Report

By Seemant Kaushal

Context About the data set

"This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail .The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers."

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom

Brief Description about the Data

Variable	Description
InvoiceNo	Code representing each unique transaction. If this code starts with letter 'c', it indicates a cancellation.
StockCode	Code uniquely assigned to each distinct product.
Description	Description of each product.
Quantity	The number of units of a product in a transaction.
InvoiceDate	The date and time of the transaction.
UnitPrice	The unit price of the product in sterling.
CustomerID	Identifier uniquely assigned to each customer.
Country	The country of the customer.

STATISTICS Summary (Numerical Data)

```
df.describe()
```

	Quantity	UnitPrice	CustomerID
count	541909.000000	541909.000000	406829.000000
mean	9.552250	4.611114	15287.690570
std	218.081158	96.759853	1713.600303
min	-80995.000000	-11062.060000	12346.000000
25%	1.000000	1.250000	13953.000000
50%	3.000000	2.080000	15152.000000
75%	10.000000	4.130000	16791.000000
max	80995.000000	38970.000000	18287.000000

INFERENCE ABOUT THE DATASET

The dataset contains **541,909** transactions with an average of **9.55** units sold per transaction, although there is high variability as indicated by a standard deviation of **218.08** units. The minimum value of **-80,995** units suggests there may be errors in the data, possibly due to returns or incorrect entries. The median number of units sold per transaction is 3, with the maximum reaching up to **80,995** units, which appears to be an outlier.

For unit prices, the dataset also includes 541,909 transactions with an average price of **$4.61** per unit. Similar to the quantity, there is significant variability in unit prices, with a standard deviation of $96.76. The minimum unit price recorded is -$11,062.06, again indicating potential data errors. The median unit price is $2.08, while the maximum price is $38,970, another possible outlier.

There are 406,829 unique customers in the dataset, with customer IDs averaging around 15,287.69 and a standard deviation of 1,713.60. Customer IDs range from 12,346 to 18,287.

CATEGORICAL DATA

	InvoiceNo	StockCode	Description	InvoiceDate	Country
count	541909	541909	540455	541909	541909
unique	25900	4070	4223	23260	38
top	573585	85123A	WHITE HANGING HEART T-LIGHT HOLDER	10/31/2011 14:41	United Kingdom
freq	1114	2313	2369	1114	495478

Description:

There are **4223** unique product descriptions.
The most frequent product description is "**WHITE HANGING HEART T-LIGHT HOLDER**", appearing **2369** times.
There are some missing values in this column which need to be treated.

Country:

The transactions come from **38** different countries, with a dominant majority of the transactions - (approximately 91.4%) originating from the United Kingdom.

StockCode:

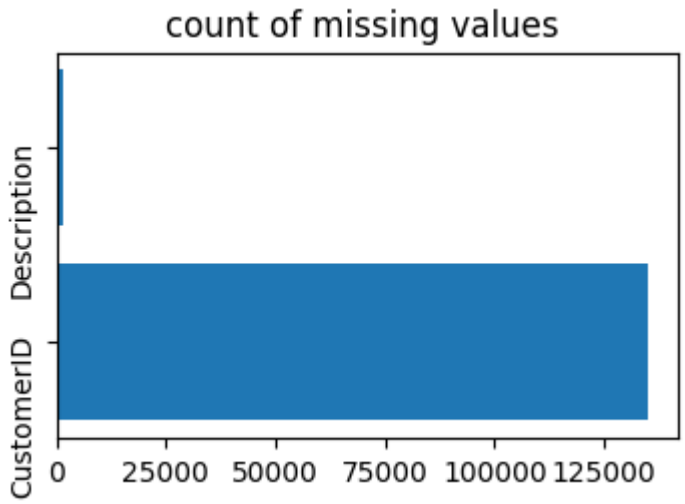
There are **4070** unique stock codes representing different products.
The most frequent stock code is **85123A**, appearing **2313** times in the dataset.

