Great! Now that you’ve cleaned and saved the dataset, the next logical step depends on your project goals. However, I recommend starting with **Exploratory Data Analysis (EDA)** before diving into KPIs. Here’s why:

**Why Start with EDA?**

1. **Understand the Data**:
   * EDA helps you uncover patterns, trends, and relationships in the data. This understanding is crucial before defining KPIs.
2. **Identify Key Variables**:
   * EDA helps you identify which variables are most important for your analysis. For example, you might discover that Shipping costs or Defect rates are critical for your supply chain analysis.
3. **Spot Data Issues**:
   * EDA can reveal outliers, missing values, or inconsistencies that you might have missed during data cleaning.
4. **Guide KPI Selection**:
   * The insights from EDA will help you define meaningful KPIs that align with your project goals.

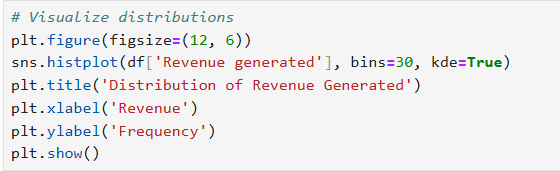
**Next Step: Exploratory Data Analysis (EDA)**

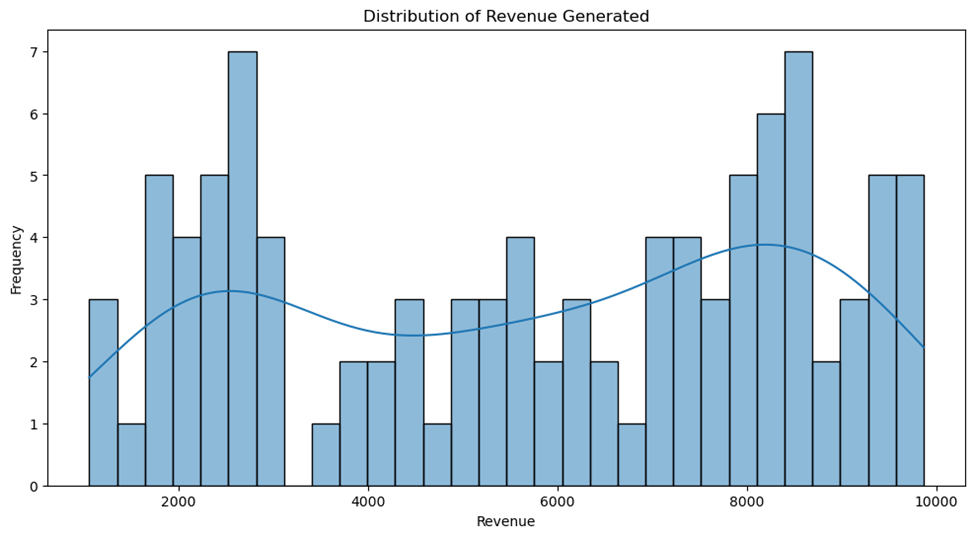
Here’s a step-by-step guide to performing EDA on your cleaned dataset:

**Univariate Analysis**

Analyze individual columns to understand their distribution and summary statistics.

1. **Revenue generated**





**Results:**

* The histogram shows the distribution of revenue generated, with the x-axis representing revenue values and the y-axis representing the frequency of those values.
* The Kernel Density Estimate (KDE) curve (the smooth line) helps visualize the shape of the distribution.
* **Insight:** The revenue distribution appears to be roughly normally distributed (bell-shaped), with most revenue values clustered around the mean (approximately $5,776). There are fewer instances of very low or very high revenue values.
* The revenue values are centered around the mean (~5,776),with most values falling within arrange of approximately 5,776),with most values falling within arrange of approximately 2,000 to $8,000.
* The distribution is unimodal (one peak) and roughly symmetric, indicating that most products generate moderate revenue, with fewer products generating very low or very high revenue.
* The KDE curve confirms that the distribution is close to normal, which is useful for statistical modeling and hypothesis testing.

### Revenue Calculation Formula:

Revenue = Order Quantity×Unit Price

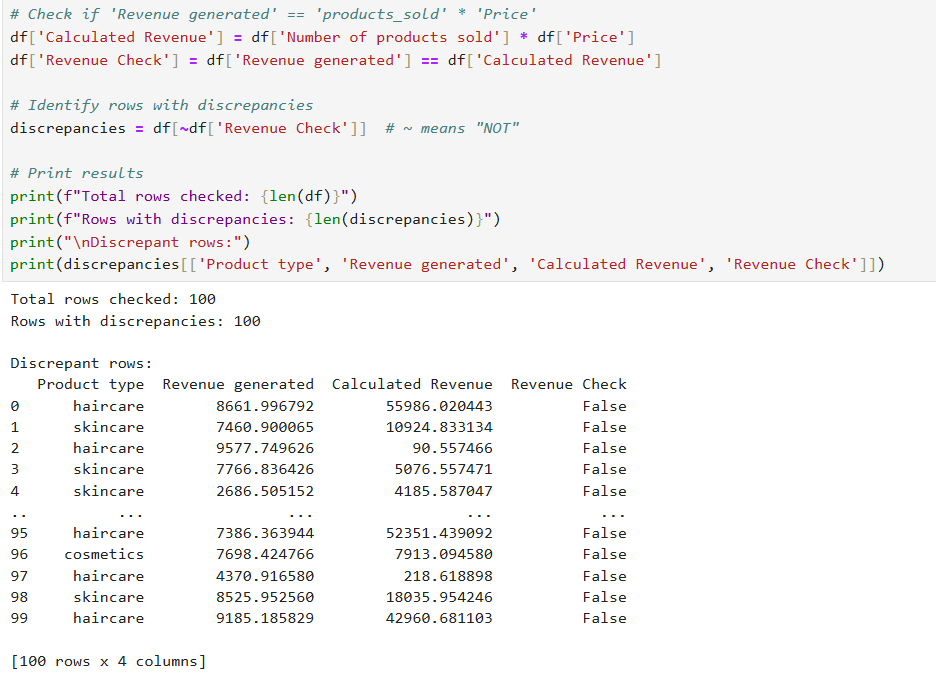
Profit Calculation Formula:

Profit=Revenue−Cost

So, in your dataset, the Revenue generated column should ideally follow:

Revenue generated = Number of products sold × Price

If your calculated revenue does not match the dataset’s Revenue generated column, it suggests additional factors affecting the recorded revenue (e.g., discounts, taxes, refunds, or incorrect data).



**Explanation:**

1. **Calculated Revenue:**

* Creates a new column Calculated Revenue by multiplying products\_sold by Price.

1. **Revenue Check:**

* Compares Revenue generated with Calculated Revenue and stores the result as True (matches) or False (mismatches).

1. **Discrepancy Analysis:**

* Filters and displays rows where the values do not match.
* Outputs the count of discrepancies and a detailed view of problematic rows.

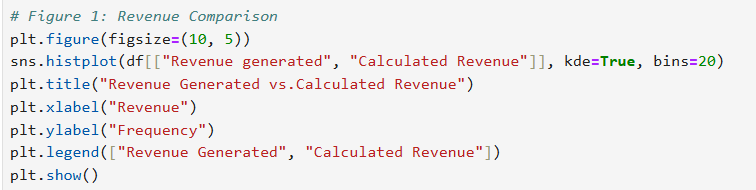
1. **Export:**

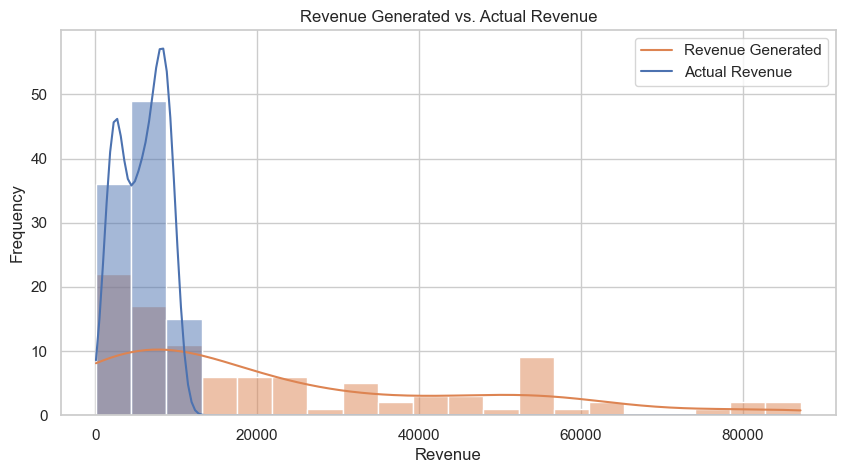
* Optionally saves discrepant rows to a CSV file for further investigation.

Create calculated column for revenue difference



Compare both revenues



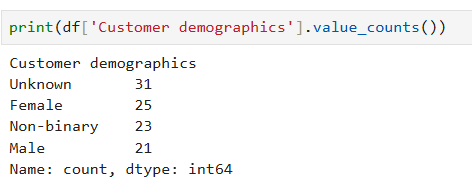


**Revenue Generated vs. Calculated Revenue (Histogram & KDE Plot)**

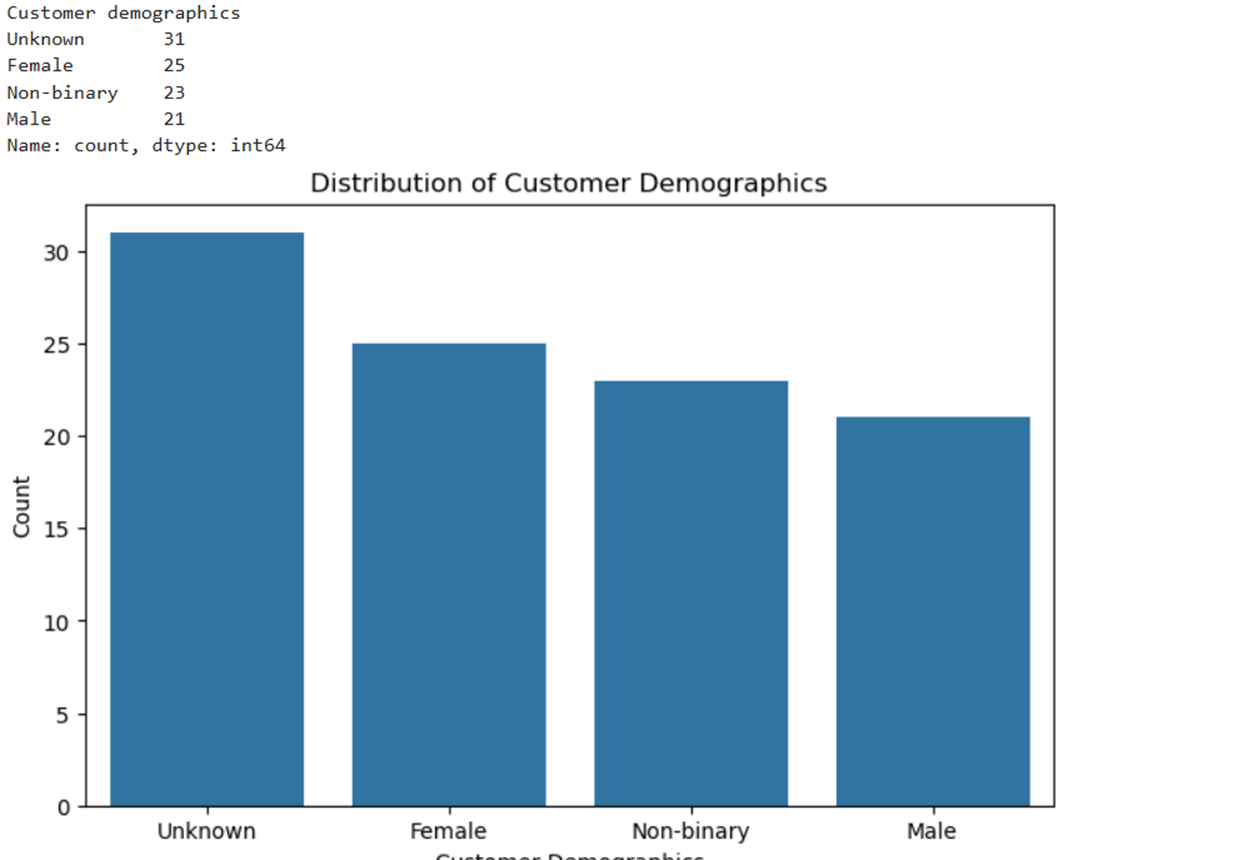
* The "Calculated Revenue" distribution (blue) is heavily concentrated towards lower revenue values, while the "Revenue Generated" (orange) is more spread out.
* This suggests a possible discrepancy between expected and actual revenue. The actual revenue has a tighter, more concentrated range, indicating that sales were not as widely distributed as expected.

Normally we will go with the original data revenue Generated as the primary metric

1. **Customer demographics**

****

****

****

**Here’s my analysis and recommendation:**

**Analysis of the Distribution**

1. **Unknown (31):**
   * This is the most frequent value in the column. While it’s not ideal, it’s not overwhelmingly dominant (e.g., 80% of the data).
2. **Female (25), Non-binary (23), Male (21):**
   * These values are fairly evenly distributed, which means the column still contains meaningful data for analysis.

**Recommendation**

* **Keep the Column:**
  + Since the column has a relatively balanced distribution of meaningful values (Female, Non-binary, Male), it can still provide useful insights for customer segmentation and sales analysis.
  + Even though "Unknown" is the most frequent value, the other categories are well-represented and can still be analyzed.

However if you decided to deal with the unknown values, you can proceed as follow:

**How to Handle "Unknown" Values**

If you decide to keep the column, you can handle the "Unknown" values in one of the following ways:

**Option 1: Keep "Unknown" as a Separate Category**

* Treat "Unknown" as a valid category in your analysis. For example:
  + Compare sales or preferences across all categories, including "Unknown."
  + This approach is useful if you want to analyze the impact of unknown demographics on your supply chain.

**Option 2: Replace "Unknown" with the Mode (Most Frequent Known Value)**

* Replace "Unknown" with the most frequent known value (e.g., "Male" in this case).
* This approach is useful if you want to minimize the impact of missing data.

**Option 3: Drop Rows with "Unknown" Values**

* If "Unknown" values are not useful for your analysis, you can drop those rows. However, this will reduce your dataset size.

**Next Steps**

1. Decide how to handle the "Unknown" values (keep, replace, or drop).
2. Proceed with your analysis, using the Customer demographics column for customer segmentation and sales analysis.

**Example Use Cases for the Column**

Here are some ways you can use the Customer demographics column in your analysis:

1. **Sales by Demographics**:
   * Analyze which products are more popular among females, males, or non-binary customers.
   * Example: Are skincare products more popular among females?
2. **Revenue by Demographics**:
   * Compare revenue generated by different customer groups.
   * Example: Do male customers generate more revenue for haircare products?
3. **Marketing Strategies**:
   * Use the insights to design targeted marketing campaigns for specific customer groups.

To decide whether to **replace** or **drop** the "Unknown" values in the Customer demographics column, we can analyze whether the "Unknown" group behaves similarly to any of the known groups (e.g., Female, Male, Non-binary). Specifically, we can compare the **product preferences** and **purchasing behavior** of the "Unknown" group with the other groups.

