**EXPLORATORY DATA A NALYSIS ON**

**BIKE SALES DATASET**

****

**INTRODUCTION**

This analysis focuses on a comprehensive exploration of a bike sales dataset containing 100,000 rows and 11 columns, specifically designed to uncover insights into customer behavior, sales trends, and performance metrics. The dataset has been meticulously curated, ensuring there are no missing or duplicate values, which enhances the reliability of the findings.

Initially, the dataset included columns such as Sale\_ID, Customer\_ID, and Salesperson\_ID, which were deemed unnecessary for this analysis and subsequently removed. The remaining columns include Bike\_Model, Price, Quantity, Store\_Location, Payment\_Method, Customer\_Age, and Customer\_Gender. To ensure accurate data processing, data types were thoroughly checked and adjusted as necessary.

To derive deeper insights, feature engineering was employed, resulting in the creation of several new variables. These include total sales, categorized age groups, and the extraction of year and month names from the sales data. By focusing on the relevant data, this analysis aims to uncover insights related to sales trends, customer demographics, and payment preferences within the bike sales landscape. Utilizing the power of pandas, we will explore various dimensions of the dataset to identify patterns that can inform business strategies and enhance customer engagement.

**AIM**

The aim of this analysis is to explore the bike sales dataset to uncover useful insights that can improve sales and customer satisfaction. By looking at important details such as bike models, prices, and customer information, this report will identify trends that affect buying choices. We will also examine how factors like store location and payment methods relate to sales volume. The insights gained from this analysis will help stakeholders make better decisions, allowing them to create effective marketing strategies and manage inventory more efficiently in the bike sales industry.

**COLUMNS**

1. Sale\_ID : A unique identifier for each sales transaction.

2. Date : The date when the sale occurred.

3. Customer\_ID : A unique identifier for each customer.

4. Bike\_Model : The specific model of the bike sold.

5. Price : The price at which the bike was sold.

6. Quantity : The number of bikes sold in the transaction.

7. Store\_Location : The location of the store where the sale took place.

8. Salesperson\_ID : A unique identifier for the salesperson involved in the sale.

9. Payment\_Method : The method used for payment (e.g., credit card, cash).

10. Customer\_Age : The age of the customer making the purchase.

11. Customer\_Gender : The gender of the customer.

12. Age\_cat : A categorical representation of customer age (e.g., youth, adult, senior).

13. Total\_Sales : The total revenue generated from the sale (calculated as Price × Quantity).

14. Month\_Name : The name of the month in which the sale occurred.

15. Year : The year in which the sale took place.

**OBJECTIVES**

Here are some key objectives for conducting exploratory data analysis (EDA) on the bike sales dataset:

**1.Understand Data Distribution:**

Analyze the distribution of key variables like `Price`, `Quantity`, `Customer\_Age`, and `Total\_Sales` to identify patterns or anomalies.

**2. Identify Sales Trends:**

Examine trends over time, such as monthly or yearly sales volumes, to understand seasonality and high-performing sales periods.

**3. Analyze Customer Demographics:**

Explore customer demographics, including `Customer\_Age`, `Customer\_Gender`, and `Age\_cat`, to gain insights into the customer profile.

**4. Evaluate Store Performance:**

Assess `Store\_Location` data to identify top-performing locations and understand the impact of location on sales.

**5. Examine Payment Preferences:**

Analyze `Payment\_Method` to determine the preferred payment options among customers, which can inform payment-related strategies.

**6. Explore Product Preferences:**

Investigate the popularity of different `Bike\_Model` types to understand which models contribute most to sales.

**7. Calculate Total Sales:**

Use `Total\_Sales` to evaluate overall revenue and profitability per transaction, enabling revenue trend analysis.

**8. Seasonal and Temporal Insights:**

Leverage `Month\_Name` and `Year` to discover seasonal and yearly changes in sales, identifying any long-term growth or decline patterns.

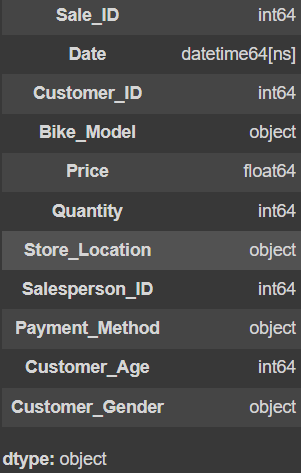
**9. Identify High-Value Customers:**

Examine customers who make high-value purchases, which can help in creating targeted marketing strategies.

