Customer Retention Risk & Revenue Stability Analysis

1. Project Overview

This project analyzes customer behavioral and transactional data to identify customers at risk of churn and quantify the potential impact on long-term revenue stability. The objective is to proactively detect high-value customers exhibiting declining retention behavior and determine product categories contributing significantly to churn-risk revenue.

2. Dataset Summary

-Rows: 3,900

- Columns: 18

  - Key Features: - Customer demographics (Age, Gender, Location, Subscription Status)
  - Purchase details (Item Purchased, Category, Purchase Amount, Season, Size, Color)
  - Shopping behavior (Discount Applied, Promo Code Used, Previous Purchases, Frequency of Purchases, Review Rating, Shipping Type)

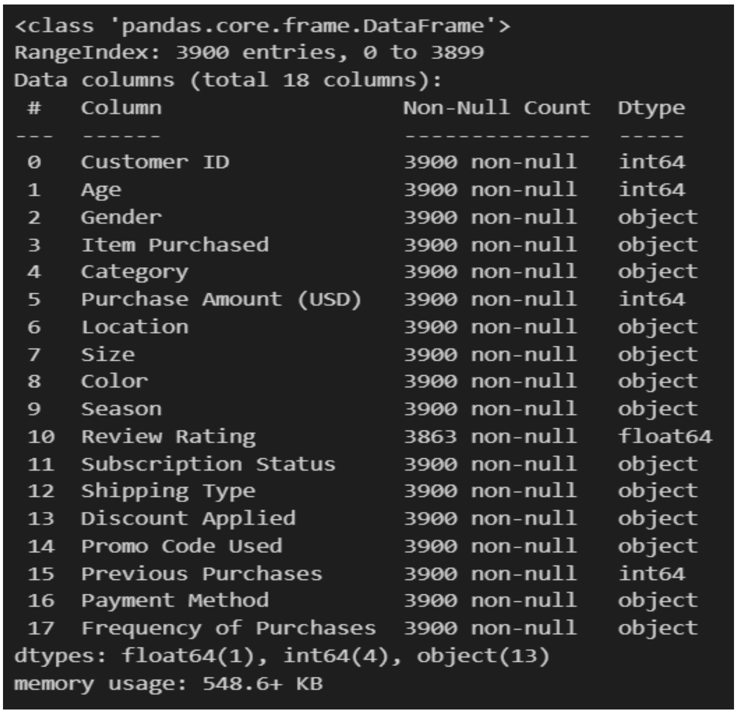
- Missing Data: 37 values in Review Rating column

3. Exploratory Data Analysis using Python

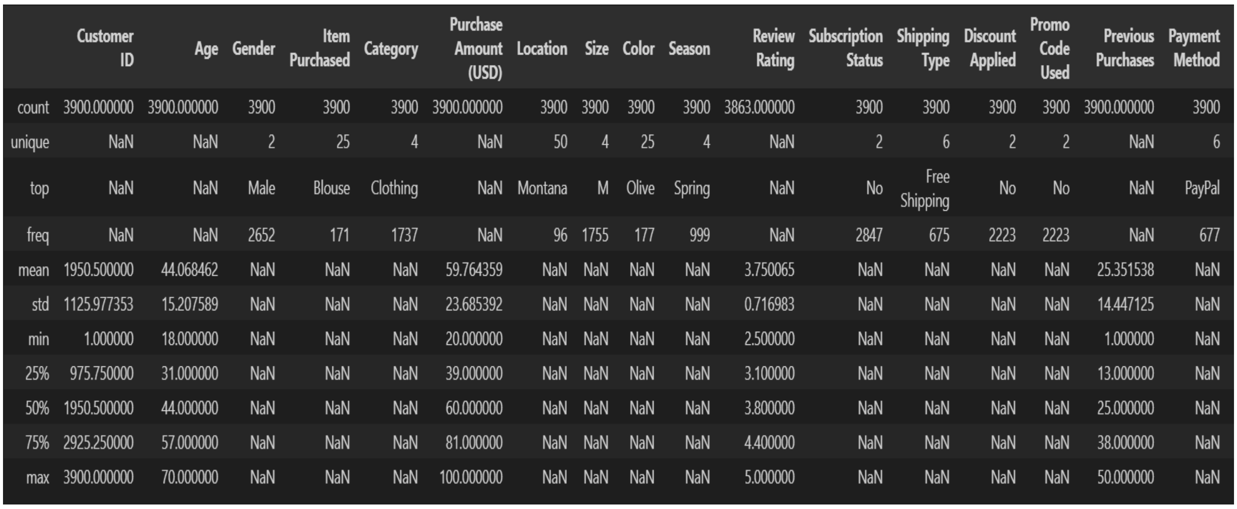
We began with data preparation and cleaning in Python:

● **Data Loading:** Imported the dataset using pandas.

