**Data Analysis and Interpretation**

This chapter presents the step-by-step analysis of the customer churn dataset using descriptive analytics techniques. The objective is to explore customer behavior, identify churn patterns, and highlight significant trends using Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn.

**1. Overview of the Dataset**

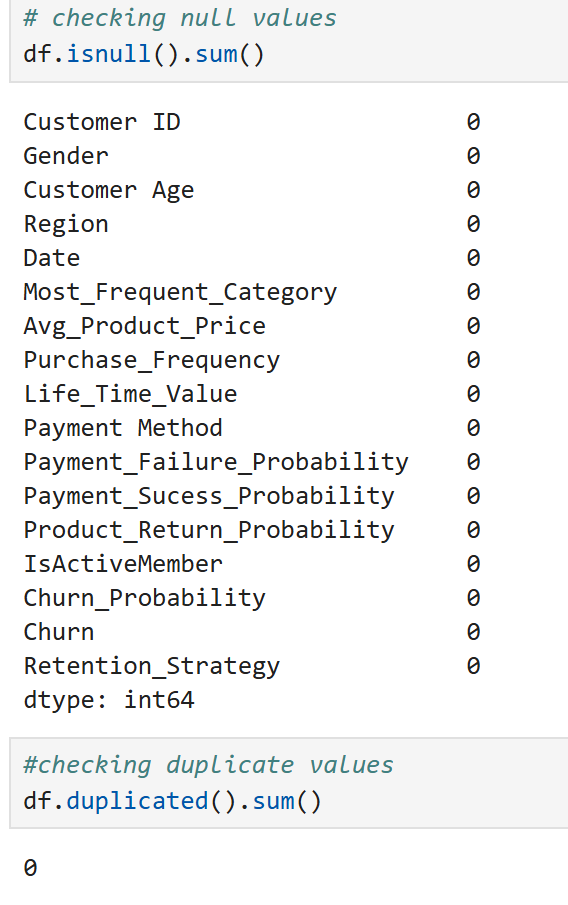
The DataFrame df contains a diverse set of features relevant to customer behavior and churn analysis. These include:

* **Identification:** Customer ID for unique identification.
* **Demographics:** Gender, Customer Age, Region.
* **Transaction History:** Date, Most\_Frequent\_Category, Avg\_Product\_Price, Purchase\_Frequency, Life\_Time\_Value.
* **Payment Information:** Payment Method, Payment\_Failure\_Probability, Payment\_Success\_Probability.
* **Product Interaction:** Product\_Return\_Probability.
* **Engagement:** IsActiveMember.
* **Target Variable & Related Prediction:** Churn\_Probability, Churn, Retention\_Strategy.

The data type of the index is 'object', indicating that the column names are stored as strings. This initial look at the columns suggests the dataset is rich with information that can be used to understand and predict customer churn.

**2 Checking for Null and Duplicate Values**

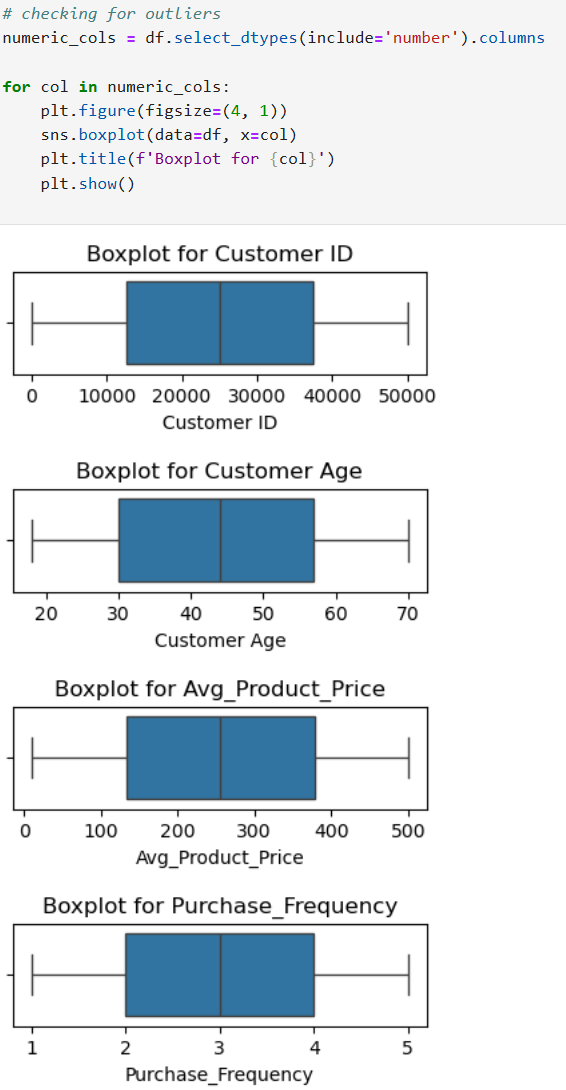
Before performing any analysis, it is important to ensure the quality and integrity of the dataset by checking for null (missing) values and duplicate entries.

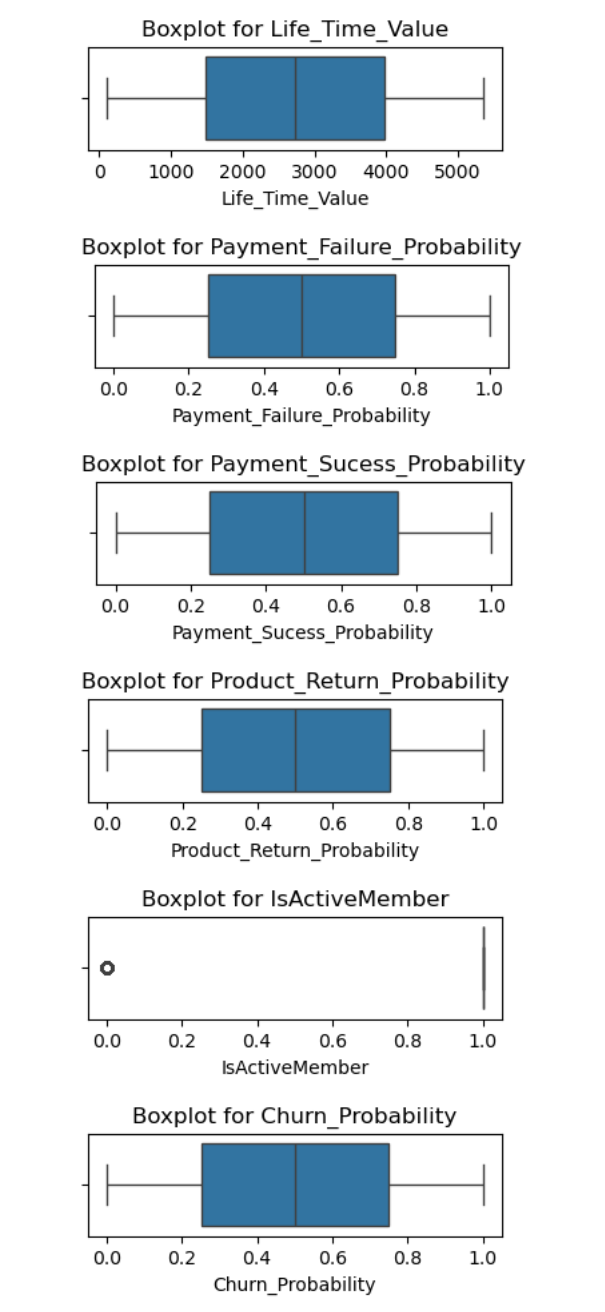
**Output:** The output revealed that there were no null values and no duplicate entries in the dataset, indicating that the data is complete, clean, and each customer record is unique—particularly important due to the presence of the Customer ID column.

