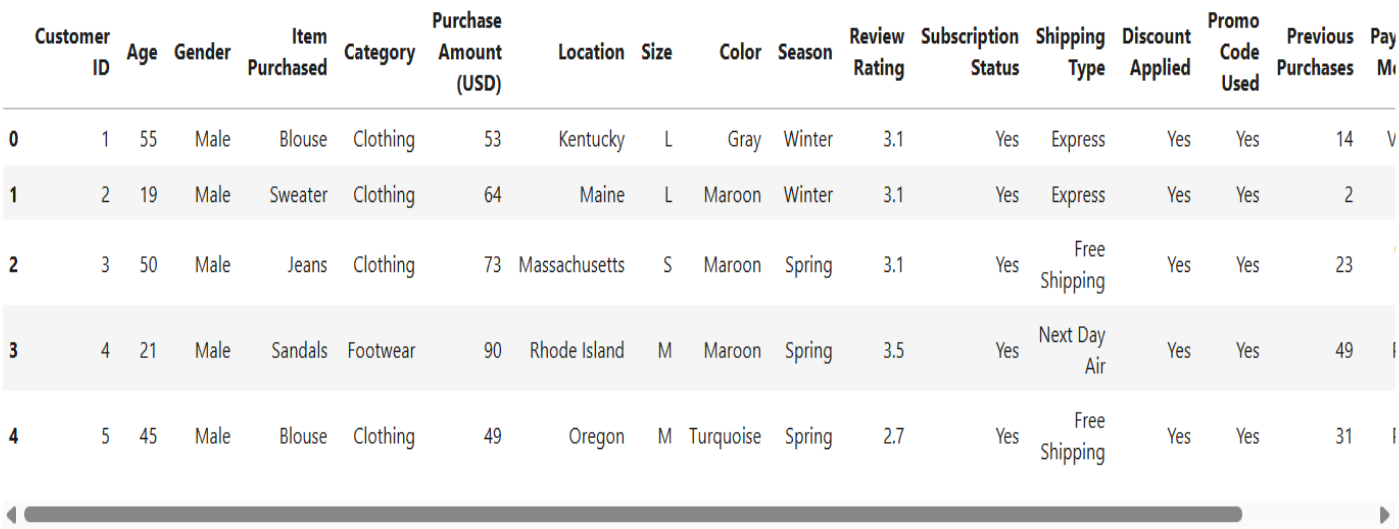
Customer Shopping Behavior Analysis

Objective:

To perform a comprehensive analysis of customer shopping behavior by leveraging **data cleaning, SQL-based querying, and interactive visualization.** The goal is to uncover actionable insights that inform **business strategies, demonstrating proficiency** in end-to-end data analytics processes.

Dataset Summary:

- **Rows:** 3900.
- **Columns:** 18.
- **Key Features:**
  - Customer Demographics (Age, Gender, Location).
  - Purchase Details (Item Purchased, Category, Purchased Amount, Seasons, Size, Color).
  - Shopping Behaviour (Discounts Applied, Promo Code Used, Previous Purchase, Frequency Of Purchase, Review Rating, Shipping Type).
- **Missing Data**
  - The Review Rating Column contained 37 missing values.

