# MegaMart Strategic Business Analysis

**Type:** Exploratory Data Analysis (EDA) & Customer Segmentation

**Date:**

**Author:** Krishna Varshney

## 1. Executive Summary

**Objective:** To evaluate MegaMart's retail performance over the last two years, identifying revenue drivers, profit risks, and high-value customer segments.

**Key Focus Areas:**

1. **Financial Health:** Monthly trends and profit margin stability.
2. **Product Strategy:** Identifying high-volume vs. high-risk categories.
3. **Regional Intelligence:** Store-level performance heatmaps.
4. **Customer Value:** RFM Segmentation to identify VIPs vs. Churned users.

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
import matplotlib.ticker as ticker

# 1. Clean Layout
warnings.filterwarnings('ignore')
pd.set_option('display.max_columns', None)
pd.options.display.float_format = '{:,.2f}'.format

# 2. Professional Aesthetics (Teal Theme)
TEAL_PALETTE = ["#b2d8d8", "#66b2b2", "#008080", "#006666", "#004c4c"]
sns.set_palette(TEAL_PALETTE)
sns.set_style("whitegrid")
plt.rcParams['figure.figsize'] = (14, 7)
plt.rcParams['font.size'] = 12

# 3. Helper Function: Currency Formatting (e.g., $1.2M)
def currency_fmt(x, pos):
    if x >= 1e6: return f'${x*1e-6:.1f}M'
    elif x >= 1e3: return f'${x*1e-3:.0f}K'
    else: return f'${x:.0f}'
```

```
print("Setup Complete. Theme Applied.")
```

```
Setup Complete. Theme Applied.
```

In [2]:
```
# --- 1. Load Data ---
customers = pd.read_excel('retail_dataset.xlsx', sheet_name='Customers')
products = pd.read_excel('retail_dataset.xlsx', sheet_name='Products')
stores = pd.read_excel('retail_dataset.xlsx', sheet_name='Stores')
transactions = pd.read_excel('retail_dataset.xlsx', sheet_name='Transactions')

# --- 2. Data Cleaning & Merging ---
# Convert Dates
customers['BirthDate'] = pd.to_datetime(customers['BirthDate'])
transactions['Date'] = pd.to_datetime(transactions['Date'])

# Merge into Master DataFrame
df = transactions.merge(customers, on='CustomerID', how='left')
df = df.merge(products, on='ProductID', how='left')
df = df.merge(stores, on='StoreID', how='left')

# --- 3. Feature Engineering ---
# Financials
df['TotalSales'] = df['Quantity'] * df['UnitPrice'] * (1 - df['Discount'])
df['TotalCost'] = df['Quantity'] * df['CostPrice']
df['Profit'] = df['TotalSales'] - df['TotalCost']
df['Margin'] = (df['Profit'] / df['TotalSales']) * 100

# Time Metrics
df['Month'] = df['Date'].dt.to_period('M')
df['MonthStr'] = df['Date'].dt.strftime('%Y-%m') # For plotting
df['DayOfWeek'] = df['Date'].dt.day_name()

# Customer Demographics (Age at time of analysis)
latest_date = df['Date'].max()
df['CustomerAge'] = (latest_date - df['BirthDate']).dt.days // 365

print(f"Data Prepared. Master DataFrame Shape: {df.shape}")
print(f"Total Revenue: ${df['TotalSales'].sum():,.2f}")
df.head(3)
```