**10. Detect Outliers or Unusual Patterns:**

Spot any unusual data points that may indicate data entry errors, exceptional sales, or unique customer behaviors.

**DATA OVERVIEW**



**ANALYSIS**

Sales Performance by Model and Location: Summarize total sales revenue and quantity sold for each bike model across different store locations.

Customer Demographics: Analyze the average age and gender distribution of customers for different bike models.

Payment Method Impact: Evaluate how different payment methods affect total sales and average transaction values.

Temporal Trends: Identify sales trends over time by month and year, focusing on total revenue and quantities sold.

Datatype :

1. df['Date']=pd.to\_datetime(df['Date'],format='%d-%m-%Y')

Outlier :

for col in df:

if df[col].dtype=='int' or df[col].dtype=='float':

print(col)

print('---------------')

print(f'Mean: {df[col].mean()}')

print(f'Median: {df[col].median()}')

print(f'Mode: {df[col].mode()}')

print(f'Minimum of {col}: {df[col].min()}')

print(f'Maximum of {col}: {df[col].max()}')

print(f'Variance of {col}: {df[col].var()}')

print(f'Standard Deviation: {df[col].std()}')

print()

plt.figure(figsize=(7,4))

plt.boxplot(df[col])

plt.title(col)

plt.show()

print('\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_')

print()

Feature Engineering :

1. df.drop(columns=['Sale\_ID','Customer\_ID','Salesperson\_ID'],inplace=True)
2. df['Age\_cat']=df['Customer\_Age'].apply(lambda x:'Seniors' if x>=60 else 'Youth' if x<=25 else 'Adults')
3. df['Total\_Sales'] = df['Price'] \* df['Quantity']
4. df['Month\_Name']=df['Date'].dt.month\_name()
5. df['Year']=df['Date'].dt.year
6. df['Is\_weekend'] = df['Date'].dt.dayofweek.isin([5,6]).astype(int)

Findings :

1. TOTAL COUNT OF EACH MODEL

df['Bike\_Model'].value\_counts()

The dataset shows a balanced distribution of bike sales across different models, with BMX bikes being the most popular and mountain bikes the least, indicating diverse customer preferences across types.

1. AGE DISTRIBUTION

df['Age\_cat'].value\_counts()

Adults make up the majority of bike purchases, followed by seniors and youth, suggesting that adults are the primary customer demographic for bike sales.

1. F M COUNT

df['Customer\_Gender'].value\_counts()

Bike purchases are almost evenly split between male and female customers, indicating a balanced gender interest in biking.

1. PAYMENT COUNT

df['Payment\_Method'].value\_counts()

Customers use a wide variety of payment methods fairly evenly, with a slight preference for digital options like Apple Pay and Debit Cards, indicating diverse but balanced payment preferences.

1. BIKE WITH HIGHEST REVENUE

df[['Bike\_Model','Total\_Sales']].groupby('Bike\_Model').sum().sort\_values(by='Total\_Sales',ascending=False)

The Hybrid Bike model generates the highest revenue, closely followed by BMX and Cruiser bikes, indicating strong sales performance and potential market demand for these models.

1. TOTAL SALES IN EACH LOCATION

df[['Store\_Location','Total\_Sales']].groupby('Store\_Location').sum().sort\_values(by='Total\_Sales',ascending=False)

New York leads in total bike sales, followed closely by Phoenix and Philadelphia, suggesting a strong market presence in these urban areas. The high sales figures across major cities indicate a robust demand for bikes, likely influenced by urban commuting trends and lifestyle choices.

1. AVERAGE PRICE OF EACH MODEL OF BIKE

df[['Bike\_Model','Price']].groupby('Bike\_Model').mean().sort\_values(by='Price',ascending=False)

The average prices of the bike models are closely clustered, with BMX bikes being the most expensive at approximately $2,608, while Folding Bikes are the least expensive at around $2,590. This small price variation suggests that consumers may prioritize features and quality over price when selecting a bike model.

1. TOTAL SALE BY AGE CATEGORY

df[['Age\_cat','Total\_Sales']].groupby('Age\_cat').sum().sort\_values(by='Total\_Sales',ascending=False)

Adults contribute the most significantly to total bike sales, with nearly $500 million, followed by seniors and youth, who account for $162 million and $116 million, respectively. This data highlights that adults are the primary consumers in the bike market, reflecting their greater purchasing power or interest in biking.

