ANLY482 ANALYTICS PRACTICUM

Interim Report

Oh Peng Ho | Yap Jessie | Tan Jhun Boon

APSpoilMarket

MARKET BASKET ANALYSIS

LITERATURE REVIEW

WHAT IS MARKET BASKET ANALYSIS

Market Basket Analysis (MBA) or also known as Association Analysis has been one of the most popular data analysis techniques used for marketing (Marakas, 2003). MBA was first introduced by Agrawal, Imielinski, and Swami (1993). The name is derived from analysing items bought together in a market basket by customers while at a grocery store or supermarket. The aim of the analysis is to identify that when a customer purchases a particular item, a second particular item will be predictably purchased as well. Historically, a classic example of MBA is the purchasing of “beers” and “diapers”, 2 items seemingly unrelated but shown to be often bought together, having a high association. In recent times, common application of MBA can be found in online book stores where customers are recommended “associated products” when they purchase a particular item.

MARKET BASKET ANALYSIS IN FOOD AND BEVERAGE INDUSTRY

While MBA has been used extensively in fast-food chains to initiate cross-selling and up-selling opportunities between food and beverage products, little application of MBA has been carried out on more expensive food products. A rare application of MBA is used on a prix fixe menu of a Japanese-style chain store restaurant in central Taiwan (Ting, Pan, Chou, 2010). In that example, a simplified application of MBA is carried out using Microsoft Excel’s PivotTables. The team takes into consideration the methodology carried out by the above experiment when considering MBA within sets, essentially taking on the role of a prix fixe menu. However, the team takes the application further and employs a full Association Analysis on all food items found on the menu.

OBJECTIVE IN MARKET BASKET ANALYSIS

At the end of the day, the analysis seeks to provide the user with information of products that are commonly bought together. The value of identifying such information allows for cross-selling and up-selling opportunities. Furthermore, it allows the store to shake the “undecided” customer – someone who cannot decide whether to buy a particular product or not to a customer who buys the product from the store. MBA helps to create associations in the mind of the “undecided” customer that is analysis-driven. MBA allows retailers to carry out calculated marketing – if two products have high association it would mean that:

1. Putting a product on discount will see a rise in the sales of the other product
2. By placing products that have high association together, they will see an increase in sales volume.

DATA EXPLORATION

Looking at both MW and RP's product sales, the sale of Main - Meal has a decreasing trend, together with Main - Drink. This is likely due to the introduction of Set Menus, where customers tend to prefer purchasing sets rather than ala carte. The indirect relationship between Main - Meal and Set Menu is a lot more obvious in RP. We can see that the moment Set Menu was introduced in November, Main - Meal sales started dropping.

The most popular Main - Meal in both outlets would be the Kaisendon, which is sold 60 times and 50 to 60 times daily on average in MW and RP respectively. It's relating set is the Seafood Feast which averages 40 and 24 times daily in MW and RP respectively.

We can also see that the sale of Main - Onigiri and Main - Fried have relatively stagnant to decreasing trends in both outlets. This means that the onigiris and fried items may not be very popular items. In order to boost sales of onigiris and fried items, Teppei Syokudo may want to consider introducing onigiri sets and fried item sets.

