RETAIL SALES BUSINESS INTELLIGENCE PROJECT

EXECUTIVE SUMMARY

This project analyzes a synthetic retail sales dataset to uncover insights that can guide business decisions aimed at maximizing revenue. Using SQL, Google Sheets, Excel, and Tableau, the analysis focuses on key performance indicators (KPIs) including sales performance, customer demographics, and purchasing trends. Insights reveal significant seasonal trends, customer spending patterns, and the impact of gender and age on purchasing decisions. The project includes recommendations for targeted promotions, seasonal discounts, and inventory management, with interactive Tableau visualizations to present findings clearly to both technical and non-technical stakeholders.

Key Concepts: KPIs, Interactive Dashboard, Statistical Tests (t-test, welch's test, chi-square test), EDA

Technologies: MySQL Workbench, Tableau Public 2024, Google Sheets (Pivot table, VLookup, SumIf, Min/Max, Index, Match), Python (Jupyter Notebook, Scipy, Pandas), PowerBI

INTRODUCTION

In this project, I utilize a synthetic retail sales dataset to simulate a real-world scenario where business insights are derived from company data. Specifically, I analyze key performance indicators (KPIs) including sales performance, customer demographics, and purchasing trends, by leveraging SQL, Google Sheets, and Python. I identify and address relevant business questions for each KPI category and then present the findings through an interactive Tableau dashboard. Ultimately, the goal of this project is to provide actionable recommendations to help the "company" maximize their total revenue and to present the insights in a clear and accessible format that can easily be understood by both technical and non-technical stakeholders.

I address the following questions for each KPI category –

Sales Performance:

1. What were the total sales per fiscal quarter and which quarter generated the highest total sales?

2. Is there a statistically significant difference between the means of the total sales of the two fiscal quarters with the highest total sales?
3. What are the total sales per product category (beauty, clothing, electronics), and which category has the highest total sales?

Customer Demographics:

1. What percentage of total sales comes from each age group?
2. Is there an association between gender and product category preference in terms of the number of products bought in that product category?
3. What is the total amount spent per product category by each age group and gender combination and which age group and gender combinations have the highest and lowest total spending?
4. What percentage of total sales comes from each gender/age group combination?

Purchasing Trends:

1. What is the most frequently purchased product category in each fiscal quarter?
2. What percentage does each product category contribute to total revenue?
3. Which product category has the highest average price per unit?
4. What was the highest single transaction amount, and what product(s) contributed to it?

DATA OVERVIEW

The dataset used in this project is a synthetic retail sales dataset sourced from Kaggle. It consists of 1,000 unique transactions and includes the following fields:

1. **Transaction ID** (int): A unique identifier for each transaction.
2. **Date** (date): The date when the transaction occurred.
3. **Customer ID** (varchar): A unique identifier for each customer. (No customer is repeated).
4. **Gender** (varchar): The gender of the customer (Male/Female).
5. **Age** (int): The age of the customer.
6. **Product Category** (varchar): The category of the purchased product (Electronics, Clothing, or Beauty).
7. **Quantity** (int): The number of units of the product purchased.
8. **Price per Unit** (decimal): The price of one unit of the product.
9. **Total Amount** (decimal): The total monetary value of the transaction (aka total sales).

To prepare the dataset for analysis, I used the following SQL queries to load and validate the data:

```sql
DROP TABLE IF EXISTS sales_data;

CREATE TABLE sales_data (
    Transaction_ID INT,
    Date DATE,
    Customer_ID VARCHAR(50),
    Gender VARCHAR(10),
    Age INT,
    Product_Category VARCHAR(50),
    Quantity INT,
    Price_Per_Unit DECIMAL(10, 2),
    Total_Amount DECIMAL(10, 2)
);


LOAD DATA INFILE 'C:/ProgramData/MySQL/MySQL Server
8.0/Uploads/retail_sales_dataset.csv'
INTO TABLE sales_data
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES
(Transaction_ID, Date, Customer_ID, Gender, Age, Product_Category, Quantity,
Price_Per_Unit, Total_Amount);

SELECT * FROM sales_data;
```

Below is a sample of the dataset:

Transaction_ID	Date	Customer_ID	Gender	Age	Product_Category	Quantity	Price_Per_Unit	Total_Amount
1	2023-11-24	CUST001	Male	34	Beauty	3	50.00	150.00
2	2023-02-27	CUST002	Female	26	Clothing	2	500.00	1000.00
3	2023-01-13	CUST003	Male	50	Electronics	1	30.00	30.00
4	2023-05-21	CUST004	Male	37	Clothing	1	500.00	500.00
5	2023-05-06	CUST005	Male	30	Beauty	2	50.00	100.00

METHODOLOGY AND RESULTS

OVERVIEW

The methodology begins with exploratory data analysis (EDA) to identify potential biases or anomalies in the dataset. Then, the specific KPI questions mentioned above are addressed using SQL, Google Sheets, and/or Python and, where applicable, statistical tests.

EXPLORATORY DATA ANALYSIS (EDA)

Assessing data completeness –

1. Checking for missing or null values in the dataset:

```
SELECT
   SUM(CASE WHEN Transaction_ID IS NULL OR Transaction_ID = 0 THEN 1 ELSE 0
END) AS Invalid_Transaction_ID,
   SUM(CASE WHEN Date IS NULL THEN 1 ELSE 0 END) AS Missing_Date,
   SUM(CASE WHEN Customer_ID IS NULL OR Customer_ID = '' THEN 1 ELSE 0 END)
AS Invalid_Customer_ID,
   SUM(CASE WHEN Gender IS NULL OR Gender = '' THEN 1 ELSE 0 END) AS
Invalid_Gender,
   SUM(CASE WHEN Age IS NULL OR Age = 0 THEN 1 ELSE 0 END) AS Invalid_Age,
   SUM(CASE WHEN Product_Category IS NULL OR Product_Category = '' THEN 1 ELSE
0 END) AS Invalid_Product_Category,
   SUM(CASE WHEN Quantity IS NULL OR Quantity = 0 THEN 1 ELSE 0 END) AS
Invalid_Quantity,
   SUM(CASE WHEN Price_Per_Unit IS NULL OR Price_Per_Unit = 0 THEN 1 ELSE 0
END) AS Invalid_Price_Per_Unit,
   SUM(CASE WHEN Total_Amount IS NULL OR Total_Amount = 0 THEN 1 ELSE 0
END) AS Invalid_Total_Amount
FROM sales_data;
```

Output: None found

2. Checking for duplicate transactions:

```
SELECT
   Transaction_ID,
   COUNT(*) AS Record_Count
FROM sales_data
GROUP BY Transaction_ID
HAVING COUNT(*) > 1;
```

Output: None found

3. Checking for invalid values:

a. Negative quantities or prices:

```
SELECT *
FROM sales_data
WHERE Quantity < 0 OR Price_Per_Unit < 0;
```

Output: None found

b. Age outside of a reasonable range:

```
SELECT *
FROM sales_data
WHERE Age < 0 OR Age > 100;
```

Output: None found

Analyzing basic statistics/distributions of the age, gender, product category, and date variables –

Please Note: Graphs were generated through Google Sheets

1. Age:
    a. Basic stats:

```
SELECT
   MIN(Age) AS Min_Age,
   MAX(Age) AS Max_Age,
   AVG(Age) AS Avg_Age,
   STDDEV(Age) AS StdDev_Age
FROM sales_data;
```

Output:

Min_Age	Max_Age	Avg_Age	StdDev_Age
18	64	41.3920	13.674587233258626

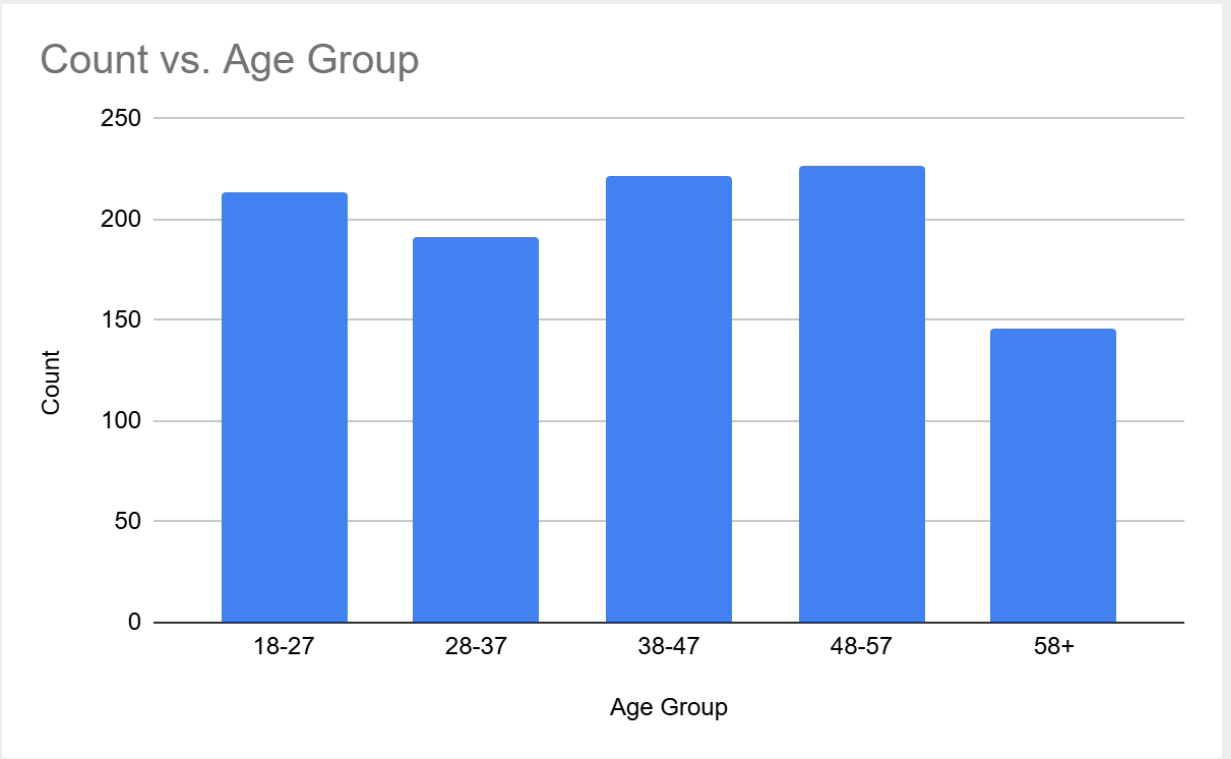
b. Distribution:

```
SELECT
```

```
    CASE
        WHEN Age BETWEEN 18 AND 27 THEN '18-27'
        WHEN Age BETWEEN 28 AND 37 THEN '28-37'
        WHEN Age BETWEEN 38 AND 47 THEN '38-47'
        WHEN Age BETWEEN 48 AND 57 THEN '48-57'
        WHEN Age > 57 THEN '58+'
    END AS Age_Group,
    COUNT(*) AS Count
FROM sales_data
GROUP BY Age_Group
ORDER BY Age_Group;
```

Output:

Age_Group	Count
18-27	214
28-37	191
38-47	222
48-57	227
58+	146

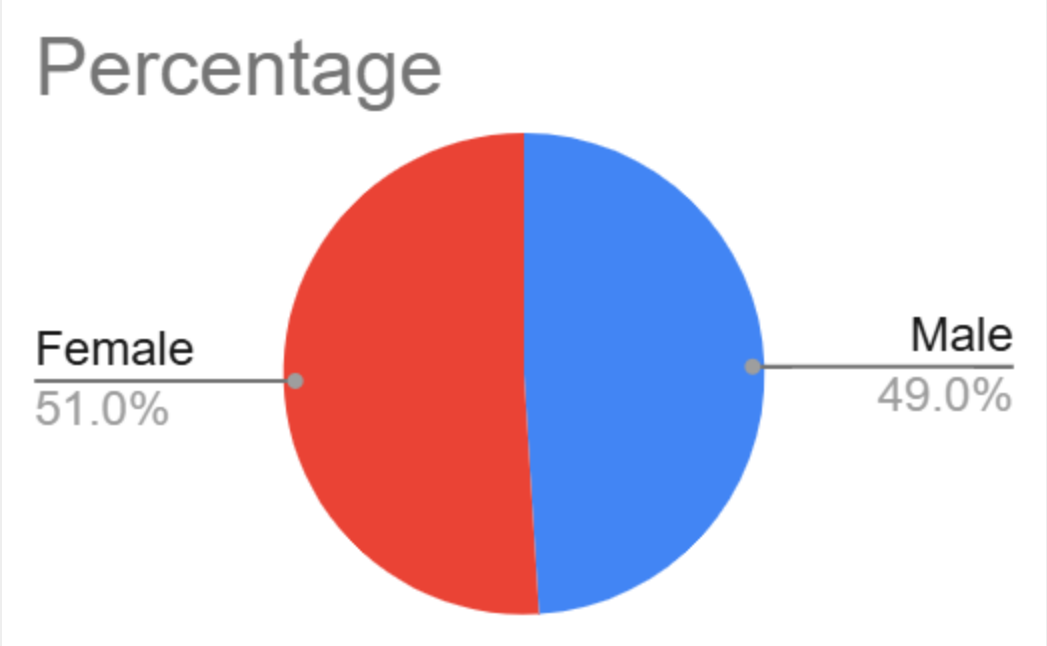


2. Gender:

```
SELECT
    Gender,
    COUNT(*) AS Count,
    ROUND(COUNT(*) * 100.0 / (SELECT COUNT(*) FROM sales_data), 2) AS Percentage
FROM sales_data
GROUP BY Gender;
```

Output:

Gender	Count	Percentage
Male	490	49.00
Female	510	51.00

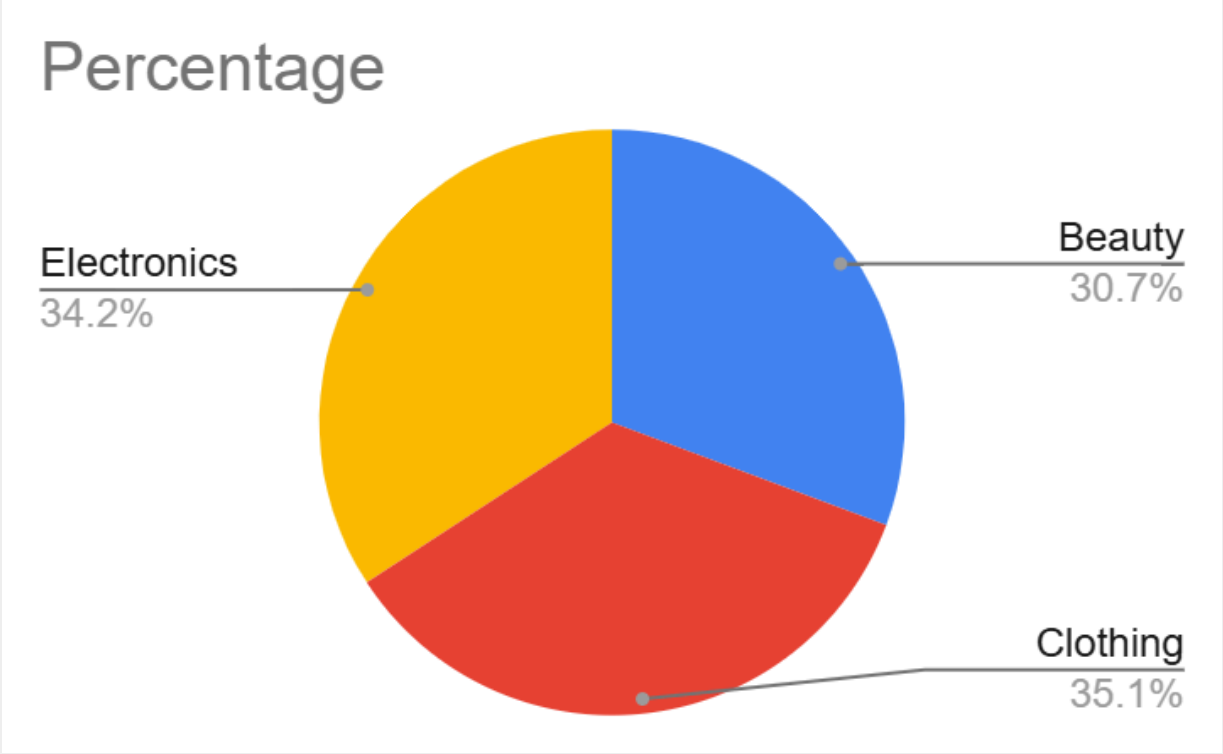


3. Product category:

```
SELECT
    Product_Category,
    COUNT(*) AS Count,
    ROUND(COUNT(*) * 100.0 / (SELECT COUNT(*) FROM sales_data), 2) AS Percentage
FROM sales_data
GROUP BY Product_Category;
```

