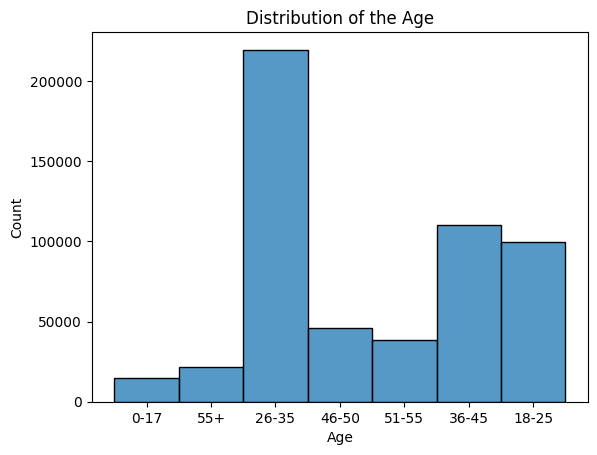
**Report**

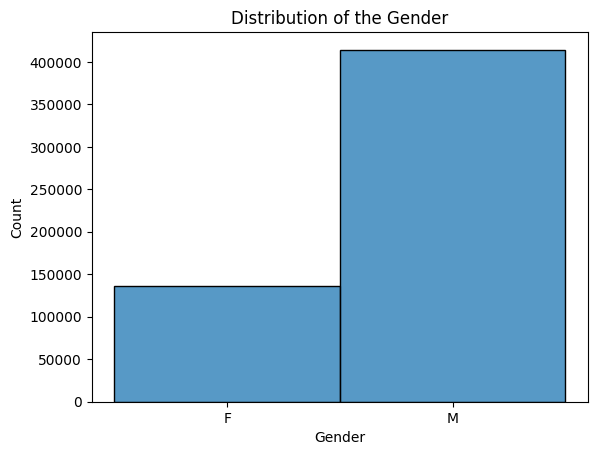
**Possible EDA cases –**

**1. Understanding User Demographics:**

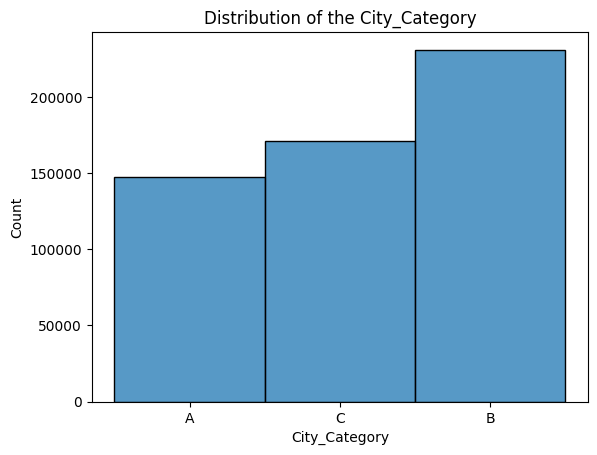
* **Distribution of Users: Analyze the distribution of users across Age groups, Gender, City Category, and Marital Status. Use histograms or bar charts to visualize the spread.**



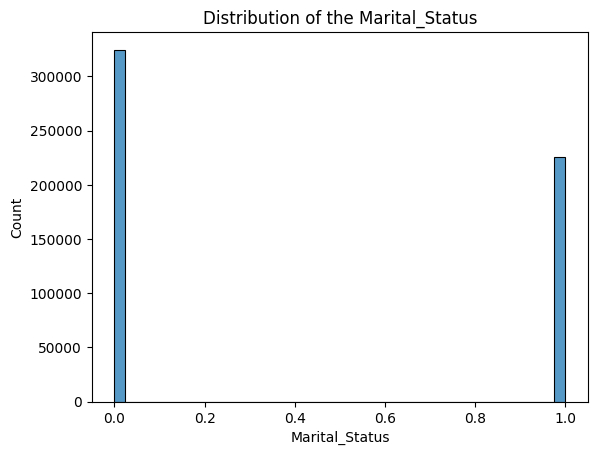
The bar chart titled "Distribution of the Age" illustrates the population count across different age groups, showing a significant concentration in the younger and middle-aged categories. The 26-35 age group has the highest count, exceeding 200,000 individuals, followed by the 36-45 group with slightly above 100,000. The 18-25 group also shows a substantial count of around 100,000. There is a noticeable decline in population for age groups 46-50 and 51-55, each slightly above 50,000, and the numbers drop further for those aged 55+, with counts below 50,000. The 0-17 age group is the least populous, with under 50,000 individuals. This distribution highlights a demographic skewed towards younger and middle-aged populations, with a significant decline in older age groups.



