Amazon Electronic Sales Analysis: Graduation Project Documentation

Submitted to: Digital Egypt Pioneers Initiative (DEPI)

Project Title: Amazon Electronic Sales Analysis

Team Information:

Item	Detail
Team Name	The Hunters
Team Number	175
Group Name	Dynamixx - DKH0_DAT0_S0
Instructor/Supervisor	Zyad Ayman

Team Members:

Abdelmonem Ayman Abdelmonem Mostafa Mohamed Ali Khaled sheref ahmed Mohamed Anis Ali 23 Omar islam Abdullah

---

Amazon Electronic Sales Analysis: Graduation Project Documentation

Executive Summary

This project presents a comprehensive analysis of Amazon’ s electronic sales data, spanning from data processing and cleaning to the extraction of strategic, actionable insights, all supported by interactive visualizations. The analysis revealed that the Adult age segment is the primary revenue driver, necessitating a focused marketing strategy. Furthermore, specific brands like Anker and Apple dominate profitability, underscoring the need for a robust inventory management strategy to ensure their availability during peak demand seasons. Operationally, Cash remains the most common payment method; thus, incentives are recommended to encourage a shift towards electronic payments to reduce operational costs. Additionally, investment in logistics efficiency is crucial in high-revenue cities like Alexandria and Cairo to reduce the average delivery time and enhance customer satisfaction. This project provides a clear, data-driven framework for decision-making, aiming to increase sales, improve profit margins, and set the foundation for future predictive analysis.

---

1. Project Planning and Methodology

The project was structured into a four-week timeline, following a systematic methodology that progressed from data preparation to final visualization and reporting.

1.1. Project Timeline and Deliverables

Week	Phase	Key Tasks	Tools Used	Deliverables
1	Data Preparation	Build Data Model, Data Cleaning and Preprocessing.	Excel, SQL	Cleaned dataset ready for analysis. Data preprocessing notebook.
2	Analysis Questions	Determine all possible analysis questions that can be deduced from the given dataset and would be of interest to the organization’ s decision makers (e.g., impact of product category and regions on sales performance).	SQL, Python (pandas, Matplotlib)	Set of analysis questions that can be answered via the dataset.
3	Forecasting Questions	Determine a set of forecasting questions and answer them using the trends found in the given dataset.	Python (scikit-learn, pandas, Matplotlib)	Visualization plots answering forecasting questions.
4	Visualization & Presentation	Build a Visualization Dashboard that visualizes the answers to all answered questions. Prepare a report and presentation summarizing the project work.	Power BI	Visualization dashboard. Final report and presentation.

1.2. Key Metrics and Performance Indicators (KPIs)

The analysis was driven by a set of Key Performance Indicators (KPIs) and core metrics, directly derived from the SQL queries executed against the dataset. These metrics provide a holistic view of the business performance.

1.2.1. Defined Key Performance Indicators

Category	Key Metric / KPI	Description	SQL Query Focus
Financial Performance	Total Revenue	The aggregate monetary value of all sales transactions.	`SUM(Total_Revenue_Capped)`
Financial Performance	Total Profit	The total net gain realized from all sales.	`SUM(Profit_Capped)`
Financial Performance	Average Final Price	The average selling price across all products.	`AVG(Final_Price)`
Financial Performance	Total Revenue per Year	Revenue trend over time.	`GROUP BY YEAR(YearMonth)`
Financial Performance	Total Revenue per City	Revenue contribution segmented by geographical location.	`GROUP BY City`
Sales & Operations	Total Orders	The total count of transactions recorded.	`COUNT(Order_ID)`
Sales & Operations	Average Unit Price	The average cost per unit across all products.	`AVG(Unit_Price)`
Sales & Operations	Orders per Year and Quarter	The breakdown of orders by time.	`GROUP BY YEAR(Order_Date), DATEPART(qq, Order_Date)`
Sales & Operations	Orders per Delivery Status	Distribution of orders by delivery status.	`GROUP BY Delivery_Status`

Category	Key Metric / KPI	Description	SQL Query Focus
Sales & Operations	Orders per Payment Type	Distribution of orders by payment method.	`GROUP BY Payment_Type`
Sales & Operations	Orders per Sales Channel	Distribution of orders by sales channel.	`GROUP BY Sales_Channel`
Sales & Operations	Orders per Sales Level	Distribution of orders by sales level (High, Medium, Low).	`GROUP BY Sales_Level`
Customer & Demographic	Total Customers	The total number of unique customers in the dataset.	`COUNT(Customer_ID)`
Customer & Demographic	Average Customer Age	The mean age of the customer base.	`AVG(Age)`
Customer & Demographic	Customers per City	Distribution of customers by geographical location.	`GROUP BY City`
Customer & Demographic	Customers per Age Band	The demographic breakdown of the customer base.	`GROUP BY Age_Band`
Customer & Demographic	Customers per Gender	The demographic breakdown of the customer base.	`GROUP BY Gender`
Customer & Demographic	Total Revenue per Age Band	The financial contribution of different customer age segments.	`GROUP BY Age_Band`
Product Performance	Total Products	The total number of unique products.	`COUNT(Product_ID)`

Category	Key Metric / KPI	Description	SQL Query Focus
Product Performance	Average Warranty Months	The mean duration of product warranties.	`AVG(Warranty_Months)`
Product Performance	Total Revenue per Brand	Revenue contribution segmented by brand.	`GROUP BY Brand`
Product Performance	Units Sold per Sub-Category	The volume of sales for specific product types.	`GROUP BY Sub_Category`
Product Performance	Total Profit per Brand	The profitability analysis for each major brand.	`GROUP BY Brand`
Product Performance	Final Price vs Profit	Detailed view of final price and profit per brand and sub-category.	`ORDER BY Brand`

