THE REPUBLIC OF TURKEY

BAHCESEHIR UNIVERSITY

DISCOVERING USERS' USAGE PATTERNS OF WEB LOG THROUGH ASSOCIATION RULES MINING METHODOLOGY

Master's Thesis

AHMAD HISHAM ARNAOUT

ISTANBUL, 2021

THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL BIG DATA ANALYTICS AND MANAGEMENT MASTER'S PROGRAM

DISCOVERING USERS' USAGE PATTERNS OF WEB LOG THROUGH ASSOCIATION RULES MINING METHODOLOGY

Master's Thesis

AHMAD HISHAM ARNAOUT

Thesis Supervisor: Assist. Prof. TAMER UÇAR

ISTANBUL, 2021

THE REPUBLIC OF TURKEY

BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL

...../...../.....

MASTER THESIS APPROVAL FORM

Program Name:	Master of Big Data Analytics
Student's Name and Surname:	Ahmad Arnaout
Name of The Thesis:	DISCOVERING USERS' USAGE PATTERNS OF WEB LOG THROUGH ASSOCIATION RULES MINING METHODOLOGY
Thesis Defense Date:	

This thesis has been approved by the Graduate School, which has fulfilled the necessary conditions as a Master thesis.

Assoc. Prof. Dr. Burak KÜNTAY

Institute Director

This thesis was read by us, quality, and content as a Master's thesis has been seen and accepted as sufficient.

	Title/Name	**Signature**
Thesis Advisor's	**TAMER UÇAR**	
Member		
Member		

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor, Dr. Tamer Uçar for giving me the chance to conduct research and providing me with precious guidance throughout this process. His vision, genuineness, and motivation have all left an indelible impression on me. He advised me how to conduct research and present my findings in the clearest and concise manner possible. Working and studying under his direction was a wonderful honour and privilege. I am appreciative of everything he has done for me. I would also like to express my gratitude for his friendship, empathy, and high spirits.

I owe my parents a debt of gratitude for their love, prayers, care, and sacrifices in teaching and supporting me in preparing for the future. My fiancée deserves special thanks for her love, patience, prayers, and unwavering support in helping me finish this research project. I also want to thank my sisters, brothers, for their help and prayers.

I am extending my thanks to Dr. Serkan Ayvaz, for his kindness and support throughout my study period. Finally, I want to express my gratitude to everyone who has helped me complete the research work, either directly or indirectly.

Istanbul, 2021 Ahmad Arnaout

ABSTRACT

DISCOVERING USERS' USAGE PATTERNS OF WEB LOG THROUGH ASSOCIATION RULES MINING METHODOLOGY

Ahmad Hisham Arnaout

Master of Big Data Analytics

Thesis Supervisor: Assist. Prof. Tamer Uçar

June 2021, 64 Pages

Web usage mining is an approach to discover user usage patterns and extract hidden knowledge from the weblog transaction dataset, this research aimed toward web usage mining analysis for weblog transactions gathered from an online menu of a restaurant. This research focuses on finding the correlation of frequently visited items that will impact on enhance the marketing strategies for the business. This is done using K-means clustering algorithm for grouping transactions into different groups based on their similarity within the transactions and two different methods used for evaluating the clustering process, Elbow and Silhouette coefficient method deployed for that purpose. Then association rule mining was applied to discover the correlation between the generated itemset and the algorithm Apriori and FP Growth has been used for finding frequent itemset. The result obtained from this research depends on the defined parameters provided by the user. This research is implemented using a Python programming language (3.8), MySQL and Apache HTTP Server.

Keywords: Web Usage Mining, Association Rule Mining, Clustering, User Usage Pattern, Weblog.

ÖZET

KULLANICILARIN WEB LOG KULLANIM ŞEKİLLERİNİN İLİŞKİLİ KURALLAR MADENCİLİK METODOLOJİSİYLE KEŞFİ

Ahmad Hisham Arnaout

Büyük Veri Analitiği ve Yönetimi Programı

Tez Yöneticisi: Asst. Prof. Dr. Tamer Uçar

Haziran 2021, 64 Sayfa

Web kullanımı madenciliği, kullanıcı kullanım kalıplarını keşfetmek ve web günlüğü işlem veri kümesinden gizli bilgileri çıkarmak için kullanılan bir yöntemdir. Bu araştırma, bir restoranın online menüsünden toplanan web günlüğü işlemleri ile yapılan web kullanımı madenciliği analizine yöneliktir. Bu araştırma, iş için pazarlama stratejilerini etkileyecek veya geliştirecek olan sık ziyaret edilen öğelerin korelasyonunu bulmaya odaklanmaktadır. Bunu yapmak için işlemler K-means kümeleme algoritmasıyla işlem benzerliklerine göre farklı gruplara ayırılmıştır ve kümeleme sürecinde değerlendirme için iki farklı yöntem, Elbow ve Silhouette katsayısı yöntemi kullanılmıştır. Sonra, Apriori algoritması ve öğe kümesi arasındaki korelasyonun tespiti için birliktelik kural çıkarımı (assossiation rule mining) uygulanmıştır ve sık öğe kümesini bulmak için FP Growth algoritması kullanılmıştır. Bu araştırmadan elde edilen sonuç, kullanıcı tarafından sağlanan tanımlanmış parametrelere bağlıdır. Bu araştırma bir Python programlama dili olan(3.8), MySQL ve Apache HTTP Sunucusu kullanılarak yapılmıştır.

Anahtar Kelimeler: Web Kullanımı Madenciliği, Birliktelik Kural Çıkarımı, Kümeleme, Kullanıcı Kullanım Kalıbı, Web Günlüğü.

TABLE OF CONTENTS

TABLES

FIGURES

EQUATION

1. INTRODUCTION

With the recent high heavy turnout for the companies and organization due to the Covid pandemic toward using various online solution and platform internally and externally to overcomes the restriction physical presence of the employee and serve the clients. Consequently, the business transactions explosive growth of information gathered from the various online channels.

Since this gathered information creates a new challenge for the business for extracting hidden knowledge by analyzing the data and getting informative and rich insights that can be used and reflected in different enhancement and business development areas, the knowledge of users' behaviors can be discovered by analyzing these data (Shih and Huang 2015).

(Zaw 2018) has mentioned that, From the business point of view, knowledge obtained from the web usage patterns could be directly applied to efficiently manage activities related to marketing strategies, web page personalization, etc.

In addition, (Persson 2017) concludes that, having indicators that signify user behavior related to high user retention rates can be useful in many ways to the company behind a product.

The project purpose is to discover user's usage patterns of web Log through association rules mining methodology, toward implement web usage mining technique on data that collected and granting from a company provide digital transformation solution for the franchise organization. In the analysis I conducted, we focused on a single restaurant data for the transactions that have been processed within 6 months, from September 2020 until March 2021.

The goals are to finding insights regarding the visited items to improve the process of enhancement the marketing campaigns in order to support the business marketing department to design a retargeting campaign based on their customers' visiting behavior.

Web mining (WM) is one of the data mining applications for the aim of extracting insights based on data generated from various levels and types of the web. WM is classified into three categories: Web Content Mining (WCM), Web Structured Mining (WSM), and Web Usage Mining (WUM) (Zaw 2018). Further, now worldwide-web usage applications are expanding day by day and consider as one of the biggest data sources (Sangavi, Suvetha and Umashankari 2016; Shih and Huang 2015).

Besides, improving the decision-making process for online businesses based on WUM application attracted high attention related to the current literature. WUM is an application used to analyze web log data of user's usage to discover interesting patterns and insights (Kasliwal and Katkar 2015). In this project clustering using K means clustering algorithm to group the transactions based on their similarities, two methods were used to evaluate the clustering algorithm Elbow and Silhouette Methods.

Then, associating rule mining using the Apriori algorithm and FP Growth were implemented for each group, the generated association rule demonstrates the correlation among the visited items under each group that will provide a better understanding for the decision-maker and enhance the campaign designing process.

This paper is organized as follows:

a. In Section 2, related work is discussed.
b. In Section 3: a detailed report about the collected dataset, the proposed data mining methods are presented, data preprocessing and implantation.
c. In Section 4: finding is concluded.
d. In Section 5: the conclusion.

2. LITERATURE REVIEW

Previous research pointed out that discovering and extracting hidden knowledge from weblog data provides a better understanding of online visitor's behavior in various areas. For example, Kasliwal and Katkar (2015) measure the frequent user visits to the website during the specific period using WUM and association rule mining by RapidMiner framework. They use the result to optimize the website and provide a recommendation for website visitors based on their behaviors.

In addition, Shih and Huang (2015) employ k-means clustering and the Apriori algorithm for WUM. They characterize Web users based on required criteria into different groups and that support the marketing decision based on each group's behavioral finding.

Also, Sharma, Khan, Singh, and Tiwari (2015) performs comparing D-APRIORI and DFP algorithm on data clustered using the Divisive Analysis method. The analysis study which algorithm is more efficient based on the memory usage and progress time of creating the association role.

Other studies claim that the enhancement of the recommendation system based on WUM application by performs comparing between KNN and Improved KNN classification are positively enhanced the accuracy as stated by Sangavi, Suvetha, and Umashankari (2016). Another instance is when Brilliant, Handoko, and Sriyanto (2017) concluded that by implementing data mining with the association rule method can help the company in finding consumer patterns.

It is expected that the company can create a list of entertainment services and product packages that can be offered to consumers based on the rules generated at competitive prices. They provide a result about customer selecting criteria of when regular tents select, a regular sound system it will also choose the decoration.

In addition, Persson (2017) investigates whether the usage patterns of retained users differ from the usage patterns of non-returning users, based on the usage during the first few sessions. He concludes that using association rule mining on identifying early usage patterns was significantly increased retention rates in a mobile web browser. The majority of the rules analyzed in this work imply a retention rate increase between 150 percent and 200 percent. At the same time, Mehrban (2017) proposes WUM using FP-growth on log files provides faster results on large data from other algorithms.

Furthermore, Kaur (2017) predicts user behavior via analyzing web log files using k means clustering and different association rule algorithms. He observed that the Apriori algorithm is easy to implement and k-means clustering produces tighter clusters whereas the FP-Growth pattern is faster than the Apriori algorithm.

Moreover, Zaw (2018) improves the new website based on pattern discovery using association rule mining on clustered data. He discovered how users are using the current website based on frequently accessed content together and that greatly improves the design of the new website layout by putting those content closed to each other.

Similarly, Herwanto and Ningtyas (2018) propose two-level recommendations based on association rules and topic similarity. Besides, Khan and Pandhare (2018) evaluate the usage pattern recognition using Apriori and FP tree algorithm. The output of the system was in terms of memory usage and speed of producing association rules.

Also, Mehra (2018) performs a data cleaning technique on server logs for reducing the number of records and therefore enlarges the quality of the available data. This research demonstrates the most popular page according to page access frequency.
Moreover, Serin (2018) examine the WUM using Fuzzy C-means clustering and the Apriori association rule algorithm. The result specifies the user's frequent browsing behavior and predicts the next visiting web page.

