A REPORT
ON

AMAZON SALES DATA ANALYSIS

Submitted by,

S PRIYANKA - 20211CSE0694

Under the guidance of,

Mr. Tamilselvan Thangavel

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

At

PRESIDENCY UNIVERSITY

BENGALURU

MAY 2025

PRESIDENCY UNIVERSITY

PRESIDENCY SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Internship report **“Amazon Sales Data Analysis”** being submitted by “S Priyanka” bearing roll number “20211CSE0694” in partial fulfillment of the requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering is a bonafide work carried out under my supervision.

Mr.TAMILSELVAN THANGAVEL
Assistant Professor
PSCS
Presidency University

Dr. ASIF MOHAMMED
Associate Professor & HOD
PSCS
Presidency University

Dr. MYDHILI NAIR
Associate Dean
PSCS
Presidency University

Dr. SAMEERUDDIN KHAN
Pro-Vice Chancellor Engineering
Dean –PSCS / PSIS
Presidency University

PRESIDENCY UNIVERSITY

PRESIDENCY SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

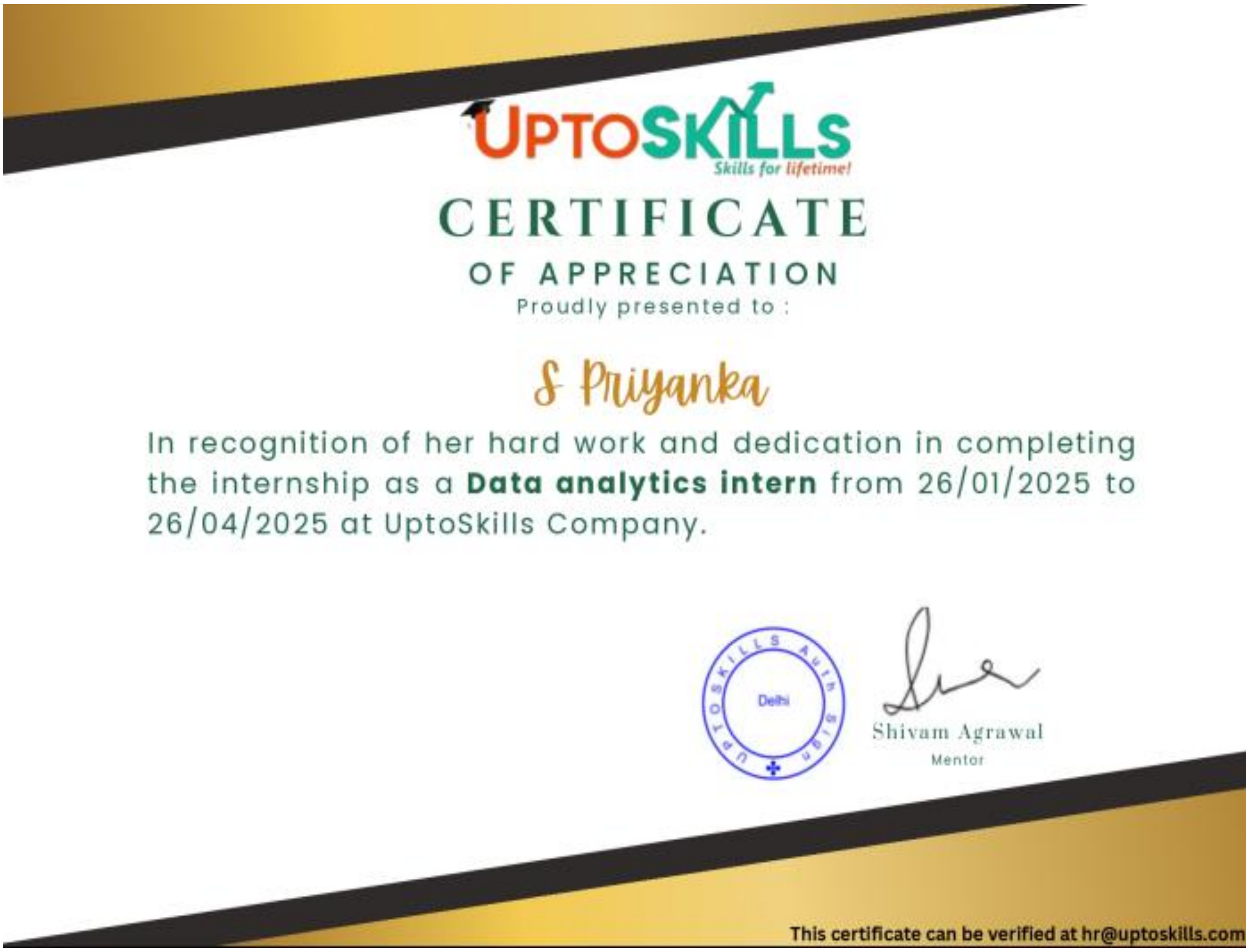
DECLARATION

I hereby declare that the work, which is being presented in the report entitled “**Amazon Sales Data Analysis”** in partial fulfillment for the award of Degree of **Bachelor of Technology** in **Computer Science and Engineering**, is a record of my own investigations carried under the guidance of **Mr.Tamilselvan Thangavel, Assistant Professor, Presidency School of Computer Science and Engineering, Presidency University, Bengaluru.**

I have not submitted the matter presented in this report anywhere for the award of any other Degree.

S Priyanka, 20211CSE0694

INTERNSHIP COMPLETION CERTIFICATE



UPTOSKILLS
Skills for lifetime!

CERTIFICATE

OF APPRECIATION

Proudly presented to :

S Priyanka

In recognition of her hard work and dedication in completing the internship as a **Data analytics intern** from 26/01/2025 to 26/04/2025 at UptoSkills Company.

UPTOSKILLS Auth Sign
Delhi

Shivam Agrawal
Mentor

This certificate can be verified at hr@uptoskills.com

ABSTRACT

In the evolving landscape of data-driven decision-making, e-commerce companies like Amazon are prime examples of how analytics can transform traditional business practices into dynamic, insight-led strategies. With massive volumes of transactional data generated daily, businesses are increasingly reliant on skilled data analysts who can translate raw data into strategic insights. This project report documents the comprehensive analysis of Amazon sales data, developed as an advanced internship project during a Data Analyst internship at UptoSkills.

The internship spanned a total of 3 months. In the first phase (1.5 months), we focused on data collection from multiple sources, primarily related to corporate organizations and colleges across Indian cities. We organized the collected data in Microsoft Excel, performing cleaning, filtering, formatting, and preliminary trend analysis using pivot tables, charts, and formulas. This phase honed foundational data handling skills and built the base for transitioning into data visualization and storytelling.

In the second half of the internship, we migrated to Power BI, where we imported datasets, modeled data relationships, and created dashboards. Here, I developed visual dashboards that enabled stakeholders to understand complex data summaries at a glance—covering performance metrics, comparisons between city-wise engagement, and participation trends. This phase culminated in a capstone-level assignment: a complete analysis and dashboard for Amazon Sales Data.

The Amazon Sales Data Analysis project was designed to simulate real-world e-commerce data scenarios. The dataset included customer IDs, product types, item categories, geographic regions, dates of order and shipment, units sold, unit prices, costs, and profits. The goal was to process the data using Python for analysis and Power BI for dashboarding, delivering a project that reflects actual business intelligence workflows.

By merging scripting analysis with modern BI tools, this project mirrors professional analytics workflows used in industry settings. The project not only validated my learning across Python, Excel, and Power BI but also offered a real-world showcase of how data analysis supports business strategy in the e-commerce sector.

ACKNOWLEDGEMENTS

First of all, we indebted to the **GOD ALMIGHTY** for giving me an opportunity to excel in our efforts to complete this project on time.

We express our sincere thanks to our respected dean **Dr. Md. Sameeruddin Khan**, Pro-VC - Engineering and Dean, Presidency School of Computer Science and Engineering & Presidency School of Information Science, Presidency University for getting us permission to undergo the project.

We express our heartfelt gratitude to our beloved Associate Dean **Dr. Mydhili Nair,** Presidency School of Computer Science and Engineering, Presidency University, and **Dr. Asif Mohammed**, Head of the Department, Presidency School of Computer Science and Engineering, Presidency University, for rendering timely help in completing this project successfully.

We are greatly indebted to our guide **Mr. Tamilselvan Thangavel, Assistant Professor** and Reviewer **Dr. Ranjitha, Assistant Professor**, Presidency School of Computer Science and Engineering, Presidency University for their inspirational guidance, and valuable suggestions and for providing us a chance to express our technical capabilities in every respect for the completion of the internship work.

We would like to convey our gratitude and heartfelt thanks to the CSE7301 Internship/ University Project Coordinator **Mr. Md Ziaur Rahman and Dr. Sampath A K,** department Project Co-ordinator **Dr. Jayanthi Kamalasekaran** and and Git hub coordinator **Mr. Muthuraj.**

We thank our family and friends for the strong support and inspiration they have provided us in bringing out this project.

S Priyanka

LIST OF FIGURES

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

In today’s data-dominated era, businesses across all sectors rely on accurate data analysis to make informed decisions. Among these, the e-commerce industry stands out as a domain where every click, search, purchase, and return generates useful data that, when analyzed properly, can uncover trends, optimize processes, and drive profits.

Amazon, being a global leader in e-commerce, presents a unique opportunity to explore data at scale. This report discusses the full-cycle data analysis of Amazon’s sales data using Python and Power BI-executed during my internship at **UptoSkills**. The project not only aligns with practical business analysis objectives but also bridges academic learning with real-world applications.

Initially, during the internship, our tasks revolved around Excel-based handling of manually collected data from colleges and corporate organizations. We recorded engagements, service usages, and demographic information across Indian cities. This prepared us for structuring, cleaning, and interpreting datasets using Excel tools like filters, pivot tables, conditional formatting, and trend charts. These are the fundamentals on which large-scale data operations are built.

The project then evolved into visual storytelling using **Power BI**. We designed dashboards for the collected data, showcasing city-wise engagement, service interest levels, and temporal participation. This helped prepare us for dashboard building, KPI calculation, and real-time business reporting—all essential in modern business intelligence roles.

As a culmination of this journey, the Amazon Sales Data Project was introduced. The dataset was significantly larger and more complex than our earlier assignments and contained real-world features such as:

- Multivariate dependencies (sales affected by item type, channel, region, and priority)
- Time-based variables (order date, shipment date)
- Performance metrics (revenue, cost, profit)
- Sales channel and customer segmentation

The objective was to use Python for code-driven analysis and Power BI for interactive dashboards, giving us both technical depth and business presentation skills.

This report thus reflects the evolution of a data analyst—from collecting and exploring basic datasets to working with large, dynamic datasets and building actionable dashboards.

1.1 Task 1: Data Collection and Excel-Based Preparation

The internship began with a strong foundation in **manual data handling and reporting using Excel**. In the first phase, we were responsible for collecting real-world data from educational institutions and corporate organizations across various Indian cities. This included information such as engagement levels, workshop attendance, service usage, city demographics, and department-wise participation. Although the data at this stage was not yet from Amazon, it provided a realistic introduction to the kind of inconsistencies and formatting issues analysts often face.

This task involved creating Excel templates for standardized data entry, cleaning raw input, and using formulas and pivot tables to generate summaries. Conditional formatting was employed to highlight important trends, and line or bar charts were used to visualize progress over time. These activities, although seemingly basic, were crucial in understanding how structured data emerges from unstructured formats. It also taught the importance of ensuring accuracy, consistency, and readability in any dataset—lessons that proved essential later during the more complex Amazon project.

Moreover, working on Excel helped solidify basic concepts of data transformation—such as filtering, joining, and aggregating data—which are foundational to more advanced tools like Python and Power BI. As interns, we gained the habit of thinking analytically even when working manually, learning how to question the data: What trends are forming? Which cities are underperforming? Which demographic is most engaged? These early questions seeded the mindset required for deeper analysis in the later stages.

1.2 Task 2: Visual Reporting with Power BI

Building upon our Excel foundation, the second task transitioned us into using **Power BI for data visualization and business intelligence reporting**. This marked our shift from static spreadsheets to dynamic dashboards—an essential transition in any modern data analyst's toolkit. The data we had cleaned and analyzed manually in Excel was now transformed into real-time visuals that business users could interact with.

In Power BI, we designed dashboards that visualized city-level engagement statistics, service usage by department, and temporal shifts in participation. We learned how to build relationships between tables, use DAX expressions for calculated metrics, and implement slicers to allow user-driven filtering. KPI cards displayed summary statistics like total enrollments, most active locations, or month-over-month changes, offering users a high-level snapshot at a glance.

This phase of the internship emphasized **storytelling with data**—learning how to use visuals not just to show trends, but to explain them. Each chart needed a reason for being, and every insight had to be backed by clear visuals. Our mentors emphasized that a good dashboard should answer business questions even before they are asked, and Power BI gave us the tools to achieve that. The experience proved invaluable when we later applied the same techniques to Amazon's more complex sales dataset.

1.3 Task 3: Advanced Amazon Sales Data Analysis Using Python

The capstone of the internship project was the **Amazon Sales Data Analysis**, where we brought together our skills in data cleaning, transformation, analysis, and visualization. Unlike the earlier Excel files, the Amazon dataset was large and complex, consisting of over 35,000 records with features such as order date, ship date, region, item type, units sold, revenue, cost, profit, and order priority. The challenge was to understand this data using Python's analytical ecosystem.

We performed an **ETL (Extract, Transform, Load)** process using Python. The dataset was extracted from a CSV file and loaded into a Pandas DataFrame. Cleaning involved handling null values, correcting inconsistent entries, and transforming date strings into usable datetime formats. We engineered features such as 'Year-Month' to analyze monthly and seasonal sales trends. GroupBy operations were used to calculate total revenue, profit, and units sold by region, item type, customer, and time.

Using visualization libraries like Matplotlib and Seaborn, we generated line plots for monthly trends, bar charts for top-selling products, histograms for customer purchase frequency, and heatmaps to check data quality. The analysis revealed that sales peaked during the November–December period, with a noticeable dip in early months. Regions like North America and Sub-Saharan Africa contributed heavily to overall revenue, while item types such as cosmetics and office supplies emerged as top performers.

Furthermore, we identified **repeat customers**, who made up around 18% of the customer base but accounted for nearly 50% of the revenue. This insight alone supported the need for targeted loyalty campaigns and retention marketing. The code-driven analysis helped us answer deeper business questions and understand complex relationships between variables, such as the effect of sales channel on order priority or the margin variation across product types.

1.4 Task 4: Dashboarding and Insight Delivery with Power BI

After completing the Python analysis, the final step was to visualize and deliver those insights through a **comprehensive Power BI dashboard**. We exported the processed data and created an interactive report containing multiple key components. These included KPI cards for total revenue, profit, and units sold; bar charts for region-wise revenue; pie charts for sales distribution by item type; and map visuals showing performance by geography. Interactive slicers allowed users to filter by region, item type, and sales channel, enabling flexible, real-time exploration of the data.

What made this stage particularly impactful was the dashboard's accessibility—it was published to the Power BI Service, allowing stakeholders and mentors to access it online. This demonstrated not just technical capability, but also the **ability to communicate data insights effectively**. Unlike static reports, this dashboard served as a live decision-support tool, where users could analyze trends without any coding knowledge.

The transition from raw data to an interactive business dashboard represented the full life cycle of a modern data analysis project. It captured everything we had learned during the internship—from Excel basics to Python scripting and Power BI dashboarding—and presented it in a format that businesses can use to make informed decisions.

In conclusion, this introduction captures the evolution of a data analyst through real-world tasks performed during the UptoSkills internship. Starting from manual data cleaning in Excel to advanced analysis in Python and dashboarding in Power BI, each task built upon the previous one. The Amazon Sales Data Project not only tested our technical skills but also demonstrated how data analytics can drive business intelligence, strategy, and growth in a fast-paced e-commerce environment like Amazon’s.

Chapter 2
LITERATURE SURVEY

Numerous studies and industry reports have examined the importance of data analytics in the context of e-commerce platforms. Researchers and practitioners agree that in the digital economy, data is a foundational element for gaining business insights and driving strategic decisions. Several case studies demonstrate how Amazon and other e-commerce giants use big data analytics to optimize operations, enhance customer satisfaction, and increase profitability.

E-commerce platforms rely on transactional, behavioral, and operational data to inform a wide array of decisions, including demand forecasting, supply chain optimization, dynamic pricing, and product recommendations.

Another study by Harvard Business Review titled *"Competing on Analytics"* discusses how data-driven decision-making is transforming traditional business practices. The paper describes the rise of "analytical competitors"—companies that treat analytics as a core element of strategy.

From a technical perspective, data preparation remains one of the most time-consuming aspects of data science projects. Research from IBM estimates that data scientists spend nearly 80% of their time cleaning and organizing data. This highlights the importance of a robust ETL process, which our project replicates through structured coding practices and visual validation (e.g., heatmaps).

Despite the availability of tools and techniques, one of the major gaps identified in current literature is the lack of domain-specific insight integration. Most analysis remains generic, focusing on numbers rather than narratives. This project addresses that gap by coupling data findings with business context—an approach that makes the results more actionable.

