ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 13 (2023) No. 6, pp. 97-110
DOI:10.33168/JSMS.2023.0606

Implementation of K-Means Clustering for Classification of Total Transaction and Seasonal Correlation on Online Retail Shop

Marchello Yoloan[1], Andre Setiawan Wijaya[1], Ford Lumban Gaol[2]

[1] Computer Science Department, BINUS Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480

[1]Computer Science Department, BINUS Graduate Program - Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480

marchello.yoloan@binus.ac.id, andre.wijaya002@binus.ac.id, fgaol@binus.edu

Abstract. E-commerce growth has been increasing rapidly for the past few years, and it's already become an important part for customers or retailers to sell and buy products with no bound of distances. With e-commerce having a lot of transaction data, it will be hard to make the most effective selling strategy by only using bare eyes. Thus, in this study, we adopt k-means clustering technique for clustering analysis to gain useful patterns and insights about total transaction and seasonal correlation on online retail shop dataset with the aim of giving these retail shops a more strategic sales plan. The result shows that different countries had its peak sales in different seasons, these insights can be applied for many e-commerce stores to give a deeper sales strategy to various target market.

Keywords: Data mining, E-commerce, K-Means clustering, sales transaction, clustering analysis, retail shop

1. Introduction

Data mining can be defined as a process that aims to generate knowledge from data and present survey results to users in a comprehensive way (Schuh, et al., 2019). Knowledge generation in the context of data mining can be transformed into discovering new important patterns, relationships, and trends in data that are useful to the user. Data mining has a lot of approaches and one of them is called clustering. Clustering is a well-known data mining approach for identifying relevant patterns in a huge database of data. But, because data sets in data mining typically comprise categorical values, working primarily with numeric values limits its application in data mining (Ali & Kadhum, 2017).

This type of approach can help us classify groups of data based on its attributes. In this case, we want to classify a dataset about an online retail shop. Online shops have an advantage when compared to offline stores, where the desire to avoid the bigger drawbacks of the alternative may drive the decision to shop online or in store, rather than the perceived advantages of one channel over the other (Harris, Riley, Riley, & Hand, 2017).

During this pandemic, a lot of our behaviors and daily routine changed to adapt. A study concluded that consumer behavior has been significantly disrupted because of the lockdown and social separation used to battle the covid-19 virus. All consumption is constrained by time and location. Consumers have learned to adapt in new and novel ways due to time flexibility but location rigidity. People now work, study, and relax at home, blurring the lines between work and personal life. Because the customer is unable to visit the store, the store must travel to the customer (Sheth, 2020).

A study by Rose and Dolega (2021) found that weather is a crucial impact in many facets of retail sector decision-making and is regarded as an influential factor on consumer purchase behavior. As a result, many retailers and other stakeholders will benefit from a greater knowledge of the scale and nature of the impact of varied UK weather conditions (Rose & Dolega , 2021). Knowing that there are hundreds of thousands of unprocessed data and to analyze it manually would be inefficient and time consuming, thus, in this study, we will implement k-means clustering to group countries' transactions according to its season to gain insights whether a particular season affect buyer's interest

The problems that we like to solve on this paper is:

1. Classifying customers from countries based on their spending rates and finding out which have the most customers with above average spending.
2. Finding out if seasons have any correlation with customers' spending.

The limitations of our study as follow:

1. We focus only on the k-means clustering algorithm.
2. The dataset that we used only from Carrie (2017)
3. We focus only classify large datasets to help find out which countries have the most customers with above spending.
4. We focus on four seasons for factors in customers' spending.

2. Literature Review

2.1. Data Mining

Data mining, also called knowledge discovery in databases, is used to get useful insights from analysing large amounts of data. These insights are usually in the form of various patterns or relationships between each data that haven't been seen before. Data mining helps users to create predictive or descriptive models that can be generalized into new data (Chen, et al., 2015). There are 4 types of data mining categories based on its models: clustering, classification, regression, association (Siguenza-Guzman, Saquicela, Avila-Ordóñez, Vandewalle, & Cattrysse, 2015). The detail of data mining methods shown on Figure 1.

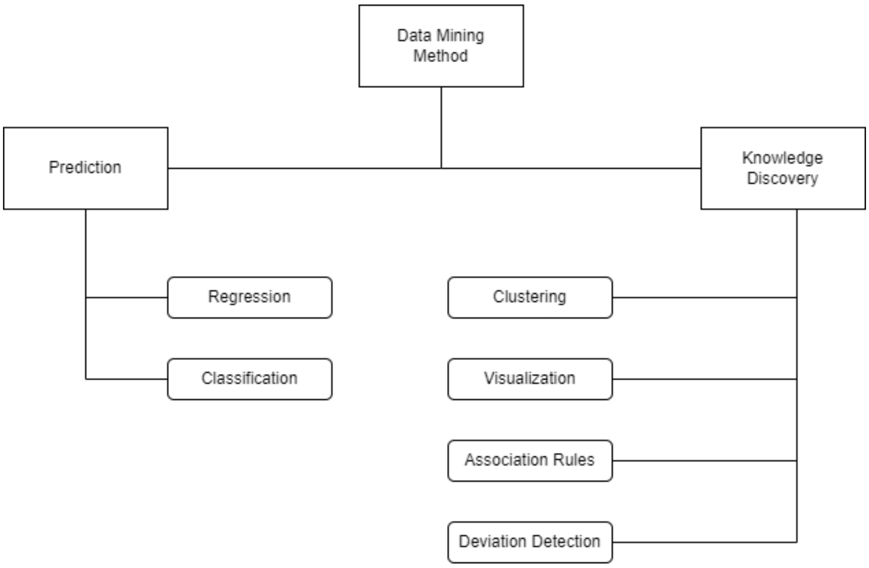


Fig.1: Data mining methods (Chen, et al., 2015)

Clustering is used to group similar data items into a cluster. Cluster analysis is the construction of a set of patterns into a cluster based on the similarity they had, Clustering is suitable for grouping unlabeled data, explaining pattern-analysis, decision making, data mining, and pattern classification (S & E, 2014).

