International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 11 (2020), pp. 3444-3455
© International Research Publication House. https://dx.doi.org/10.37624/IJERT/13.11.2020.3444-3455

Using Big Data Analytics to Design an Intelligent Market Basket-Case Study at Sameh Mall

Farah Almaslamani[1], Raneem Abuhussein[1], Hanan Saleet[1,*], Laith AbuHilal[3], Nader Santarisi[1]

[1] Industrial Engineering Department, Applied Science Private University, Jordan.

[2] Chief Executive Officer, Sameh Mall, Jordan. *[3]General Manager, Sameh Mall, Jordan.*

**Corresponding author*

**ORCID: 0000-0003-3780-0231*

Abstract

The long term social, economic and health impacts of the COVID-19 pandemic are still unknown. Retailers should think about the impact this pandemic will have on the customer relationship. Another factor that is rigorously influencing the retail industry is the digital transformation. With the digital transformation worldwide, coupled with the exponential growth of the use of big data analytics, retailers can use intelligent market basket analysis to help in shoring up customer relationships. This study uses big data analytics to design and analyze intelligent market basket in one top retailer in Jordan, "Sameh Mall". It aims to help managers to improve customer relationship while increasing sales. Customers' behavioral similarities analysis results in different baskets, which contain items commonly bought together. Such baskets are displayed physically in stores and are displayed online as promotions. This study results are interesting, enabling Sameh Mall to send recommendations to VIP customers through their account on the online application; and recommendations for physical cross promotional or cross merchandising leading to increases in basket size, increase in sales, as well as increase in customer satisfaction.

Keywords: Intelligent market basket analysis, customized basket, general basket, smart sales and marketing system.

I. INTRODUCTION

"People like buying in store, even millennials and Gen Z who have all the apps on their phones, still like to shop in stores" [1]. This is true even though the long term social, economic and health impacts of the COVID-19 pandemic are still unknown; still people are willing to go shopping in the physical stores. Nevertheless, retailers should think about the impact this pandemic will have on the customer and the customer relationship [2]. The coronavirus pandemic is forcing the retail industry to quickly innovate in a race that's likely to challenge customer relationships. Because of this pandemic, many changes has emerged such as shift in the buying process, changes in the customer experience, and new marketing approaches were adapting [3]. Utilizing intelligent systems, retailers can learn a lot from how people shop and react to offers on store shelves [4]; consequently, these smart systems can help retailers in shoring up the customer relationship.

According to Grand View Research, developing intelligent systems had a market value of $272 million during 2016 which is expected to hit $10.2 billion in 2025 [5]; mainly because of the growth of big data analytics. Big data analytics is flourishing because of the following: the increase in data storage capacity, accurate and fast computing power, and the ability to stream and manage hundreds of billions of complex datasets [6]. Retail stores are continually improving their operations through the adoption and application of such intelligent systems which utilizes big data analytics [7] [8] [9].

This study aims to shoring up the customer relationship by designing and analyzing an intelligent system. It utilizes big data analytics to study the market basket analysis for one of the top retail series in Jordan, called "Sameh Mall".

The retail industry is one the largest sectors as it provides the vast majority of required items in one stop shop. Competition among retailers led to intensive utilization of digitization, which led to complete transformation of this industry. One important tool that must be used by retailers is market basket analysis. Market basket analysis studies purchasing patterns by grouping items which are frequently bought together. In other word a customer who buy a certain item is more likely to buy another item such as coffee and chocolate. Some of these correlations (similarities) can be easily figured out. Nevertheless, usually there are huge number of items in retail stores with hidden correlations; one way to discover this correlation is using market basket analysis.

Market basket analysis helps managers in understanding their customers and improving their experience [10] [11]. Marketers then can infer the profiles of customers in each group and propose management strategies appropriate to each group [12] [13]; it is known that retaining customers makes good business sense and costs less than attracting new ones. According to a report by Content Marking Institute, 90% of top-performing B2B content marketers prioritize audiences' information needs. Integrated Market Basket Analysis in B2B help in optimization of campaigns and promotions, increase sales and optimize ROI, assist in the optimization of in-store operations, increase in the market share and help in analyzing customers' behaviors [14].

Different types of models can be constructed for big data analytics including classification, regression, association and clustering models [15] - [21]. Big data analytics can also be used as a tool to provide the retailers with strategies to

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 11 (2020), pp. 3444-3455
© International Research Publication House. https://dx.doi.org/10.37624/IJERT/13.11.2020.3444-3455

effectively enhance direct marketing and therefore affecting customer behavior. Big data analytics is used to analyze market basket. It is defined as a data mining tool used to extract important information from existing data and enable better decision making throughout an organization. It identifies the correlation between the items in large databases; it examines the buying habits of the customers by identifying the associations among the items purchased by the customers in their baskets.

Many retailers use the different data analytics techniques to achieve their business goals. Among the most well-known of them is Wal-Mart, which uses Market Basket Analysis to understand the specific purchasing patterns of their customers, developing directed targeting of customers for new products, at a fraction of the cost. In addition to Master Card and Burger King that use data mining techniques in combination with the available technology to better understand their customers [22].

Prasad and Mourya [23] provided lot of case studies about the Association rules and the existing data analytics algorithms usage for market basket analysis. In [24], Uninorms were used as an alternative measure to aggregate support and confidence in analyzing market basket data using the UK grocery retail sector as a case study. Hemalatha [25] used market basket analysis in Indian retail industries.

