# Contents

# Segmenting Customers via K-means Clustering: a machine learning application in online retail.

## 1.Introduction:

Knowing customers is one of the most fundamental aspects in a business, particularly when customers' habits and preferences are dynamic (Saumendra & Janmenjoy, 2022).

Thus, extracting insights from customers can help businesses enhance their performance. In order to do that, training a model to automatically classify customers into different segments is imperative. The goal is to assist businesses in understanding customers and hence to tailor appropriate marketing content and pricing strategies accordingly, and to improve customer experience.

There is a plethora of articles researching about machine learning application in classifying customer into segments, specifically with K-means clustering (Kumar, 2023; Gandiki *et al*., 2022; Yadegaridehkordi *et al*., 2021; Monil *et al*., 2020; Alkhayrat *et al*., 2020; Ahani *et al*., 2019), while few articles focus on regressions such as Multivariate Linear Regression and Logistic Regression (Ozan, 2018), or decision tree (Yadegaridehkordi *et al*., 2021). Generally, the method selection is adopted depending on the dataset and its context.

Define a learning problem: The task of implementing a model to automatically classify customers into different segments is based on the online retail dataset, the problem is an **unsupervised learning problem** because the focal point is to find patterns or similarities in a group of customers without having any predefined categories or labelled data indicating which segment each customer belongs to.

This report will, therefore, analyse a systematic approach based on machine learning classification with unsupervised learning methods to gain customers' hindsight, and to enhance automation in classifying customers into segments.

Data & Source Code: [dt8-nguyen / 17037529_UFCFMJ-15-M_MACHINE LEARNING AND PREDICTIVE ANALYTICS · GitLab (uwe.ac.uk)]

# 2. Dataset overview:

The data is collected from Kaggle. The dataset consists of 8 attributes regarding customers' purchase history and information, particularly:

- Invoice number, Stock Code, Description of the stock, Quantity, Date of Invoice, Price per unit, customers' ID, and Country.

In addition, there are 541909 observations (samples) in total.

Three fundamental techniques were executed to explore the data: detecting, sorting out missing values and outliers to ensure data cohesion, and adding data columns based on existing data to gain more meaningful insights.

## 2.1. Data Preparation & Quality:

Data preparation:
Missing value:

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

Outliers:

```
Outliers: Quantity         36.0
UnitPrice          8.5
CustomerID     17905.0
Name: 0.95, dtype: float64
```

## 2.2 Data Quality & Limitation:

The dataset seemingly lacks essential values such as age, and income from the customers because with these types of values, it can necessarily help us see more patterns, such as people with a particular age range normally buy this product; or people with a particular range of income rarely purchase another particular product.

Therefore, in order to extract further meanings from the data, the dataset is added two columns: turnover (based on Quantity * Price per Unit) and purchase frequency (based on CustomerID & Invoice Number)

## 2.3. Central tendencies of the dataset:

+) The mean (average value):

```
The mean of Quantity: 9.55224954743324
The mean of Unit Price: 4.611113626088513
```

Quantity (the average number of items purchased): 9.5
Price per unit (the average cost of a single item): 4.6
-> Turnover average: 43.7

+) Standard deviation:

```
Std of Quantity: 218.08115785023384
Std of Unit Price: 96.75985306117938
```

-> This means that there is a wide range of variation in the quantity of purchasing items, and in the prices for each item

