Data Insights Report

**Team Name:** Data Hunters

**Team Members:** Sarkar Tuhassinul Arnob, MD Shajada Sazid

**Challenge:** 1st Data Innovators Challenge 2024

# 1. Introduction

* **Objective:** The objective of this analysis is to understand customer behavior, identify patterns that contribute to customer loyalty (return customers), and provide insights into customer satisfaction, payment methods, and purchase trends. By analyzing the dataset, we aim to uncover factors influencing return customers and examine various attributes like purchase amount, time spent on the website, device usage, and payment methods.

**Key Goals:**

* Identify the factors contributing to customers being classified as return customers.
* Investigate how payment methods impact customer satisfaction and return rates.
* Analyze the relationship between customer location and both purchase amount and delivery time.
* Uncover actionable insights to improve customer retention and satisfaction.

# 2. Data Overview

**Dataset Information**:

* **Number of Entries (Rows)**: The dataset contains **10000** entries representing individual customer records.
* **Number of Features (Columns)**: There are **16 columns** or features in the dataset, each representing different aspects of customer behavior and transactions.

**Features Description:**

1. **Customer ID:** Unique identifier for each customer.
2. **Age:** Age of the customer.
3. **Gender:** Gender of the customer.
4. **Location:** The geographical location of the customer.
5. **Product Category:** The category of the product purchased by the customer.
6. **Purchase Amount ($):** The amount spent by the customer during the purchase.
7. **Time Spent on Website (min):** The time (in minutes) the customer spent on the website before purchasing.
8. **Device Type:** The device used to make the purchase (e.g., mobile, desktop, tablet).
9. **Payment Method:** The payment method used (e.g., credit card, digital wallet, etc.).
10. **Discount Availed:** Indicates whether the customer availed a discount during the purchase.
11. **Number of Items Purchased:** The number of items bought during the transaction.
12. **Return Customer:** Boolean value indicating if the customer is a returning customer.
13. **Review Score (1-5):** The review score given by the customer after the purchase.
14. **Delivery Time (days):** The time taken (in days) to deliver the product.
15. **Subscription Status:** Indicates if the customer is a free or premium subscriber.
16. **Customer Satisfaction:** A score representing the overall satisfaction of the customer.

# 3. Data Cleaning Process

**Step 1**: **Handling Missing Values**

* + There were no missing values in the dataset, so no imputation or removal of missing data was necessary.

**Step 2**: **Removing Duplicates**

* + No duplicates were found. Typically, removing duplicates is essential to prevent skewing the analysis, especially if the same customer data appears more than once.

**Step 3**: **Data Type Conversions**

* + All columns were already correctly typed. If required, conversions would involve changing **numerical columns** to their proper types (e.g., float or int).

**Final Dataset**:

* The final dataset contains **10000** **rows** and **16 columns**, with all features in their correct formats, ready for analysis.

# **4. Analysis and Insights**

## Level 1: Basic Insights (14 Marks):

**Q1**: Find Mean, Median, and Mode (Age)

* **Mean**: Mean value of Age of the dataset is: **43.7899**
* **Median**: Median value of Age of the dataset is: **44.0**
* **Mode**: Mode value of Age of the dataset is: **51**

**Q2:** Find Variance, Standard Deviation, and Z-Score (Purchase Amount)

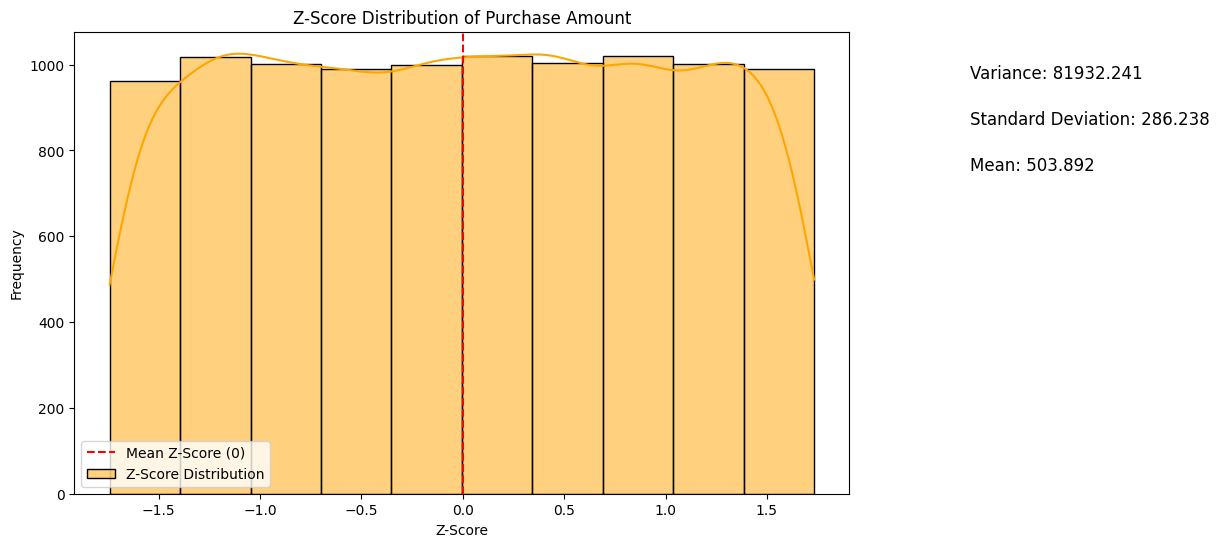
* **Variance of Purchase Amount:** 81932.241
* **Standard Deviation of Purchase Amount**: 286.238

