Vol 3, Issue 6, June 2023 E-ISSN: 2582-9734
International Journal of Engineering, Science, Technology and Innovation (IJESTI)

MACHINE LEARNING-BASED BIG DATA APPLICATIONS FOR E-COMMERCE LOGISTICS

Abdul Khayyum Farooqui

Department of Computer Science Engineering, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana.

Email: abdulkhayyum.519@gmail.com

ABSTRACT

The impetus for business growth and a location to conduct electronic commerce to reach and communicate with customers The sales performance for e-commerce needs to be enhanced and improved. One of the challenging areas of research today is the use of data science and machine learning to handle big data while maintaining the transparency of data records and enhancing e-commerce. The approaches to improving grocery sales performance using data science and machine learning are discussed in this paper. To identify distinct customer segments, customer behaviour, and relevant products, as well as the most popular purchases, this paper can perform sales analysis on datasets.

E-commerce data is cumulative. Depending on the action taken, such as adding a new product, creating a new customer, or making daily sales, new data can arrive every day, in a minute, in a second, or even in milliseconds. As a result, we are expanding the project. For auditing, we'll use a shell script that includes all the details from our analysis. And we set up a job to perform this analysis at a particular time (for example every 2h). For that, we'll use crontab, and we'll use Linux of course. For scheduling and training/testing, we'll use colab.

Key Words: *Sales Analysis, Customer Behaviour, Customer Segmentation, Relevant Products. Incremental Data, Machine Learning, E-Commerce.*

1. INTRODUCTION

The expansion of the e-commerce business globally has been fueled by the expansion of Internet accessibility through computers and mobile devices. "In this case, the consumer has constant access to a vast range of products from many sellers." By equating prices and, in certain cases, invading regional boundaries, these traits in the offer increase market competition and benefit consumers (Allen and Fjermestad, 2001). However, the development of this activity benefits the industry as well as the consumers. Compared to traditional commerce, this sort of trading requires significantly less infrastructure and manpower, providing a business opportunity with more effective cost control (Galinari et al., 2015).

The efficiency of the delivery operation, for example, may suffer as a result of increased demand, which might have an adverse effect on the level of service and customer satisfaction. Given the correlation between freight conditions and customer happiness, it is plausible to argue that logistic management affects the client's decision-making during the purchasing process (Ramanathan, 2010; Pyke et al., 2001). Therefore, in order to continue the growth of e-commerce, it is crucial to investigate the possibilities for

improving logistical strategy. Using consumer profiles and behavioural data may also be a helpful tool in this process. (2008) Mentzer et al. and Leeflang et al. (2014) assert that the new paradigms in the digital domain are altering marketing as an area of study and bringing the outcomes and predictability of sales closer to the consumer. These data may thus be used to manage supply chains in a manner that better satisfies market expectations. In order to predict sales and automatically learn characteristics, Zhao and Wang (2017) analyze raw data from client online behaviour using a convolutional neural network (CNN). This approach outperformed deep neural networks, the auto regressive integrated moving average (ARIMA), and gradient booster regression trees for feature extraction (DNN).

In notable publications like Agatz et al. (2008), which concentrated on multi-channel distribution, and Pyke et al. (2001), which explicitly addressed those themes in the context of the furniture business, several authors have recorded those e-fulfillment challenges. This paper's contribution can be distinguished from that of earlier studies because I it approaches the issue from a multidisciplinary viewpoint; (ii) it provides a systematic appraisal of the study; and (iii) it is not merely descriptive but also lays the groundwork for a novel solution.

1.1 Data Science, Machine Learning

To develop methods, processes, and algorithms with software applications to scientifically extract useful and timely information from user-provided structured and unstructured data collected from e-commerce websites, the Data Science (DS) and Machine Learning (ML) approaches are in fact necessary. Big data and data mining expertise are also utilized to examine business and market trends over time. Additionally, it unites programming practices, statistical techniques, machine learning algorithms, and data engineering. Artificial intelligence requires and greatly benefits from significant knowledge and competence in the fields of mathematics, statistics, information sciences, and computer science. ‘Machine learning’ is a subfield of ‘Computer Science,’ or ‘Artificial Intelligence,’ which is the study of stochastic theory-based algorithms that are competent at doing tasks without requiring explicit programme instructions by relying on patterns and inference. Machine learning techniques are used to build the mathematical model of the training data, or sample data, to make decisions about the data.

2. EVALUATING RESEARCH DONE FOR THE OPERATION OF DISTRIBUTING E-COMMERCE

This part, which is focused exclusively on the distribution process for e-commerce, gives a study evaluation that is broken down into six categories: Manufacturing, Line-Haul cargo, Last Mile delivery, synchronized system, E-commerce traffic sources, E-commerce Buyer Journey Three of those issues were mentioned and linked to a set of keywords to direct a database search. Each of those categories comprises a variety of obstacles and possibilities for study.

2.1 E-Commerce Traffic Sources

Companies frequently look for more effective ways to communicate with their customers; these options will vary by industry because the performance of the channels will vary depending on the situation. Despite this, the idea and research surrounding digital sources are still fairly new. For instance, the first publication in Springer search on the subject of social media happened in 2004.

2.2 E-Commerce Buyer Journey

With suggestion tools based on similarities or buyer profiles, it is possible to assist the visitor in finding the suitable product after the first interaction with the business (Ogawa et al., 2008). Reliability and Data Privacy are essential in the e-commerce market at this moment since they go beyond the product value proposition that the firm must provide. That might be the reason the Data Privacy subject produced the majority of the results in this category; this shows a significant worry owing to the widespread use of the internet in daily life. From a technological perspective, several research are ongoing to determine the system requirements to allow consumers to navigate safely while online shopping (Cheung and Chanson, 2001).

2.3 Manufacturing

An intricate issue, resource allocation within a given time period is produced by the production schedule, which incorporates a number of elements. The research suggests a variety of solutions to this issue, including data-driven (Kuck et al., 2016) and protocol-based multi-agent techniques (Kaihara and Fujii, 2005), among others. Additionally, a comprehensive knowledge base for the industrial scheduling issue is offered in Varela et al. (2005). "In addition to production scheduling, additional ideas like smart manufacturing maintenance management and fog computing are becoming more relevant with industry 4.0 applications." An integrated planning for production and maintenance is presented by Rdseth et al. (2017). Fog computing allows for this kind of local processing demand by enabling data processing at the network's edge without sending superfluous data to the cloud. In this area, research is being done on methods to facilitate resource discovery and the creation of a fog orchestrator (Skarlat et al., 2017; Velasquez et al., 2018).

2.4 E-Commerce Data Is Incremental

Means new data comes daily/in a minute/in sec/ or milliseconds depending on action performed like new product added or new customer created or daily sales.

3. METHOD

On datasets, we will perform sales analysis to discover distinct customer segments, consumer behaviour, and pertinent products, most frequent purchases, etc. E-commerce data is cumulative. Depending on the action taken, such as adding a new product, creating a new customer, or making daily sales, new data can arrive every day, in a minute, in a second, or even in milliseconds. As a result, we are expanding the project. For auditing, we'll use a shell script that contains all of the details about our analysis, and we'll set up a job to run it at a certain time (for example every 2h).

For that, we'll use crontab, and we'll use Linux of course. For scheduling and training/testing, we'll use colab.