InvoiceNo:

There are 25900 unique invoice numbers, indicating **25900** separate transactions.
The most frequent invoice number is **573585**, appearing **1114** times, possibly representing a large transaction or an order with multiple items.

HANDLING MISSING VALUES

```
InvoiceNo            0
StockCode            0
Description       1454
Quantity             0
InvoiceDate          0
UnitPrice            0
CustomerID      135080
Country              0
order_amt            0
year                 0
month                0
Day                  0
Time                 0
dtype: int64
```



The CustomerID column contains nearly a **quarter** of missing data. Imputing such a large percentage of missing values might introduce significant bias or noise into the analysis.

The Description column has a **minor** percentage of missing values. However, it has been noticed that there are inconsistencies in the data where the same StockCode does not always have the same Description. This indicates data quality issues and potential errors in the product descriptions.

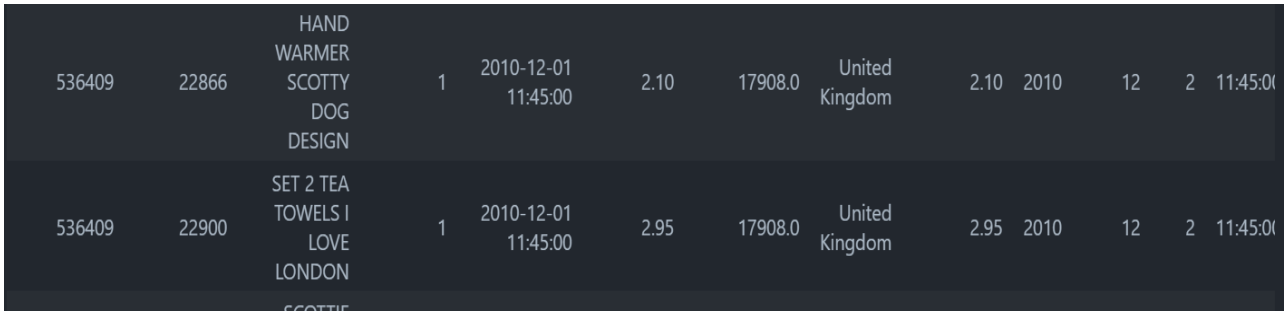
The dataset includes negative values in the quantity and unit price columns, indicating the presence of return transactions and potential data errors. The high variability and presence of extreme values suggest that data cleaning and further investigation are necessary to handle these issues appropriately.

So we created a new column as Transaction_Status, which have values as 'completed 'and 'Cancelled'

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	order_amt
141	C536379	D	Discount	-1	2010-12-01 09:41:00	27.50	14527.0	United Kingdom	-27.50
154	C536383	35004C	SET OF 3 COLOURED FLYING DUCKS	-1	2010-12-01 09:49:00	4.65	15311.0	United Kingdom	-4.65
235	C536391	22556	PLASTERS IN TIN CIRCUS PARADE	-12	2010-12-01 10:24:00	1.65	17548.0	United Kingdom	-19.80