**Z-Score Explanation**:  
The z-score, or standard score, is a measure that indicates how many standard deviations a data point is from the mean of the dataset. It helps identify how extreme or typical a particular value is within the context of the entire dataset.

* A z-score close to 0 indicates that the data point is close to the mean.
* A positive z-score means the data point is above the mean.
* A negative z-score indicates that the data point is below the mean.

In our dataset:

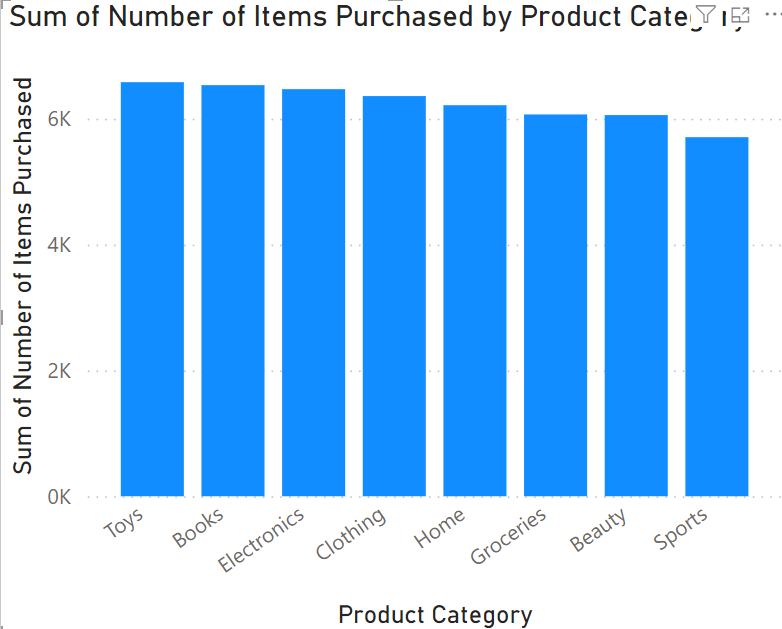
* **Minimum Z-Score**: -1.724
* **Maximum Z-Score**: 1.733



This indicates that the purchase amounts in the dataset range from being **1.724 standard deviations below the mean** to **1.733 standard deviations above the mean**, showing that most purchases fall within a relatively typical range from the average.

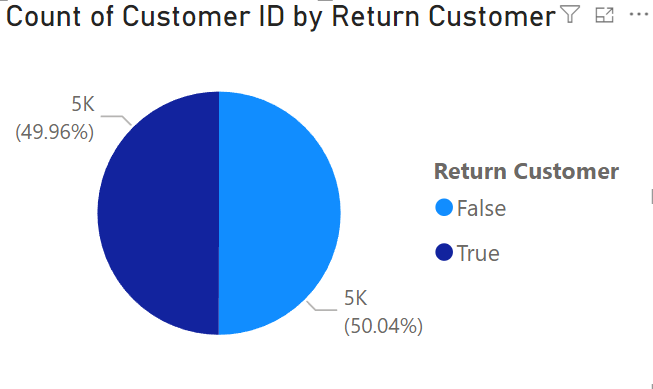
**Q3**: Top Three Product Categories Based on Purchases