Requirements:

- Linux/Windows WS
- Python3.7 or Higher

Step-1: Install required python libraries.

```
$ pip install -r requirements.txt
```

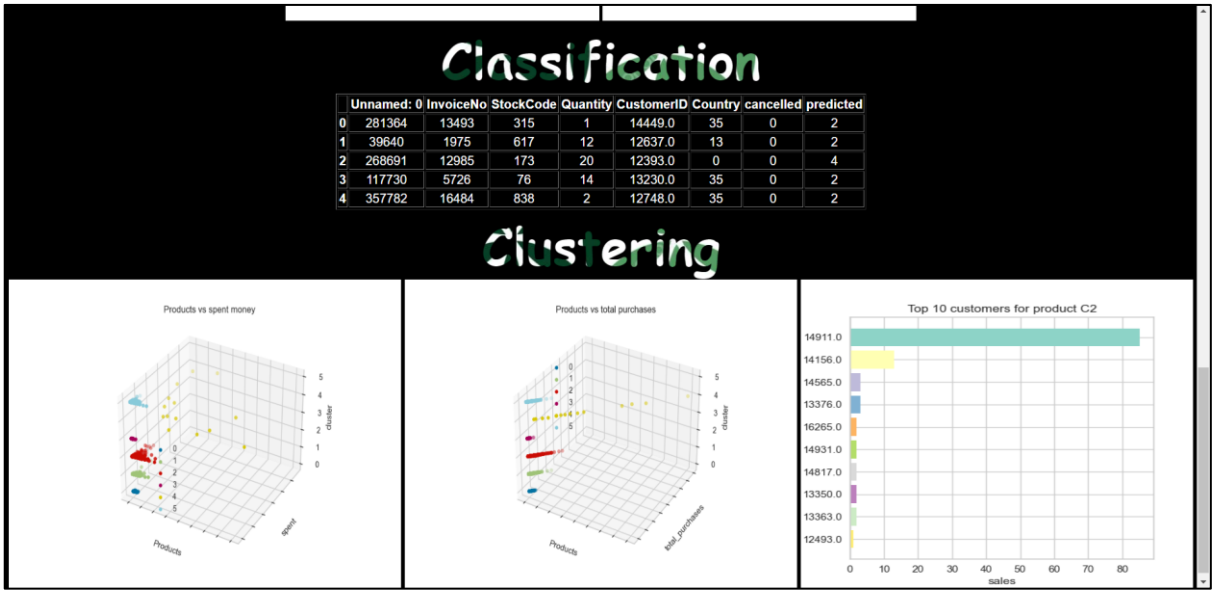
Step-2: run the ecom_datascience.py script.

```
$ python3 ecom_datascience.py
```

Step-3: run app.py script.

```
$ python3 app.py
```

Step-4: Now go to http://127.0.0.1:5000 or localhost:5000



All the graph/visualization and Data sample can be found on the web page.

4. PROBLEM DEFINITION

There are two major streams in an e-commerce transaction: informational and physical. These are often autonomous and their sequential planning and operation jeopardise efficiency and service quality. A major potential for development still exists in the prospect of integrating both using customer behaviour. The phases of the buyer journey may, in fact, be tied to the information flow. For instance, it is anticipated that a visitor to an e-commerce website would execute a certain action related to his stage in the purchasing process. For instance, if a user is a new visitor, the company's initial concern is to deanonymize the user and collect their personal data in order to associate information such as surfing or social media behaviour; usually, the email address is the information utilized for this. "It is feasible to comprehend each person's profile and where they are in their buyer's journey via a methodical processing." As a result, people in charge of internet marketing may leverage consumer information to create value for their customers (Allen and Fjermestad, 2001). More specifically, it is possible to determine whether the user is just beginning their journey and discovering their needs, or whether they have identified a problem and are at the point of decision-making where they only need a few justifications to complete a transaction. In this step, recommendation algorithms can help the user find more products that are relevant to their interests. A third possibility is that the buyer is already in the closing stage, likely already intends to make the purchase, and is just considering the alternatives that are suitable for his needs, such as quick delivery. The process does not stop after the first contract is closed; rather, it must be sustained in order to increase client loyalty and ensure future deals.

5. DATA CLEANING

Implementing mistake avoidance measures is the main component of data cleansing (see data quality control procedures later in the document). However, error-prevention techniques may only significantly minimize frequent mistakes; hence, many data errors will be found accidentally while doing tasks like:

- When collecting or entering data
- When transforming/extracting/transferring data
- When exploring or analyzing data
- When submitting the draft report for peer review

Even with the best error prevention techniques in place, it will still be necessary to actively and methodically look for errors/problems, identify them, and take planned steps to fix them. Repeated cycles of screening, diagnosing, treating, and documenting this process are required for data cleaning. Data collection and entry procedures should be modified as error patterns are found to address them and prevent recurrence.

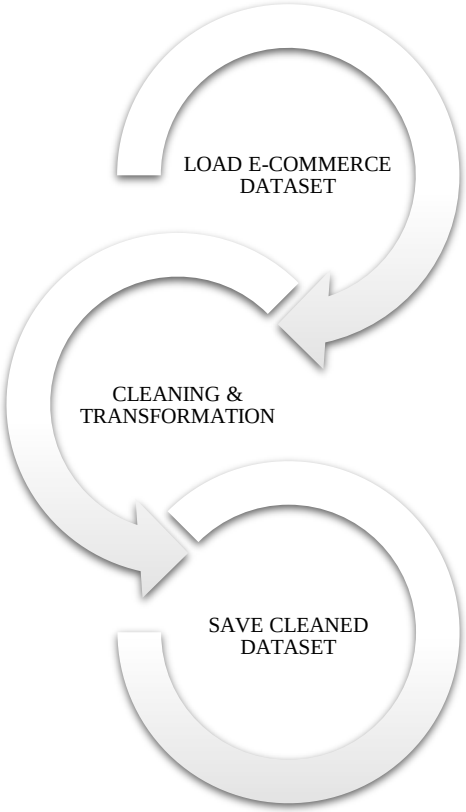


Figure 1: Data Cleaning

The data cleaning processes for cleaning data the dataset's section has been defined as a cleaned data output with the following fields: Invoice No, Stock code, Description, Quantity, Invoice Date, Unit Price, and Country.

```
import pandas as pd

df = pd.read_csv('ecommerce-data.csv', encoding='ISO-8859-1')

df.head()
```

Out [1]:

	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

In [ ]:

```
#Check missing values

df.isnull().sum()
```

Out [ ]:

```
InvoiceNo        0
StockCode        0
Description      1454
Quantity         0
InvoiceDate      0
UnitPrice        0
CustomerID       135080
Country          0
dtype: int64
```

For the description, fill NAs values with an empty string.

- For the Costomer ID, as it is an identifier remove those rows.

In [ ]:

```
df['Description'].fillna('', inplace=True)

df.dropna(0, inplace=True)
```

5.1.1 Transformations

The term 'data transformation' refers to changing the format, structure, or values of data. When used for data analytics projects, data can be altered twice along the data pipeline. Data transformation serves as the intermediary step in the ETL (extract, transform, load) methodology used by businesses with on-premises data warehouses. The computational and storage capacity of cloud-based data warehouses may be expanded with a delay measured in seconds or minutes.

Organizations can load raw data directly into data warehouses using ELT (extract, load, transform), without making any preload adjustments, and convert the data when a query is received. Data migration, warehousing, integration, and wrangling are a few examples of operations where data transformation may be used.