Nonetheless, Joshi (2019) amplifies the prediction of user behavior based on weblog using the Parallel FP-Growth Algorithm (IPFP). The results show that the IPFP algorithm is feasible and a higher mining efficiency and can meet the rapidly growing needs of frequent item sets mining for massive, small files datasets.

Conversely, research for discovering an association rule in WUM using the FP Tree algorithm by Thu and Khine (2019) has been conducted. The main result has displayed the Relationships of web server log to provide information to better accommodate the website based on user's needs. The mentioned studies explore and demonstrate how companies have become increasingly interested in the specific effects of extract hidden knowledge.

Besides, discovering hidden patterns from their data is another eager interest due to various types of scenarios.

As a result, this research attempts to analyze customers' transactions of Mbsher Company weblog dataset to understand visitors' behaviors' using K means clustering method with association rule mining and the algorithm Apriori and FP Growth for generating frequent itemset for answering the following question: Is there any strong correlation between the visited items and what is the measurement indicators of defining the power of that correlation.

The answers to those questions from discovering users' usage patterns provide better measurement on the impact enhancement of marketing campaigns in order to support the business marketing department to design a retargeting campaign based on their customers' transaction behavior.

Table 2.1: Papers of Reference

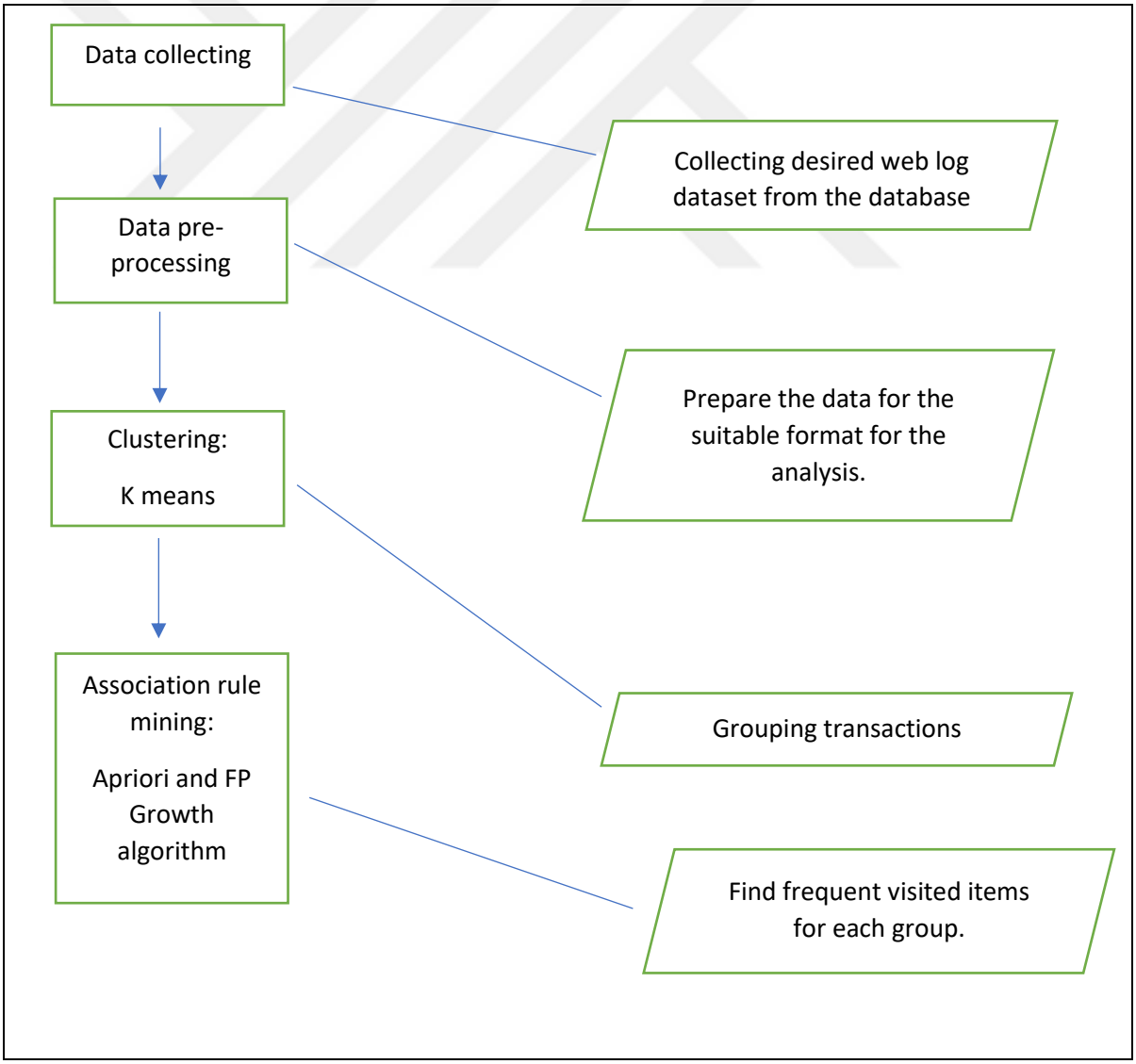
No.	Title	Auther	Year	Method
1	Web Usage mining for Predicting User Access Behaviour	Amit Dipchandji Kasliwal And Girish S. Katkar	2015	- Association rule mining RapidMiner framework
2	Characterizing Web Users Based on Their Required Criteria	Ming-Yi Shih and Syun-Sian Huang	2015	- K-means algorithm Apriori algorithm
3	An Analysis on Web Usage Mining For Internet Users	Priyanka Sharma, Sumayya Khan, Shilpa Singh and Pooja Tiwari	2015	- Divisive Analysis - D-APRIORI algorithm DFP algorithm
4	Web usage mining using Improved KNN Algorithm	Sangavi. S Suvetha. T Umashankari. T	2016	- KNN classification Improved KNN classification
5	Implementation of Data Mining Using Association Rules for Transactional Data Analysis	Muhamad Brilliant, DwiHandoko and Sriyanto	2017	FP-Growth algorithm
6	Identifying early usage patterns that increase retention rates in a mobile web browser	Pontus Persson	2017	- FP-Growth algorithm - DBSCAN clustering - Mini-batch K-Means clustering - Pearson's c2 Test - Manual Feature Selection No selection
7	WEB USAGE MINNING USING PATTERNS WITH DIFFERENT ALGORITHMS	SOBIA MEHRBAN	2017	- Apriori Algorithm. - FP- Growth Algorithm K-means Algorithm
8	Analysis of Web Usage Mining techniques to predict the user behavior from Web Server Log Files	Anmol Kaur	2017	- K means Algorithm - Apriori Algorithm FP-Growth algorithm

9	PATTERN DISCOVERY USING ASSOCIATION RULE MINING ON CLUSTERED DATA	HTUN ZAW OO	2018	- Density-based clustering Apriori algorithm
10	Recommendation system for web article based on association rules and topic modelling	Guntur Budi Herwantoa and Annisa Maulida Ningtyas	2018	- Latent Dirichlet Allocation (LDA) Apriori algorithm
11	Web Usage Mining and User Behaviour Prediction	Asiya N. Khan and Pallavi S. Pandhare	2018	- Apriori Algorithm - FP Tree Algorithm
12	An Effective method for Web Log Preprocessing and Page Access Frequency using Web Usage Mining	Jayanti Mehra	2018	- Data preprocessing algorithem in JAVA
13	Clustering based Association Rule Mining to discover user behavioural pattern in Web Log Mining	Serin . J	2018	- Fuzzy C-means clustering - Apriori association rule Algorithm
14	Enhancing Prediction of User Behavior on the basic of Web Logs	Mrunmayee R. Joshi	2019	- Apriori Algorithm - KNN Algorithm - FP-Growth (FP) algorithm - Improved Parallel FP-Growth Algorithm (IPFP)
15	Discovering Generalized Association Rule in Web Usage Mining by FP Tree	Han Ni Ni Myint Thu and Khine Khine Oo	2019	- FP-Growth algorithm

3. MATERIAL AND METHODS

This research intends to implement web usage mining for discovering hidden patterns and meaningful frequent visited items from a weblog of visitors transactions with the combination of clustering analysis using k means algorithm and association rule mining using the Apriori algorithm. The proposed workflow is shown in the following figure:

Figure 3.1: Proposed Workflow



The proposed workflow has four main steps: web log data collection, data pre-processing, clustering analysis using K means clustering and association rule mining.

3.1 DATA GATHERING AND PREPROCESSING

3.1.1 Data Gathering

The database granted from the Mbsher company. Mbsher is a company based in Saudi Arabia, they provide digital transformation solutions for franchises and hospitality. The database contains all their clients' data and distributed within separate tables where each table contains different information.

The Tables listed as following:

a. Places
b. Category
c. Items
d. Visits
e. Visits Info

3.1.1.1 Places table

The places table contains information regarding each tenant's information subscribe with the Mbhser solution. The total number of registered tenants: 96. The following figure shows the fields and types under the table:

Table 3.1: Display of table Place

Field	Type
Id	Bigint(20) unsigned
Domain	Varchar(255)
Foodics_Id	Varchar(255)
User_Id	Int(11) unsigned
Name	Text
Description	Text
Address	Longtext
Type	Varchar(255)
Currency	Text
Color1	Varchar(255)
Color2	Varchar(255)
Color3	Varchar(255)
Cover	Varchar(255)
Logo	Varchar(255)
Font	Varchar(255)
Contacts	Text
WithOrders	Tinyint(1)
SocialMedia	Text
VatValue	Int(11)
Takeaway	Tinyint(1)
Delivery	Tinyint(1)
TablesCount	Int(11)
Theme	Varchar(255)
Status	Int(11)
Foodics_Token	Text
Hyperpay_Token	Text
Hyperpay_EntitiyId_Mada	Text
Hyperpay_EntitiyId_Visa	Text
Hyperpay_Extra_Fees	Text
Cash_Method	Int(11)
Pos_Method	Int(11)
Online_Method	Int(11)
Deleted_At	Timestamp
Created_At	Timestamp
Updated_At	timestamp

3.1.1.2 Category table

The category table contains the category information defined under each tenant and it's linked with the places table through the Place_id filed. The following table for description for the category table fields:

Table 3.2: Display of table Category

Field	Type
Id	Bigint(20) unsigned
Parent_Id	Bigint(20)
Foodics_Id	Varchar(255)
User_Id	Int(11) unsigned
Place_Id	Int(11) unsigned
Name	Text
Foodics_Name	Text
Foodics_Ref	Varchar(255)
Description	Text
Img	Varchar(255)
Cover	Varchar(255)
Order	Int(11)
Main	Tinyint(1)
Status	Int(11)
Deleted_At	Timestamp
Created_At	Timestamp
Updated_At	timestamp

3.1.1.3 Items table

The items table list all the items defined for the provided services as a landing page for the product or service under the Mbhser solution that belong to every tenants' account and it linked with the tenants' services through places table and categorized through category table, as it defined under the table visits under the filed visit type, these items are linked with the mentioned table through the fields: place_id for Places table and category_id for category table.

