In our paper we will analyze a data set which contains transactional data of an unknown store. We will be implementing market basket analysis techniques. Our paper will be a demonstration of the use of market basket analysis and the implementation of report writing in the CRISP-DM process. After reading this paper and the corresponding learning module the student should be able to write a cross industry data mining report using both market basket analysis and predictive modelling techniques. The two data sets we will be working with are a transactional data set and a bank marketing data set by using market basket analysis and predictive modelling respectively. In this paper we use the subheading “Data Processing”, which consists of data understanding and preparation.
# Table of Contents

Market Basket Analysis: ........................................................................................................... 2  
Business Understanding: ....................................................................................................... 2  
Data Processing: ..................................................................................................................... 2  
Data Modelling.......................................................................................................................... 3  

Sequence Analysis..................................................................................................................... 7  
  Confidence ................................................................................................................................. 8  
  Lift ........................................................................................................................................... 8  
  Support ..................................................................................................................................... 9  

Concluding Remarks: ............................................................................................................... 9  
  Suggestions ............................................................................................................................... 9
Market Basket Analysis:

Business Understanding:

The first phase of the CRISP-DM process is the business understanding phase. In this phase we aim to define the business (in this case the “store”), and the business objectives.

In our data mining analysis we set out to analyze the transactions of an unknown grocery store to discover any relationships and associations in our data. Particularly, we want to know what products are more likely to be bought together, the occurrence of multiple product purchases and the correlation between products.

By discovering associations in the products begin purchased we are able to make suggestions that could be implemented in the store to meet specific store objectives specified by the client ie; Store Manager. We understand that the objectives of the store are to increase revenue, and better customer service therefore our interpretations, evaluations and suggestions will be aimed at satisfying that condition.

Data Processing:

The second phase of our cross industry data mining process is the data understanding phase. We begin by creating a new project in SAS Enterprise miner, creating a new library and defining our data source. It is important to note that since we are dealing with a transactional data set we change the data set role to transaction and the variable roles as shown below.

The variables of “customerpurchases.sas7bdat” data set are represented in the following table:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model Role</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product</td>
<td>Target</td>
<td>Product purchased.</td>
</tr>
<tr>
<td>Time</td>
<td>Rejected</td>
<td>Sequence in which products are purchased.</td>
</tr>
<tr>
<td>Customer</td>
<td>ID</td>
<td>Customer Identification.</td>
</tr>
</tbody>
</table>

We reject the time variable as it has no use in the association analysis. This will later be modified in order to implement a sequence analysis, but for the time being we reject the variable. Customer is set to ID and our target variable is the products purchased.

By using the explore feature in SAS Enterprise Miner we can illustrate certain key aspects and characteristics of the data which will help us further in our data processing. The output presents the following results:

- There are 3 variables in our data and 7007 customer transactions.
- We have no missing values in the data set which makes our data processing job much easier because we will not need to filter the data in order to neutralize those values.
- We have 20 different products that are being recorded.
Data Modelling

Our next task is to create a new analysis diagram in order to perform an association analysis on the data, as shown in figure 1. Since we have already created our data source we can drag the transaction data set to our diagram. We then will insert an association node into our diagram as well. We create a process flow by linking the data node and the association node. Examining the association node and its properties panel we set it to export rule by ID. We must note that this step is very important in order for the Rule-by-ID data to be exported from the node and so that the rule description table will be available for display in the results window, as shown in figure 2.
After this is done we run the diagram from the association node and view the results. This produces various outputs, but we are mainly interested in the table of association rules as shown in figure 3.
We notice all the rules have quite strong measures. All the association rules produced are at least 5.11 times more likely to happen than a customer chosen at random, with the rule *sardines & apples → peppers & avocado* having the highest lift of 5.67. All the rules produced indicate a strong positive correlation.

We also notice that the support values for the rules produced are very low with the highest value being 11.59% and the lowest 8.99%. This is unfortunate as it indicates that even though the rules are strong and products within each rule are positively correlated it is unlikely that all 4 products in each rule will be bought at the same time, so it will be hard to make progress in satisfying our business objectives.