Furthermore, it is essential for any organisation wishing to use its data to offer pertinent business insights. As the volume of data has increased, organisations require a trustworthy method for utilising it to use it effectively for their operations. Utilizing this data requires data transformation because, when done correctly, it guarantees that the information is available, consistent, secure, and eventually acknowledged by the targeted business users. Data transformation carried out using the cancelled column If the invoice number begins with the letter 'C,' a cancellation means that corresponds. So, it's clear that there have been a lot of cancellations if we count them. 'cancelled' = 'Invoice No' in the database. str. Starts with('C'). as type('int32')

In [ ]:
```
df[df['InvoiceNo'].str.startswith('C')]['InvoiceNo'].count()
```

Out [ ]:
```
8905
```

In [ ]:
```
#add column cancelled

#If the Invoice number starts with 'C' means that corresponds to a cancellation. So,

#if we count the number of cancellations, we can see that there are many cancellations.



df['cancelled'] = df['InvoiceNo'].str.startswith('C').astype('int32')

df['cancelled'].unique()
```

Out [ ]:
```
array([0, 1])
```

In [ ]:
```
df.to_csv('cleaned_data.csv', index=False)
```

5.2 Data Exploration

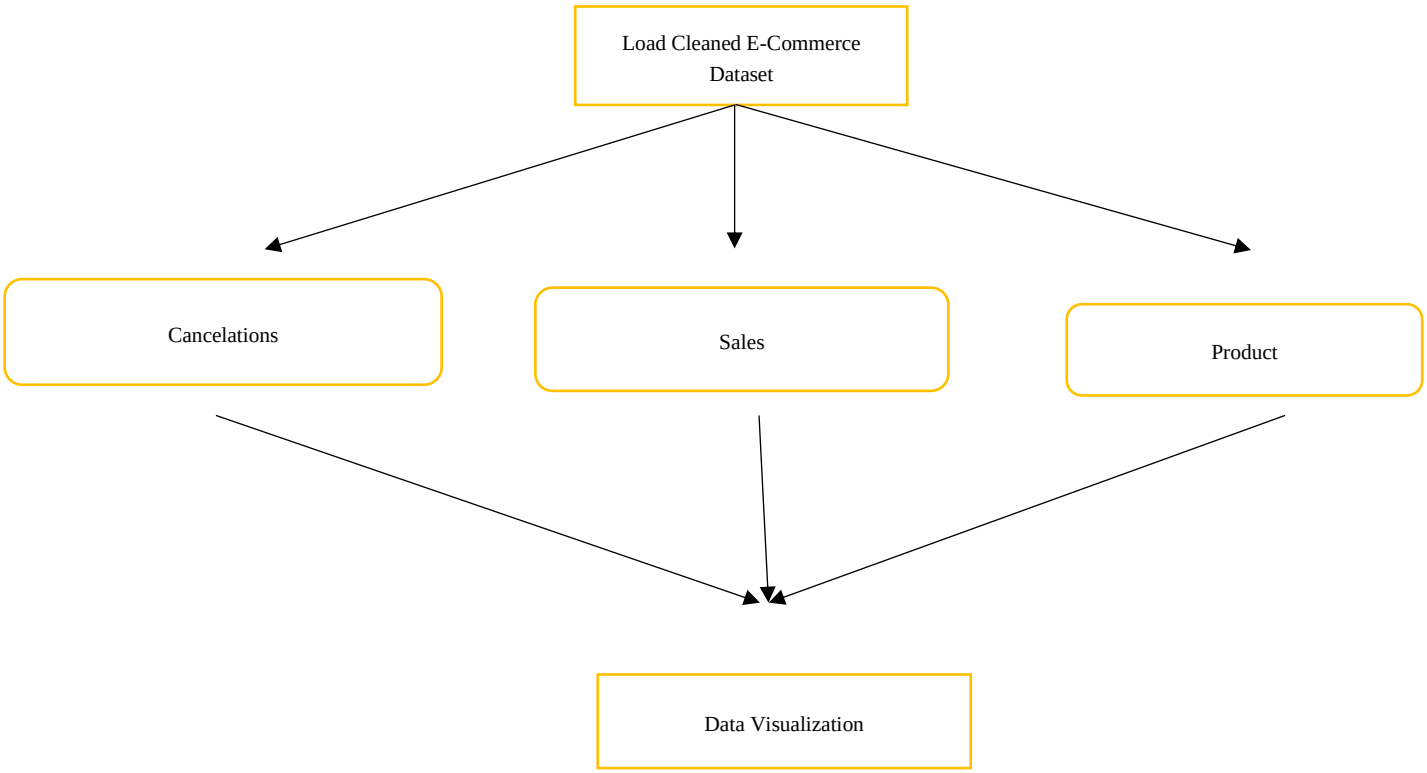
Users examine and comprehend their data using statistical and visualisation techniques during data exploration, also referred to as exploratory data analysis (EDA). Choosing which model or algorithm to use in the following steps, as well as identifying patterns and issues in the dataset, are all aided by this step. For data exploration, we will load the cleaned e-commerce dataset and present cancellations, sales, products, and data visualisation to draw a conclusion.

5.2.1 Cancellations

The import pandas line of the code instructs Python to load the pandas data analysis library into your current environment. The as pd part of the code directs Python to give pandas the pd alias. You can utilise pandas functions by using pd. function name rather than pandas. The import pandas line of the code instructs Python to load the pandas data analysis library into your current environment. The as pd part of the code directs Python to give pandas the pd alias. You can utilise pandas functions by using pd. function name rather than pandas. Using matplotlib. Pyplot, import plt. Set the figure size, as well as the padding within and outside the subplots. Create x and y data points using Numpy. Plot the x and y data points

using the plot () function. Use the show () function to display the figure. The open-source project GeoPandas makes working with geographical data in Python much easier. The datatypes used by Pandas are expanded by GeoPandas to provide geometric types for spatial operations. Shapely performs geometric operations. Additionally, Geopandas makes use of fiona for file access and matplotlib for charting.

Matplotlib is a Python package that facilitates graph charting. It is used for data visualisation and graphical charting. Before using it, matplotlib must be installed. By calling the statement import numpy as np, you can shorten the word 'numpy' to 'np,' making your code easier to read. Additionally, it helps to avoid namespace issues. Tkinter and ttk are a nice example of what may happen if you do have that issue. The import seaborn section of the code instructs Python to import the Seaborn library into your current environment. The as sns command in the code tells Python to give Seaborn the sns alias. You may utilise Seaborn functions by using sns. function name rather than seaborn. The colormap class allows for the creation of colormap objects from a list of colours. Both normal mapping and direct indexing into colormaps may be done using this to create distinctive colormaps.



In [83]:

```
import pandas as pd
import matplotlib.pyplot as plt
import geopandas
from matplotlib import cm
import numpy as np
import seaborn as sns
from matplotlib.colors import ListedColormap, LinearSegmentedColormap
import countries as countries_utils
```

In [109]:

```
df = pd.read_csv('cleaned_data.csv', encoding='ISO-8859-1')

df.head()
```

Out [109]:

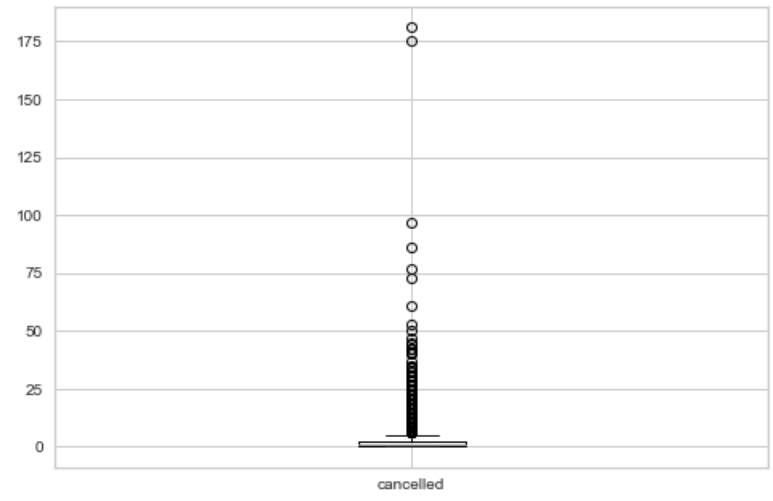
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	cancelled
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom	0
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom	0
2	536365	84406B	CREAM CUPID HEARTS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

In [85]:

```
# amount of cancellations per stock

per_stockcode = df[['StockCode','cancelled','Quantity']].groupby(['StockCode']).sum().reset_index()

per_stockcode.boxplot(column=['cancelled'])
```

