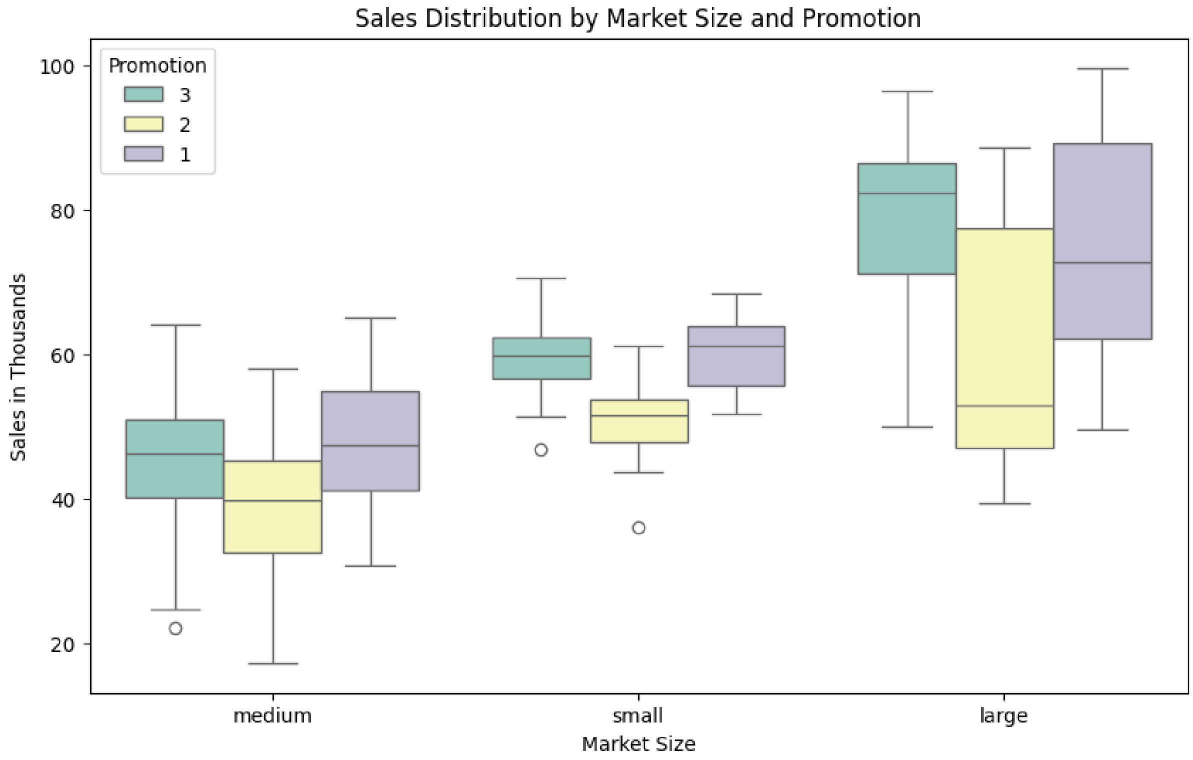
1.Key Components of the Box Plot:



1. **Box (Middle Rectangle)**:
   - Represents the **interquartile range (IQR)** — middle 50% of the data.
   - The bottom of the box = **Q1 (25th percentile)**.
   - The top of the box = **Q3 (75th percentile)**.
   - So, the height of the box shows **how spread out the middle values are**.
2. **Horizontal Line Inside the Box**:
   - This is the **median (50th percentile)** — the middle value of the data.
   - It helps you know whether the data is **skewed or symmetric**.
3. **Whiskers (Lines extending from the box)**:
   - Show the **range of the data**, excluding outliers.
   - They go from the **lowest** to the **highest non-outlier values**.
4. **Circles Outside the Whiskers**:
   - These are **outliers** — unusually high or low values that differ significantly from other observations.