Classification is used to group a class of objects within their characteristics (Oleiwi & Adebayo, 2019). Classification is used to classify each item in a data set into a predetermined set of groups or classes. The purpose of using classification is to precisely forecast each case in the data to their respective target class. For example, classification models can be used to identify loan applicants with a classification of low, medium, and high credit risk (Kesavaraj & Sukumaran, 2013).

Regression analysis is used to explain the relationship between one dependent variable and one or more independent variables. Independent variables in data mining are the attributes that are already known, on the other hand, the dependent variables are variables that's about to be predicted (Ramageri, 2010).

Association is used to discover relations that look unrelated in the dataset / database. An example statement of association rule is such: “If a customer buys a dozen eggs, he is 80% likely to also purchase a milk”. There are two parts in association rules. The antecedent part (if) is the item that can be found in the data, and the consequent part (then) is the item that can be found in the combination of the antecedent item. The purpose of association is to gain insights from a large amount of data. For example, the association rule can find information of a customer who buys a keyboard and tends to buy a mouse at the same time (Singh & Jassi, 2017).

The stages of data mining as shown on Figure 2 usually consist of (Susanto & Meiryani, 2019):

1. Data cleaning: steps where noisy datas are deleted
2. Data integration: putting broken data source together
3. Data selection: return the data related analysis to the database
4. Data transformation: transforming data into the right format
5. Knowledge discovery: process of extracting data patterns using intelligent methods
6. Pattern evolution: explain attractive pattern related to interesting behavior that contain useful information
7. Knowledge presentation: presenting interesting information that has been obtained using visualization.

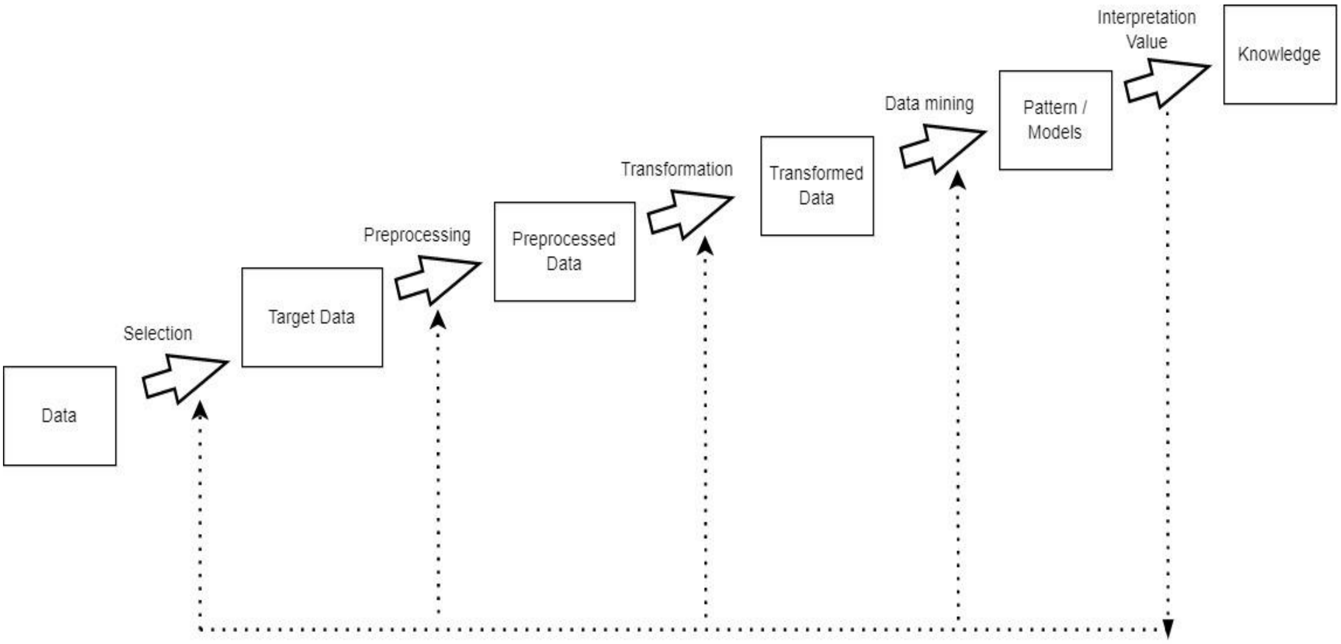


Fig.1: Data mining process (Kesavaraj & Sukumaran, 2013).

2.2. Clustering

Pande et al. (2012) defines clustering as a process of categorizing a set of objects (often represented as points in a multidimensional space) into groups of related things. In data analysis, cluster analysis is a critical tool. It is a set of approaches for automatically clustering a group of patterns based on their similarity. Patterns belonging to the same cluster appear to be more similar than patterns belonging to separate clusters. The distinction between clustering (unsupervised classification) and supervised classification must be understood (Pande, Sambare, & Thakre, 2012).

Data cluster appraisal is a necessary step in the process of discovering knowledge and mining data. Clustering can be done in an unsupervised, semi-supervised, or supervised fashion (Hossain, Akhtar, Ahmad, & Rahman, 2019). All clustering approaches use the same concept of classifying cluster centers to represent each cluster. In this paper we are going to do clustering in an unsupervised method, where it deals with the finding of a structure in unlabeled data sets (Faizan, Zuhairi, Ismail, & Sultan, 2020).

The following as shown on Figure 3 are some of the most common uses for unsupervised learning:

- Aggregate variables with comparable qualities to simplify datasets.
- Finding anomalies that don't fit into any of the categories.
- Datasets are segmented based on some common characteristics.