**Conclusion:** The dataset is clean and ready for analysis without the need for data imputation or de-duplication processes.

**3. Outlier Detection (Using Boxplots)**

The boxplots provide the following insights about potential outliers:

****



**Output:**

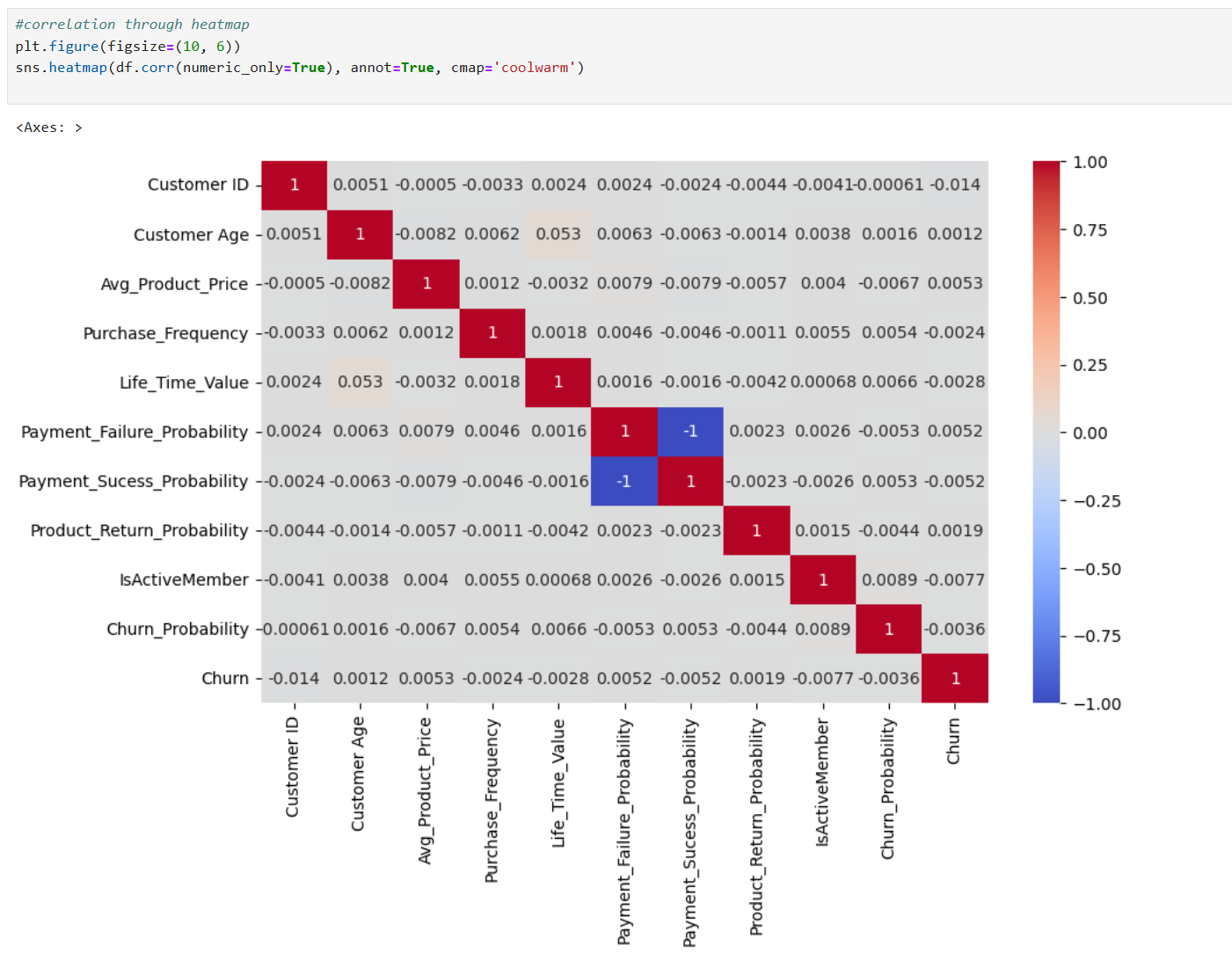
* Most numerical variables such as Customer Age, Avg\_Product\_Price, Purchase\_Frequency, Life\_Time\_Value, Payment\_\*\_Probability, Product\_Return\_Probability, and Churn\_Probability show no visible outliers in their boxplots.
* Customer ID is a unique identifier and does not require outlier analysis.
* A minor outlier is detected in the IsActiveMember column, indicating a small group of inactive customers (value = 0), which is valid and expected.

**Conclusion:**

The dataset is clean and well-behaved with **no significant outliers** that require treatment. All numerical features are **within acceptable ranges**, making the dataset ready for further exploratory or predictive analysis without additional preprocessing.

**4. Correlation Analysis**

To understand the relationships between numerical features in the dataset, a **correlation heatmap** was used. This visual representation helps identify how strongly pairs of variables are related, using the Pearson correlation coefficient, where values range from -1 (perfect negative correlation) to +1 (perfect positive correlation).

****

**Observations:**

* Most features exhibit very weak or near-zero correlation with one another, indicating that the dataset has minimal multicollinearity.
* A perfect negative correlation (-1) is observed between:
  + Payment\_Failure\_Probability and Payment\_Success\_Probability, which is expected since they represent opposite outcomes of the same event.
* Other features such as Churn\_Probability, Life\_Time\_Value, and Purchase\_Frequency show very low correlation with the target variable Churn, suggesting that customer churn may depend on a combination of subtle factors rather than any single strong predictor.
* Customer ID has no analytical relevance (correlation of 0 or near 0) and can be excluded from modeling.

**Conclusion:**

The correlation heatmap indicates that the dataset is well-balanced with no strong linear dependencies between features, except for one expected inverse relationship. Therefore, no immediate feature reduction is needed based on correlation, and all features may be retained for further analysis or modeling.

**5. Churn Rate Analysis**

To understand the proportion of customers who have discontinued using the service, we calculated the **churn rate** using the following formula:

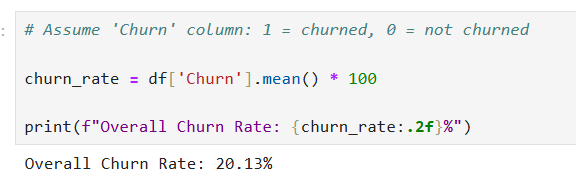
**Churn Rate=( Number of Churned Customers​ / Total Number of Customers)×100**

In the dataset, the Churn column is binary:

* 1 represents a churned customer,
* 0 represents an active customer.

