DATA ANALYTICS CAPSTONE PROJECT

AUSTINE MANDELA

February 2025

More analysis on the python script file

01. Data Cleaning and Preparation

Overview

- Data QualityAssessment: Inspecting the dataset for missing values, duplicates, or inconsistent data types
- Feature Engineering: Create the following columns "Month-Year" from DATE column

Dataset info.

Entries

→ 329880 rows

Total Columns

→ 8

8 NULL values on the Unit Price columns:
Product-ccbc, Product 3d7f, Product 7eed, Product 84a5, Product dfc8, Product 15e0, Product 15f3, Product 9204

```
# pd.options.display.max_rows = 100
kt_df.describe()
```

	DATE	QUANTITY	UNIT PRICE
count	329881	329881.000000	329873.000000
mean	2024-07-18 11:49:34.170867200	2.321507	2319.016579
min	2024-01-01 05:54:00	0.000000	0.000000
25%	2024-04-30 16:34:00	1.000000	1420.000000
50%	2024-07-29 18:40:00	1.000000	1840.000000
75%	2024-10-14 21:32:00	2.000000	2750.000000
max	2024-12-31 18:24:00	359.000000	16136.000000
std	NaN	3.767796	1582.578700

```
# Checking for the data types to determine whether there are any inconsistencies.
# The UNIT PRICE column has Null Values
kt_df.info()
```

