# UK Online Store Retail Transactions

Dataset Variable Information:

1. InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
2. StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
3. Description: Product (item) name. Nominal.
4. Quantity: The quantities of each product (item) per transaction. Numeric.
5. InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.
6. UnitPrice: Unit price. Numeric, Product price per unit in sterling.
7. CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
8. Country: Country name. Nominal, the name of the country where each customer resides.

## Establishing Python Library Packages

Show code

## Dataset Overview

Show code

| | InvoiceNo | StockCode | Description | Quantity | InvoiceDate | UnitPrice | CustomerID | Cou |
|---|---|---|---|---|---|---|---|---|
| **0** | 536365 | 85123A | WHITE HANGING HEART T-LIGHT HOLDER | 6 | 2010-12-01 08:26:00 | 2.55 | 17850.0 | U King |
| **1** | 536365 | 71053 | WHITE METAL LANTERN | 6 | 2010-12-01 08:26:00 | 3.39 | 17850.0 | U King |
| **2** | 536365 | 84406B | CREAM CUPID HEARTS | 8 | 2010-12-01 08:26:00 | 2.75 | 17850.0 | U King |

## Dataset Summary Overview

Show code

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  datetime64[ns]
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 33.1+ MB
```

## // Observations

Show code

Observation

- Dataset has 8 columns
- Max Row numbers: 541,909
- "Description" and "CustomerID" have lesser row count; possibly null values
- Description = 540,455 total rows
- CustomerID = 406,829 total rows

- "CustomerID" datatype is float64; convert into str object

# **CLEANING** | Null Values

## Counting nulls

Show code

| | 0 |
|---|---|
| **InvoiceNo** | 0 |
| **StockCode** | 0 |
| **Description** | 1454 |
| **Quantity** | 0 |
| **InvoiceDate** | 0 |
| **UnitPrice** | 0 |
| **CustomerID** | 135080 |
| **Country** | 0 |

**dtype:** int64

## .describe(): 'Description' overview

Show code

| | InvoiceNo | StockCode | Description | Country |
|---|---|---|---|---|
| **count** | 541909 | 541909 | 540455 | 541909 |
| **unique** | 25900 | 4070 | 4223 | 38 |
| **top** | 573585 | 85123A | WHITE HANGING HEART T-LIGHT HOLDER | United Kingdom |
| **freq** | 1114 | 2313 | 2369 | 495478 |

## /// Observations

Show code

Observation

where using 'StockCode' as identifier:

- 4070 unique rows on 'StockCode'
- 1454 null values on 'StockCode'
- 'StockCode' = 541,909 total row
- 'Description' = 540,455 total rows
- // most likely, 1,454 'StockCode' rows have no corresponding 'Description', (541,909 - 540,455)

## ~~ Investigate: [Description] Null Values

Show code

// Objective: generate a dataframe with 3 columns:

1. 'StockCode' = lists out unique rows
2. 'Count' = shows the number of occurrences of each unique 'StockCode'
3. 'Description' = provides the corresponding description for each 'StockCode'

// Method: create specific dataframes then concatenate ON unique 'StockCode'

.

## Create Dataframe: unique 'StockCode' & corresponding counts

Show code

## Create Dataframe: excluding 'Description' nulls on 'raw'

Show code

```
<class 'pandas.core.frame.DataFrame'>
Index: 540455 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    540455 non-null  object
 1   StockCode    540455 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     540455 non-null  int64
 4   InvoiceDate  540455 non-null  datetime64[ns]
 5   UnitPrice    540455 non-null  float64
 6   CustomerID   406829 non-null  float64
```