The table describes the items table field:

Table 3.3: Display of table Items

Field	Type
Id	Bigint(20) unsigned
Foodics_Id	Varchar(255)
Foodics_Sku	Varchar(255)
Foodics_Barcode	Varchar(255)
User_Id	Int(11) unsigned
Place_Id	Int(11) unsigned
Category_Id	Int(11) unsigned
Foodics_Category_Id	Varchar(255)
Name	Text
Foodics_Name	Text
Description	Text
Foodics_Description	Text
Contents	Text
Img	Varchar(255)
Cover	Varchar(255)
Price	Double
Cal	Double
Orders	Int(100)
Status	Int(11)
ExtraFields	Longtext
Deleted_At	Timestamp
Created_At	Timestamp
Updated_At	timestamp

3.1.1.4 Visits table

The table visits contain basic information for each transaction session created from a visit done by one of the tenants' clients. The defined transaction starts from the moment one of the users enter the main page of any tenants' profile under the Mbsher solution and for each visited page rather it's category or item, for each navigation process attempt a transaction record is created under the visit table.

Table 3.4: Display of table Visits.

Field	Type
Id	Bigint(20) unsigned
Visitor_Id	Int(11) unsigned
Visitable_Id	Int(11) unsigned
Visitable_Type	Varchar(255)
Created_At	Timestamp
Updated_At	timestamp

The visits table defines the level of visit under the field: visitable_type, based on the visited page and it can one of the defined levels: Place, category, or item. Moreover, the visitable_id field shows the id for the visited page rather than its place "Main Page" from the table "Place", category from the table "Category", or an item from the table Items. Where for each single visit for any page, a transaction record created under visits table.

3.1.1.5 Visits info table

The table visits info contains weblog data for all tenant visitors' transactions.
The field of visits info table is designed and defined as follows:
a. Id: the primary key and demonstrate the id for each single transaction visit.
b. visit_id: the linked visit id with the visits table.

c. User_id: the linked visitor id with the user table.

d. Asn: is a unique identifier as an autonomous system number.

e. User_agent: the browser used by the visitors.

f. Language: the language of the visitors' devices.

g. Screen_resolution: the resolution of the devices.

h. Platform: the operation system OS of the devices.

i. Adblock: checking if visitors using AdBlock software.

j. Touch_support: checking if the used device's touch supported.

k. Ip: the IP address

l. Version: version of IP address

m. Region, city, country, country name, lat and lng, and utc_offset: information about the location of where the visitors browsing from.

n. Currency_name: the name of the currency used to pay or offer items.

o. Start and created_at: both fields contain the value of when the session transaction is started.

p. End and updated_at: demonstrate the date and time the session is ended.

This study will focus on a tenant: Chef's Homemade Burger Gourmet with the Place-ID: 46, based on information provided by the Mbsher company. Chef Burger is a restaurant uses a solution of the online menu as a service for their customers for booking their orders directly without engaging with the restaurant employee and even before they arrived the to the restaurant Branche to pick up the order.

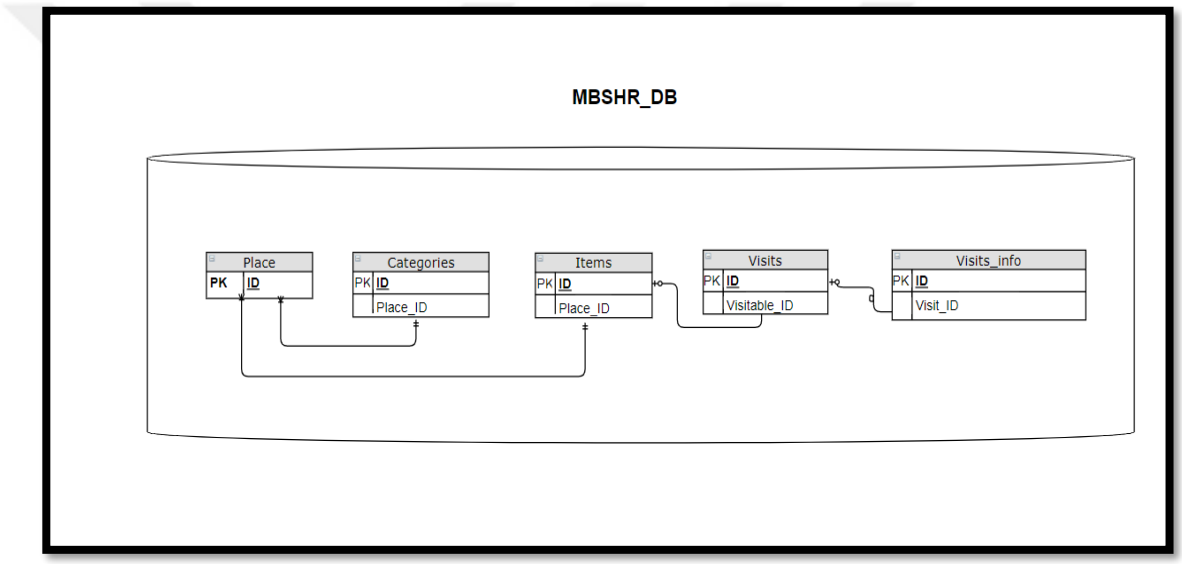
The transaction records are generated from the customer transactions on the QR menu. Each transaction demonstrates a single item visit for a single user at a particular time.

Thus, the desired dataset from the database will include only the data that belong to Chef's Homemade Burger Gourmet restaurant. Consequently, a sequence of SQL queries was implemented to collecting the required data.

The desired dataset will collect from table visits_info. As mentioned earlier, the table visits_info contains all detailed transaction information for all tenants' clients who used Mbsher solutions.

The figure demonstrates the tables with the relational on how the table records connected.

Figure 3.2: Mbsher database Tables

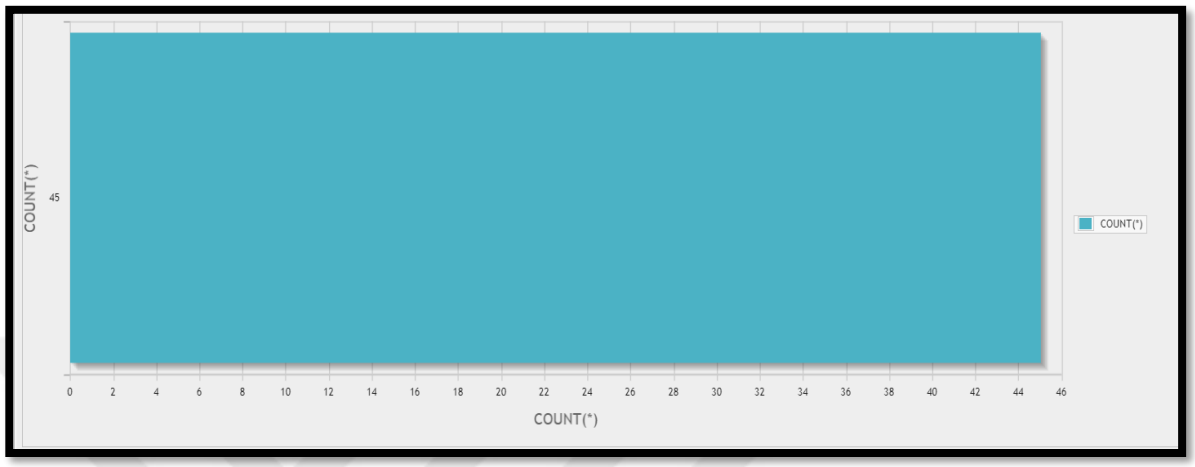


Since not all the characteristics are important for the research topic where the focusing of analysis is consist of the correlation between visited items. The transaction records for the place and categories are excluded.

The first step is selecting all the items that belong to Chef Burger restaurant with the place id: 46 from the table items into a new table with the title: Items_46.

The total number of items belonging to Chef Burger restaurant is: 45.

Figure 3.3 Count of Items for place Id:46



Second, in order to select all the transaction records attempted to the items belong to Chef burger restaurant only, a selecting query implemented for the records from the visits table with the condition defined of when the visitable id field under visits table is equal to the Id field under the recent created table items_46 Items. The query will select only the transaction record for the visited item for the Chef burger restaurant into a new table with the title: visits_46_items.

The table demonstrates a sample from the visits_46_items:

Table 3.5 Visitst_46_items caption

id	user_id	visitable_id	visitable_type	created_at	updated_at
40	1	2105	Item	9/22/2020 11:19	10/13/2020 14:06
41	1	2107	Item	9/22/2020 11:19	10/13/2020 14:06
56	129	2198	Item	9/22/2020 11:28	9/22/2020 11:28
59	129	2110	Item	9/22/2020 11:28	9/22/2020 11:28

60	129	2192	Item	9/22/2020 11:28	9/22/2020 11:29
61	129	2107	Item	9/22/2020 11:29	9/22/2020 11:29
67	129	2111	Item	9/22/2020 11:29	9/22/2020 11:29
68	129	2196	Item	9/22/2020 11:29	9/22/2020 11:29
70	129	2105	Item	9/22/2020 11:30	9/22/2020 11:30
74	129	2208	Item	9/22/2020 11:30	9/22/2020 11:30
77	129	2105	Item	9/22/2020 11:30	9/22/2020 11:30
91	129	2105	Item	9/22/2020 11:32	9/22/2020 11:32
92	129	2110	Item	9/22/2020 11:32	9/22/2020 11:32
93	129	2192	Item	9/22/2020 11:32	9/22/2020 11:32

Then, selecting all transactions records from visits_info into a new table with the title: visits_info_46, for all items that belong to Chef's restaurant through implementing the selecting query at when the visits_id field under the visits_info table is equal to the id value under the visits_46_items table.

Taking into consideration that the visits_info table does not contain the visitable id column that we define the visitable id in the selecting query.