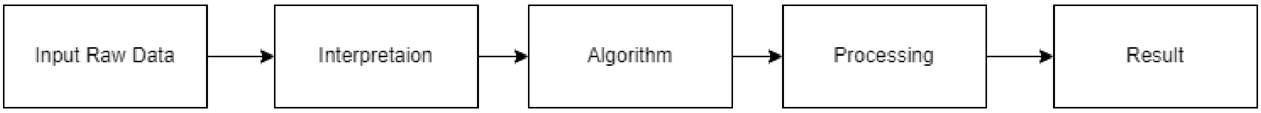


Fig. 2: Unsupervised learning model (Pande, Sambare, & Thakre, 2012).

All clustering approaches use the same concept of classifying cluster centers to represent each cluster. K-means clustering is a cluster analysis method that involves viewing and splitting data points into k clusters, with each observation belonging to the mean cluster closest to it (Faizan, Zuhairi, Ismail, & Sultan, 2020). The illustration of clustering example is shown on Figure 4.

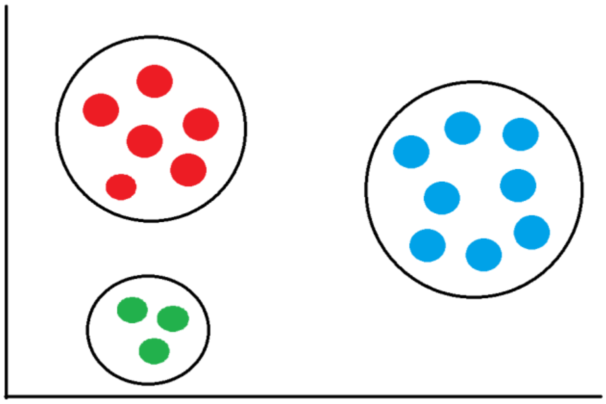


Fig.3: Clustering example (Faizan, Zuhairi, Ismail, & Sultan, 2020).

Because the clustering method is subjective, it is a versatile tool that can be utilized to achieve a variety of goals. Every methodology adheres to a set of rules and regulations that govern how data points are similar. There are often over 100 clustering techniques available. However, only a few of these algorithms are widely employed. The following as shown on Table 1 are some of the clustering approaches (Faizan, Zuhairi, Ismail, & Sultan, 2020).

Table 1. Clustering Methodologies

Typical Clustering Methodologies	
Method	**Algorithm**
Distance-based method	• "K-means, K-medians, K-medoids" are partitioning algorithms. • Hierarchical algorithms, "Agglomerative, Divisive method." These algorithms are extremely simple to learn and run iteratively to identify the local optima, however they lack scalability when dealing with huge datasets.
Grid-based method	• Individual sections of the data space are generated into a grid-like structure using a grid-based method. These methods divide the entire issue domain into cells using a single consistent grid mesh. A collection of statistical qualities from the items is used to represent the data objects within a cell.
Density-based method	• DBSCAN stands for Density-Based Spatial Clustering of Applications with Noise. • OPTICS order points to determine clustering structure These algorithms search the data space for places with various densities of data points. Within the same cluster, it isolates different density zones and distributes data points to those regions.

Probabilistic and generative models	• Modeling data from a generative process using the expectation-maximization algorithm. Over-fitting is a common problem with these models. The Expectation-Maximization algorithm, which uses normal multivariate distributions, is a well-known example of such models.

2.3. K-Means Clustering

The K-Means clustering algorithm was invented by Mac Queen in 1967. The K-Means clustering algorithm divides data into k groups using a partitioning clustering method (Dunham, 2006). Because the initial cluster center may shift, the algorithm requires accurate values for determining the number of clusters k, as this event may result in unstable data grouping (Likas, Vlassis, & Verbeek, 2003).

The K-means clustering algorithm works by using 2 processes. First process is randomly selecting the k center, which k has a fixed value. Second process is choosing each data object to the nearest center. K-means algorithms usually also use euclidean distance to calculate distances between data objects and the center of clusters. Initial grouping is finished if several clusters already have all data objects. Then the algorithm will recalculate the average of the initial formed cluster. This process is repeated until the creation function is minimized.

Assuming that target object is x and x_i shows the mean of cluster c_i, the criterion function will be:

$$E = \sum_{i=1}^{k} \sum_{x \in Ci} | x - x_i |^2 \quad (1)$$

E equals the sum of squared error of all objects in the database, and criterion function distance is equal to euclidean distance. Euclidean distance of vector $x = (x_1, x_2,.. x_n)$ and vector $y=(y_1,y_2,..y_n)$, or can be written as $d(x_1,y_2)$ can be calculated with this formula (Na, Xumin, & Yong, 2010):

$$d(x_i, y_i) = [\sum_{i=1}^{n} (x_i - y_i)^2]^{\frac{1}{2}} \quad (2)$$

The most significant advantage of the K-Means algorithm in data mining applications is its efficiency in clustering large data sets. K-means and its different variants have a computation time complexity that is linear in the number of records but is assumed to discover inferior clusters (Faizan, Zuhairi, Ismail, & Sultan, 2020).