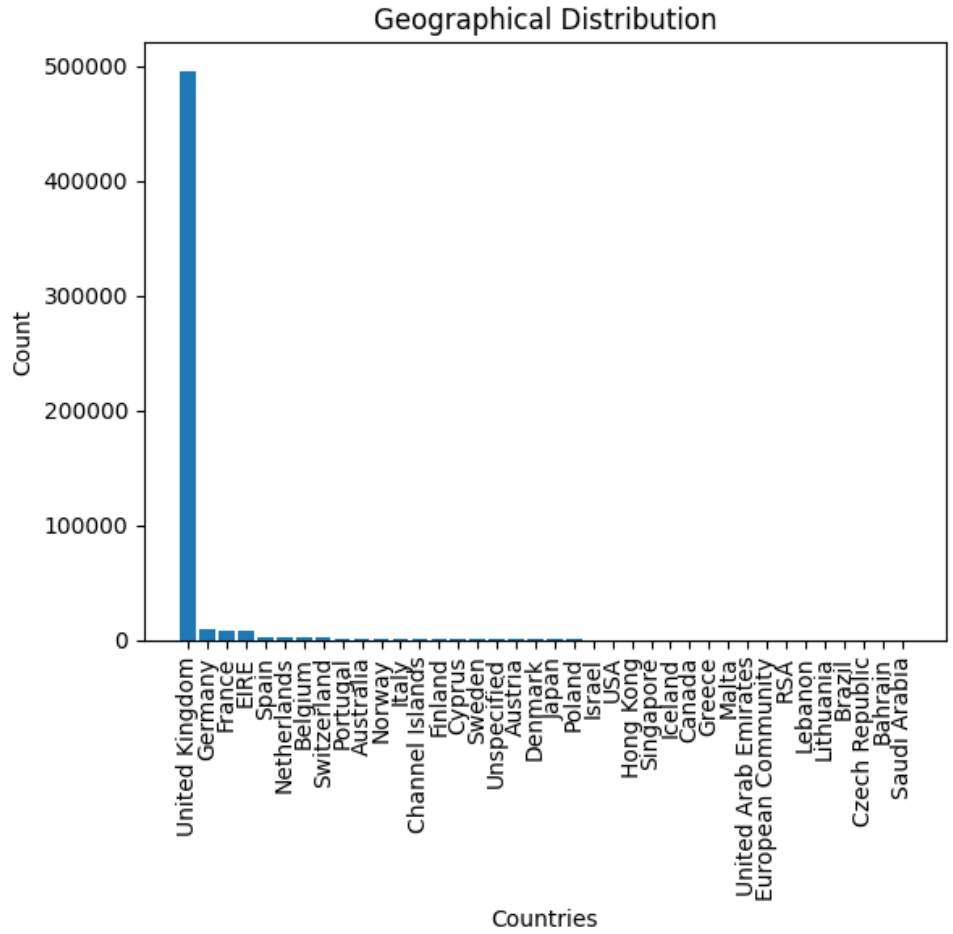
+) The spread between the min and the max value:

The min and max value of Invoice Date

```
Min Invoice Date 2010-12-01 08:26:00
Max Invoice Date 2011-12-09 12:50:00
```