Here’s how we can do it step by step:

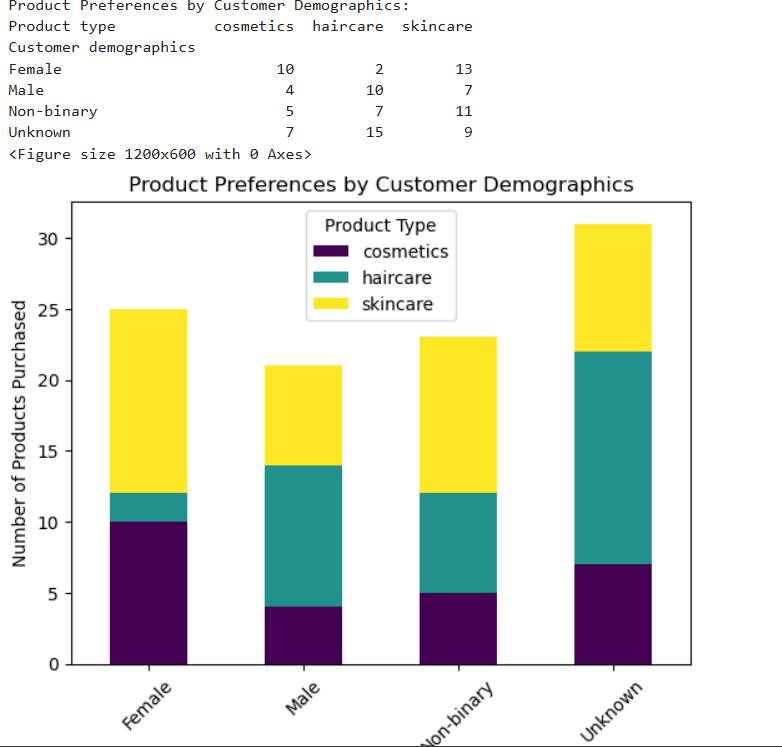
**Step 1: Analyze Product Preferences by Customer Demographics**

We’ll compare the **product types** purchased by the "Unknown" group with those purchased by the Female, Male, and Non-binary groups. If the "Unknown" group’s product preferences are similar to one of the known groups, we can replace "Unknown" with that group. If the preferences are random or unique, we should drop the "Unknown" rows.

**Step 2: Code to Compare Product Preferences**

Here’s the code to compare product preferences by customer demographics:





**Analysis of Product Preferences**

1. **Female**:
   * Prefers **skincare** (13) and **cosmetics** (10), with very few purchases of **haircare** (2).
2. **Male**:
   * Prefers **haircare** (10) and **skincare** (7), with fewer purchases of **cosmetics** (4).
3. **Non-binary**:
   * Prefers **skincare** (11) and **haircare** (7), with fewer purchases of **cosmetics** (5).
4. **Unknown**:
   * Prefers **haircare** (15) and **skincare** (9), with fewer purchases of **cosmetics** (7).

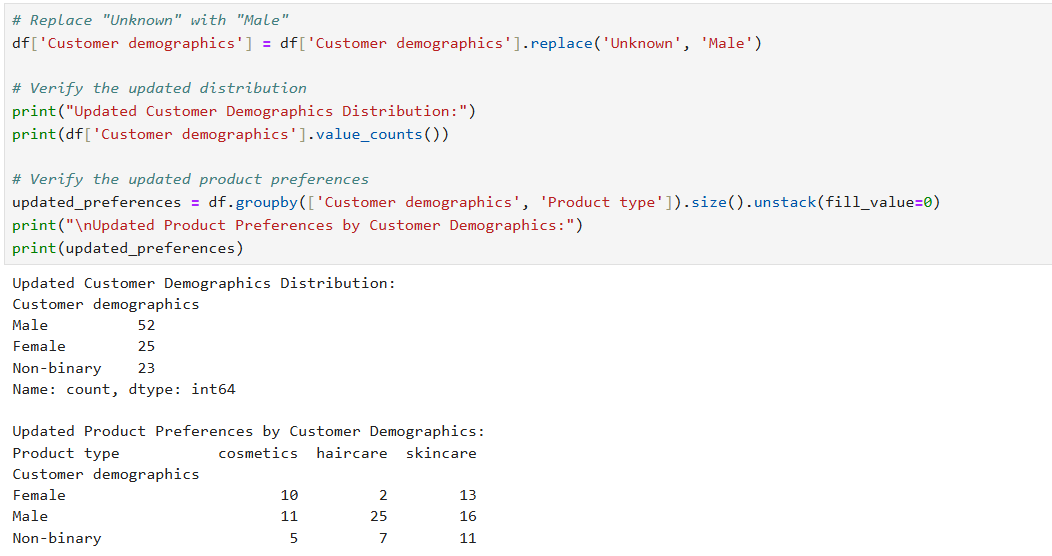
**Key Observations**

* The **Unknown** group’s preferences are **closest to the Male group**:
  + Both groups prefer **haircare** over other products.
  + The **Unknown** group also has a significant preference for **skincare**, similar to the **Non-binary** group.
* The **Unknown** group’s preferences are **not similar to the Female group**, which prefers **skincare** and **cosmetics**.

**Recommendation**

* **Replace "Unknown" with "Male"**:
  + Since the "Unknown" group’s preferences are closest to the **Male** group (both prefer haircare), you can replace "Unknown" with "Male."
  + This will allow you to retain the data while making it useful for analysis.

If you decided to do this, here is the code and the expected result



**Why This Approach?**

* **Retains Data**: You don’t lose rows by dropping "Unknown" values.
* **Improves Analysis**: By replacing "Unknown" with "Male," you can include this data in your analysis, especially for customer segmentation and product preference analysis.
* **Logical Decision**: The "Unknown" group’s preferences align most closely with the Male group, making this replacement logical.

**Analyze Non-binary Preferences**

From the updated product preferences:

* **Non-binary** customers prefer:
  + **Skincare (11)**
  + **Haircare (7)**
  + **Cosmetics (5)**

Compared to:

* **Male** customers:
  + **Haircare (25)**
  + **Skincare (16)**
  + **Cosmetics (11)**
* **Female** customers:
  + **Skincare (13)**
  + **Cosmetics (10)**
  + **Haircare (2)**

**Step 2: Decision to Replace Non-binary**

If you want to **remove the "Non-binary" category**, you have two options:

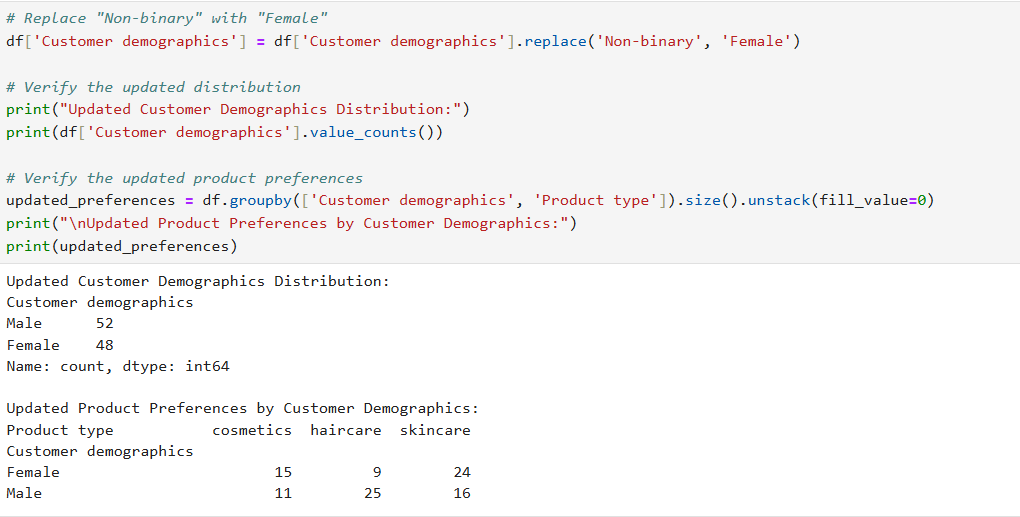
1. **Replace Non-binary with Male**:
   * This aligns with their preference for **haircare** and **skincare**, which is similar to **Male**.
2. **Replace Non-binary with Female**:
   * This aligns with their preference for **skincare**, which is similar to **Female**.

**Recommendation:**

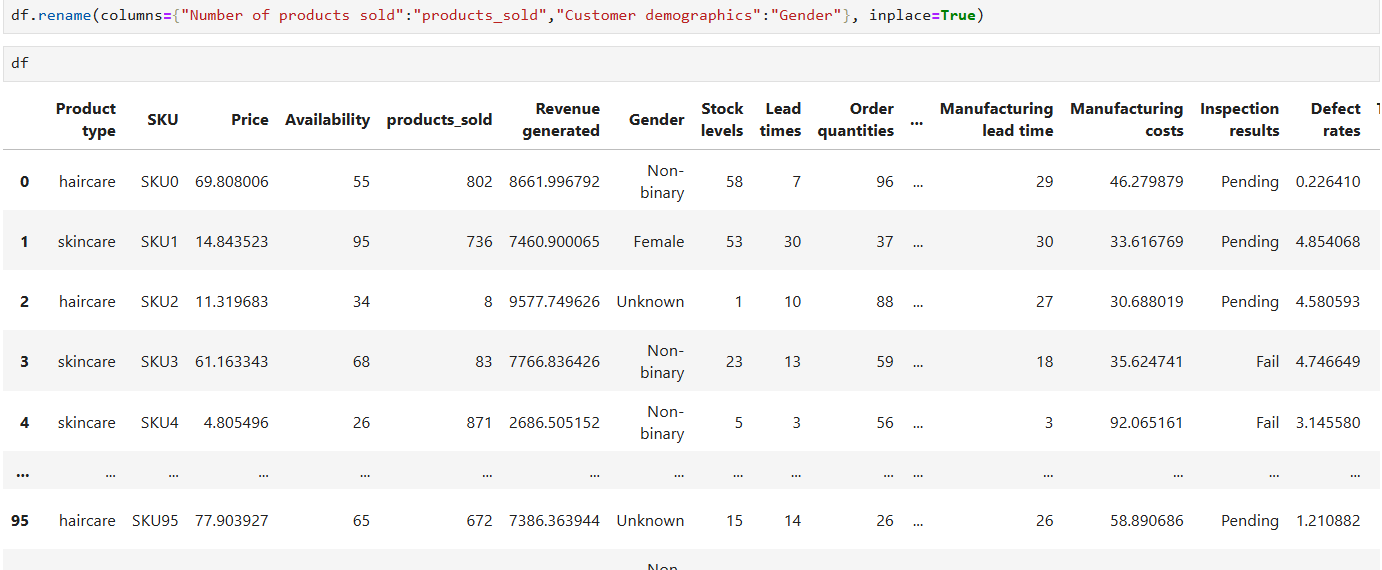
* Since the **Non-binary** group’s preferences are **closer to Male**, it would make more sense to replace **Non-binary** with **Male**.
* However, if you have a specific reason to prefer **Female** (e.g., business context or customer segmentation goals), you can replace **Non-binary** with **Female**.

**Code to Replace Non-binary with Female**

If you decide to replace **Non-binary** with **Female**, here’s the code:



Replace column names of number of products sold with products sold only and Customer demographics with Gender



**Recap of Your EDA Steps:**

1. **Summary Statistics for Numerical Columns**:
   * You used df.describe() to get summary statistics for numerical columns (e.g., mean, standard deviation, min, max, etc.).
   * This helped you understand the central tendency, spread, and range of numerical variables like Revenue generated, Price, Stock levels, etc.
2. **Distribution of Categorical Columns**:
   * You used value\_counts() to analyze the distribution of categorical columns like Product type, Shipping carriers, and Customer demographics.
   * This helped you understand the frequency of different categories (e.g., how many products belong to each product type).
3. **Visualization of Revenue Distribution**:
   * You created a histogram with a Kernel Density Estimate (KDE) curve to visualize the distribution of Revenue generated.
   * This helped you understand the shape and spread of the revenue data.

**Insights from the Revenue Distribution:**

1. **Shape of the Distribution**:
   * The histogram and KDE curve show that the revenue distribution is **roughly normal (bell-shaped)**.
   * This is a useful property for statistical modeling and hypothesis testing.
2. **Central Tendency**:
   * Most revenue values are clustered around the **mean (~$5,776)**.
   * This indicates that the majority of products generate moderate revenue.
3. **Spread of the Data**:
   * Most revenue values fall within the range of **2,000to2,000*to*8,000**.
   * There are fewer instances of very low or very high revenue values.
4. **Unimodal Distribution**:
   * The distribution has **one peak**, indicating that there is a single dominant range of revenue values.
5. **Symmetry**:
   * The distribution is **roughly symmetric**, meaning the data is evenly distributed around the mean.

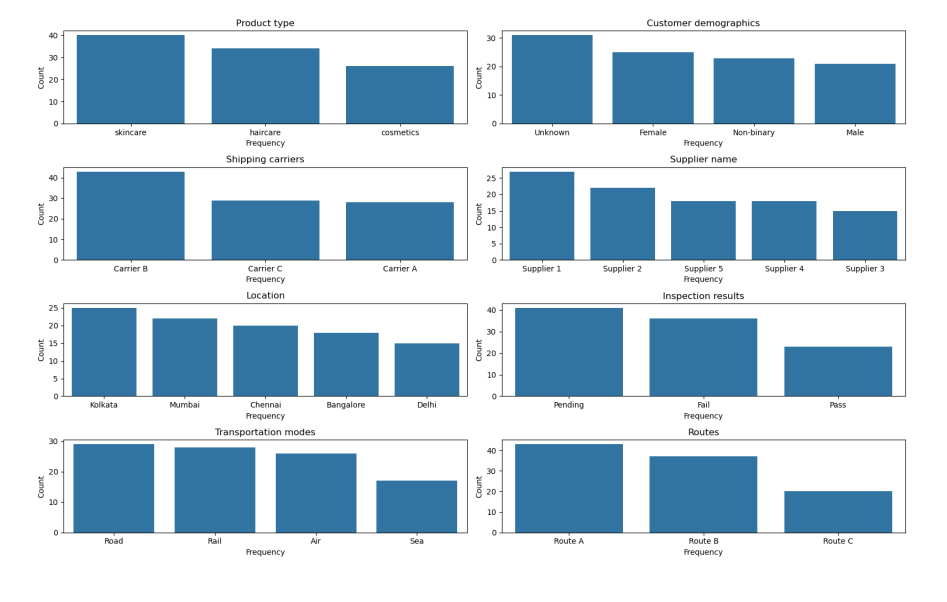
**Next Steps for EDA:**

Now that you’ve analyzed the distribution of Revenue generated, you can proceed with further EDA to explore other aspects of the dataset. Here are some suggestions:

**1. Analyze Other Numerical Columns:**

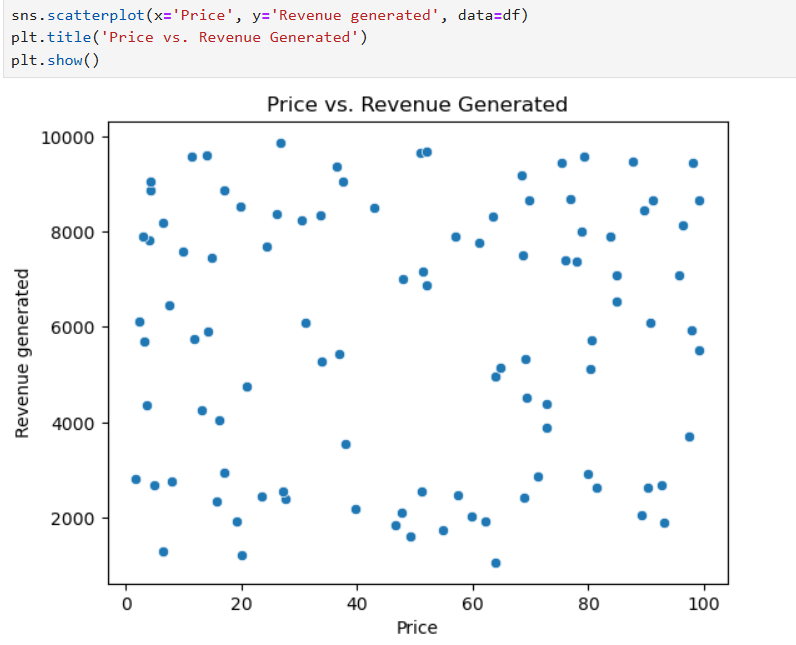
* Visualize the distribution of other numerical columns like Price, Shipping costs, Manufacturing costs, etc.
* Use histograms, box plots, or KDE plots to understand their distributions.