1. TOTAL SALES BY FEMALE IN EACH CATEGORY

female\_sales=df[df['Customer\_Gender']=='Female']

female\_sales[['Age\_cat','Total\_Sales']].groupby('Age\_cat').sum().sort\_values(by='Total\_Sales',ascending=False)

Total sales among female customers reveal that adults lead with approximately $253.6 million, indicating a strong preference for biking among adult women.

1. TOTAL SALES BY MALE IN EACH CATEGORY

male\_sales=df[df['Customer\_Gender']=='Male']

male\_sales[['Age\_cat','Total\_Sales']].groupby('Age\_cat').sum().sort\_values(by='Total\_Sales',ascending=False)

Total sales among male customers show that adults account for approximately $245.9 million, reflecting a significant interest in biking within the adult male demographic.

1. COUNT OF EACH AGE CATEGORY FOR CUSTOMERS WHO USE CASH AS PAYMENT

cash\_payments = df[df['Payment\_Method'] == 'Cash']

cash\_payments['Age\_cat'].value\_counts()

Among customers who used cash as a payment method, adults represent the largest group with 10,596, followed by seniors at 3,526 and youth at 2,570, indicating that cash transactions are predominantly made by adults.

1. AVG SALE IN EACH MONTH OF 2023

df\_23=df[df['Date'].dt.year==2023]

df\_23[['Month\_Name','Total\_Sales']].groupby('Month\_Name').mean().sort\_values(by='Total\_Sales',ascending=False)

Monthly average sales in 2023 show that January leads with approximately $7,953, followed closely by February and August, while sales peak during the first half of the year. The data suggests a decline in sales as the year progresses, particularly noticeable in the latter months, indicating potential seasonal trends in bike purchasing behavior.

1. AVG SALES IN EACH YEAR

df.groupby(df['Date'].dt.year)['Total\_Sales'].mean()

Average sales show a relatively stable trend from 2020 to 2024, with 2021 experiencing a slight peak at approximately $7,809, while the figures for subsequent years remain close, indicating consistent demand in the bike market over this period. Overall, the data suggests that sales have remained resilient despite potential market fluctuations.

1. USAGE OF PAYMENT METHOD BY EACH GENDER

df.groupby(['Customer\_Gender','Payment\_Method']).size().unstack(fill\_value=0)

Females predominantly use Apple Pay, Debit Card, and PayPal, while males show slightly higher usage of Cash and Credit Card. Overall, the data indicates a balanced but distinct preference in payment methods between genders, with females leaning towards digital options.

1. USAGE OF PAYMENT METHOD BY EACH AGE CATEGORY

df.groupby(['Age\_cat','Payment\_Method']).size().unstack(fill\_value=0)

Adults are the primary users of all payment methods, with the highest usage for Debit Card and Apple Pay, while seniors and youth show significantly lower engagement across all options. This trend suggests that adults have a stronger preference and greater accessibility to various payment methods compared to the other age categories.

1. WEEKEND v/s WEEKDAY SALES

df.groupby('Is\_weekend')['Total\_Sales'].mean()

Weekday and weekend sales are almost identical, with weekdays showing a slight edge, suggesting a steady demand for bikes throughout the week without a strong weekend preference.

1. GROUP BY BIKE MODEL AND CALCULATE THE AVERAGE PRICE AND QUANTITY

df.groupby('Bike\_Model').agg({'Price': 'mean', 'Quantity': 'sum'})

The average price and quantity sold across bike models are relatively consistent, with only slight variations, indicating steady demand and similar pricing across models. This suggests that price is less of a differentiator, with consumers likely choosing based on model features and personal preference.

