Data Analysis, is all about trying different approaches and selecting which approach is the closest fit to answering your problems.

Firstly, studying similar problems to understand the spectrum of solutions that the paper can focus on, which would involve conducting an exhaustive literature review. After which the discusses about the data at hand identifying underlying relationships and trends in the data that would help a retailer understand the problem much better and further point them in the right direction.

This could be done by data visualization by using various tools such as R

Once the intricacies of the data are well understood, the process would start by developing a model that would solve the problem. After the development of the model the next process would be to implement this model using the R interface and test its performance on a small section of the data. If the model meets the required evaluating parameters it would be replicated throughout the data.

To summarize these steps, they can be divided into different phases as the paper progresses with the project

* Analyzing the data/Exploratory Analysis
* Model design
* Model testing using an interface such as R
* Model Implementation

# **Scope, Assumptions and Limitations**

The scope of the project is to apply principles of data analytics for studying the behavior of the “grocery on the go” customers. The study aims to understand the shopping nature of the customers by analyzing the underlying data patterns. In the study, the paper focuses on finding the association among the products bought by the customers and thereby helping the retailers to optimize the cross-selling feature.

**Assumption**

The underlying assumption in market basket analysis is that joint occurrence of two or more products in most baskets imply that these products are complements in purchase, therefore, purchase of one will lead to purchase of others. In the present study, the paper discusses into two item rules.

**Limitation**

* The analysis is based on the anonymized dataset provided by the corresponding company and the paper is not checking the validity of the data.
* The paper is not performing any kind of survey among the users to have the results verified.

# **Literature Review**

**Market Basket Analysis (MBA)**

Market basket analysis has been a buzz word in retail space for a very long time and recently its relevance has increasingly become popular in online retail space. Market basket analysis is a data mining technique commonly used in marketing and retailing to help the retailers understand the consumers purchase behavior. The term market basket analysis is named after the whole idea of a customer putting the purchased items into a shopping cart for the duration of the shopping (Vats, 2015). Understanding the purchase behavior helps organizations to make better business decisions. Organizations use these information for attracting more customers and increasing the overall sales by performing cross selling, upselling, designing store layout and in creating promotional offers. Analyzing a customer’s shopping behavior and deriving rules for creating predictive systems, which can help retailers in determining the cross-selling strategies and thereby increasing the overall sales is the main idea behind the MBA technique (Decker & Monien, 2003). The technique uses an assumption that the joint occurrence of two or more products in majority of the baskets implies the correlation between the items and hence gives a high probability of being bought together (Kamakura, 2012). Market basket analysis can be considered as an application of data mining technique**.** Usage of Market basket analysis originated in the field of marketing but it’s been effectively used in other areas such as bioinformatics, education, nuclear science, immunology and geophysics. Market basket technique can be grouped as an application of data mining technique. Data mining also known as knowledge discovery is the process of gathering useful information from large amounts of data following a step wise process of data selection, data cleaning, applying data mining techniques and interpretation of results. (Kaur & Kang, 2016). Market basket analysis can be performed by two alternative methods such as Association Rule mining and Time series Clustering.

**Association Rule Mining**

Association analysis is an unsupervised data mining technique which can be used in identifying relationship between the entities in a large data set. The output of the analysis will be rules that can be used to identify hidden relationship between the items in the data set. From the association rules the set of frequent items in a transaction can be identified. The technique is widely used for performing market basket analysis due to its ability in determining the relationship between items in the transaction data.

The techniques of association rule mining start by identification of most frequent item sets from the transaction dataset. Items are the base objects on which the analysis process is carried out. Transactions indicate the co-occurrence of a group of items together (Yali, 2012). Using the frequent item sets as the starting point, association rules are derived. The output of the market basket analysis will be a set of rules which can be used to predict the output of a customer purchase. Apriori algorithm is the most commonly used algorithm to perform the association analysis. Since the paper explains the basics of association rule, its mandatory to introduce the terms Antecedent and Resultant. For example, in an association rule A🡪B, A is the antecedent and B is the resultant. Few other terms that are commonly used in the explaining the efficiency of association rules are Confidence, Support and Lift ratio. Support value for a rule indicates the impact of a rule in the entire data set. For example, if we say the support for a rule “If Milk and Sugar then Butter” is 60%. It means that this rule affects 60% of the total data set. If the support value of the rule is less, then it indicates that the effect of the rule on the data set is considerably less. Support is a key measure in determining the rule as a low value of support means that the rule occurred just by chance. Considering the business perspective, it is not profitable to promote items that are rarely bought by the customers. Hence, the support measure is often used as a filter in removing irrelevant rules. Confidence of a rule can be defined as the likelihood of the occurrence of two items together. For example, confidence of the rule A🡪B can be defined as the probability of occurrence of A&B together to the probability of occurrence of A in the transaction. Hence confidence can be considered as a measure which determines the relevance of a rule. The lift ratio indicates the likelihood of an item Y to be purchased with the purchase of item X considering the popularity of item Y. Lift provides inference as what is the chance of an item, consider Y, to be purchased with another item (X) with respect to general buying rate of Y. Rules with value of lift ratio greater than one is considered as relevant (Gupta, Kumar, & Shaikh, 2015). The results of the association rule mining are sorted and selected with much caution using the measures of support, confidence and lift to ascertain the reliability of the rule in predicting the customer behavior. The process of association rule mining can be split into two major steps or subtasks as determining the frequently bought items and rule generation. From the first step, as the name infers, the frequently bought item sets are retrieved. In the second step using these frequent item sets high confidence rules known as strong rules are extracted.