**E-Commerce Customer Churn Analysis**

**Overview:**

This project focuses on data preprocessing and data exploration of a raw MySQL e-commerce dataset by solving 18 queries and performing data preprocessing. Each query addresses issues such as missing values, inconsistent data, and incorrect data types to transform the dataset into a clean and structured form. The assignment provides practice in MySQL query skills and prepares the data for further analysis, such as identifying customer churn patterns.

**Problem Statement:**In the e-commerce industry, businesses face the significant challenge of losing customers, which directly impacts revenue, growth, and long-term sustainability. Understanding why customers stop buying, identifying patterns in their behavior, and detecting inconsistencies in customer data are essential for addressing churn effectively. Without proper analysis and data preparation, companies cannot implement informed strategies to retain customers, improve loyalty, and enhance overall business performance.

**Objective:**

To clean and prepare the uncleaned e-commerce customer dataset in MySQL, including columns such as **Tenure, Satisfaction Score, Preferred Order Cat** and **Hour Spent on App,** ensuring the dataset is well-structured, consistent, and suitable for detailed analysis, while retaining long-term relationships with customers, enhancing loyalty, and improving business growth.

**Data description:**

**Dataset source:** The dataset was provided as part of the course materials by the instructor. The original source of the data is not specified.

**Key columns:** Customer ID,Tenure, Prefered Order Cat, Preferred Payment Mode, Satisfaction Score, Complain, Hours Spend on App,Order Amount Hike from Last Year, Day Since Last Order and Coupon used.

**Number of Rows and Columns:**

20 Columns 5642 Rows

**Data types:** 13 Integers 2 BIT 4 Varchar 1 Enumeration

**Data Preprocessing:**

**Handling Null Values:**  
Null values in the columns “Ware House to Home”, “Hour Spent on App”, “Order Amount Hikefrom Last Year”, and “Day Since Last Order” were imputed using the **mean (AVG () function)**. Values were rounded to the nearest integer for consistency.

Null values in the columns “Tenure”, “Coupon Used”, and “Order Count” were imputed using the **mode**, calculated with GROUP BY ORDER BY COUNT () to preserve the most frequent value in each column.

**Handling Outliers:**  
Outliers in the “Warehouse to Home” column were addressed by deleting rows where values exceeded 100

**Dealing with Inconsistencies:** To ensure uniformity and consistency.

* Occurrences of “Phone” in the Preferred Login Device column and “Mobile” in the Preferred Order Category column were replaced with “Mobile Phone”.
* In the Preferred Payment Mode column, “COD” was replaced with “Cash on Delivery” and “CC” was replaced with “Credit Card”.

**Data Transformation** **(Column renaming, creating new columns & dropping existing columns)**

Column renaming: was performed to correct spelling errors and ensure consistent naming conventions across the dataset.

* Renamed “Prefered Order Cat” to “Preferred Order Cat”.
* Renamed “Hour Spend on App” to “Hour Spent on App”.

Creating new columns: These transformations were implemented to make the dataset more structured and suitable for better analysis.

* A new column “Complaint Received” was created with values set to “**Yes”** if the corresponding value in the “Complain”column was **1**, and “**No”** otherwise.
* Similarly, a new column “Churn Status” was created, assigning churned if the value in the **“Churn”** column was **1**, and “**Active”** otherwise.
* Customers were grouped based on their distance from the warehouse, and a new column called “**Distance”** was added to reflect this classification Very Close Distance: ≤ 5 km Close Distance: > 5 km and ≤ 10 km Moderate Distance: > 10 km and ≤ 15 km Far Distance: > 15 km

Dropping existing columns: The columns “Churn” and “Complain” were dropped from the dataset to remove redundant information and avoid unnecessary columns, as their values are now represented in the newly created “Churn Status” and “Complaint Received” columns.

**Data Exploration and Analysis:**

Data exploration and analysis were performed by addressing 18 questions using MySQL queries. This process enabled a comprehensive understanding of each column in the dataset, supporting the identification of necessary adaptations, innovations, and techniques, as well as highlighting practices and elements to avoid. The analysis also revealed patterns and reasons behind customers ceasing online purchases, providing a clear foundation for strategies aimed at improving customer retention and business performance. As a result, the dataset is now clean, structured, and fully prepared for further analysis and visualization using BI tools such as Power BI or Tableau.

Examples:

* Display the average tenure and total cashback-amount of customers who churned?

select sum (Cash Back Amount) from customer churn where Churn Status in ("Churned"); select avg (Tenure) from customer churn where Churn Status in ("Churned"); select round (3.1762);

* Identify the city tier with the highest number of churned customers whose

preferred order category is Laptop & Accessory? Query: select City Tier, count (\*)

from customer churn

where Churn Status in ("Churned")

and Preferred Order Cat in ("Laptop & Accessory")

group by City Tier;

* List the top 3 preferred order categories with the highest average cashback amount? Query: select Preferred Order Cat, avg (Cash Back Amount) as Categories from customer churn group by Preferred Order Cat order by Categories desc;

Reference for the steps involved in data preprocessing and data exploration, along with analysis conducted through solving 18 analytical questions. <https://drive.google.com/file/d/1hSpaBnyamhp3O-n1CUa8nbUvmsySqxNs/view?usp=drive_link>

**Cleaned Dataset Summary**

**Key differences between the cleaned and the uncleaned dataset:**

**Columns and Rows:** Number of Rows and Columns in uncleaned Dataset: 20 Columns 5642 Rows. The dataset originally contained 20 columns and 5,642 rows. A new column named “Distance” was added, and 14 rows were removed due to outliers

Number of Rows and Columns in cleaned Dataset: 21 Columns 5628 Rows

Data types: Data types in uncleaned Dataset: 13 Integers 2 BIT 4 Varchar 1 Enumeration. Two BIT columns (churn and complain) were converted into VARCHAR columns named “Churn Status” and “Complaint Received”. Additionally, one new VARCHAR column, “Distance” was created Data types in cleaned Dataset: 13 Integers

7 Varchar 1 Enumeration

**Summarization:** 5628 Rows, 21 Columns, 13 Integers, 7 Varchar, 1 Enumeration The Integer fields are: CustomerID, Churn Status, Tenure, Day Tier, Warehouse To Home, Distance, Hour Spent on App, Number of Device Registered, Satisfaction Score, Number of Address, Complain Received, Order Amount Hike from Last Year, Coupon Used, Order Count, and Day Since Last Order. The Enumeration field is: Gender. The Varchar fields are: Preferred Login Device, Preferred Payment Mode, Preferred Order Cat, Marital Status, and Cashback Amount.

Finally, a copy of the cleaned dataset has been exported from the database as a CSV file to serve as a reference for future use, documentation, or any subsequent analysis that may be required. This ensures that a structured and reliable version of the dataset is readily available for further processing or review <https://drive.google.com/file/d/1DTdhh_iiZi1fevlKcVgwVA5cB0yi5EVt/view?usp=drive_link>

**Learning Outcomes**

This project is the module-end project of the Data Analytics course. Working on cleaning and analyzing an unclean dataset allowed me to develop advanced skills in MySQL, particularly in writing complex queries and managing databases efficiently. As part of this project, I solved 18 analytical questions, which helped me apply my theoretical knowledge to practical scenarios and strengthen my problem-solving abilities. This process enhanced my confidence in storing, retrieving, and manipulating large datasets, and in performing database operations accurately and efficiently. Overall, this project has significantly improved my technical skills in MySQL and data handling, preparing me to handle database-related tasks in data analytics and data science roles. This project was completed as part of the Data Analytics certification course from **Illinois Tech** through the **Entri Elevate** online learning platform.