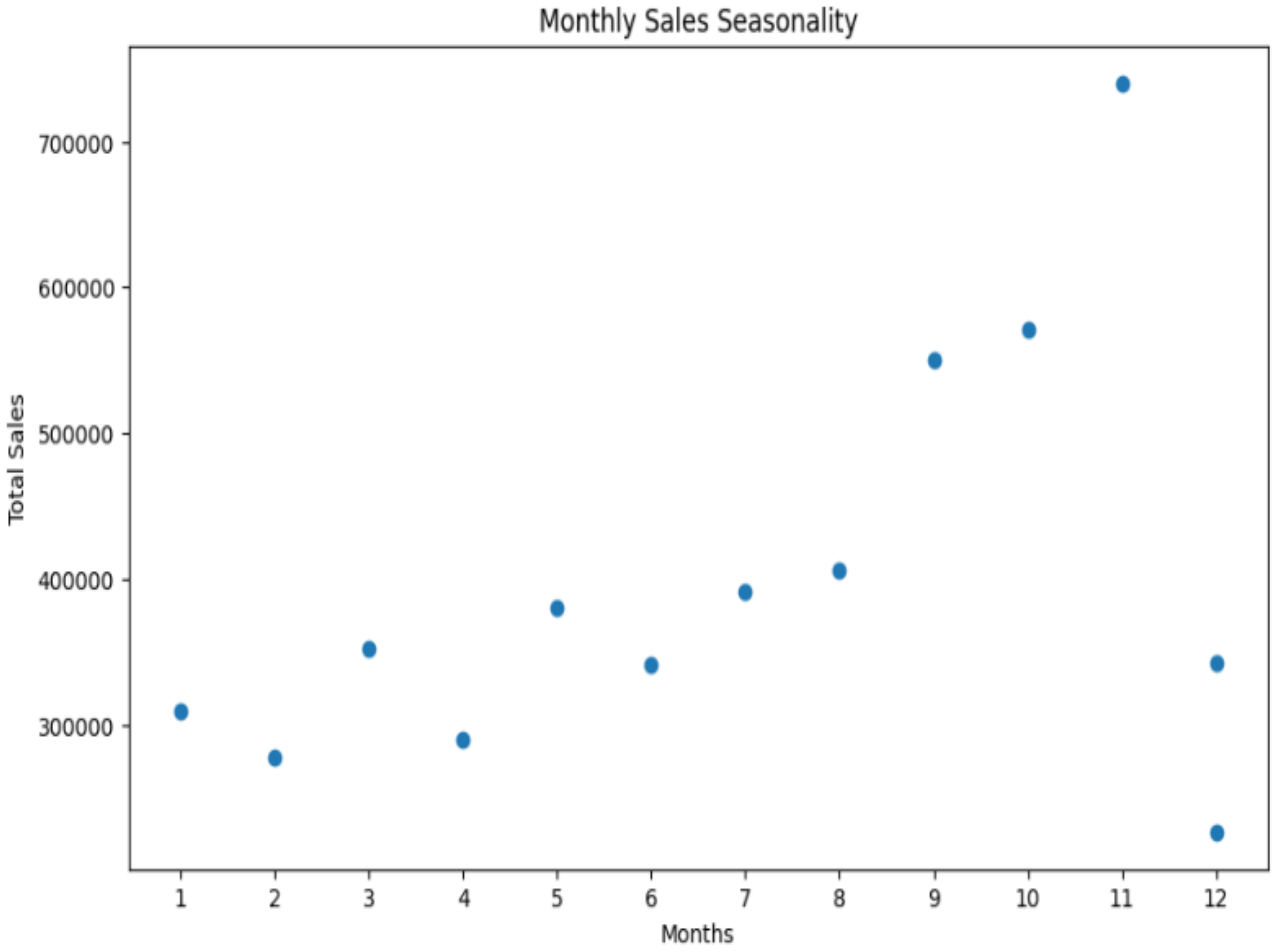
-> it shows roughly one year period of operating the retail

Combining with the mean value. It indicates that over one year period, the average turnover of an online retail is: 43.
+) Other central tendencies of the dataset:
Geographical distribution




This shows that the UK is one of the most important markets for online retail compared to other European countries due to the highest number of customers from the UK.

Seasonality: a significant increase or decrease of sales in any particular month (based on quantity * price per unit, month from the invoice date attributes).