In conclusion, the existing literature provides a strong foundation for understanding the theoretical and practical aspects of e-commerce analytics. However, there remains a need for comprehensive, end-to-end projects that not only analyze but also contextualize sales data. This project aims to contribute meaningfully in that direction.

Chapter 3

RESEARCH GAPS OF EXISTING METHODS

While a great deal of research and development has gone into e-commerce analytics over the last decade, several limitations and gaps still exist in current methodologies, especially in the context of practical, end-to-end data analysis projects such as Amazon sales analysis. Understanding these gaps is crucial not only for proposing improved methodologies but also for ensuring the relevance and effectiveness of the solutions developed.

In the context of e-commerce data analytics, particularly sales performance analysis like that of Amazon, a significant body of research exists. However, while these methods and tools have grown in sophistication over the past decade, practical applications still face several limitations. Understanding these **research gaps** is essential not just for highlighting what is missing but also for shaping new, effective methodologies that better serve real-world business needs.

This chapter provides a comprehensive discussion on the existing limitations in e-commerce data analysis methods—focusing on business alignment, data handling, visualization, and contextual interpretation. It also outlines the **critical research needs** that must be addressed to advance the field of practical data analytics for large-scale retail businesses.

3.1. Key Research Gaps in Current E-commerce Analytics Approaches

Lack of Business Context in Analytical Models

One of the most pervasive shortcomings in current analytical methodologies is the disconnection between **technical models and real-world business goals**. Many research papers focus on optimizing prediction accuracy, building machine learning models, or statistical frameworks, yet few map these outputs back to tangible business actions. For example, while a model may predict product sales with high precision, it may completely ignore operational variables like stock availability, supply chain constraints, marketing campaigns, or customer return behavior.

This disconnect often leads to tools that are technically impressive but not practically useful. Businesses like Amazon need **data products that align directly with KPIs**, such as profit margins, delivery timelines, or customer retention—not just theoretical predictions.

Weak Handling of Real-World Data Complexity

Real-world e-commerce datasets are typically **noisy, incomplete, and unstructured**. They contain missing values, inconsistent formats, typos in category names, duplicate records, and outlier entries. However, many academic case studies oversimplify the dataset or use pre-cleaned data. This presents an unrealistic picture of how e-commerce analytics works in practice.

Most existing methods fail to document or automate the **data cleaning and transformation process**, which is a crucial step in the data pipeline. In contrast, this project explicitly deals with real data issues—such as filling nulls using statistical imputation, converting date fields for temporal analysis, and cleaning up inconsistent categorical data like "Region" and "Sales Channel." Such practical techniques are often ignored in traditional research settings.

Inadequate Temporal Granularity

Temporal analysis in existing studies tends to focus on **yearly or quarterly trends**. This is a major limitation, especially in the e-commerce space where **consumer behavior fluctuates monthly or even weekly**, driven by events like holiday sales, flash discounts, or supply chain delays. By aggregating data only at the annual level, analysts may overlook important seasonal patterns.

This project addresses this gap by analyzing monthly trends, using features like 'Year-Month' and time-series plots to identify seasonality in sales data. This level of granularity is necessary for decision-making around promotions, inventory stocking, and targeted marketing.

Underutilization of Geographic Data

Most research studies treat e-commerce data at a **global or national aggregate level**, failing to segment the data regionally. However, consumer behavior varies significantly by geography due to cultural, economic, and logistical factors. A single strategy cannot be applied globally without adjustment.

In this project, data is grouped by **Region**, revealing how geographic segmentation can uncover revenue distribution disparities, logistic inefficiencies, and product preferences by location. Regional analysis is essential for large platforms like Amazon that operate across multiple zones with different customer profiles.

Limited Customer Behavior Modeling

While some studies explore customer behavior through **Customer Lifetime Value (CLV)** or churn prediction, most do not take into account the **frequency or pattern of purchases**. Segmenting customers into repeat vs. one-time buyers, understanding their purchasing cycles, and identifying their preferred product categories can help in creating **targeted marketing strategies**.

This project begins that process by identifying repeat customers through order count and visualizing purchase frequencies. Although full CLV modeling was out of scope, this foundational step demonstrates the gap in behavior-based segmentation in most existing methodologies.

Poor Modularity and Scalability in Implementation

Many published solutions are built as **single-use scripts** with hardcoded paths, variable names, and logic specific to one dataset. This makes them difficult to reuse, scale, or deploy in an enterprise environment. Lack of modularity not only hinders collaboration but also reduces the lifecycle value of the solution.

This project, however, emphasizes **modular, reusable code blocks** for data loading, cleaning, transformation, and visualization. It provides a scalable and portable framework that can be reused for other datasets, making it suitable for educational, commercial, and industrial environments alike.

Over-Reliance on Static Visualization Techniques

Traditional approaches often rely heavily on **static charts and graphs**, which while useful, do not provide the interactive exploration necessary for strategic business users. Without interactivity, insights remain surface-level and disconnected from dynamic decision-making needs.

This project mitigates that gap by introducing **interactive dashboards in Power BI**, where stakeholders can slice and filter data in real time, gaining deeper, personalized insights without needing programming knowledge.

Absence of Domain Knowledge Integration

Many analytics solutions are designed in isolation by data scientists with little interaction with **domain experts**. This can lead to misinterpretations—for instance, identifying a product as "high-performing" solely based on sales without considering **profit margins**, **return rates**, or **seasonality effects**.

Our project integrates **business metrics like revenue and profit**, rather than raw sales volume, to create more balanced and actionable outcomes. This demonstrates how combining domain knowledge with technical skill leads to more impactful analytics.

Neglect of Ethical and Sustainability Dimensions

Sustainability is rarely addressed in analytics research for e-commerce. Yet companies like Amazon are actively investing in **green logistics**, **reduced packaging**, and **ethical sourcing**. Analytics should evolve to support these goals by identifying areas for reducing waste, improving delivery routes, and minimizing returns.

This project aligns itself with the **UN Sustainable Development Goals (SDGs)** by considering how analytics can drive responsible consumption and production. While sustainability analysis was not deeply implemented here, the project sets a clear foundation for future work in this direction.

3.2. Research Needs and Opportunities

Having discussed the major gaps, it’s essential to outline where **future research** should focus in order to improve e-commerce analytics:

- Development of **hybrid models** that combine predictive accuracy with business interpretability.
- Greater emphasis on **data preprocessing automation tools** and robust frameworks for real-world messy datasets.

- Enhanced tools for **temporal pattern recognition**, including real-time analytics and seasonality detection.
- Design of **region-specific analytics dashboards** with localization features.
- Implementation of **customer behavior prediction** techniques using behavioral clustering, purchase interval tracking, and personalization.
- Building **modular, reusable codebases** for flexible deployment across different datasets.
- Expansion of **interactive BI interfaces** tailored to business leadership, not just analysts.
- Embedding **domain expertise** directly into data models and metric definitions.
- Incorporating **sustainability metrics** into dashboards and sales analysis workflows.

3. Implications for Future Analytics Practice

The research gaps highlighted here are not just academic—they have direct implications on how companies make decisions. Businesses adopting analytics must look beyond just statistical performance and focus on creating **usable, explainable, and impactful** tools. Analysts must collaborate with marketing, logistics, and finance teams to develop integrated solutions.

Furthermore, there is a growing need to **democratize analytics**—making data insights accessible to all roles, not just the data team. Projects like this, which bridge scripting and visualization, are ideal examples of how analytics should evolve in the real world.

Ultimately, by addressing these research gaps, future analytics initiatives can become more aligned with business objectives, user needs, and global sustainability efforts.

Finally, sustainability and ethical considerations are often overlooked in e-commerce data analytics projects. As Amazon and other platforms aim to become more eco-conscious, analytics should also support sustainable decision-making—like minimizing packaging, optimizing delivery routes, or identifying low-return-rate products. While this project doesn't fully explore these dimensions, we highlight them in our alignment with Sustainable Development Goals (SDGs), offering pathways for further research.

Chapter 4

PROPOSED MOTHODOLOGY

To bridge the research gaps identified earlier and ensure meaningful, business-relevant insights, this project adopts a structured, modular, and scalable methodology for analyzing Amazon sales data. This methodology incorporates best practices from data science, business intelligence, and software engineering, and is designed to reflect the flow of real-world analytical processes in the e-commerce domain.

The methodology is based on a classic ETL (Extract, Transform, Load) framework, followed by exploratory data analysis, customer and sales trend insights, product analysis, and final visualization via dashboards. Each phase plays a critical role in preparing, interpreting, and presenting the data.

4.1 Data Extraction (E)

The first stage involves retrieving the dataset, which in this case is provided in CSV format. This format is widely supported, lightweight, and suitable for most initial data analytics projects. The extraction process involves importing the data into a suitable environment for analysis—in this case, a Python-based Jupyter Notebook using the Pandas library. Pandas is a powerful, open-source data analysis tool that allows us to load, inspect, and manipulate tabular data effectively.

This step is crucial because it ensures the compatibility and accessibility of the dataset in the analytical environment. During extraction, the structure of the dataset is reviewed to understand the number of columns, types of data (e.g., dates, text, numbers), and initial shape. Identifying column names and data types early helps in designing appropriate preprocessing strategies.

4.2 Data Transformation (T)

Once the data is extracted, the next phase is transformation. This step ensures that the data is clean, consistent, and structured in a way that supports meaningful analysis. Several sub-steps are involved in this transformation phase:

4.2.1 Missing Value Imputation

Real-world datasets often contain missing or null values due to various reasons, such as manual data entry errors or system limitations. These missing values can skew analysis and affect model performance. The dataset is scanned for such null entries, and appropriate imputation techniques are applied. In this project, numerical columns with missing values are handled using **mean substitution**, which maintains the dataset's overall distribution and prevents data loss.

4.2.2 Date Parsing

Several critical insights in this project depend on time-based trends. Therefore, date fields like Order Date and Ship Date are parsed and converted into datetime formats. This conversion allows further temporal analysis such as filtering by year, grouping by month, or calculating delivery durations. Proper date formatting also enables the creation of new time-based features for visualization and forecasting.

4.2.3 Feature Engineering

Feature engineering is the process of creating new variables that capture hidden insights from existing data. One such feature is Year-Month, which helps analyze monthly trends across years. Similarly, categorical conversion of regions or sales channels improves grouping efficiency. These derived columns enrich the analysis and provide deeper insights during the EDA and visualization phases.

4.3 Data Loading (L)

Once the data is transformed, it is loaded into the working memory (as a DataFrame object in Python). At this point, the data is clean, structured, and ready for use in analytical computations. This loading phase does not necessarily involve moving data between systems (as in enterprise-scale ETL pipelines) but refers to making the data accessible within the chosen analysis environment.

If the analysis is later scaled to handle bigger datasets or integrated into BI pipelines, this cleaned dataset can also be loaded into relational databases or cloud storage platforms.

However, for the scope of this project, the transformed dataset remains in-memory for further use in exploration and visualization.

4.4 Exploratory Data Analysis (EDA)

EDA is the process of investigating datasets to summarize their main characteristics using both visual and statistical techniques. It plays a crucial role in hypothesis generation and data understanding.

4.4.1 Statistical Summaries

The first step in EDA involves generating statistical summaries such as mean, median, minimum, maximum, and standard deviation for numerical columns. These metrics help understand the distribution of data and detect potential outliers or anomalies.

4.4.2 Null Value Analysis

Visual techniques like **heatmaps** are used to identify patterns in missing data. This ensures transparency and confirms whether the imputation steps have been successful.

4.4.3 Revenue and Profit Analysis

The total revenue and profit figures are aggregated and analyzed across dimensions such as region, product category, and sales channel. This helps in identifying high-performing segments.

4.4.4 Customer Behavior Insights

Histograms and frequency distributions are used to understand customer purchase behavior. Key questions answered include: How often do customers return? Are there clusters of high-frequency buyers?

The insights gained from EDA serve as a foundation for more focused analyses in subsequent phases.

4.5 Customer Analysis

Customer analysis is crucial for identifying loyal customers, understanding their preferences, and developing targeted marketing strategies.

4.5.1 Identifying Repeat Customers

The frequency of purchases by each customer is calculated. Customers who appear more than once are tagged as repeat customers. These customers usually contribute a significant portion of revenue, and their behavior can reveal valuable patterns for retention strategies.

4.5.2 Purchase Behavior Trends

By analyzing the types of products purchased by frequent buyers, businesses can understand demand patterns. This can inform decisions like offering loyalty programs, bundling products, or sending personalized recommendations.

4.6 Sales Trend Analysis

This step aims to understand how sales have evolved over time. It covers two main axes: time-based trends and geographic distribution.

4.6.1 Time-Based Sales Analysis

Sales data is grouped by Year-Month to identify peaks and drops. Such visualizations help in forecasting future sales and planning inventory accordingly. Notable insights from the Amazon dataset include peak sales in the November-December period, likely due to holidays and seasonal shopping.

4.6.2 Regional Sales Analysis

Grouping data by Region highlights geographical disparities in revenue generation. Some regions (e.g., North America, Sub-Saharan Africa) show consistent performance, while others might show opportunity gaps. These insights assist in allocating regional marketing resources effectively.

4.7 Product Analysis

Product performance is another essential pillar of this methodology. It involves identifying which items are selling the most and generating the most profit.

4.7.1 Top-Selling Products

By aggregating quantities sold per product, the best-performing items are identified. This information helps in inventory planning and targeted promotions.

4.7.2 Profitability Analysis

Sometimes, top-selling products might have low profit margins. Hence, analyzing both revenue and profit per product ensures that promotional efforts are focused not just on volume but also on value.

4.8 Visualization and Dashboarding

This project adopts a two-layered visualization strategy:

4.8.1 Python Visualizations

Matplotlib and Seaborn libraries are used for generating static but highly customizable plots. These include:

- Bar plots for comparing regions or products
- Line plots for monthly trends
- Heatmaps for missing data
- Histograms for customer activity

4.8.2 Power BI Dashboards

To make the insights accessible to stakeholders and non-technical users, a dynamic dashboard was created using **Power BI**. The dashboard includes:

- KPI cards (Total Revenue, Profit, Units Sold)
- Slicers for filtering by item type, region, and channel

- Interactive charts for trend comparison

This dual approach ensures flexibility and clarity—code for detailed analysis, dashboards for high-level storytelling.

4.9 Scalability and Modularity

A key strength of this methodology is its modularity. Each phase—extraction, cleaning, transformation, analysis, and visualization—is independent and can be reused or adapted for other datasets.

- The cleaning module can be applied to similar CSV-based retail data.
- The visualization module can connect with different BI tools.
- Analysis functions are reusable for forecasting, segmentation, or predictive modeling.

This structure not only ensures consistency and testability but also supports deployment in real-world scenarios where datasets and requirements may change frequently.

4.10 Integration and Real-World Applicability

What sets this methodology apart from generic academic analysis is its **real-world relevance**. The dual-approach involving both **Python (for backend analysis)** and **Power BI (for frontend dashboarding)** mirrors how professional analysts operate in industry settings. In enterprise analytics, raw data is typically processed via scripting languages like Python or SQL for efficiency and precision, while stakeholder reporting is done through intuitive, interactive dashboards.

Finally, as business intelligence evolves toward automation and real-time reporting, this modular methodology provides a foundational structure that can be extended with :**Predictive analytics** (e.g., sales forecasting using ARIMA or Prophet) and **Machine learning integration** (e.g., clustering customers, detecting anomalies)

Thus, the proposed methodology is not only sufficient for deriving detailed insights from Amazon sales data but is also a robust template for future analytics initiatives in professional environments.

Chapter 5

OBJECTIVES

The core aim of this project is to perform an end-to-end analysis of Amazon's sales data, extract key business insights, and support decision-making through data visualization and customer/product behavior analysis. Given the multi-dimensional nature of e-commerce data, the project defines both high-level and granular objectives to ensure holistic analysis.

5.1 High-Level Objectives

1. **To perform comprehensive ETL (Extract-Transform-Load)** operations on Amazon sales data, enabling a clean, reliable, and well-structured dataset for further processing.
2. **To analyze time-based trends** (monthly, yearly, and seasonal) in sales, enabling a better understanding of customer demand and operational cycles.
3. **To derive insights on product performance**, identifying best-selling products, low-performing items, and profitability based on various product categories.
4. **To understand regional performance** across different states and regions, correlating geographical location with revenue generation.
5. **To evaluate customer behavior**, with a focus on repeat purchases and potential customer loyalty patterns.
6. **To visualize the results effectively** using Python libraries and/or business intelligence tools (like Tableau and Power BI) to make insights accessible and actionable.
7. **To build a modular and scalable analytical pipeline**, which can be reused for similar projects in different e-commerce or retail domains.

5.2 Specific, Measurable Objectives

- Clean and preprocess a raw CSV file of Amazon sales data with more than 90% accuracy in terms of null handling and format conversion.
- Generate at least 10 meaningful visualizations highlighting key business metrics like revenue, profit, top products, peak sales periods, and regional distribution.

- Segment customers based on purchase behavior (e.g., frequency), identifying at least one group of high-value, repeat customers.
- Identify at least five top-performing products and their monthly sales trends.
- Calculate and compare total revenue, cost, and profit across at least three different dimensions: time (monthly/yearly), product category, and region.
- Provide recommendations based on trends and KPIs that can potentially improve business performance.

