# ****Consumer Behaviour and Shopping Habits Dataset****

# Aim of the Project:

The Aim of this project is to analyse consumer behaviour and shopping habits using a comprehensive dataset that includes various attributes related to customers, their purchases, and other relevant demographic and transactional information. This analysis will provide insights into the factors influencing consumer purchasing decisions, patterns, and preferences, ultimately aiding in the development of targeted marketing strategies, product recommendations, and improved customer experience.

**Project Description:**

Retail businesses need a comprehensive analysis of consumer behaviour to identify key patterns and trends, understand the factors driving purchase decisions, and predict future buying behaviours. The challenge lies in effectively analyzing a multi-dimensional dataset that includes customer demographics, transactional data, product preferences, payment methods, and feedback to derive meaningful insights.

###### **Dataset Glossary (column wise):**

**Customer ID:** A unique identifier assigned to each individual customer, facilitating tracking and analysis of their shopping behaviour over time.

**Age:** The age of the customer, providing demographic information for segmentation and targeted marketing strategies.

**Gender:** The gender identification of the customer, a key demographic variable influencing product preferences and purchasing patterns.

**Item Purchased:** The specific product or item selected by the customer during the transaction.

**Category:** The broad classification or group to which the purchased item belongs (e.g., clothing, electronics, groceries).

**Purchase Amount (USD):** The monetary value of the transaction, denoted in United States Dollars (USD), indicates the cost of the purchased item(s).

**Location:** The geographical location where the purchase was made.

**Size:** The size specification (if applicable) of the purchased item, relevant for apparel, footwear, and certain consumer goods.

**Colour:** The colour variant or choice associated with the purchased item, influencing customer preferences and product availability.

**Season:** The seasonal relevance of the purchased item (e.g., spring, summer, fall, winter).

**Review Rating:** A numerical or qualitative assessment provided by the customer regarding their satisfaction with the purchased item.

**Subscription Status**: Indicates whether the customer has opted for a subscription service.

**Shipping Type:** Specifies the method used to deliver the purchased item (e.g., standard shipping, express delivery).

**Discount Applied:** Indicates if any promotional discounts were applied to the purchase.

**Promo Code Used:** Notes whether a promotional code or coupon was utilized during the transaction.

**Previous Purchases:** Provides information on the number or frequency of prior purchases made by the customer.

**Payment Method:** Specifies the mode of payment employed by the customer (e.g., credit card, cash).

**Frequency of Purchases:** Indicates how often the customer engages in purchasing activities, a critical metric for assessing customer loyalty and lifetime value.

# ****INSIGHTS FROM DATA:****

# ****PIE CHART****

# 

# 

# 

Pie chart illustrates the distribution of purchases based on gender. The chart is divided into two segments, representing the proportion of purchases made by males and females.

* **Yellow Segment (67.7%)**: Represents the percentage of purchases made by male customers.
* **Blue Segment (32.3%)**: Represents the percentage of purchases made by female customers.

Higher Purchase Rate Among Males:

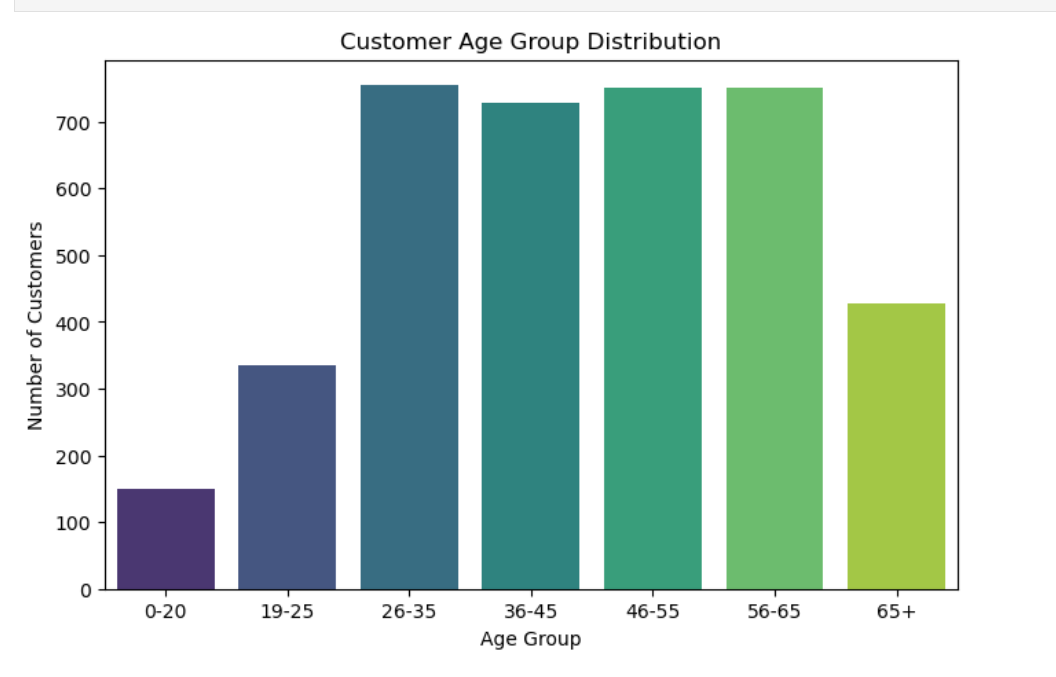
* 67.7% of purchases are made by male customers.
* Males are more active shoppers in the dataset under analysis.
* Females account for only 32.3% of purchases.
* There is potential to increase engagement and sales among female customers.

Gender-Specific Marketing:

* Males dominate the purchase statistics.
* Businesses should consider tailored marketing campaigns and product offerings that cater to male preferences and purchasing behaviours.
* This insight provides a basis for businesses to develop targeted strategies aimed at engaging female customers more effectively, ultimately striving for a balanced and inclusive market presence.

**BAR CHART**

Bar chart illustrates the distribution of customers across different age groups



The bar chart illustrates the distribution of customers across different age groups. The x-axis represents various age groups, while the y-axis indicates the number of customers in each group.

* **0-20 Age Group**: Approximately 150 customers
* **19-25 Age Group**: Approximately 350 customers
* **26-35 Age Group**: Approximately 750 customers
* **36-45 Age Group**: Approximately 700 customers
* **46-55 Age Group**: Approximately 700 customers
* **56-65 Age Group**: Approximately 700 customers
* **65+ Age Group**: Approximately 400 customers

Dominant Age Group:

* The 26-35 age group has the highest number of customers, with around 750 individuals.
* This age group is the most active in terms of shopping, making them a key target demographic for marketing strategies.

Balanced Middle Age Groups:

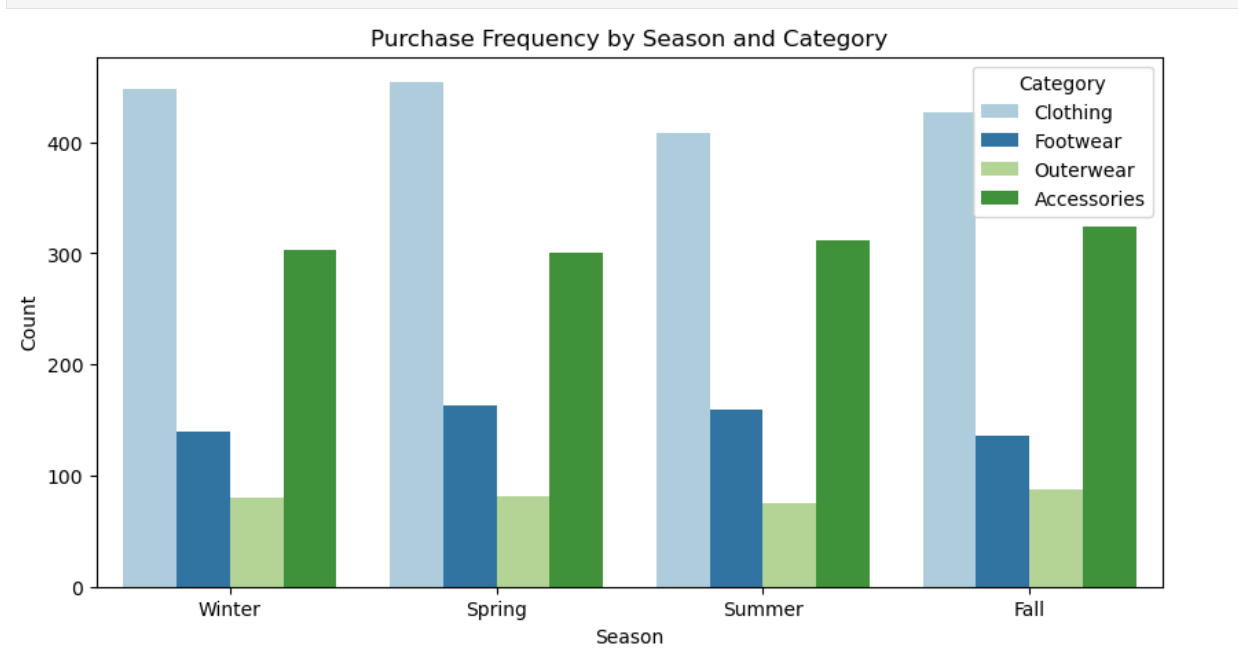
* The 36-45, 46-55, and 56-65 age groups each have around 700 customers.
* These age groups are fairly balanced and also represent significant portions of the customer base, indicating steady engagement across these segments.
* he 0-20 age group has the lowest number of customers, with about 150 individuals.