Out [85]:

```
<AxesSubplot:>
```

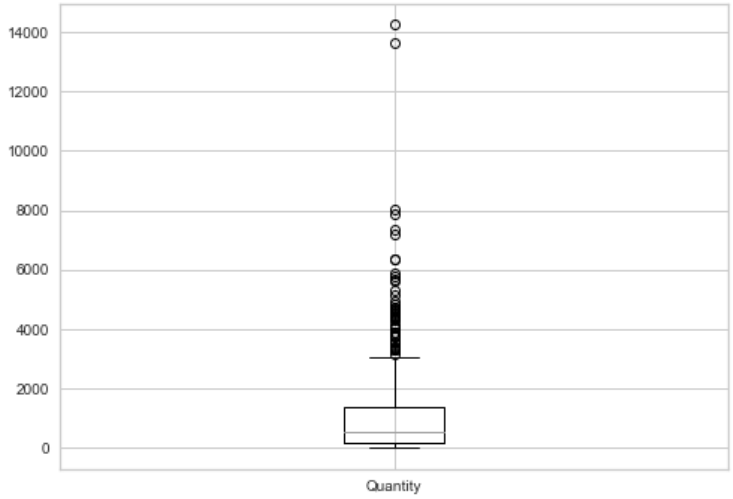


In [86]:

```
# amount of items of each stockcode where cancelled

cancelled_per_stockcode = per_stockcode[per_stockcode['cancelled']==1]

cancelled_per_stockcode.boxplot(column=['Quantity'])
```

Out [86]:

```
<AxesSubplot:>
```



5.2.2 Analyze Outliers

An observation that differs abnormally from other values in a population-based random sample is referred to as an outlier. "In a way, this definition defers to the analyst's (or a consensus process') judgement as to what constitutes abnormal behaviour."

From the graphs we can see that there are two products are cancelled more times.

In [87]:

```
#print outliers

per_stockcode[per_stockcode['cancelled']>170]
```

Out [87]:

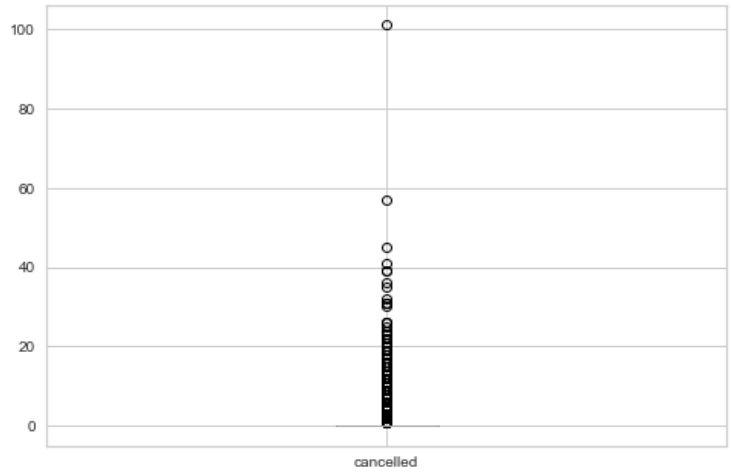
	StockCode	cancelled	Quantity
1292	22423	181	11555
3681	M	175	3184

In [88]:

```
# cancellations per customer.

per_client = df[['CustomerID','InvoiceNo','cancelled']]. groupby(['CustomerID','InvoiceNo']).sum().reset_index()

per_client.boxplot(column=['cancelled'])
```

Out [88]:

```
<AxesSubplot:>
```



From the graph, we can see that there is an outlier, a customer that cancelled more than 100 orders.

In [89]

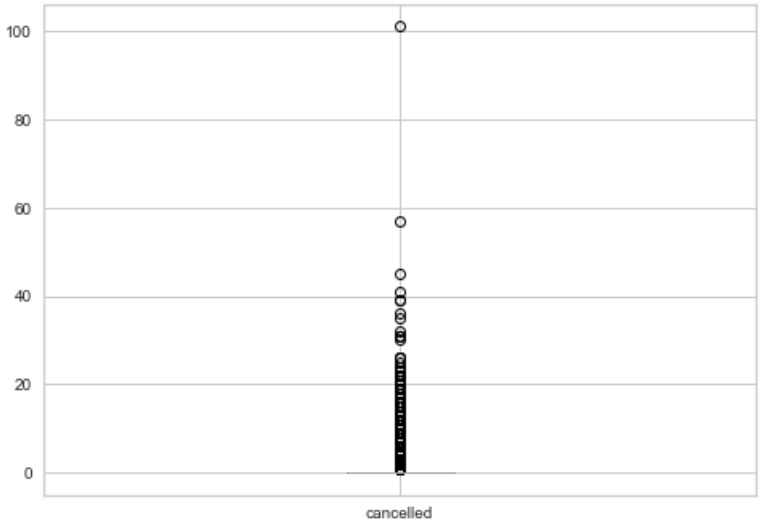
```
#print outlier

per_client[per_client['cancelled']>100]
```

Out [89]:

	CustomerID	InvoiceNo	cancelled
1004	12607.0	C570867	101

In [90]:

```
# cancellations per country.

per_country = df[['Country','InvoiceNo','cancelled']]. groupby(['Country','InvoiceNo']).sum().reset_index()

fig=per_country. boxplot(column=['cancelled']).get_figure().savefig('output.png')
```



There is a country that is outlier, we can check which is that country by filtering the data.

In [91]:

```
#print outlier

per_country[per_country['cancelled']>100]
```

Out [91]:

	Country	InvoiceNo	cancelled
2320	USA	C570867	101

5.2.3 DATA FUSION

The process of combining data from various sources to create precise, thorough, and unified data about an entity is known as data fusion. Low level, feature level, and decision level data fusion are the different classifications.

```
world = geopandas.read_file(geopandas.datasets.get_path('naturalearth_lowres'))

countries = per_country[['Country','cancelled']].groupby(['Country']).sum().reset_index()

world_values = world['name'].values

countries['is_in'] = np.array([x in world_values for x in countries['Country']])

print(countries[countries['is_in']==False])
```

```
            Country  cancelled  is_in
2           Bahrain          0  False
6     Channel Islands       10  False
8     Czech Republic          5  False
10              EIRE       247  False
11  European Community         1  False
22             Malta        15  False
27               RSA         0  False
29         Singapore          7  False
33               USA       112  False
36       Unspecified          0  False
```

Rename countries to match with world dataset

In [93]:

```
countries_utils.update_values(countries, 'Country', ['USA', 'RSA', 'Czech Republic', 'EIRE'],
                              ['United States of America', 'South Africa', 'Czechia', 'Ireland'])
```

Keep only countries that can be plotted with the world dataset

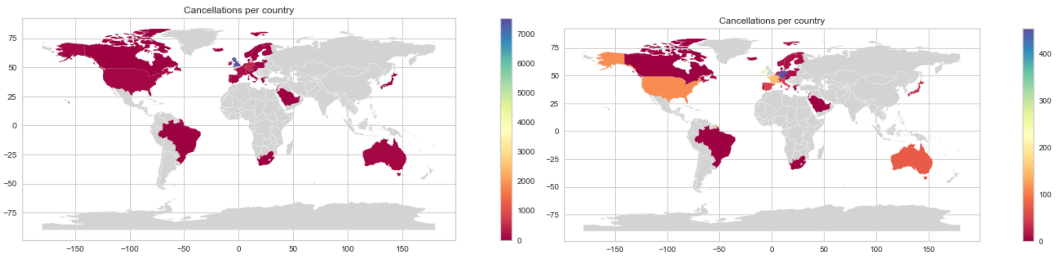
In [94]:

```
countries['is_in'] = np.array([x in world_values for x in countries['Country']])

countries = countries[countries['is_in']].copy()

countries.drop(columns=['is_in'], axis=1, inplace=True)

countries
```