Zuo, et al. proposed consumer purchasing behavior method by utilizing RFID data acquired from individuals in a Japanese supermarket [26]. They provided a time perspective on shopping in a certain area instead of the entire grocery store by conducting a multivariate normality test on the data to find an optimal solution. Yao , in his research concluded how to set down the customer-oriented business strategy in supermarket from the perspective of price, promotion, product, customer service and human resource management and finishes with suggestions on how to improve operations within the supermarket [27]. Yu concluded that store environment will also influence customer cognitive and emotional evaluations in a positive way [28]. Tayyar studied the cereal market. Weekly scanner data from 13 stores of Safeway and 16 stores of Asda for 22 competing products at individual stores levels were analyzed using multilevel modeling methods to provide a complete understanding of this market [29]. Laio et al. used the Apriori algorithm of association rules, and clustering analysis based on an ontology-based data mining approach, for mining customer knowledge from a database. Knowledge extracted from data mining results is illustrated as knowledge patterns, rules, and maps in order to propose suggestions and solutions to Taiwan Adidas for possible product promotion and sport marketing [30]. Bhargav et al. highlighted that there is a problem related to Apriori algorithm; it requires a repeated scan of the entire database of customer transactions to find candidate sets and to find frequent item- sets. Their work focused on the use of artificial neural network technique to overcome these problems. They proposed a single layer feed-forward partially connected neural network technique, which reduces the time taken in repeated scanning of the database and also increases the efficiency of the algorithm [31]. Dorismond investigated how a retail store layout can influence customers' decisions using historical purchase data in order to understand the customers shopping patterns, and identifying customers' plan and unplanned purchases, and then how to build a data driven model that finds the optimal layout of traditional retail store, by increasing the visibility of "impulse items" and how to optimize the product placement of items in promotional areas [32].

This study uses big data analytics to analyze and design the intelligent market basket for one of the top retailer in Jordan, Sameh Mall. The results guide the managers to adjust their sales and marketing strategy by designing customized and general market baskets.

II. INTELLIGENT SYSTEM DESIGN

II.I Sameh Mall

Sameh Mall is one of the leading retailers in Jordan. It started as a commercial investment group that started in 2001 as retail outlets and shopping centers. It widely expanded in 16 years bringing its branches up to 30 branches and hypermarkets all around the country. Sameh Mall adopts an integration strategy to provide all the customer needs in one place by offering diversity of food products and household items. It provides a wide variety of items for the end customer such as: groceries, fresh section, dairy products, appetizers, beverages, personal care, electronics, home care and frozen food. Sameh Mall also provide customers with an online shopping experience via Sameh Mall application; show in Figure 1 .

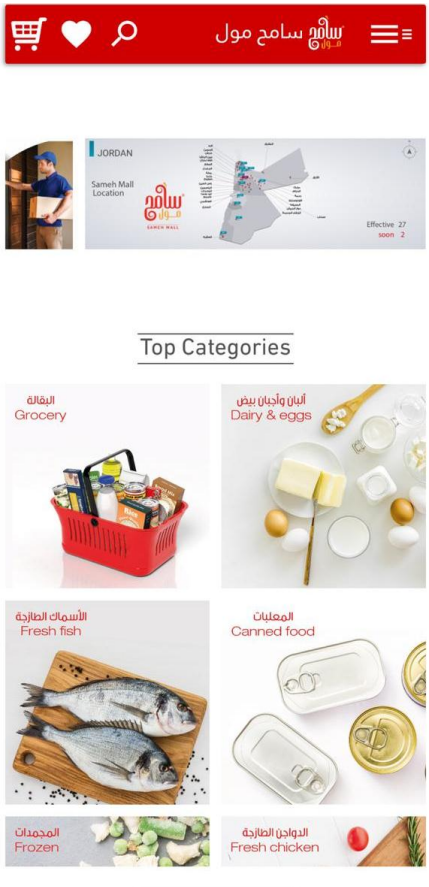

Figure 1. Sameh Mall mobile application

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 11 (2020), pp. 3444-3455
© International Research Publication House. https://dx.doi.org/10.37624/IJERT/13.11.2020.3444-3455

Sameh Mall is growing their business rapidly in respond to strong competition and seeking for a greater market share in Jordan. They are looking for using big data analytics to help management take important decisions. For managers at Sameh Mall, it is critical to know if two customers have similar shopping behavior; so that they may recommend new items to those customers. In addition, knowing the items that are usually bought together is important for both improving layout design and for deciding on the discounts on certain items; cross promotions and cross merchandizing. Figure 2 shows layout design for the physical store of Sameh Mall.

Figure 2. Current general baskets at Sameh Mall

II.I System Description

When someone goes shopping, he/she usually buys a number of items. Those items which are bought together are called a market basket. This study aims to explore customers' behavior such that those customers who have similar baskets are said to have similar shopping behavior. In addition, the items which are usually bought together within one basket are said to form a general basket. In other words, Market basket analysis is a technique to study the patterns of customers' purchasing (buying habits of the customers) by identifying the similarities (correlation) among the items purchased by the customers in their baskets. Market basket analysis helps managers in understanding their customers and improving their experience.

