Retail analysis Report

🔎 1. General Overview

- Shape of the dataset (rows × columns).
  - `(9994 Rows , 23 Columns )`
- Missing values per column.

```
Row ID          0
Order ID        0
Order Date      0
Ship Date       0
Ship Mode       0
Customer ID     0
Customer Name   0
Segment         0
Country         0
City            0
State           0
Postal Code     0
Region          0
Product ID      0
Category        0
Sub-Category    0
Product Name    0
Sales           0
Quantity        0
Discount        0
Profit          0
```

- Summary statistics (mean, median, min, max, std).

Row ID	Postal Code	Sales	Quantity	Discount	Profit
count	9994.000000	9994.000000	9994.000000	9994.000000	9994.000000
mean	4997.500000	55190.379428	229.858001	3.789574	0.156203
std	2885.163629	32063.693350	623.245101	2.225110	0.206452
min	1.000000	1040.000000	0.444000	1.000000	0.000000
25%	2499.250000	23223.000000	17.280000	2.000000	0.000000
50%	4997.500000	56430.500000	54.490000	3.000000	0.200000
75%	7495.750000	90008.000000	209.940000	5.000000	0.200000
max	9994.000000	99301.000000	22638.480000	14.000000	0.800000

1. Sales

- Mean = **229.9**, Std = **623.2** → **Std > Mean** 🚨
- Min = 0.44, Max = **22,638.48**
- Median (50%) = 54.49 → way lower than mean (229.9).

👉 Interpretation:

- **Right-skewed distribution** → Most sales are small, but some extremely large sales drive the average way up.
- Mean is not representative; **median** is a better "typical" value here.
- Outliers (like 22,638) inflate variability.

---

2. Quantity

- Mean = 3.79, Std = 2.22
- Min = 1, Max = 14

  👉 Distribution is fairly tight around the mean.

  Most purchases are **small quantities (1–5 items)**.

  This looks reasonable and not very skewed.

---

3. Discount

- Mean = 0.156 (~15.6%)
- Std = 0.206
- Min = 0, Max = 0.8 (80%)
- 50% of values = 0.20 or below

👉 Discounts are mostly **0–20%**, but some go as high as **80%**.

This suggests promotions/clearances exist, but most discounts are modest.

---

4. Profit

- Mean = **28.66**, Std = **234.26** → Std ≫ Mean 🚨
- Min = **–6599.98 (huge losses)**
- Max = **8399.98 (huge gains)**
- Median = **8.66** → much lower than mean.

👉 Interpretation:

- Profits are **highly variable**, sometimes **negative**.
- Most orders make only small profits (median = 8.66), but a few orders cause very large profits or very large losses.
- The distribution is extremely **skewed and heavy-tailed**.
- Mean is misleading here; median and **profit margin analysis by category** would be more meaningful.

🔎 Insights

1. **Sales and Profit are highly skewed** → better use log transformation or percentiles when modeling.
2. **Std > Mean for Sales & Profit** → strong variability, mean is not a good summary measure.
3. **Profit can be negative** → not all sales are profitable. You'll want to segment by **category, sub-category, or discount levels** to see why.
4. **Discounts impact profitability** → check if higher discounts correlate with losses.
5. **Quantity is stable** → most orders are small, so variability is mainly in **price/profit**, not in quantity.

✅ So: This dataset tells a story of **mostly small transactions, a few very large ones, and significant risks (losses) when discounts are high.**

That's why this dataset is often used for **profitability, discount strategy, and customer segmentation analysis**.

- Data types check.
  -

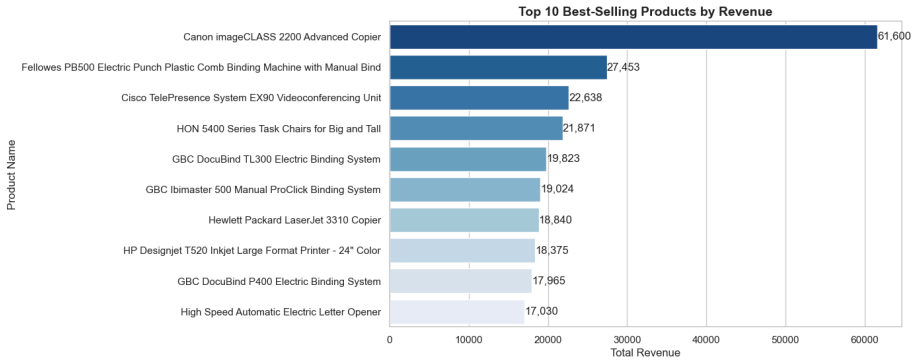
```
RowID              int64
OrderID           object
OrderDate     datetime64[ns]
ShipDate          object
ShipMode          object
CustomerID        object
CustomerName      object
Segment           object
Country           object
City              object
State             object
PostalCode         int64
Region            object
ProductID         object
Category          object
Sub-Category      object
ProductName       object
Sales            float64
Quantity           int64
Discount         float64
Profit           float64
Month              int32
Year               int32
```