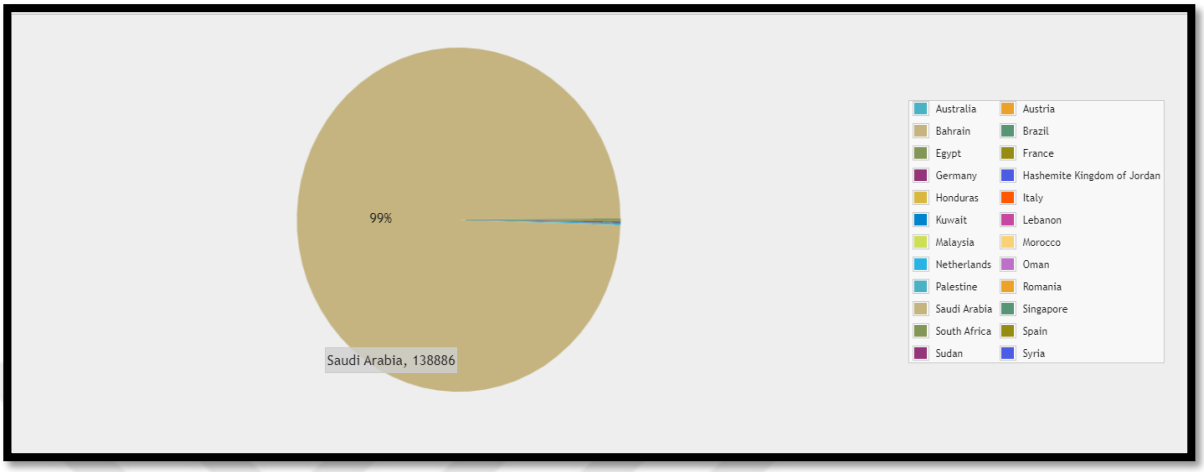
Table 3.6 Visits_info_46 caption

Field	Type
Id	Bigint(20) unsigned
Visit_Id	Int(11) unsigned
User_Id	Int(11) unsigned
Asn	Varchar(255)
User_Agent	Text
Language	Varchar(255)
Screen_Resolution	Text
Platform	Text
Adblock	Int(11)

Touch_Support	Text
Ip	Varchar(255)
Version	Varchar(255)
Region	Varchar(255)
City	Varchar(255)
Country	Varchar(255)
Country_Name	Varchar(255)
Lat	Varchar(255)
Lng	Varchar(255)
Utc_Offset	Double
Currency_Name	Varchar(255)
Start	Datetime
End	Datetime
Duration	Double
ExtraFields	Longtext
Created_At	Timestamp
Updated_At	timestamp
Visitable_Id	Int(11) unsigned

In addition, many records observed that it contains transaction records with countries out of Saudi Arabia, more with a blank value under the country name and other fields. As Mbsher technical department explains, this might happen due to several reasons and they recommended selecting the records when the country is: Saudi Arabia. The following figure 3.4 shows the records grouped by country name.

Figure 3.4: Records grouped by country = Saudi Arabia



The final data collected with the 138886 transactions records into a new the table with the title: visits_info_items_46 exported as CSV and uploaded to python for starting data pre-processing.

3.1.2 Data preprocessing

The data pre-processing phase is proposed through three steps:

a. Data cleaning
b. User Identification
c. Session identification

3.1.2.1 Data cleaning

The collected dataset contains 26 fields or columns as mentioned earlier in chapter 3 under the materials. The first step of data preprocessing was implemented on the table before exported as a CSV file when the transactions under the country name: Saudi Arabia was selected.

The company that provides the explained that due to continuous development process and upgrade on their product features many of transactions contains null value since there were

features added recently and that drive to missing for many of characteristics. For instance, platform, language, and user_agent.

These fields that contain missing value and based on the information provided by Mbsher the selected data based on country as Saudi Arabia provide the most efficient data to use. The features were selected upon the purpose of the research, to finding the frequently visited items by clients.

Consequently, the fields user_id, updated_at, and vistable_id was selected as well.
After the field was chosen, all not necessary fields were dropped, and the shape of data was modified as explained in the following table:

Table 3.7: Shape of dataset before and after selecting

Data set Shape	
Orginal data set	After Fields Selected
(138886, 26)	(138886, 3)

Moreover, the following table, display the first 10 records from the dataset that demonstrate for each record presents a single visit for an item.

Table 3.8: caption from dataset for the first 10 records

User_Id	Updated_At	Visitable_Id
1	10/13/2020 14:06	2105
1	10/13/2020 14:06	2107
129	9/22/2020 11:28	2198
129	9/22/2020 11:28	2110
129	9/22/2020 11:29	2192
129	9/22/2020 11:29	2107
129	9/22/2020 11:29	2111
129	9/22/2020 11:29	2196
129	9/22/2020 11:30	2105
129	9/22/2020 11:30	2208

For instance, the result shows that for the user with id:1 there are two transactions on the same date each one is for a different item, and same for the user with the id:129 there are Eight transactions records, each record is a visit for a different item.

3.1.2.2 User identification

The user identification is the process of identifying every single user. For this purpose, the user_id field was selected instead of the IP field for the reason that the user_id is representing the client ID that defined with the client account on the system when their register on the system as a unique ID as explained by the company.

3.1.2.3 Session identification

The session identification is the process of identifying every single visit for each client as one session that will include all the dates of the session for all visited items by each client. For the session identification process, a method is implemented to select all visited items within the same session and for each user in one record.

The method for user and session identification was defined through converting the field updated_at from a timestamp type to date type and for every single day and for particular user_id it considered as a single session. Consequently, the visited items through each transaction were fit into a single record and by this way as an output of the method, each session record will demonstrate the id of the visited user, the date of visit, and the visited items id.

The table demonstrates the original dataset structure:

Table 3.9: Original dataset structure

User_Id	Updated_At	Visitable_Id
1	10/13/2020 14:06	2105
1	10/13/2020 14:06	2107
129	9/22/2020 11:28	2198
129	9/22/2020 11:28	2110
129	9/22/2020 11:29	2192
129	9/22/2020 11:29	2107
129	9/22/2020 11:29	2111
129	9/22/2020 11:29	2196

The following table, demonstrates the head of dataset after users and session identification process have been implemented:

Table 3.10: Dataset structure after session identification

user_id	New_date	visitable_id
1	2020-10-13	[2105 2107]
1	2020-10-15	[6744 4905 6744 6744]
129	2020-09-22	[2198 2110 2192 2107 2111 2196 2105 2208 2105 2105 2110 2192 2192 2196 2111]
145	2020-09-22	[2105 2107 2111 2107 2192 2198 2107 2107 2109 2109]
151	2020-09-23	[2105 2109 2109 2110]
151	2020-10-09	[2110 2415 4905]
155	2020-11-15	[7072 2209 4905 2415 2209 7072]
162	2020-11-16	[2209 2194 2109 2105]
200	2020-09-22	[2194 2198 2192 2192 2105 2194 2192 2194]

The other method was defined to add all items id as a feature, each item as a single feature, and under each feature assign the value of ONE for the transaction value that contains that item under the visited items features.

In other words, a new field was modified for each item and the value:1 was assigned for the visited item and 0 assigned for the non-visited item within each transaction.

The following table, demonstrate an example for record, the method assigned value 1 under the visited items attribute:

Table 3.11: Caption of assigning value to attributes

user_id	New_date	visitable_id	2104	2105	2106	2107	2108
1	2020-10-13	[2105 2107]	0	1	0	1	0
1	2020-10-15	[6744 4905 6744 6744]	0	0	0	0	0
129	2020-09-22	[2198 2110 2192 2107 2111 2196 2105 2208 2105 2105 2110 2192 2192 2196 2111]	0	1	0	1	0
145	2020-09-22	[2105 2107 2111 2107 2192 2198 2107 2107 2109 2109]	0	1	0	1	0
151	2020-09-23	[2105 2109 2109 2110]	0	1	0	0	0
151	2020-10-09	[2110 2415 4905]	0	0	0	0	0

Then, the date and user id fields are dropped, and the final format dataset shape consists in (16046, 45) as shown in the table: in order to prepare the dataset for suitable structure for clustering and association rule algorithms.

Table 3.12: Caption of final dataset structure

2104	2105	2106	2107	2108	2109	2110	2111	2112	2113
0	1	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	1	0	1	0	0	1	1	0	0
0	1	0	1	0	1	0	1	0	0
0	1	0	0	0	1	1	0	0	0

In addition, the number of transaction records has reduced after implementing the pre-processing steps as shown in the following table:

Table 3.13: Shape of dataset before and after pre-processing.

Data set Shape	
Before Pre-processing	After Pre-processing
(138886, 3)	(16046, 45)

3.2 METHODS

This section explains the analysis method used in implementing the project. The chosen clustering algorithm is K means algorithm. As it is one of the most popularly used clustering algorithms in Web usage mining applications (Shih aand Huang 2015).

Moreover, two evaluation methods Elbow, and the Silhouette method employed to define the optimal number of clusters needed since using K means clustering algorithm requires to define the number of clusters to group data based on it. Moreover, an association rule mining Apriori and FP Growth algorithm were deployed in this research as there are of the most well-known algorithms for association analysis according to (Shih and Huang 2015) and (Persson 2017), and by observing the number of overall items from the gathered data set is 45, Apriori algorithm is considered as a more efficient algorithm for this research.

In addition, the FP growth algorithm is deployed as on of the fastest algorithm for generate the frequent item set as defined by (Mehrban 2017).

3.2.1 K means Clustering Algorithm

K means clustering is an unsupervised machine learning algorithm that used to group data based on their similarity by measure the distance between the point and the defined number of clusters centroid (Mayo 2020).

The input in k means model is basically a set of inputs points as X1, X2…. Xn. In addition to K value that defines the number of clusters. The model starts by place the centroid location

randomly then compute the distance between every data point and every single centroid and assigning the points with the nearest distance for each centroid based on the number of K.

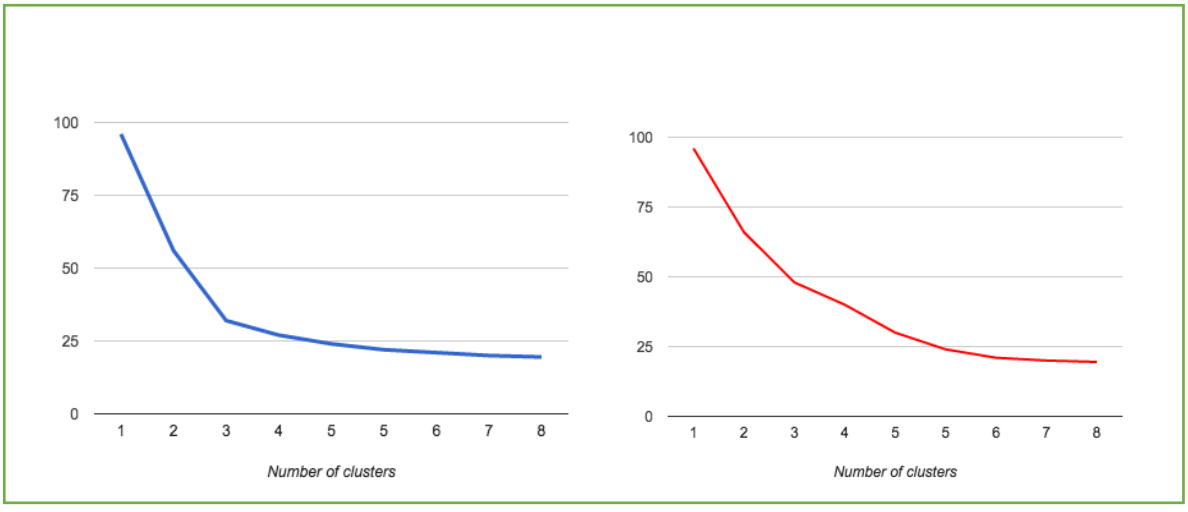
Then the model recomputes the centroid position by taking the vectors of all the points that belong for each cluster from the previous step and compute the average and the average points will be the new centroid for the cluster. This process is repeated until no change on the centroid location and no point moved from one cluster to another.