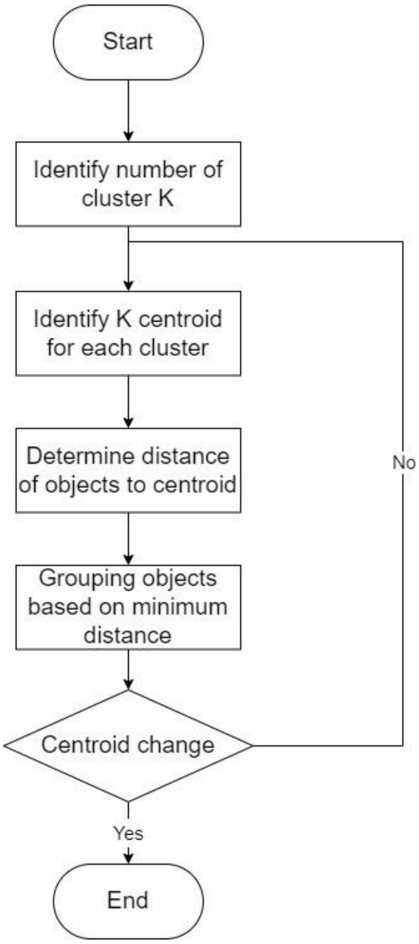

Start
Identify number of cluster K
Identify K centroid for each cluster
Determine distance of objects to centroid
Grouping objects based on minimum distance
Centroid change
No
Yes
End

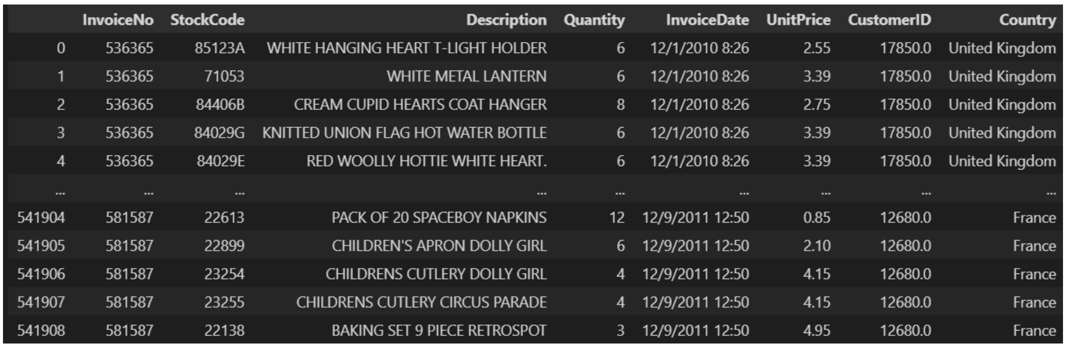
Fig. 4: K-Means Clustering Algorithm

3. Research Methodology

The Research methodology that we used base on K-Means clustering algorithm as shown on Figure 5.

3.1. Data extraction

We first read the csv data from Carrie (2017) on jupyter notebook, and the initial data looks as shown on Figure 6.



	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
...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	12/9/2011 12:50	0.85	12680.0	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	12/9/2011 12:50	2.10	12680.0	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	12/9/2011 12:50	4.15	12680.0	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	12/9/2011 12:50	4.15	12680.0	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	12/9/2011 12:50	4.95	12680.0	France

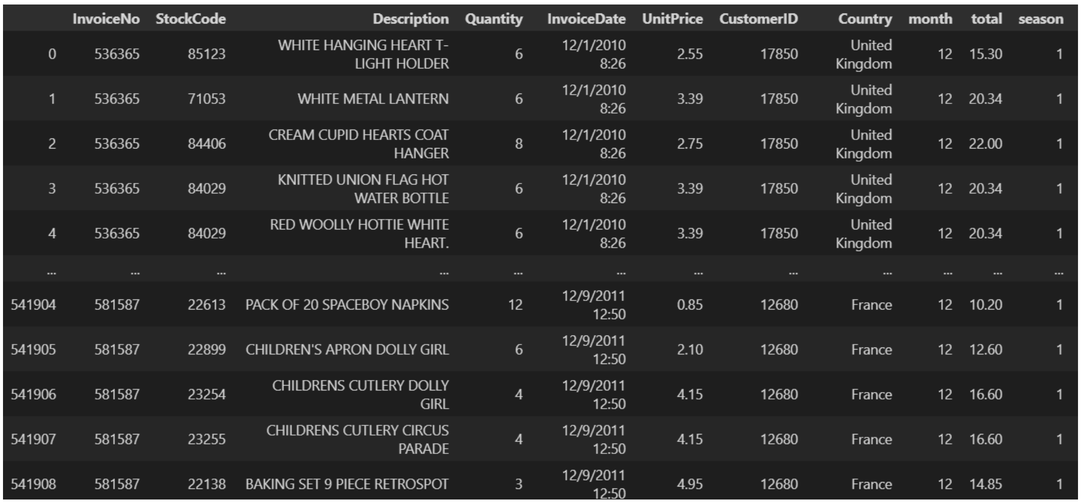
Fig.5: Initial Dataset

3.2. Data Cleaning

We clean the datasets that will hinder the clustering process such as dropping rows with empty values and dropping rows with values that are too extreme for clustering. After the data is cleaned, the number of rows in the dataset dropped from 541909 rows to 396500 rows.

3.3. Feature Selection

The first feature that we want to take is total price, which is taken from the multiplication between the Quantity and UnitPrice column so we can figure out how much the item costs during each transaction. The second feature is to find out what season did the customers spend the most and the country of origin of the buyer, this feature can be taken from the InvoiceDate column. Figure 7 is shown with after creating total and seasons columns.