The bar chart titled "Distribution of the Gender" reveals a significant gender imbalance, with males numbering close to 400,000 and females slightly above 100,000. This indicates that males constitute the majority of the population in the dataset, outnumbering females by approximately four times. This distribution is important for analyzing gender representation, planning gender-specific programs, and addressing gender disparities in various contexts.

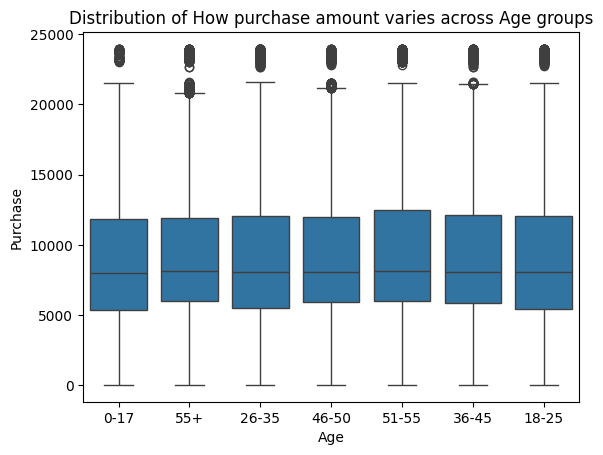


The bar chart titled "Distribution of the City\_Category" shows that City Category B has the highest population, with over 200,000 individuals, followed by City Category C with around 175,000, and City Category A with the lowest population slightly above 150,000. This indicates a clear hierarchy in population distribution, with the majority residing in City Category B. This data is valuable for targeting the audience.



The dataset shows a higher number of unmarried users compared to married users. The difference between the two groups is noticeable but not overwhelmingly large, indicating that both married and unmarried users are significant segments of Walmart's customer base. This insight can be useful for targeted marketing strategies, product placements, and understanding the customer demographics better.

* **User Age and Purchase: Investigate how purchase amount varies across Age groups. Consider boxplots or scatter plots to identify trends.**



### Analysis by Age Group:

1. **0-17 Years**:
   * Median purchase amount is around 7,500.
   * The IQR ranges approximately from 4,000 to 12,000.
   * There are several outliers above 20,000.
2. **18-25 Years**:
   * Median purchase amount is similar to the 0-17 group, around 7,500.
   * The IQR is roughly from 4,000 to 13,000.
   * This group also has outliers above 20,000.
3. **26-35 Years**:
   * Median purchase amount is slightly higher, around 9,000.
   * The IQR ranges from approximately 5,000 to 14,000.
   * Outliers are present above 20,000.
4. **36-45 Years**:
   * Median purchase amount is similar to the 26-35 group, around 9,000.
   * The IQR ranges from 5,000 to 14,000, indicating a similar spending pattern.
   * Outliers are also present above 20,000.
5. **46-50 Years**:
   * Median purchase amount is around 9,000.
   * The IQR ranges from about 5,000 to 14,000.
   * Outliers are observed above 20,000.
6. **51-55 Years**:
   * Median purchase amount is approximately 9,000.
   * The IQR is from 5,000 to 14,000.
   * Outliers are present above 20,000.
7. **55+ Years**:
   * Median purchase amount is around 9,000.
   * The IQR ranges from 5,000 to 14,000.
   * Outliers are present above 20,000.

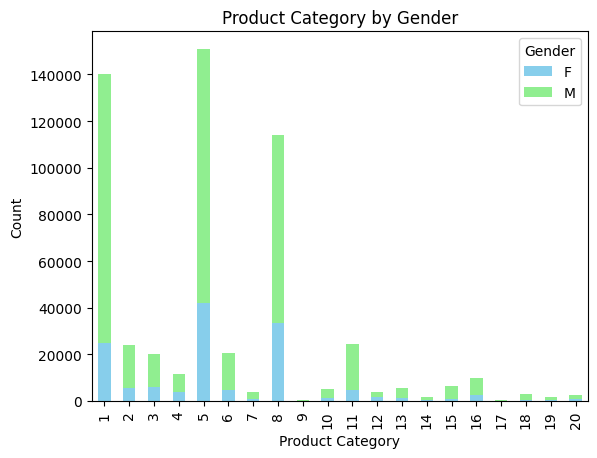
### General Observations:

* **Median Purchase Amount**: The median purchase amount across all age groups is relatively consistent, around 7,500 to 9,000.
* **Interquartile Range (IQR)**: The IQR is also consistent across age groups, ranging roughly from 4,000 to 14,000.
* **Outliers**: Each age group has several outliers, indicating that some users make significantly higher purchases than the majority.