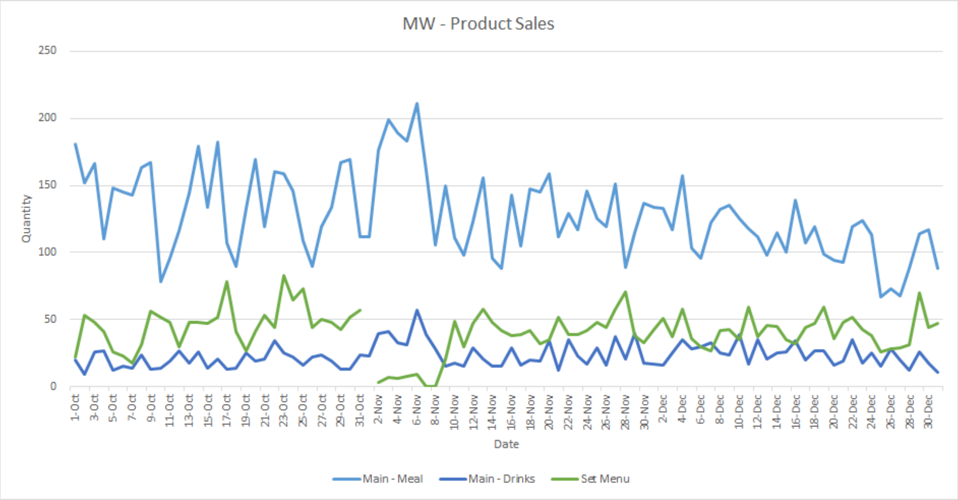


Figure 1. MW - Product Sales – Meal, Drinks and Sets

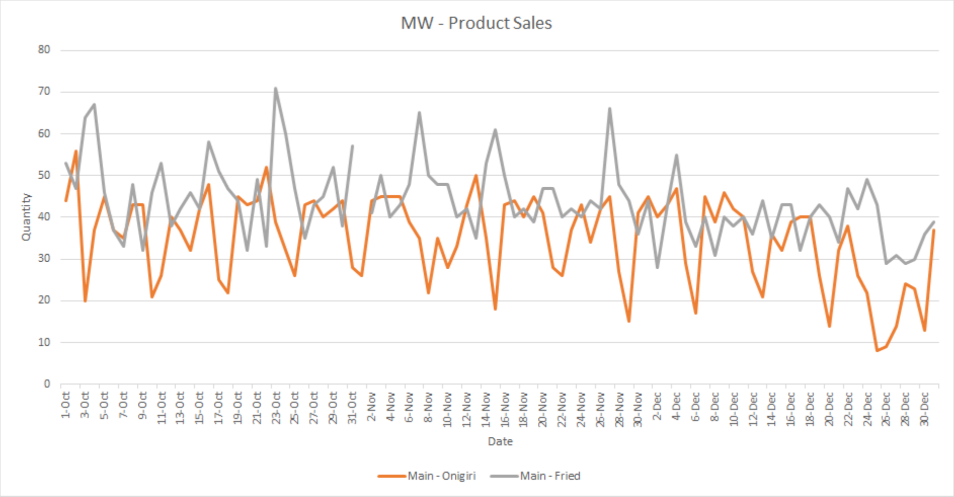


Figure 2. MW - Product Sales – Onigris and Fried

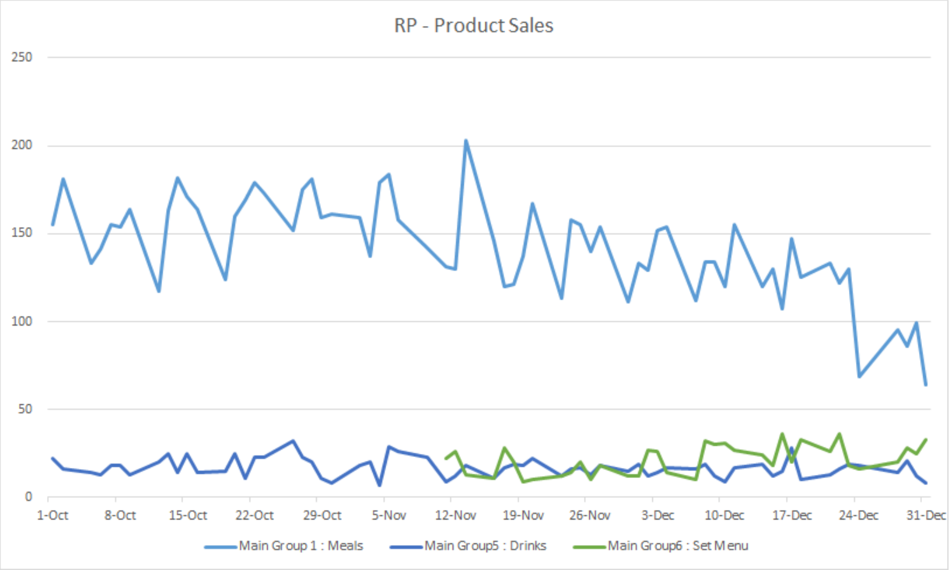


Figure 3. RP - Product Sales - Meals, Drinks and Sets

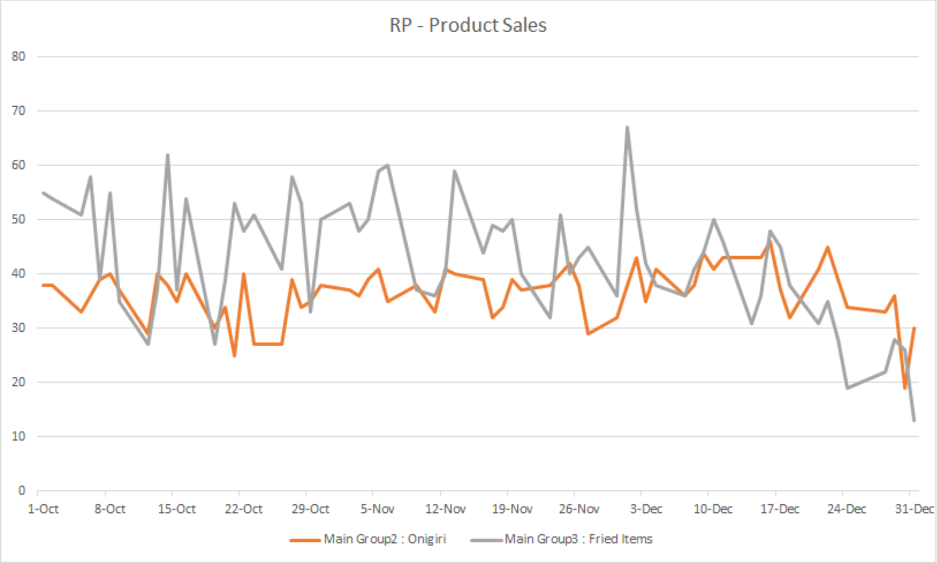
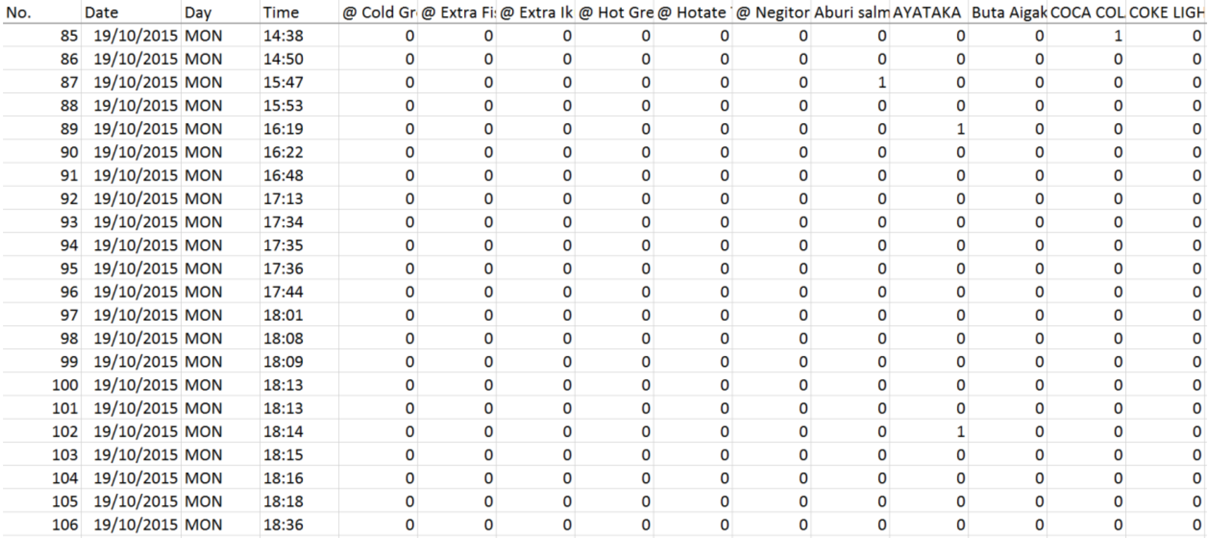


Figure 4. RP - Product Sales - Onigris and Fried

DATA CLEANING & PREPARATION

Receipt data had to be recoded to rows of transaction data. The following dataset is prepared before data analysis can be carried out:



No.	Date	Day	Time	@ Cold Gr	@ Extra Fi:	@ Extra Ik	@ Hot Gre	@ Hotate '	@ Negitor	Aburi salm	AYATAKA	Buta Aigak	COCA COL	COKE LIGH
85	19/10/2015	MON	14:38	0	0	0	0	0	0	0	0	0	1	0
86	19/10/2015	MON	14:50	0	0	0	0	0	0	0	0	0	0	0
87	19/10/2015	MON	15:47	0	0	0	0	0	0	1	0	0	0	0
88	19/10/2015	MON	15:53	0	0	0	0	0	0	0	0	0	0	0
89	19/10/2015	MON	16:19	0	0	0	0	0	0	0	1	0	0	0
90	19/10/2015	MON	16:22	0	0	0	0	0	0	0	0	0	0	0
91	19/10/2015	MON	16:48	0	0	0	0	0	0	0	0	0	0	0
92	19/10/2015	MON	17:13	0	0	0	0	0	0	0	0	0	0	0
93	19/10/2015	MON	17:34	0	0	0	0	0	0	0	0	0	0	0
94	19/10/2015	MON	17:35	0	0	0	0	0	0	0	0	0	0	0
95	19/10/2015	MON	17:36	0	0	0	0	0	0	0	0	0	0	0
96	19/10/2015	MON	17:44	0	0	0	0	0	0	0	0	0	0	0
97	19/10/2015	MON	18:01	0	0	0	0	0	0	0	0	0	0	0
98	19/10/2015	MON	18:08	0	0	0	0	0	0	0	0	0	0	0
99	19/10/2015	MON	18:09	0	0	0	0	0	0	0	0	0	0	0
100	19/10/2015	MON	18:13	0	0	0	0	0	0	0	0	0	0	0
101	19/10/2015	MON	18:13	0	0	0	0	0	0	0	0	0	0	0
102	19/10/2015	MON	18:14	0	0	0	0	0	0	0	1	0	0	0
103	19/10/2015	MON	18:15	0	0	0	0	0	0	0	0	0	0	0
104	19/10/2015	MON	18:16	0	0	0	0	0	0	0	0	0	0	0
105	19/10/2015	MON	18:18	0	0	0	0	0	0	0	0	0	0	0
106	19/10/2015	MON	18:36	0	0	0	0	0	0	0	0	0	0	0

Figure 5. Cleaned Data

The data was segregated between set and non-set orders to try to find patterns in both groups of data. In the non-set data, sets and their components are treated as one entire item. For sets, we have chosen to analyse the components within the sets - identifying popular choices from non-popular choices.

DATA ANALYSIS METHODOLOGY

Market Basket Analysis is broken down to two broad steps: frequent itemset generation and the creation of association rules. Popular algorithms employed are such as the Apriori and FP-Growth algorithms.

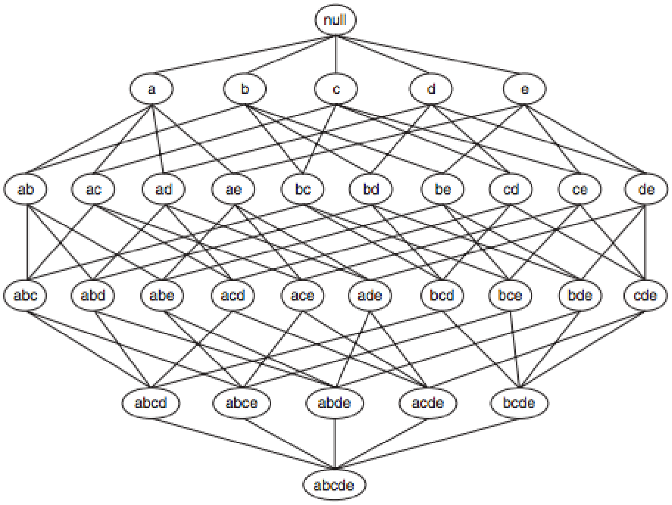


Figure 6. Itemset Lattice Example

Consider the above lattice – each of these are itemsets. Algorithms have to identify the most efficient way to traverse the lattice and identify if a particular itemset is frequent. There are various ways of generating candidates for frequent itemsets and pruning, and this is determined by the algorithm used to carry out association analysis. The way the itemsets are generated and association rules created determine how computationally complex the analysis will be. Therefore, considerations affecting the computational complexity of an algorithm have to be determined when dealing with mining association rules for large datasets. These include factors such as transaction width, number of products, minimum support level and max itemset size (Tan, Michael, Kumar, 2005). Since the transaction width and number of products are predetermined, the team has chosen to specifically focus on the latter 2 factors to refine for our analysis - association thresholds and the max itemset size.

An important aspect of association analysis is the generation of frequent itemsets (or the elimination of infrequent itemsets). The minimum support (minsup) and minimum confidence (minconf) is taken into account. These are thresholds used to determine if for A -> B whether the itemsets A and B are frequent itemsets and whether A -> B is an acceptable association rule. While the team has explored algorithms to determine the optimal minimal support and minimal confidence levels such as applying Particle Swarm Optimization, the team has examined the data spread to determine appropriate minimum support and confidence levels.