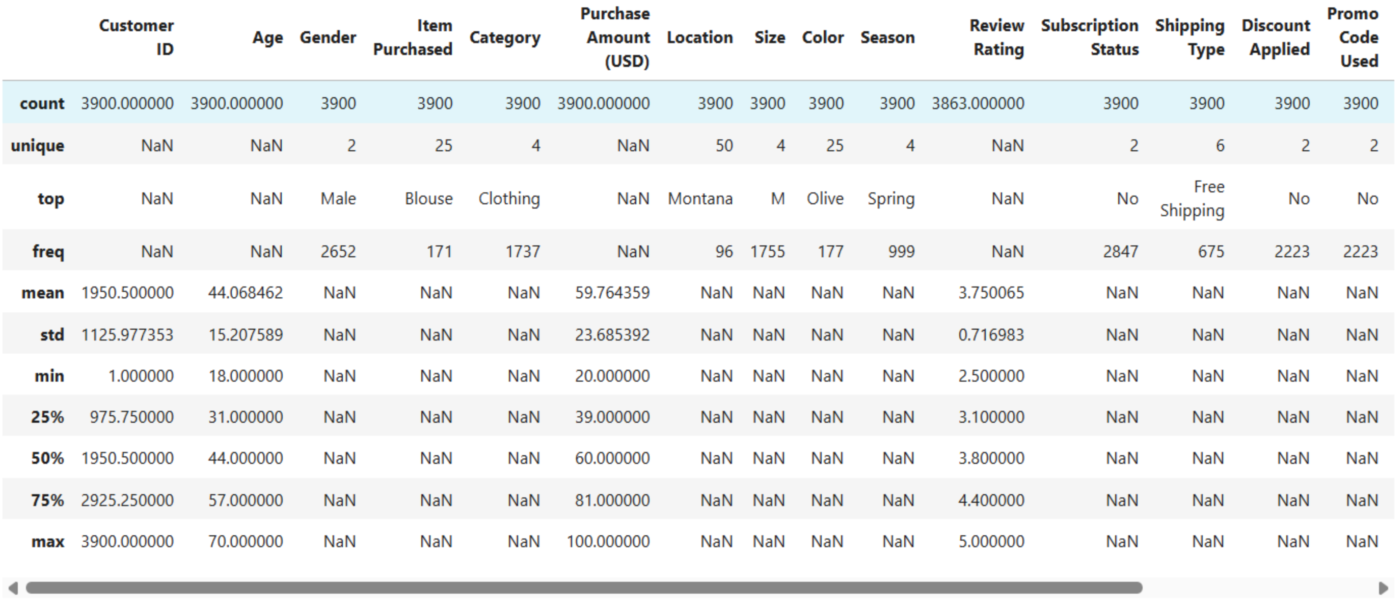


	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied	Promo Code Used	Previous Purchases	Pay M
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express	Yes	Yes	14	V
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Express	Yes	Yes	2	
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Shipping	Yes	Yes	23	
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Day Air	Yes	Yes	49	I
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Free Shipping	Yes	Yes	31	I

Exploratory Data analysis with Python:

We began with data preparation and cleaning with python.

- **Data Loading:**
  - loaded the data set using Pandas.
- **Initial Exploration:**
  - Checked the structure and summary of the data set.



	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied	Promo Code Used
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	3900	3900
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	2	2
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN	No	Free Shipping	No	No
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN	2847	675	2223	2223
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750065	NaN	NaN	NaN	NaN
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	NaN	NaN
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	NaN	NaN
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	NaN	NaN
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	NaN	NaN
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	NaN	NaN
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	NaN	NaN

- **Missing Data:**
  - Checked for missing data, the data set had 37 missing values in the Review Rating column.
  - Then imputed the missing values with the median rating of each Product Category.
- **Data Consistency Check:**
  - Verified if Discount Applied Columns and Promo Code Used were redundant
  - Dropped Promo Code Used.
- **Column Standardization:**
  - Renamed column names to Snake Case for better readability and documentation.
- **Feature Engineering:**
  - Created age_grouped column by binning customer age
  - Created mapped_frequency_of_purchase column from purchase data
- **Database Integration:** Connected Python script to MySQL and loaded the cleaned DataFrame into the database for SQL analysis.

Executive Summary:

- **Revenue** is primarily driven by **Male customers** and **Young-Adult age groups**.
- The customer base shows **strong loyalty**, but **subscription adoption remains low**, even among **high-frequency buyers**, indicating **limited perceived value**.
- **Clothing** and **Accessories** are the top **revenue-generating categories**.
- Several **high-revenue products** receive **low customer ratings**, creating potential **retention and brand risk**.
- **Shipping experience** has a meaningful impact on **customer satisfaction**.
- **Seasonality** significantly affects **purchasing behavior** across **age groups** and **locations**.

- **Discounts** do not reduce **average order value**, suggesting they can be used **strategically** without harming **revenue**.