3.2.2 Elbow Method

The Elbow method runs k means algorithm multiple times within a range of defined candidate value of K e.g.,1 to 10. It calculates the average distance from the centroid to each value across all points (Franklin 2019).

Then check when the average distance of the centroid stops falling suddenly, which means the decrease in dissimilarity is leveling out. At that point, it can be considering the value of K is the optimal number of clusters.

Figure 3.5: example of Elbow method plot



Source: (ERIK RANBY 2016)

3.2.3 Silhouette Method

Silhouette methods calculate the Silhouette Coefficient for each data point p as we define a(p) = average distance of p to other points under the same cluster as the average dissimilarity between p and the other data points in the same cluster, and b(p) = average distance to the nearest other cluster points as the lowest average dissimilarity between p and data points in other clusters (Ranby 2016). Then compute Silhouette Coefficient for each p calculated as the following formula:

Equation 3-1: Silhouette method formula

$$s(p) = b(p) - a(p) \,/\, max(a(p), b(p)) \quad (3.1)$$

Source:(ERIK RANBY 2016)

The ideal value of Silhouette closest = 1 and worst possible closest = -1, where the value of Silhouette closest to +1 indicate there is a long distance between the point and from the neighbor cluster and the value closest to -1 indicate that there is a very short distance between that point and the neighbor cluster (Chaudhary 2020).

3.2.4 Association Rule Mining

Association rule mining is method associations and correlations among variables in a large dataset. Let the rule discovered be X → Y, where X and Y are items in the dataset, and the left side X is the antecedent of the rule, and on the right side Y is the consequent of the rule.

The rule defines an association between items by measuring the frequent itemset by compute the support, confidence, and left. The support is computing for the frequency of occurrence of antecedent X or combination of item X and Y as an antecedent in the overall dataset items. The confidence computes how often the union of antecedent and consequent items occur in transactions that contain the antecedent items.

Left formula defined as follows: Let X as antecedent and Y as consequent, the lift is the confidence X to Y divided by the probability of Y: P(Y), where P(Y) is the percentage of transactions contains Y.

Consequently, Left provides better assessment or probability of association between items (Steinbach, Tan, and Kumar 2005).

3.2.5 Apriori Algorithm

Alpriori is one of the most well-known algorithms used to identify frequent itemsets. The algorithm used for generated frequent itemsets. It uses a breadth-first leveled approach for generating candidate itemsets by iteratively creating larger and larger item sets (Persson 2017).

The algorithm starts with a check of each candidate single item with sufficient support as defined in the model in the first iteration and considers these items as frequent items in the first level.

Consequently, in the second iteration, the algorithm combines the chosen frequent items from the first iteration and combines each two together, and considers each subset of items from

the previous iteration as a single candidate. For instance, let's consider the candidate in the first iteration as {A}, {B}, {C}, {D} and E.

Assuming in the first iteration the candidate {A}, {C} and {E} consider as an item with sufficient support. In other words, the support for these items is equal to or higher than the given support.

Thus, in the second iteration, the subset candidate will define as: {A, C}, {A, E} and {C, E}. Then the candidate with sufficient support will be chased as a frequent itemset. This process will continue until there is no possible frequent item set that can generate.

The confidence also may use to create the frequent itemset based on support and confidence as well. Finally, the items that found infrequent based on support or confidence or both together consider as pruned subsets, and its excluded from the generated frequent item set (Steinbach, Tan, and Kumar 2005).

3.2.6 FP Growth

FP growth or frequent pattern tree algorithm is another algorithm for generate frequent item set. The algorithm works on way of generates only the frequent item without a candidate items generation (Han 2014). The first step is created sorted list of items set that represent each single item with the support count. Then a FP tree created with a null root nude, the first tree left node is start with the item with the highest support count. Then the sequence of nodes represents based on the sequence of the item in each single transaction. For any cross combination between items in different transactions, no new root created if there an exist root with same items is already created. As well as, for each item represent within single root many times, the count of that item in the tree is increased one by one.

Consequently, after all transaction's items defined in the FP Tree. The conditional pattern base is defined for each item, starting from the below of the of sorted list with the least support count, for each item the root for that item the count of that item only.

Then, for the conditional FP tree, the total count for each item under single root is calculated, the items with countless that the minimum support is excluded.

Finally, for generate the frequent items, a joint for each single item in the root with the item in the leaf is applied. In case there were item represent in two different roots, the sum of count is calculated for that particular item (Charu and Han 2014).

3.3 IMPLEMENTATION

The implementation process will consist of several phases. data pre-processing as explained in section 3.1.2, that cover the process of preparing the dataset for the suitable format for applying clustering and association rules models which includes the feature re-engineering process.

Clustering analysis: after setting the data a cluster analysis will implement to group clients based on their transaction's similarity. Finally, generate the rules and finding similarities within each group that will provide insights and better understanding for clients visiting behaviors.

3.3.1 Clustering Analysis

The clustering analysis is one of the major steps in web usage mining. Clustering was used in transactions segmentation to find the transactions with similar characteristics in terms of

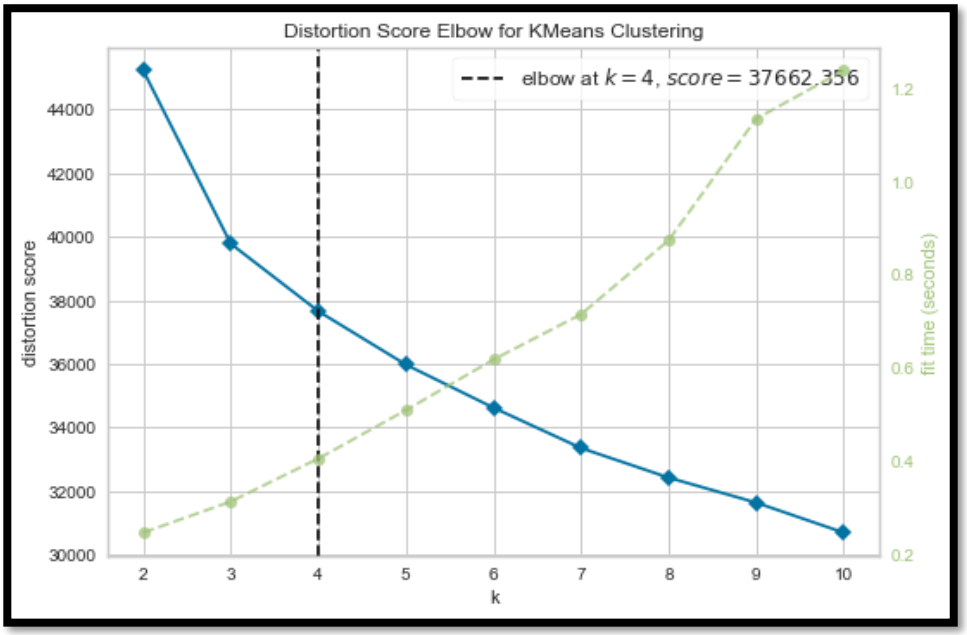
visited items. As mentioned in section 3, the K-means clustering algorithm will deploy in this paper. K-mean is one of the famous clustering methods and easy to implement, and as a fundamental step of unsupervised learning algorithm, it is very important to define the optimal number of clusters since with the K-means algorithm the number of clusters K must be specified before clustering.

Consequently, as mentioned in section 3, two methods to investigate and discover the optimal number of clusters will implement, the Elbow method and the Silhouette Coefficient method.

3.3.2 Elbow Methods

The Elbow method help in selecting the optimal number of clusters based on the average distance from the centroid to each value across all points. The function KElbowVisualizer from library yellow brick. a cluster was used to define the k-means model with a range number of K clusters defined from 2 to 9.
The Figure 4.18 demonstrates the plot of the Elbow method with a K range between 2 to 9.

Figure 3.6: Elbow Method plot



From the previous figure, the elbow method plot shows the distortion score differentiates between K=4 and K=5 start decreased significantly and at this point the concluded optimal number of clusters is K=4.

Moreover, to empower the decision of a selected optimal number of clusters the method silhouette coefficient implemented.

3.3.3 Silhouette Coefficient

As explained in section 3.2.3, the Silhouette Coefficient is a method to evaluate the clustering model though measure the separation distance between the points of the resulting cluster.

The following table, shows the plot of the silhouette coefficient average score with different no of K:

Table 3.14: Silhouette Coefficient Average Score

No of Clusters	Silhouette Average Score
2	0.346
3	0.253
4	0.222
5	0.218
6	0.171
7	0.167
8	0.157
9	0.144

Where the silhouette coefficient average score observed significantly decreased when K between 4 and 5, it conducted the optimal number of K based on silhouette coefficient method is k=4. As observed with the Elbow method. Thus, the K means cluster model was defined with a K number of clusters equal to 4.

3.3.4 K means Algorithm

The k means algorithm model defined with setting the number of clusters K = 4 based on the result from the methods elbow and silhouette. Consequently, the dataset fit in the model and cluster label was assigned for every single transaction, then a method to divide the samples based on its labeled cluster into four different data frames were implemented to prepare the data for the association rule mining algorithm on each cluster to discover the correlation between the visited items.

In addition, the K means have an attribute called: random estate. This attribute provides the ability to define the value of the initial centroid and its value can be defined as an integer it could be 0 or any other value. The aspect of define this value is based mainly on better understanding the data and experience in the domain is a major factor.

In this model, a sequence number of integers was used it with silhouette and the best result was with the random estate equal 9. In case this variable not defined manually, the model will set the initial centroid randomly and that will derive the model every time it runs will provide a different result with clustering the data. The following table, provide the shape of each cluster:

Table 3.15: Shape of clusters

DF Name	Shape of DF
cluster_one	(3972,45)
cluster_two	(2783,45)
cluster_three	(6892,45)
cluster_four	(2399,45)

3.3.5 Association Rule Mining

This section will discuss the final step of the process by implanting the association rule algorithm on the result obtained on each cluster created as explained in the previous section. The method Apriori algorithm and FP Growth were used to create the frequent item set. Then association rule algorithm will be implemented on the generated frequent item set to investigate the association between the items that have been visited based on the result of the support, confidence, and the left value for each fined association.

The first step is to generate the frequent user item set using the Apriori algorithm and FP Growth. The Apriori algorithm generates the frequent itemset by defining the minimum support value that calculates the frequency of the given items among the total number of transactions for the given data and pruning the items that considered as infrequent itemset with the fraction value under the defined minimum support ratio.