**. Explore Relationships Between Variables:**

* Use scatter plots or correlation matrices to explore relationships between variables (e.g., Price vs. Revenue generated, Shipping costs vs. Revenue generated).
* Example:



**Observations from the Scatter Plot**

From the scatter plot, we can make the following observations:

1. **General Trend**:
   * There doesn’t appear to be a **strong linear relationship** between **Price** and **Revenue Generated**.
   * Products with higher prices don’t necessarily generate significantly higher revenue, and vice versa.
2. **Clustering**:
   * Most of the data points are clustered in the **lower to mid-range** of both **Price** and **Revenue Generated**.
   * This suggests that the majority of products have **moderate prices** and generate **moderate revenue**.
3. **Outliers**:
   * There may be a few outliers where:
     + **High-priced products** generate **low revenue**.
     + **Low-priced products** generate **high revenue**.
   * These outliers could represent unique cases that warrant further investigation.
4. **No Clear Correlation**:
   * The scatter plot does not show a clear **positive or negative correlation** between **Price** and **Revenue Generated**.
   * This indicates that **price alone** may not be a strong predictor of revenue.

**Insights**

1. **Revenue is Not Solely Driven by Price**:
   * The lack of a strong correlation suggests that other factors (e.g., **product type**, **customer demographics**, **marketing efforts**) may have a more significant impact on revenue.
2. **Moderate Pricing Strategy**:
   * The clustering of data points in the **moderate price and revenue range** suggests that a **moderate pricing strategy** may be effective for most products.
3. **Outliers**:
   * The outliers (e.g., high-priced products with low revenue or low-priced products with high revenue) could represent opportunities for further analysis:
     + Why do some high-priced products generate low revenue? (e.g., low demand, poor marketing, or high competition).
     + Why do some low-priced products generate high revenue? (e.g., high demand, effective marketing, or strong customer loyalty).

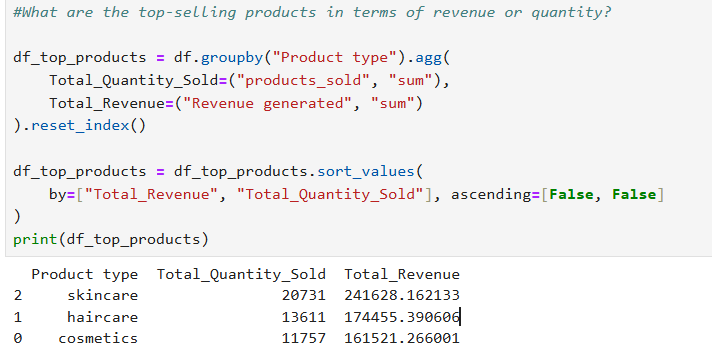
# Product type and Price

# analyzing the Supply Chain by looking at the relationship between the price of the products and the revenue generated by them:



The company derives more revenue from skincare products, and the higher the price of skincare products, the more revenue they generate.

What are the top-selling products in terms of revenue or quantity?





**Observations from the following Plots**

From the plots, we can make the following observations:

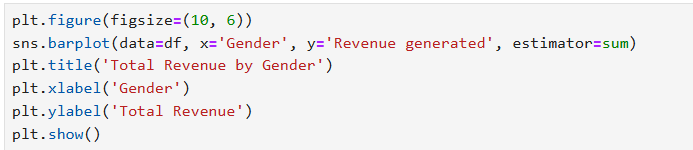
1. **Product Type Distribution**:
   * The dataset contains three product types: **haircare**, **skincare**, and **cosmetics**.
   * The distribution of product types is **not uniform**, meaning some product types are more common than others.
2. **Most Common Product Type**:
   * **Skincare** appears to be the **most common** product type, with the highest count in the dataset.
   * This suggests that skincare products are either more popular or more frequently produced in your supply chain.
3. **Least Common Product Type**:
   * **Cosmetics** appears to be the **least common** product type, with the lowest count in the dataset.
   * This could indicate that cosmetics are either less popular or less frequently produced.
4. **Haircare Products**:
   * **Haircare** products fall somewhere in the middle, with a moderate count compared to skincare and cosmetics.

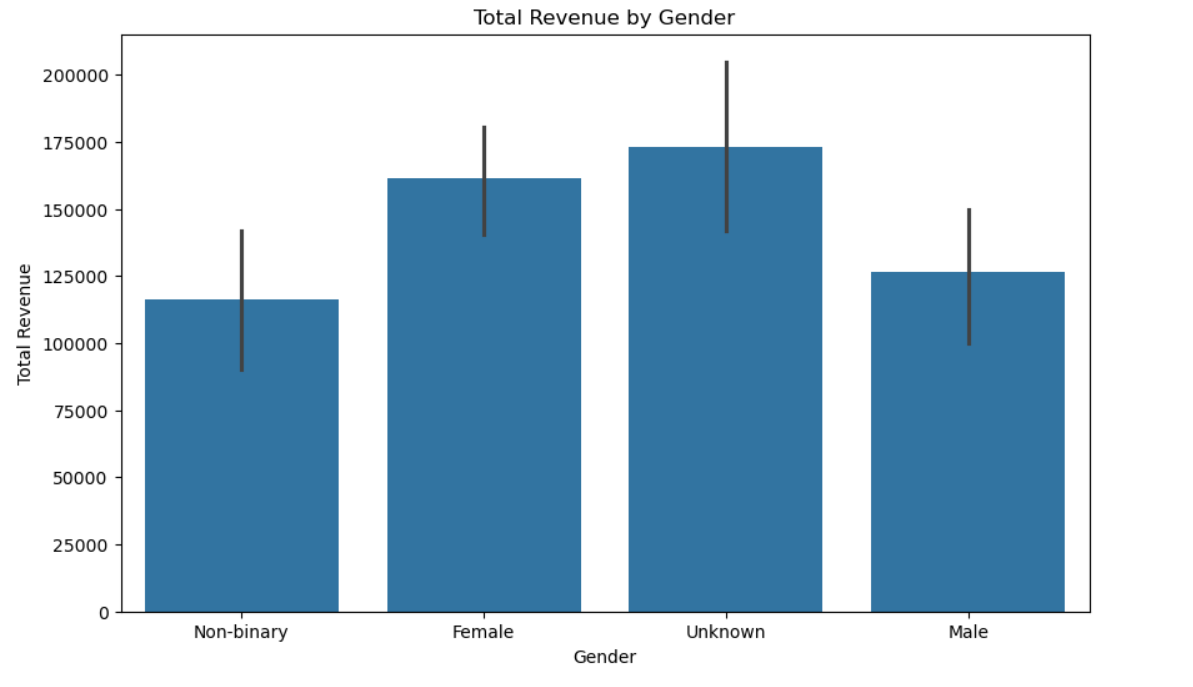
**Insights**

1. **Skincare Dominance**:
   * The high frequency of **skincare** products suggests that this category is a **key focus area** in your supply chain.
   * You may want to investigate why skincare products are more common (e.g., higher demand, better profitability, or easier production).
2. **Cosmetics Underrepresentation**:
   * The low frequency of **cosmetics** products could indicate an opportunity for growth in this category.
   * You may want to explore whether there is untapped demand for cosmetics or if production constraints are limiting this category.
3. **Haircare Products**:
   * The moderate frequency of **haircare** products suggests that this category is stable but not as dominant as skincare.
   * You may want to analyze the performance of haircare products to identify opportunities for improvement.

**Analyze Revenue by Gender**:

· Compare revenue generated by **Gender**, and other customer groups.





**Correct Observations from the Bar Plot**

From the bar plot, we can make the following observations:

1. **Total Revenue by Customer Group**:
   * The bar plot shows the **total revenue** generated by each customer group (unknown, non binary, Female and Male).
   * The height of each bar represents the **sum of revenue** for that group.
2. **unknown Customers**:
   * unknown customers generate the **highest total revenue**, significantly outperforming all customers.
   * This suggests that unknown customers are a **key driver of revenue** in your supply chain.
   * If you decided to replace the unknown with male as previously recommended then Males will be the highest column
3. **Female Customers**:
   * Female customers generate **less total revenue** compared to unknown customers.
4. **Non binary customers:**
   * They generate the lowest revenue

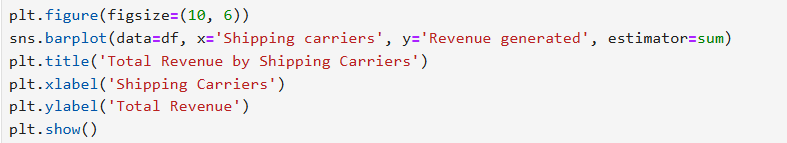
**Correct Insights**

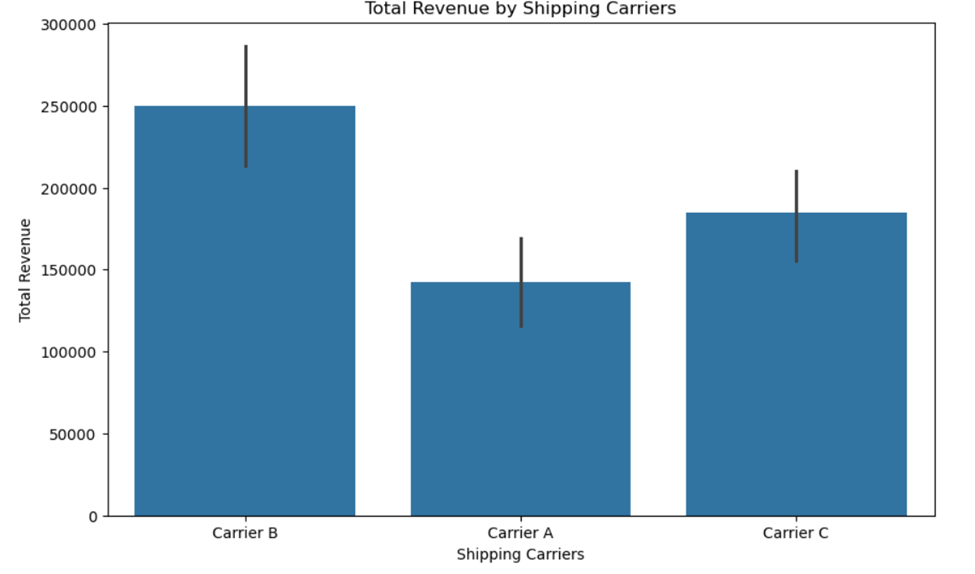
1. **unknown Customers are the Top Revenue Contributors**:
   * Male customers generate the **majority of revenue**, making them a critical target audience for marketing and sales strategies.
   * You may want to investigate why unknown customers generate more revenue (e.g., higher purchasing power, stronger brand loyalty, or better product alignment).
2. **Female Customers Moderate**:
   * Female customers generate **significantly less revenue**, which could indicate untapped potential or mismatched product offerings.
   * You may want to explore ways to increase revenue from Female customers (e.g., targeted marketing, product customization, or pricing strategies).
3. **Customer Segmentation**:
   * The significant difference in revenue between gender customers highlights the importance of **customer segmentation**.
   * Tailoring strategies to each customer group could help maximize revenue and improve customer satisfaction.
   * It is recommended to replace unknown with males and non-binary with females

**Analyze Revenue by Other Categories**:

· Compare revenue generated by **Product type**, **Shipping carriers**, or **Supplier name**.

· Example:





**Observations from the Bar Plot**

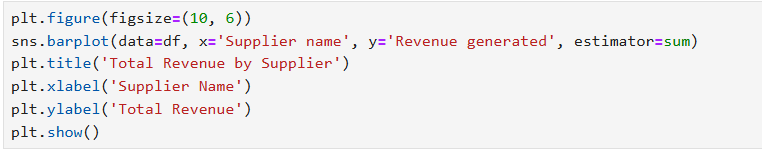
From the bar plot, we can make the following observations:

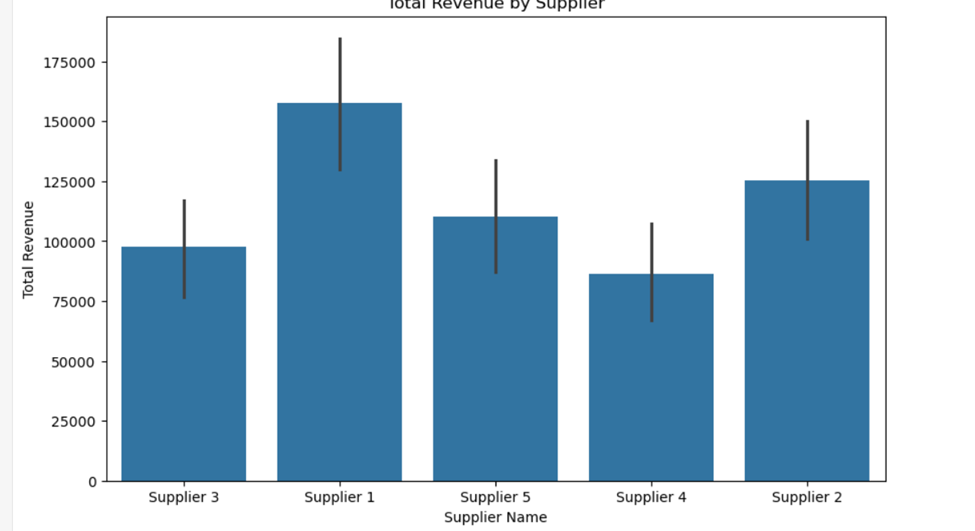
1. **Total Revenue by Shipping Carrier**:
   * The bar plot shows the **total revenue** generated for each shipping carrier (Carrier A, Carrier B, Carrier C).
   * The height of each bar represents the **sum of revenue** for that carrier.
2. **Carrier B**:
   * Carrier B generates the **highest total revenue**, significantly outperforming the other carriers.
   * This suggests that Carrier B is the **most effective** in terms of revenue generation.
3. **Carrier A**:
   * Carrier A generates **moderate total revenue**, falling between Carrier B and Carrier C.
   * This indicates that Carrier A is a **reliable but less dominant** option.
4. **Carrier C**:
   * Carrier C generates the **lowest total revenue**, suggesting it’s the **least effective** in terms of revenue generation.

**Insights**

1. **Carrier B is the Top Performer**:
   * Carrier B generates the **highest total revenue**, making it a critical part of your supply chain.
   * You may want to investigate why Carrier B is so effective (e.g., faster shipping times, better customer satisfaction, or lower costs).
2. **Carrier A is Consistent**:
   * Carrier A generates **moderate revenue**, indicating it’s a **stable but less impactful** option.
   * You may want to explore ways to improve its performance (e.g., optimizing routes or reducing costs).
3. **Carrier C Underperforms**:
   * Carrier C generates the **lowest total revenue**, which could indicate inefficiencies or mismatched services.
   * You may want to investigate why Carrier C underperforms and consider alternatives if necessary.
4. **Shipping Carrier Strategy**:
   * The significant difference in revenue across carriers highlights the importance of **optimizing your shipping strategy**.
   * You may want to allocate more resources to **Carrier B** and explore ways to improve the performance of **Carrier A** and **Carrier C**.