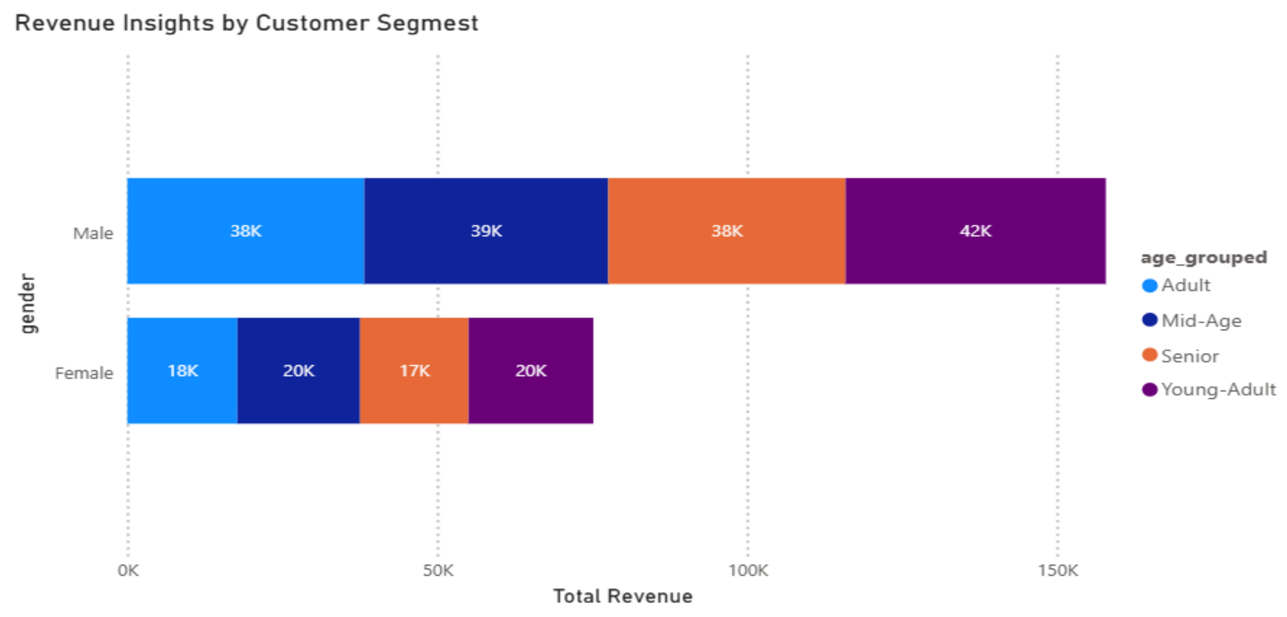
Key KPIs:

- **Total Revenue:** $233,081
- **Average Order Value:** $59.76
- **Male Revenue Share:** 67.7%
- **Loyal Customers:** 80.5%
- **Subscription Revenue Share:** 26.9%
- **Top Category:** Clothing (44.7%)

Customer Demographics & Value

- **Gender:**
  Male customers generate **$157,890 (67.7%)** of total revenue, while female customers contribute **$75,191 (32.3%)**, indicating a strong male-driven revenue skew.
- **Age Groups:**
  Young-Adults lead revenue contribution (**$62,143**), followed by **Mid-Age ($59,197)**, **Adults ($55,978)**, and **Seniors ($55,763)**. Revenue is relatively balanced across age groups, with a slight advantage among younger customers.
- **Subscriptions:**
  Non-subscribers contribute 73.1% of total revenue, while subscribers account for **26.9%**, despite both groups having nearly identical average order values (~**$59.7**).
- **Loyalty:**
  Loyal customers represent 80.5% of the customer base, confirming strong customer retention.
- **Repeat Buyers:**
  Among high-frequency buyers, **72.4% are non-subscribers**, highlighting a significant gap between repeat purchasing behavior and subscription adoption.

Insight:
Revenue is primarily driven by male and young-adult customers. While loyalty levels are high (80.5%), subscriptions fail to capture high-frequency buyers, limiting their incremental business value.

