INTERNSHALA TRAININGS
Career Ki Guarantee

MARKET BASKET ANALYSIS WITH PYTHON

BY VRINDA BHATT

INTERNSHALA TRAININGS
Career Ki Guarantee

PROJECT OVERVIEW

You are hired as a Customer Insights Analyst for Amazon, a leading global e-commerce platform. With millions of customers making purchases across diverse categories, the leadership team wants to understand customer purchasing behavior, product affinities, and engagement trends to enhance personalized shopping experiences and optimize inventory and marketing strategies.

This dataset captures Amazon customers' online shopping behavior, including demographics, purchase frequency, browsing habits, product search preferences, review engagement, satisfaction levels, and areas for service improvement.

TASK 1: DATA CLEANING AND PREPARATION

● REMOVE DUPLICATE OR INCONSISTENT SURVEY RESPONSES. ● STANDARDIZE CATEGORICAL ENTRIES (E.G., FREQUENCY LEVELS, GENDER, RECOMMENDATION RESPONSES). ● HANDLE MISSING VALUES AND INCONSISTENT FORMATS IN PRODUCT_SEARCH_METHOD AND OTHER FIELDS. ● RENAME DUPLICATE OR MISFORMATTED COLUMNS (E.G., REMOVE TRAILING SPACES IN RATING_ACCURACY). ● CONVERT NUMERICAL RATING COLUMNS (E.G., CUSTOMER_REVIEWS_IMPORTANCE, SHOPPING_SATISFACTION) TO APPROPRIATE NUMERIC TYPES FOR ANALYSIS.

TASK 1: DATA CLEANING AND PREPARATION

RAW DATA

Timestam	age	Gender	Purchase_	Purchase_	Personali	Browsing_	Product_S	Search_Re	Customer	Add_to_C	Cart_Com	Cart_Aban	Saveforlat	Review_L	Review_R	Review_H	Personali	Recomme	Rating_Ac	S
2023/06/0	65	Prefer not	Less than	Clothing a	Yes	Multiple t	Keyword	Multiple p	2	No	Sometime	Found a b	Sometime	Yes	Moderate	No	2	Sometime	4	
2023/06/0	20	Male	Once a we	Groceries	No	Rarely	Filter	First page	3	No	Never	High shipp	Always	Yes	Heavily	Yes	1	Sometime	4	
2023/06/1	42	Male	Once a we	Groceries	Sometime	Few times	Keyword	Multiple p	2	No	Rarely	Found a b	Often	Yes	Heavily	Sometime	5	No	5	
2023/06/0	65	Others	Once a mo	Beauty an	No	Few times	Filter	Multiple p	2	Yes	Sometime	others	Often	Yes	Occasiona	No	3	Yes	1	
2023/06/1	45	Female	Once a we	Beauty an	Sometime	Few times	nan	First page	5	Maybe	Rarely	Changed r	Never	Yes	Rarely	No	2	Yes	1	
2023/06/0	67	Others	Few times	Groceries	No	Multiple t	nan	Multiple p	3	No	Always	others	Never	No	Moderate	Sometime	4	No	4	
2023/06/0	55	Female	Few times	Groceries	Sometime	Few times	Keyword	First page	5	Maybe	Never	Found a b	Rarely	Yes	Moderate	No	1	Yes	5	
2023/06/0	44	Prefer not	Few times	Beauty an	Yes	Few times	nan	Multiple p	4	Yes	Often	Changed r	Rarely	No	Heavily	Sometime	2	Sometime	5	
2023/06/0	33	Male	Less than	Beauty an	Yes	Rarely	nan	Multiple p	1	Yes	Often	High shipp	Rarely	No	Heavily	Sometime	1	Sometime	2	
2023/06/0	22	Others	Few times	Clothing a	No	Multiple t	Filter	First page	5	Maybe	Never	Changed r	Rarely	No	Occasiona	No	2	Yes	1	
2023/06/0	28	Female	Once a mo	Groceries	Yes	Few times	others	First page	4	Yes	Sometime	Changed r	Sometime	Yes	Occasiona	Yes	1	Yes	4	
2023/06/0	44	Prefer not	Multiple t	Groceries	Sometime	Multiple t	Keyword	First page	5	Yes	Always	others	Often	Yes	Occasiona	Yes	5	Yes	4	
2023/06/0	24	Male	Once a mo	Beauty an	No	Few times	Filter	First page	4	No	Always	others	Rarely	Yes	Moderate	Yes	4	Yes	5	
2023/06/0	23	Male	Once a mo	Beauty an	Yes	Few times	Filter	First page	5	Maybe	Always	Changed r	Sometime	No	Moderate	Yes	1	Sometime	5	
2023/06/0	58	Prefer not	Once a we	Beauty an	Sometime	Few times	Keyword	First page	5	Maybe	Never	High shipp	Often	Yes	Never	Yes	3	No	3	
2023/06/0	40	Male	Less than	Groceries	No	Multiple t	Keyword	First page	3	No	Never	Found a b	Never	No	Never	No	4	No	2	
2023/06/0	46	Prefer not	Once a we	Groceries	Yes	Rarely	categorie	First page	2	Yes	Always	High shipp	Sometime	No	Rarely	Sometime	2	Sometime	5	