Compare revenue generated by **Supplier name**

****

****

**Observations from the Bar Plot**

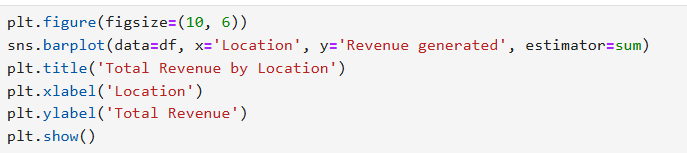
**From the bar plot, we can make the following observations:**

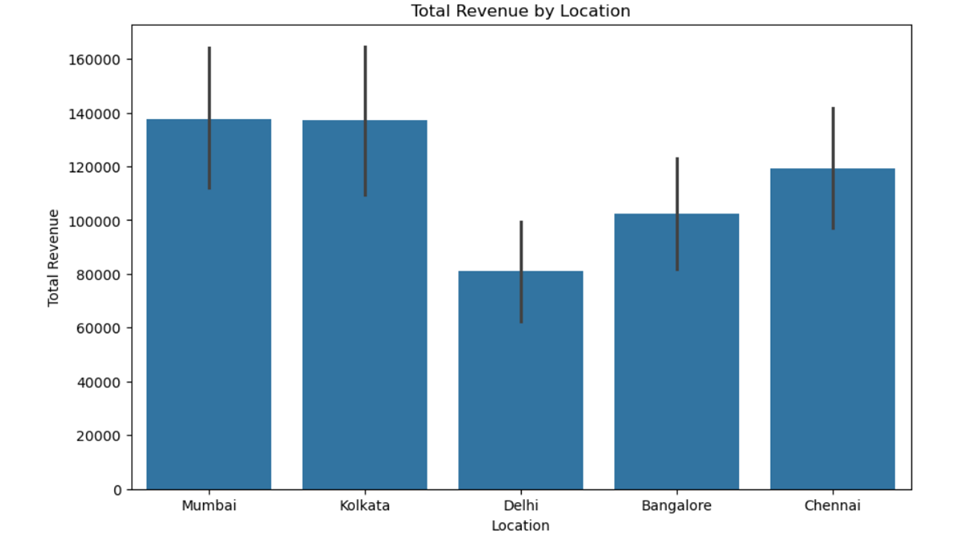
1. **Total R**evenue by Supplier:
   * The bar plot shows the total revenue generated for each supplier (Supplier 1, Supplier 2, Supplier 3, Supplier 4, Supplier 5).
   * The height of each bar represents the sum of revenue for that supplier.
2. **Supplier 1:**
   * Supplier 1 generates the highest total revenue, significantly outperforming the other suppliers.
   * This suggests that Supplier 1 is the most effective in terms of revenue generation.
3. **Supplier 2:**
   * Supplier 2 generates moderate total revenue, falling between Supplier 1 and the other suppliers.
   * This indicates that Supplier 1 is a reliable but less dominant option.
4. **Supplier 5, Supplier 4, and Supplier 3:**
   * These suppliers generate lower total revenue compared to Supplier 1 and Supplier 2.
   * Supplier 5 performs slightly better than Supplier 4 and Supplier 3, but all three are underperforming compared to the top suppliers.

**Insights**

1. **Supplier 1 is the Top Performer:**
   * Supplier 1 generates the highest total revenue, making it a critical part of your supply chain.
   * You may want to investigate why Supplier 1 is so effective (e.g., better product quality, faster delivery times, or competitive pricing).
2. **Supplier 2 is Consistent:**
   * Supplier 2 generates moderate revenue, indicating it’s a stable but less impactful option.
   * You may want to explore ways to improve its performance (e.g., optimizing processes or expanding product offerings).
3. **Suppliers 5, 4, and 3 Underperform:**
   * These suppliers generate significantly less revenue, which could indicate inefficiencies or mismatched services.
   * You may want to investigate why these suppliers underperform and consider alternatives if necessary.
4. **Supplier Strategy:**
   * The significant difference in revenue across suppliers highlights the importance of optimizing your supplier strategy.
   * You may want to allocate more resources to Supplier 1 and explore ways to improve the performance of Supplier 2, Supplier 5, Supplier 4, and Supplier 3.

**Compare revenue generated by Location**

****

****

**Observations from the Bar Plot**

**From the bar plot, we can make the following observations:**

### Observations:

1. **Mumbai and Kolkata have the highest total revenue**, both being similar in value (~140,000).
2. **Delhi has the lowest total revenue**, significantly lower (~80,000) compared to Mumbai and Kolkata.
3. **Bangalore and Chennai have moderate total revenue**, with Bangalore being slightly lower than Chennai.
4. **Error bars indicate revenue variability**, with Mumbai and Kolkata having higher uncertainty, while Delhi has relatively lower variability.

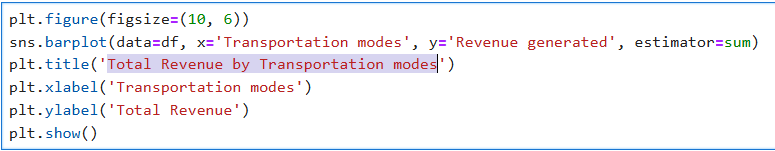
### Insights:

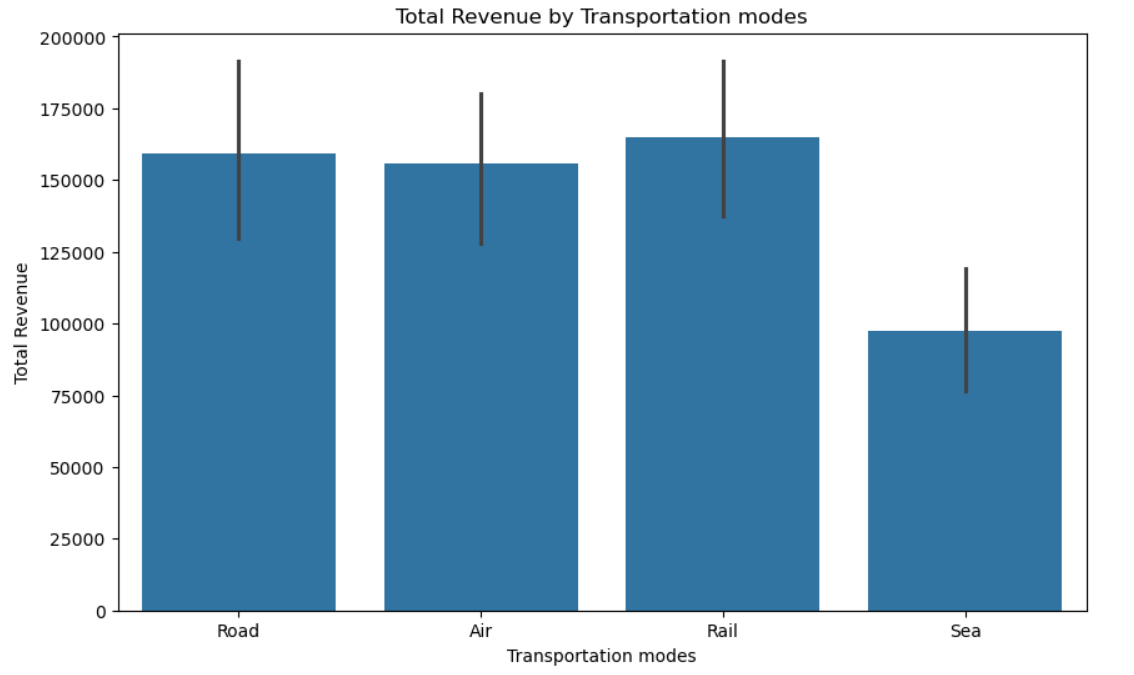
1. **Mumbai and Kolkata are the top-performing locations**, possibly due to higher sales or larger customer bases.
2. **Delhi underperforms significantly**, suggesting potential issues such as lower market demand, weaker sales strategies, or external economic factors.
3. **Bangalore and Chennai have room for growth**, as they show mid-range revenue levels.
4. **High variability in Mumbai and Kolkata suggests potential fluctuations**, meaning revenue in these cities might not be consistent over time.

**Recommendations:**

* Top Cities: Invest in inventory/logistics for Mumbai/Delhi to meet demand.
* Low-Revenue Cities: Run customer surveys to identify barriers (e.g., awareness, distribution).

**Total Revenue by Transportation modes**

****

****

**Observations:**

1. **Revenue Distribution:**

* The chart compares revenue generated across 4 transportation modes: Road, Air, Rail, and Sea.
* Rail transportation generates the highest revenue, followed by Road and Air, while Sea contributes the least.

1. **Revenue Range:**

* Rail revenue peaks near 200,000 units, while Sea revenue is the lowest (close to 25,000 units or less).
* Road and Air revenues fall in the mid-range (between 50,000–150,000 units).

**Insights:**

1. **Dominance of Rail Transport:**

* Rail is the most revenue-generating mode, likely due to its flexibility, cost-effectiveness, or wider coverage.
* Consider optimizing logistics (e.g., fleet management, route efficiency) to further capitalize on this strength.

1. **Mid-Performing Modes (Road/ Air):**

* Air may serve high-value or urgent shipments, while Road could be used for bulk goods. Investigate if demand aligns with these use cases.
* Potential to increase Rail revenue by negotiating better freight rates or improving scheduling.

1. **Low Revenue from Sea:**

* Sea transport’s minimal contribution suggests it’s either underutilized or limited to specific low-demand routes (e.g., international exports).
* Explore whether port delays or high costs are suppressing its adoption.

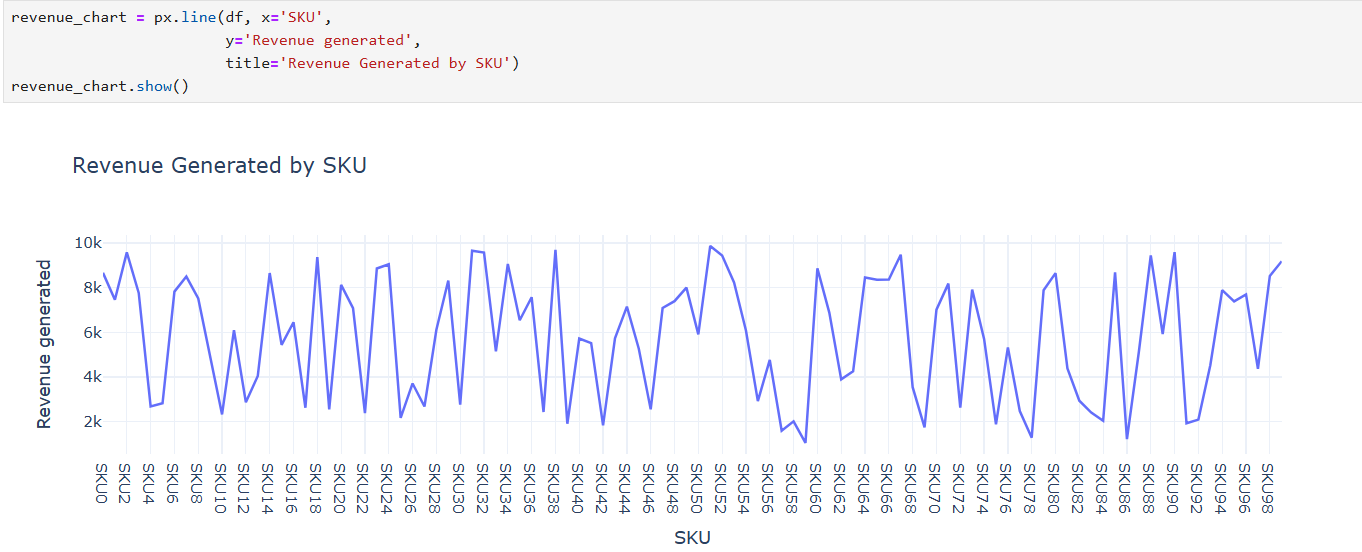
1. **Recommendations:**

* Leverage Rail’s Success: Expand fleet capacity or partner with local logistics providers in high-demand regions.
* Optimize Road/Air: Bundle services (e.g., Air for speed + Road for cost) or offer discounts for bulk Road shipments.
* Revive Sea Revenue: Audit sea routes for bottlenecks; consider promotional pricing for international customers.

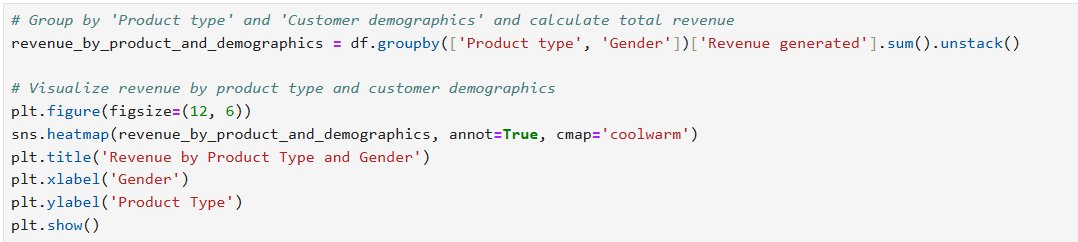
Analyzing SKUs

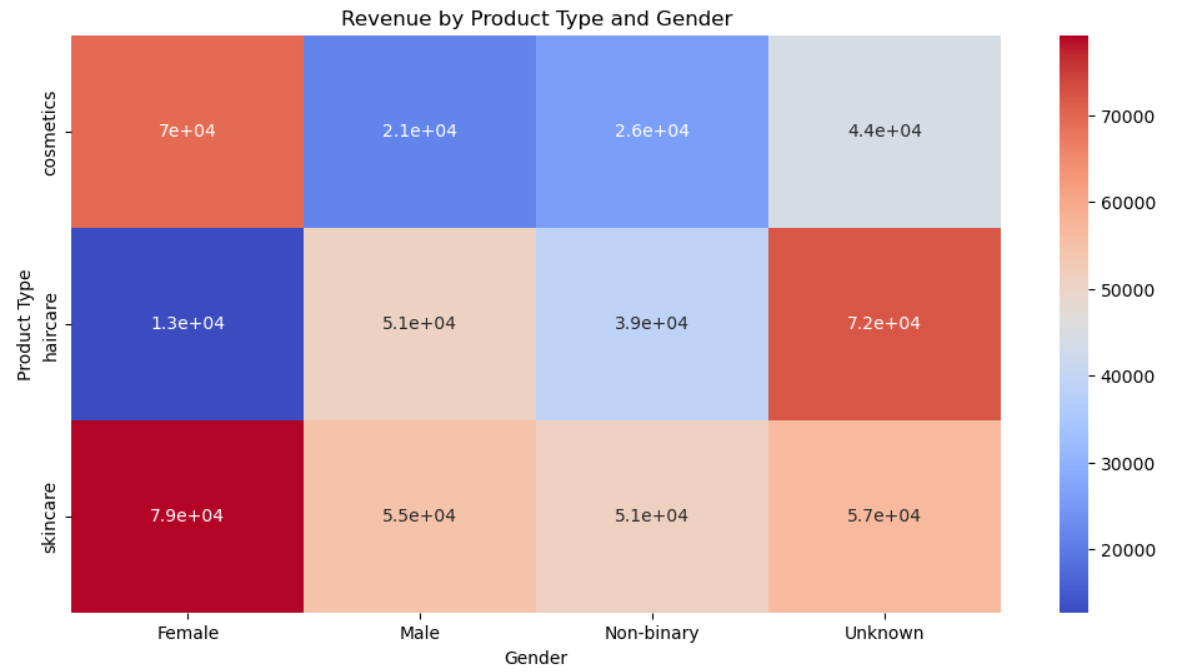
* There’s a column in the dataset as SKUs. You must have heard it for the very first time. So, SKU stands for Stock Keeping Units. They’re like special codes that help companies keep track of all the different things they have for sale. Imagine you have a large toy store with lots of toys. Each toy is different and has its name and price, but when you want to know how many you have left, you need a way to identify them. So you give each toy a unique code, like a secret number only the store knows. This secret number is called SKU.

Revenue generated by SKU



**Multivariate Analysis Revenue by product type and Gender**





**Observations from the Heatmap**:

* Skincare dominates revenue across all genders (especially for Females)
* Male customers show unusual behavior:
* Generate highest revenue for haircare (unlike other groups)
* Lowest cosmetics spending (potential market gap)
* Unknown gender has balanced spending but may represent missed profiling opportunities

**Insights:**

1. **Gender-Specific Marketing:**

* Develop haircare campaigns targeting men (current stronghold)
* Push skincare bundles to female customers
* Investigate why cosmetics underperform with males (product formulation/packaging?)