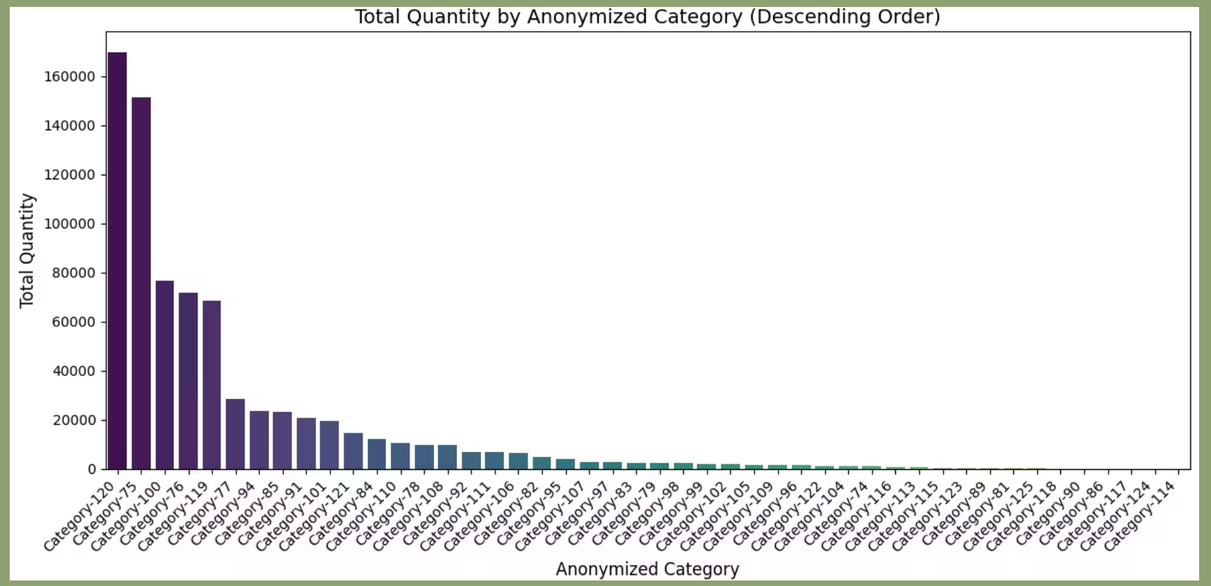
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 329881 entries, 0 to 329880
Data columns (total 8 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   DATE                 329881 non-null  datetime64[ns]
 1   ANONYMIZED CATEGORY  329881 non-null  object
 2   ANONYMIZED PRODUCT   329881 non-null  object
 3   ANONYMIZED BUSINESS  329881 non-null  object
 4   ANONYMIZED LOCATION  329881 non-null  object
 5   QUANTITY             329881 non-null  int64
 6   UNIT PRICE           329873 non-null  float64
 7   Month-Year           329881 non-null  object
dtypes: datetime64[ns](1), float64(1), int64(1), object(5)
memory usage: 20.1+ MB
```

02. Exploratory Data Analysis

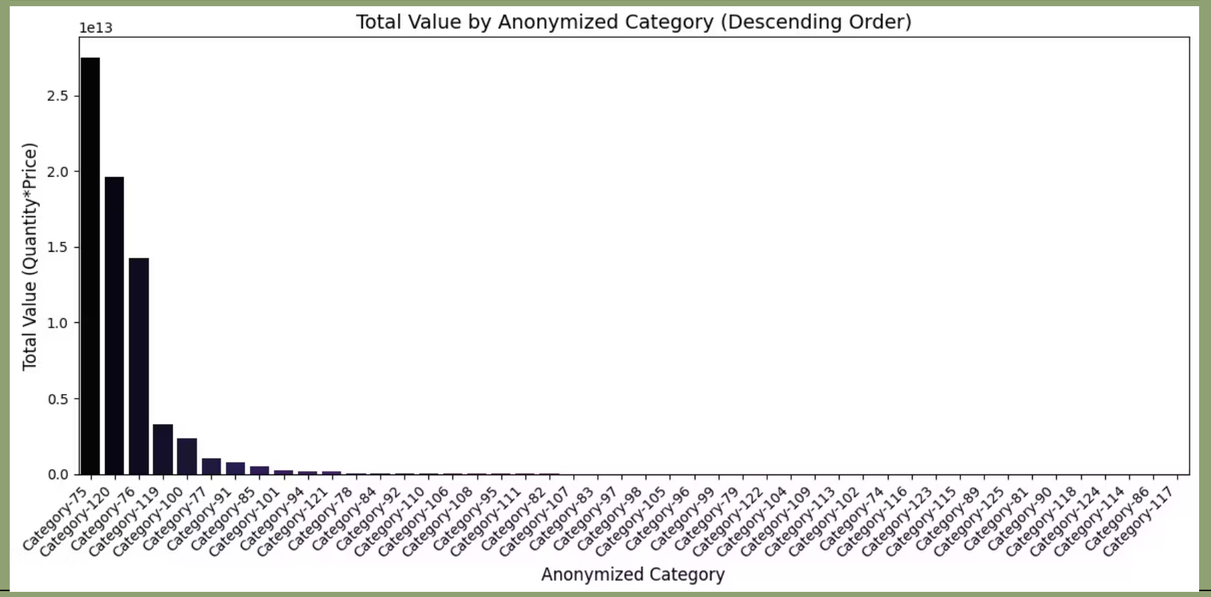
Overview

- Sales Overview: Quantity and Value grouped by Anonymized Category and Anonymized Business
- Trends Over Time: Analysis of sales trends by Month-Year. Time series showing seasonal patterns in sales performance.
- Performance Analysis: Top most frequently purchased products by Quantity & Top most valuable products based on Value