CLEAN DATA

Timestam	age	Gender	Purchase_	Purchase_	Personali	Browsing_	Product_S	Search_Re	Customer	Add_to_C	Cart_Com	Cart_Aban	Saveforlat	Review_L	Review_R	Review_H	Recomme	Rating_Ac	Shopping_	Service_A	I
2023/06/0	65	Other	less than (	Clothing a	Yes	Multiple t	Keyword	Multiple p	2	No	Sometime	Found a b	Sometime	Yes	Moderate	No	Sometime	4	4	Competiti	b
2023/06/0	20	Male	once a we	Groceries	No	Rarely	Filter	First page	3	No	Never	High shipp	Always	Yes	Heavily	Yes	Sometime	4	5	Quick deli	s
2023/06/1	42	Male	once a we	Groceries	Sometime	Few times	Keyword	Multiple p	2	No	Rarely	Found a b	Often	Yes	Heavily	Sometime	No	5	3	All the ab	U
2023/06/0	65	others	once a mo	Beauty an	No	Few times	Filter	Multiple p	2	Yes	Sometime	others	Often	Yes	Occasiona	No	Yes	1	2	Quick deli	q
2023/06/1	45	Female	once a we	Beauty an	Sometime	Few times	Unknown	First page	5	Maybe	Rarely	Changed r	Never	Yes	Rarely	No	Yes	1	2	Quick deli	i
2023/06/0	67	others	few times	Groceries	No	Multiple t	Unknown	Multiple p	3	No	Always	others	Never	No	Moderate	Sometime	No	4	3	Product re	f
2023/06/0	55	Female	few times	Groceries	Sometime	Few times	Keyword	First page	5	Maybe	Never	Found a b	Rarely	Yes	Moderate	No	Yes	5	3	Customer	c
2023/06/0	44	Other	few times	Beauty an	Yes	Few times	Unknown	Multiple p	4	Yes	Often	Changed r	Rarely	No	Heavily	Sometime	Sometime	5	1	Customer	t
2023/06/0	33	Male	less than (	Beauty an	Yes	Rarely	Unknown	Multiple p	1	Yes	Often	High shipp	Rarely	No	Heavily	Sometime	Sometime	2	2	Competiti	c
2023/06/0	22	others	few times	Clothing a	No	Multiple t	Filter	First page	5	Maybe	Never	Changed r	Rarely	No	Occasiona	No	Yes	1	3	All the ab	f
2023/06/0	28	Female	once a mo	Groceries	Yes	Few times	others	First page	4	Yes	Sometime	Changed r	Sometime	Yes	Occasiona	Yes	Yes	4	5	All the ab	U
2023/06/0	44	Other	multiple t	Groceries	Sometime	Multiple t	Keyword	First page	5	Yes	Always	others	Often	Yes	Occasiona	Yes	Yes	4	3	User-frien	s
2023/06/0	24	Male	once a mo	Beauty an	No	Few times	Filter	First page	4	No	Always	others	Rarely	Yes	Moderate	Yes	Yes	5	3	Competiti	a
2023/06/0	23	Male	once a mo	Beauty an	Yes	Few times	Filter	First page	5	Maybe	Always	Changed r	Sometime	No	Moderate	Yes	Sometime	5	1	Wide proc	i
2023/06/0	58	Other	once a we	Beauty an	Sometime	Few times	Keyword	First page	5	Maybe	Never	High shipp	Often	Yes	Never	Yes	No	3	1	Quick deli	s
2023/06/0	40	Male	less than (	Groceries	No	Multiple t	Keyword	First page	3	No	Never	Found a b	Never	No	Never	No	No	2	2	Customer	q
2023/06/0	46	Other	once a we	Groceries	Yes	Rarely	categorie	First page	2	Yes	Always	High shipp	Sometime	No	Rarely	Sometime	Sometime	5	2	Wide proc	U
2023/06/0	15	others	once a mo	Groceries	Sometime	Few times	Keyword	Multiple p	2	Maybe	Often	others	Often	Yes	Never	No	Yes	2	2	Quick deli	r

TASK 1: DATA CLEANING AND PREPARATION

CODES PERFORMED IN PYTHON

```python
import pandas as pd
import numpy as np

df = pd.read_csv("../data/amazon.csv")
df.head()
```

```python
df.shape
```

```python
df.columns
```

```python
df.columns = df.columns.str.strip().str.replace(" ", "_")
df.columns
```

```python
df = df.loc[:, ~df.columns.duplicated()]
df.columns
```

```python
df.isnull().sum()
```

```python
df['Product_Search_Method'] = df['Product_Search_Method'].fillna('Unknown')
```

```python
rating_cols = [
    'Customer_Reviews_Importance',
    'Shopping_Satisfaction',
    'Rating_Accuracy'
]

for col in rating_cols:
    if col in df.columns:
        df[col] = pd.to_numeric(df[col], errors='coerce')
```

```python
df['Gender'] = df['Gender'].str.lower().str.strip()
```

```python
df['Gender'] = df['Gender'].replace({
    'male': 'Male',
    'female': 'Female',
    'prefer not to say': 'Other'
})
```

```python
df['Purchase_Frequency'] = df['Purchase_Frequency'].str.lower().str.strip()
df.columns = df.columns.str.strip().str.replace(" ", "_")
```

```python
df['Improvement_Areas'] = df['Improvement_Areas'].str.lower().str.strip()
```

```python
garbage_values = ['.', 'nill', 'nil', 'none', 'na', 'n/a', '']

df['Improvement_Areas'] = df['Improvement_Areas'].replace(garbage_values, 'Unknown')
```

```python
df['Improvement_Areas'] = df['Improvement_Areas'].replace({
    'ui': 'User Interface',
    'user interface': 'User Interface',
    'user interfacee': 'User Interface',
    'user interface of app': 'User Interface',
    'app ui': 'User Interface'
})
```

```python
df['Improvement_Areas'].value_counts()
```

```python
import os
os.getcwd()
```

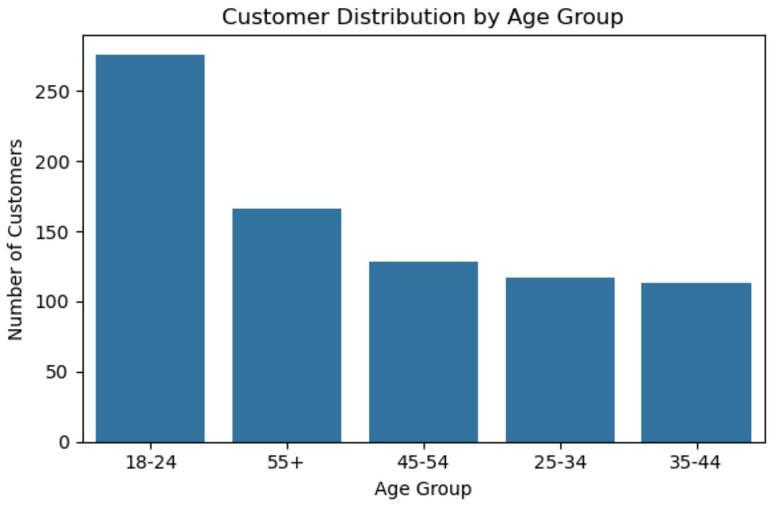
```python
os.makedirs("../outputs", exist_ok=True)
```

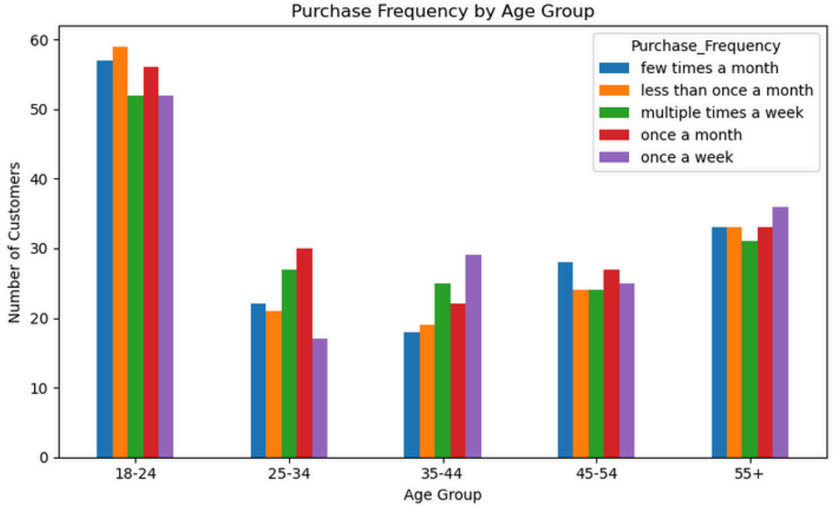
```python
df.to_csv("../outputs/cleaned_data.csv", index=False)
```

