Available online at www.scientiaresearchlibrary.com

Scientia Research Library

ISSN 2348-0424
USA CODEN: JETRB4

Journal of Engineering And Technology Research, 2019, 7 (3):1-10

(*http://www.scientiaresearchlibrary.com/arhcive.php*)

MINING CAFÉ SALES DEMAND USING DEPENDENCY ASSOCIATION AND TIME SERIES TECHNIQUES

FERNANDO S*. VIRAY JR.
[1]*Pangasinan State University*
Corresponding Author: [1]skyfher@gmail.com

ABSTRACT

Creating product combinations or pairing with an intention to increase sales revenue while at the same time properly control product inventory stock balance is key in the successful operation of any enterprise. This study focused on analyzing a substantial previous sales transactions with the use of dependency association, affinity, time series, and visualization techniques of data mining in identifying saleable coffee shop product pairs, determining the appropriate product stock balance, and validating such findings with the use of common data mining statistical techniques. The results can help a local café in innovating marketing strategies and providing superior product value to customers. A comparative analysis of sales transactions' history was performed to validate findings and establish a significant degree of accuracy.

Keywords : coffee shop product pairing, café sales analysis, dependency association, affinity model

INTRODUCTION

The coffee shop industry has boomed in the past decades from being a humble local coffeehouse now to a lucrative international coffee shop chain. In the United States alone, the industry comprises of around 30,000 stores with a pooled annual sales of $20billion while one of its biggest coffee chain which has outlets also here in the Philippines, Starbucks, is owning and licensing 12,000 stores located in 75 countries worldwide. Meanwhile, aside from the U.S., Germany, Brazil and Japan topped the world's largest consumers of coffee.[1]

Coffee shops sell coffee beverages and other food items for consumption within the premises of the store or for customers' takeout. Most café outlets specialize in coffee complementary products such as donuts, bagels, ice cream, and yogurt. The demand for café products are driven by personal consumer taste and income. [2] The ability of coffee shop owners to be established in a prime location, provide high-quality café products, and maneuver store traffic dictate high profits in individual companies. In order for small coffee shops to compete effectively with larger cafés which have significant edge in terms of financing, marketing and purchasing, to offer specialty products, innovative promotional packages, excellent customer service are of a high degree of necessity.[3]

Currently, many enterprises are trying to pair their products by creating “combos” or product combinations with the goal of boosting sales revenues, have a steady cash flow and efficiently manage inventories. Product combos implemented with innovative marketing strategy can help coffee shops attract new consumers and amplify demand among loyal ones by incorporating superior value in the product pairing.[4] However, although creating product combos may be promising, the process of identifying product pairing can easily go wrong without a thorough knowledge of surefire methodology.[5]

The objectives of this study are: (1) identify the most saleable coffee shop products and the most saleable coffee shop beverage product paired or associated with a coffee shop food product; (2) determine the most applicable inventory stock balance for café products; and (3) validate product pairings – by using dependency association, affinity, time series and visualization methodologies of data mining over sixteen (16) months of actual sales transactions of a local coffee shop. All these goals are set to purposely benefit the local coffee shop business owner who freely provided integral data to accomplish the goals of this research.

Related Works

The use of data mining has been proven in computing and processing large datasets to extract relevant information after careful analysis and validation by implementing common techniques and methodologies of database systems, statistics, machine learning and artificial intelligence.[6] After every visualization of dataset using several software-based operators, database queries, and statistical formulae, a data miner can draw interpretations and conclude on actual results. Meanwhile, to fully harness the power of data mining, one can further develop a model to predict possible outcomes in case similar instance of data will be subjected on the same operators and procedures used in the first dataset. To be fully convinced of the results of predictions, the model will be tested for accuracy by using the initial dataset and check whether the model has correctly processed each instance of the initial dataset.[7]

Among the earliest methods of detecting patterns in datasets involve regression analysis and Bayes' theorem. As datasets grew larger and larger as time passes by and as available technology in the field of computing and computer hardware resources increased incrementally while having lower cost, digital data gathering, storage and processing has also shown a paralleled dramatic boost. Currently Knowledge Discovery in Databases or KDD has been an established field in computer science, augmented the growth of support vector machines, cluster analysis, neural networks, decision trees, genetic algorithms and decision rules.[8]

