# Evoastra Ventures Intern Assessment Task

**Duration: 35 Minutes | Total Points: 100**

---

## Candidate Information

**Fill in your details below:**

- **Name:**
- **Email:**
- **Phone:**
- **College/University:**
- **Course/Branch:**
- **Start Time:**
- **End Time:**

## Dataset for Assessment

**E-commerce Customer Behavior Dataset**

- **Direct Link:** https://www.kaggle.com/datasets/shriyashjagtap/e-commerce-customer-for-behavior-analysis

**Dataset Columns:**

- **Customer ID:** Unique identifier for each customer
- **Customer Name:** Name of the customer
- **Customer Age:** Age of the customer
- **Gender:** Gender of the customer
- **Purchase Date:** Date of each purchase
- **Product Category:** Category of the purchased product
- **Product Price:** Price of the purchased product
- **Quantity:** Quantity of product purchased
- **Total Purchase Amount:** Total amount spent in each transaction
- **Payment Method:** Payment method used (credit card, PayPal, etc.)
- **Returns:** Whether customer returned products (0 = No, 1 = Yes)
- **Churn:** Whether customer has churned (0 = Retained, 1 = Churned)

```
import pandas as pd
import os
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
```

```
import warnings
from scipy.stats import chi2_contingency

warnings.filterwarnings("ignore")

sns.set_theme(
    style = "whitegrid",
    palette = "colorblind",
    context = "notebook"
)

plt.style.use("seaborn-v0_8")

plt.rcParams.update({
    "axes.titlesize" : 12,
    "axes.labelsize" : 8,
    "axes.grid" : True,
    "font.size" : 4,
    "lines.linewidth" : 3
})

np.set_printoptions(precision = 4 , suppress = True)

pd.set_option("display.float_format" , "{:.4f}".format)

os.getcwd()
```

```
'C:\\Users\\KISHORE\\Downloads\\ecommerce_customer_data_large.csv'
```

```
os.chdir("C:/Users/KISHORE/Downloads/
ecommerce_customer_data_large.csv")

e_com_data = pd.read_csv("ecommerce_customer_data_large.csv")
e_com_data.head()
```

```
   Customer ID        Purchase Date Product Category  Product Price
Quantity  \
0        44605  2023-05-03 21:30:02             Home            177
1
1        44605  2021-05-16 13:57:44      Electronics            174
3
2        44605  2020-07-13 06:16:57            Books            413
1
3        44605  2023-01-17 13:14:36      Electronics            396
3
4        44605  2021-05-01 11:29:27            Books            259
4

   Total Purchase Amount Payment Method  Customer Age  Returns
Customer Name  \
0                  2427         PayPal            31   1.0000   John
```

```
Rivera
1                   2448          PayPal              31   1.0000   John
Rivera
2                   2345    Credit Card               31   1.0000   John
Rivera
3                    937            Cash              31   0.0000   John
Rivera
4                   2598          PayPal              31   1.0000   John
Rivera

   Age  Gender  Churn
0   31  Female      0
1   31  Female      0
2   31  Female      0
3   31  Female      0
4   31  Female      0
```

---

# SECTION A: Data Understanding & Basic Analysis (25 Points - 10 Minutes)

## Question 1 (8 points)

Based on the dataset structure, identify the **data types** for the following columns and explain why each classification is important for analysis:

- **Customer Age**
- **Gender**
- **Total Purchase Amount**
- **Churn**

```
# Your Answer for Question 1
# Customer Age:
'''
- Customer Age is a integer data type.
- it contains the age of customers from 18 years to 70 years.
- if we observe basic statistics below, distribution of the customer
age is low variance (becouse small Standard deviation ~15).
- mean and median (that is 50%) in very close , this suggest no
outliers in customer age column
- Hence modest of the data is around the typical value (that is median
~ 44 Age).So, it is very important feature for our churn analysis.
'''
# Gender:
'''
- Gender columns is a object(categorical) data type(male/female).
- below countplot shows the destribution of this column , where both
```

```
Male and Females customers are almost same.
- It is also a important feature becouse , the two categories helps to easily label churn customer and non churn customers by also considering other properties.
'''
# Total Purchase Amount:
'''
- Total Purchase Amount is a integer type column.
- we can see in basic statistics , this feature does not have any extream values becouse mean and median almost same values.data varing from center value by 15.3649 standard deviation.
- Most of the data around the center value
- Similarly Age column , it also important feature for our customer churn analysis.
'''

# Churn:
'''
- it is a integer data type(0/1) , 0 means customer not churn , 1 means churn respectivily.
- Churn is the Target feature , this we need to predict.
-  Hence ii is mandatory in our analysis
- also , Below countplot absorve , Class imbalence.non chrun customers (0) records are ~200000 this larger than churn customer(1) records ~50000 only in E-commerce Customer Behavior Dataset.
'''

#code
e_com_data.info()

e_com_data.describe()
```