* **1st**: Toys
* **2nd**: Books
* **3rd**: Electronics



**Q4**: Number of Return Customers

* **Total Return Customers**: 4996

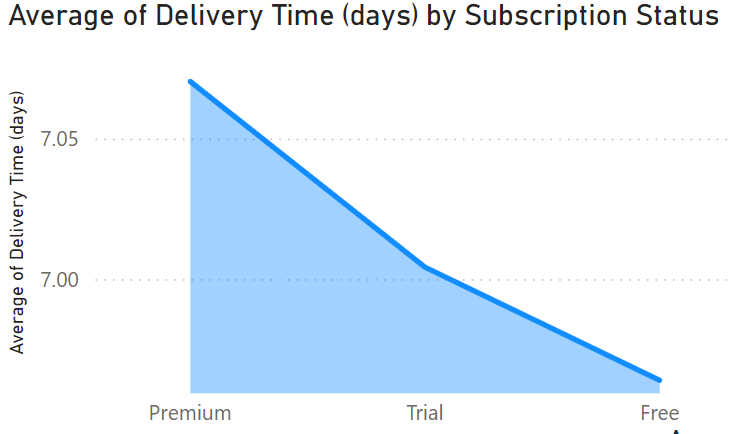


**Q5**: Average Review Score Given by Customers

* **Average Review Score**: 3.0

**Q6**: Average Delivery Time (Free vs. Premium Subscription)

* **Free Subscription**: 6.96 Days
* **Premium Subscription**: 7.03 Days



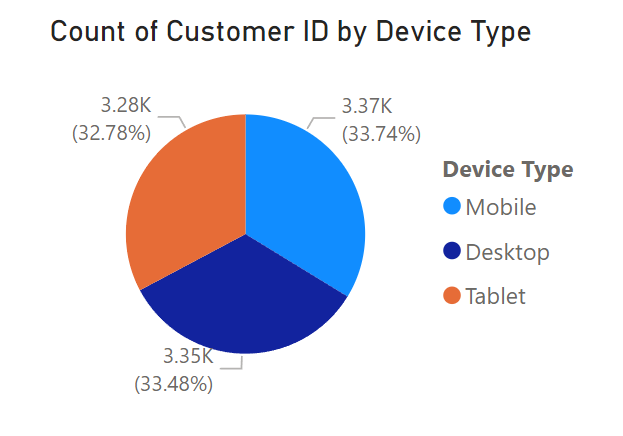
* **Insight**:
  + The data shows that the **Premium Subscription** (7.03 days) has a slightly **longer delivery time** than the **Free Subscription** (6.96 days). The slightly longer delivery time for premium subscribers could be due to additional services (e.g., better packaging or tracking), more exclusive product availability, or geographic differences. Despite the minor difference, both delivery times are quite similar overall.

**Q7**: Total Number of Subscribed Customers

* **Subscribed Customers**: 3369

**Q8**: Percentage of Customers Using Devices (Mobile, Desktop, Tablet)

* **Mobile**: 33.74 %
* **Desktop**: 33.48 %
* **Tablet**: 32.78 %

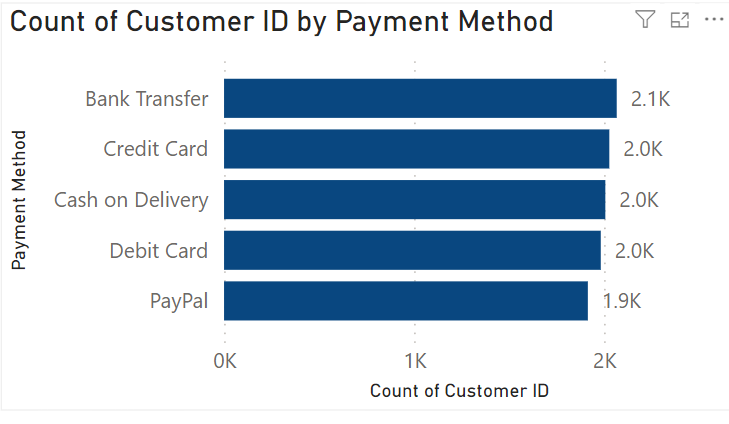


**Q9**: Average Purchase Amount (Discounted vs. Non-Discounted)

* **Discounted**: 505.26$
* **Non-Discounted**: 502.51$

**Q10**: Most Common Payment Method

* **Payment Method**: Bank Transfer



## Level 2: Intermediate Insights (21 Marks)

**Q1**: Average Review Scores of Users of the Most Common Payment Method

* **Average Review Score of most Common Payment Method**: 3.0
* **Payment Method**: Bank Transfer

**Q2:** What is the correlation between time spent on the website and purchase amount? Do customers who spend more time on the website purchase more items**?**

**Correlation Coefficient**: 0.010025

The **correlation coefficient** is a statistical measure that quantifies the strength and direction of a relationship between two variables. It ranges from **-1 to 1**:

* **+1** indicates a perfect positive correlation (as one variable increases, the other also increases).
* **-1** indicates a perfect negative correlation (as one variable increases, the other decreases).
* **0** means no correlation (the variables do not have a linear relationship).
* **Insight**:
  + The correlation coefficient of **0.010025** between **time spent on the website** and **purchase amount** indicates a **very weak positive correlation**. This means that spending more time on the website does not significantly affect how much a customer buys. This could suggest that other factors, like Location, Delivery Time, or Discount Avail, play a larger role in influencing purchases than just time spent browsing.