There are three most commonly used data mining process models, these are: the KDD, CRISP-DM and SEMMA models. The CRISP-DM model or CRoss-Industry Standard Process for Data Mining (in Figure 1) involves: the systematic understanding of the application domain; the creation of applicable target dataset which is composed of relevant data that pattern/s will be discovered; the methodologies and techniques of data pre-processing (also known as data cleansing or cleaning) or the removal and/or replacement of unwanted, insignificant, null values; data reduction which is the result of data cleansing; selection of appropriate data mining function depending on the results expected or pattern needed to be discovered; selection and application of existing data mining algorithm or the creation of an algorithm if one is not yet available; the implementation of data mining procedure; the interpretation of the mined or discovered results; and finally, the use of discovered knowledge. [9]

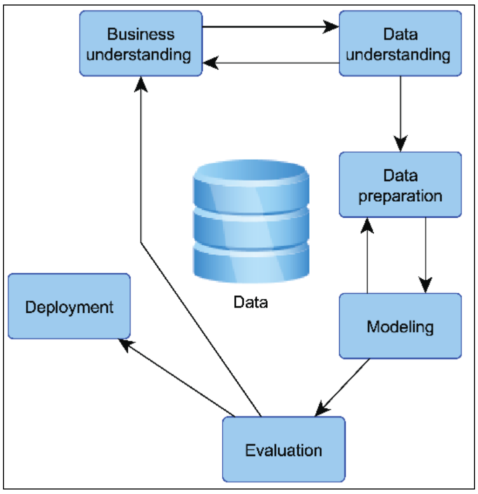

Figure 1. CRISP-DM Model

As per the CRISP-DM model, the selection of existing data mining techniques, functions and algorithm is key in data mining. Among these data mining techniques, functions and algorithms are Market-Basket Analysis, Apriori Algorithm and the Time Series Technique. Market-basket analysis, which is also called association mining, is a data mining affinity analysis technique that refers to the discovery of patterns, relations, and correlations among each instance of item in a dataset.[10]

The Market Basket Analysis is an application of the Apriori Algorithm which is used to mine frequent itemset and learn association rules for datasets stored in transactional databases. Its pseudocode was proposed by Rakesh Agrawal and Ramakrishnan Srikant in 1994 as follows: [11]

```
Apriori(T, ε)
    L_1 ← {large 1 − itemsets}
    k ← 2
    while L_{k−1} ≠ ∅
        C_k ← {a ∪ {b} | a ∈ L_{k−1} ∧ b ∉ a} − {c | {s | s ⊆ c ∧ |s| = k − 1} ⊈ L_{k−1}}
        for transactions t ∈ T
            C_t ← {c | c ∈ C_k ∧ c ⊆ t}
            for candidates c ∈ C_t
                count[c] ← count[c] + 1
        L_k ← {c | c ∈ C_k ∧ count[c] ≥ ε}
        k ← k + 1
    return ⋃_k L_k
```

Meanwhile, time series analysis is composed of methodologies for investigating data that occurred periodically and aims to mine and infer significant characteristics and statistics of data. The data to be subjected under this analysis is most of the times plotted in line charts to visually give face to data as they occur over time.[12]

Many researchers worldwide considered mining knowledge and information from arbitrary large datasets as a key subject matter in machine learning and database systems. Commercial industries and corporations with sufficient resources are driven by major revenue and are currently taking advantage in collecting knowledge from discovered patterns from these datasets.[13]

MATERIALS AND METHOD

The KDD model was used by the researchers in this study as model of data mining. Sixteen (16) months of sales transactions, together with its corresponding sales details, were gathered from a local coffee shop in the aim to find out and identify the actual coffee shop products bought by local consumers. The data are exported from a MySQL relational database management system as a selfcontained dump file format with a .sql file extension. To pre-process the data and since the dump file contains the structure of the schema where it came from, the researchers imported the said .sql file in a MySQL server to recreate the schema, developed and executed queries, and produced a 12-thousand-plus rows of dataset which was exported as a comma-delimited text file (.csv format).