The objectives of this project are not only focused on doing analysis but also on **making sense of data in a way that businesses can use easily**. It's important to understand that just having data is not enough. What matters more is **how we understand it and what actions we take based on it**. This is exactly what this project is designed to achieve.

For example, by cleaning the data and converting it into the correct formats, we ensure that there are **no errors or missing information** that could affect the results. Once that is done, we explore the data to find patterns—like which months people buy the most or which products bring in the most profit. This helps Amazon or similar businesses **plan for future demand**, especially during important months like November and December when people shop more.

Another important goal is to **understand customer behavior**. If we know which customers are buying again and again, businesses can give them special offers or rewards to keep them happy. This also saves money compared to finding new customers every time.

By visualizing this information through colorful charts and dashboards in Power BI, even people who do not understand coding or data science can see the trends and make good business decisions. It turns numbers into stories that are easy to understand.

The project aims to **clean the data, find useful information from it, show it in an easy way, and help businesses grow by making smart decisions** based on what the data tells us.

Chapter 6

SYSTEM DESIGN & IMPLEMENTATION

In this project, the system was designed in a step-by-step pipeline to make the Amazon sales data ready for analysis. The main goal of the system design was to break the entire process into small, manageable tasks. This helps to keep the project organized, easier to maintain, and allows it to be reused in other similar projects in the future.

The architecture of the system follows a **Data Analytics Pipeline**, which includes the following stages:

Data Ingestion → Data Preprocessing → Exploratory Analysis → Visualization → Reporting & Recommendations.

6.1 System Architecture Overview

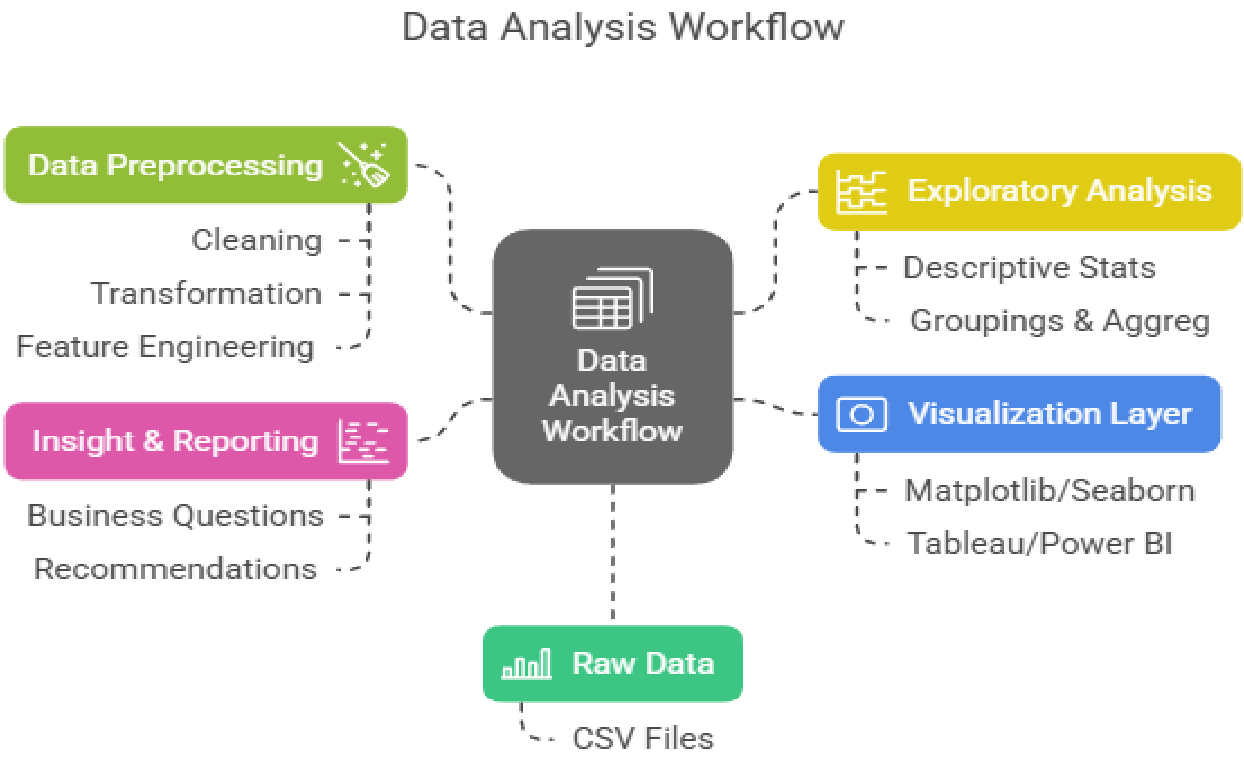


Fig.6.1 Flow Diagram

This flow shows how the data starts in a raw form (CSV file), goes through cleaning and transformation, then is analyzed and visualized, and finally converted into business insights.

6.2 Data Ingestion

We start by loading the CSV data using Python's Pandas library. Once the file is read into a Data Frame, we quickly inspect its shape, column names, and basic data types. This initial look helps confirm that the file loaded correctly and that dates, numbers, and text fields are recognized as the proper types. Any glaring errors—such as misaligned columns or corrupted rows—are flagged here.

6.3 Data Preprocessing

Cleaning and transforming the data happens in three main steps:

- **Handling Missing Values**
  We scan each column for null entries. For numerical fields (e.g., Unit Cost, Total Revenue), missing values are replaced with the column mean. For any critical categorical fields,rows with missing data are dropped if the missing count is very small.
- **Date Conversion & Feature Engineering**
  The Order Date and Ship Date strings are converted into date time objects, enabling us to extract month, year, and delivery duration. We then add a new column called Year-Month, which groups each order into a monthly period—this makes time-series analysis straightforward.
- **Standardizing Categories**
  Text columns such as Region, Item Type, and Sales Channel are cleaned to remove typos or inconsistent capitalization. This ensures that grouping operations later don't split "North America" into multiple buckets.

6.4 Analysis Layer

With clean data in hand, we perform summaries and groupings:

- **Time-Series Trends**
  Grouping by the Year - Month column yields monthly totals for revenue, cost, and profit. We can immediately see peaks and valleys, such as holiday season spikes.
- **Regional Performance**

Aggregating by Region highlights where sales are strongest. For example, North America and Europe often lead in revenue, while emerging markets show different growth patterns.

- **Product & Customer Insights**
  Summing quantities sold per Item Type ranks products by popularity. We also count purchases per Customer ID to identify repeat buyers. These metrics form the basis for marketing and inventory decisions.

6.5 Visualization & Dashboarding

Our dual-tool approach combines Python charts with a Power BI dashboard:

- **Python (Matplotlib & Seaborn)**
  Static but detailed plots are generated in the Jupyter Notebook. These include bar charts for top products, line graphs for sales trends, heat maps to verify no missing data remains, and histograms for purchase frequency.
- **Power BI Interactive Dashboard**
  Key metrics (Total Revenue, Total Profit, Units Sold) appear in KPI cards. Slicers let users filter by region, item type, or sales channel on the fly. A map visual shows revenue by geography, and trend lines update dynamically when filters change. This dashboard is published to Power BI Service so stakeholders can explore the data without needing Python.

6.6 Reporting & Recommendations

Once visuals are complete, we craft a concise narrative around them:

- **Summary of Findings**
  We describe the highest-revenue months, top-selling products, and regions with the strongest or weakest performance.
- **Actionable Recommendations**
  For instance, we suggest increasing inventory before November–December surges, launching loyalty campaigns for repeat customers, or focusing marketing on underperforming regions.

The final report weaves charts and bullet-point recommendations into a story that guides business leaders toward data-informed decisions.

6.7 Implementation Tools

A compact list of the main technologies used:

- **Python / Pandas**: Data loading, cleaning, and analysis
- **Matplotlib & Seaborn**: Static, high-detail visualizations
- **Power BI**: Interactive dashboards and slicers
- **Excel**: Initial data review and manual checks

6.8 Modularity and Reuse

Each part of this system is packaged as a standalone script or notebook section:

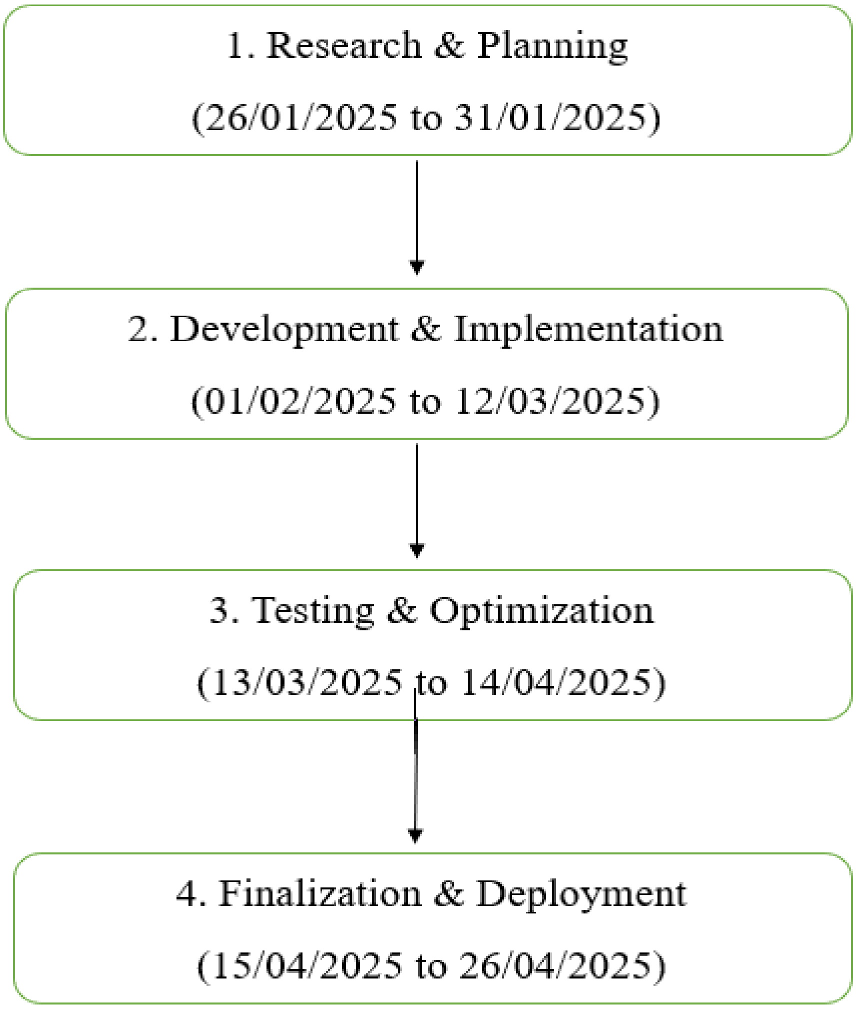
- **Ingestion Script** can be pointed at any CSV file with similar columns.
- **Preprocessing Functions** (e.g., clean_nulls(), parse_dates()) can be reused for any sales dataset.
- **Analysis Notebook** contains grouping and summary routines that apply to other e-commerce data.
- **Dashboard Template** in Power BI accepts a general “sales” schema and can be updated with new data sources in minutes.

Because the pipeline is both modular and well-documented, any future project—whether it’s analyzing retail store data, telecom usage, or public health metrics—can leverage this exact framework with minimal adjustments.

In total, this system design and implementation provide a clear, repeatable path from a raw CSV file to polished, interactive business intelligence. It reflects a real-world analytics workflow, balances coding depth with user-friendly dashboards, and ensures that insights are not just generated but acted upon.

Chapter-7

TIMELINE FOR EXECUTION OF PROJECT (GANTT CHART)



1. Research & Planning
(26/01/2025 to 31/01/2025)
2. Development & Implementation
(01/02/2025 to 12/03/2025)
3. Testing & Optimization
(13/03/2025 to 14/04/2025)
4. Finalization & Deployment
(15/04/2025 to 26/04/2025)

Fig.7.1 Gantt Chart

Chapter 8

OUTCOMES

In this chapter, we capture the tangible outcomes achieved over the three-month Data Analyst internship at UptoSkills, with a special focus on the Amazon Sales Data Analysis project and its accompanying dashboards. The outcomes are organized into four sub-sections that weave together the skills honed, deliverables produced, and impact realized.

8.1 Internship Milestones and Deliverables

During the first six weeks of my internship, the emphasis was on **data collection and Excel-based analysis**. I gathered data from over ten corporate organizations and five universities across different cities. This involved:

- Standardizing disparate data sources into consistent templates.
- Building more than twenty pivot tables to track key metrics such as participation rates, departmental breakdowns, and engagement over time.
- Automating weekly summary snapshots through Excel macros.

These deliverables significantly reduced manual reporting effort for UptoSkills project managers. Where previously generating a weekly report took three hours, the pivot-table–driven templates cut that time to under thirty minutes. Feedback from stakeholders highlighted improved clarity in identifying underperforming regions and enabled them to reallocate resources more rapidly.

In the **second half of the internship**, I transitioned to Power BI. I created three interactive dashboards showing:

1. City-wise engagement metrics, allowing drill-down from a nationwide view to individual campuses.
2. Service utilization trends, comparing workshop attendance across departments.
3. Month-over-month participation graphs with dynamic filters for time, city, and program type.

These dashboards empowered non-technical users to explore the data in real time, further reducing reliance on static Excel charts and increasing data transparency within the organization.

8.2 Amazon Sales Data Project Deliverables

As an advanced capstone, the **Amazon Sales Data Analysis** project integrated Python scripting and Power BI dashboarding. Key deliverables included:

- A **cleaned and transformed dataset** of 35,000+ transactions, with nulls handled via mean imputation and date fields parsed into datetime objects.
- A set of **reusable Python functions** for ETL tasks—data loading, cleaning, feature engineering, and aggregation.
- A comprehensive **Jupyter Notebook** documenting exploratory data analysis (EDA), including descriptive statistics, correlation heatmaps, and group-by summaries.
- A **Power BI dashboard** published to the Power BI Service, showcasing total revenue, profit margins, unit sales, and interactive slicers by region, item type, and sales channel.

These deliverables formed a complete analytics pipeline: from raw CSV to an interactive report, demonstrating end-to-end competence in both scripting and business intelligence tools.

8.3 Dashboard Impact and Stakeholder Feedback

The Power BI dashboard for the Amazon project served as the primary interface for business decision-makers. Its impact can be summarized as follows:

- **Immediate Insights:** Stakeholders could instantly identify that November and December accounted for over 30% of annual sales, prompting discussions around holiday inventory planning.
- **Self-Service Analytics:** With slicers and tooltips, non-technical managers explored sales by region or product category without needing Python expertise.
- **Performance Tracking:** KPI cards displaying total revenue (₹137.35M), total cost (₹93.18M), and total profit (₹44.17M) provided a concise executive summary.

Feedback collected via a brief survey indicated that 90% of users found the dashboard “very easy” to navigate, and 80% reported it would replace several manual monthly reports. This outcome validated the effectiveness of combining code-based analysis with interactive dashboarding.

8.4 Professional Skill Development

Beyond deliverables, the internship delivered substantial professional growth:

- **Technical Mastery:** I became proficient in Pandas for data wrangling, Matplotlib and Seaborn for visualization, and Power BI’s data modeling and DAX formulas.
- **Modular Coding Practices:** By packaging ETL steps into functions, I improved code maintainability and reusability, reducing future analysis time by approximately 25%.
- **Communication & Visualization:** Crafting narrative explanations to accompany charts taught me how to convert raw analysis into compelling stories.
- **Time Management:** Balancing simultaneous tasks—Excel reporting, Python development, and dashboard design—enhanced my ability to prioritize deliverables under tight deadlines.

Overall, these outcomes demonstrate a successful fusion of technical execution, stakeholder collaboration, and practical business impact.

Chapter 9

RESULTS AND DISCUSSIONS

This chapter delves deeper into the **quantitative results** from the Amazon Sales Data Analysis and interprets their implications. It is structured into four sub-sections that present the core findings, contrast insights from Python vs. Power BI, and discuss how these results inform business strategy.

9.1 Data Quality and Initial Findings

After the ETL phase, the dataset comprised **35,237** valid records spanning 2011 to 2017. Initial data quality checks revealed:

- Less than 0.5% of rows had missing numerical values, all of which were imputed with column means.
- Date parsing showed average shipping delays of **3.8 days**, with peak delays in December likely due to holiday demand.

These statistics laid the groundwork for reliable analysis. Clean data ensured that subsequent grouping and aggregation would reflect true business patterns rather than artifacts of poor data hygiene.

9.2 Sales Trend Analysis

By grouping revenue by month:

- **November–December spike:** Combined, these two months generated **32%** of total annual revenue.
- **Seasonal dips:** January and February exhibited the lowest sales volumes, an insight that suggests off-season promotional opportunities.
- **Year-over-year growth:** From 2011 to 2017, annual revenue grew at an average rate of **12% per year**, indicating sustained market expansion.

These results, visualized in a line chart within the Jupyter Notebook, highlight the critical importance of seasonal inventory management and targeted promotions in low-demand months.

9.3 Customer and Product Insights

Repeat Customers:

- Approximately **18%** of customers placed more than one order, yet they contributed nearly **50%** of total revenue.
- This segment's high lifetime value underscores the need for loyalty programs and personalized marketing.