Total Quantity and Value grouped by Anonymized Category

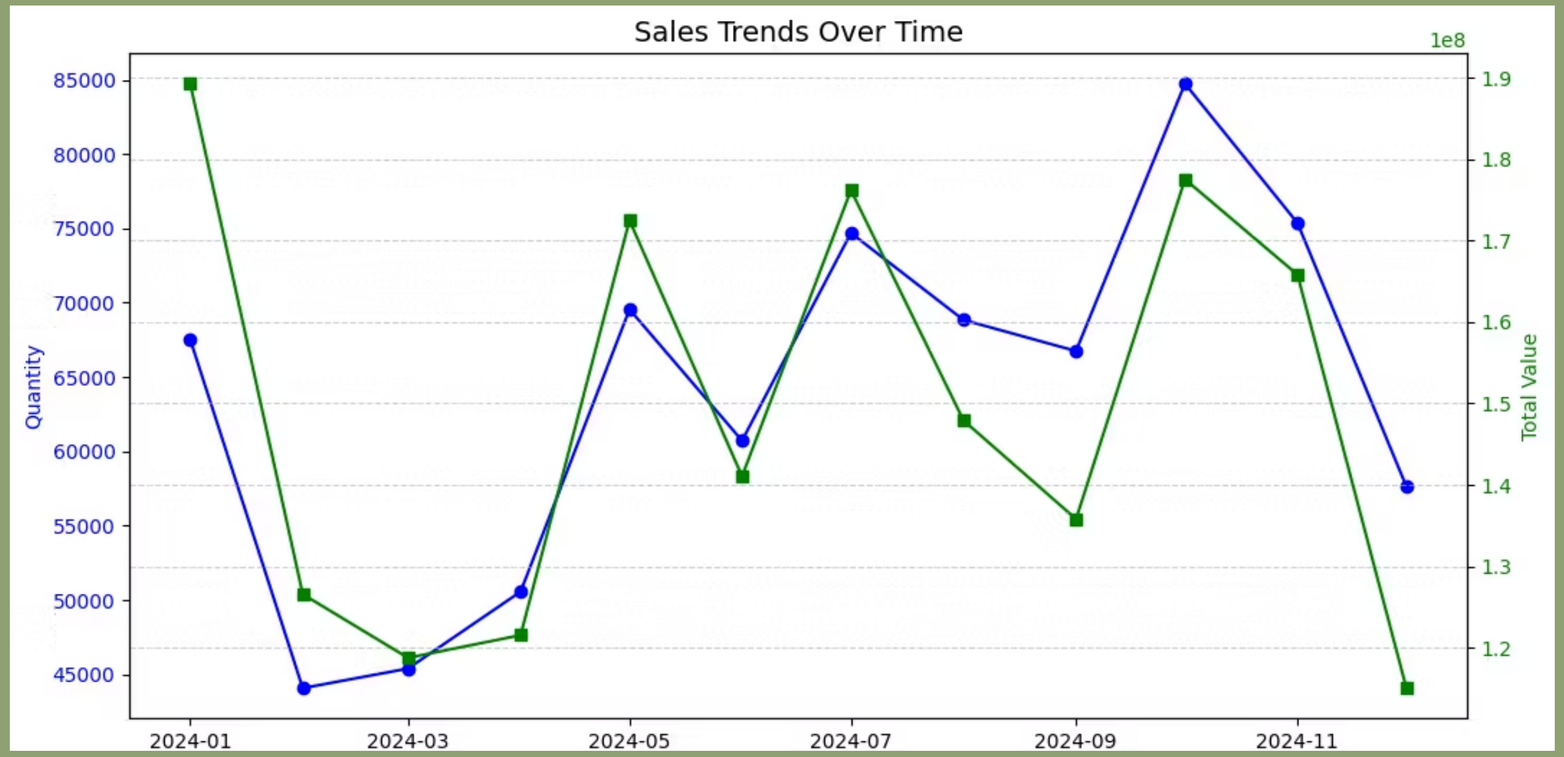


Total Quantity and Value grouped by Anonymized Category



Trends Over Time: Analysis of sales trends (Value and Quantity) by Month-Year

Time series plot showing seasonal patterns / changes in sales performance



Dual y-axis used due to difference in scales; Total Value is way larger than quantity

Sales Overview Table: Top 15 Businesses

Anonymized Business	Total Quantity	Total Unit Price
Business-978e	13991	4976687
Business-6068	8214	3407380
Business-fe7d	6743	2600735
Business-7a03	6318	2543926
Business-07de	6065	3023871
Business-ba13	5533	1939899
Business-468e	5450	1483889
Business-1e3e	4981	443257
Business-cb1f	4636	2032885
Business-80b3	4303	2612746
Business-0e5b	4289	4374082
Business-5613	4089	2037734
Business-f4f4	3852	2901220
Business-d72e	3835	2747470
Business-8119	3788	1091718

03. Performance Analysis

Overview

- Top 5 most frequently purchased products based on Quantity
- Top 5 most valuable products based on Value

Top five most frequently purchased products based on Quantity