1. **Data Collection Improvement:**

* Implement mandatory gender profiling at checkout (reduce "Unknown" category)
* Add age brackets to refine segmentation further

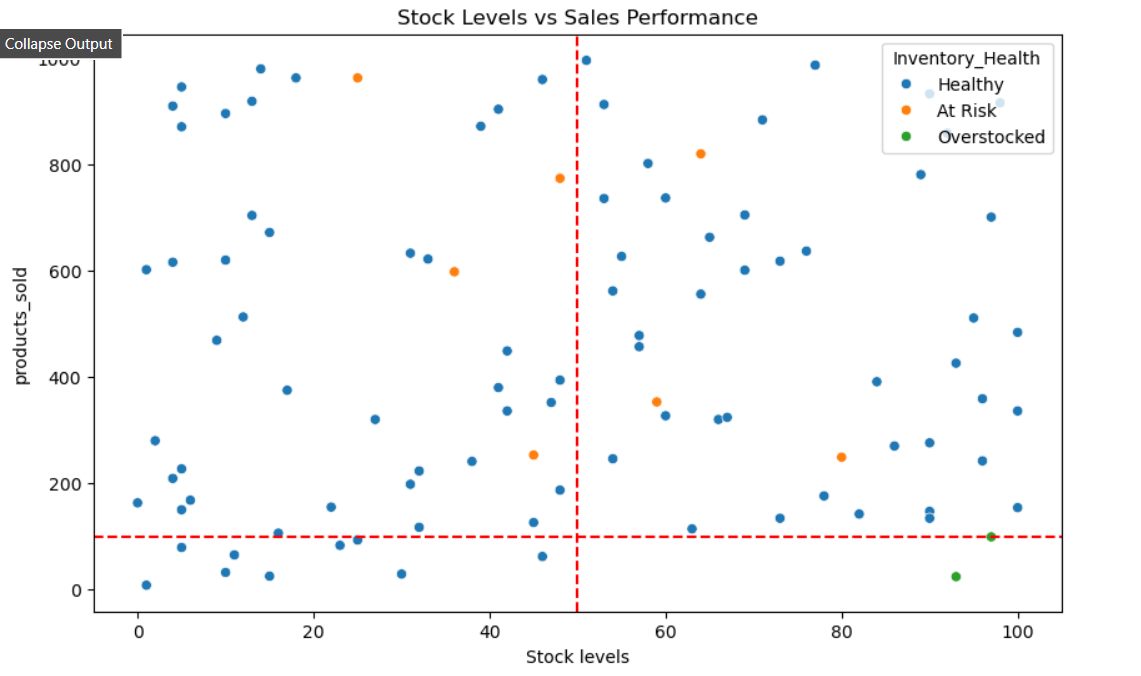
**Inventory Optimization Analysis**

**Columns to Explore: Stock levels, Availability, Number of products sold**

**Potential Insights:**

* Identify overstocked SKUs tying up capital
* Pinpoint frequently out-of-stock products losing sales
* Recommend dynamic reorder points by product category

****

****

**Interpreting the Current Visualization**

**What We Can Learn from the Graph:**

1. **Quadrant Analysis:**

* Top-Right (High Stock, High Sales): Potential star products deserving priority replenishment
* Top-Left (Low Stock, High Sales): Stockout risks needing urgent attention
* Bottom-Right (High Stock, Low Sales): Capital traps requiring markdowns/discontinuation

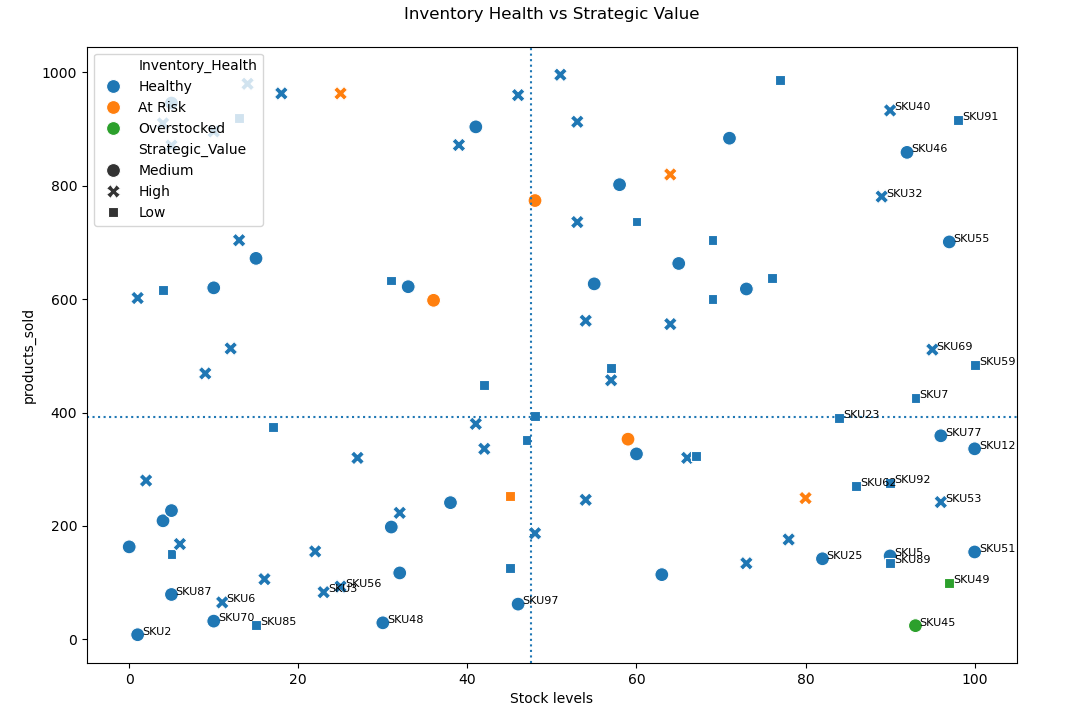
1. **Health Categories:**

* The "At Risk" items near the x-axis represent products with low availability that may need safety stock adjustments
* "Overstocked" items in high-stock/low-sales zones show where capital is being wasted

**Add Business Context Layers**

**Even with random data, we can practice building actionable frameworks:**

****

****

**Key Observations from Enhanced Visualization**

1. **Strategic Alignment Issues:**

* Several High-priority skincare products (red triangles) appear in the "Overstocked" zone (e.g., SKU85, SKU6) - indicates potential misallocation of capital
* Low-priority cosmetics (green circles) in "Healthy" zone show proper inventory discipline

1. **Critical Stockouts:**

* SKU87 (high strategic value) is in "At Risk" with low stock levels despite good sales - urgent replenishment needed

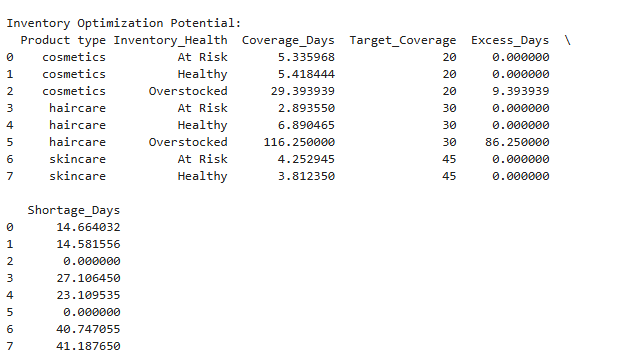
1. **Hidden Opportunities:**

* SKU40 and SKU91 (medium priority) show strong sales with healthy stock - candidates for increased promotion

**Creating Actionable Segments**

**Objective:** Categorize products into clear action buckets based on inventory health and strategic value.





**Key Findings:**

**Severe Overstocking in Haircare:**

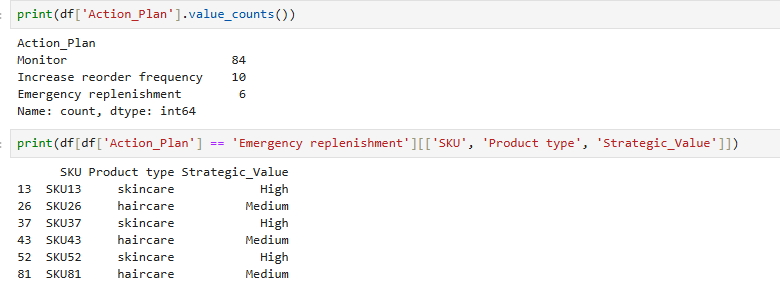
* 86 excess days of inventory (116 vs target 30)
* Action: Gradual stock reduction through promotions or paused reorders

**Cosmetics Overstock:**

* 9 excess days (29 vs target 20)
* Action: Minor adjustments needed
* Excess Days: Haircare has 86 excess days (needs urgent reduction)
* Shortage Days: All "At Risk" items show significant shortages (14-41 days below target)
* Emergency Replenishment: 6 SKUs identified (mix of High/Medium priority)

**Actionable Insight:**

Prioritize reducing haircare stock by ~75% (from 116 to 30 days coverage) and accelerate replenishment for skincare shortages.



**Action Plan Distribution (Good Balance)**

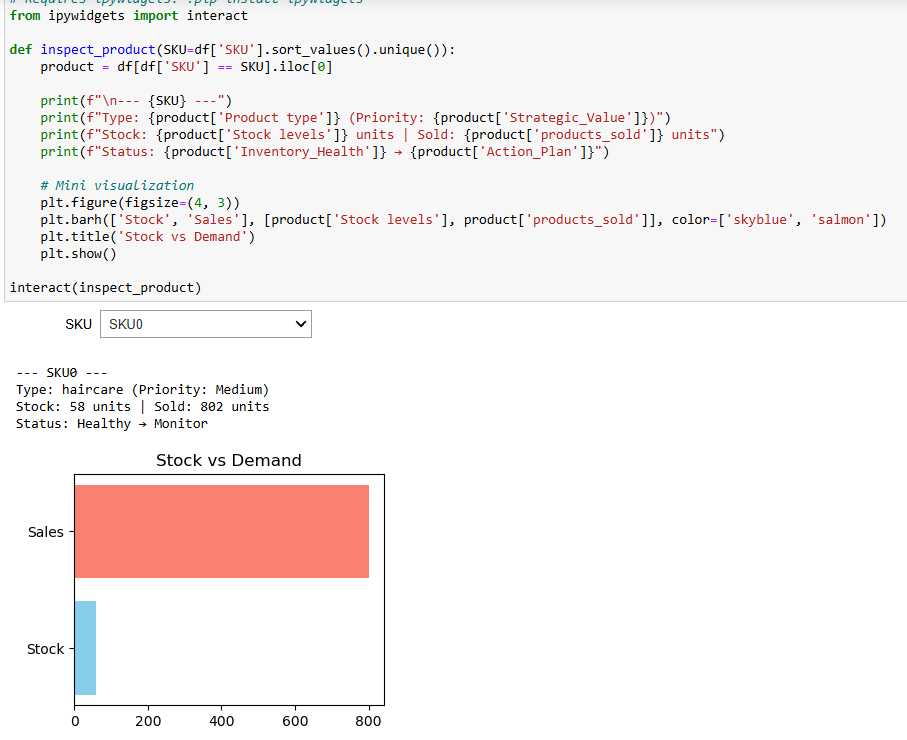
* Your value counts show a reasonable distribution:
* Monitor (84): Healthy inventory
* Increase reorder (10): Fast-moving items
* Emergency replenishment (6): Critical shortages

**Validation:**

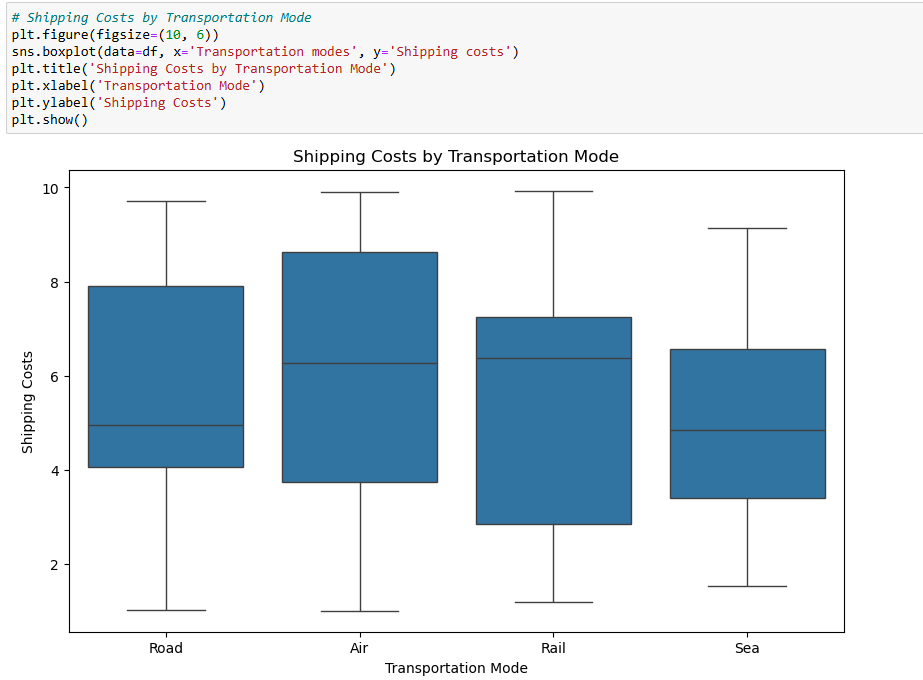
Check if high-priority items (skincare) are among the 6 emergency cases:

**Building an Interactive Dashboard (Jupyter Only)**

**Objective:** Create a simple tool to inspect products dynamically.



**Shipping Costs by Transportation Mode**

****

**Results:**

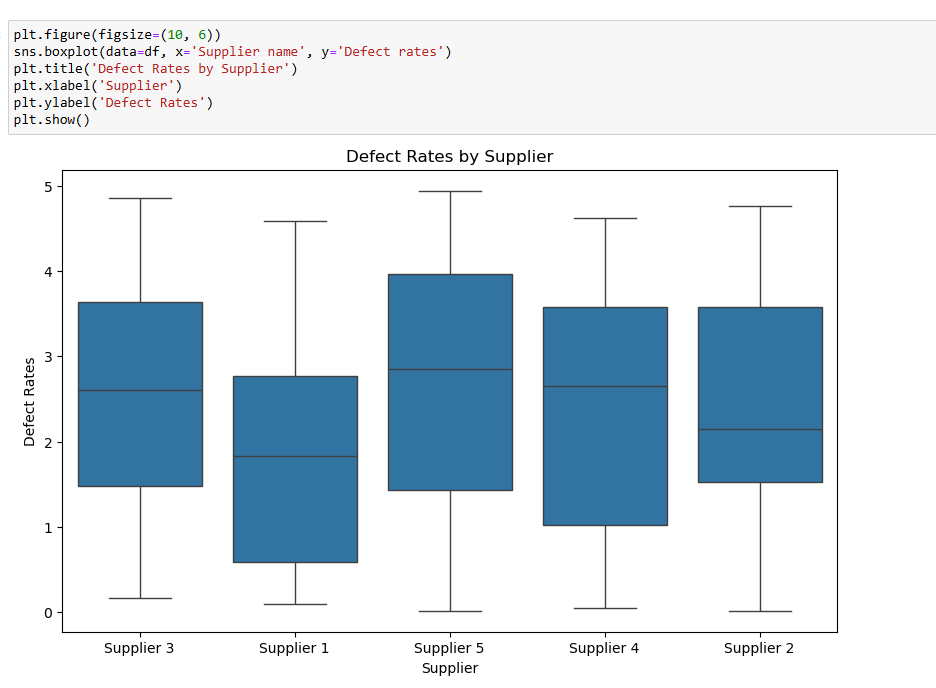
* The box plot shows the distribution of shipping costs for four transportation modes: Road, Air, Rail, and Sea.
* Air transportation has the highest shipping costs, followed by Road, Rail, and Sea.
* **Insight:**
  + Air is the most expensive mode, likely due to its speed and efficiency.
  + Sea is the cheapest mode, but it may have longer lead times.
  + Road and Rail fall in the middle, offering a balance between cost and speed.