---

2. Data Sources and Collection

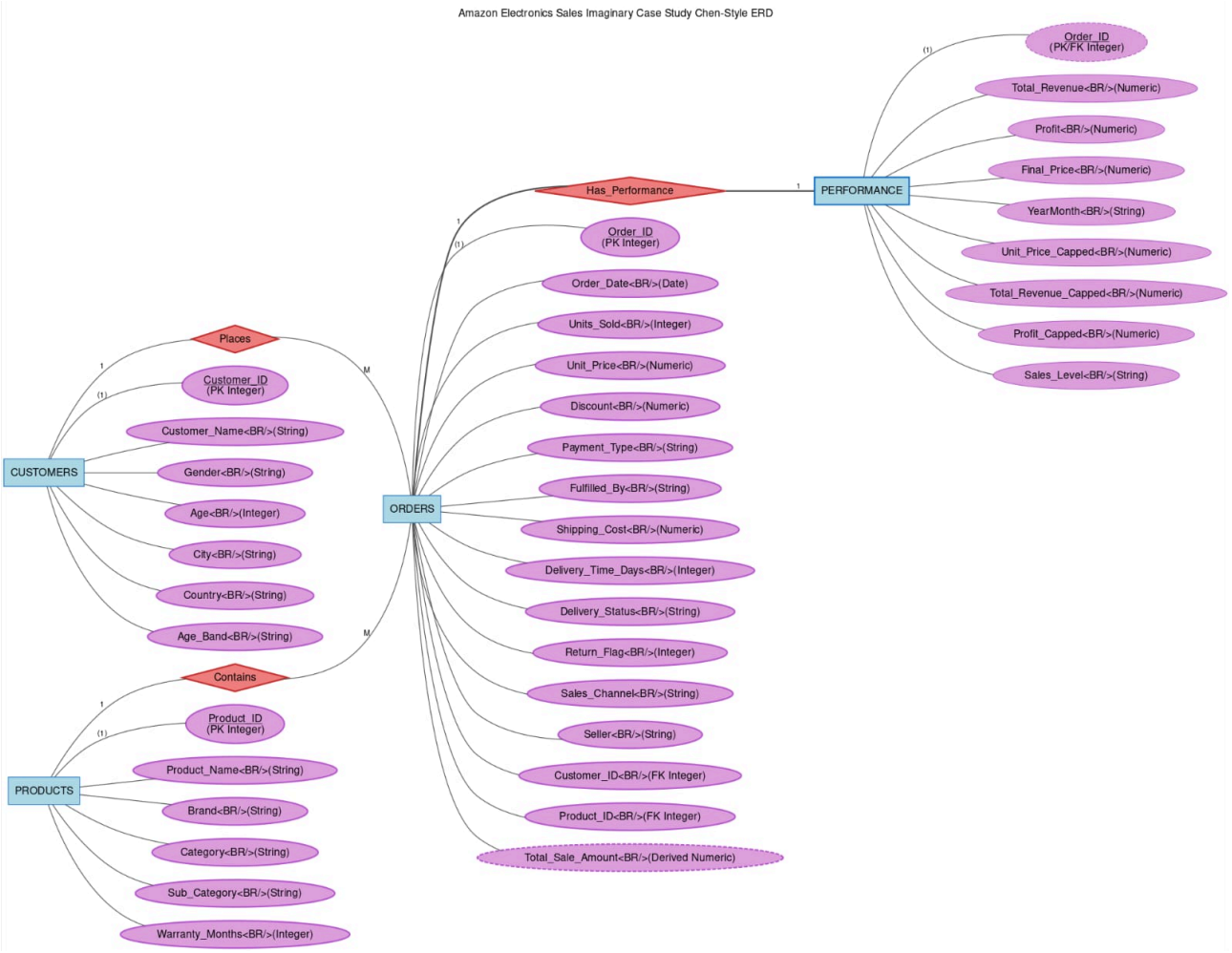
2.1. Data Source

The data utilized for this analysis was sourced exclusively from a publicly available dataset on Kaggle. This dataset contains historical transaction records for Amazon electronic sales, providing a rich, pre-processed foundation for the analysis. The reliance on a single, structured public dataset ensured data consistency and facilitated the initial data modeling phase.

2.2. Data Model (Entity-Relationship Diagram)

The underlying structure of the data is represented by the Entity-Relationship Diagram (ERD). This Chen-style diagram illustrates the relationships between the core entities:

CUSTOMERS, PRODUCTS, ORDERS, and PERFORMANCE, which is crucial for understanding the data flow and the foundation of the SQL queries.



2.3. Data Attributes

The final, unified dataset comprised the following key attributes:

Category	Attribute Name	Description
Product Info	ProductID , Product_Category , Brand , Unit_Price	Unique identifier, product type, manufacturer, and cost per unit.
Sales Info	OrderID , Quantity , Total_Amount , Discount_Amount , Order_Date	Unique transaction ID, units sold, final transaction value, applied discount, and timestamp.
Customer Info	CustomerID , City , Gender , Age	Unique customer ID, geographical location, demographic information.
Additional Data	Payment_Method , Customer_Rating , Delivery_Time_Days	Method of payment, customer satisfaction score, and logistics duration.

---

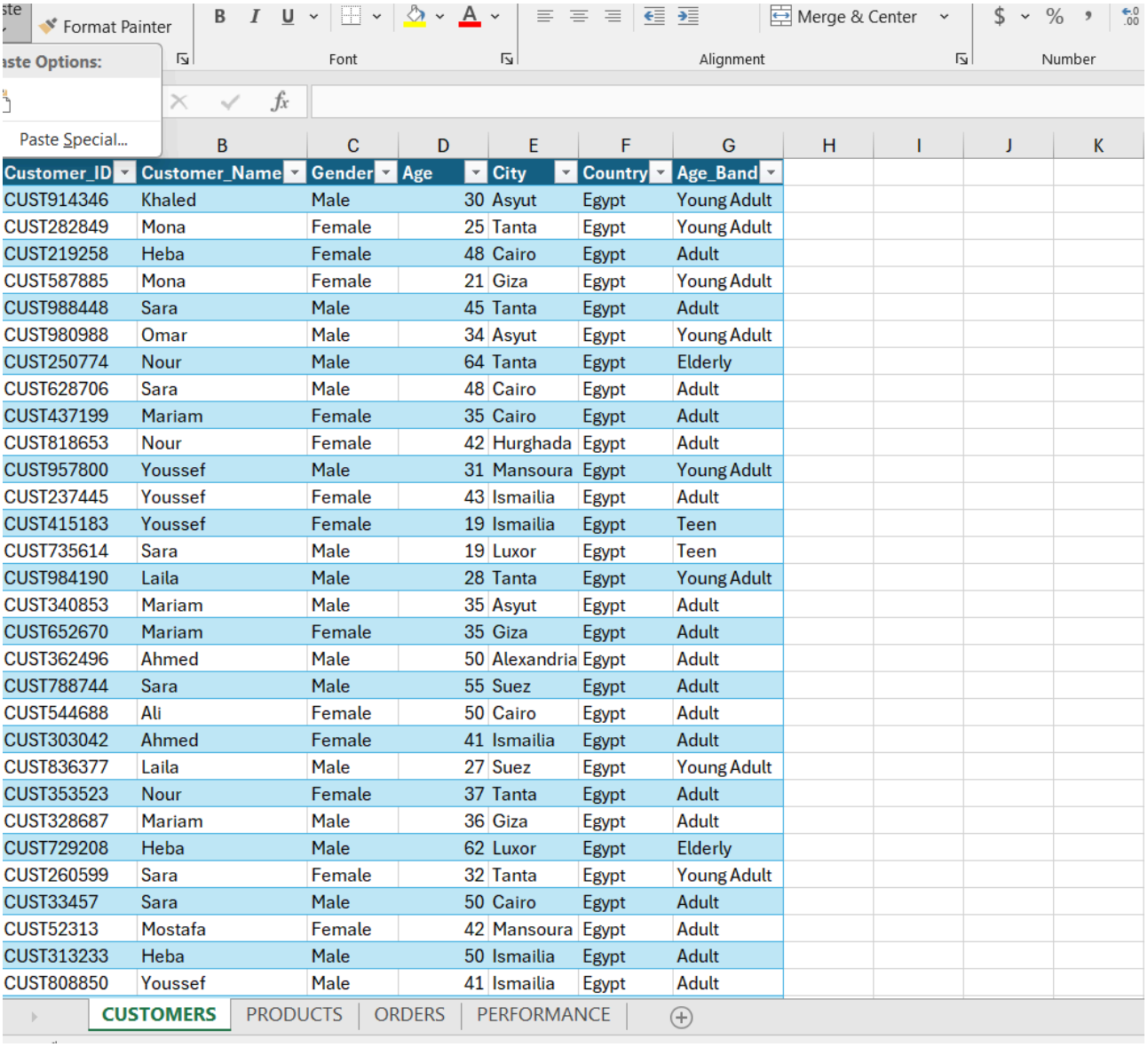
3. Data Cleaning and Preparation

The data cleaning and preparation phase was critical to ensure the quality and reliability of the subsequent analysis. Microsoft Excel was used in the initial stages of the Data Cleaning process before transitioning to more specialized tools like SQL and Python, primarily to facilitate quick visual inspection of the data and identify preliminary issues. A systematic approach was followed to address common data issues.