● **Initial Exploration:** Used df.info() to check structure and .describe() for summary statistics.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 18 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Customer ID             3900 non-null   int64
 1   Age                     3900 non-null   int64
 2   Gender                  3900 non-null   object
 3   Item Purchased          3900 non-null   object
 4   Category                3900 non-null   object
 5   Purchase Amount (USD)   3900 non-null   int64
 6   Location                3900 non-null   object
 7   Size                    3900 non-null   object
 8   Color                   3900 non-null   object
 9   Season                  3900 non-null   object
 10  Review Rating           3863 non-null   float64
 11  Subscription Status     3900 non-null   object
 12  Shipping Type           3900 non-null   object
 13  Discount Applied        3900 non-null   object
 14  Promo Code Used         3900 non-null   object
 15  Previous Purchases      3900 non-null   int64
 16  Payment Method          3900 non-null   object
 17  Frequency of Purchases  3900 non-null   object
dtypes: float64(1), int64(4), object(13)
memory usage: 548.6+ KB
```



	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	Payment Method
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	3900	3900	3900.000000	3900
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	2	2	NaN	6
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN	No	Free Shipping	No	No	NaN	PayPal
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN	2847	675	2223	2223	NaN	677
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750065	NaN	NaN	NaN	NaN	25.351538	NaN
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	NaN	NaN	14.447125	NaN
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	NaN	NaN	1.000000	NaN
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	NaN	NaN	13.000000	NaN
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	NaN	NaN	25.000000	NaN
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	NaN	NaN	38.000000	NaN
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	NaN	NaN	50.000000	NaN

● **Missing Data Handling:** Checked for null values and imputed missing values in the Review Rating column using the median rating of each product category**.**

● **Column Standardization:** Renamed columns to snake case for better readability and documentation.

● **Feature Engineering:**

- Created **age_group** by segmenting customers into demographic cohorts.
- Mapped purchase frequency into **purchase_frequency_days** for retention analysis.
- Developed a **retention_score** to measure customer engagement level.
- Segmented customers into **Highly Retained, Moderate Risk,** and **High Churn Risk groups**.
- Flagged Moderate and High Churn Risk customers as **Revenue At Risk.**
- Estimated **Customer Lifetime Value (CLV)** to identify high-value customers.

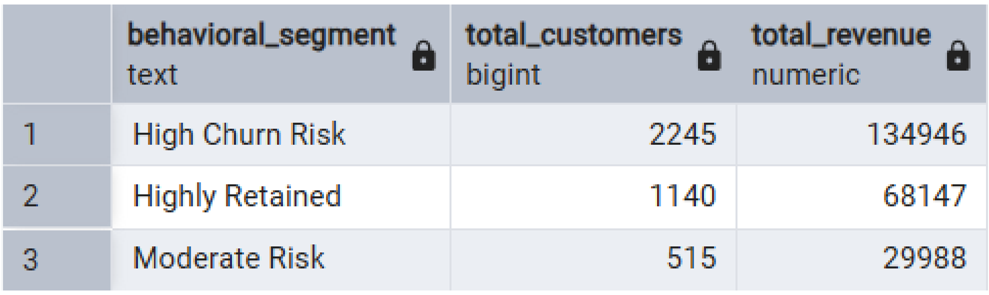
● **Database Integration:** Connected Python script to PostgreSQL and loaded the cleaned DataFrame into the database for SQL analysis.

4. Data Analysis using SQL (Business Transactions)

We performed structured analysis in PostgreSQL to answer key business questions:

1. Revenue Distribution Across Behavioral Retention Segments

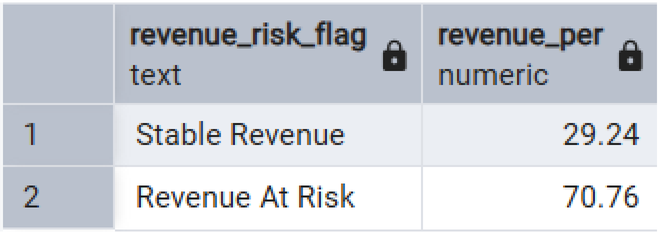
Analyzed how total revenue is distributed among customer engagement groups (Highly Retained, Moderate Risk, High Churn Risk) to evaluate revenue stability.



	behavioral_segment text	total_customers bigint	total_revenue numeric
1	High Churn Risk	2245	134946
2	Highly Retained	1140	68147
3	Moderate Risk	515	29988

2. Percentage of Revenue Dependent on Churn-Risk Customers

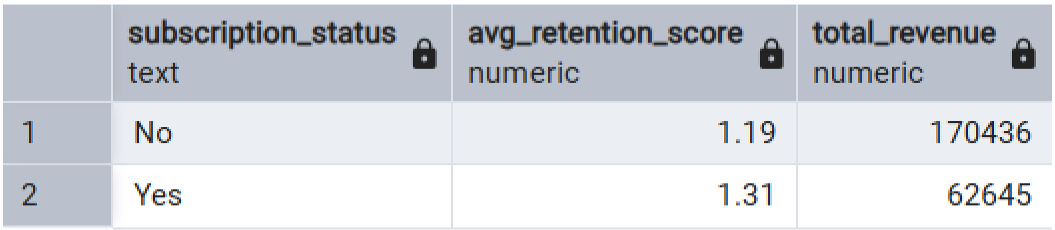
Calculated the proportion of total revenue contributed by customers identified as Moderate Risk and High Churn Risk to estimate potential revenue vulnerability.



	revenue_risk_flag text	revenue_per numeric
1	Stable Revenue	29.24
2	Revenue At Risk	70.76

3. Retention Comparison: Subscribers vs Non-Subscribers

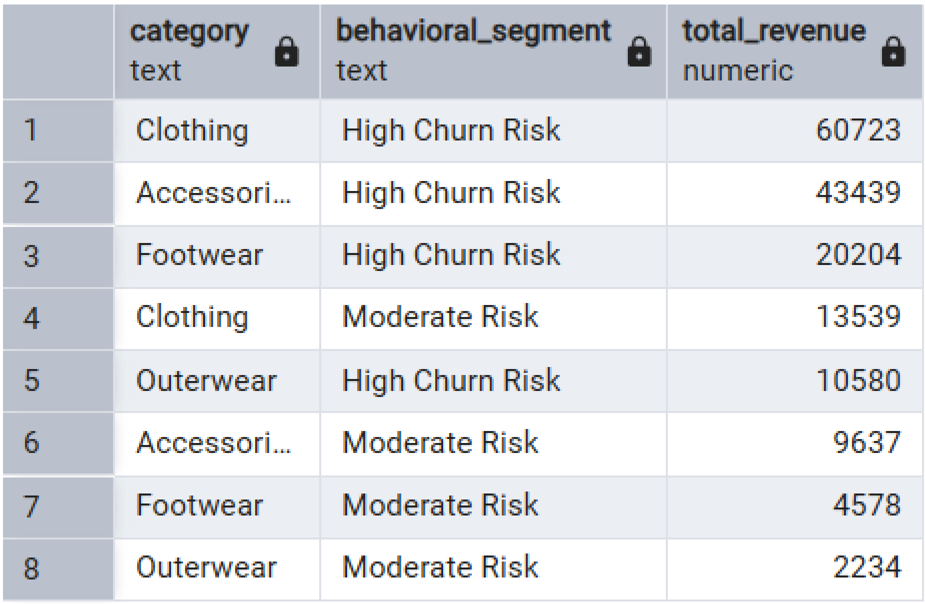
Compared the average retention score between subscribed and non-subscribed customers to assess the impact of subscription-based engagement on customer loyalty.



	subscription_status text	avg_retention_score numeric	total_revenue numeric
1	No	1.19	170436
2	Yes	1.31	62645

4. Product Categories Contributing to Churn-Risk Revenue

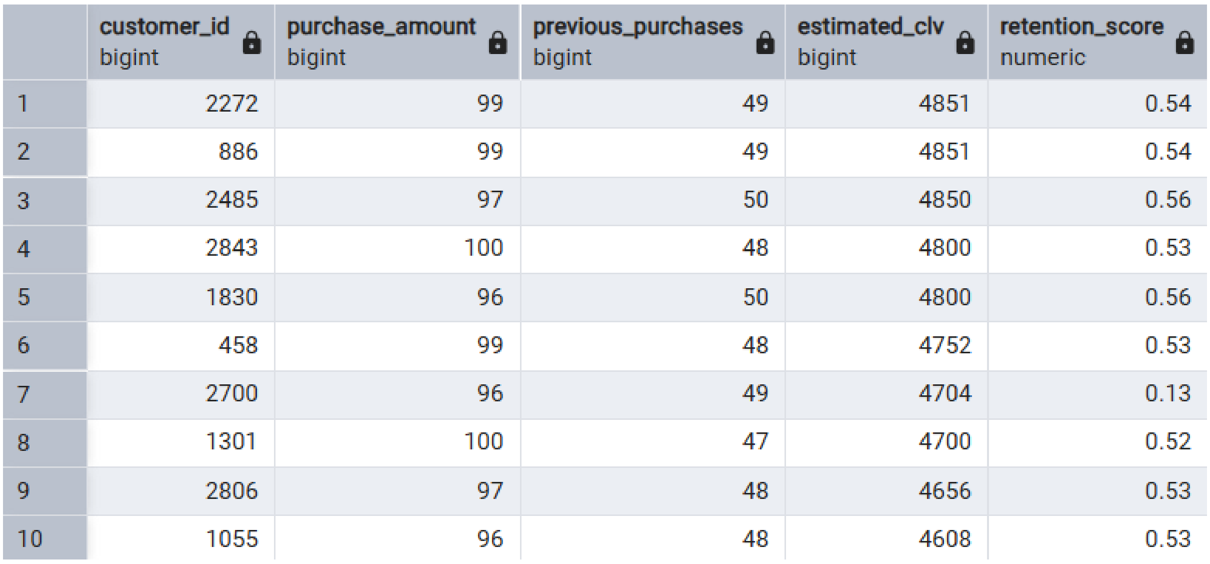
Identified product categories generating the highest revenue from churn-risk customers to detect segments requiring immediate retention interventions.



	category text	behavioral_segment text	total_revenue numeric
1	Clothing	High Churn Risk	60723
2	Accessori...	High Churn Risk	43439
3	Footwear	High Churn Risk	20204
