Detailed Project Report

# Customer Segmentation using k-prototype algorithm

Revision Number: 1.3

Last date of revision: 23/05/2022

**Rajdeep Mondal**

**Arpan Das**

# Abstract

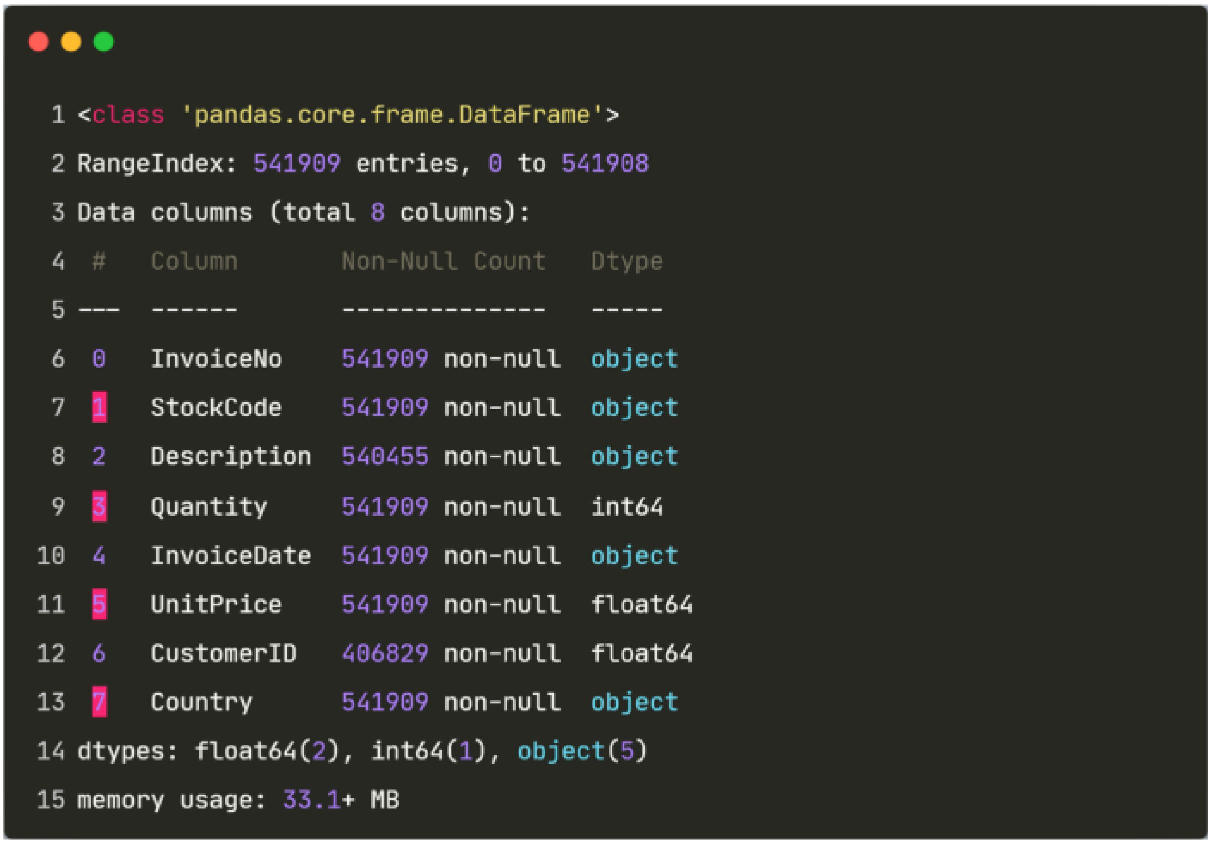
This project aims to analyze E-Commerce data that list purchases made by nearly 4000 customers from December 2010 to December 2021. Based on this database we performed Exploratory Data Analysis with Statistical Methods for gaining data-driven insights with machine learning. Here we used Unsupervised techniques with Python for grouping the customers by their behavioral patterns.

## Dataset Description

This data contains 8 columns —

1. **InvoiceNo**: This is the Invoice number. There are 25,900 unique invoice data. It is a six-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'C', it indicates a cancellation.