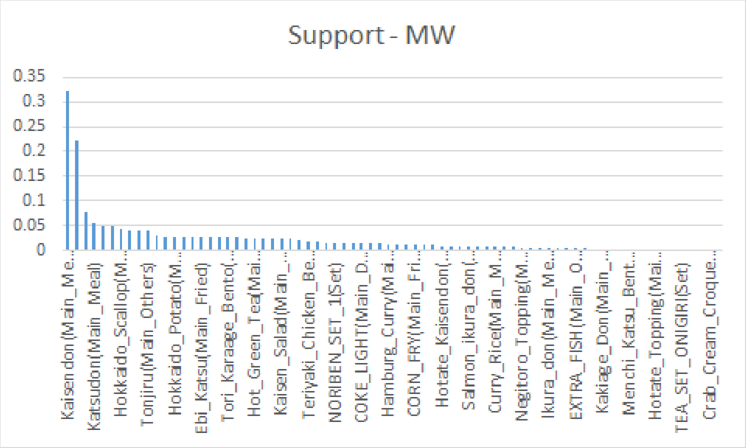


Figure 7. Support - MW

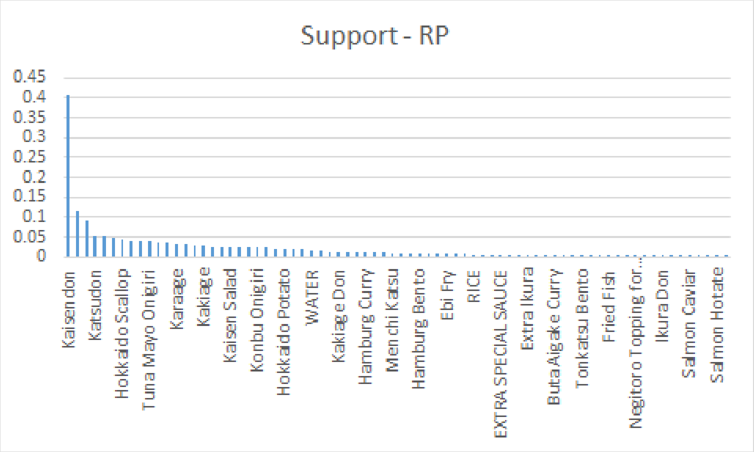


Figure 8. Support – RP

Based on the support levels seen in our dataset we can see that most products have a rather low support level of less than 0.05. This is because most customers of the store purchase particular products - namely the Kaisendon and sets containing the Kaisendon. Hence a minimum support level of 0.005 is selected as compared to conventional levels that are 0.1 and higher. Although having a low minimum support and confidence level might create a higher computational complexity, currently the computational complexity of the data mining is low due to the low transaction width and the number of products.

A relatively low minimum confidence level of 0.1 is also selected. A max itemset size of 2 is set since most transactions have a low transaction width of 1.51 (MW) and 1.49 (RP).

DATA ANALYSIS TOOLS

In carrying out MBA, certain considerations have to be made. One important factor is the software or tool used to carry out MBA. Based on the client requirements in this project, the tool used must be one that is open-source and easy to use. While the team understands that there are far greater utility in employing paid software such as Clementine (SPSS), Enterprise Miner (SAS), GhostMiner 1.0, Quadstone or XLMiner, this requirement essentially narrows down the tools that the team is able to use (Haughton et. al., 2003). The tools that are open-source are narrowed down into 3 tools: RapidMiner, R and Tanagra.

Software and Package	Pros		Cons
R (arules package)	• Free to use • Easy to install • Association Analysis tutorial document easily available	• Flexibility and Customizability	• Difficulty learning curve for using Software • Difficulty in manipulating input dataset • Programming Background required
Rapidminer (FP-growth operator)		• Extensive Interestingness Measures • Gentle learning curve for using Software	• Set number of operators that can be used • Set Operator-based processes limits the customizability of processes
Tanagra (Apriori PT Component)		Gentle learning curve for using Software	• Limited interestingness measures • Lack of customizability in set software processes

Table 1. Market Basket Analysis Software Analysis

After evaluating the 3 tools, the team realized that though R provided measures and customizability, the learning curve to use R is extremely steep and may not be best for the client based on the non-programming nature of their background. Both RapidMiner and Tanagra is extremely lightweight and easy to use, however the presence of extensive interestingness measures caused the team to choose in favour of RapidMiner.

DATA ANALYSIS MEASURES

Using Rapidminer, six different interestingness measures is collected - Support, Confidence, LaPlace, p-s, Lift, Conviction.

Measure	Formula	Range
Support	$P(A, B)$	0 … 1
Confidence	$\max(P(B\|A), P(A\|B))$	0 … 1
Lift	$\mathrm{lift}(X \Rightarrow Y) = \frac{\mathrm{supp}(X \cup Y)}{\mathrm{supp}(X) \times \mathrm{supp}(Y)}$	0 … 1
LaPlace	$\max\left(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2}\right)$	0 … 1
Leverage[1]	$P(A, B) - P(A)P(B)$	-0.25 … 0 … 0.25
Conviction	$\max\left(\frac{P(A)P(B)}{P(A\overline{B})}, \frac{P(B)P(A)}{P(B\overline{A})}\right)$	0.5 … 1 … ∞

Table 2. Market Basket Analysis Measures

By analysing the interestingness measures based on three key properties (Piatetsky-Shapiro, 1991) that determines a good measure:

- *M* = 0 if *A* and *B* are statistically independent;
- *M* monotonically increases with *P(A, B)* when *P(A)* and *P(B)* remain the same;
- *M* monotonically decreased with *P(A)* (or *P(B)*) when the rest of the parameters (*P(A,B)* and *P(B)* or *P(A)*) remain unchanged.

[1] Leverage, Piatetsky-Shapiro Measure (p-s)

Below is an analysis of the 5 measures based on the above 3 properties:

Measure	Property 1	Property 2	Property 3
Support	No	Yes	No
Confidence	No	Yes	No
Lift2	Yes*	Yes	Yes
LaPlace	No	Yes	No
Leverage	Yes	Yes	Yes
Conviction	No	Yes	No

Table 3. Market Basket Analysis Measure Analysis

The analysis shows that lift is a good interestingness measure if the data is normalized and Leverage is generally a good interestingness measure. In essence both lift and leverage serves our purpose in interpreting the analysis results - to measure how many more units of an itemset is sold together than expected from the independent sales. Leverage identifies the difference of *A* and *B* occurring together in transactions in a dataset and what would be expected if they were **statistically dependent** (Piatetsky-Shapiro, 1991). Lift conversely measures directly how many times more often does *A* and *B* occur together than expected if they were **statistically independent**. Consequently, the two measures provides the same ranking or ordering of products since the meaningfulness of these two measures are essentially the same.

[2] Yes if measure is normalized

ANALYSIS RESULTS

After understanding the interestingness measure to analyse the result set, the team examines the analysis results. The results are broken into 2 sections: sets and non-sets.

SETS

Within sets, we look at the association between main courses and their drinks or side toppings. Since main dishes within a set does not vary, they're the independent variable, the premise and the side dishes or drinks are the dependent variable and the conclusions. We've broken down set components to the various main dishes.

Premises	Conclusion	Leverage	Lift
Kaisendon(Meal)	Hot_Green_Tea(Drink)	0.0493	1.1339
Kaisendon(Meal)	Hotate_Topping(Topping)	0.0464	1.1453
Kaisendon(Meal)	Cold_Green_Tea(Drink)	0.0424	1.1453
Kaisendon(Meal)	EXTRA_FISH(Topping)	0.0346	1.1453
Kaisendon(Meal)	Negitoro_Topping(Topping)	0.0247	1.1453
Kaisendon(Meal)	Hot_Green_Tea(Drink), Hotate_Topping(Topping)	0.0237	1.1453

Table 4. MW Sets Analysis Results

We can see that the most popular topping is the Hotate Topping and the Hot Green Tea. (We couldn't analyse the data for RP since there is only one set's data collected and this provides an inaccurate measure of the set's components' association with each other.