The .csv-format dataset was imported to the latest (as of this writing) versions of RapidMiner and Orange, open-source data mining computer tools both implementing graphical interface programming, both with built-in operators, statistical processes, and visualization features thru advanced charts.

The market-basket analysis was employed in order to achieve the research's first and second objectives – identify the most saleable coffee shop products and its appropriate stock balance. Once the frequency of a coffee shop item is analyzed, its corresponding volume of consumption (units sold per instance in the frequency) can be analyzed and its subsequent average and mean can be easily computed which can be a basis to solve the second objective of this research in the form of product's reorder level. Products identified are tabulated together with their percentages of support and confidence.

The time series analysis has been used by the researchers to validate the degree of accuracy and confidence on the solutions and interpretations of visualizations of data as they repeatedly show similar results over different periods as time series analysis are mostly applied to evaluate data which are produced consecutively over a series of time and find out certain interestingness and regularities.

Meanwhile, an SQL algorithm similar to the Apriori algorithm was developed to fully identify the actual coffee shop food products bought and paired by consumers with their choice of coffee shop beverage products. Since Apriori algorithm accounts for all instances in a dataset, the developed SQL algorithm excludes single item transactions and covered only those sales transactions with 2 or more items of sale transactions and such items must both have beverage and food items. If sales transaction items are more than two, each product is permuted with one another. The volume of orders were also noted as ranking in terms of instance frequency have different inference and conclusive details with that of quantitative ranking in terms of units sold per instance. The results of these rankings were also subjected to the utilized data mining software for visualizations. Below is the pseudo code of the SQL pairing algorithm developed by the researchers in identifying the actual food and beverage product pairing.

```
For Each SalesTransaction S In All SalesTransactions ST
  If S.Items.Count=1 Then Go To NextTransaction
  If HasFoodNDrink(S.Items)=False Then Go to NextTransaction
  For x = LBound(S.Items) to UBound(S.Items)-1
    If (S.Items(x).Type=Food AND S.Items(x+1).Type=Drink) OR
       (S.Items(x).Type=Drink AND S.Items(x+1).Type=Food)Then
ID=GetArrayPairID(S.Item(x),S.Item(x+1))
      If ID=-1 Then
ArrayPair.Insert(S.Item(x), S.Item(x+1))
      Else
 ArrayPair(ID).Count += 1
      End If
    End If
   Next x
NextTransaction:
Next S
```

RESULTS AND DISCUSSION

Market Basket Analysis. Figure 2 exhibits the graph of coffee shop product frequency among the

8,065 sales transactions recorded from January 1, 2016 to April 30, 2017 (484 sales days). It is notable that among 219 products of the local coffee shop, 16 of it have stood out to be the most frequent café products consumed by customers and among these 16 products, nachos products remained consistently as food item frequently ordered by customers in the said period. These top saleable products are shown according to the highest frequency count in Figure 3.

To highlight and differentiate food from beverage products, Figure 4 shows the frequency graphs of the top 7 coffee shop food items and top 7 beverage items. The graph highlights that the nachos products have been the most frequent café product which are most of the times being partnered with the local café's own concoction of milk-based tea beverage and their home-style brewed coffee.

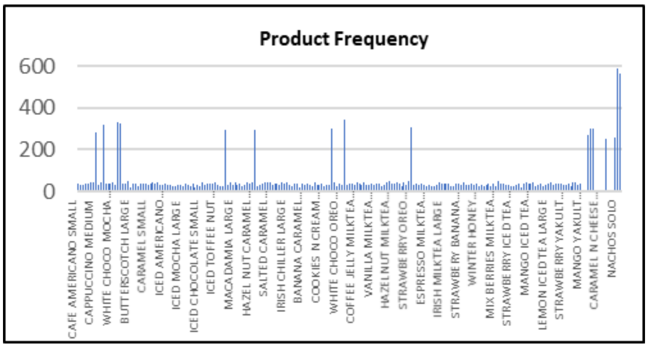

Figure 2. Graph of Product Frequency from Jan. 1, 2016 to Apr 30, 2017.

