Customer Revenue & Segmentation Analysis

Karim Hamada

Overview

In today’s competitive market environment, organizations must shift from intuition-driven decisions to data-driven revenue optimization strategies. Despite having access to transactional and customer-level sales data, many companies struggle to identify their most valuable customers and the product categories that truly drive profitability This project aims to transform raw sales data into strategic business insights by analyzing customer purchasing behavior, revenue contribution patterns, and product performance metrics. Through structured data analysis and customer segmentation techniques, the study identifies high-value customer segments, evaluates category-level profitability, and uncovers behavioral differences across demographic groups.

Methodology

Problem Statement

Data Collection

Inspect & Handel Missing data

Eploratory data analysis

Problem Statement

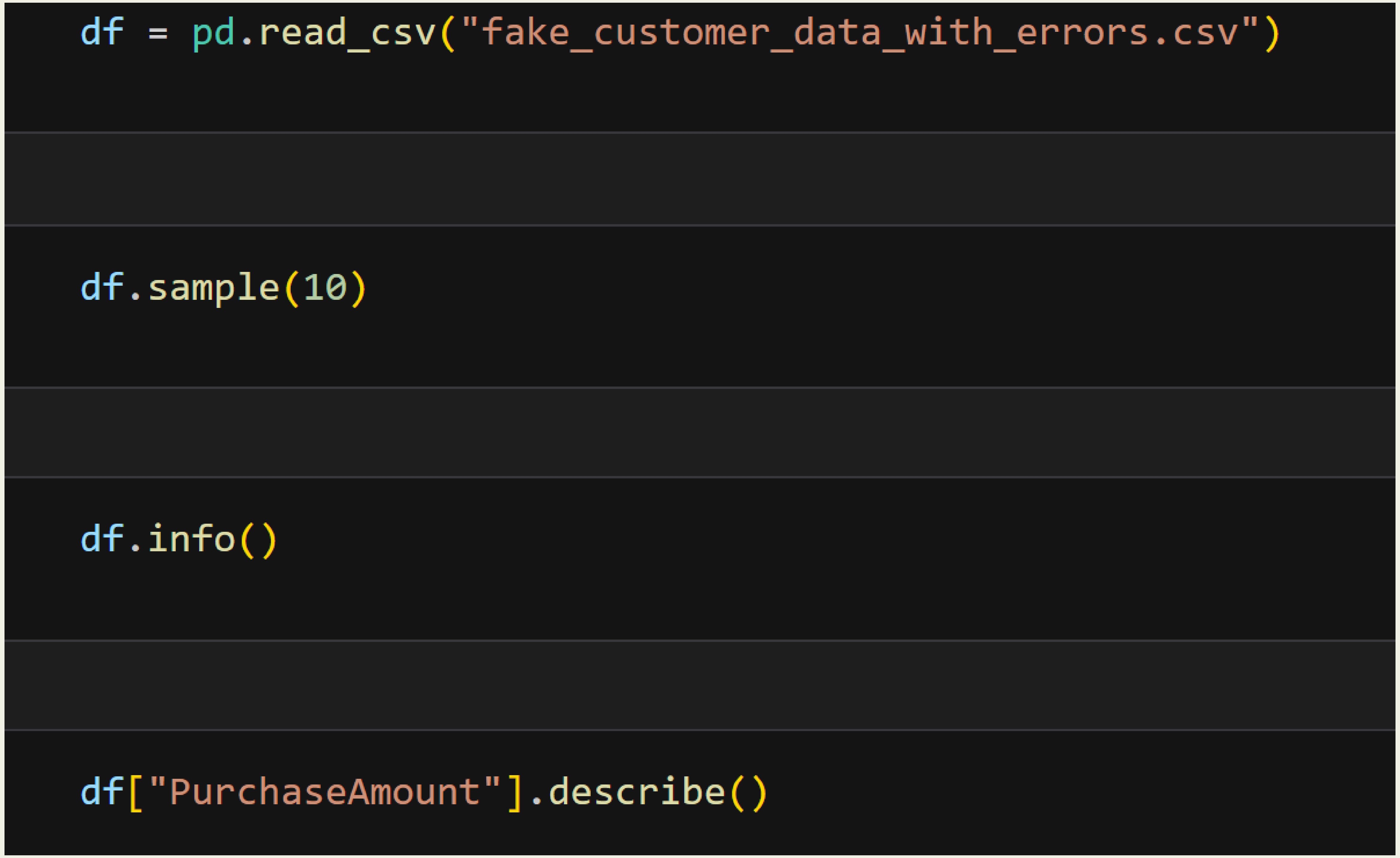
The company possesses extensive transactional data; however, it lacks analytical visibility into customer profitability drivers and segment-level revenue contribution. Without identifying high-value customers and high-performing product categories, marketing investments and strategic initiatives risk being inefficient and misaligned with revenue optimization goals. Therefore, a structured customer segmentation and profitability analysis is required to uncover revenue concentration patterns, behavioral differences, and growth opportunities that can directly enhance business performance.

“How can the company increase revenue by identifying high-value customers and profitable product categories?”