By selecting the association node and modifying the sort criterion in the property panel we can adjust our rules to be sorted by highest support in order to examine which products are highly likely to be purchased together. This is shown in the following figure.
We notice that the rules Heineken → cracker and cracker → Heineken have the highest support values of 36.56% indicating that the it is highly likely that a customer will purchase Heineken and Crackers together. In other words, 36.56% of customers purchased Heineken and crackers. By examining the confidence values we notice that 75% of customers who bought crackers also bought Heineken but only 61% of customers who bought Heineken bought crackers.

Similarly rules 5 and 6 have the same lift and support values. Looking at the confidence values we notice that a higher percentage of customers who bought baguettes also bought Heineken than the other way around. In comparison with our first comparison of rules 1 and 2 we can make the decision that rule 2 is the most interesting rule.

The next set of rules that we find interesting are rules 7 and 8 (soda/Heineken). Similarly to what we did before, comparing the confidence values we choose rule 7 (soda→Heineken) is the more interesting rule as it has the higher confidence value.

Another interesting rule is soda → cracker. This is actually the rule with the highest lift indicating the highest correlation. Again we compared it with the rule cracker → soda, and by examining the confidence of each we determined that soda → cracker was the more likely rule.
In order to perform sequence analysis we attach another association node to the dataset. It is important to note that we renamed the node to “sequence” for clarity.

In the association analysis we chose to ignore the time variable. In the sequence analysis we do not reject the time variable; instead we use it and set its role to sequence. Similarly to the association analysis we attached a second association node to the dataset in order to run a sequence analysis, this time using the time variable. The following figure is the sequence report produced with all the sequence rules:
We notice that the rules change after including the sequence variable. This indicates to us the importance of the sequence variable in deriving more accurate rules.

Confidence

Immediately, 3 rules stand out in our sequence analysis. By looking at the Confidence values we notice the three extremely high values.

- **Soda → cracker → Heineken**, with a confidence value of 99.09%. In other words, 99.09% of customers who bought soda bought crackers and customers who bought crackers bought Heineken.

- **Baguette → herring → Heineken**, with a confidence of 95%. In other words, 95% of customers who bought baguettes bought herring and bought Heineken.

- **Avocado → artichoke → Heineken**, with a confidence of 94.69%

Lift
The 3 rules with the highest lift values indicating rule strength are:

- Coke $\rightarrow$ ice cream, with a lift value of 2.34.
- Avocado $\rightarrow$ artichoke, with a lift of 1.87.
- Herring $\rightarrow$ corned_b $\rightarrow$ olives with a lift of 1.82.

**Support**

The 3 rules with the highest support, which means that they are the most occurring rules in the dataset:

- Cracker $\rightarrow$ Heineken with a value of 33.67%.
- Herring $\rightarrow$ Heineken with a value of 23.48%.
- Olives $\rightarrow$ Bourbon with a value of 23.28%.

**Concluding Remarks:**

Based on our analysis several key characteristics of our data can be identified which allow us to make suggestions that would meet business objectives.

- 99.09% of customers who bought soda, bought crackers and Heineken.
- 95% of customers who bought baguettes also bought herring and Heineken.
- 94.69% of customers who bought avocado also bought artichoke and Heineken.

We also identified products with the highest correlation:

- A customer who bought coke and ice cream was 2.34 times more likely to do so than a customer chosen at random.
- A customer who bought avocado and artichoke was 1.87 times more likely to do so than a customer chosen at random.
- A customer who bought herring, corned beef and olives was 1.82 times more likely to do so than a customer chosen at random.

We also identified most occurring purchases:

- 33.67% of the customers bought crackers and Heineken.
- 23.48% of the customers bought herring and Heineken.
- And 23.28% of the customers bought olives and bourbon.

**Suggestions**

We can make the following suggestions:

1. Bundle soda and crackers in order to increase the occurrence (support of the rule) of customers who bought soda, crackers and Heineken.
2. We could also bundle baguettes and herring in order to increase the occurrence of baguette, herring and Heineken sales.
3. By placing avocados, artichoke and Heineken in close proximity we can increase the sale of all 3 items together.
4. Place soda, crackers, baguettes, herring, avocados and artichoke in the same aisle as Heineken.
5. Place coke and ice cream in close proximity to each other.
6. Place olives and bourbon near each other for better customer service.

Please note that these are only suggestions and the deployment of our concluding results is determined by the client (store manager).