**Q3**: Percentage of Satisfied (Rating 4 or 5) and Return Customers

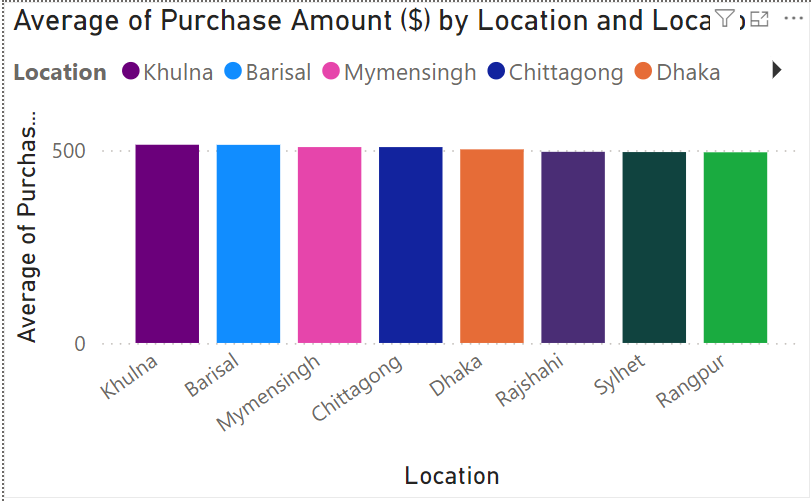
* Total Satisfied Customer (assuming only rating 4 or 5 as per question) is: 4004
* Satisfied and Return Customer: 2008
* **Percentage of Satisfied Return Customers**: 50.15%

**Q4:** What is the relationship between the number of items purchased and customer satisfaction?

* **Insight:**
  + I calculated the correlation between the **number of items purchased** and **customer satisfaction**, resulting in a value of **-0.003408**. This correlation is very close to **zero**, indicating a **very weak negative relationship** between the two variables.
* **Explanation:**
  + This means that the number of items a customer purchases **has little to no impact on their satisfaction leve**l. Whether a customer buys more or fewer items does not appear to significantly affect how satisfied they are. Based on this result, it's likely that other factors such as product quality, service experience, or pricing may have a greater influence on customer satisfaction than the number of items they purchase.

**Q5: Location with the 2nd Highest Average Purchase Amount**

* **Location:** Barisal
* **Average Purchase Amount:** 513.67$

****

# Level 3: Critical Thinking Insights

**Q1**: What factors contribute most to a customer being classified as a return customer?

To identify the key factors that contribute to a customer being classified as a return customer, we conducted a thorough analysis using three statistical methods: **correlation analysis** and **chi-square test**s for categorical variables.

**1. Key Findings from P-Values**

* **Return Customer**: P-value of **0.000000**, indicating a highly significant factor that directly classifies customers as return customers.
* **Discount Availed**: P-value of **0.013128**, suggesting that customers who take advantage of discounts are significantly more likely to return.
* **Gender**: P-value of **0.087425**, indicating a moderate association with return rates but not statistically significant at the conventional level.
* **Delivery Time**: P-value of **0.163882**, showing no significant impact on return status.
* **Review Score (1-5)**: P-value of **0.381226**, indicating no significant influence on return behavior.
* **Time Spent on Website**: P-value of **0.529676**, suggesting minimal impact on return likelihood.
* **Purchase Amount ($)**: P-value of **0.847462**, showing no significant relationship with return status.

**2. Correlation Analysis**

Age -0.008255

Purchase Amount ($) 0.001924

Time Spent on Website (min) -0.006286

Discount Availed 0.025003

Number of Items Purchased 0.000815

Review Score (1-5) 0.008757

Return Customer 1.000000

The correlation analysis yielded the following insights:

* **Age**: Correlation of **-0.008255**, indicating a negligible relationship with return likelihood.
* **Purchase Amount ($)**: Correlation of **0.001924**, suggesting no significant influence on return behavior.
* **Time Spent on Website**: Correlation of **-0.006286**, indicating little effect on return rates.
* **Discount Availed**: Correlation of **0.025003**, indicating a slight positive relationship with return likelihood. And it is the Highest Relation.
* **Number of Items Purchased**: Correlation of **0.000815**, showing negligible impact.
* **Review Score (1-5)**: Correlation of **0.008757**, also suggesting little effect.