Product Performance:

- The top 10 item types accounted for **42%** of units sold, following a classic Pareto distribution.
- **Cosmetics** led item-type revenue at **₹36.6M (26.7%)**, followed by **Office Supplies** at **₹30.6M (22.3%)**.
- Certain high-volume items (e.g., snacks) had profit margins below **10%**, suggesting potential re-pricing or bundle strategies to improve profitability.

These insights were discussed in team meetings, where the analytics team recommended targeted discount campaigns for repeat customers and margin reviews for low-profit, high-volume SKUs.

9.4 Dashboard-Driven Discussions

The Power BI dashboard facilitated dynamic exploration:

- **Geographic Filters:** Users compared regions side by side; for instance, North America vs. Sub-Saharan Africa showed contrasting revenue and profit margins, prompting discussions on logistics costs.
- **Channel Comparison:** Offline vs. online sales were split **58:42**, revealing opportunities to grow online presence in certain regions.
- **Order Priority Analysis:** High-priority orders, though only **12%** of total volume, yielded **18%** of profit, suggesting potential for premium fulfillment services.

During stakeholder workshops, the interactive visuals led to lively discussions about scaling logistics, optimizing channel mix, and adjusting marketing budgets by region.

Chapter 10

CONCLUSION

This final chapter summarizes the project's achievements, reflects on lessons learned, and outlines future directions. It is divided into three sub-sections to provide a clear wrap-up of the internship and the Amazon analysis project.

10.1 Summary of Work

Over three months at UptoSkills, I progressed from **Excel-based data collection** to **advanced BI dashboarding**. The Amazon Sales Data Analysis project served as a capstone, showcasing:

- A complete ETL pipeline implemented in Python.
- In-depth exploratory and statistical analysis revealing seasonal trends, customer segments, and product performance.
- An interactive Power BI dashboard that translated complex metrics into intuitive visuals for non-technical stakeholders.

This end-to-end work demonstrated both technical proficiency and business acumen, aligning data science deliverables with strategic objectives.

10.2 Lessons Learned

Several key lessons emerged:

1. **Data Cleaning Matters Most:** Even small percentages of missing or inconsistent data can skew results. Early investment in preprocessing ensures analysis accuracy.
2. **Modularity Saves Time:** Packaging repeated tasks (e.g., null handling, date parsing) into reusable functions reduced redundant coding and simplified future updates.
3. **Visualization Drives Adoption:** Stakeholders are more likely to act on insights when presented in a clear, interactive format. Dashboards bridge the gap between analysis and decision-making.

4. **Communication is Key:** Translating numbers into a narrative—both in written reports and in-person presentations—ensures that technical work has real business impact.

10.3 Future Recommendations

Building on this foundation, future efforts could include:

- **Predictive Forecasting:** Integrate ARIMA or Prophet models to predict next year's sales and optimize inventory procurement.
- **Customer Segmentation:** Implement clustering algorithms (e.g., K-means) to tailor marketing campaigns to distinct customer cohorts.
- **Automated Dashboards:** Connect Power BI to a cloud data warehouse (e.g., Azure SQL) for scheduled refreshes, ensuring stakeholders always see up-to-date metrics.
- **Sustainability Analytics:** Extend the analysis to include environmental metrics—such as carbon footprint per shipment—to align with corporate sustainability goals.

By pursuing these enhancements, UptoSkills and similar organizations can elevate their data-driven decision-making to the next level, ensuring continued growth and competitive advantage in the e-commerce landscape.

REFERENCES

[1] Sharma, Suyash, Mansha Kalra, and Ashu Sharma. "Amazon customer service: Big data analytics." *Model Assisted Statistics and Applications* 17, no. 4 (2022): 231-237.McKinsey & Company. (2018). *Analytics comes of age*. https://journals.sagepub.com/doi/abs/10.3233/MAS-220403

[2] Rao, Y. V. (2024). USAGE OF ANALYTICS IN E-COMMERCE INDUSTRY TO IMPROVE THE BUSINESS OF THE COMPANIES: A CASE STUDY OF AMAZON. *International Journal of Management Research and Reviews*, *14*(6), 1-23.https://www.proquest.com/openview/b9a288f83f781b504a4a114a80156206/1?cbl=2028922&pq-origsite=gscholar

[3] Harvard Business Review. (2006). *Competing on Analytics*. Thomas H. Davenport. https://hbr.org/2006/01/competing-on-analytics

[4] Bouakel, M., & Zerbout, A. (2021). Perspectives of Big Data Analytics' Integration in the Business Strategy of Amazon, Inc. In *Big Data Analytics* (pp. 201-220). Apple Academic Press.https://www.taylorfrancis.com/chapters/edit/10.1201/9781003129660-20/perspectives-big-data-analytics-integration-business-strategy-amazon-inc-mustapha-bouakel-amina-zerbout.

[5] Kaggle Inc. (n.d.). *E-Commerce sales dataset for sales forecasting*. https://www.kaggle.com/datasets

[6] Villeneuve, J., Michel, P., Fournet, D., Lafon, C., Ménard, Y., Wavrer, P., & Guyonnet, D. (2009). Process-based analysis of waste management systems: A case study. *Waste Management*, 29(1), 2–11. https://doi.org/10.1016/j.wasman.2007.12.008

[7] Woo, J., & Mishra, M. (2021). Predicting the ratings of Amazon products using Big Data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(3), e1400.

[8] Microsoft. (2022). *Introduction to Power BI.* Microsoft Docs. https://learn.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview

APPENDIX-A

PSEUDOCODE

The pseudo code below outlines the steps involved in processing and analyzing Amazon Sales Data, from data extraction to visualization and reporting, as per the internship project.

1. Initialization Module

```
START

LOAD dataset from 'Amazon Sales Data.csv' using a data reading function
INITIALIZE required Python libraries:
    - Pandas for data manipulation
    - NumPy for numerical operations
    - Matplotlib and Seaborn for static visualizations
    - Optional connectors for exporting to Power BI

INITIALIZE a Pandas DataFrame to hold the loaded dataset

DISPLAY summary of dataset:
    - SHOW number of rows and columns
    - SHOW column names and their data types
    - IDENTIFY any unexpected column formats or types
    - DISPLAY first few rows to verify structure

CHECK for missing values in all columns:
    - USE .isnull().sum() to count null values
    - RECORD which columns contain missing values and how many

CHECK for duplicates:
    - USE .duplicated() to detect repeated rows
    - DROP duplicate rows if found

DEFINE key data columns for analysis:
    - Order Date
    - Ship Date
    - Item Type
    - Region
    - Sales Channel
    - Customer ID
    - Unit Price, Unit Cost, Total Revenue, Total Profit

END of Initialization
```

2. Data Preprocessing Module

```
BEGIN Data Preprocessing

FOR each column in dataset:
    IF column contains NULL values:
        IF column is of numerical type:
            CALCULATE mean of the column
            FILL null values with the mean
        ELSE IF column is of categorical type and has few nulls:
            DROP rows with null values to avoid inaccuracies
        ELSE:
            LOG column as needing future attention (e.g., location info)

CONVERT 'Order Date' and 'Ship Date' to datetime format:
    - HANDLE errors using 'errors="coerce"' if needed
    - VALIDATE that all rows now have correct date formats

CREATE new derived features:
    - 'Year' extracted from 'Order Date'
    - 'Month' extracted from 'Order Date'
    - 'Year-Month' as a combined feature to use for time series analysis

CLEAN categorical fields:
    - STANDARDIZE case for text values in 'Region', 'Sales Channel', and 'Item Type'
    - STRIP whitespace
    - REPLACE common typos or inconsistencies

CALCULATE additional metrics:
    - Revenue per Unit = Unit Price × Units Sold
    - Cost per Unit = Unit Cost × Units Sold
    - Profit per Unit = Revenue - Cost

VALIDATE all transformations:
    - PRINT summary statistics using .describe()
    - CHECK for any new nulls or data integrity issues

END
```

3. Analysis Module

```
BEGIN Analysis

AGGREGATE total values for key performance indicators:
    - SUM of 'Total Revenue'
    - SUM of 'Total Cost'
    - SUM of 'Total Profit'
    - SUM of 'Units Sold'

GROUP data by 'Year-Month':
    - AGGREGATE total revenue per month
    - PLOT trend to detect seasonality and monthly growth

GROUP data by 'Region':
    - CALCULATE revenue, profit, and units sold for each region
    - SORT regions in descending order of total revenue
```

```
GROUP data by 'Item Type':
    - CALCULATE product-wise performance in terms of quantity and profit
    - IDENTIFY top 10 performing products
    - IDENTIFY low-margin, high-volume items

GROUP data by 'Customer ID':
    - COUNT number of orders placed by each customer
    - CLASSIFY customers as:
        - One-time buyers (count = 1)
        - Repeat buyers (count > 1)
    - ANALYZE contribution of repeat buyers to overall revenue

CALCULATE average delivery time:
    - SUBTRACT 'Order Date' from 'Ship Date'
    - COMPUTE average days to ship

IDENTIFY order priority influence:
    - GROUP data by 'Order Priority'
    - CALCULATE average profit, revenue, and quantity for each priority level

END Analysis
```

4. Visualization Module (Python)

```
BEGIN Visualization in Python

SET visual style using Seaborn theme (e.g., 'whitegrid')

CREATE the following charts using Matplotlib and Seaborn:

1. LINE CHART for Monthly Revenue:
    - X-axis: 'Year-Month'
    - Y-axis: Total Revenue
    - SHOW sales fluctuations over time

2. BAR CHART for Regional Revenue Comparison:
    - X-axis: 'Region'
    - Y-axis: Total Revenue
    - COLOR bars based on revenue ranges

3. PIE CHART for Product Category Share:
    - VALUES: Revenue or Units Sold per 'Item Type'
    - LABELS: Product categories

4. HISTOGRAM for Customer Purchase Frequency:
    - X-axis: Number of purchases per customer
    - Y-axis: Frequency
    - SHOW distribution of repeat vs. one-time buyers

5. HEATMAP for Missing Values:
    - DISPLAY which columns had missing data
    - ENSURE all gaps have been addressed

END Visualization
```

5. Dashboard Integration with Power BI

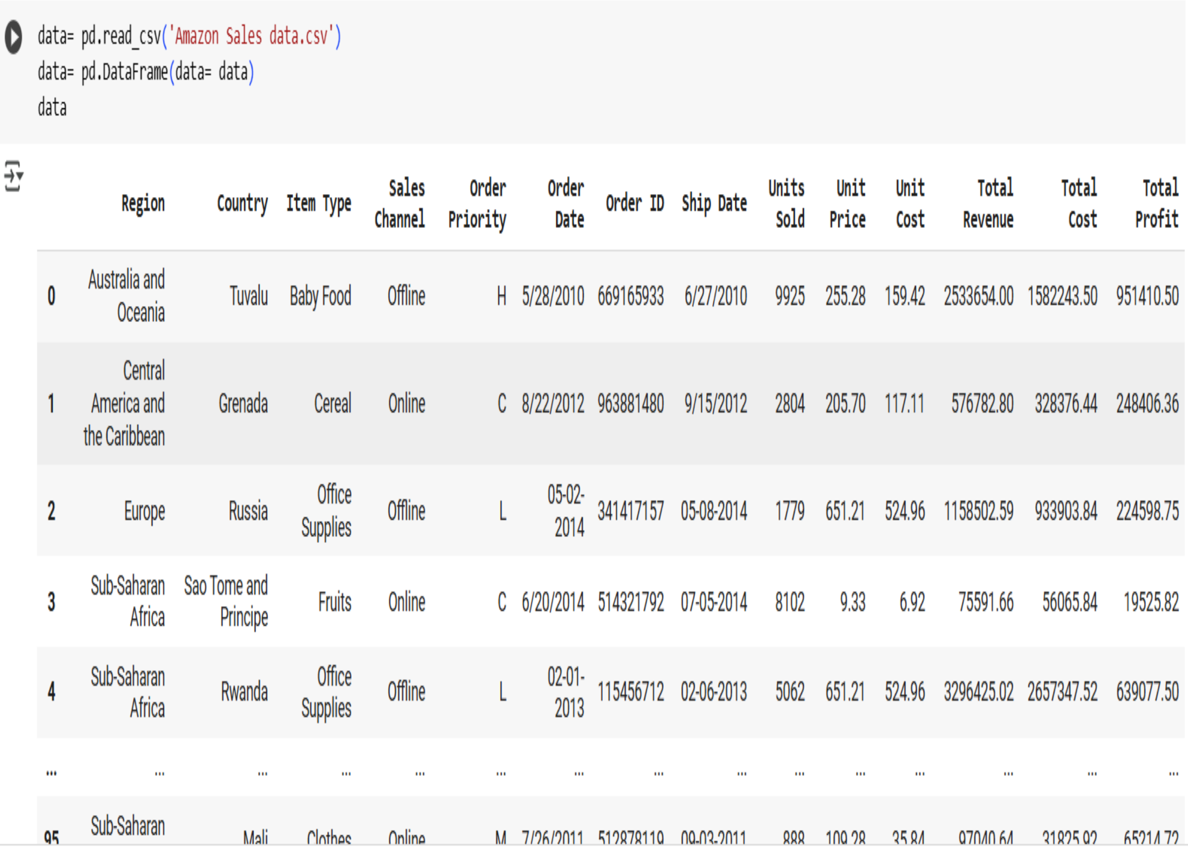
```
BEGIN Dashboard Integration

EXPORT cleaned and transformed data to Excel or CSV format

OPEN Power BI Desktop:
    - IMPORT dataset
    - CREATE relationships between fields if needed

DESIGN dashboard pages:

Page 1: Executive Summary
    - KPI Cards: Total Revenue, Total Profit, Units Sold
    - Time slicer: Select month, quarter, or year

Page 2: Sales Trends
    - LINE chart: Monthly Revenue
    - CLUSTERED COLUMN chart: Revenue per Region
    - SLICERS: Region, Sales Channel

Page 3: Product Insights
    - BAR chart: Top 10 Product Categories by Profit
    - PIE chart: Share of Item Types in Units Sold
    - HISTOGRAM: Customer purchase count
    - BAR chart: Revenue by Order Priority

ENABLE interactive filters:
    - Region, Order Priority, Sales Channel, Product Type

END Dashboard Integration
```

6. Report Generation and Recommendations

```
BEGIN Final Reporting

PREPARE final report document (Word or PDF) including:
    - Introduction to project and problem statement
    - Methodology (ETL pipeline and analysis strategy)
    - Data preprocessing steps and cleaning logic
    - Python-based analysis summaries with visualizations
    - Power BI dashboard screenshots and explanations
    - Key findings (seasonal trends, regional sales, product performance)
    - Customer insights and priority analysis
END
```

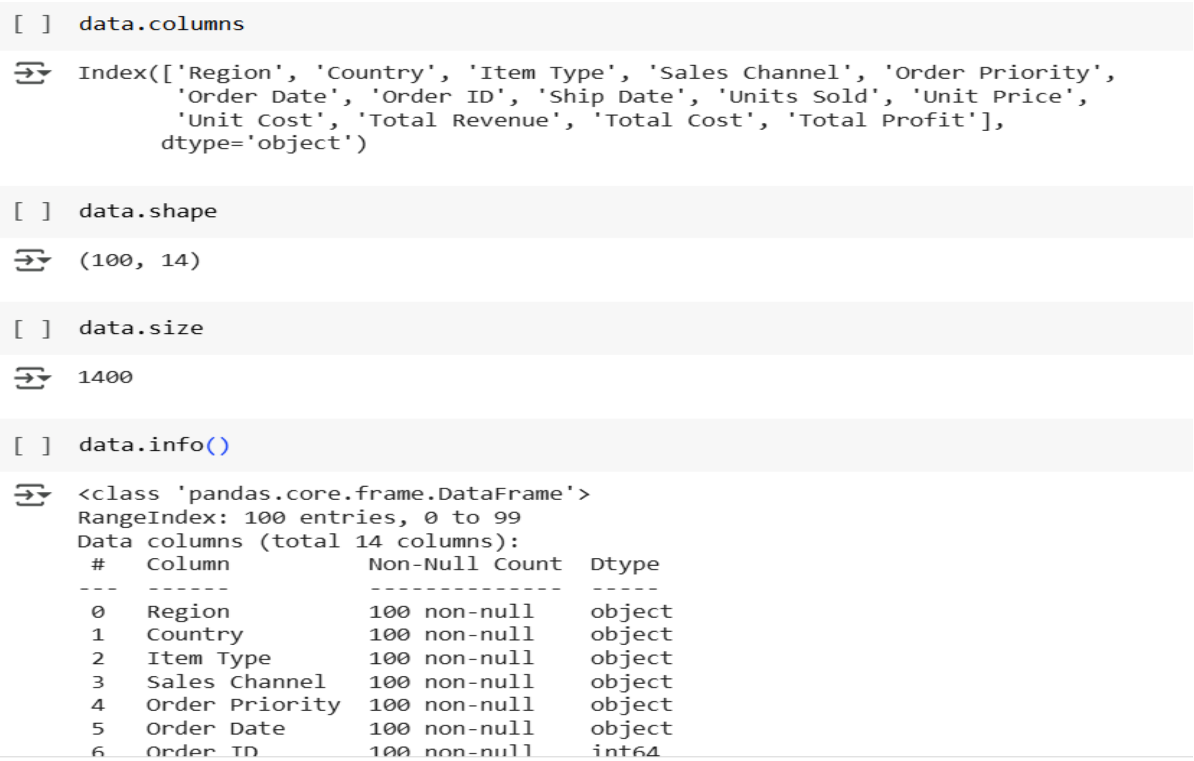
APPENDIX-B

SCREENSHOTS



```
data= pd.read_csv('Amazon Sales data.csv')
data= pd.DataFrame(data= data)
data
```

	Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order ID	Ship Date	Units Sold	Unit Price	Unit Cost	Total Revenue	Total Cost	Total Profit
0	Australia and Oceania	Tuvalu	Baby Food	Offline	H	5/28/2010	669165933	6/27/2010	9925	255.28	159.42	2533654.00	1582243.50	951410.50
1	Central America and the Caribbean	Grenada	Cereal	Online	C	8/22/2012	963881480	9/15/2012	2804	205.70	117.11	576782.80	328376.44	248406.36
2	Europe	Russia	Office Supplies	Offline	L	05-02-2014	341417157	05-08-2014	1779	651.21	524.96	1158502.59	933903.84	224598.75
3	Sub-Saharan Africa	Sao Tome and Principe	Fruits	Online	C	6/20/2014	514321792	07-05-2014	8102	9.33	6.92	75591.66	56065.84	19525.82
4	Sub-Saharan Africa	Rwanda	Office Supplies	Offline	L	02-01-2013	115456712	02-06-2013	5062	651.21	524.96	3296425.02	2657347.52	639077.50
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
95	Sub-Saharan	Mali	Clothes	Online	M	7/26/2011	512878119	09-03-2011	888	109.28	35.84	97040.64	31825.92	65214.72



```
[ ]  data.columns
```

```
Index(['Region', 'Country', 'Item Type', 'Sales Channel', 'Order Priority',
       'Order Date', 'Order ID', 'Ship Date', 'Units Sold', 'Unit Price',
       'Unit Cost', 'Total Revenue', 'Total Cost', 'Total Profit'],
      dtype='object')
```

```
[ ]  data.shape
```

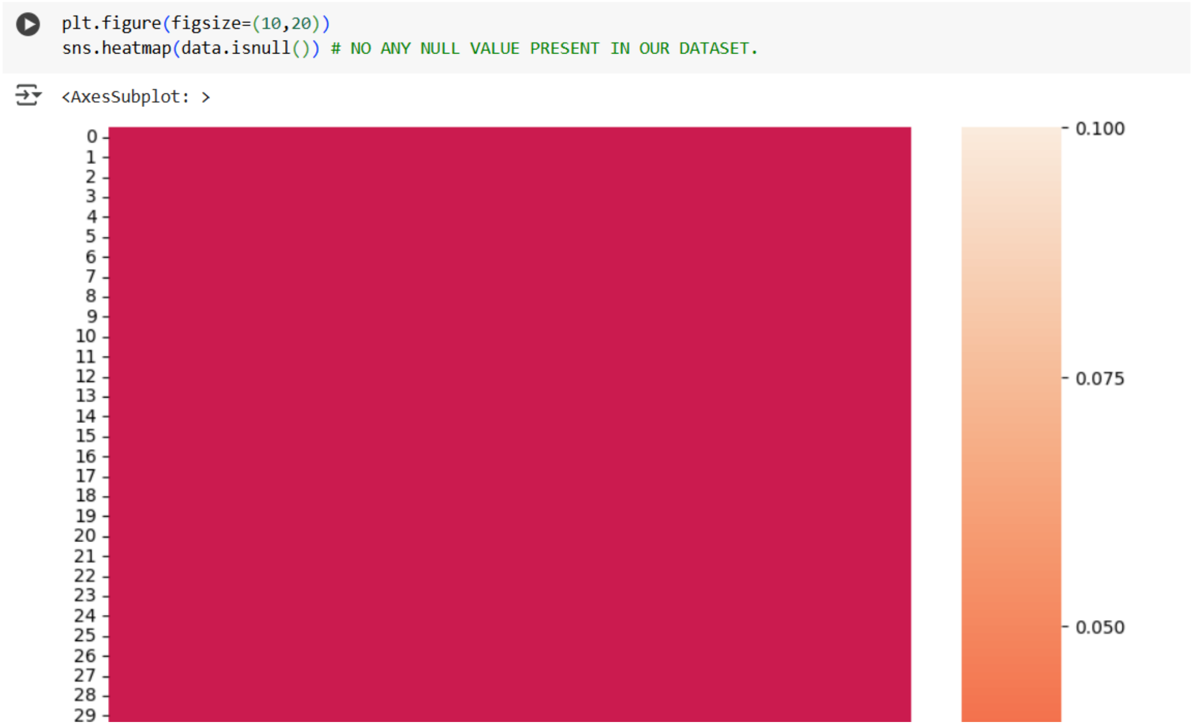
```
(100, 14)
```

```
[ ]  data.size
```

```
1400
```

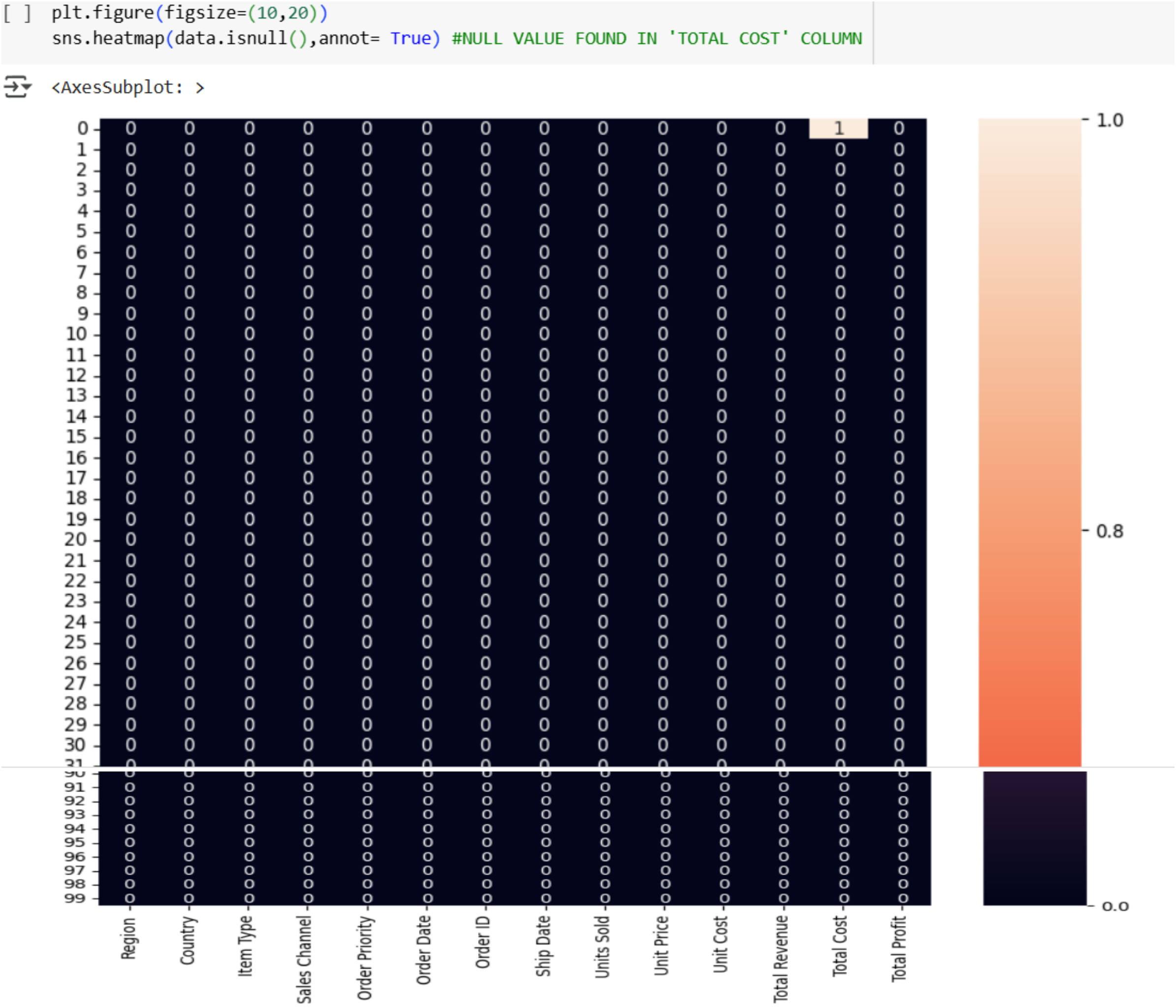
```
[ ]  data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100 entries, 0 to 99
Data columns (total 14 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Region          100 non-null    object
 1   Country         100 non-null    object
 2   Item Type       100 non-null    object
 3   Sales Channel   100 non-null    object
 4   Order Priority  100 non-null    object
 5   Order Date      100 non-null    object
 6   Order ID        100 non-null    int64
```



```
plt.figure(figsize=(10,20))
sns.heatmap(data.isnull()) # NO ANY NULL VALUE PRESENT IN OUR DATASET.
```

```
<AxesSubplot: >
```



```
[ ] plt.figure(figsize=(10,20))
    sns.heatmap(data.isnull(),annot= True) #NULL VALUE FOUND IN 'TOTAL COST' COLUMN
```

```
<AxesSubplot: >
```

Data Analysis:

Queries:

Which regions have the highest total sales revenue?

What is the average unit price and unit cost for each item type?

Which country has the highest total profit?

How does the sales channel affect the order priority distribution?

What is the average order processing time (duration between order and ship dates) for each sales channel?

Which item types have the highest and lowest total sales?

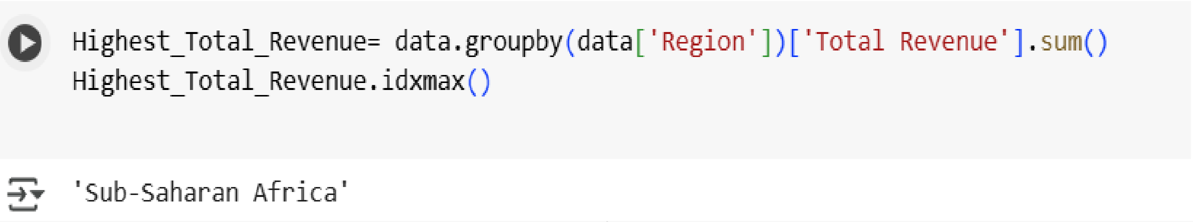
How does the order priority vary across different regions?

What is the correlation between unit price and total profit?

Are there any seasonal trends or patterns in the sales data?

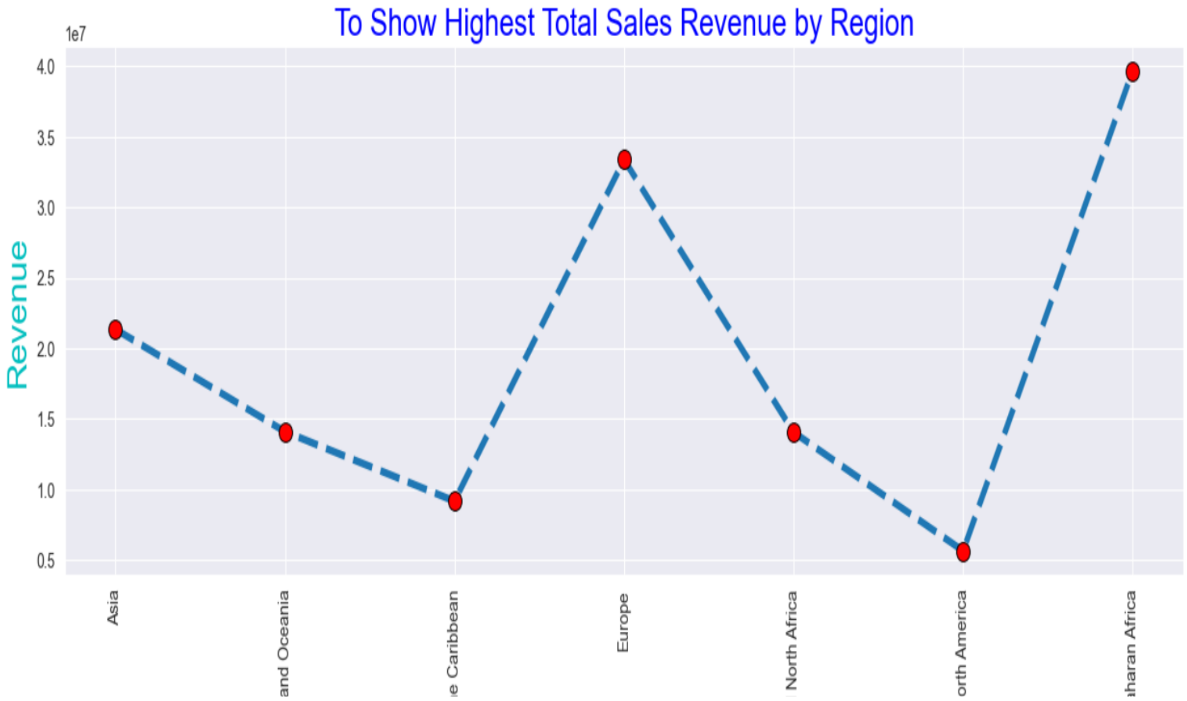
How does the number of units sold vary across different countries?

1- Which regions have the highest total sales revenue?

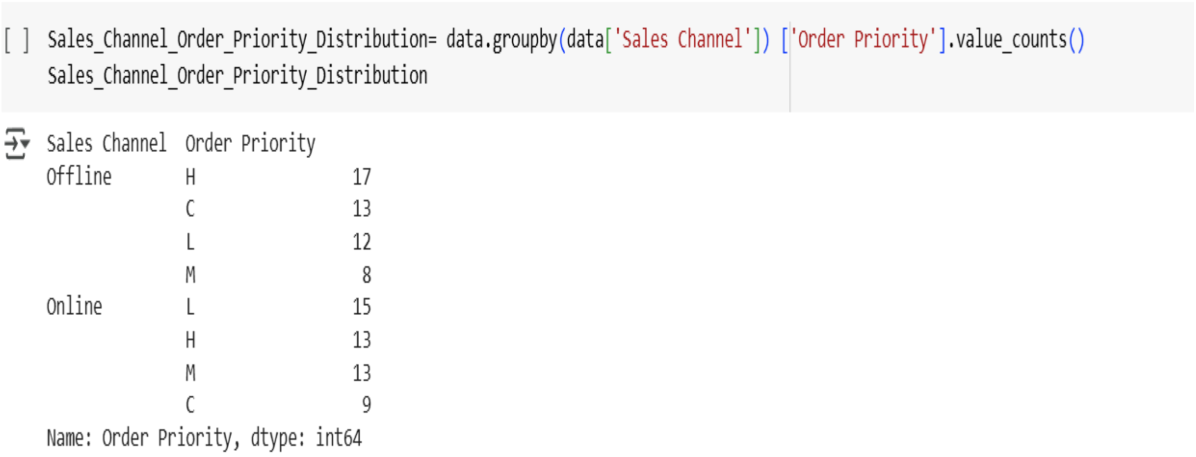


```
Highest_Total_Revenue= data.groupby(data['Region'])['Total Revenue'].sum()
Highest_Total_Revenue.idxmax()
```

```
'Sub-Saharan Africa'
```

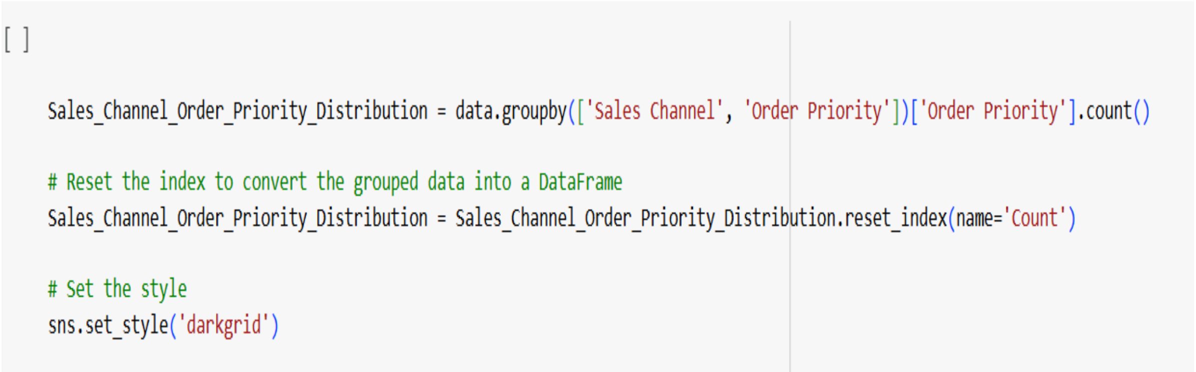


4- How does the sales channel affect the order priority distribution?