2. **StockCode**: This is the Product (item) code. There are 4,070 unique StockCode values. It is a five-digit integral number uniquely assigned to each distinct product. For some data, it contains special code like — `D`, `POST`, `M`, `C2`, `CRUK`, `Discount`, `POSTAGE`, `Manual`, `CARRIAGE`, `CRUK`, `Commission`.
3. **Description**: This describes the product, ie — Product Name. There are 4224 unique descriptions.
4. **Quantity**: This represents the quantities of each product (item) per transaction. It is a Numeric column.
5. **InvoiceDate**: This displays the Invoice Date and time which was generated when each transaction was completed. It is a Numeric column.
6. **UnitPrice**: This represents the Unit price of each product. It is a Numeric column.

7. **CustomerID**: This represents the unique Customer number. It is a five-digit integral number uniquely assigned to each customer.
8. **Country**: This represents the Country name where each customer resides.



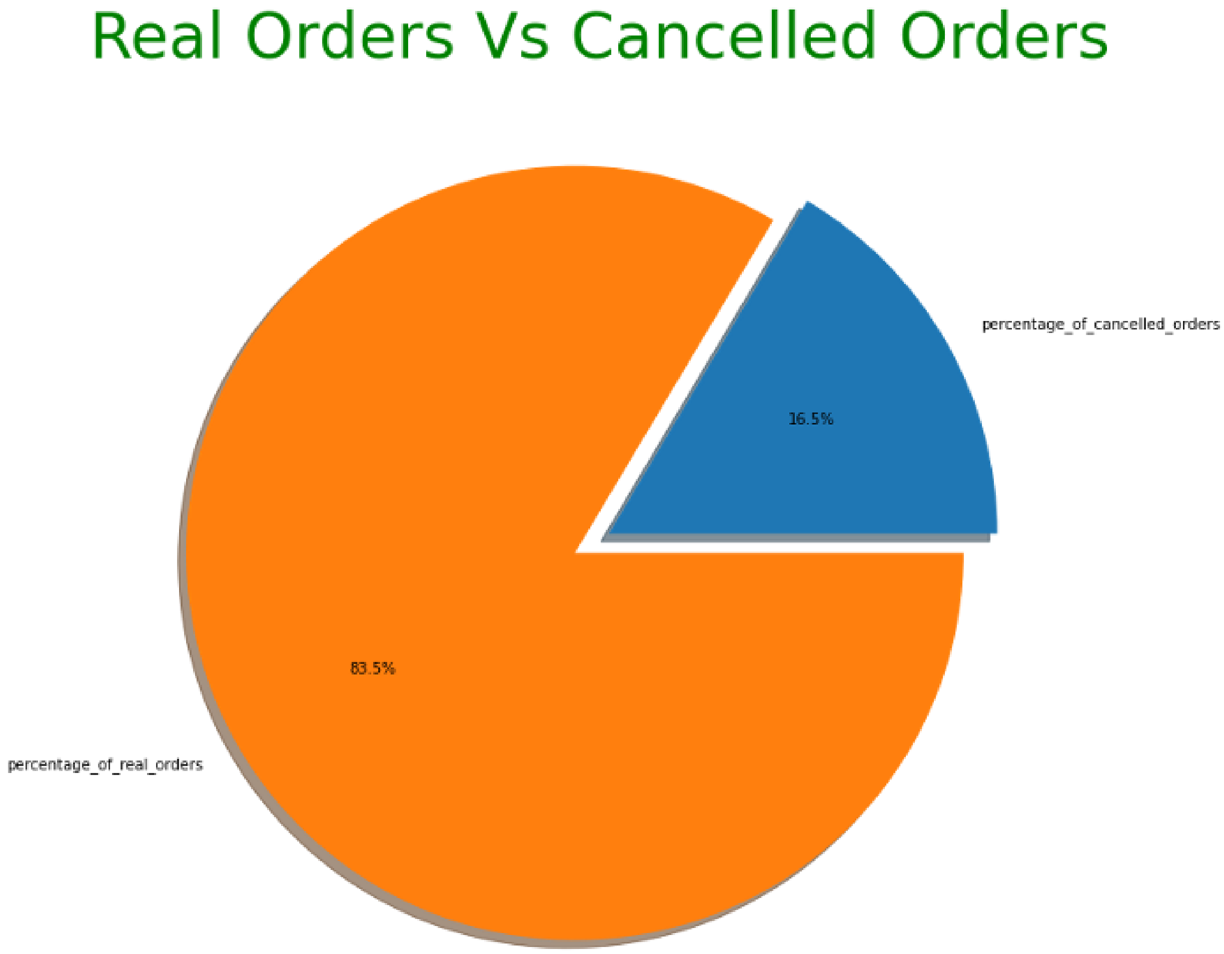
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  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
```

# Some Details we Fetched

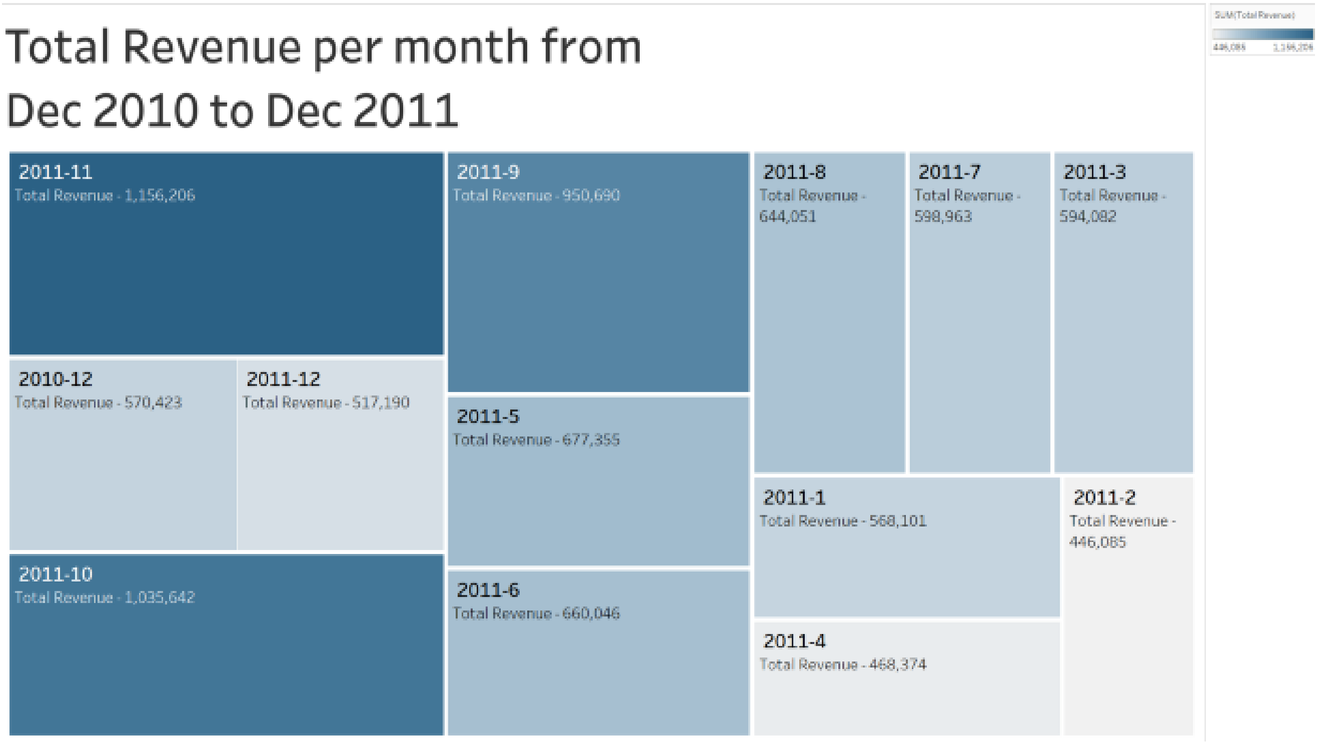
## 1. What was the total revenue?

The total revenue was `£ 8.73M`