	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	month	total	season
0	536365	85123	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850	United Kingdom	12	15.30	1
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850	United Kingdom	12	20.34	1
2	536365	84406	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850	United Kingdom	12	22.00	1
3	536365	84029	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850	United Kingdom	12	20.34	1
4	536365	84029	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850	United Kingdom	12	20.34	1
...	...	...	...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	12/9/2011 12:50	0.85	12680	France	12	10.20	1
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	12/9/2011 12:50	2.10	12680	France	12	12.60	1
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	12/9/2011 12:50	4.15	12680	France	12	16.60	1
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	12/9/2011 12:50	4.15	12680	France	12	16.60	1
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	12/9/2011 12:50	4.95	12680	France	12	14.85	1

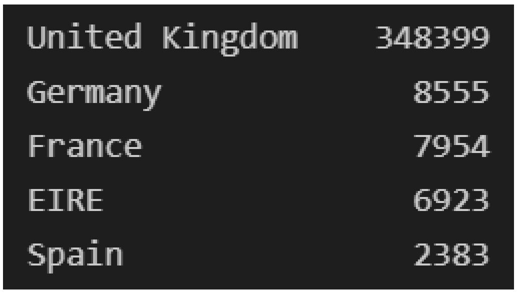
Fig.7: After creating total and season columns

3.4. Transformation

After applying K-Means Clustering on the dataset with 2 features: total and season and using 3 clusters based on the elbow method, we got the result: first cluster with the below average spending has 388850 transactions, second cluster with the average spending has 7406 transactions, and third cluster with above average spending has 244 transactions. On the third cluster, we found out that the top 5 countries are the United Kingdom (222), Ireland (6), Spain (5), Netherlands (4), and Japan (3) which is the cluster with high spending per transaction.

4. Results and Discussion

After the transformation process, we found out that on all the clusters, all top 5 countries' transactions in those clusters spiked during season fall. Because most of these transactions are from the United Kingdom, we display each cluster with two graphs, one graph where we include the United Kingdom and one graph where we exclude the United Kingdom, so that we can show other countries' transactions more clearly. This result also has a high accuracy with the silhouette score of 0.91 as shown on Figure 8.



```
United Kingdom    348399
Germany             8555
France              7954
EIRE                6923
Spain               2383
```