Senior Engagement:

* The 65+ age group has around 400 customers.
* While this group is smaller than the middle-aged groups, it still represents a significant portion of the customer base, suggesting opportunities for specialized marketing.

The bar chart reveals that the 26-35 age group is the most active segment in terms of customer numbers, followed closely by the 36-45, 46-55, and 56-65 age groups. The minimal engagement from the 0-20 age group indicates a potential area for growth, while the 19-25 and 65+ age groups also present opportunities for targeted marketing. Businesses should focus on maintaining engagement with the dominant age groups while exploring strategies to increase participation from younger and older customers.

**GROUPED BAR CHART**

****

The grouped bar chart that shows the purchase frequency by season and category. Each bar represents the count of purchases within a specific category for each season.

* **X-Axis (Season)**: Represents the seasons (Winter, Spring, Summer, Fall).
* **Y-Axis (Count)**: Represents the number of purchases.

The different categories of items purchased, color-coded as follows

**Clothing:** Light blue

**Footwear**: Dark blue

**Outerwear**: Light green

**Accessories:** Dark green

**Clothing**:

* Consistently has the highest number of purchases across all seasons.
* Winter and Spring have particularly high counts, indicating that these might be peak seasons for clothing purchases.

**Footwear**:

* The number of purchases remains relatively stable across seasons, with slight increases in Winter and Fall.
* Significantly fewer purchases compared to clothing but more consistent across the year.

**Outerwear**:

* Has the lowest number of purchases in all seasons.
* Purchases are relatively stable but very low compared to other categories.

**Accessories**:

* Shows a noticeable increase in purchases during Spring, Summer, and Fall.
* Winter has the lowest count for accessories purchases.

**Seasonal Trends**:

* Clothing purchases peak during Winter and Spring, possibly due to seasonal sales, holidays, or the start of new fashion seasons.
* Accessories see higher purchases in warmer seasons (Spring, Summer, Fall), suggesting they might be more popular as complementary items to seasonal outfits.
* Footwear has a consistent purchase pattern across seasons, indicating a stable demand throughout the year.
* Outerwear has the least variability and the lowest count, suggesting it might be a niche category or one with less frequent purchases.

