**Exploration and Analysis of Data**

* **Pre-processing data before Analysis: -**

1. During data import from CSV to SQL Server, missing fields were initially blank. These were later updated to **NULL** or tagged as **Non-member customers** for better categorization and analysis.
2. The **Transaction Date** column was not in the correct date format, so its data type was converted to **Date Time**. This column had **50,000 missing values**, which defaulted to the date **'1901-01-01'**. These were later updated to **Null** for accuracy.
3. The columns **Transaction Amount**, **Quantity**, **Discount Percent**, and **Loyalty Points** were initially stored as **Varchar** data types. These have been converted to **numeric** data types for accurate calculations and analysis.
4. To address the issue of filling missing values in the **region** and **payment method** fields based on the highest number of transactions done by a customer in specific segments.

* **Facts And Exploration of data:**

1. The dataset comprises **500,000 records** and **19 columns**, indicating a substantial volume of data for analysis.
2. Around **10% of data is missing** across important customer attributes, including:
   1. Customer IDs, Age, and Gender.
   2. Buying preferences like Payment Method, Product Name, and Region.
3. A total of **48,995 unique customers** have completed **500,000 transactions**, **indicating an average of approximately 10 transactions per customer**.
4. The data we have is from January 1st, 2022, to December 14th, 2022, with the exception of missing dates, which we have considered as January 1st, 1753.
5. Transactions with an amount **less than 0 are considered returns,** while transactions with an amount **greater than 0 are considered purchases**.
6. A **maximum discount of 50 percent** has been offered, meaning that customers can receive a discount up to half the original price of the product or service.

* **Customer Behaviour Segmentation:**

1. The method ensures that the missing values for **region** and **payment method** are aligned with the customer’s historical transaction preferences.
2. Customers are segmented by their transactional dominance (e.g., most-used region and payment method), improving the data's integrity for subsequent analyses.

* **Actionable Steps in SQL:**

1. Ensure all transactions are grouped by customer ID.
2. Compute the most frequent **region** and **payment method** for each customer.
3. Impute the missing values in the **region** and **payment method** columns for each transaction based on the highest frequency calculated for that customer.
4. For customers with insufficient data, consider using aggregate preferences (e.g., region and payment method with the highest count across all customers).
5. I have considered the maximum age from the customer age bucket and used it to fill in the missing values in the customer age field in the dataset.

* **Bucketing:**

1. **One Timer – Multi Timer Tagging:** The customers who have visited only once are considered one-time buyers, while customers who have visited more than once are tagged as multi-time buyers.
2. **Spend Bucket Tagging:** I created these spend buckets to categorize customers based on their total spending, which helps us better understand their purchasing power. By grouping customers into ranges, we can identify patterns, such as high-spend customers who may be key targets for premium offerings, or low-spend customers who might need different marketing strategies.
3. **Age Bucket Tagging:** I created age buckets to group customers into specific age ranges, making it easier to analyse customer behaviour based on their age group. This helps us understand which age segments are more active or valuable in specific regions or product categories.
4. **Reactivation tagging (In days):** I created buckets for reactivation days to analyse how long it takes for customers to visit again. This helps us identify patterns, such as how quickly customers return or when they start becoming inactive. By grouping reactivation days into buckets, we gain insights into when customers are most likely to engage again, allowing us to time win-back strategies effectively.

* On average, the total monthly sales amount to **₹80.32 Cr**. However, sales in **February, June, September, and November** consistently fell below this average, indicating underperformance during these months. In contrast, **August** emerged as an exceptionally high-performing month, significantly surpassing the average sales, suggesting a strong seasonal or promotional impact driving sales during this period.
* **Payment method:**

1. **UPI** transactions processed the highest **total quantity (9,14,282)**, yet total sales from UPI are lower than Debit Card and Cash. This suggests UPI transactions might involve lower-priced items or smaller purchase sizes. (refer excel sheet)
2. The **26–35** age group has the highest total transactions (**79,001**), indicating this segment as the most active and potentially the most valuable target audience for marketing efforts.
3. The **18–25** age group shows a slight inclination towards **Online** shopping (**7.01%**) compared to In-Store (**6.98%**), suggesting a potential opportunity to expand digital engagement for this segment.