Moreover, each fraction frequency for the subset from the itemset that defined as a frequent itemset measured with the defined minimum support to assign the subset itemset as frequent or infrequent. This process continues until no more itemset were generated.
Both the Apriori algorithm and FP Growth required to define the parameter of the minimum support while the algorithm will calculate the confidence that defines the correlation between each antecedent and the consequently generated item in addition to the third measure that is Lift to provide a better indicator for the correlation between visited items. If the minimum support defines with low value, the generated items with low frequency will be generated as well.

The confidence defines how strong the correlation between two different items and the left will empower the process of select the strongest association rule from each cluster.

3.3.6 Association Rules using Apriori

3.3.6.1 Association rules using Apriori on first cluster

The first cluster shape is (3972,45). The Apriori algorithm first defined with minimum support equal to 0.7. the following table 3.16 demonstrates the generated itemset:

Table 3.16: Generated itemset using Apriori with minimum support = 0.7

Support	Itemsets
0.749748	2209
0.927996	2415
0.860775	4905
0.794310	4905, 2415

But when the association rule algorithm implemented to discover the correlation between the generated itemset there were no correlation between any itemset.

The minimum support value decreased to 0.5, the following table 3.17, demonstrates the generated items set:

Table 3.17: Generated itemset using Apriori with minimum support = 0.5

Support	Itemsets
0.749748	2209
0.927996	2415
0.860775	4905
0.632931	7072
0.678499	2209, 2415

Consequently, the association rule implemented to discover the correlation between items and there was only one association was discovered as demonstrated in the following table 3.18:

Table 3.18: Discovered association rules using Apriori with minimum support = 0.5

Antecedents	Consequents	Antecedent support	Consequents support	Supprot	Confidence	Lift
7072	2209	0.632	0.749	0.519	0.821	1.095

In addition, setting the minimum support to 0.3. the following association with lift higher than 1.2 as shown in the table 3.19:

Table 3.19: Discovered association rules using Apriori with minimum support = 0.3

Antecedents	Consequents	Antecedent support	Consequents support	Supprot	Confidence	Lift
2209	2106	0.749	0. 359	0.303	0.405	1.127

Moreover, the minimum support decrease to 0.23. an association between 17 itemsets were discovered the following table 3.20, shows the association with lift higher than 2:

Table 3.20:Discovered association rules using Apriori with minimum support = 0.23

Antecedents	Consequents	Antecedent support	Consequents support	Supprot	Confidence	Lift
2106	2108	0.359	0.324	0.236	0.656	2.025

The minimum support ratio that given to the Apriori algorithm to investigate the correlation between the generated itemset with higher value does not provide interest association when it started with 0.7 since it didn't provide any association at all.

Consequently, when the minimum support decreased to 0.5 and 0.3 there was an association founded but the measurement of lift value was not good since it was near to one and as

explained in section 3 as much as the lift was higher in parallel with confidence and support as much as it can consider as strong correlation.

Thus, when the minimum support defined as equal to 0.23, the association founded as demonstrated in the table conduct there is a strong correlation between the items.

3.3.6.2 Association rules using Apriori on second cluster

The second cluster two contains a smaller number of samples compared with the first cluster. The shape of cluster two is: (2783,45). The process sequence for implanting the Apriori algorithm in the first cluster was applied in the second cluster. It started with defining the minimum support equal to 0.7, 0.5, 0.2, and 0.07.

The following table 3.21, shows the generated itemset when minimum support 0.7, without any association rule discovered:

Table 3.21: Generated itemset using Apriori with minimum support = 0.7

Support	Itemsets
0.996	2106

The following table 3.22, shows the generated itemset when minimum support 0.5, and same with the minimum support equal 0.7, no association rule discovered.

Table 3.22: Generated itemset using Apriori with minimum support = 0.7

Support	Itemsets
0.996	2106
0.586	2108
0.582	2106, 2108

When define the minimum support equal to 0.2, The following table 3.23, shows the generated itemset:

Table 3.23: Generated itemset using Apriori with minimum support = 0.2

Support	Itemsets
0.996	2106
0.586	2108
0.225	2105
0.225	2105, 2106
0.582	2106, 2108

And single association rule was found as shown in the following table 3.24:

Table 3.24: Discovered association rules using Apriori with minimum support = 0.2

Antecedents	Consequents	Antecedent support	Consequents support	Supprot	Confidence	Lift
2105	2106	0.225	0.996	0. 225	1	1.003

Consequently, when the minimum support defined as equal to 0.07, the number of the generated frequent itemset is 53 and the association rule discovered was 51. The following table 3.25, demonstrates the discovered association with higher left and confidence:

Table 3.25: Discovered association rules using Apriori with minimum support = 0.07

Antecedents	Consequents	Antecedent support	Consequents support	Supprot	Confidence	Lift
2107	2105	0.137	0.225	0.079	0.578	2.567
2106,2107	2105	0,137	0.225	0.079	0.578	2.567
2106,2105	2107	0.225	0.137	0.079	0.352	2.567
2197	2108	0.145	0.586	0.127	0.874	1.490
2107	2106	0.137	0.996	0.137	1	1.003
2110	2106	0.091	0.996	0.091	1	1.003
2111	2106	0.141	0.996	0.141	1	1.003
2192	2106	0.114	0.996	0.114	1	1.003
2194	2106	0.119	0.996	0.119	1	1.003
2198	2106	0.074	0.996	0.074	1	1.003
4905	2106	0.140	0.996	0.140	1	1.003
6744	2106	0.083	0.996	0.083	1	1.003
7072	2106	0.116	0.996	0.116	1	1.003
2105, 2107	2106	0.079	0.996	0.079	1	1.003
2105, 2108	2106	0.137	0.996	0.137	1	1.003
2107, 2108	2106	0.081	0.996	0.08	1	1.003
2111, 2108	2106	0.094	0.996	0.094	1	1.003
2192, 2108	2106	0.073	0.996	0.073	1	1.003
2194, '2108	2106	0.075	0.996	0.075	1	1.003
4905, 2108	2106	0.097	0.996	0.097	1	1.003
2196	2106	0.111	0.996	0.111	0.996	1.000
2197	2108	0.145	0.586	0.127	0.874	1.490
2106,2197	2108	0.141	0.586	0.123	0.870	1.484
2197	2106, 2108	0.145	0.582	0.123	0.847	1.453
2415	2108	0.149	0.586	0.121	0.812	1.385
2106, 2415	2108	0.149	0.586	0.121	0.812	1.383
2415	2106, 2108	0.149	0.582	0.121	0.808	1.386

The implementation of association rule one the second cluster does not provide any correlation between the itemset when the minimum support defined as 0.7 and 0.5. In addition, when the minimum support set equal to 0.2 there was only a single correlation discovered with lift 1.003 and confidence equal to 1.

Taking into consideration that when lift value is higher than 1 as much as it considers strong correlation but also the value of confidence conducts the correlation as strong one.

Deeper investigation to figure higher lift, the minimum support equal to 0.07 shows that there was more association found as shown in the table but with lower confidence that conduct the association figured when minimum support equal to 0.2 is much better.
Also, the conservation from the previous table shows that the items with the Id 2106 and 2108 a very strong correlation between them when it assigned with other items with confidence equal to 1 in most cases.

Thus, if the data contained the name of the items it might be more clear to understand the reason for item 2106 got the optimal confidence value for the correlation as 1 and 2108 very near to that as well.

3.3.6.3 Association rules using Apriori on third cluster

The Apriori algorithm defined minimum support with values 0.7 and 0.5 does not generate any frequent itemset. Moreover, at minimum support equal to 0.2 generate only a single itemset as shows in the table:

Table 3.26: Generated itemset using Apriori with minimum support = 0.2

Support	Itemsets
0.266	2105

On the other hand, a frequent itemset were generated when the minimum support value equal to 0.05. The association rules discovered with the higher value of lift that greater than 3 as shows in the following table:

Table 3.27: Discovered association rules using Apriori with minimum support = 0.05

Antecedents	Consequents	Antecedent support	Consequents support	Supprot	Confidence	Lift
2196	2111	0.124	0.134	0.063	0.511	3,812
2107	2105	0.143	0.266	0.078	0.545	2.04
2194	2105	0.159	0.266	0.057	0.362	1.360

In addition, setting the minimum support to 0.03 provided good association with high value of confidence and lift as the result demonstrated in the following table:

Table 3.28: Discovered association rules using Apriori with minimum support = 0.03

Antecedents	Consequents	Antecedent support	Consequents support	Supprot	Confidence	Lift
2111, 2196	2112	0.063	0.063	0.033	0.519	8.228
2111,2112	2196	0.042	0.124	0.033	0.783	6.293
2112	2196	0.063	0.124	0.044	0.705	5.669
2196, 2112	2111	0.044	0.134	0.033	0.742	5.533
2112	2111	0.063	0.134	0.042	0.668	4.984

Even though the third cluster is a bigger one as a number of samples, where the shape of the third cluster is (6892,45). But when the Apriori algorithm minimum support defined with a value higher than 0.07 there was no association rule generated. Moreover, the lift value from the association rules generated from the third cluster was higher compared with the previous two.

3.3.6.4 Association rules using apriori on forth cluster

The Forth cluster shape is (2399,45), the minimum support defined on the fourth cluster equal to 0.6. The associations discovered from the cluster for 18 strong correlations between itemset, with high support greater than 0.6, confidence greater than 0.7, and lift greater than 1. The following table, demonstrates the discovered associations rules:

Table 3.29: Discovered association rules using Apriori with minimum support = 0.6

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
2107	2105	0.828	0.879	0.765	0.924	1.051
2105	2109	0.879	0.734	0.659	0.750	1.020
2194	2105	0.780	0.879	0.687	0.880	1.002
2107	2109	0.828	0.734	0.659	0.796	1.084
2107	2110	0.828	0.832	0.695	0.839	1.007
2109	2110	0.734	0.832	0.656	0.892	1.071
6744	2110	0.756	0.832	0.684	0.904	1.086
2194	2192	0.780	0.805	0.677	0.867	1.077
2192	6744	0.805	0.756	0.636	0.790	1.044
2194	6744	0.780	0.756	0.604	0.774	1.022
2105, 2107	2109	0.765	0.734	0.613	0.801	1.090
2107, 2109	2105	0.659	0.879	0.613	0.929	1.057
2105, 2109	2107	0.659	0.828	0.613	0.930	1.123
2105, 2107	2110	0.765	0.832	0.642	0.839	1.008
2107, 2110	2105	0.695	0.879	0.642	0.924	1.051

2105, 2110	2107	0.725	0.828	0.642	0.885	1.069
2192, 2107	2105	0.661	0.879	0.614	0.929	1.057
2192, 2105	2107	0.704	0.828	0.614	0.871	1.052

3.3.6.5 Association rules using Apriori on original data

The final approach of association rule was implemented on the original dataset with the overall number of samples without clustering. When minimum support defined as 0.7, 0.5, and 0.4 the Apriori algorithm does not generate any frequent itemset. Consequently, the minimum support value decreased to 0.3 only four frequent itemsets generated but without any association rule found.