**Calculations:**

* **Total Customers:** 47,687
* **Churned Customers:** 9,599
* **Churn Rate:**



**Interpretation:**

* Approximately **1 out of every 5 customers** has churned.
* This rate is **moderate**, indicating a potential area for improvement in customer retention.
* Further segmentation and behavior analysis can help identify high-risk customers and implement proactive retention strategies.

### **5.1 Churn Rate by Gender**

To analyze churn behavior based on gender, the dataset was grouped by Gender and Churn status. The results were then visualized using a bar chart to show how churn is distributed among male and female customers.

**Observation:**

The churn distribution across genders shows that male customers have a slightly higher churn rate (20.71%) compared to female customers (19.55%). Although the difference is marginal, it suggests that male customers may require slightly more attention in customer retention strategies.

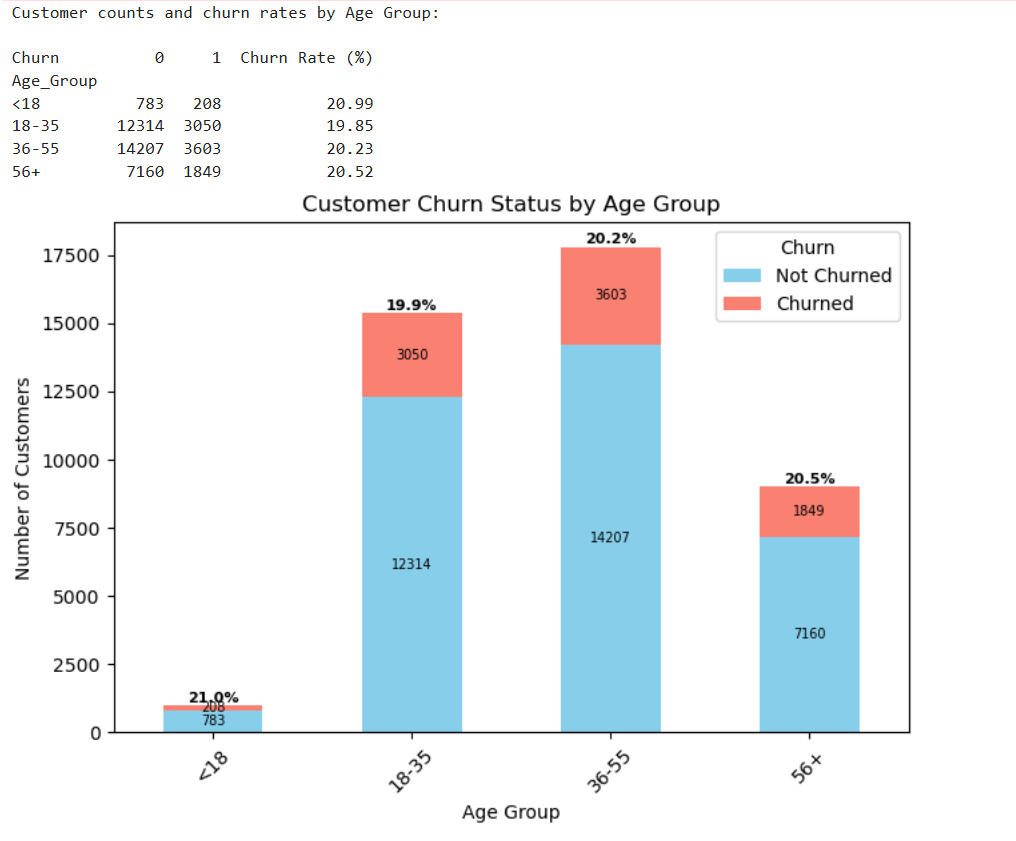
**5.2 Churn Rate by Gender**

In this analysis, we aim to understand **customer churn patterns based on gender**. Churn refers to customers who have stopped using a service, and analyzing churn by gender helps identify whether **male or female customers are more likely to leave**.

This insight is crucial for:

* Designing **targeted retention strategies**
* Improving **customer experience** based on gender behavior
* Making **data-driven business decisions** to reduce churn and improve customer loyalty

By grouping customers by gender and their churn status (churned or not), we can visualize and compare how churn is distributed between male and female customers.



**Observation:**

The churn rate across age groups is relatively consistent, ranging from **19.85% to 21.0%**.

* The **<18** age group has the **highest churn rate** at **21.0%**, indicating potentially lower brand loyalty or relevance.
* The **18–35** age group has the **lowest churn rate** at **19.85%**, suggesting stronger engagement or satisfaction.
* The **36–55** and **56+** age groups show moderate churn rates at **20.23%** and **20.52%**, respectively.

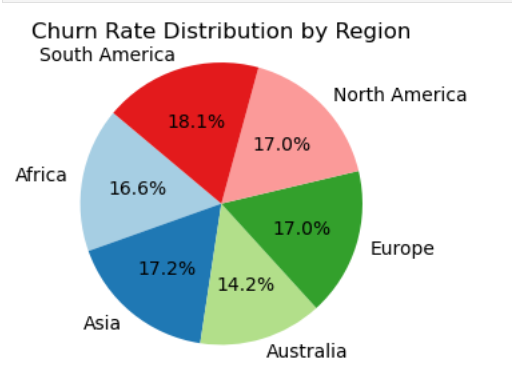
Overall, while churn rates are quite balanced, targeted retention efforts may be more necessary for the youngest and oldest customer segments.

**5.3 Churn Rate Distribution by Region**

This pie chart visualizes the **percentage of customer churn from different regions**. Understanding churn distribution across regions helps businesses identify:

* **Geographical areas with higher churn rates**
* Potential **regional issues** such as service quality, competition, or customer dissatisfaction
* Where to **focus retention efforts or marketing strategies** to reduce churn

By analyzing churn regionally, companies can tailor their actions to improve customer satisfaction and minimize churn based on regional behaviors and needs.

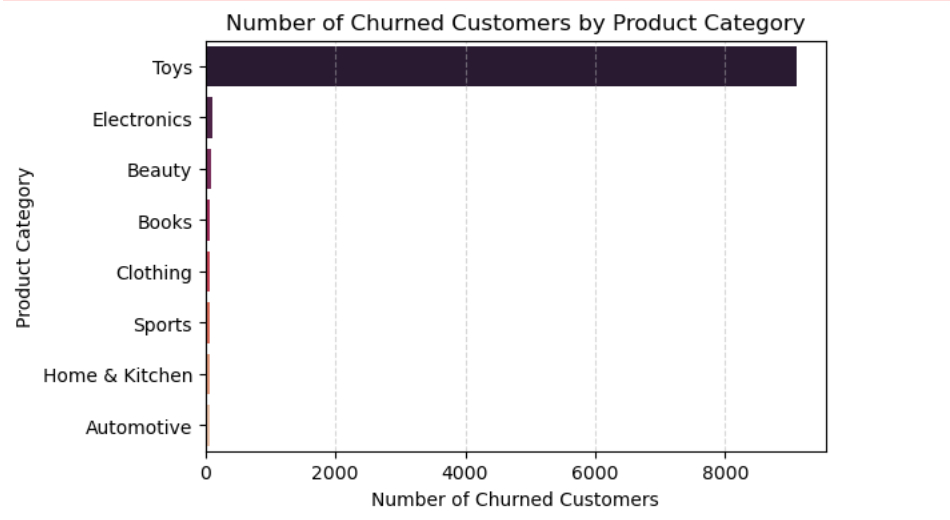


**Observations:**

1. **South America** has the **highest churn rate** at **18.1%**, indicating a significant loss of customers in that region.
2. **Asia (17.2%)**, **Europe (17.0%)**, and **North America (17.0%)** have **moderate churn rates**, suggesting room for improvement in customer retention.
3. **Africa** shows a churn rate of **16.6%**, slightly lower than the average.
4. **Australia** has the **lowest churn rate** at **14.2%**, indicating relatively better customer satisfaction or loyalty in that region.

