**ISSN** 1816-6075 (Print), 1818-0523 (Online)
**Journal of System and Management Sciences**
Vol. 12 (2022) No. 6, pp. 511-531
DOI:10.33168/JSMS.2022.0630

# Click Analysis: How E-ommerce Companies Benefit from Exploratory and Association Rule Mining

Lokman Hakim Bin Ahmad Sabri, Amy Hui Lan Lim[+] , Hui-Ngo Goh

Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100, Cyberjaya, Selangor, Malaysia

1171303755@student.mmu.edu.my, amy.lim@mmu.edu.my (corresponding author), hngoh@mmu.edu.my

**Abstract** Electronic commerce (henceforth referred to as e-commerce) has attracted many people to buy things online because of its convenience. With Covid-19 pandemic, the popularity of e-commerce increases as many people are working from home. Ability to understand customers' surfing and buying behavior on the e-commerce platform provides competitive advantage to e-commerce companies by being able to devise specific marketing plans to increase their market coverage and subsequently revenues from online sales of products. This paper discusses how the results derived from both, the exploratory data analysis (EDA) and association rule mining (ARM) can assist e-commerce companies to design specific marketing plans. The methodology consists of data understanding, data pre-processing, EDA, ARM, and analysis of results. A public dataset that is made available in the year 2020 consisting of clickstream data that are collected in 2018 from a popular fashion e-commerce website is used as a case study to prove the viability of the methodology in deriving results that can be used to design specific marketing plans. This study proves that it is possible to use clickstream data consisting of customers' surfing and buying behavior and apply the methodology to derive analysis and devise better marketing plans.

**Keywords:** e-commerce, exploratory data analysis, association rule mining

# 1. Introduction

E-commerce is defined by buying and selling of products on an online platform. The main goal of e-commerce companies is to reach as many customers as possible at the right time to increase the company's sales and profitability. Some of the activities that are performed on the platform include buying and selling goods over the Internet. Statistics have revealed that sales volume is predicted to be USD$ 6542 billion by the year 2023 (Coppola, 2021). Furthermore, the Covid-19 crisis has caused people in many countries to significantly limit physical interactions. This explains why during the Covid-19 crisis, the digital economy is booming. As people embraced social distance, they have turned to online shopping more than ever.

For e-commerce companies to reach out to many customers, marketing plans must be devised to popularize the e-commerce platforms. One way to devise the so-called appropriate and "right" plans is by understanding how the customers interact in e-commerce platforms. Their interactions in e-commerce platforms subsequently left footprints and their footprints are logged by the e-commerce servers as clickstream data.

The term clickstream refers to the interactions consisting of pages being surfed or clicked by a user in e-commerce platform. Clickstream data also shows the pages visited by the users, the length of time that they have stayed in the e-commerce platform, the order in which the pages are viewed, and the time stamps when they have visited and left the page. This makes clickstream data, a rich data on customers that can be mined to understand customers' buying behaviour. However, the e-commerce servers usually log the clickstream data in a raw format that cannot be directly understood or passed to any data mining algorithms for analysis. There is a need to pre-process the clickstream data so that it can be easily interpreted or passed to any data mining algorithms for analysis.

Exploratory data analysis (EDA) can be used to derive analysis from pre-processed data but depending entirely on the results from EDA to devise specific marketing plans is insufficient as EDA cannot derive complex patterns from clickstream data. The use of data mining algorithm such as association rules algorithms to derive more complex analysis from the underlying clickstream data is essential to compliment EDA in deriving associative patterns in the pre-processed data.

The insights from EDA and association rule mining (ARM), can be used by e-commerce companies to understand customers' surfing and buying habits, improve their e-commerce websites, and discover market trends. e-commerce companies can adapt their marketing plans to their customers' needs, develop new products that meet customers' needs, and provide superior customer service while at the same time increase conversion rate which means increase the probability of buying goods from e-commerce platforms.

This paper consists of following sections. Section 2 describes the literature review. Section 3 describes the methodology; Section 4 describes the results and analysis. Section 5 describes the conclusion. The list of references is available in the reference section.

## 2. Literature Review

The following describes the related work in predictive analytics and ARM. Many recent studies of clickstream analysis have focused on predicting customer behaviour by applying and comparing the performance of predictive machine learning algorithms.

The authors (Gumber et al., 2021) have proposed the XGBoost classifier on an online dataset from Kaggle to understand customers' surfing patterns. The result reveals an accuracy of 85.9% with a precision of 80.69% and a recall of 91.04%.

The authors (Naser et al., 2021) have proposed a framework for mining the weighted clickstream pattern named Weighted Pre-order Linked Web Access Pattern (WPLWAP) Tree with Maximum Possible Weighted Support (MaxPWS), a pruning technique for finding frequent patterns from weighted clickstream databases for the first time. They also compare the performance of WPLWAP to Weighted Sequential Pattern Mining (WSpan) and Weighted Frequent Sequential Pattern Mining (WFSPM). Experimental results show that WPLWAP is superior to WSpan and WFSPM in terms of required runtime and memory with larger datasets and lower thresholds.

The authors (Requena et al., 2020) have conducted two types of studies: classification of arbitrarily long click sequences and the early prediction of limited-length sequences by comparing two radically different strategies involving two alternative algorithmic approaches, using classification algorithms on a set of predefined hand-crafted features, and learning the features via deep learning (DL). The results show that they can produce fast and accurate classifications, highlighting that purchase prediction is reliable even for extremely short observation windows.

The authors (Vijayarani et al., 2022) have proposed and compared Scan-Reduced Indexing and Matrix algorithm to populate frequent itemsets from streaming data. The result shows that Scan-Reduced Indexing performs better in generating frequent itemsets.

The authors (Luna et al., 2008) have investigated the relation between support, confidence and lift for association rules that are generated from 30 datasets. This study can assist researchers in identifying the right measures to be used when interpreting the association rules.