When identifying general baskets, the following can be achieved:

- The layout design of the physical store can be revised to allocated items, which are included in one general basket, close to each other. Thus, customers will be happy to find the items that are usually bought together in the same location; especially under the current COVID-19 pandemic where customers strive to shorten their stay time inside the stores.
- Placing items close to each other helps promoting items that are considered new to the market. For example, locating a new brand of cereal close to milk will bring attention of customers to this new brand. This is true since people tend to buy cereal and milk together; which results in promoting and increasing the sales of cereal.
- In addition, placing items that are usually bought together close to each other would remind the customers about the items that they need to buy.
- To increase the sales of items, retailers place discounts on certain items. Using big data analytics will discover correlated items that are included in one general baskets. Thus instead of placing discounts on all correlated items, it is enough to place discount on some items; called cross merchandizing. This is true because customers usually buy those items together; coffee and chocolate or milk and cereal. Using this method increases sales while saving the retail store money that would be lost if discounts was placed on many items within one general basket.
- For online shopping, Sameh Mall has website and it has recently developed an application. Those general baskets can also be displayed online.

As mentioned earlier studying customers' behavior results in identifying customized baskets; the following can be achieved.

- If a customer usually buys a certain item every two weeks for example, a notification will be sent to that customer through his/her account on the online application every two weeks to remind him/her to buy this item.
- Customers that has similar shopping behavior, have nearly similar baskets. Sameh Mall sends those customers recommendations to buy new items. Those new items are the items found in one customer's basket but not the others. Thus increasing the size of the customized baskets.
- The results of market basket analysis are totally dependent on seasons and time and it needs to be performed repeatedly. This would promote the idea that Sameh Mall takes care of its customers by being close to them all the time; which would increase customers' satisfaction.

III. INTELLIGENT MARKET BASKET ANALYSIS

This study adopts a process that involves the following main decision steps for analyzing the market basket.

1-Retrieve data from the ERP system: studying customers' behavior needs data about items that a customer buys. This

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 11 (2020), pp. 3444-3455
© International Research Publication House. https://dx.doi.org/10.37624/IJERT/13.11.2020.3444-3455

information is found in invoices. Invoices are stored in the ERP system for Sameh Mall. The accounting module has the invoices which has rich data about the different transactions such as:

- COMP_ID: company identification.
- BRANCH_ID: Branch identification.
- BRANCH_DESC_A: branch description (branch location).
- RECEIPT_NO: receipt number.
- PHONE: Customers phone number.
- TERMINAL_NO: Terminal number (cash number).
- TRANS_DATE: transaction date.
- TRANS_TIME: Transaction time.
- ITM_NO: Item number.
- ITM_EQUIVELENT_QTY: Quantity of equivalent item.
- PAYMENT: Payment method (cash, visa).
- ITM_PART: Sub sub sub category number.
- ITM_A_DESC: Sub sub sub category description.
- SUPDEPT_NO: General category number.
- DESC_A: General category description.
- ITM_SUB_CAT2: Sub category number.
- SUB2_DESC_A: Sub category description.
- ITM_SUB_CAT3: Sub sub category number.
- SUB3_DESC_A: Sub sub category description.
- ITM_GROUP_CODE: Items are divided into groups such as beauty, biscuits and detergents.
- QTY: Defined quantity for the pack (-1 means that this item is returned).
- AMT: Payment amount.

Table 1. Dataset "dataset.csv" contains data from the invoices taken from the ERP system in Sameh Mall.

InvoiceNo	StockCode	Quantity	InvoiceDate	UnitPrice	CustomerID
9984	136027	1	191020	1.1	785455143
9984	238705	1	191020	0.75	785455143
9984	238705	1	191020	0.75	785455143
9984	285897	3	191020	2.37	785455143
9984	285878	3	191020	1.17	785455143
9984	269461	1	191020	1.89	785455143
9984	206165	1	191020	1.99	785455143
9984	206165	1	191020	1.99	785455143
9984	267569	1	191020	1.35	785455143
9981	135621	0.659	191115	0.389	785758343
9981	25618	0.43	191115	0.641	785758343
9981	239865	0.515	191115	2.57	785758343
9981	269255	1	191115	1.48	785758343
9981	269256	1	191115	1.48	785758343
9981	285887	10	191115	3.9	785758343
9981	270101	1	191115	1.49	785758343
9981	25672	0.951	191115	0.941	785758343
9981	30422	3.885	191020	27.156	785758343
9981	135656	0.75	191115	1.493	785758343
9981	230479	0.465	191115	0.46	785758343
9981	135580	1.62	191115	1.426	785758343
9981	25607	1.2	191115	0.468	785758343
9981	269253	1	191115	1.89	785758343
9981	269253	1	191115	1.89	785758343
9981	269253	-1	191115	-1.89	785758343
9981	124308	1	191115	0.15	785758343

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 11 (2020), pp. 3444-3455
© International Research Publication House. https://dx.doi.org/10.37624/IJERT/13.11.2020.3444-3455

The main data needed for this study is displayed in Table 1, which contains part of the data found in invoices. Table 1 displays a sample of the dataset used in this study.

2-Similarity indices evaluation: using data in Table 1, similarity indexes can be calculated. The similarity indexes represent the correlation between two customers or correlation between two items. Several methods have been proposed to calculate similarity indexes. The proposed solution uses cosine similarity indexes; which are calculated for each pair of customers and for each pair of items.

3-Based on the analysis of the similarity indexes, both sales and marketing departments at Sameh Mall can build important strategic decisions.

4-Location assignment: based on the cosine similarity between items, decisions about relocation of items can be taken.

The four decision steps mentioned above belong to a hierarchical process that involves several steps. the results of each step depends on the quality of the input data. For example, how effective the decision making process is depends on the quality and significance of the similarity index adopted, whose values must be evaluated correctly.

III.I Big Data Analytics

This section introduces the big data analytics mechanism used to deal with the market basket analysis problem. The data analytic mechanism used in this study involves the following four main stages; shown in Figure 3.