3.1. Initial Data Inspection (Excel Screenshots)

The following screenshots illustrate the structure and initial state of the data as viewed in Microsoft Excel during the preliminary cleaning and inspection phase:

Figure 3.1: Customers Data Snapshot



Customer_ID	Customer_Name	Gender	Age	City	Country	Age_Band
CUST914346	Khaled	Male	30	Asyut	Egypt	Young Adult
CUST282849	Mona	Female	25	Tanta	Egypt	Young Adult
CUST219258	Heba	Female	48	Cairo	Egypt	Adult
CUST587885	Mona	Female	21	Giza	Egypt	Young Adult
CUST988448	Sara	Male	45	Tanta	Egypt	Adult
CUST980988	Omar	Male	34	Asyut	Egypt	Young Adult
CUST250774	Nour	Male	64	Tanta	Egypt	Elderly
CUST628706	Sara	Male	48	Cairo	Egypt	Adult
CUST437199	Mariam	Female	35	Cairo	Egypt	Adult
CUST818653	Nour	Female	42	Hurghada	Egypt	Adult
CUST957800	Youssef	Male	31	Mansoura	Egypt	Young Adult
CUST237445	Youssef	Female	43	Ismailia	Egypt	Adult
CUST415183	Youssef	Female	19	Ismailia	Egypt	Teen
CUST735614	Sara	Male	19	Luxor	Egypt	Teen
CUST984190	Laila	Male	28	Tanta	Egypt	Young Adult
CUST340853	Mariam	Male	35	Asyut	Egypt	Adult
CUST652670	Mariam	Female	35	Giza	Egypt	Adult
CUST362496	Ahmed	Male	50	Alexandria	Egypt	Adult
CUST788744	Sara	Male	55	Suez	Egypt	Adult
CUST544688	Ali	Female	50	Cairo	Egypt	Adult
CUST303042	Ahmed	Female	41	Ismailia	Egypt	Adult
CUST836377	Laila	Male	27	Suez	Egypt	Young Adult
CUST353523	Nour	Female	37	Tanta	Egypt	Adult
CUST328687	Mariam	Male	36	Giza	Egypt	Adult
CUST729208	Heba	Male	62	Luxor	Egypt	Elderly
CUST260599	Sara	Female	32	Tanta	Egypt	Young Adult
CUST33457	Sara	Male	50	Cairo	Egypt	Adult
CUST52313	Mostafa	Female	42	Mansoura	Egypt	Adult
CUST313233	Heba	Male	50	Ismailia	Egypt	Adult
CUST808850	Youssef	Male	41	Ismailia	Egypt	Adult