Data Collection

kaggle™

Inspect & Handel Missing data



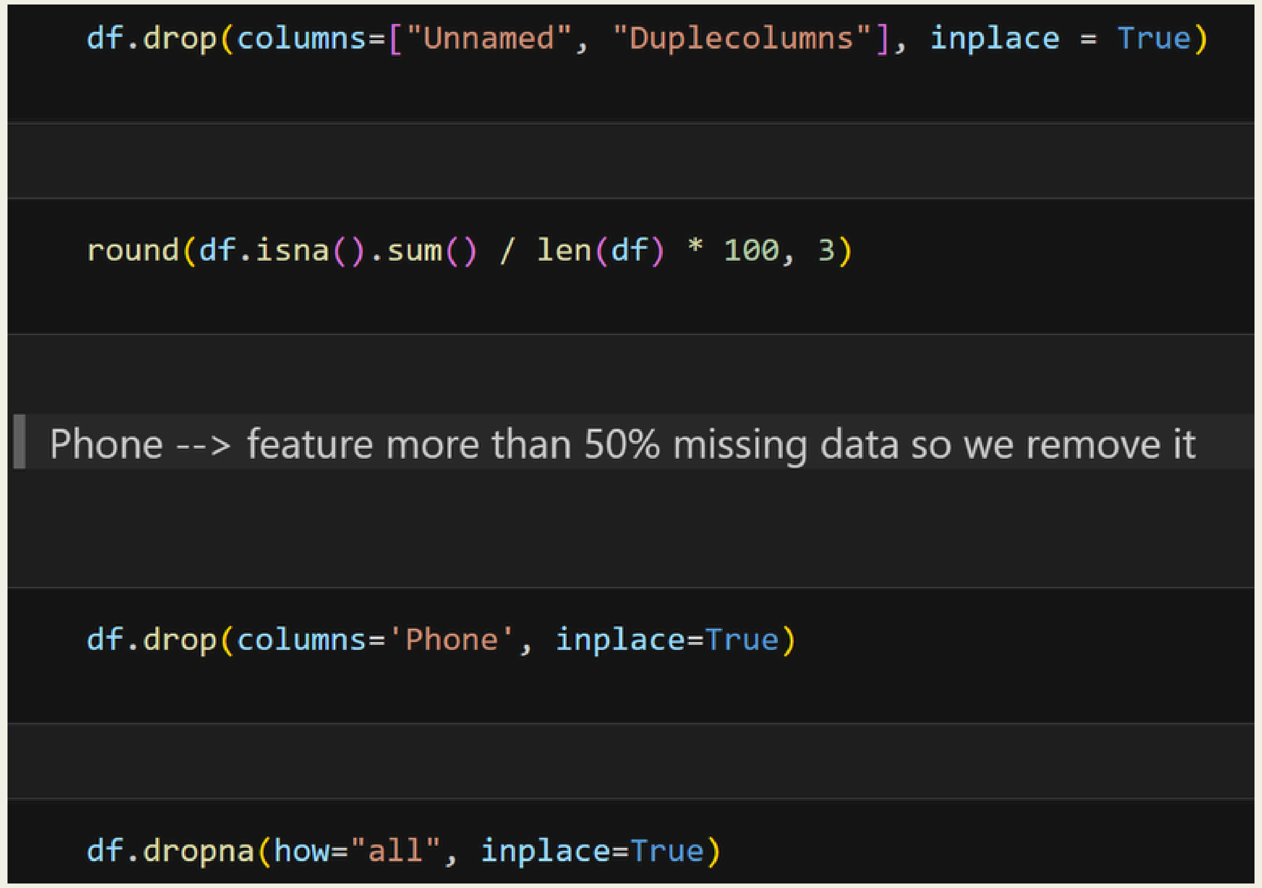
```
df = pd.read_csv("fake_customer_data_with_errors.csv")

df.sample(10)

df.info()

df["PurchaseAmount"].describe()
```

Inspect & Handel Missing data



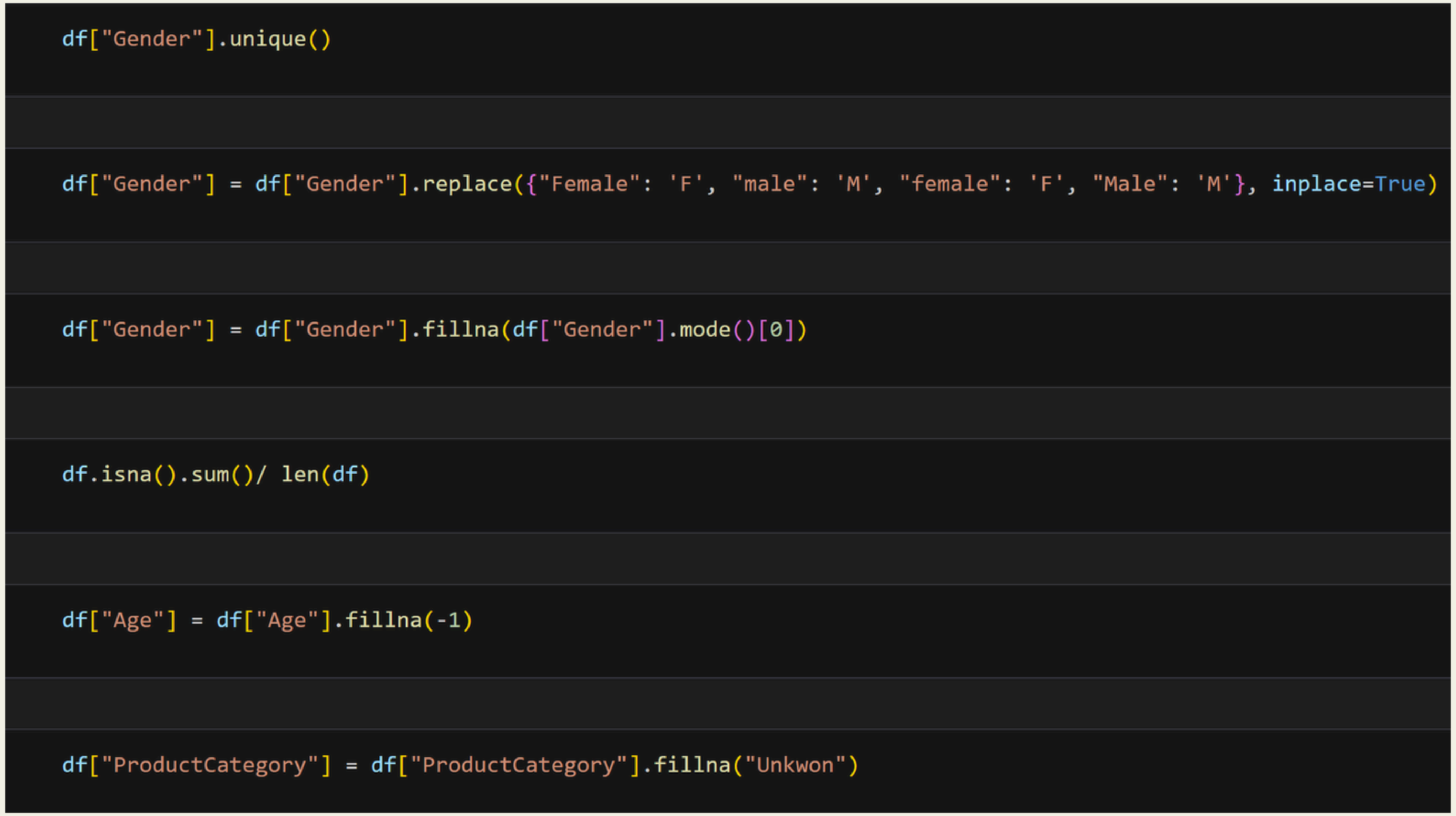
```
df.drop(columns=["Unnamed", "Duplecolumns"], inplace = True)
```

```
round(df.isna().sum() / len(df) * 100, 3)
```