Let’s take **"large market with promotion level 3"**:

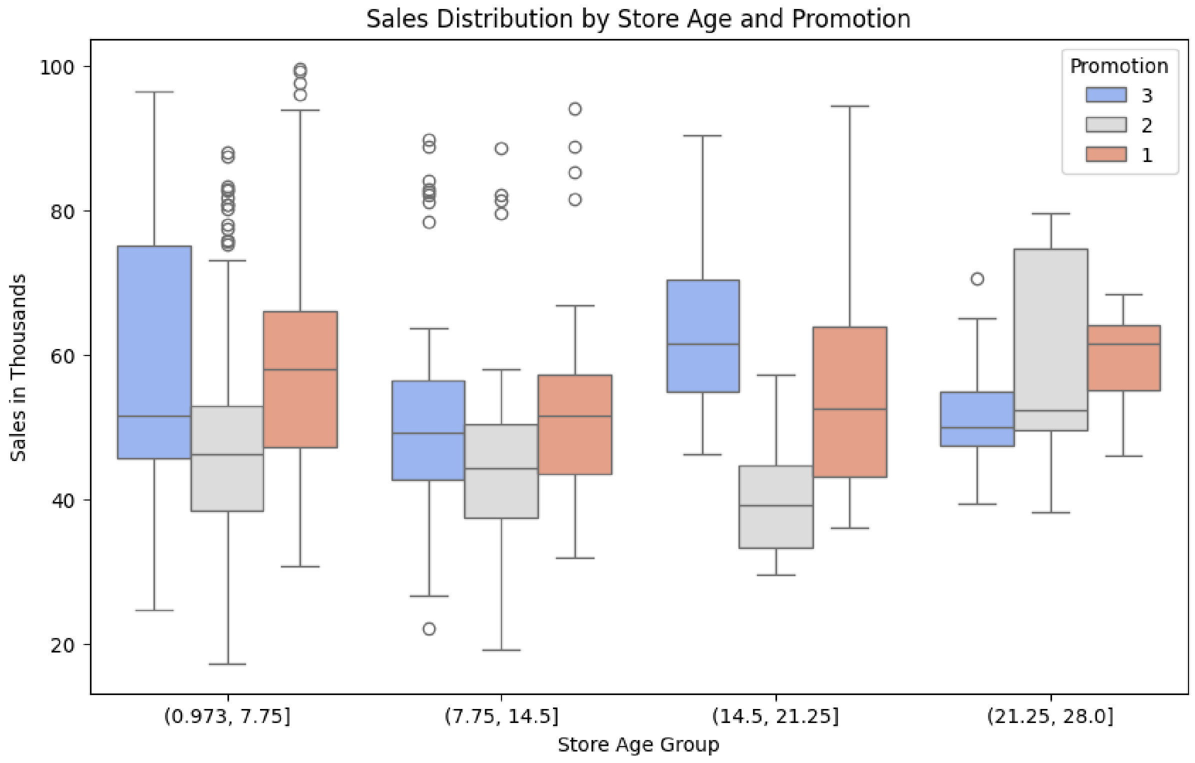
- The **box is high** on the y-axis → this group has **higher overall sales**.
- The **median line is near the top of the box** → most sales are skewed **toward the lower end** of the range.
- The **whiskers are long** → more variability in performance.
- **No outliers** → data is relatively consistent.

Compare that to **"medium market with promotion level 2"**:

- **Low box** → generally lower sales.
- **Outliers present** → a few stores performed much worse or better than the average.

---

2. 🔍 **Step-by-Step: How to Understand the Graph**



✅ **1. Read the Axes**

- **X-axis** → Store Age Group (4 bins from youngest to oldest stores).
- **Y-axis** → Sales (in thousands).
- Each **box plot** shows how sales are distributed within a particular promotion strategy (1, 2, or 3) for that store age group.

✅ **2. Understand the Box Plot Components**

- **Box (middle 50%)** → This is the interquartile range (IQR: Q1 to Q3).
- **Middle line** → Median (Q2), the central tendency of sales.
- **Whiskers** → Range of typical values (usually 1.5 × IQR).
- **Dots beyond whiskers** → Outliers.