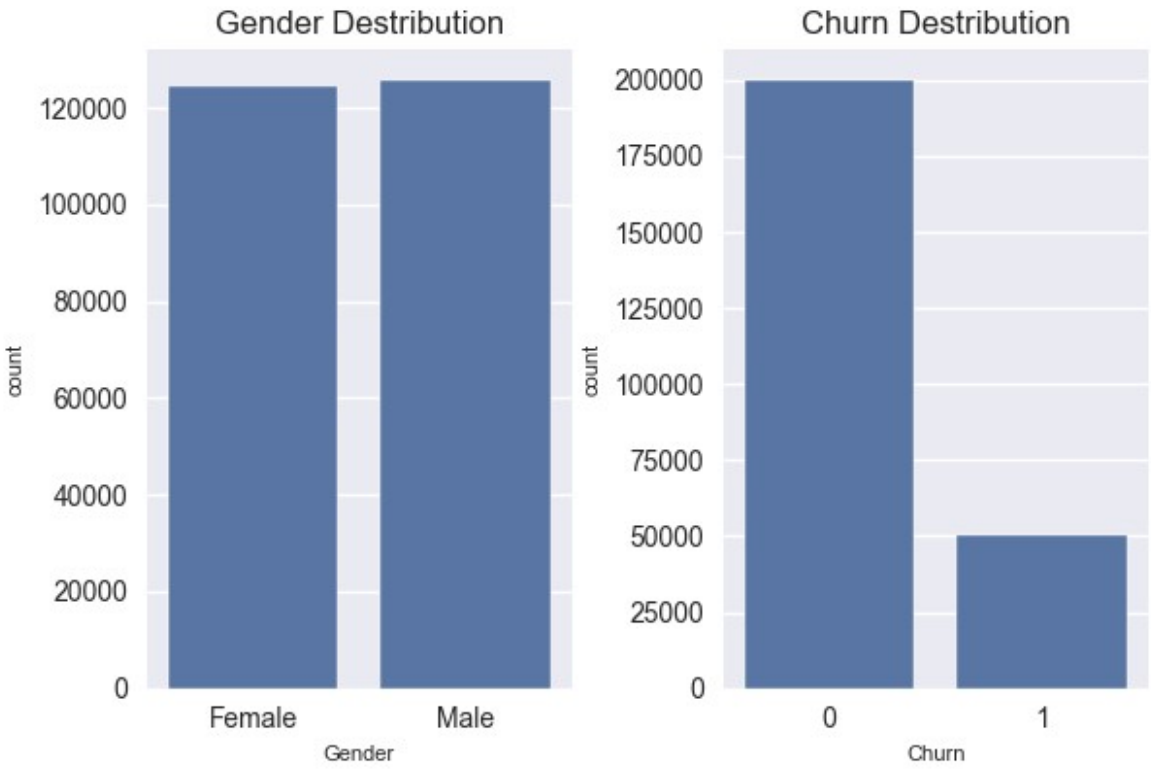
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250000 entries, 0 to 249999
Data columns (total 13 columns):
 #   Column                 Non-Null Count   Dtype  
---  ------                 --------------   -----  
 0   Customer ID            250000 non-null  int64  
 1   Purchase Date          250000 non-null  object 
 2   Product Category       250000 non-null  object 
 3   Product Price          250000 non-null  int64  
 4   Quantity               250000 non-null  int64  
 5   Total Purchase Amount  250000 non-null  int64  
 6   Payment Method         250000 non-null  object 
 7   Customer Age           250000 non-null  int64  
 8   Returns                202618 non-null  float64
 9   Customer Name          250000 non-null  object 
 10  Age                    250000 non-null  int64  
 11  Gender                 250000 non-null  object 
 12  Churn                  250000 non-null  int64
```