> Phone --> feature more than 50% missing data so we remove it

```
df.drop(columns='Phone', inplace=True)
```

```
df.dropna(how="all", inplace=True)
```



```
df["Gender"].unique()
```

```
df["Gender"] = df["Gender"].replace({"Female": 'F', "male": 'M', "female": 'F', "Male": 'M'}, inplace=True)
```

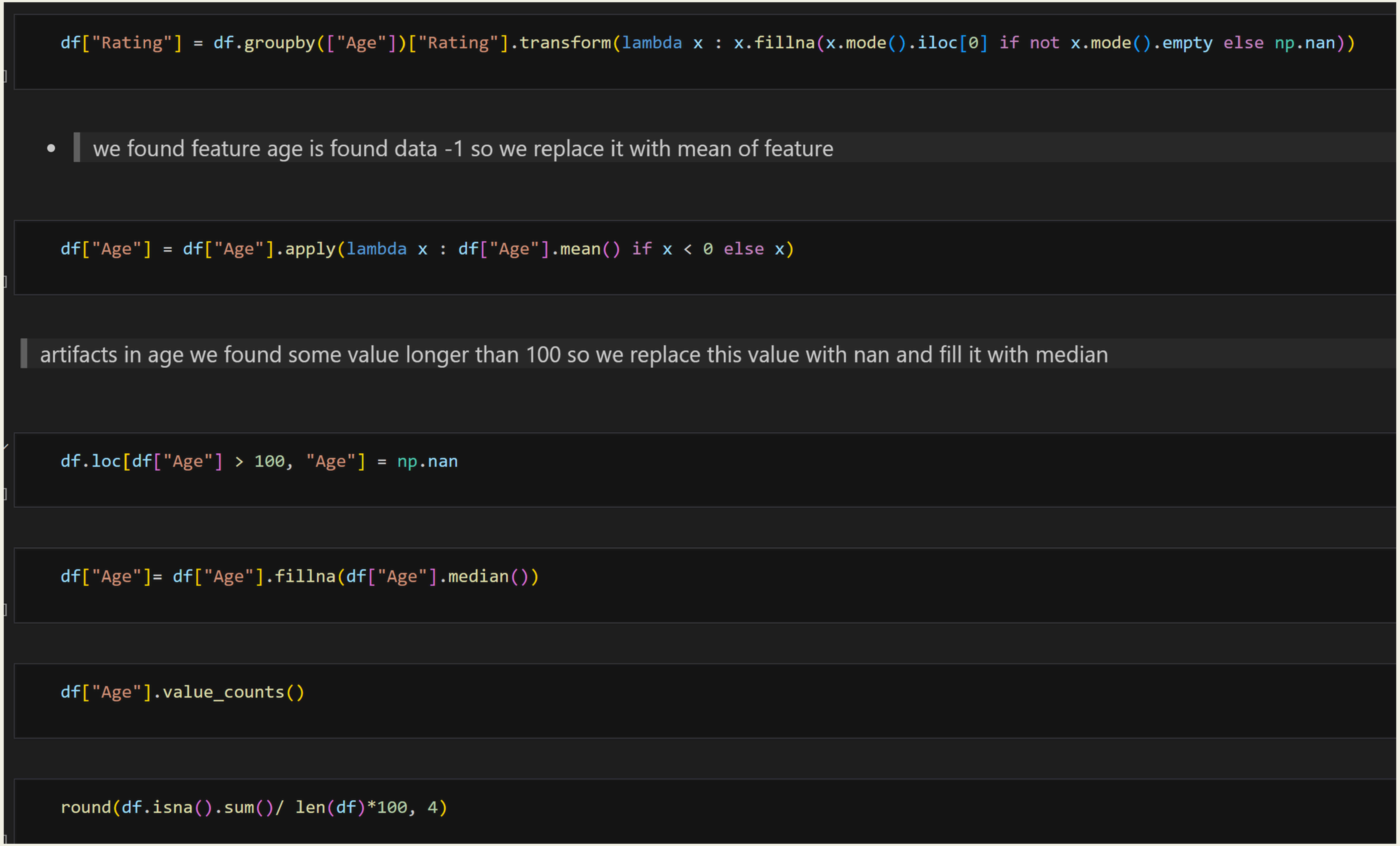
```
df["Gender"] = df["Gender"].fillna(df["Gender"].mode()[0])
```

```
df.isna().sum()/ len(df)
```

```
df["Age"] = df["Age"].fillna(-1)
```

```
df["ProductCategory"] = df["ProductCategory"].fillna("Unkwon")
```

