```python
import pandas as pd

# Load CSV file (replace with your actual filename if different)
df = pd.read_csv("online_retail_ii.csv")

# Quick preview
print("Dataset Preview:")
display(df.head())

# Shape of dataset
print(f"\nDataset shape: {df.shape}")

# Null values summary
null_report = pd.DataFrame({
    "Null Count": df.isnull().sum(),
    "Null Percentage": (df.isnull().sum() / len(df)) * 100
}).sort_values(by="Null Count", ascending=False)

print("\nNull Values Report:")
display(null_report)
```

```
Dataset Preview:
```

	Invoice	StockCode	Description	Quantity	\
0	489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	
1	489434	79323P	PINK CHERRY LIGHTS	12	
2	489434	79323W	WHITE CHERRY LIGHTS	12	
3	489434	22041	RECORD FRAME 7" SINGLE SIZE	48	
4	489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	

	InvoiceDate	Price	Customer ID	Country
0	2009-12-01 07:45:00	6.95	13085.0	United Kingdom
1	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
2	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
3	2009-12-01 07:45:00	2.10	13085.0	United Kingdom
4	2009-12-01 07:45:00	1.25	13085.0	United Kingdom

```
Dataset shape: (1067371, 8)

Null Values Report:
```

	Null Count	Null Percentage
Customer ID	243007	22.766873
Description	4382	0.410541
Invoice	0	0.000000
StockCode	0	0.000000
Quantity	0	0.000000
InvoiceDate	0	0.000000
Price	0	0.000000
Country	0	0.000000

```python
import pandas as pd

# Load your dataset
df = pd.read_csv("online_retail_ii.csv")

# 1. Drop rows with missing Customer ID (important for retention analysis)
df = df.dropna(subset=['Customer ID'])

# 2. Fill missing product descriptions with "Unknown"
df['Description'] = df['Description'].fillna("Unknown")

# 3. Save cleaned dataset as a new CSV
df.to_csv("online_retail_ii_cleaned.csv", index=False)

print("✅ Cleaned dataset saved as 'online_retail_ii_cleaned.csv'")
print(f"Final shape: {df.shape}")
```

```
✅ Cleaned dataset saved as 'online_retail_ii_cleaned.csv'
Final shape: (824364, 8)
```

```python
print("Dataset Preview:")
display(df.head())

# Shape of dataset
print(f"\nDataset shape: {df.shape}")

# Null values summary
null_report = pd.DataFrame({
    "Null Count": df.isnull().sum(),
    "Null Percentage": (df.isnull().sum() / len(df)) * 100
}).sort_values(by="Null Count", ascending=False)

print("\nNull Values Report:")
display(null_report)
```

```
Dataset Preview:
```

	Invoice	StockCode	Description	Quantity	\
0	489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	
1	489434	79323P	PINK CHERRY LIGHTS	12	
2	489434	79323W	WHITE CHERRY LIGHTS	12	
3	489434	22041	RECORD FRAME 7" SINGLE SIZE	48	
4	489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	

	InvoiceDate	Price	Customer ID	Country
0	2009-12-01 07:45:00	6.95	13085.0	United Kingdom
1	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
2	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
3	2009-12-01 07:45:00	2.10	13085.0	United Kingdom
4	2009-12-01 07:45:00	1.25	13085.0	United Kingdom

```
Dataset shape: (824364, 8)

Null Values Report:

             Null Count  Null Percentage
Invoice               0              0.0
StockCode             0              0.0
Description           0              0.0
Quantity              0              0.0
InvoiceDate           0              0.0
Price                 0              0.0
Customer ID           0              0.0
Country               0              0.0
```

Data Cleaning Decisions

1. Dropping Missing Customer ID

The Customer ID is critical for customer-level analysis such as retention, churn, and cohort analysis.

Transactions without a customer identifier cannot be linked to any individual, which would:

Distort retention calculations (since those customers can’t belong to a cohort).

Inflate order/revenue counts without contributing to meaningful customer insights.

 Decision: We dropped rows with missing Customer ID to ensure only identifiable customers are used in retention and churn analysis.

1. Filling Missing Description with "Unknown"

The Description field provides product details, but each product is still uniquely identifiable using StockCode.

Only a small fraction (~0.4%) of rows had missing descriptions, so dropping them would cause unnecessary data loss.

 Decision: We replaced null values in Description with "Unknown", preserving transactions while keeping the dataset consistent for product-level analysis.