The boxplot indicates that purchase behavior in terms of median and IQR is quite similar across different age groups. The presence of outliers in all age groups suggests that there are always some users with high purchase amounts, regardless of age.

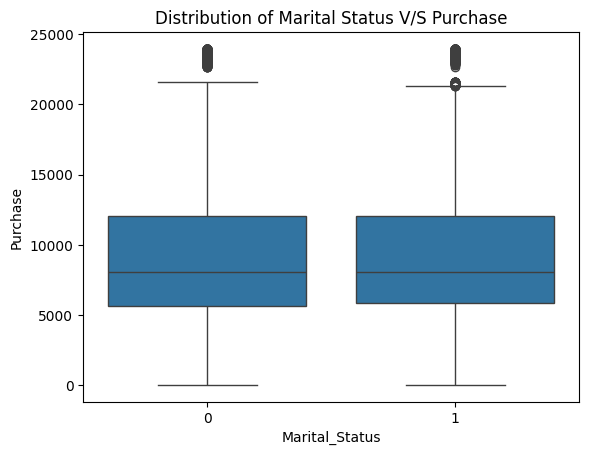
This analysis can help Walmart understand that while the average spending is consistent across age groups, there are high-value customers in every segment that might warrant targeted marketing strategies.

* **Gender and Product Category: Explore if there are any significant differences in the Product Category purchased between genders. You can use stacked bar charts or compare means.**



The bar chart "Product Category by Gender" reveals that males dominate purchases in most product categories, especially in categories 1, 5, and 8, which have the highest counts. While males lead in almost all categories, categories 10 to 20 have notably low purchase counts for both genders. These insights suggest that marketing strategies should focus on targeting males in high-purchase categories and investigating the reasons behind low engagement in the lower-purchase categories to boost overall sales.

* **Marital Status and Purchase: Analyze if Marital Status has any influence on the amount spent. This could be visualized with boxplots or scatter plots with Purchase amount on the Y-axis.**



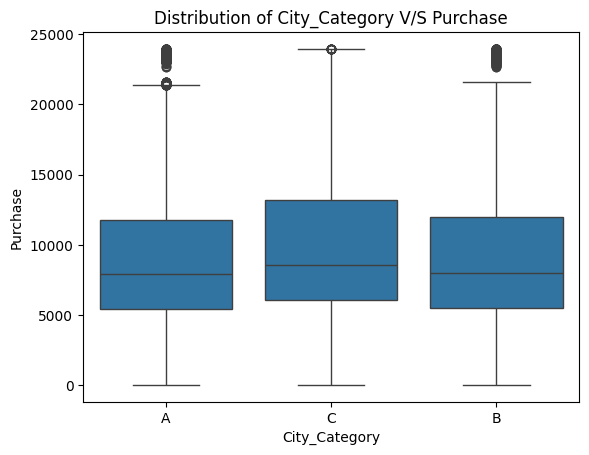
* + **Marital Status 0 (Single)**:
    - The median purchase amount is around 8000.
    - The IQR ranges approximately from 4000 to 13000.
    - There are several outliers above approximately 20000.
  + **Marital Status 1 (Married)**:
    - The median purchase amount is slightly lower than that for singles, around 7500.
    - The IQR ranges approximately from 4000 to 12500.
    - There are outliers above approximately 20000, similar to the single category.

**Comparison**:

* + The overall distribution of purchase amounts is similar for both single and married individuals.
  + Both groups have similar median purchase amounts, with a slightly lower median for married individuals.
  + The range of typical purchase amounts (IQR) is also similar for both groups.
  + Both groups exhibit outliers with high purchase amounts above 20000.

In summary, the box plot shows that marital status has a negligible effect on the distribution of purchase amounts at Walmart, with both single and married individuals exhibiting similar purchasing behaviors.