Figure 3.2: Products Data Snapshot

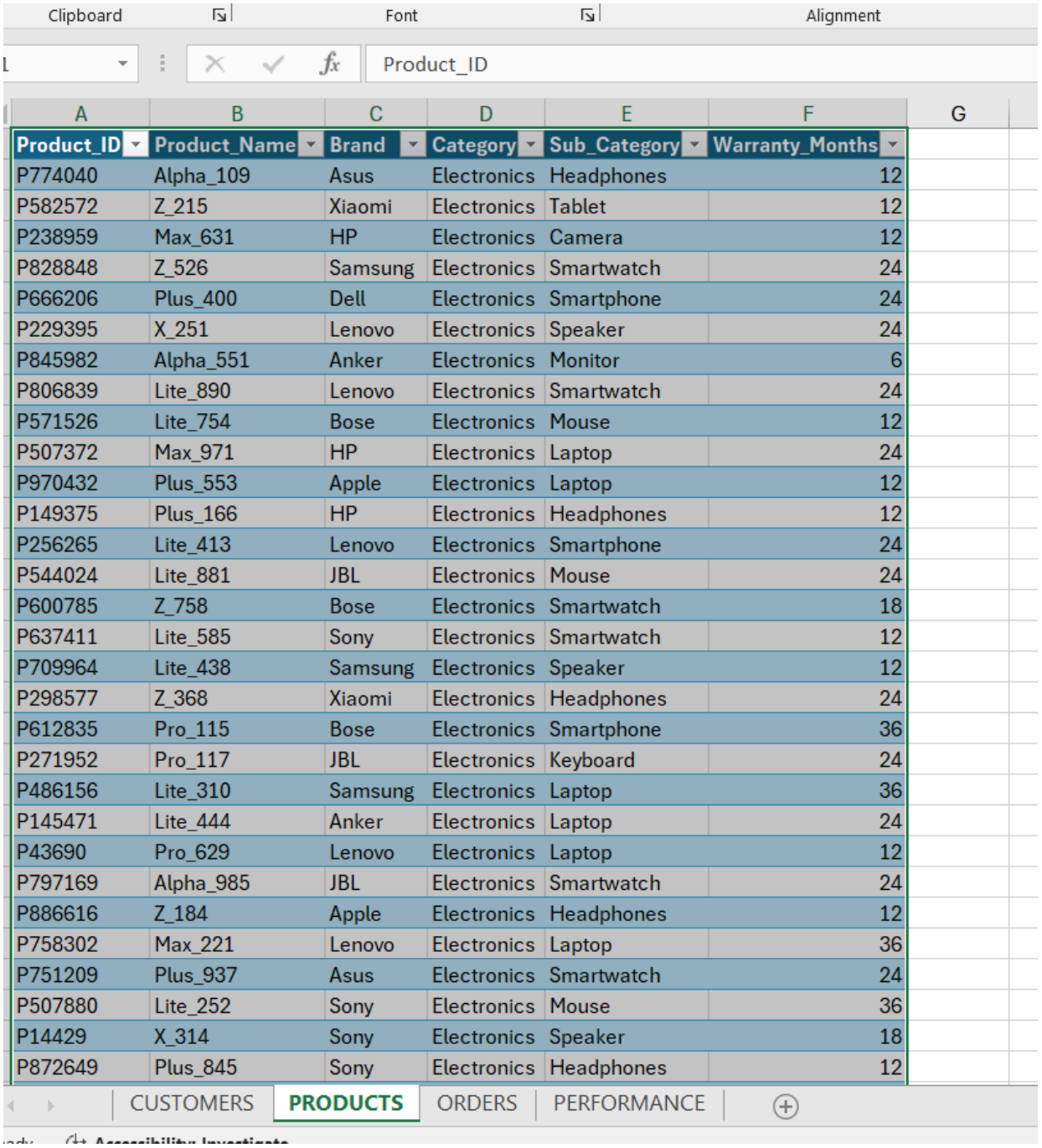
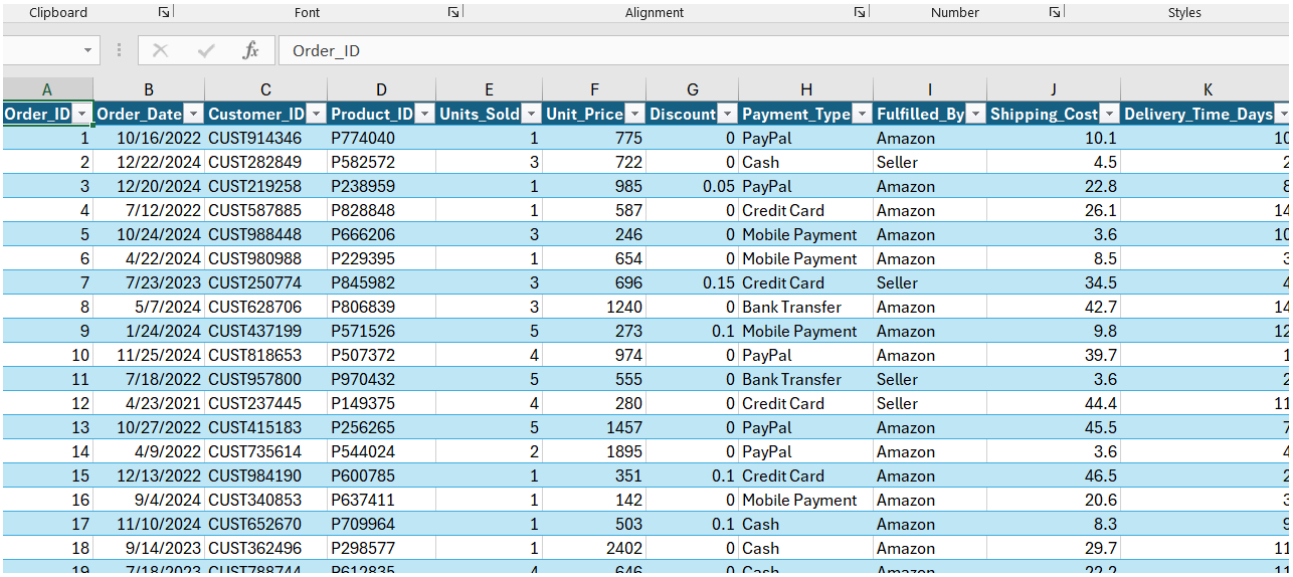
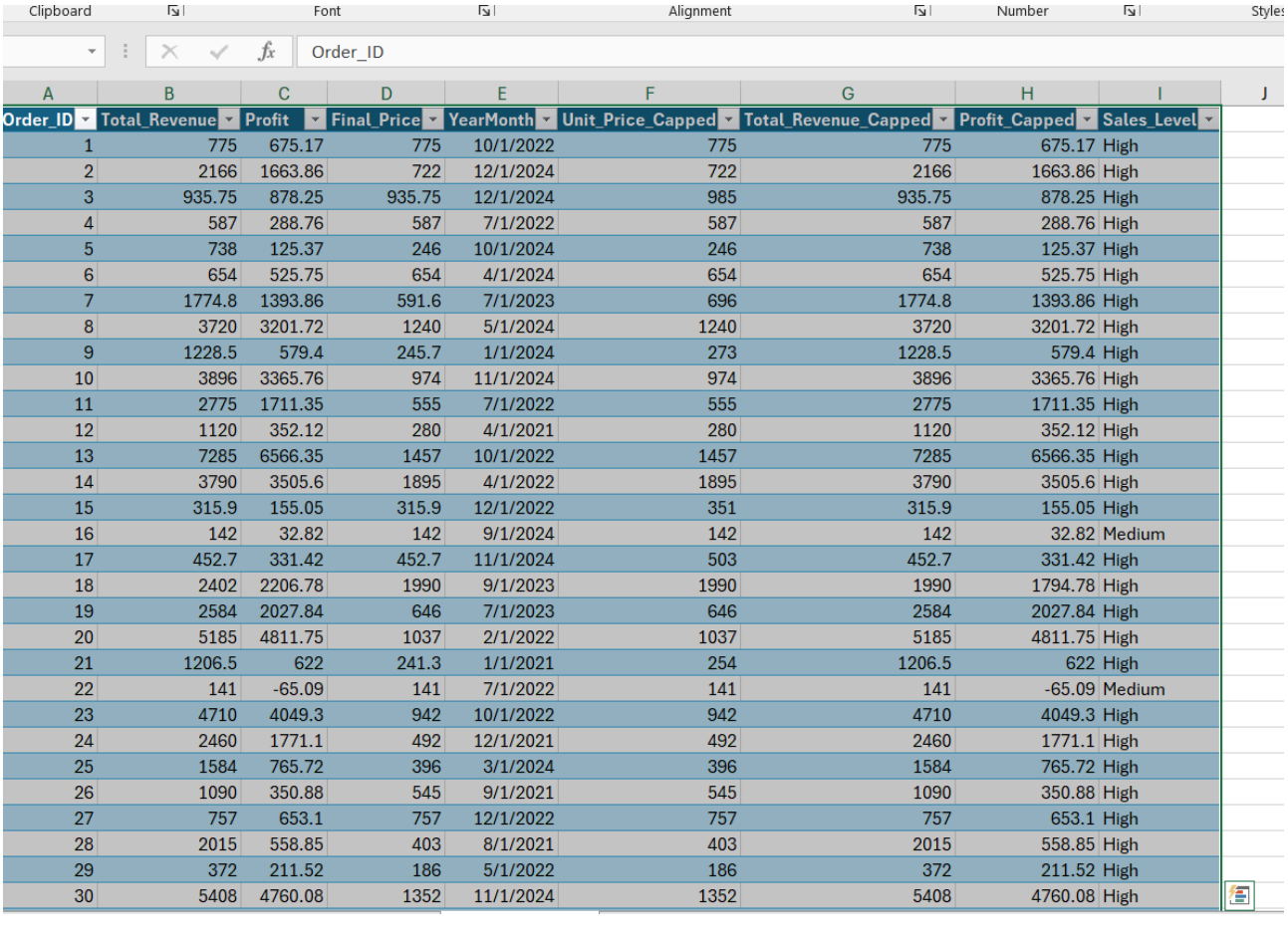