Output:

Product_Category	Count	Percentage
Beauty	307	30.70
Clothing	351	35.10
Electronics	342	34.20



4. Date:

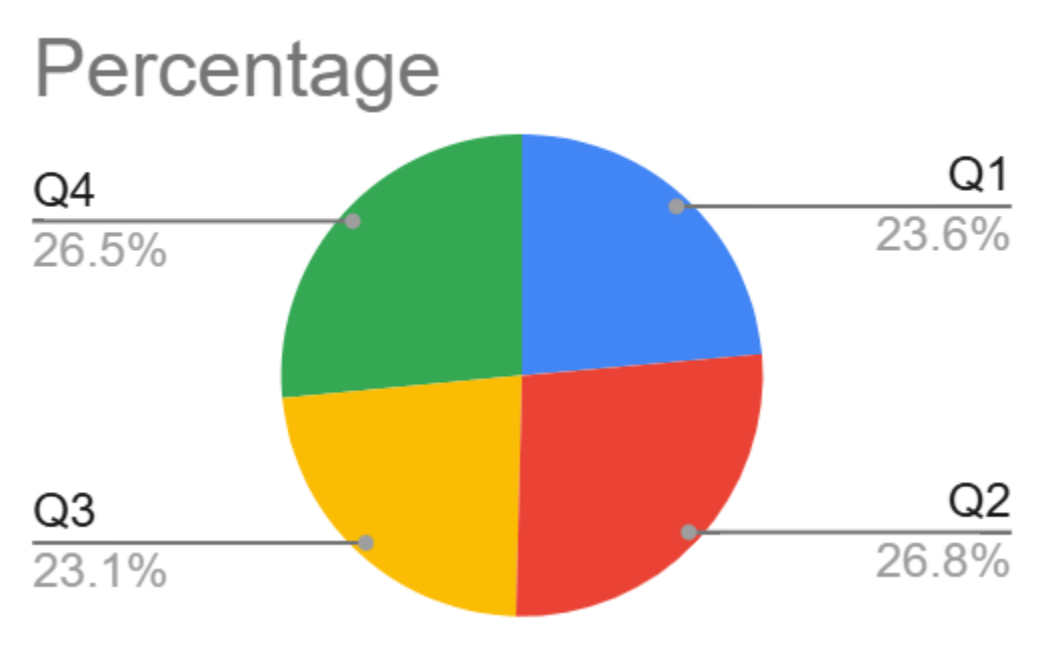
   Please Note: The dates were converted into fiscal quarters to make the analysis more relevant to a business environment. Fiscal quarters are a three-month period in a company's financial year used for reporting earnings and paying dividends. Analyzing data by fiscal quarters offers valuable insights into seasonal trends and performance cycles, which can guide inventory and marketing strategies. They are typically defined as January, February, and March being the first quarter, April, May, and June being the second quarter, July, August, and September being the third quarter, and October, November, and December being the fourth quarter.

```
SELECT
  CASE
    WHEN MONTH(Date) IN (1, 2, 3) THEN 'Q1'
    WHEN MONTH(Date) IN (4, 5, 6) THEN 'Q2'
    WHEN MONTH(Date) IN (7, 8, 9) THEN 'Q3'
    WHEN MONTH(Date) IN (10, 11, 12) THEN 'Q4'
  END AS Fiscal_Quarter,
  COUNT(*) AS Count,
  ROUND(COUNT(*) * 100.0 / (SELECT COUNT(*) FROM sales_data), 2) AS Percentage
FROM sales_data
```

```
GROUP BY Fiscal_Quarter
ORDER BY FIELD(Fiscal_Quarter, 'Q1', 'Q2', 'Q3', 'Q4');
```

Output:

Fiscal_Quarter	Count	Percentage
Q1	236	23.60
Q2	268	26.80
Q3	231	23.10
Q4	265	26.50



Detecting outliers or anomalies in transaction amounts or quantities –

1. Transaction amount outliers:

```
SELECT *
FROM sales_data
WHERE Total_Amount > (SELECT AVG(Total_Amount) + 3 * STDDEV(Total_Amount)
FROM sales_data)
```

```
  OR Total_Amount < (SELECT AVG(Total_Amount) - 3 * STDDEV(Total_Amount) FROM
sales_data);
```

Output: None found

2. Quantity outliers:

```
SELECT * FROM sales_data WHERE Quantity > (SELECT AVG(Quantity) + 3 *
STDDEV(Quantity) FROM sales_data) OR Quantity < (SELECT AVG(Quantity) - 3 *
STDDEV(Quantity) FROM sales_data);
```

Output: None found

3. Price per unit outliers:

```
SELECT * FROM sales_data WHERE Price_Per_Unit > (SELECT AVG(Price_Per_Unit) + 3
* STDDEV(Price_Per_Unit) FROM sales_data) OR Price_Per_Unit < (SELECT
AVG(Price_Per_Unit) - 3 * STDDEV(Price_Per_Unit) FROM sales_data);
```

Output: None found

EDA Findings Summary –

The exploratory data analysis revealed no missing or null values, duplicate transactions, or invalid entries in the dataset. The distributions across age, gender, product category, and fiscal quarter categories appear well-balanced in terms of total transactions. While the “58+” age group was slightly under-represented, this likely reflects the narrower age range within this category, given that the maximum age in the dataset is 64, rather than any inherent age bias. Additionally, no outliers were detected in transaction amounts, quantities, or prices per unit. The absence of significant anomalies or biases ensures that the dataset is reliable and representative, eliminating the need for further cleaning or adjustments before question formation or analysis.

KPI QUESTION ANALYSIS

Sales Performance –

1. **What were the total sales per fiscal quarter and which quarter generated the highest total sales?**

Significance: This question would allow us to see how much revenue is generated for each quarter of the year and which quarter is the most profitable allowing the “company” to plan their financial strategies accordingly.

Analysis method: SQL will be used to aggregate total sales per fiscal quarter and determine which quarter had the highest sales because SQL is ideal for large queries.

```
SELECT
   CASE
      WHEN MONTH(Date) IN (1, 2, 3) THEN 'Q1'
      WHEN MONTH(Date) IN (4, 5, 6) THEN 'Q2'
      WHEN MONTH(Date) IN (7, 8, 9) THEN 'Q3'
      WHEN MONTH(Date) IN (10, 11, 12) THEN 'Q4'
   END AS Fiscal_Quarter,
   SUM(Total_Amount) AS Total_Sales
FROM sales_data
GROUP BY Fiscal_Quarter
ORDER BY Total_Sales DESC;
```

Output:

Fiscal_Quarter	Total_Sales
Q4	126190.00
Q2	123735.00
Q1	110030.00
Q3	96045.00

Takeaway: It seems as though quarter 4 had the highest total sales which could indicate increased consumer spending during the holiday season (October - December).

2. **Is there a statistically significant difference between the means of the total sales of the two fiscal quarters with the highest total sales?**

Significance: This helps determine if the difference between the sales of the two leading quarters is significant or if the difference is due to random fluctuations.

Analysis method: Python will be used for conducting a statistical test to compare the means of the total sales in the top two quarters since it has built-in libraries for statistical tests.

From the question above, it can be seen that the top 2 fiscal quarters in terms of highest total sales were Q4 and Q2. Hence, this information is used in the statistical tests below:

2-sample t-test –

Checking requirements:

- ☑ Independence: Given this dataset was synthetically created, I assume that the samples are independent as there doesn't seem to be any known relationship between Q4 and Q2.
- ☑ Random sampling: Given this dataset was synthetically created, I assume that the samples were selected randomly as there doesn't seem to be any apparent bias in the data.
- ☑ Continuous data: The data in the "Total Amount" column in the dataset is continuous.
- ☐ Normal distribution: This requirement was not met (see image below).
- ☑ Equal variances: This requirement was met (see image below).

