Association Rules and Unsupervised Clustering-based on Goods Analysis in Online Retail Selling

Jirayu Petchhan*
Department of Electrical Engineering
National Taiwan University of Science and Technology
Taipei, Taiwan
D10907801@mail.ntust.edu.tw

蔡承哲(Cheng-Zhe Cai)*
Department of Industrial Management
National Taiwan University of Science and Technology
Taipei, Taiwan
M11001105@mail.ntust.edu.tw

林依儒 (Yi-Ju Lin)*
Department of Finance
National Taiwan University of Science and Technology
Taipei, Taiwan
M10918004@mail.ntust.edu.tw

***Abstract*— As the application of big data becomes more and more widespread. Every day, we witness new applications of big data, allowing people to derive actual value from it. The analysis of big data has an impact on many businesses and individuals, but how can big data help people tap into it? Thus, we have selected our data cleaning & preprocessing as RFM segmentation which stands for recency, frequency and monetary of user's purchasing by our evaluating method there are three kinds for testing i.e., data analysis, association rules and finding the relationship between RFM segmentation via K-Means clustering method to explore and develop the business strategy for the retailer. Our results have been displayed that several relationships and make a call easier and more convenient to customer purchasing and/or brick and mortar's owner to adapt to procurement and distribution including planning for future trades.**

Keywords— Data Science, Association Rules, K means clustering, Basket analysis

I. Introduction

Big Data refers to the observation, tracking, recording and collection of various behaviors of target groups through a large number of illogical or structured sources, and then analyzing the data to meet the applications. In terms of business applications, Big Data can be applied in a wide range of ways, such as finding out the most time-consuming labor items and the most streamlined way to handle goods within a company, finding out the information of customer groups who buy products externally, analyzing through Big Data that the buyers of product A will also be interested in product B. The following are two examples of actual applications of Big Data in enterprises.

Facebook: An invisible algorithm

The social platform is not new to everyone, where you can see the posts of friends you haven't seen for a long time, the current status of your family, etc. However, the phrase "the most expensive for free" can be fully applied to Facebook. Whenever users use Facebook, they agree that they can track and collect the websites you have visited, the products you have purchased, and even the types of friends you have made on your phone or computer. The more frequently they are presented, the more they can arouse the interest of users, who can then buy the products and achieve advertising benefits.

In addition, Facebook will also analyze which people are highly associated with users and recommend them to become friends according to the process of confirming users as friends by recommending similar posts and increasing interaction with friends, which can make users more dependent on Facebook and make advertising on the platform more valuable for companies, thus forming a positive cycle.

Amazon: Items are shipped before orders are placed

Amazon is one of the largest online retailers in the world, and it was ranked 44th on Fortune magazine's list of the world's 500 largest corporations in 2016. Key data plays a big role in how Amazon controls orders and grows despite the odds throughout the pandemic.

Consumers love to compare prices when shopping online, and Amazon uses big data to know the prices of the same products on different platforms, so it can create marketing strategies with the highest discounts or super-low prices, luring customers into Amazon as if they were walking into a supermarket and wanting to make crazy purchases, resulting in higher customer unit prices.

Amazon can examine the products that customers have purchased and browsed, as well as the products that customers who have purchased the same products have also seen by applying big data. Amazon has a "predictive shopping" patent that allows it to "predict the consumer's desire to buy" and deliver those products with a high probability of being purchased to the nearest courier warehouse before the consumer places an order, in addition to allowing consumers to see many of the products they may be interested in as soon as they enter the Amazon page.

As a result, when a customer places an order, the warehouse is able to pack the goods in a short amount of time, reducing the time it takes for the goods to be forwarded and shipped across different regions, and achieving high efficiency in delivering the purchased goods to the customer's doorstep immediately. This method not only cuts shipping time in half, but it also satisfies customers' desire for speedy delivery.

Amazon utilizes big data to implement a three-pronged strategy of low prices, predictive shopping sites, and early delivery of products that customers are likely to purchase, making Amazon even more desirable to customers.

This paper is structured as follows. Section II gives a brief overview of common data analysis methods used in basket analysis. Section III presents the approach for machine learning model description and training dataset. Section IV provides the experimentation methodology and results. Finally, in Section V, the conclusion of the article is summarized and concluded.