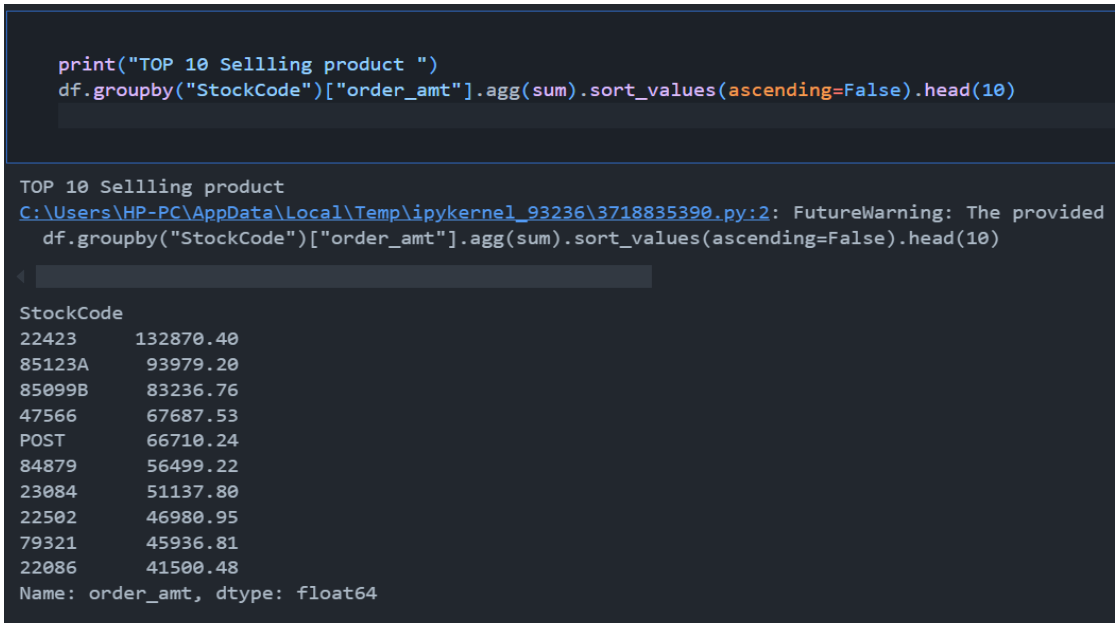
DUPLICATES RECODES

We observe that the dataset contains multiple duplicate records. These duplicates could be due to various reasons, such as data entry errors or system glitches, rather than the same order being placed twice. The presence of these duplicate records can significantly impact our analysis, particularly when detecting customer purchasing behaviour.



536409	22866	HAND WARMER SCOTTY DOG DESIGN	1	2010-12-01 11:45:00	2.10	17908.0	United Kingdom	2.10	2010	12	2	11:45:0(
536409	22900	SET 2 TEA TOWELS I LOVE LONDON	1	2010-12-01 11:45:00	2.95	17908.0	United Kingdom	2.95	2010	12	2	11:45:0(

IDENTIFY THE TOP 10 BEST-SELLING PRODUCTS BY REVENUE.



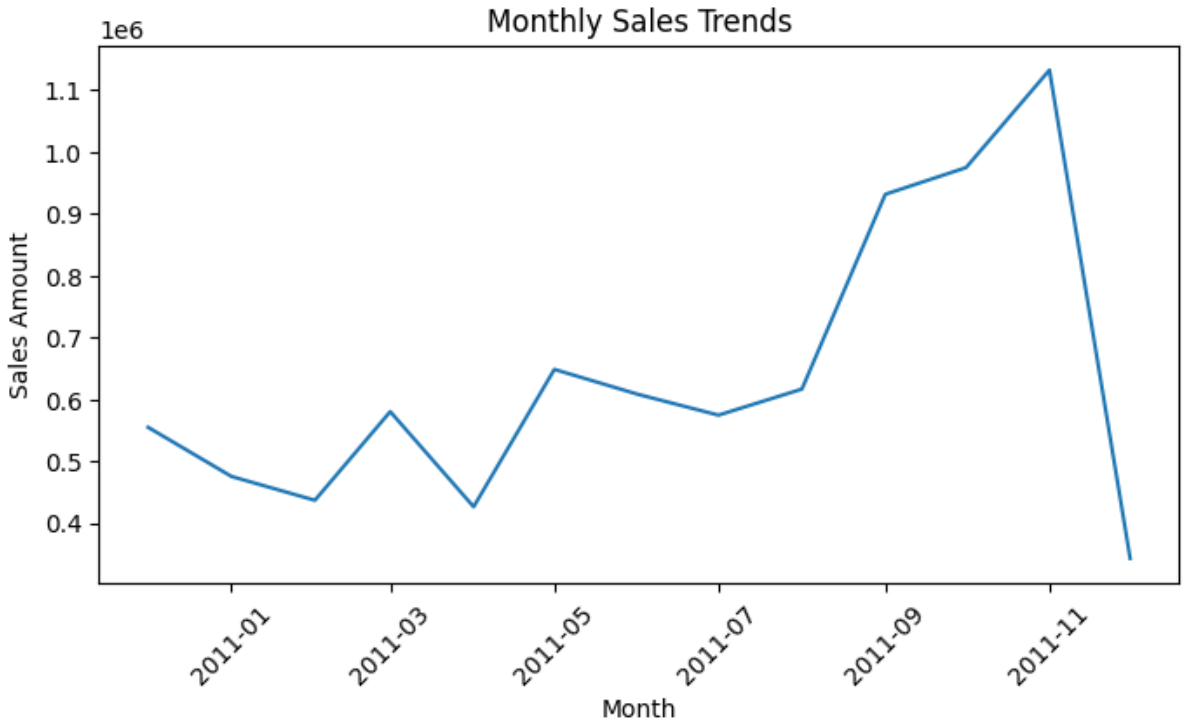
```
print("TOP 10 Sellling product ")
df.groupby("StockCode")["order_amt"].agg(sum).sort_values(ascending=False).head(10)
```

```
TOP 10 Sellling product
C:\Users\HP-PC\AppData\Local\Temp\ipykernel_93236\3718835390.py:2: FutureWarning: The provided
  df.groupby("StockCode")["order_amt"].agg(sum).sort_values(ascending=False).head(10)

StockCode
22423     132870.40
85123A     93979.20
85099B     83236.76
47566      67687.53
POST       66710.24
84879      56499.22
23084      51137.80
22502      46980.95
79321      45936.81
22086      41500.48
Name: order_amt, dtype: float64
```