```
dtypes: float64(1), int64(7), object(5)
memory usage: 24.8+ MB
```

```
       Customer ID  Product Price    Quantity  Total Purchase
Amount  \
count  250000.0000    250000.0000 250000.0000             250000.0000

mean    25017.6321       254.7427      3.0049               2725.3852

std     14412.5157       141.7381      1.4147               1442.5761

min         1.0000        10.0000      1.0000                100.0000

25%     12590.0000       132.0000      2.0000               1476.0000

50%     25011.0000       255.0000      3.0000               2725.0000

75%     37441.2500       377.0000      4.0000               3975.0000

max     50000.0000       500.0000      5.0000               5350.0000
```

```
       Customer Age     Returns          Age       Churn
count   250000.0000 202618.0000 250000.0000 250000.0000
mean        43.7983      0.5008     43.7983      0.2005
std         15.3649      0.5000     15.3649      0.4004
min         18.0000      0.0000     18.0000      0.0000
25%         30.0000      0.0000     30.0000      0.0000
50%         44.0000      1.0000     44.0000      0.0000
75%         57.0000      1.0000     57.0000      0.0000
max         70.0000      1.0000     70.0000      1.0000
```

```
cols = ["Gender" , "Churn"]

fig , axes = plt.subplots(1 , 2 , figsize = (6  , 4))
axes = axes.flatten()

for i , col in enumerate(cols):
    sns.countplot(x = e_com_data[col] , ax = axes[i])
    axes[i].set_title(f"{col} Destribution")
plt.tight_layout()
plt.show()
```




## Question 2 (8 points)

Which **analytical technique** would be most appropriate for each business question below?

a) "Which product categories generate the highest revenue?"

b) "Can we predict customer churn based on purchase behavior?"

c) "What is the relationship between customer age and spending patterns?"

d) "Which payment methods are preferred by different customer segments?"

```
# Your Answer for Question 2
# a) Product categories with highest revenue:
'''
- below we can see,Home is generate the highest revenue 48130856
comparet to other product categories.
- then Clothing wiht 47977746 and Electronics with 47801925 , both are
not have that much of difference in revenue
- low revenue from Books category by 47578138 respectively.
'''
# b) Predicting customer churn:
'''
- when we talk about customer purchase behavior there are three main
features they are Product Category ,Quantity ,Payment Method
respectively.
- if we use Contingency Table for purchase behavior feature vs Churn ,
```

```
it give this frequency distribution of the variables. it follows in
below
- There clearly shows , there is no differences between desctribution
of any of this  three feature against to the churn , they are same in
distribution frequencies.
- Hence, we Conclude "We Can not predict customer churn based on
purchase behavior (Product Category ,Quantity ,Payment Method)" , it
do model give unreliable predictions.
'''
# c) Age and spending relationship:
'''
- Below we can see the relationship between Age and spending.
- both are with positive correlation with 0.0566 value , this value is
near to zero.
- does't have strong positive correlation between this two features.
'''

# d) Payment method preferences:
'''
- if we segment customers by Customer Age "Young"(18-30) , "Middle-
aged"(31-50) , "Senior"(51+) respectively.
- the Contingency Table shows , payment methods are preferred by
different customer segments.
- Credit Card method preferred 81995 number of times and Cash 81398
times, PayPal 81819 times.
- Credit Card is highly preferred as Payment methods by different
customer segments.
'''

#code
e_com_data["revenue"] = e_com_data["Product Price"] *
e_com_data["Quantity"]

e_com_data.groupby("Product Category")
["revenue"].sum().sort_values(ascending = False)
```

```
Product Category
Home           48130856
Clothing       47977746
Electronics    47801925
Books          47578138
Name: revenue, dtype: int64
```

```
cols = ["Product Category" , "Quantity" , "Payment Method"]

for col in cols:
    print(f"Contingency Table for {col}")
    print(end = "\n")
    print(pd.crosstab(e_com_data["Churn"] , e_com_data[col] , margins
= True) , end = "\n\n\n")
```

```
Contingency Table for Product Category

Product Category  Books  Clothing  Electronics   Home     All
Churn
0                 49748     49988        50077  50057  199870
1                 12499     12593        12553  12485   50130
All               62247     62581        62630  62542  250000


Contingency Table for Quantity

Quantity      1      2      3      4      5     All
Churn
0         39857  39724  39898  40166  40225  199870
1         10032   9961  10049  10095   9993   50130
All       49889  49685  49947  50261  50218  250000


Contingency Table for Payment Method

Payment Method   Cash  Credit Card  PayPal     All
Churn
0               66200        67120   66550  199870
1               16812        16427   16891   50130
All             83012        83547   83441  250000



num_cols = e_com_data.select_dtypes(include = np.number).columns
print(e_com_data[num_cols].corr()["Total Purchase
Amount"].sort_values(ascending = False))

plt.figure(figsize = (6 , 4))
sns.scatterplot(x = "Customer Age" , y = "Total Purchase Amount" ,
data = e_com_data)
plt.title("Customer Age vs Total Purchase Amount")
plt.show()