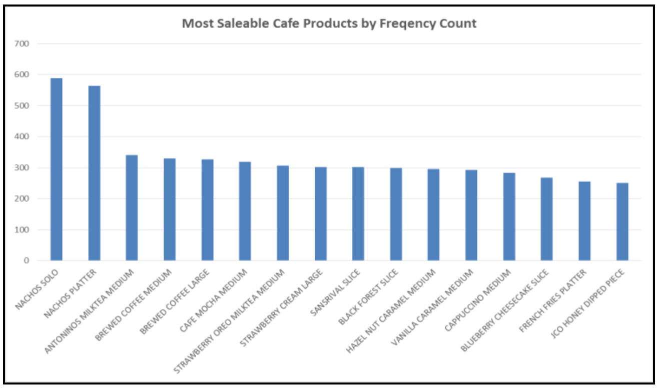

Figure 3. Graph of the most saleable café products consumed by customers from January 1, 2016 to April 30, 2017 based on frequency analysis.

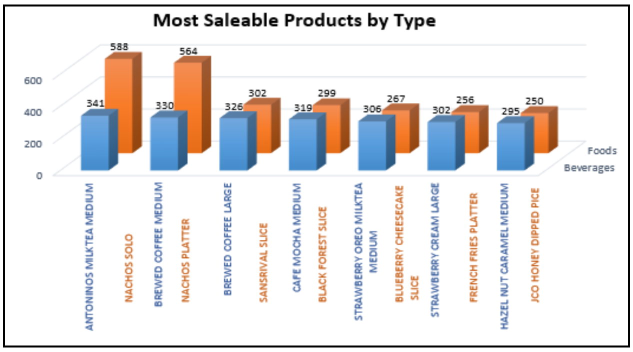

Figure 4. Frequency Graph of Top 7 Foods and Top 7 Beverages.

Figure 5 exhibits the average daily sales as to product type from January 1, 2016 to April 30, 2017. The graph shows that among café food products, finger food and pastries topped the sales while the home-made coffee chillers and iced teas of the local café topped sales among the beverage items.

Table 1 presents the association/affinity rules generated using the association operator of RapidMiner as part of the Affinity/Market Basket Analysis conducted. Focusing on the rules that have the highest degree of confidence (1.000), it is indicative that if a customer will order Sanrival Slice, it is most likely that the said customer will also order Antonino's Milktea Medium. If a customer will order, a platter of French fries, the customer will most likely order Vanilla Caramel Medium as a complementing beverage. Meanwhile, the rules also stated that if a customer will order a Hazel Nut Milk Tea Medium, the customer will most likely order either Vanilla Caramel Medium or French Fries Platter.

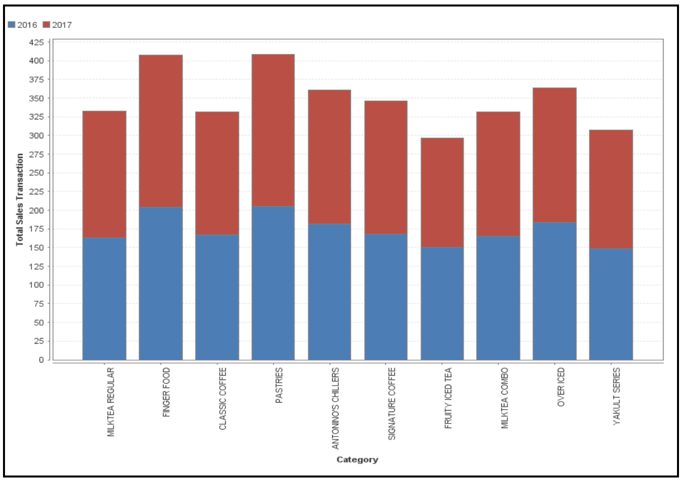

Figure 5. Stack graph of average daily sales as to product type from Jan. 1- Apr 30, in the years 2016 and 2017.