NON-SETS

Premises	Conclusion	Leverage	Lift
Ebi_Katsu(Main_Fried)	Hokkaido_Scallop(Main_Fried)	0.004	6.141
Hokkaido_Scallop(Main_Fried)	Ebi_Katsu(Main_Fried)	0.004	6.141
Salmon_Caviar_(Ikura)_Onigiri(Main_Onigiri)	Salmon_Onigiri(Main_Onigiri)	0.004	3.900
Salmon_Onigiri(Main_Onigiri)	Salmon_Caviar_(Ikura)_Onigiri(Main_Onigiri)	0.004	3.900
AYATAKA(Main_Drink)	Katsu_Curry(Main_Meal)	0.003	1.913
Hot_Green_Tea(Main_Drink)	Kaisendon(Main_Meal)	0.001	1.142
AYATAKA(Main_Drink)	Kaisendon(Main_Meal)	0.000	0.986
Tonjiru(Main_Others)	Kaisendon(Main_Meal)	-0.002	0.815
Aburi_salmon_don(Main_Meal)	Kaisendon(Main_Meal)	-0.003	0.656
Katsudon(Main_Meal)	Kaisendon(Main_Meal)	-0.005	0.593
Katsu_Curry(Main_Meal)	Kaisendon(Main_Meal)	-0.009	0.511

Table 5. MW Non-Sets Analysis Results

Based on the association found, we can make the following recommendations for MW:

1. Provide a set for Ebi Katsu and Hokkaido Scallop
2. Provide a set for Salmon Caviar Onigir and Salmon Onigri

3. Buddy Meals consisting the following:
    - Aburi Salmon Don & Kaisendon
    - Katsudon & Kaisendon
    - Katsu Curry & Kaisendon

Premises	Conclusion	Leverage	Lift
Tuna Mayo Onigiri	Salmon Onigiri	0.0067	4.9548
Konbu Onigiri	Salmon Onigiri	0.0051	5.9488
Negitoro Topping	Kaisendon	0.0034	2.8412
Tonkatsu (Pork Cutlet)	Kaisendon	-0.0007	0.9385
Kakiage	Kaisendon	-0.0010	0.8905
Ebi Katsu	Kaisendon	-0.0016	0.7669
Hokkaido Cheese	Kaisendon	-0.0024	0.7002
Tonjiru	Kaisendon	-0.0029	0.7721
Karaage	Kaisendon	-0.0037	0.6607
AYATAKA	Kaisendon	-0.0042	0.6802
Hokkaido Scallop	Kaisendon	-0.0043	0.6825

Table 6. RP Non-Sets Analysis Results

Based on the association found, we can make the following recommendations for RP:

1. Provide a set for Tuna Mayo Onigri and Salmon Onigri
2. Provide a set for Konbu Onigir and Salmon Onigri
3. Buddy Meals consisting the following:
    - Tonkatsu & Kaisendon

OBSERVATIONS

We noticed varying product focus from the products associated in RP as compared to MW. For example, the salmon onigri has a higher association with the salmon caviar onigri in MW and for RP it has a higher association with the Konbu onigri. A general observation of the support of products purchased by customers in MW shows that the preference of food is generally more expensive and there's more focus at quality of product while for RP there is more focus on sets and value for money. A guess would be that the customers at RP are more likely to be office staffs (RP is closed on weekends) and as for MW, casual shoppers that tend to have a higher spending power.

RECOMMENDATIONS

Ultimately the analysis of the data provides 3 recommendations:

1. Highly associated items should be placed near each other or in a set to drive sales
2. For items frequently bought together, giving a discount for an item will drive the sales of the other significantly.

For the items such as the onigris that we've identified earlier or even some of the fried dishes, by putting them nearby or giving discount to one or the other, an increase in sales can potentially be driven. Similarly, putting these items in a set will see an increase sales volume as well.

3. Avoiding the **Profitable Product Death Spiral** - we should not eliminate unprofitable products that are attracting profitable customers

We see that some of the side dishes that were bunched together with the Kaisendon are rather unprofitable and are also low in sales volume. However these product are attracting customers to purchase the Kaisendon and hence we should not exclude these seemingly unprofitable items.

LIMITATIONS

Market Basket Analysis does not account for the quantity of items in a transaction – this affects the strength of association within products that is not captured.

K-MEANS CLUSTERING

LITERATURE REVIEW

We looked at applications of k-means on analysing and predicting students' performance in school. Islam & Haque used k-means to evaluate students' quiz and examination results so that teachers can be well-informed early on of their students' performance and take appropriate action to improve class performance. Through k-means clustering, Islam & Haque were able to identify 3 groups of students - high, medium and low GPA students. Oyelade, Oladipupo & Obagbuwa, also used k-means clustering to predict students' academic performance. They analysed students' overall performance using k = 3, k = 4, and k = 5.

Being able to use k-means clustering to evaluate students' academic performance would be largely similar to using it to evaluate staff / labour performance. Both forms of evaluation aim to identify individuals who are performing well or poorly.

To know what poor or excellent performance is, a benchmark (KPI) is needed. We intend to identify each sales data point for each individual staff for each hour in each outlet, and peg this to the average hourly sales for the respective outlet. This is so as to normalise sales during peak and non-peak hours. Then, we will use this scoring to cluster individual staff based on k-means.

The next thing would be to determine k. We will use the Average Within Centroid Distance and Davies-Bouldin Index to identify the appropriate k. The Average Within Centroid Distance calculates the average distance between centroids in a cluster. The smaller average distance will be a good evaluation of an appropriate k. As for the Davies-Bouldin Index, a small Davies-Bouldin Index will indicate which an appropriate k to use is.

DATA EXPLORATION

1.2 PEAK HOUR SALES

In exploring the sales-labour data, we first found the mean sales per hour from both MW and RP. It can be seen that the mean sales for MW peak from 11:00 to 13:00 during lunchtime and 18:00 to 20:00 during dinnertime. Likewise, RP's mean sales peak from 11:00 to 13:00, and from 17:00 to 19:00.

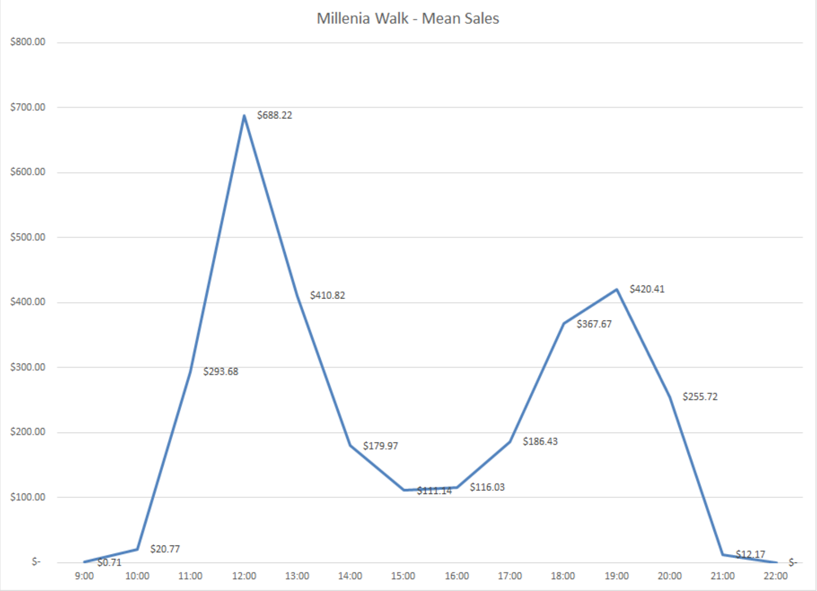


Figure 9. MW Sales over Hours

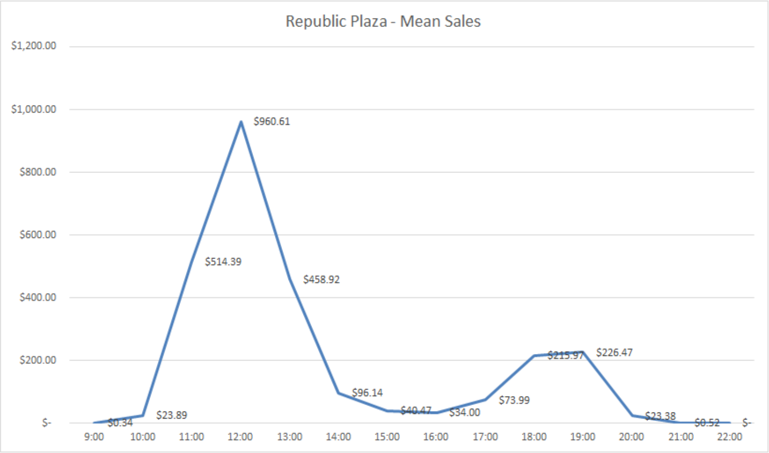


Figure 10. RP Sales over Hours

STAFF PERFORMANCE

We hypothesized that staff who are more productive will perform better on average as compared to the shop sales on an hourly basis. We explored staff performance by looking at a common measure of labour performance, which is staff productivity.

We attributed hourly sales to each of the staff present in the shop at that hour. We then took an average of the attributed sales for each of the staff by dividing his total sales with the total number of hours that he worked.

However, we realized that there was an hourly effect on retail sales, which affects the labour productivity of the staff. This means that on an absolute basis, if Staff A and B both work the same number of hours, but A works during peak hours, and B works during non-peak hours, A’s labour productivity (Store Sales / Number of hours worked) will be higher than B. This might lead to a possible misrepresentation because Staff A might be poorer at customer service or upselling as compared to B. This leads to a need for data standardization on an hourly basis. For more information on the methodology used, please refer to the Data preparation section.