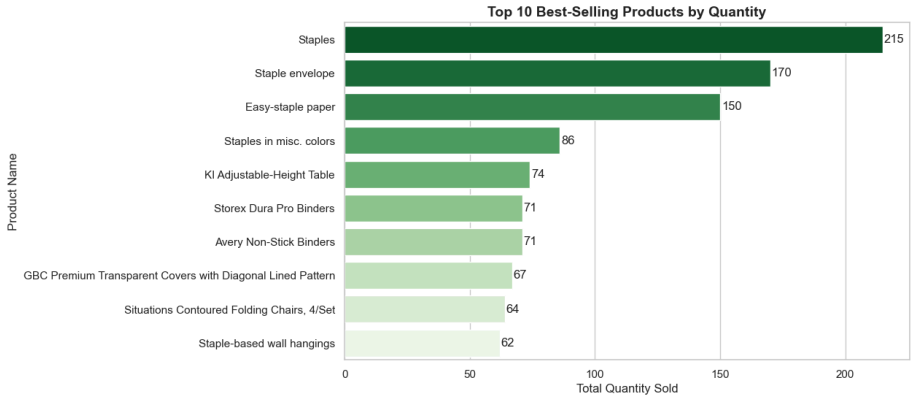
📊 2. Sales Performance Insights

- **Top-selling products** (by revenue & quantity).

```
Top_ selling products by revenue

ProductName
Canon imageCLASS 2200 Advanced Copier                                  61599.824
Fellowes PB500 Electric Punch Plastic Comb Binding Machine with Manual Bind    27453.384
Cisco TelePresence System EX90 Videoconferencing Unit                  22638.480
HON 5400 Series Task Chairs for Big and Tall                           21870.576
GBC DocuBind TL300 Electric Binding System                             19823.479
GBC Ibimaster 500 Manual ProClick Binding System                       19024.500
Hewlett Packard LaserJet 3310 Copier                                   18839.686
HP Designjet T520 Inkjet Large Format Printer - 24" Color              18374.895
GBC DocuBind P400 Electric Binding System                              17965.068
High Speed Automatic Electric Letter Opener                            17030.312


Top_selling products by quantity

ProductName
Staples                                 215
Staple envelope                         170
```

Easy-staple paper	150
Staples in misc. colors	86
KI Adjustable-Height Table	74
Storex Dura Pro Binders	71
Avery Non-Stick Binders	71
GBC Premium Transparent Covers with Diagonal Lined Pattern	67
Situations Contoured Folding Chairs, 4/Set	64
Staple-based wall hangings	62



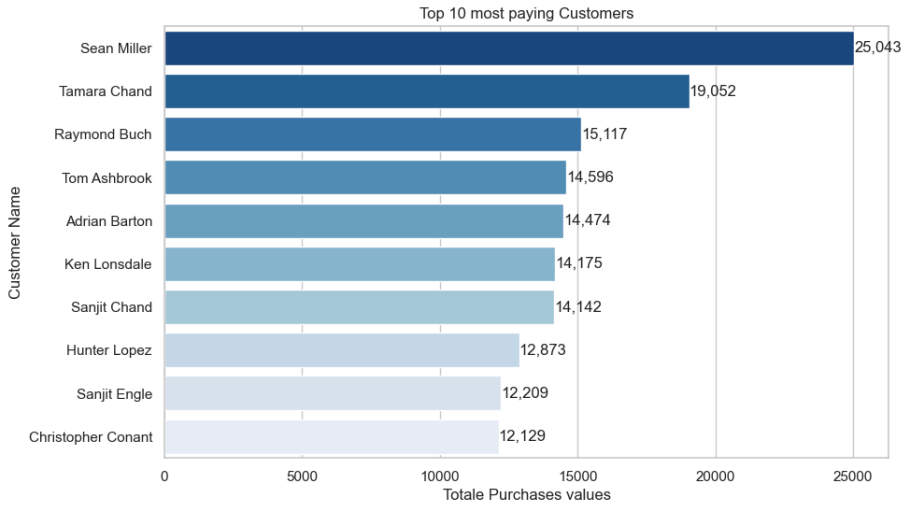


- **Top customers** (by total purchases).

Top 10 customers by total purchases:

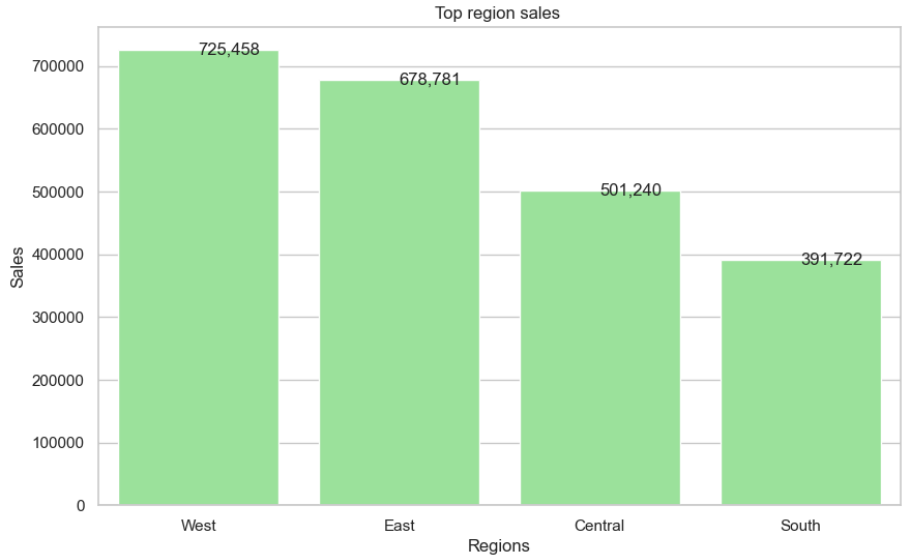
CustomerName	
Sean Miller	25043.050
Tamara Chand	19052.218
Raymond Buch	15117.339
Tom Ashbrook	14595.620
Adrian Barton	14473.571
Ken Lonsdale	14175.229
Sanjit Chand	14142.334
Hunter Lopez	12873.298

```
Sanjit Engle         12209.438
Christopher Conant   12129.072
```



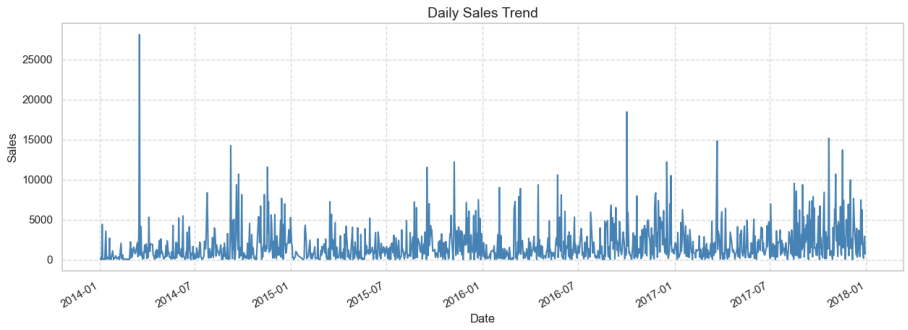
- **Top countries/regions** (by sales).

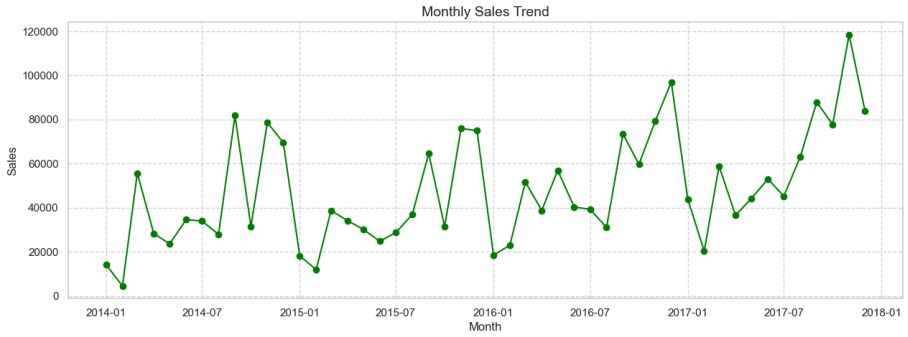
```
Top Regions by sales
Region
West      725457.8245
East      678781.2400
Central   501239.8908
South     391721.9050
```



📈 3. Time-Series Analysis

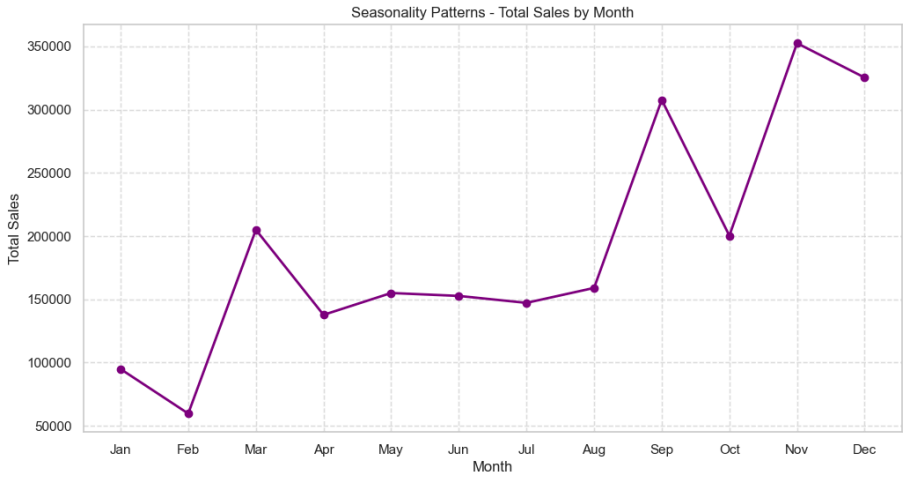
- Sales trend over time (daily, monthly).