II. Literature review

The purpose of data analysis is to focus, extract, and distill information that is hidden in a mass of seemingly disorganized data in order to identify the intrinsic laws of the subject under study. In practical terms, data analysis assists people in making judgments so that they can take appropriate actions. Data analysis is the process by which an organization purposefully collects data, analyzes it, and turns it into information. This process is a supporting process for the quality management system. The data analysis process is used throughout the product life cycle, from market research to after-sales service and final disposal, to enhance effectiveness.

TABLE I. Universal strategy in data analysis

Type	Function	Algorithm	Application
Classification Prediction	Classification	Decision Tree, Neural Network Classification, Discriminant Analysis, Logistic Regression, Probability Regression	Risk Analysis, Customer Retention Analysis, Fraud Detection
	Predict	Linear Regression, Nonlinear Regression	Yield Analysis, Income Forecast, Credit Value Forecast, Customer Potential Value Forecast
Clustering	Cluster Analysis	K-Means, Neural Network, Clustering	Customer Segmentation
Association Rules	Correlation Analysis	Statistic, Set Theory	Cross-Selling, Bundling
	Sequence Association Analysis	Statistic, Set Theory	Cross-Selling
	Similar Time Series Anaylsis	Statistic, Set Theory	Product Lifecycle
Predict	Time series forecasting	Satistical Time Series Models, Neural Networks	Sales Forecast, Interest Rate Forecast, Loss Forecast

In general, when we conduct business data analysis, we always expect to find more business insights through the data we have obtained in the past. When a retailer makes a sale, it also leaves a record of the customer's purchases, and if the manufacturer has a membership system, the sales data may be linked to the consumers' personal information. In this section, we shall present a data analysis concept: "association rules." Through this technique, we can extract the important key information from the past sales data and use it as the basis for product recommendations, which helps us to make management decisions.

Basket Analysis: Association Rule

Regarding sales, it is not always possible to determine who the consumer of a product is in all situations, such as when shopping at a supermarket or convenience store. In this scenario, how should the business offer item to consumers or plan products based on the consumer's buying history? At this stage, we can analyze the consumer's purchase details through the "association rule" algorithm, which is often referred to as "shopping cart analysis." For instance, if we see that a consumer has purchased both product A and product B, we may utilize this framework to identify which product we should promote to the user. The most classic case is beer and diapers. By analyzing the data of the U.S. supermarket, the data analysts found that the correlation between these two products is very high, so they put the display position of the two products together, increasing the revenue of the store. Such an algorithm, also known as the Apriori algorithm, is shown in the following figure.

According to the findings, market basket analysis utilizing the K. Apriori algorithm for shops does increases total income. (Annie, L. C. M., & Kumar, A. D. ,2012). In the Apriori algorithm, when calculating the association of a product A with other products, there are three important statistics that we can use as the basis for judgment.

i. Support: the proportion of purchase details containing Product A to the total purchase details
ii. Confidence: the proportion of product B in the purchase details of product A
iii. Lift: The formula is Confidence (A → B) / Support(B), and this value indicates the correlation between Product A and Product B. If this value is close to 1, then the two are independent. If it is greater than one, then it is negative correlation, and vice versa, it is positive correlation, as shown below.

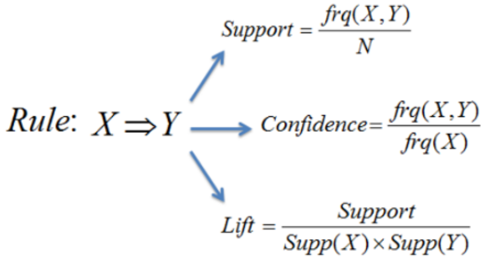

$Support = \frac{frq(X,Y)}{N}$
$Rule: X \Rightarrow Y$
$Confidence = \frac{frq(X,Y)}{frq(X)}$
$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$

Fig. 1. Formula in Associaiton Rule

Fig. 2. The exemplary imagination of shopping basket recommendation to increase sales by modeling evaluation.

Target Customer: K-means Clutering

K-means is an unsupervised learning algorithm that divides a group of data into k-clusters. The algorithm uses the distance between data as the basis for classifying clusters, and the similar data will form a cluster and find the centroids through weighted calculation or simple averaging. The distance from the centroid can be used to determine which subgroup the test data belongs to, or it can be further used for data compression to represent specific types of data to achieve issues such as noise reduction or gap filling. This is one of the partitional clustering methods, in which the error target value is repeatedly reduced until the target function value no longer changes or is lower, and the finale result of clustering is achieved. The process of K-means operation steps.