Monthly Sales trends

We have observed the month wise Sales of the company, which can see from the below line graph for the different Months.



Customer Segmentation based on the Mean and standard Deviation Spending

We have created the following segmentation based on the value of mean and standard deviation.

- **Loyal Premium Customers**: Customers with high mean spending and low variability in spending, indicating they consistently spend a lot.
- **Unpredictable Bargain Shoppers**: Customers with low mean spending but high variability, indicating they tend to make irregular purchases at lower price points.
- **High Value but Volatile Customers**: Customers who spend a lot but have high variability, indicating that their spending is unpredictable. They may respond to targeted promotions.
- **Low Risk Bargain Shoppers**: Customers with low spending and low variability, indicating consistent but low-value purchases.
- **Other**: Customers who don't fit neatly into the defined segments.

	CustomerID	Segment
0	12346.0	Low Risk Bargain Shoppers
1	12347.0	Other
2	12348.0	High Value but Volatile Customers
3	12349.0	Other
4	12350.0	Other
5	12352.0	Other
6	12353.0	Other
7	12354.0	Other
8	12355.0	Loyal Premium Customers
9	12356.0	High Value but Volatile Customers

Using this analysis we can get the following insights

- **Loyal Premium Customers**: Offer exclusive deals, loyalty programs, or premium products.
- **Unpredictable Bargain Shoppers**: Engage them with promotions and discounts to increase frequency.
- **High Value but Volatile Customers:** Use targeted marketing campaigns to stabilize their spending.
- **Low Risk Bargain Shoppers**: Encourage larger purchases or bundle offers to increase average order value.

THE MOST VALUABLE CUSTOMERS (TOP 10% OF CUSTOMERS BY TOTAL SPEND).

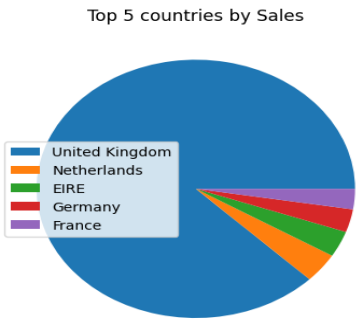
Monthly_Spending_Mean: *This is the average amount a customer spends monthly. It helps us gauge the general spending habit of each customer. A higher mean indicates a customer who spends more, potentially showing interest in premium products, whereas a lower mean might indicate a more budget-conscious customer*