- Seasonality patterns (peak months).

```
Month
1     94924.8356
2      59751.2514
3     205005.4888
4     137762.1286
5     155028.8117
6     152718.6793
7     147238.0970
8     159044.0630
9     307649.9457
10    200322.9847
11    352461.0710
12    325293.5035
```

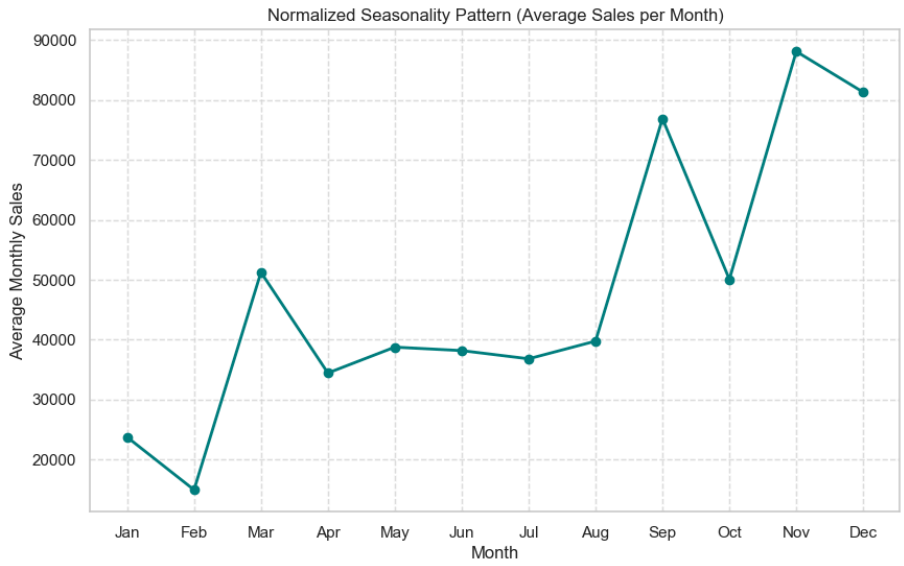


However If we just sum sales by month across all years:

- Years with more customers or more invoices will dominate.
- Example: If 2011 has twice as many transactions as 2010, December 2011 will "inflate" the December total, making it look like December is always the strongest month.

👉 To truly detect seasonality, we want to remove the effect of different year sizes

Normalized Seasonality Pattern (Average Sales per Month



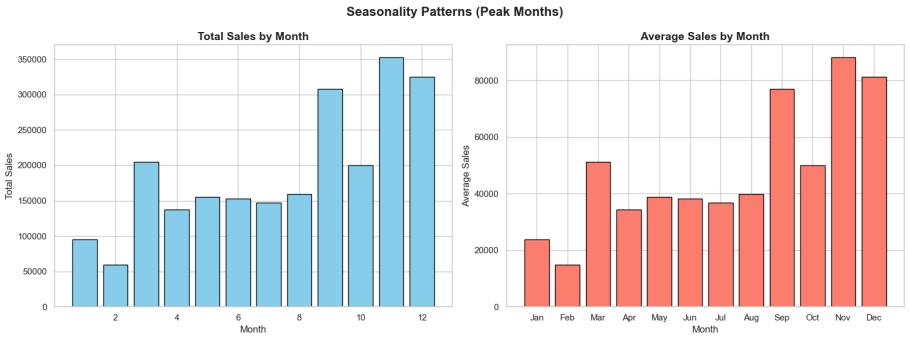
I placed **two bar plots side by side**:

- **Left:** Total monthly sales across all years (so you see absolute peaks).
- **Right:** Average monthly sales per month (so you see normalized seasonality across years).

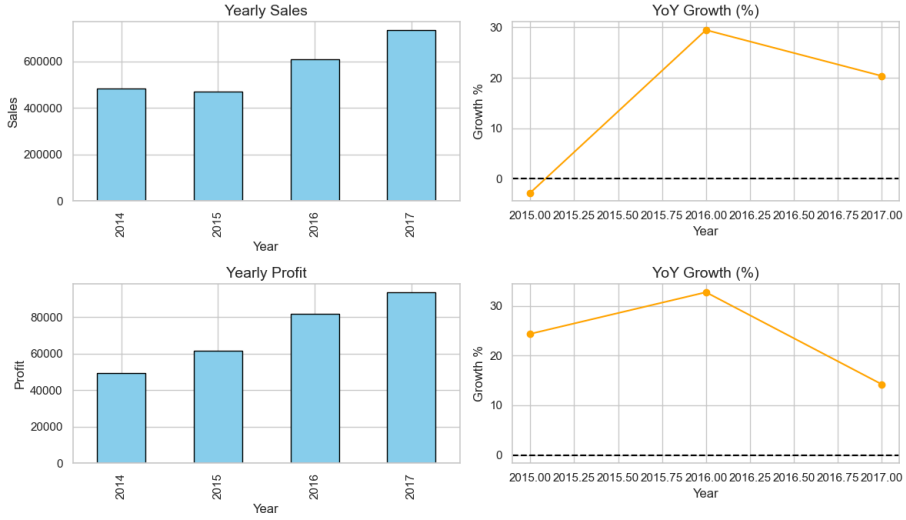
Insights:

- **November (11)** is the strongest peak, followed by **December (12)** and **September (9)** → very strong seasonal effect.
- **February (2)** is consistently the weakest month.

