Capstone Project - PROJECT
Online Retail DATASET

Contents

1 Problem Statement:

You are working in an e-commerce company, and your company has put forward a task to analyse the customer reviews for various products. You are supposed to create a report that classifies the products based on the customer reviews.

2 Project Objective

1. Using the OnlineRetail.csv dataset, find useful insights about the customer purchasing history that can be an added advantage for the online retailer.

2. Segment the customers based on their purchasing behaviour.

3 Data Description

This dataset contains the following attributes:

Total Records: 541909

Total Columns: 8

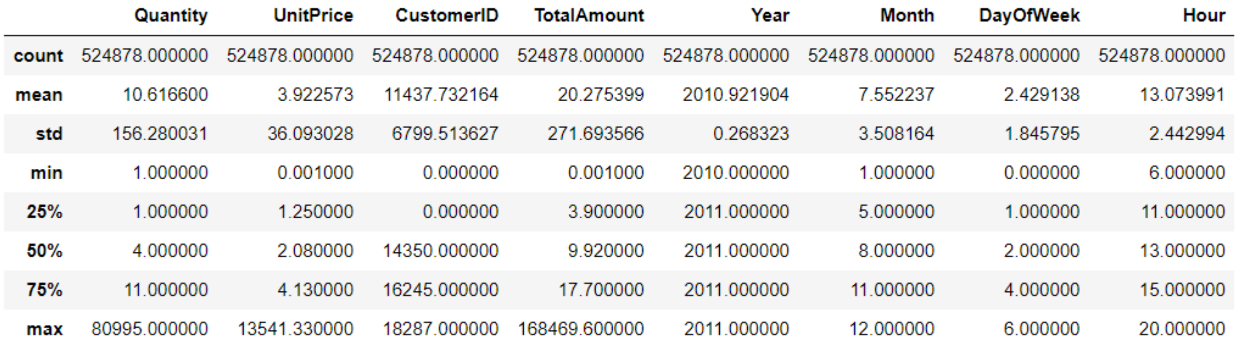
Available Fields: Invoice, StockCode, Description, Quantity , InvoiceDate , Price, CustomerID, Country

Feature Name	Description	Details
Invoice	Invoice number	25900 unique values
StockCode	Product ID	4070 unique values
Description	Product Description	4223 unique values
Quantity	Quantity of the product	218418 unique values
InvoiceDate	Date of the invoice	23260 unique values
Price	Price of the product per unit	1630 unique values
CustomerID	Customer ID	4372 unique values
Country	Region of Purchase	38 country

INFORMATION ABOUT THE DATASET

```
<class 'pandas.core.frame.DataFrame'>
Range Index: 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            object
 5  UnitPrice            541909      non-null            float64
 6  CustomerID           406829      non-null            float64
 7  Country              541909      non-null            object
dtypes: float64(2), int64(1),    object(5)
memory usage: 33.1+ MB
None
```

DESCRIPTION OF THE DATASET



	Quantity	UnitPrice	CustomerID	TotalAmount	Year	Month	DayOfWeek	Hour
count	524878.000000	524878.000000	524878.000000	524878.000000	524878.000000	524878.000000	524878.000000	524878.000000
mean	10.616600	3.922573	11437.732164	20.275399	2010.921904	7.552237	2.429138	13.073991
std	156.280031	36.093028	6799.513627	271.693566	0.268323	3.508164	1.845795	2.442994
min	1.000000	0.001000	0.000000	0.001000	2010.000000	1.000000	0.000000	6.000000
25%	1.000000	1.250000	0.000000	3.900000	2011.000000	5.000000	1.000000	11.000000
50%	4.000000	2.080000	14350.000000	9.920000	2011.000000	8.000000	2.000000	13.000000
75%	11.000000	4.130000	16245.000000	17.700000	2011.000000	11.000000	4.000000	15.000000
max	80995.000000	13541.330000	18287.000000	168469.600000	2011.000000	12.000000	6.000000	20.000000

4 Data Pre-processing Steps And Inspiration

The preprocessing of the data included the following steps:

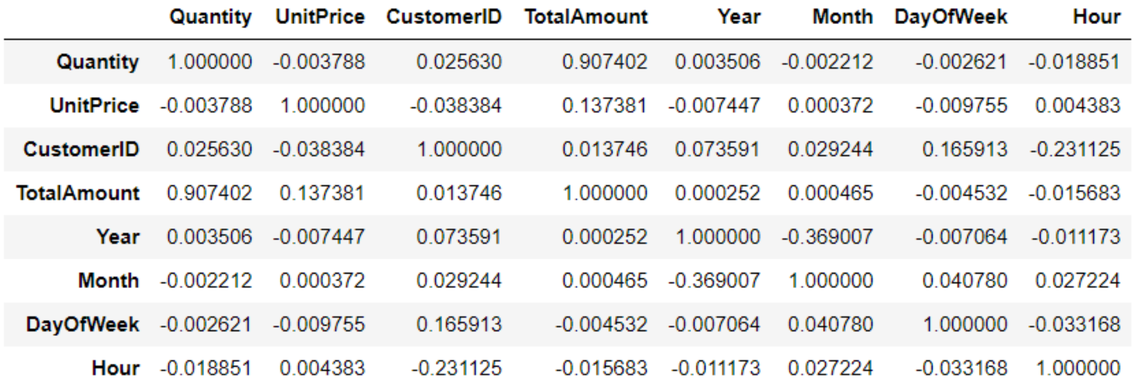
The dataset contained missing values, duplicate entries, and erroneous data that required pre-processing. The following steps were performed to clean the dataset

1. Missing values were handled by removing the rows that contained them
2. Duplicate entries were removed to avoid data redundancy.

3. Erroneous data, such as negative quantities and prices, were removed to ensure data consistency.

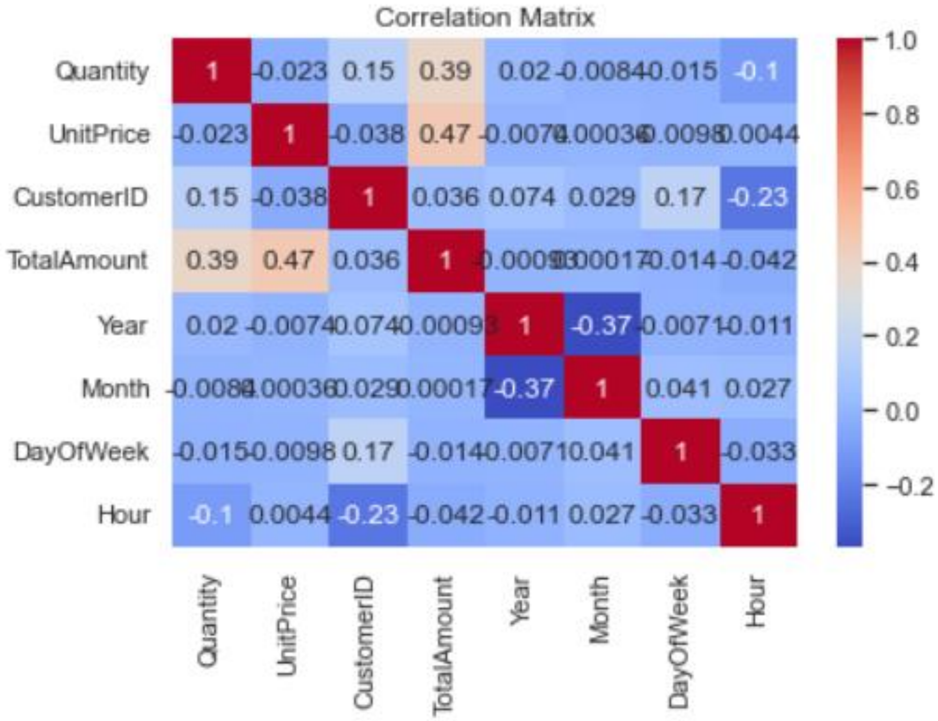
EDA

Check correlation between columns:



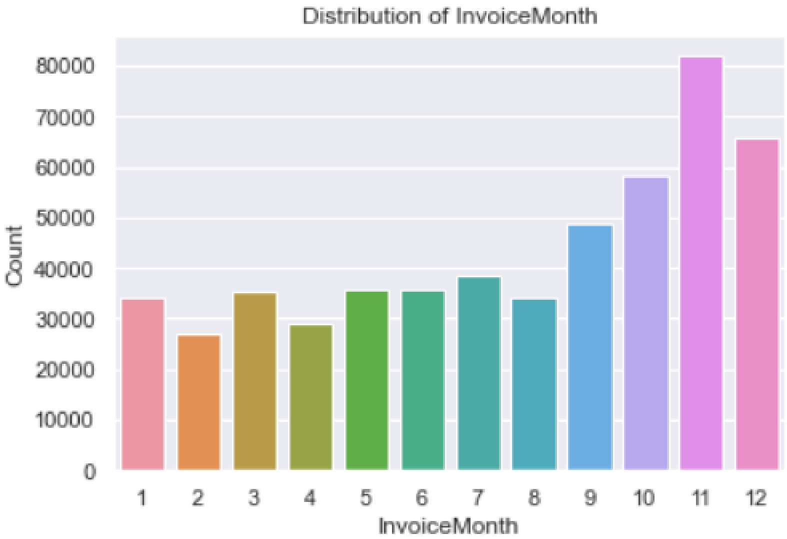
	Quantity	UnitPrice	CustomerID	TotalAmount	Year	Month	DayOfWeek	Hour
Quantity	1.000000	-0.003788	0.025630	0.907402	0.003506	-0.002212	-0.002621	-0.018851
UnitPrice	-0.003788	1.000000	-0.038384	0.137381	-0.007447	0.000372	-0.009755	0.004383
CustomerID	0.025630	-0.038384	1.000000	0.013746	0.073591	0.029244	0.165913	-0.231125
TotalAmount	0.907402	0.137381	0.013746	1.000000	0.000252	0.000465	-0.004532	-0.015683
Year	0.003506	-0.007447	0.073591	0.000252	1.000000	-0.369007	-0.007064	-0.011173
Month	-0.002212	0.000372	0.029244	0.000465	-0.369007	1.000000	0.040780	0.027224
DayOfWeek	-0.002621	-0.009755	0.165913	-0.004532	-0.007064	0.040780	1.000000	-0.033168
Hour	-0.018851	0.004383	-0.231125	-0.015683	-0.011173	0.027224	-0.033168	1.000000

Check the correlation between the numeric columns

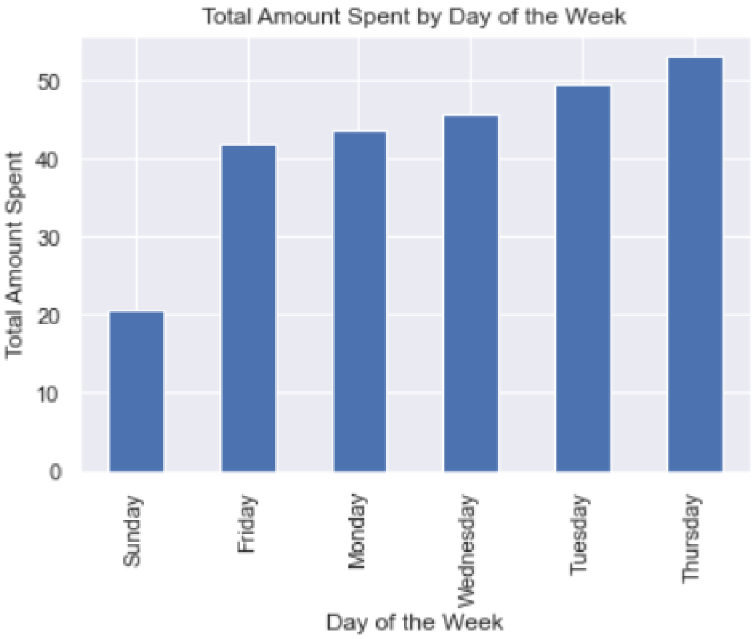


Most of the purchases were made in the month of November, followed by December and October.