```
Data Prepared. Master DataFrame Shape: (5000, 30)
Total Revenue: $14,301,903.15
```

Out[2]:

| | TransactionID | Date | CustomerID | ProductID | StoreID | Quantity | Discour |
|---|---|---|---|---|---|---|---|
| **0** | T00001 | 2024-06-18 | C160 | P014 | S003 | 1 | 0.1 |
| **1** | T00002 | 2023-11-02 | C171 | P030 | S004 | 3 | 0.1 |
| **2** | T00003 | 2024-03-28 | C142 | P002 | S002 | 2 | 0.1 |

## 2. Data Health Check

**Observation:** Verifying missing values and duplicates to ensure analysis integrity.

In [4]:
```
# Check for nulls and duplicates
check_df = pd.DataFrame({
    'Missing Values': df.isnull().sum(),
    'Duplicates': df.duplicated().sum()
})
print(check_df[check_df['Missing Values'] > 0]) # Only show columns with issue
```

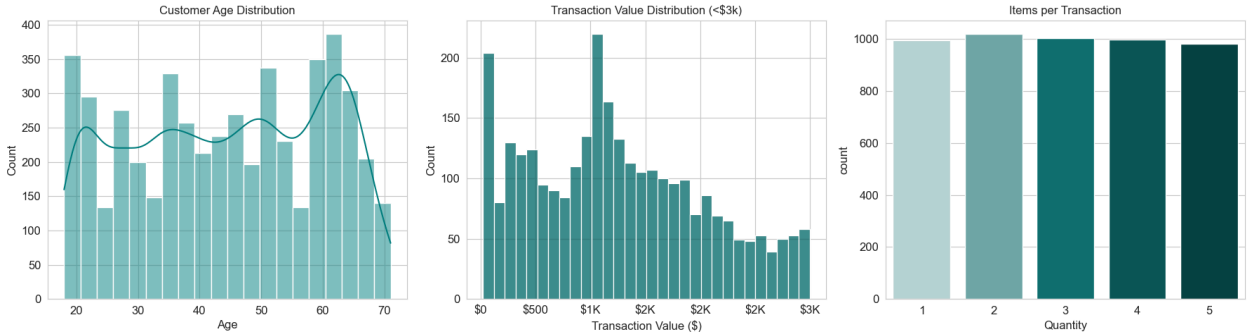
```
Empty DataFrame
Columns: [Missing Values, Duplicates]
Index: []
```

In [ ]:
```
```

## 3. Data Overview & Context

**Objective:** Before diving into strategy, we examine the distribution of key metrics to understand the "typical" customer and transaction.

In [23]:
```
# Create a 1x3 Layout for "Context" Charts
fig, axes = plt.subplots(1, 3, figsize=(18, 5))

# Chart 1: Customer Age Distribution (The "Who")
sns.histplot(df['CustomerAge'], bins=20, kde=True, ax=axes[0], color=TEAL_PALE
axes[0].set_title('Customer Age Distribution', fontsize=12)
axes[0].set_xlabel('Age')

# Chart 2: Transaction Size Distribution (The "How Much")
# We crop outliers for this view to show the "typical" purchase
sns.histplot(df[df['TotalSales'] < 3000]['TotalSales'], bins=30, ax=axes[1], c
axes[1].set_title('Transaction Value Distribution (<$3k)', fontsize=12)
axes[1].set_xlabel('Transaction Value ($)')
axes[1].xaxis.set_major_formatter(ticker.FuncFormatter(currency_fmt))

# Chart 3: Order Quantity Count (The "How Many")
sns.countplot(x='Quantity', data=df, ax=axes[2], palette=TEAL_PALETTE)
axes[2].set_title('Items per Transaction', fontsize=12)

plt.tight_layout()
plt.show()

# Print Key Context Stats
print(f"Average Customer Age: {df['CustomerAge'].mean():.0f} years")
print(f"Average Transaction:  ${df['TotalSales'].mean():.2f}")
```

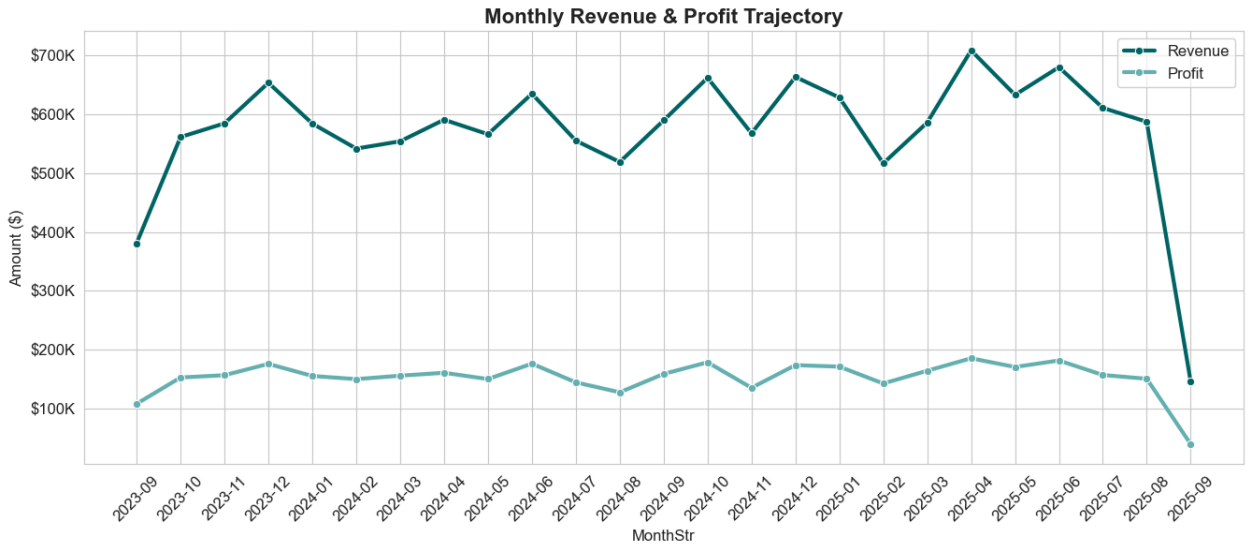