Checking normality and equaliy of variances requirements for a 2-sample t-test:

```
from scipy.stats import shapiro, levene

# Filter sales data for Q4 (October - December) and Q2 (April - June)
df['Date'] = pd.to_datetime(df['Date'])  # Ensure the Date column is in datetime format
df['Month'] = df['Date'].dt.month       # Extract the month from the date

# Q4: Filter for months October (10), November (11), and December (12)
q4_sales = df[df['Month'].isin([10, 11, 12])]['Total Amount']

# Q2: Filter for months April (4), May (5), and June (6)
q2_sales = df[df['Month'].isin([4, 5, 6])]['Total Amount']

# Shapiro-Wilk test for normality
q4_normality_test = shapiro(q4_sales)
q2_normality_test = shapiro(q2_sales)

print("Shapiro-Wilk Test Results:")
print(f"Q4: Statistic = {q4_normality_test.statistic}, p-value = {q4_normality_test.pvalue}")
print(f"Q2: Statistic = {q2_normality_test.statistic}, p-value = {q2_normality_test.pvalue}")

# Interpretation for Shapiro-Wilk:
if q4_normality_test.pvalue > 0.05:
    print("Q4 appears to be normally distributed (p > 0.05).")
else:
    print("Q4 does not appear to be normally distributed (p ≤ 0.05).")

if q2_normality_test.pvalue > 0.05:
    print("Q2 appears to be normally distributed (p > 0.05).")
else:
    print("Q2 does not appear to be normally distributed (p ≤ 0.05).")

# Levene's test for equality of variances
levene_test = levene(q4_sales, q2_sales)
print("\nLevene's Test Results:")
print(f"Statistic = {levene_test.statistic}, p-value = {levene_test.pvalue}")

# Interpretation for Levene's test:
if levene_test.pvalue > 0.05:
    print("The variances of Q4 and Q2 are equal (p > 0.05).")
else:
    print("The variances of Q4 and Q2 are not equal (p ≤ 0.05).")
```

```
Shapiro-Wilk Test Results:
Q4: Statistic = 0.7594879269599915, p-value = 2.1484601156735604e-19
Q2: Statistic = 0.7523835897445679, p-value = 8.97643023681077e-20
Q4 does not appear to be normally distributed (p ≤ 0.05).
Q2 does not appear to be normally distributed (p ≤ 0.05).

Levene's Test Results:
Statistic = 0.03499498827340892, p-value = 0.8516776148399059
The variances of Q4 and Q2 are equal (p > 0.05).
```

Given that all the conditions except the normal distribution condition were met for the 2-sample t-test, we are going to use a Welch's t-test for this analysis –

We will use the following hypotheses for the test:

H0: There is no statistically significant difference between the means of Q4 and Q2 total sales.

H1: There is a statistically significant difference between the means of Q4 and Q2 total sales.

Conducting Welch's T-test: ¶

```
from scipy.stats import ttest_ind

# Conduct Welch's t-test
welch_ttest = ttest_ind(q4_sales, q2_sales, equal_var=False)

# Output the results
print("Welch's t-test Results:")
print(f"Statistic = {welch_ttest.statistic:.4f}, p-value = {welch_ttest.pvalue:.4f}")

# Interpretation
alpha = 0.05  # Significance level
if welch_ttest.pvalue > alpha:
    print(f"Fail to reject H₀ (p = {welch_ttest.pvalue:.4f} > {alpha}).")
    print("There is no statistically significant difference between the means of Q4 and Q2 total sales.")
else:
    print(f"Reject H₀ (p = {welch_ttest.pvalue:.4f} ≤ {alpha}).")
    print("There is a statistically significant difference between the means of Q4 and Q2 total sales.")
```

```
Welch's t-test Results:
Statistic = 0.2937, p-value = 0.7691
Fail to reject H₀ (p = 0.7691 > 0.05).
There is no statistically significant difference between the means of Q4 and Q2 total sales.
```

Takeaway: Given the results shown above, we fail to reject the null hypothesis concluding that there is no statistically significant difference between the means of the Q4 and Q2 total sales.

3. **What are the total sales per product category (Beauty, Clothing, Electronics), and which category has the highest total sales?**

Significance: Identifying the best-performing product categories can help the “company” with product line decisions and inventory allocation.

Analysis method: SQL will be used to aggregate sales by product category.

```
SELECT
   Product_Category,
   SUM(Total_Amount) AS Total_Sales
FROM
   sales_data
GROUP BY
   Product_Category
```

```
ORDER BY
    Total_Sales DESC;
```

Output:

Product_Category	Total_Sales
Electronics	156905.00
Clothing	155580.00
Beauty	143515.00

Takeaway: The product category that seems to have the highest total sales is electronics.

Customer Demographics –

1. **What percentage of total sales comes from each age group?**

   Significance: This question gives insights into how much each age group contributes to sales, helping with targeted marketing and promotional strategies.

   Analysis method: SQL will be used to calculate the percentage of total sales for each age group.

```
SELECT
    CASE
        WHEN Age BETWEEN 18 AND 27 THEN '18-27'
        WHEN Age BETWEEN 28 AND 37 THEN '28-37'
        WHEN Age BETWEEN 38 AND 47 THEN '38-47'
        WHEN Age BETWEEN 48 AND 57 THEN '48-57'
        ELSE '58+'
    END AS Age_Group,
    SUM(Total_Amount) AS Total_Sales,
    (SUM(Total_Amount) / (SELECT SUM(Total_Amount) FROM sales_data) * 100) AS
Percentage_of_Total_Sales
FROM
    sales_data
GROUP BY
```

```
    Age_Group
ORDER BY
    Percentage_of_Total_Sales DESC;
```

Output:

Age_Group	Total_Sales	Percentage_of_Total_Sales
18-27	107915.00	23.665570
38-47	96710.00	21.208333
28-37	95870.00	21.024123
48-57	93825.00	20.575658
58+	61680.00	13.526316

Takeaway: For this "company", it seems as though the most profitable age group is 18-27.

2. **Is there an association between gender and product category preference in terms of the number of products bought in that product category?**

   Significance: Understanding whether gender influences product preferences can help tailor product recommendations and marketing efforts.

   Analysis method: Python's built-in libraries will be used to perform a Chi-square test of independence to assess whether there's a statistically significant association between gender and product category.

   Chi-square test –

   Checking requirements:

   - ☑ Categorical variables: Gender is categorical.
   - ☑ Random Sampling: Given this dataset was synthetically created, I assume that the samples were selected randomly as there doesn't seem to be any apparent bias in the data.
   - ☑ Expected frequencies: expected frequencies are greater than or equal to 1 (see image below).

Checking expected frequencies requirement for Chi-square test:

```
from scipy.stats import chi2_contingency

# Create a contingency table for Gender and Product_Category
contingency_table = pd.crosstab(df['Gender'], df['Product Category'])

# Perform the Chi-square test to get the expected frequencies
chi2, p, dof, expected = chi2_contingency(contingency_table)

# Convert the expected frequencies to a DataFrame for easier interpretation
expected_df = pd.DataFrame(expected,
                           index=contingency_table.index,
                           columns=contingency_table.columns)

# Check the condition that all expected frequencies are >= 1
condition_met = (expected >= 1).all()

print("Contingency Table (Observed):")
print(contingency_table)
print("\nExpected Frequencies:")
print(expected_df)
print(f"\nCondition (All expected frequencies >= 1): {'Yes' if condition_met else 'No'}")
```

```
Contingency Table (Observed):
Product Category  Beauty  Clothing  Electronics
Gender
Female               166       174          170
Male                 141       177          172

Expected Frequencies:
Product Category  Beauty  Clothing  Electronics
Gender
Female            156.57    179.01       174.42
Male              150.43    171.99       167.58

Condition (All expected frequencies >= 1): Yes
```

Give that all the requirements for the Chi-square test were met, we will proceed with conducting it:

We will use the following hypotheses for the test:

H_0: There is no association between gender and product category preference. Gender and product category are independent.

