Assignment 2 – Exploring the Data

- List the features and identify their types (nominal, numeric, etc.)

```python
# Load data
import pandas as pd
from datetime import datetime

data = {
    'CustomerID': [101, 102, 103, 104, 105],
    'Age': [25, 34, 45, 23, 35],
    'Gender': ['F', 'M', 'F', 'F', 'M'],
    'ProductCategory': ['Electronics', 'Clothing', 'Clothing', 'Groceries', 'Electronics'],
    'PurchaseAmount': [120.50, 45.00, 78.90, 13.50, 220.00],
    'PaymentMethod': ['Credit Card', 'Cash', 'Credit Card', 'Debit Card', 'Credit Card'],
    'StoreRegion': ['North', 'South', 'North', 'West', 'East'],
    'PurchaseDate': pd.to_datetime(['2024-06-01', '2024-06-01', '2024-06-02', '2024-06-03', '2024-06-03'])
}
df = pd.DataFrame(data)
df.head()
```

	CustomerID	Age	Gender	ProductCategory	PurchaseAmount	PaymentMethod	StoreRegion	PurchaseDate
0	101	25	F	Electronics	120.5	Credit Card	North	2024-06-01
1	102	34	M	Clothing	45.0	Cash	South	2024-06-01
2	103	45	F	Clothing	78.9	Credit Card	North	2024-06-02
3	104	23	F	Groceries	13.5	Debit Card	West	2024-06-03
4	105	35	M	Electronics	220.0	Credit Card	East	2024-06-03

Next steps: Generate code with df | View recommended plots | New interactive sheet

🔍 Describing the Dataset

Feature List and Data Types		
Feature	**Data Type**	**Description**
Customer ID	Numeric (Int)	Unique ID for each customer
Age	Numeric (Int)	Customer's age
Gender	Nominal (object)	Gender of the customer
ProductCategory	Nominal (object)	Product category purchased
PurchaseAmount	Numeric (float)	Total amount spent
PaymentMethod	Nominal (object)	Method used for payment
StoreRegion	Nominal (object)	Region where purchase was made
PurchaseDate	DateTime	Date of the purchase

- Summary Statistics (.describe() output)

```
# Summary stats
df.describe()
```

	CustomerID	Age	PurchaseAmount	PurchaseDate
count	5.000000	5.000000	5.000000	5
mean	103.000000	32.400000	95.580000	2024-06-02 00:00:00
min	101.000000	23.000000	13.500000	2024-06-01 00:00:00
25%	102.000000	25.000000	45.000000	2024-06-01 00:00:00
50%	103.000000	34.000000	78.900000	2024-06-02 00:00:00
75%	104.000000	35.000000	120.500000	2024-06-03 00:00:00
max	105.000000	45.000000	220.000000	2024-06-03 00:00:00
std	1.581139	8.820431	80.117145	NaN

Count: 5 rows

Mean Purchase Amount: 95.58

Min Purchase Amount: 13.5

Max Purchase Amount: 220.0

Standard Deviation: 80.12

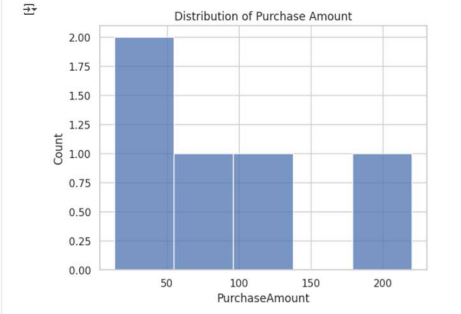
25% Quartile (Q1): 45.0

Median (Q2): 78.9

75% Quartile (Q3): 120.5

- Range, Mean and Standard deviation of PurchaseAmount

```
sns.histplot(df['PurchaseAmount'], bins=5)
plt.title('Distribution of Purchase Amount')
plt.show()
```

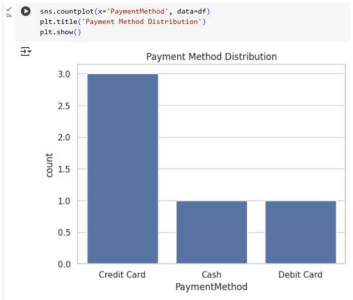


Range: 220.0 - 13.5 = 206.5

Mean: 95.58 (from .describe())

Standard Deviation: 80.12 (from .describe())

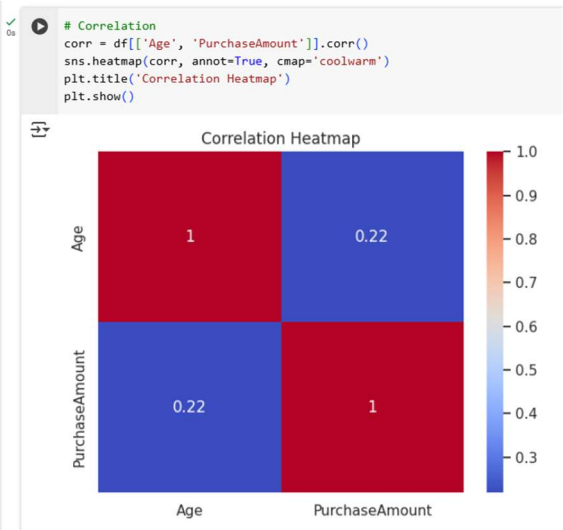
📊 Visualize the Data

- Include or describe two charts:
  - Chart1: Histogram of PurchaseAmount
    - **Why used**: Shows the distribution of purchase values
    - **Reveals**: Most purchases are clustered in the lower range, with one high outlier
  - Chart2: Bar Chart of PaymentMethod
    - 

    - **Why used**: Categorical variable best visualized with bar chart.
    - **Reveals**: Majority of purchases are made with credit card.

🔗 Explore Relationships

- Correlation Heatmap



  - Explored correlation between **Age** and **Purchase Amount**
  - Correlation Coefficient: 0.22 (weak positive correlation)
- Interpretation
  - There is a slight positive relationship between age and spending. Older customers may tend to spend a bit more.

- This weak correlation suggests age might be a minor factor in predicting purchase amount, but it shouldn't be the only feature considered for modeling.