Figure 3.3: Orders Data Snapshot



Order_ID	Order_Date	Customer_ID	Product_ID	Units_Sold	Unit_Price	Discount	Payment_Type	Fulfilled_By	Shipping_Cost	Delivery_Time_Days
1	10/16/2022	CUST914346	P774040	1	775	0	PayPal	Amazon	10.1	10
2	12/22/2024	CUST282849	P582572	3	722	0	Cash	Seller	4.5	2
3	12/20/2024	CUST219258	P238959	1	985	0.05	PayPal	Amazon	22.8	8
4	7/12/2022	CUST587885	P828848	1	587	0	Credit Card	Amazon	26.1	14
5	10/24/2024	CUST988448	P666206	3	246	0	Mobile Payment	Amazon	3.6	10
6	4/22/2024	CUST980988	P229395	1	654	0	Mobile Payment	Amazon	8.5	3
7	7/23/2023	CUST250774	P845982	3	696	0.15	Credit Card	Seller	34.5	4
8	5/7/2024	CUST628706	P806839	3	1240	0	Bank Transfer	Amazon	42.7	14
9	1/24/2024	CUST437199	P571526	5	273	0.1	Mobile Payment	Amazon	9.8	12
10	11/25/2024	CUST818653	P507372	4	974	0	PayPal	Amazon	39.7	1
11	7/18/2022	CUST957800	P970432	5	555	0	Bank Transfer	Seller	3.6	2
12	4/23/2021	CUST237445	P149375	4	280	0	Credit Card	Seller	44.4	11
13	10/27/2022	CUST415183	P256265	5	1457	0	PayPal	Amazon	45.5	7
14	4/9/2022	CUST735614	P544024	2	1895	0	PayPal	Amazon	3.6	4
15	12/13/2022	CUST984190	P600785	1	351	0.1	Credit Card	Amazon	46.5	2
16	9/4/2024	CUST340853	P637411	1	142	0	Mobile Payment	Amazon	20.6	3
17	11/10/2024	CUST652670	P709964	1	503	0.1	Cash	Amazon	8.3	9
18	9/14/2023	CUST362496	P298577	1	2402	0	Cash	Amazon	29.7	11

Figure 3.4: Performance Data Snapshot



Order_ID	Total_Revenue	Profit	Final_Price	YearMonth	Unit_Price_Capped	Total_Revenue_Capped	Profit_Capped	Sales_Level
1	775	675.17	775	10/1/2022	775	775	675.17	High
2	2166	1663.86	722	12/1/2024	722	2166	1663.86	High
3	935.75	878.25	935.75	12/1/2024	985	935.75	878.25	High
4	587	288.76	587	7/1/2022	587	587	288.76	High
5	738	125.37	246	10/1/2024	246	738	125.37	High
6	654	525.75	654	4/1/2024	654	654	525.75	High
7	1774.8	1393.86	591.6	7/1/2023	696	1774.8	1393.86	High
8	3720	3201.72	1240	5/1/2024	1240	3720	3201.72	High
9	1228.5	579.4	245.7	1/1/2024	273	1228.5	579.4	High
10	3896	3365.76	974	11/1/2024	974	3896	3365.76	High
11	2775	1711.35	555	7/1/2022	555	2775	1711.35	High
12	1120	352.12	280	4/1/2021	280	1120	352.12	High
13	7285	6566.35	1457	10/1/2022	1457	7285	6566.35	High
14	3790	3505.6	1895	4/1/2022	1895	3790	3505.6	High
15	315.9	155.05	315.9	12/1/2022	351	315.9	155.05	High
16	142	32.82	142	9/1/2024	142	142	32.82	Medium
17	452.7	331.42	452.7	11/1/2024	503	452.7	331.42	High
18	2402	2206.78	1990	9/1/2023	1990	1990	1794.78	High
19	2584	2027.84	646	7/1/2023	646	2584	2027.84	High
20	5185	4811.75	1037	2/1/2022	1037	5185	4811.75	High
21	1206.5	622	241.3	1/1/2021	254	1206.5	622	High
22	141	-65.09	141	7/1/2022	141	141	-65.09	Medium
23	4710	4049.3	942	10/1/2022	942	4710	4049.3	High
24	2460	1771.1	492	12/1/2021	492	2460	1771.1	High
25	1584	765.72	396	3/1/2024	396	1584	765.72	High
26	1090	350.88	545	9/1/2021	545	1090	350.88	High
27	757	653.1	757	12/1/2022	757	757	653.1	High
28	2015	558.85	403	8/1/2021	403	2015	558.85	High
29	372	211.52	186	5/1/2022	186	372	211.52	High
30	5408	4760.08	1352	11/1/2024	1352	5408	4760.08	High

3.2. Data Quality Management

- **Handling Missing Values:** Missing numerical values (e.g., `Age`) were imputed using the mean, while critical missing values (e.g., `OrderID`) led to the removal of the corresponding rows to maintain data integrity.

- **Removing Duplicates:** The dataset was scanned for fully duplicated records (based on `OrderID` and `ProductID`) and removed to prevent inflation of sales figures.
- **Outlier Treatment:** Statistical techniques, such as the Interquartile Range (IQR), were applied to identify and investigate outliers in fields like `Unit_Price` and `Quantity`. Outliers were either corrected or removed based on whether they represented data entry errors or genuine, albeit extreme, transactions.

3.3. Data Transformation

- **Data Type Conversion:** The `Order_Date` field was converted from text to a proper `DateTime` format for time-series analysis.
- **Standardization:** Category and brand names were standardized to ensure consistency (e.g., unifying variations of the same name).
- **Feature Engineering:** New fields were derived to deepen the analysis, such as `Month`, `Year`, `DayOfWeek` for seasonal analysis, and `Price_Range` for product categorization.

---

4. Data Analysis

A range of analytical techniques was applied to explore the data and extract actionable insights, directly addressing the analytical questions derived from the SQL queries.

4.1. Analytical Questions

The analysis focused on answering key questions across three domains:

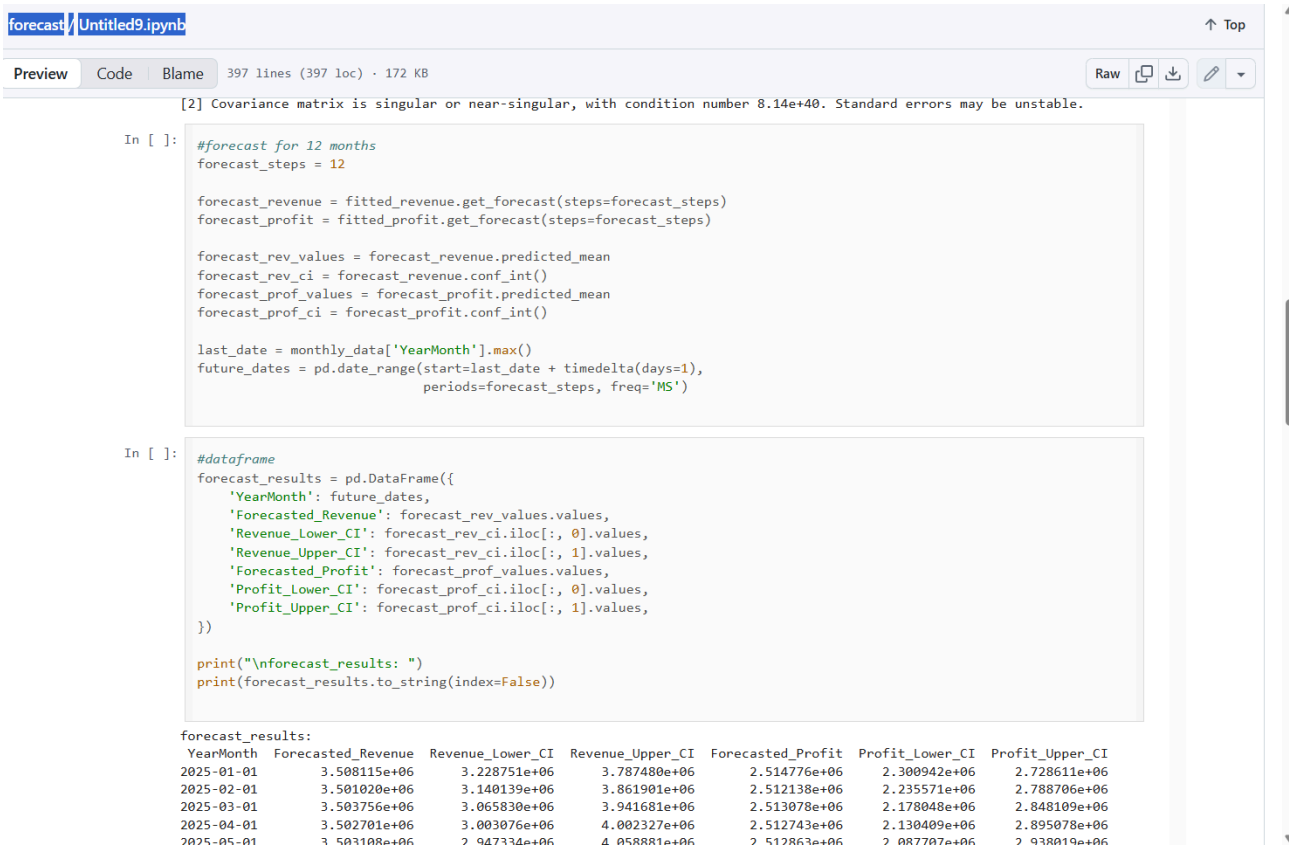
Domain	Key Analytical Questions
Customer	What is the total number of customers? What is the distribution of customers by city, age band (Adult, Young Adult, Elderly, Teen), and gender? What is the total revenue generated by each age band?
Product	What is the total revenue and profit for each brand? How many units were sold per sub-category? How does the final price compare to the profit for each brand and sub-category?
Order & Performance	What is the total number of orders, total revenue, and total profit? How are orders distributed across years and quarters? What is the distribution of orders by payment type (e.g., Cash, Credit Card) and sales channel (e.g., Online, Marketplace)?

4.2. Key Findings

- **Customer Demographics:** The Adult age band (approximately 40% of customers) is the most significant segment, followed by the Young Adult segment. Sales are heavily concentrated in major cities, with Alexandria, Asyut, and Cairo showing the highest customer counts.
- **Product Performance:** Brands such as Anker, Apple, and Bose are the top contributors to both revenue and profit. The highest volume of units sold belongs to categories like Laptop and Smartwatch.
- **Operational Trends:** The analysis of order distribution over time reveals clear seasonal peaks, particularly towards the end of the year. Cash payment remains the dominant transaction method, and the Online channel accounts for the majority of sales.

4.3. Sales Forecasting

A time-series forecasting model was applied to predict future revenue and profit for the next 12 months. The results, including the 95% confidence interval (CI), are presented below as a screenshot from the analysis notebook.


forecast / Untitled9.ipynb

↑ Top

Preview | Code | Blame | 397 lines (397 loc) · 172 KB | Raw

```
[2] Covariance matrix is singular or near-singular, with condition number 8.14e+40. Standard errors may be unstable.
```

In [ ]:

```
#forecast for 12 months
forecast_steps = 12

forecast_revenue = fitted_revenue.get_forecast(steps=forecast_steps)
forecast_profit = fitted_profit.get_forecast(steps=forecast_steps)

forecast_rev_values = forecast_revenue.predicted_mean
forecast_rev_ci = forecast_revenue.conf_int()
forecast_prof_values = forecast_profit.predicted_mean
forecast_prof_ci = forecast_profit.conf_int()

last_date = monthly_data['YearMonth'].max()
future_dates = pd.date_range(start=last_date + timedelta(days=1),
                             periods=forecast_steps, freq='MS')
```

In [ ]:

```
#dataframe
forecast_results = pd.DataFrame({
    'YearMonth': future_dates,
    'Forecasted_Revenue': forecast_rev_values.values,
    'Revenue_Lower_CI': forecast_rev_ci.iloc[:, 0].values,
    'Revenue_Upper_CI': forecast_rev_ci.iloc[:, 1].values,
    'Forecasted_Profit': forecast_prof_values.values,
    'Profit_Lower_CI': forecast_prof_ci.iloc[:, 0].values,
    'Profit_Upper_CI': forecast_prof_ci.iloc[:, 1].values,
})

print("\nforecast_results: ")
print(forecast_results.to_string(index=False))
```

```
forecast_results:
 YearMonth  Forecasted_Revenue  Revenue_Lower_CI  Revenue_Upper_CI  Forecasted_Profit  Profit_Lower_CI  Profit_Upper_CI
2025-01-01        3.508115e+06      3.228751e+06      3.787480e+06       2.514776e+06     2.300942e+06     2.728611e+06
2025-02-01        3.501020e+06      3.140139e+06      3.861901e+06       2.512138e+06     2.235571e+06     2.788706e+06
2025-03-01        3.503756e+06      3.065830e+06      3.941681e+06       2.513078e+06     2.178048e+06     2.848109e+06
2025-04-01        3.502701e+06      3.003076e+06      4.002327e+06       2.512743e+06     2.130409e+06     2.895078e+06
2025-05-01        3.503108e+06      2.947334e+06      4.058881e+06       2.512863e+06     2.087707e+06     2.938019e+06
```

5. Data Visualization

To present the complex analytical findings in an accessible and intuitive manner, an interactive dashboard was developed using Power BI. The dashboard is structured into four main pages: Customers, Products, Orders, and Performance.