H_1: There is an association between gender and product category preference. Gender and product category are not independent.

Conducting Chi-square test:

```python
from scipy.stats import chi2_contingency

# Perform the Chi-square test
chi2, p, dof, expected = chi2_contingency(contingency_table)

# Print the results
print("Chi-square Test Results:")
print(f"Chi-square statistic: {chi2:.4f}")
print(f"P-value: {p:.4f}")
print(f"Degrees of freedom: {dof}")
print("\nExpected Frequencies:")
print(expected_df)

# Interpret the results
alpha = 0.05  # Significance level
if p < alpha:
    print("\nConclusion: Reject the null hypothesis (H₀).")
    print("There is a statistically significant association between gender and product category preference.")
else:
    print("\nConclusion: Fail to reject the null hypothesis (H₀).")
    print("There is no statistically significant association between gender and product category preference.")
```

```
Chi-square Test Results:
Chi-square statistic: 1.6738
P-value: 0.4330
Degrees of freedom: 2

Expected Frequencies:
Product Category  Beauty  Clothing  Electronics
Gender
Female            156.57    179.01       174.42
Male              150.43    171.99       167.58

Conclusion: Fail to reject the null hypothesis (H₀).
There is no statistically significant association between gender and product category preference.
```

Takeaway: As it can be seen, we fail to reject the null hypothesis. This indicates that there is no statistically significant association between gender and product category preference.

3. **What is the total amount spent per product category by each age group and gender combination and which age group and gender combinations have the highest and lowest total spending?**

   Significance: This question can reveal which demographic groups are the biggest spenders for each product category, guiding targeted marketing and inventory decisions.

   Analysis method: Google sheets will be used to create a pivot table and then MAX and MIN will be used to find the necessary values.

	A	B	C	D	E	F
1	*Demographic Grouping*		*Total Amount Spent by Product Category*			
2	*Age Group*	*Gender*	Beauty	Clothing	Electronics	Grand Total
3	18-27	Female	16450	21440	16235	54125
4		Male	21300	18010	14480	53790
5	18-27 Total		37750	39450	30715	107915
6	28-37	Female	17035	14285	17950	49270
7		Male	12680	21430	12490	46600
8	28-37 Total		29715	35715	30440	95870
9	38-47	Female	19200	14915	15505	49620
10		Male	18770	11350	16970	47090
11	38-47 Total		37970	26265	32475	96710
12	48-57	Female	14490	18535	18245	51270
13		Male	11480	14090	16985	42555
14	48-57 Total		25970	32625	35230	93825
15	58+	Female	7655	12100	8800	28555
16		Male	4455	9425	19245	33125
17	58+ Total		12110	21525	28045	61680
18	**Grand Total**		**143515**	**155580**	**156905**	**456000**
19						

Google Sheets Formulas: =MAX(F3:F4, F6:F7, F9:F10, F12:F13, F15:F16), =MIN(F3:F4, F6:F7, F9:F10, F12:F13, F15:F16)
Outputs: 54125, 28555

Takeaway: From the results, we can see that the demographic group that spends the most are 18-27 year old females and the demographic group that spends the least is 58+ females.

4. **What percentage of total sales comes from each gender/age group combination?**

Significance: This question allows us to see exactly what percentage each age and gender combination contributes to the total sales of the “company” allowing for a more clear view of how much each demographic group is spending in this “business”.

Analysis method: Using our pivot table from the previous question, we can use VLOOKUP and/or INDEX & MATCH formulas to calculate the necessary values.
Please note: While vlookup and index/match formulas were not necessary for these calculations, I wanted to demonstrate their use in this project.

General vlookup formula used:

=((VLOOKUP(F3, $F$3:$F$16, 1, FALSE))/456000)*100

I double checked the results of my vlookup by using index/match. Again, while this wasn't necessary, I wanted to demonstrate its usage in this project.

General index/match formula used:
=((INDEX($F$3:$F$16, MATCH(F3, $F$3:$F$16, 0)))/456000)*100

Output:

Demographic Grouping		Total Amount Spent by Product Category				Calcuating Percentage Contributions to Total Sales by Age/Gender Combinations:	
Age Group	Gender	Beauty	Clothing	Electronics	Grand Total		
18-27	Female	16450	21440	16235	54125	11.86951754	11.86951754
	Male	21300	18010	14480	53790	11.79605263	11.79605263
18-27 Total		37750	39450	30715	107915	23.66557018	23.66557018
28-37	Female	17035	14285	17950	49270	10.80482456	10.80482456
	Male	12680	21430	12490	46600	10.21929825	10.21929825
28-37 Total		29715	35715	30440	95870	21.02412281	21.02412281
38-47	Female	19200	14915	15505	49620	10.88157895	10.88157895
	Male	18770	11350	16970	47090	10.32675439	10.32675439
38-47 Total		37970	26265	32475	96710	21.20833333	21.20833333
48-57	Female	14490	18535	18245	51270	11.24342105	11.24342105
	Male	11480	14090	16985	42555	9.332236842	9.332236842
48-57 Total		25970	32625	35230	93825	20.57565789	20.57565789
58+	Female	7655	12100	8800	28555	6.262061404	6.262061404
	Male	4455	9425	19245	33125	7.264254386	7.264254386
58+ Total		12110	21525	28045	61680	#N/A	#N/A
Grand Total		**143515**	**155580**	**156905**	**456000**	#N/A	#N/A
						(vlookup) ^	(index/match) ^

Takeaway: From the table above, we can see the percentage that each age/gender combination contributes to the total sales of the company and that the results from both our vlookup and index/match formulas matched. Also, next to 18-27 year old females contributing the most at around 11.87%, we have 18-27 year old males contributing the second most at around 11.80% and next to 58+ year old females contributing the least at around 6.26%, we have 58+ year old males contributing the second least at around 7.26%.

Purchasing Trends –

1. **What is the most frequently purchased product category in each fiscal quarter?**

Significance: Understanding product category popularity per quarter can inform promotional strategies, inventory decisions, and sales forecasting.

Analysis method: SQL will be used to identify the most frequently purchased product category per fiscal quarter by counting transactions and grouping by quarter and category.

```
WITH QuarterCategoryCounts AS (
    SELECT
        QUARTER(Date) AS FiscalQuarter,
        Product_Category,
        COUNT(*) AS TransactionCount
    FROM
        sales_data
    GROUP BY
        QUARTER(Date), Product_Category
),
RankedCategories AS (
    SELECT
        FiscalQuarter,
        Product_Category,
        TransactionCount,
        ROW_NUMBER() OVER (PARTITION BY FiscalQuarter ORDER BY
TransactionCount DESC) AS RowNum
    FROM
        QuarterCategoryCounts
)
SELECT
    FiscalQuarter,
    Product_Category AS MostFrequentlyPurchasedCategory,
    TransactionCount
FROM
    RankedCategories
WHERE
    RowNum = 1;
```

Output:

FiscalQuarter	MostFrequentlyPurchasedCategory	TransactionCount
1	Clothing	97
2	Clothing	101
3	Electronics	89
4	Electronics	102

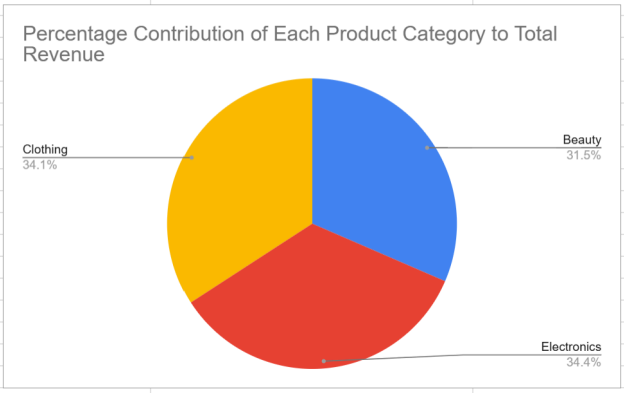
Takeaway: From the table above, it can be seen that for quarters 1 and 2, clothing is the most frequently purchased product category and in quarters 3 and 4, electronics is the most frequently purchased product category.

2. **What percentage does each product category contribute to total revenue?**

Significance: This highlights the relative importance of each product category in driving total sales, guiding inventory and pricing decisions.