**2. User Location and Purchase Behavior:**

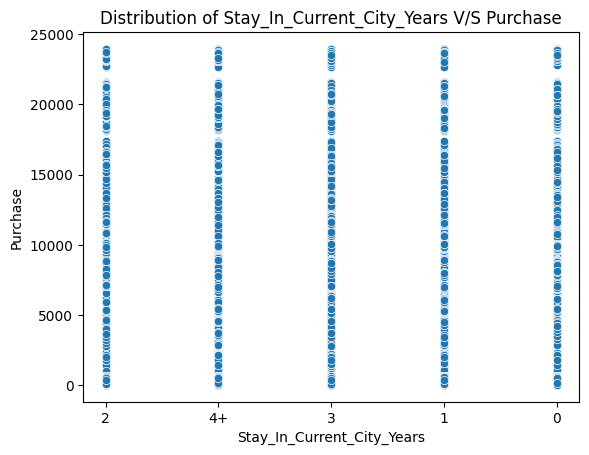
* **City Category and Purchase: See if there are spending differences between City Category groups (A, B, C). You can use boxplots or ANOVA tests to compare means.**
  + **City Category A**:
    - The median purchase amount is around 8000.
    - The IQR ranges approximately from 4000 to 13000.
    - There are several outliers above approximately 20000.
  + **City Category B**:
    - The median purchase amount is slightly lower, around 7000.
    - The IQR ranges approximately from 4000 to 12000.
    - There are outliers above approximately 20000.
  + **City Category C**:
    - The median purchase amount is around 8000.
    - The IQR ranges approximately from 4000 to 13000.
    - There are outliers above approximately 20000, similar to the other categories.

**Comparison**:

* + The overall distribution of purchase amounts is similar across all three city categories.
  + City Category A and C have slightly higher median purchase amounts compared to City Category B.
  + The range of typical purchase amounts (IQR) is also similar for all three groups.
  + All three groups exhibit outliers with high purchase amounts above 20000.

In summary, the box plot shows that the city category does not have a significant effect on the distribution of purchase amounts at Walmart, with all three city categories (A, B, and C) exhibiting similar purchasing behaviors.

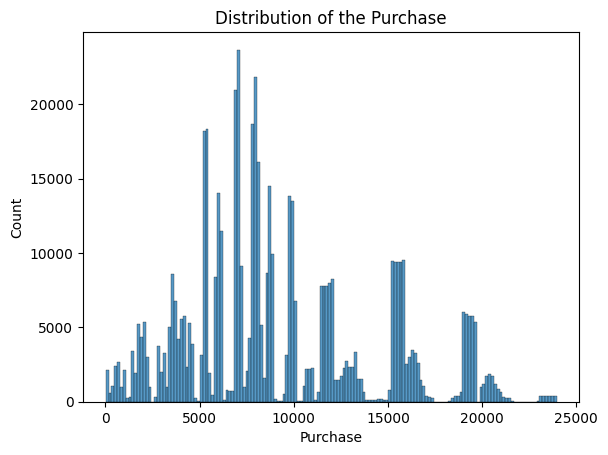
* **Length of Stay and Purchase:** Investigate if the number of years a user has stayed in their current city (StayInCurrentCityYears) impacts purchase amount. Utilize scatter plots or correlation analysis.



Based on the graph, it appears there is a weak positive correlation between the two variables. This means that as the number of years a user stays in their current city increases, the purchase amount also tends to increase. However, the data points are quite scattered, so the correlation is weak. There are many users who have stayed in the city for a long time but who also have low purchase amounts, and vice versa.

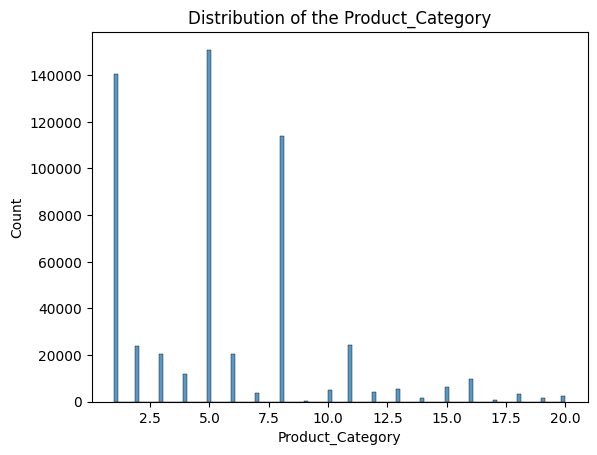
Overall, the scatter plot suggests a possible positive correlation between length of stay and purchase amount, but the correlation is weak and there are some outliers that may be affecting the results.

**User Purchase Behavior:**

* **Purchase Distribution:** Analyze the distribution of purchase amounts. Histograms are useful to identify patterns and potential outliers.