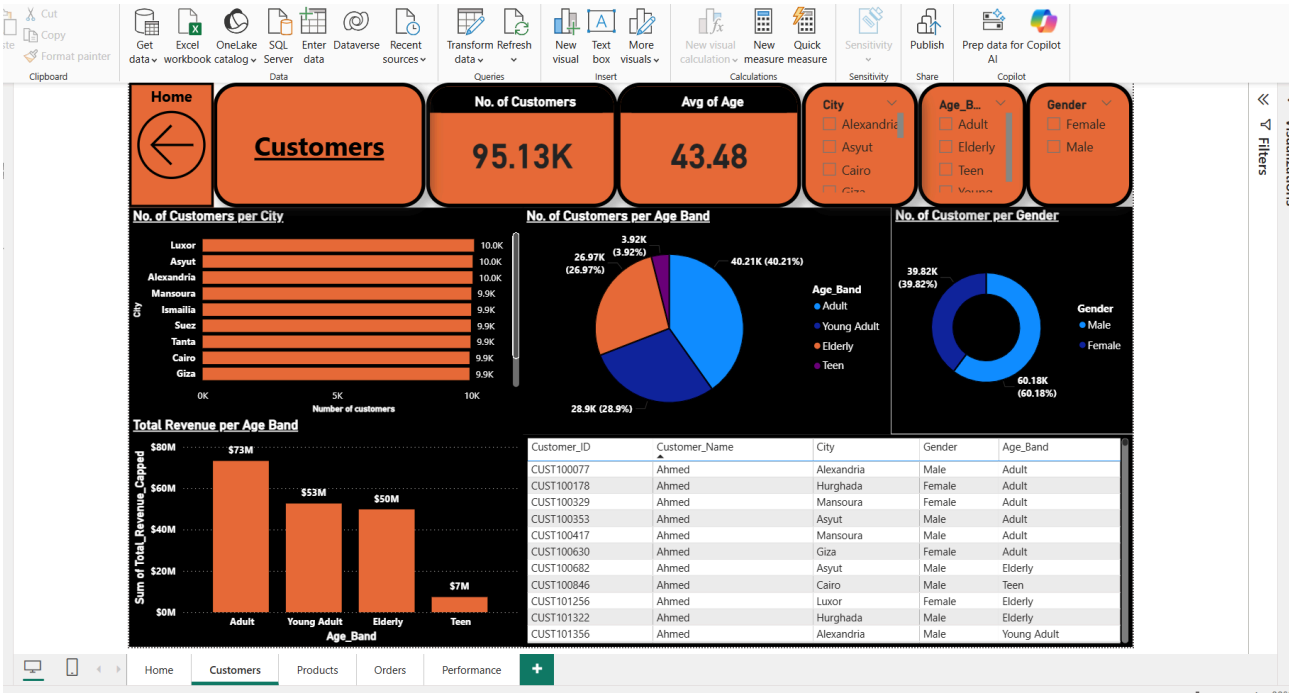
5.1. Visualization Types and Decision Support

The dashboard utilized a variety of visualization types, each serving a specific purpose in supporting decision-making (KPI Cards, Bar Charts, Donut/Pie Charts, Line Charts, Scatter Plots, and Treemaps).

5.2. Dashboard Structure and Visuals

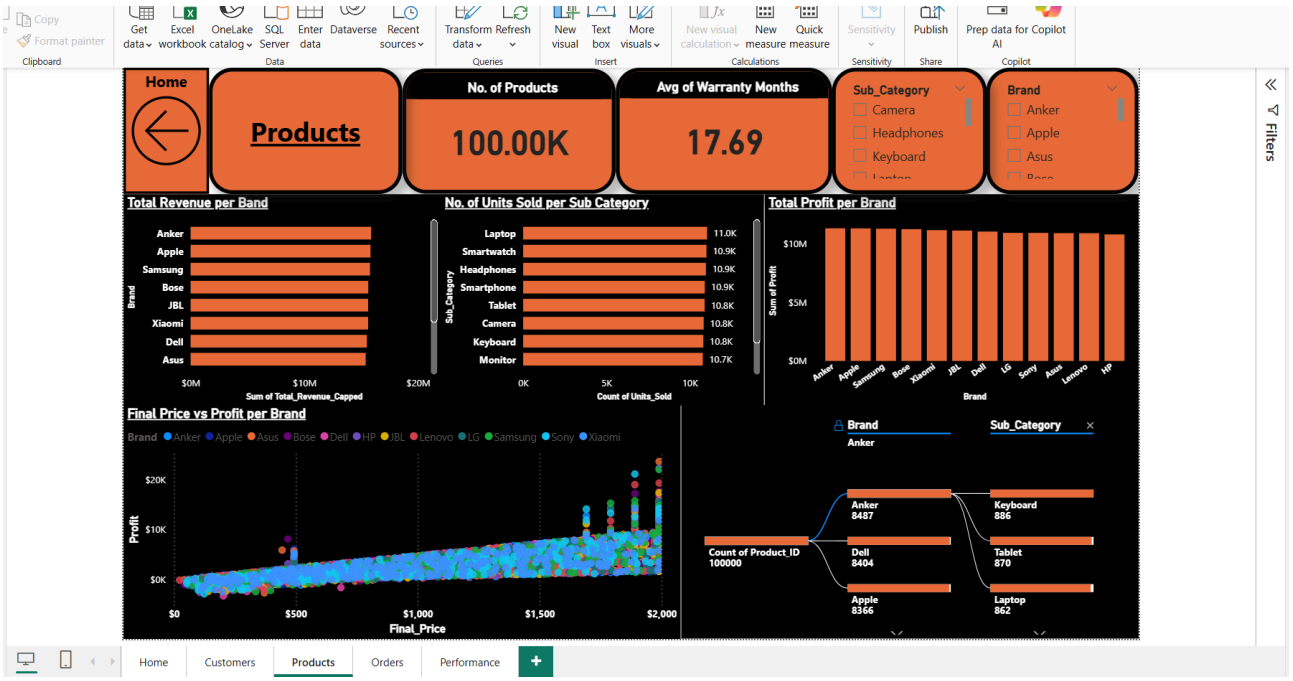
5.2.1. Customers Page

Focuses on demographic analysis, showing customer counts by city, age band, and gender, and linking these to revenue generation.