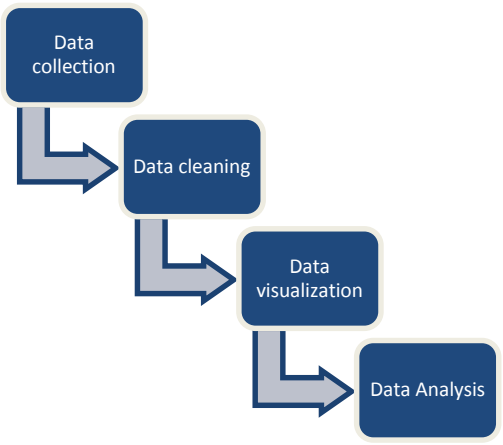

Figure 3. General steps in big data analytics

III.I.I Data collection

The invoices contain the items that has be bought by a certain customer. As depicted in Table 1, the dataset contains a unique ID for each invoice (InvoiceNo); each invoice includes a basket that contains certain items (StockCode). In addition, the dataset contains a unique ID for each customer (CustomerID)

-Only the customers who are VIP are taken into consideration since they always buy from Sameh Mall

-The Customer identification number is his/her unique phone number; it is saved in the ERP system. But for the confidentiality purposes, the ID is coded.

III.I.II Data cleaning

-Table 1 which contains the raw data has some negative numbers. These numbers are errors in data entry or they are some bugs in the system; thus they need to be eliminated from the dataset. In addition, some data points are not available NaN; those are also eliminated.

-If an invoice includes an item that is (are) repeated more than once, only just one record will be kept. In this study, the similarity or correlation between items is considered regardless of their quantity.

III.I.III Data visualization

-Data about CustomerID and StockCode that is listed in Table 1 is transform into a matrix *inm,* equation (1), that contains CustomerID in the rows and StockCode in the columns. The entries for this matrix are 0 or 1 depending on the fact that the customer bough this item or not.

$$inm = \begin{bmatrix} a_{11}\ a_{12}\ \dots\ a_{1n} \\ a_{21}\ a_{22}\ \dots\ a_{1n} \\ \vdots\ \ \vdots\ \ \ddots\ \ \vdots \\ a_{m1}\ a_{m2}\ a_{m3}\ \dots\ a_{mn} \end{bmatrix} \dots (1)$$

Where m is total number of customers in this dataset, and n is total number of stock items

III.I.IV Data analysis

Big data analytics includes calculating cosine similarity. Cosine similarity presents the Euclidean L2-normalized dot product of vectors. For customer1 and customer2 their cosine similarity is defined as:

$$\text{customer1} = [a_{11}\ a_{12}\ \cdots a_{1n}\ ] \dots (2)$$

$$\text{customer2} = [a_{21}\ a_{22}\ \cdots a_{2n}\ ] \dots (3)$$

$$cosSim = \frac{\sum_{i=1}^{n} a_{1n}\ a_{2n}}{\sqrt{\sum_{i=1}^{n} a_{1n}^2}\ \sqrt{\sum_{i=1}^{n} a_{2n}^2}} \dots (4)$$

Details about big data analytics is presented in the next section.

IV. SOLUTION ILLUSTRATION

This section presents big data analytics of the market basket using an illustrative example.

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 11 (2020), pp. 3444-3455
© International Research Publication House. https://dx.doi.org/10.37624/IJERT/13.11.2020.3444-3455

As mentioned above data is summarized into a matrix called *inm*; as shown in Table 2

CustomerIDs are located in rows: 1,2,3,4,5. Stock codes are in the columns: 201, 202, 203 and 204.

Table 2. Illustrative example

	StockCode			
CustomerID	201	202	203	204
1	1	0	1	1
2	1	1	1	0
3	0	0	1	0
4	1	0	1	1
5	0	1	0	1

In vector form

Customer1 = [1 0 1 1]

Customer2 = [1 1 1 0]

Recall the similarity index formula in equation (4), the following calculations are needed:

$$\sum_{i=1}^{n} a_{1i}\, a_{2i} = 1 * 1 + 0 * 1 + 1 * 1 + 1 * 0 = 2$$

$$\sqrt{\sum_{i=1}^{n} a_i^2} = \sqrt{1^2 + 0^2 + 1^2 + 1^2} = 3$$

$$\sqrt{\sum_{i=1}^{n} b_i^2} = \sqrt{1^2 + 1^2 + 1^2 + 0^2} = 3$$

$$cosSim = \frac{2}{3 * 3} = 0.6667$$

The output is a square matrix *m x m,* shown in Table 3 that displays the similarity between each pair of customers.
The following observations can be concluded from Table 3:

- The numbers in the diagonal are all 1's since the similarity between one customer and him/her self is 100%
- Customer1 and customer4 baskets are the same, thus their similarity index is 1
- Customer3 and customer5 has totally different basket, thus their similarity is 0.

Table 3. Pairwise cosine similarity indexes between two customers

	CustomerID				
CustomerID	1	2	3	4	5
1	1	0.666667	0.57735	1	0.408248
2	0.666667	1	0.57735	0.666667	0.408248
3	0.57735	0.57735	1	0.57735	0
4	1	0.666667	0.57735	1	0.408248
5	0.408248	0.408248	0	0.408248	1

Customers who has nearly similar baskets as customer1, for example, can be found by sorting column one in Table 3

Table 4. Customers that have baskets nearly similar to customer1 basket

CustomerID	Similarity index
Customer4	1
Customer2	0.666667
Customer3	0.57735
Customer5	0.408248

The following conclusions can be drawn form Table 4

1- Since customer1 and customer4 has same baskets, their shopping behavior will be monitored so that any changes that may create differences in their baskets in the future can be recorded. This will help the marketing department to send promotions to these customers.

