Customer Behaviour Analysis Project

Task 1
Objective: Customer Behaviour Analysis
Company: Alfido Tech
Tech Stack: Python(Pandas, Seaborn, Matplotlib)

01

Overview

Introduction

This project aims to analyze customer behavior to understand purchasing patterns, preferences, and trends.
The analysis helps uncover underlying factors influencing customer decisions in a competitive market environment.

Problem Statement

The primary goals are to identify and analyze key behavioral trends, segment customers by demographics and purchasing patterns, and generate strategic insights. These insights enable targeted marketing campaigns, optimized product offerings, and improved customer relationship management, ultimately driving business growth and competitive advantage in a dynamic market environment.

02

Data Cleaning

Importing Libraries and Reading Data from CSV File

```
import numpy as np
import pandas as pd
import seaborn as sns
```

```
# Reading data from the CSV file and Creating a copy of the dataframe
df=pd.read_csv('ecommerce_customer_data_large.csv')
df_copy = df.copy()
```

We import NumPy, Pandas, and Seaborn to handle numerical operations, data manipulation, and visualization. The CSV file is read into a Pandas DataFrame, and a copy is created to preserve the original dataset for safe analysis.

Gathering Information about the Dataset

```
# Gathering information about the dataset

print(df_copy.info())
print(df_copy.shape)
print(df_copy.describe())
```

The dataset is first explored to gain a clear understanding of its structure and contents. Using info(), we can check the column names, data types, and presence of null values. The shape attribute provides the number of rows and columns, helping us gauge the dataset's size. Finally, describe() generates summary statistics such as mean, median, minimum, maximum, and standard deviation for numerical columns, giving an initial overview of the data distribution and helping identify any irregularities or patterns.

Data Cleaning Code 1

```
# Correcting Data types of columns
# Purchase Date col is not in correct format

df_copy['Purchase Date'] = pd.to_datetime(df_copy['Purchase Date'],dayfirst=True,errors='coerce')
```

```
# Checking for NaN values and filling (if any)
df_copy.isnull().sum()
# Null values found in the returns column
# As from the basic logic a customer ignores return if he/she has not returned the product

# Replacing NaN values
df_copy.fillna(0,inplace=True)

# Now check if the NaN values have been replaced
df_copy.isnull().sum() # NaN values successfully removed
```

In this step, the dataset is cleaned to make it ready for analysis. The Purchase Date column is converted into a proper datetime format to ensure consistency. Next, the dataset is checked for any missing values using isnull().sum(), and null entries are handled by replacing them with zeros using fillna(0, inplace=True). This ensures that the data remains complete, avoids calculation errors, and maintains logical consistency, especially in cases like returns where a missing value can reasonably be treated as zero.

Data Cleaning Code 2

```
# Checking for duplicate values and Replacing (if any)
df_copy.duplicated().any()

# No duplicates present
```

```
# Saving the cleaned data into a new xlsx file
df.to_excel("Cleaned Dataset Excel.xlsx")
```

The dataset is further cleaned by checking for duplicate records using **duplicated().any()**, ensuring that each entry is unique and reliable. Since no duplicates are found, the data is already consistent. Finally, the cleaned dataset is saved into a new Excel file using **to_excel()**, making it ready for future analysis and reporting.

03

Data Visualization and Analysis

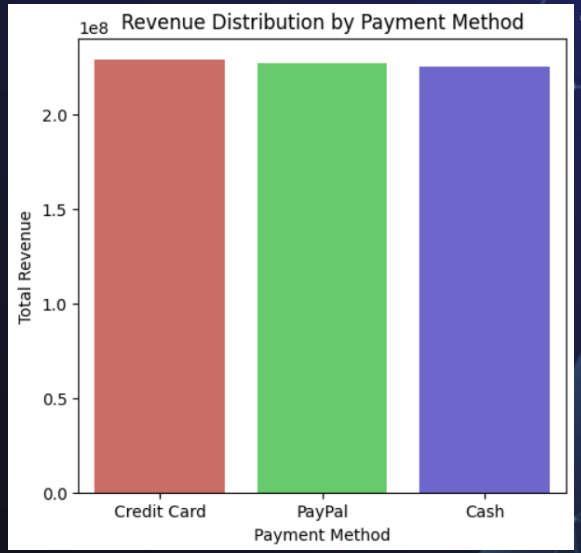
Visuals Used

The analysis incorporated various visual tools including bar charts, pie charts, KPI’s, tables, and scatter plots to illustrate customer segmentation, purchase behaviors, and Payment mode preferences. These visualizations enhance data comprehension by clearly highlighting significant trends, relationships, and outliers, thereby facilitating informed decision-making and strategic planning.