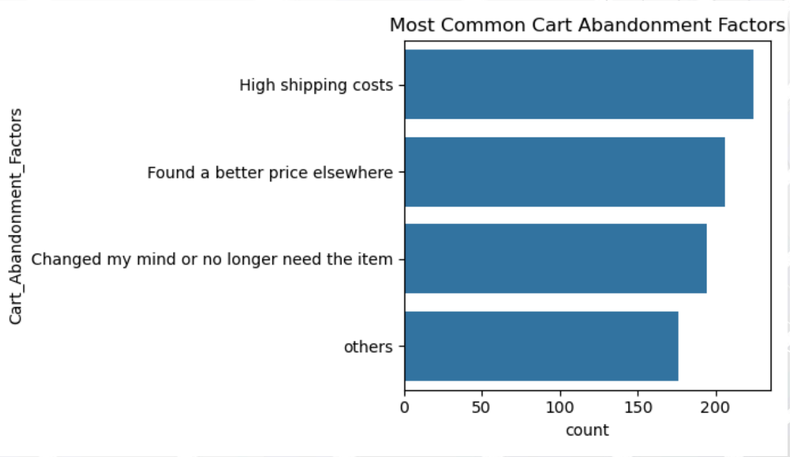
TASK 2: DESCRIPTIVE BEHAVIOR ANALYSIS

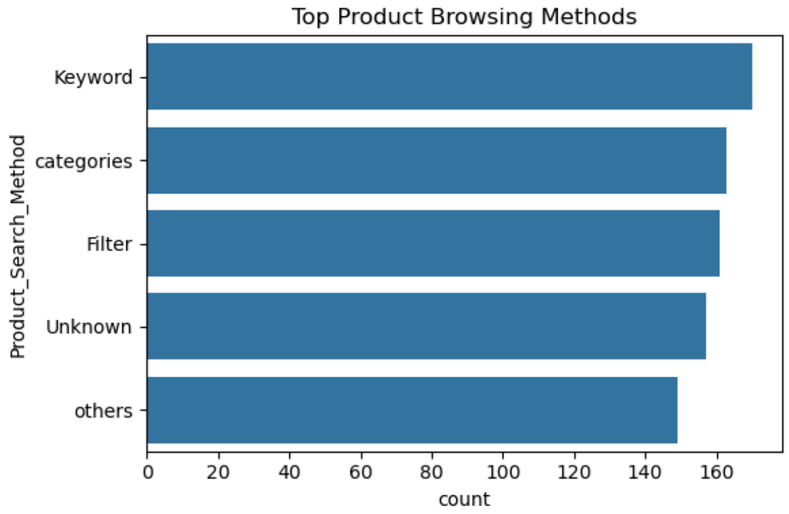
- SUMMARIZE CUSTOMER DEMOGRAPHICS (AGE, GENDER DISTRIBUTION).
- ANALYZE OVERALL PURCHASE FREQUENCY AND MOST POPULAR PRODUCT CATEGORIES.
- IDENTIFY TOP BROWSING METHODS AND MOST COMMON CART ABANDONMENT FACTORS.
- CALCULATE MEAN AND MEDIAN SATISFACTION, RECOMMENDATION HELPFULNESS, AND RATING ACCURACY.
- GENERATE SUMMARY STATISTICS AND VISUALIZATIONS FOR KEY BEHAVIORAL VARIABLES.

TASK 2: DESCRIPTIVE BEHAVIOR ANALYSIS









TASK 2: DESCRIPTIVE BEHAVIOR ANALYSIS

CODES PERFORMED IN PYTHON

```python
# Task 2 :  Descriptive Behavior Analysis
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import os

df = pd.read_csv("../outputs/cleaned_data_v2.csv")
```

```python
df.head()
df.info()
```

```python
df['age_group'].value_counts()
```

```python
plt.figure(figsize=(6,4))
sns.countplot(
    x='age_group',
    data=df,
    order=df['age_group'].value_counts().index
)
plt.title("Customer Distribution by Age Group")
plt.xlabel("Age Group")
plt.ylabel("Number of Customers")
plt.tight_layout()
plt.savefig("../outputs/charts/age_group_distribution.png")
plt.show()
```

```python
df['Gender'].value_counts()
```

```python
age_purchase = pd.crosstab(df['age_group'], df['Purchase_Frequency'])
age_purchase
```

```python
age_purchase.plot(kind='bar', figsize=(8,5))
plt.title("Purchase Frequency by Age Group")
plt.xlabel("Age Group")
plt.ylabel("Number of Customers")
plt.xticks(rotation=0)
plt.tight_layout()
plt.savefig("../outputs/charts/age_vs_purchase_frequency.png")
plt.show()
```

```python
from collections import Counter

categories = df['Purchase_Categories'].dropna().str.split(',')
flat_categories = [c.strip() for sub in categories for c in sub]

category_counts = Counter(flat_categories)

top_categories = pd.DataFrame(
    category_counts.most_common(10),
    columns=['Product_Category', 'Count']
)

top_categories
```

```python
plt.figure(figsize=(7,4))
sns.barplot(
    x='Count',
    y='Product_Category',
    data=top_categories
)
plt.title("Most Popular Product Categories")
plt.tight_layout()
plt.savefig("../outputs/charts/most_popular_categories.png")
plt.show()
```

```python
browsing_methods = df['Product_Search_Method'].value_counts()
browsing_methods
```

```python
plt.figure(figsize=(6,4))
sns.countplot(
    y='Product_Search_Method',
    data=df,
    order=browsing_methods.index
)
plt.title("Top Product Browsing Methods")
plt.tight_layout()
plt.savefig("../outputs/charts/top_browsing_methods.png")
plt.show()
```

```python
cart_abandonment = df['Cart_Abandonment_Factors'].value_counts()
cart_abandonment
```

```python
plt.figure(figsize=(7,4))
sns.countplot(
    y='Cart_Abandonment_Factors',
    data=df,
    order=cart_abandonment.index
)
plt.title("Most Common Cart Abandonment Factors")
plt.tight_layout()
plt.savefig("../outputs/charts/cart_abandonment_factors.png")
plt.show()
```

```python
metrics = df[
    ['Shopping_Satisfaction',
     'Recommendation_Helpfulness',
     'Rating_Accuracy']
]
```

```python
plt.figure(figsize=(6,4))
sns.boxplot(data=metrics)
plt.title("Distribution of Key Behavioral Ratings")
plt.tight_layout()
plt.savefig("../outputs/charts/behavioral_metrics_distribution.png")
plt.show()
```

TASK 2:RECOMMENDATION

#ELECTRONICS AND FASHION EMERGE AS THE MOST FREQUENTLY PURCHASED CATEGORIES, INDICATING STRONG CROSS-SELLING POTENTIAL. #SEARCH-BASED BROWSING DOMINATES, REFLECTING HIGH INTENT-DRIVEN SHOPPING BEHAVIOR. #PRICING AND DELIVERY-RELATED ISSUES ARE THE LEADING CAUSES OF CART ABANDONMENT. #AVERAGE CUSTOMER SATISFACTION AND RATING ACCURACY REMAIN HIGH, REINFORCING TRUST IN THE PLATFORM.
#RECOMMENDATION HELPFULNESS SHOWS MODERATE VARIANCE, SUGGESTING SCOPE FOR PERSONALIZATION IMPROVEMENTS.