	CustomerID	Total_spends
1703	14646.0	279489.02
4233	18102.0	256438.49
3758	17450.0	187482.17
1895	14911.0	132572.62
55	12415.0	123725.45
1345	14156.0	113384.14
3801	17511.0	88125.38
3202	16684.0	65892.08
1005	13694.0	62653.10
2192	15311.0	59419.34

GRAPHICAL ANALYSIS:

Country Wise Total Sales: This shows the sales of product from the different geographical Area, This feature identifies the country where each customer is located. Including the country data can help us understand region-specific buying patterns and preferences. Different regions might have varying preferences and purchasing behaviours which can be critical in personalizing marketing strategies and inventory planning. Furthermore, it can be instrumental in logistics and supply chain optimization, particularly for an online retailer where shipping and delivery play a significant role.

	Country	Total sales
35	United Kingdom	6767873.394
23	Netherlands	284661.540
10	EIRE	250285.220
14	Germany	221698.210
13	France	196712.840



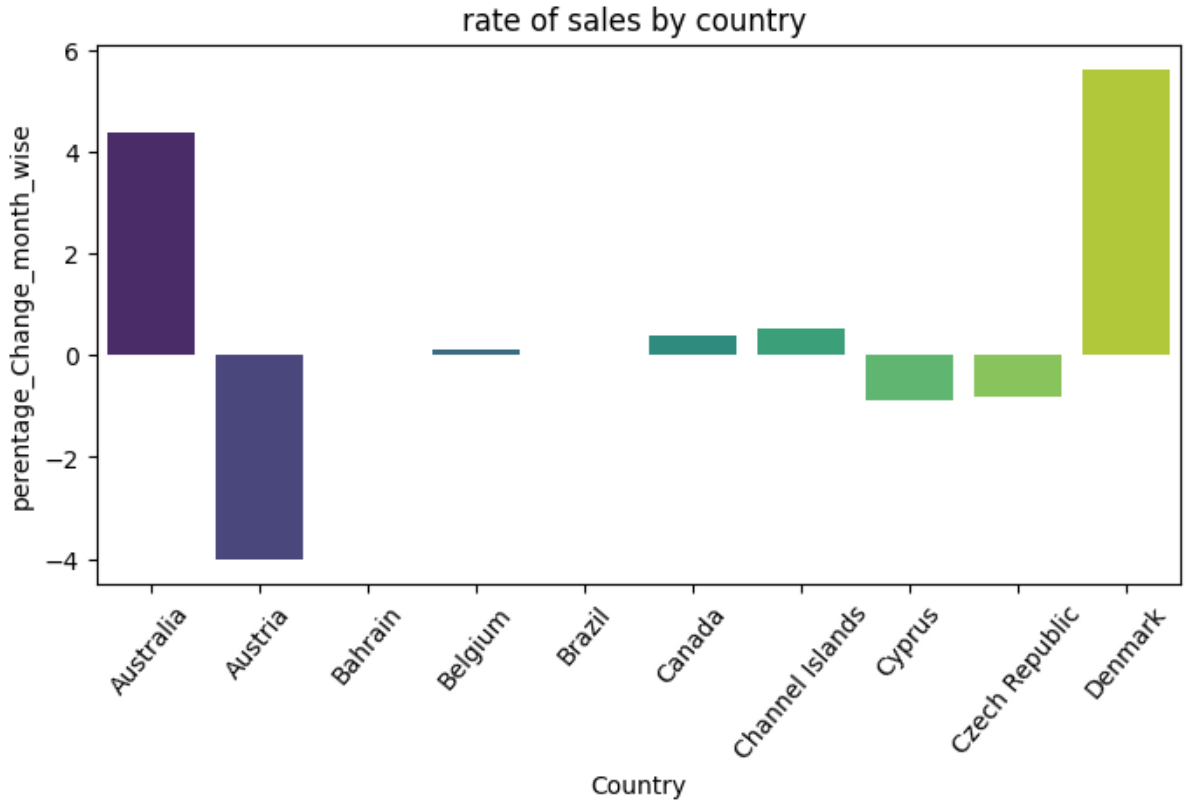
Country	
United Kingdom	0.889509
Germany	0.023339
France	0.020871
EIRE	0.018398
Spain	0.006226
Name: proportion, dtype: float64	

• **Dominant Market**: The **United Kingdom** is the clear leader in sales, contributing nearly **89%** of the total sales proportion. This suggests that a significant amount of business comes from this market.