Visual 1

```
# 1.Most Preferred and Least Preferred Payment Method
# Bar Plot
# Group By
revenue_by_pmt_mthd = df.groupby('Payment Method')['Total Purchase Amount'].sum().reset_index()
revenue_by_pmt_mthd = revenue_by_pmt_mthd.sort_values(by='Total Purchase Amount',ascending=False)
print(revenue_by_pmt_mthd)

plt.figure(figsize=(5,5))
sns.barplot(data=revenue_by_pmt_mthd,x='Payment Method',y='Total Purchase Amount',palette='hls')
plt.title('Revenue Distribution by Payment Method')
plt.xlabel('Payment Method')
plt.ylabel('Total Revenue')
plt.tight_layout()

# This shows that all the payment methods are almost equally preferred.
```

✓ 0.5s



From this visualization, we can infer that all three payment methods—Credit Card, PayPal, and Cash—generate nearly the same amount of revenue. This indicates that customers do not have a strong preference for any specific payment method, and all options are almost equally popular. It also suggests that offering multiple payment methods helps cater to diverse customer choices without significantly impacting overall revenue distribution.

Visual 2

```
# 2.Top Category in terms of Revenue

# Creating a dataframe which contains grouped by data of top category
top_category = (df.groupby('Product Category')['Total Purchase Amount']
                .sum().reset_index().sort_values(by='Total Purchase Amount',ascending=False).head(1))

# Extracting the values of category and revenue
category = top_category['Product Category'].iloc[0]
revenue = top_category['Total Purchase Amount'].iloc[0]


# Creating a card visual
plt.figure(figsize=(1,1))
plt.text(0.5, 0.5, f"{category}\nRevenue: ${revenue:,.0f}",
         fontsize = 10, ha='center', va='center')
plt.axis('off')
plt.title('Top Category by Revenue', fontsize = 11)

# Hence we found that Home is the best selling Product Category
```

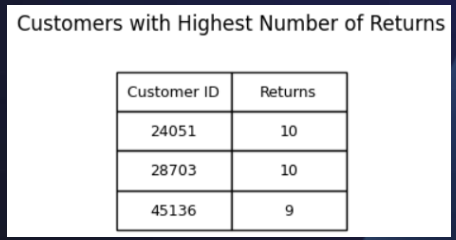
✓ 0.1s

The analysis shows that the Home category generates the highest revenue, amounting to $171,138,916. This indicates that products under the Home category are the most popular and contribute the most to overall sales, making it the best-performing category in terms of revenue.

Visual 3

```python
# 3. Top 3 customers who returned highest number of products by volume

# Grouping By Customer ID and saving the data
returns_by_customer = (df.groupby('Customer ID')['Returns'].sum().reset_index().sort_values(by='Returns',ascending=False).head(3))

# Creating a table
fig, ax = plt.subplots(figsize=(1,0.1))
ax.axis('off')

table = ax.table(
    cellText = returns_by_customer.values,
    colLabels = returns_by_customer.columns,
    colLoc='center',
    cellLoc='center'
)

table.scale(2.5, 2)  # Adjust size

# Title
plt.title("Customers with Highest Number of Returns", pad=20)
plt.show()

# From this we can ifer that Customer ID 24051, 28703 and 45136 return the products very frequently
```

✓ 0.1s

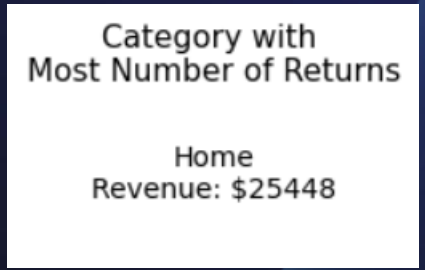


The insight from this analysis shows that customers with IDs 24051, 28703, and 45136 have the highest number of product returns, with 10, 10, and 9 returns respectively. This indicates that these customers frequently return items, which may suggest dissatisfaction, product quality issues, or unusual buying behavior. Monitoring such customers and identifying potential reasons behind frequent returns can help improve customer experience and reduce return-related losses.

Visual 4

```
# 4. Category with Highest No. of Returns

# Grouping by Category
returns_per_cat = (df.groupby('Product Category')['Returns'].sum().reset_index().sort_values(by='Returns',ascending=False))

cat = returns_per_cat['Product Category'].iloc[0]
returns = returns_per_cat['Returns'].iloc[0]

plt.figure(figsize=(1,1))
plt.text(0.5, 0.5, f"{category}\nRevenue: ${returns:}", fontsize=10, ha ='center', va = 'center')
plt.axis('off')
plt.title('Category with \nMost Number of Returns', fontsize=11)

# We can clearly see that the category Home gets the Highest number of returns
```