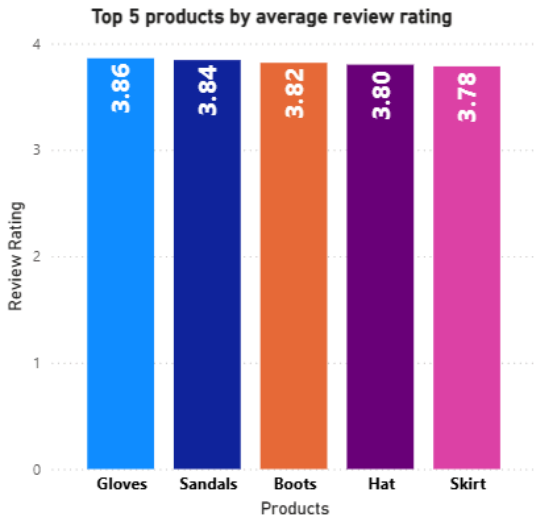


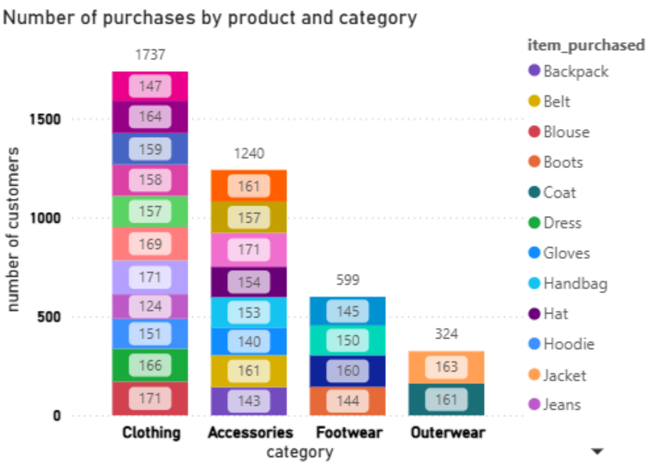
Product & Category Performance:

- **Top Categories:**
  Clothing ($104,264) is the highest revenue-generating category, followed by **Accessories ($74,200)**, **Footwear ($36,093)**, and **Outerwear ($18,524)**.
- **Top Rated Products:**
  Gloves, Sandals, Boots, Hats, and Skirts maintain high average review ratings (approximately **3.8**) when evaluated across all transactions, indicating positive overall customer sentiment.
- **Repeat Purchases:**
  Dresses (98.8%), **Scarves (98.7%)**, and **Boots (98.6%)** show the highest repeat purchase rates, reflecting strong product-market fit.
- **Risk Products:**
  Coats and Skirts exhibit high average purchase values (**$97 and $81 respectively**) but significantly lower review ratings (**2.6–2.8**) when analyzed within high-spend transactions only, indicating declining satisfaction as price expectations increase.
- **Category Leaders:**
  - Clothing: **Blouses, Pants, Shirts**
  - Accessories: **Jewelry, Sunglasses**
  - Footwear: **Sandals**
  - Outerwear: **Jackets**

Insight:
While Clothing and Accessories drive the majority of revenue, certain high-spend products (Coats

and Skirts) show a price–expectation mismatch, posing quality or value perception risks. High-rated and repeat-purchase products offer strong opportunities for cross-selling and bundling.

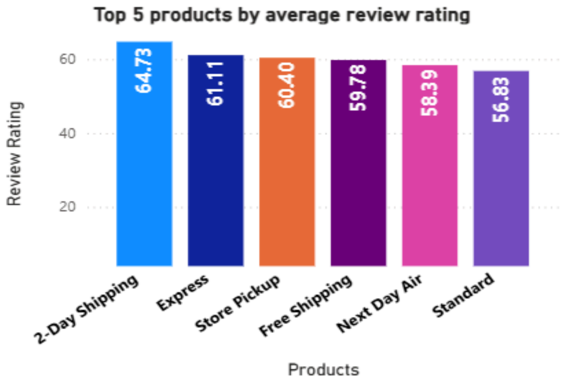


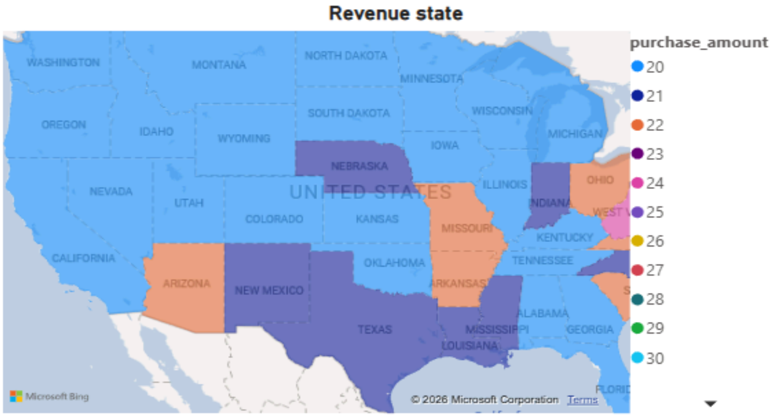


Location & Shipping Insights:

- **Top Revenue States:**
  Montana, Illinois, California, and Idaho each generate approximately **$5.6k** in revenue, indicating relatively even geographic performance among top states.
- **Highest AOV:**
  Alaska records the highest average order value at $67.60, suggesting fewer but higher-value purchases.
- **Shipping Preferences:**
  Shipping preferences vary by location, with **Store Pickup** most common in Illinois and **Free Shipping** favored in Montana.
- **Best-Rated Shipping:**
  Standard Shipping achieves the highest average customer rating (**3.82**), outperforming faster or alternative fulfillment methods.