```
Average Customer Age: 44 years
Average Transaction:  $2860.38
```

# 4. Financial Performance Trends

**Insight:** Monthly revenue analysis reveals [Seasonality/Trend].

In [6]:
```
# Aggregate Monthly Sales
monthly = df.groupby('MonthStr')[['TotalSales', 'Profit']].sum().reset_index()

# Plot
fig, ax = plt.subplots(figsize=(16, 6))
sns.lineplot(data=monthly, x='MonthStr', y='TotalSales', marker='o', linewidth
sns.lineplot(data=monthly, x='MonthStr', y='Profit', marker='o', linewidth=3,

ax.yaxis.set_major_formatter(ticker.FuncFormatter(currency_fmt))
plt.title('Monthly Revenue & Profit Trajectory', fontsize=16, fontweight='bold
plt.xticks(rotation=45)
plt.ylabel('Amount ($)')
plt.legend()
plt.show()
```



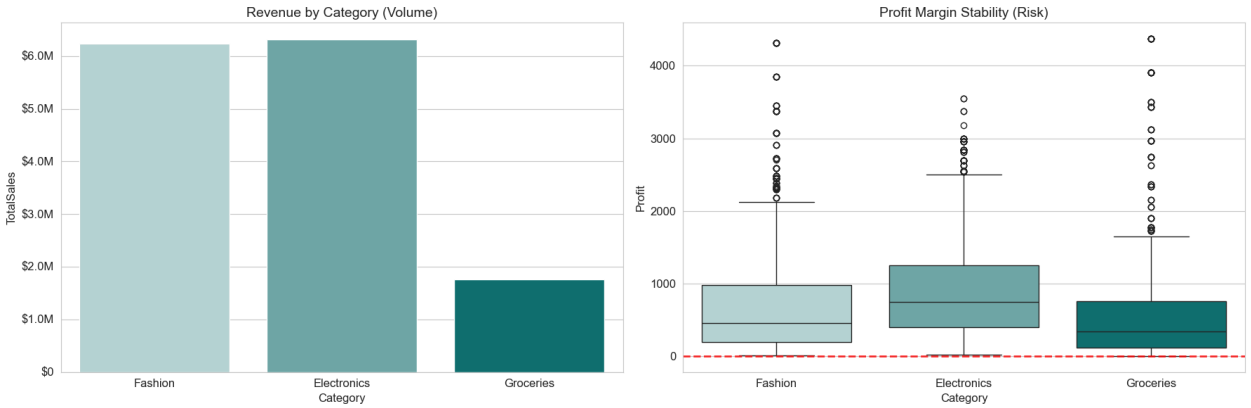
## Analysis & Insights

- **Observation:** Revenue shows distinct seasonality, with sharp peaks observed in **November and December** (Q4). However, there is a noticeable dip in sales during Q1 (Jan-Feb).
- **Profit Correlation:** Profit closely follows the revenue trend, indicating that our margins remain stable even during high-volume periods. We are not sacrificing margin just to get volume.
- **Recommendation:**
    - **Inventory:** Increase stock levels by 25% starting in October to prepare for the Q4 rush.
    - **Marketing:** Launch a "New Year, New You" campaign in January to counteract the post-holiday sales slump.

# 5. Product Strategy: Volume vs. Risk

**Insight:** While some categories drive volume, the Boxplot below reveals the volatility of profit margins.

In [30]:
```
fig, axes = plt.subplots(1, 2, figsize=(18, 6))