- Averaging confirms that these peaks hold across years (not just one outlier year).



- Year-over-year growth.



Sales			Profit			Profit_Margin	
Category	Furniture	Office Supplies	Technology	Furniture	Office Supplies	Technology	Furni
Year							
2014	157192.8531	151776.412	175278.233	5457.7255	22593.4161	21492.8325	0.034
2015	170518.2370	137233.463	162780.809	3015.2029	25099.5338	33503.8670	0.017
2016	198901.4360	183939.982	226364.180	6959.9531	35061.2292	39773.9920	0.034
2017	215387.2692	246097.175	271730.811	3018.3913	39736.6217	50684.2566	0.014

- **2015 was a weak year** → Sales dipped slightly (−2.83%). This could be due to fewer customers, seasonal issues, or external market conditions. Worth investigating which categories or regions underperformed.

  - **Sales Growth (2014 → 2015)**
    - **Furniture**: ↑ **8.5%** (157,193 → 170,518)
    - **Office Supplies**: ↓ **9.6%** (151,776 → 137,233)
    - **Technology**: ↓ **7.1%** (175,278 → 162,781)
    - **Total Sales**: ↓ **~3.5%** overall

So the *decline wasn't across the board* → Furniture grew, but Office Supplies and Technology dragged down the total.

overall sales dipped mainly due to Office Supplies (-9.6%) and Technology (-7.1%). Furniture actually grew but at the cost of lower margins. This suggests that the company faced weaker demand in non-furniture categories, possibly because of market trends (digitization, delayed tech upgrades). However, higher profit margins in Office Supplies and Tech indicate the company may have tightened pricing and controlled discounts to preserve profitability, even at the cost of lower sales.

Profit Growth (2014 → 2015)

Technology

- **Sales**: 175,278 → 162,780 (↓ dropped)
- **Profit**: 21,492 → 33,503 (↑ increased a lot)
- **Profit Margin**: 12.3% → 20.6% (↑ stronger efficiency)

👉 Even though **sales volume decreased**, the company was **selling more profitable products** or **managing costs better** in 2015.

Office Supplies

- **Sales**: 151,776 → 137,233 (↓ dropped)
- **Profit**: 22,593 → 25,100 (↑ increased)
- **Profit Margin**: 14.9% → 18.3% (↑ improved efficiency)

👉 Same situation: fewer sales, but **higher margins**, so total profit still increased.

Furniture

- **Sales**: 157,192 → 170,518 (↑ increased)
- **Profit**: 5,457 → 3,015 (↓ dropped a lot)
- **Profit Margin**: 3.5% → 1.8% (↓ margin collapsed)

👉 Furniture went the opposite way: more sales, but **very low profitability**, maybe due to heavy discounts, higher costs, or low-margin items.

📌 **Conclusion**:

- For **Technology & Office Supplies**, 2015 was a year of **"quality over quantity"** — fewer sales but **better margins**, leading to **higher profits**.
- For **Furniture**, they sold more but made less profit → could mean **bad pricing strategy** or **cost inefficiencies**.

- **2016 was a breakout year** → Sales grew nearly **30% YoY**, the highest in the dataset. This might be linked to successful product launches, promotions, or expansion into new markets.
  - **1. Drill down by category or region (2016 growth driver)**
    - Sales jumped **~29% in 2016**.
    - To confirm what drove this:
    - **Group sales by Category and Year** → look for which categories grew the most in 2016.

Category	Furniture	Office Supplies	Technology
Year			
2014	NaN	NaN	NaN
2015	8.477093	-9.581824	-7.130049
2016	16.645257	34.034351	39.060729
2017	8.288444	33.792106	20.041435

- Group Profit by Category and Year → look for which categories grew the most profit in 2016.

Category	Furniture	Office Supplies	Technology
Year			
2014	NaN	NaN	NaN
2015	-44.753489	11.092248	55.883907
2016	130.828682	39.688767	18.714631
2017	-56.632017	13.334936	27.430650

Contribution to 2016 Growth

- **Total sales growth (2015 → 2016):** +29.5% (from your YoY calc).
- Breaking down by category:
  - **Furniture:** 8.47% → 16.64% → **+8.17%**
  - **Office Supplies:** -9.58% → 34.03% → **+43.61%**
  - **Technology:** -7.13 → 39.06% → **+46.19%**

✅ Insight:

The **2016 jump was mainly driven by Technology (+46.19%) and Office Supplies (+43.61%)**, while Furniture grew more moderately (+8.17%).

So, the surge wasn't evenly spread—it was largely a **Tech & Supplies boom**.

- **2017 continued strong growth** → Another **20% increase**, showing momentum. Growth slowed compared to 2016, but still indicates a healthy business trajectory.
- **Trend**: Overall, from 2014 → 2017, sales grew from **~484K → ~733K**, a **total increase of ~52%** in 4 years. That's a solid long-term upward trend.