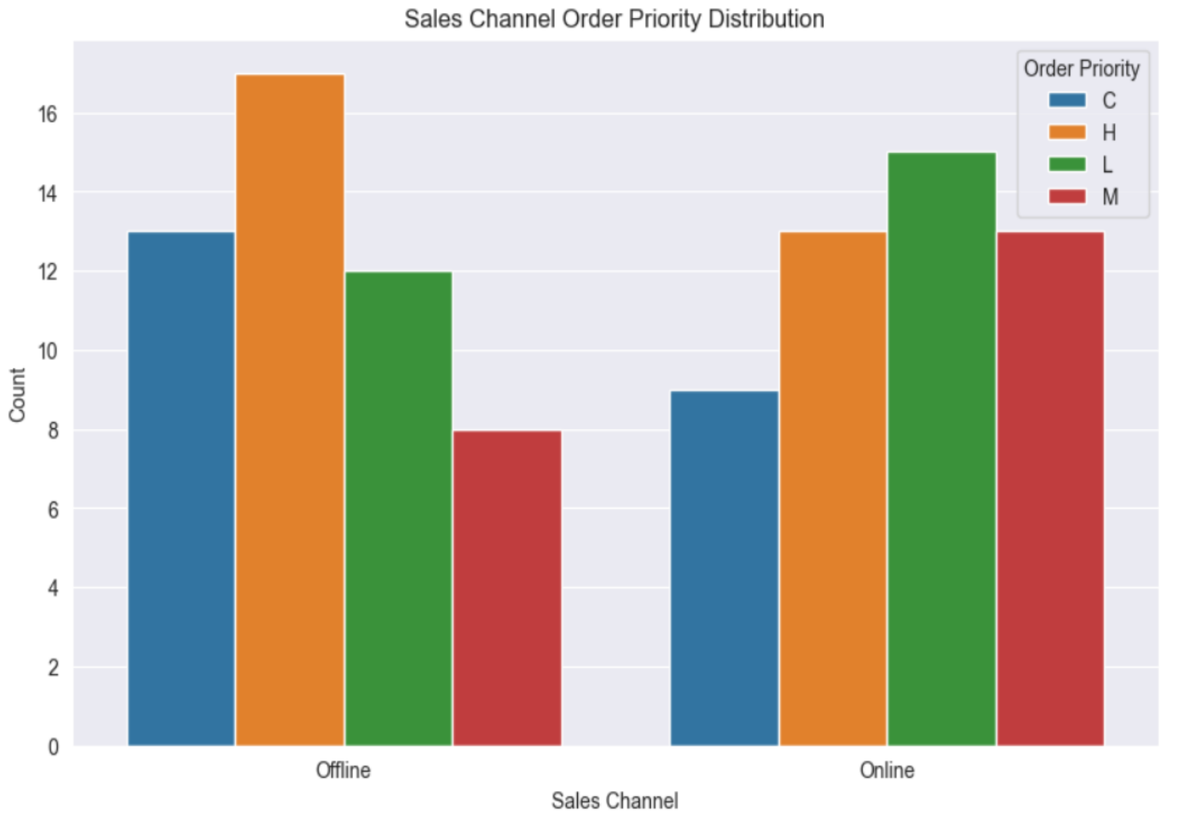


```
Sales_Channel_Order_Priority_Distribution= data.groupby(data['Sales Channel']) ['Order Priority'].value_counts()
Sales_Channel_Order_Priority_Distribution
```

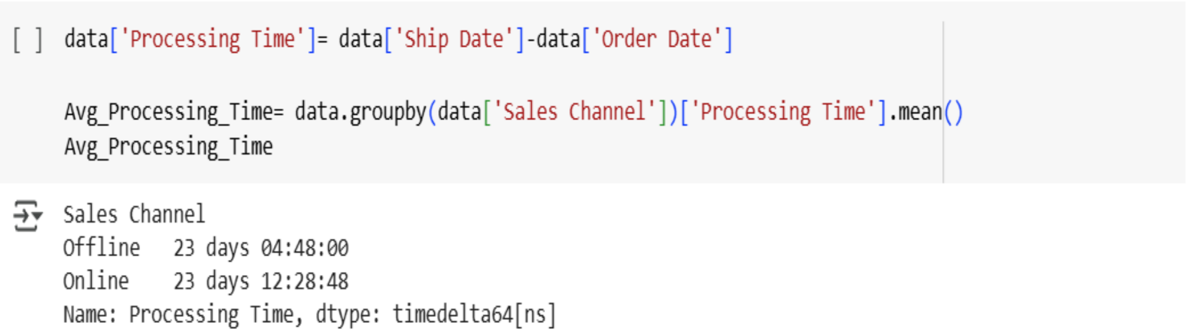
```
Sales Channel  Order Priority
Offline        H                 17
               C                 13
               L                 12
               M                  8
Online         L                 15
               H                 13
               M                 13
               C                  9
Name: Order Priority, dtype: int64
```



```
Sales_Channel_Order_Priority_Distribution = data.groupby(['Sales Channel', 'Order Priority'])['Order Priority'].count()

# Reset the index to convert the grouped data into a DataFrame
Sales_Channel_Order_Priority_Distribution = Sales_Channel_Order_Priority_Distribution.reset_index(name='Count')

# Set the style
sns.set_style('darkgrid')
```

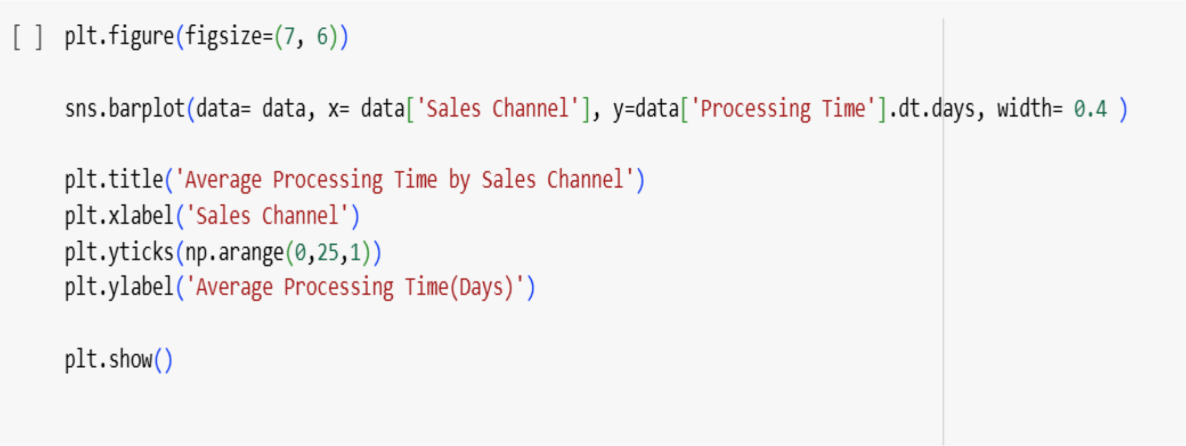


5- What is the average order processing time (duration between order and ship dates) for each sales channel?

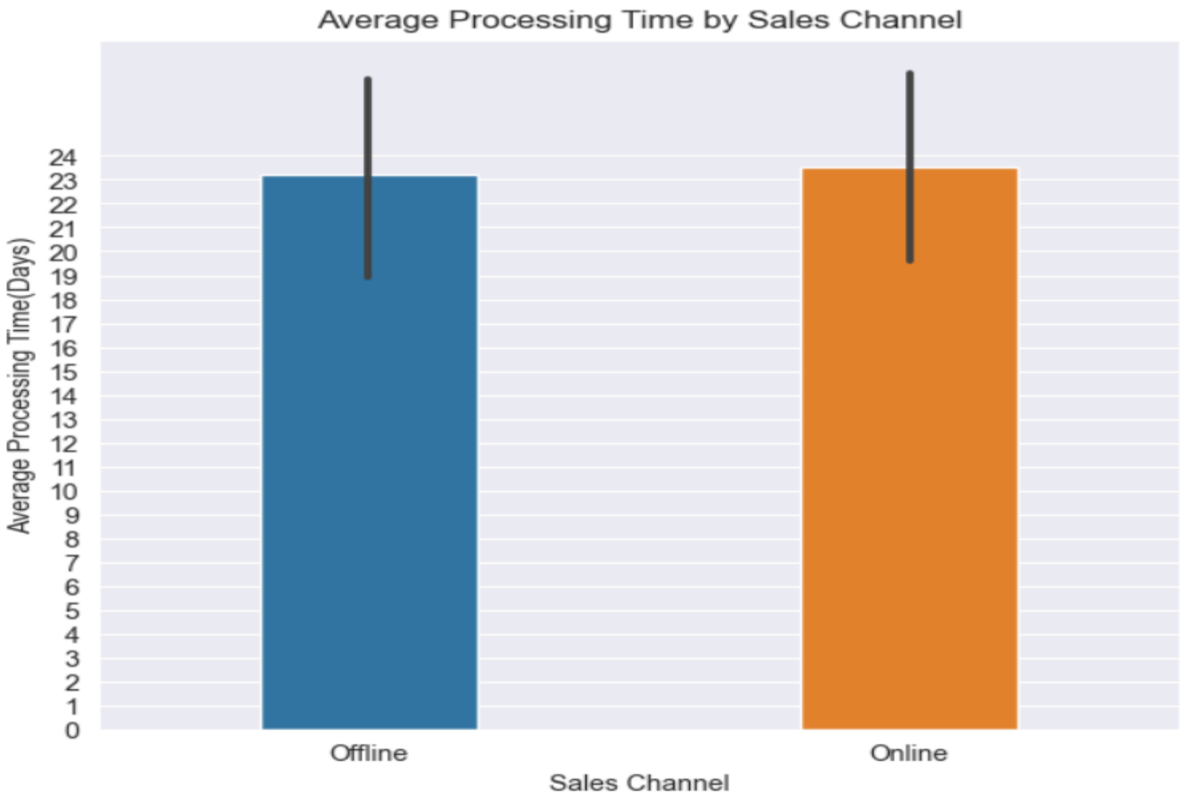


```
[ ] data['Processing Time']= data['Ship Date']-data['Order Date']

    Avg_Processing_Time= data.groupby(data['Sales Channel'])['Processing Time'].mean()
    Avg_Processing_Time
```

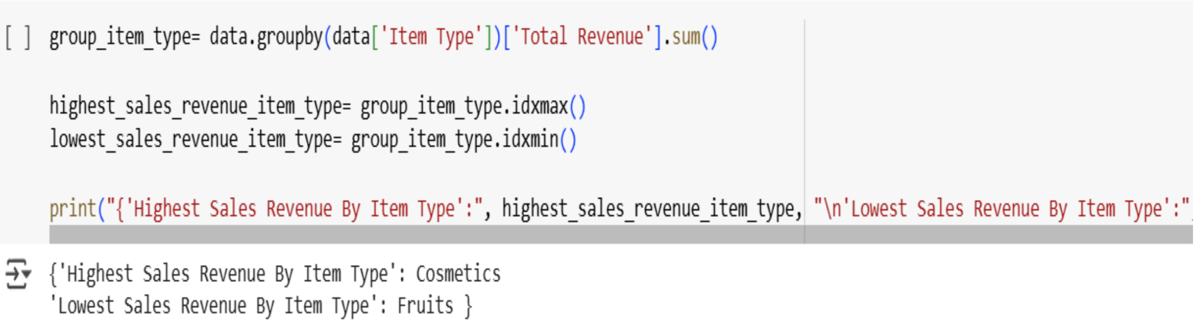
```
Sales Channel
Offline    23 days 04:48:00
Online     23 days 12:28:48
Name: Processing Time, dtype: timedelta64[ns]
```



```
[ ] plt.figure(figsize=(7, 6))

    sns.barplot(data= data, x= data['Sales Channel'], y=data['Processing Time'].dt.days, width= 0.4 )

    plt.title('Average Processing Time by Sales Channel')
    plt.xlabel('Sales Channel')
    plt.yticks(np.arange(0,25,1))
    plt.ylabel('Average Processing Time(Days)')

    plt.show()
```

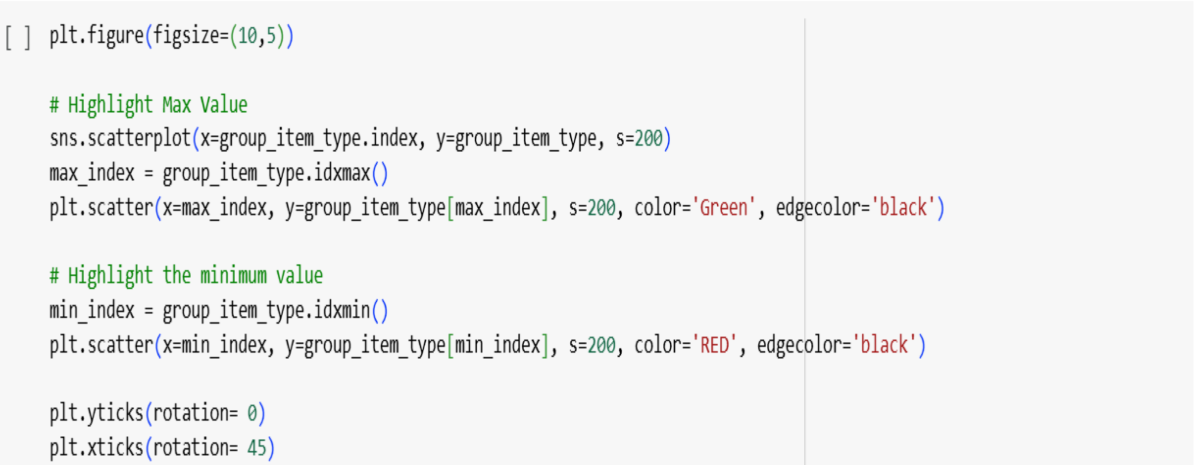


6- Which item types have the highest and lowest total sales?

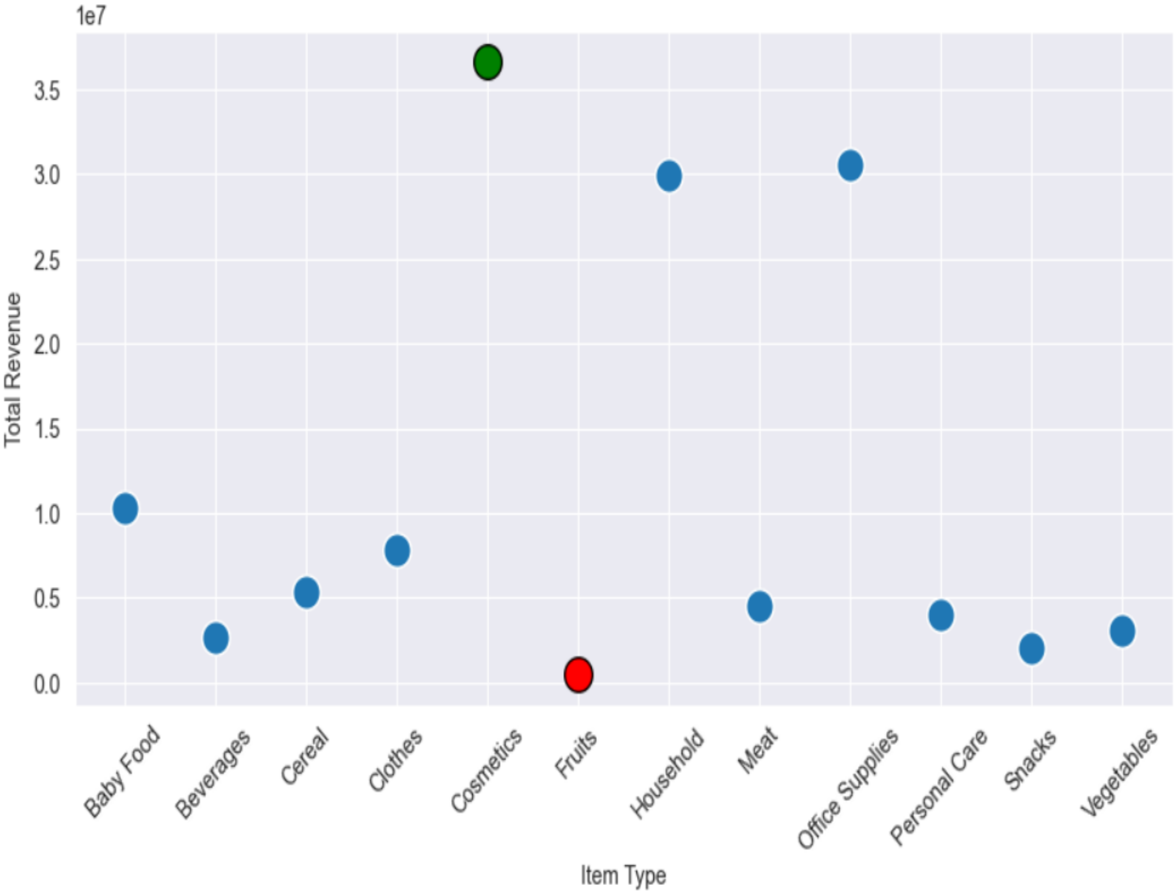


```
[ ] group_item_type= data.groupby(data['Item Type'])['Total Revenue'].sum()

    highest_sales_revenue_item_type= group_item_type.idxmax()
    lowest_sales_revenue_item_type= group_item_type.idxmin()

    print("{'Highest Sales Revenue By Item Type':", highest_sales_revenue_item_type, "\n'Lowest Sales Revenue By Item Type':"
```