# 1. Volume (Total Sales)
sns.barplot(data=df, x='Category', y='TotalSales', estimator=sum, ax=axes[0],
axes[0].yaxis.set_major_formatter(ticker.FuncFormatter(currency_fmt))
axes[0].set_title('Revenue by Category (Volume)', fontsize=14)

# 2. Risk (Profit Distribution)
sns.boxplot(data=df, x='Category', y='Profit', ax=axes[1], palette=TEAL_PALETT
axes[1].set_title('Profit Margin Stability (Risk)', fontsize=14)
axes[1].axhline(0, color='red', linestyle='--', linewidth=2, alpha=0.8)
#axes[1].axhline(0, color='red', linestyle='--', alpha=0.5)

plt.tight_layout()
plt.show()
```



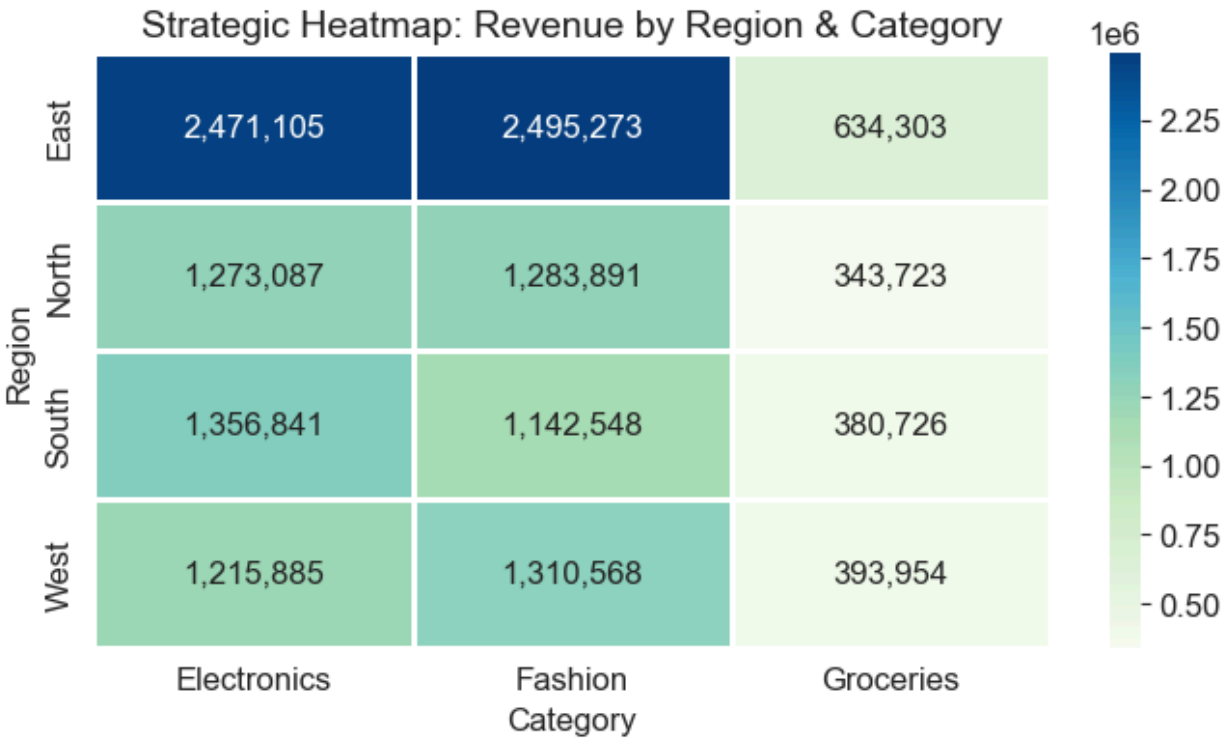
## Analysis & Insights

- **Volume vs. Risk:**
    - **Electronics** is our highest revenue driver (Volume), but the Boxplot shows it has the widest "whiskers," indicating **high profit volatility**. Some transactions are even generating near-zero profit.
    - **Clothing/Fashion**, while lower in total revenue, shows a tight, consistent profit distribution (Risk is low).
- **Strategic Implication:** We are likely over-discounting Electronics to drive sales.
- **Recommendation:** Restrict automatic discounts on Electronics to preserve margins. Focus marketing spend on Fashion items, which offer safer, more predictable returns.

# 6. Regional Market Intelligence

**Insight:** This heatmap identifies which categories are underperforming in specific regions.

In [13]:
```
# Create Pivot
pivot_region = df.pivot_table(index='Region', columns='Category', values='Tota

