

Ministry of high education,

Culture and science city at Oct 6,

The High institute of computer Science & information systems



المعهد العالي لعلوم الحاسب ونظم المعلومات

Graduation Project:

Data analytics (mining) for Market basket analysis.

Prepared by

41171- كمال الدين برعي يوسف

41160- عمر جلال رمضان علي

41192- محمد حمدي أحمد حسين

41178- ماجد عفيفي ابوالعلا عفيفي

41235- محمود عبدالباسط محمود

41201- محمد صبحي عبدالغني دمرdash

41247- مروان إسماعيل جودة

41236- محمود عبدالله عبدالخالق عبدالله

42163- مصطفى محمود محمد محمد

Supervised by

Prof.Dr: Mohammed Khafagy

Assistant:

Andrew Nader

Project No: 33

Academic Year :2019/2018

Contents

1 INTRODUCTION.....	9
1.1 What is data mining.....	10
1.2 Data mining steps.....	11
1.3 What Kind of Data	11
1.4 Data Mining Functionalities—What Kinds of Patterns Can Be Mined	12
1.5.4 Association Rules.....	16
1.7 REQUIREMENT	16
1.7.1 Software.....	16
1.7.2 Hardware.....	16
2 DATA PREPARATION AND PREPROCESSING	17
2.1 Business Understanding	18
2.2 Data understanding	18
2.2.1 Data Understanding: Quantity.....	18
2.4.1 Data preparation.....	20
2.4.2 Data Preparation methods	20
2.4.3 Data Normalization	20
2.4.4 Removing outliers	21
2.5 Data Preprocessing	22
2.5.1 Why Preprocess the Data	22
2.5.2 Why Is Data Preprocessing Important.....	22
2.6 Multi-Dimensional Measure of Data Quality	23
2.7 Data Cleaning Steps	23
2.7.1 Missing Data.	23
2.7.2 Dealing with missing values	24
2.7.3 Noisy Data	24
2.7.4 Source of Noisy data	25
2.7.5 How to handle noisy data	25
2.8 Inconsistent Data.....	26

2.9 Data Integration:	28
2.9.1 Data Integration	28
2.10 Data Reduction	29
3 Modeling	31
3.1 Modeling	32
3.2 Clustering Algorithms.....	33
3.2.1 What is Clustering.....	33
3.2.2 The Goal of clustering	34
3.2.3 Requirements	34
3.4 The classification algorithm used in the project.....	34
3.4.1 K-means clustering Algorithm.....	34
3.4.2 goals of this algorithm	35
3.4.3 Key Concepts	35
3.4.4 Cluster Inertia.....	36
3.5 Algorithm Steps	36
3.6 K-Means Hyperparameters	38
3.7 Challeges of K-Means.....	38
3.8 Points to be Considered When Applying K-Means.....	38
3.9 How to Choose the Right K Number	38
3.10 Elbow Method	39
3.11 K-Means Limitations	40
3.12 RESEARCH METHODOLOGY	40
3.12.1 Association Rule.....	40
3.12.2 Definition of Association Rule	41
3.12.3 Process of Association Rule	44
3.13 We can list the associations for two algorithms as follows.	45
3.13.1 Apriori.....	45
3.13.2 carma.....	45
4 IMPELIMENTIONS	47
4.1 Data Preprocessing.....	48
Step 1.....	48
Step 2.....	48
Step 3:.....	49

Step 4.....	49
4.2 Implement k-means algorithm	50
Step 1	50
Step 2	50
Step 3	51
Step 4	51
Neural Network result.	53
Step 5	53
Step 6	54
Step 7	55
Step 8	56
4.3 Data analysis and results	57
4.3.1 SPSS Modeling Process.....	57
Step 1.....	57
Step 2.....	57
Step 3.....	58
Step 4.....	59
4.3.2 Results	59
4.3.3 Most Interesting Association Rules.....	61
6 REFERENCE	62

Contents Figure

Figure 1 data mining as a step the process of knowledge discovery.....	10
Figure 2. data preparation	20
Figure 3 data preprocessing	22
Figure 4 Inconsistent Data	26
Figure 5 Data Integration	29
Figure 6 Clustering Algorithms.....	33
Figure 7 K -means clustering algorithm flowchart	36
Figure 8 K-Means Hyperparameters.....	37
Figure 9 Elbow Method	39
Figure 10 K-Means Limitations	40

ACKNOWLEDGMENT

In the performance of our mission, we had to get the help and direction of some respectable people, who deserved our greatest gratitude. Completing this task gives us a lot of fun.

As well as the Dean of the Institute (Sami Abdel Moneim) who gave me the golden opportunity to do this wonderful project on the subject (data mining for market basket analysis)

Also, thank you Professor Mohamed Khafagy who gave us the methodology of work, which was his passion for the permanent impact on the infrastructure.

We would like to express our gratitude to Mr. Andrew Nader, teacher of the course, for the Higher Institute of Computer Science and Information Systems for giving us good guidance for duty during many consultations. We would also like to express our deep gratitude to all those who have addressed this task directly and indirectly

PROJECT OBJECTIVES

Using data mining tools in BI to Understand what items are purchased together and leverage the insights for cross sell, pricing, promotion and merchandizing decisions.

