18]:
            product_id  product_name  category  launch_date  base_price  supplier_code

        No results indicate that all values are non-negative
```

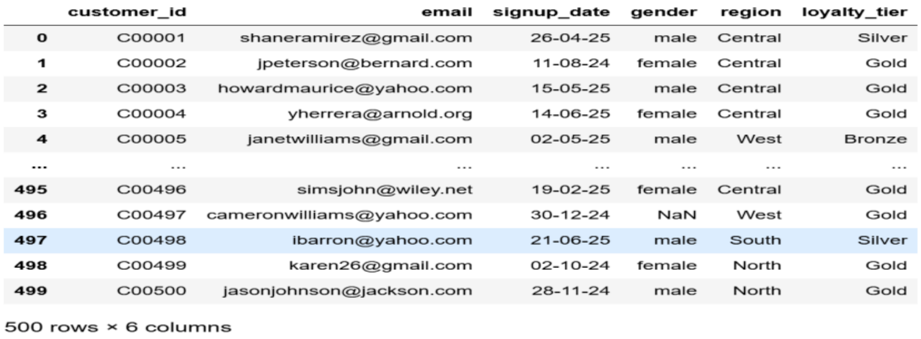
SALES DATASET



	order_id	customer_id	product_id	quantity	unit_price	order_date	delivery_status	payment_method	region	discount_applied
0	O966977	C00397	P0022	3	39.25	06-07-2025	Delivered	PayPal	Central	0.00
1	O696648	C00236	P0023	5	18.92	06-07-2025	Delayed	Credit Card	North	0.00
2	O202644	C00492	P0011	1	29.68	07-07-2025	Delivered	Bank Transfer	North	0.15
3	O501803	C00031	P0003	1	32.76	08-07-2025	Cancelled	Credit Card	Central	0.20
4	O322242	C00495	P0016	1	47.62	08-07-2025	Delayed	Credit Card	West	0.20
...	...	...	...	...	...	...	...	...	...	...
2995	O868860	C00233	P0001	5	43.40	29-05-2025	Delivered	Bank Transfer	West	0.20
2996	O949709	C00246	P0029	4	34.04	29-05-2025	Delayed	Bank Transfer	West	0.20
2997	O763639	C00182	P0026	1	42.34	29-05-2025	Delivered	Credit Card	South	0.00
2998	O753958	C00074	P0003	5	35.96	29-05-2025	Delivered	Credit Card	Central	0.00