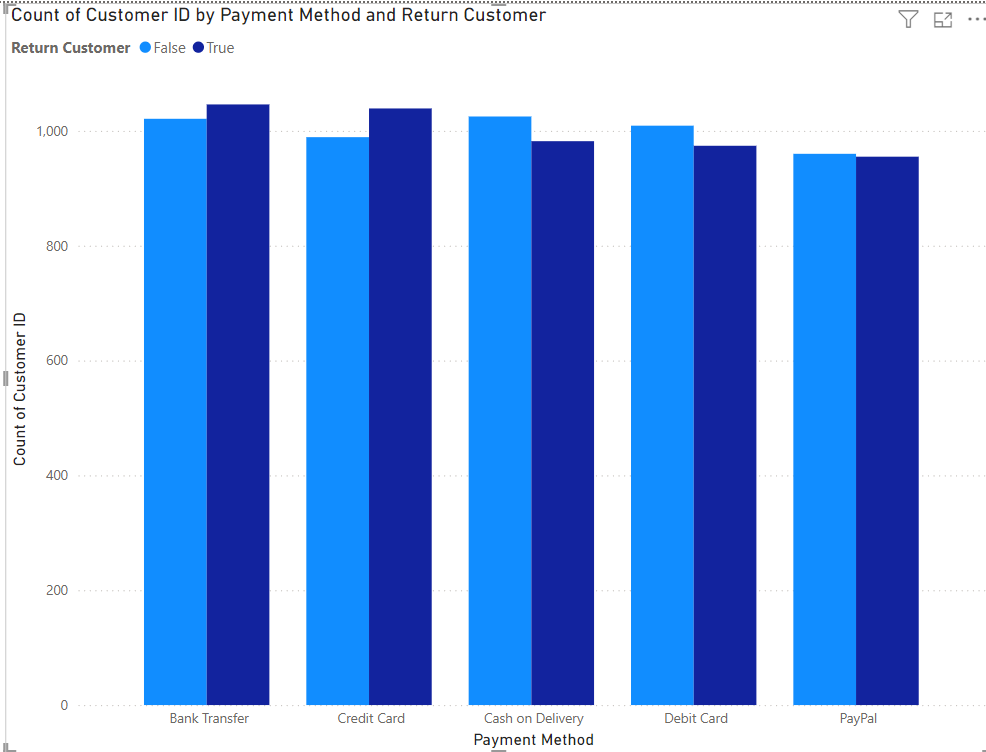
**Conclusion and Final Answer**

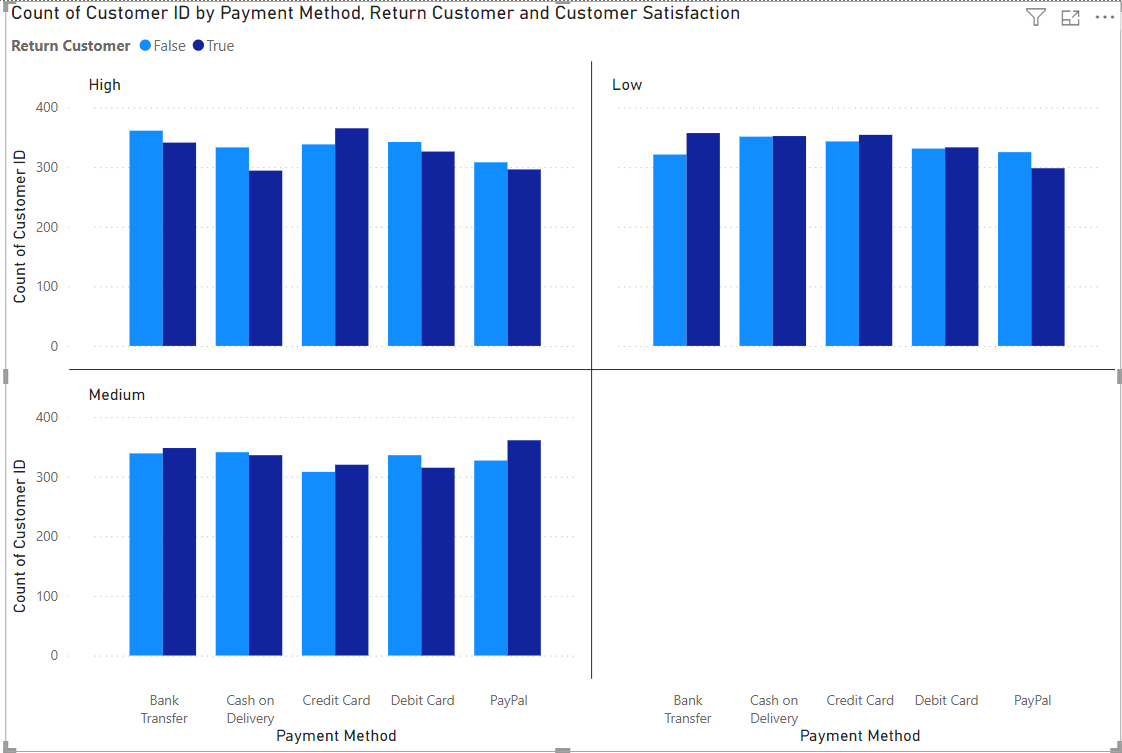
The analysis indicates that the most significant factors contributing to customers being classified as return customers are:

* **Discount Availed**: Customers utilizing discounts tend to return at a higher rate.

Factors like age, purchase amount, and time spent on the website show minimal to no correlation with return status, highlighting that strategic efforts to enhance customer retention should focus on discount offers and understanding the dynamics of return customer behavior.

**Q2:** How do payment methods influence customer satisfaction and return rates?





1. **Payment Method Distribution:**
   * The total payments across methods are relatively balanced, with the top five payment methods (Bank Transfer, Credit Card, Cash on Delivery, Debit Card, and PayPal) each accounting for around 20% of total transactions.
2. **Customer Satisfaction Levels:**
   * **High Satisfaction:**
     + Bank Transfer has the highest number of satisfied customers (702), accounting for **33.96%** of its total users.
     + Credit Card follows closely with **34.66%** high satisfaction.
   * **Medium Satisfaction:**
     + Cash on Delivery shows the highest medium satisfaction percentage (**33.73%**) among its users.
   * **Low Satisfaction:**
     + All payment methods have a similar distribution of low satisfaction, ranging from **32.53%** (PayPal) to **35.03%** (Cash on Delivery).
3. **Return Rates:**
   * **Returning Customers:**
     + Bank Transfer leads with **50.60%** return customer percentage, indicating that customers using this method are more likely to return for subsequent purchases.
     + Credit Card users also show a high return rate at **51.23%**.
   * **Non-Returning Customers:**
     + Similar trends are observed in the non-returning customer percentages, where Credit Card users account for **48.77%**, and Debit Card users represent **50.88%** of non-returning customers.