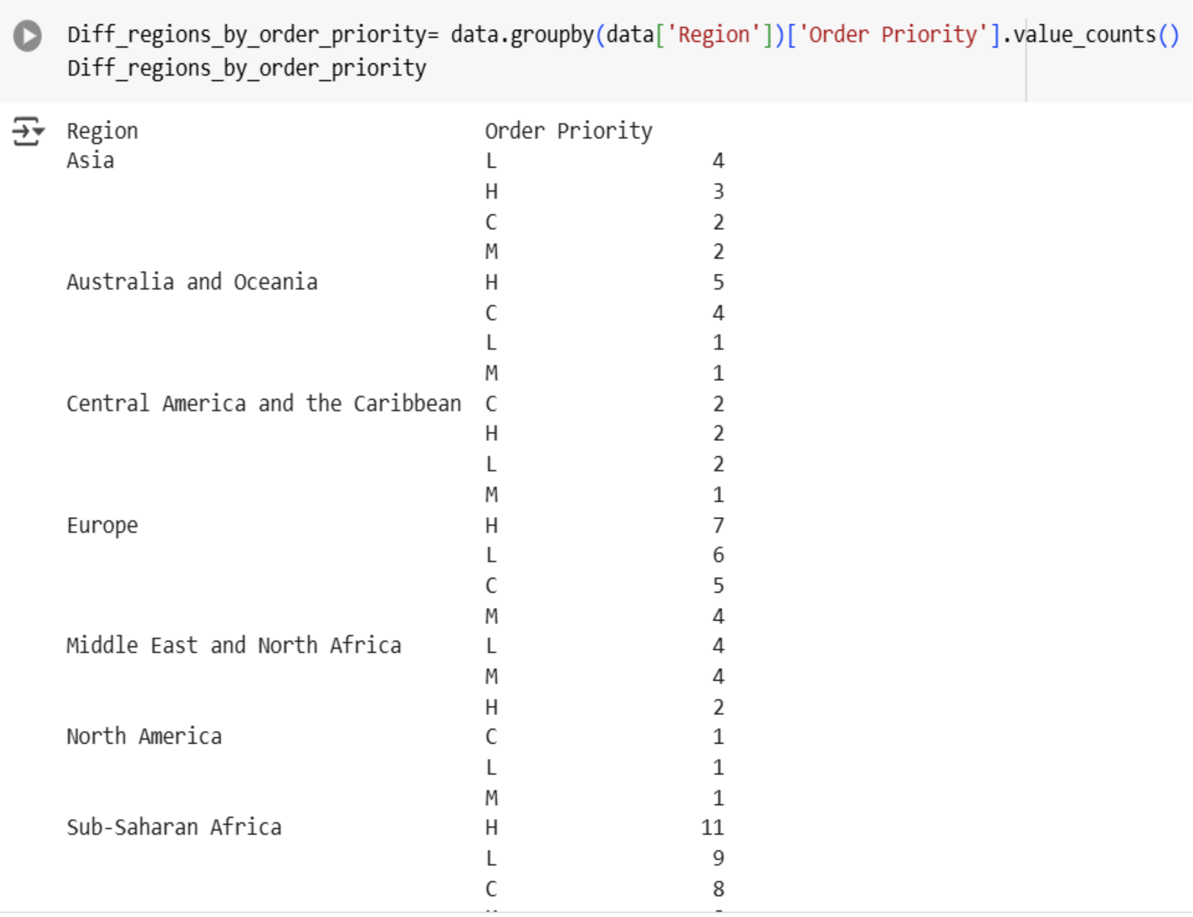
```
{'Highest Sales Revenue By Item Type': Cosmetics
'Lowest Sales Revenue By Item Type': Fruits }
```



```
[ ] plt.figure(figsize=(10,5))

    # Highlight Max Value
    sns.scatterplot(x=group_item_type.index, y=group_item_type, s=200)
    max_index = group_item_type.idxmax()
    plt.scatter(x=max_index, y=group_item_type[max_index], s=200, color='Green', edgecolor='black')

    # Highlight the minimum value
    min_index = group_item_type.idxmin()
    plt.scatter(x=min_index, y=group_item_type[min_index], s=200, color='RED', edgecolor='black')

    plt.yticks(rotation= 0)
    plt.xticks(rotation= 45)
```

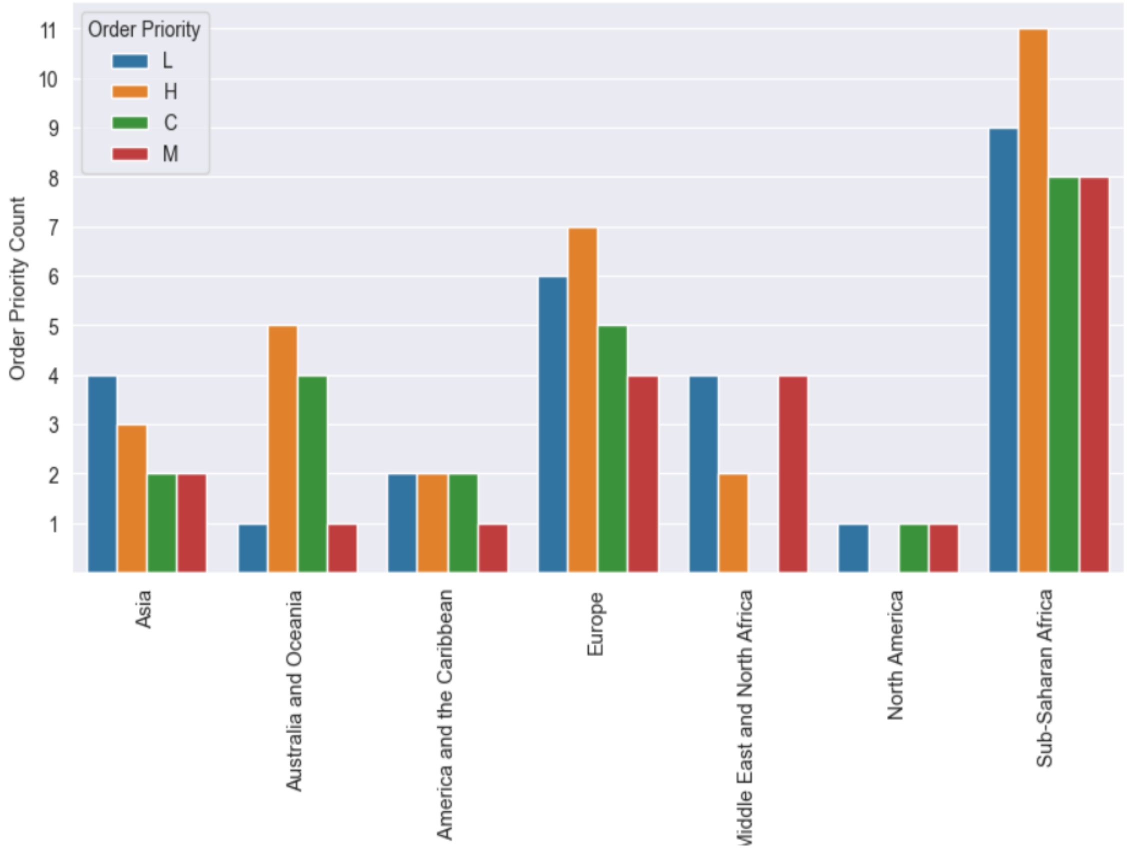


7- How does the order priority vary across different regions?

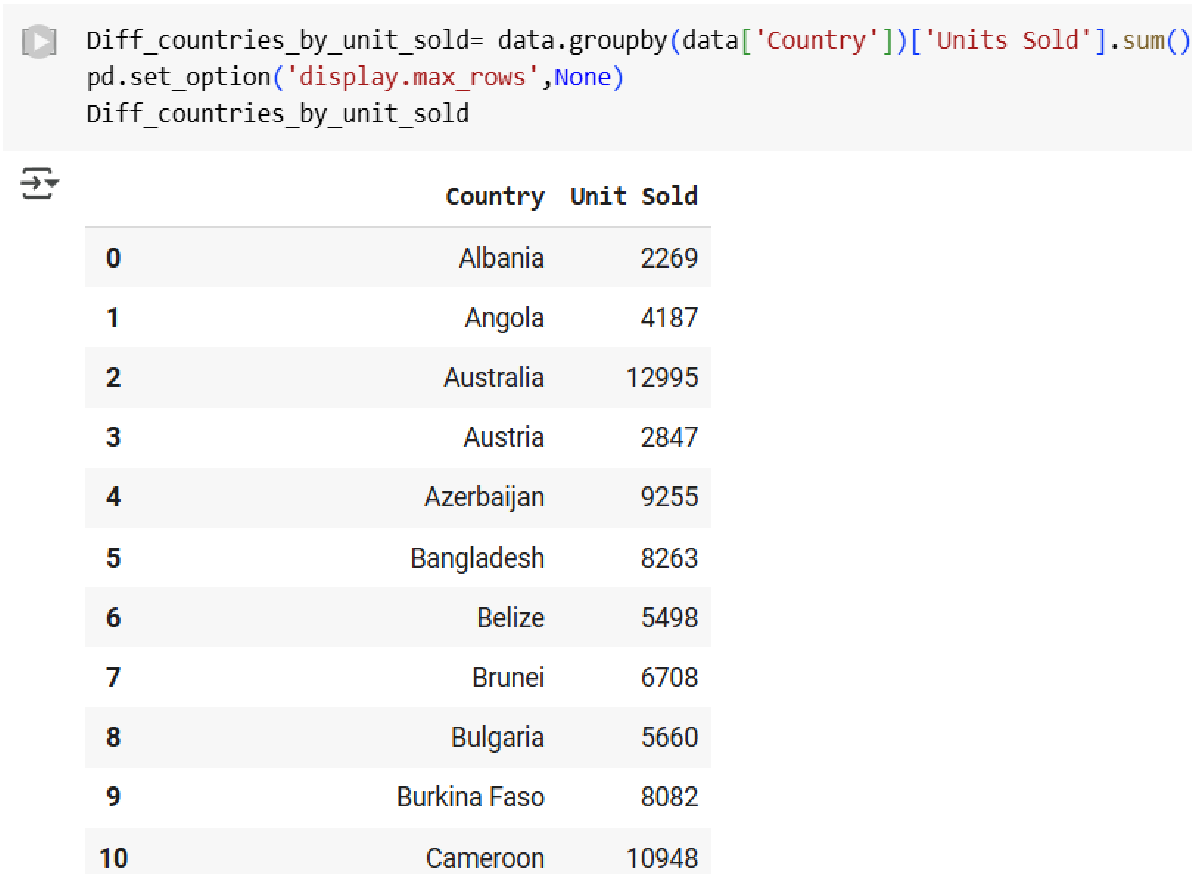


```
Diff_regions_by_order_priority= data.groupby(data['Region'])['Order Priority'].value_counts()
Diff_regions_by_order_priority
```

```
Region                            Order Priority
Asia                              L                  4
                                  H                  3
                                  C                  2
                                  M                  2
Australia and Oceania             H                  5
                                  C                  4
                                  L                  1
                                  M                  1
Central America and the Caribbean C                  2
                                  H                  2
                                  L                  2
                                  M                  1
Europe                            H                  7
                                  L                  6
                                  C                  5
                                  M                  4
Middle East and North Africa      L                  4
                                  M                  4
                                  H                  2
North America                     C                  1
                                  L                  1
                                  M                  1
Sub-Saharan Africa                H                 11
                                  L                  9
                                  C                  8
```

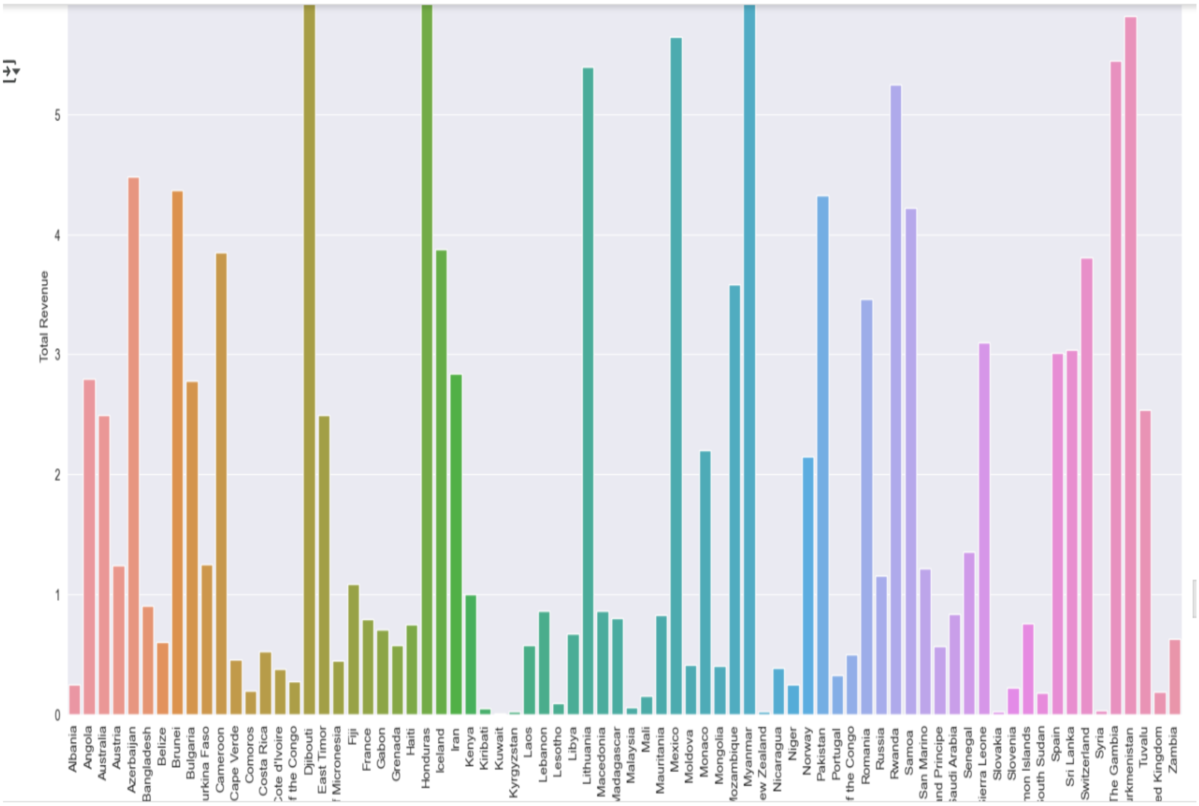


10- How does the number of units sold vary across different countries?



```
Diff_countries_by_unit_sold= data.groupby(data['Country'])['Units Sold'].sum()
pd.set_option('display.max_rows',None)
Diff_countries_by_unit_sold
```

	Country	Unit Sold
0	Albania	2269
1	Angola	4187
2	Australia	12995
3	Austria	2847
4	Azerbaijan	9255
5	Bangladesh	8263
6	Belize	5498
7	Brunei	6708
8	Bulgaria	5660
9	Burkina Faso	8082
10	Cameroon	10948



Other Queries:

How does the total sales revenue vary across different countries?

What is the distribution of unit prices for each item type?

Which sales channel has the highest average unit price?

Are there any outliers in the total cost distribution?

How does the total profit vary across different item types?