Table 1
Association Rules Generated using RapidMiner

PRODUCT / PAIRING			CONFIDENCE
CHOCOMINT MILKTEA SMALL	->	NACHOS PLATTER	0.250
BUTTERSCOTCH MEDIUM	->	NACHOS PLATTER	0.250
BUTTERSCOTCH MEDIUM	->	STRAWBERRY CREAM LARGE	0.250
HAZELNUT MILKTEA MEDIUM	->	NACHOS SOLO	0.333
STRAWBERRY OREO MILKTEA SMALL	->	ANTONINOS MILKTEA MEDIUM	0.375
STRAWBERRY OREO MILKTEA SMALL	->	SANSRIVAL SLICE	0.375
CHOCOMINT MILKTEA SMALL	->	NACHOS SOLO	0.429
NACHOS SOLO	->	STRAWBERRY OREO MILKTEA MEDIUM	0.477
NACHOS PLATTER	->	STRAWBERRY CREAM LARGE	0.500
NACHOS PLATTER	->	CAFÉ MOCHA MEDIUM	0.504
NACHOS SOLO	->	BREWED COFFEE MEDIUM	0.524
BUTTERSCOTCH MEDIUM	->	BLUEBERRY CHEESECAKE SLICE	0.667
BUTTERSCOTCH MEDIUM	->	HAZELNUT CARAMEL MEDIUM	0.667
CAFÉ MOCHA MEDIUM	->	NACHOS PLATTER	0.969
ANTONINOS MILKTEA MEDIUM	->	SANSRIVAL SLICE	0.971
VANILLA CARAMEL MEDIUM	->	FRENCH FRIES PLATTER	0.981
BREWED COFFEE LARGE	->	BLACK FOREST SLICE	0.983
BREWED COFFEE MEDIUM	->	NACHOS SOLO	0.984
HAZELNUT CARAMEL MEDIUM	->	BLUEBERRY CHEESECAKE SLICE	0.985
STRAWBERRY OREO MILKTEA MEDIUM	->	NACHOS SOLO	0.986
CAPPUCCINO MEDIUM	->	JCO HONEY DIPPED PIECE	0.988
STRAWBERRY CREAM LARGE	->	NACHOS PLATTER	0.989
BLACK FOREST SLICE	->	BREWED COFFEE LARGE	0.993
BLUEBERRY CHEESECAKE SLICE	->	HAZELNUT CARAMEL MEDIUM	0.996
JCO HONEY DIPPED PIECE	->	CAPPUCCINO MEDIUM	0.996
SANSRIVAL SLICE	->	ANTONINOS MILKTEA MEDIUM	1.000
FRENCH FRIES PLATTER	->	VANILLA CARAMEL MEDIUM	1.000
HAZELNUT MILKTEA MEDIUM	->	VANILLA CARAMEL MEDIUM	1.000
HAZELNUT MILKTEA MEDIUM	->	FRENCH FRIES PLATTER	1.000

Association rules or affinity rules generated using RapidMiner Studio among the 219 coffee shop products consumed by customers from January 1, 2016 to April 30, 2017 based on market-basket analysis.

Figure 6 showcases the affinity graph of the most frequently sold café products and their relationship to other products. The graph shows that Nachos Solo, Nachos Platter, Butterscotch Medium, and Antonino's Milk Tea, respectively, are among the most correlated products that coffee shop clients consumed with most products linked to it.

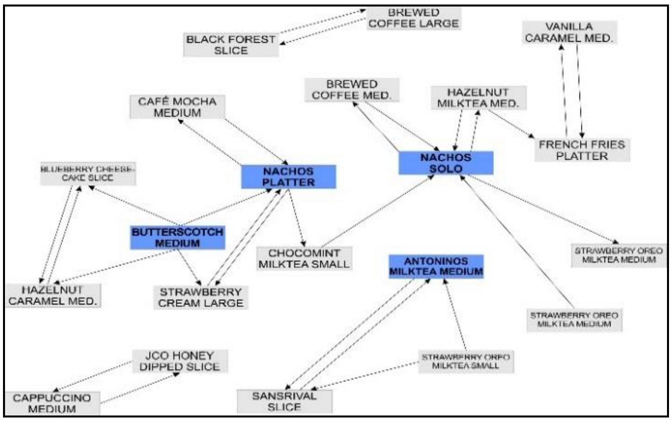

Figure 6. The affinity graph of products and their correlation to other items.