Insight:
Customer satisfaction is driven more by shipping reliability than speed. Regional shipping preferences present opportunities for location-based fulfillment optimization.

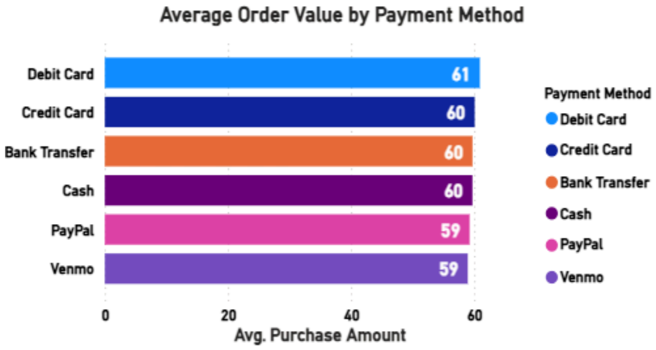


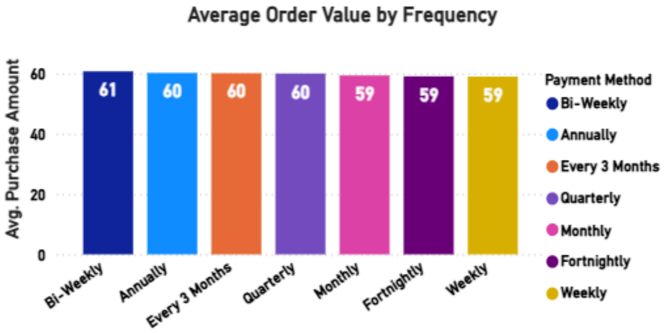


Discounts, Payments & Behavior:

- **Discount Impact:**
  Discounted orders exceed the overall average order value (**$59.76**), demonstrating that discounts increase basket size rather than reduce revenue.
- **Discount Hotspots:**
  Highest discount usage occurs in **Rhode Island (39.7%)**, **Vermont (38.8%)**, and **Virginia (37.7%)**, indicating regional price sensitivity.
- **Payments:**
  Debit card transactions show the highest average order value ($60.92), slightly outperforming other payment methods.
- **Purchase Cadence:**
  Customers purchasing on a **quarterly basis** exhibit the highest discount usage (**44.2%**), making them the most promotion-responsive segment.

Insight:
Discounts are most effective for mid-frequency shoppers and do not negatively impact spending. Payment behavior suggests debit card users are higher-value customers.

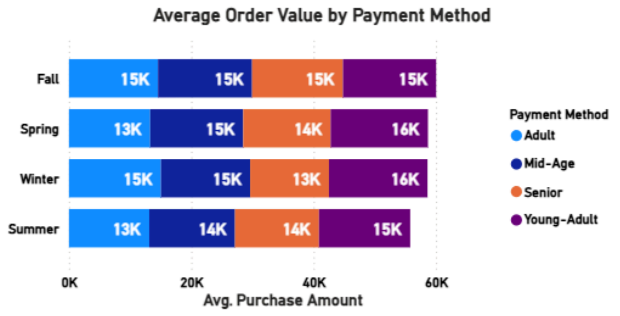


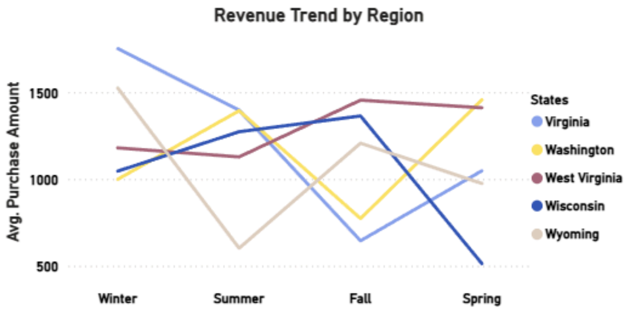


Seasonality & Engagement:

- **Regional Peaks:**
  Revenue peaks in **Fall for Colorado** and **Spring for Ohio**, reflecting region-specific seasonal demand patterns.
- **Age-Based Peaks:**
  - **Winter:** Young-Adults and Adults generate the highest seasonal revenue (each exceeding **$15k**)
  - **Fall:** Mid-Age and Seniors dominate seasonal revenue (approximately **$15k** each)