1. Sales Trends (2014 → 2017)

- **Furniture**: steady growth → **157k → 215k** (≈ +37%).
- **Office Supplies**: strong growth, especially in 2017 → **152k → 246k** (≈ +62%).
- **Technology**: also grows strongly → **175k → 272k** (≈ +55%).

✅ Technology consistently leads in sales, followed by Office Supplies (surging in 2017), then Furniture.

---

2. Profit Trends

- **Furniture**: fluctuates, low profit compared to sales:
  - 2014: 5.4k → peak in 2016 (6.9k) → drop to 3.0k in 2017.
- **Office Supplies**: climbs consistently → **22.6k → 39.7k**.

- **Technology**: very profitable → **21.5k → 50.7k**.

✅ Technology dominates profit, Office Supplies is solid, Furniture struggles.

3. Profit Margin (Profit ÷ Sales)

- **Furniture**: very low → 1.7%–3.5% (barely profitable).
- **Office Supplies**: strong → ~15%–19%.
- **Technology**: strongest margins → ~12%–20%.

📉 **Furniture looks like a red flag** → big sales volume but razor-thin margins.

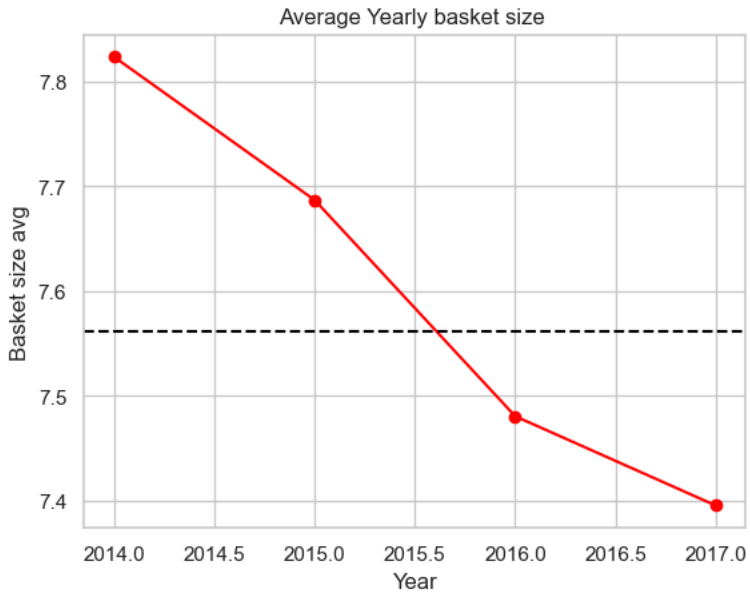
📈 **Office Supplies & Technology drive profitability**.

4. Insights

- **Best performer**: Technology → high sales, high profit, strong margins.
- **Hidden gem**: Office Supplies → smaller sales but excellent margins.
- **Weak spot**: Furniture → poor margins, profit barely grows despite sales growth.

👥 4. Customer Behavior

- Average basket size (# of items per invoice).
  - Average Basket Size: 7.56 items per invoice
  - By Year → see if customers started buying more/less items per basket over time.



Shrinking Basket Size

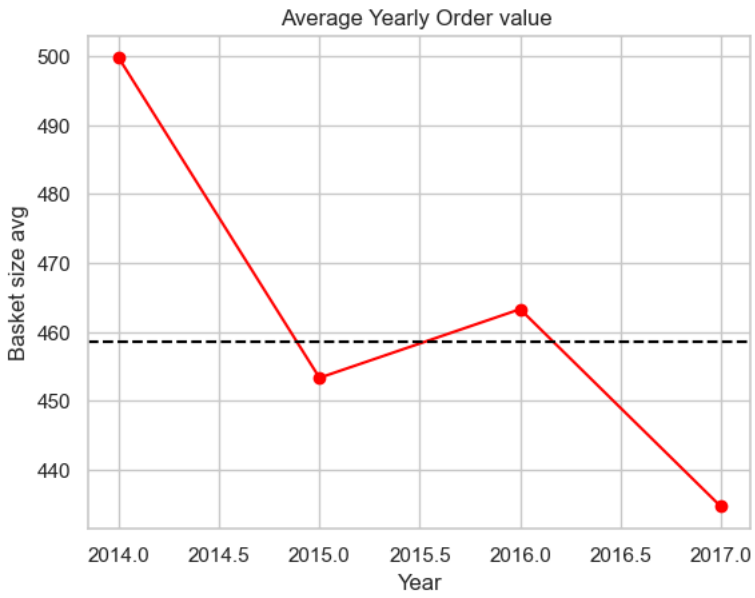
- The **average basket size** decreased from **7.82 items (2014)** to **7.39 items (2017)** (↓ 5.5%).

- Customers are buying **fewer items per invoice**.

👉 Interpretation:

- Customers may be **purchasing only essentials** instead of bulk shopping.
- Rise of **single-item transactions** (possibly tech products bought individually).

- Average order value.
  - Average Order Value : 458.61 sale per invoice



Declining Order Value

- In **2014**, the **average order value (AOV)** was ~$500 .
- By **2017**, it dropped to ~$435 (↓ 13%).
- This suggests that customers are spending **less per transaction over time**.

