Week 2 - Project: Sales & Customer Behaviour Insights – Green Cart Ltd.

Please write the answers in the **'Project Coversheet'** and refer to the dataset provided for completing the tasks.

Deliverables

1. **Jupyter Notebook (.ipynb):** A clean, well-structured notebook with merged and cleaned dataset, feature engineering, summary tables, visualisations, and markdown explanations throughout.
2. **PDF Report (max 1500 words):** A professional summary including key insights, tables and charts, business questions answered, and clear recommendations (optional screenshots of outputs. No code screenshots required).

Business Scenario

You've joined 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 your manager on the Data & Insights team has asked you to investigate sales and customer behaviour across regions and product lines.

You've been given access to sales, product, and customer datasets, and your job is to:

- Clean and merge the data
- Create new features
- Analyse patterns and performance
- Present insights using charts and summary tables

The findings will inform upcoming marketing and operational strategies.

Dataset

1. sales_data.csv
2. product_info.csv
3. customer_info.csv

Columns Include:

1. **sales_data.csv**
   - order_id
   - customer_id
   - product_id
   - quantity

- unit_price
- order_date
- delivery_status (e.g., Delivered, Delayed, Cancelled - inconsistent casing expected)
- payment_method
- region
- discount_applied

2. **product_info.csv**
   - product_id
   - product_name
   - category
   - launch_date
   - base_price
   - supplier_code

3. **customer_info.csv**
   - customer_id
   - email
   - signup_date
   - gender
   - region
   - loyalty_tier (Bronze, Silver, Gold — inconsistent formatting expected)

Tasks

1. Load the Data

- Load all three CSVs into separate Pandas DataFrames

2. Clean the Data

Clean each DataFrame individually before merging:

- Standardise text formatting using .str.strip(), .str.lower() or .str.title():
  - e.g., " DELAYED " → "Delayed", "gold " → "Gold"
- Convert date columns (order_date, signup_date, launch_date) to datetime using pd.to_datetime()
- Handle missing values:
  - Use .isnull().sum() to identify missing values
  - Fill where appropriate (e.g., discount_applied = 0.0)
  - For categories, fill with "Unknown" or "Other" if appropriate
  - Drop only if necessary (explain your decision in markdown)
- Remove duplicates:

- Use .duplicated() and .drop_duplicates() (e.g., by order_id)

- Validate numeric columns:
  - Ensure quantity, unit_price, and discount_applied are all non-negative
  - Optionally filter or correct records with invalid values

3. Merge the Data

After cleaning, perform the following merges:

- Merge sales_data with product_info using product_id
- Then merge the result with customer_info using customer_id
- Use how='left' to preserve all sales transactions
- Save the merged dataset as merged_df
- Inspect merged_df using .info() and .head() to confirm merge success

4. Feature Engineering

Create the following new columns:

- 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 (e.g., gmail.com)
- is_late = True if delivery_status is "Delayed"

Use functions like .dt, .apply(), .map(), or pd.cut() where appropriate.

5. Create Summary Tables

Using groupby(), agg(), pivot_table(), or query():

- Weekly revenue trends by region
- Product category performance (revenue, quantity, discount)
- Customer behaviour by loyalty_tier and signup_month
- Delivery performance by region and price_band
- Preferred payment methods by loyalty_tier

6. Visual Exploration

Create at least **6 clear and labelled visualisations** using **Matplotlib and Seaborn**. Suggested charts:

1. Line plot - weekly revenue trends by region
2. Bar chart - top 5 categories by revenue
3. Boxplot - quantity vs discount across categories
4. Heatmap - correlation between revenue, discount, and quantity
5. Countplot - orders by loyalty tier (with hue = region)

6. Stacked bar or pie - delivery status by price band

7. Business Questions to Answer (In PDF Report)

1. Which product categories drive the most revenue, and in which regions?
4. Do discounts lead to more items sold?
5. Which loyalty tier generates the most value?
6. Are certain regions struggling with delivery delays?
7. Do customer signup patterns influence purchasing activity?

8. Optional Stretch Tasks

Try these if you finish early:

- Use .query() to extract:
  - Customers who signed up in Q2
  - Placed an order within 14 days
  - Received a discount > 20%

- Use MinMaxScaler to normalise revenue or price
- Flag underperforming products (low quantity, high discount, delayed deliveries)

Report Structure (Submit as a PDF file)

Your report should be clear, concise, and well-structured, following the format below (maximum 1500 words):

1. Introduction
   - Briefly describe the business task you were given and dataset
   - Include a short explanation of what the report aims to show (e.g., revenue trends, customer behaviour, delivery performance)
2. Data Cleaning Summary
   - Explain what you cleaned and how
   - Mention duplicates removed, missing data handled, and Inconsistent labels standardised (e.g., delivery_status, loyalty_tier) (optional: include screenshot of output)
3. Feature Engineering Summary
   - List and briefly explain the new features you created, such as: revenue, order_week, price_band, email_domain, is_late
4. Key Findings & Trends
   - Write 2–3 key observations or trends based on your analysis (e.g., "Revenue peaked in Week XX, mostly driven by high-value orders in the ‘XYZ’ category.")
   - Include relevant screenshots of summary tables or visualisations from your notebook
5. Business Question Answers
   - Clearly answer each question with short explanations

- Use visual or tabular outputs to support your answers (include screenshots where needed).

6. Recommendations
   - Suggest 2-3 ideas based on your findings (e.g., Focus promotions on the 'Personal Care' category in high-performing regions, Improve shipping reliability in regions with high delay rates)
7. Data Issues or Risks
   - Identify one data quality issue you encountered (e.g., inconsistent regions, missing loyalty tiers)
   - Suggest how it could be fixed at the source (e.g., improve data entry validation, add automated checks)

Note: Include screenshots from your jupyter notebook where required.

Submission Checklist

Before you submit, ensure you have:

✅ Your completed Jupyter Notebook (.ipynb)
✅ A PDF report following the structure above
✅ Answers written in the **'Project Coversheet'** as instructed

Final Tip: Remember, your audience is non-technical. Avoid jargon. Your job is to tell a story with the data that helps the business make better decisions, clear, actionable, and relevant.