Inspect & Handel Missing data



```
df["Rating"] = df.groupby(["Age"])["Rating"].transform(lambda x : x.fillna(x.mode().iloc[0] if not x.mode().empty else np.nan))
```

- > we found feature age is found data -1 so we replace it with mean of feature

```
df["Age"] = df["Age"].apply(lambda x : df["Age"].mean() if x < 0 else x)
```

> artifacts in age we found some value longer than 100 so we replace this value with nan and fill it with median

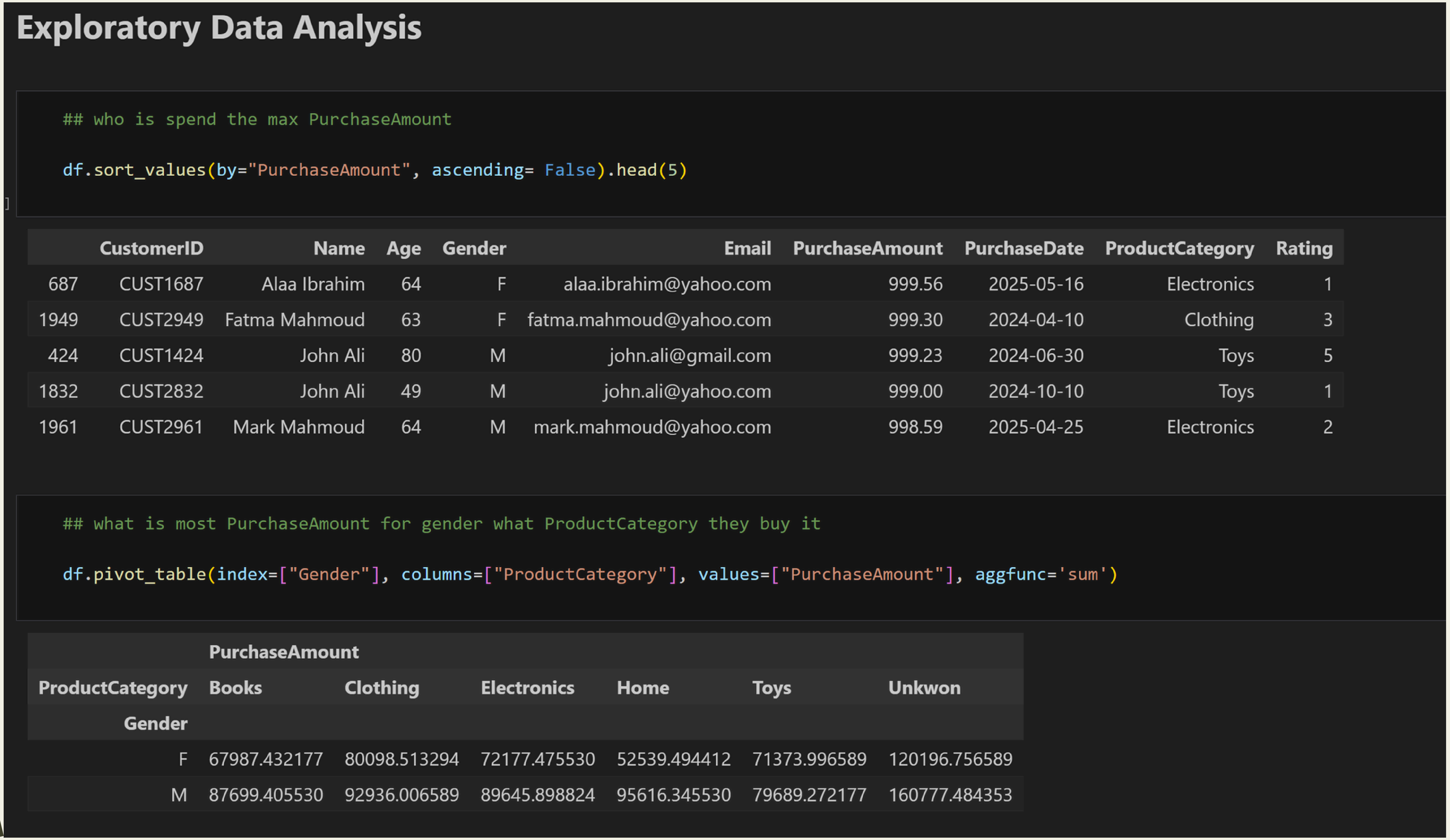
```
df.loc[df["Age"] > 100, "Age"] = np.nan
```

```
df["Age"]= df["Age"].fillna(df["Age"].median())
```

```
df["Age"].value_counts()
```

```
round(df.isna().sum()/ len(df)*100, 4)
```

EDA



Exploratory Data Analysis

```
## who is spend the max PurchaseAmount

df.sort_values(by="PurchaseAmount", ascending= False).head(5)
```