1. First set the number of groups to be divided into: K
2. Set K group centers randomly in the feature space.
3. Calculate the distance from each data point to each of the K group centers (basically, L2 distance is used, but it can be changed to something else.)
4. Assign the data point to the group center with the closest distance.
5. After all the data points are assigned, each group updates the group centers by averaging the data points just assigned to it.
6. Finally, repeat 3-5 until the group center has been converted (the group center does not change much after each update).

Overall, K means clustering, which takes a collection of frequent item sets from the data and then extracts the rules with the highest information content for distinct groups of customers by splitting the consumers into different clusters, in order to enhance total sales contribution levels. (Lim, T. ,2021)

III. Methodology

After we surveyed and reviewed the contribution of several existing work. We can see that the Unsupervised K-Means model is being implemented which is a beneficial network to adapt into any system with not having to rely on label data from dataset. And explicitly that we have been worked on Basket Analysis dataset to predict recommendation to create useful recommendations for work and investment in the retail business and distribution.

TABLE II. The exemplars of RFM and user ID after data-cleaning (Unit: recency - Days, frequency - Amount, Monetary - $)

CustomerID	Recency (Days)	Frequency (Amount)	Monetary (Price)
12413	66	3	758.10
12437	1	18	4951.41
12441	366	1	173.55
12488	9	3	1298.66
12489	335	1	334.93

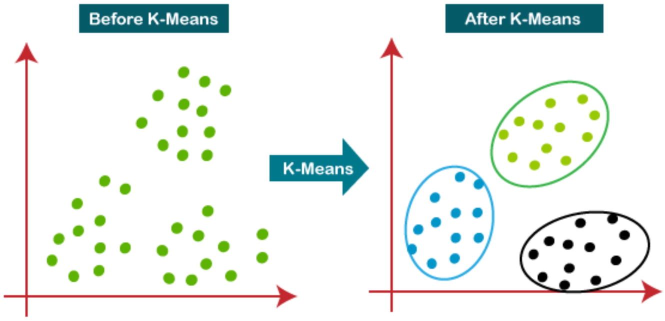


Fig. 3. The exemplary clustering processing from K-Means baseline model

A. Dataset

The dataset we implement to adapt to our contribution which is an "E-Commerce Data" (Carrie, 2017) with availability on Kaggle. The dataset had included 7 feature attributes i.e., Invoice Number, StockCode, Item Description, Quantity, InvoiceDate, CustomerID, and Country. Mainly, the dataset frequently included the object datatype which could make invulnerable for implementing processing. Hence, we need to filter the data as number type. This dataset is so beneficial to decide of several offering to customers and others who get the benefit in sales which is able to make a goods bundle in available inventory or warehouse from retail to increase the sales and motivate customers more easily.

Especially, we significantly rely on three factors essentially based on RFM Segmentation. RFM is a data modeling method used to analyze customer value. It stands for recency, frequency, and monetary, which are just three metrics that describe what your customers did. i.e., recency(R), frequency(F) and monetary(M). Recency of a customer is days since the last purchase. Frequency (F) of the bookings/turnover of a customer is several purchases, e.g., in 6 months. Monetary (M) is the total turnover of a customer: Sum of sales, e.g., in 6 months. Besides, for in-depth study, we strongly think on the internal relationship for calculating the probability between frequency, recency and momentary. By every RFM vulnerable is referred and related on each customerID information comparable to the person who pays the goods for him/herself.