The authors (Dogan et al., 2022) have implemented Fuzzy Association Rule Mining (FARM) on a sales dataset belonging to a B2C company where association

rules are derived to include sales amount as opposed to traditional ARM that only display occurrence as itemsets.

The authors (Orogun & Onyekwelu, 2019) have applied data pre-processing to select important attributes in the dataset before applying association rules algorithm. The dataset is sourced from UCI Machine Learning repository consisting of e-commerce transactions between 1/12/2010 and 9/12/2011.

In work by Sudirman et al. (2021), the authors have applied association rules to understand customers' shopping behaviour in a supermarket that have few branches in Bandung, Indonesia across three different time spans within a month.

Based on the description of the above literatures, most of the work focuses on predictive analytics (Gumber et al., (2021), Naser et al., (2021) and Requena et al., (2020)). The remaining work as described in Vijayarani et al. (2022), Luna et al. (2008), Dogan et al., (2022) and Orogun & Onyekwelu (2019) focus on ARM but the results from work by Vijayarani et al. (2022) and Luna et al. (2008) are not based on clickstream dataset that we intend to use in this paper. In fact, work by Vijayarani et al. (2022) is using Connect dataset from UCI Machine Learning repository while work by Luna et al. (2008) is using 30 datasets from Keel-dataset repository. The work by Dogan et al. (2022) and Sudirman et al. (2021) are not replicable due to the use of proprietary dataset. For work by Orogun & Onyekwelu (2019), the authors use e-commerce dataset between year 2010 to year 2011.

Overall, the dataset used in the work by Requena et al. (2020) is chosen to be used because it is a recent, real-life dataset that is publicly available in year 2020. Furthermore, the published work in clickstream analysis is either using older datasets, proprietary datasets, or synthetically generated datasets. The reason as to why older datasets are not used for clickstream analysis in this paper is because they do not reflect recent or near-recent customers' behaviours in e-commerce platforms. For work published using proprietary dataset, it is not possible to replicate its experiments due to non-disclosure of the content of the dataset.

# 3. Research Methodology

The project implementation is done by running a series of python codes in a Jupyter Notebook. This project follows the methodology as follows: Data Understanding, Data Pre-processing, Data Transformation, EDA, ARM and finally Analysis of Results. Figure 1 shows the graphical representation of the description as above.

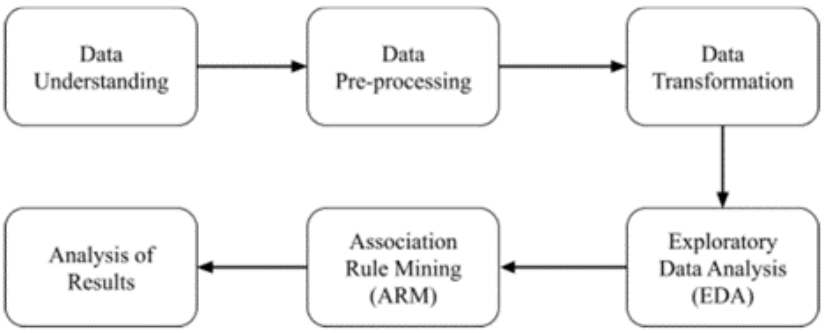


Fig. 1: Methodology

## 3.1. Data Understanding

The dataset that is used in this paper consists of clickstream data that are collected in 2018 from a popular fashion e-commerce website Requena et al. (2020). The dataset contains 5433611 individual events. The following describes the columns in the dataset.

*session_id_hash*: is an identifier of the shopping session that is hashed into a long string. The duration of a session is 30 minutes, and it can contain more than one event. If the same user returns to the same website after 31 minutes from the last interaction, a new session identifier is assigned. A repetition of a unique identifier may occur because each row may have only one *product_action.*

*event_type*: is an enumerated data consisting of pageview and event. This column describes the next column, *product_action.* For example, an "add" in *product_action* can happen on a page load (*event_type*: pageview) or a stand-alone event (*event_type*: event).

*product_action*: is an enumerated data consisting of 'detail', 'add', 'purchase', 'remove', and 'click'. This column shows the user's action in a session. If the field is null, the event is a simple page view without associated products. For example, the field is empty if the user is at homepage or at the FAQ page where there are no products involved.

*product_skus_hash*: is the hashed identifier of the products, given that the *product_action* is a product related event, such as detail, add, purchase, and remove. If *product_action* is null, then *product_skus_hash* will be null too.

*server_timestamp_epoch_ms*: is the epoch time, in milliseconds. The Unix epoch (or Unix time or POSIX time or Unix timestamp) is the number of seconds that have elapsed since January 1, 1970 (midnight UTC/GMT), not counting leap seconds. The epoch time for this data set has been shifted in time to further anonymize the data.

*hashed_url*: is the url of the current web page that is hashed into a long string to further anonymize the data.

## 3.2. Data Pre-Processing

After downloading the raw dataset, the dataset is read into the Jupyter Notebook. Then the column name *session_id_hash* is changed to *session_id*. This is because there is a need to refactor the *session_id* because the values are too long and hard to comprehend and visualize. The column is refactored to the respective unique integer starting from 0 up to 443,659 which means there are 443, 660 unique *session_id*.

After that the missing values in column *product_action* is replaced by the character 'V'. As mentioned in Section 3.1, if this field is empty, the event is a simple page view without associated products. For example, the field is empty if the user is at homepage or at the FAQ page where there are no products' pages being surfed. Thus, the character 'V' is chosen to generalize the missing values in *product_action*.

In the next step, the values in the *product_action* column are replaced with abbreviations to make them easier to handle. The values 'detail', 'click', 'add', 'purchase', and 'remove' are replaced with *D*, *C*, *A*, *P*, and *R* accordingly.

Then, the column name *product_skus_hash* is changed to *product_id* because of the same reason as the *session_id_hash*; the values are too long and not suitable for visualization. The values are refactored to the respective unique integer starting from -1 as a replacement for the missing values. This is because there is no product description page being viewed. Users at the e-commerce site are currently viewing pages that are not related to products in that session.