👉 Possible reasons:

- Discounts or promotions leading to lower invoice totals.
- Customers buying cheaper products (shift in product mix).
- Increased price sensitivity.

Combined Effect

- **Both metrics (AOV & basket size) are declining**, meaning:
  - Customers are spending less **and** buying fewer items.
  - Could signal **increased competition**, **economic slowdown**, or **shifts in purchasing behavior**.

But Important Note

- Even though **order-level metrics are dropping**, **total sales by year kept growing (2015–2017)**.

- This implies **higher order volumes** (more customers or more frequent purchases) are compensating for the decline per transaction.

✅ **Insight Summary:**

- From 2014 → 2017, **customers buy less per order (in value & items)**.
- Growth in **total sales** is being driven by **more transactions**, not by larger/more expensive baskets.
- This shift highlights the importance of **customer acquisition** and **order frequency**, rather than relying on upselling in each transaction.

Lets see if discount effect the drop of order value and basket size

```
Year
2014    32.555212
2015    31.511561
2016    30.442586
2017    30.718435
```

Average Discount is Falling

- From **32.55% (2014) → ~30.7% (2017)**, the discounts have slightly decreased.
- This means customers are paying closer to the **full price** than before.

🔎 What This Means

- **Customer behavior is changing** → they're becoming **more price-sensitive** and **buying smaller baskets**.
- smaller discounts likely contributed to lower basket sizes and reduced order value
- This could suggest:
  - Stronger competition (customers compare more and buy less in one place).
  - Economic/environmental factors reducing purchasing power.
  - A shift toward **Technology** products (higher profit but fewer items per basket).

```
yearly_order_counts Year
2014     969
2015    1038
2016    1315
2017    1687

customers_count Year
2014    595
2015    573
2016    638
2017    693
```

✅**Final interpretation**:

The business is attracting **more repeat purchases** (good for loyalty), but**each order is worth less**

. The reduction in discounts may have caused customers to spread purchases out into smaller, lower-value baskets.

Growth in total sales is being driven by more transactions, and increase in customers count not by larger/more expensive baskets.

- Customer segmentation: frequent vs. one-time buyers.

```
Frequent Buyers: 781
One-Time Buyers: 12
Frequent Buyers Contribution: 2292033.2202999997
One-Time Buyers Contribution: 5167.639999999999
```

1. **Almost all revenue comes from repeat customers**
   - 781 frequent buyers generated **>99.7% of sales**.
   - One-time buyers are negligible (just ~0.22%).
2. **Customer Loyalty is extremely strong**
   - This suggests that once a customer buys, they are very likely to return.
   - Could indicate strong **customer relationships, brand trust, or recurring business needs** (like B2B).
3. **Strategic implications**
   - Retention > Acquisition: It's more profitable to **retain existing buyers** than to chase new ones.

RFM Analysis (Recency, Frequency, Monetary value).

- Segment "555" = Champions
- Segment "111" = Lost Customers
- Score closer to 15 = higher value.

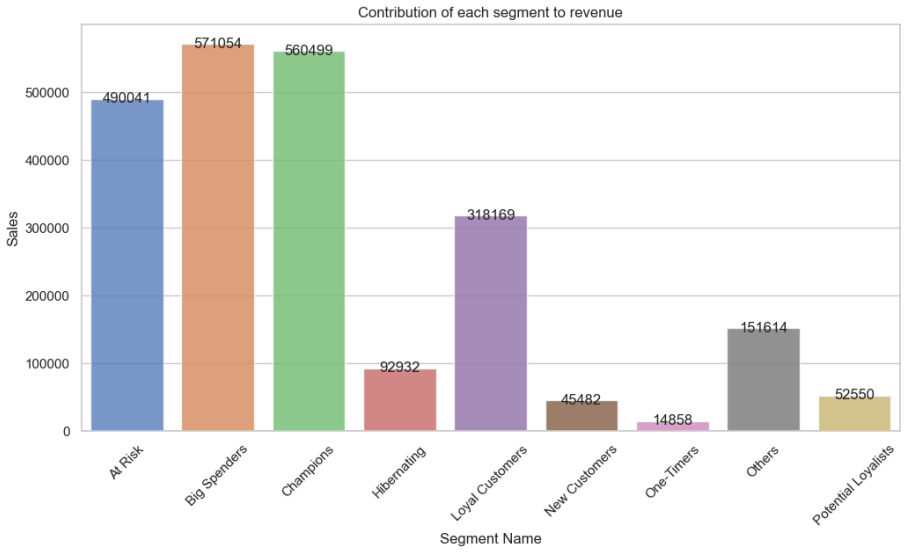
Map Segments

Define business-friendly groups:

