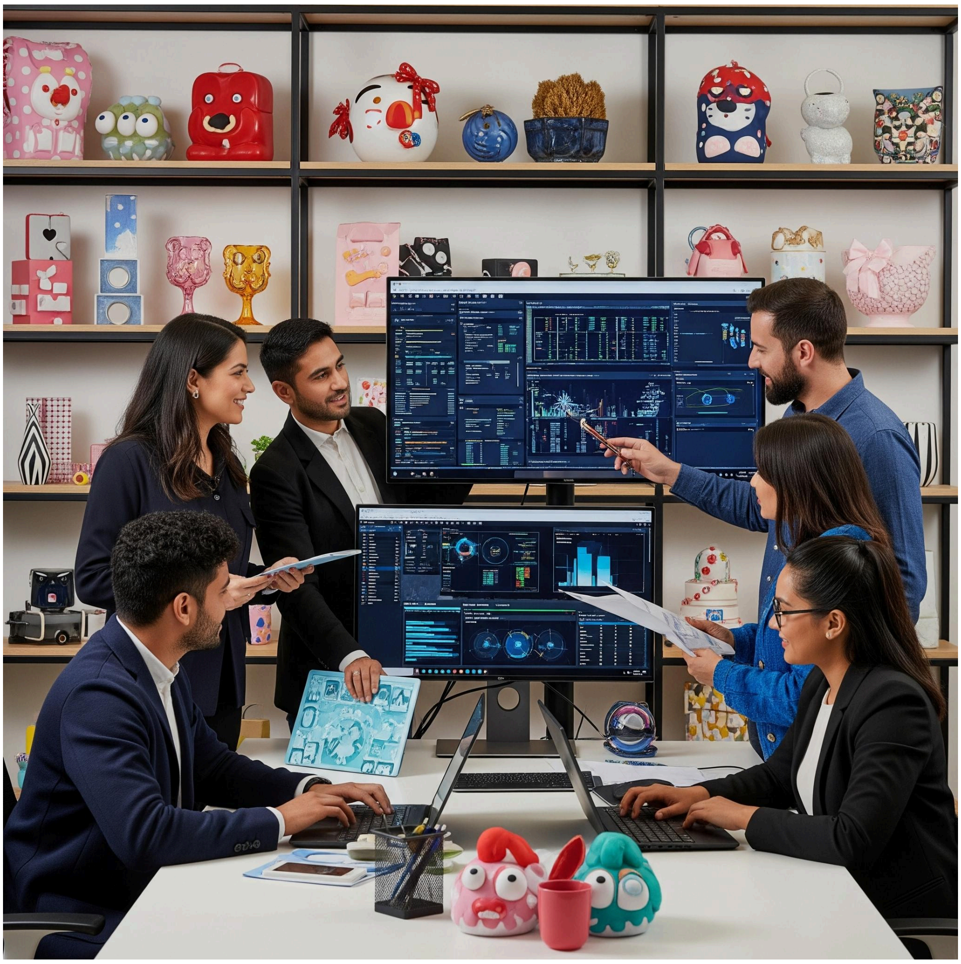
IJSE®

Institute of Software Engineering

ITS 2122: Python for Data Science & AI - Group Project Specification (Semester 3, 2025)

Project Title: Strategic Growth Analysis for a UK-Based E-Commerce Retailer



Part 1: The Business Mandate & Project Overview

1.1 Introduction: Your Role as Data Science Consultants

For this capstone project, your group will operate as a team of data science consultants. You have been engaged by the executive board of "**Unique Gifts Ltd.**," a UK-based, non-store online retailer that has carved a niche in the market by selling unique, all-occasion giftware. The company has enjoyed several years of organic growth and has amassed a significant volume of transactional data. However, their strategic decision-making process has been largely intuitive rather than data-driven.