After that, the column *server_timestamp_epoch_ms* is converted from milliseconds (Unix) to datetime format and put into a new column timestamp. From this timestamp, the month, day, and hour are extracted and put into their respective columns which are *month_of_date*, *day_of_date*, and *hour_of_date*. The original *server_timestamp_epoch_ms* column is then dropped completely.

Finally, the columns *day_of_date* and *hour_of_date* are binned into a specific number of labels. The column *day_of_date* is binned into four weeks as labels (1, 2, 3 and 4) to indicate four weeks in a month. The column *hour_of_date* is binned into five sessions as labels (0, 1, 2, 3 and 4) that follow (early morning, morning, afternoon, evening, and night) as sessions in a day. The need to do binning is to create various time frames to explore the popular surfing time of the users. Figure 2 shows the snapshot of the pre-processed data extracted from the outcome from running the Python notebook.

| | session_id | event_type | product_action | product_id | hashed_url | timestamp | month_of_date | day_of_date | hour_of_date | session | week |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | pageview | V | -1 | da99729686aff70a02733b6cd69ee7df35622d9302347e... | 2018-12-10 19:26:36.111 | 12 | 10 | 19 | 3 | 2 |
| 1 | 1 | pageview | V | -1 | e2f7e0cee4272e804f0d323a3513dd01716a5a40ab9abf... | 2018-12-25 11:39:19.865 | 12 | 25 | 11 | 1 | 4 |
| 2 | 2 | pageview | V | -1 | ea7b2493be61ff454f8cce412f9dc281e605daec8c43b5... | 2018-12-15 21:20:35.402 | 12 | 15 | 21 | 4 | 2 |
| 3 | 2 | pageview | V | -1 | ea7b2493be61ff454f8cce412f9dc281e605daec8c43b5... | 2018-12-15 21:20:47.263 | 12 | 15 | 21 | 4 | 2 |
| 4 | 2 | pageview | D | 0 | 8fa1ecf31ececb27ebe9c529966f3d1f907542fe138d5d... | 2018-12-15 21:23:55.879 | 12 | 15 | 21 | 4 | 2 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 5433606 | 443656 | pageview | D | 1879 | a498ddf07e784d7d00df18a8f51a81e45aedf57ef8a45a... | 2018-12-23 13:20:37.554 | 12 | 23 | 13 | 2 | 3 |
| 5433607 | 443656 | pageview | V | -1 | aaf8c8581bfc7373f63dbdde27304c619bc7db1755ec50... | 2018-12-23 13:20:43.349 | 12 | 23 | 13 | 2 | 3 |
| 5433608 | 443657 | pageview | V | -1 | 54f2670e3703a7b85cf5015dc130bc6c1011d7f2fce07c... | 2018-12-25 14:28:24.469 | 12 | 25 | 14 | 2 | 4 |
| 5433609 | 443658 | pageview | V | -1 | f0ba8800a3e7fc0a3ea6904ad219ec44b964656817c52d... | 2018-12-11 13:48:33.737 | 12 | 11 | 13 | 2 | 2 |
| 5433610 | 443659 | pageview | V | -1 | 9ada061f4a8755d3409d8e45378a64108849e60ada89e7... | 2018-12-22 15:14:02.625 | 12 | 22 | 15 | 2 | 3 |

Fig. 2: Pre-processed dataset

## 3.3. Data Transformation

Before performing EDA, the pre-processed dataset must undergo several transformations. Examples include grouping the *product_id* by unique *session_id* to find the number of products' pages visited for each unique *session_id*. Before implementing ARM, the dataset is transformed into lists by grouping the *product_action* by unique *session_id*. Then it is transformed again into a dataframe by using an encoding method so that it fits the parameter of the ARM algorithm.

## 3.4. EDA

EDA refers to the process of conducting an initial observation of data to discover patterns, detect anomalies, test hypotheses, and use summary statistics and graphs to test assumptions. It is always a good idea to first understand the data and then try to get as much insight as possible from it. The results are presented and explained in Section 4.1.

## 3.5. ARM

ARM (Agrawal et al., 1993) is commonly used to understand customers' buying habits. By looking for correlations and associations between different items that customers place in their physical or virtual shopping carts, repeating patterns can be drawn. The purpose of applying this technique to the clickstream data is to observe the associations and correlations between the events that happen when the customers visit e-commerce platforms.

An association rule is an expression of the form of a rule as in Equation (1) where:

$$A \rightarrow B \tag{1}$$

This means that whenever A appears, B also tends to appear. A and B are itemsets. An itemset is a collection of database items. Some common metrics have

been developed to assess the "interestingness" of association rules such as support, confidence, lift and conviction.

The support metric is defined for itemsets. It refers to the frequency of occurrence of the itemsets in the database. The support's value ranges between 0 and 1 (both inclusive). The Equation (2) shows the formula to calculate the support for association rule of $A \rightarrow B$.

$$\text{support}(A \rightarrow B) = \text{support}(A \cup B) \quad (2)$$

The confidence of a rule is the probability of seeing the consequent in a sequence given that it also contains the antecedent. The metric is not symmetric or directed. If the consequent and antecedent always occur together (every time), the confidence for the association rule is 1. Like support, its value ranges from 0 to 1 (inclusive). Equation (3) shows the formula to calculate the confidence of $A \rightarrow B$.

$$\text{conf}\,(A \rightarrow B) = \text{support}(A \rightarrow B) / \text{support}(A) \quad (3)$$

The lift metric is used to measure how much more often the antecedent and consequent of a rule occur together than expected if they are statistically independent. Referring to Equation (1), if A and B are independent, then the value for lift will be exactly 1. If the value is greater than 1 then there is some usefulness to the rule. The larger the value for lift, the greater the strength of the association. The value for lift is minimally 0 and there is no maximum threshold for its value. Equation (4) shows the formula to calculate the lift for association rule of $A \rightarrow B$.

