UK Retail Analytics

Why this data set - The UCI Online Retail dataset is famous for being messy and realistic — missing values, duplicates, and anomalies are common in retail transactions. This gives you a chance to showcase practical data cleaning and data wrangling

Project mindset → raw CSV → cleaning → EDA → insights → dashboards → recommendations.

Data Understanding

- **Number of rows**
- Total Transactions = COUNTROWS('Online Retail')
- Number of rows is 542,000
- **Number of columns** → 8 columns
- **Invoiceno**
  Data Type → Mixed data type (123 – Integer , ABC – Text)
  Purpose → Invoice no is a unique ID for each transaction , it helps track individual purpose
  Obsevation → mixed data type : Numerical (like 536365) & Alphanumeric
- **StockCode**
  Data Type → Mixed data type (123 – Integer , ABC – Text)
  StockCode represent the unique code for each products sold
  Used to identify products in the inventory and sales report
  Together with the Description it tells what the product is
- **Description**
  Data Type → Data type is Text
  This column describes the product or service sold in each transaction. It helps identify what item the StockCode refers to in plain language. Used for reporting, grouping, and product insights.
- **Quantity**
  Data Type → Data Type is whole number
  Purpose → Must be representning the number of units sold for each products in each transaction, used for calculating total sales, returns, and inventory tracking.
  Typical Values → Positive intergers (1, 2, 3, 4) for sales and negative integers (-24, -23, -12) indicates product returns or cancellations
- **InvoiceDate**
  Data Type is Date/Time Format
  Indicates when each invoice (transaction) was issues
  Cruital for time series analysis sales, seasonal patterns , customer purchase frequency and other time-based insights.
- **UnitPrice**
  Data Type → Decimal
  Quantity of the products sold
- **CustomerID**
- Identify the customer placing each order
- Helps group transaction, track repat buyers, segments by customer.
- Can have some missing values
- Country
- Indicates the customers country of purchase
- Geographic location better for geographical analysis , sales by region, and market segmentation.

Data Exploration

- **Invoiceno**

  Nulls → No null values in the data set
  Duplicates → Multiple rows per invoice valid (multiple – item orders)
  Alphanumeric pattern → C prefix means cancellation, new IsCancelled column added
  Action → will use IsCancelled for net sales analysis later.

- **Stockcode**

- Null → No null values in the data set
- StockCode contains both numeric and text-based codes.
- Special patterns
  1. Pure number → Standard products
  2. Single letters → Possible special item or service type
  3. POST → Postage free
  4. Gift_0001_50 → Gitcard or promotional items
  5. Mixed numeric + letter suffix → Product sub types or variants

- **Description**

  5 true nulls in Description
  ~1000+ rows have empty or whitespace-only descriptions
  2 Unique values
  Possible typos/junk not yet found — will check after cleaning blanks
  Inspected text patterns to identify invalid placeholders and inconsistent formats.
  Added DescriptionLength helper column to detect empty string and whitespaces only entries
  Filter and sorted the column to spot blaks, symbols, and junk text

- **Quantity**

- Null values → No null values in the column
- Basic Stats → Max → 216, Min → 24, Unique values → 14, Distinct values → 34, Average → 11.52, Standard Deviation → 19.28, Even → 784, Odd → 216
- No Zeros present in the column , confirms valid sales/returns data
- Most common quantaties – 12 units : 40 % of total , 1 unit 14% and 2 units 12%
- This shows a realistic mix single – itmes purchase and wholesales packs
- Even odd and even spread is normal
- No Outliers spike observed in the frequency distribution

- **InvoiceDate**

- **Null Values** →No null values
- **Data Type** → Date Time Format
- **Range** → Recordas are from - **01-12-2010** 08:26:00 to **09-12-2011** 12:50:00
- The Range coves 1 year of records
- Time info → Time portion varis across rows, Its meaningful and can be use for hourly trends, peak hours and details time based visuals.
- Duplicates → Not checked but will handle in Data Cleaning

- **UnitPrice**

- Null Values → No null values in the column
- Basic Stats Min → 0.72, Max → 20.79, Average → 1.89, Std.Dev → 2.622, Unique values → 231, Distinct Values → 23
- No Zeros and negative valid for sales data
- No outliers , highest value is 20.79 and most frequent price is 0.85 (35%) and 0.83 (14%) indicates valid bulk low-cost items.
- Data type → Decimal , Correct for currency values.
- Distribution → Price are mostly low clustered – typically for retail micro-transactions.