Fig.8: Top 5 Countries in first cluster

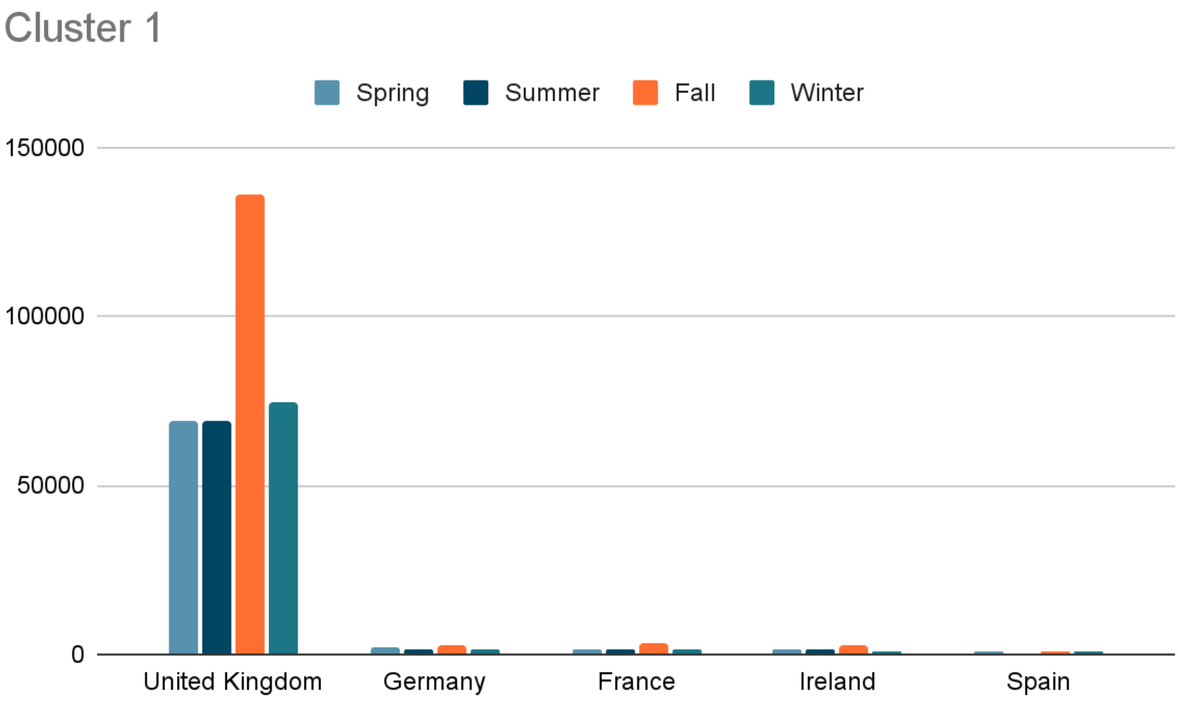


Fig.9: Top 5 countries in first cluster based on season

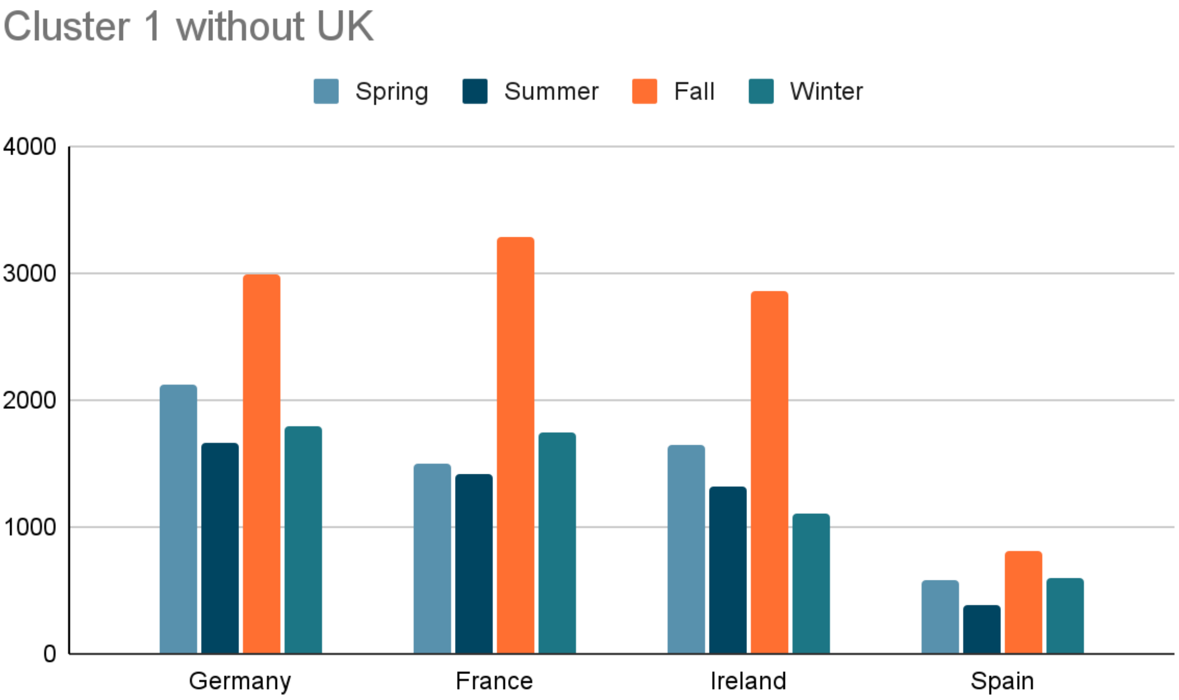
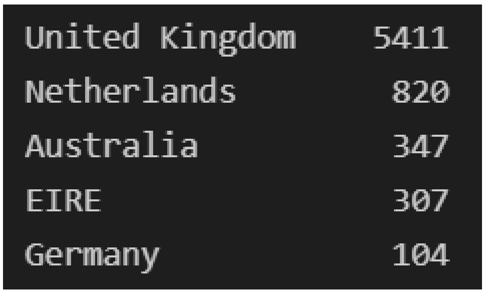


Fig.10: Top 5 countries in first cluster based on season without United Kingdom

From figure 9 and 10, result shows that in this cluster, United Kingdom, Germany, France, Ireland, and Spain has its peak transaction at fall season. France, Spain, and United Kingdom second peak transaction is at winter season, on the other hand Germany, and Ireland has its second best total transaction at spring season.



```
United Kingdom    5411
Netherlands        820
Australia          347
EIRE               307
Germany            104
```

