Project Title: Online Retail Data Cleaning & Quality Analysis Using MySQL

Objective:

- Clean raw online retail transaction data
- Handle missing, duplicate, and invalid records
- Improve data quality for reliable business intelligence

Tools Used:

- MySQL Workbench & MySQL CLI
- Dataset from Kaggle (online_retail_II.csv)

Dataset Overview

Dataset Name: online_retail_II.csv
Rows (Raw): 1,067,371
Columns: 8
InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country

---

Data Import Steps

```
CREATE DATABASE RetailDB;

USE RetailDB;


CREATE TABLE online_retail (
  InvoiceNo VARCHAR(20),
  StockCode VARCHAR(20),
  Description TEXT,
  Quantity INT,
  InvoiceDate DATETIME,
  UnitPrice DECIMAL(10,2),
  CustomerID INT,
  Country VARCHAR(100)
```

);

```
LOAD DATA LOCAL INFILE ' E:/Retail_Data_Project/data/online_retail_II.csv'
INTO TABLE online_retail
FIELDS TERMINATED BY ',' ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 ROWS;
SELECT * FROM online_retail LIMIT 10;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	2009-12-01 07:45:00	6.95	13085	United Kingdom
489434	79323P	PINK CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085	United Kingdom
489434	79323W	WHITE CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085	United Kingdom
489434	22041	RECORD FRAME 7" SINGLE SIZE	48	2009-12-01 07:45:00	2.10	13085	United Kingdom
489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	2009-12-01 07:45:00	1.25	13085	United Kingdom
489434	22064	PINK DOUGHNUT TRINKET POT	24	2009-12-01 07:45:00	1.65	13085	United Kingdom
489434	21871	SAVE THE PLANET MUG	24	2009-12-01 07:45:00	1.25	13085	United Kingdom
489434	21523	FANCY FONT HOME SWEET HOME DOORMAT	10	2009-12-01 07:45:00	5.95	13085	United Kingdom

---

Identified Data Quality Issues

1. Missing Values

```
SELECT COUNT(*) FROM online_retail WHERE CustomerID IS NULL;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

COUNT(*)
0

```
SELECT COUNT(*) FROM online_retail WHERE Description IS NULL;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

COUNT(*)
0

2. Duplicates

```
SELECT COUNT(*) FROM online_retail;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

COUNT(*)
1067371

```
SELECT COUNT(*) FROM (
  SELECT DISTINCT * FROM online_retail
) AS unique_rows;
```

3. Invalid Values

```
SELECT COUNT(*) FROM online_retail WHERE Quantity <= 0;
```


```
SELECT COUNT(*) FROM online_retail WHERE UnitPrice <= 0;
```

Data Cleaning Steps

Step 1: Remove Exact Duplicate Rows

```
CREATE TABLE retail_cleaned AS
SELECT DISTINCT * FROM online_retail;
```

Step 2: Remove Rows With NULL CustomerID

```
DELETE FROM retail_cleaned WHERE CustomerID IS NULL;
```

Step 3: Remove Rows With Invalid Quantity or Unit Price

```
DELETE FROM retail_cleaned
WHERE Quantity <= 0 OR UnitPrice <= 0;
```

Step 4: Handle NULL Descriptions

```
UPDATE retail_cleaned
SET Description = 'Unknown Product'
WHERE Description IS NULL;
```

Step 5: Final Row Count After Cleaning

SELECT COUNT(*) FROM retail_cleaned;

COUNT(*)
1007896

Analytical Queries After Cleaning

Unique Customer Count

SELECT COUNT(DISTINCT CustomerID) FROM retail_cleaned;

COUNT(DISTINCT CustomerID)
5879

Top 10 Products by Sales Volume

SELECT Description, SUM(Quantity) AS Total_Quantity

FROM retail_cleaned

GROUP BY Description

ORDER BY Total_Quantity DESC

LIMIT 10;