**5.4 Churned Customers by Product Category**

This visualization, a horizontal bar chart, was utilized to illustrate and compare the absolute number of customers who churned within each of the identified product categories. The objective was to determine which product segments experienced the greatest magnitude of customer attrition, thereby facilitating a targeted approach to investigate category-specific churn drivers and prioritize retention efforts.

****

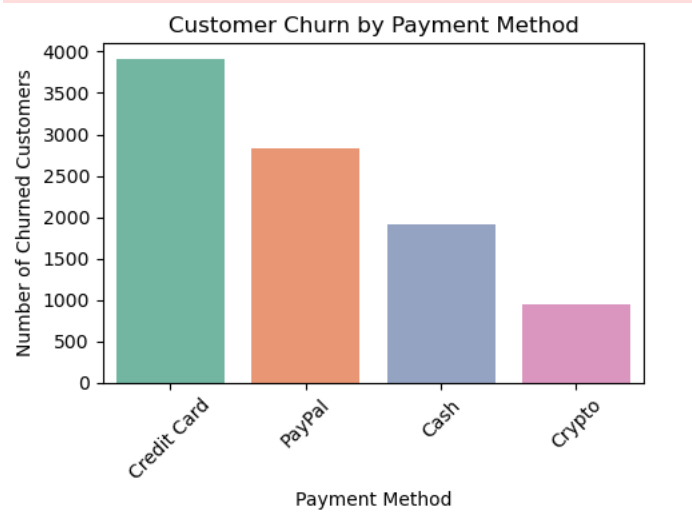
**Observations:**

* The "**Toys**" category exhibits a significantly higher number of churned customers **(over 9000)** compared to all other product categories.
* The number of churned customers in categories such as "Electronics," "Beauty," "Books," "Clothing," "Sports," "Home & Kitchen," and "Automotive" is substantially lower (below 200 in each).
* Retention efforts or investigations into churn drivers might be particularly beneficial if focused on customers who frequently purchase or interact with the "Toys" category.

The "Toys" category exhibits a conspicuously higher number of churned customers compared to all other product categories within the dataset. In stark contrast, the "Electronics," "Beauty," "Books," "Clothing," "Sports," "Home & Kitchen," and "Automotive" categories demonstrate significantly lower counts of customer attrition. This statistically notable disparity suggests a potential association between the "Toys" product category and an increased likelihood of customer churn. Consequently, future research should focus on exploring factors unique to the "Toys" customer segment, such as product-specific issues, customer expectations, or competitive pressures, to develop effective and targeted retention strategies aimed at mitigating the observed elevated churn rate within this high-volume attrition area.

**5.5 Churned Customers by Payment Method**

The reason for analyzing the distribution of churned customers by payment method is to **identify potential correlations or associations between the payment method chosen by a customer and their likelihood of churning.**

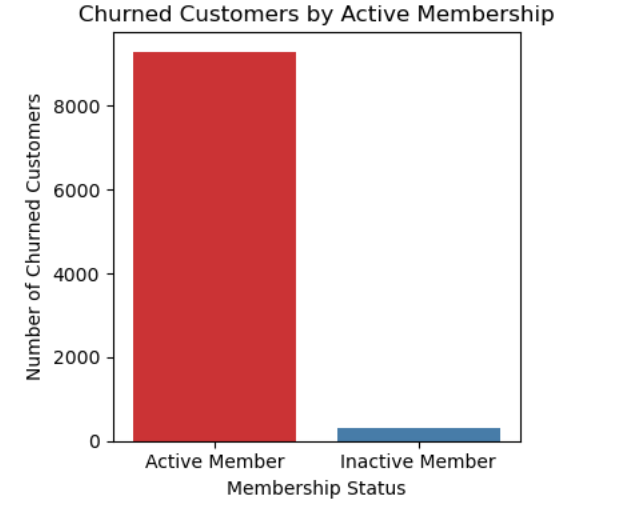
****

**Observations:**

* **Credit Card Dominance:** The "Credit Card" payment method has the highest number of churned customers, significantly exceeding other methods.
* **PayPal Second Highest:** "PayPal" accounts for the second-highest number of churned customers, though considerably less than "Credit Card."
* **Lower Churn with Cash and Crypto:** "Cash" and "Crypto" payment methods show substantially lower numbers of churned customers compared to "Credit Card" and "PayPal."