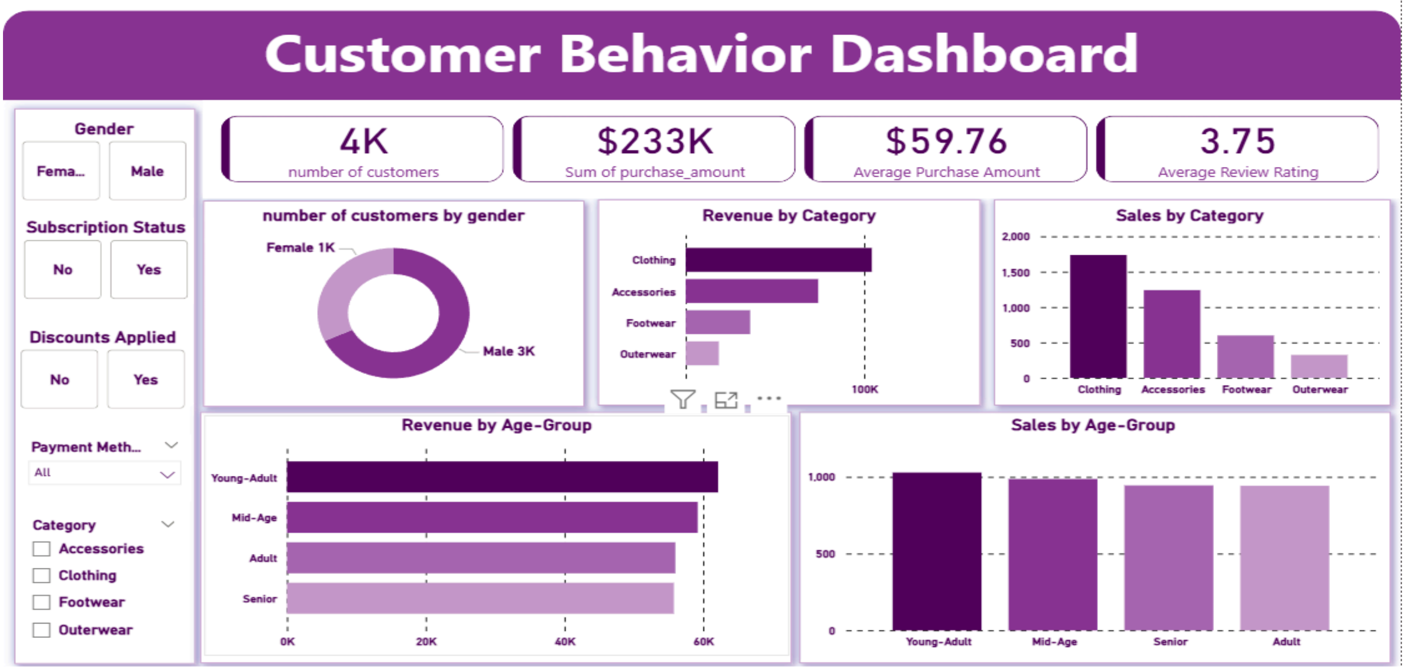
Insight:
Seasonal engagement varies by both age and geography. Aligning campaigns and inventory with these patterns can significantly improve seasonal performance and demand forecasting.





Dashboard in Power BI:

Built an Interactive Dashboard in Power BI to present insights.



Conclusion:

- Redesign subscription offerings to better align with the needs of high-frequency customers and increase adoption.
- Improve quality and value perception for high-revenue but low-rated products to reduce churn risk.
- Prioritize Clothing and Accessories while leveraging high-repeat products for cross-selling and bundling strategies.
- Promote Standard Shipping as the default option to enhance customer satisfaction.
- Target discounts toward mid-frequency shoppers to maximize incremental revenue.
- Align marketing campaigns and inventory planning with age-based and regional seasonal demand patterns.

Recommendations:

- Convert high-frequency non-subscribers with value-led subscription incentives.
- Improve quality of high-revenue, low-rated products.
- Target discounts strategically at mid-frequency shoppers.
- Promote Standard Shipping to enhance satisfaction.
- Align marketing and inventory with seasonal demand and age groups.

The Path Forward:

- **Strengthen loyalty programs** to retain the core 80% of revenue-driving customers.
- **Apply discounts strategically** on high-response products to boost sales without eroding margins.
- **Promote top-rated items** to enhance customer satisfaction and increase conversion rates.
- **Increase subscription adoption** by targeting repeat buyers with tailored value propositions.
- **Leverage convenience-based upsells** such as express shipping to enhance the customer experience and drive additional revenue.