The leadership team recognizes that to sustain growth, optimize operations, and increase profitability in a competitive market, they must leverage their data as a strategic asset. Your team's primary mandate is to conduct a thorough analysis of their historical sales data, uncover actionable insights, and provide a clear, data-backed strategic roadmap for the future. You are expected to move beyond simple reporting and deliver high-value business intelligence that can directly inform marketing, inventory, and customer relationship management strategies.

1.2 The Core Challenge: From Data to Decisions

The management team at Unique Gifts Ltd. has provided you with two full years of raw transactional data. They have identified several critical blind spots in their understanding of the business and have tasked your team with addressing the following fundamental questions:

1. **Sales Performance & Seasonality:** What are the overarching trends in our sales performance on a monthly and yearly basis? Are there discernible seasonal patterns or peak shopping periods that we can capitalize on with targeted promotions or inventory planning?
2. **Product Portfolio Optimization:** Which products are our top performers, and which are lagging? Crucially, how should we define "performance"—is it by sales volume (quantity

sold) or by revenue generated (total value)? A clear distinction is needed to manage our "bread-and-butter" products versus our "cash cow" products.

3. **Geographic Footprint:** Where are our most valuable customers located? What is the balance between our domestic UK market and our international sales, and where are the key opportunities for international growth?
4. **Customer Segmentation:** Can we move beyond a one-size-fits-all approach to our customers? Is it possible to segment our customer base into meaningful groups based on their purchasing behavior? How can we tailor marketing campaigns and retention efforts to the unique characteristics of each segment?
5. **Wholesaler vs. Retail Analysis:** We know that a significant portion of our customer base consists of wholesalers who buy in bulk. How does their purchasing behavior (e.g., frequency, order value) differ from that of individual retail customers? Should these two groups be managed with distinct strategies?

1.3 Project Objectives

This project is designed to be a comprehensive application of the skills and concepts covered in the ITS 2122 module. Successful completion will require your team to demonstrate proficiency in the entire data analysis lifecycle. Specifically, you will be expected to:

- **Data Ingestion and Cleaning:** Apply robust data cleaning, validation, and preprocessing techniques using the Pandas library to handle real-world data imperfections, such as missing values and incorrect entries. This directly assesses skills from **Module 5: File Handling** and **Module 7: Python Libraries and Data Processing**.
- **Exploratory Data Analysis (EDA):** Conduct a comprehensive EDA using Pandas and NumPy to dissect the data, identify underlying patterns, and formulate data-driven answers to the core business questions outlined above. This assesses **Module 7: Python Libraries and Data Processing**.
- **Data Visualization:** Create a portfolio of clear, insightful, and professionally formatted data visualizations using Matplotlib and Seaborn to effectively communicate your findings to a non-technical audience. This assesses **Module 7: Data Visualization**.

- **Advanced Analytics and Modeling:** Implement a customer segmentation model using the Recency, Frequency, Monetary (RFM) framework. This will require the application of Python's core data structures, user-defined functions, and advanced Pandas manipulations, covering **Modules 2, 3, and 7**.
- **Professional Communication:** Synthesize all analytical findings into a persuasive, professionally structured business report that presents not just data, but actionable, evidence-based strategic recommendations.
- **Data Enrichment:** Demonstrate proficiency in acquiring and integrating external data by using a web API to enrich the primary dataset. This task assesses skills from **Module 8: APIs, Data Collection & Debugging**.

Part 2: The Primary Asset: Data & Tools

2.1 The Dataset: Online Retail II

The foundation of your analysis will be the "**Online Retail II**" dataset, a public and well-documented collection of transactional data from the UCI Machine Learning Repository [1]. This dataset is highly relevant to the business scenario, containing over one million records of transactions that occurred between December 1, 2009, and December 9, 2011, for a UK-based online giftware retailer.

Your first task, which directly tests your file handling capabilities (**Module 5**), will be to load the provided data file (in .csv format) into a Pandas DataFrame to begin your analysis.

2.2 Data Dictionary

A precise understanding of the data is critical for any meaningful analysis. The following table provides a definitive dictionary for the dataset's columns, compiled from the official documentation and preliminary data exploration. Your team must familiarize itself with these fields and their specific nuances before proceeding.