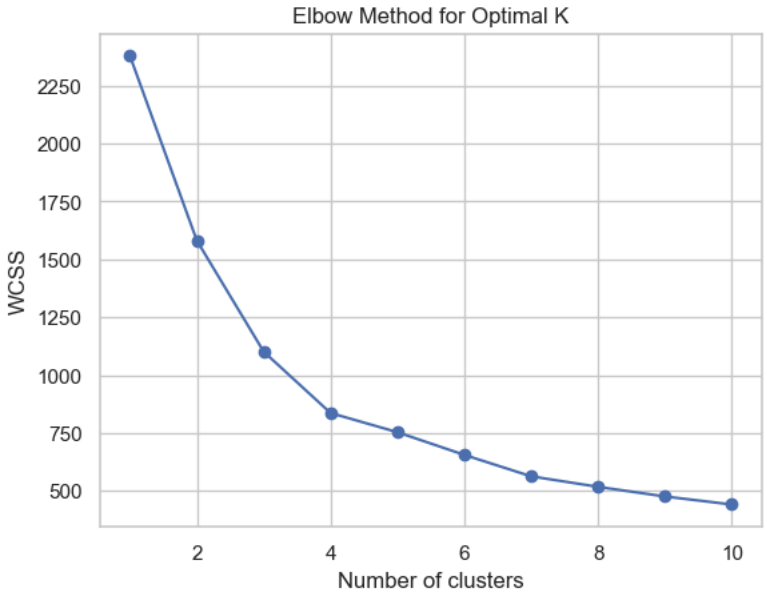
RFM Score Range	Segment Name	Meaning
13–15	**Champions**	Recent, frequent, high spenders
10–12	**Loyal Customers**	Frequent, moderate spend
7–9	**Potential Loyalists**	Could become champions
4–6	**At Risk**	Not recent, low frequency
1–3	**Lost Customers**	Haven't purchased in a long time

• Contribution of each segment to revenue.

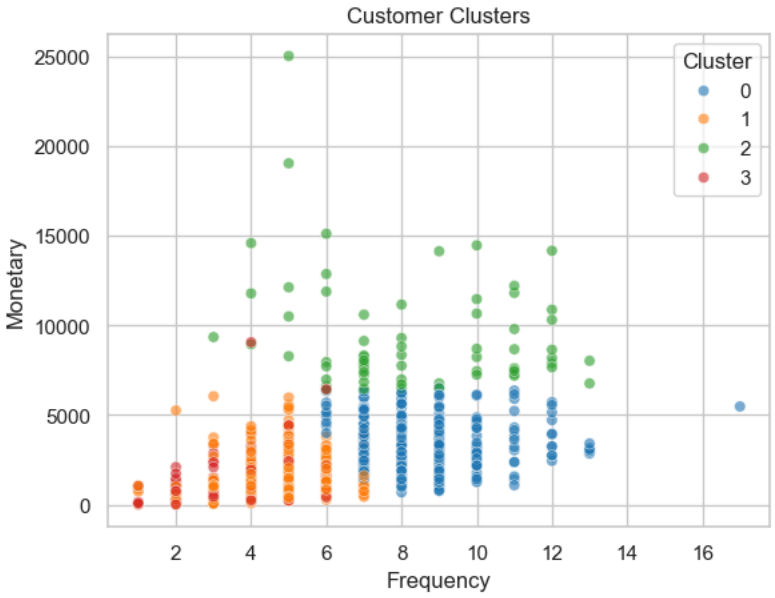
```
SegmentName
At Risk              490041.0701
Big Spenders         571054.4098
Champions            560498.8171
Hibernating           92932.0525
Loyal Customers      318169.3881
New Customers         45482.0778
One-Timers            14858.0844
Others               151614.4945
Potential Loyalists   52550.4660
```



Clustering customers (KMeans).



Cluster	Recency	Frequency	Monetary	Count
0	72.741611	8.516779	3322.222985	298
1	101.197015	4.731343	1669.688290	335
2	123.718750	8.296875	9479.545687	64
3	559.489583	3.697917	1470.228226	96



📊 Cluster Profiles

1. **Cluster 0 (298 customers)**
   - Recency ≈ 73 → They purchased relatively recently.
   - Frequency ≈ 8.5 → Medium purchase frequency.
   - Monetary ≈ 3,322 → Good spenders.
     🔷 Likely **loyal customers**.
2. **Cluster 1 (335 customers)**
   - Recency ≈ 101 → Longer since last purchase.
   - Frequency ≈ 4.7 → Low to medium frequency.
   - Monetary ≈ 1,670 → Lower spend.
     🔷 These are **at-risk or occasional customers**.
3. **Cluster 2 (64 customers)**
   - Recency ≈ 124 → Last purchase was a while ago.
   - Frequency ≈ 8.3 → Fairly frequent.
   - Monetary ≈ 9,480 → Very high spenders.
     🔷 These are **high-value VIP customers**.
4. **Cluster 3 (96 customers)**
   - Recency ≈ 559 → Extremely long since last purchase.
   - Frequency ≈ 3.7 → Very low frequency.
   - Monetary ≈ 1,470 → Low spenders.
     🔷 These are **churned customers**.

```
         Recency  Frequency     Monetary  Count           Segment
Cluster
```

0	72.741611	8.516779	3322.222985	298	Champions
1	101.197015	4.731343	1669.688290	335	Needs Attention
2	123.718750	8.296875	9479.545687	64	Big Spenders at Risk
3	559.489583	3.697917	1470.228226	96	Lost