Out [94]:

5.3 DATA VISUALIZATION

```
countries_utils.plot(countries, 'Cancellations per country', 'Country', 'cancelled', True)


#without United Kingdom since it has a lot of cancellations and it might be the outlier that we saw in the other graph

countries_utils.plot(countries[countries['Country'] != 'United Kingdom'], 'Cancellations per country', 'Country', 'cancelled', True)
```



5.3.1 Sales

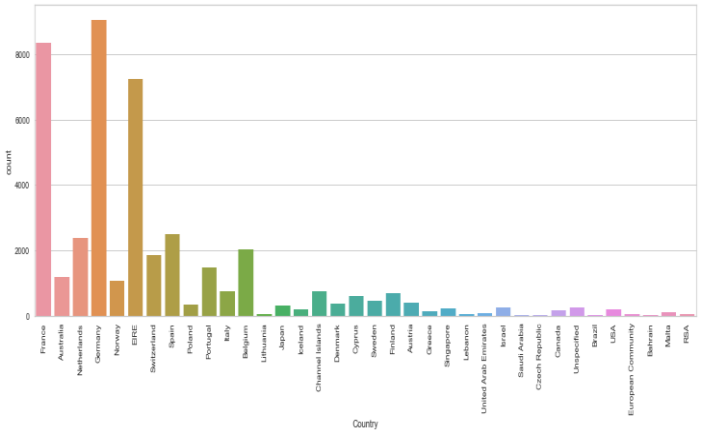
In [97]:

```
plt.figure(figsize=(18,6))

sns.countplot(df[(df['Country'] != 'United Kingdom') & (df['cancelled']==0)]['Country'])

plt.xticks(rotation=90)

plt.savefig('invoice_per_country.png')
```

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```

In [107]:

```
df[(df['Country'] != 'United Kingdom') & (df['cancelled']==0)]
```



5.3.2 Products

For the products, we can show the top 10 products for each country (the most popular products). We use the quantity field, which indicates how many times the product was purchased in the whole purchase. So, we should have to group by country and product and make the sum.

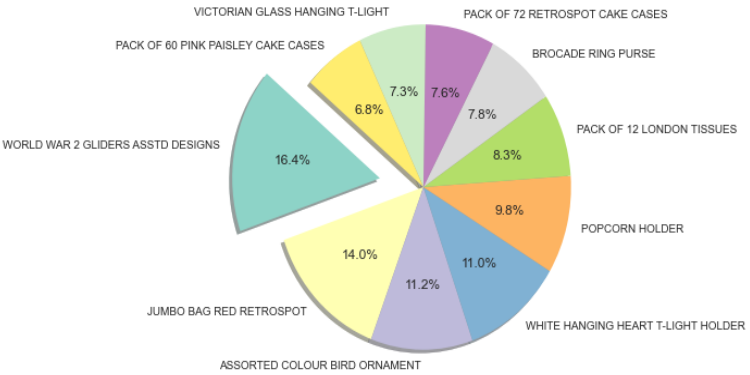
In [98]:

```
products = df[['StockCode', 'Country', 'Quantity','Description']].groupby(['StockCode', 'Country','Description']).sum().reset_index()
```

If we want to see the top 10 products for a specific country, the code would be

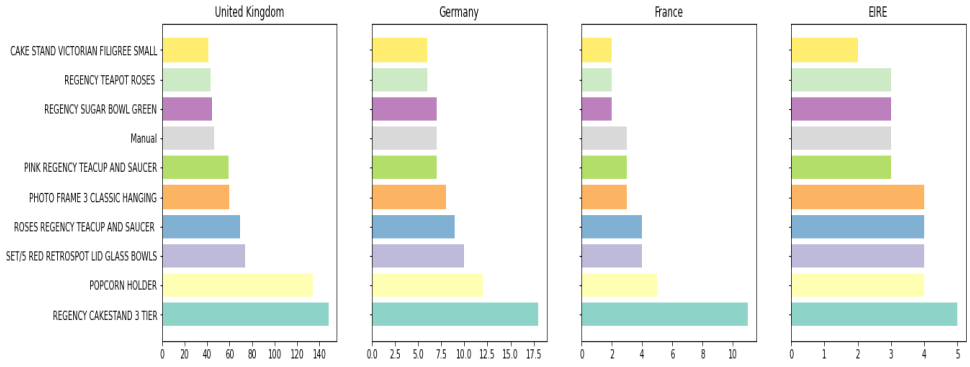
In [99]:

```
UK_products = products[products['Country']=='United Kingdom'].sort_values('Quantity',

ascending=False).reset_index().head(10)


UK_products
```



See the most cancelled products for those countries

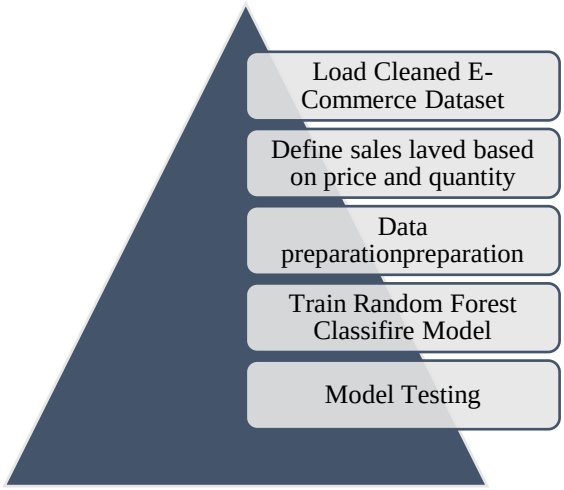
```
df['cancelled'] = df['InvoiceNo'].str.startswith('C').astype('int32')

cancelled_products = df[['Description', 'Country', 'cancelled']].groupby(['Description', 'Country']).sum().reset_index()

cancelled_products = get_top_n(cancelled_products, 'Country', 'cancelled', 10)


f, axs = plt.subplots(1,4, figsize=(20,5), sharex=False, sharey=True)

i = 0

for country in top_4['Country']:

  axs[i] = plot_filter_bar(axs[i],cancelled_products, 'Country', 'cancelled', 'Description', country, 10, country)

  i+=1


plt.show()
```



6. CLASSIFICATION

Classification is the process of categorizing things on the basis of properties. "Organisms are grouped together when they have common features." The classification of living things includes seven levels such as kingdom, phylum, class, order, family, genus, and species.

We will divide the data frame while classifying sales as follows: After splitting the dataset into tests and training runs, import model selection from Sklearn.



6.1 Classify Sales

Out[39]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	cancelled
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom	0
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom	0
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom	0
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom	0
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom	0
5	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/1/2010 8:26	7.65	17850.0	United Kingdom	0
6	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/1/2010 8:26	4.25	17850.0	United Kingdom	0
7	536366	22633	HAND WARMER UNION JACK	6	12/1/2010 8:28	1.85	17850.0	United Kingdom	0
8	536366	22632	HAND WARMER RED POLKA DOT	6	12/1/2010 8:28	1.85	17850.0	United Kingdom	0
9	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	12/1/2010 8:34	1.69	13047.0	United Kingdom	0

In [45]:

```
from sklearn.ensemble import RandomForestClassifier
clf = RandomForestClassifier(random_state=202)
clf.fit(X_train, y_train)
print('Score:', clf.score(X_test, y_test))
predictions = clf.predict(X_test)
df_p = pd.DataFrame()
df_p['Actual'] = y_test
df_p['Predicted'] = predictions
df_p.head()
```

```
Score: 0.8960868175896566
```