2- Customer2 basket is nearly 66.67% similar to customer1's basket. When investigating the baskets of both customers, the different items in the baskets can be found. This can help the marketing department to recommend new items for both customers

Table 5. Mapping of the stock code for the items in the basket for customer1 and customer2

Stock Code	201	202	203	204
Stock description	Citrus detergent, 1 Liter	Indomie fried noodles	Matrix orange juice 200ml	Orange type 1
Customer1 basket	1	0	1	1
Customer2 basket	1	1	1	0

3- Table 5 shows the mapping of the stock code for the items in the basket for customer1 and customer2. The items that are not found in customer1 basket but are found in customer2 basket which is "indomie fried noodles" can be recommended

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 11 (2020), pp. 3444-3455
© International Research Publication House. https://dx.doi.org/10.37624/IJERT/13.11.2020.3444-3455

to customer1. Likewise, orange type 1 can be recommended to customer2

The information about items that are usually bought together within the same invoice can be used to redesign the layout of the physical store. This will save time for customers who usually need to pick items, pay for them, and leave as soon as possible; to minimize the chance of coming in contact with many people and then being infected by COVID-19. This is one face of the coin. The other face, especially for the highly correlated items, some items with high demand can be located far apart. Locating them far apart enforces the customer to walk through the aisles looking for the other item.

To continue with the previous example, the cosine similarity indexes can be calculated between stock items as shown in Table 6

Table 6. Pairwise cosine similarity indexes between two items

StockCode	201	202	203	204
201	1	0.408248	0.866025	0.666667
202	0.408248	1	0.353553	0.408248
203	0.866025	0.353553	1	0.57735
204	0.666667	0.408248	0.57735	1

Sorting column 1 in Table 7 , results in emphasizing that there is a strong correlation between SockCode 201 and SockCode 203. Thus they need to be included within the same general basket, and then they need to be located close to each other in the physical stores. The management need to decide how close those items to each other since StockCode 201 is a detergent and StockCode 203 is a canned food. In addition, they need to be displayed in one general basket when promoted on the online application.

Table 7. Items that can be included in one general basket with StockCode 201

StockCode	Item description	Similarity index
203	Matrix orange juice 200ml	0.866025
204	Orange type 1	0.666667
202	Indomie fried noodles	0.408248

V. CASE STUDY: RESULTS AND DISCUSSION

Every business model is unique in the sales and marketing strategy. Sameh Mall cooperated by sharing the data about their market baskets and allowing thorough discussions with different resources within the company including IT department, sales and marketing department, and logistics department. The steps followed to apply big data analytics in this study is displayed in Figure 4. The steps are summarized in the following sections.

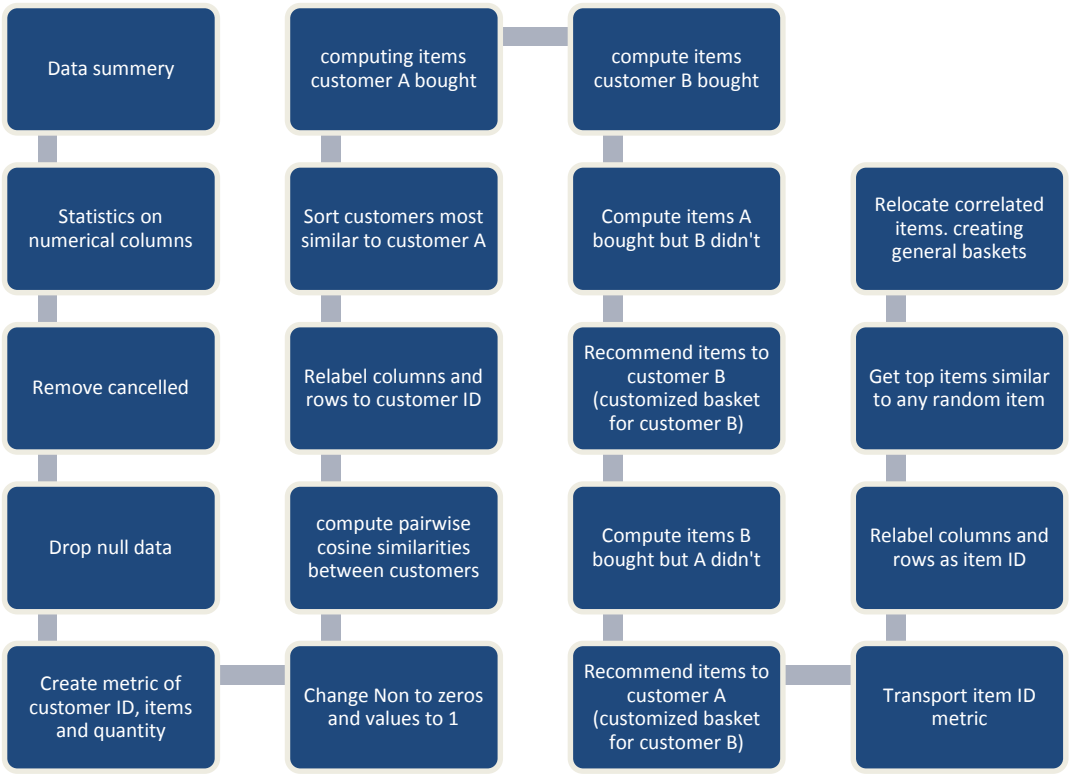

Figure 4. Steps followed in this study to apply big data analytics to study the intelligent market basket in Sameh Mall

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 11 (2020), pp. 3444-3455
© International Research Publication House. https://dx.doi.org/10.37624/IJERT/13.11.2020.3444-3455

V.I Generating customized market baskets

For the purpose of this study, the dataset is built using the information taken from the invoices that are saved in the accounting module within the ERP system in Sameh Mall. The dataset "dataset.csv" contains 135,290 record. Table 1 shows part of the dataset. All personal information such as customers' phone number were coded. The table shows the following features: invoices No, stock codes, quantity, unit price and customer ID (coded phone number).