**4.5.6 Churned Customers by Active Membership Status**

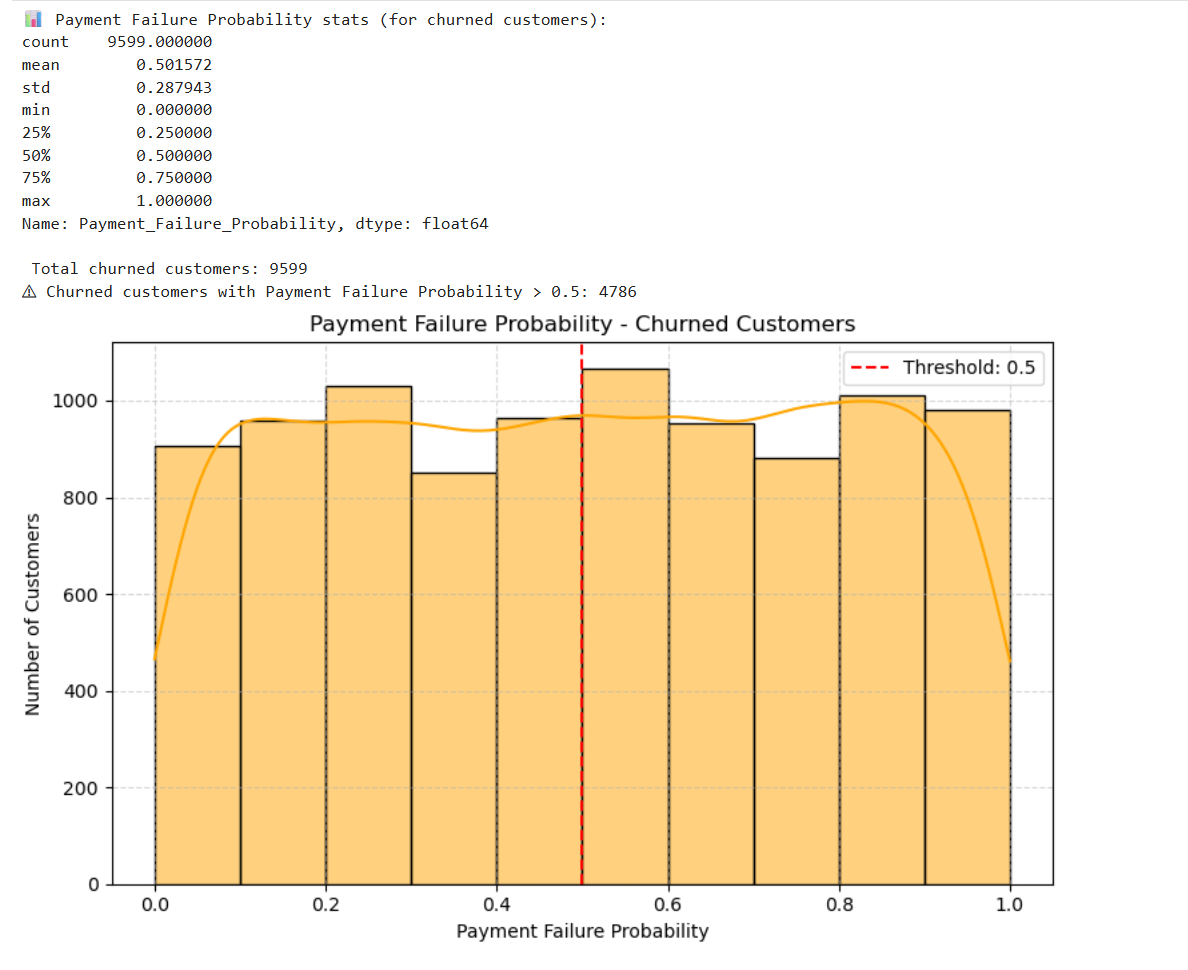
The reason for analyzing the number of churned customers based on their active membership status is to **understand the relationship between customer engagement (as indicated by active membership) and the likelihood of churn.** This analysis aims to determine if active members are less likely to churn compared to inactive members, or if the opposite is true. Understanding this relationship is crucial for evaluating the effectiveness of membership programs and identifying whether engagement is a significant factor in customer retention. It can also highlight potential areas for intervention to improve retention among both active and inactive segments.



**Observations:**

* **Higher Churn Among Active Members:** The number of churned customers who were active members is significantly higher than the number of churned customers who were inactive members.
* **Substantially Fewer Inactive Churners:** The count of churned inactive members is very low in comparison to churned active members.

**5.7 Distribution of Payment Failure Probability Among Churned Customers**

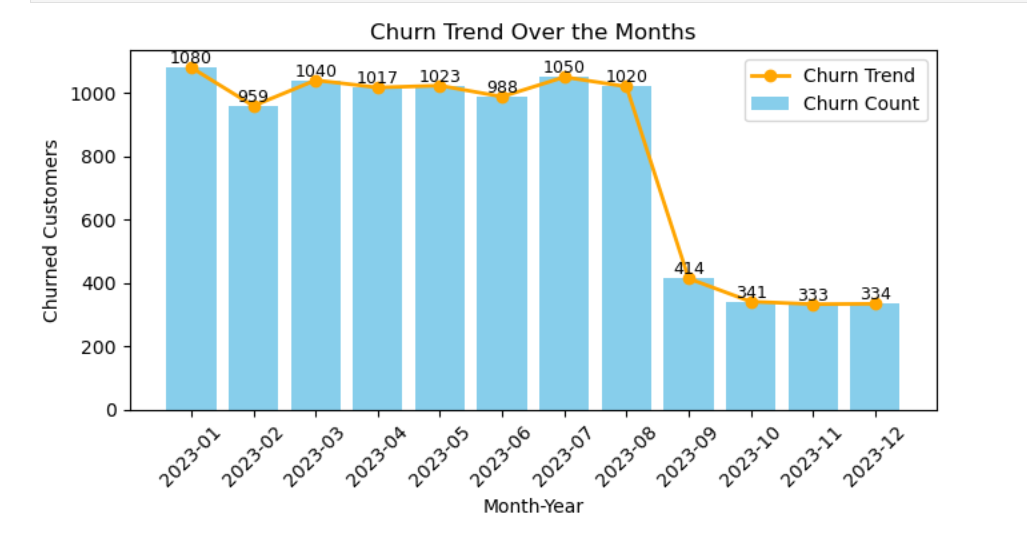
****The reason for analyzing the distribution of payment failure probability among churned customers is to **investigate the potential relationship between payment issues and customer attrition.** By examining the range and frequency of payment failure probabilities for customers who ultimately churned, we aim to understand if a higher likelihood of payment failure is associated with an increased propensity to churn. This can help identify whether payment-related friction points contribute to customer dissatisfaction and eventual departure.

**Observations:**

* **Relatively Uniform Distribution:** The histogram shows a relatively even distribution of payment failure probabilities across the entire range (0.0 to 1.0) for churned customers.
* **Mean Around 0.5:** The mean payment failure probability for churned customers is approximately 0.50, suggesting an average likelihood of payment failure at the midpoint of the probability scale.
* **Substantial Number Above 0.5:** A significant portion of churned customers (4786 out of 9599) had a payment failure probability greater than 0.5.

**6. Monthly Customer Churn Trend in 2023**

The reason for analyzing the churn trend over time is to **identify temporal patterns and fluctuations in customer attrition.** By examining the number of customers churning each month, we can detect periods of high or low churn, understand the stability of customer retention, and potentially correlate these trends with specific events, campaigns, or seasonal factors that might influence customer behavior. This longitudinal perspective is crucial for evaluating the effectiveness of retention strategies and predicting future churn patterns.



**Observations:**

* **Stable High Churn (Jan-Jul):** The number of churned customers remained relatively high and fluctuated within a narrow range (approximately 959 to 1080) during the first seven months of 2023.
* **Significant Drop in August:** A substantial and abrupt decrease in churn occurred in August 2023.
* **Sustained Low Churn (Aug-Dec):** Following the drop in August, the number of churned customers remained consistently low (between 333 and 414) for the remainder of the year.

