# Data Report Before Preprocessing

## Dataset Overview

Our dataset consists of 536641 rows and 9 columns.

## Dataset Information

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 536641 entries, 0 to 536640

Data columns (total 9 columns):

```
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    536641 non-null  object
 1   StockCode    536641 non-null  object
 2   Description  535187 non-null  object
 3   Quantity     536641 non-null  int64
 4   InvoiceDate  536641 non-null  object
 5   UnitPrice    536641 non-null  float64
 6   CustomerID   401604 non-null  float64
 7   Country      536641 non-null  object
 8   Active       536641 non-null  object
dtypes: float64(2), int64(1), object(6)
memory usage: 36.8+ MB
```

## Summary Statistics

| Statistic | Quantity | UnitPrice | CustomerID |
|---|---|---|---|
| count | 536641.00 | 536641.00 | 401604.00 |
| mean | 9.62 | 4.63 | 15281.16 |
| std | 219.13 | 97.23 | 1714.01 |
| min | -80995.00 | -11062.06 | 12346.00 |
| 25% | 1.00 | 1.25 | 13939.00 |
| 50% | 3.00 | 2.08 | 15145.00 |

| | | | |
|---|---|---|---|
| 75% | 10.00 | 4.13 | 16784.00 |
| max | 80995.00 | 38970.00 | 18287.00 |

**Active Column Distribution**