Total Purchase Amount    1.0000
Age                      0.0566
Customer Age             0.0566
Customer ID              0.0013
Quantity                 0.0012
Returns                  0.0010
Churn                    0.0007
revenue                  0.0004
Product Price           -0.0013
Name: Total Purchase Amount, dtype: float64
```


```
bins = [18 , 30 , 50 , 70]
labels = ["Young" , "Middle-aged" , "Senior"]

e_com_data["age_group"] = pd.cut(e_com_data["Customer Age"] , bins =
bins , labels = labels)

pd.crosstab(e_com_data["Payment Method"] , e_com_data["age_group"],
margins = True)
```

| age_group | Young | Middle-aged | Senior | All |
|---|---|---|---|---|
| Payment Method | | | | |
| Cash | 19226 | 31174 | 30998 | 81398 |
| Credit Card | 19374 | 31284 | 31337 | 81995 |
| PayPal | 19625 | 31268 | 30926 | 81819 |
| All | 58225 | 93726 | 93261 | 245212 |

## Question 3 (9 points)

**Data Quality Assessment:** What are the top 3 potential data quality issues you would check for in this e-commerce dataset before starting analysis? For each issue, suggest one method to detect it.

```
# Your Answer for Question 3
# Issue 1:
'''
- Duplicate records
'''
```

```
# Detection Method 1:
'''
Use: DataFrame.duplicated()
example: e_com_data[e_com_data.duplicated()]
this give the duplicate records in your dataset
'''

# Issue 2:
'''
- Null Values or Empty record
'''

# Detection Method 2:
'''
there are three ways to detect the null values in our dataset
way 1: info() this simply suggest which columns have null records by
giving number of nun-null records
ex: e_com_data.info()

way 2: isnull() or isnull().sum()
ex:e_com_data[e_com_data.isnull()] , e_com_data.isnull().sum()
(recomended)

way 3: missingno library
- it used to see null records in dataset in visuvaly
ex: missingno.bar(e_com_data)
'''

# Issue 3:
'''
- Outliers or extream values
'''

# Detection Method 3:
'''
Visuvalization method:
- Box plot

Statistical Method:
- Z-Score Method (Standard Deviation Method) (if data is nurmaly
described)
- IQR Outlier Detection
- MAD (Median Absolute Deviation) (recomended)
'''

'\nVisuvalization method:\n- Box plot\n\nStatistical Method:\n- Z-
Score Method (Standard Deviation Method) (if data is nurmaly
described)\n- IQR Outlier Detection\n- MAD (Median Absolute
Deviation) (recomended)\n'
```

---

# SECTION B: Customer Analysis & Business Intelligence (35 Points - 15 Minutes)

## Scenario: E-commerce Revenue Analysis

Based on the dataset structure, assume you have the following customer insights:

**Customer Segments by Age:**

- **Young (18-30):** 40% of customers, Average Purchase Amount: ₹850, Return Rate: 12%
- **Middle-aged (31-50):** 45% of customers, Average Purchase Amount: ₹1,200, Return Rate: 8%
- **Senior (51+):** 15% of customers, Average Purchase Amount: ₹950, Return Rate: 15%

**Additional Information:**

- Average customer acquisition cost: ₹180
- Platform profit margin: 20% of purchase amount
- Customer churn rates: Young (25%), Middle-aged (15%), Senior (30%)

## Question 4 (15 points)

**Calculate and analyze:**

a) Which customer segment generates the highest **net profit per customer** (considering returns)? Show your calculations. (8 points)

b) Which segment has the **best customer lifetime value** considering churn rates? Provide reasoning. (7 points)

```
# Your Calculations for Question 4

# a) Net profit per customer calculations:
'''
Net Profit per Customer = [(Average Purchase Amount × (1 − Return
Rate)) × Platform profit margin] − Average customer acquisition
cost(CAC)
'''
# Young Customers:
'''
Give that:
Average Purchase Amount: ₹850
Return Rate: 12%
'''

young_customers_Net_Profit = ((850 * (1 - 0.12)) * 0.20) - 180
print("Young Customers Net Profit per Customer is : " ,
```

```
round(young_customers_Net_Profit , 4))

# Middle-aged Customers:
'''
Give that:
Average Purchase Amount:  ₹1,200
Return Rate: 8%
'''

Middle_customers_Net_Profit = ((1200 * (1 - 0.08)) * 0.20) - 180
print("Middle-aged Customers Net Profit per Customer is : " ,
round(Middle_customers_Net_Profit , 4))