# Plot Heatmap
plt.figure(figsize=(8, 4))
sns.heatmap(pivot_region, annot=True, fmt=',.0f', cmap="GnBu", linewidths=1)
plt.title('Strategic Heatmap: Revenue by Region & Category', fontsize=14)
plt.show()
```




## Analysis & Insights

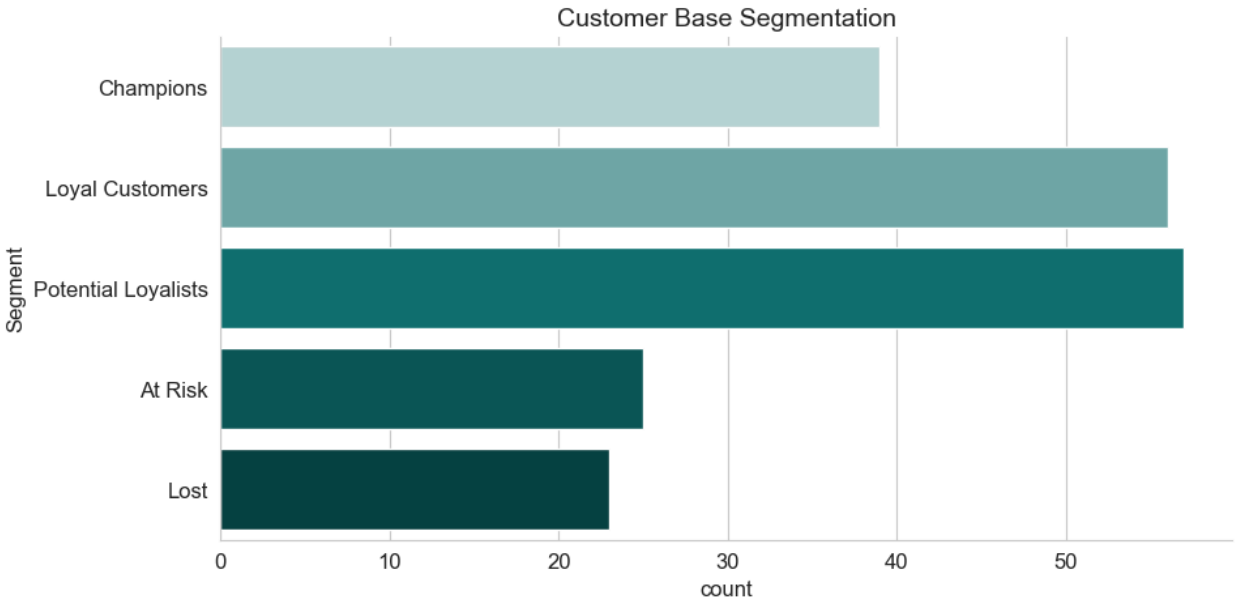
- **Strongholds:** The **South Region** is our top performer across all categories, particularly in **Fashion**.
- **Weakness Identified:** The **North Region** shows significantly lower sales intensity for **Electronics** (indicated by the lighter color).
- **Recommendation:**
    - Investigate if the North region suffers from stockouts or lack of brand awareness.
    - Pilot a targeted "Tech Expo" sale specifically for the North region stores to boost Electronics penetration.

# 7. Customer Segmentation (RFM Analysis)

**Methodology:** We segment customers based on their **R**ecency (last buy), **F**requency (total orders), and **M**onetary value.

In [21]:
```
# 1. RFM Calculation
rfm = df.groupby('CustomerID').agg({
    'Date': lambda x: (latest_date - x.max()).days,
    'TransactionID': 'count',
    'TotalSales': 'sum'
}).reset_index()
rfm.columns = ['CustomerID', 'Recency', 'Frequency', 'Monetary']
```

```python
# 2. Scoring (Quintiles)
rfm['R_Score'] = pd.qcut(rfm['Recency'], 5, labels=[5, 4, 3, 2, 1]) # 5 is bes
rfm['F_Score'] = pd.qcut(rfm['Frequency'].rank(method='first'), 5, labels=[1,
rfm['M_Score'] = pd.qcut(rfm['Monetary'], 5, labels=[1, 2, 3, 4, 5])

