Section 1: **Online Retail II, RFM Analysis & Customer Segmentation**

GitHub (code):
https://github.com/wanadzhar913/rfm-analysis-using-online-retail2

Variable Name	Description
InvoiceNo	Invoice number. 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. 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. The day and time when a transaction was generated.
UnitPrice	Product price per unit in sterling (£).
CustomerID	Customer number. A 5-digit integral number uniquely assigned to each customer.
Country	Country name. The name of the country where a customer resides.

Table 1. Variables available in the dataset

An introduction to the data

No. of transactions: 1,067,371

Dates: 1/12/2009 - 09/12/2011

The Online Retail II data set contains all the transactions occurring for a UK-based, registered, non-store online retail company.

They mainly sell unique all-occasion gift-ware. Many customers of the company are wholesalers.

	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
470	489521	21646	NaN	-50	2009-12-01 11:44:00	0.00	NaN	United Kingdom
3114	489655	20683	NaN	-44	2009-12-01 17:26:00	0.00	NaN	United Kingdom
3161	489659	21350	NaN	230	2009-12-01 17:39:00	0.00	NaN	United Kingdom
3731	489781	84292	NaN	17	2009-12-02 11:45:00	0.00	NaN	United Kingdom
4296	489806	18010	NaN	-770	2009-12-02 12:42:00	0.00	NaN	United Kingdom
4566	489821	85049G	NaN	-240	2009-12-02 13:25:00	0.00	NaN	United Kingdom
6378	489882	35751C	NaN	12	2009-12-02 16:22:00	0.00	NaN	United Kingdom
6555	489898	79323G	NaN	954	2009-12-03 09:40:00	0.00	NaN	United Kingdom
6576	489901	21098	NaN	-200	2009-12-03 09:47:00	0.00	NaN	United Kingdom
6581	489903	21166	NaN	48	2009-12-03 09:57:00	0.00	NaN	United Kingdom
7204	490015	21982	NaN	467	2009-12-03 12:29:00	0.00	NaN	United Kingdom
7205	490016	21982	NaN	-1012	2009-12-03 12:30:00	0.00	NaN	United Kingdom
7559	490055	20620	NaN	-25	2009-12-03 13:22:00	0.00	NaN	United Kingdom
8553	490084	85064	NaN	-89	2009-12-03 15:37:00	0.00	NaN	United Kingdom
9249	490123	84508B	NaN	184	2009-12-03 18:08:00	0.00	NaN	United Kingdom

Table 2. Sample data with missing and irregular values

Due to the inconsistencies, I've decided to **drop these rows** so as to not affect the following analyses. Hence, **greater focus ought to be bought on data quality** to avoid from dropping rows due to incompleteness/incorrect data entry.

Additional analysis can also be done to identify **differences between transactions with a Customer ID versus those that don't.**