B. Association Rules

Market Basket Analysis could explain a basic concept along with using in the Association Rule [9]: Support, Confidence, and Lift. this section of the article will be a sequel to one of the

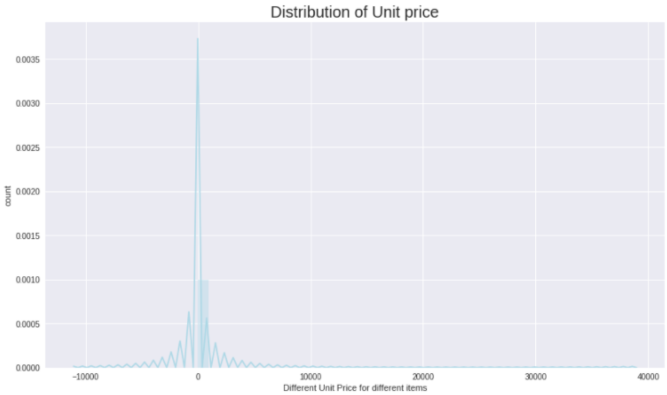


Fig. 4. The unit price distribution for different goods

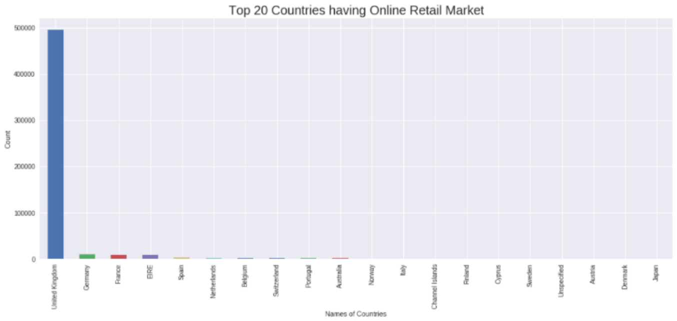


Fig. 5. Top 20 countries having online retail market

algorithms for creating relationship rules. by popular method It is a technique called Apriori Algorithm. the process of finding association rules (Association Rule Mining) can be divided into two sub-tasks: firstly, Frequent Itemset Generation is to find item sets that meet the minimum support criteria. These item sets are referred to as “Frequent item sets”. Secondly, Rule Generation is to find relationship rules with high Confidence or Lift, also known as “Strong Rules”. These rules are created from Frequent item sets obtained from the previous step.

In these two processes, Apriori Algorithm will play a part in helping to increase the efficiency and speed of relationship rule generation with large amounts of items and transactions. The focus of this article will be on implementing the Apriori Algorithm step by step for finding frequent item sets.

C. Unsupervised Learning: K-Means Clustering

We have proceeded the data to analyze on K-means [8] baseline model. Like many of the related jobs mentioned in the literature surveyed. K-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

The most common algorithm uses an iterative refinement technique. Due to its ubiquity, it is often called "the k-means algorithm". It is also referred to as Lloyd's algorithm, particularly in the computer science community. It is sometimes also referred to as "naïve k-means" because there exist much faster alternatives. So, we aim this model look able to access the preliminary data analysis to understand more So we have chosen this baseline model.

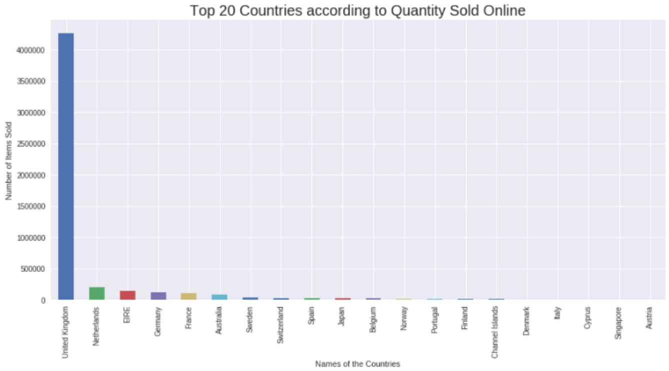


Fig. 6. Top 20 countries according to quantity sold online

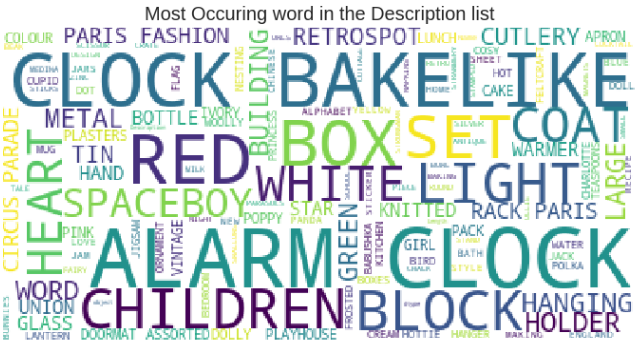


Fig. 7. Most Occuring word in the description list which was mostly mentioned.