```python
# Identifying the top 5 most frequently purchased products (Based on Quantity)
top_products = kt_df.groupby('ANONYMIZED PRODUCT')['QUANTITY'].sum().nlargest(5).reset_index()
print(top_products)
```

```
  ANONYMIZED PRODUCT  QUANTITY
0      Product-66e0     46957
1      Product-e805     42602
2      Product-8f75     37566
3      Product-29ee     35940
4      Product-4156     28487
```

Top five most valuable products based on value

```python
# Identifying the top 5 most valuable products (Bases on Value)
top_products_value = kt_df.groupby('ANONYMIZED PRODUCT')['UNIT PRICE'].mean().nlargest(5).reset_index()
print(top_products_value)
```

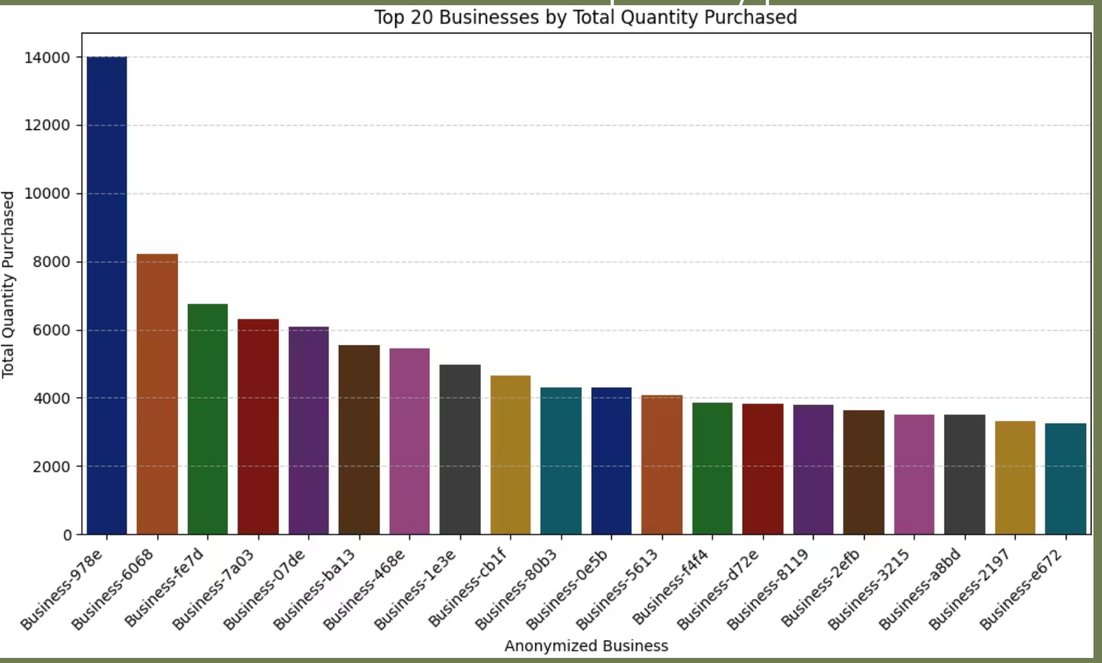
```
  ANONYMIZED PRODUCT    UNIT PRICE
0      Product-f084  15925.584906
1      Product-5a61  10920.000000
2      Product-6666   7990.000000
3      Product-41c3   7819.485294
4      Product-d527   7340.000000
```

04. Advanced Analysis

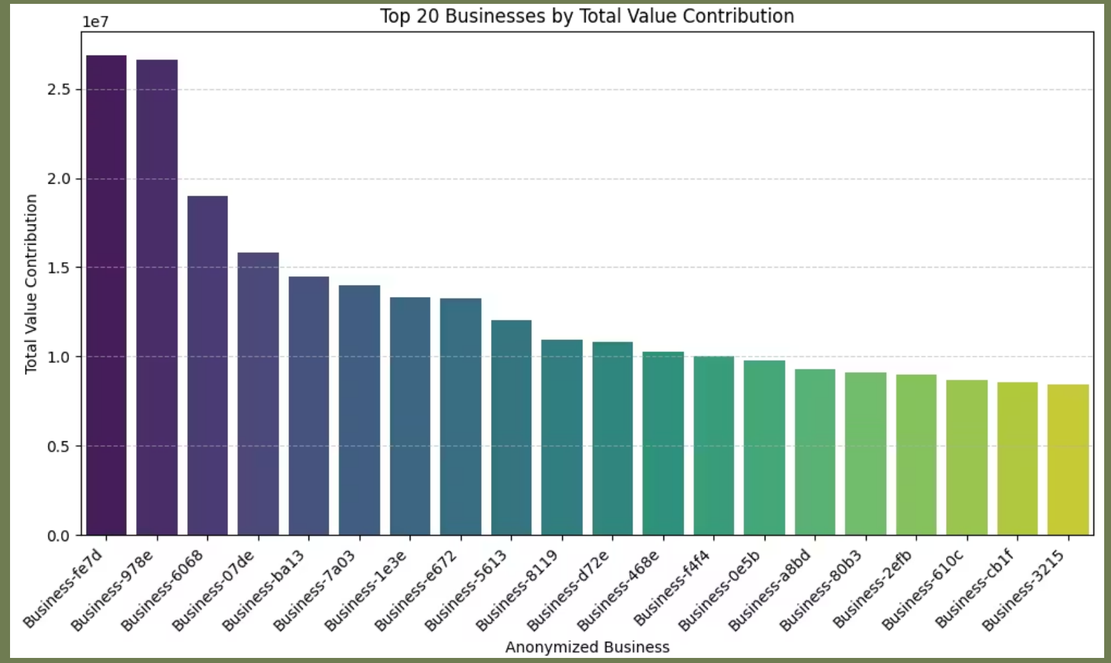
Overview

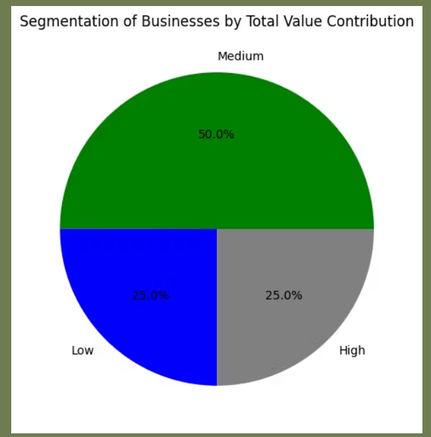
- **Customer Segmentation:** Segmentation analysis of businesses based on their purchasing behavior
- Classify businesses into 3 groups: High value, Medium value, Low value with a recommendations for engagement with each group.
- Forecasting
- Anomaly Detection
- Correlation Analysis

Sorted businesses based on quantity purchased



Sorted businesses based on total value contribution

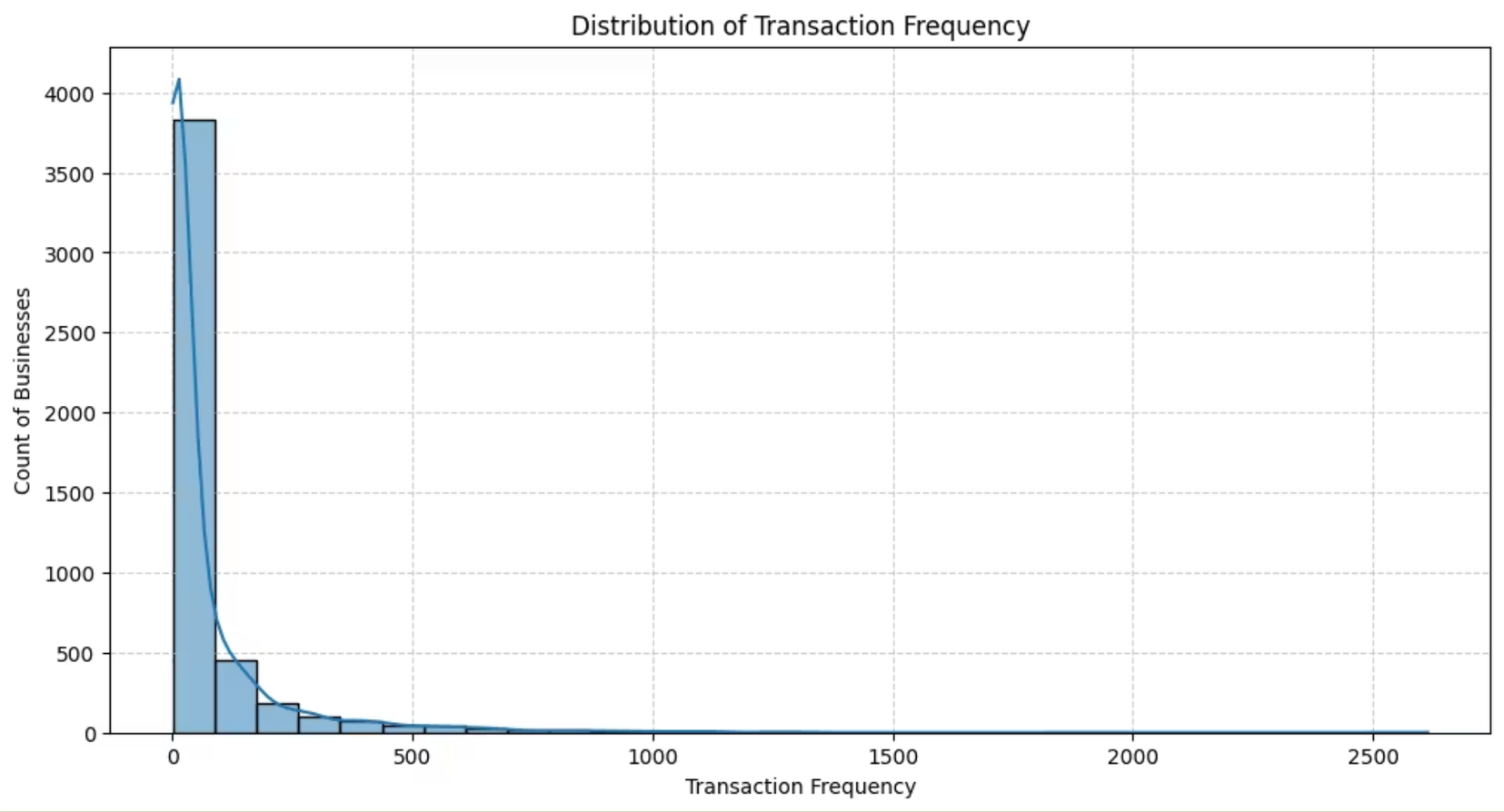




Segmentation of business based total value contribution

Product Groups	Engagement Strategies for each group