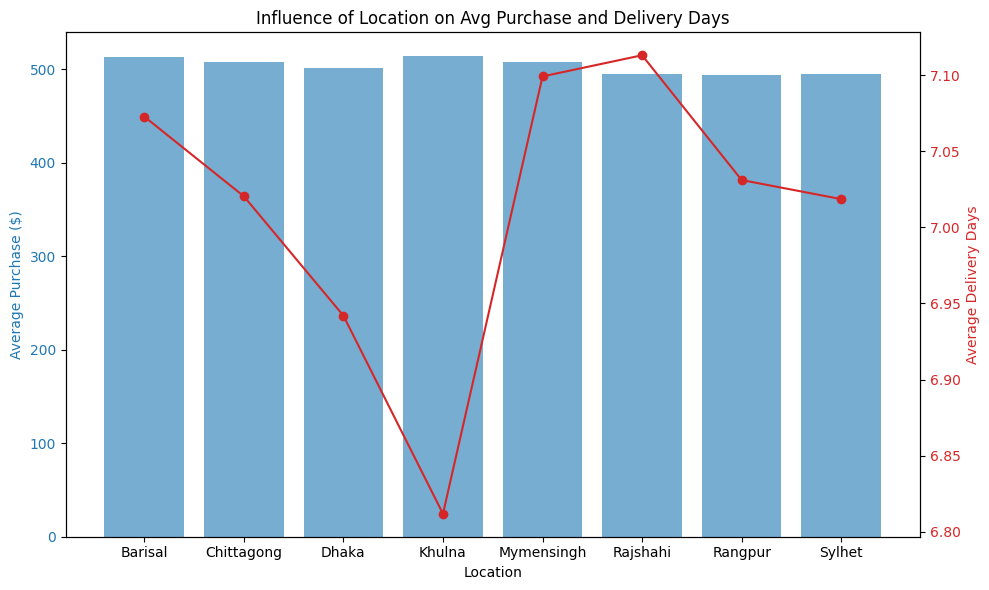
**Analysis**

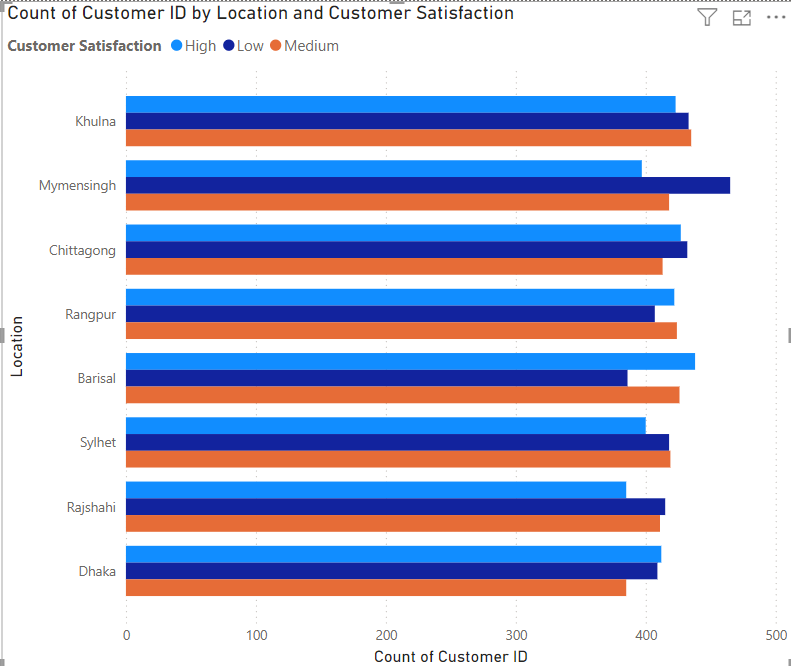
* **Payment Method Impact:**
  + Payment methods that yield higher customer satisfaction also correspond to higher return rates. For instance:
    - **Bank Transfer** not only has a high satisfaction level but also leads in the percentage of returning customers, suggesting a positive relationship.
    - **Credit Card**, with its slightly lower high satisfaction percentage, still maintains a robust return rate, indicating that customers find this method reliable.
* **Satisfaction and Return Correlation:**
  + The data suggest that payment methods significantly influence both satisfaction and return rates. Customers who are satisfied with their payment experience are more likely to return, reinforcing the importance of a smooth and trusted payment process.

**Recommendations**

* **Enhance Payment Experiences:**
  + Focus on improving the payment process for methods with lower satisfaction levels, particularly Debit Card and Cash on Delivery, to boost overall customer satisfaction and return rates.
* **Targeted Marketing:**
  + Leverage the successful payment methods (like Bank Transfer and Credit Card) in marketing strategies to encourage new customers and retain existing ones.

**Q3:** How does the location influence both purchase amount and delivery time?





**Major Insights from the Data Analysis**

1. **Max and Min Purchase Amounts**:
   * The maximum purchase amounts are highest in **Sylhet (999.98)** and **Dhaka (999.86)**, suggesting these areas have customers willing to make high-value purchases. Conversely, the minimum purchase amount in **Barisal (5.19)** indicates that there is a diverse customer base, including budget-conscious shoppers. This diversity in spending behavior reflects the economic variations across locations, which can influence targeted marketing strategies.
2. **Volume of Purchases**:
   * **Chittagong** shows the highest number of purchases (**1272**), indicating strong customer engagement or a larger customer base. On the other hand, **Rajshahi** has the lowest number of purchases (**1211**). This disparity may suggest that Chittagong has more effective marketing efforts or product availability, whereas Rajshahi may require improved customer outreach and engagement strategies.
3. **Average Purchase Amount**:
   * The average purchase amounts are highest in **Khulna (513.94)** and **Barisal (513.67)**, indicating that customers in these areas tend to spend more per transaction. This insight suggests that higher-value products or better customer service may exist in these locations, warranting further investigation to replicate these successes in areas with lower average purchases like **Rajshahi (495.54)** and **Rangpur (494.37)**.
4. **Delivery Times**:
   * Average delivery times are relatively consistent, with **Khulna (6.81 days)** experiencing the shortest delivery time, suggesting efficient logistics in that area. Conversely, **Rajshahi (7.11 days)** has the longest delivery time, which could affect customer satisfaction negatively. Addressing these delivery time discrepancies is essential for maintaining customer satisfaction, particularly in locations with longer wait times.
5. **Customer Behavior Trends**:
   * The analysis highlights potential trends in customer behavior based on location. Areas with higher average purchases and faster delivery times tend to correlate with higher customer satisfaction and retention. Conversely, areas with lower average purchases and longer delivery times may benefit from strategic marketing initiatives and enhanced logistics.