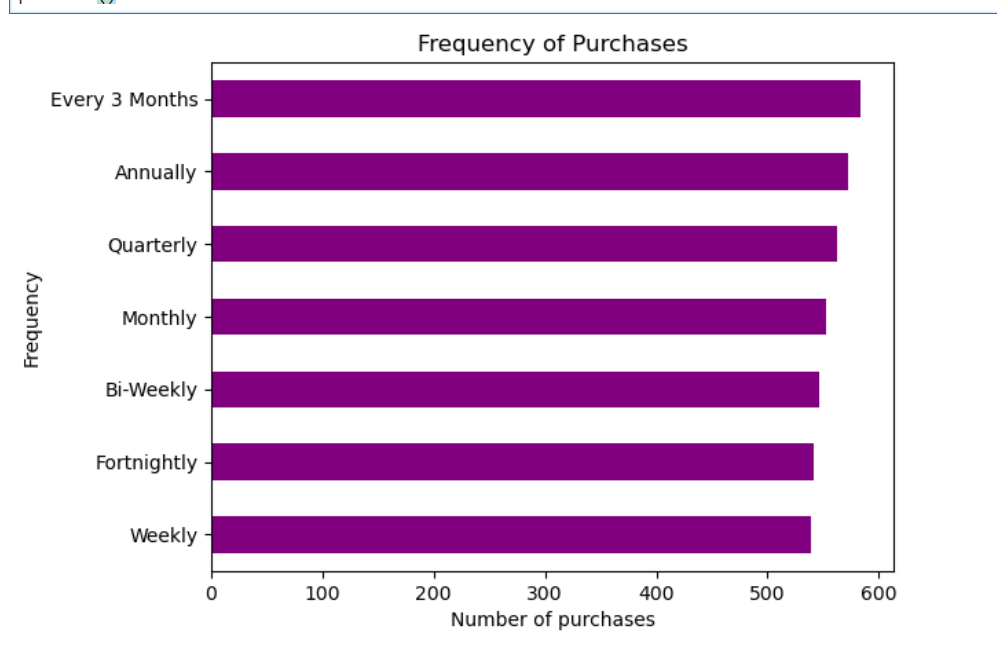
**Overall Purchase Patterns**:

* Clothing dominates the purchase frequency in all seasons, indicating it is the most popular category among the consumers in the dataset.

**HORIZONTAL BAR CHART**

**Frequency of Purchases**

The chart provided is a horizontal bar chart illustrating the frequency of purchases over various time intervals. The y-axis represents different frequency categories, while the x-axis represents the number of purchases. Each bar corresponds to a specific frequency category and its length indicates the total number of purchases made within that time frame.

The frequencies included are:

* Weekly
* Fortnightly
* Bi-Weekly
* Monthly
* Quarterly
* Annually
* Every 3 Months

**Uniform Purchase Distribution**

The chart shows that each frequency category has an equal number of purchases, with each bar being the same length. This suggests that the data is uniformly distributed across all time intervals, with no single category dominating in terms of purchase frequency.

**Number of Purchases**

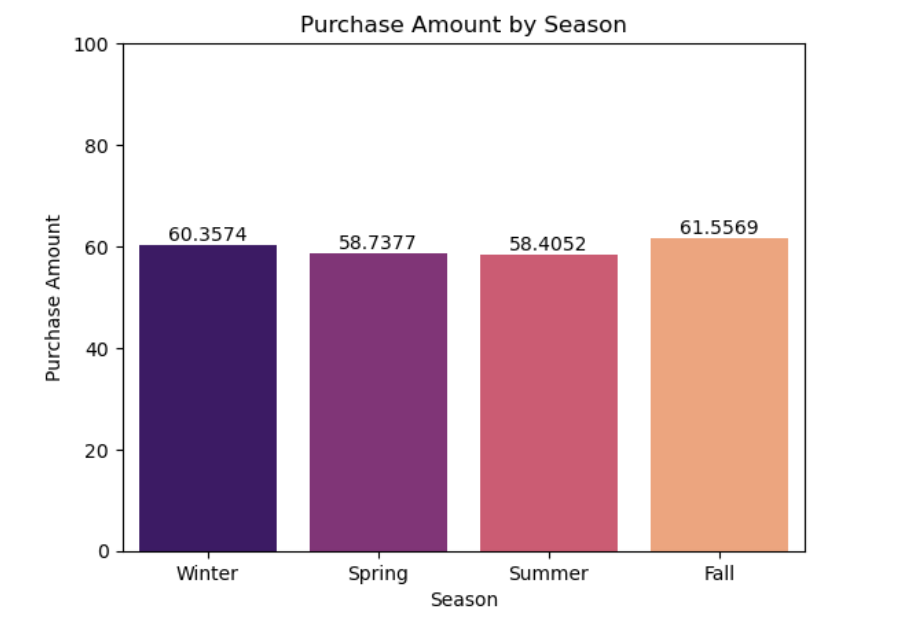
The number of purchases in each frequency category appears to be around 600, indicating a consistent purchasing pattern regardless of the time interval.