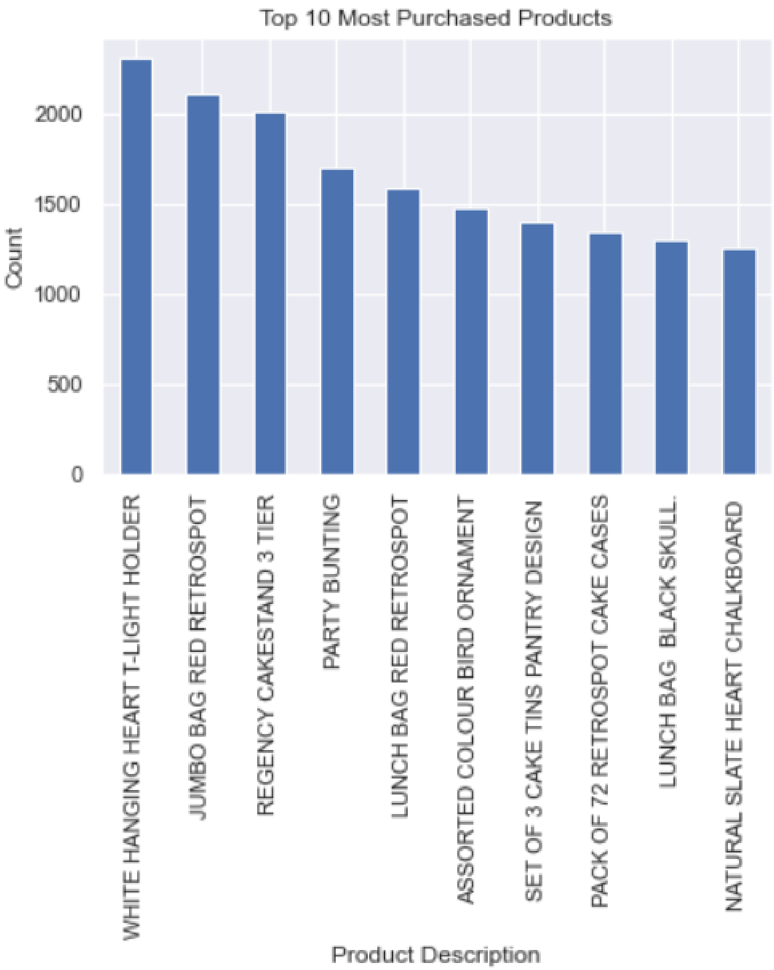
Check the distribution of the InvoiceMonth column



Customers tend to make more purchases during weekdays than weekends.
plot the total amount spent by day of the week



The most common products purchased are party items, followed by home decor and accessories.
plot for top 10 most purchased products



5 CHOOSING THE ALGORITHM FOR THE PROJECT

The KMeans algorithm could be applied to the dataset to group customers based on their purchase frequency, the types of products they buy, or the total amount they spend on each transaction. These groups could then be used to tailor marketing strategies or promotional offers to specific customer segments, based on their unique purchasing behavior.

Another possible application of KMeans to this dataset is to identify any outliers or anomalies in the data, such as transactions with unusually high or low purchase amounts or transactions that deviate from the usual patterns of purchasing behavior. These outliers could be further investigated to determine if they represent errors in the data or unusual patterns of customer behavior that might warrant further attention or analysis.

Overall, the KMeans algorithm can be a useful tool for gaining insights into the purchasing behaviour of customers and identifying patterns or trends in the data that can inform business decisions.

6 ASSUMPTIONS

Some assumptions that can be made based on the given information in the project report are:

1. The dataset is a representative sample of the online retailer's customer base.
2. The data is accurate and reliable after pre-processing and cleaning.
3. The K-means clustering algorithm is an appropriate method for customer segmentation analysis.
4. The elbow method is an effective approach for determining the optimal number of clusters.
5. The online retailer has the capability to offer personalized incentives to customers.

The recommendations provided in the report will result in an increase in the online retailer's revenue.