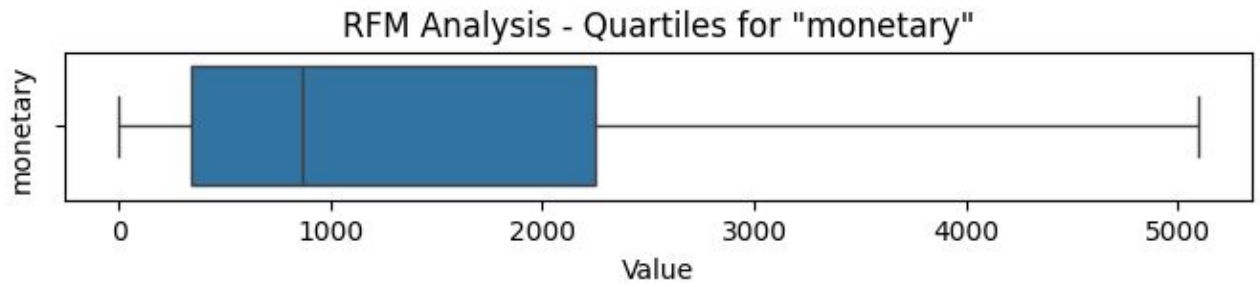
Challenges

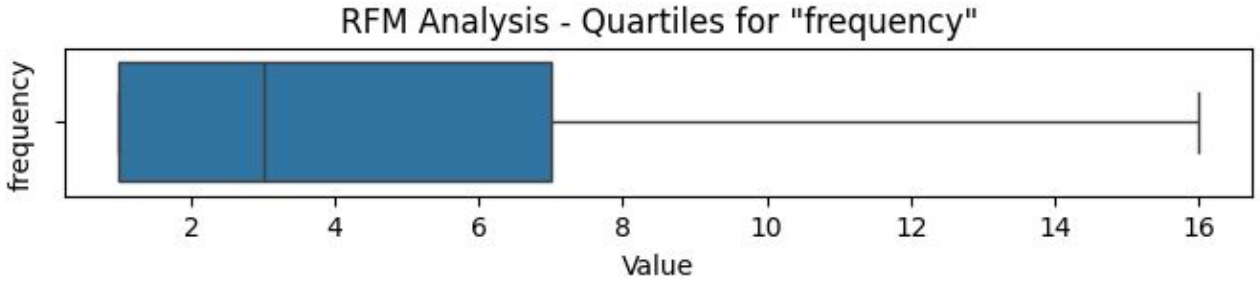
Missing values in the *Descriptions* & *Customer ID* column (likely due to Return/Cancelled items)

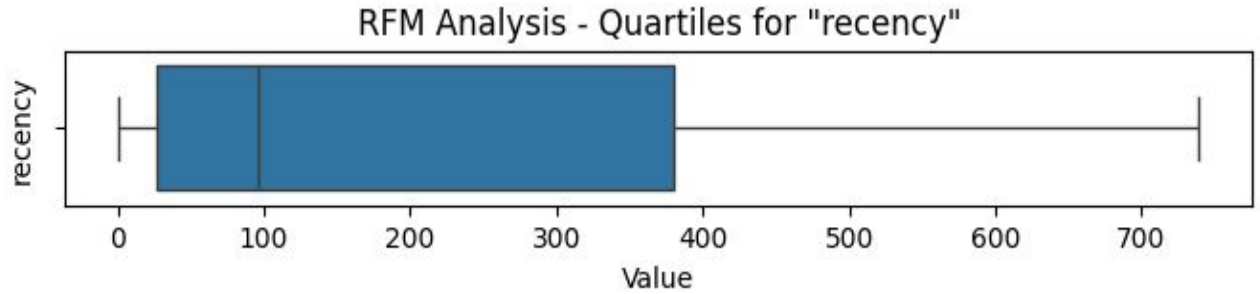
No or improper labelling for Return/Cancelled items/orders. This makes it hard for us to analyse the return/cancellation rate for or attribute them to system errors, etc.

22% of the data have missing values in the *Customer ID* column while it's 0.41% for *Description*

1.7% of rows in the *Quantity* column are negative while this is 2.5% of rows in the *Price* column.







Figures 1-3. Box-plots (with outliers removed) of the extracted features

RFM Analysis

To effectively group and segment customers according to their purchasing habits, we had to fashion the following variables:

Recency: The number of days since the last purchase for each customer.

Frequency: How often each customer makes a purchase.

Monetary: The total amount of money each customer has spent.

From the figures, **all the variables are positive-skewed,** so there's potentially a lot of outliers!

approach & outcome.

Preprocessing steps

1. **Remove rows where either *Quantity* or *Price* is negative or 0.**
2. **Drop rows where the *Description* & *Customer ID* is missing.**
3. Remove *StockCodes* with "TEST" in their text as a cautionary step.
4. Applied a Log-transform on the RFM dataset to **avoid results being skewed by outliers,** followed by a Standard Scaler to ensure dataset is on a similar scale.
5. Determined the required number of clusters (4) using the Elbow method & Silhouette score.

Outcome

cluster	recency	monetary	frequency
0	234.98	1915.75	4.97
1	27.37	10464.66	18.89
2	28.09	837.08	3.07
3	389.99	311.87	1.31

Table 3. Mean values of Recency, Monetary & Frequency in each cluster

clusters breakdowns & marketing suggestions.

Cluster 1 is the best cluster although it is the 2nd smallest. This group spends more money and made the most transactions, and in average had their last transactions 27 days ago. **These are the regular customers** so **loyalty programmes & brand awareness ads that encourage long-term patronage** is a suitable campaign for them.

Cluster 0 is the next best cluster, because they make the **2nd most amount of money.** However the **customers in that group haven't come for a while**. Hence, more targeted marketing efforts focused on **building brand awareness and engagement is the recommendation for the long term.** Over time, promotions e.g., free shipping or discounts can eventually be made to induce loyalty and patronage.

Cluster 2 can be characterised as those that **spend less money, but order relatively frequently** (as their last transactions a month ago). Much like Cluster 1, loyalty programmes & brand awareness ads that encourage long-term patronage is recommended. Additionally, **promotions and discounts are likely to work well for this group to encourage them to encourage them to spend more seasonally.**

Customers in Cluster 3 have likely churned as they haven't spent much and haven't stopped by in a long time. This is **worrisome as they make up the largest cluster.**