**Customer Behaviour**

The uniform distribution suggests that customers are equally likely to make purchases at any of these frequencies. This could indicate a diverse customer base with varied purchasing habits or a well-balanced sample.

The chart indicates a balanced distribution of purchases across different frequencies, which can be beneficial for developing consistent strategies in marketing, sales, and inventory management. Understanding this uniform behaviour can lead to more effective and tailored approaches to meet the needs of customers who shop at various intervals.

**LABELED BAR CHART**



The chart is a bar graph that shows the average purchase amounts across four different seasons: Winter, Spring, Summer, and Fall. Each bar represents the purchase amount for a respective season.

-X-axis: Represents the seasons (Winter, Spring, Summer, and Fall).

-Y-axis: Represents the purchase amounts, with values ranging from 0 to 100.

-Bars: Each bar is color-coded and labeled with the exact purchase amount for each season.

Fall is the season with the highest average purchase amount, while Spring is the season with the lowest. There is a noticeable but not dramatic variation in purchase amounts across the seasons, with Fall and Winter being slightly higher compared to Spring and Summer.

* Highest and Lowest Purchase Amounts

- Fall has the highest average purchase amount, indicating that consumers tend to spend more during this season.

- Summer has the lowest average purchase amount, indicating a possible decrease in consumer spending during this time.

* Winter and Fall:

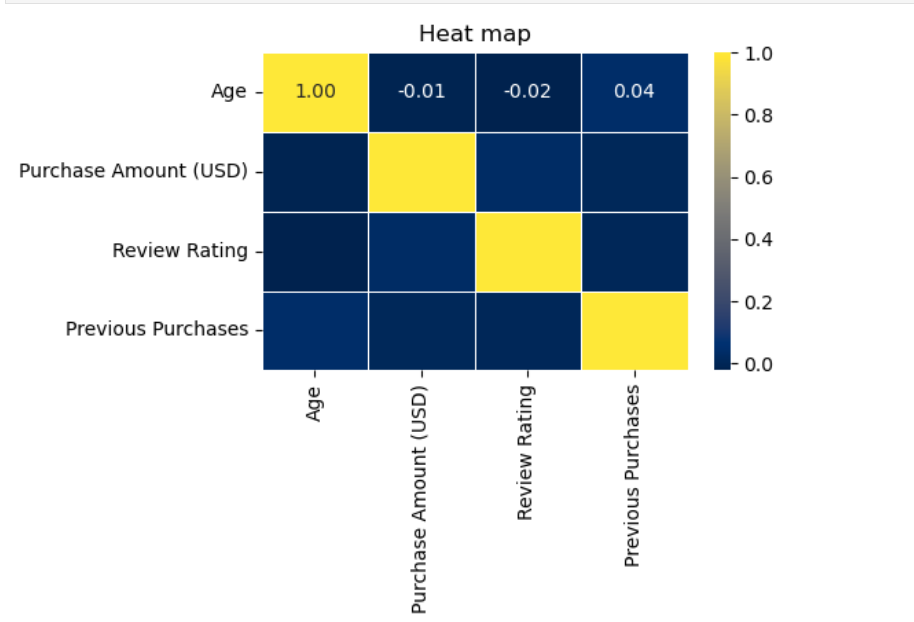
- Both Winter and Fall show higher purchase amounts compared to Spring and Summer. This could be due to several factors such as holidays, sales events, or climatic conditions that encourage spending.

* Spring and Summer:

- These seasons have relatively lower purchase amounts, which could be due to different consumer priorities or economic conditions during these times.

Understanding these trends can help businesses and marketers tailor their strategies according to seasonal demand, optimizing inventory, marketing efforts, and promotions.

**HEAT MAP**

****

The heatmap provided represents the correlation matrix between four variables: Age, Purchase Amount (USD), Review Rating, and Previous Purchases. The colour intensity indicates the strength and direction of the correlation, with yellow representing positive correlation, blue representing negative correlation, and the numbers indicating the exact correlation coefficients.