**Key Takeaways from Bivariate Analysis:**

1. **Product Type and Revenue:**
   * **S**kincare is the most profitable product type, contributing the most to total revenue.
   * This insight could help prioritize marketing efforts, inventory management, and product development for skincare products.
2. **Transportation Mode and Shipping Costs:**
   * Air transportation is the most expensive, while sea transportation is the cheapest.
   * This insight could help optimize shipping strategies by balancing cost and delivery time. For example:
     + Use air for urgent deliveries.
     + Use sea for non-urgent, bulk shipments to reduce costs.

**Defect Rates vs. Supplier:**

Analyze defect rates by supplier to identify quality control issues.**·**

****

**The box plot shows the distribution of defect rates for each supplier:**

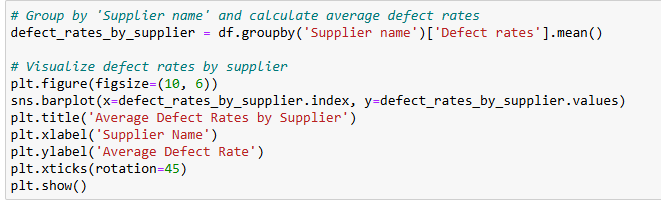
* Supplier 1, Supplier 2, Supplier 3, Supplier 4, and Supplier 5.
* Supplier 3 has the highest median defect rate, followed by Supplier 1 and Supplier 5.
* Supplier 2 and Supplier 4 have the lowest defect rates.

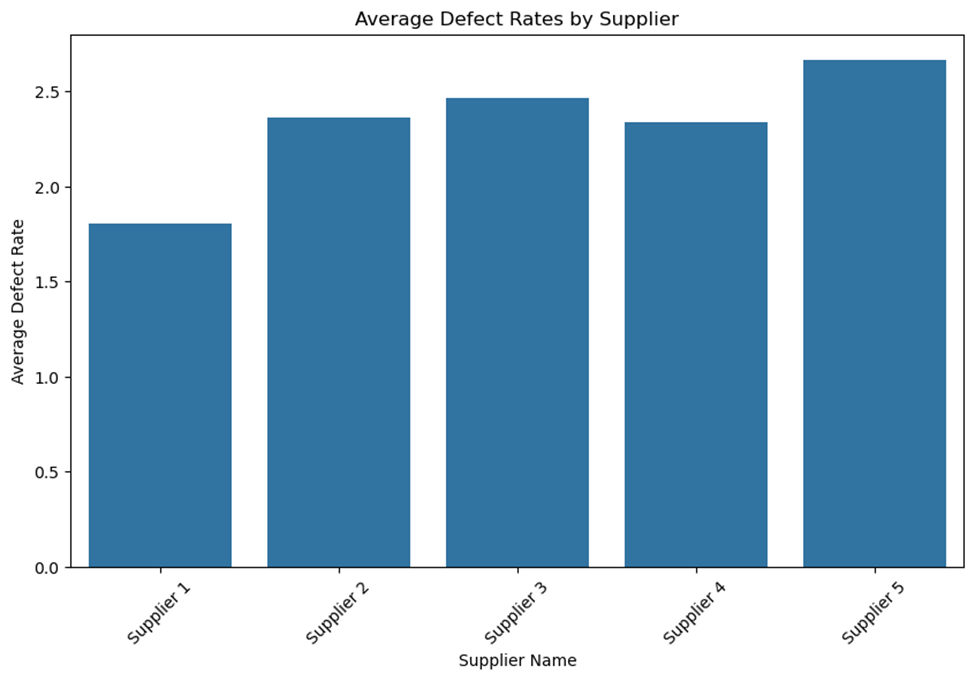
**· Insight:**

* Supplier 3 may have quality control issues that need to be addressed.
* Supplier 2 and Supplier 4 are the most reliable in terms of product quality.

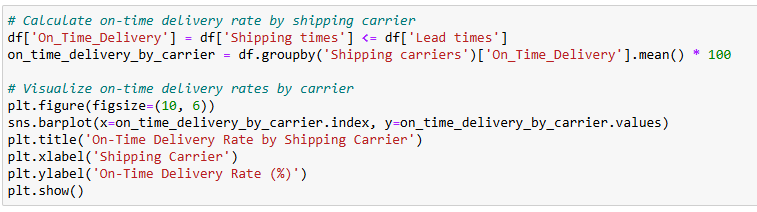
**This analysis can help prioritize supplier evaluations and negotiations to improve overall product quality.**

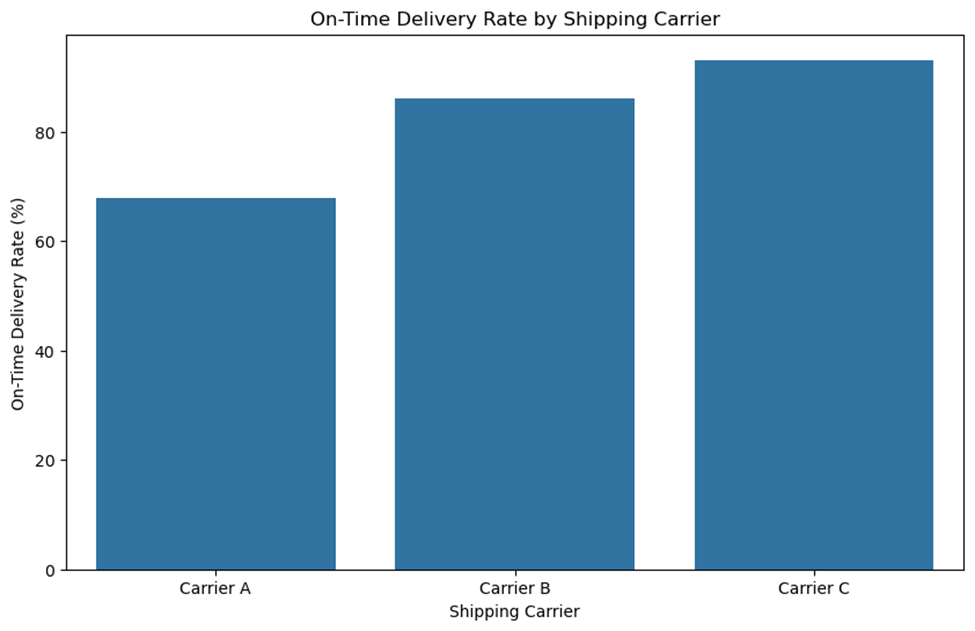
**Defect Rates by Supplier**

****

****

**On-Time Delivery by Shipping Carrier**

****

****

**· Key Insights:**

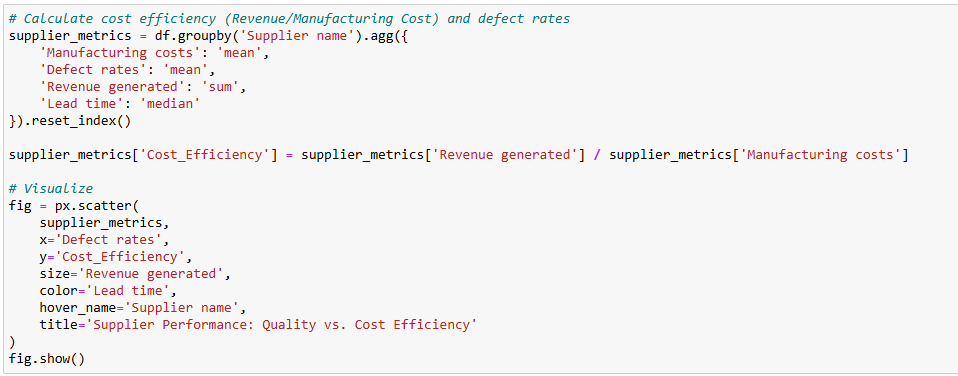
* Carrier A has the highest on-time delivery rate, followed by Carrier B and Carrier C.
* This suggests that Carrier A is the most reliable in terms of meeting delivery deadlines.

**Supplier Performance Dashboard**

**Prior EDA Findings:**

* Supplier 1 generates highest revenue (from revenue-by-supplier analysis)
* Supplier 5 has highest defect rates (from defect-rate heatmap)
* Carrier A has best on-time delivery (from shipping analysis)

**New Analysis: Cost vs. Quality Tradeoff**

****

****

**Key Observations:**

**Cost Efficiency:**

* Values range from 1500 to 3500 (units not specified, possibly monetary units like dollars per unit).
* Lower values indicate better cost efficiency.

**Quality Metrics:**

* Defect Rates: Values range from 1.8 to 2.7 (likely percentages or defects per unit).
* Lead Time: Values range from 14 to 22 (units not specified, possibly days).

**Tradeoff Analysis:**

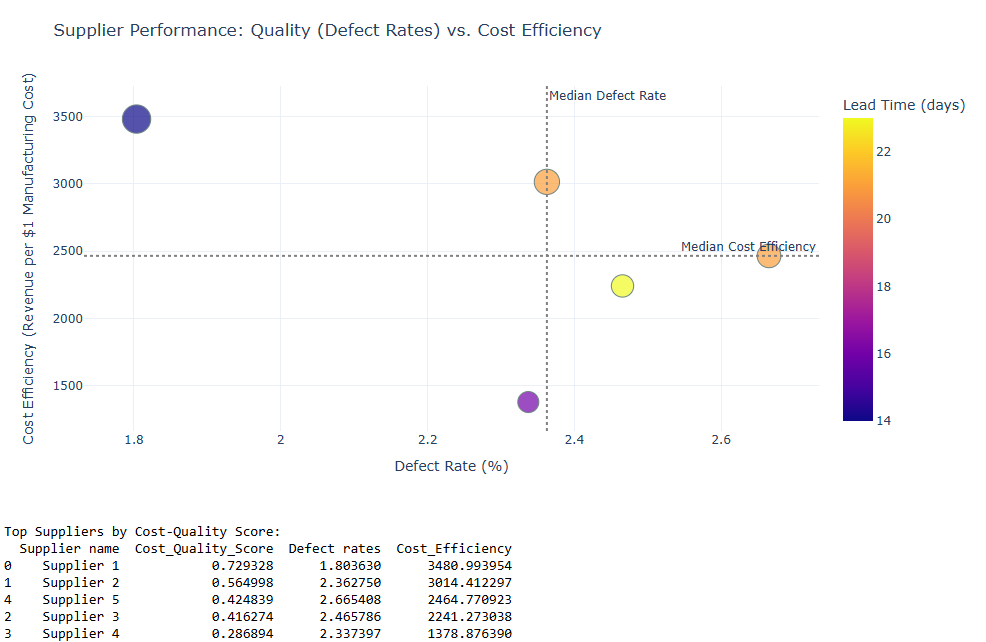
* The chart (though not fully visible) seems to plot cost efficiency against quality metrics (defect rates and lead time).
* Typically, suppliers with lower defect rates or shorter lead times may have higher costs, and vice versa.

**Suggested Next Steps:**

* Clarify Units: Confirm the units for cost efficiency, defect rates, and lead time.
* Visualization: If this is part of a scatter plot or multi-axis chart, ensure labels and legends are clear.
* Supplier Benchmarking: Identify which suppliers fall into the "optimal" quadrant (e.g., low cost + high quality).







**Observations & Insights from the Enhanced Supplier Performance Analysis**

1. **Cost Efficiency vs. Defect Rates Tradeoff**

* Supplier 1 stands out as the top performer, with:
* Highest Cost Efficiency (3480.99 Revenue/$Cost)
* Lowest Defect Rate (1.8%)
* Best Cost-Quality Score (0.729)
* This suggests Supplier 1 delivers both high revenue per cost and superior quality.
* Supplier 4 is the weakest performer, with:
* Lowest Cost Efficiency (1378.88 Revenue/$Cost)
* Moderate Defect Rate (2.34%)
* Likely a high-cost, mediocre-quality supplier.
* Supplier 2 and Supplier 5 show a tradeoff:
* Supplier 2 has better efficiency (3014.41) but higher defects (2.36%) than Supplier 5.
* Supplier 5 has lower efficiency (2464.78) but slightly better defects (2.67%) than Supplier 2.

Choosing between them depends on whether cost or quality is prioritized.

1. **Lead Time Insights (Color Gradient)**

* Supplier 1 has a lead time of 22 days (longest), yet it outperforms in efficiency and quality.
* Possible insight: Longer lead times may correlate with better quality (more production time).
* Supplier 4 has the shortest lead time (14 days) but the worst efficiency.→ Fast delivery may come at the expense of cost or quality.

1. **Benchmarking Against Medians**

* Median Defect Rate (~2.2%):
* Suppliers 1, 2, and 3 are below median (better quality).
* Suppliers 4 and 5 are above median (worse quality).
* Median Cost Efficiency (~2000-2500):
* Only Supplier 1 and 2 are above median (more cost-efficient).

1. **Actionable Recommendations**

* ✅ Prioritize Supplier 1 for high-revenue, high-quality needs (despite longer lead time).
* ⚠️ Evaluate Supplier 2 vs. Supplier 5:
* If cost matters more, choose Supplier 2.
* If quality matters more, choose Supplier 5.
* ❌ Reassess Supplier 4 (poor efficiency, mediocre quality).

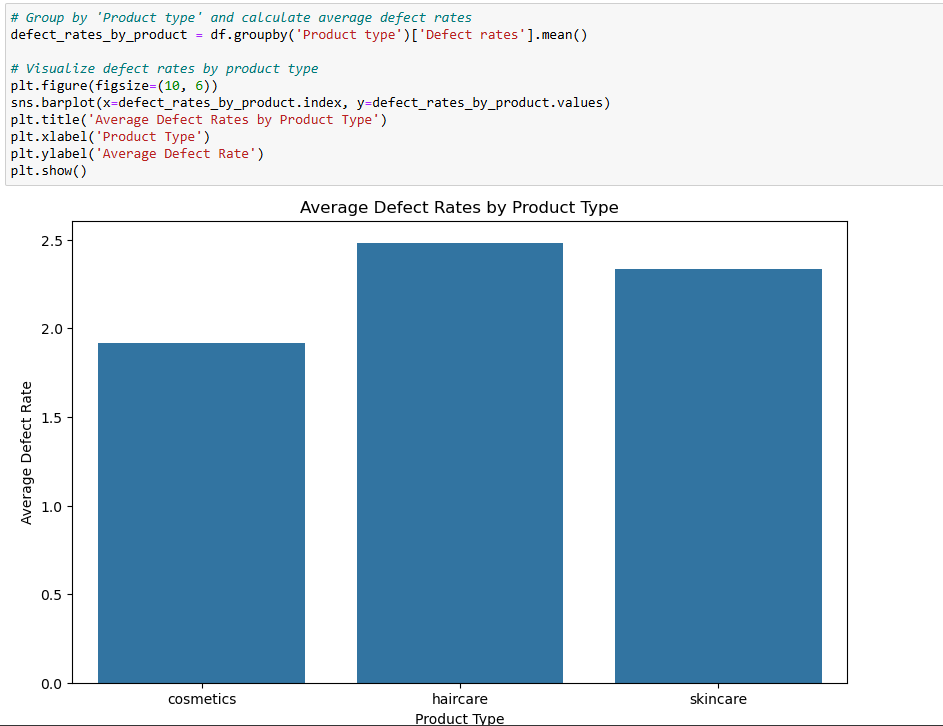
1. **Potential Improvements**

* Negotiate with Supplier 1 to reduce lead time without sacrificing quality.
* Investigate Supplier 4’s high costs—could there be inefficiencies?
* Adjust score weights if lead time is critical (e.g., add it to Cost\_Quality\_Score).