Fig.11: Top 5 countries in second cluster

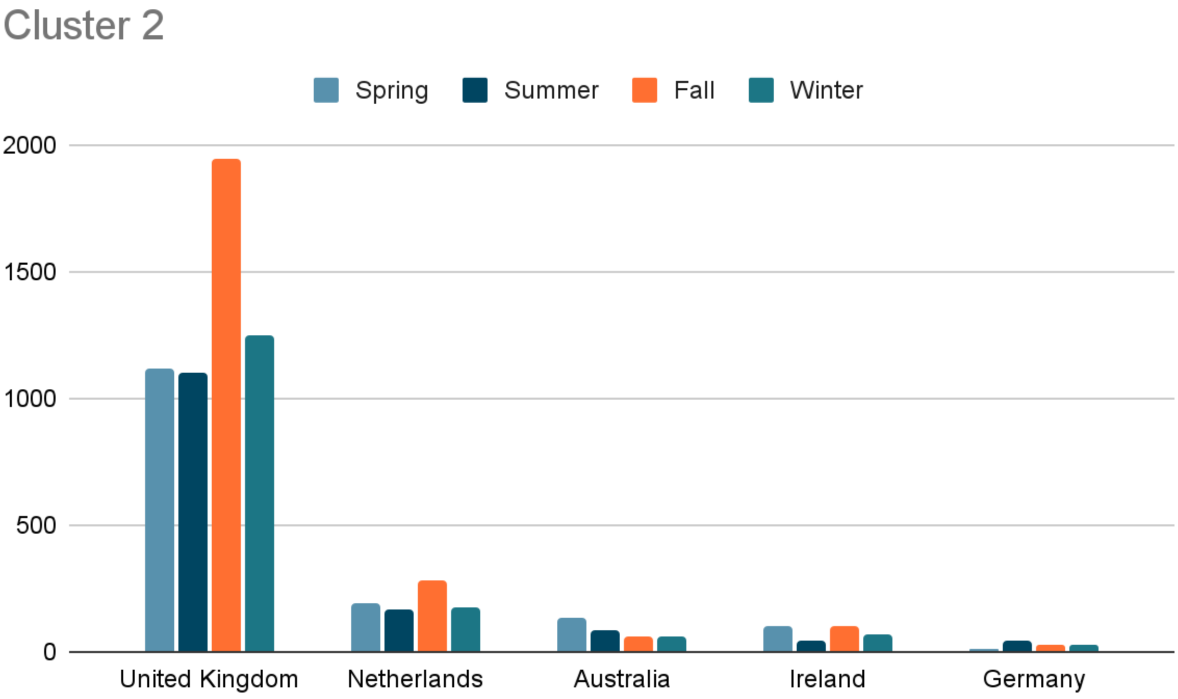


Fig.12: Top 5 countries in second cluster based on season

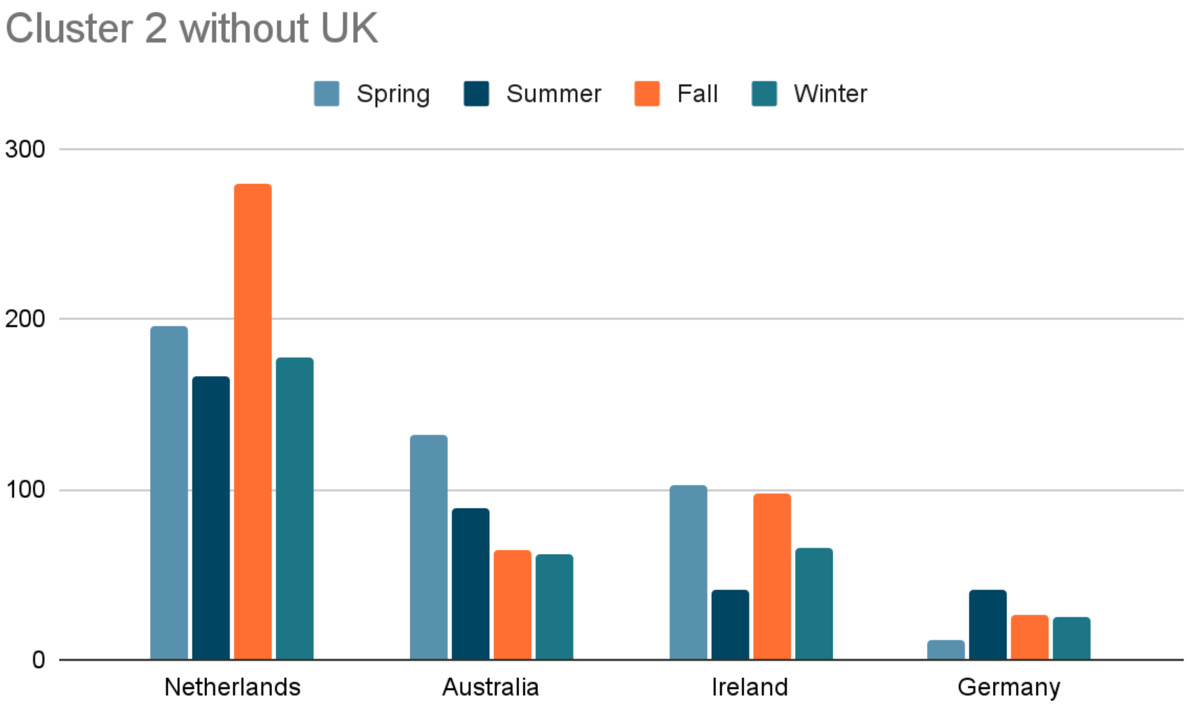
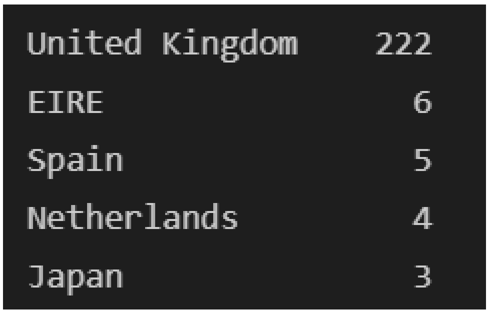


Fig.13: Top 5 countries in second cluster based on season without United Kingdom

Based on Figure 11, Figure 12, and Figure 13, the cluster shows that each country peak at various season. United Kingdom and Netherlands peaked at fall season, Australia and Ireland peaked at spring season, and Germany peaked at summer season.



```
United Kingdom    222
EIRE                6
Spain               5
Netherlands         4
Japan               3
```

Fig.14: Top 5 countries in third cluster

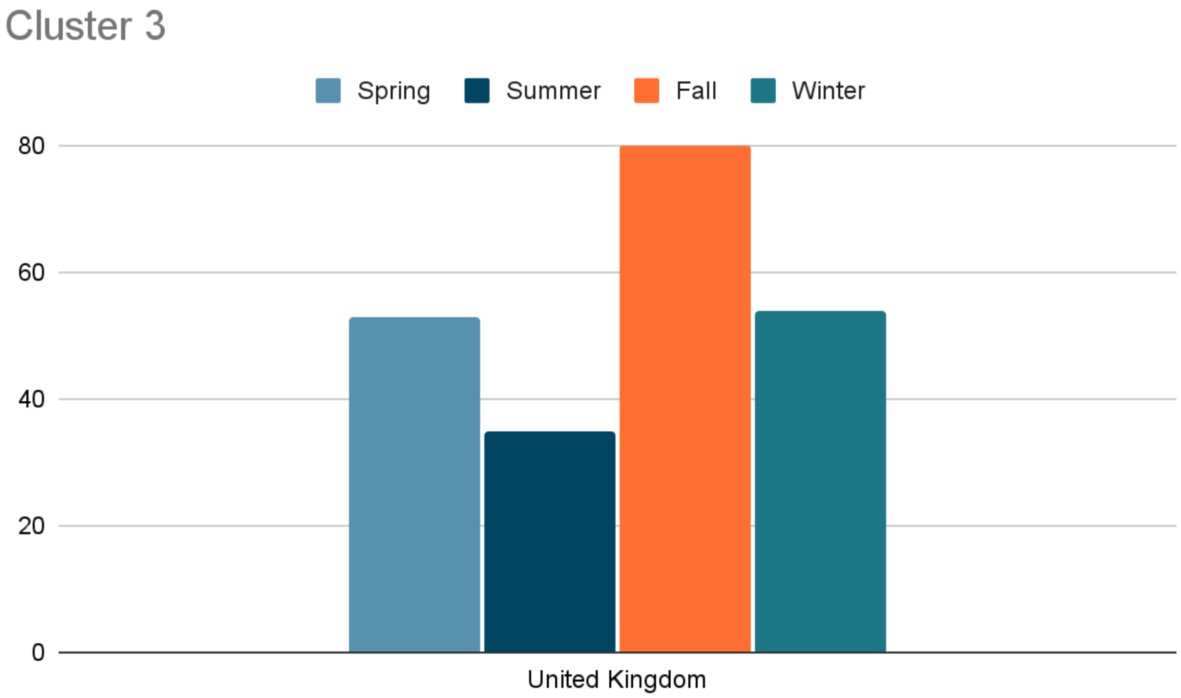


Fig.15: Third cluster based on season

Because other countries besides United Kingdom have too little transaction in this cluster (6 transactions from Ireland, 5 transactions from Spain, 4 transactions from Netherlands, and 3 transactions from Japan), we decided not to put them into the graph. The detail of results are show on Figure 14 and Figure 15.