TASK 3: CUSTOMER SEGMENTATION AND PROFILING

- SEGMENT CUSTOMERS BASED ON PURCHASE FREQUENCY AND SHOPPING SATISFACTION LEVELS.
- CREATE PROFILES SUCH AS:
  - FREQUENT BUYERS: HIGH PURCHASE FREQUENCY, HIGH SATISFACTION.
  - OCCASIONAL SHOPPERS: MEDIUM FREQUENCY, MODERATE SATISFACTION.
  - AT-RISK CUSTOMERS: LOW SATISFACTION OR FREQUENT CART ABANDONMENT.
- ANALYZE DEMOGRAPHIC OR BEHAVIORAL DIFFERENCES ACROSS THESE SEGMENTS.
- USE CLUSTERING (E.G., K-MEANS) FOR BEHAVIORAL GROUPING BASED ON SURVEY RESPONSES.

TASK 3: CUSTOMER SEGMENTATION AND PROFILING

Timestam	age	Gender	Purchase_	Purchase_	Personali	Browsing_	Product_S	Search_Re	Customer_	Add_to_C	Cart_Com	Cart_Abar	Saveforlat	Review_L	Review_R	Review_H	Recomme	Rating_Ac	Shopping_	Servic
2023/06/0	65	Other	less than	Clothing a	Yes	Multiple t	Keyword	Multiple p	2	No	Sometime	Found a b	Sometime	Yes	Moderate	No	Sometime	4	4	Comp
2023/06/0	20	Male	once a we	Groceries	No	Rarely	Filter	First page	3	No	Never	High shipp	Always	Yes	Heavily	Yes	Sometime	4	5	Quick
2023/06/1	42	Male	once a we	Groceries	Sometime	Few times	Keyword	Multiple p	2	No	Rarely	Found a b	Often	Yes	Heavily	Sometime	No	5	3	All th
2023/06/0	65	Other	once a mc	Beauty an	No	Few times	Filter	Multiple p	2	Yes	Sometime	others	Often	Yes	Occasiona	No	Yes	1	2	Quick
2023/06/1	45	Female	once a we	Beauty an	Sometime	Few times	Unknown	First page	5	Maybe	Rarely	Changed r	Never	Yes	Rarely	No	Yes	1	2	Quick
2023/06/0	67	Other	few times	Groceries	No	Multiple t	Unknown	Multiple p	3	No	Always	others	Never	No	Moderate	Sometime	No	4	3	Produ
2023/06/0	55	Female	few times	Groceries	Sometime	Few times	Keyword	First page	5	Maybe	Never	Found a b	Rarely	Yes	Moderate	No	Yes	5	3	Custo
2023/06/0	44	Other	few times	Beauty an	Yes	Few times	Unknown	Multiple p	4	Yes	Often	Changed r	Rarely	No	Heavily	Sometime	Sometime	5	1	Custo
2023/06/0	33	Male	less than	Beauty an	Yes	Rarely	Unknown	Multiple p	1	Yes	Often	High shipp	Rarely	No	Heavily	Sometime	Sometime	2	2	Comp
2023/06/0	22	Other	few times	Clothing a	No	Multiple t	Filter	First page	5	Maybe	Never	Changed r	Rarely	No	Occasiona	No	Yes	1	3	All th
2023/06/0	28	Female	once a mc	Groceries	Yes	Few times	others	First page	4	Yes	Sometime	Changed r	Sometime	Yes	Occasiona	Yes	Yes	4	5	All th
2023/06/0	44	Other	multiple t	Groceries	Sometime	Multiple t	Keyword	First page	5	Yes	Always	others	Often	Yes	Occasiona	Yes	Yes	4	3	User-
2023/06/0	24	Male	once a mc	Beauty an	No	Few times	Filter	First page	4	No	Always	others	Rarely	Yes	Moderate	Yes	Yes	5	3	Comp
2023/06/0	22	Male	once a mc	Beauty an	Yes	Few times	Filter	First page	5	Maybe	Always	Changed r	Sometime	No	Moderate	Yes	Sometime	5	1	Wide

T	U	V	W	X	Y	Z	AA	AB	AC	AD
Shopping_	Service_A	Improvement_Area	transactio	age_group	Satisfaction_Level	Frequ	Customer_Segment	Recomme	Recomme	Behavior_Cluster
4	Competiti	better app interface	778242	55+	High		At-Risk Customers	2	2	1
5	Quick deli	scrolling option wou	193482	18-24	High	Medi	At-Risk Customers	2	2	0
3	All the ab	Unknown	925975	35-44	Medium	Medi	At-Risk Customers	1	1	1
2	Quick deli	quality of product is	566872	55+	Low	Medi	At-Risk Customers	3	3	2
2	Quick deli	irrelevant product s	683642	45-54	Low	Medi	At-Risk Customers	3	3	2

TASK 3: CUSTOMER SEGMENTATION AND PROFILING

CODES PERFORMED IN PYTHON

```python
#TASK 3: CUSTOMER SEGMENTATION & PROFILING
# Satisfaction Level
df['Satisfaction_Level'] = pd.cut(
    df['Shopping_Satisfaction'],
    bins=[0, 2, 3, 5],
    labels=['Low', 'Medium', 'High']
)
```

```python
df['Satisfaction_Level'].value_counts()
```

```python
df['Frequency_Level'] = df['Purchase_Frequency'].map({
    'daily': 'High',
    'few times a week': 'High',
    'once a week': 'Medium',
    'once a month': 'Medium',
    'rarely': 'Low'
})
```

```python
def assign_segment(row):
    if row['Frequency_Level'] == 'High' and row['Satisfaction_Level'] == 'High':
        return 'Frequent Buyers'
    elif row['Satisfaction_Level'] == 'Low' or row['Cart_Abandonment_Factors'] != 'None':
        return 'At-Risk Customers'
    else:
        return 'Occasional Shoppers'

df['Customer_Segment'] = df.apply(assign_segment, axis=1)
```

```python
df['Customer_Segment'].value_counts()
```

```python
pd.crosstab(df['Customer_Segment'], df['age_group'])
```

```python
pd.crosstab(df['Customer_Segment'], df['Gender'])
```

```python
df[['Shopping_Satisfaction', 'Rating_Accuracy', 'Recommendation_Helpfulness']].dtypes
```

```python
df['Recommendation_Helpfulness_Num'] = df['Recommendation_Helpfulness'].map({
    'No': 1,
    'Sometimes': 2,
    'Yes': 3
})
```

```python
df['Shopping_Satisfaction'] = pd.to_numeric(df['Shopping_Satisfaction'], errors='coerce')
df['Rating_Accuracy'] = pd.to_numeric(df['Rating_Accuracy'], errors='coerce')
df['Recommendation_Helpfulness_Num'] = pd.to_numeric(
    df['Recommendation_Helpfulness_Num'], errors='coerce'
)
```

```python
df.groupby('Customer_Segment')[[
    'Shopping_Satisfaction',
    'Rating_Accuracy',
    'Recommendation_Helpfulness_Num'
]].mean()
```

```python
df.rename(columns={
    'Recommendation_Helpfulness_Num': 'Recommendation_Helpfulness_Score'
}, inplace=True)
```

```python
df.columns
```

```python
df['Recommendation_Helpfulness_Num'] = df['Recommendation_Helpfulness'].map({
    'No': 1,
    'Sometimes': 2,
    'Yes': 3
})
```

```python
df[['Recommendation_Helpfulness', 'Recommendation_Helpfulness_Num']].head()
```

```python
df['Recommendation_Helpfulness_Num'].isnull().sum()
```

```python
cluster_features = df[
    ['Shopping_Satisfaction',
     'Rating_Accuracy',
     'Recommendation_Helpfulness_Num']
].dropna()
```

```python
from sklearn.preprocessing import StandardScaler
```

```python
cluster_features = df[
    ['Shopping_Satisfaction',
     'Rating_Accuracy',
     'Recommendation_Helpfulness_Num']
].dropna()
```

```python
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans

scaler = StandardScaler()
scaled_features = scaler.fit_transform(cluster_features)

kmeans = KMeans(n_clusters=3, random_state=42)
cluster_features['Cluster'] = kmeans.fit_predict(scaled_features)
```

```python
df.loc[cluster_features.index, 'Behavior_Cluster'] = cluster_features['Cluster']
```