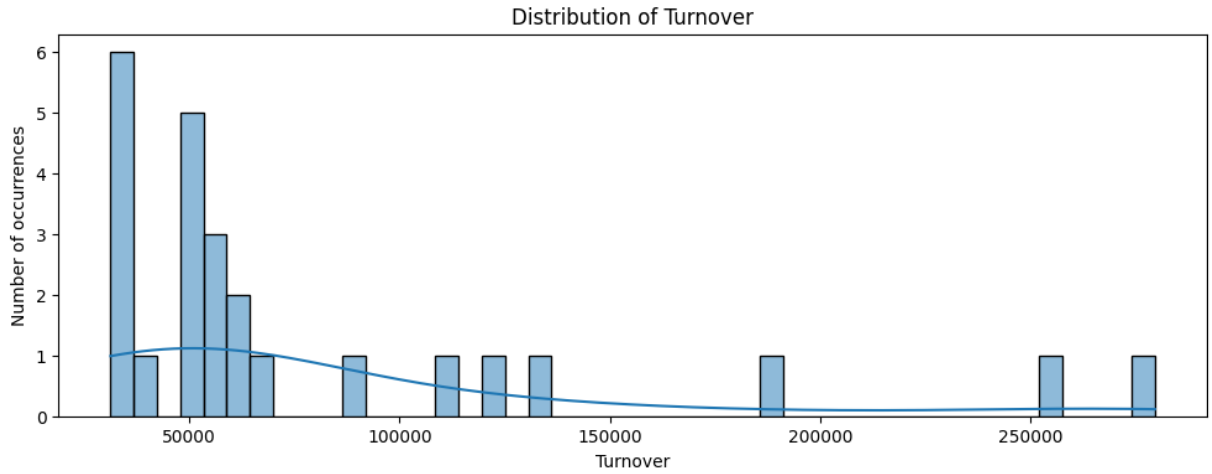



There is a clear upward trend of sales from January, especially till November. This noteworthy point suggests that customers may buy more during special events such as Black Friday, Cyber Monday and the sales drop dramatically afterwards.

Sort out the most valuable customer based on their ID & turnover:

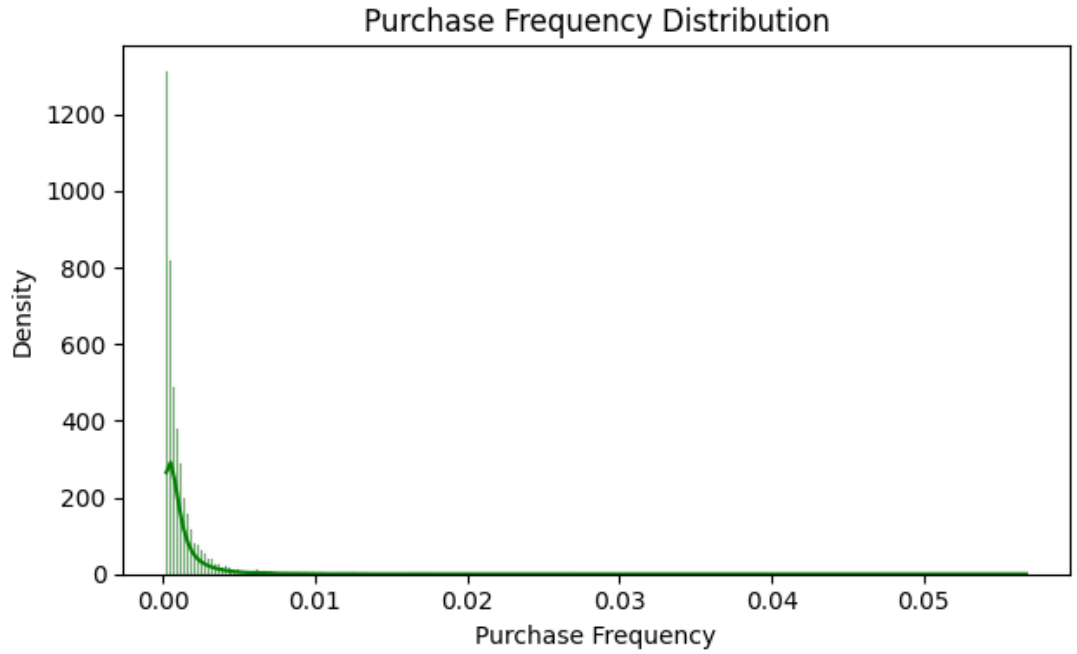
```
Top 25 CustomerIDs based on Turnover: CustomerID
14646.0    279489.02
18102.0    256438.49
17450.0    187482.17
14911.0    132572.62
12415.0    123725.45
14156.0    113384.14
17511.0     88125.38
16684.0     65892.08
13694.0     62653.10
15311.0     59419.34
13089.0     57385.88
14096.0     57120.91
15061.0     54228.74
17949.0     52750.84
15769.0     51823.72
16029.0     50992.61
14298.0     50862.44
14088.0     50415.49
17841.0     40340.78
13798.0     36351.42
16422.0     33805.69
12931.0     33462.81
16013.0     33366.25
15838.0     33350.76
17389.0     31300.08
Name: Turnover, dtype: float64
```

Turnover Distribution:




The bar exhibits a positively skewed distribution suggesting that there is an unequivocal group of customers who spend less money than others.