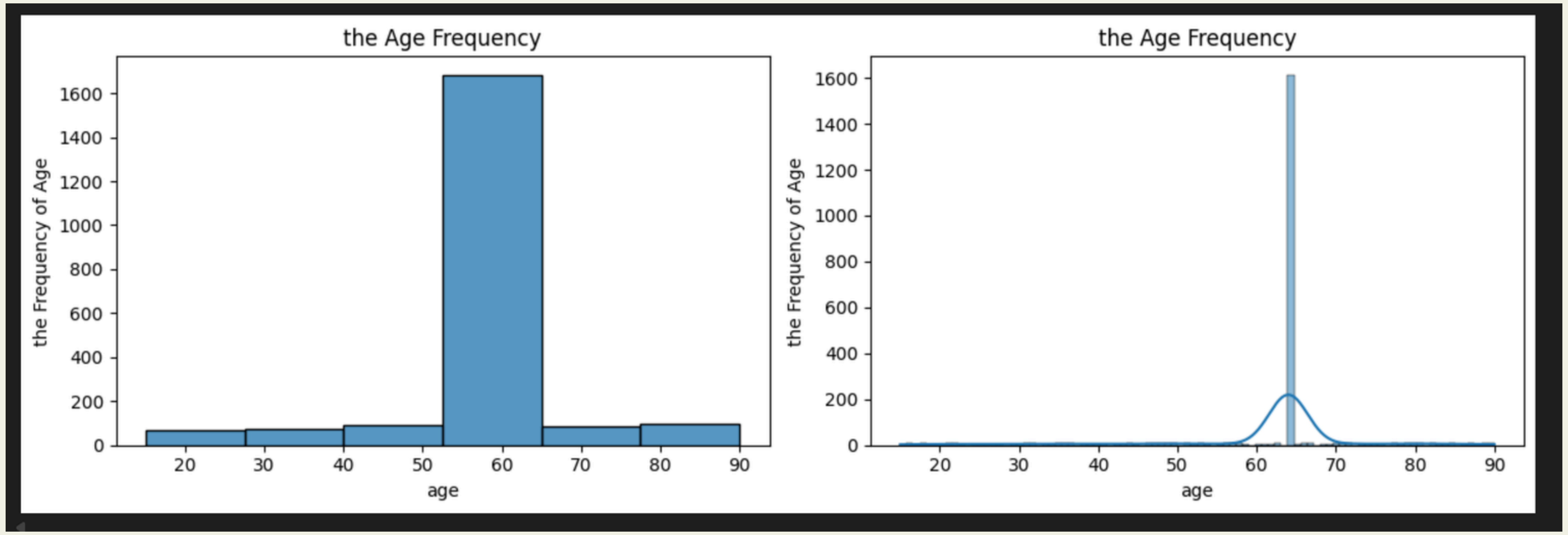
	CustomerID	Name	Age	Gender	Email	PurchaseAmount	PurchaseDate	ProductCategory	Rating
687	CUST1687	Alaa Ibrahim	64	F	alaa.ibrahim@yahoo.com	999.56	2025-05-16	Electronics	1
1949	CUST2949	Fatma Mahmoud	63	F	fatma.mahmoud@yahoo.com	999.30	2024-04-10	Clothing	3
424	CUST1424	John Ali	80	M	john.ali@gmail.com	999.23	2024-06-30	Toys	5
1832	CUST2832	John Ali	49	M	john.ali@yahoo.com	999.00	2024-10-10	Toys	1
1961	CUST2961	Mark Mahmoud	64	M	mark.mahmoud@yahoo.com	998.59	2025-04-25	Electronics	2

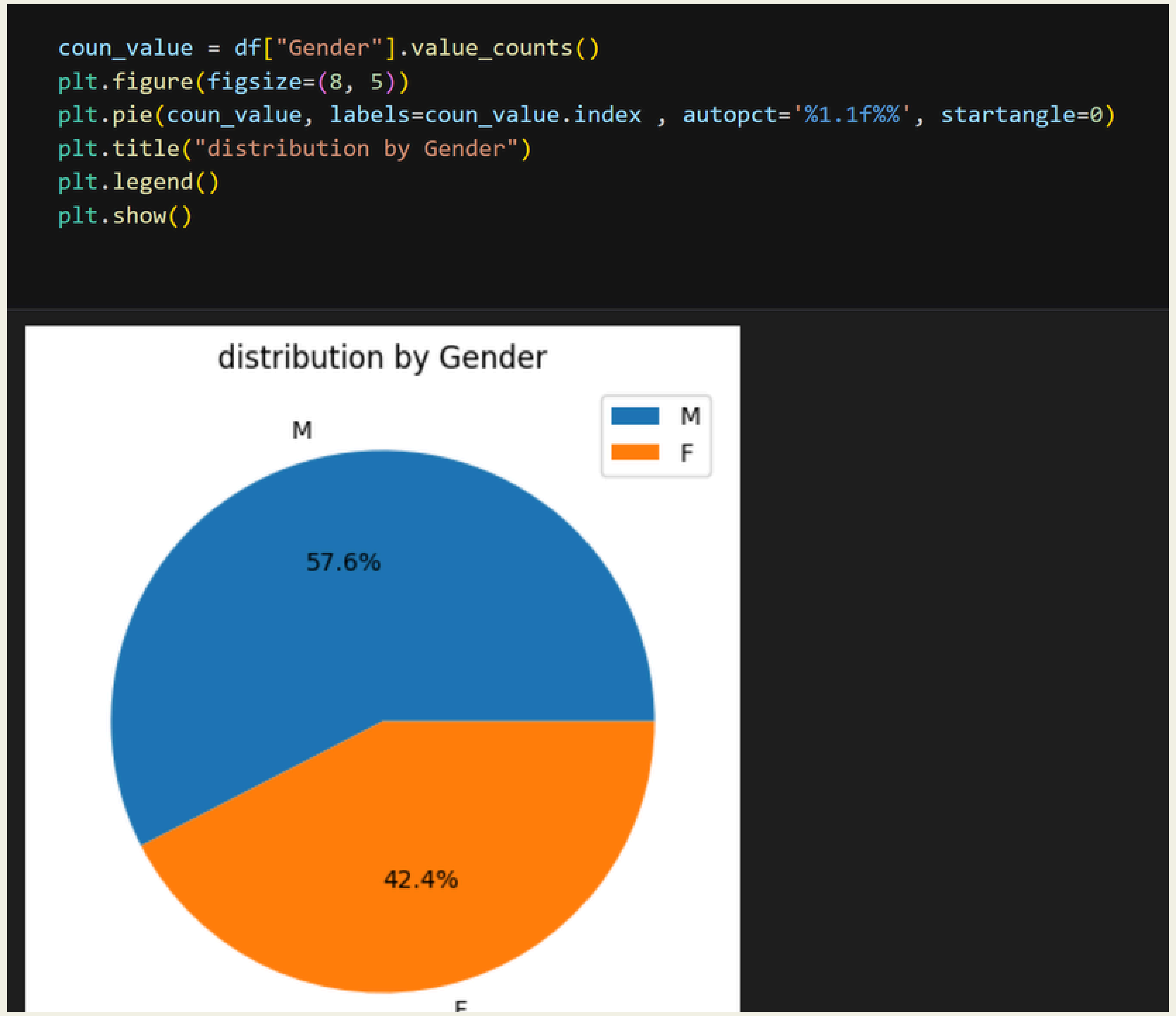
```
## what is most PurchaseAmount for gender what ProductCategory they buy it

df.pivot_table(index=["Gender"], columns=["ProductCategory"], values=["PurchaseAmount"], aggfunc='sum')
```