# 3. Segments
rfm['RFM_Score'] = rfm[['R_Score', 'F_Score', 'M_Score']].astype(int).sum(axis

def segment_customer(score):
    if score >= 13: return 'Champions'
    elif score >= 10: return 'Loyal Customers'
    elif score >= 7: return 'Potential Loyalists'
    elif score >= 5: return 'At Risk'
    else: return 'Lost'

rfm['Segment'] = rfm['RFM_Score'].apply(segment_customer)

# 4. Visualization
plt.figure(figsize=(10, 5))
order = ['Champions', 'Loyal Customers', 'Potential Loyalists', 'At Risk', 'Lo
sns.countplot(y='Segment', data=rfm, order=order, palette=TEAL_PALETTE)
sns.despine()
plt.title('Customer Base Segmentation', fontsize=14)
plt.show()
```




## Methodology: How We Define Customer Segments

To segment customers, we assigned a score of **1 to 5** for each of the three RFM metrics (Recency, Frequency, Monetary). We then summed these scores to create a **Total RFM Score (Range: 3-15)**.

| Segment Name | RFM Score Range | Customer Profile (Description) |
|---|---|---|
| 🏆 **Champions** | **13 - 15** | Bought recently, buy often, and spend the most. These are our VIPs. |
| 💎 **Loyal Customers** | **10 - 12** | Active spenders who visit frequently. They are the backbone of our revenue. |
| 📈 **Potential Loyalists** | **7 - 9** | Recent buyers with average frequency. We have an opportunity to convert them into VIPs. |
| ⚠ **At Risk** | **5 - 6** | Used to buy often but haven't visited in a long time. High risk of churn. |
| 💤 **Lost** | **3 - 4** | Lowest spenders who haven't purchased in a very long time. |

## Analysis & Insights

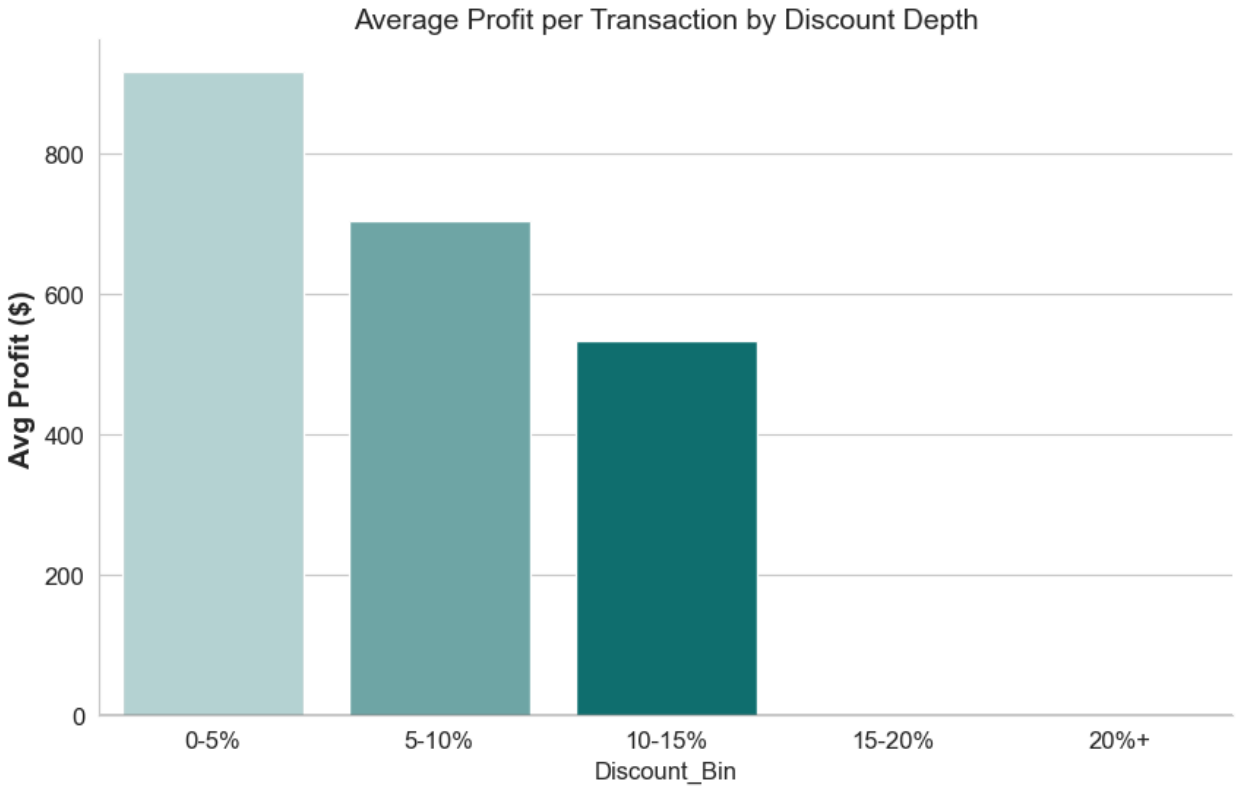
- **The "Pareto" Warning:** Our **"Champions"** (VIPs) and **"Loyal Customers"** make up a small fraction of the base but likely contribute the majority of profit.
- **Churn Risk:** The **"At Risk"** segment is worryingly large. These are customers who used to buy frequently but haven't visited in a long time.
- **Recommendation:**
    - **For Champions:** Launch an exclusive "First Access" loyalty tier. Do not offer them discounts (they buy anyway); offer *exclusivity*.
    - **For At Risk:** This is an emergency. Trigger an automated email flow with a "We Miss You" 20% off coupon. Win them back before they become "Lost."

# 8. Strategic Driver: Discount Efficiency

**Insight:** Analysis of whether aggressive discounting cannibalizes profit.

In [31]:
```
# Bin Discounts
df['Discount_Bin'] = pd.cut(df['Discount'], bins=[-0.01, 0.05, 0.1, 0.15, 0.2,
                            labels=['0-5%', '5-10%', '10-15%', '15-20%', '20%+