**Conclusion**

Overall, the analysis reveals significant variations in purchasing behavior, delivery times, and customer engagement across different locations. These insights suggest that tailored marketing strategies, improved logistics, and customer engagement initiatives should focus on the unique characteristics and needs of each region. Addressing the areas of concern, particularly in locations with lower performance metrics, can help enhance overall customer satisfaction and drive sales growth effectively.

**Q4:** What major insights are found? Give a proper explanation.

**1. Customer Loyalty and Return Customers**

* **Key Finding**: Approximately **49.96% of customers are return customers**, indicating a substantial portion of customers are willing to make repeat purchases.
* **Explanation**: This high percentage suggests that the business has successfully created a customer base that finds value in their offerings, which is crucial for long-term success. The analysis indicates that strategies focusing on enhancing customer experiences can further increase this loyalty.

**2. Influence of Discounts on Return Rates**

* **Key Finding**: **Discount Availed** has a significant p-value of **0.013128**, indicating customers who utilize discounts are more likely to return.
* **Explanation**: This insight highlights the effectiveness of promotional strategies. By offering discounts, the business can attract more customers to return, thus boosting customer retention. This suggests that implementing regular discount offers could be a strategic approach to improve loyalty.

**3. Minimal Impact of Time Spent on Website**

* **Key Finding**: The correlation coefficient between **time spent on the website** and **purchase amount** is **0.010025**, indicating a very weak positive correlation.
* **Explanation**: This suggests that simply spending more time on the website does not significantly influence the amount customers purchase. Other factors such as product quality, availability, and marketing strategies may play more significant roles in influencing purchase behavior. Therefore, optimizing user experience and guiding customers more effectively could enhance purchasing outcomes.

**4. Payment Methods and Customer Satisfaction**

* **Key Finding**: **Bank Transfer** is identified as the most preferred payment method, with a **return customer percentage of 50.60%** and a high satisfaction level.
* **Explanation**: This finding underscores the importance of a seamless payment experience. Customers are more likely to return if they have positive interactions with payment methods. The findings suggest that optimizing these payment channels can increase satisfaction and retention rates, reinforcing the need for a reliable and user-friendly payment system.

**5. Customer Satisfaction and Review Scores**

* **Key Finding**: The average review score across the dataset is **3.0**, suggesting moderate satisfaction levels.
* **Explanation**: This average indicates there is room for improvement in customer satisfaction. Given the significant relationship between satisfaction and return rates, the business should consider strategies to enhance product quality and customer service to boost review scores, thereby fostering loyalty.

**6. Demographic Insights**

* **Key Finding**: The dataset shows that factors like **age, gender, and location** have minimal impact on customer return rates, with p-values indicating weak associations.
* **Explanation**: This suggests that return behavior is not significantly influenced by demographic factors, which could imply that the business has a broad appeal across various customer segments. Thus, marketing efforts should be targeted more toward behavioral aspects rather than demographic ones.

**7. Delivery Time and Subscription Type**

* **Key Finding**: The average delivery time for **premium subscribers (7.03 days)** is slightly longer than for **free subscribers (6.96 days)**.
* **Explanation**: This counterintuitive result may indicate that premium services might involve additional processing times due to features like special handling or enhanced delivery options. Understanding customer perceptions about delivery expectations can help tailor premium services to better meet customer needs.

**8. Correlation Between Items Purchased and Customer Satisfaction**

* **Key Finding**: The correlation between the number of items purchased and customer satisfaction is **-0.003408**, indicating a very weak negative relationship.
* **Explanation**: This suggests that buying more items does not necessarily lead to higher satisfaction. It may reflect that customers prioritize quality and service over quantity. Thus, focusing on enhancing the shopping experience and product quality might be more beneficial than merely encouraging larger purchases.

**Recommendations:**

1. **Leverage Discounts**: Create strategic discount campaigns to boost return rates.
2. **Enhance Payment Experiences**: Focus on improving the user experience with payment methods that have lower satisfaction rates to improve overall satisfaction and loyalty.
3. **Optimize Customer Experience**: Address the factors influencing moderate satisfaction scores, aiming for consistent improvements in service and product quality.
4. **Analyze Delivery Dynamics**: Investigate the reasons behind longer delivery times for premium subscriptions to better manage customer expectations.