$$\text{lift}\,(A \rightarrow B) = \text{conf}\,(A \rightarrow B) / \text{support}(A) \quad (4)$$

A high value for conviction means that the consequent is highly dependent on the antecedent. For instance, in the case of a perfect confidence value of 1, the denominator becomes 0 for which the value for conviction is infinity and is defined as 'inf'. Like lift, if items are independent, the value for conviction is 1. The nearer the value for conviction to 0, the consequent is considered more independent on the antecedent. Equation (5) shows the formula to calculate the conviction for association rule of .

$$\text{conv}\,(A \rightarrow B) = 1 - \text{support}(B) / 1 - \text{conf}\,(A \rightarrow B) \quad (5)$$

# 4. Results and Analysis

The findings from EDA and ARM will be reported and discussed in this section

## 4.1. Results from EDA

The following is the summary of analysis. There is a total of 387,492 unique sessions where the events in each of the sessions is not associated with any product. There is a total of 276,293 unique sessions where events in each of the sessions is associated with events related to *product_actions* and viewing product description page. There is a total of 220,125 unique sessions having events related to combination of both as above. There are a total of 38344 unique products in the dataset. From implementing data transformation as described in Section 3.3, Figure 3 reveals the sessions sorted by the highest number of products' pages that are visited. On average, a session visits 6-7 product pages. The sessions are logged in December 2018.

| | session_id | no_of_product |
|---|---|---|
| **625** | 1016 | 978 |
| **164220** | 263843 | 503 |
| **179801** | 289012 | 478 |
| **231903** | 372380 | 464 |
| **130810** | 210235 | 431 |
| ... | ... | ... |
| **125567** | 201782 | 1 |
| **85184** | 136735 | 1 |
| **44856** | 71815 | 1 |
| **196553** | 315757 | 1 |
| **209537** | 336634 | 1 |

Fig. 3: Number of products' pages visited per unique session.

For each uniquely assigned *session_id*, the number of *product_actions* per product varies from minimum of 1 to maximum of 50 as in Figure 4 with an average of 1.2 *product actions* per unique *session_id*.

The product with the *product_id* of 11 is the most popular product with the highest number of visits of 10,327 as in Figure 5. There are several products that have only 1 visit which is the lowest number of visits. The average number of visits per product is 49.2.

| | session_id | product_id | no_of_product_action |
|---|---|---|---|
| 903725 | 263843 | 2720 | 50 |
| 1459327 | 425421 | 9343 | 47 |
| 1355504 | 395296 | 13856 | 43 |
| 1144426 | 334297 | 2819 | 36 |
| 1428674 | 416441 | 20540 | 35 |
| ... | ... | ... | ... |
| 553414 | 161907 | 1192 | 1 |
| 553413 | 161907 | 875 | 1 |
| 553412 | 161907 | 13 | 1 |
| 553411 | 161905 | 9932 | 1 |
| 1520731 | 443656 | 17986 | 1 |

Fig. 4: Number of *product_actions* per unique session and per unique product

| | product_id | no_of_visit |
|---|---|---|
| 11 | 11 | 10327 |
| 191 | 191 | 8279 |
| 277 | 277 | 5382 |
| 77 | 77 | 4963 |
| 78 | 78 | 4814 |
| ... | ... | ... |
| 25734 | 25734 | 1 |
| 25735 | 25735 | 1 |
| 11324 | 11324 | 1 |
| 25739 | 25739 | 1 |
| 38343 | 38343 | 1 |

Fig. 5: Number of visits per unique product.

The product with product id 2280 is the most purchased product with a total of 14 times being purchased. Figure 6 shows the products that have at least been purchased once.

| | product_id | no_of_purchase |
|---|---|---|
| 29 | 2280 | 14 |
| 540 | 10522 | 12 |
| 83 | 4745 | 11 |
| 674 | 11326 | 10 |
| 48 | 3061 | 9 |
| ... | ... | ... |
| 3270 | 21605 | 1 |
| 3269 | 21602 | 1 |
| 3268 | 21600 | 1 |
| 3267 | 21597 | 1 |
| 9159 | 38342 | 1 |

Fig. 6: Number of purchases per unique product

The total number of URLs being surfed that are not associated with product-related events is 194,797. The number of non-product URLs being surfed per unique session ranges from 1 up to 7,483. The average number of non-product URLs visited per unique session is averagely 9.1.

| | session_id | url_count |
|---|---|---|
| 385626 | 441519 | 7483 |
| 274601 | 314386 | 3944 |
| 21952 | 25256 | 3593 |
| 262596 | 300590 | 3584 |
| 243173 | 278445 | 3571 |
| ... | ... | ... |
| 199864 | 228932 | 1 |
| 199868 | 228936 | 1 |
| 199873 | 228941 | 1 |
| 199881 | 228950 | 1 |
| 387491 | 443659 | 1 |

Fig. 7: Number of non-product URLs being surfed per unique session

The most visited non-product URL has a total of 688,361 visits. The second most visited non-product URL has 242,152 total visits and the third most visited non-product URL has a total of 160,282 visits. This can be seen in Figure 8. A logical explanation for the most number of visits is that the URL is most likely the homepage of the e-commerce website.

| | hashed_url | visit_count |
|---|---|---|
| 138695 | b6133a35ba149e49bcfa0f6481add6f2b88b947731ca02... | 688361 |
| 172822 | e2f7e0cee4272e804f0d323a3513dd01716a5a40ab9abf... | 242152 |
| 101382 | 855f65e9dccdfab68c12b29ea3ed95adaa70567644183c... | 160282 |
| 49955 | 41c5bc438612280333556a4f7ffdfa85bd1bd5a6e4b2c3... | 95485 |
| 192422 | fcf3dfd02260cd45d86842fd4f160a149249807eb31d95... | 47914 |
| ... | ... | ... |
| 103551 | 882e1a0fc1fcf32e748b6174831bbe53173a251e62bccc... | 1 |
| 103550 | 882dd6b93d0a9ab960226f5e9bfaf8b6bc45a78e9f57f0... | 1 |
| 103549 | 882c9c52b2b4819f7871de19d80c9a01f43a1581381006... | 1 |
| 36300 | 2fb68ee7a910bc6191ae1cefcdee55253fa0b851eb7c04... | 1 |
| 0 | 00002bf113e07e87752d36493d810a672ea4030bcea678... | 1 |

Fig. 8: Number of visits per non-product URLs

Based on Figure 9, the frequency of non-product and product-related sessions follows the pattern of the frequency of total session. This is because both types of sessions are the subsets of the total sessions. From the graph, the time at approximately 1200 and 1300 hours of the day is the most popular times for the users to surf the e-commerce platform.