# Senior Customers:
'''
Give that:
Average Purchase Amount: ₹950
Return Rate: 15%
'''

Senior_customers_Net_Profit = ((950 * (1 - 0.15)) * 0.20) - 180
print("Senior Customers Net Profit per Customer is : " ,
round(Senior_customers_Net_Profit , 4))

# b) Customer Lifetime Value Analysis:
young_CLV = young_customers_Net_Profit * (1 / 0.25)
Middle_aged_CLV = Middle_customers_Net_Profit * (1 / 0.15)
Senior_CLV = Senior_customers_Net_Profit * (1 / 0.30)

print(end = "\n\n")

print("Yuung Customers CLV : " , round(young_CLV , 4))
print("Middle-aged Customers CLV : " , round(Middle_aged_CLV , 4))
print("Senior Customers CLV : " , round(Senior_CLV , 4))
```

```
Young Customers Net Profit per Customer is :  -30.4
Middle-aged Customers Net Profit per Customer is :  40.8
Senior Customers Net Profit per Customer is :  -18.5


Yuung Customers CLV :  -121.6
Middle-aged Customers CLV :  272.0
Senior Customers CLV :  -61.6667
```

## Question 5 (10 points)

**Strategic Recommendations:** Based on your analysis, what would be your **top 2 marketing strategies** to maximize overall profitability? Consider customer acquisition, retention, and return rates.

```
# Your Answer for Question 5

# Strategy 1:
'''
Focus Acquisition & Retention on Middle-aged Customers

Increase marketing budget targeting the 31–50 age group, as they
generate the highest net profit and CLV.
Use:
- Personalized email campaigns
- Loyalty programs and rewards
- Cross-selling and upselling premium products
- Retention initiatives (since churn is only 15%) will further
increase lifetime value.

Impact: Maximizes ROI by acquiring and retaining the most profitable
segment.
'''

# Strategy 2:
'''
Reduce Return Rates and Acquisition Costs for Young & Senior Segments

-These segments are loss-making due to high returns and CAC.

Actions:
- Improve product descriptions, size guides, and customer reviews to
reduce returns
- Offer exchange instead of refund
- Use low-cost digital channels (organic social, referrals) to reduce
CAC
- Target only high-intent customers using behavioral segmentation

Impact: Converts loss-making segments into profitable ones by lowering
costs and return losses.
'''

'\nReduce Return Rates and Acquisition Costs for Young & Senior
Segments\n\n-These segments are loss-making due to high returns and
CAC.\n\nActions:\n- Improve product descriptions, size guides, and
customer reviews to reduce returns\n- Offer exchange instead of
refund\n- Use low-cost digital channels (organic social, referrals) to
reduce CAC\n- Target only high-intent customers using behavioral
segmentation\n\nImpact: Converts loss-making segments into profitable
ones by lowering costs and return losses.\n'
```

## Question 6 (10 points)

**Churn Prevention:** You notice that customers who make purchases in the "Electronics" category have a 35% churn rate, while "Fashion" category customers have only 18% churn rate. What

**specific data analysis** would you conduct using the available dataset columns to understand this difference, and what **action plan** would you recommend?

```
e_com_data["Product Category"].unique()
```

```
array(['Home', 'Electronics', 'Books', 'Clothing'], dtype=object)
```

```
# Your Answer for Question 6
'''
- perform Chi-Square Test of Independence
'''
# Data Analysis Plan:
'''
- define Hypothesis:

H0: Not statistically significant
H1: statistically significant

- Select only Electronics and Fashion customers.
ex : df_sub = e_com_data[e_com_data['Product
Category'].isin(['Electronics', 'Fashion'])]
- Create Contingency Table
ex : contingency_table = pd.crosstab(df_sub['Product_Category'],
df_sub['Churn'])
- Perform Chi-Square Test
ex : chi2, p_value, dof, expected =
chi2_contingency(contingency_table)

-If p-value < 0.05 → The churn difference between Electronics and
Fashion is statistically significant.