Purchase Frequency Distribution (based on Customer ID, Invoice Number):




The graph shows a positively skewed distribution indicating that there is a certain group of customers who make more frequent purchases than others.

## 2.4. Interesting Findings:

### 2.4.1. Variance pairs:

1. Quantity & Price per Unit

```
Variance: Quantity     47559.391409
UnitPrice     9362.469164
dtype: float64
```

There are relatively high variances among Quantity & Price per Unit due to large figures, this means that there is a wide range of purchasing behaviour for a wide range of prices for the products.

-> Insight: Quantity * Unit Price = Turnover. Thus, the turnover source is diversified and not concentrated on any products.

2. Quantity & Customer's ID:

```
Variance: Quantity      4.755939e+04
CustomerID    2.936426e+06
dtype: float64
```

-> There is a diversity of quantity sold and ID is distributed quite broadly indicating a large customer base, further analysis is needed.

### 2.4.2. Correlation between variables:

Using correlation helps us identify feature significance (namely, whether or not a pair has a strong positive or negative relationship):

```
Correlation between Quantity & Turnover 0.8866810911791445
Correlation between Quantity & Date of Invoice 0.0006502493075075393
Correlation between Price per Unit & Turnover -0.1620285932476276
```

Quantity and Turnover: positive (the higher the quantity, the higher the turnover or the lower the quantity, the lower the turnover) -> **inventory caution**.
Quantity & Date of Invoice: no correlation -> insignificant.
Price per Unit & Turnover: negative (the higher the price, the lower the turnover or the lower the price, the higher the turnover) -> **price sensitivity**.

To recap the learning problem: it's the unsupervised learning problem, along with multivariate analysis because the dataset examination is mostly based on the relationships and patterns of more than one attribute.

# 3. Algorithm selection:

There are other clustering techniques such as hierarchical clustering, density-based spatial clustering, yet current research with the same scope has found out that K-means clustering is one of the best clustering methods (Turkmen, 2022). Therefore, K-means clustering will be used to **empirically test**. The other algorithm such as the hierarchical, density-based cluster will be theoretically used for comparison due to the same category as clustering method within unsupervised learning approach.

## 3.1. Suitability:

Two major criteria are taken into account:

3.1.1. Explainability: K-means clustering assists in classification simplification. For instance, the algorithm will divide a customer into a group, namely, a cluster according to similarities in purchasing behaviour. Specifically, it can be based on turnover (the price per unit and the quantity of items), and frequency (customer ID and invoice number).

3.1.2. Dimensionality reduction: The dataset needs this approach because the more attributes are included to analyse and visualise, the harder it is to extract patterns or relationships among those attributes. Manual attribute selection is adopted based on domain knowledge and correlation of variables. However, it's important for future work with Principal Component Analysis (PCA) technique to methodically extract pivotal attributes.

## 3.2. Cost:

It can be represented by how effective clustering divides customers into groups (clusters). This can be evaluated through the combination of two ways: one is through error evaluation metrics; another one is through visualisation of clusters which shows whether or not each cluster is tightly packed. If the data point stays close to each other, it will indicate common characteristics.

## 3.3. Algorithm selection:

- K-means clustering: It's a non-parametric method of grouping objects with similarity together based on their proximity.
- Hierarchical clustering: This is similar to K-means clustering in terms of its type and purpose, however, it groups objects based on hierarchy structure.

- Density-based spatial clustering (DBSC): This is also analogous to K-means clustering regarding its type and purpose, yet it groups objects based on the density of data points rather than a particular number of clusters.