From Figure 10, the x-axis is labelled as 0, 1, 2, 3 and 4 which represents early morning, morning, afternoon, evening, and night accordingly as sessions of a day. The most popular session of the day is the early morning, followed by the evening session. A reasonable explanation is that the duration of early morning session is 8 hours long, from 12 a.m. to 8 a.m. and the duration of evening session is 5 hours long, from 3 p.m. to 8 p.m. These two sessions have more hours per session hence having more active sessions in total.

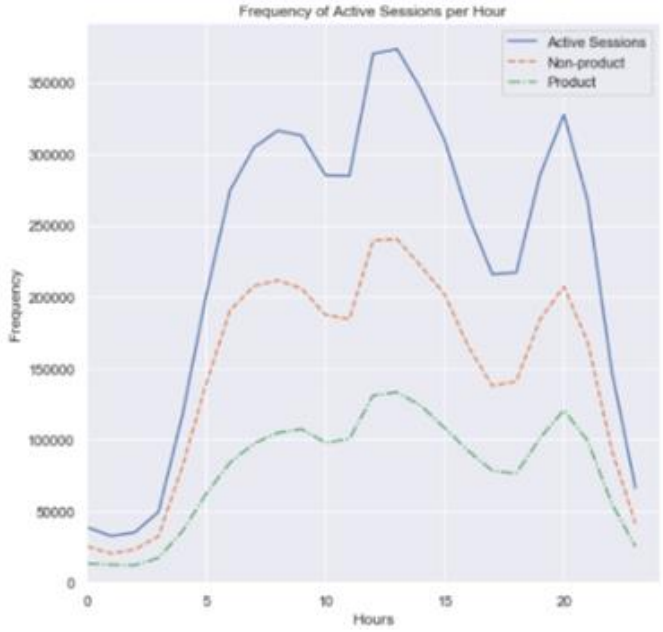


Fig. 9: Frequency of active sessions per hour

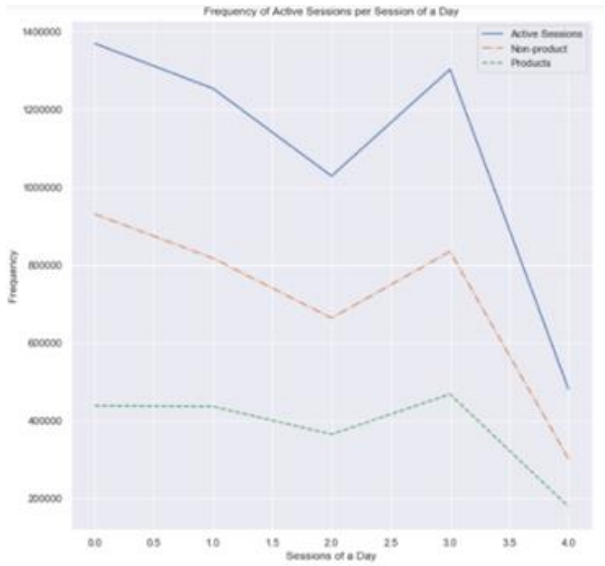


Fig. 10: Frequency of active sessions per session of a day

Figure 11 shows that the number of active sessions starts to accelerate from the 8th day of the month. The day with the highest frequency of active sessions is the 10th day.

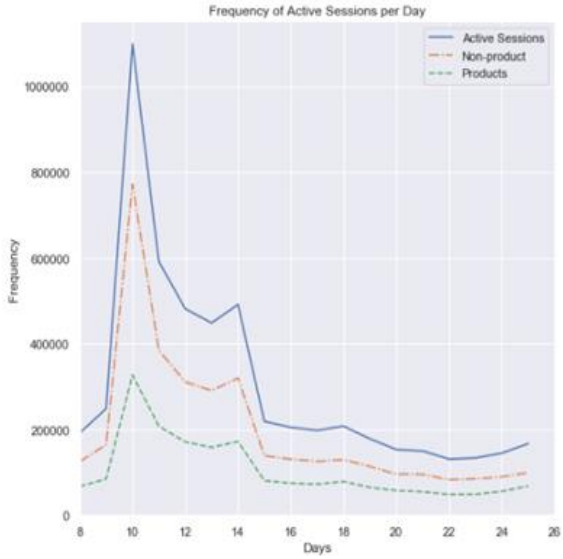


Fig. 11: Frequency of active sessions in a month

Figure 12 shows that the number of active sessions increases and peak at Week 2 of the month while Figure 13 shows the frequency of sessions with purchase peaks at 8 a.m. to 9 a.m. and 12 p.m. to 1 p.m. This indicates that users are more likely to purchase in these time spans compared to the other time of the day.