The market basket is analyzed using Python **[*33*]**; a high level programming language that hosts many third party modules. The dataset is imported as a dataframe using pandas module.

```
import pandas as pd
df = pd.read_csv("C:\\dataset.csv", encoding = 'ISO-8859-1')
```

Collected data in the dataset were statistically analyzed mainly to study the different statistics such as the maximum, minimum and average values; shown in Table 8

```
# Statistics on the numeric columns
df.describe()
```

Table 8. statistics about the different values contained in the dataset

	InvoiceNo	StockCode	Quantity	InvoiceDate	UnitPrice	CustomerID
count	135290.000000	135290.000000	135290.000000	135290.000000	135290.000000	1.352900e+05
mean	5174.875992	159749.969517	1.162702	191059.977278	1.606814	7.919523e+08
std	2778.942283	105273.972195	1.220778	44.844555	3.791096	2.991504e+07
min	1.000000	4.000000	-44.000000	190928.000000	-111.000000	6.569339e+07
25%	2754.000000	25625.000000	1.000000	191020.000000	0.396000	7.906183e+08
50%	5346.000000	172811.000000	1.000000	191028.000000	0.950000	7.957809e+08
75%	7465.000000	261106.000000	1.000000	191107.000000	1.890000	7.975700e+08
max	9998.000000	289226.000000	120.000000	191118.000000	441.600000	8.000000e+08

Table 1 and Table 8 shows that there are some data points which are considered to be outliers; which may cause shifting in the results. And then they need to be eliminated from the dataset before starting the market basket analysis. Some of these data points are negative values for "Quantity" and "UnitPrice", such records represent items that are returned or cancelled; and then they need to be eliminated. In addition, some records have empty data points. Those records need to be eliminated as well. If an invoice includes an item that is (are) repeated more than once, only just one record will be kept. In this study, the similarity between items is considered regardless of their quantity.

```
# Removing cancelled orders (shown as negative values in Quantity or unit price)
df = df.loc[df['Quantity'] > 0]
df = df.loc[df['UnitPrice'] > 0]
df.loc[df['CustomerID'].isna()].head()
df = df.dropna(subset=['CustomerID'])
```

```
# Number of records and shape after dropping missing values
df.shape
```

Out:(131118, 8)

The number of records has dropped to 131118 after removing the negative and/or missing values.

Create a matrix that contains the customer IDs as the index, and each stock code as a column. Use the pivot function to use the CustomerID as the index and use the StockCode as columns .

```
customer_item_matrix = df.pivot_table(index='CustomerID', columns='StockCode', values='Quantity',aggfunc='sum')
```

As mentioned earlier, the quantity of each item bought by a customer in each invoice is not important. Thus, if the invoice contains a certain item (the shopping basket contains an item), the value will be 1; otherwise, it will be 0. As show in Table 9

```
customer_item_matrix = customer_item_matrix.applymap(lambda x: 1 if x > 0 else 0)
```

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 11 (2020), pp. 3444-3455
© International Research Publication House. https://dx.doi.org/10.37624/IJERT/13.11.2020.3444-3455

customer_item_matrix.head()# displays the first five records

Table 9. customerID in rows and stock code in columns

StockCode	4	178	200	201	214	225	226	227	231	232	...
CustomerID											
65693391	0	0	0	0	0	0	0	0	0	0	...
78870697	0	0	0	0	0	0	0	0	0	0	...
79055317	0	0	0	0	0	0	0	0	0	0	...
79333402	0	0	0	0	0	0	0	0	0	0	...
111111111	0	0	0	0	0	0	0	0	0	0	...

The number of unique customerID in the dataset is 5624. Since these customrs' phone number is saved in the ERP system of Sameh Mall, they are considered VIP customers. The number of unique stock items that are included in this dataset is 8940. As show in the following output:

customer_item_matrix.shape

Out:(5624, 8940)

Cosine similarity index is calculated to figure out customers that have similar baskets. Thus, the cosine similarity function from sklearn is used to compute the pairwise index. Rows and columns were relabeled to customer ID to observe relationship between customers as shown in Table 10 .

#from sklearn.metrics.pairwise import cosine_similarity

user_user_sim_matrix = pd.DataFrame(cosine_similarity(customer_item_matrix))

user_user_sim_matrix

The matrix should be square. Its size equals the number of customers (5624); use python to check the shape of the matrix

user_user_sim_matrix.shape

Out:(5624, 5624)

Table 10. Cosine similarity indexes between two customers

	0	1	2	3	4	5	6	7	8	9	...
0	1.000000	0.0	0.0	0.0	0.0	0.106600	0.166667	0.0	0.000000	0.0	...
1	0.000000	1.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.000000	0.0	...
2	0.000000	0.0	1.0	0.0	0.0	0.044455	0.000000	0.0	0.000000	0.0	...
3	0.000000	0.0	0.0	1.0	0.0	0.019462	0.000000	0.0	0.000000	0.0	...
4	0.000000	0.0	0.0	0.0	1.0	0.000000	0.000000	0.0	0.000000	0.0	...
...	...	...	...	...	...	...	...	...	...	...	...