* **1.0**: Perfect positive correlation (the variables move in the same direction).
* **-0.01**: Very weak negative correlation (almost no relationship).
* **-0.02**: Very weak negative correlation (almost no relationship).
* **0.04**: Very weak positive correlation (almost no relationship).
* Weak Correlations

- Most of the variables show very weak correlations with each other, with correlation coefficients close to zero. This indicates that there is little to no linear relationship between these variables.

* Age

- Age has a very weak negative correlation with Purchase Amount (USD) and Review Rating, and a very weak positive correlation with Previous Purchases.

- This suggests that age does not significantly influence purchase amount, review rating, or the number of previous purchases.

* Purchase Amount (USD)

- Purchase Amount (USD) has no correlation with Review Rating and Previous Purchases.

- This indicates that the amount spent by customers does not affect their review ratings or the number of previous purchases

* Review Rating

- Review Rating shows no correlation with Purchase Amount (USD) and Previous Purchases.

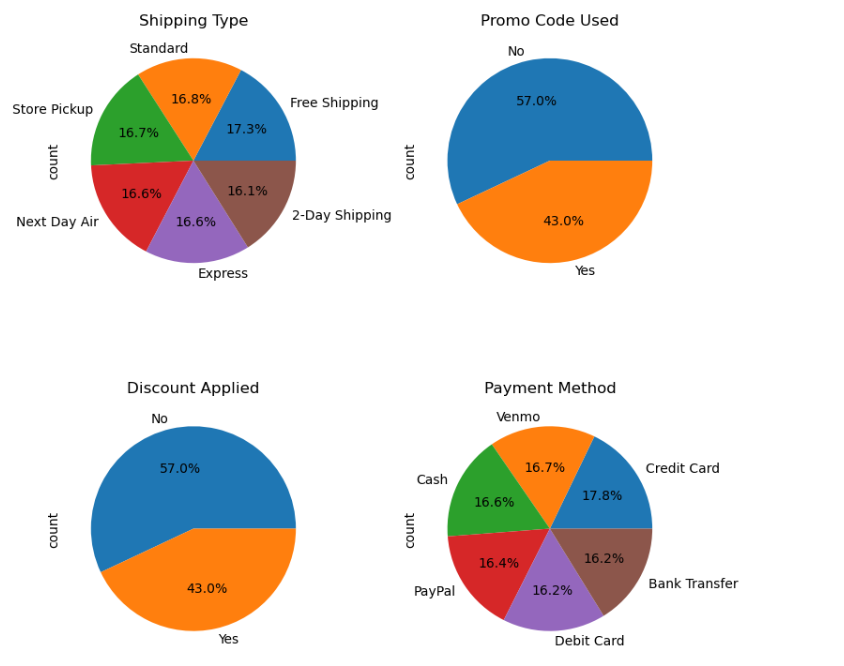
- This implies that customers' ratings do not depend on how much they spend or how many times they have purchased before.

* Previous Purchases

- Previous Purchases have no correlation with Purchase Amount (USD) and Review Rating.

- This means that the frequency of past purchases does not affect the amount spent or the ratings given by customers.

**PIE CHART**



The charts provide a distribution of customer preferences in terms of shipping types, usage of promo codes, application of discounts, and payment methods. The distribution is relatively even across most categories, with slight variations in preferences.

**Shipping Type**

* Even Distribution

The choice of shipping type is relatively evenly distributed among the different options, with no single method dominating significantly.

* Free Shipping Preference

Free shipping is slightly more preferred than other options, indicating that customers are likely attracted to cost-saving opportunities.

* Variety in Shipping Choices

The nearly equal preference for other shipping options (Standard, Store Pickup, Next Day Air, 2-Day Shipping, and Express) suggests that customers value a variety of shipping methods depending on their needs.

**Promo Code Used**

* More Non-Users

A higher percentage of customers (57%) did not use promo codes, suggesting either a lack of awareness or availability of promo codes.

* Significant Usage

Despite this, a substantial 43% did use promo codes, indicating that a large segment of customers is motivated by discounts and special offers.

**Discount Applied**

* Similar to Promo Code Usage

The distribution is identical to the promo code usage, reinforcing the idea that nearly half of the customer base is responsive to discounts.

* Potential for Increased Engagement

The 43% of customers who received discounts shows that there is a potential to increase engagement through more targeted discount strategies.

**Payment Method**

* Credit Card Dominance

Credit cards are the most commonly used payment method, highlighting their convenience and widespread acceptance.