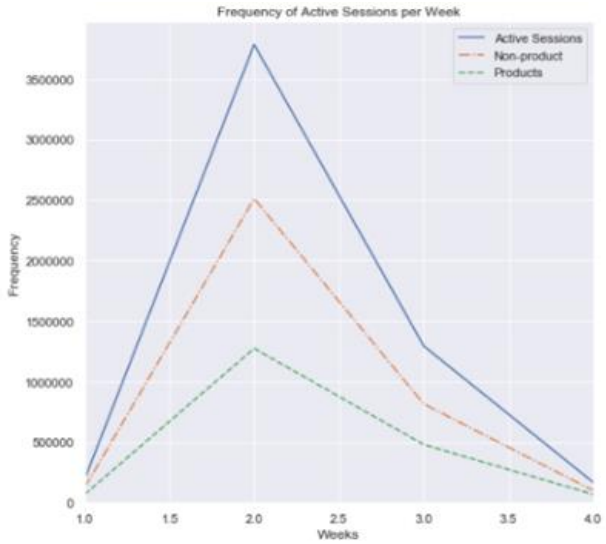


Fig. 12: Frequency of weekly active sessions.

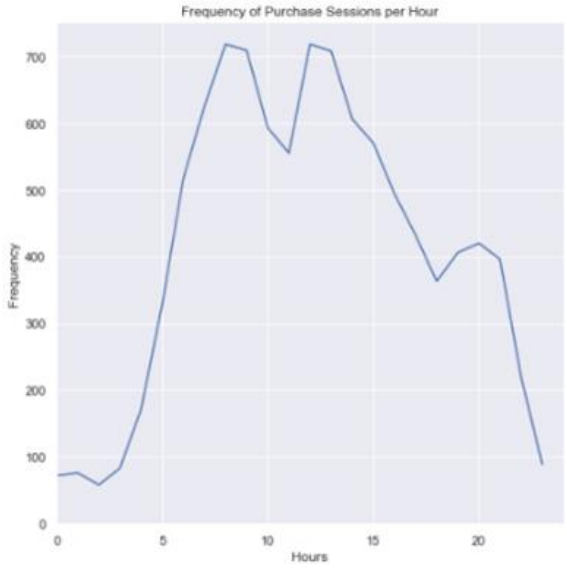


Fig. 13: Frequency of sessions with purchase by hour

Figure 10 shows the same downtrend pattern as Figure 14. The only difference is the frequency range is much lower indicating that only a small percentage of total sessions are performing purchase.

Figure 15 also has the same trend as Figure 11, where the highest frequency is on the 10th day, then the frequency starts plummeting as the days go by. This proves that most purchases are done on the day of the month that has the highest total active sessions.

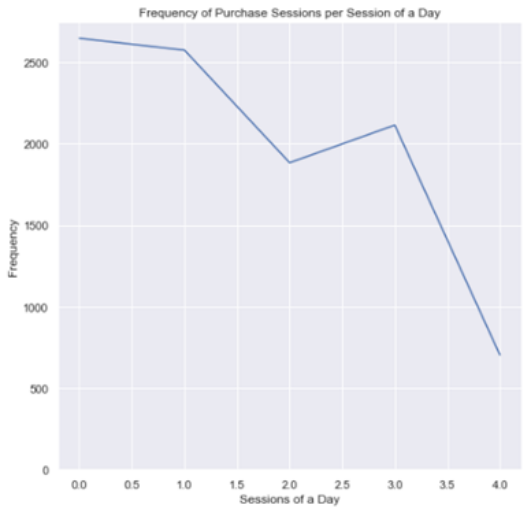


Fig. 14: Frequency of sessions with purchase per sessions of a day

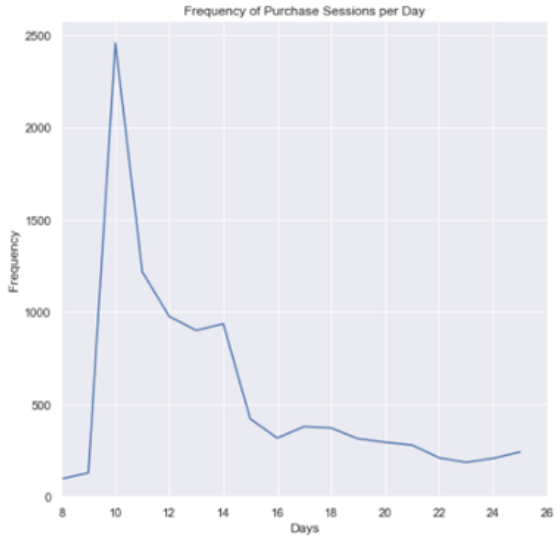


Fig. 15: Frequency of sessions with purchase in a day

Figure 16 follows the same trend as in Figure 12, where the 2nd week is the most popular in terms of active sessions frequency. There are over 7,000 purchases in the 2nd week.

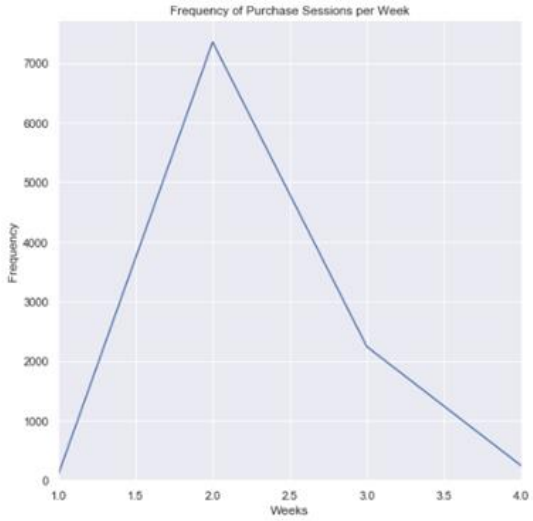


Fig. 16: Frequency of sessions with purchase per week

From the earlier EDA analysis, there are few marketing plans that can be devised by e-commerce companies to boost the likelihood that purchase occurs within the sessions as described below.

From our general observation on Figure 5 and Figure 6, a product that is popularly purchased is not the product whose page is highly visited. Similarly, the

vice versa. The e-commerce companies can boost the likelihood of the highly visited product being purchased by creating a combo package, where customers who bought the highly purchased product can buy the product that is highly visited at a certain offer price.