## 2. What is the percentage of canceled orders and real orders?

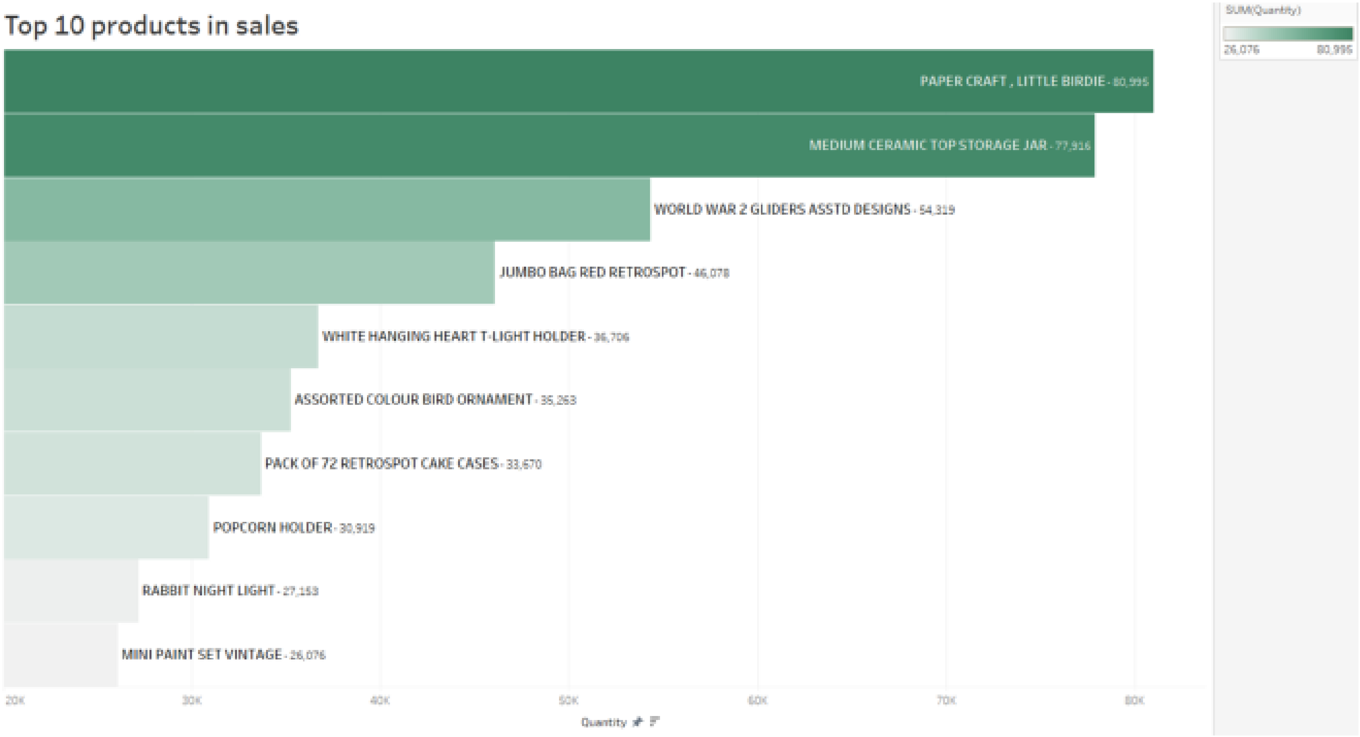