Additional Analysis (if significant)
- Compare average purchase amount by category
- Compare return rates
'''

# Action Plan:
'''
If Chi-square shows a significant difference:
1)Improve Retention for Electronics Customers
2)Post-purchase support (installation guides, service reminders)
3)Extended warranty and customer service follow-ups
4)Increase Repeat Purchases
5)Offer accessory bundles and upgrade discounts
6)Reduce Dissatisfaction & Returns
7)Re-engagement Campaigns
'''
```

```
'\nIf Chi-square shows a significant difference:\n1)Improve Retention
for Electronics Customers\n2)Post-purchase support (installation
guides, service reminders)\n3)Extended warranty and customer service
```

```
follow-ups\n4)Increase Repeat Purchases\n5)Offer accessory bundles and upgrade discounts\n6)Reduce Dissatisfaction & Returns\n7)Re-engagement Campaigns\n'
```

---

# SECTION C: Research Methodology & Predictive Analytics (25 Points - 8 Minutes)

## Scenario: Churn Prediction Model Development

Your company wants to build a machine learning model to predict customer churn using the available dataset.

## Question 7 (15 points)

**Model Development Plan:** Create a comprehensive approach including:

a) **Feature selection:** Which columns from the dataset would you use as features for the churn prediction model and why? (5 points)

b) **Data preprocessing steps:** What preprocessing would you apply to prepare the data? (5 points)

c) **Model evaluation metrics:** Which metrics would you use to evaluate model performance for this business problem? (5 points)

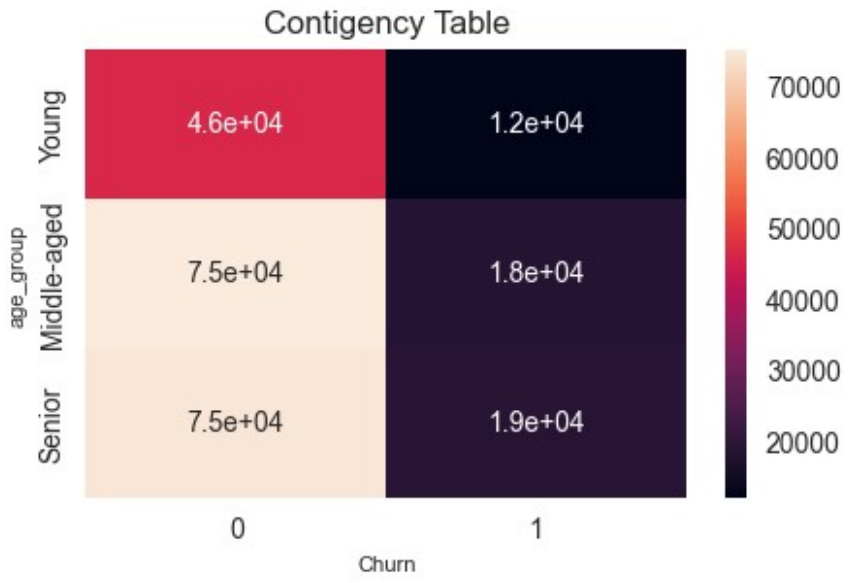
```
e_com_data[num_cols].corr()["Churn"].sort_values(ascending = False)
```

```
Churn                    1.0000
Product Price            0.0012
revenue                  0.0011
Total Purchase Amount    0.0007
Customer ID              0.0004
Quantity                -0.0014
Customer Age            -0.0023
Age                     -0.0023
Returns                 -0.0041
Name: Churn, dtype: float64
```

```
def chie2_test(group1 , group2):
    cross = pd.crosstab(group1 , group2)
    stat , p , df , evidence = chi2_contingency(cross)

    n = cross.to_numpy().sum()
    cramers_v = np.sqrt(stat / (n * (min(cross.shape) - 1)))
    result = {
        "statistic" : stat,
        "P value" : p,
```

```
        "df" : df,
        "cramers_v" : cramers_v
    }
    segnificance_level = 0.05

    if p < segnificance_level:
        print("Reject Null Hypotheses , there is significant
difference between two groups")
    else:
        print("Accept Null Hypotheses , there is no significant
difference bewtween two groups")
    plt.figure(figsize = (5 , 3))
    sns.heatmap(cross , annot = True , annot_kws = {"size" : 10})
    plt.title("Contigency Table")
    plt.show()

    return result

chie2_test(e_com_data["age_group"] , e_com_data["Churn"])
```

```
Reject Null Hypotheses , there is significant difference between two
groups
```




```
{'statistic': np.float64(25.72951925706938),
 'P value': np.float64(2.587652074012287e-06),
 'df': 2,
 'cramers_v': np.float64(0.01024341990782592)}
```

```
# Your Answer for Question 7

# a) Feature Selection:
```

```
'''
Product Price : it have correlation value +0.0012. it is not storng
but it is positive relation with Churn.
revenue : it is also same , it correlation value +0.0011 with Churn.
Total Purchase Amount : it is also have positive relation with churn
of +0.0007
age_group : it has difference in its distribution agaist churn and
that streanth low but still it is useful , it can separate the churn
and not churn customers.
'''