2999	O929624	C00405	P0004	3	43.23	09-05-2025	Delivered	Credit Card	West	0.10

3000 rows × 10 columns

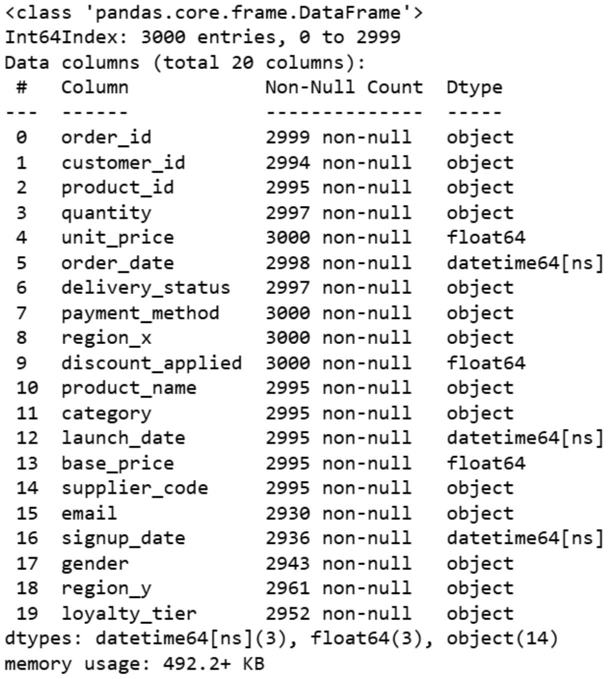
CUSTOMER INFO DATASET



	customer_id	email	signup_date	gender	region	loyalty_tier
0	C00001	shaneramirez@gmail.com	26-04-25	male	Central	Silver
1	C00002	jpeterson@bernard.com	11-08-24	female	Central	Gold
2	C00003	howardmaurice@yahoo.com	15-05-25	male	Central	Gold
3	C00004	yherrera@arnold.org	14-06-25	female	Central	Gold
4	C00005	janetwilliams@gmail.com	02-05-25	male	West	Bronze
...	...	...	...	...	...	...
495	C00496	simsjohn@wiley.net	19-02-25	female	Central	Gold
496	C00497	cameronwilliams@yahoo.com	30-12-24	NaN	West	Gold
497	C00498	ibarron@yahoo.com	21-06-25	male	South	Silver
498	C00499	karen26@gmail.com	02-10-24	female	North	Gold
499	C00500	jasonjohnson@jackson.com	28-11-24	male	North	Gold

500 rows × 6 columns

After all cleaning activity, the three datasets were merged into a single dataframe with about 20 columns.

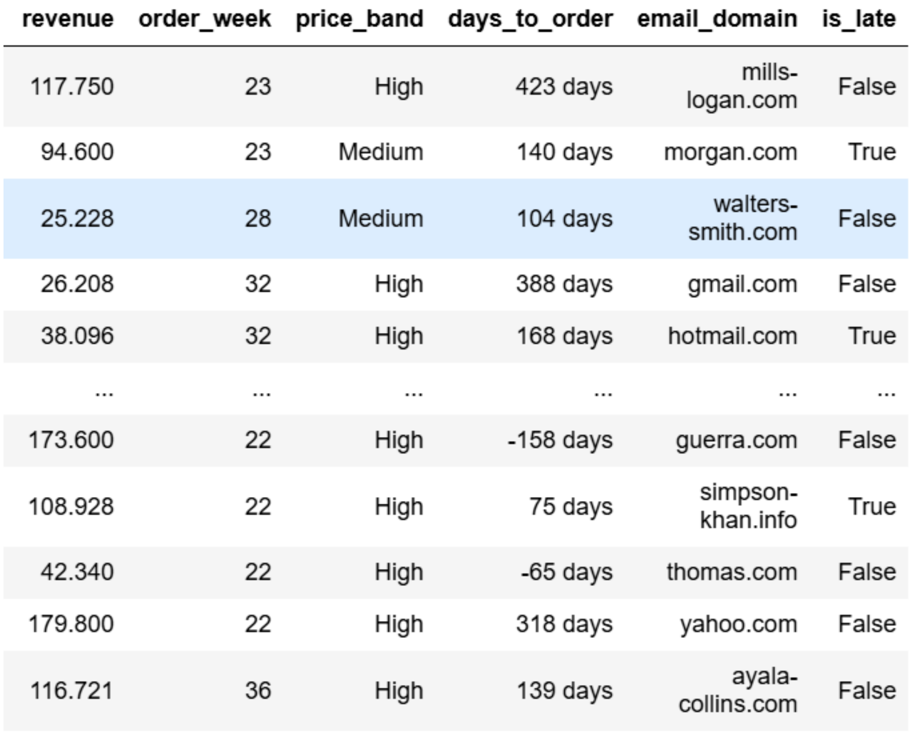