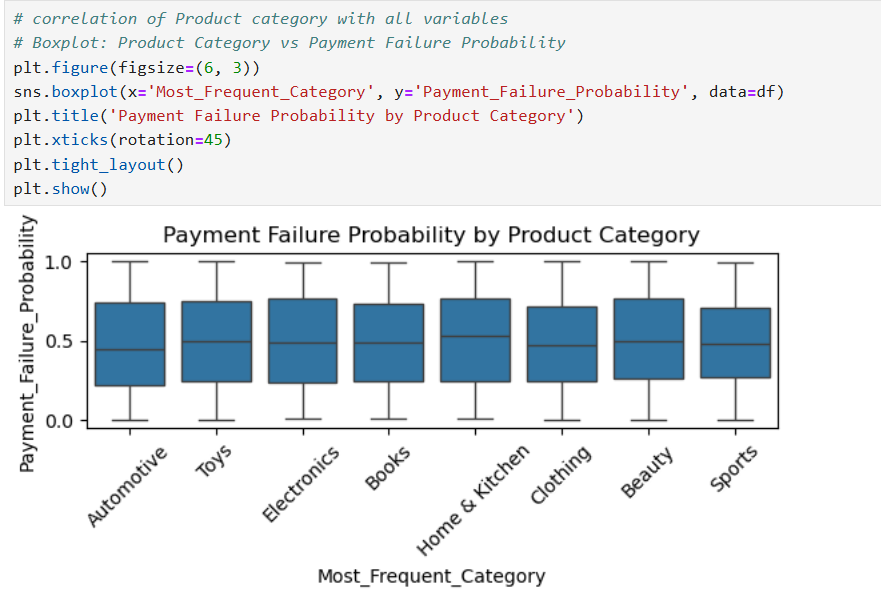
**7. correlation of Product category with all variables**

The reason for examining the relationship between the "Most\_Frequent\_Category" and other variables is to **identify potential associations or dependencies that might influence customer behavior, including churn.** By analyzing how different customer attributes (like payment failure probability, age, region, gender, and payment method) are distributed across various product categories, we can uncover insights such as:

* **Are customers who frequently buy certain types of products more likely to experience payment failures?**
* **Does the age or gender demographic of customers vary significantly based on their preferred product category?**
* **Are there regional preferences for certain product categories, and does this correlate with other behaviors?**
* **Do customers who favor specific product categories tend to use particular payment methods?**

Understanding these correlations can help in segmenting customers more effectively, tailoring marketing efforts, identifying potential risk factors for churn within specific product-focused groups, and ultimately developing more targeted and effective retention strategies. It allows for a more nuanced understanding of the customer base beyond simple demographic or transactional data.

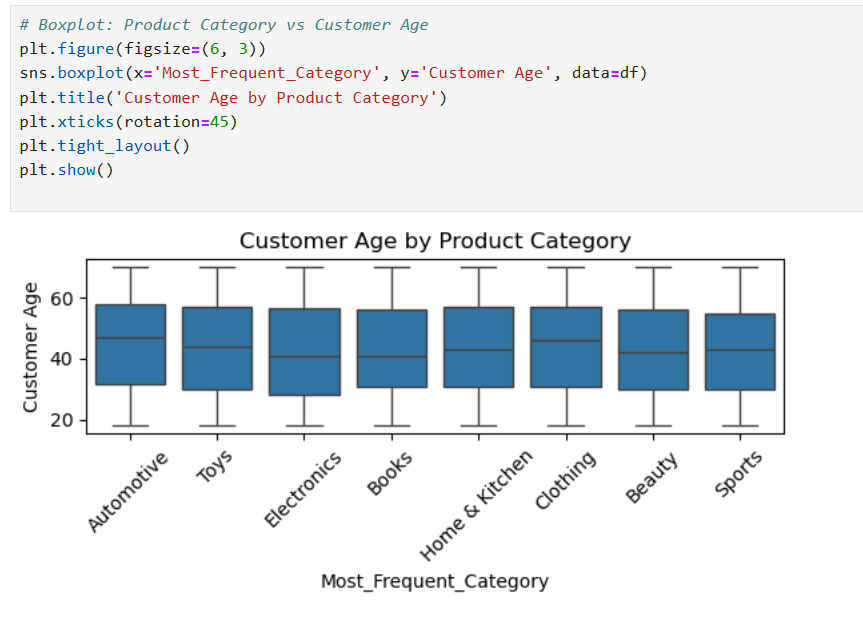
**7.1 Payment Failure Probability Distribution Across Product Categories**

The reason for analyzing the distribution of payment failure probability across different product categories is to **determine if there is a relationship between the type of product a customer frequently purchases and the likelihood of their payment failing.** Understanding such a relationship could indicate potential issues with payment processes specific to certain product categories or highlight customer segments with different payment behaviors based on their preferred purchases.

**Observations:**

* Median payment failure probability is consistent across product categories.
* Variability (IQR) in payment failure probability is similar across categories.
* Outliers in payment failure probability exist in most categories.
* Minor differences in the range of payment failure probability are observed for some categories (e.g., "Books," "Home & Kitchen").

**7.2 Customer Age Distribution Across Product Categories**

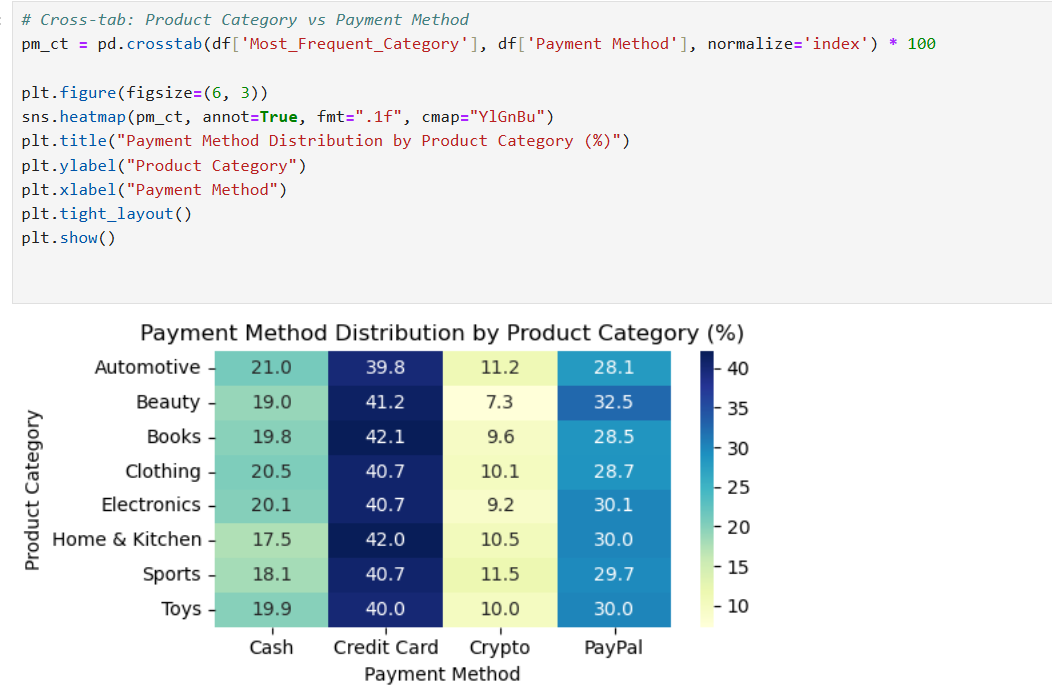
The reason for analyzing the distribution of customer age across different product categories is to **identify if there are any significant age-related preferences or patterns in the types of products customers frequently purchase.** Understanding the age demographics associated with different product categories can inform targeted marketing strategies, product development efforts, and customer segmentation. It helps determine if certain age groups are more inclined towards specific product types.