# b) Data Preprocessing Steps:
'''
In our final dataset have both numerical and categorical columns.

1)Encoding:

- Encode the Categorical Column (age_group) using encoding techniques

2) standardization :
- we need to standardization dataset it is important for
classification models using StandardScaler , etc.
- some models are require this like knn , decision tree , ect.
'''

# c) Model Evaluation Metrics:
'''
- we aldready saw in our analysis , In  E-commerce Customer Data ,
target variable "Churn" is Imbalenced , consider this. So Model
Evaluation Metrics for Churn classification are:
1) Accuracy (not recommended becouse dataset is imbalenced)
2) Precision & Recall & F1-score(very important)
3) ROC-AUC
4) Confusion Matrix
5) KS Statistic (Kolmogorov–Smirnov) - Measures maximum separation
between Cumulative distribution of churners,Non-churners
'''
```

```
'\n- we aldready saw in our analysis , In  E-commerce Customer Data ,
target variable "Churn" is Imbalenced , consider this. So Model
Evaluation Metrics for Churn classification are:\n1) Accuracy (not
recommended becouse dataset is imbalenced)\n2) Precision & Recall &
F1-score(very important)\n3) ROC-AUC \n4) Confusion Matrix\n5) KS
Statistic (Kolmogorov–Smirnov) - Measures maximum separation between
Cumulative distribution of churners,Non-churners\n'
```

## Question 8 (10 points)

**Business Impact Analysis:** Identify 3 potential business challenges in implementing a churn prediction model and propose one **data-driven solution** for each challenge using insights from the customer behavior dataset.

```
# Your Answer for Question 8

# Challenge 1:
'''
Incorrect Predictions (False Positives / False Negatives)

Problems
    Model may predict some customers will churn, but they actually
won't.
    Company may waste money on unnecessary offers.
    Or miss real churn customers.

# Data-driven Solution 1:

Analyze model performance metrics (precision, recall, ROC).
Use customer behavior features like:
    Low usage
    High complaints
    Payment delays
Focus retention offers only on high-risk probability customers (e.g.,
churn probability > 0.7).
'''
# Challenge 2:
'''
Data Quality Issues

Problems
    Customer data may be:
        Missing (null values)
        Incorrect
        Outdated
    Poor data leads to wrong churn predictions.

# Data-driven Solution 2:

Perform data quality analysis:
Handle missing values (imputation)
Remove duplicates
Update recent customer activity (last login, last purchase)
Create features like:
    Recency (days since last activity)
    Frequency (number of transactions)
    Monetary value
'''
```

```
# Challenge 3:
'''
Customer Behavior Changes Over Time (Model Drift)

Problem
Customer habits change.
Model trained on old data becomes less accurate.

# Data-driven Solution 3:

Monitor model performance regularly.
Retrain model monthly/quarterly using latest customer behavior data.
Track metrics like:
    Accuracy
    AUC
'''
```

```
'\nCustomer Behavior Changes Over Time (Model Drift)\n\nProblem\
nCustomer habits change.\nModel trained on old data becomes less
accurate.\n\n# Data-driven Solution 3:\n\nMonitor model performance
regularly.\nRetrain model monthly/quarterly using latest customer
behavior data.\nTrack metrics like:\n    Accuracy\n    AUC\n'
```

---

# SECTION D: Professional Communication & Problem-Solving (15 Points - 2 Minutes)

## Question 9 (8 points)

**Crisis Management:** While analyzing the dataset, you discover that 40% of customers who returned products (Returns = 1) also churned within the same month. However, your initial analysis showed returns don't strongly correlate with churn. As a team member, describe your immediate approach to investigate this discrepancy and communicate findings to stakeholders (60-80 words).

```
# Your Answer for Question 9 (60-80 words)
'''
- I can first validate the data by checking return and churn
definitions, time alignment, and potential data quality issues.
- Next, I would perform segmented analysis (by customer type, product
category, and return reasons) to identify hidden patterns that may not
appear in overall correlation.
- I would also test statistical significance and interaction effects.
Then, I would communicate to stakeholders that while overall
correlation is weak,
- specific customer segments show higher risk, recommending targeted
```

```
investigation and retention actions.
'''

'\n- I can first validate the data by checking return and churn definitions, time alignment, and potential data quality issues. \n- Next, I would perform segmented analysis (by customer type, product category, and return reasons) to identify hidden patterns that may not appear in overall correlation.\n- I would also test statistical significance and interaction effects. Then, I would communicate to stakeholders that while overall correlation is weak, \n- specific customer segments show higher risk, recommending targeted investigation and retention actions.\n'
```

## Question 10 (7 points)

**Leadership Scenario:** If selected as team lead for analyzing this e-commerce customer dataset, what would be your **top 3 priorities** to ensure effective team collaboration and delivery of actionable business insights?