```
 7   Country      540455 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 37.1+ MB
```

## // Observations

Show code

Observation

- 540,455 max rows as per excluding NAs (original 541,909 max rows)
- 'StockCode' = 540,455 total rows (previously 541,909)
- 1,454 rows are 'Description' nulls as per calculation and section: counting nulls

## Create Dataframe: unique 'StockCode' & corresponding 'Description'

Show code

| StockCode | Description |
|---|---|
| **10002** | INFLATABLE POLITICAL GLOBE |
| **10080** | GROOVY CACTUS INFLATABLE |
| **10120** | DOGGY RUBBER |
| **10123C** | HEARTS WRAPPING TAPE |
| **10124A** | SPOTS ON RED BOOKCOVER TAPE |

**dtype:** object

## Concatenate Dataframe: unique StockCode + Counts + 'Description'

Show code

| | StockCode | Count | Description |
|---|---|---|---|
| **3536** | 85123A | 2313 | WHITE HANGING HEART T-LIGHT HOLDER |
| **1348** | 22423 | 2203 | REGENCY CAKESTAND 3 TIER |
| **3515** | 85099B | 2159 | JUMBO BAG RED RETROSPOT |
| **2733** | 47566 | 1727 | PARTY BUNTING |
| **180** | 20725 | 1639 | LUNCH BAG RED RETROSPOT |
| **...** | ... | ... | ... |
| **885** | 21854 | 0 | NaN |
| **886** | 21858 | 0 | NaN |
| **2786** | 62095B | 0 | NaN |
| **937** | 21923 | 0 | NaN |
| **2593** | 35951 | 0 | NaN |

4070 rows × 3 columns

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

## // Observations

Show code

Observation

- 4070 unique 'StockCode' values (consistent with section:.describe(): 'Description' overview)
- highest count at 2,313 = 'StockCode' 85123A, WHITE HANGING HEART T-LIGHT HOLDER

## Merge DataFrames: 'unique_stocks' and 'raw'

Show code

```
<class 'pandas.core.frame.DataFrame'>
Index: 541909 entries, 160128 to 40383
Data columns (total 10 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   InvoiceNo      541909 non-null  object
 1   StockCode      541909 non-null  object
 2   Description_x  540455 non-null  object
 3   Quantity       541909 non-null  int64
```

```
 4   InvoiceDate    541909 non-null  datetime64[ns]
 5   UnitPrice      541909 non-null  float64
 6   CustomerID     406829 non-null  float64
 7   Country        541909 non-null  object
 8   Count          541909 non-null  int64
 9   Description_y  541797 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(2), object(5)
memory usage: 45.5+ MB
```

## // Observations

Show code

Observation

- Description_x = 540,455 total rows (from 'raw)
- Description_y = 541, 797 total rows (from 'unique_stocks')
- CustomerID = 406, 829 total rows
- CustomerID datatype = float64 (must be converted into 'object')
- 541, 909 max total rows

```
# Refining updated dataframe
'''
> dropping 'Description_x' from 'raw'
> dropping 'Count' from 'unique_stocks'
> renaming 'Description_y
'''
merged_data = merged_data.drop(['Description_x', 'Count'], axis=1).rename(columns={'Descrip


# Renaming dataframe back as 'raw'
raw = merged_data.copy()
raw.head()
```

| | InvoiceNo | StockCode | Quantity | InvoiceDate | UnitPrice | CustomerID | Country | Descrip |
|---|---|---|---|---|---|---|---|---|
| **0** | 536365 | 85123A | 6 | 2010-12-01 08:26:00 | 2.55 | 17850.0 | United Kingdom | WH HANG HEAF LI HOL |
| **1** | 536365 | 71053 | 6 | 2010-12-01 08:26:00 | 3.39 | 17850.0 | United Kingdom | WH ME LANT |
| **2** | 536365 | 84406B | 8 | 2010-12-01 08:26:00 | 2.75 | 17850.0 | United Kingdom | CR CU HEA |

## Counting remaining nulls

Show code

| | 0 |
|---|---|
| **InvoiceNo** | 0 |
| **StockCode** | 0 |
| **Quantity** | 0 |
| **InvoiceDate** | 0 |
| **UnitPrice** | 0 |
| **CustomerID** | 135080 |
| **Country** | 0 |
| **Description** | 112 |

**dtype:** int64

## // Observations

Show code

Observation

- There are still 112 nulls on 'Description'
- 135,080 nulls on 'CustomerID'

## ~~Investigate: [Description] Remaining Null Values

Show code

// Objective: examine nature of nulls on [Description]; specifically, those that could pose as irrelevant rows for the sales transaction analysis

.

// Method: identify nature of 'Description' nulls accounting corresponding values on the following: (1) 'UnitPrice' (2) 'Quantity' (3) 'CustomerID'

```
zero_unitprice = raw[raw['UnitPrice'] == 0][['UnitPrice','Description', 'Quantity', 'Custom
null_zero_unitprice = zero_unitprice[zero_unitprice['Description'].isna()].sort_values(by =
null_zero_unitprice.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 112 entries, 1259 to 14
Data columns (total 5 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   index        112 non-null    int64
 1   UnitPrice    112 non-null    float64
 2   Description  0 non-null      object
 3   Quantity     112 non-null    int64
 4   CustomerID   0 non-null      float64
dtypes: float64(2), int64(2), object(1)
memory usage: 5.2+ KB
```