**Observations:**

* Median customer age is consistent across product categories (mid-30s to early 40s).
* Age variability within each product category is comparable.
* Most categories show a broad range of customer ages (20s to 60s/70s).
* Minor differences exist in the extreme age ranges for some categories (e.g., "Toys" potentially having a younger lower age).

**7.3 Payment Method Distribution by Product Category (%)**

The reason for analyzing the distribution of payment methods across different product categories is to **identify if there are preferred payment methods associated with specific product types.** Understanding these preferences can inform marketing strategies, optimize payment options offered for different product categories, and potentially reveal insights into the customer segments purchasing those products.

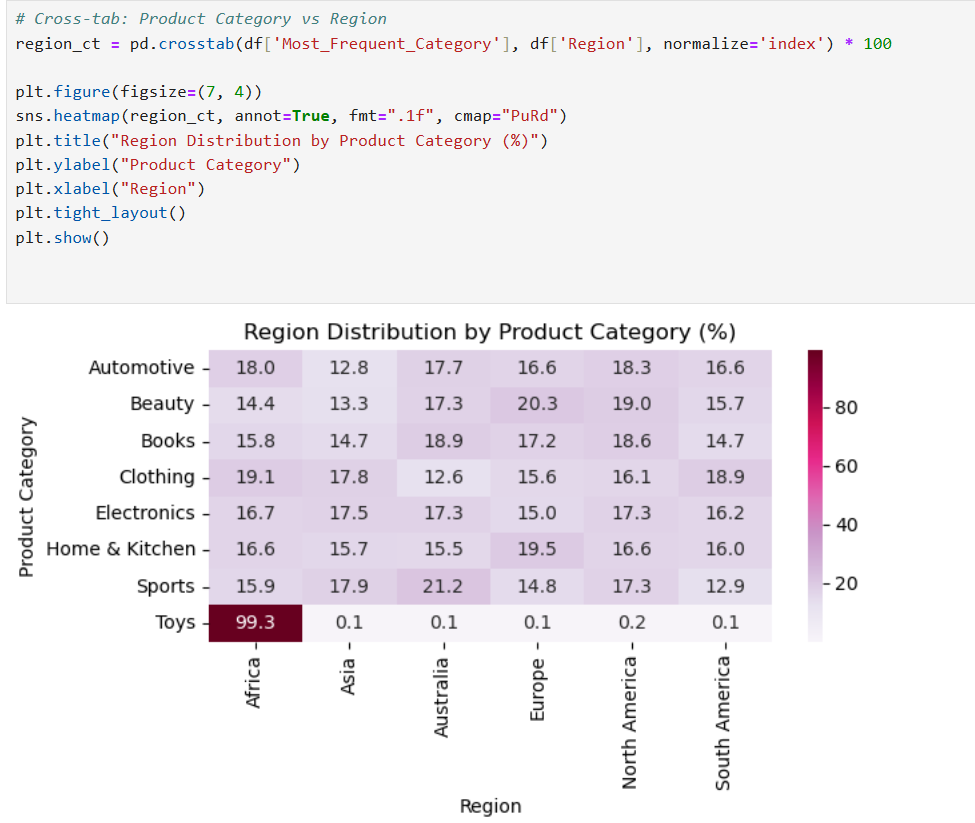
****

**Observations:**

* **Credit Card Dominance:** "Credit Card" is a consistently popular payment method across all product categories, generally ranging from approximately 39.8% to 42.1%.
* **Varied Cash Usage:** The use of "Cash" varies more significantly, with "Automotive" showing the highest percentage (21.0%) and "Home & Kitchen" the lowest (17.5%).
* **Low Crypto Adoption:** "Crypto" consistently has the lowest usage across all product categories, ranging from 7.3% to 11.5%.
* **Consistent PayPal Usage:** "PayPal" usage is relatively stable across categories, ranging from 28.1% to 32.5%.

**7.4 Region Distribution by Product Category (%)**

The reason for analyzing the distribution of regions across different product categories is to **identify if there are geographical preferences or concentrations for specific product types.** Understanding these regional patterns can be valuable for tailoring marketing campaigns, optimizing supply chain logistics, and gaining insights into market-specific demands for different product categories.

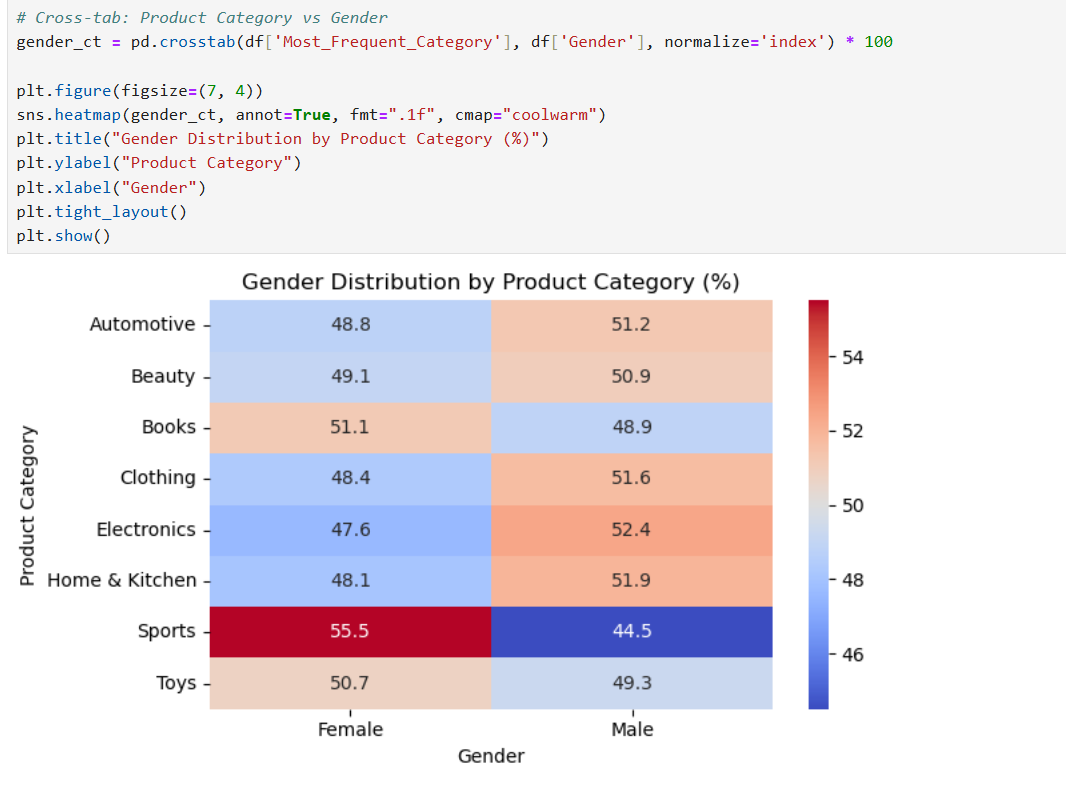
****

**Observations:**

* **Dominant "Toys" in Africa:** An overwhelmingly high percentage (99.3%) of "Toys" purchases are attributed to the "Africa" region.
* **Relatively Even Distribution for Most Categories:** For most other product categories (Automotive, Beauty, Books, Clothing, Electronics, Home & Kitchen, Sports), the distribution across different regions is more balanced, with no single region dominating.
* **Some Regional Skews:** While mostly balanced, some minor regional skews exist. For example, "Beauty" shows a slightly higher percentage in "Europe" (20.3%), and "Sports" has a relatively higher percentage in "Australia" (21.2%).
* **Low Presence in Asia for "Toys":** The "Asia" region shows a very low percentage (0.1%) for the "Toys" category.