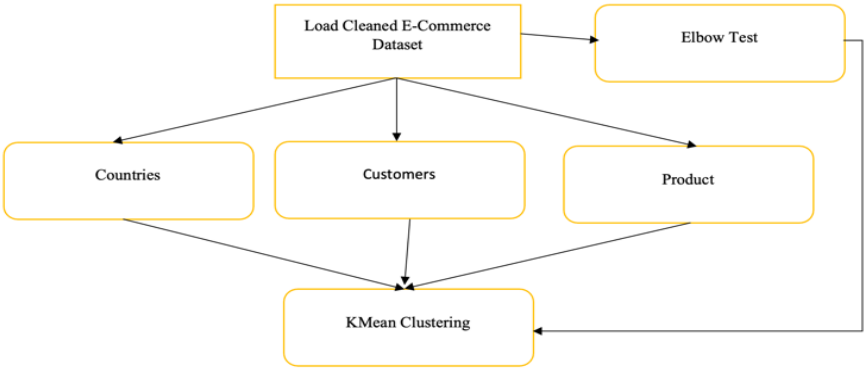
Out[45]:

	Actual	Predicted
281364	2	2
39640	2	2
268691	4	4
117730	2	2
357782	2	2

7. CLUSTERING

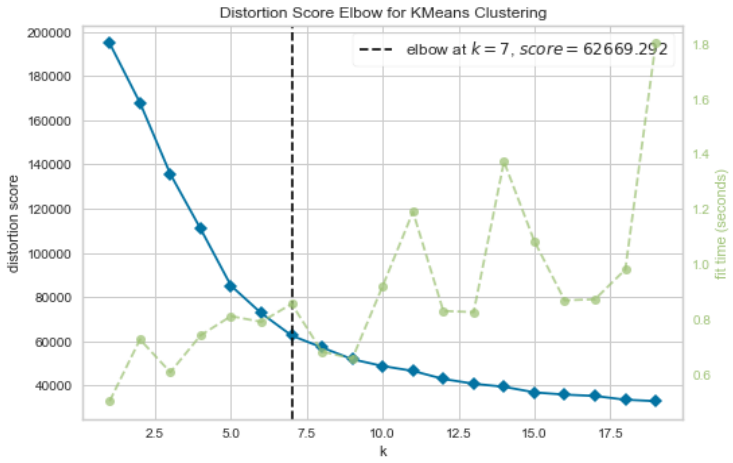
Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups it is done by K-means. We will import Numpy as np and Pandas as pd from Sklearn for Clustering with K-means. kneed import from cluster import KneeLocator imports countries as countries utils from sklearn. preprocessing, imports matplotlib. pyplot as plt from matplotlib, imports scale, and imports matplotlib. pyplot as plt from mpl toolkits. mplot3d. wordcloud import Axes3D import

WordCloud imports nltk and TfidfVectorizer from Sklearn. feature extraction. text. cluster import Yellowbrick KMeans. import KElbow Visualizer cluster.



7.1 Countries

In order to find the optimal number of clusters, we can perform the elbow test. We will create year and month columns from the date when clustering countries, remove any unnecessary columns when grouping by other columns, and take the sum when converting Stock Code to integers. We will investigate clusters without the United Kingdom because we have established that it is an outlier. The data that will be used as input for the model must be scaled. "We can use the elbow test to determine the ideal number of clusters. We can run the model now that we have the ideal number of clusters." We can plot the clusters onto the map to see if there is a relationship between the countries in order to visualize the results by adding a column that identifies the cluster assigned.



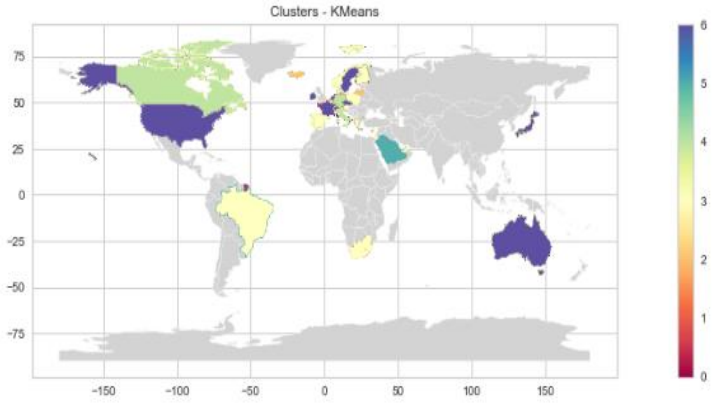
Now that we have the optimal number of clusters, we can perform the model.

```
km = KMeans(n_clusters=n_clusters, random_state=20)

y = km.fit_predict(x)

#add column that determines the cluster assigned

df_aux_1['cluster'] = km.labels_
```

In order to visualize the results, we can plot the clusters into the map to see if there is a relationship with the countries.

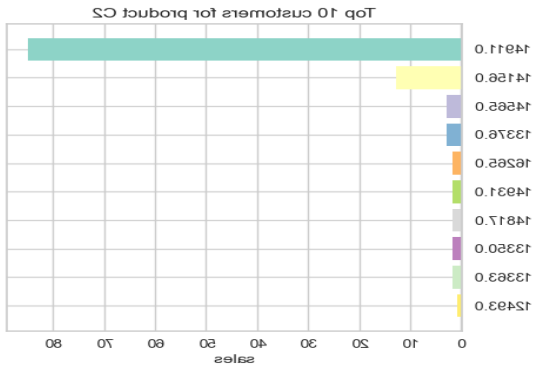
In [8]:

```
countries_utils.plot(df_aux_1, 'Clusters - KMeans', 'Country', 'cluster', True)
```



7.2 Customers

We can see how the clients are divided into categories based on their purchases. observe who purchased more of each product and maintain the top 10 sales.

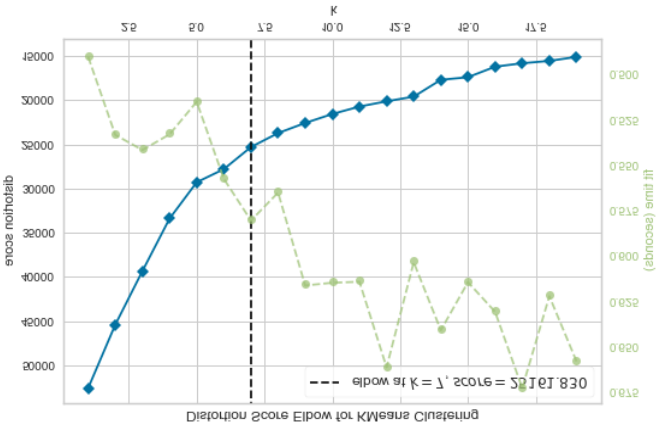


Keep the relevant information for each customer:

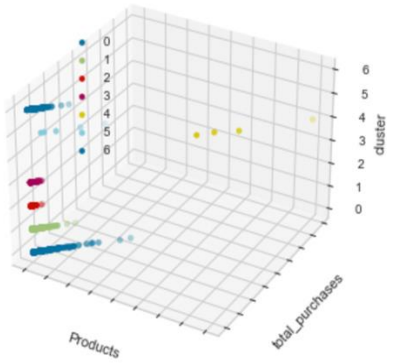
- Total amount of purchases
- Start date
- How many different products the customer bought
- Total money spent
- Last date purchase

sum of all transactions with customers = df [df['cancelled']==0] [['InvoiceNo','CustomerID']] clients = clients.groupby(['CustomerID']) .count().reset index() customers.