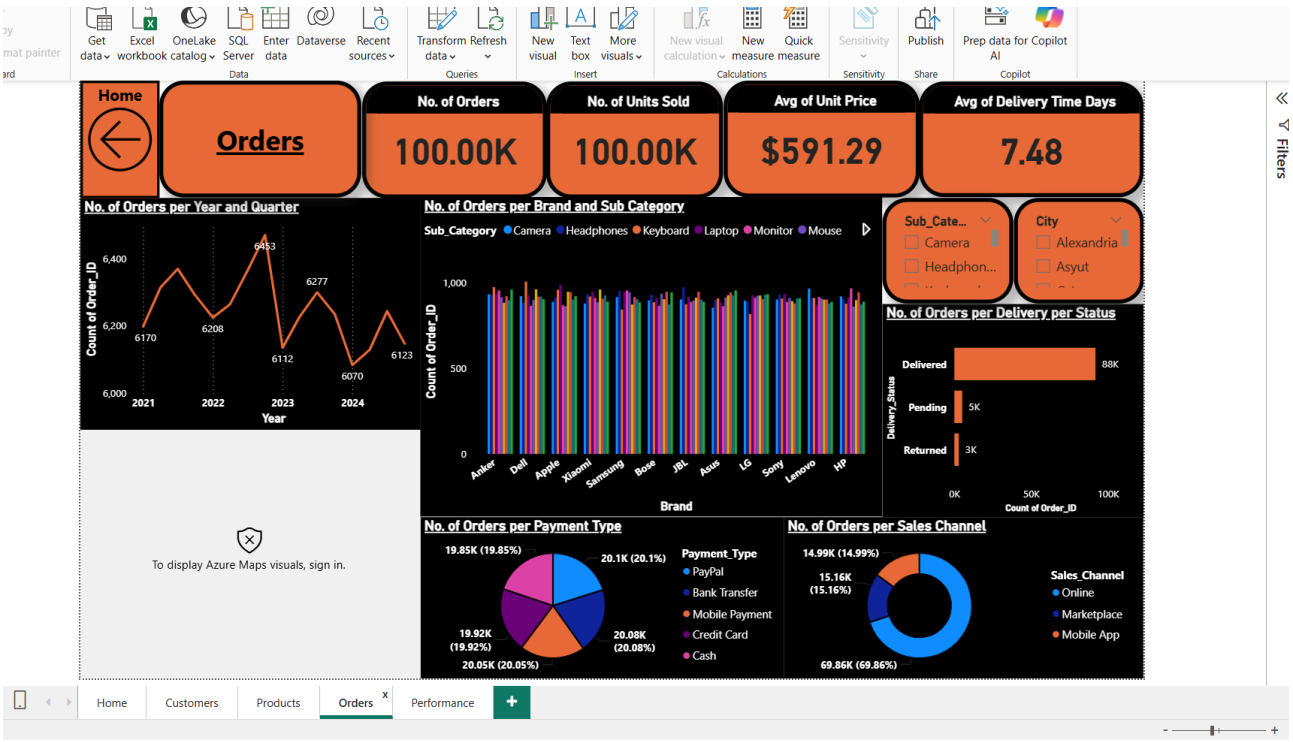
5.2.2. Products Page

Dedicated to product performance, comparing revenue, units sold, and profit across brands and sub-categories.



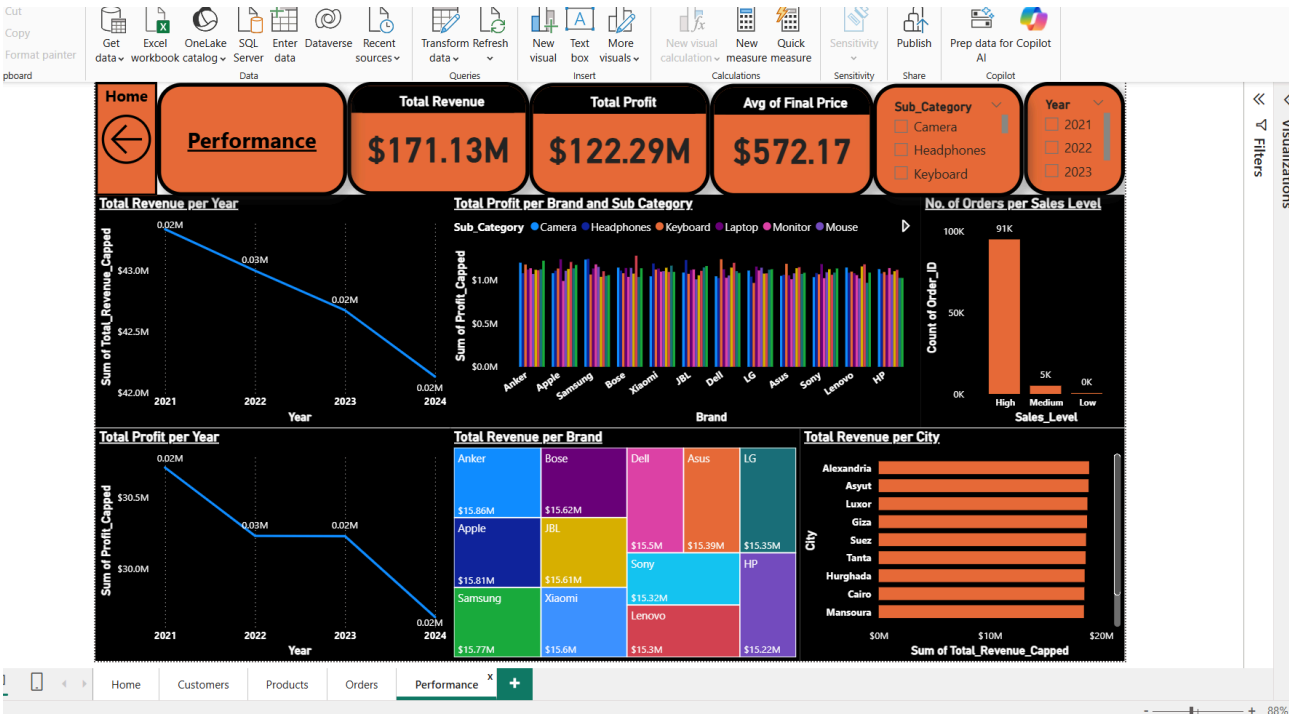
5.2.3. Orders Page

Details operational aspects, including order counts over time, delivery status, payment type distribution, and sales channel usage.



5.2.4. Performance Page

Consolidates key financial metrics (Revenue, Profit) and tracks their annual trends, alongside a detailed breakdown of revenue and profit by brand and city.



6. Insights and Recommendations

6.1. Key Insights Summary

Domain	Insight	Supporting Data
Customer	Adult Segment Dominance: The Adult age band is the most valuable segment, driving the highest revenue and customer count.	Bar chart of Total Revenue by Age Band.
Product	Brand Concentration: A small number of brands (Anker, Apple, Bose) contribute disproportionately to the total profit and revenue.	Treemap and Bar charts of Revenue/Profit by Brand.
Operations	Cash Payment Preference: Cash remains the preferred payment method, indicating a reliance on traditional transaction methods.	Donut chart of Orders by Payment Type (Cash: 60%).
Performance	Seasonal Peaks: Sales volume exhibits clear seasonal peaks, particularly towards the end of the year (Q4).	Line chart of Number of Orders per Year and Quarter.
Geography	Urban Sales Concentration: Sales and customer base are heavily concentrated in major urban centers (Alexandria, Cairo, Asyut).	Bar chart of No. of Customers by City.

6.2. Strategic Recommendations

1. **Targeted Marketing for Adult and Young Adult Segments:** Design marketing campaigns focused on the needs of the Adult and Young Adult segments, emphasizing product utility for professional and educational use.
2. **Optimized Inventory Management for Top Brands:** Implement a robust inventory forecasting model for high-profit brands (Anker, Apple) to prevent stockouts during seasonal peaks.
3. **Incentivize Electronic Payments:** Offer small discounts or loyalty points for customers who opt for electronic payment methods (Credit Card, Mobile App) to gradually shift the payment preference away from cash.
4. **Enhance Logistics in Key Cities:** Invest in improving the supply chain and delivery infrastructure in high-revenue cities to reduce the average delivery time

(currently 2.5 days) and boost customer satisfaction.

5. **Review Low-Profit Product Strategy:** Conduct a deeper analysis of products with low-profit margins to either adjust pricing, bundle them with high-demand items, or consider phasing them out.

---

7. Conclusion

This project successfully delivered a comprehensive analysis of Amazon electronic sales data, transitioning from raw data to strategic insights. The project provides a clear, data-driven foundation for operational and strategic decision-making.

- **Project Impact:** The findings offer actionable strategies to optimize inventory, target high-value customer segments, and improve operational efficiency.