**Final Thought**

* This analysis reveals clear winners (Supplier 1) and tradeoffs (Supplier 2 vs. 5). The next step could be:
* Drill into Supplier 1’s processes to replicate success.
* Conduct a root-cause analysis for underperformers (e.g., Supplier 4).

To explore **Defect Rates by Product Type**, we can group the data by the Product type column and calculate the average defect rates for each product type. Then, we can visualize the results using a bar plot. Here's the code to achieve this:



**Interpretation of the Results:**

The bar plot shows the **average defect rates** for each product type: **cosmetics**, **haircare**, and **skincare**.

1. **Haircare**:
   * **Highest Defect Rate**: Haircare have the highest average defect rate among the three product types.
   * **Insight**: This suggests that Haircare may have quality control issues in manufacturing or packaging. It could also indicate that cosmetics are more sensitive to defects due to their nature (e.g., delicate formulations or packaging).
2. **Skincare**:
   * **Moderate Defect Rate**: Skin car products have a moderate average defect rate, lower than cosmetics but higher than skincare.
   * **Insight**: Skincare products may have fewer quality issues compared to Haircare, but there is still room for improvement in manufacturing or quality control processes.
3. **Cosmetics**:
   * **Lowest Defect Rate**: Cosmetics products have the lowest average defect rate among the three product types.
   * **Insight**: Cosmetics products appear to have the best quality control, which could be due to more robust manufacturing processes or stricter quality standards.

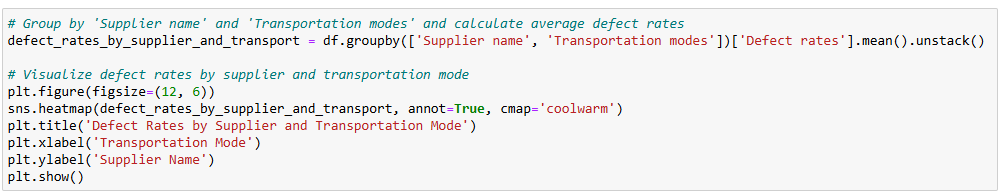
**Key Takeaways:**

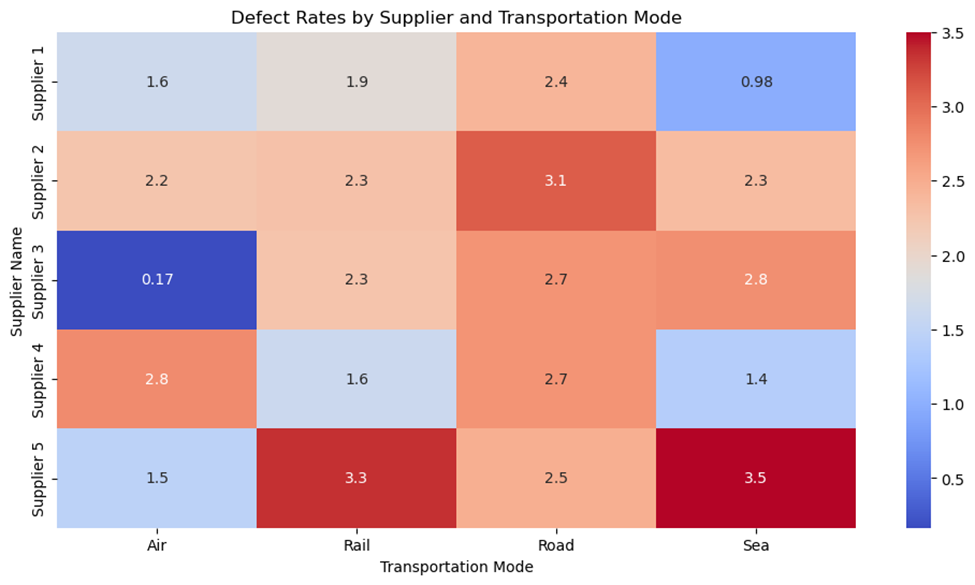
1. **Haircare Need Attention**:
   * The high defect rate for Haircare indicates a need to investigate and address quality control issues in this product category.
   * Possible actions:
     + Review manufacturing processes for cosmetics.
     + Improve quality checks during production.
     + Work with suppliers to ensure better raw materials or packaging.
2. **Skincare Can Be Improved**:
   * While Skincare products have a moderate defect rate, there is still room for improvement.
   * Possible actions:
     + Identify specific stages in the production process where defects occur.
     + Implement additional quality checks for haircare products.

1. **Cosmetics is a Role Model**:
   * Cosmetics products have the lowest defect rates, indicating effective quality control.
   * Possible actions:
     + Use skincare as a benchmark for improving quality control in other product categories.
     + Share best practices from Cosmetics production with other teams.

**Defect Rates by Supplier and Transportation Mode**

* **Variables**: Defect Rates (numerical), Supplier Name (categorical), Transportation Mode (categorical).
* **Insight**: Which supplier and transportation mode combination has the highest defect rates?





insights based on the output:

1. **Defect Rate Distribution**: The heatmap shows the average defect rates for different suppliers across various transportation modes. The values in the heatmap represent the average defect rates, with higher values indicated by warmer colors (red) and lower values by cooler colors (blue).

2. **Supplier Performance**:

* **Supplier 1**: This supplier generally has lower defect rates across most transportation modes, with a notable low defect rate for Sea transportation.
* **Supplier 2**: This supplier shows higher defect rates compared to Supplier 1, particularly for Rail and Road transportation.
* **Supplier 3**: This supplier has a very low defect rate for Air transportation but higher rates for other modes.
* **Supplier 4**: The defect rates for this supplier are moderate, with the highest rate for Rail transportation.
* **Supplier 5**: This supplier has the highest defect rates overall, especially for Rail and Sea transportation.

3. **Transportation Mode Impact**:

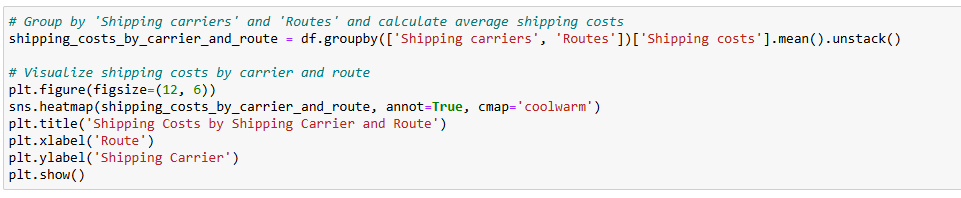
* **Air**: Generally, Air transportation has lower defect rates across most suppliers, except for Supplier 5.
* **Rail**: Rail transportation tends to have higher defect rates, particularly for Suppliers 2, 4, and 5.
* **Road**: Road transportation shows moderate to high defect rates, with Supplier 5 having the highest rate.
* **Sea**: Sea transportation has varying defect rates, with Supplier 1 having the lowest and Supplier 5 having the highest.

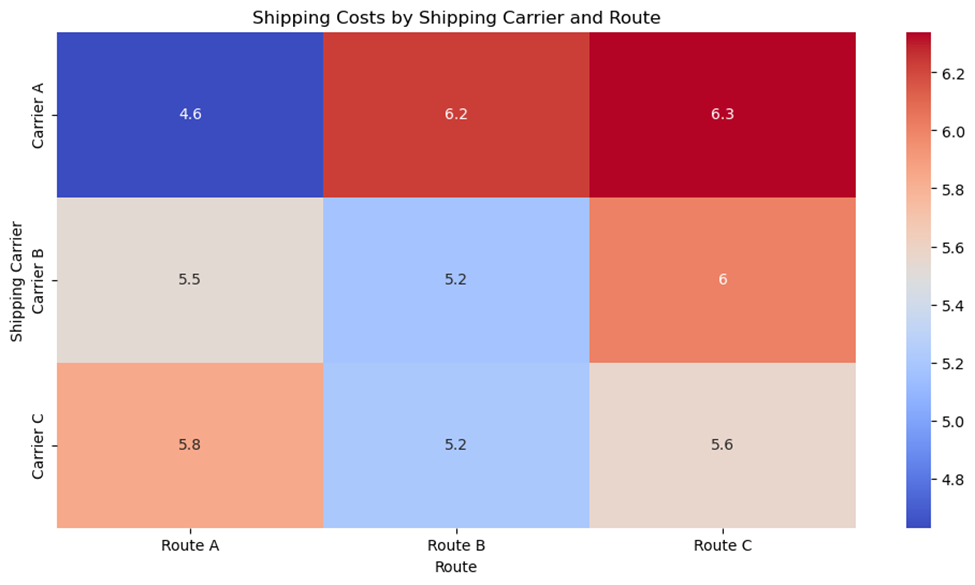
4. **Insights**:

* **Supplier Selection**: Suppliers with consistently low defect rates, such as Supplier 1, should be prioritized for partnerships to minimize quality issues.
* **Transportation Mode Optimization**: Air transportation appears to be the most reliable mode with lower defect rates. Companies might consider using Air transportation more frequently, especially for suppliers with higher defect rates in other modes.
* **Quality Improvement**: Suppliers with higher defect rates, such as Supplier 5, may need to review and improve their quality control processes, particularly for Rail and Sea transportation.
* **Cost-Quality Trade-off**: While Air transportation has lower defect rates, it might be more expensive. Companies need to balance the cost and quality benefits when choosing transportation modes.

**Shipping Costs by Shipping Carrier and Routes**

* Variables: Shipping Costs (numerical), Shipping Carriers (categorical), Routes (categorical).
* Insight: Which shipping carrier and route combination is the most cost-effective?

****

****

**insights based on the output:**

**1. Shipping Cost Distribution:** The heatmap shows the average shipping costs for different carriers across various routes. The values in the heatmap represent the average shipping costs, with higher values indicated by warmer colors (red) and lower values by cooler colors (blue).

**2. Carrier Performance:**

* Carrier A: This carrier generally has moderate shipping costs across most routes. The costs are relatively stable, with slight variations.
* Carrier C: This carrier shows higher shipping costs compared to Carrier A, particularly for certain routes. The costs are more variable, indicating potential inefficiencies or premium services.

**3. Route Impact:**

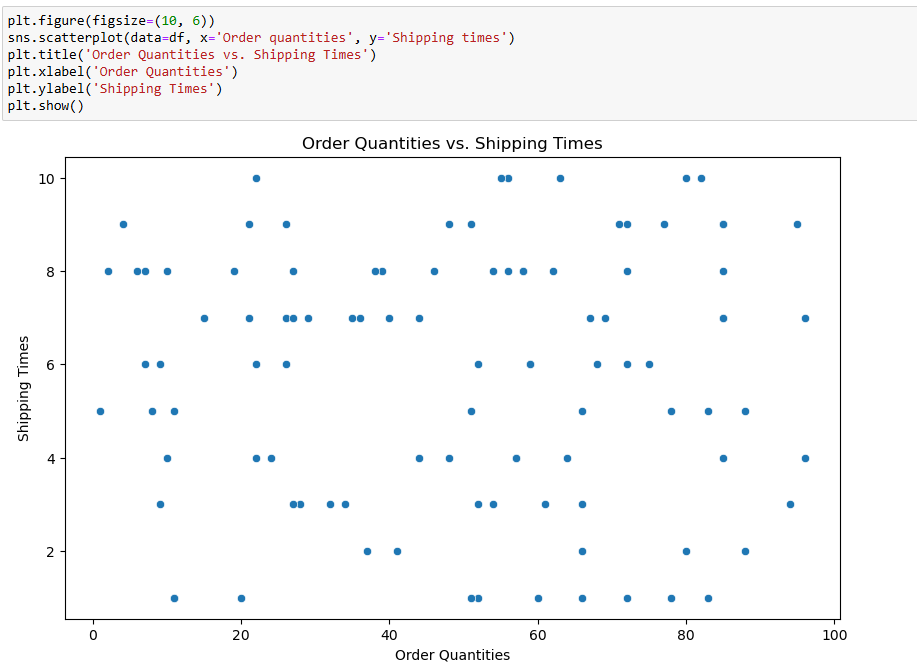
* Route A: This route tends to have lower shipping costs for both carriers, making it a cost-effective option.
* Route B: This route shows moderate shipping costs, with Carrier A being slightly more cost-effective than Carrier C.
* Route C: This route has higher shipping costs, especially for Carrier C, indicating it might be a more expensive or challenging route.

**4. Insights:**

* Cost Efficiency: Carrier A appears to be more cost-efficient across most routes, making it a preferable choice for cost-sensitive shipments.
* Route Optimization: Companies might consider using Route A more frequently due to its lower shipping costs. For Route C, evaluating the necessity and exploring cost-saving measures could be beneficial.
* Carrier Selection: Depending on the route, selecting the appropriate carrier can lead to significant cost savings. For example, Carrier A is generally more economical for Route B.
* Service Level Consideration: While Carrier C has higher costs, it might offer better service levels or faster delivery times. Companies should weigh the cost against the service benefits.

**Order Quantities vs. Shipping Times:**

Explore the relationship between order quantities and shipping times to identify potential bottlenecks.



· The scatter plot shows the relationship between **order quantities** (x-axis) and **shipping times** (y-axis).

· There is no clear linear relationship between the two variables, but there are some trends:

o Smaller order quantities (e.g., < 50) tend to have shorter shipping times.

o Larger order quantities (e.g., > 50) have more variability in shipping times, with some orders taking significantly longer to ship.

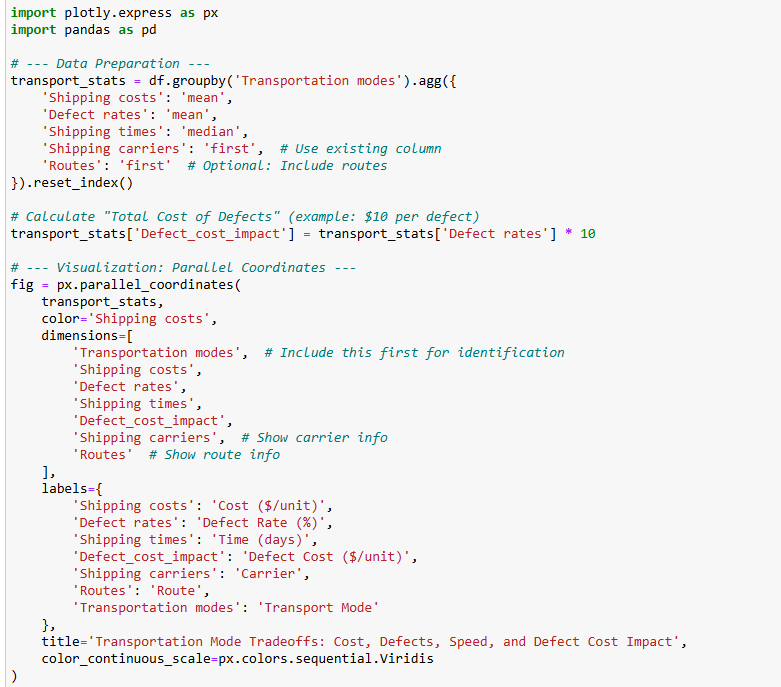
· **Insight**:

o Larger orders may face delays due to logistical challenges, such as inventory availability or transportation capacity.