✅ 3. Compare Boxes Within Each Store Age Group

Look **vertically** to compare **Promotions 1, 2, and 3** within each store age bin:

- Higher **median line** → Higher central sales value.
- Taller boxes → More **variability**.
- More **outliers** → Unusual performances (either very good or bad).

🧠 How This Helps You Understand Your Summary:

👉 Younger Stores (0.97 – 7.75 years)

- Promotion **3** has high **upper whisker** and **many outliers** → some stores do **really well**, but performance **varies**.
- Promotion **1** is more stable → smaller box, less risk. ✅ From graph: The height and number of outliers help you see this.

👉 Mid-aged Stores (7.75 – 14.5 years)

- Promotion **1** and **3** have **similar medians**, but Promotion **3** has fewer outliers → more **consistent**.
- Promotion **2** has **lowest median**. ✅ From graph: Median lines are nearly equal, but box shape/outliers show difference.

👉 Older Stores (14.5 – 21.25 years)

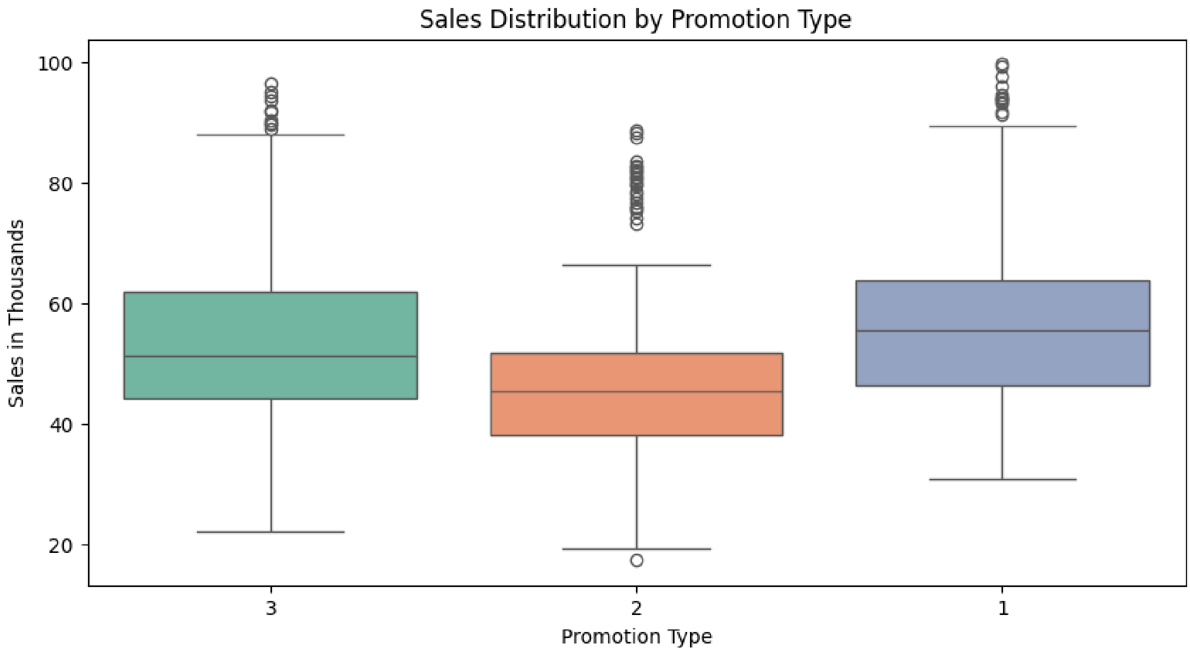
- Promotion **3** has the **highest median** and larger spread → best performer.
- Promotion **2** again has low median → least effective. ✅ From graph: Median of Promotion 3 is clearly higher than others.

👉 Mature Stores (21.25 – 28 years)

- Promotion **1** shows **highest median**.
- Promotion **3** has tighter spread → more consistent.

- Promotion **2** still low. ✅ From graph: Median line of Promotion 1 is highest here, and box of Promotion 3 is smaller.