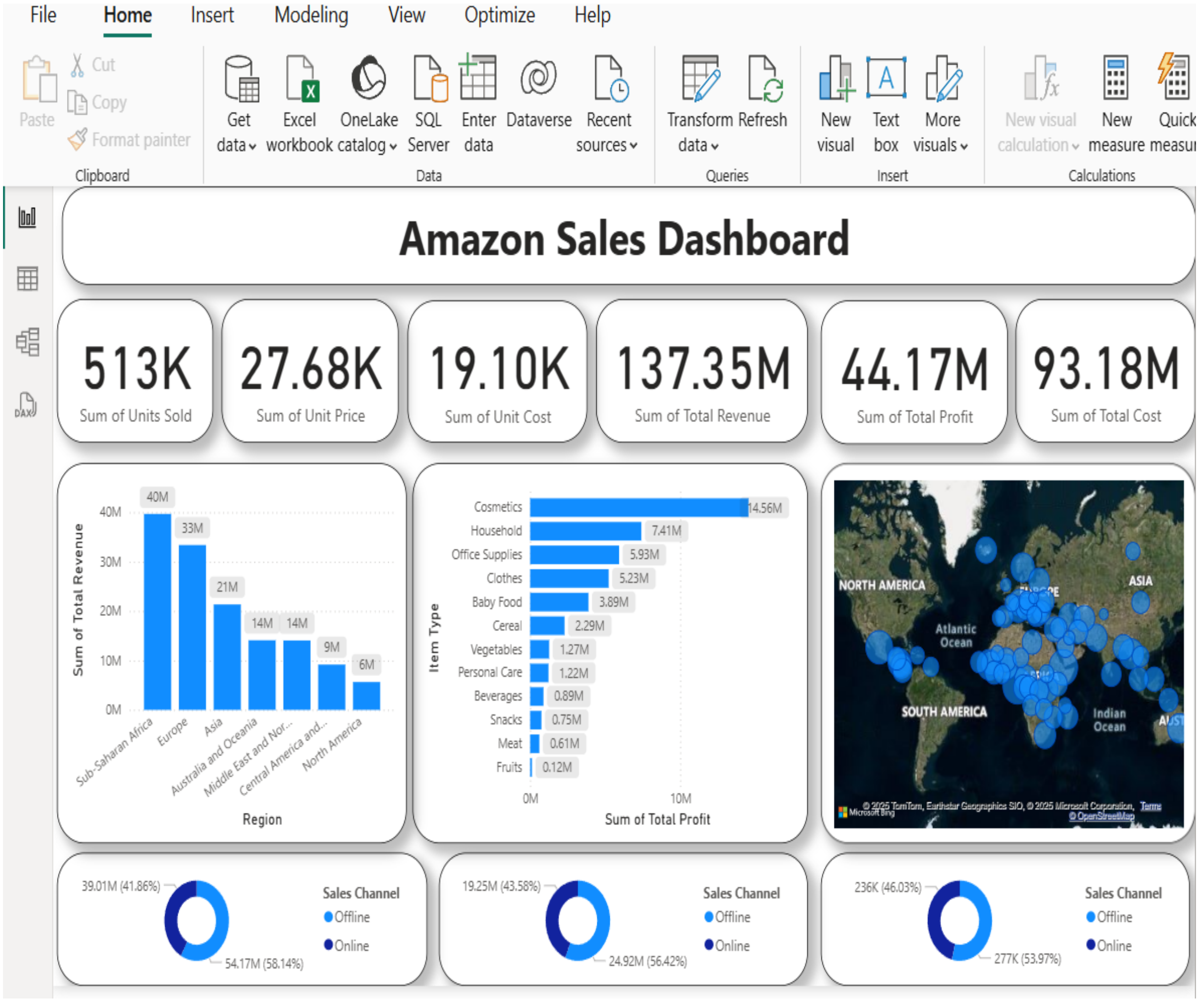
What is the average order processing time for each country?

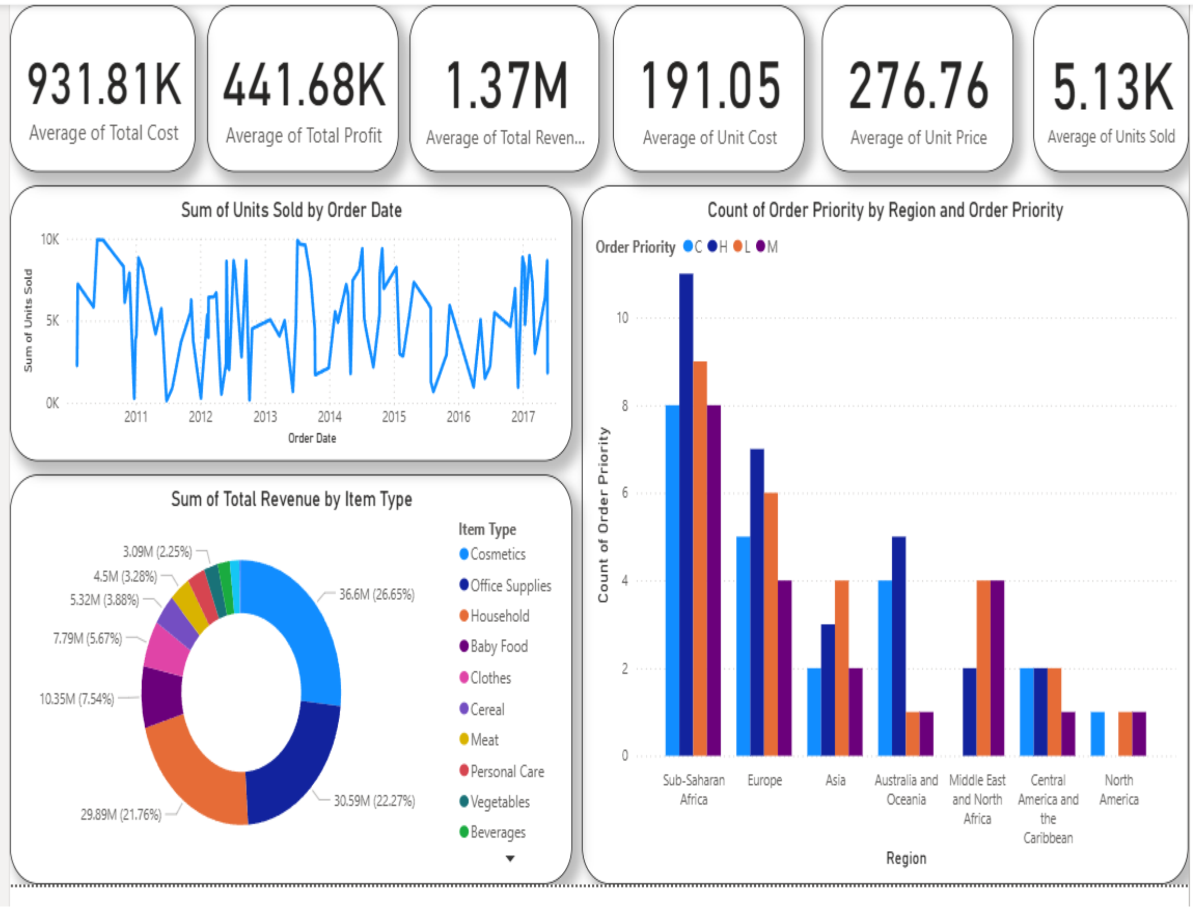
Which region has the highest average total revenue per order?

Is there a relationship between the number of units sold and the total profit?

How does the order priority vary based on the item type?

Are there any trends or patterns in the order dates?





931.81K Average of Total Cost

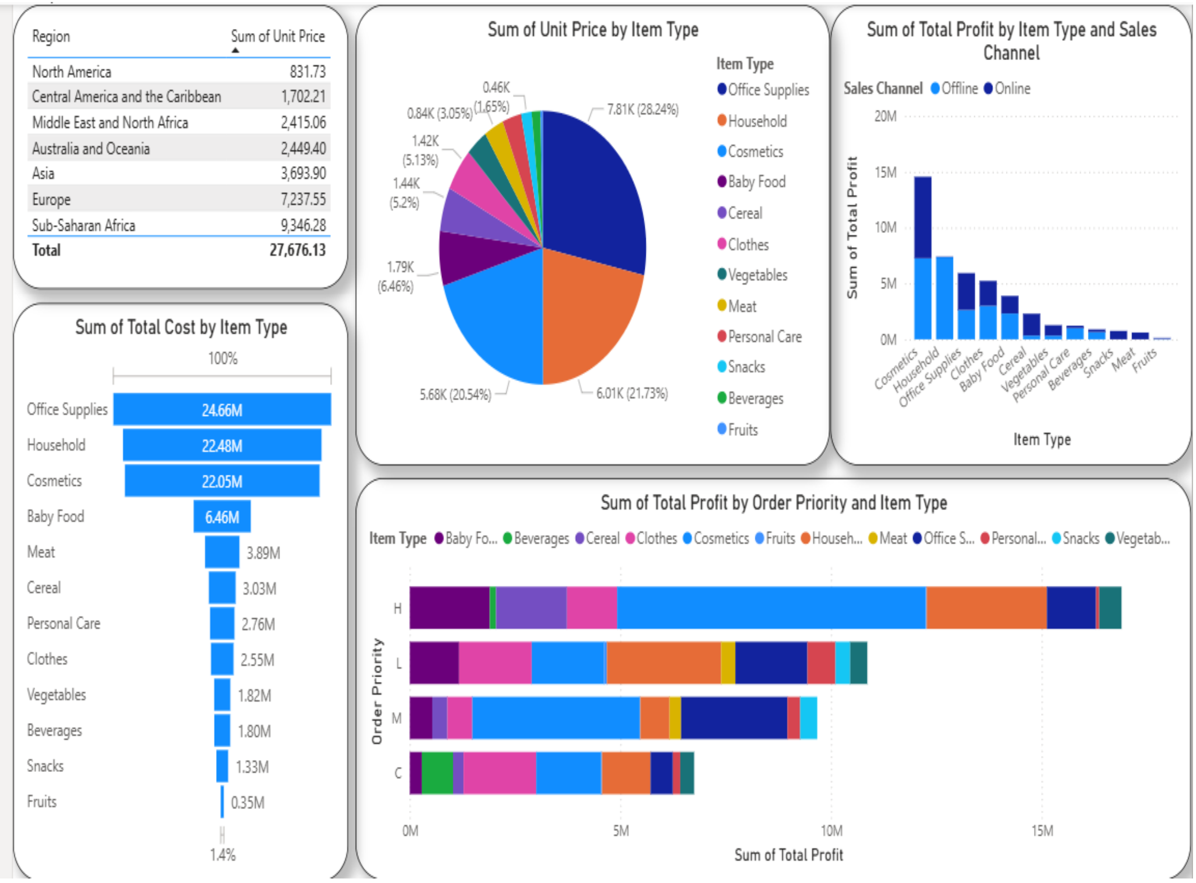
441.68K Average of Total Profit

1.37M Average of Total Reven...

191.05 Average of Unit Cost

276.76 Average of Unit Price

5.13K Average of Units Sold



Region	Sum of Unit Price
North America	831.73
Central America and the Caribbean	1,702.21
Middle East and North Africa	2,415.06
Australia and Oceania	2,449.40
Asia	3,693.90
Europe	7,237.55
Sub-Saharan Africa	9,346.28
Total	**27,676.13**

APPENDIX-C

ENCLOSURES

SUSTAINABLE DEVELOPMENT GOALS

SUSTAINABLE DEVELOPMENT GOALS
1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS
SUSTAINABLE DEVELOPMENT GOALS

This project aligns with the following SDGs:

Goal 9: Industry, Innovation, and Infrastructure

- Promotes data-driven business decision-making in the digital economy.

Goal 12: Responsible Consumption and Production

- Helps optimize inventory management and reduce overproduction.

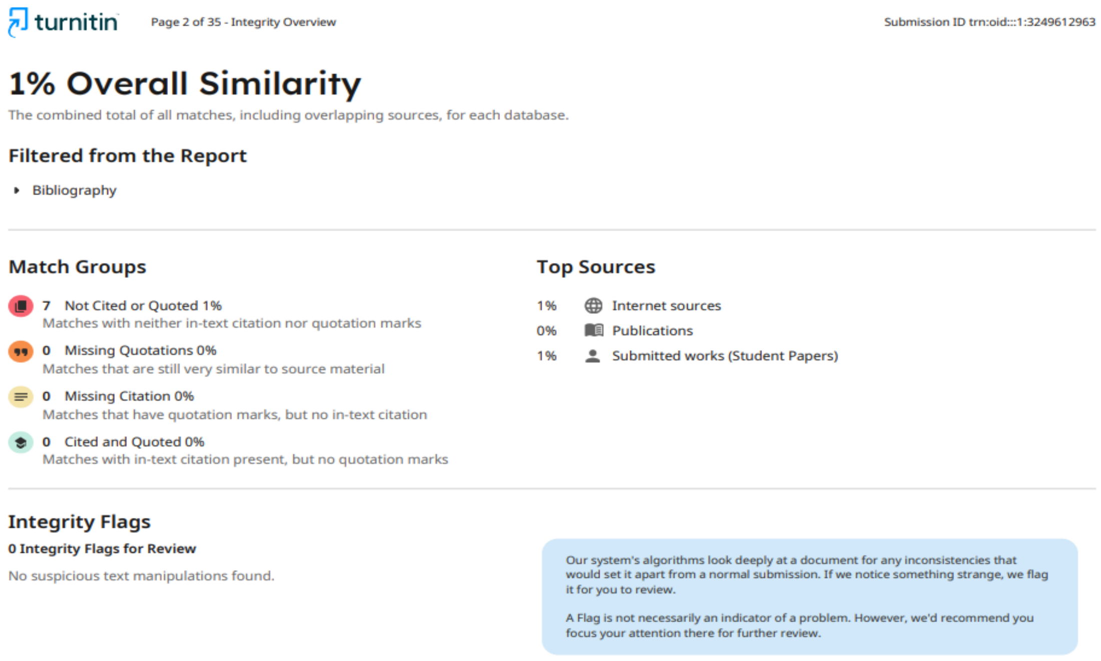
Goal 8: Decent Work and Economic Growth

- Provides insights to improve operational efficiency and revenue.

Goal 13: Climate Action

- Encourages better forecasting to reduce carbon footprints from over-shipping and returns.

PLAGIARISM REPORT



turnitin Page 2 of 35 - Integrity Overview Submission ID trn:oid:::1:3249612963

1% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography

Match Groups

- 7 Not Cited or Quoted 1%
  Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%
  Matches that are still very similar to source material
- 0 Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

1%	Internet sources
0%	Publications
1%	Submitted works (Student Papers)

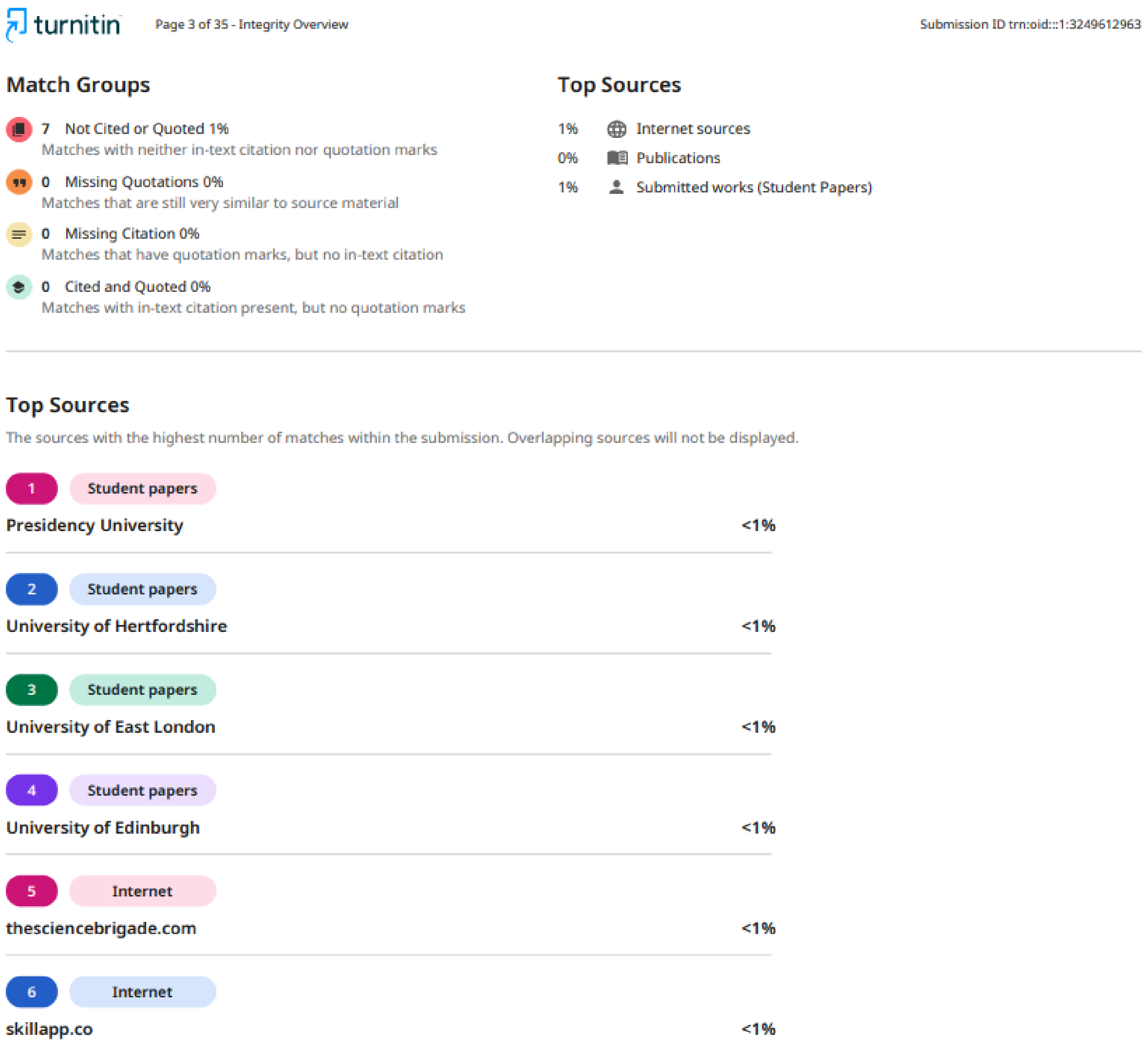
Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.



turnitin Page 3 of 35 - Integrity Overview Submission ID trn:oid:::1:3249612963

Match Groups

- 7 Not Cited or Quoted 1%
  Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%
  Matches that are still very similar to source material
- 0 Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

1%	Internet sources
0%	Publications
1%	Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Student papers	Presidency University	<1%
2	Student papers	University of Hertfordshire	<1%
3	Student papers	University of East London	<1%
4	Student papers	University of Edinburgh	<1%
5	Internet	thesciencebrigade.com	<1%
6	Internet	skillapp.co	<1%