1. YEARLY SALES (PLOT)

df\_year=df.sort\_values('Year')

res=df\_year[['Year','Total\_Sales']].groupby('Year').agg('mean')

plt.figure(figsize=(10,4))

plt.plot(res.index,res)

plt.title('Year v/s Mean Total Sales')

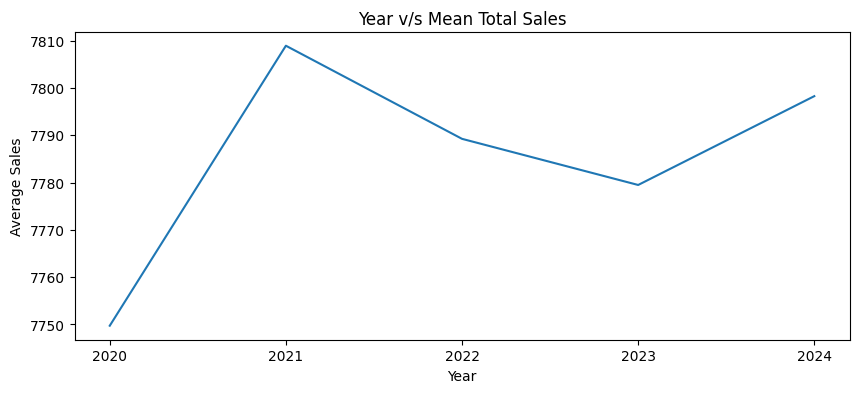
plt.xlabel('Year')

plt.ylabel('Average Sales')

plt.xticks(range(2020,2025)

plt.show()

Increasing Sales: The chart shows a general upward trend in average sales over the years. There is a significant increase from 2020 to 2021, followed by a slight dip in 2022 and 2023. However, the sales recover in 2024 and surpass the previous years' average sales.



1. MONTHLY SALE IN EACH YEAR

# Monthly Sales Trend

monthly\_sales = df.groupby(['Year', 'Month\_Name'])['Total\_Sales'].sum().reset\_index()

# Plotting the monthly sales trend

plt.figure(figsize=(14, 7))

sns.lineplot(data=monthly\_sales, x='Month\_Name', y='Total\_Sales', hue='Year', marker='o')

plt.title('Monthly Sales Trend (by Year)')

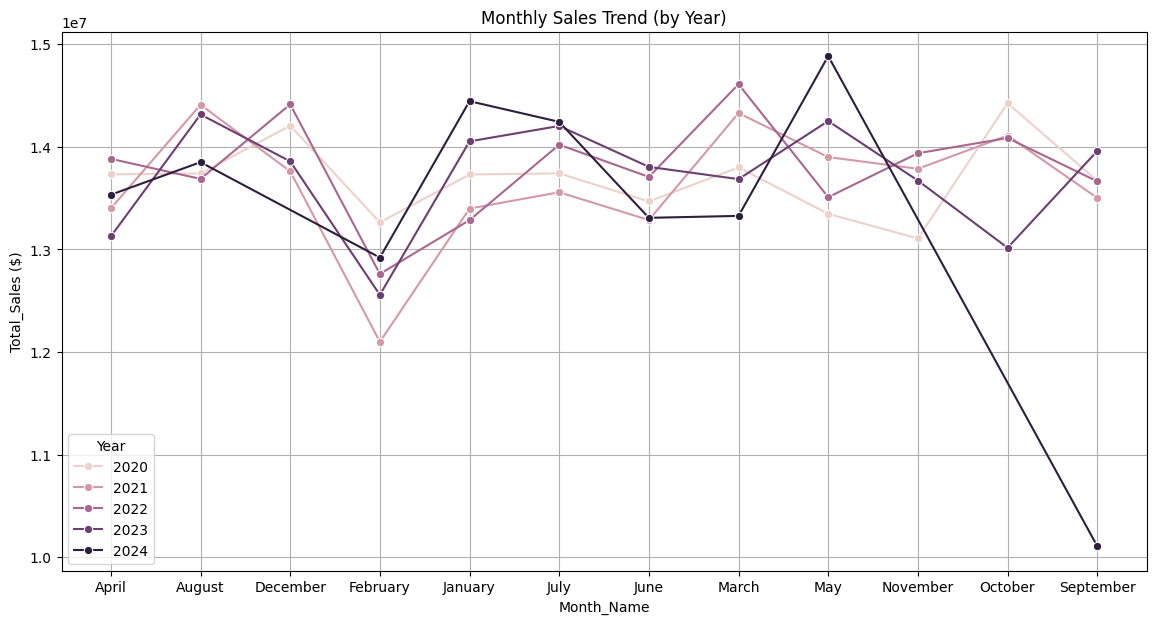
plt.xlabel('Month\_Name')

plt.ylabel('Total\_Sales ($)')

plt.legend(title='Year')

plt.grid(True)

plt.show()



Overall Trend:

The overall trend is slightly upward, with some fluctuations. There is a significant dip in sales in 2024, especially in September.