```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3000 entries, 0 to 2999
Data columns (total 20 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   order_id          2999 non-null   object
 1   customer_id       2994 non-null   object
 2   product_id        2995 non-null   object
 3   quantity          2997 non-null   object
 4   unit_price        3000 non-null   float64
 5   order_date        2998 non-null   datetime64[ns]
 6   delivery_status   2997 non-null   object
 7   payment_method    3000 non-null   object
 8   region_x          3000 non-null   object
 9   discount_applied  3000 non-null   float64
 10  product_name      2995 non-null   object
 11  category          2995 non-null   object
 12  launch_date       2995 non-null   datetime64[ns]
 13  base_price        2995 non-null   float64
 14  supplier_code     2995 non-null   object
 15  email             2930 non-null   object
 16  signup_date       2936 non-null   datetime64[ns]
 17  gender            2943 non-null   object
 18  region_y          2961 non-null   object
 19  loyalty_tier      2952 non-null   object
dtypes: datetime64[ns](3), float64(3), object(14)
memory usage: 492.2+ KB
```

3. FEATURE ENGINEERING:

Feature engineering is the process of transforming raw data into meaningful features that would help in improving the performance. For example, revenue was computed as a function of ‘quantity’, ‘unit_price’, and ‘discount_applied’, helping in measuring the sales performance. Date variables such as ‘order_week’, ‘signup_month’, and ‘days_to_order’ were converted into a ‘datetime’ format to ensure uniformity in the data. These also helped in analyzing trends over time and across product categories. Categorical features such as ‘price_band’, ‘email_domain’, and ‘is_late’ provided more information on customer and orders.

CREATION OF NEW FEATURES – revenue, order_week, price_band, days_to_order, email_domain, is_late

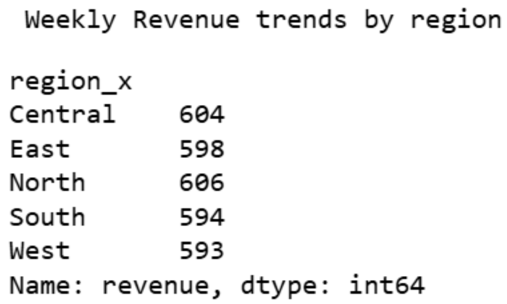


revenue	order_week	price_band	days_to_order	email_domain	is_late
117.750	23	High	423 days	mills-logan.com	False
94.600	23	Medium	140 days	morgan.com	True
25.228	28	Medium	104 days	walters-smith.com	False
26.208	32	High	388 days	gmail.com	False
38.096	32	High	168 days	hotmail.com	True
...	...	...	...	...	...
173.600	22	High	-158 days	guerra.com	False
108.928	22	High	75 days	simpson-khan.info	True
42.340	22	High	-65 days	thomas.com	False
179.800	22	High	318 days	yahoo.com	False
116.721	36	High	139 days	ayala-collins.com	False

4. KEY FINDINGS AND TRENDS:

A description of the summary tables is listed below;

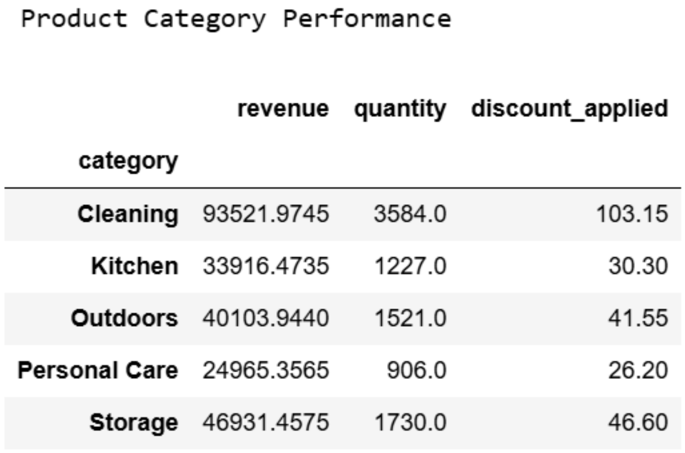
- **Weekly revenue trends by region:** Revenue analysis revealed significant regional differences, with some regions consistently outperforming others in weekly revenue.