Analysis method: Google Sheets will be used for this analysis as it would allow for a pie chart representing the percentages to be created easily.

Total Revenue ("=SUM(L:L)")	Beauty Total Revenue ("=SUMIF(H2:H1000, "Beauty", L2:L1000)")	Electronics Total Revenue	Clothing Total Revenue
456000	143515	156785	155580
	Beauty Total Revenue Percentage Contribution ("(N2/M2)*100")	Electronics Total Revenue Percentage Contribution:	Clothing Total Revenue Percentage Contribution:
	31.47258772	34.38267544	34.11842105
	Beauty	31.47258772	
	Electronics	34.38267544	
	Clothing	34.11842105	



Takeaway: From the pie chart above, it can be seen that each product category contributes approximately equally to the total revenue with beauty contributing slightly less than clothing and electronics.

3. **Which product category has the highest average price per unit?**

Significance: Identifying which categories have the highest price points helps with pricing strategies and inventory management, ensuring profitability.

Analysis method: SQL will be used to calculate the average price per unit for each product category

```
SELECT
   `Product_Category`,
   AVG(`Price_Per_Unit`) AS `Average Price per Unit`
FROM
   sales_data
GROUP BY
   `Product_Category`
ORDER BY
   `Average Price per Unit` DESC
```

Output:

Product_Category	Average Price per Unit
Beauty	184.055375
Electronics	181.900585
Clothing	174.287749

Takeaway: From the table above, we can see that beauty has the highest average price per unit, followed by electronics, and then clothing.

4. **What was the highest single transaction amount, and what product(s) contributed to it?**

   Significance: Identifying the highest single transaction can provide insights into high-value customers and successful product bundles, helping with promotional efforts.

   Analysis method: SQL will be used to find the highest transaction by identifying the transaction with the maximum total amount, followed by a query to identify the products involved.

```
SELECT
   `Transaction_ID`,
   `Product_Category`,
```

```
    `Quantity`,
    `Price_Per_Unit`,
    `Total_Amount`
FROM
    sales_data
WHERE
    `Total_Amount` = (SELECT MAX(`Total_Amount`) FROM sales_data);
```

Output:

(First few rows)

Transaction_ID	Product_Category	Quantity	Price_Per_Unit	Total_Amount
15	Electronics	4	500.00	2000.00
65	Electronics	4	500.00	2000.00
72	Electronics	4	500.00	2000.00
74	Beauty	4	500.00	2000.00
89	Electronics	4	500.00	2000.00
93	Beauty	4	500.00	2000.00
109	Electronics	4	500.00	2000.00
118	Electronics	4	500.00	2000.00

From our SQL query, tt seems like there were multiple transactions with the highest amount of $2000 so now we will see of these transactions, what was the most popular product category:

```
SELECT
    `Product_Category`,
    COUNT(*) AS `Frequency`
FROM
    sales_data
WHERE
    `Total_Amount` = (SELECT MAX(`Total_Amount`) FROM sales_data)
GROUP BY
    `Product_Category`
ORDER BY
    `Frequency` DESC
LIMIT 1;
```

Output:

Product_Category	Frequency
Electronics	19

Takeaway: From the output above, we can see that amongst the most expensive transactions, electronics was the product category with the highest frequency.

DISCUSSION

From the KPI questions analysis above, several key insights emerge (listed below) that can guide this "company's" strategies to maximize their total revenue while understanding their customer base and purchasing trends more holistically.

1. **Seasonal Trends and Product Preferences**
   Quarter 4 outperformed other quarters in total sales, which suggests heightened consumer spending during the holiday season. Electronics emerged as the top-performing product category in total sales, which aligns with trends where high-value items like electronics are purchased more frequently during the holiday season. This indicates a significant opportunity to focus marketing and promotional efforts on electronics in Q4, particularly in holiday campaigns and bundle offers, and an opportunity for strategic inventory management to ensure sufficient stock of popular items during peak demand. Additionally, clothing was the most frequently purchased category in Quarters 1 and 2, showing a seasonal preference that might be linked to spring wardrobe refreshes. Strategic discounts or promotions in Q1 and Q2 could bolster sales in these quarters to offset the natural lull compared to Q4. Lastly, addressing supply chain efficiency for seasonal restocking could ensure smooth operations and mitigate stockouts.
2. **Customer Demographics and Spending Patterns**
   Age and gender analysis revealed that younger demographics, particularly 18-27-year-old females, contribute the most to total sales. This group shows strong purchasing power, especially in categories like clothing and beauty. Complementing this finding, 18-27-year-old males rank second in total contribution, further underscoring the importance of targeting this age group. Interestingly, while older demographics (58+) contribute less overall, their higher spending on certain high-value items like electronics suggests a niche market that could be explored further through tailored advertising or loyalty programs.
3. **Gender Neutrality in Product Preferences**
   Statistical testing revealed no significant association between gender and product category preference, indicating that marketing campaigns can focus more on highlighting

product value and features rather than gender-specific messaging. This insight reinforces the opportunity to deploy broad, inclusive advertising strategies that emphasize quality and utility.

4. **Revenue Diversification Across Product Categories**
   Although electronics generated the highest revenue overall, the contributions of beauty and clothing were not far behind, showcasing an evenly distributed revenue stream across categories. This balance provides stability for the company, as it reduces dependency on any single product line. Beauty products, having the highest average price per unit, present a unique opportunity to increase profitability by encouraging bundle purchases or leveraging exclusivity to justify premium pricing.
5. **High-Value Transactions as a Revenue Driver**
   Analysis of the highest single transactions highlighted electronics as the dominant category, reaffirming its role as a revenue driver. Promoting high-ticket items through installment payment plans, extended warranties, or limited-time offers could further capitalize on this trend.

Recommendations

To capitalize on these insights and maximize revenue, the company could implement the following strategies:

- **Targeted Promotions**: Develop targeted marketing campaigns for the 18-27 age group, emphasizing beauty and clothing. Consider leveraging social media platforms to reach younger audiences effectively.
- **Seasonal Discounts and Bundling**: Offer seasonal discounts or bundle deals during key quarters, particularly for electronics in Q4 (for holidays) and clothing in Q2 (for spring wardrobe refreshes).
- **Loyalty Programs for Older Demographics**: Introduce loyalty programs or targeted advertising campaigns to engage older demographics, focusing on high-value electronics purchases.
- **Cross-Promotions Among Product Categories**: Promote cross-category purchases, such as offering discounts on beauty products with the purchase of electronics, to increase average transaction value.
- **Universal Campaigns**: Design gender-neutral marketing campaigns that highlight product quality and utility to appeal to a wide audience.