IV. RESULT

We have evaluated within three sections i.e., data analysis for first part, association rules for second part and the last one is K-means clustering evaluating on preprocessed RFM segmentation. Our data evaluate on Google Colaboratory [7] which is available memory for free training.

A. Data Analysis

We have analyzed the data before data-cleaning. We could observe that several results shown that distribution of unit price normalized and shown at zero-point distribution in Figure 4. Next, checking the different values for country in the dataset. The evaluation result has been presented top 20 countries having online retail market at United Kingdom. Also, the same country i.e., United Kingdom having at the top countries in terms of quantities according to the countries in Figure 5 and 6 respectively. By our investigating, we observe that most of the goods that people are interested in are which displaying in large-scale size words in Figure 7.

When having analyzed in-depth of United Kingdom. We could see from the displayed information in December 2012 that got the peak demand of goods purchasing. This might depict those new year’s influences the spending and a lot of sales from many people in the country due to the festive season and long holidays shown in Figure 8. Besides, the analyzation of most popular stock codes that portray us that top 3 goods in stock plot its might be the best one for customer/client. Whether it's ordering, selling, stockpiling, claiming, or repairing, these 3 product codes are ideal for selling.

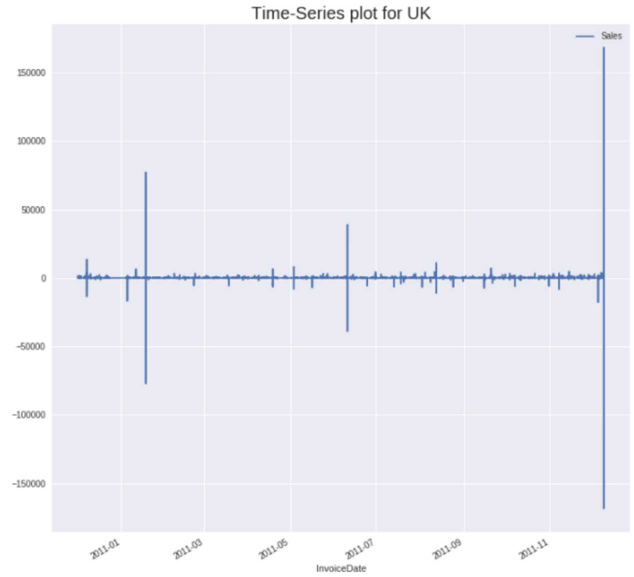
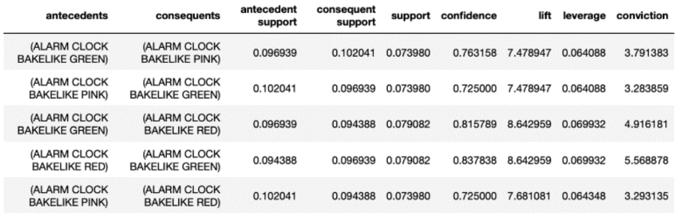


Fig. 8. Time-Series plot for United Kingdom

TABLE III. THE TABULATED EXEMPLAR OF APRIORI ALGORITHM PROCESSING SO AS TO THE ASSOCIATION RULE MINING.



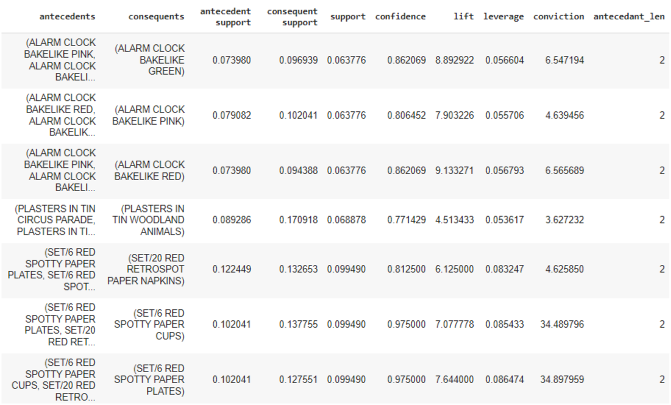
antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE PINK)	0.096939	0.102041	0.073980	0.763158	7.478947	0.064088	3.791383
(ALARM CLOCK BAKELIKE PINK)	(ALARM CLOCK BAKELIKE GREEN)	0.102041	0.096939	0.073980	0.725000	7.478947	0.064088	3.283859
(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE RED)	0.096939	0.094388	0.079082	0.815789	8.642959	0.069932	4.916181
(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE GREEN)	0.094388	0.096939	0.079082	0.837838	8.642959	0.069932	5.568878
(ALARM CLOCK BAKELIKE PINK)	(ALARM CLOCK BAKELIKE RED)	0.102041	0.094388	0.073980	0.725000	7.681081	0.064348	3.293135