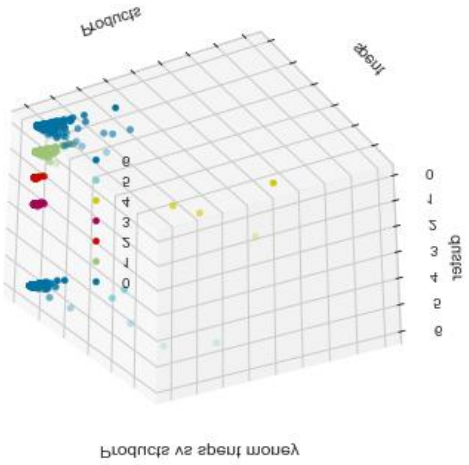
Rename the column ‘InvoiceNo’ to ‘total purchases’ with ‘inplace=‘True’ How many different products the customer purchased after this start date, along with the total amount spent and the date of the last purchase Following that, we can use k-means. look over this data and observe how the clients are categorised. after that, add a column that identifies the group of customers who made comparable numbers of purchases and similar types of purchases. Although cluster 3 is more dispersed, customers typically purchase similar goods. The only cluster with customers who spent more money is Cluster 3.



Product vs Total Purchase Made

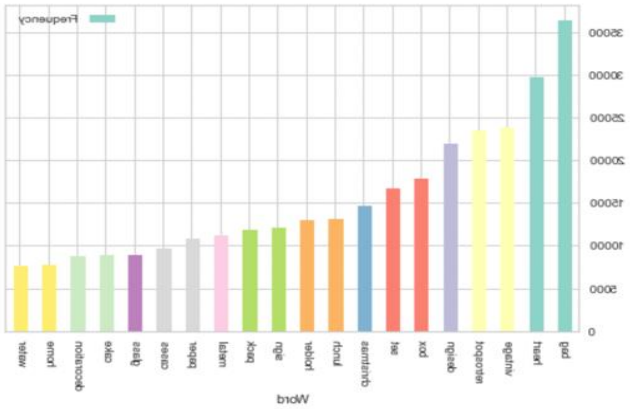


Product vs Spent money



7.3 Products

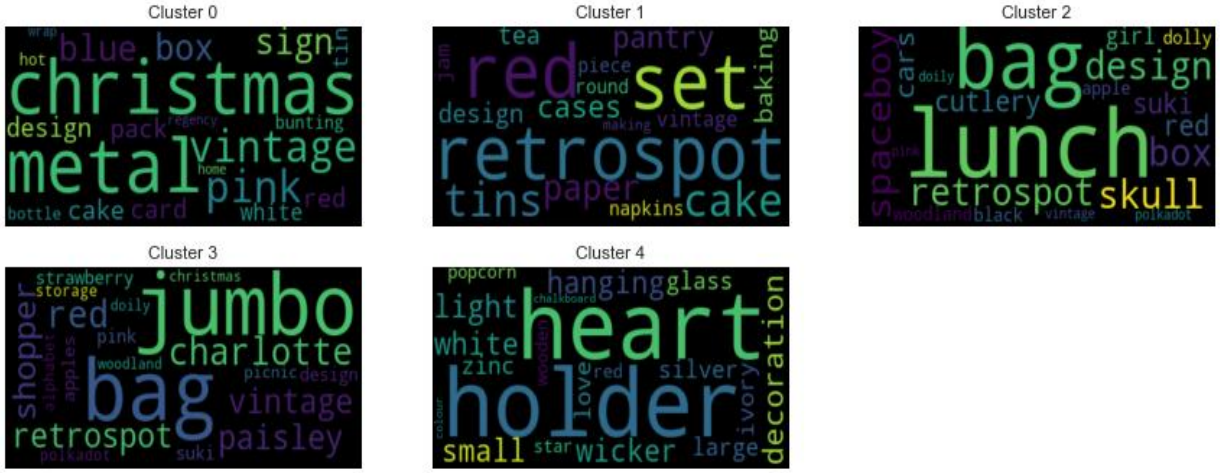
Taking into account the description, we can also see how the products are organized. As well as plotting the words into the space and sizing them differently to represent frequency, we can also see the frequency of the words in a bar plot. We can take a sample of the entire set of data and run the model from there. sample df = df Reset index () df sample. head ['Description'].sample (frac=0.5, random state=1) ().



KMeans
KMeans(n_clusters=5, random_state=0)

Taking into account the description, we can also see how the products are organized. Verify codes with just letters. As well as plotting the words into the space and sizing them differently to represent frequency, we can also see the frequency of the words in a bar plot. We can take a sample of the entire set of data and run the model from there. Order the cluster centres once you have them. Each centroid is transformed into a sorted by the argsort()[:,::-1] line. list of the most important words in each column.

Cluster output with section was presented



8. RECOMMENDER SYSTEM

The goal is to develop a recommender system that can identify potential products that a client might be interested in purchasing. import preprocessing from surprise import and import pandas as pd from sklearn KNNWithMeans of unexpected import import of the surprise. model selection dataset GridSearchCV from unexpected import.

We will keep columns in recommendations based on customer interests. Calculate the rating based on the quantity, customer ID, and View various Ratings while Print the shape, then collect another sample to determine the best settings using GridSearchCV. Create the best predictor possible using the following example: predict the score for customer 1 and stock code 273, then look for stock codes that user 1 hasn't purchased yet. then look up the recommended ratings for each of those items and keep the highest possible.

Out[62]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	cancelled
2875	536743	16012	FOOD/DRINK SPONGE STICKERS	24	12/2/2010 13:32	0.21	17964.0	United Kingdom	0
20595	538888	16012	FOOD/DRINK SPONGE STICKERS	48	12/14/2010 16:23	0.21	17912.0	United Kingdom	0
21555	539036	16012	FOOD/DRINK SPONGE STICKERS	24	12/15/2010 14:35	0.21	14913.0	United Kingdom	0
22525	539213	16012	FOOD/DRINK SPONGE STICKERS	24	12/16/2010 12:23	0.21	12877.0	United Kingdom	0
32982	540672	16012	FOOD/DRINK SPONGE STICKERS	24	1/10/2011 15:51	0.21	15281.0	United Kingdom	0
38463	541500	16012	FOOD/DRINK SPONGE STICKERS	48	1/18/2011 15:25	0.21	18041.0	United Kingdom	0
46974	542586	16012	FOOD/DRINK SPONGE STICKERS	24	1/28/2011 15:45	0.21	12377.0	Switzerland	0
50678	543030	16012	FOOD/DRINK SPONGE STICKERS	24	2/2/2011 15:22	0.21	12437.0	France	0
67679	545053	16012	FOOD/DRINK SPONGE STICKERS	24	2/27/2011 12:41	0.21	17516.0	United Kingdom	0
71477	545514	16012	FOOD/DRINK SPONGE STICKERS	96	3/3/2011 12:06	0.21	15053.0	United Kingdom	0
71478	545514	16012	FOOD/DRINK SPONGE STICKERS	24	3/3/2011 12:06	0.21	15053.0	United Kingdom	0
71917	545550	16012	FOOD/DRINK SPONGE STICKERS	72	3/3/2011 15:11	0.21	18041.0	United Kingdom	0
78882	546415	16012	FOOD/DRINK SPONGE STICKERS	48	3/13/2011 10:27	0.21	17975.0	United Kingdom	0