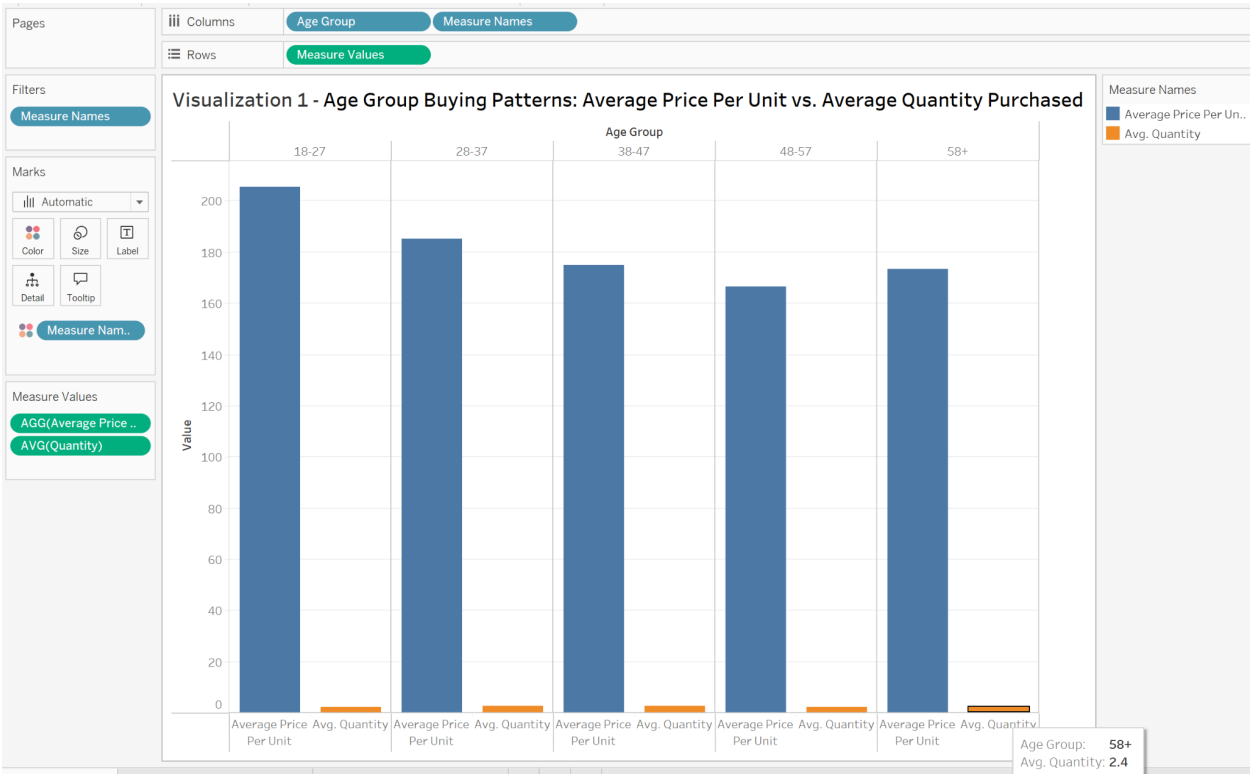
Tableau Visualizations

From our analysis, it is clear that investigating the following trends are important: The age group buying patterns (more specifically, their quantity purchased vs. price per unit) since as we saw previously, it could be the case that 58+ year olds buy less items but more expensive items, the revenue that each product category brings in during each fiscal quarter, and the total spending of

each age/gender combination in each product category. To present these trends effectively in a way that both technical and non-technical stakeholders could understand, I created the following interactive Tableau visualizations:

1. Age group buying patterns - quantity purchased vs. price per unit:

   For this trend, I created visualizations with bar charts for each age group; One set of bar charts representing the average quantity purchased for each age group and the other set representing the average price per unit of items bought for each age group. This way, stakeholders can easily compare the general buying pattern of each age group - whether they buy fewer but more expensive items, more but cheaper items, fewer and cheaper items, etc. Hovering over the graph allows the user to see exact values.



   Takeaway: From this visualization, it can be seen that on average, all age groups bought relatively the same amount of items. However, the average price per unit of the items bought by each age group (i.e. the overall "value" of the products bought), seem to differ a bit between the age groups. First, we can see that on average, it seems as though 18-27 year olds buy the most expensive products and following them, 58+ year olds (which seems to be aligned with our original belief that they buy fewer items (their average quantity was the lowest) but more expensive items.

Link to visualization on Tableau Public

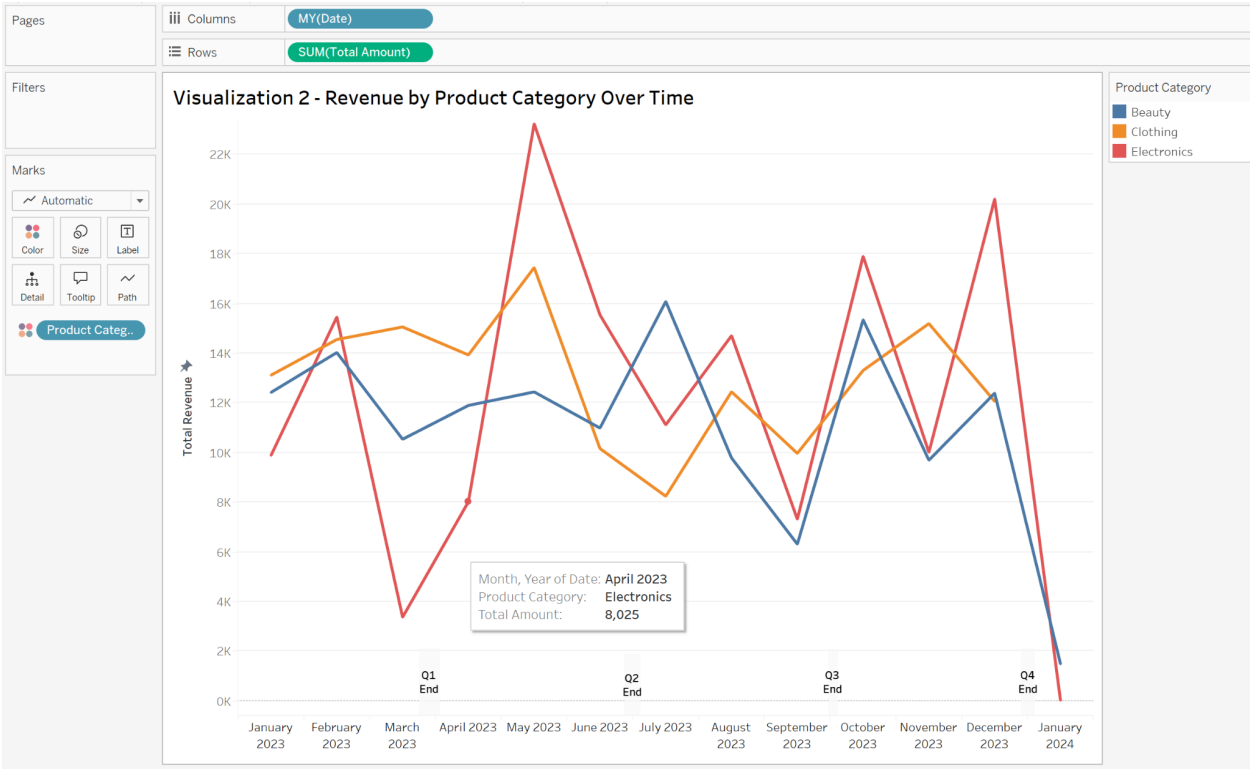
(please go to the visualization 1 tab)

2. The revenue that each product category brings in during each fiscal quarter:

   For this trend, I created a graph with color-coded line plots for each product category with labels for each month/year and the end of the quarters on the x axis and the total revenue that each product category brings in on the y axis so that stakeholders can easily see the trend for the total revenue generated from each product category over time. Hovering over the graph allows the user to see exact values.

   Entered the following SQL query to create “Fiscal Quarter” as a calculated field:

   IF MONTH([Date]) IN (1, 2, 3) THEN "Q1" ELSEIF MONTH([Date]) IN (4, 5, 6) THEN "Q2" ELSEIF MONTH([Date]) IN (7, 8, 9) THEN "Q3" ELSE "Q4" END