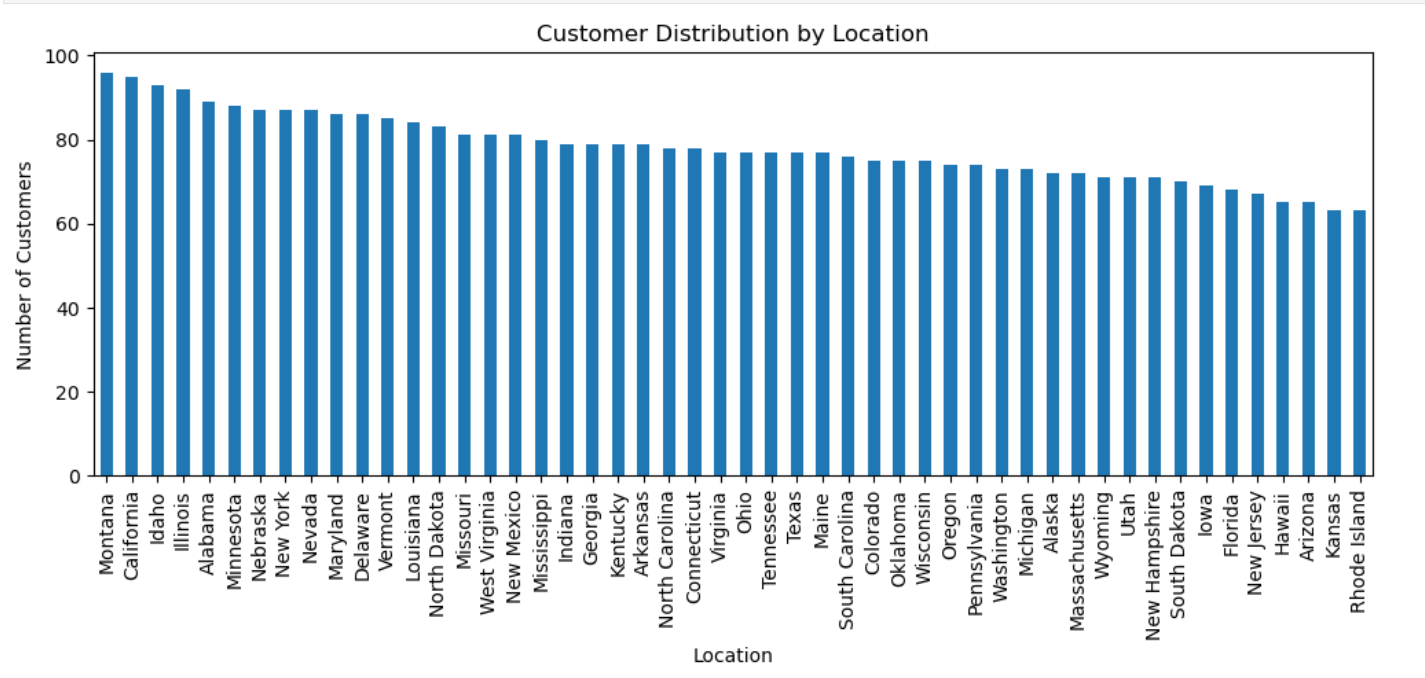
* Diversified Payment Preferences

The nearly equal distribution among Venmo, Cash, PayPal, Bank Transfer, and Debit Card shows a diversity in customer payment preferences.

* Importance of Payment Options

Offering multiple payment methods is important to cater to different customer preferences and enhance the shopping experience.

**BAR PLOT**

****

* X-Axis (Location)

- Represents different states in the United States.

* Y-Axis (Number of Customers)

- Represents the number of customers in each state.

- The scale ranges from 0 to 100 customers.

* Bars

- Each bar represents the number of customers from a specific state.

- The height of the bar indicates the customer count for that state.

* Montana and California Lead

- Montana and California have the highest number of customers, both close to 100.

* Gradual Decrease

- There is a gradual decrease in the number of customers as we move from left to right on the chart.

* Lower Customer Counts

- States like Kansas and Rhode Island have the lowest number of customers, though still significant.