→ High Contribution Businesses (Top 25%)	• Exclusive Loyalty Programs: Offer priority services, special discounts, or cashback incentives to encourage repeat purchases. • Dedicated Account Managers: Assign a personal relationship manager for tailored services and faster issue resolution. • Early Access to Offers: Provide exclusive first-hand access to new products, premium support, and special pricing deals. • Volume-Based Discounts: Encourage bulk purchases by offering progressive discounts based on quantity ordered. • Co-Branding or Partnership Opportunities: Explore joint ventures or co-marketing strategies to deepen relationships
→ Medium Contribution Businesses (Middle 50%)	• Targeted Promotions & Bundled Offers: Provide tailored offers based on their purchasing history (e.g., "Buy X, Get Y Free"). • Flexible Payment Terms: Consider installment plans or credit options to encourage higher spending. • Regular Check-Ins & Upselling: Use personalized email campaigns or calls to introduce related products/services. • Training & Support Programs: Offer webinars, guides, or industry insights to help them optimize product use. • Referral Incentives: Reward them for bringing in new customers.
→ Low Contribution Businesses (Bottom 25%)	• Onboarding & Education: Provide step-by-step onboarding materials, product demos, and guides. • Freemium or Trial Offers: Offer introductory discounts or free samples to increase engagement. • Automated Follow-Ups: Use email campaigns and reminders to encourage re-purchases. • Exit Surveys & Feedback Requests: Understand why they are not buying more and address their concerns. • Segment-Based Upsell Strategies: Gradually introduce them to higher • value offerings through personalized recommendations.

KDE Plot showing Distribution of Transcation Frequency



Transaction Freqeuncy = Count of Transactions