✓ 0.1s

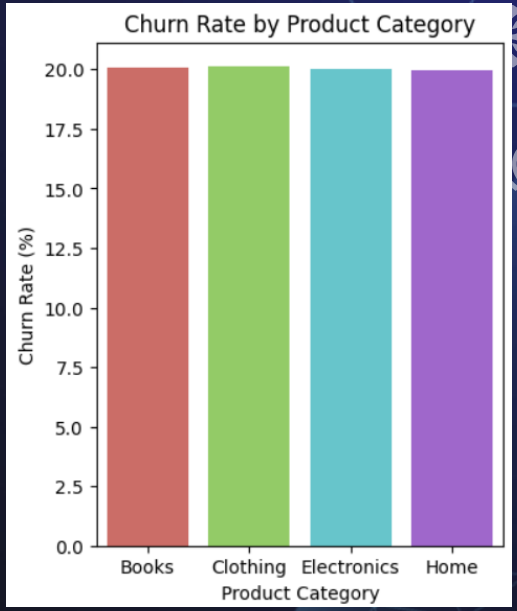


The analysis reveals that the Home category records the highest number of returns, with a return value of $25,448. While this category also generates the highest revenue, it simultaneously faces the most product returns, which could indicate potential issues such as product quality concerns, mismatched customer expectations, or logistics challenges. This insight highlights the need for deeper investigation into why returns are concentrated in this category.

Visual 5

```
# 5.Churn wrt Product Category

# Grouping by Category

churn_per_category = (df.groupby('Product Category')['Churn'].mean().reset_index().rename(columns={'Churn':'Churn Rate(%)'}))

churn_per_category['Churn Rate(%)'] = churn_per_category['Churn Rate(%)'] * 100

# Plotting a Bar Plot

plt.figure(figsize=(4,5))
sns.barplot(data=churn_per_category, x='Product Category', y='Churn Rate(%)', palette='hls')
plt.title('Churn Rate by Product Category')
plt.ylabel('Churn Rate (%)')
plt.show()

# This shows that the churn is constant along all the product categories
```

✓ 0.2s

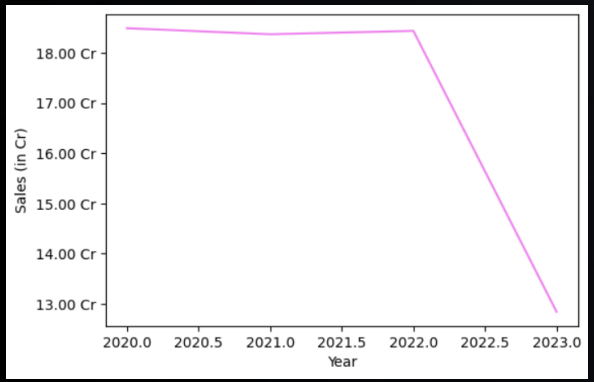


The churn analysis across product categories shows that the churn rate is nearly constant at around 20% for all categories—Books, Clothing, Electronics, and Home. This suggests that customer drop-off behavior is not strongly influenced by the type of product purchased, and other factors such as pricing, service quality, or overall customer experience might be driving churn uniformly across categories.

Visual 6

```python
# 6. Total Sales by Date

# Grouping by Date and the total sales
sale_by_date = (df.groupby('Purchase Date')['Total Purchase Amount'].sum().reset_index().sort_values(by='Purchase Date', ascending=True))

print(sale_by_date)

# Creating a line chart
plt.figure(figsize=(6,4))
plt.plot(sale_by_date['Purchase Date'],sale_by_date['Total Purchase Amount'], color='Violet')
plt.gca().yaxis.set_major_formatter(plt.FuncFormatter(lambda x, _: f'{x/1e7:.2f} Cr'))
plt.xlabel('Year')
plt.ylabel('Sales (in Cr)')

# As we can clearly see the Sales have drastically decreased in year 2022.
```

✓ 0.3s

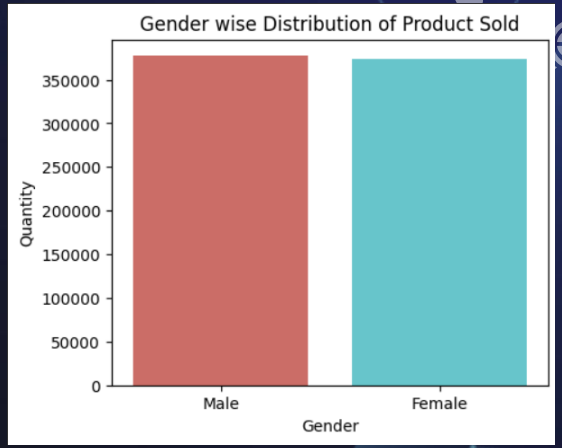


The trend of total sales over time shows that revenues remained fairly steady from 2020 through 2021, with only slight fluctuations. However, there is a sharp decline in sales during 2022, indicating a major drop in customer purchases. This downward trend could be due to external factors such as market shifts, reduced demand, or operational challenges, and it highlights the need for deeper investigation to understand the cause of this decline.