Yearly Performance:

2020: Sales started strong and peaked in August. There was a decline towards the end of the year.

2021: Sales were relatively consistent throughout the year, with a slight dip in February and a peak in August.

2022: Sales were strong in the first half of the year, peaking in June. There was a decline in the second half, with a recovery in November and December.

2023: Sales were relatively stable throughout the year, with a slight dip in February and a peak in May.

2024: Sales were strong in the first half of the year, peaking in June. There was a significant decline in the second half, with a sharp drop in September.

There seems to be a seasonal pattern with sales peaking in the summer months (June-August) and dipping in the winter months (December-February).

The fluctuating bike sales likely result from a combination of seasonal factors, economic conditions, marketing efforts, and technological advancements. Favorable weather in spring and summer, economic fluctuations influencing consumer spending, aggressive marketing campaigns, and the introduction of new bike models or features can all contribute to these cycles.

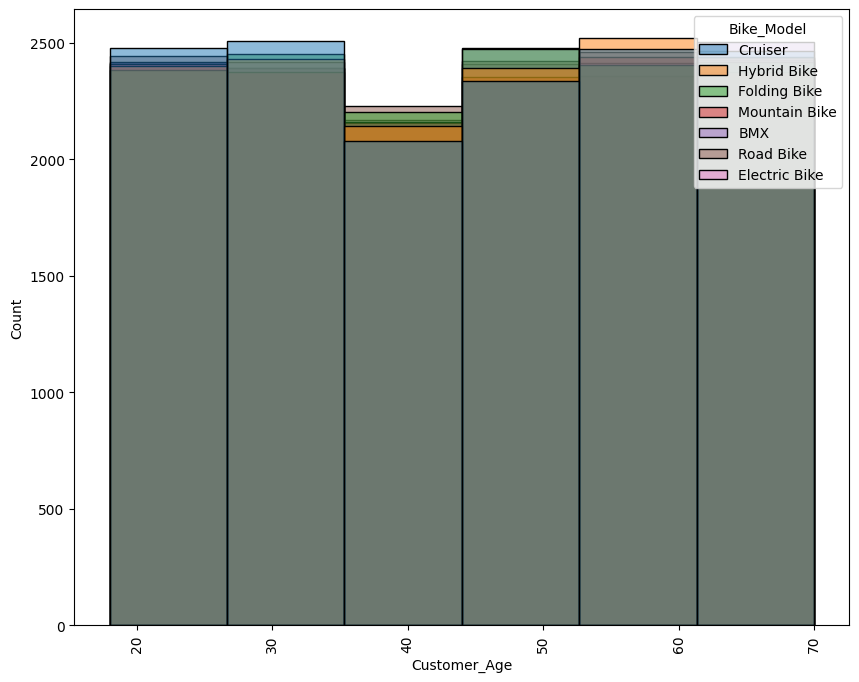
1. HISTPLOT

plt.figure(figsize=(10,8))

sns.histplot(data=df,x='Customer\_Age',bins=6,hue='Bike\_Model')

plt.xticks(rotation=90)

plt.show()



Bike Model Insights:

Mountain Bikes: Most popular across all age groups, especially among younger customers.

Road Bikes: Second most popular, favored by younger and middle-aged customers.

Cruiser Bikes: Gain popularity among middle-aged customers.

Hybrid Bikes: Gain popularity among older customers, especially those aged 50+.

Electric Bikes: Becoming increasingly popular, especially among older customers.

BMX, Folding Bikes: Less popular overall, with some demand from younger customers.

The graph suggests a shift in preferences as customers age. Younger customers tend to prefer performance-oriented bikes like Mountain Bikes and Road Bikes, while older customers lean towards comfort and convenience with Hybrid Bikes and Electric Bikes.

The increasing popularity of Electric Bikes could be attributed to various factors like convenience, accessibility, and environmental concerns.

Understanding these trends can help bike retailers tailor their inventory and marketing strategies to cater to different customer segments.