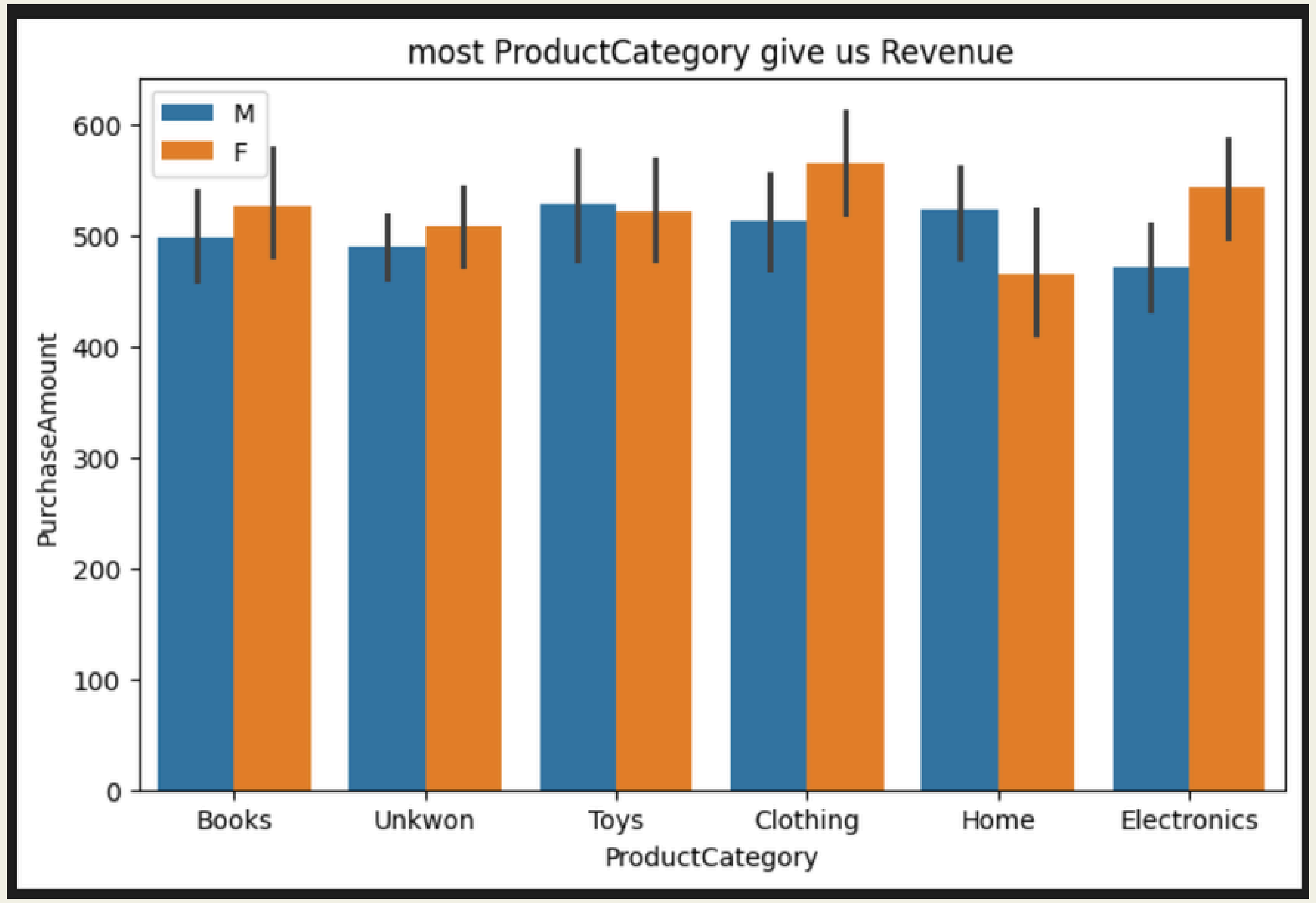
	PurchaseAmount					
ProductCategory	Books	Clothing	Electronics	Home	Toys	Unkwon
Gender						
F	67987.432177	80098.513294	72177.475530	52539.494412	71373.996589	120196.756589
M	87699.405530	92936.006589	89645.898824	95616.345530	79689.272177	160777.484353