The sales and marketing department at Sameh Mall uses the information in Table 10 to build the customized baskets for the VIP customers. First, the information about most similar pair of customers are extracted. Using python, the following code is used to display the ranking of customers that are similar to customer with ID: 65693391

*user_user_sim_matrix.loc[65693391].sort_values(ascending =**False**)*

Out:

CustomerID

65693391 1.000000

798104708 0.471405

797921588 0.408248

777952682 0.384900

770779775 0.384900

...

796233233 0.000000

796232125 0.000000

796230951 0.000000

796230939 0.000000

795736819 0.000000

Name: 65693391, Length: 5624, dtype: float64

The customer that have the most similar shopping behavior to the customerID 65693391 is the customer with ID 798104708. Thus the sales and marketing department looks in the baskets of the two customers and recommend to them to buy the items that are different between the two baskets. Using python, the following code will find the items that are different between the two baskets; and then the items that should be recommended to the two customers; shown in Table 11 and Table 12.

items_bought_by_65693391 = set(customer_item_matrix.loc[65693391].iloc[customer_item_matrix .loc[65693391].to_numpy().nonzero()].index)

items_bought_by_798104708 = set(customer_item_matrix.loc[798104708].iloc[customer_item_matrix.loc[798104708].to_numpy().nonzero()].index)

items_to_recommend_to_798104708 = items_bought_by_65693391 - items_bought_by_798104708

items_to_recommend_to_65693391 = items_bought_by_798104708 - items_bought_by_65693391

Table 11. Items Recommended for Customer ID 798104708

StockCode	Description
285205	Stain remover
159057	Orange Fruit
288340	Pan
191397	Fresh Meat
270103	Indomi Soup
283976	Orange Juice

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 11 (2020), pp. 3444-3455
© International Research Publication House. https://dx.doi.org/10.37624/IJERT/13.11.2020.3444-3455

Table 12. Items Recommended for Customer 65693391

StockCode	Description
288071	Facial tissue Fine
159483	Tuna Cans
288871	Sugar
227650	Black Pepper
253250	Nuts

Thus customized basket for each customer is created. Using this method Sameh Mall will be able to create a basket for each customer based on their behavior.These items shown in Table 11 and Table 12 are sent as recommendation to the customers' account on the online application; which enlarges the market basket size for both customers and then increase sales.

V.II Generating genral market baskets

In the current situation, with COVID-19 pandemic, even though customers like to go physically to the Sameh Mall to buy items, they would like to find items that they usually buy together (called general basket) close to each other. Thus the layout design should be revised. In this way, customers will find items that they are looking for in one place. This will minimize the chance that they need to explore the entire store looking for certain items, and then come into contact with many people in the store. In addition, the presence of a general basket where items that are frequently bought together are placed in one location, reminds the customers with the complement items that they need to buy; which enlarges the market basket size for both customers and then increase sales. Furthermore, as mentioned earlier, general baskets also increases sales of items that are new to the market and also saves the retailer money by guiding the process of cross merchandizing.

A general basket can be generated using the stockCode as a pivot instead of customer ID. The resulting matrix is a square matrix of size 8940 x 8940 ; 8940 is the number of stock items in the dataset.

Using similar analysis as above, the items that are usually bought together can be included in one general basket.Table 13 shows a list of nine items that are correlated to stockCode 253250 (item description is Nuts). Thus Sameh Mall displays those items close to each other. Note that this list has food (cheese, chips, …) and it has chemicals (body lotion, babies shampoo). The management should decide on layout design; they need to decide on how close these items to each other. This will give customers a more satisfying experience in shopping at Sameh Mall because, they will find their items easily. It is important to notice that market basket analysis is totally dependent on seasons and time; thus the previous analysis need to be performed frequently. This would enable the updating of recommendations about both the customized and the general market baskets; increasing customer satisfaction. Finally, this study put forward managerial implications for retail companies on how to adjust their strategies to create an environment that attracts customers and influence their behavior.

Table 13. General Basket Recommendation

Stock Code	Description
253250	Nuts
285045	Cheese
146913	Chips
272920	Honey
1811	Smoked Tuna
268362	Chicken Breast
21154	Body Lotion
288719	Glass Plates
266501	Babies Shampoo
276619	Grilled Turkey Roll

VI. CONCLUSION

This study presents an intelligent market basket analysis using big data analytics. It is applied in Sameh Mall, a top retail store in Jordan. It presents recommendation for the management to adjust marketing, sales and layout design strategy to cope up with the explosion in the digitization revolution and to account for the negative effects of COVID-19 pandemic. Big data analytics utilizes cosine-similarity indexes to find correlations. Customized baskets are generated based on customers' shopping behavior; when calculating similarity index between customers. General baskets are generated based on correlation between stock items which are usually bought together; within same basket. Both physical and online merchandizing is adjusted; where customized marketing which targets VIP customers is recommended. This gives customers a more satisfying experience in shopping at Sameh Mall because, they easily find items; which enlarges the market basket size and then increases sales.

In this study, the association between items is considered regardless of their quantity. Further study could be conducted to include the quantity of each item. In addition, future research should aim to include different measures related customer stay time, distance moved, storage capacity, etc. this can help improve understanding of in-store behavior which is one important factor impacting customers' shopping behavior.

ACKNOWLEDGEMENTS

The authors are grateful to the Applied Science Private University, Amman, Jordan for the full financial support granted to this research.