```
 Weekly Revenue trends by region

region_x
Central    604
East       598
North      606
South      594
West       593
Name: revenue, dtype: int64
```

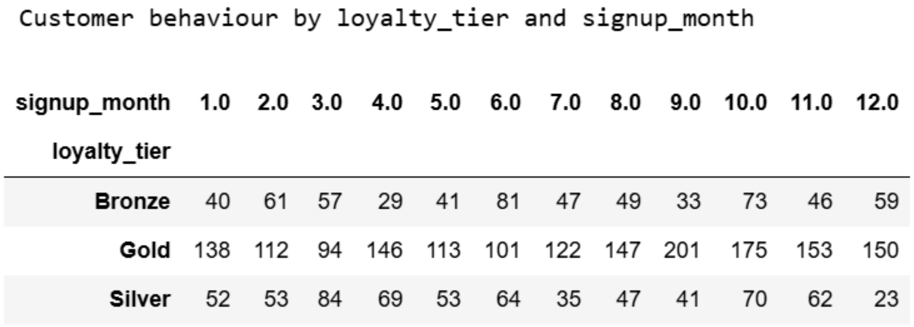
- **Product category performance (revenue, quantity, discount_applied):** This highlighted a small number of categories generating the majority of revenue, showing potential focus areas for product strategy.



Product Category Performance

	revenue	quantity	discount_applied
category			
Cleaning	93521.9745	3584.0	103.15
Kitchen	33916.4735	1227.0	30.30
Outdoors	40103.9440	1521.0	41.55
Personal Care	24965.3565	906.0	26.20
Storage	46931.4575	1730.0	46.60

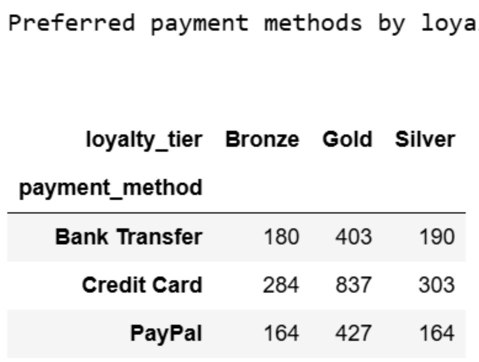
- **Customer behaviour by loyalty_tier and signup_month:** Customer behaviour analysis showed loyalty tier signups fluctuating by month, suggesting marketing campaigns could be timed for better results.



Customer behaviour by loyalty_tier and signup_month

signup_month	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0	11.0	12.0
loyalty_tier												
Bronze	40	61	57	29	41	81	47	49	33	73	46	59
Gold	138	112	94	146	113	101	122	147	201	175	153	150
Silver	52	53	84	69	53	64	35	47	41	70	62	23

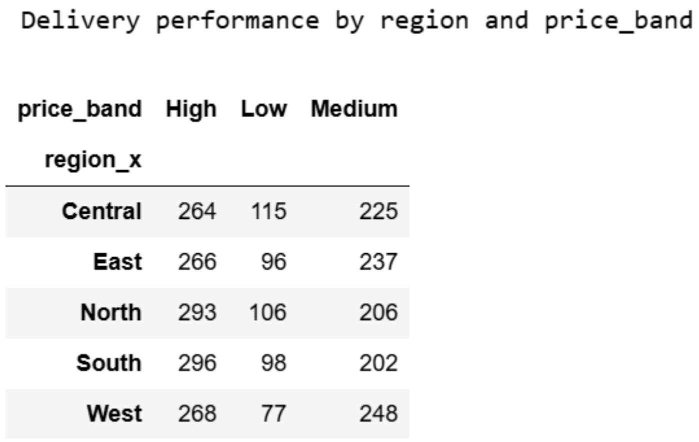
- **Preferred payment methods by loyalty tier:** Payment method preferences also varied by tier, indicating that certain customer groups leaned toward specific payment channels.



```
Preferred payment methods by loyalty tier
```

loyalty_tier	Bronze	Gold	Silver
payment_method			
Bank Transfer	180	403	190
Credit Card	284	837	303
PayPal	164	427	164

- **Delivery performance by region and price_band:** Delivery delays were more frequent in particular regions and price bands, pointing to logistical or operational inefficiencies.

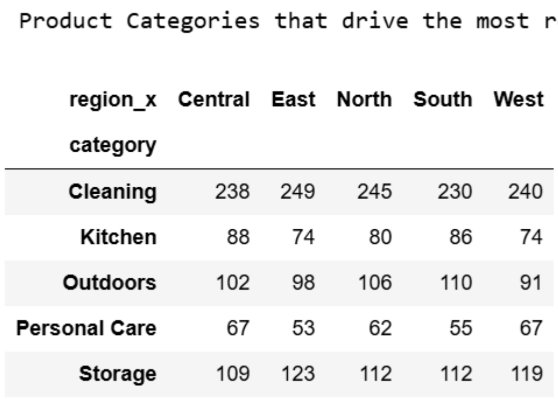