## 3.4. Error valuation metric for K-means clustering:

The clusters will be assessed without the ground truth because we don't have predefined labels or categories of the customer's data.
Hence, the following metrics are presented:

Silhouette Coefficient: it indicates how well each data point fits within its assigned cluster, and usually ranges from -1 to 1 with 1 as the impeccable.

Davies-Bouldin index: it measures distance between clusters based on their centroids.

Calinski-Harabasz Index: it measures the quality of clusters based on its dispersion.

For this research, the goal is to ensure that each data point is assigned properly to clusters clarifying to which category customers will belong, hence, the Silhouette Coefficient is used for assessment. However, since Silhouette is merely defined as a numeric value, this report will further use visual inspection which means the utilisation of scatter plots because it helps us to see directly how well the clusters are represented.

## 3.5. Non-selection algorithms:

### 3.5.1. Naive Bayes

Naive Bayes's limitation is that attributes are independent of each other. In the case of the online retail dataset, each attribute in fact has strong interaction with each other (eg: Quantity, Price per Unit, Turnover). Therefore, this algorithm may not precisely capture the relationship among these attributes.

### 3.5.2. Decision Trees

Decision Trees can be interpretable, however, due to the fact that the online retail dataset consists of 541909 observations, hence, this algorithm can become overly complex, in other words, it's prone to overfitting. For instance, it might end up creating numerous branches and conditions based on the large number of observations (samples), making it difficult to follow.

# 4. Analysis & Discussion:

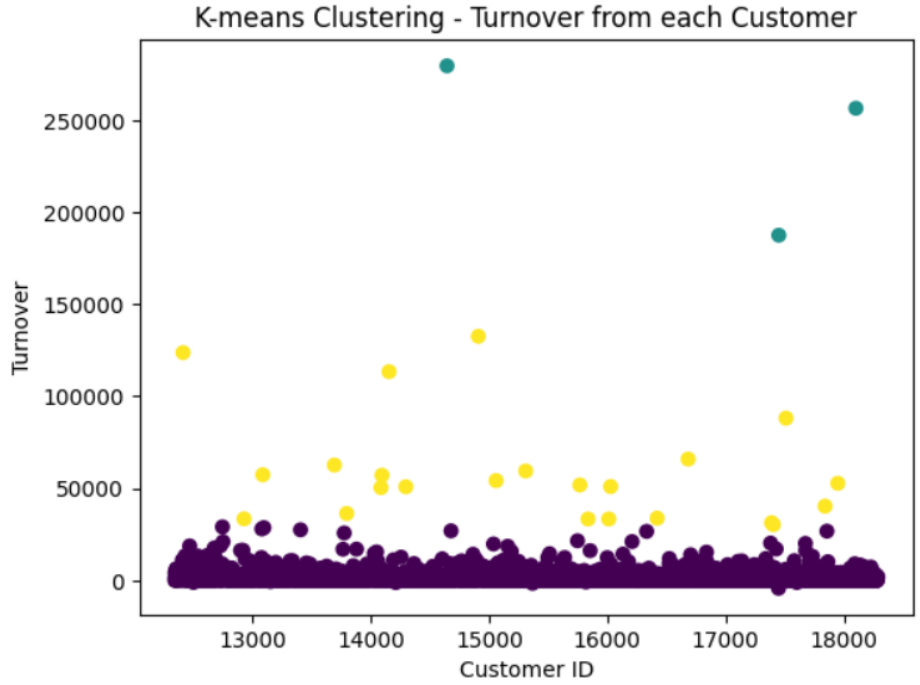
## 4.1. Pipeline Summary:

# 4.2. Implementation, Evaluation and Insight:

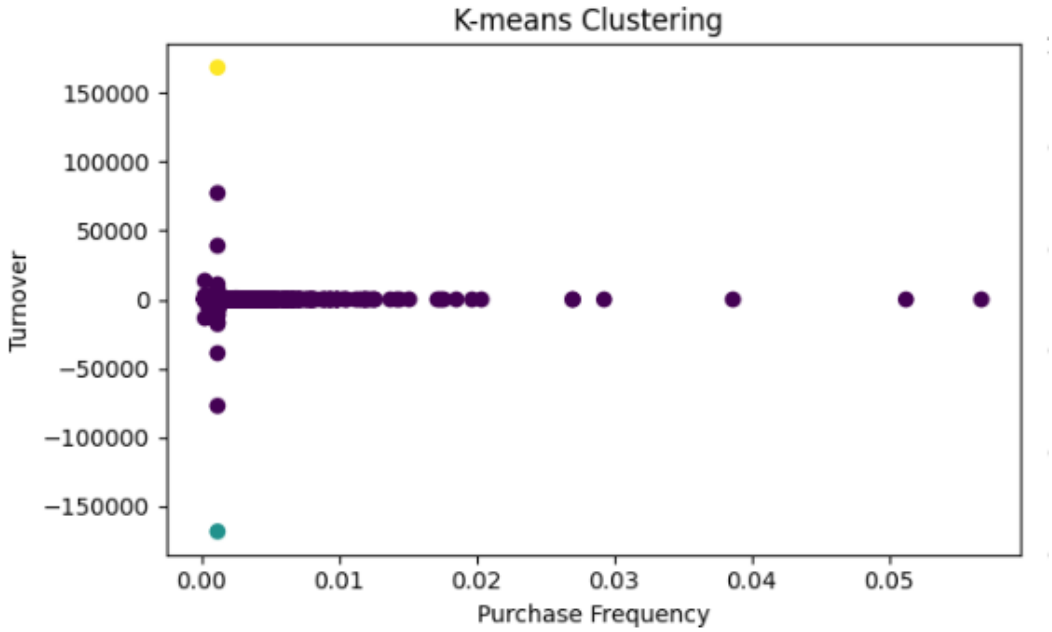
## 4.2.1. Implementation:

- K -means Clustering results:

Categorising customers based on turnover (Turnover, Customer ID):




Categorising customers based on Turnover and Purchase Frequency:




- Hierarchical clustering: Turkmen (2022) indicated that although this clustering method didn't require us to identify the optimal number of clusters, the experimental result showed that it failed from the business perspective due to a lack of variation to segment. Also, it is sensitive to outliers and that means it can be easily distorted by noisy data, for instance, some customers spend more money to buy stuff than the others from the online retail dataset. As a result, this metric may not fully capture the pattern effectively.
- DBCS: the method was suspected due to the large amount of noise in the clusters despite its advantage in dealing with an immense database (ibid). Besides, varying density can be a challenge as this metric assumes clusters having similar densities, however, customer behaviour can be varied. For example, some customers make frequent purchases while the others make irregular purchases. Consequently, it can be hard to gain insight from the clustering result.

### 4.2.2. Evaluation:

Statistical Significance of the model: Silhouette Coefficient (the higher the better, an impeccable score of 1):

1) Cluster based on Turnover & CustomerID:

```
Silhouette Score for Turnover with 3 clusters: 0.999827020922339
```

This means that K-means clustering nearly perfectly categorised the customers into three distinct groups based on Turnover & Customer ID criteria.

Visual Inspection: This plot showed clear results of 3 groups suggesting that more customers are spending from a low to moderate level of money, however, the purple group has some noise near 0 value, and the blue group is quite sparse.

2) Cluster based on Turnover & Purchase Frequency:

```
Silhouette Score for Frequency & Turnover with 3 clusters: 0.9998270209216649
```

This suggests that K-means clustering nearly perfectly categorised the customers into three distinct groups based on Turnover & Frequency.

Visual Inspection: This plot did not clearly show the 3 clusters; however, it represented a moderately explicit result of customers who make more frequent purchases but spend less money for the retail. It's noteworthy that there is an outlier in the plot (the blue dot) due to the negative value of turnover with nearly zero frequency which doesn't make much sense. Hence, it indicates anomalies of using K-means clustering.