B. Association Rules: AR Data Commentary

We tried to run our data processing on Apriori library framework and algorithm as association rules for calculating how to create a bundle of products of selling for retail that shows Applying the Apriori algorithm, the following table illustrates the support, confidence, and lift of each product portfolio. Finally, if clients purchase "Alarm Clock Bakelike Green," the suggested system may suggest that they also purchase "Alarm Clock Bakelike Pink" or "Alarm Clock Bakelike Red."

Investigate more extensively, assuming the buyers buy two goods, what the suggested systems will propose they buy. The table below illustrates two types of bundling portfolios: “PLASTERS IN TIN SPACEBOY, PLASTERS IN TIN CIRCUS PARADE, PLASTERS IN TIN WOODLAND ANIMALS” set or “/6 RED SPOTTY PAPER CUPS, SET/6 RED SPOTYY PAPER PLATES, SET/20 RED RETROSPOT PAPER NAPLINS” set. Finally, the shop manager might utilize these two portfolios to develop a campaign to attract people to purchase as displayed in table 4.

TABLE IV. THE TABULATED EXEMPLAR OF APRIORI ALGORITHM PROCESSING SUPPORTING DESCRIPTION IN MENTIONED PREVIOUS PARAGRAPH.



antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	antecedant_len
(ALARM CLOCK BAKELIKE PINK, ALARM CLOCK BAKELI...	(ALARM CLOCK BAKELIKE GREEN)	0.073980	0.096939	0.063776	0.862069	8.892922	0.056604	6.547194	2
(ALARM CLOCK BAKELIKE RED, ALARM CLOCK BAKELIK...	(ALARM CLOCK BAKELIKE PINK)	0.079082	0.102041	0.063776	0.806452	7.903226	0.055706	4.639456	2
(ALARM CLOCK BAKELIKE PINK, ALARM CLOCK BAKELI...	(ALARM CLOCK BAKELIKE RED)	0.073980	0.094388	0.063776	0.862069	9.133271	0.056793	6.565689	2
(PLASTERS IN TIN CIRCUS PARADE, PLASTERS IN TI...	(PLASTERS IN TIN WOODLAND ANIMALS)	0.089286	0.170918	0.068878	0.771429	4.513433	0.053617	3.627232	2
(SET/6 RED SPOTTY PAPER PLATES, SET/6 RED SPOT...	(SET/20 RED RETROSPOT PAPER NAPKINS)	0.122449	0.132653	0.099490	0.812500	6.125000	0.083247	4.625850	2
(SET/6 RED SPOTTY PAPER PLATES, SET/20 RED RET...	(SET/6 RED SPOTTY PAPER CUPS)	0.102041	0.137755	0.099490	0.975000	7.077778	0.085433	34.489796	2
(SET/6 RED SPOTTY PAPER CUPS, SET/20 RED RETRO...	(SET/6 RED SPOTTY PAPER PLATES)	0.102041	0.127551	0.099490	0.975000	7.644000	0.086474	34.897959	2

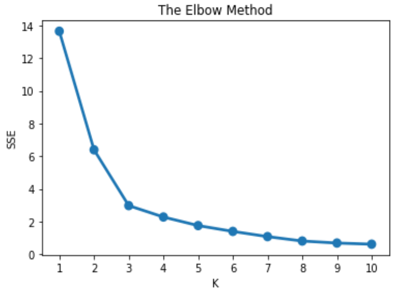


Fig. 9. Elbow method to evaluate which K-value better for applying. (K=3 is selected for us in testing).

According to the results a customer is 9 times more likely to buy pink alarm clock bakelike and green one than an average customer (lift) if he/she buys ALARM CLOCK BAKELIKE GREEN and ALARM CLOCK BAKELIKE PINK . This rule is "true" in 86% of the cases (confidence).This can be used as insight to recommend ALARM CLOCK BAKELIKE RED for those who bought the Green one and pink one.