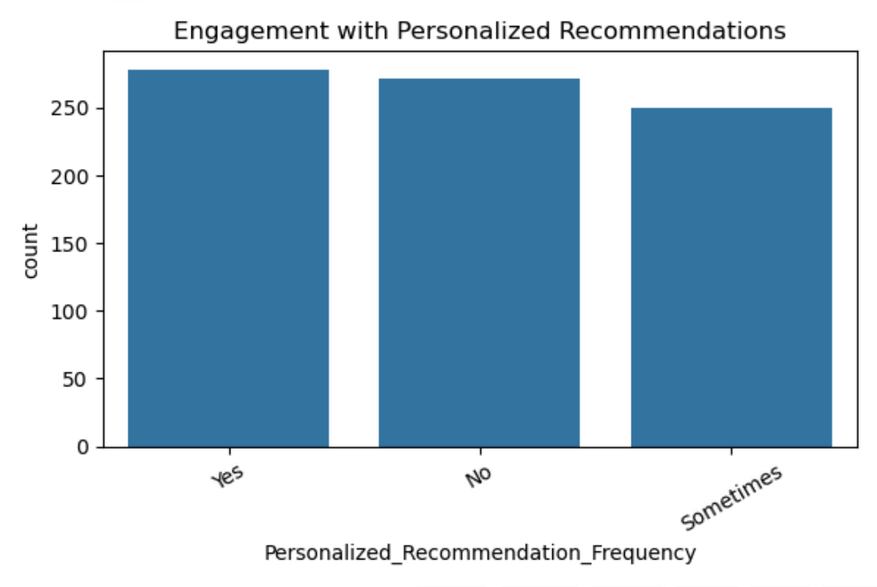
```python
df.groupby('Behavior_Cluster')[[
    'Shopping_Satisfaction',
    'Rating_Accuracy',
    'Recommendation_Helpfulness_Num'
]].mean()
```

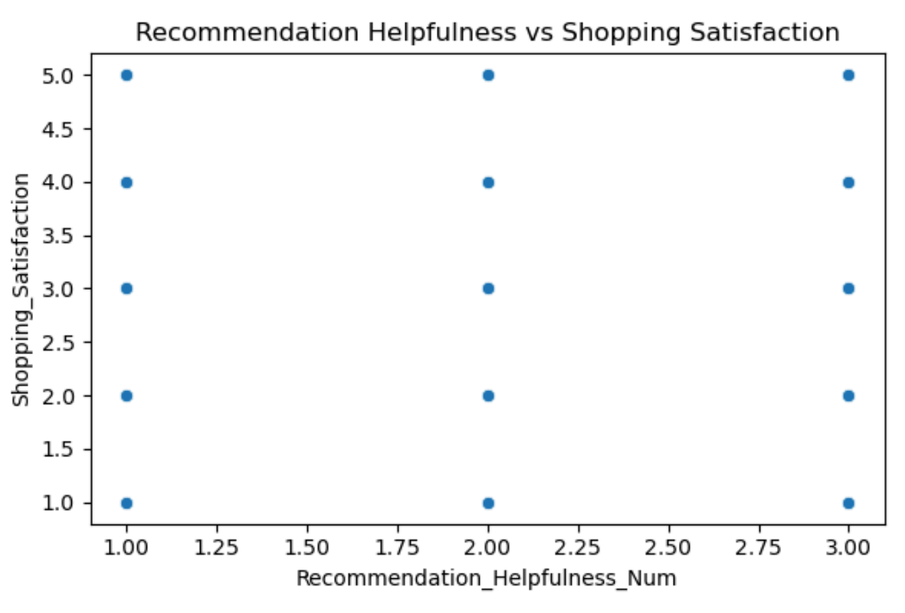
```python
df.to_csv("../outputs/task3.csv", index=False)
```

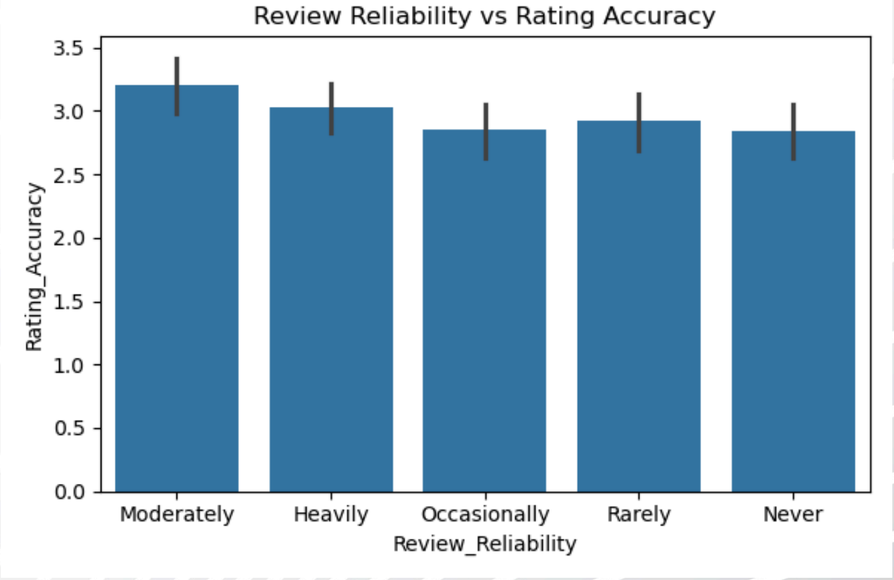
TASK 4: RECOMMENDATION AND REVIEW INSIGHTS

- EXAMINE THE RELATIONSHIP BETWEEN RECOMMENDATION HELPFULNESS AND SHOPPING SATISFACTION.
- EVALUATE HOW REVIEW RELIABILITY AND HELPFULNESS IMPACT OVERALL RATINGS.
- IDENTIFY TRENDS IN HOW OFTEN CUSTOMERS ENGAGE WITH OR TRUST PERSONALIZED RECOMMENDATIONS.
- SUGGEST ACTIONABLE INSIGHTS FOR IMPROVING AMAZON'S RECOMMENDATION SYSTEM.