As an alternative, the highly visited products can be put on promotional offer without being tied to the highly purchased products. Example of promotional offers can be discounts varying according to the quantity of the products being purchased.

Based on Figure 4, there is a need to devise customer retention plan where the customers whose sessions have recorded a high number of product pages visits should be offered promotions through rebates, discount vouchers such as free shipping fees and etc. so that they will continue to return to the same e-commerce platform and increase their chance of purchasing products.

Based on Figure 8, the promotional banners or advertisements should be placed at the non-product URLs that are highly visited to attract potential customers to surf product pages and increase their chance of purchasing products.

Marketing plans can also be devised according to timestamps. Based on Figure 9 and Figure 13, we can see an almost similar trends where the active sessions that leads to purchase generally occur during non-working hours: between 8 am to 9 am, 12 pm to 1pm and 8 pm. Figure 10 and Figure 14, confirm that active sessions and sessions that ended up with purchase occur in early morning (before working hours) and evening (after working hours). A marketing plan can be devised on non-working hours by providing discounts or offers for purchases within those periods of times. There should be more exciting promotional offers such as greater discounts for purchase during 12pm – 1pm which is a lunch break.

Referring to Figure 12 and Figure 16, the active sessions and sessions that leads to purchase occur in week 2. By zooming further to Figure 11 and Figure 15, we can observe that the active sessions and sessions leading to purchase occur at Day 10 of the month. From our earlier analysis description of EDA, this dataset captures the customers' clickstream that occur in December. We guess that Week 2 of December has highest purchase as people are preparing to celebrate Christmas which falls on 25th December. For e-commerce companies, festive season is an opportunity to do great sales, promotions and offers, continuously and actively done on the e-commerce platforms during the first two weeks of December up to the days leading to Christmas.

## 4.2. Results from ARM

This section presents the results and analysis from ARM. Figure 17 shows the snapshot of association rules generated.

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift | leverage | conviction |
|---|---|---|---|---|---|---|---|---|---|
| 370 | (C, R) | (A, D, P) | 0.005114 | 0.016885 | 0.001871 | 0.365800 | 21.664770 | 0.001784 | 1.550166 |
| 359 | (A, D, P) | (C, R) | 0.016885 | 0.005114 | 0.001871 | 0.110800 | 21.664770 | 0.001784 | 1.118854 |
| 444 | (C, V, R) | (A, D, P) | 0.005114 | 0.016885 | 0.001871 | 0.365800 | 21.664770 | 0.001784 | 1.550166 |
| 424 | (A, V, D, P) | (C, R) | 0.016885 | 0.005114 | 0.001871 | 0.110800 | 21.664770 | 0.001784 | 1.118854 |
| 437 | (A, D, P) | (C, V, R) | 0.016885 | 0.005114 | 0.001871 | 0.110800 | 21.664770 | 0.001784 | 1.118854 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 48 | (V) | (A, D) | 0.873313 | 0.089228 | 0.087626 | 0.100337 | 1.124499 | 0.009701 | 1.012348 |
| 42 | (A) | (V) | 0.090330 | 0.873313 | 0.088525 | 0.980013 | 1.122179 | 0.009638 | 6.338467 |
| 43 | (V) | (A) | 0.873313 | 0.090330 | 0.088525 | 0.101367 | 1.122179 | 0.009638 | 1.012281 |
| 1 | (D) | (V) | 0.619386 | 0.873313 | 0.492916 | 0.795813 | 0.911257 | -0.048002 | 0.620446 |
| 0 | (V) | (D) | 0.873313 | 0.619386 | 0.492916 | 0.564421 | 0.911257 | -0.048002 | 0.873810 |

602 rows × 9 columns

Fig. 17: Association rules sorted by lift values

From Figure 17, the following association rules can be extracted:

$$(C, R) \rightarrow (A, D, P) \tag{6}$$

$$(A, D, P) \rightarrow (C, R) \tag{7}$$

The association rules as in Equation (6) and Equation (7) have the same values for support, lift and leverage, but they have different values for confidence and conviction. Based on the difference in the values for confidence, the association rule as in Equation (6) is three times more likely to be found in the dataset compared to association rule as in Equation (7).

Refer to the third paragraph of Section 3.2, 'add' is represented as *A*, 'detail' is represented as *D*, 'purchase' is represented as *P*, 'click' is represented as *C* and 'remove' is represented as *R*. It makes sense to have these five events highly dependent on each other. Based on Equation (6), the association rule can be interpreted as follows: IF customers 'click'(*C*) on a link (maybe a hyperlink to the shopping cart page), 'remove'(*R*) a product from shopping cart, THEN 'add'(*A*) another product, check its 'details'(*D*) before 'purchase'(*P*). A marketing plan can be devised to ensure that the customers do not leave the e-commerce platform after purchase by re-routing them to the page where the promotional banners and advertisements are located.

Based on Equation (7), the association rule can be interpreted as follows: IF customers 'add'(*A*) a product to the shopping cart, check its 'details'(*D*) and 'purchase'(*P*), THEN the customers 'click'(*C*) on hyperlink (maybe a hyperlink to the shopping cart page) and remove(*R*) the product from shopping cart. A marketing plan should be designed to "stop" the removal of other products in shopping cart by providing more offers so that customers continue with more purchases.