3.Step 1: Read the Axes



- **X-axis** → Promotion Types (1, 2, 3)
- **Y-axis** → Sales in Thousands

Each box plot summarizes the **distribution of sales** under each promotion type.

📦 Step 2: Know the Box Plot Anatomy

- **Box** = Middle 50% of data (Interquartile Range: Q1–Q3)
- **Line inside the box** = Median (middle sales value)
- **Whiskers** = Range of most values (not including outliers)
- **Dots beyond whiskers** = Outliers (unusual performance)

📈 Step 3: Analyze Each Promotion

🟦 Promotion 1

- **Highest median** → Best average performance
- **Wide box** → High variability in results
- **Many high outliers** → Some stores perform **exceptionally well**

✅ **Conclusion**: High potential, but performance varies store-to-store.

🟧 Promotion 2

- **Lowest median** → Weakest performance overall
- **Narrower spread** → Sales are more consistent (but consistently low)
- **Some low-end outliers** → A few stores did really poorly

⚠️ **Conclusion**: Not very effective. Needs rework or reevaluation.

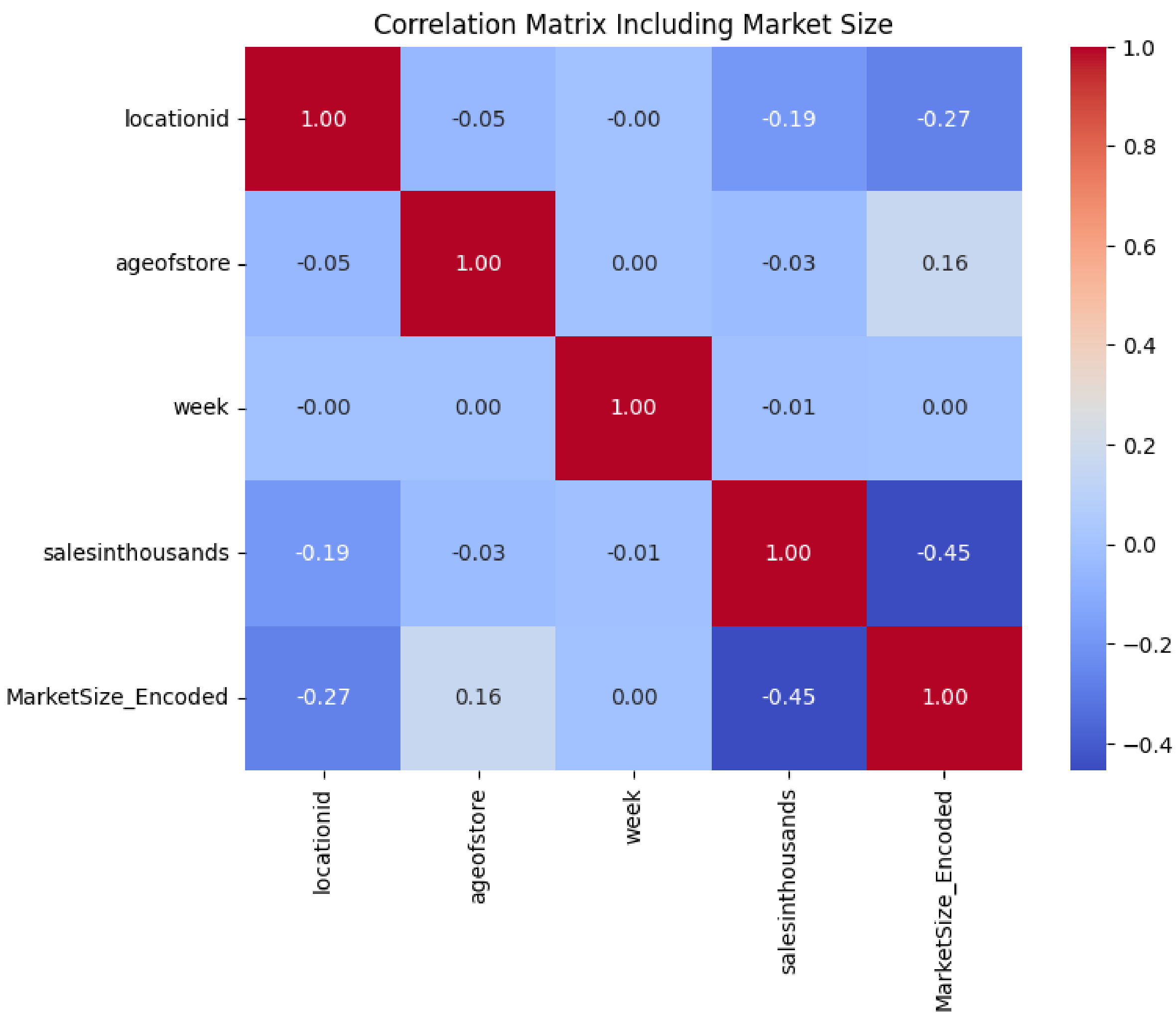
🟩 Promotion 3

- **Median slightly lower than Promotion 1**, but better than Promotion 2
- **Moderate spread** → Balanced performance across stores
- **Fewer outliers** → More consistent results

👍 **Conclusion**: Solid and stable performer.

💡 Business Interpretation

- **Promotion 1**: Best overall but unpredictable. Great for high-risk, high-reward strategy.
- **Promotion 3**: Reliable performer. Best choice if consistency matters.
- **Promotion 2**: Underperforming. Should be improved or replaced.
-



A **correlation matrix heatmap** is a graphical representation of the relationships (correlations) between multiple variables in a dataset. It uses colors to indicate the strength and direction of these relationships.

What is Correlation?

Correlation measures how two variables are related:

- **Positive correlation**: When one variable increases, the other also increases. Values range from 0 to 1.
- **Negative correlation**: When one variable increases, the other decreases. Values range from 0 to -1.
- **No correlation**: When there is no relationship between the two variables. The value is 0.

Understanding the Heatmap in This Image

1. **Matrix Layout**:
    - Each row and column represents a variable (e.g., locationid, ageofstore, etc.).