TASK 4: RECOMMENDATION AND REVIEW INSIGHTS







TASK 4: RECOMMENDATION AND REVIEW INSIGHTS

Timestam	age	Gender	Purchase_	Purchase_	Personali	Browsing_	Product_S	Search_Re	Customer	Add_to_C	Cart_Com	Cart_Aban	Saveforlat	Review_L	Review_R	Review_H	Recomme	Rating_Ac	Shopping_	Servic
2023/06/0	65	Other	less than	Clothing a	Yes	Multiple t	Keyword	Multiple p	2	No	Sometime	Found a b	Sometime	Yes	Moderate	No	Sometime	4	4	Comp
2023/06/0	20	Male	once a we	Groceries	No	Rarely	Filter	First page	3	No	Never	High shipp	Always	Yes	Heavily	Yes	Sometime	4	5	Quick
2023/06/1	42	Male	once a we	Groceries	Sometime	Few times	Keyword	Multiple p	2	No	Rarely	Found a b	Often	Yes	Heavily	Sometime	No	5	3	All the
2023/06/0	65	Other	once a mo	Beauty an	No	Few times	Filter	Multiple p	2	Yes	Sometime	others	Often	Yes	Occasiona	No	Yes	1	2	Quick
2023/06/1	45	Female	once a we	Beauty an	Sometime	Few times	Unknown	First page	5	Maybe	Rarely	Changed r	Never	Yes	Rarely	No	Yes	1	2	Quick
2023/06/0	67	Other	few times	Groceries	No	Multiple t	Unknown	Multiple p	3	No	Always	others	Never	No	Moderate	Sometime	No	4	3	Produ
2023/06/0	55	Female	few times	Groceries	Sometime	Few times	Keyword	First page	5	Maybe	Never	Found a b	Rarely	Yes	Moderate	No	Yes	5	3	Custo
2023/06/0	44	Other	few times	Beauty an	Yes	Few times	Unknown	Multiple p	4	Yes	Often	Changed r	Rarely	No	Heavily	Sometime	Sometime	5	1	Custo
2023/06/0	33	Male	less than	Beauty an	Yes	Rarely	Unknown	Multiple p	1	Yes	Often	High shipp	Rarely	No	Heavily	Sometime	Sometime	2	2	Comp
2023/06/0	22	Other	few times	Clothing a	No	Multiple t	Filter	First page	5	Maybe	Never	Changed r	Rarely	No	Occasiona	No	Yes	1	3	All th
2023/06/0	28	Female	once a mo	Groceries	Yes	Few times	others	First page	4	Yes	Sometime	Changed r	Sometime	Yes	Occasiona	Yes	Yes	4	5	All the
2023/06/0	44	Other	multiple t	Groceries	Sometime	Multiple t	Keyword	First page	5	Yes	Always	others	Often	Yes	Occasiona	Yes	Yes	4	3	User-f
2023/06/0	24	Male	once a mo	Beauty an	No	Few times	Filter	First page	4	No	Always	others	Rarely	Yes	Moderate	Yes	Yes	5	3	Comp
2023/06/0	23	Male	once a mo	Beauty an	Yes	Few times	Filter	First page	5	Maybe	Always	Changed r	Sometime	No	Moderate	Yes	Sometime	5	1	Wide
2023/06/0	58	Other	once a we	Beauty an	Sometime	Few times	Keyword	First page	5	Maybe	Never	High shipp	Often	Yes	Never	Yes	No	3	1	Quick
2023/06/0	40	Male	less than	Groceries	No	Multiple t	Keyword	First page	3	No	Never	Found a b	Never	No	Never	No	No	2	2	Custo
2023/06/0	46	Other	once a we	Groceries	Yes	Rarely	categories	First page	2	Yes	Always	High shipp	Sometime	No	Rarely	Sometime	Sometime	5	2	Wide
2023/06/0	15	Other	once a mo	Groceries	Sometime	Few times	Keyword	Multiple p	2	Maybe	Often	others	Often	Yes	Never	No	Yes	2	2	Quick

Service_A	Improvem	transactio	age_group	Satisfactic	Frequency	Customer	Recomme	Recomme	Behavior_Cluste	
Competiti	better app	778242	55+	High		At-Risk Cu	2	2	1	
Quick deli	scrolling c	193482	18-24	High	Medium	At-Risk Cu	2	2	0	
All the ab	Unknown	925975	35-44	Medium	Medium	At-Risk Cu	1	1	1	
Quick deli	quality of	566872	55+	Low	Medium	At-Risk Cu	3	3	2	
Quick deli	irrelevant	683642	45-54	Low	Medium	At-Risk Cu	3	3	2	
Product re	nothing	461923	55+	Medium		At-Risk Cu	1	1	1	

TASK 4: RECOMMENDATION AND REVIEW INSIGHTS

CODES PERFORMED IN PYTHON

```python
#TASK 4: RECOMMENDATION & REVIEW INSIGHTS
df[['Recommendation_Helpfulness_Num', 'Shopping_Satisfaction']].corr()
```

```python
import os

os.makedirs("../outputs/charts_task_4", exist_ok=True)
```

```python
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(6,4))
sns.scatterplot(
    x='Recommendation_Helpfulness_Num',
    y='Shopping_Satisfaction',
    data=df
)
plt.title("Recommendation Helpfulness vs Shopping Satisfaction")
plt.tight_layout()
plt.savefig("../outputs/charts_task_4/recommendation_helpfulness_vs_satisfaction.png")
plt.show()
```

```python
df['Review_Reliability'].value_counts()
```

```python
review_rating = df.groupby('Review_Reliability')['Rating_Accuracy'].mean()
review_rating
```

```python
import os
os.makedirs("../outputs/charts_task_4", exist_ok=True)

plt.figure(figsize=(6,4))
sns.barplot(
    x='Review_Reliability',
    y='Rating_Accuracy',
    data=df,
    estimator='mean'
)
```

```python
    estimator='mean'
)
plt.title("Review Reliability vs Rating Accuracy")
plt.tight_layout()
plt.savefig("../outputs/charts_task_4/review_reliability_vs_rating_accuracy.png")
plt.show()
```

```python
df['Personalized_Recommendation_Frequency'].value_counts()
```

```python
plt.figure(figsize=(6,4))
sns.countplot(
    x='Personalized_Recommendation_Frequency',
    data=df,
    order=df['Personalized_Recommendation_Frequency'].value_counts().index
)
plt.title("Engagement with Personalized Recommendations")
plt.xticks(rotation=30)
plt.tight_layout()
plt.savefig("../outputs/charts_task_4/recommendation_engagement_frequency.png")
plt.show()
```

```python
df.groupby('Personalized_Recommendation_Frequency')['Shopping_Satisfaction'].mean()
```

```python
df.groupby('Customer_Segment')['Recommendation_Helpfulness_Num'].mean()
```

```python
df.to_csv("../outputs/task_4.csv", index=False)
```