After standardizing the data, we proceeded to rank them based on their standardized labour productivity and took the top 5 performers, as well as the bottom 5 performers for each store, and plotted their hourly sales, compared to the shops’ average sales.

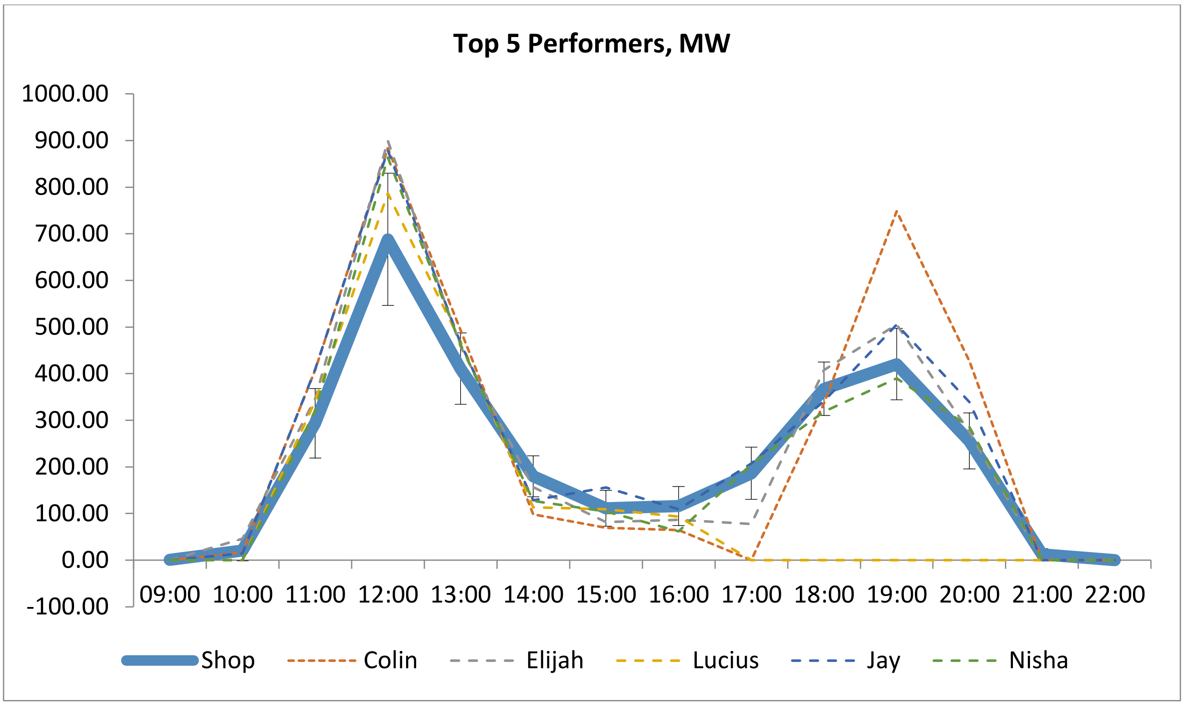


Figure 11. MW Top 5 Performers

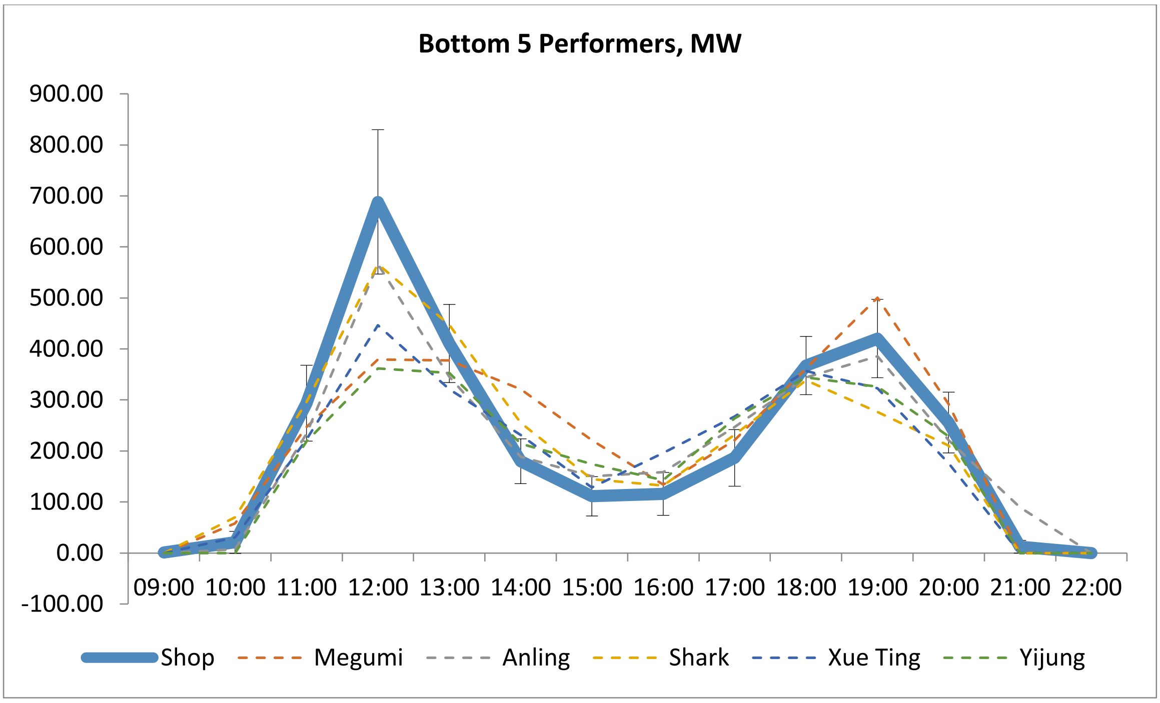


Figure 12. MW Bottom 5 Performers

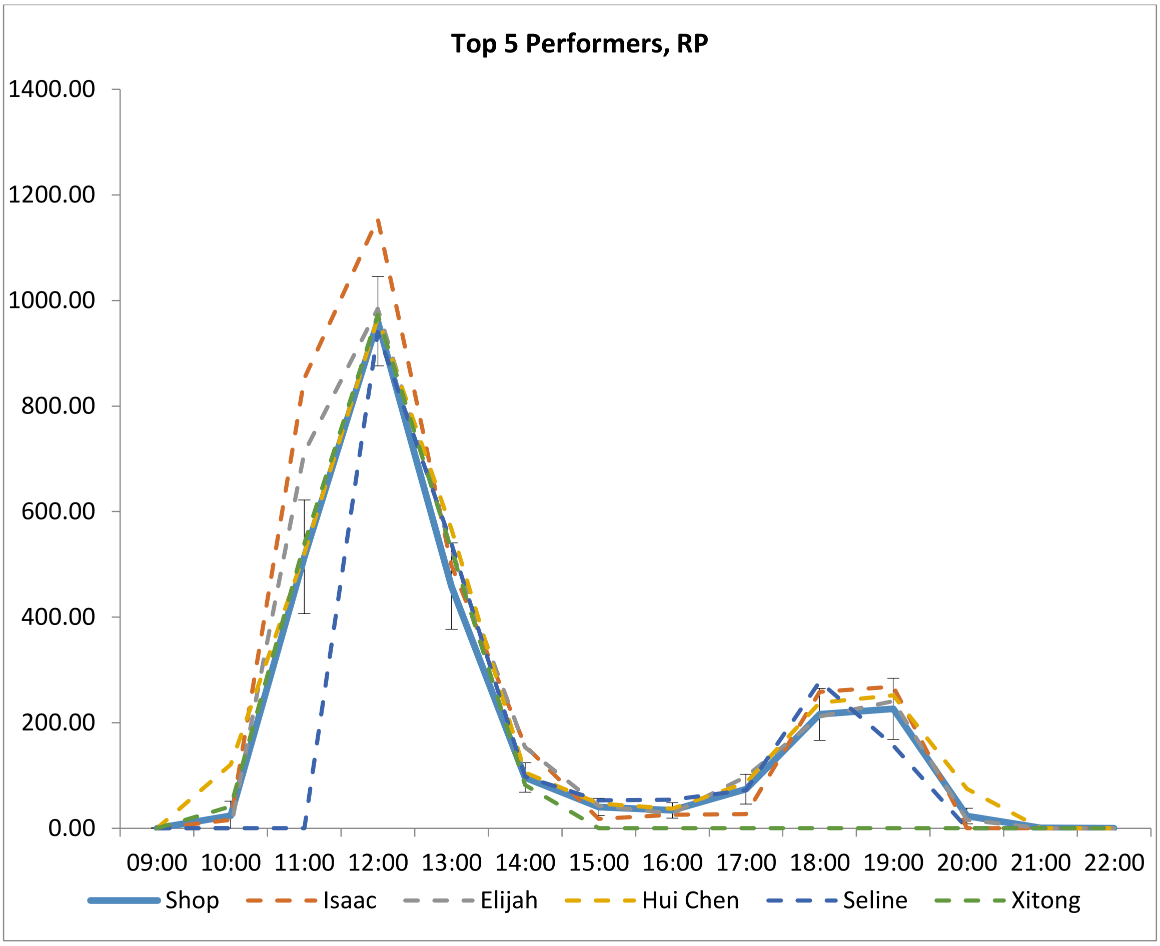


Figure 13. RP Top 5 Performers

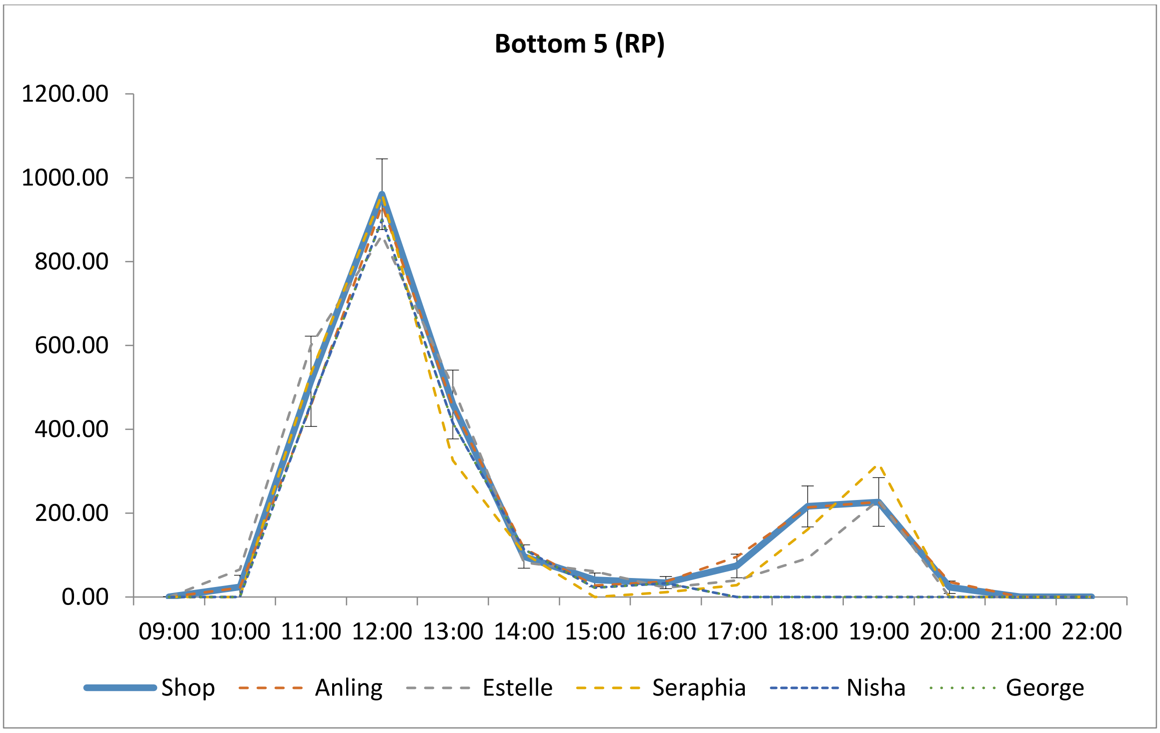


Figure 14. RP Bottom 5 Performers

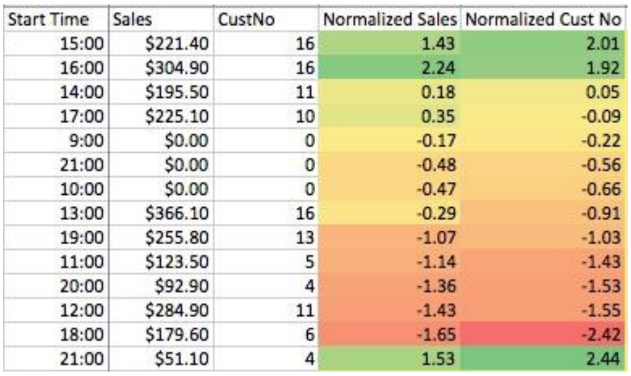
Our hypothesis is partially true because the top 5 performers almost perform higher than the mean shop sales but mostly during peak hours. The bottom 5 performers also perform lower than the mean shop sales but mostly during peak hours.

This implies that there is value in identifying high performers that perform on a consistent basis. Firstly, we can benchmark staff performance using the top performers. Secondly, we can qualitatively assess the behavior of top performers that affect sales and develop means to train the rest of the staff to be like them.

DATA PREPARATION

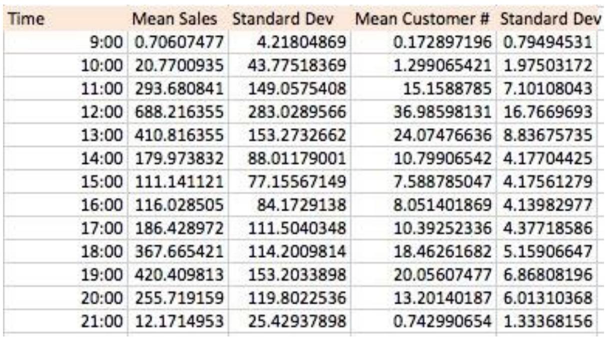
DATA STANDARDIZATION

We standardized the sales and customer numbers on an hourly basis, using hourly means and standard deviations. This allows us to better compare staff performance based on how far is a person's performance from the hourly average. A good staff will have a high positive deviation from the mean average sales, and vice versa.



Start Time	Sales	CustNo	Normalized Sales	Normalized Cust No
15:00	$221.40	16	1.43	2.01
16:00	$304.90	16	2.24	1.92
14:00	$195.50	11	0.18	0.05
17:00	$225.10	10	0.35	-0.09
9:00	$0.00	0	-0.17	-0.22
21:00	$0.00	0	-0.48	-0.56
10:00	$0.00	0	-0.47	-0.66
13:00	$366.10	16	-0.29	-0.91
19:00	$255.80	13	-1.07	-1.03
11:00	$123.50	5	-1.14	-1.43
20:00	$92.90	4	-1.36	-1.53
12:00	$284.90	11	-1.43	-1.55
18:00	$179.60	6	-1.65	-2.42
21:00	$51.10	4	1.53	2.44

Figure 15. Standardized Sales based on Hourly Means and Standard Deviations 1



Time	Mean Sales	Standard Dev	Mean Customer #	Standard Dev
9:00	0.70607477	4.21804869	0.172897196	0.79494531
10:00	20.7700935	43.77518369	1.299065421	1.97503172