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift | leverage | conviction |
|---|---|---|---|---|---|---|---|---|---|
| 350 | (A, C, D, R) | (P) | 0.004722 | 0.020809 | 0.001871 | 0.396181 | 19.039193 | 0.001773 | 1.621665 |
| 306 | (A, C, R) | (P) | 0.004722 | 0.020809 | 0.001871 | 0.396181 | 19.039193 | 0.001773 | 1.621665 |
| 380 | (A, C, V, R) | (P) | 0.004722 | 0.020809 | 0.001871 | 0.396181 | 19.039193 | 0.001773 | 1.621665 |
| 410 | (D, R, A, C, V) | (P) | 0.004722 | 0.020809 | 0.001871 | 0.396181 | 19.039193 | 0.001773 | 1.621665 |
| 292 | (C, V, R) | (P) | 0.005114 | 0.020809 | 0.001927 | 0.376818 | 18.108651 | 0.001821 | 1.571277 |
| 200 | (C, R) | (P) | 0.005114 | 0.020809 | 0.001927 | 0.376818 | 18.108651 | 0.001821 | 1.571277 |
| 320 | (C, V, D, R) | (P) | 0.005114 | 0.020809 | 0.001927 | 0.376818 | 18.108651 | 0.001821 | 1.571277 |
| 278 | (C, D, R) | (P) | 0.005114 | 0.020809 | 0.001927 | 0.376818 | 18.108651 | 0.001821 | 1.571277 |
| 124 | (A, V, R) | (P) | 0.027474 | 0.020809 | 0.008464 | 0.308065 | 14.804589 | 0.007892 | 1.415148 |
| 104 | (A, R) | (P) | 0.027474 | 0.020809 | 0.008464 | 0.308065 | 14.804589 | 0.007892 | 1.415148 |

Fig. 18: Association rules with only 'purchase'(*P*) as consequent and sorted by values of lift

Based on Figure 18, we can cluster the association rules into three clusters based on the interests' values. The first cluster consists of association rules at indices 350, 306, 380 and 410. The second cluster consists of association rules at indices 292, 200, 320 and 278. The third cluster consists of association rules at indices 124 and 104. For each rule in each cluster, there is a high dependency among the events which is indicated by high values of lift. The values for conviction are positive indicating that the 'purchase'(P) is highly dependent on the series of other values in *product_action* as shown in the antecedent part of the association rule.

Within each cluster, we can sort the association rules according to its length and we can observe that the events in antecedent of the shortest length of association rule is present in the antecedents of the remaining association rules. There are two points to note:

- Firstly, this is consistent with the downward closure principle, which states that all subsets of a frequent itemset must also be frequent (LZP, 2019).
- Secondly, for the first cluster, the common subset is (A, C, R), while the second cluster is (C, R) and for the third cluster is (A, R).

With the above observations, whenever these common subsets are present within the sessions, they will eventually lead to a 'purchase'(*P*). We advise the e-commerce companies to do targeted marketing and customer retention by identifying patterns within the sessions that matches any of the subsets and do promotion by providing offers such as vouchers to increase customers' likelihood of 'purchase'(P).

# 5. Conclusion

In conclusion, it is possible to understand customers' shopping patterns and behaviour in e-commerce websites using the clickstream data and design specific marketing plans to boost the chance of products being purchased in e-commerce platforms. The clickstream data can be mined by applying the methodology consisting of Data Understanding, Data Pre-processing, Data Transformation, EDA, ARM, and Analysis of Results. This paper discusses how the results derived from

both, the exploratory data analysis (EDA) and association rule mining (ARM) can assist e-commerce companies to design specific marketing plans. By using the anonymized public dataset, we have proven the viability of applying the above-mentioned methodology to extract results from EDA and ARM for analysis. We have also proposed possible marketing plans that can be devised with results from EDA and ARM which can capture more customers, increase engagement and rate of purchase.

## Authors' Contributions

All authors contribute to the preparation until the completion of the manuscript.

## Acknowledgements

We would like to thank the staff of Faculty of Computing and Informatics, our family members and friends for their support and encouragement towards the completion of the manuscript.

## References

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, 207-216. DOI:10.1145/170036.170072

Coppola, D. (2021). eMarketer, number of digital buyers worldwide from 2014 to 2021 (in billions). https://www.statista.com/statistics/251666/number-of-digital-buyers-worldwide/

Dogan, O., Kem, F. C. & Oztaysi, B. (2022). Fuzzy association rule mining approach to identify e-commerce product association considering sales amount. *Complex & Intelligent Systems*, 1-10. DOI: DOI:10.1007/s40747-021-00607-3

Gumber, M., Jain, A., & Amutha, A. L. (2021). Predicting customer behavior by analyzing clickstream data. 5th International Conference on Computer. *Communication and Signal Processing, (ICCCSP)*, 1-6. DOI:DOI:10.1109/ICCCSP52374.2021.9465526

Luna, J. M., Ondra, M., Fardoun, H. M. & Ventura, S. (2008). Optimization of quality measures in association rule mining: An empirical study. *International Journal of Computational Intelligence Systems*, 12(1), 59 - 78. DOI:10.2991/ijcis.2018.25905182

LZP, J. (2019). What is the downward closure principle? What is the downward closure principle? | by Jason LZP | https://lzpdatascience.medium.com/what-is-the-downward-closure-principle-42e2f084f13e," Medium

Naser, A., Muhammad, Sultana, N., Islam, M. A., & Ovi, J. A. (2021). Weighted clickstream mining using pre-order linked web-access pattern tree. *2nd International Conference for Emerging Technology (INCET)*, 1-11

Orogun, A., & Onyekwelu, B. (2019). Predicting consumer behaviour in digital market: A machine learning approach. *International Journal of Innovative Research in Science, Engineering and Technology*, 8(8), 8391-8402. DOI:10.15680/IJIRSET.2019.0808006

Requena, B., Cassani, G., Tagliabue, J., Greco, C., & Lacasa, L. (2020). Shopper intent prediction from clickstream e-commerce data with minimal browsing information. *Scientific reports*, 10(1), 1-23

Sudirman, I. D., Bahri, R. S., Utama, I. D., & Ratnapuri, C. I. Using association rule to analyze hypermarket customer purchase patterns. *In Proceedings of the Second Asia Pacific International Conference on Industrial Engineering and Operations Management*, 12-23

Vijayarani, S., Sivamathi, C. & Prassanalakshmi, R. (2022). Frequent items mining on data streams using matrix and scan reduced indexing algorithms. *ASEAN Journal of Science and Engineering*, 3(2), 123-138. DOI: 10.17509/ajse.v3i2.45345