The histogram analysis reveals a distribution of purchase amounts primarily clustered between 0 and 10,000 units, with a notable peak around 5,000 units, suggesting a common purchase amount. While the spread extends to approximately 25,000 units, there are discernible gaps and spikes beyond 10,000 units, indicating varying purchasing patterns possibly linked to pricing tiers or product categories. The right-skewed distribution suggests a prevalence of lower-value purchases, though high-value outliers, especially beyond 20,000 units, highlight opportunities in premium products or bulk purchases. This data provides valuable insights for businesses, indicating potential pricing strategies and opportunities for catering to diverse customer preferences and behaviors.

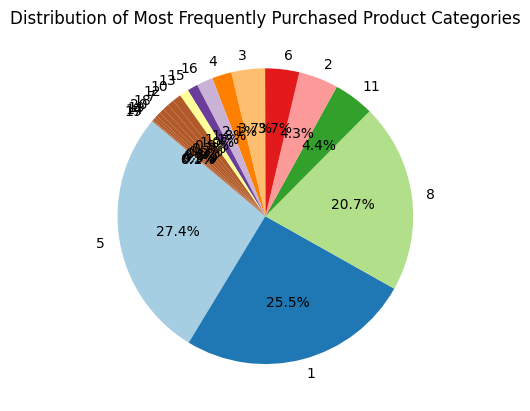
* **Frequency of Purchase:** If there's a time component in the data, use it to calculate purchase frequency per user and explore its distribution.



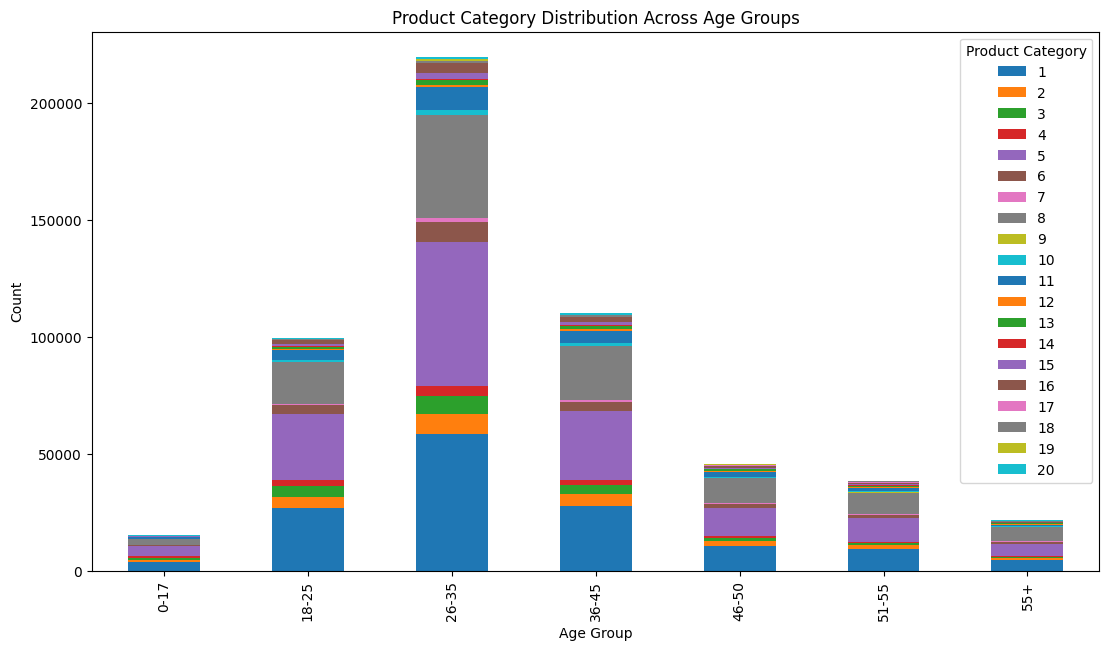
The bar chart "Distribution of the Product\_Category" shows a clear preference for certain product categories, with Category 5 leading at over 140,000 purchases, followed by Category 1 at around 120,000, and Category 8 at about 80,000. Categories 11 and 2 have moderate purchase counts, while the remaining categories have low engagement, each below 10,000 purchases. This data suggests that businesses should prioritize high-purchase categories in their marketing and inventory strategies, while investigating ways to boost engagement in less popular categories.

**4. Product Category Analysis:**

* **Distribution of Product Categories:** Identify the most frequently purchased product categories. Use bar charts or pie charts for visualization.