### 4.2.3. Insights:

Other critical factors: inventory caution, price sensitivity, UK country.

| Categories | Turnover | Purchase Frequency | Customers | Recommendation |
|---|---|---|---|---|
| 1 | **Low** | **High** | Spend less money with high frequency | Stock availability should be highly prioritised, especially with the affordable and during the peak season (November).<br><br>Economies of Scale should be applied with promotion and discount to enhance customer loyalty |
| 2 | **Medium** | **Medium** | Spend moderately with moderate frequency | Offer a diverse range of products based on their preferences and needs in addition to bundle deals to encourage spending per transaction. |
| 3 | **High** | **Low** | Spend a lot of money and less frequency | Offer limited edition and exclusive products (from medium to high-end) based on their preferences and needs in addition to VIP membership package to enhance their experience. |

# 5. Conclusion:

This research shows an application of Machine Learning (unsupervised learning method) towards a real-word online retail dataset, supporting business decisions to enhance performance through tailoring appropriate strategies for different customer segmentation. K-means clustering can be considered as one of the most potential algorithms to segment customers into distinct groups despite some minor limitations such as containing minor abnormalities. Results show that the most potential group of customers is the one spending less money with high frequency. The retail business should also pay attention to their stock availability, price change and local preferences to improve and sustain customers' experience.

## 5.1. Limitations:

This report shows implementation of K-means clustering and justification of other non-selection algorithms theoretically and not experimentally, nevertheless. Moreover, this report doesn't show comprehensively the PCA technique, the distribution of other critical factors such as recency and the optimal number of clusters (k) due to a constraint of coding knowledge. The K-mean results contained some abnormalities despite a good indication of Silhouette score.

## 5.2. Further work:

future work may include further exploration of data in terms of demographics, product categories and preferences to expand the scope of customer segmentation in addition to implementing other critical factors such as recency and the optimal number of clusters (k) to explore further analysis.

More experimental assessment metrics should be included to provide a comprehensive evaluation of K-means clustering's effectiveness due to the experimental results that K-means contained some outliers. For example, of further assessment: testing the stability and consistency of K-means clustering through iterations, or through WCSS (within-cluster sum of squares) measuring how far each data point is from the centre of its own cluster.

# 6. References

Ahani, A., Nilashi, M., Ibrahim, O., Sanzogni, L., & Weaven, S. (2019). Market segmentation and travel choice prediction in Spa hotels through TripAdvisor's online reviews. *International Journal of Hospitality Management*, *80*, pp. 52-77.

Alkhayrat, M., Aljnidi, M., & Aljoumaa, K. (2020). A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA. *Journal of Big Data*, *7*, pp. 1-23.

Das, S., & Nayak, J. (2022). Customer segmentation via data mining techniques: state-of-the-art review. *Computational Intelligence in Data Mining: Proceedings of ICCIDM 2021*, pp. 489-507.

Gankidi, N., Gundu, S., viqar Ahmed, M., Tanzeela, T., Prasad, C. R., & Yalabaka, S. (2022, June). Customer Segmentation Using Machine Learning. In *2022 2nd International Conference on Intelligent Technologies (CONIT)*, pp. 1-5.

Kumar, A. (2023). Customer Segmentation of Shopping Mall Users Using K-Means Clustering. In *Advancing SMEs Toward E-Commerce Policies for Sustainability* (pp. 248-270). IGI Global.

Monil, P., Darshan, P., Jecky, R., Vimarsh, C., & Bhatt, B. R. (2020). Customer Segmentation Using Machine Learning. *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, *8*(6), pp. 2104-2108.

Ozan, Ş. (2018, September). A case study on customer segmentation by using machine learning methods. In *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)* (pp. 1-6). IEEE.

Turkmen, B. (2022). Customer Segmentation with Machine Learning for Online Retail Industry. *The European Journal of Social & Behavioural Sciences*.

Yadegaridehkordi, E., Nilashi, M., Nasir, M. H. N. B. M., Momtazi, S., Samad, S., Supriyanto, E., & Ghabban, F. (2021). Customers segmentation in eco-friendly hotels using multi-criteria and machine learning techniques. *Technology in Society*, *65*, 101528.