Product Pairing. Figure 7 displays the top 20 product pairs as the result of the modified Apriori pairing algorithm developed by the researchers which counted and ranked the frequency of actual paired food and beverage coffee shop products as consumed by the customers of a local coffee shop. The algorithm identified 898 beverage-food pairs. Unlike the association results of data mining software tools which leaves to the user the interpretation of the affinity outcome, details of Figure 7 proved the actual beverage and food products pairing as preferred by consumers.

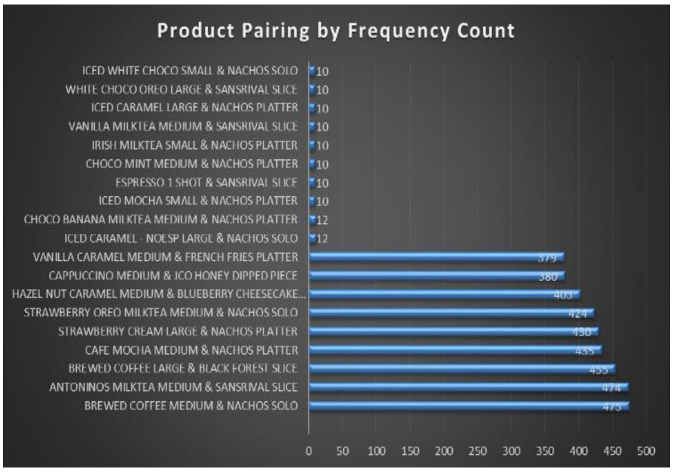

Figure 7. Results of the Pairing Algorithm ranking the most frequently paired food and beverage café products in terms of frequency count.

In comparison to Figure 7, Figure 8 exhibits the result of the pairing algorithm which summed the quantity sold per product for each instance of sales transactions. This provides clarity and basis with regards to the relationship of ranking café products in terms of frequency count and ranking them in terms of the volume of quantity sold. It is indicative that although a product pair may rank first in terms of frequency count, its rank may change if ranking will be based to volume of quantity sold.

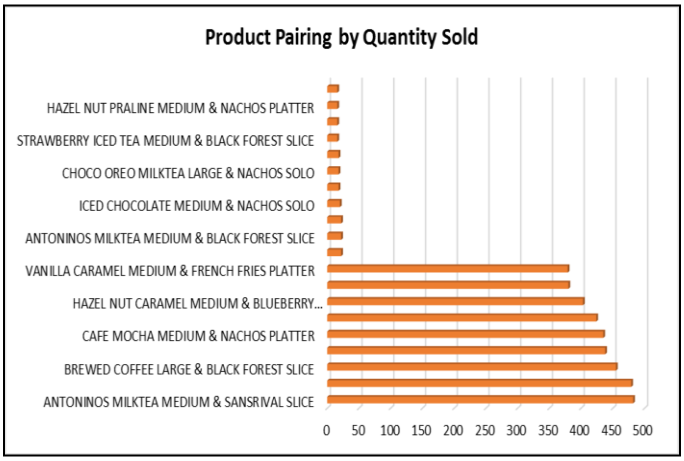

Figure 8. Results of the pairing algorithm ranking the most frequently paired food and beverage café products in terms of quantity sold.

Stock Balance. Since café beverage products are to be prepared as per order basis unlike café food products, and considering also that the ingredients of one beverage product is also an ingredient of another beverage, this study focused only in identifying the average stock balance of coffee shop food products which are sold per piece since it will require food costing data in order to actually compute stock balance for each café beverage stock balance per ingredient. Figure 8 shows the average units sold per day among café food products from January 1, 2016 to April 30, 2017. Judging from these figures, it is safe to interpret that stock balance of food products shown in Figure 9 should be maintained high or above reorder level as indicated.