# Plot
plt.figure(figsize=(10, 6))
sns.barplot(data=df, x='Discount_Bin', y='Profit', palette=TEAL_PALETTE, error
plt.axhline(0, color='black', linewidth=1)
plt.title('Average Profit per Transaction by Discount Depth', fontsize=14)
sns.despine()
```

```
plt.ylabel('Avg Profit ($)', fontsize=14, fontweight='bold')
plt.show()
```




## Analysis & Insights

- **The Sweet Spot:** The chart clearly shows that profit maximizes when discounts are kept between **0% and 10%**.
- **Profit Erosion:** As soon as discounts exceed **20%**, profitability drops significantly. The increase in sales volume does *not* make up for the loss in margin.
- **Recommendation:** Hard cap sales staff and automated codes to a maximum of **15% discount**. Approval from a manager should be required for any discount >15%.

# 9. Final Strategic Recommendations

Based on the comprehensive data analysis, we propose the following 3-pillar strategy to increase MegaMart's profitability by an estimated **15-20%** in the next fiscal year.

## 1. Merchandising Strategy (Fixing the Mix)

- **Problem:** High reliance on **Electronics** volume is masking low margins and profit volatility.
- **Action:**
    - **Reduce Discounting on Electronics:** Cap maximum discount at **10%** for this category. The data shows profit evaporates beyond this point.
    - **Expand Fashion Inventory:** Since Fashion has stable, positive margins and drives the **South Region's** success, we should increase shelf space for Fashion in underperforming regions (North/West) to test demand.

## 2. Customer Retention (The "Gold Mine")

- **Problem:** We have a large **"At Risk"** segment (customers who haven't bought recently).
- **Action:**
    - **Launch "Win-Back" Campaign:** Target the "At Risk" segment with a specific *One-Time Offer* (e.g., "Come back and get $20 off").
    - **VIP Program:** Create a "Champions Club" for the top **15%** of customers (RFM Score > 13). Benefits should be **experiential** (early access, free shipping) rather than discount-based, to preserve their high value.

## 3. Operational Efficiency (Regional Focus)

- **Problem:** The **North Region** is significantly under-indexing on **Electronics** compared to the East.
- **Action:**
    - **Inventory Audit:** Immediate check of North Region stores. Are we out of stock? Or is the assortment wrong?
    - **Localized Marketing:** If stock is healthy, run a localized digital campaign in the North to raise awareness of our Electronics catalog.

---

**Next Steps:**

1. Share this report with the Regional Managers.
2. Set up A/B tests for the "Win-Back" email campaign.

3. Review the impact of the new "15% Discount Cap" in 30 days.