# 2 Sample Independent t \_test

A two-sample independent t-test compares the means of two independent groups to determine if there is a statistically significant difference between them. It assumes that the samples are randomly selected, the populations are normally distributed, and the variances are equal. The test calculates the t-statistic, which measures the difference between the group means relative to the variability of the samples. A p-value is then derived from the t-statistic to assess the significance of the observed difference. If the p-value is below a chosen significance level (e.g., 0.05), the null hypothesis of no difference is rejected.

 **Null Hypothesis (H0)**: The null hypothesis states that there is no significant difference in the mean purchase amounts between the male and female groups.

H0:μ male=μ female

# ****Alternative Hypothesis (H1)****: The alternative hypothesis states that there is a significant difference in the mean purchase amounts between the male and female groups.

# H1:μmale≠μfemale

* **T-Statistic: -0.87**

# P-Value: 0.38

**T-Statistic**: The t-statistic is -0.87. This value measures how many standard errors the sample mean difference is away from the null hypothesis mean difference (which is 0 in this case). A t-statistic close to 0 suggests that the sample means are close to each other.

**P-Value**: The p-value is 0.38. This value indicates the probability of obtaining a test statistic as extreme as, or more extreme than, the observed one, assuming the null hypothesis is true.

### Decision Making

* **Significance Level (α)**: Typically, a significance level of 0.05 is used.
* **Compare P-Value to α**: Since the p-value (0.38) is greater than the significance level (0.05), we fail to reject the null hypothesis.

### Conclusion of T\_Test

* Since the p-value (0.3805) is greater than alpha (0.05), we do not have enough evidence to reject the null hypothesis.
* The results are not statistically significant at the 0.05 significance Therefore, we fail to reject the null hypothesis.

# Chi Square Test

# The chi-square test assesses whether there is a significant association between categorical variables. It compares the observed frequencies in each category to the expected frequencies, which are calculated under the assumption that the variables are independent. The test statistic is calculated by summing the squared differences between observed and expected frequencies, divided by the expected frequencies. A large chi-square value indicates a greater discrepancy between observed and expected frequencies, suggesting an association. The significance is determined by comparing the test statistic to a chi-square distribution with appropriate degrees of freedom.

The chi-square test was conducted to determine if there is a significant association between two categorical variables (e.g., "Shipping Type" and "Preferred Payment Method"). Here's how to interpret and conclude based on the results:

1. **Null Hypothesis (H0)**: The null hypothesis states that there is no significant association between the variables; they are independent.
2. **Alternative Hypothesis (H1)**: The alternative hypothesis suggests that there is a significant association between the variables; they are not independent.

### Interpretation of Results:

* **Chi Square Statistic**: The calculated chi-square statistic of 19.01 indicates the magnitude of discrepancy between the observed and expected frequencies under the null hypothesis.
* **Degrees of Freedom (DOF)**: With 25 degrees of freedom, calculated based on the number of categories in the variables being compared.
* **p-value**: The p-value of 0.7968 is greater than the typical significance level of 0.05. This high p-value suggests that there is no significant evidence to reject the null hypothesis.

### Conclusion:

Given the p-value of 0.7968:

* We fail to reject the null hypothesis (H0).
* There is significant association between "Shipping Type" and "Preferred Payment Method" .

**FINAL ANALYSIS OF THE PROJECT**

* **Customer Base and Geographic Insights**

**Montana**: Montana has the highest number of purchases, indicating a substantial portion of our customer base resides here.

**High-Spending States**: Alaska and Pennsylvania customers exhibit the highest average and median purchase amounts, suggesting they spend the most money while shopping.

* **Shipping Preferences**

**Free Shipping**: Free shipping services are the most popular among customers, closely followed by standard shipping services. This trend suggests that keeping shipping costs low can attract more customers.

* **Seasonal Purchasing Trends**

**Seasonal Purchases**: Customers tend to make more purchases during winter and fall compared to spring and summer.

**Promo Codes**: The usage of promo codes does not seem to have a significant impact on purchase behaviour.

* **Product Popularity**

**Clothing**: Clothing is the most popular product category across all consumer demographics.

**Accessories**: Accessories are equally popular across age groups, except for those aged 15-25 and 65-75.

**Footwear**: Footwear is particularly popular among the 45-55 age group.