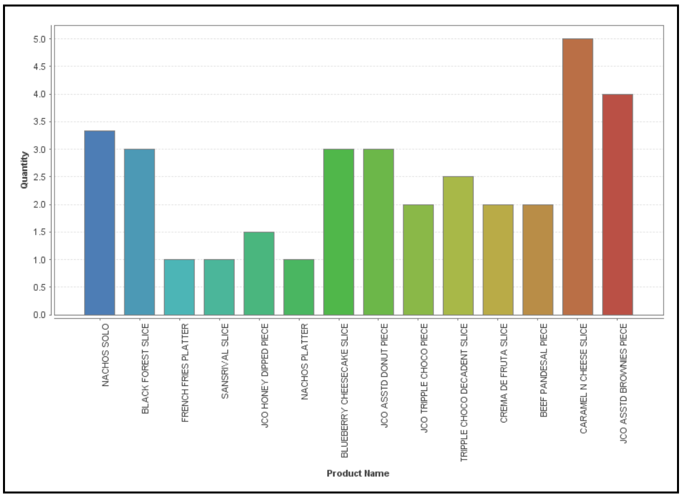

Figure 9. Average units sold per day of coffee shop food products from January 1, 2016 to April 30, 2017.

Meanwhile, Figure 10 reveals that every Tuesdays (from 3:00 PM - 9:00PM) and every Saturdays (from 2:00PM-9:00PM), sales traffic in the business premises occurred and most customers preferred cold beverages and ChipsNDips types of food products. It is indicative that on these particular days of the week and time period, café products stock balance should be kept above reorder level.

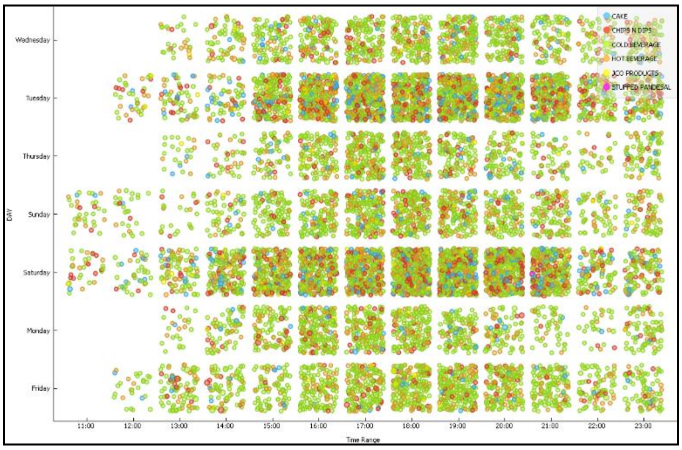

Figure 10. Sales in terms of day and time period to identify when stock balance should be high.

Validation Using Time Series Techniques. To validate the occurrence of sales among coffee shop products, the researchers employed time series technique by comparing sales from the different periods. Figure 11 exhibits the daily average sales of café products for the first 4 months of 2016 (in blue series line) and the first 4 months of 2017 (in orange series line). The graph is conclusive that saleable products from the previous year remain to be the top saleable products on the same period on the preceding year. These results validated the correctness of that results of the data mining techniques undertaken by the researchers and add up to the degree of confidence and accuracy of such results.

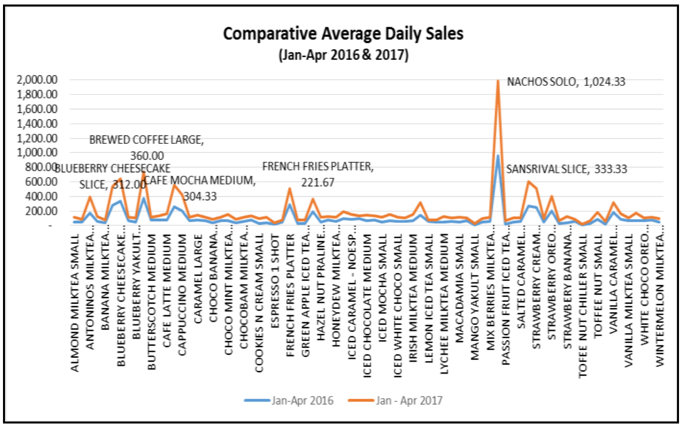

Figure 11. Comparative average daily sales of coffee shop products for the first 4 months of the years 2016 and 2017.