11:00	293.680841	149.0575408	15.1588785	7.10108043
12:00	688.216355	283.0289566	36.98598131	16.7669693
13:00	410.816355	153.2732662	24.07476636	8.83675735
14:00	179.973832	88.01179001	10.79906542	4.17704425
15:00	111.141121	77.15567149	7.588785047	4.17561279
16:00	116.028505	84.1729138	8.051401869	4.13982977
17:00	186.428972	111.5040348	10.39252336	4.37718586
18:00	367.665421	114.2009814	18.46261682	5.15906647
19:00	420.409813	153.2033898	20.05607477	6.86808196
20:00	255.719159	119.8022536	13.20140187	6.01310368
21:00	12.1714953	25.42937898	0.742990654	1.33368156

Figure 16. Standardized Sales based on Hourly Means and Standard Deviations 2

Another benefit of standardization is the ability to remove outliers in the data. We define outliers as data points that have sales or customer numbers that are more than three standard deviations from the mean in both directions. This will prevent an underperformer from being shown to have a high performance due to an anomaly, and vice versa.

WEEKENDS AND PUBLIC HOLIDAYS FOR RP

In the data given for RP, public holiday and weekend information was included with blank spaces even though the store was not open on those days. As these values affected the store mean and standard deviation, there was a need to remove these data points. We identified the public holidays from the Ministry of Manpower's website for the time period and removed them, as well as the weekends from RP's dataset.

DATA ANALYSIS METHODOLOGY

DATA ANALYSIS METHOD: K-MEANS CLUSTERING

Our objective is to help the Teppei managers to differentiate the workers by performance. A common means of doing so would be to use a bell curve to grade the workers. This approach, which is most commonly used in education is a statistical method of assigning grades designed to yield a pre-determined distribution of grades among the students in a class. While bell-curving is a convenient approach to measure employees against other employees, it requires the manager to impose arbitrary benchmarks that may be too rigid at times. For example, if a manager classifies top performers as staff who fall within the top 25 percentile, he may risk excluding staff who perform much better than the rest of the staff but fall right outside the top 25 percentile.