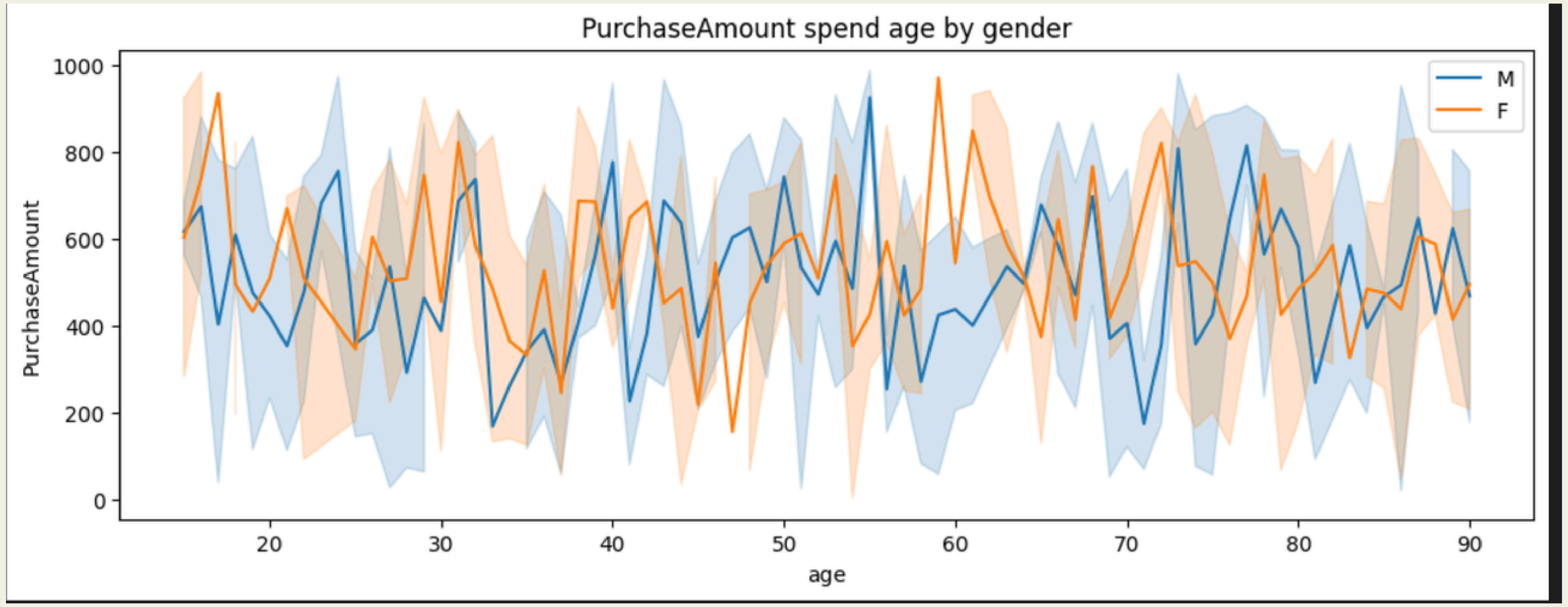
EDA



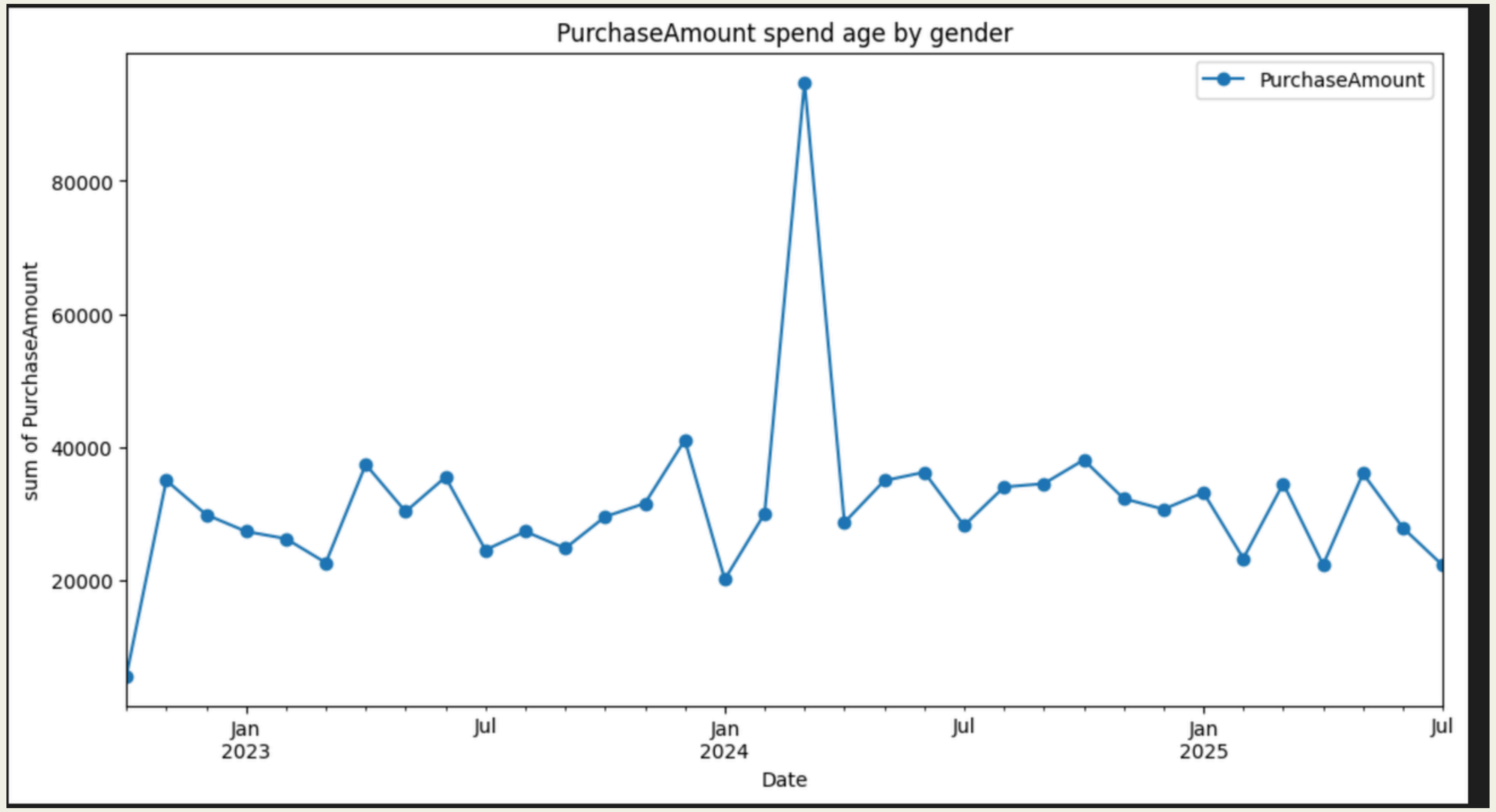


```
coun_value = df["Gender"].value_counts()
plt.figure(figsize=(8, 5))
plt.pie(coun_value, labels=coun_value.index , autopct='%1.1f%%', startangle=0)
plt.title("distribution by Gender")
plt.legend()
plt.show()
```