Total Quantity Purchased = Sum of Quantity

Average Order Value = Mean of Quantity * Unit Price

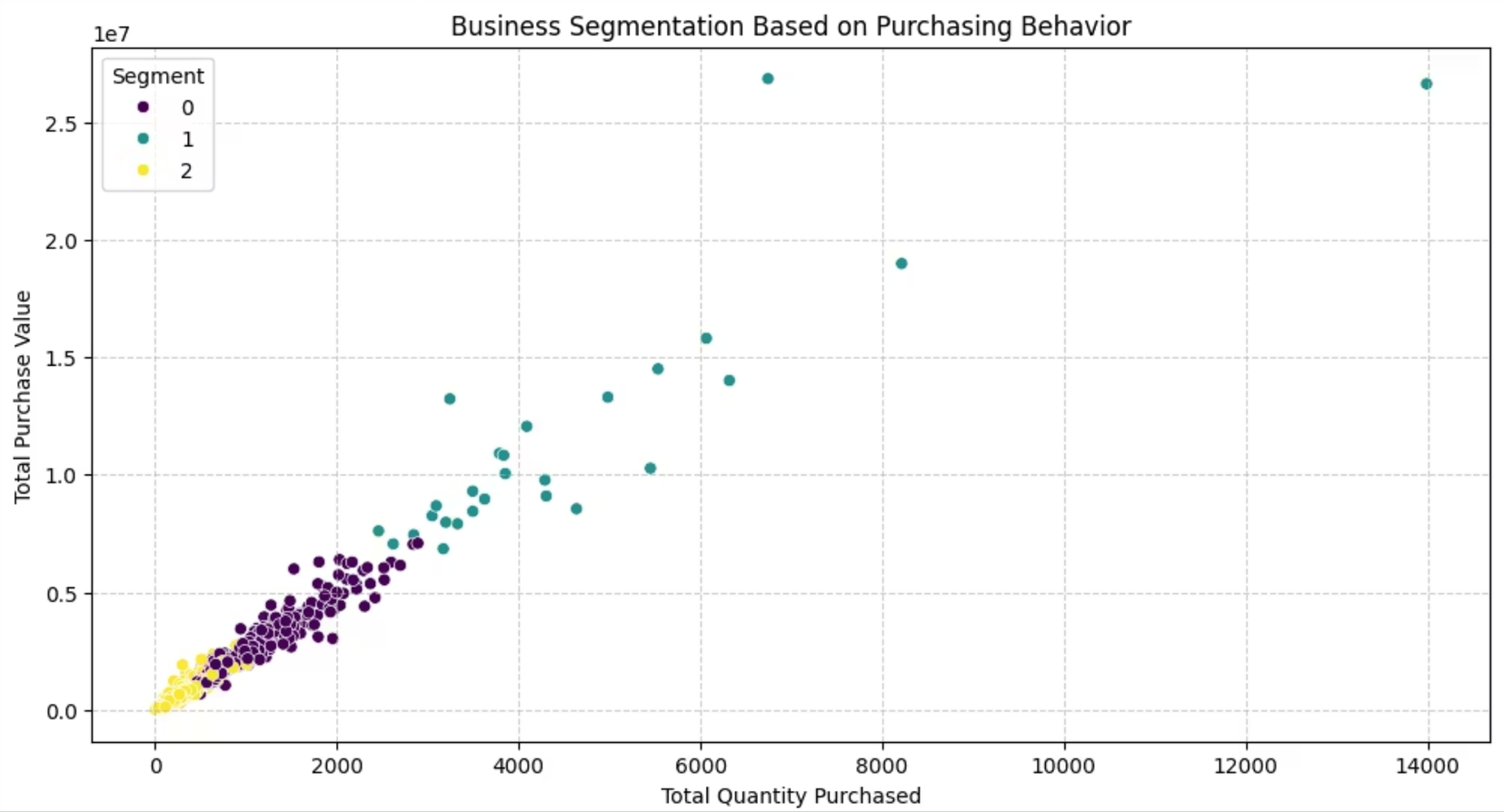
Total Purchase Value = Sum of Quantity * Unit Price

- The highest bar on the left shows that many businesses have few transactions
- A long tail towards the right shows that some businesses have significantly higher transactions
- This is suggests a right-skewed distribution, where a small number of businesses contribute to a large portion of transactions

Reasons for low transaction frequency

- Are these seasonal buyers?
- Are these new customers?
- Has transaction frequency changes over months?
- Next steps: Target high-frequency businesses for possible premium businesses

Scatter Plot using K-means clustering to segment business based on their purchasing behavior



Segment 0 : Represents a large number of businesses with lower purchase quantities and values.

Segment 1 : Businesses with moderate to high purchasing behavior.

Segment 2 : Businesses with very low purchase quantities and values.Trend:

There is a clear positive correlation.

Businesses purchasing more quantity tend to have a higher total purchase value

A few outliers in Segment 1 show extremely high purchases compared to others.

Scatter Plot with regression line showing relationship between quantity and total value

There is a positive correlation between quantity sold and total value

There is some variability whereby some data points deviate from the expected trend, which could indicate that other factors such as price fluctuations also influence this.

Actionable insights:

- If some high-quality sales have lower total value, consider revising pricing models