* Although marginal, the **66–75** age group has a **negative percent difference (-0.02%)**, showing a slightly higher preference for **In-Store** shopping compared to Online, which might suggest a preference for traditional shopping experiences in older demographics.
* The **66-75** age group shows the lowest percentages, particularly for products like Notebook (**15.47%**) and Laptop (**15.42%**), suggesting lower engagement or spending from this older demographic.
* **Spending Bucket Tag:**

1. The **>3L bucket** includes **7,344 customers (15%)**, indicating a strong presence of high-value customers. So, we can Consider these 15% bucket as revenue driver of the business.
2. Customers in the **1L-1.5L (8,737)** and **1.5L-2L (8,927)** buckets together account for **36% of the total customers** and these is Core customer base.
3. The **0-50K bucket** includes only **3,981 customers (8%)**, is the smallest segment.
4. The **2.5L-3L bucket (5,391 customers)** is significantly smaller compared to the **1.5L-2L bucket (8,927)**, showing a potential challenge in transitioning mid-spenders to the highest-spending tier. Customers in the 2.5L-3L range might stagnate or drop out if they're not adequately incentivized to cross into the >3L segment.
5. There is no significant difference in spending power bifurcated by gender wise.
6. **East region shares highest contribution in the >3L bucket (4.08%)**, indicating strong high-spending customers from east region.
7. The **>3L bucket share (3.90%)** is lower in **North** compared **to East and South.**

* **Loyalty Points: -**

1. The largest bucket (6001-8000 points) has 41,393 customers, just 290 more than the smallest bucket (0-2000 points), indicating no major dominance of any bucket.
2. The highest transaction count (100,520) is observed in the 4001-6000 bucket, indicating high engagement from customers in this tier.
3. Despite accumulating high loyalty points, customers in the greater than 8000 points buckets exhibit a lower transaction frequency (2.42) and reduced engagement compared to other tiers.

* **Product:**

1. Lucknow ranks second among cities in terms of the number of transactions, demonstrating strong customer engagement. However, Delhi takes the second spot in total quantities purchased, indicating higher purchase volumes per transaction in the region
2. The feedback distribution across all products and scores remains consistent, with no significant variance observed. This indicates that feedback scores are not contributing to any measurable differentiation between products.
3. Conversely, **Ahmedabad** and **Mumbai** show relatively lower transaction counts across categories, suggesting potential opportunities to improve customer engagement.

* **Customer retention duration: -**

1. **Majority Reactivate Quickly:** Over half (53.13%) of the customers reactivated within the first 0-60 days, indicating a high initial re-engagement rate.
2. 45.30% of customers reactivated within 61-180 days. This shows a significant portion of customers take slightly longer to re-engage.
3. Only 1.57% of customers reactivated after 180 days (>240 days + 181-240 days). This indicates that reactivation rates drop drastically after six months, suggesting a need to revise or intensify strategies for long-term inactive customers.

* **Region\_wise:**

1. The South region accounts for the highest number of transactions (159,653) and total sales (₹319.07 crore), making it the strongest-performing region in terms of customer activity and revenue generation.
2. The **East region** has the second-highest number of transactions (128,328) but leads in **total quantity purchased** (960,384 units).
3. Both **West** and **North regions** show nearly identical figures for transactions (~101,000), distinct customers, and sales. This reflects balanced engagement, but slightly lower performance compared to South and East regions.
4. **South and East region** has the highest number of returns in quantity (13,428 units), but it still maintains strong overall sales. Analyze the root causes of high returns in the South and East regions. This may involve revising quality checks, return policies, or customer support processes.

* **Key Insights and Summary:**

1. **Data Pre-processing**: Missing fields were updated to NULL or categorized (e.g., Non-member customers), while incorrect data types (dates, amounts) were converted for accuracy.
2. **Dataset Overview**: Comprising 500,000 transactions from 48,995 unique customers between Jan 1, 2022, and Dec 14, 2022, with 10% missing customer attributes.
3. **Customer Segmentation**: Customers were tagged by purchase frequency (one-timer/multi-timer), spending power, age, and reactivation patterns for targeted strategies.
4. **Spending Insights**: High-value customers (>₹3L bucket) account for 15%, making them revenue drivers. Core customer base lies in ₹1L-₹2L buckets (36%).
5. **Regional Performance**: The South leads in sales (₹319.07 Cr), while the East leads in total units sold. Both regions exhibit high returns, requiring quality improvements.
6. **Loyalty Points Engagement**: Customers with 4,001–6,000 points show the highest transaction activity, while higher-point tiers (>8,000) display reduced engagement.
7. **Payment Method Patterns**: UPI dominates in quantity but not value; the 26–35 age group shows the highest activity, while 18–25 leans toward online shopping.
8. **Seasonal Trends**: Monthly sales average ₹80.32 Cr; February, June, September, and November underperform, while August shows exceptional growth.
9. **Reactivation Trends**: Over 53% of customers return within 0–60 days, while only 1.57% reactivate after six months, highlighting the need for long-term win-back strategies.
10. **Product and City Insights**: Lucknow ranks high in transactions, but Delhi outpaces in purchase volumes. Ahmedabad and Mumbai show lower engagement, indicating potential for growth.