Moreover, when the minimum support defined as equal to 0.2, 15 frequent item sets were generated as the estate in the following table:

Table 3.30: Generated itemset using Apriori with minimum support = 0.2

Support	Itemset
0.367	2415
0.353	4905
0.330	2105
0.329	2106
0.297	2209
0.276	2108
0.270	7072
0.258	4905, 2415
0.236	2107
0.234	2194
0.227	2110
0.218	2192
0.215	2415, 2209
0.207	2106, 2108
0.204	6744

and three strong association rules were discovered as demonstrate in table:

Table 3.31: Discovered association rules using Apriori with minimum support = 0.2

Antecedents	Consequents	Antecedent support	Consequents support	Supprot	Confidence	Lift
2106	2108	0.276	0.329	0.207	0.7512	2.278
4905	2415	0.353	0.367	0.258	0.7300	1.985
2209	2415	0.297	0.367	0.215	0.7242	1.969

Even though the value of minimum support for the full data was not too low, but still there was a frequent dataset generated at 0.2. with high confidence and support.

Lastly, from the result obtained from association rule mining on each cluster the conducted correlation for the overall visited items with the highest value for confidence and lift were demonstrated in the following table:

Table 3.32: discovered associations among clusters

No.	Antecedent	Consequent	Cluster
1	7072	2209	1
2	2106	2108	1
3	2105	2106	2
4	2107	2105	2
5	2106,2107	2106	2
6	2106,2105	2107	2
7	2196	2111	3
8	2111,2112	2196	3
9	2112	2196	3
10	2105	2109	4
11	2194	2105	4
12	2107	2109	4
13	2107	2110	4

14	2109	2110	4
15	6744	2110	4
16	2194	2192	4
17	2192	6744	4
18	2194	6744	4
19	2105, 2107	2109	4
20	2105, 2107	2110	4
21	2192, 2107	2105	4

3.3.7 Association Rules using FP Growth

The FP Growth algorithm defined with the same values of minimum support for each cluster as implemented in using the Apriori algorithm.

3.3.7.1 Association rules using fp growth on first cluster

The defined minimum support equal to 0.7, 0.5, 0.3 and 0.23 as demonstrates in the following tables.

Table 3.33: Generated item set using FP Growth at minimum support = 0.7

Support	Itemsets
0.927	2415
0.860	4905
0.749	2209
0.794	4905, 2415

And no association rule was discovered.

Table 3.34: Generated item set using FP Growth at minimum support = 0.5

Support	Itemsets
0.749748	2209
0.927996	2415
0.860775	4905
0.632931	7072
0.678499	2209, 2415

And only one association was discovered as demonstrated in the following table:

Table 3.35: Discovered association rules using FP Growth at minimum support = 0.5

Antecedents	Consequents	Antecedent support	Consequents support	Supprot	Confidence	Lift
7072	2209	0.632	0.749	0.519	0.821	1.095

Table 3.36: Discovered association rules set using FP Growth at minimum support = 0.3

Antecedents	Consequents	Antecedent support	Consequents support	Supprot	Confidence	Lift
2209	2106	0.749	0. 359	0.303	0.405	1.127

Table 3.37:Discovered association rules using FP Growth at minimum support = 0.23

Antecedents	Consequents	Antecedent support	Consequents support	Supprot	Confidence	Lift
2106	2108	0.359	0.324	0.236	0.656	2.025

3.3.7.2 Association rules using fp growth on second cluster

The following table, shows the generated item set using FP growth when minimum support 0.7, without any association rule discovered:

Table 3.38: Generated item set using FP Growth at minimum support = 0.7

Support	Itemsets
0.996	2106

The following table shows the generated item set using FP Growth, when minimum support 0.5, and no association rule discovered.

Table 3.39: Generated itemset using FP Growth at minimum support = 0.5

Support	Itemsets
0.996	2106
0.586	2108
0.582	2106, 2108

The following table 3.40, shows the generated item set using FP Growth, when minimum support equal to 0.2:

Table 3.40: Generated itemset using FP Growth at minimum support = 0.2

Support	Itemsets
0.996	2106
0.586	2108
0.225	2105
0.225	2105, 2106
0.582	2106, 2108

And only a single association rule was found as shown in the following table:

Table 3.41: Discovered association rules using FP Growth at minimum support = 0.2

Antecedents	Consequents	Antecedent support	Consequents support	Supprot	Confidence	Lift
2105	2106	0.225	0.996	0. 225	1	1.003

Consequently, the following table, demonstrates the discovered association with higher left and confidence when the minimum support defined as equal to 0.07.

Table 3.42: Discovered association rules using FP Growth at minimum support = 0.07

Antecedents	Consequents	Antecedent support	Consequents support	Supprot	Confidence	Lift
2107	2105	0.137	0.225	0.079	0.578	2.567
2106,2107	2105	0,137	0.225	0.079	0.578	2.567
2106,2105	2107	0.225	0.137	0.079	0.352	2.567
2197	2108	0.145	0.586	0.127	0.874	1.490
2107	2106	0.137	0.996	0.137	1	1.003
2110	2106	0.091	0.996	0.091	1	1.003
2111	2106	0.141	0.996	0.141	1	1.003
2192	2106	0.114	0.996	0.114	1	1.003
2194	2106	0.119	0.996	0.119	1	1.003
2198	2106	0.074	0.996	0.074	1	1.003
4905	2106	0.140	0.996	0.140	1	1.003
6744	2106	0.083	0.996	0.083	1	1.003
7072	2106	0.116	0.996	0.116	1	1.003
2105, 2107	2106	0.079	0.996	0.079	1	1.003
2105, 2108	2106	0.137	0.996	0.137	1	1.003
2107, 2108	2106	0.081	0.996	0.08	1	1.003
2111, 2108	2106	0.094	0.996	0.094	1	1.003
2192, 2108	2106	0.073	0.996	0.073	1	1.003

2194, '2108	2106	0.075	0.996	0.075	1	1.003
4905, 2108	2106	0.097	0.996	0.097	1	1.003
2196	2106	0.111	0.996	0.111	0.996	1.000
2197	2108	0.145	0.586	0.127	0.874	1.490
2106,2197	2108	0.141	0.586	0.123	0.870	1.484
2197	2106, 2108	0.145	0.582	0.123	0.847	1.453
2415	2108	0.149	0.586	0.121	0.812	1.385
2106, 2415	2108	0.149	0.586	0.121	0.812	1.383
2415	2106, 2108	0.149	0.582	0.121	0.808	1.386

3.3.7.3 Association rules using FP growth on third cluster

The FP Gwoth algorithm defined minimum support with values 0.7 and 0.5 does not generate any frequent itemset. Moreover, at minimum support equal to 0.2 generate only a single itemset as shows in the table:

Table 3.43: Generated itemset using FP Growth at minimum support = 0.2

Support	Itemsets
0.266	2105

On the other hand, a frequent itemset were generated when the minimum support value equal to 0.05. The association rules discovered with the higher value of lift that greater than 3 as shows in the following table:

Table 3.44: Discovered association rules using FP Growth at minimum support = 0.05

Antecedents	Consequents	Antecedent support	Consequents support	Supprot	Confidence	Lift
2196	2111	0.124	0.134	0.063	0.511	3,812
2107	2105	0.143	0.266	0.078	0.545	2.04
2194	2105	0.159	0.266	0.057	0.362	1.360

In addition, setting the minimum support to 0.03 provided good association with high value of confidence and lift as the result demonstrated in the following table:

Table 3.45: Discovered association rules using FP Growth at minimum support = 0.03

Antecedents	Consequents	Antecedent support	Consequents support	Supprot	Confidence	Lift
2111, 2196	2112	0.063	0.063	0.033	0.519	8.228
2111,2112	2196	0.042	0.124	0.033	0.783	6.293
2112	2196	0.063	0.124	0.044	0.705	5.669
2196, 2112	2111	0.044	0.134	0.033	0.742	5.533
2112	2111	0.063	0.134	0.042	0.668	4.984

3.3.7.4 Association rules using fp growth on forth cluster

The following table demonstrates the discovered associations rules using FP Growth algorithm, when the minimum support defined equal to 0.6.

Table 3.46: Discovered association rules using FP Growth at minimum support = 0.6

Antecedents	Consequents	Santecedent Support	Consequent Support	Support	Confidence	Lift
2107	2105	0.828	0.879	0.765	0.924	1.051
2105	2109	0.879	0.734	0.659	0.750	1.020
2194	2105	0.780	0.879	0.687	0.880	1.002

2107	2109	0.828	0.734	0.659	0.796	1.084
2107	2110	0.828	0.832	0.695	0.839	1.007
2109	2110	0.734	0.832	0.656	0.892	1.071
6744	2110	0.756	0.832	0.684	0.904	1.086
2194	2192	0.780	0.805	0.677	0.867	1.077
2192	6744	0.805	0.756	0.636	0.790	1.044
2194	6744	0.780	0.756	0.604	0.774	1.022
2105, 2107	2109	0.765	0.734	0.613	0.801	1.090
2107, 2109	2105	0.659	0.879	0.613	0.929	1.057
2105, 2109	2107	0.659	0.828	0.613	0.930	1.123
2105, 2107	2110	0.765	0.832	0.642	0.839	1.008
2107, 2110	2105	0.695	0.879	0.642	0.924	1.051
2105, 2110	2107	0.725	0.828	0.642	0.885	1.069
2192, 2107	2105	0.661	0.879	0.614	0.929	1.057
2192, 2105	2107	0.704	0.828	0.614	0.871	1.052

3.3.7.5 Association rules using fp growth on orginal data

The following table demonstrate the generated item set when minimum support defined equal to 0.2.

Table 3.47: Generated itemset using FP growth at minimum support = 0.2

Support	Itemset
0.367	2415
0.353	4905
0.330	2105
0.329	2106
0.297	2209
0.276	2108
0.270	7072
0.258	4905, 2415

0.236	2107
0.234	2194
0.227	2110
0.218	2192
0.215	2415, 2209
0.207	2106, 2108
0.204	6744

And three strong association rules were discovered as demonstrate in table 3.48:

Table 3.48: Discovered association rules using FP Growth at minimum support = 0.2

Antecedents	Consequents	Antecedent Support	Consequents Support	Supprot	Confidence	Lift
2106	2108	0.276	0.329	0.207	0.7512	2.278
4905	2415	0.353	0.367	0.258	0.7300	1.985
2209	2415	0.297	0.367	0.215	0.7242	1.969

4. FINDINGS

Based on the experimental observations. The result obtained from applying the associating rule mining shows that there are 21 correlations between visited items based on the optimal value defined for minimum support in order with the highest value of confidence and lift together.

The best of it that generated from cluster One and cluster Four. Cluster one a single strong association rule was discovered with minimum support 0.5 and confidence higher than 0.7 as well as Lift higher than 1.