4	Clothing	Moderate Risk	13539
5	Outerwear	High Churn Risk	10580
6	Accessori...	Moderate Risk	9637
7	Footwear	Moderate Risk	4578
8	Outerwear	Moderate Risk	2234

5. Top 10 High-Value Customers at High Churn Risk

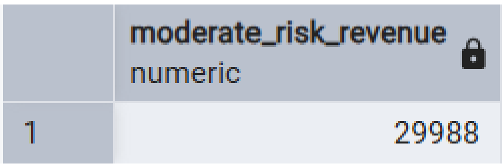
Detected high lifetime-value customers currently exhibiting low retention behavior to prioritize targeted retention campaigns.



	customer_id bigint	purchase_amount bigint	previous_purchases bigint	estimated_clv bigint	retention_score numeric
1	2272	99	49	4851	0.54
2	886	99	49	4851	0.54
3	2485	97	50	4850	0.56
4	2843	100	48	4800	0.53
5	1830	96	50	4800	0.56
6	458	99	48	4752	0.53
7	2700	96	49	4704	0.13
8	1301	100	47	4700	0.52
9	2806	97	48	4656	0.53
10	1055	96	48	4608	0.53

6. Revenue Generated by Moderately Engaged Customers

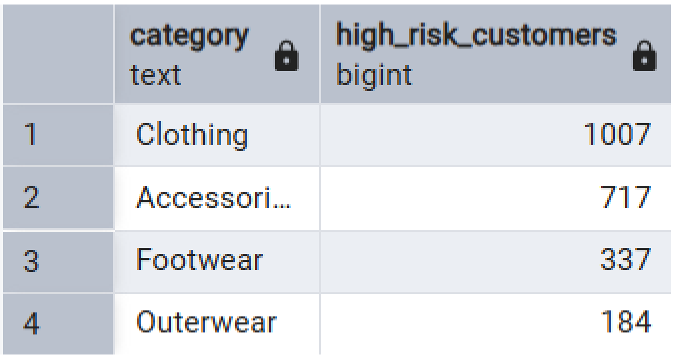
Measured total revenue contributed by moderately engaged customers representing a preventable churn-risk pool.



	moderate_risk_revenue numeric
1	29988

7. Product Categories with Highest High-Risk Customer Concentration

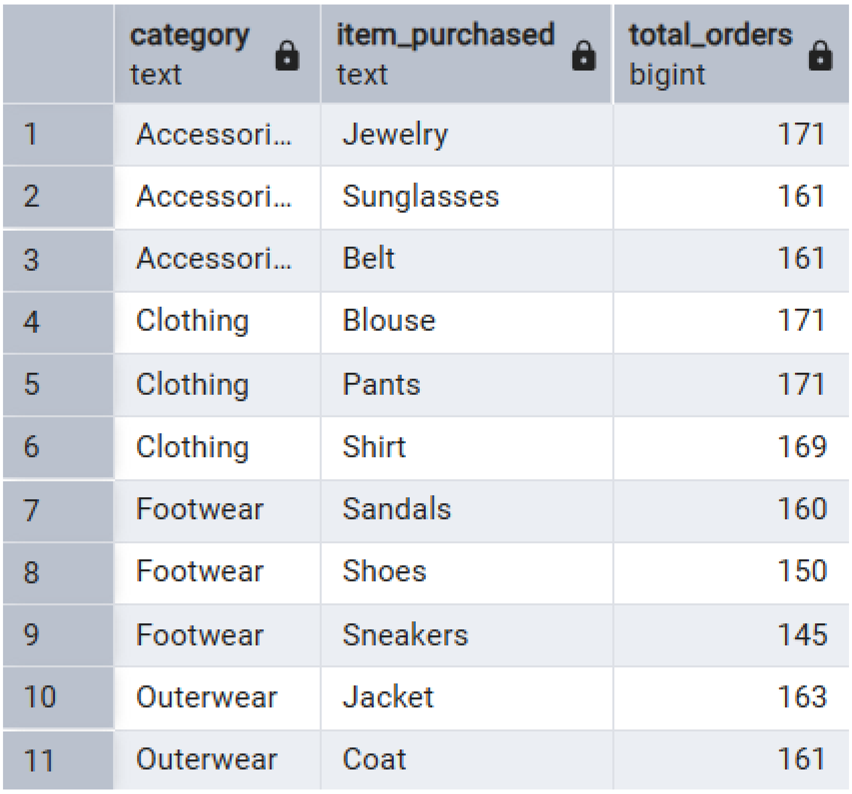
Identified product categories with the highest number of high churn-risk customers to determine areas of revenue exposure.



	category text	high_risk_customers bigint
1	Clothing	1007
2	Accessori...	717
3	Footwear	337
4	Outerwear	184

8. Top 3 Most Purchased Products within Each Category

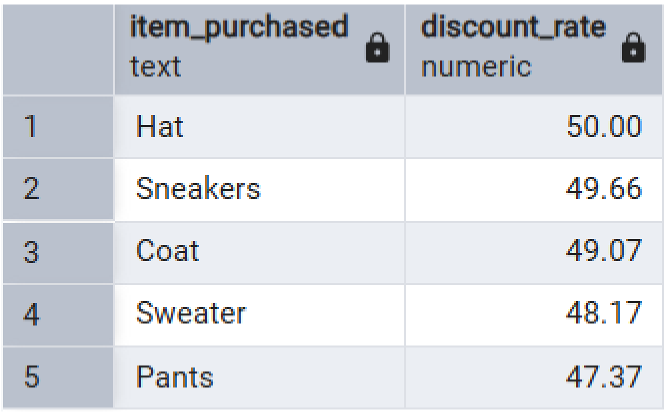
Ranked the most frequently purchased products within each category to support category-level retention strategies.



	category text	item_purchased text	total_orders bigint