```
null_zero_unitprice.head()
```

| | index | UnitPrice | Description | Quantity | CustomerID |
|---|---|---|---|---|---|
| **1259** | 201756 | 0.0 | NaN | -102 | NaN |
| **327** | 55319 | 0.0 | NaN | -61 | NaN |
| **755** | 139064 | 0.0 | NaN | -45 | NaN |
| **312** | 52094 | 0.0 | NaN | -45 | NaN |
| **1617** | 280661 | 0.0 | NaN | -39 | NaN |

## // Observations

Show code

Observation

- 112 rows are zero in 'UnitPrice' and null in both 'Description' and 'CustomerID'
- Considering these 112 as irrelevent rows hence be removed

Start coding or generate with AI.

# **CLEANING** | Duplicated Rows

[ ] ↳ *2 cells hidden*

# **DATA TYPE** | Convertion

## Converting [CustomerID] float into object

Show code

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  541909 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  datetime64[ns]
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   541909 non-null  object
 7   Country      541909 non-null  object
dtypes: datetime64[ns](1), float64(1), int64(1), object(5)
memory usage: 33.1+ MB
```

## Counting Duplicate Rows

Show code

```
5268
```

## Removing Duplicate Rows

Show code

## Overviewing updated dataset

Show code

```
(536641, 8)
```

## info(): Updated DataFrame

Show code

```
<class 'pandas.core.frame.DataFrame'>
Index: 536641 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    536641 non-null  object
 1   StockCode    536641 non-null  object
 2   Description  536641 non-null  object
 3   Quantity     536641 non-null  int64
 4   InvoiceDate  536641 non-null  datetime64[ns]
 5   UnitPrice    536641 non-null  float64
 6   CustomerID   536641 non-null  object
 7   Country      536641 non-null  object
dtypes: datetime64[ns](1), float64(1), int64(1), object(5)
memory usage: 36.8+ MB
```

## shape(): Updated DataFrame

Show code

```
(536641, 8)
```

# OBSERVATIONS

Current Dataset Status:

- 536,641 = current max row range (from 541,909 rows)

Performed Data Manipulation:

- "Description" and "CustomerID" have null values: 1454, 135080 respectively; replaced with 'NA'
- "CustomerID" 'float64' data type converted into 'object'

- 5,268 duplicate rows were removed

## SUMMARY

- There are 8 Columns: InvoiceNo, StockCode, Description, Quantity, InvoiveDate, UnitPrice, CustomerID, Country
- Raw Max Rows = 541, 909
- Raw: Nulls on 'Description' = 1454
- Raw: Nulls on 'CustomerID' = 135, 080
- [Description] and [CustomerID] Nulls were replaced "NA"
- [Unit Price] negative (-) , at -11062.06 as min value
- [Quantity] negative (-) , at -80995.00 as min value
- [InvoiceNo] unique values at 25900 ~ invoice numbers are duplicated; there are transactions having the same invoice numbers
- [StockCode] unique values at 4070,
- [Description] unique values at 4224 -> indicates that there are stockcodes with varying derscriptions
- [CustomerID] unique values at 4373 ~ distinct customer identity; having total rows of 541,909, there are repeat customer including the 'null' 135,080 customers
- 10147 rows are duplicates of distinct rows
- 4879 distinct rows are duplicated nth times
- 5268 rows are exact duplicates of the 4879 distinct rows; HENCE, be removed. ~ where 5268 = 10147 - 4879

Notes:

1. [Unit Price] Why a negative (-) , at -11062.06 as min value?
2. [Quantity] Why a negative (-) , at -80995.00 as min value?
3. [InvoiceNo] unique values at 25900 ~ invoice numbers are duplicated; there are transactions having the same invoice numbers
4. [StockCode] unique values at 4070,
5. [Description] unique values at 4224 -> indicates that there are stockcodes with varying derscriptions
6. [CustomerID] unique values at 4373 ~ distinct customer identity; having total rows of 541,909, there are repeat customer including the 'null' 135,080 customers

---

# Working Dataset: Overview

## describe(): 'object'

Show code

| | InvoiceNo | StockCode | Description | CustomerID | Country |
|---|---|---|---|---|---|
| **count** | 536641 | 536641 | 536641 | 536641 | 536641 |
| **unique** | 25900 | 4070 | 4224 | 4373 | 38 |
| **top** | 573585 | 85123A | WHITE HANGING HEART T-LIGHT HOLDER | NA | United Kingdom |
| **freq** | 1114 | 2301 | 2357 | 135037 | 490300 |