Variable Name	Data Type	Description	Key Considerations & Notes
Invoice	object (str)	A 6-digit number uniquely assigned to each transaction.	If this code starts with the letter 'C', it indicates a cancellation. These records are not sales and must be handled appropriately during the cleaning phase.
StockCode	object (str)	A 5-digit number (or sometimes with a letter at the end) uniquely assigned to each distinct product.	Some codes may represent non-product charges like postage, bank fees, or other adjustments. These may need to be identified and filtered out.
Description	object (str)	The name of the product or item.	This field may contain inconsistencies in capitalization, spacing, or wording that require standardization for accurate product-level analysis.
Quantity	int64	The number of units of each product per transaction.	Can be negative. Negative quantities correspond to cancelled orders (identified by an InvoiceNo starting with 'C') and must be addressed.
InvoiceDate	object (str)	The date and time when the transaction was generated.	This column must be converted from its raw object (string) format to a proper datetime type to enable any time-series analysis.

Price	float64	The price of one unit of the product in British Pounds (£).	Some records have a UnitPrice of 0.0. These may represent promotional gifts, data entry errors, or other non-revenue events and should be carefully evaluated.
Customer ID	float64	A 5-digit number uniquely assigned to each customer.	This ID is **essential** for any customer-level analysis, including RFM. The key challenge is that many rows have missing 'Customer ID's. These specific rows must be filtered out before conducting the RFM analysis. The column should then be converted to an integer type.
Country	object (str)	The name of the country where the customer resides.	The dataset is heavily skewed towards the "United Kingdom," but includes a long tail of dozens of other nations, presenting an opportunity for geographic analysis.

2.3 Required Python Libraries

To complete this project, you will utilize the core data science stack taught in this module. Your Jupyter Notebook environment should import the following essential libraries:

- **pandas**: The primary tool for data loading, manipulation, cleaning, and analysis.
- **numpy**: For numerical operations, which is often used under the hood by Pandas but may be needed for specific calculations.

- **matplotlib** and **seaborn**: The standard libraries for creating static, annotated, and aesthetically pleasing data visualizations.
- **requests**: Required only for the task involving API integration.

Part 3: The Analytical Blueprint: A Phased Approach

Follow this structured, phased approach to guide your analysis from raw data to strategic recommendations. Each phase builds upon the last and is designed to ensure a comprehensive and methodologically sound project.

3.1 Phase 1: Data Sanitation and Preprocessing

The quality of your insights is directly dependent on the quality of your data. This initial phase is critical for transforming the raw, messy dataset into a clean, reliable, and analysis-ready format.

Tasks:

1. **Load Data:** Ingest the given '**online_retail.csv**' file into a Pandas DataFrame.
2. **Initial Assessment:** Conduct a preliminary data audit. Use methods like **.info()**, **.describe(include='all')**, and **.isnull().sum()** to get a high-level overview of the DataFrame, including data types, summary statistics for both numerical and categorical columns, and the extent of missing values.
3. **Handle Duplicates:** Identify and remove any duplicate rows to ensure each transaction is unique.
4. **Handle Missing Data:** The most significant issue is the missing '**Customer ID**'s. Develop a strategy to handle these records. The most common approach is to remove them, but you must explicitly state this decision in your report and justify it.
5. **Clean Transactional Data:**
   - ○ Identify and remove all cancelled orders (transactions where the Invoice No begins with 'C').
   - ○ Investigate and filter out records where the '**Price**' is 0.0. Justify why these records

are being excluded from the sales analysis (e.g., they do not represent a commercial transaction).

- ○ Some **StockCode** s may represent non-product charges like postage (‘POST’), manual entries (‘M’), bank fees, or other adjustments. These may need to be identified and filtered out.