5. Conclusion

Based on the implementation of k-means clustering on online retail dataset above with the process of data extraction, feature selection, data cleaning, and transformation, we can conclude that there are 3 parts of clusters with the help of total transaction and season features. The first cluster is the cluster that consists of countries with the highest spending transactions, the second cluster is the cluster that consists of countries with average spending transactions, and the third cluster is the cluster that consists of countries with the lowest spending transactions. With the help of k-means clustering, we also know the top countries with the highest total transaction for each cluster. In the first cluster, the country that has the most spending is the United Kingdom with a total of 348399 transactions. In the second cluster, the country that has the most spending is the United Kingdom with a total of 5411 total transactions, and in the cluster, the country with the most spendings is the United Kingdom with a total of 222 transactions. Additionally, using the result of k-means clustering, and inputting the results into the graphs, help us know that certain seasons affect the behavior of customers in each country, as we can see that transactions of each country will peak at various season depending on the customer spendings, this information can be used for online shop to prepare their sales strategies in advance, so they can make various plans depending on the season and country, for example utilizing discount promos to boost sales at its peak season, or other promotion strategies to boost their sales in seasons that have low transaction. The country that has the most record of transactions is the United Kingdom this could possibly happen because the dataset that being used in this study comes from a shop based in the United Kingdom.

6. Future Works

In future work, it will be beneficial for researcher to use large dataset from e-commerce that have category column, and product column in its dataset so researcher can cluster what categories, and items which are frequently sold at certain season. Also, by having category, and product columns in the dataset, implementing algorithms like apriori algorithm can be practical to find itemsets that are frequently sold in the dataset, to give the retailer much more detailed strategic sales plan.

References

Ali, H. H., & Kadhum, L. E. (2017). K-means clustering algorithm applications in data mining and pattern recognition. *International Journal of Science and Research (IJSR), 6*(8), 1577-1584. doi:10.21275/ART20176024

Carrie. (2017, August 17). *E-Commerce Data*. Retrieved November 16, 2021, from Kaggle: https://www.kaggle.com/carrie1/ecommerce-data

Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A. V., & Rong, X. (2015). Data mining for the internet of things: literature review and challenges. *International Journal of Distributed Sensor Networks, 11*(8), 431047. doi:10.1155/2015/431047

Dunham, M. H. (2006). *Data mining: Introductory and advanced topics.* Pearson Education India.

Faizan, M., Zuhairi, M. F., Ismail, S., & Sultan, S. (2020). Applications of Clustering Techniques in Data Mining: A Comparative Study. *ALGORITHMS, 11*(12). doi:10.14569/ijacsa.2020.0111218

Harris, P., Riley, F. D., Riley, D., & Hand, C. (2017). Online and store patronage: a typology of grocery shoppers. *International Journal of Retail & Distribution Management.* doi:10.1108/ijrdm-06-2016-0103

Hossain, M. Z., Akhtar, M. N., Ahmad, R. B., & Rahman, M. (2019). A dynamic K-means clustering for data mining. *Indonesian Journal of Electrical engineering and computer science, 13*(2), 521-526. doi:10.11591/ijeecs.v13.i2.pp521-526

Kesavaraj, G., & Sukumaran, S. (2013). A study on classification techniques in data mining. *2013 fourth international conference on computing, communications and networking technologies (ICCCNT),* 1-7.

Likas, A., Vlassis, N., & Verbeek, J. (2003). The global k-means clustering algorithm. *Pattern recognition, 36*(2), 451-461. doi:10.1016/s0031-3203(02)00060-2

Na, S., Xumin, L., & Yong, G. (2010). Research on k-means clustering algorithm: An improved k-means clustering algorithm. *In 2010 Third International Symposium on intelligent information technology and security informatics*. doi:10.1109/IITSI.2010.74

Oleiwi , A. A., & Adebayo, A. O. (2019). ata Mining Application Using Clustering Techniques (K-Means Algorithm) In The Analysis Of Student's Result. *Journal of Multidisciplinary Engineering Science Studies, 5*(5), 2587-2593.

Pande, S. R., Sambare, M. S., & Thakre, V. M. (2012). Data clustering using data mining techniques. *International Journal of advanced research in computer and communication engineering, 1*(8), 494-9.

Ramageri, B. M. (2010). DATA MINING TECHNIQUES AND APPLICATIONS. *Indian Journal of Computer Science and Engineering, 1*(4), 301-305.

Rose, N., & Dolega , L. (2021). It's the Weather: Quantifying the Impact of Weather on Retail Sales. *Applied Spatial Analysis and Policy,* 1-26.

S, M., & E, M. (2014). An Analysis on Clustering Algorithms inData Mining. *International Journal of Computer Science and Mobile Computing, 3*(1), 334-340.

Schuh, G., Reinhart, G., Prote, J. P., Sauermanna, F., Horsthofer, J., Oppolzer, F., & Knoll, D. (2019). Data mining definitions and applications for the management of production complexity. *Procedia CIRP, 81*, 874-879.

Sheth, J. (2020). Impact of Covid-19 on consumer behavior: Will the old habits return or die? *Journal of business research, 117*, 280-283.

Siguenza-Guzman, L., Saquicela, V., Avila-Ordóñez, E., Vandewalle, J., & Cattrysse, D. (2015). Literature review of data mining applications in academic libraries. *The Journal of Academic Librarianship, 41*(4), 499-510.

Singh, G., & Jassi, S. (2017). A Review Paper: A Comparative Analysis on Association Rule Mining Algorithms. *International Journal of Recent Technology and Engineering, 6*(2), 1-3.

Susanto, A., & Meiryani. (2019). Functions, Processes, Stages And Application Of Data Mining. *International Journal of Scientific & Technology Research, 8*(7), 136-140.