1	Accessori...	Jewelry	171
2	Accessori...	Sunglasses	161
3	Accessori...	Belt	161
4	Clothing	Blouse	171
5	Clothing	Pants	171
6	Clothing	Shirt	169
7	Footwear	Sandals	160
8	Footwear	Shoes	150
9	Footwear	Sneakers	145
10	Outerwear	Jacket	163
11	Outerwear	Coat	161

9. Products with Highest Discount Dependency

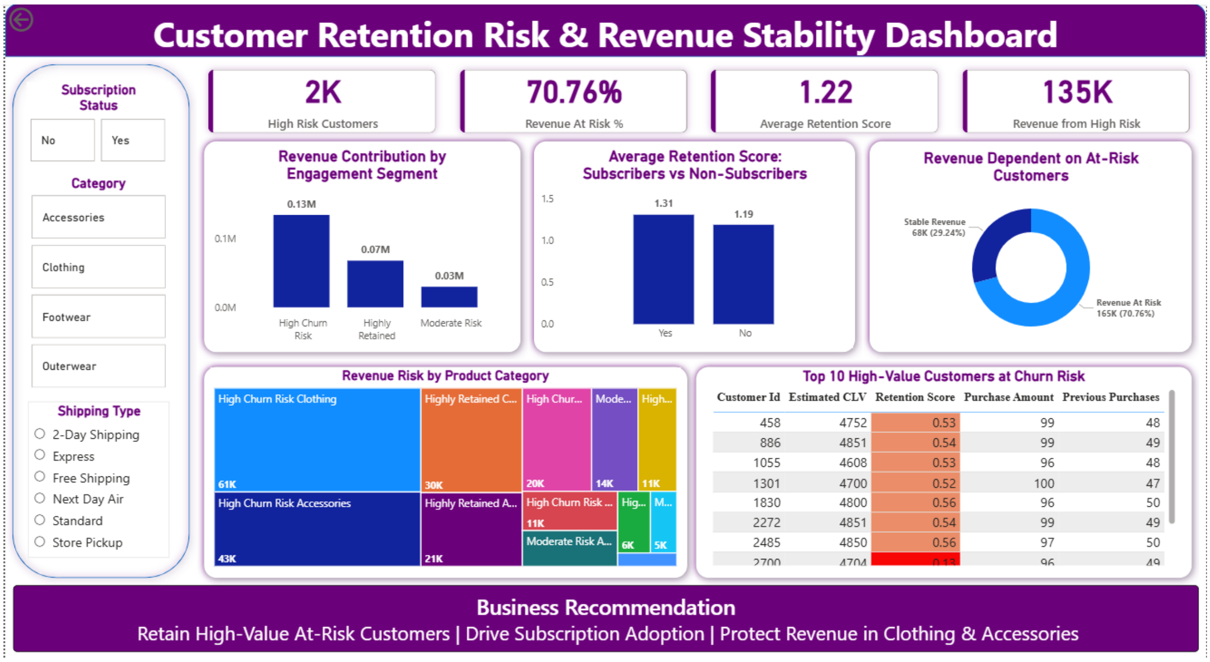
Evaluated which products have the highest percentage of discounted purchases to assess promotional reliance.



	item_purchased text	discount_rate numeric
1	Hat	50.00
2	Sneakers	49.66
3	Coat	49.07
4	Sweater	48.17
5	Pants	47.37

5. Dashboard in Power BI

Finally, we built an interactive dashboard in Power BI to present insights visually.



Customer Retention Risk & Revenue Stability Dashboard

2K High Risk Customers

70.76% Revenue At Risk %

1.22 Average Retention Score

135K Revenue from High Risk

Top 10 High-Value Customers at Churn Risk

Customer Id	Estimated CLV	Retention Score	Purchase Amount	Previous Purchases
458	4752	0.53	99	48
886	4851	0.54	99	49
1055	4608	0.53	96	48
1301	4700	0.52	100	47
1830	4800	0.56	96	50
2272	4851	0.54	99	49
2485	4850	0.56	97	50
2700	4704	0.13	96	49

Business Recommendation

Retain High-Value At-Risk Customers | Drive Subscription Adoption | Protect Revenue in Clothing & Accessories

6. Business Recommendations

- Retain high-value customers identified at churn risk to prevent revenue loss.
- Promote subscription adoption to improve customer retention.
- Focus retention efforts on Clothing & Accessories categories with high churn-risk revenue.
- Engage moderately retained customers to reduce future churn risk.