```
Delivery performance by region and price_band
```

price_band	High	Low	Medium
region_x			
Central	264	115	225
East	266	96	237
North	293	106	206
South	296	98	202
West	268	77	248

5. <u>BUSINESS QUESTION ANSWERS:</u>

- Which product categories drive the most revenue, and in which regions? Analysis of product categories across regions showed that a few categories consistently contributed higher revenue. For example, products in the Cleaning category, demonstrated higher revenue in all regions, followed by Storage and then Outdoors.

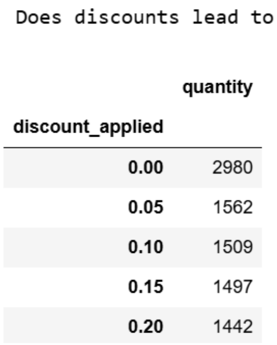


```
Product Categories that drive the most revenue among the regions in UK
```

region_x category	Central	East	North	South	West
Cleaning	238	249	245	230	240
Kitchen	88	74	80	86	74
Outdoors	102	98	106	110	91
Personal Care	67	53	62	55	67
Storage	109	123	112	112	119

- Do discounts lead to more items sold?

  Not necessarily. Higher discounts, led to a considerable increase in the quantity of products sold. However, the highest quantity of products sold were the ones which had no discount.

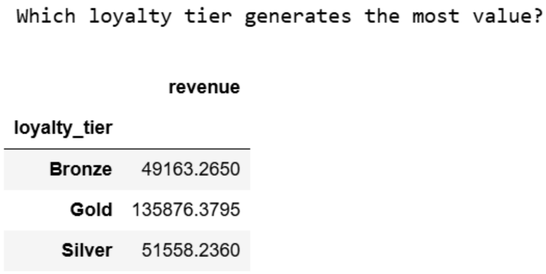


```
Does discounts lead to more items sold?
```

discount_applied	quantity
0.00	2980
0.05	1562
0.10	1509
0.15	1497
0.20	1442

- Which loyalty tier generates the most value?

  The Gold tier emerged as the most valuable segment, generating the highest revenue among all loyalty tiers. This suggests that customers in the Gold tier are more profitable, making them a key focus group.

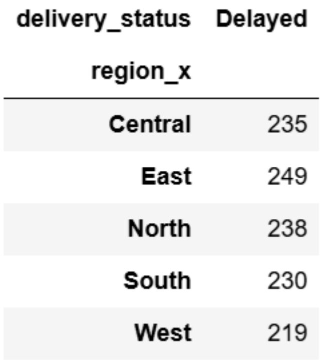


```
Which loyalty tier generates the most value?
```

loyalty_tier	revenue
Bronze	49163.2650
Gold	135876.3795
Silver	51558.2360

- Are certain regions struggling with delivery delays?

  Almost all regions showed delays in deliveries, with the East region having the most followed by the North region.



delivery_status	Delayed
region_x	
Central	235
East	249
North	238
South	230
West	219

- Do customer signup patterns influence purchasing activity?

  Customers signing up in certain months contributed more significantly to both revenue and quantity sold. This suggests that the signup month /time affects future purchasing behaviour, making it valuable to align marketing and onboarding strategies with these peak signup months.

6. RECOMMENDATIONS:

Based on the findings, businesses should prioritize categories and regions that consistently drive revenue and develop targeted campaigns around them. Reducing delivery delays in almost all regions can directly improve customer satisfaction and brand trust. Loyalty programs should emphasize retaining Gold tier customers while also providing incentives to Silver and Bronze members to increase their spending. Finally, discount strategies should be optimized to increase sales volume without decreasing the overall profits.

7. DATA ISSUES OR RISKS:

One simple issue that I encountered during the creation of a new revenue feature was that the column ‘quantity’ was of type ‘object’, so the multiplication with the given formula couldn’t be done. It had to be converted into float before performing the operation. Another issue is that there were certain entries missing in the region column. When analyzing user behavior or trends, ensuring all entries present especially in crucial columns like region is something to be taken into account in the future.