6. **Feature Engineering:**
   - ○ Create a new column named ‘**TotalPrice**’ by multiplying the ‘**Quantity**’ and ‘**Price**’ columns. This is the fundamental metric for measuring revenue.
   - ○ To facilitate temporal analysis, parse the InvoiceDate column. Create new columns for **Year**, **Month**, **DayOfWeek**, and **HourOfDay**.
7. **Data Type Conversion:** Verify and correct the data types of all columns. This is a critical step.

- ● **InvoiceDate:** Convert this from an object (string) to a datetime64[ns] object using pd.to_datetime(). This is essential for all time-based analysis.
- ● **Customer ID:** After removing the missing values, convert this column from float64 to int. This is its correct, final data type.
- ● **StockCode:** Ensure this column is treated as an object or str, as it contains alphanumeric characters and is a unique identifier, not a number for calculation.

3.2 Phase 2: Exploratory Data Analysis (EDA) & Insight Generation

With a clean dataset, your team can now dive into exploration to uncover patterns and answer the board's initial questions. Your goal is to move from observation to interpretation.

Tasks:

1. **Temporal Analysis:**
   - ○ Generate visualizations (e.g., line charts) to show total sales revenue over time. Plot this on a monthly basis across the two years to identify growth trends and potential seasonality.
   - ○ Some analyses of this dataset have noted sales dips in specific months like February and April. Investigate this. Is it a consistent pattern? Propose and discuss potential

business reasons, such as post-holiday spending fatigue, the absence of major marketing campaigns, or inventory issues.

  - Analyze sales patterns by **DayOfWeek** and **HourOfDay**. Create bar charts to identify the peak shopping times. This information is highly valuable for scheduling promotions, server maintenance, and customer service staffing.

2. **Geographic Analysis:**
   - Determine the top 10 countries by total sales revenue.
   - Calculate the exact percentage of total revenue generated by the UK market versus all other countries combined.
   - Create a bar chart to visualize the revenue contribution of the top countries, providing a clear picture of the company's international market presence.
3. **Product Performance Analysis:**
   - This task requires a nuanced approach to the concept of "top products." You must produce two distinct analyses:
     - **Top 10 Products by Quantity Sold:** Identify the products that are sold most frequently. These are your high-volume, "bread-and-butter" items.
     - **Top 10 Products by Total Revenue:** Identify the products that contribute the most to the company's bottom line. These are your high-value "cash cows."
   - In your report, present both lists and write a detailed comparison. Are the lists similar or different? What does this tell you about the product portfolio? For example, a low-cost item might sell in high volume but contribute less revenue than a high-cost, lower-volume item. This distinction is critical for strategic inventory and marketing decisions.

3.3 Phase 3: Advanced Analytics - RFM Customer Segmentation

This phase moves from descriptive to diagnostic analytics. You will implement the RFM (Recency, Frequency, Monetary) model, a classic marketing analytics technique used to quantitatively rank and segment customers based on their purchasing behavior.

Tasks:

1. **Calculate RFM Metrics:** Create a new DataFrame summarized by CustomerID. For each unique customer, you must calculate:
    - ○ **Recency (R):** The number of days between the customer's last purchase and a "snapshot" date. The snapshot date should be set to one day after the most recent InvoiceDate in the entire dataset to ensure all recency values are positive.
    - ○ **Frequency (F):** The total count of unique invoices (transactions) made by the customer.
    - ○ **Monetary (M):** The sum of TotalPrice across all purchases made by the customer.
2. **Assign RFM Scores:**
    - ○ For each of the three metrics (R, F, and M), use the pandas.qcut() function to divide the customers into five equal-sized groups (quintiles).
    - ○ Assign a score from 1 to 5 to each customer for each metric. **Note:** For Recency, a smaller value (fewer days since last purchase) is better. Therefore, you must assign scores so that the most recent purchasers receive a score of 5. For Frequency and Monetary, higher values are better, so the highest spenders/most frequent buyers should receive a score of 5.
3. **Combine Scores:** Concatenate the individual R, F, and M scores to create a combined RFM_Segment string (e.g., a customer in the top quintile for all three would be '555').
4. **Map to Descriptive Segments:** To make the scores interpretable, create a Python dictionary or a custom function to map the numerical RFM scores to descriptive segment names (e.g., 'Champions', 'At-Risk Customers').