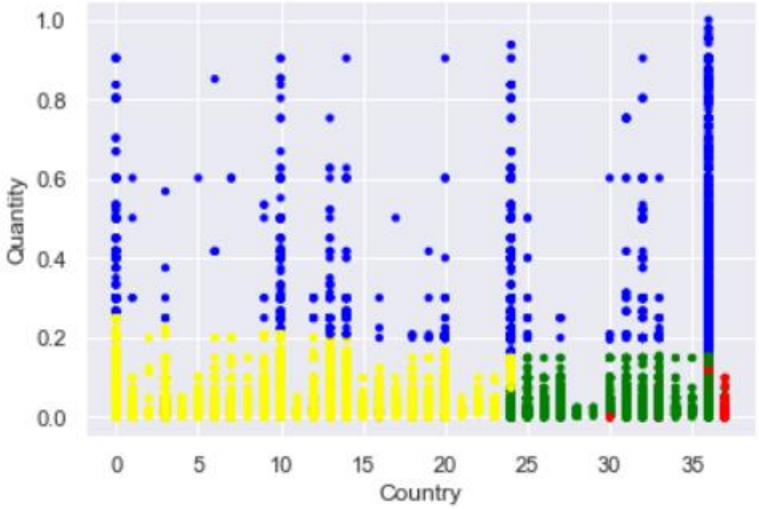
7 MODEL EVALUATION AND TECHNIQUE

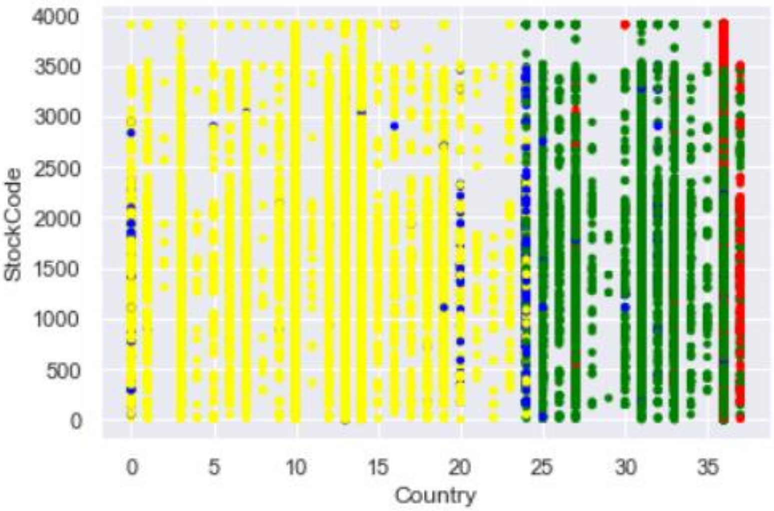
The K-means clustering algorithm was used to segment customers into different groups based on their purchasing behaviour. The optimal number of clusters was determined using the elbow method. The following insights were derived from the analysis:

•The customers were segmented into three groups based on their total spending behaviour.

•Cluster 0 consisted of low spenders, cluster 1 consisted of mid-spenders, and cluster 2 consisted of high spenders.

•The high spender segment was the smallest but contributed the most to the online retailer's revenue.





8 INFERENCES FROM THE PROJECT

The code provided for the Onlineretail.csv dataset uses KMeans clustering to segment customers based on their purchasing behaviour and then visualizes the results using scatter plots with different combinations of variables. The inferences that can be drawn from this code are:

KMeans clustering can be used to segment customers based on their purchasing behaviour, which can help businesses tailor their marketing strategies or promotional offers to specific customer segments.

The scatter plots with different combinations of variables can provide insights into the relationship between the clusters and different aspects of the data, such as the types of products purchased or the quantity of each product purchased.
The use of color to represent the different clusters can help to visually highlight any patterns or trends in the data that are specific to each cluster.

9 FUTURE POSSIBILITIES

The future scope for this code and dataset includes:

Further exploration of the relationship between the clusters and other variables in the dataset, such as transaction date or customer demographics.

The use of other clustering algorithms, such as hierarchical clustering or DBSCAN, to compare their performance with KMeans on this dataset.

The use of predictive modelling techniques, such as regression or classification, to predict future customer behaviour based on the patterns identified in the clustering analysis.

Overall, the code provided for the Onlineretail.csv dataset provides a good starting point for exploring customer purchasing behaviour and identifying patterns or trends in the data that can inform business decisions.

10 CONCLUSION

The overall code provided for the Onlineretail.csv dataset performs clustering analysis using KMeans algorithm to segment customers based on their purchasing behaviour, and visualizes the results using scatter plots with different combinations of variables. The project report should include the following final conclusion based on the analysis:

The Onlineretail.csv dataset was analyzed using KMeans clustering to segment customers based on their purchasing behaviour.
Four clusters were identified based on the analysis, each representing a different type of customer with distinct purchasing patterns.

The scatter plots created using different combinations of variables helped to visualize the relationship between the clusters and different aspects of the data.
The use of color to represent the different clusters helped to visually highlight any patterns or trends in the data that were specific to each cluster.
The insights gained from this analysis can be used by businesses to tailor their marketing strategies or promotional offers to specific customer segments.
Future work may include further exploration of the relationship between the clusters and other variables in the dataset, such as transaction date or customer demographics, and the use of other clustering algorithms or predictive modelling techniques.