## 3. What is the total revenue per month from December 2010 to December 2011?

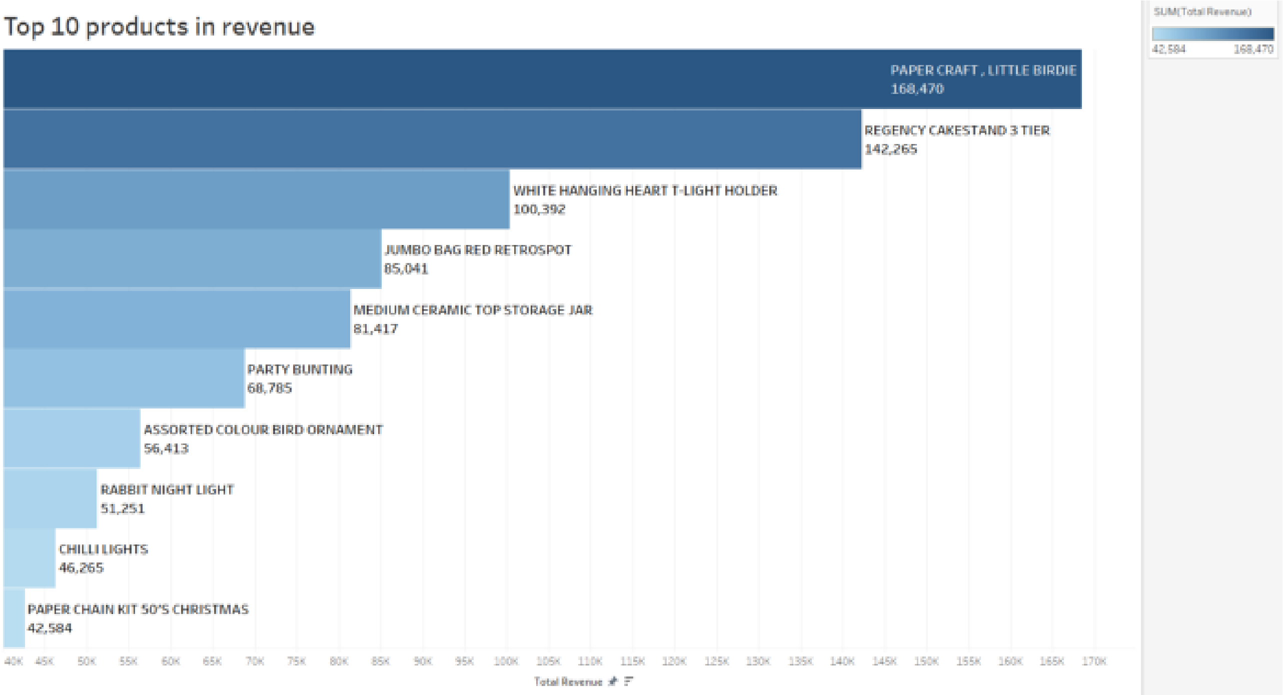


# 4. What are the top 10 products in terms of sales and revenue?