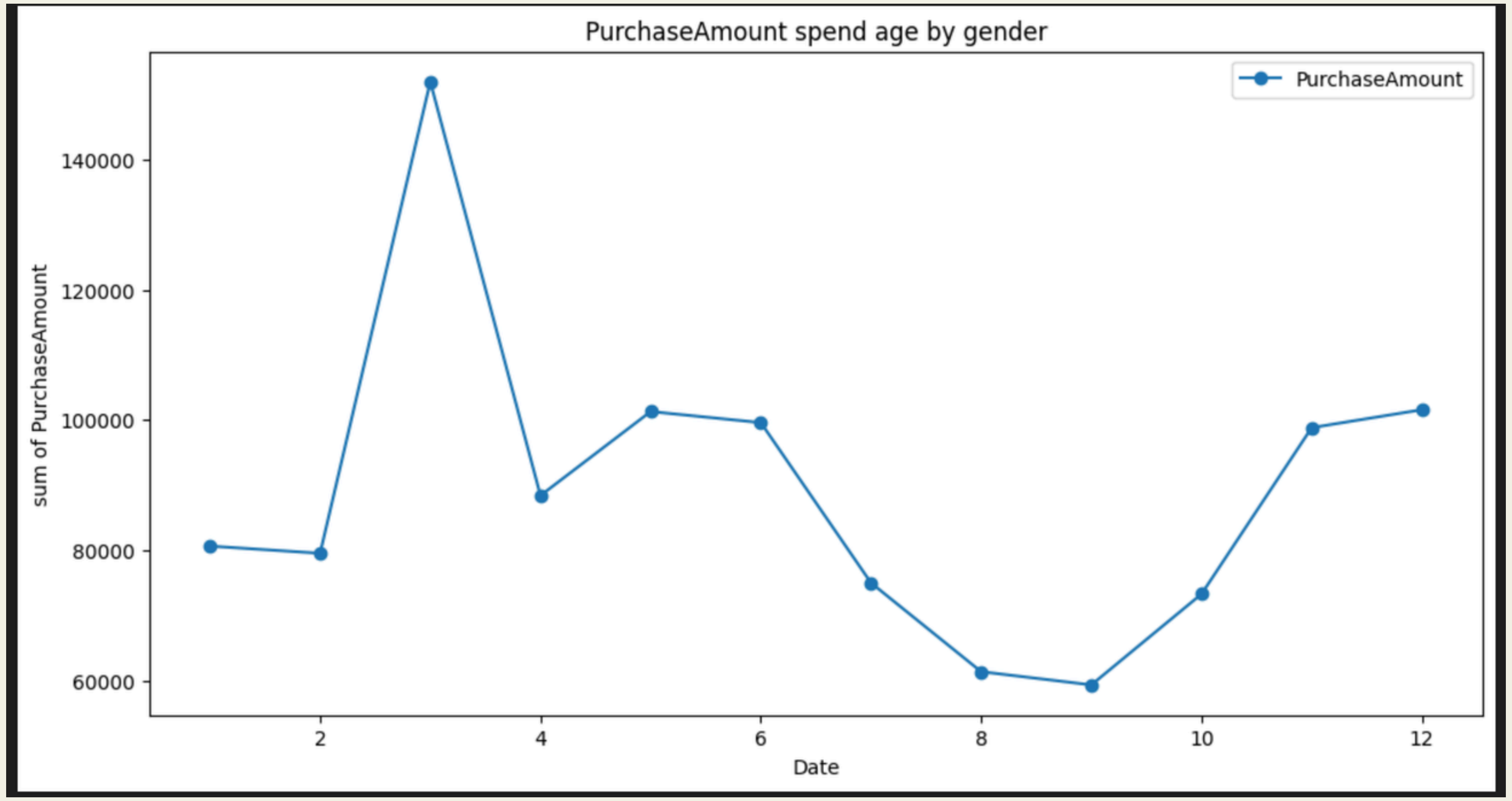




EDA



PurchaseAmount spend age by gender
PurchaseAmount
80000
60000
40000
20000
sum of PurchaseAmount
Jan 2023
Jul
Jan 2024
Jul
Jan 2025
Jul
Date



PurchaseAmount spend age by gender
PurchaseAmount
140000
120000
100000
80000
60000
sum of PurchaseAmount
2
4
6
8
10
12
Date

Karim Hamada

karimahamda221@gmail.com

Karim Hamada