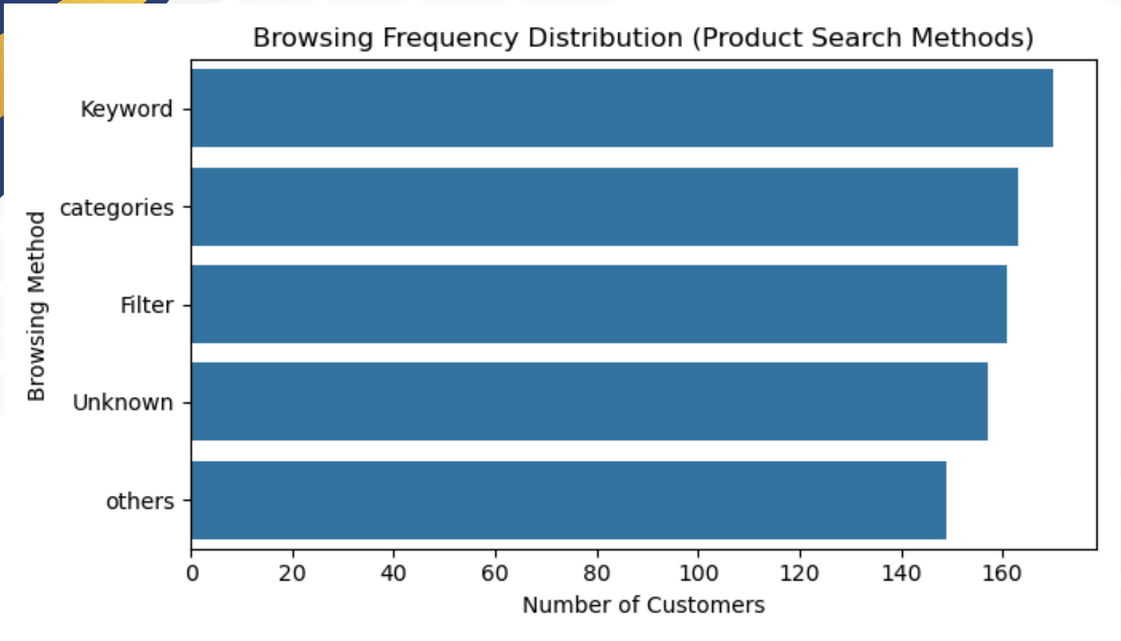
TASK 4: RECOMMENDATION

#IMPROVE RECOMMENDATION RELEVANCE FOR AT-RISK CUSTOMERS TO REDUCE CHURN #OPTIMIZE RECOMMENDATION FREQUENCY TO AVOID USER FATIGUE #PRIORITIZE VERIFIED AND RELIABLE REVIEWS IN RECOMMENDATION RANKING #USE BEHAVIORAL CLUSTERS TO PERSONALIZE RECOMMENDATIONS #ENHANCE UI/UX TO MAKE RECOMMENDATIONS CLEARER AND LESS CLUTTERED