Description	Total_Quantity
WORLD WAR 2 GLIDERS ASSTD DESIGNS	106139
WHITE HANGING HEART T-LIGHT HOLDER	94658
PAPER CRAFT , LITTLE BIRDIE	80995
ASSORTED COLOUR BIRD ORNAMENT	80082
MEDIUM CERAMIC TOP STORAGE JAR	78033
JUMBO BAG RED RETROSPOT	77699
BROCADE RING PURSE	70369
PACK OF 60 PINK PAISLEY CAKE CASES	56061
60 TEATIME FAIRY CAKE CASES	54028

Top Countries by Transactions

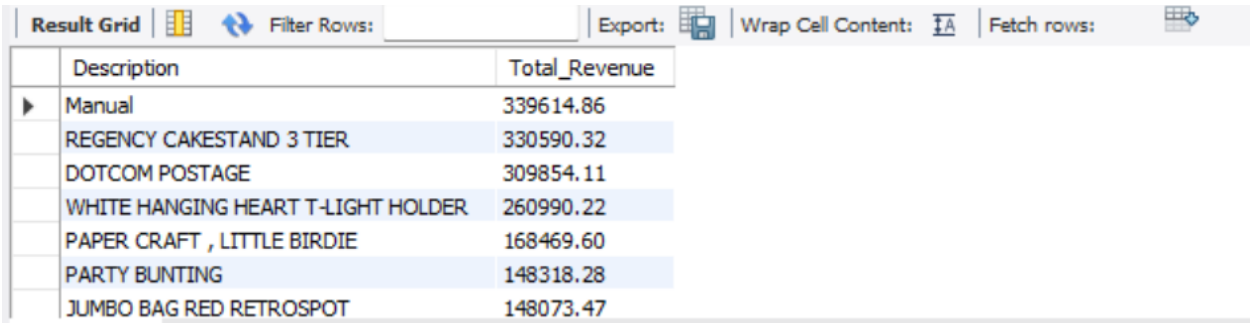
```
SELECT Country, COUNT(*) AS Num_Transactions
FROM retail_cleaned
GROUP BY Country
ORDER BY Num_Transactions DESC;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Country	Num_Transactions
United Kingdom	926022
EIRE	17154
Germany	16431
France	13639
Netherlands	5085
Spain	3662
Switzerland	3122

Most Profitable Products

```
SELECT Description, SUM(Quantity * UnitPrice) AS Total_Revenue
FROM retail_cleaned
GROUP BY Description
ORDER BY Total_Revenue DESC
LIMIT 10;
```



Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

Description	Total_Revenue
Manual	339614.86
REGENCY CAKESTAND 3 TIER	330590.32
DOTCOM POSTAGE	309854.11
WHITE HANGING HEART T-LIGHT HOLDER	260990.22
PAPER CRAFT , LITTLE BIRDIE	168469.60
PARTY BUNTING	148318.28
JUMBO BAG RED RETROSPOT	148073.47

Monthly Sales Trend

```
SELECT MONTH(InvoiceDate) AS Month, SUM(Quantity * UnitPrice) AS Monthly_Sales
FROM retail_cleaned
GROUP BY Month
ORDER BY Month;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Month	Total_Sales
1	1340966.72
2	1074423.85
3	1547130.52
4	1215843.74
5	1427002.11
6	1510084.32
7	1366886.39
8	1453093.29
9	1978132.18

```
SELECT COUNT(DISTINCT CustomerID) AS Unique_Customers FROM retail_cleaned;
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

COUNT(DISTINCT CustomerID)
5879

Summary:

- Original Rows: 1,067,371

- Final Cleaned Rows: 1,007,896

- Unique Customers: 5879

- Most Sold Product: 'WORLD WAR 2 GLIDERS ASSTD DESIGNS'

- Top Countries by Transaction: 'United Kingdom', 'EIRE', 'Germany'

- Highest Profitable Product: 'Manual'

- Month Wise Sale: 1340966.72, 1074423.85

Conclusion

This project helped clean over 1 million raw rows of transaction data and prepare them for meaningful analytics. The experience solidified foundational SQL skills and gave confidence to handle real-world datasets.