```
# Your Answer for Question 10

# Priority 1:
'''
Clear Business Understanding & Goal Alignment

Priority
Ensure the team clearly understands the business objectives before starting analysis.

Why important
Without business clarity, the team may produce technical results that are not useful.

Actions

Discuss with stakeholders:
    What is the goal? (e.g., increase revenue, reduce churn, improve retention)
    Define key business questions:
    Which customers generate most revenue?
    Who is likely to churn?
    Which products sell the most?

Define success metrics:
    Increase retention by 10%
    Identify high-value customers
Outcome
Team works on business-driven analysis, not just data exploration.
'''
# Priority 2:
```

```
'''
Structured Team Collaboration & Task Management

Priority
Organize the team workflow for smooth collaboration and timely
delivery.

Actions
    Divide work into modules:
    Data Cleaning
    Exploratory Data Analysis (EDA)
    Feature Engineering
    Modeling
    Visualization/Reporting

Use tools:
    Git/GitHub for version control

Conduct:
    Regular stand-up meetings
    Progress reviews
    Code reviews

Outcome
Better coordination, fewer errors, faster project completion.
'''

# Priority 3:
'''
Focus on Actionable Insights & Business Communication

Priority
Ensure the final output provides clear, actionable recommendations,
not just technical results.

Actions
    Convert analysis into business insights:
    Identify high-value customers
    Detect churn-risk customers
    Find top-selling product categories

Create dashboards/reports using:
    Power BI / Tableau / Streamlit

Present insights in simple business language:
    “Customers inactive for 30+ days have 3x higher churn risk”
    “Top 20% customers contribute 60% revenue”

Outcome
```

```
Stakeholders can take decisions based on insights.
'''

'\nFocus on Actionable Insights & Business Communication\n\nPriority\
nEnsure the final output provides clear, actionable recommendations,
not just technical results.\n\nActions\n    Convert analysis into
business insights:\n    Identify high-value customers\n    Detect
churn-risk customers\n    Find top-selling product categories\n\
nCreate dashboards/reports using:\n    Power BI / Tableau / Streamlit\
n\nPresent insights in simple business language:\n    “Customers
inactive for 30+ days have 3x higher churn risk”\n    “Top 20%
customers contribute 60% revenue”\n\nOutcome\nStakeholders can take
decisions based on insights.\n'
```

---

## Self-Assessment Section

```
# Time Management Check
# Did you complete all sections within 35 minutes? (Yes/No):Yes

# Which section took the most time?
# SECTION B: Customer Analysis & Business Intelligence

# Which section was most challenging?
# Section C (Research Methodology)

# Confidence Level (1-10 scale):
# Section A (Data Understanding):9
# Section B (Business Analysis):8
# Section C (Research Methodology):9
# Section D (Communication):8

# Additional Comments:The assessment helped improve my ability to
connect data insights with business decisions. I was able to manage
time effectively, though business interpretation and methodological
thinking required more attention. Overall, I am confident in my
analytical and communication approach.
```

---

## Submission Instructions

1. **Save this notebook** with the filename: `YourName_Evoastra_Assessment.ipynb`
2. **Ensure all code cells have been executed** and answers are visible
3. **Double-check** that all sections are completed

**Submission Confirmation:**

- I confirm that I have completed this assessment independently

- All my responses are my own original work

**Digital Signature:** _____Tirumani Kishore_____
**Final Submission Time:** _____18 Feb 2026, 11: 30 AM_____

---

# Evaluation Criteria

**Scoring Breakdown:**

- **Section A (Data Understanding):** 25 points
- **Section B (Business Analysis):** 35 points

- **Section C (Research Methodology):** 25 points
- **Section D (Communication):** 15 points
- **Total:** 100 points

**Team Selection Criteria:**

- **Team Lead:** Score ≥ 75 points + Strong Section D performance
- **Co-Lead:** Score ≥ 65 points + Good Section D performance
- **Team Member:** Successful completion of assessment

---

*Good luck with your assessment! Focus on clear reasoning, accurate calculations, and practical business applications.*