• **Minor Markets**: The other countries (Germany, France, Ireland, and Spain) contribute relatively small proportions to the total sales. Together, they account for just about **11%** of the total sales.

• **Growth Opportunities**: Given the small proportions from Germany, France, Ireland, and Spain, there may be opportunities for growth and increased marketing efforts in these markets to boost sales.

COUNTRIES WITH SIGNIFICANT SALES GROWTH OR DECLINE



Positive Growth Markets:

- Denmark has the highest percentage change at +5.62%, suggesting a strong upward trend in sales. This may indicate effective marketing, successful product launches, or favorable market conditions.
- Australia also shows a robust growth rate of +4.37%, indicating healthy sales performance in this market.

Mild Growth:

- Several countries, including Belgium (+0.12%), Canada (+0.39%), and Channel Islands (+0.52%), are experiencing slight positive growth, indicating stable sales performance but with limited growth potential.

Declining Markets:

- Austria shows a notable decline at -4.00%, which may require immediate attention to understand the causes (e.g., increased competition, market saturation).
- Other countries with negative changes include Cyprus (-0.88%) and Czech Republic (-0.82%), indicating challenges in these markets that may need strategic interventions to address.

No Change:

- Bahrain and Brazil both show 0.00% change, indicating stability but also a lack of growth. This could be a signal to explore new strategies to stimulate sales.

Potential Recommendations

Focus on High-Growth Markets: Capitalize on the positive trends in Denmark and Australia. Consider increasing marketing budgets or resources in these areas.

Address Declining Sales: For countries experiencing negative changes, such as Austria, Cyprus, and Czech Republic, analyse market conditions and customer feedback to identify the underlying issues. Implement strategies to regain customer interest or address competition.

Stabilization Strategies: For markets showing no growth, like Bahrain and Brazil, consider innovation in product offerings or targeted promotions to stimulate sales.

Machine Learning Model for the prediction of purchase in Next month by customer

Model Overview

This report outlines the performance of a logistic regression model developed to predict customer purchases based on the following features: **Recency**, **Frequency**, and **Monetary Value**. The target variable represents whether a customer will make a purchase in the next month.

Model Performance Summary

Accuracy

- **Overall Accuracy**: The model achieved an accuracy of **99.31%**, indicating that it correctly predicted customer purchases in approximately 993 out of every 1000 cases. This demonstrates high performance

Confusion Matrix

The confusion matrix provides insight into the model's predictions:

```
[[1103    9]
[   0  200]]
```

- **True Negatives (TN)**: 1103 – Correctly predicted non-purchases.
- **False Positives (FP)**: 9 – Incorrectly predicted purchases.
- **False Negatives (FN)**: 0 – Failed to predict any purchases that actually occurred.
- **True Positives (TP)**: 200 – Correctly predicted purchases.

Classification Report:

	precision	recall	f1-score	support
0.0	1.00	0.99	1.00	1112
1.0	0.96	1.00	0.98	200
accuracy			0.99	1312
macro avg	0.98	1.00	0.99	1312
weighted avg	0.99	0.99	0.99	1312

Recommendations:

Based on your analysis, provide three actionable recommendations for improving sales performance.

1: To decrease the return Items this will definitely help in increase in the rate of revenue of the company

2: Focus on bestselling items, Increase the stock availability of the Top 10 selling products, in order to maximize the revenue.

3: To add discounts over the products which are having less sales. This will give the boost to the overall sales of the company.

4: Try to enhance the sales in the season, Provides offer in season to increase the sales of products.