Visual 7

```
# 7. Gender wise distribution of Product Qty Sold

# Grouping by Gender on Quantity
gender_wise_qty = (df.groupby('Gender')['Quantity'].sum().reset_index().sort_values(by='Quantity',ascending=False))

# Creating Bar Chart

plt.figure(figsize=(7,5))
sns.barplot(data=gender_wise_qty, x = 'Gender', y = 'Quantity', palette='hls')
plt.title('Gender wise Distribution of Product Sold')
plt.xlabel('Gender')
plt.ylabel('Quantity')
```

✓ 0.2s

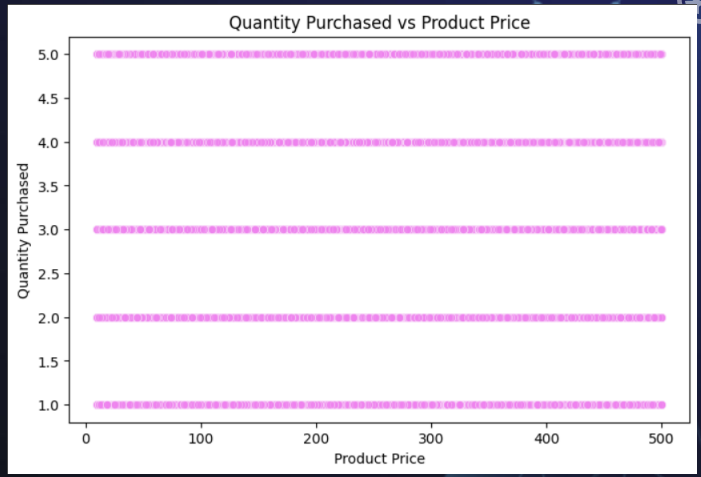


The gender-wise distribution of product sales shows that both male and female customers contributed almost equally to the total quantity sold, with only a slight edge for male customers. This indicates a balanced demand across genders, suggesting that products appeal broadly without strong gender-specific preferences. Such insights can help businesses focus on inclusive marketing strategies rather than targeting only one demographic.

Visual 8

```python
# 8. Relation between Quantity Purchased and Product Price

plt.figure(figsize=(8,5))
sns.scatterplot(
    data=df,
    x='Product Price',
    y='Quantity',
    alpha=0.6,
    color='Violet'
)
plt.title('Quantity Purchased vs Product Price')
plt.xlabel('Product Price')
plt.ylabel('Quantity Purchased')
plt.show()

# This clearly shows customers are purchasing the products regardless of price
```

✓ 1.2s



The relationship between product price and quantity purchased reveals that customers tend to buy products consistently across different price ranges. The scatterplot shows no significant correlation between higher prices and lower quantities, indicating that demand remains steady regardless of product cost. This suggests that customer purchase decisions are not heavily influenced by price, highlighting strong product appeal or brand loyalty.

04

Insights and Decision Making

Key Insights

1. Customers use all payment methods (Credit Card, PayPal, Cash) almost equally, showing no dominant preference.

2. The Home category generates the highest revenue ($171M+), but also records the largest number of returns ($25K+).

3. A few specific customers are responsible for frequent product returns, indicating possible dissatisfaction or unusual buying behavior.

4. Churn rate is steady at around 20% across all product categories, suggesting customer drop-off is not category-specific.

5. Sales remained stable during 2020–2021 but experienced a sharp decline in 2022, raising concerns about retention or market changes.

6. Gender-wise sales distribution shows near-equal contributions from male and female customers, reflecting balanced demand across demographics.

7. Price vs. Quantity analysis indicates no significant link between higher prices and lower purchase quantities, suggesting demand is not strongly price-sensitive.

Recommendations for Decision Makers

1. Investigate the high return rate in the Home category by improving product quality checks, enhancing descriptions, and strengthening after-sales service.

2. Continue offering multiple payment options to maintain customer convenience and inclusivity.

3. Address the consistent churn rate by introducing loyalty programs, personalized offers, and better engagement initiatives to retain customers across categories.

4. Examine the 2022 sales decline through customer feedback, competitive analysis, and market research, followed by corrective actions such as targeted promotions or pricing adjustments.

5. Leverage the balanced gender demand by designing inclusive marketing campaigns that appeal to both male and female customers.

6. Take advantage of the price-insensitive demand by introducing premium product lines, bundling strategies, or upselling opportunities, while ensuring strong brand positioning and product quality.

Thank you!

If you have any questions?
Connect on Socials.