RFM Segment	**Characteristics**
Champions	Bought recently, buy often, and spend the most.
Loyal Customers	Buy on a regular basis. Responsive to promotions.
Potential Loyalists	Recent customers with average frequency.
New Customers	Bought recently, but not often.
At-Risk Customers	Haven't purchased in a while. Used to be frequent buyers.
Hibernating	Low recency, low frequency. Last purchase was long ago.

3.4 Phase 4: Strategic Recommendations

A technical RFM model is only useful if its output is translated into actionable business strategy. This phase requires you to think like a marketing strategist.

Task:

Investigate the Wholesaler Hypothesis: The dataset documentation notes that many customers are wholesalers. This is a critical piece of domain knowledge. Before finalizing your segments, investigate this. Create a histogram or box plot of the Monetary value per customer. Do you see a skewed distribution or evidence of two distinct groups (a large group of low-spending retail customers and a smaller group of very high-spending customers)? Comment on this finding. This analysis provides crucial context for your recommendations, as the strategy for a wholesaler is fundamentally different from that for a retail customer.

3.5 Phase 5: Data Enrichment via API Integration

This task allows your team to demonstrate advanced skills in data acquisition and integration, as covered in **Module 8**.

Task:

Select a free currency conversion API (e.g., ExchangeRate-API, Open Exchange Rates). Write a Python script using the ‘requests’ library to make API calls to fetch the daily exchange rates for USD and EUR relative to GBP. Apply this to the top 100 transactions (by TotalPrice) in your dataset. Create two new columns, **TotalPrice_USD** and **TotalPrice_EUR**, and populate them with the converted values. In a separate section of your notebook, document the process, include the code, and explain the business value of this enrichment (e.g., for financial reporting to international stakeholders or for creating regional pricing strategies).

Part 4: Final Deliverables

Your final submission will consist of two distinct components: a professional business report and a technical appendix.

4.1 The Strategic Insights Report (PDF Format)

This document is the primary deliverable for the executive board of Unique Gifts Ltd. It should be a polished, professional report written in clear business language. **It is not a code report.** The focus must be on insights and recommendations, with visualizations used to support your narrative. The main body of your report must be between **2,500 and 3,000 words**. This is a strict requirement. The word count does not include the title page, table of contents, or appendix.

Required Structure:

1. **Executive Summary:** A concise, one-page overview of the project's purpose, your most critical findings, and your top three strategic recommendations. This is the most important page of the report.
2. **Introduction:** Briefly state the project's objectives and the business problem as defined in the mandate.
3. **Data & Methodology:** Briefly describe the dataset used and outline the key steps (Phase 1) of your methodology (e.g., data cleaning approach).
4. **Sales Performance Analysis:** Present your findings from the temporal and geographic EDA (Phase 2). Use your best visualizations to illustrate sales trends, seasonality, and the geographic distribution of revenue.
5. **Product Portfolio Insights:** Discuss your analysis of the top-performing products by both volume and revenue (Phase 2). Explain the strategic implications of the differences between these two views.
6. **Customer Deep Dive: Segmentation & Strategy:** This is the core analytical section of your report. Present the results of your RFM segmentation, including a summary of each segment's size and characteristics (Phase 3). Discuss your findings on the wholesaler hypothesis (Phase 4).

7. **Data Enrichment via API Integration:** Describe the results you obtained through the currency exchange API integration in Phase 5.

4.2 The Technical Appendix (Jupyter Notebook)

Submit a single, well-organized Jupyter Notebook (.ipynb) file that contains all the Python code used to perform your analysis. This notebook serves as the technical documentation for your project and must be reproducible.