- Identify products which contribute most to revenue and prioritize them
- Data points that deviate signficantly from the trend could help understand external factors affecting sales.

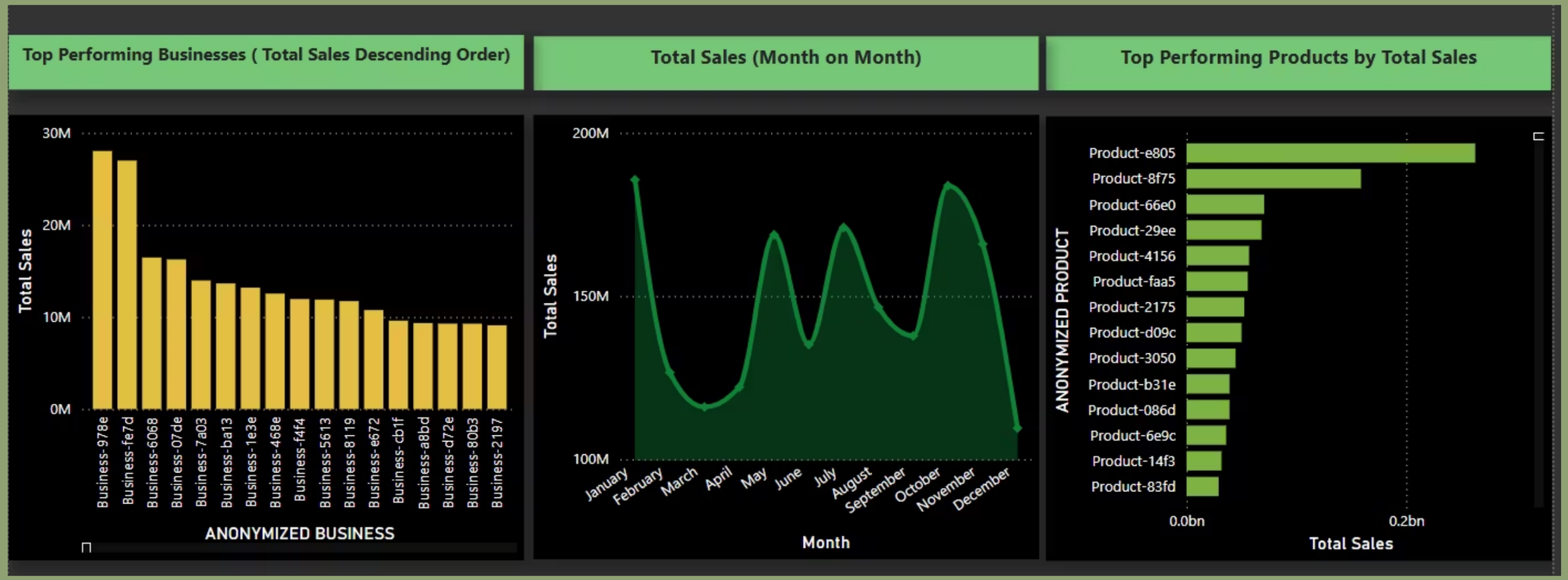
05. Dashboard and Reporting

Overview

- Total Quantity and Value by Anonymized Category
- Top-performing products and businesses
- A time-series chart of sales trends
- A segmentation summary of customer groups

Dashboard Overview

Sales breakdown over time



A segmentation summary of customer groups

A Segmentation summary of customer Groups

Customer Segment	Total Transactions	Avg Purchase Value	Total Sales	ANONYMIZED CATEGORY
High-Value Customer	10725	4,116.75	44152103	Category-91
High-Value Customer	4844	2,161.17	10468723	Category-92
High-Value Customer	8303	2,017.44	16750815	Category-94
High-Value Customer	3032	2,462.71	7466932	Category-95
High-Value Customer	1017	2,211.82	2249424	Category-96
High-Value Customer	1589	1,654.06	2628309	Category-97
High-Value Customer	1041	2,420.46	2519695	Category-98
High-Value Customer	1036	1,534.25	1589480	Category-99
Low-Value Customer	87	828.29	72061	Category-81
Low-Value Customer	175	782.00	136850	Category-89
Mid-Value Customer	907	512.09	464463	Category-102
Mid-Value Customer	238	1,787.23	425360	Category-115
Mid-Value Customer	445	949.99	422745	Category-116
Mid-Value Customer	60	4,951.00	297060	Category-125
Occasional Buyer	1			
Occasional Buyer	3	2,866.67	8600	Category-114
Occasional Buyer	4	387.50	1550	Category-117
Occasional Buyer	9	840.00	7560	Category-118
Occasional Buyer	4	2,515.00	10060	Category-124
Occasional Buyer	7	474.29	3320	Category-86
Occasional Buyer	15	1,050.00	15750	Category-90
High-Value Customer	**96704**	**18,292.67**	**1768974198**	