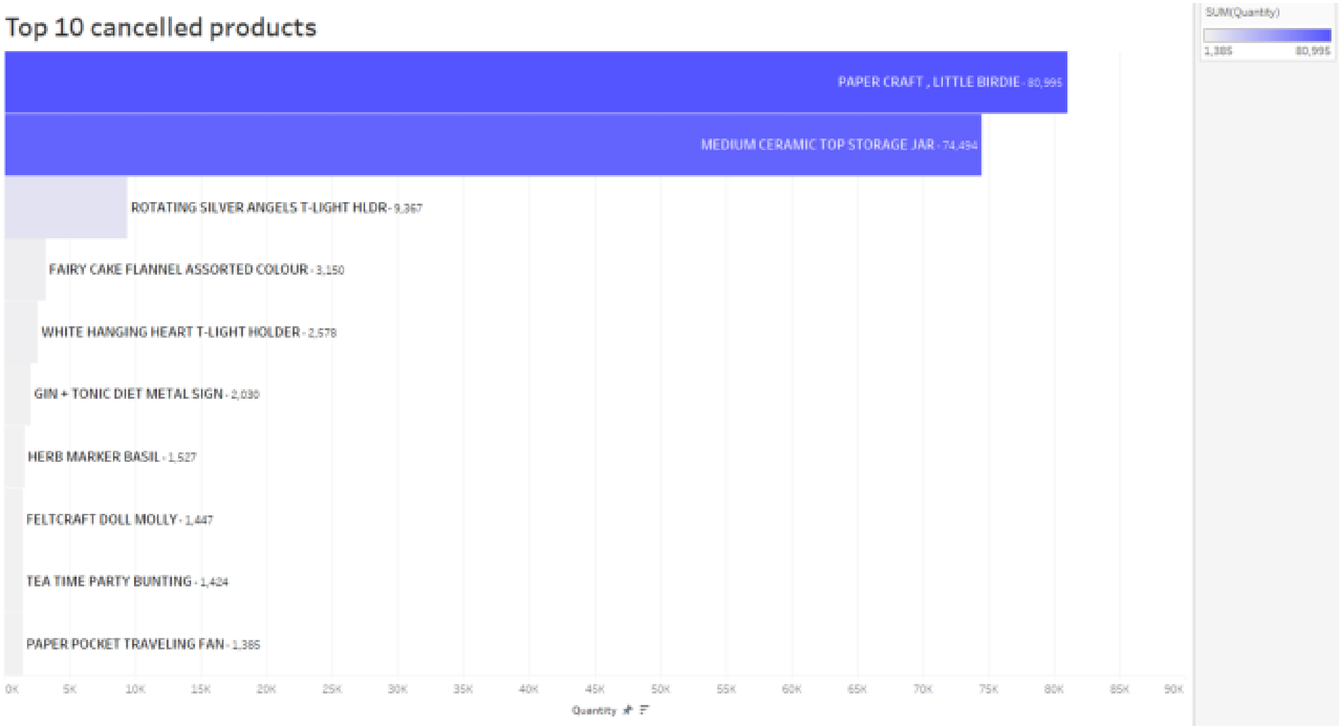
Top 10 products in sales



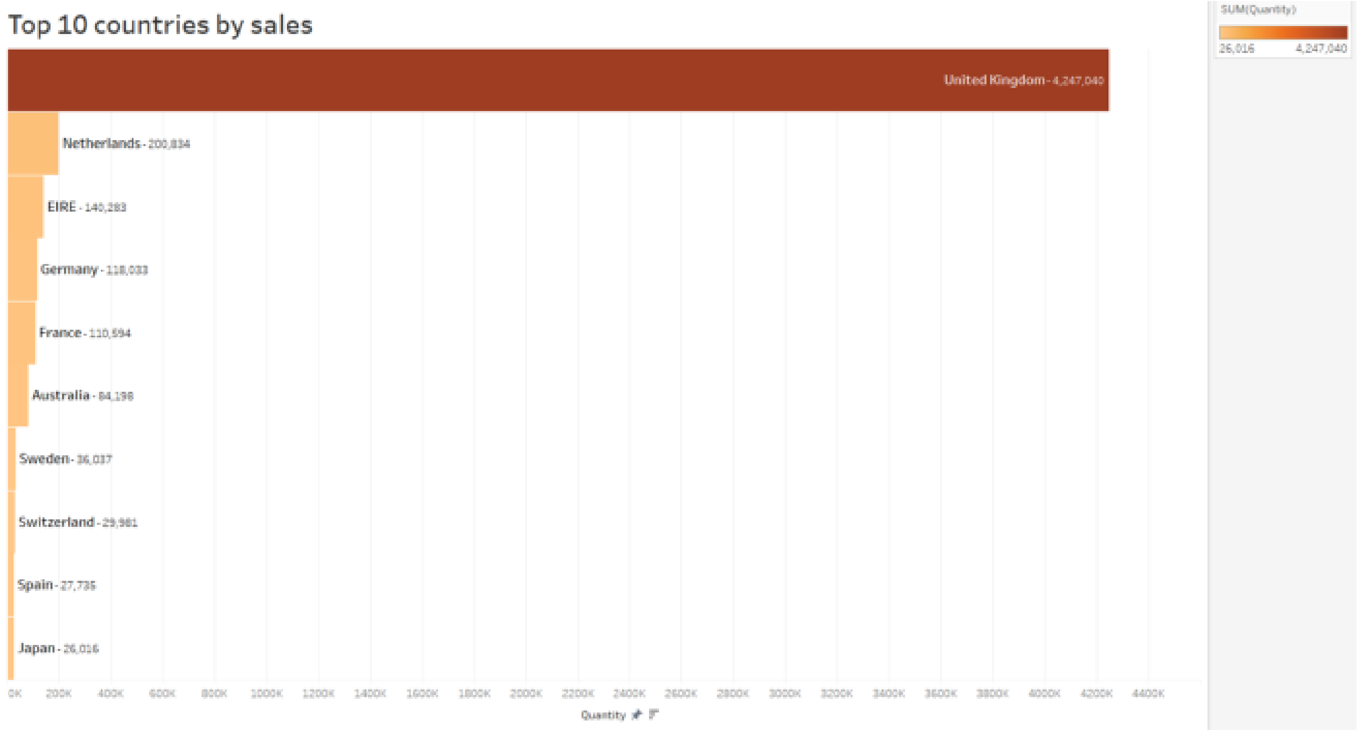
Top 10 products in revenue



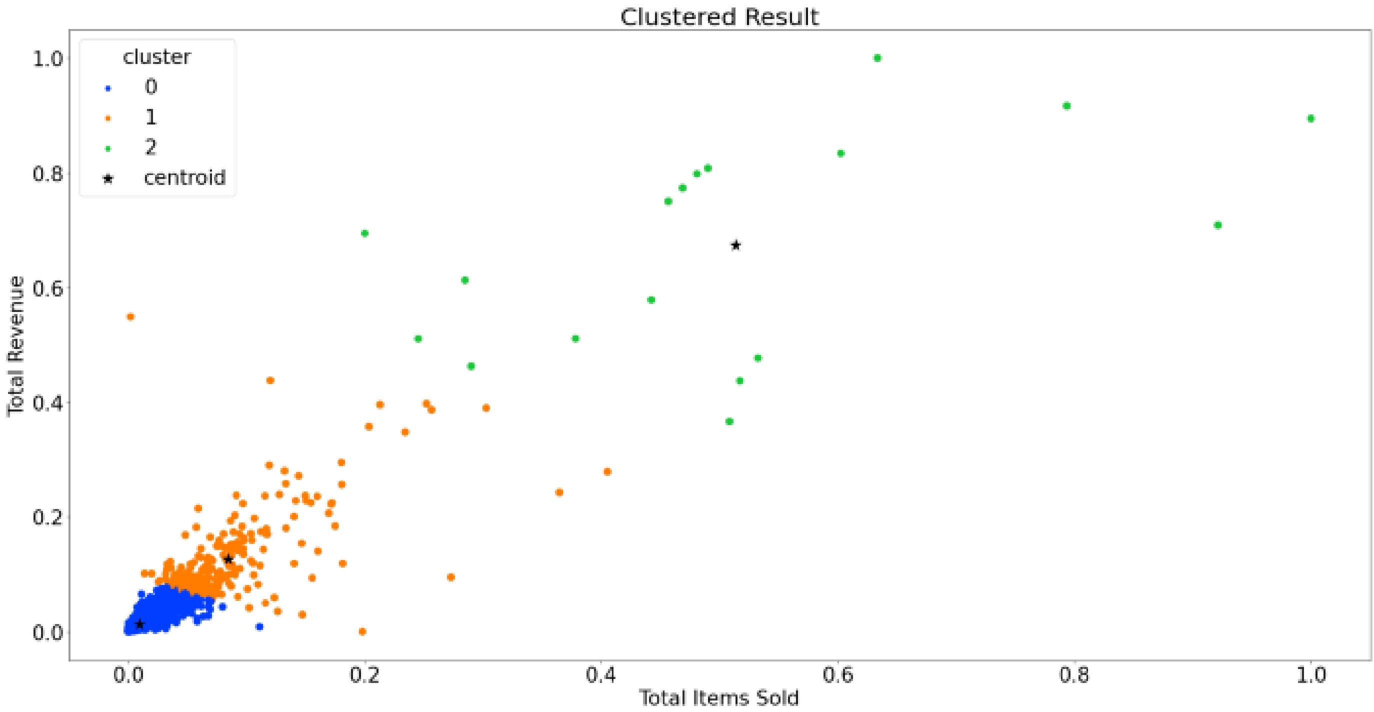
# 5. Which products were returned more frequently?



# 6. What are the top 10 countries that purchased the most?



# 7. Result after Clustering


Clustered Result
cluster
centroid
Total Revenue
Total Items Sold