Requirements:

- **Clean and Logical Flow:** The notebook should follow the phased structure of the project (Phase 1 to Phase 5).
- **Well-Documented Code:** Code cells must be accompanied by Markdown cells that explain the rationale behind your actions. Do not just write code; explain *why* you are performing a certain step, what choices you made (e.g., why you dropped certain rows), and how you interpret the output of a code cell.
- **Reproducibility:** The notebook should run from top to bottom without errors, assuming the data file is in the correct location.

Part 5: Assessment & Evaluation Rubric

5.1 Grading Breakdown

The project will be graded out of 100 points, distributed as follows:

- **Strategic Insights Report (PDF):** 50%
- **Technical Appendix (Jupyter Notebook):** 40%
- **Teamwork & Collaboration (Peer Evaluation):** 10%

Please note that this will not be calculated on a project-wise basis. All of the above will be assessed based on individual performance in the final presentation, viva, and your contribution to the project. Therefore, members of the same group may receive different marks.

5.2 Detailed Rubric

Your submission will be evaluated against the following criteria to ensure a fair and transparent assessment process.

Criteria	**Weight**	**Unsatisfactory (0-49%)**	**Satisfactory (50-69%)**	**Good (70-89%)**	**Excellent (90-100%)**
Code Quality & Correctness	20%	Code is non-functional, contains major logical errors, is disorganized, or shows evidence of plagiarism.	Code runs but is inefficient, poorly structured, lacks comments, and is difficult to follow.	Code is functional, mostly correct, and includes some documentation explaining its purpose.	Code is efficient, well-structured, extensively documented with Markdown explanations, and follows Python best practices (e.g., PEP 8).
Data Cleaning & Preprocessing	10%	Fails to address or incorrectly handles major data quality issues like nulls, duplicates, or cancellations.	Basic cleaning is performed (e.g., dropping nulls), but without clear justification or analysis	All required cleaning steps are completed correctly with adequate justification provided	Cleaning is thorough, and choices are expertly justified. The analysis shows consideration for potential

			of the impact on the dataset.	in the notebook.	biases introduced during cleaning.
Data Analysis & Depth	20%	Analysis is superficial, contains significant errors in calculation or interpretation, or fails to address the business questions.	Basic analysis is performed (e.g., simple totals and averages), but lacks depth, interpretation, or a connection to business context.	Correctly performs all required EDA and RFM analysis, generating valid results and drawing logical conclusions from them.	Analysis is deep and insightful, uncovering non-obvious patterns (e.g., the wholesaler effect, nuanced seasonality) and forming strong, data-supported hypotheses.
Data Visualization	10%	Visualizations are missing, misleading, inappropriate for the data type, or completely unreadable.	Plots are generated but are poorly labeled, cluttered, lack titles, and fail to communica	Plots are clear, correctly labeled, have appropriate titles and axes, and are	Visualizations are compelling, aesthetically pleasing, and effectively tell a story, making

			te a clear point.	suitable for the data being presented.	complex data easy to understand for a non-technical audience.
Report & Communication	30%	Report is poorly written, unstructured, and fails to communicate findings. It reads like a code dump rather than a business document.	Report is a simple summary of the notebook's output with little business context or strategic thinking. Recommendations are generic or absent.	Report is well-structured and clearly communicates the analytical findings. Recommendations are present and linked to the analysis.	Report is a persuasive, professional business document with a strong narrative, a compelling executive summary, and highly specific, actionable, data-driven recommendations.
Teamwork & Collaboration	10%	Based on peer evaluation, there is no evidence of collaboration. Work is clearly	Minimal evidence of collaboration. Contributions are	Clear evidence of shared work and meaningful contributio	Excellent collaboration is demonstrated, with evidence of

		siloed or completed by a single member.	highly unbalanced.	n from all team members as reported in peer evaluations.	integrated work and balanced, high-quality contributions from all members.

References:

[1]. Online Retail II data set: https://archive.ics.uci.edu/dataset/502/online+retail+ii