```
1 Customer Segment =
2     SWITCH(
3         TRUE(),
4         [Total Sales Per Customer] >= 500000, "High-Value Customer",
5         [Total Sales Per Customer] >= 200000, "Mid-Value Customer",
6         [Total Sales Per Customer] >= 50000, "Low-Value Customer",
7         "Occasional Buyer"
8     )
9
```

Above is a screenshot from DAX script showing the criteria I used to segment customers based on transactions and total sales

- High-value
- Mid-value
- Low-value

06. Strategic Insights & Recommendations

Overview

- Product Strategy
- Customer Retention
- Operational Efficiency

Product Strategy

The first step is identifying the top-performing product category based on the following criteria

- High Total Value - Ensures strong financial impact
- High Quantity Sold - Indicates strong demand
- Growth Potential - Consistent or increasing sales trends*

Category_75 is the top category based on total value, and we prioritize based on:

- It drives the most revenue
- It has a strong demand of 151,330 quantity sold, indicating a strong customer preference
- High potential for growth with further marketing
- Competitive advantage because if it has higher margins, promotions can increase profitability

Marketing Strategy Recommendations.

- Run targeted promotions such as offering deals and promotions to boost demand
- Feature the product in ads and on digital platforms
- Make the product more available in more locations or online platforms to expand distribution channels
- Analyze customer segments to identify key buyers and tailor promotions accordingly

Customer Retention

To identify businesses with reduced purchase frequency we do the following:

- Track the purchase frequency over time for each business
- Compare early vs recent transactions to detect declines
- Provide re-engagement strategies to win back customers

Re-engagement strategies:

- Personalized offers and discounts to the identified lost customers
- Follow-up emails and surveys to get to know what improvements these customers expect
- Product recommendations based on past purchases by suggesting new products or bundle deals
- Direct sales outreach; If high-value customers are dropping, assign sales rep to call them

Operational Efficiency

To optimize inventory management and supply chain processes:

- Analyze product performance trends (fast vs. slow-moving items)
- Identify seasonal demand fluctuations to adjust stock levels.
- Suggest actionable improvements to reduce stockouts and excess inventory

- Fast-moving products require frequent restocking and supply chain efficiency
- Slow-moring products may cause excess inventory, therefore consider promotions or stock reductions

- For the months where demand spikes, adjust inventory stocking and supplier leads times
- For the months where there's demand drops, reduce stocks and offer discounts to help clear inventory

Recommendations for inventory and supply chain optimization.

- Implement demand forecasting using historical sales data to predict future demand and avoid excess stocks or stockouts. e.g If sales peaks in December, order stock in October to avoid delays
- Adopt just-in-time inventory to keep minimum stock levels and order supplies only when needed
- Identifying suppliers with consistent delivery times and negotiate shorter lead times, you can consider secondary suppliers to avoid delays incase of unreliability
- Reduce deadstock for the slow-moving products using discounts, bundle offers, and limited-time promotions to create urgency
- Improve warehouse organization by storing fast-moving products near dispatch areas for faster order fulfillment