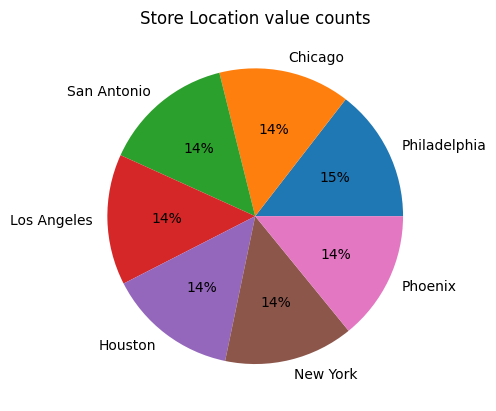
1. PIE PLOT

plt.pie(df['Store\_Location'].value\_counts(),labels=df['Store\_Location'].unique(),autopct="%1.0f%%")

plt.title("Store Location value counts")

plt.plot()

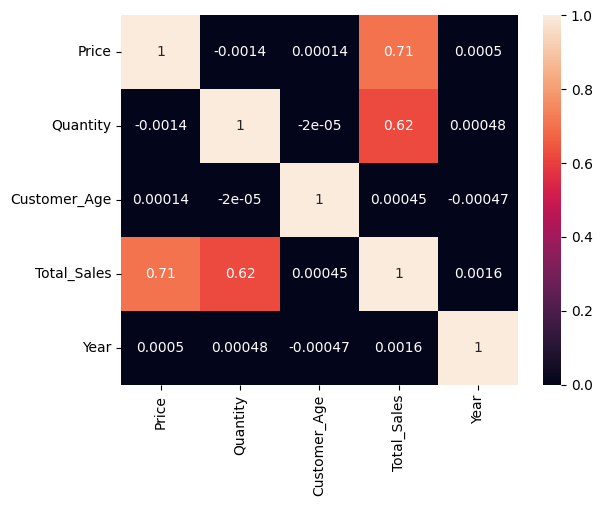
The pie chart illustrates the distribution of bike sales across different store locations. Philadelphia holds the largest share with 15%, followed closely by Chicago, Phoenix, and San Antonio, each contributing around 14% to the total sales. Los Angeles, Houston, and New York each account for 14% of the sales, indicating a relatively even distribution across these locations.



1. HEAT MAP

df.corr(numeric\_only=True)

sns.heatmap(df.corr(numeric\_only=True),annot=True)



The heatmap displays the correlation between numerical variables in the dataset. We observe a strong positive correlation between 'Price' and 'Total\_Sales', indicating that higher prices generally lead to higher total sales. Similarly, 'Quantity' and 'Total\_Sales' show a positive correlation, suggesting that larger quantities sold contribute to higher total sales. Interestingly, there is a weak negative correlation between 'Customer\_Age' and 'Total\_Sales', implying that older customers might tend to spend less. The other variables have negligible correlations with each other.

**CONCLUSION**

1. Model Popularity : BMX and Hybrid bikes lead in sales and revenue, indicating strong demand in versatile models.

2. Demographics : Adults dominate the customer base, while seniors and youth show less engagement, suggesting a need for targeted marketing.

3. Gender Balance : Sales are nearly equal between males and females, with females preferring digital payment methods, highlighting potential for gender-specific marketing.

4. Seasonal Trends : Higher sales in early months indicate opportunities for promotions during peak times and strategies to boost sales in slower months.

5.Payment Preferences : Adults favor a variety of payment methods, particularly Debit Cards and Apple Pay, suggesting enhancements in digital payment options could be beneficial.

Reasons for Current Trends

1. Market Demand: The popularity of certain bike models and payment methods reflects broader consumer trends, including the growing interest in cycling for fitness, commuting, and leisure.

2.Target Demographics: The substantial adult market suggests a more established consumer base, while seniors and youth may have differing needs or accessibility issues that limit their engagement.

3.Economic Factors: The stable average sales indicate resilience in the bike market, possibly due to increased focus on health and outdoor activities following the pandemic.

Recommendations for Improving Sale

1. Targeted Marketing : Develop campaigns aimed at seniors and youth, emphasizing biking benefits that resonate with their lifestyles.

2. Seasonal Promotions : Launch discounts and incentives during peak and slower sales months to encourage purchases.

3. Engage Female Customers : Focus on campaigns that appeal to female riders through community-oriented events and testimonials.

4.Community Involvement : Host local events and workshops to promote cycling, increasing brand visibility and attracting new customers.

5. Feedback and Adaptation : Regularly gather customer feedback to adapt products and services to evolving consumer preferences.