According to the results a customer is 7 times more likely to buy a pink alarm clock bakelike than an average customer (lift) if he/she buys ALARM CLOCK BAKELIKE GREEN. This rule is "true" in 77% of the cases (confidence). This can be used as insight to recommend ALARM CLOCK BAKELIKE PINK for tthose who bought the Green one.

C. The relationship between RFM valuable via clustering and analysis

We tried our techniques through K-Means Clustering for RFM Segmentation. K-group processing from Elbow method i.e., K=3 shown dramatically reduce error from K at 1 and 2. Subsequently, up to K=4 slightly reduce error after then. So, we implement K=3 for out testing.

Besides, considering for each relationship, we have gained insights that higher expenditure related to frequency of user’s purchasing reduced when merchandise is more & more expensive (Shown in Figure 10). Besides, the recency of product has little effect on the purchase causing a lot of sales at low but frequent sales since the product newness may not yet be understood (Shown in Figure 11).

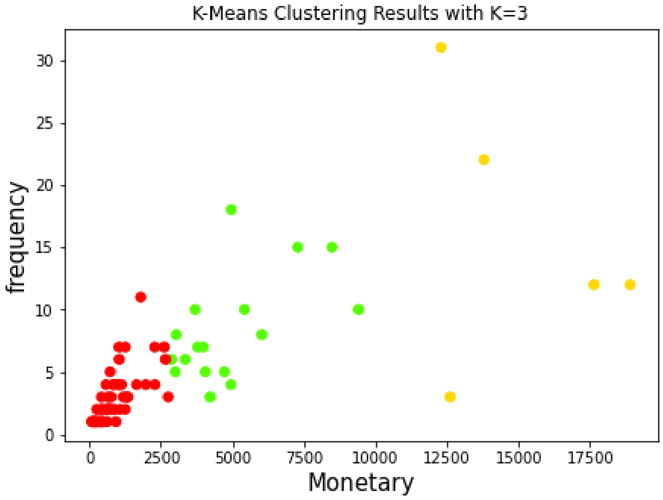


Fig. 10. The relationship between frequency (F) and monetary (M) (given the exemplar plot on France country) - best view in color.

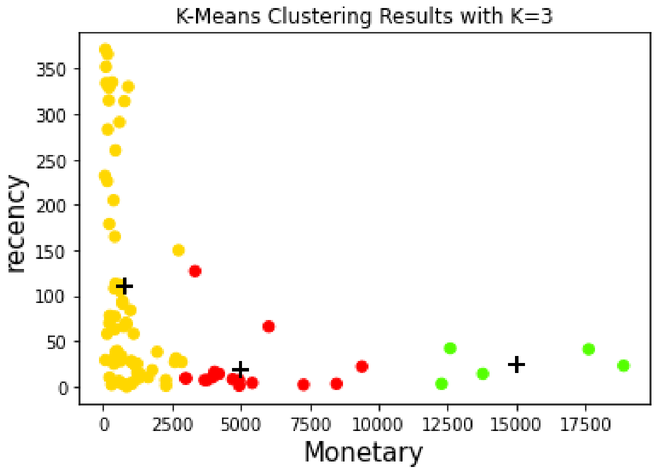
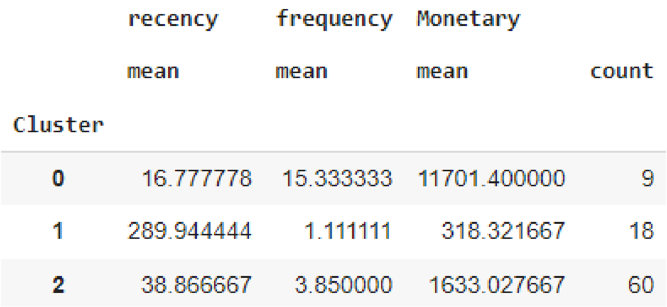


Fig. 11. The relationship between recency (R) and monetary (M) (given the exemplar plot on France country) - best view in color.

TABLE V. THE RFM ANALYSIS RESULTS



	recency	frequency	Monetary	
	mean	mean	mean	count
Cluster				
0	16.777778	15.333333	11701.400000	9
1	289.944444	1.111111	318.321667	18
2	38.866667	3.850000	1633.027667	60

In the part of RFM analysis, as can be seen from the table 5, Cluster 0 is the most loyal customer. Their visits and total spend are much higher than those of the other two cluster customers. Therefore, we can give some special offers to the customers of cluster 0 to attract them to spend.