We propose the use of clustering to differentiate staff by performance for three reasons. Firstly, clustering will allow a more dynamic way of classifying workers by performance. As staff are grouped according to their attributes, staff who perform similarly will be grouped together instead of being separated by arbitrary performance benchmarks. Secondly, clustering is useful when there are multiple types of attributes in a dataset, making it difficult to compare data points (Morissette & Sylvain, 2013). In our case where staff can be graded on their ability to bring in more customers or more sales, clustering will be able to differentiate staff who perform better in the form of bringing in more customers, and staff who perform better in the form of bringing in more sales. This is not possible in a single attribute bell-curve. Lastly, clustering analysis gives the manager a good balance of flexibility and rigor in his analysis by allowing him to control the desired number of groups, and at the same time finding the optimal groupings of staff who perform similarly. This is a more valuable alternative as compared to trying to visualize groups of staff using a scatter plot.

There are multiple ways to use clustering. We chose to use K-means clustering as it is more applicable in this scenario for two reasons. Not only does it allow managers to explore different numbers (K-values) of ways to cluster their staff, it also allows them to fix an optimal number of clusters and implement an incentive program that is tied to each cluster.

K-means clustering analyses the Euclidean distance between data points and form K clusters of data points. It first assigns K number of data points to become centroids and calculate the Euclidean distance from the rest of the points (non-centroids) to these centroids. It then assigns non-centroids to be clustered with the nearest centroids. The process is re-iterated for a fixed number of times for K clusters to try to maximise the inter-cluster distance (distance between each of the K centroids) and the intra-cluster distance (distance between non-centroids and their cluster centroids).

ANALYSIS TOOL: RAPID MINER

The client has specified that they would like to execute or modify the analysis methodology and therefore preferred if we used an open-source software that requires minimal coding.

We chose Rapid Miner over SAS, JMP and R because it is open-source, user-friendly to managers who have no technical background and has a wide range of functions, including K-means clustering.

DATA ANALYSIS MEASURES: SILHOUETTE INDEX AND THE DAVIES-BOULDIN INDEX

In order to perform K-means cluster, the manager has to set the optimal K number of clusters before he proceeds with his analysis. Even if the manager has a K value in mind, he is encouraged to use cluster performance measures to measure the quality of the clusters that are formed with respect to different K values.

We have selected two different Data Analysis measures that are commonly used for K-means clustering. We chose to use two measures instead of one to test the consistency of both measures and increase the rigor of our evaluation. Ideally, the optimal number of clusters should score the highest on both indexes.

The Silhouette Index evaluates the consistency within clusters of data by measuring the similarity of each data point to the data points within and without its cluster. If a cluster's Silhouette Index, which is the average Silhouette score of all its data points, is high, it means that its data points are very similar to each other, and dissimilar to other data points in other clusters, and vice versa.

The Davies-Bouldin index (DB) measures how good the clustering scheme is by taking the ratio of the average scatter within a cluster to the average distance between clusters (in this case, Euclidean distance). If a cluster's Davies-Bouldin index is low, it means that its data points are tightly clustered relative to their distance from the other clusters.

To select an optimal K value, a manager should plot the K-values with respect to the Silhouette Index and the Davies-Bouldin index. The K value with the highest Silhouette index and the lowest Davies-Bouldin index will yield the most optimal clusters.

DATA ANALYSIS RESULTS

SELECTING THE K-VALUE

It would be intuitive for a manager to set K = 2 or 3, in order to differentiate between the good, average and poor performers. However, we tested a range of K = 2 to K =5 so as to fully utilize the ability of the clustering analysis method to achieve a deeper level of differentiation. We did not go higher than K = 5 as we recognized that the benefits that ensue from marginally better clusters will not justify the diseconomies of scale in the form of a complicated incentive program for too many groups.

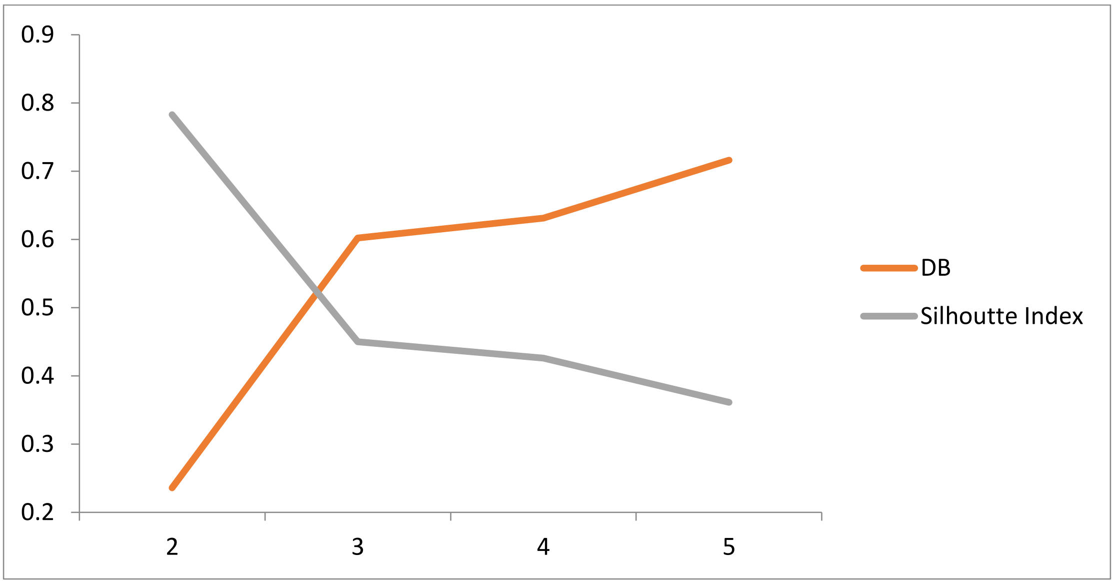


Figure 17. MW Cluster Performance Measures

K	DB	Silhouette Index
2	0.236	0.783
3	0.602	0.45
4	0.631	0.426
5	0.716	0.361

Figure 18. MW Cluster Performance Measures

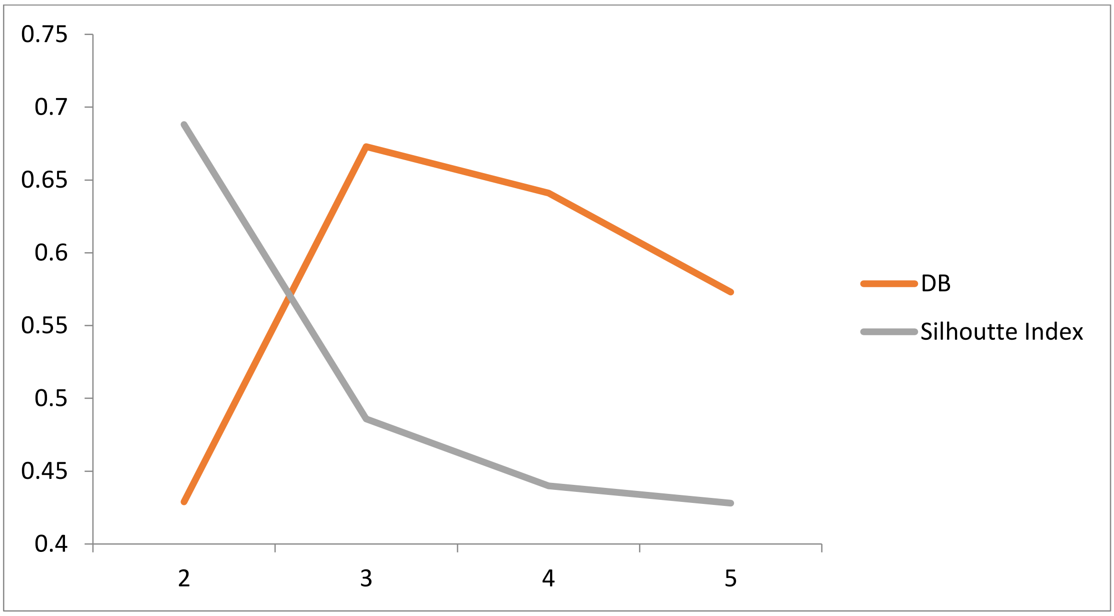


Figure 19. RP Performance Measure

K	DB	Silhouette Index
2	0.429	0.688
3	0.673	0.486
4	0.641	0.44
5	0.573	0.428

Figure 20. RP Performance Measures

It is clear that the Silhouette index decreases for both stores, as the number of clusters increase. DB however, shows different results for MW and RP. DB increases as K increases for MW. For RP, DB increases as K increases to K = 3, and then decreases thereafter.

Based on the definitions of both performance measures, it is clear that K =2 is the most optimal, with the smallest DB and the largest Silhouette index value. The second most optimal K is K = 3 for MW and K = 5 for RP.