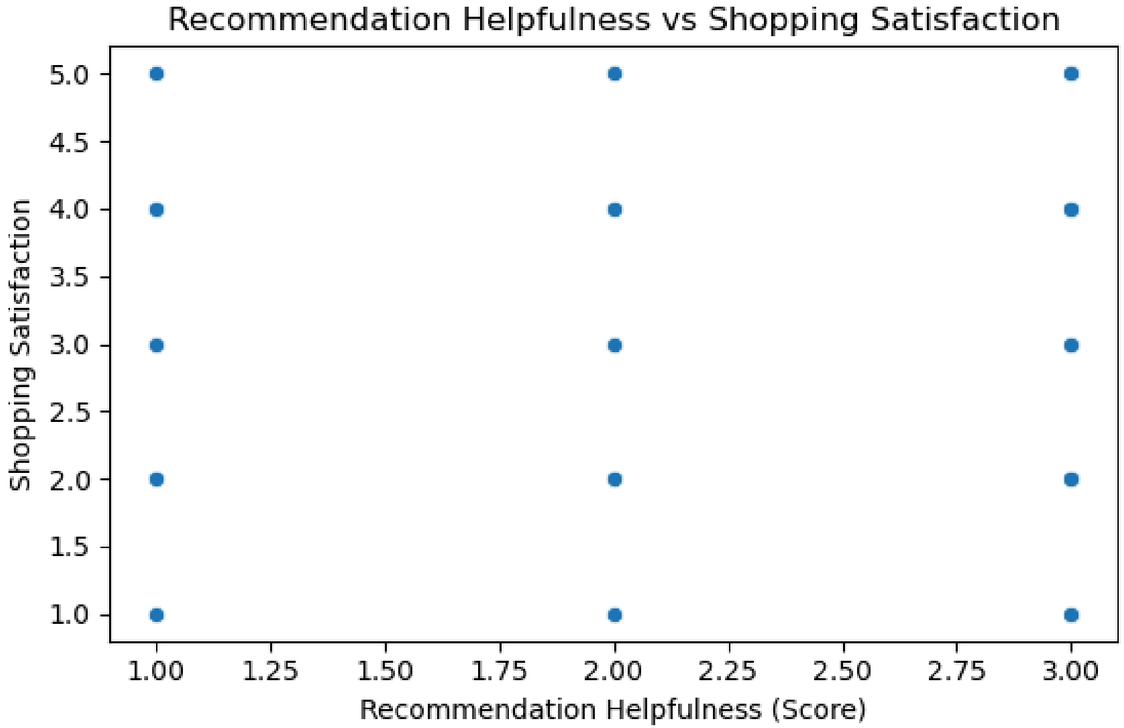
TASK 5: VISUALIZATION AND REPORTING

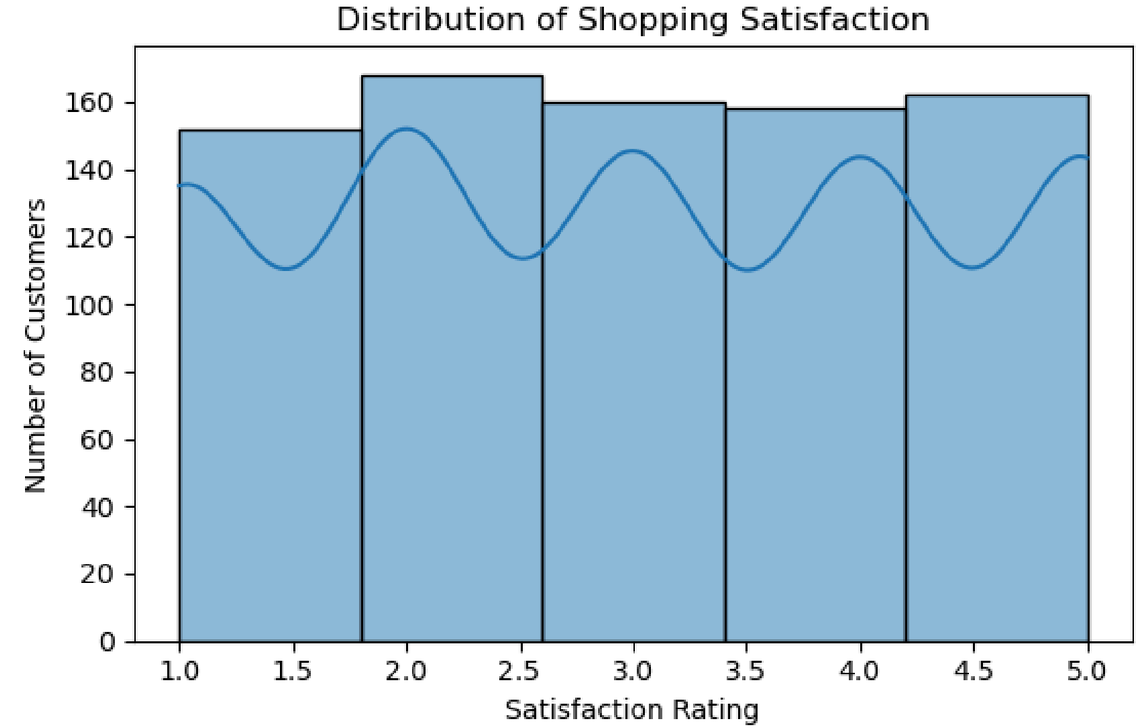
- CREATE ATTRACTIVE VISUALIZATIONS (BAR CHARTS, HEATMAPS, PIE CHARTS) FOR: ○ PURCHASE CATEGORIES ○ BROWSING FREQUENCY DISTRIBUTION ○ SATISFACTION LEVELS ○ CORRELATION BETWEEN RECOMMENDATION USEFULNESS AND SATISFACTION
- SUMMARIZE FINDINGS IN A CLEAR AND VISUALLY APPEALING DASHBOARD OR REPORT FORMAT.

TASK 5: VISUALISATION



TASK 5: VISUALISATION





TASK 5: VISUALISATION

CODES PERFORMED IN PYTHON

```python
#task 5 : Visualization and Reporting
import os
os.makedirs("../outputs/charts_task_5", exist_ok=True)
```

```python
from collections import Counter
import seaborn as sns
import matplotlib.pyplot as plt

categories = df['Purchase_Categories'].dropna().str.split(',')
flat_categories = [c.strip() for sub in categories for c in sub]

category_counts = Counter(flat_categories)
top_categories = dict(category_counts.most_common(8))

plt.figure(figsize=(7,4))
sns.barplot(x=list(top_categories.values()), y=list(top_categories.keys()))
plt.title("Top Purchased Product Categories")
plt.xlabel("Number of Customers")
plt.ylabel("Category")
plt.tight_layout()
plt.savefig("../outputs/charts_task_5/purchase_categories.png")
plt.show()
```

```python
df['Product_Search_Method'].value_counts()
```

```python
import seaborn as sns
import matplotlib.pyplot as plt
import os

os.makedirs("../outputs/charts_task_5", exist_ok=True)

plt.figure(figsize=(7,4))
sns.countplot(
    y='Product_Search_Method',
    data=df,
    order=df['Product_Search_Method'].value_counts().index
)
plt.title("Browsing Frequency Distribution (Product Search Methods)")
plt.xlabel("Number of Customers")
plt.ylabel("Browsing Method")
plt.tight_layout()
plt.savefig("../outputs/charts_task_5/browsing_frequency_distribution.png")
plt.show()
```

```python
import os
os.listdir("../outputs/charts_task_5")
```

```python
df['Shopping_Satisfaction'].value_counts().sort_index()
```

```python
import seaborn as sns
import matplotlib.pyplot as plt
import os

os.makedirs("../outputs/charts_task_5", exist_ok=True)

plt.figure(figsize=(6,4))
sns.histplot(
    df['Shopping_Satisfaction'],
    bins=5,
    kde=True
)
```

```python
df['Shopping_Satisfaction'].value_counts().sort_index()
```

```python
import seaborn as sns
import matplotlib.pyplot as plt
import os

os.makedirs("../outputs/charts_task_5", exist_ok=True)

plt.figure(figsize=(6,4))
sns.histplot(
    df['Shopping_Satisfaction'],
    bins=5,
    kde=True
)
plt.title("Distribution of Shopping Satisfaction")
plt.xlabel("Satisfaction Rating")
plt.ylabel("Number of Customers")
plt.tight_layout()
plt.savefig("../outputs/charts_task_5/satisfaction_levels.png")
plt.show()
```

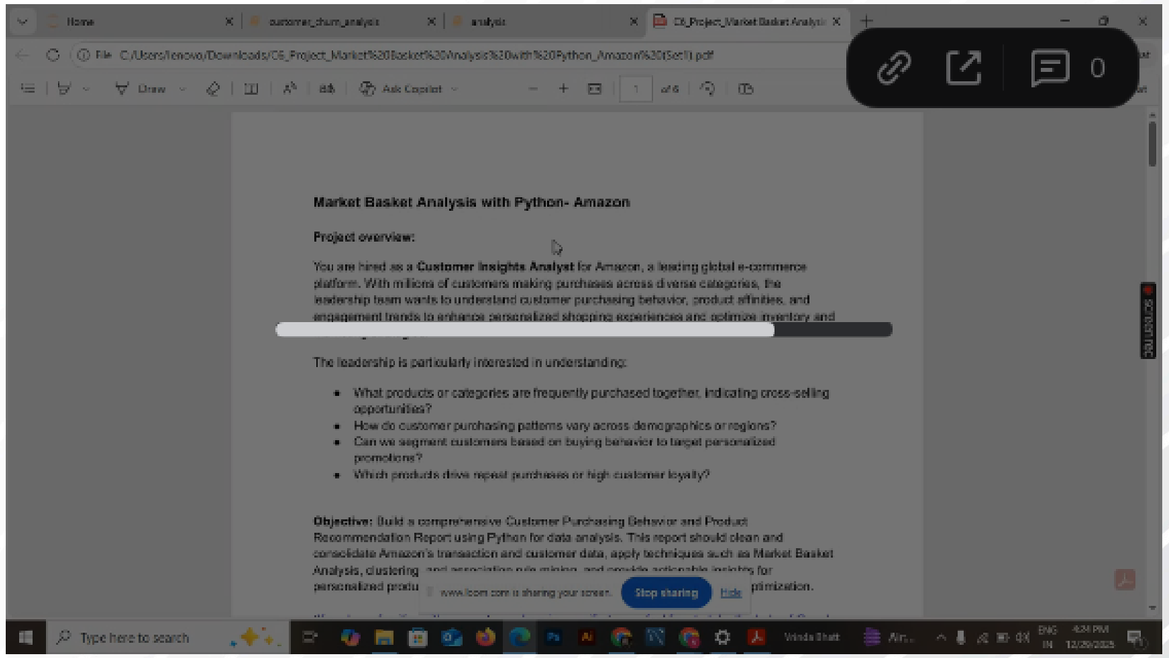
TASK 5: INSIGHTS

#MOST CUSTOMERS REPORT MEDIUM TO HIGH SATISFACTION LEVELS

#INDICATES OVERALL POSITIVE SHOPPING EXPERIENCE WITH SCOPE FOR IMPROVEMENT AMONG LOW-SATISFACTION USERS

THANK YOU

VIDEO LINK:



VIDEO LINK