CONCLUSION

By using Affinity/Market Basket Analysis techniques of data mining, the researchers were able to find relations and associations of each product with other coffee shop products which gained significant percentage of support and confidence and determined the most saleable products based on the 16-month sales transactions of the subject coffee shop by utilizing such percentages. The correlations of these products were further emphasized by the result of the modified Apriori pairing algorithm developed in which top correlated products as resulted from the utilized data mining software tool are exactly pinpointed in terms of actual food-is-to-drink pairing. Time series analysis has also proven that such products ranked as the primary items being bought by customers regularly in the first four months of the year 2016 paralleled to the first four months of the year 2017 which indicated additional degree of accuracy and confidence in the results of data mining methodologies employed in this study.

REFERENCE

[1] Dun & Bradstreet First Research. Article: *Coffee Shops Industry Profile*.

http://www.firstresearch.com/Industry-Research/Coffee-Shops.html, accessed **2017**-05-04.

[2] Godsmark, Elizabeth; Arduser, Lora; and Brown, (Author), Douglas R. (**2004**). *How to Open a Financially Successful Coffee, Espresso & Tea Shop: With Companion CD-ROM*. Atlantic Publishing Group Inc. ISBN-10: 0910627312 / ISBN-13: 978-0910627313.

[3] Scott, David Meerman (**2010**). *The New Rules of Marketing and PR.* Wiley. ISBN: 0470547812 / 9780470547816.

[4] Shankar, Venkatesh; Berry, Leonard L.; and Dotzel, Thomas. (**2009**). Hardvard Business Review. Article: *A Practical Guide to Combining Products and Services.* https://hbr.org/2009/11/a-practical-guide-to-combining-products-and-services, accessed: 2017-05-04.

[5] 50minutes.com (**2015**). *The Marketing Mix: Master the 4 Ps of Marketing*. 50 Minutes. ISBN: 2806266122 / 9782806266125.

[6] Hastie, Trevor; Tibshirani, Robert; and Friedman, Jerome (**2009**). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction 2nd Edition.* Springer Science and Business Media. ISBN: 0387848584 / 9780387848587.

[7] Thuraisingham, Bhavani (**2000**). *Article: A Primer for Understanding and Applying Data Mining.* IT Pro. IEEE. Retrieved 2017-05-04.

[8] Fayyad, U.M., Smyth, P. and Uthurusamy (**1996**). *Advances In Knowledge Discovery and Data Mining Revised Edition*. American Association for Artificial Intelligence Press. ISBN: 0262560976/ 9780262560979.

[9] O. Maimon, M. Last. (**2000**). *Knowledge Discovery and Data Mining: The Info-Fuzzy Network (IFN) Methodology*. Springer Science and Business Media. ISBN: 0792366476 / 9780792366478.

[10] Han, Jiawei (**2006**). Data Mining: Concepts and Techniques. Elsevier. ISBN:1558609016 / 9781558609013.

[11] Agrawal, Rakesh and Srikant, Ramakrishnan (**1994**). *Fast Algorithms for Mining Association Rules in Large Databases.* Morgan Kaufmann Publishers, Inc. ISBN: 1-55860-153-8.

[12] Shumway, Robert H., Stoffer, David S. (**2006**). *Time Series Analysis and Its Applications 2nd Edition.* Springer-Verlag New York. ISBN: 978-0-387-36276-2.

[13] Bramer, Max A. (**1999**). *Knowledge Discovery and Data Mining*. The Institution of Engineering and Technology (IET). ISBN: 0852967675/978085296767.