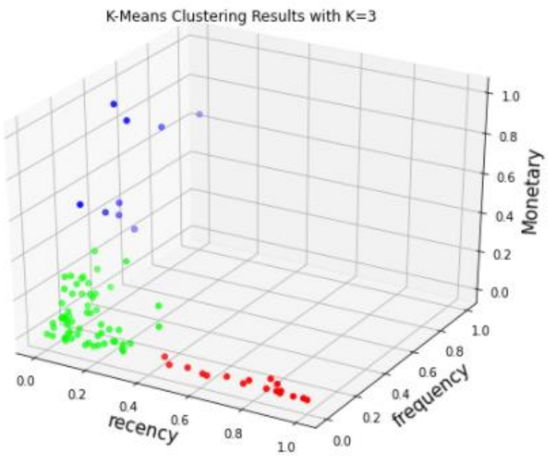


Fig. 12. The relationship between recency (R), frequency (F) and monetary (M) (given the exemplar plot on France country in three dimension) - best view in color.

V. CONCLUSION

Our menasurement instrumentality for making a bundle of goods through Apriori algorithm and finding the relationship between RFM segmentation through unsupervised K-Means clustering learning shows that it has been shown that processing contributes to decision-making and accelerates interaction between buyers and sellers in managing merchandise and making trades faster.

Likewise, raw data analyzation before data cleaning sharply help decision making This makes it more convenient to know how to penetrate the target for increasing sales or focusing on specific investments from owner brick and mortar more and more. In the one hand, with this kind of analysis from the field of mareting you can now determine which products are most often bought in combination with each other. In addition, with this knowledge it is possible to arrange the products efficiently in the store. In the best case, products that are often bought together are positioned in the opposite direction in the store so that customers are forced to walk past as many other products as possible. Furthermore, one can now consider targeted discount campaigns. If you discount a product that is often bought in combination with others, you increase the chance of buying these products in combination, whereby a small discount is granted on only one.

VI. CONTRIBUTION

Jirayu Petchhan, he has completed the part of data analysis from pre-processing data (i.e., raw online retail dataset available) including his experiments and prepared report from third to fourth session paper. 蔡承哲 (Cheng-Zhe Cai), he prepared data so as to post-processing data for RFM segmentation to fit in unsupervised clustering learning & Apriori algorithm for association rule for creating a bundle pack sales for basket online retail including conclude the finale part of paper. And 林依儒 (Yi Ju, Lin) she collected data and reviewed literature to write first and second part of paper which are particularly useful for writing on presentation slides. Besides, everyone cooperates to utilize an information from his/her parts where are achieved and composed the presentation slide together.

References

[1] MBA Library(n.d.). *Data Analysis.* https://wiki.mbalib.com/zh-tw/数据分析

[2] Huang. (n.d.). *Why use shopping cart analysis and collaborative filtering to look at business data*. Medium. https://medium.com/marketingdatascience/為什麼要用購物籃分析與協同過濾來看商務資料-c0b866785702

[3] Huang. (n.d.). *Machine Learning: clustering analysis- k-means-clustering*. Medium. https://chih-sheng-huang821.medium.com/機器學習-集群分析-k-means-clustering-e608a7fe1b43

[4] Lim, T. (2021, July). K-Means Clustering-Based Market Basket Analysis: UK Online E-Commerce Retailer. In *2021 International Conference on Information Technology (ICIT)* (pp. 126-131). IEEE.

[5] Annie, L. C. M., & Kumar, A. D. (2012). Market basket analysis for a supermarket based on frequent itemset mining. *International Journal of Computer Science Issues (IJCSI)*, *9*(5), 257.

[6] Carrie. “E-Commerce Data”. [Online]Available: https://www.kaggle.com/carrie1/ecommerce-data.

[7] Google. *Google Colaboratory*. [Online]. Available from: https://colab.research.google.com/.

[8] Scikit-Learn. K-Means clustering. [Online]Available: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

[9] Anisha Garg. Complete guide to Association Rules. [Online]Available: https://towardsdatascience.com/association-rules-2-aa9a77241654.