    - The intersection of a row and column shows the correlation between those two variables.

2. **Diagonal Values**:
    - The diagonal values are always 1 because each variable is perfectly correlated with itself.
3. **Color Scale**:
    - The color bar on the right shows the correlation values:
        - **Red (closer to 1)**: Strong positive correlation.
        - **Blue (closer to -1)**: Strong negative correlation.
        - **Light colors (closer to 0)**: Weak or no correlation.
4. **Key Observations from This Heatmap**:
    - salesinthousands and MarketSize_Encoded: Correlation is **-0.45**, indicating a moderate negative relationship.
    - ageofstore and MarketSize_Encoded: Correlation is **0.16**, showing a weak positive relationship.
    - Most other correlations are close to 0, meaning weak or no relationships between those variables.

Why Use a Correlation Matrix?

- To identify relationships between variables.
- To decide which variables might be useful for analysis or predictive modeling.
- To spot multicollinearity (when two variables are highly correlated).

In this image, the heatmap helps visualize how different factors like store age, location, sales, and market size relate to each other in terms of strength and direction of their relationships.

Observations:

- The diagonal values are all **1.00**, as each variable is perfectly correlated with itself.
- The color scale on the right represents the strength and direction of correlations:
    - Red indicates a positive correlation.
    - Blue indicates a negative correlation.
    - White or light shades indicate weak or no correlation.

Key Correlation Insights:

1. **salesinthousands vs MarketSize_Encoded**:
    - Correlation: **-0.45**
    - Moderate negative correlation, indicating that as market size increases (encoded), sales tend to decrease.
2. **locationid vs MarketSize_Encoded**:

- Correlation: **-0.27**
- Weak negative correlation.

3. **ageofstore vs MarketSize_Encoded**:
   - Correlation: **0.16**
   - Weak positive correlation.
4. Other variables (e.g., week, ageofstore, locationid) show very weak or negligible correlations with one another, as most values are close to 0.

   Purpose:

   This heatmap is useful for identifying relationships between variables, helping in feature selection or understanding data trends for predictive modeling or analysis.