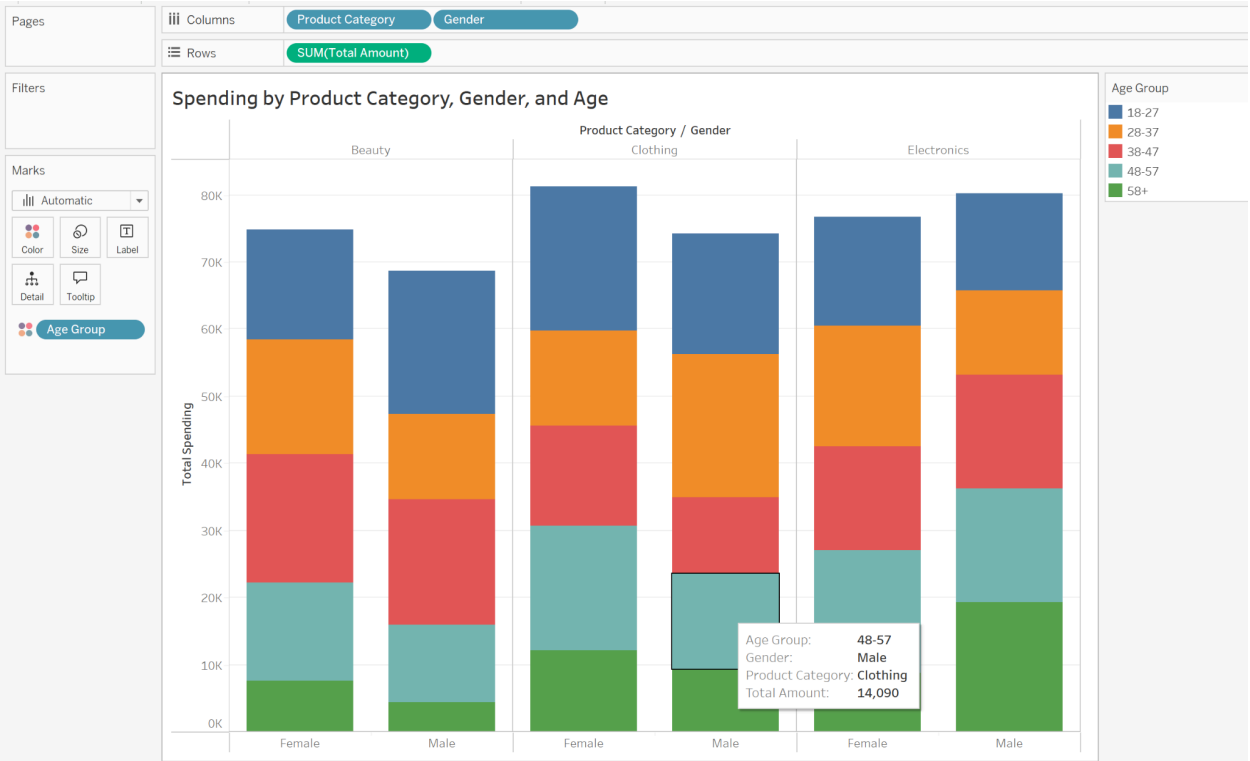


Link to visualization on Tableau Public

(please go to the visualization 2 tab)

3. The total spending of each age/gender combination in each product category:

   For this trend, I created a graph with a stacked bar for each age/gender combination grouped by the labels for each product category with the total spending on the y-axis. This way, the relationships between spending, the demographic categories, and the product categories can easily be compared. Hovering over the graph allows the user to see exact values.

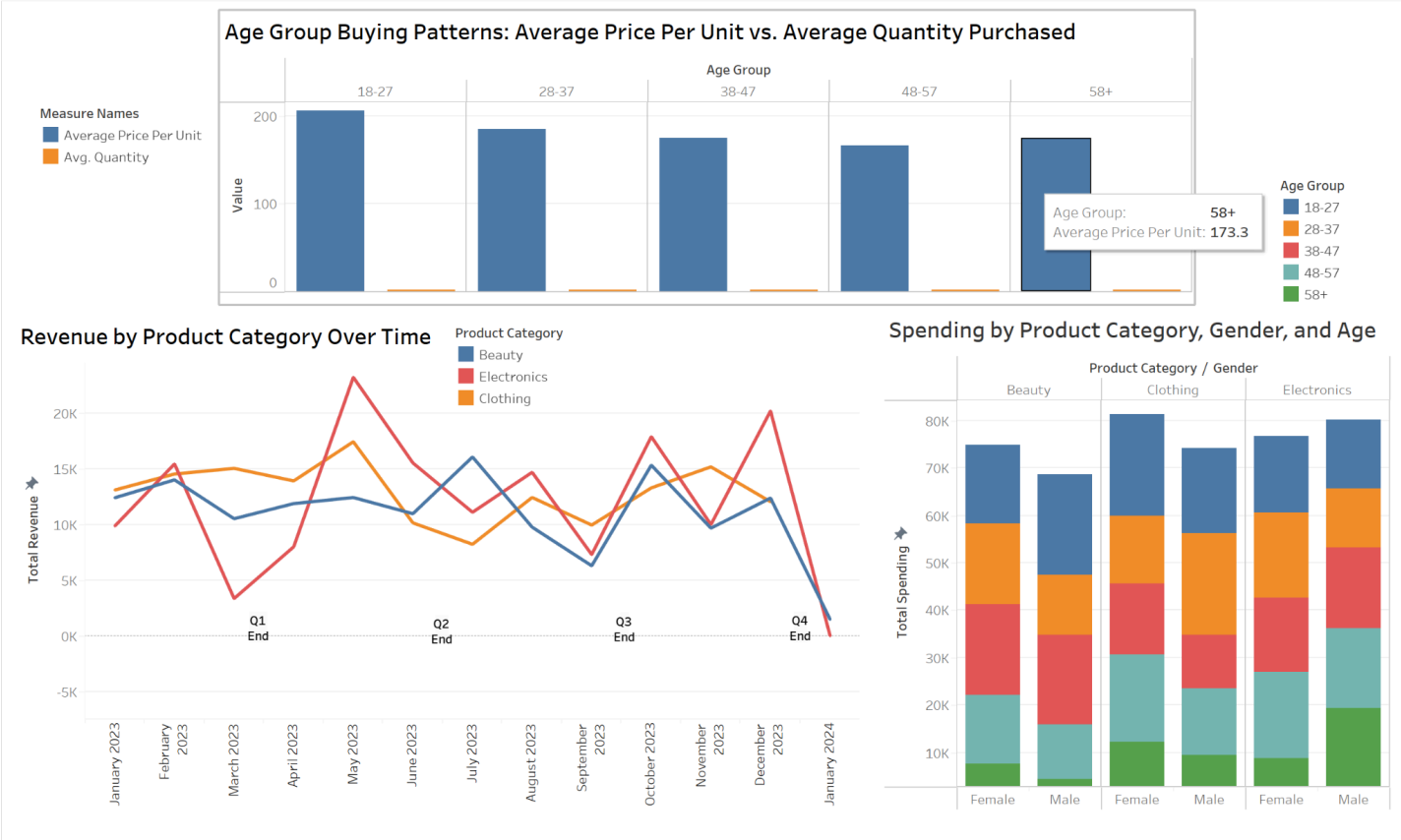


   [Link to Visualization on Tableau Public]

   (please go to the visualization 3 tab)

4. Final Dashboard:

   I combined all the visualizations into one dashboard so all the trends can be viewed and compared simultaneously by stakeholders. Hovering over the dashboard allows the user to see exact values.



[Link to Dashboard on Tableau Public]

(please go to the "final dashboard" tab)

Justification and Impact

The recommendations provided aim to maximize the company's revenue by leveraging insights into customer behavior, product performance, and seasonal trends. For example, prioritizing high-demand products in peak quarters and targeting the most profitable demographic groups can yield immediate returns. Additionally, addressing underperforming demographics and product categories ensures that potential revenue streams are not neglected. The Tableau visualizations will serve as a user-friendly tool for stakeholders, enabling them to make data-driven decisions.

This approach not only supports the company's financial goals but also enhances its ability to adapt to changing market conditions, ensuring sustained growth and customer satisfaction.

LIMITATIONS

While this project demonstrates actionable insights into retail KPIs and customer behavior, certain limitations could be addressed to enhance its real-world applicability. The dataset, being synthetic, simplifies some complexities of real-world data, such as regional variations or occasional anomalies, and excludes fields like costs and net profits, which would provide a fuller picture of business performance. Demographic analysis, though insightful, is limited to available variables and could benefit from incorporating additional factors like location, income, or customer loyalty. Assumptions about linear purchasing behavior and a focus on three product categories offer a strong starting point but may not fully capture the diversity and external influences seen in real-world business environments. Nonetheless, these limitations provide clear opportunities for refining and expanding the project to align more closely with practical, real-world scenarios.

FUTURE WORK

Future iterations of this project could enhance realism and utility by incorporating predictive modeling to forecast trends and customer behaviors. Adding variables like income, location, and loyalty data would improve demographic insights and marketing recommendations. Automating analyses, such as A/B testing and KPI calculations, would streamline reporting and enable real-time decision-making. Expanding the dataset to include more product categories and factors such as net profits could provide broader insights into sales and cross-category strategies. Finally, analyzing external factors like holidays, economic conditions, or market trends would create a more dynamic and realistic framework for business decision-making.

APPENDIX

- Original Dataset
- Google Sheets (for KPI question analysis)
- Python Notebook
- Modified Dataset (for Tableau/Power BI visualizations)
- Tableau Public Link (for all Tableau visualizations)