9. CONCLUSION

The supply chain efficiency model proposed in this paper is ambitious and makes bold claims. "Future research will involve the testing of scientific methods in virtual settings and on actual historical data sets, which will then be tested in actual operations." Even though supply chain issues have been extensively studied in the literature, the assessment shows that, despite a growing trend, problems with the application of new technologies have not been fully investigated. This presents excellent opportunities for future research. A very interesting challenge lies in processing and interpreting the vast amounts of data that the current state of e-commerce distribution operations generates into useful information. For instance, it is simple to measure potential customer actions in some sources, such as email clicks or previews, but it is more difficult to instrument and measure potential customer actions in other sources, such as social networks, where unstructured formats may contain opinions or feelings that affect the audience's perception of the brand. For some businesses, structured data may be sufficient to draw conclusions about them, whereas for others, a comprehensive analysis that includes unstructured data may be necessary to produce meaningful results. The same heterogeneous data condition exists throughout all of the other stages of the distribution operation. Finally, e-commerce logistics optimization is still a huge opportunity for the development of an integrated use of that data. After analysing the findings, we can see that people occasionally view their unhappiness as a positive. People are very hostile and suffer as a result of its negative effects. They decide against the new technology in favour of the old ways. As the population increases, so do businesses and markets. As a result, threats of fraud, theft, and other wrongdoing are made. In actuality, people don't really believe they can be trusted for any given reason. Therefore, it is obvious that going to a store in person is preferable to shopping online. This paper aids in business growth. helps clients' customers who visit the store frequently save a lot of time. convenience for the merchant and saving time for other customers. The knowledge students gain from working on a project, interacting with team members, and receiving mentorship can help improve communication skills, and finishing a project can help increase domain knowledge. Where general computing techniques are less efficient,

machine learning techniques are very helpful. But it is important to note that one method for data analysis is machine learning. This is due to the fact that machine learning is essentially an approach to data analysis that builds the analytical model automatically. "It can help in locating and highlighting any patterns in the data." Additionally, these clever and astute strategies help the business to better serve potential customers and to lead the market with timely product marketing and the introduction of product promotions.

REFERNCES

1. Agatz, N.A., Fleischmann, M., and Van Nunen, J.A. (2008). E-fulfillment and multi-channel distribution–a review. European journal of operational research, 187(2), 339–356.
2. Allen, E. and Fjermestad, J. (2001). E-commerce marketing strategies: an integrated framework and case analysis. Logistics Information Management, 14(1/2), 14–23.
3. Alvarado, U.Y. and Kotzab, H. (2001). Supply chain management: the integration of logistics in marketing. Industrial marketing management, 30(2), 183–198.
4. Arnold, F., Cardenas, I., S¨orensen, K., and Dewulf, W. (2018). Simulation of b2c e-commerce distribution in antwerp using cargo bikes and delivery points. European Transport Research Review, 10(1), 2.
5. Barnett, M. and Alexander, P. (2004). The seven-step model for e-grocery fulfilment. In Building the E-Service Society, 375–394. Springer.
6. Cheung, T.W. and Chanson, S.T. (2001). A pki-based end-to-end secure infrastructure for mobile e-commerce. In International Conference on Formal Techniques for Networked and Distributed Systems, 421–441. Springer.
7. Cheung, W.K.W. (2004). e-transformation technologies: case studies and the road ahead-a value chain perspective. In null, 510–517. IEEE.
8. Delfmann, W., Albers, S., and Gehring, M. (2002). The impact of electronic commerce on logistics service providers. International Journal of Physical Distribution & Logistics Management, 32(3), 203–222.
9. Drexl, M., Rieck, J., Sigl, T., and Press, B. (2013). Simultaneous vehicle and crew routing and scheduling for partial-and full-load long-distance road transport. Business Research, 6(2), 242–264.
10. Frazzon, E.M., Albrecht, A., Hurtado, P.A., de Souza Silva, L., and Pannek, J. (2015). Hybrid modelling approach for the scheduling and control of integrated production and logistic processes along export supply chains.
11. IFAC-PapersOnLine, 48(3), 1521–1526. Galinari, R., Cervieri J´unior, O., J´unior, T., Rodrigues, J., and Rawet, E.L. (2015). Com´ercio eletrˆonico, tecnologias m´oveis e m´ıdias sociais no brasil.
12. Halligan, B. and Shah, D. (2009). Inbound marketing: Get found using google. Social Media and Blogs. Ed. Wiley. Holliman, G. and Rowley, J. (2014). Business to business digital content marketing: marketers perceptions of best practice. Journal of research in interactive marketing, 8(4), 269–293.
13. Isasi, N.K.G., Frazzon, E.M., and Uriona, M. (2015). Big data and business analytics in the supply chain: a review of the literature. IEEE Latin America Transactions, 13(10), 3382–3391.
14. Kaihara, T. and Fujii, S. (2005). Multi-agent based robust scheduling for agile manufacturing. In Emerging Solutions for Future Manufacturing Systems, 201–208 Springer.
15. K¨uck, M., Ehm, J., Freitag, M., Frazzon, E.M., and Pimentel, R. (2016). A data-driven simulation-based optimisation approach for adaptive scheduling and control of dynamic manufacturing systems. In Advanced Materials Research, volume 1140, 449–456. Trans Tech Publ.
16. Leeflang, P.S., Verhoef, P.C., Dahlstr¨om, P., and Freundt, T. (2014). Challenges and solutions for marketing in a digital era. European management journal, 32(1), 1–12. McGarrah, R.E. (1963). Production and Logistics Management: Text and Cases. Wiley.
17. Meng, X. (2009). Developing model of e-commerce emarketing. In Proceedings. The 2009 International Symposium on Information Processing (ISIP 2009), 225. Citeseer.

18. Mentzer, J.T., Stank, T.P., and Esper, T.L. (2008). Supply chain management and its relationship to logistics, marketing, production, and operations management. Journal of Business Logistics, 29(1), 31–46.
19. Novaes, A.G. (2001). Log´ıstica e gerenciamento da cadeia de distribui¸c˜ao: Estrat´egia. Opera¸c˜ao e Avalia¸c˜ao. Rio de Janeiro: Editora Campus.
20. Ogawa, Y., Suwa, H., Yamamoto, H., Okada, I., and Ohta, T. (2008). Development of recommender systems using user preference tendencies: An algorithm for diversifying recommendation. In Towards Sustainable Society on Ubiquitous Networks, 61–73. Springer.
21. Prockl, G., Bhakoo, V., and Wong, C. (2017). Supply chains and electronic markets-impulses for value cocreation across the disciplines. Electronic Markets, 27(2), 135–140.
22. Pyke, D.F., Johnson, M.E., and Desmond, P. (2001). Efulfillment. Supply Chain Management Review, 27(5), 50–62.
23. Ramanathan, R. (2010). The moderating roles of risk and efficiency on the relationship between logistics performance and customer loyalty in e-commerce. Transportation Research Part E: Logistics and Transportation Review, 46(6), 950–962.
24. Rødseth, H., Schjølberg, P., and Marhaug, A. (2017). Deep digital maintenance. Advances in Manufacturing, 5(4), 299–310. doi:10.1007/s40436-017-0202-9. URL https://doi.org/10.1007/s40436-017-0202-9.
25. Skarlat, O., Nardelli, M., Schulte, S., Borkowski, M., and Leitner, P. (2017). Optimized iot service placement in the fog. Service Oriented Computing and Applications, 11(4), 427–443.
26. Varela, M.L.R., Apar´ıcio, J.N., and do Carmo Silva, S. (2005). A distributed knowledge base for manufacturing scheduling. In Emerging Solutions for Future Manufacturing Systems, 323–330. Springer.
27. Velasquez, K., Abreu, D.P., Assis, M.R., Senna, C., Aranha, D.F., Bittencourt, L.F., Laranjeiro, N., Curado, M., Vieira, M., Monteiro, E., et al. (2018). Fog orchestration for the internet of everything: state-of-the-art and research challenges. Journal of Internet Services and Applications, 9(1), 14.
28. Waller, M.A. and Fawcett, S.E. (2013). Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. Journal of Business Logistics, 34(2), 77–84.
29. Wolfinger, D., Tricoire, F., and Doerner, K.F. (2018). A matheuristic for a multimodal long haul routing problem. EURO Journal on Transportation and Logistics, 1–37.
30. Woudhuysen, J. (2001). E-fulfilment: The opportunities for the future: Part one. Interactive Marketing, 2(3), 219–229.
31. Zhao, K. and Wang, C. (2017). Sales forecast in ecommerce using convolutional neural network. arXiv preprint arXiv:1708.07946.