**DATA PREPARATION AND HANDLING FINAL PROJECT**

**Group Members:**

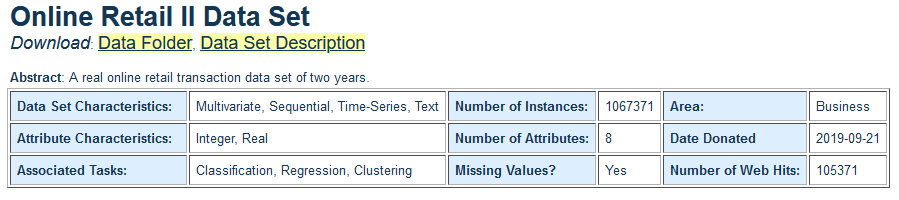
* Arvind Neelakantan (161021191)
* Parth Manish Shah (130344203)
* Rahil Wajidali Ansari (129983201)

**Dataset selected:** Online Retail II Dataset.

**Link:** [https://archive.ics.uci.edu/ml/datasets/Online+Retail+II#](https://archive.ics.uci.edu/ml/datasets/Online+Retail+II)

**Dataset Description:**

The dataset contains real online retail transaction data set of a UK-based, registered, non-store online retail.



**Attribute Information:**

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

**Preliminary Observations of the Dataset:**

1. Missing values:

* Country – represented as ‘Unspecified’.
* CustomerID – represented as blank.
* Description – represented as blank.

1. Outliers:

* Price – Negative values
* Quantity – Negative values

1. Errors in categorical variables:

* Invoice – Supposed to be a numerical column, but it contains non-numerical data for cancelled invoices (starts with the letter c). We could convert it to numerical and create a new sub category flag for cancelled invoices.
* StockCode – Supposed to be a numerical column, but it contains non-numerical data like ‘POST’ , ‘GIFT’ , ‘TEST001’ etc.