REFERENCES

[1] S. Amaro, How the coronavirus is changing the way we shop — and what we're buying, JUL 27 2020. https://www.cnbc.com/2020/07/27/the-future-of-retail-amid-covid-19.html

[2] R. Vader, P. Martin, J. Qian, The realities of retailing in a COVID-19 world, KPMG Insights, 2020. https://home.kpmg/xx/en/home/insights/2020/03/realities-of-retailing-in-covid-19-world.html

[3] D. Andrews, Shoring Up Your Marketing Strategy in Turbulent Times, customer think April 10, 2020 https://customerthink.com/shoring-up-your-marketing-strategy-in-turbulent-times/

[4] Steve Harris, Safe shopping: the impact of COVID-19 on retail, https://www.orange-business.com/en/blogs/safe-shopping-impact-covid-19-retail, August 14, 2020 , Digital Transformation

[5] M. Alshriem, The Use of Artificial Intelligence in Traffic Violation Data Analysis, International Journal of Engineering Research and Technology. ISSN 0974-3154 Vol.13, No.4 (2020), pp. 644-652.

[6] N. Al-Dmour, Using Unstructured Search Algorithms for Data Collection in IoT-Based WSN, International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 8 (2020), pp. 1992-1998.

[7] P. Yoganandhini, and G. Prabakaran, Market Basket Analysis with Enhanced Support Vector Machine (ESVM) Classifier for Key Security in Organization, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-2, December, 2019.

[8] G. Ochieng, The adoption of big data analytics by Supermarkets In Kisumu County. University of Nairobi research archives, 2015.

[9] Ahmed, S. (2004). Applications of data mining in retail business. International Conference on Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004.

[10] B. Hemalatha, and T. Velmurugan, Direct-Indirect Association Rule Mining for Online Shopping Customer Data using Natural Language Processing, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-4, 2019,

[11] F. Wu, Internal analysis of asymmetric competitive market structure using supermarket aggregate data, Doctoral Dissertation, university of Alberta, 2011.

[12] N. C. Hsieh, An integrated data mining and behavioral scoring model for analyzing bank customers. Expert systems with applications, 27(4), 2004, 623-633.

[13] K. D. Wicker, K. D., A study of customer value and loyalty in the supermarket industry (Doctoral dissertation, Capella University), 2016.

[14] V. Media, 5 Advantages of Market Basket Analysis in B2B Marketing, B2B Marketing Journal, March 2020.

[15] M. John, and H. Shaiba, Apriori-Based Algorithm for Dubai Road Accident Analysis, 16th International Learning & Technology Conference 2019, Procedia Computer Science 163, 2019, 218–227.

[16] J. Gupta, and A. Mahajan, BPSO Optimized K-means Clustering Approach for Data Analysis, International Journal of Computer Applications, 133(15), 2016, 0975 – 8887

[17] J. Han, M. Kamber, and J. Pei, Data mining concepts and techniques, 2011, San Francisco: Morgan Kaufmann .

[18] S. Chai, J. Yang, and Y. Cheng, The Research of Improved Apriori Algorithm for Mining Association Rules, International Conference on Service Systems and Service Management, 2007, doi:10.1109/icsssm.2007.4280173

[19] C. Kim, and J. Kim, A recommendation algorithm using multi-level association rules, IEEE/WIC International Conference on Web Intelligence (WI 2003). doi:10.1109/wi.2003.1241257

[20] C. ygielski, J. Wang, and D. C. Yen, Data mining techniques for customer relationship management, Technology in Society, 24(4), 2002, 483-502.

[21] Ling and Chenghui Li, Data Mining for Direct Marketing, (1998).

[22] S. Ahmed, Applications of data mining in retail business. International Conference on Information Technology: Coding and Computing, ITCC 2004. doi:10.1109/itcc.2004.1286695

[23] P. Prasad, and M. Mourya, A Study on Market Basket Analysis Using a Data Mining Algorithm, International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 6, June 2013, 361-363.

[24] R. Moodley, F. Chiclana, F. Caraffini, and J. Carter, Application of uninorms to market basket analysis, International Journal of Intelligent Systems, Vol 34, Issue 1, Jan 2019, 39-49.

[25] M. Hemalatha, Market basket analysis – a data mining application in Indian retailing, International Journal of Business Information Systems, Volume 10, Issue 1, 2012, 109-129.

[26] Y. Zuo, K. Yada, and A. S. Ali, Prediction of consumer purchasing in a grocery store using machine learning techniques, 3rd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), 2016, 18-25

[27] J. Yao, The research on supermarket business strategy based on customer (Order No. 10419522). Available from ProQuest Dissertations & Theses Global, 2009.

[28] M. Yu, The study of consumers' purchase behavior affected by the environment in general supermarkets (Order No. 10393859). Available from ProQuest Dissertations & Theses Global.

[29] N. Tayyar, Obtaining and analysing own and cross price elasticities in a breakfast ceareals market (Order No. U172288). Available from ProQuest Dissertations & Theses Global, 2002.

[30] S. H. Liao, J. L. Chen, and T. Y. Hsu, Ontology-based data mining approach implemented for sport marketing, Expert Systems with Applications, 36(8), 2009, 11045-11056.

[31] A. Bhargav, R. P. Mathur, and M. Bhargav, Market basket analysis using artificial neural network, International Conference for Convergence for Technology, 2014 (pp. 1-6).

[32] J. P. Dorismond, Data-driven models for promoting impulse items in supermarkets (Order No. 13427730). Available from ProQuest Dissertations & Theses Global, 2019.

[33] Python official website https://www.python.org