On the other hand, the generated association rules from the Fourth cluster with defined minimum support 0.6, confidence higher than 0.7 and lift higher than 1, and with 18 discovered rules.

In addition, all clusters with low minimum support value equal to 0.2 and lesser provide correlation and it might consider as a strong correlation with more investigating but be but since the minimum support is low the association result with high minimum support value was considered as the best result for this research.

Moreover, as (Zaw Oo 2018) state that, clustered association rule mining generates rules with high minimum support and confidence to the user while non-clustered association rule mining doesn't generate any rules at all, the observation from the implementation of association rule mining on nun clustered dataset does not generate any rules with high minimum support.

In addition, it observed that an association rule was created with the item as antecedent: 2107 and as confidence: 2105 with support: 0.765, confidence: 0.924 and lift: 1.051 were redundant in many clusters.

The following table 4.21, demonstrate the analysis on each cluster with the defined minimum support:

Table 4.1: Comparison of number of generated itemset and discovered associations

Cluster No.	Minimum support	Max Confidence	Max Lift	No. Of association rules	No. Of frequent itemset
1	0.7	0	0	0	4
	0.5	0.821	1.095	1	11
	0.3	0.946	1.127	16	30
	0.23	0.946	2.025	41	34
2	0.7	0	0	0	1
	0.5	0	0	0	3
	0.2	1	1.003	1	5
	0.07	1	2.567	51	53
3	0.7	0	0	0	0
	0.5	0	0	0	0
	0.2	0	0	0	1
	0.05	0.545	3.812	3	20
	0.03	0.783	8.228	14	30
4	0.8	0	0	0	4
	0.6	0.930	1.123	13	28
	0.4	0.974	1.529	834	147
	0.2	0.981	2.314	10502	1150
Full dataset	0.7	0	0	0	0
	0.5	0	0	0	0
	0.3	0	0	0	4
	0.2	0.751	2.278	3	15

Consequently, the result obtained from both algorithm Apririo and FP Growth were exact the same, except the differentiate based on the factor of algorithm running time (Sobia Mehrban 2017 and Pontus Persson 2017), and observed when define the minimum support with value equal 0.01 and lesser.

Moreover, the reason for that time takes in Apriori due to the huge number of candidates that generated while mining the frequent item set. And this is only with total number of items 45.

The following table demonstrate a comparison between the two Algorithm Apriori and FP Growth with low minimum support.

Table 4.2: Comparison between Runtime for Apriori and FP Growth

Cluster No.	Minimum support	Runtime of Algorithm in Sec.	
		Apriori	FP Growth
1	0.01	1.518	1.170
	0.001	10.125	1.679
2	0.01	1.240	1.062
	0.001	4.835	1.169
3	0.01	1.131	1.068
	0.001	1.872	1.098
4	0.01 (Memory Error)	Unable to allocate 4.77 GiB	10.762
	0.001 (Memory Error)	Unable to allocate 13.3 GiB	261.537
Full dataset	0.01 (Memory Error)	Unable to allocate 5.31 GiB	2.796
	0.001 (Memory Error)	Unable to allocate 8.06 GiB	10.905

5. CONCLUSION

Web usage mining is a very important technique for the business from different aspects, especially where the process of extracting knowledge from the clients' behavior through their transactions will invest this knowledge in the business development process.

The conducted result is beneficial for any companies as a service provider to enhance the product and to serve their clients especially marketing-wise, while the obtained result and based on the expectations of the client will provide better insight for the marketing department in designing the retargeting campaign for the existing clients.

The data was provided from a company called Mbsher and they are a solution provider of solution for franchise and hospitality organizations. The data collected from a solution as an online menu and each record present a visit transaction for a single item while when a single client visit multi-items every single visit collected as a single record in the database.

And the menu pages divided into two-level categories and under them are the items. The research focuses on the level of item visits.

Mbsher provides the full database files, the workflow starts with understanding the business model for Mbsher as a service provider and Chief Burger the tenant who provides the online menu as a digital solution for their clients. Then, there is the process of extracting the desired data from the provided database taking into consideration all the information provided by the company.

The methodology to run the web usage mining technique consists of several processes starting with data pre-processing that include the data cleaning, user and session identification, and the most important setting the dataset in a suitable format for the analysis algorithm. The next step was grouping the transactions based on their similarity through

clustering with K means algorithm in addition to using two evaluation method that supports the process of choosing the optimal number of clusters that are Elbow and Silhouette coefficient since it is required to define the number of clusters when building the model using K means algorithm.

After the data was grouped, it is split into four clusters, and the association rules mining technique was implemented and for generating the frequent itemset, the Apriori and FP Growth algorithm were used. The generated frequent itemset with higher minimum support was obtained on cluster one with minimum support 0.5 and with a single association rule, while the best result was on the Forth cluster with minimum support 0.6 and 18 association rule with confidence higher than 0.7 and lift higher than 1 as well. Cluster numbers two and three provide association as well but with lower defined minimum support comparing with the other clusters.

Moreover, both algorithm running time were almost the same except when define the minimum support with low value under 0.05.

One of the process to measure the benefit of the model is design specific campaign based on the result that obtained from this research. In addition, include more characteristics as an attribute will enhance the productivity of the clustering model, such as the language of the devices, gender, the operation system of the user device, and many more.

Some of these characteristics already exist in the database but due to an upgrade process since it is recently added as informed from the company there were a huge number of null value and for that reason this research focus only on the transaction for the visited items with excluding the other characteristics in order to keep the number of samples good for the analysis process.

Moreover, it is very important to study the design of the database for more information that could be helpful in the future for analysis. A major feature is including a value define the

user who paid directly through the payment feature through the app as it could consider a good indicator for the permanent clients for example.

In addition, having the name or titles of the items will provide many clear insights in addition to better understand based on what the correlation was defined since it's very vague using only the items ID without any additional info about the item.

REFERENCES

Books

Charu C. & Aggarwal Jiawei Han, 2014. *Frequent Pattern Mining*.

Jake VanderPlas, 2016. *Python Data Science Handbook*

Michael Steinbach, Pang-Ning Tan, & Vipin Kumar, 2005. *Introduction to Data Mining. First Ed*.

Periodicals

Amit Dipchandji Kasliwal & Girish S. Katkar, 2015. *Web Usage mining for Predicting User Access Behaviour*.

Anmol Kaur, 2017. *Analysis of Web Usage Mining Techniques to Predict the User Behavior from Web Server Log Files*.

Asiya N. Khan & Pallavi S. Pandhare, 2018. *Web Usage Mining and User Behaviour Prediction*.

Erik Ranby. 2016. *A comparison of clustering techniques for short social text messages.*

Guntur Budi Herwantoa & Annisa Maulida Ningtyas, 2018. *Recommendation System for Web Article Based on Association Rules and Topic Modelling*.

Han Ni Ni Myint Thu and Khine Khine Oo, 2019. *Discovering Generalized Association Rule in Web Usage Mining by FP Tree.*

Htun Zaw Oo, 2018. *Pattern Discovery Using Association Rule Mining on Clustered Data*.

Jayanti Mehra, 2018. *An Effective method for Web Log Preprocessing and Page Access*.

J. Han, J. Pei, and Y. Yin, 2000. *Mining frequent patterns without candidate generation*.

Ming-Yi Shih & Syun-Sian Huang, 2015. *Characterizing Web Users Based on Their Required Criteria*.

Mrunmayee R. Joshi, 2019. *Enhancing Prediction of User Behavior on the basic of Web Logs*.

Muhamad Brilliant, DwiHandoko & Sriyanto, 2017. *Implementation of Data Mining Using Association Rules for Transactional Data Analysis*.

Priyanka Sharma, Sumayya Khan, Shilpa Singh & Pooja Tiwari, 2015. *An Analysis on Web Usage Mining for Internet Users*.

Pontus Persson, 2017. *Identifying Early Usage Patterns that Increase Retention Rates in a Mobile Web Browser*.

Sangavi. S, Suvetha. T & Umashankari T, 2016. *Web usage mining using Improved KNN Algorithm.*

Serin. J, 2018. *Clustering based Association Rule Mining to Discover User Behavioural Pattern in Web Log Mining*.

Sobia Mehrban, 2017. *Web Usage Minning Using Patterns with Different Algorithms*.

Other Publications

Ajitesh Kumar, 2020. *KMeans Silhouette Score Explained With Python Example* on: https://dzone.com/articles/kmeans-silhouette-score-explained-with-python-exam.

Chinedu Pascal Ezenkwu, Simeon Ozuomba & Constance Kalu 2015. *Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services* on: https://www.researchgate.net/publication/282862569_Application_of_K-Means_Algorithm_for_Efficient_Customer_Segmentation_A_Strategy_for_Targeted_Customer_Services.

Dr. Michael J. Garbade, 2018. *Understanding K-means Clustering in Machine Learning* on: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1.

Gerald Britton, 2017. *Top 10 questions and answers about SQL Server Indexes* on: https://www.sqlshack.com/top-10-questions-answers-sql-server-indexes/

Hafeezul Kareem Shaik, 2019. *Calculate Time taken by a Program to Execute in Python* on: https://www.studytonight.com/post/calculate-time-taken-by-a-program-to-execute-in-python

Jiawei Han. *Pattern Discovery in Data Mining* on: https://www.coursera.org/learn/data-patterns.

Jiawei Han. *Pattern Discovery in Data Mining* on: https://www.coursera.org/learn/data-patterns.

Kevin Arvai 2020. *K-Means Clustering in Python: A Practical Guide* on: https://realpython.com/k-means-clustering-python.

Kovid Rathee 2020. *Indexing Very Large Tables* on: https://towardsdatascience.com/indexing-very-large-tables-569811421ee0

Matthew Mayo 2020. *Centroid Initialization Methods for k-means Clustering* on: https://www.kdnuggets.com/2020/06/centroid-initialization-k-means-clustering.html.

Mikhail Sidyakov. *Market Basket Analysis Using Association Rule Mining in Python* on: https://pyshark.com/market-basket-analysis-using-association-rule-mining-in-python/.

Mukesh Chaudhary, 2020. *K-means Clustering in machine learning* on: https://medium.com/@cmukesh8688/k-means-clustering-in-machine-learning-252130c85e23.

Mukesh Chaudhary, 2020. *Silhouette Analysis in K-means Clustering* on: https://medium.com/@cmukesh8688/silhouette-analysis-in-k-means-clustering-cefa9a7ad111

S Joel Franklin, 2019. Elbow method of K-means clustering using Python on: https://medium.com/analytics-vidhya/elbow-method-of-k-means-clustering-algorithm-a0c916adc540.

Spyder 5.0 integrated development environment, 2021. [Computer software]. open-source cross-platform, Inc, https://www.spyder-ide.org/.

XAMPP 8.0.6 web server solution, 2021. [Computer software]. Open Source, Inc. https://www.apachefriends.org/