**7.5 Gender Distribution by Product Category (%)**

The reason for analyzing the gender distribution across different product categories is to **identify if there are any significant gender-based preferences for specific product types.** Understanding these preferences can inform targeted marketing campaigns, product positioning, and the overall understanding of the customer base for each category.

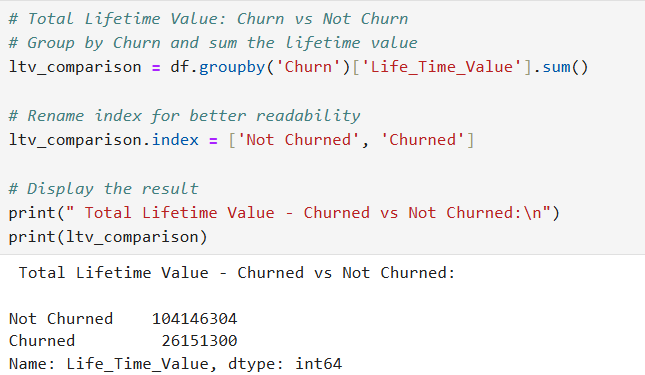


**Observations:**

* **Relatively Balanced Distribution:** For most product categories (Automotive, Beauty, Books, Clothing, Electronics, Home & Kitchen, Toys), the gender distribution is relatively balanced, with the percentage of female and male customers being close to 50%.
* **Slight Male Skew in "Electronics":** The "Electronics" category shows a slightly higher percentage of male customers (52.4%) compared to female customers (47.6%).
* **Slight Female Skew in "Sports":** The "Sports" category exhibits a slightly higher percentage of female customers (55.5%) compared to male customers (44.5%).

**8. Total Lifetime Value Comparison: Churned vs. Non-Churned Customers**

The reason for comparing the total lifetime value (LTV) of churned and non-churned customers is to **quantify the overall financial impact of customer attrition on the business.** By calculating and comparing the cumulative value these two groups of customers would have potentially contributed over their lifetime, we can directly assess the revenue lost due to churn. This provides a clear understanding of the economic consequences of customer attrition and underscores the importance of retention efforts.

****

The analysis reveals a significant financial impact resulting from customer churn. The total lifetime value of customers who churned (26,151,300) is substantially lower than that of customers who did not churn (104,146,304). This disparity underscores the direct revenue loss associated with customer attrition.

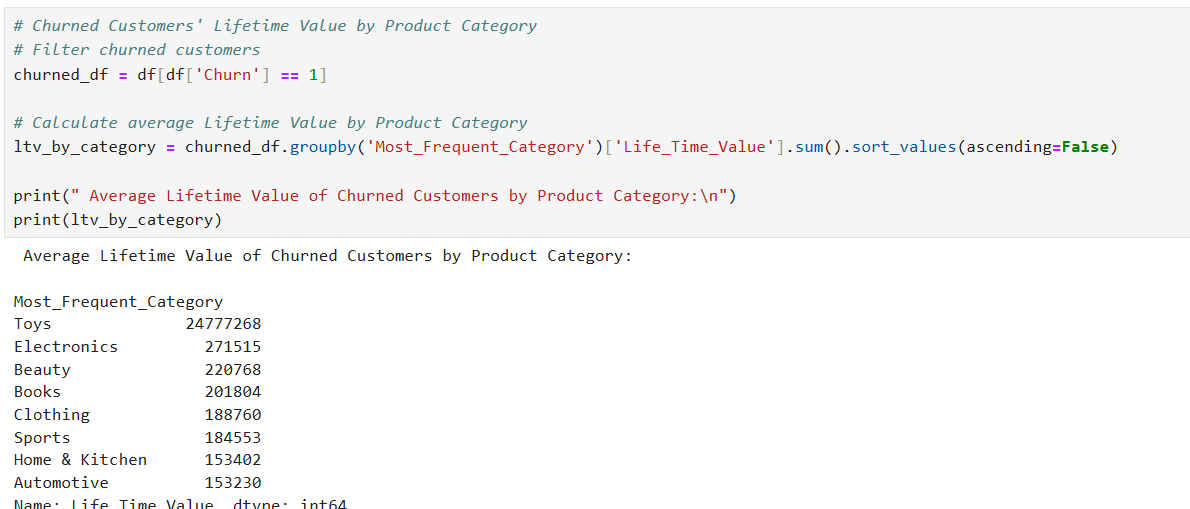
**Impacts:**

* Significant Revenue Loss: The ₹26,151,300 (assuming the currency is consistent throughout the dataset) in lost lifetime value from churned customers represents a considerable amount of potential revenue that the business will not realize.
* Reduced Overall Customer Value: Churn diminishes the overall value of the customer base. Retaining these customers would have significantly increased the total lifetime value.
* Focus on Acquisition: The loss of this value necessitates a greater emphasis on customer acquisition to compensate for the lost revenue, which is typically more expensive than retaining existing customers.
* Potential for Improvement: The large difference highlights the significant opportunity to improve the business's financial performance by implementing effective churn reduction strategies and retaining customers for longer durations.

In essence, the churned customers represent a substantial amount of unrealized potential revenue, emphasizing the critical need to understand and address the drivers of churn to mitigate these financial losses and enhance long-term profitability.

**8.1 Total Lifetime Value of Churned Customers by Product Category**

The reason for analyzing the total lifetime value of churned customers by their most frequent product category is to **identify which product segments are associated with the highest financial loss due to customer attrition.** By understanding the value of churned customers within each category, the business can prioritize retention efforts towards segments where churn has the most significant revenue impact.



The analysis of the total lifetime value of churned customers across different product categories highlights the varying financial consequences of attrition in each segment.

**Impacts:**

* **Significant Financial Risk in "Toys":** The disproportionately high total lost lifetime value associated with churned "Toys" customers signifies the greatest financial risk. Losing customers who frequently purchase toys has a much more substantial revenue impact compared to other categories.
* **Considerable Loss in "Electronics":** The relatively high lost lifetime value in the "Electronics" category also indicates a significant financial impact from churn within this segment.
* **Lower but Still Relevant Losses:** While the financial loss from churn in other categories is lower individually, the cumulative impact across all categories is still significant and contributes to the overall revenue leakage.
* **Prioritization for Retention:** The findings emphasize the need to prioritize retention efforts towards customers who frequently purchase "Toys" and "Electronics" to mitigate the most significant financial losses due to churn.

In essence, understanding the lifetime value of churned customers by product category allows the business to focus its retention strategies on the most financially valuable customer segments at risk of attrition, thereby maximizing the return on investment for churn reduction initiatives.