The pie chart "Distribution of Most Frequently Purchased Product Categories" shows a clear preference for certain product categories, with Categories 5, 1, and 8 accounting for nearly three-quarters of total purchases. Category 5 leads with 27.4%, followed by Category 1 at 25.5%, and Category 8 at 20.7%. Other categories have significantly smaller shares, indicating lower popularity. Businesses should prioritize these high-purchase categories in their marketing and inventory strategies while investigating ways to boost engagement in less popular categories.

* **Purchase by Category and Age:** Explore if there are interactions between Age groups and Product Category preferences. Consider using stacked bar charts or crosstabulation.



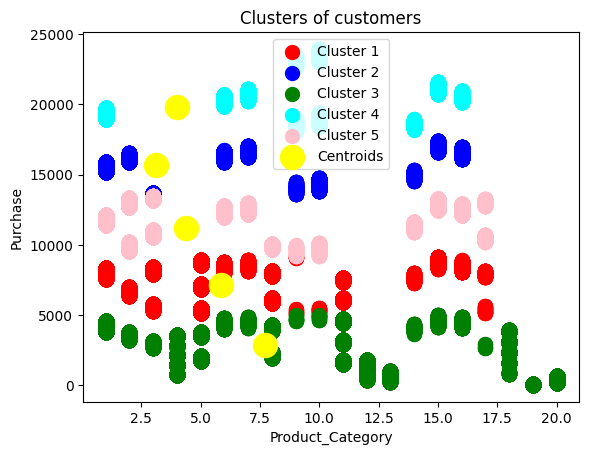
 **Age groups with similar preferences:** We can see that some age groups tend to favor similar product categories. For instance, the 26-35 and 36-45 age groups appear to purchase more products in categories 1, 3, and 10 compared to other age groups.

 **Age groups with distinct preferences:** Other age groups have distinct preferences. For example, the 0-17 age group purchases more products in category 19 compared to other age groups.

**Advanced Analysis:**

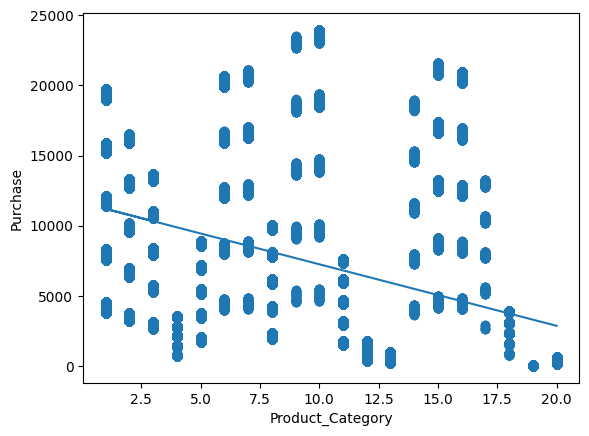
* **Segmentation:** Segment users based on demographics and purchase behavior to identify distinct customer groups. Techniques like K-Means clustering can be applied.

The graph shows five distinct customer segments based on their product category and purchase behavior.

* Cluster 1 appears in the lower left corner, indicating these customers have low scores on both product category and purchase behavior. They might be infrequent buyers who are not interested in a wide variety of products.
* Cluster 2 appears in the lower right corner, indicating these customers have a low product category score but a higher purchase behavior score. They might be frequent buyers who are not interested in a wide variety of products.
* Cluster 3 appears in the upper left corner, indicating these customers have a high product category score but a lower purchase behavior score. They might be infrequent buyers who are interested in a wide variety of products.
* Cluster 4 appears in the center of the graph, indicating these customers have moderate scores on both product category and purchase behavior. They might be regular buyers who are interested in a moderate variety of products.
* Cluster 5 appears in the upper right corner, indicating these customers have high scores on both product category and purchase behavior. They might be frequent buyers who are interested in a wide variety of products.

**Target Marketing:** This kind of segmentation can be helpful for targeted marketing campaigns. Marketers can tailor their messaging and product offerings to each customer segment based on their unique characteristics. For example, they might target Cluster 1 with special offers to try new products, and target Cluster 5 with loyalty programs and exclusive discounts.

* **Predictive Modeling:** Build models to predict purchase amount based on user characteristics or product categories. This could involve techniques like linear regression or decision trees.



* The regression line shows the average predicted purchase amount for each purchase category.
* Deviations from the line indicate how much a particular data point deviates from the average predicted value for its purchase category.

This kind of linear regression model can be a useful tool for predicting the purchase amount of a new customer based on their purchase category.