- **Customer ID**

- Identify the customer for each transaction – used to analyze repeat purchase and customer segmentation.
- Found 47 nulls – indicates some transaction have no customer id (typically for incomplete retail data)
- Basic stats → Min – 12576, Max – 18248, Distinct IDS – 153
- Repeats → Grouped Customer ID – Confirmed most customers are one time buyers, No unsual high repeat transactions, is nornal for retail data.
- All values numeric, no negative, zero or weired non-numeric entries.
- No unexpected gaps or strange patterns in ID range.
- Confirmed Data type is Whole number for clean joins and calculaions.

- **Country**

- Check for the missing values
- Found 231 missing values/null entries
- Important to adress in data cleaning
- Check for the consistency
- Sorted countries A-Z to scan for typos or inconsistent capitilixation
- All country names appeared consistent and properly capitalized
- Frequency Distribution
- Groped country and counted transactions
- United kingdom dominantes with 217 transactions
- Other country have very low transaction counts (btw 1-5) indicating limited salesoutside UK
- No suspicious or misspelled country names detected
- Confirmed data type is a set as Text data type

Data Cleaning

- **Invoiceno**

1. Some columns where having values as **C** in the starting so I have created a new column using the **First letter extraction** and then I have use the **Conditional Column** to check is the **FirstCharacter** is 'C' and flaged it as Yes and rest column as no.
2. IsCancelled → 5 – Yes and number - No

- **Stockcode**

1. Extracted first token via Split → Text Before Delimiter
2. Created helper StockCodeLength = length of that token
3. Added Conditional Column StockCodeType
   - Gift if token = "Gift"
   - Postage if token = "Post"
   - Variant if token length = 1
   - Product otherwise
4. Dropped helper columns—left with clean StockCodeType

- **Description**

1. Trimmed Values → Applied Format and trim to remove leading/ trailing spaces
2. Replace Empty String with Null → Used replace values to convert empty values to null
3. Remove Junk placeholders → Filtered and replace obvious placeholder texts (?, ??, ???, ????, ?missing, ?? missing, ???missing, ????missing with null
4. Standarized Casing → Applied Format and capatlized Each Word to unify all product names.
5. Filled Remaning Nulls → Applied a custom logic , if Description was nulll but Stockcode was available used Stockcode as a fallback and if both Descripton and stockcode were null replace with "Unknown Product"
6. Dropped Helper Column like descriptionlength and intermediate helper columns used for profiling

- Quantity

1. No invalid Quantity values found – all rows kept, Negative values are valid returns.
2. Quantity is converted to whole number
3. Created IsReturn Flag 'Yes' for negative and Quantity Returns 'No' for sales. Help filter and anlyse returns in visuals.

- **InvoiceDate**

1. Data type is converted to Date
2. Split the column to new column with names as Invoice_Date(Date only) and Invoice_Time (Invoice Time) to enable flexible time based analysis.
3. Removed duplicates rows based on Invoice_Date + Invoice_Time

- **Unit price**

1. Data Type confirmed as decimal number
2. Confirmed there is no nulls or blanks price in the column
3. Checked for the zeros using filter number and there is no negative or zerro price in the column
4. Rounded the decimal values upto 2 digits manly standaridizing the values
5. Recheck for the zeros and nagtive after rounding
6. Outlier review, maximum price was 20.79 ensures there is no unrelaistic prices found reasonable product range
7. Notes the majority of price are clustered bwteen 0.83 – 0.85 confirming consistent product pricing.

- **CustomerID**

1. Handled Missing values → Found 47 null values during data exploration and replace the numerical placeholder 99999 to represet unknown customer while keeping data type consistent
2. Confiremd Data type → Data type is now converted to Decimal , preventing any mixed types in joins or calculations
3. Sanity Check → Filtered to ensure all missing IDs are now corrected labelled 99999.
4. No unexpected null remained .

- **Country**

1. Handle missing values → Replace 231 nulls 'Unknow' to keep incomplete records traceble
2. Final Consistency Check → Verified all nulls correctly replaced no blanks left
3. Trimmed Extra Space → Removed accidental leading/ trailing spaces to ensure uniform value.