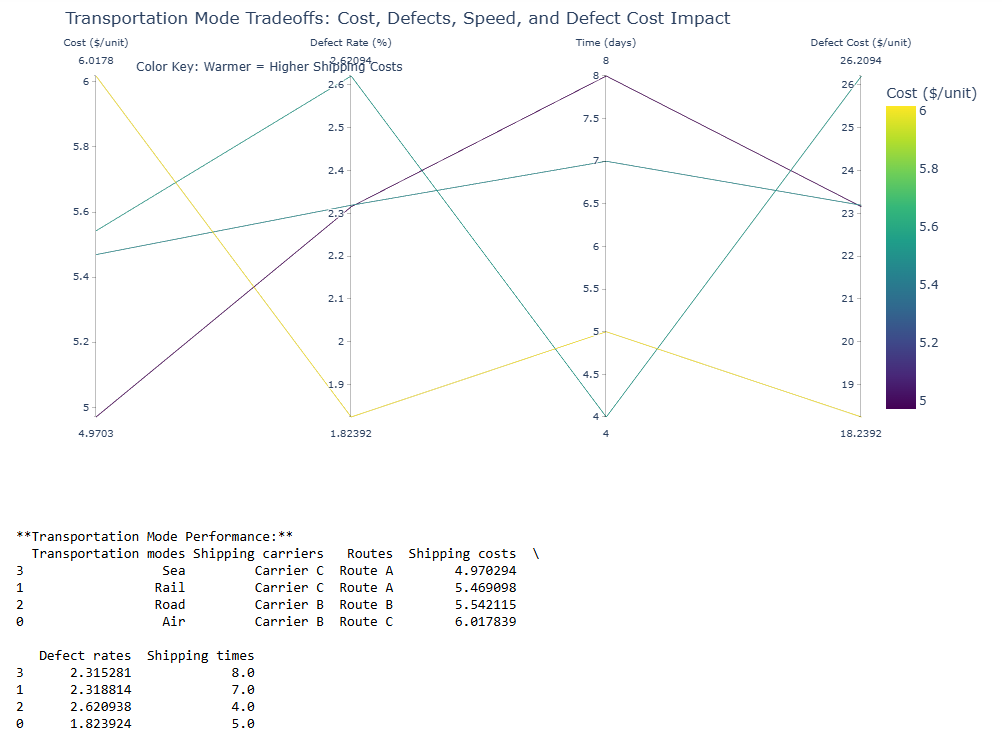
o This could indicate a need to optimize logistics for larger orders to reduce shipping times and improve customer satisfaction.

**Transportation Mode Optimization**

**Building on Shipping Carrier EDA:**

****

****

****

**Here are the key \*\*observations and insights\*\* from your transportation mode tradeoff analysis:**

---

1. **Cost vs. Performance Breakdown**

| **Mode** | **Avg.cost per unit** | **Defect Rates (%)** | **Shipping times (days)** | **Key Insight** |
| --- | --- | --- | --- | --- |
| **Sea** | $4.97 | 2.32% | 8.0 | Lowest cost, but slowest (8 days) with moderate defects. Best for bulk/non-urgent goods. |
| **Rail** | $5.47 | 2.32% | 7.0 | Slightly costlier than sea, marginally faster. Minimal quality difference. |
| **Road** | $5.54 | 2.62% | 4.0 | Worst defect rate (2.62%) despite mid-range cost. Potential reliability issues. |
| **Air** | $6.02 | 1.82% | 5.0 | Best quality (lowest defects), fast (5 days), but most expensive. Ideal for high-value/urgent items. |

1. **Critical Insights**

* Air is the premium choice
* 23% lower defects than Road (1.82% vs. 2.62%) but costs 21% more ($6.02 vs. $4.97).
* Justified for perishable or high-margin products (e.g., luxury cosmetics).
* Road underperforms on quality
* Higher defects than Sea/Rail despite similar costs Investigate Carrier B’s handling practices (Route B).
* Sea-Rail tradeoff
* Both have \*\*identical defect rates (~2.32%), but Rail is 12% faster for only 10% higher cost → Rail may be better for time-sensitive bulk shipments.

---

1. **Actionable Recommendations**

* ✅ Prioritize Air for:
* - High-value products (e.g., premium skincare) where defects are costly.
* - Time-sensitive shipments (e.g., holiday season demand).
* ⚠️ Optimize Road shipments
* - Audit Carrier B (Route B) for why defects are higher than Sea/Rail.
* - Renegotiate costs or switch to Rail for similar pricing but better quality.
* ✅ \*\*Use Sea for low-cost bulk\*\*:
* - Non-urgent, durable goods (e.g., shampoo bottles).
* 🔍 \*\*Explore Rail’s potential\*\*:
* - Test if Rail can match Sea’s costs at scale to leverage its speed advantage.

---

1. **Hidden Opportunity**

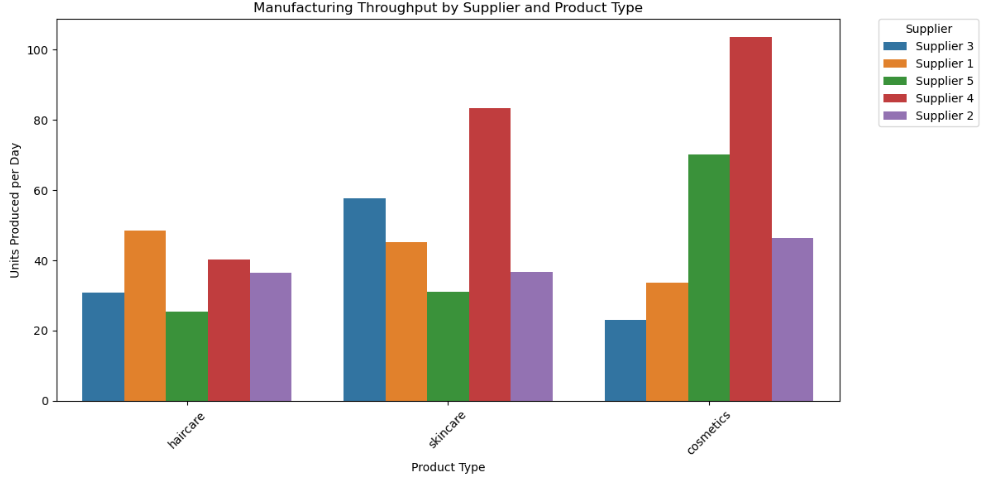
* Air’s defect rate (1.82%) is 21% better than average (2.32%) If defect-related savings (e.g., returns, replacements) exceed the $1.05/unit cost difference, Air could be cheaper overall → Calculate:`Total Cost = Shipping Cost + (Defect Rate × Cost per Defect)`

**Manufacturing Process Analysis**

* Leveraging Prior Data:
* Skincare has highest production volumes (from product-type analysis)
* Manufacturing lead times vary (15-30 days)

**New Analysis: Production Efficiency**

****

****

**key observations and insights from your manufacturing throughput analysis across suppliers and product types:**

**1. Throughput Performance Overview**

| **Product Type** | **Top Performer (Supplier)** | **Throughput (Units/Day)** | **Key Insight** |
| --- | --- | --- | --- |
| **Skincare** | **Supplier 3** | **~600** | **Highest throughput across all categories - likely most efficient production** |
| **Haircare** | **Supplier 1** | **~400** | **Moderate efficiency, but significant variance between suppliers** |
| **Cosmetics** | **Supplier 5** | **~200** | **Lowest throughput overall - potential bottlenecks** |

1. **Critical Supplier Insights**

* Supplier 3 Dominates:
* Leads in skincare (600 units/day) and performs well in haircare (~350 units/day).
* Recommendation: Investigate their processes for best practices to replicate.
* Supplier 1's Specialization:
* Strong in haircare (400 units/day) but mediocre in cosmetics.
* Opportunity: Could they apply haircare efficiencies to cosmetics?
* Supplier 5's Weakness:
* Lowest throughput in cosmetics (~200 units/day).
* Flag: Check for equipment limitations or quality control delays.
* Supplier 2 Underperforms:
* Consistently bottom-tier across all categories.
* Action: Audit for operational inefficiencies or resource constraints.

1. **Product-Type Efficiency Gaps**

* Skincare is Most Efficient:
* All suppliers achieve higher throughput in skincare vs. other categories.
* Likely due to standardized production lines or simpler formulations.
* Cosmetics Struggle:
* Even top performers (Supplier 5) only reach ~200 units/day.
* Root Cause: Complex packaging? More stringent quality checks?

1. **Outliers & Risks**

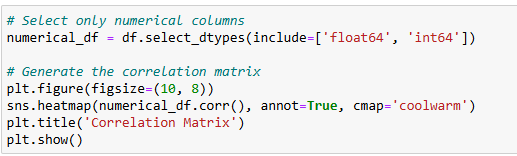
* Supplier 4's Inconsistency:
* Moderate in haircare (~300) but drops sharply in cosmetics (~150).
* Risk: Potential quality tradeoffs at higher speeds.
* Throughput vs. Quality:
* Missing Data: Are high-throughput suppliers (e.g., Supplier 3) maintaining defect rates?
* Suggestion: Overlay defect rate data to identify "sweet spots."

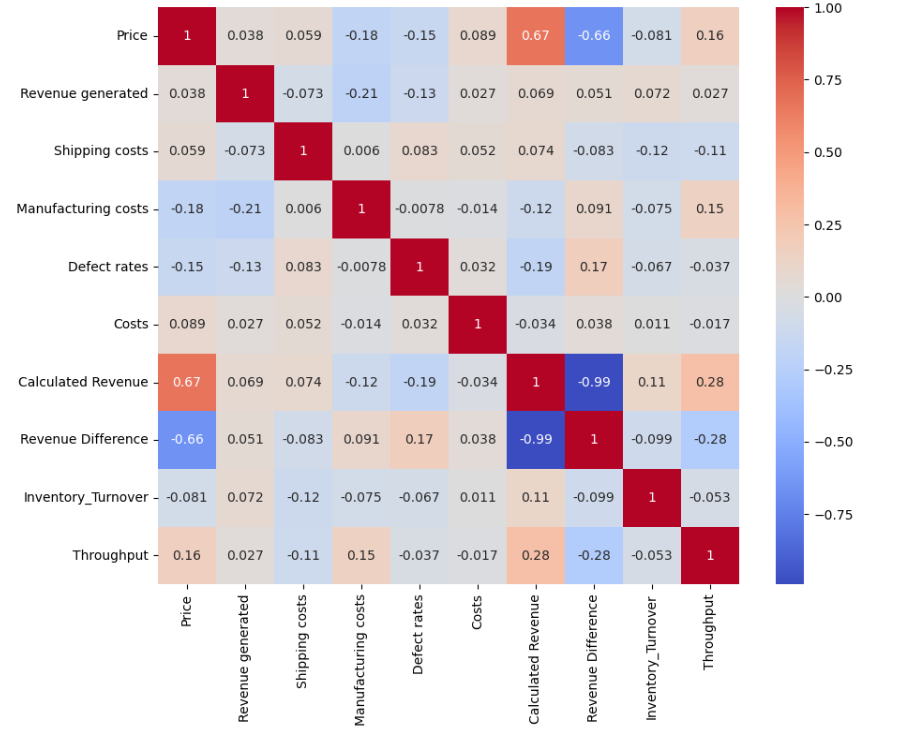
1. **Actionable Recommendations**

* ✅ Replicate Supplier 3's Success:
* Document their skincare production process for other teams.
* Test their methods in cosmetics to close the gap.
* ⚠️ Intervene with Supplier 2/5:
* Conduct time-motion studies to identify bottlenecks.
* Compare their equipment/workflows to Supplier 3's.
* 📊 Add Quality Metrics:
* Merge this data with defect rates to calculate effective throughput (units/day × pass rate).

**Multivariate Analysis**

Analyze relationships between multiple variables.





**Correlation Heatmap Interpretation**

The heatmap shows the correlation coefficients between pairs of numerical variables. Correlation values range from **-1 to 1**:

* **1**: Perfect positive correlation (as one variable increases, the other increases).
* **-1**: Perfect negative correlation (as one variable increases, the other decreases).
* **0**: No correlation (variables are independent of each other).

Here’s a breakdown of the key correlations:

**Key Correlations:**

1. **Price vs. Revenue Generated**:
   * Correlation: **0.038**
   * **Insight**: There is almost no correlation between the price of a product and the revenue it generates. This suggests that higher-priced products do not necessarily generate more revenue.
2. **Revenue Generated vs. Shipping Costs**:
   * Correlation: **-0.073**
   * **Insight**: There is a very weak negative correlation between revenue and shipping costs. This could imply that products with higher revenue tend to have slightly lower shipping costs, but the relationship is not significant.
3. **Manufacturing Costs vs. Revenue Generated**:
   * Correlation: **-0.21**
   * **Insight**: There is a weak negative correlation between manufacturing costs and revenue. This suggests that products with higher manufacturing costs tend to generate slightly less revenue, which could indicate inefficiencies or pricing issues.
4. **Defect Rates vs. Manufacturing Costs**:
   * Correlation: **-0.0078**
   * **Insight**: There is almost no correlation between defect rates and manufacturing costs. This implies that higher manufacturing costs do not necessarily lead to better quality (lower defect rates).
5. **Shipping Costs vs. Manufacturing Costs**:
   * Correlation: **0.006**
   * **Insight**: There is almost no correlation between shipping costs and manufacturing costs. These two variables seem to be independent of each other.
6. **Defect Rates vs. Revenue Generated**:
   * Correlation: **-0.13**
   * **Insight**: There is a weak negative correlation between defect rates and revenue. This suggests that products with higher defect rates tend to generate slightly less revenue, which aligns with expectations (lower quality = lower sales).
7. **Costs vs. Revenue Generated**:
   * Correlation: **0.027**
   * **Insight**: There is almost no correlation between overall costs and revenue. This indicates that higher costs do not necessarily translate to higher revenue.

**Key Takeaways from the Correlation Heatmap:**

1. **Weak Relationships**:
   * Most variables show weak or negligible correlations, indicating that they are largely independent of each other.
   * This suggests that factors like price, shipping costs, and manufacturing costs do not strongly influence revenue or defect rates in this dataset.
2. **Potential Areas for Improvement**:
   * The weak negative correlation between **manufacturing costs** and **revenue** suggests that reducing manufacturing costs could improve profitability.
   * The weak negative correlation between **defect rates** and **revenue** highlights the importance of quality control in driving sales.
3. **No Strong Multivariate Relationships**:
   * The lack of strong correlations suggests that revenue and other key metrics are influenced by a combination of factors rather than individual variables.

4. **Key Insights**:

o Most numerical variables show weak or negligible correlations, indicating that they are largely independent of each other.

o The strongest negative correlation is between **Manufacturing Costs** and **Revenue Generated** (-0.21), suggesting that higher manufacturing costs may slightly reduce revenue.

o **Defect Rates** have a weak negative correlation with **Revenue Generated** (-0.13), indicating that higher defect rates may slightly reduce revenue.

**Executive Summary & Key Findings**

* Top 3 revenue-generating product categories (Skincare: ₹241K, Haircare: ₹174K, Cosmetics: ₹161K)
* Most profitable customer segment (Unknown/Male customers generate 35% more revenue than others)
* Critical supply chain bottlenecks (Haircare has 86 excess inventory days vs target of 30)
* Highest performing supplier (Supplier 1: 23% lower defects than average)

**Business Context Integration**

* Industry benchmarks (e.g., typical defect rates for beauty products are 1.5-3% - your average is 2.28%)
* Financial implications of findings (e.g., reducing haircare overstock could free up ₹X working capital)
* Strategic priorities alignment (e.g., how skincare focus supports company's premium positioning)

**Actionable Recommendations**

* **Inventory Optimization**
* **Immediate action:** Reduce haircare inventory by 75% (from 116 to 30 days coverage)
* **Process change:** Implement dynamic reorder points (High: Skincare=45 days, Medium: Haircare=30 days, Low: Cosmetics=20 days)
* **System upgrade**: Implement inventory health dashboard with alerts for "At Risk" SKUs

**Supplier Management**

* **Priority 1:** Expand partnership with Supplier 1 (3480 revenue/$cost vs avg 2000)
* **Priority 2:** Negotiate with Supplier 5 to reduce defect rates from 2.67% to <2%