CLUSTERING LABOUR PERFORMANCE ACCORDING TO STANDARDIZED SALES AND CUSTOMER NUMBER

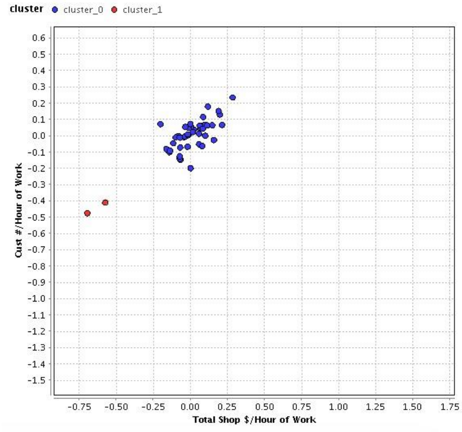


Figure 21. MW Clustering Results for k=2

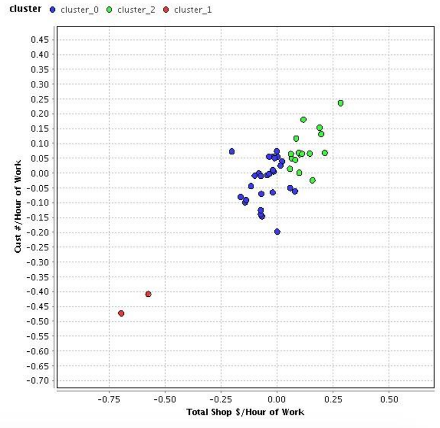


Figure 22. MW Clustering Results for k=3

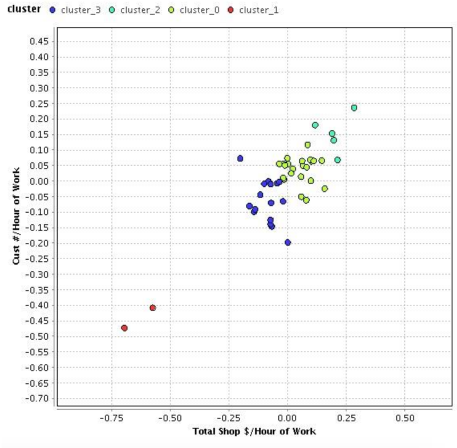


Figure 23. MW Clustering Results for k=4

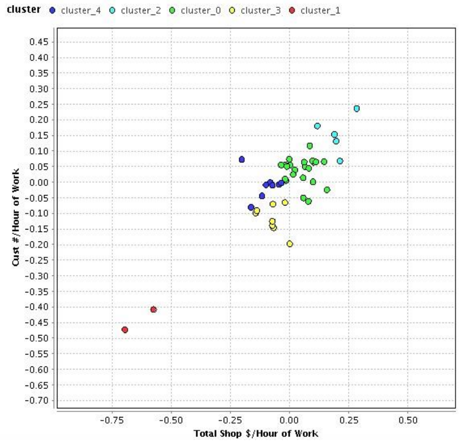


Figure 24. MW Clustering Results for k=5

With respect to the Cluster Analysis results, there are a few insights that can be drawn.

Firstly, based on both performance measures, K = 2 seems to be the optimal number of clusters. However, it does not have any useful business implications. The analysis shows that there seem to be two outliers who perform much worse than the rest of the group in terms of customer numbers and shop sales.

The two data points belong to Seraphina and Motoki, who joined Teppei on the 9th and 26th of December respectively and might be performing lower because of their lack of experience.

When K = 3, there seems to be a distinction between good performers and bad performers, with almost all of the good performers performing above shop average for Sales and Customer count.

When K =4, there seems to be a distinction between average performers and bad performers, with majority of the average performers in cluster_0 performing around the shop average for Sales and customer count, whereas the bad performers perform below shop average for both attributes.

When K = 5, not much more insights are being generated.

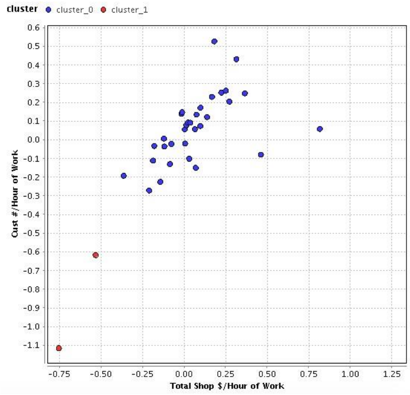


Figure 25. RP's clustering results for k=2

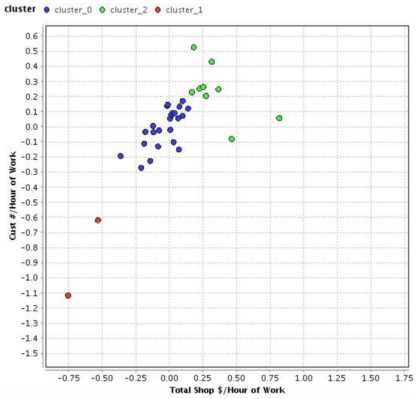


Figure 26. RP's clustering results for k=3

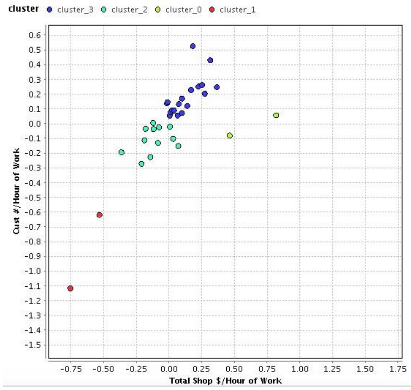


Figure 27. RP's clustering results for k=4

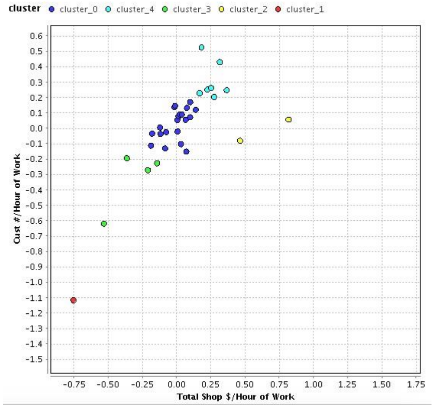


Figure 28. RP's clustering results for k=5

For RP's clustering results, the same goes for K =2, where there is not much business implications, other than the fact that there are two outliers, Megumi and Irwin.

In this case, Megumi is a veteran who has been working in MW since June, but recently got transferred to RP and spent a couple of days working there. Irwin is a newcomer who just started work at RP. This implies that transfers between stores might affect staff performance too.

For K = 3, although there is a distinction between good and bad performers, it is clear that there can be deeper differentiation as the good performers are separated between staff who sell to a lot more customers, and staff who sell to less customers but earn more revenue.

For K = 4, the differentiation can be seen between staff who are low performers who perform below the shop average sales and customer count (cluster_2) Furthermore, there is a difference between high performers who sell to more customers but is not the most productive with respect to sales (cluster_3) and high performers who is the most productive with respect to sales but not customer count. Managers can utilise this information to develop training programs for each high performer group to strengthen where they are weaker at.

When K = 5, there is further differentiation between the average performing (cluster_0) and the low performing (cluster_3). Managers can utilise this information to study the low performers and find out why they are performing poorly.

RECOMMENDATION

The clustering results clearly suggest the benefits of segmenting staff by their performance using the clustering method.

Teppei managers can utilise the results from K = 4 to identify the good performers and study their qualities. Training programs can also be developed specifically for each group to strengthen their weaknesses.

REFERENCES

Islam, H., & Haque, M. (2012). An approach of improving Student’s academic performance by using k-means clustering algorithm and decision tree. *International Journal of Advanced Computer Science and Applications*, *3*(8), 146–149. doi:10.14569/ijacsa.2012.030824

Morissette, L., Chartier, Sylvain. (2013). The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology, 9*(1), 15-24.

Oyelade, O. J., Oladipupo, O. O., Obagbuwa, I. C. (2010). Application of k-Means Clustering algorithm for prediction of Students’ Academic Performance. *International Journal of Computer Science and Information Security, 1*, 292-295.

Haughton, D., Deichmann, J., Eshghi, A., Sayek, S., Teebagy, N., & Topi, H. (2003). A review of software packages for data mining. The American Statistician, 57(4), 290-309.

Tan, Pang-Ning; Michael, Steinbach; Kumar, Vipin (2005). "Chapter 6. Association Analysis: Basic Concepts and Algorithms". Introduction to Data Mining. Addison-Wesley. ISBN 0-321-32136-7.

G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In G. Piatetsky-Shapiro and W. Frawley, editors, *Knowledge Discovery in Databases*, pages 229-248. MIT Press, Cambridge, MA, 1991.