Knowing and analyzing what products people purchase as a group can be very helpful to a retailer in particular or to any other seller in general. A retailer can use this technique to organize and place products frequently sold together into the same area. The Market can use it to determine what new products to offer their customers and evolve schemes to bundle various products and services. The primary objective of Market Basket Analysis is to improve the effectiveness of marketing and sales tactics using the customer data that is accumulated with the enterprise during the sales transaction.

PROJECT SUMMERY

The aim of the project to how can analysis products and Implementing data mining in business Intelligence and using data mining tools to find some relation between products or sort data to groups with a clustering algorithm that to Use in the marketing process and increase profit results to achieve a competitive parity market with these techniques and data mining can be effectively implemented in different business segments by systematically designing data mining modeling and implementing.

Chapter 1

1| INTRODUCTION

1.1| What is data mining?

Data mining is the process of analyzing hidden patterns of data according to different perspectives for categorization into useful information, which is collected and assembled in common areas, such as data warehouses, for efficient analysis, data mining algorithms, facilitating business decision making and other information requirements to ultimately cut costs and increase revenue.

Data mining is also known as data discovery and knowledge discovery.

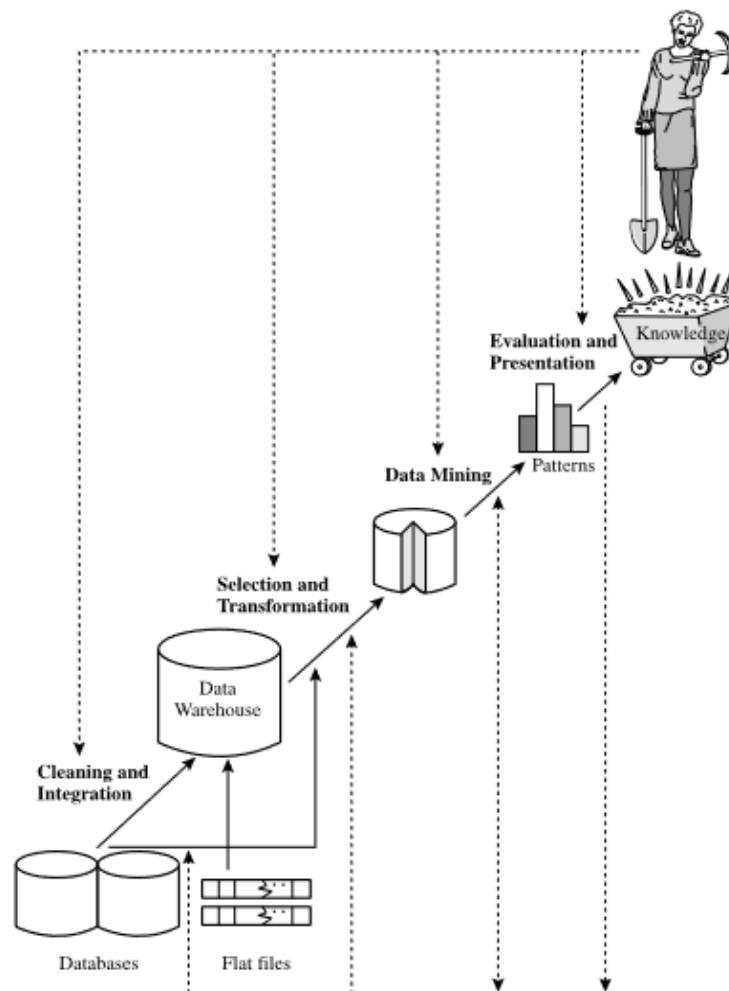


Figure 1 data mining as a step the process of knowledge discovery

1.2| Data mining steps:

1. Data cleaning
2. Data integration
3. Data selection
4. Data transformation
5. Data mining
6. Pattern evaluation
7. Knowledge presentation

1.3| What Kind of Data?

In this section, we examine many different data repositories on which mining can be performed. In principle, data mining should apply to any kind of data repository as well as to transient data such as data streams. Thus, the scope of our examination of data repositories will include relational databases data warehouses, transactional databases, advanced database, systems, flat files, data streams, and the World Wide Web.

1.4| Data Mining Functionalities—What Kinds of Patterns Can Be Mined?

Data mining functionalities specify the kind of patterns to be found in data mining tasks. In general, data mining tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions.

In some cases, users may have no idea regarding what kinds of patterns in their data may be interesting and hence may like to search for several different kinds of patterning parallel. Thus, it is important to have a data mining system that can mine multiple kinds of patterns according to different user expectations and applications. Furthermore, data mining systems should be able to discover patterns at various granularities. Data mining systems should also allow users to specify hints to guide or focus the search for interesting patterns. Because some patterns may not hold for all of the data in the measure of the database of certainty or trustworthiness is usually associated with each discovered pattern. Data mining functionalities and the kinds of patterns they can discover are described below.

Using technology to gain an edge in business is not a new idea. Whenever there is something new entrepreneurs

will be quick to try to find an application for it in the business world to make money. Data mining (DM) and business intelligence (BI) are among the information technology applications that have business value.

Data mining is the process of searching through data using various algorithms to discover patterns and correlations within a database of information. Business intelligence, on the other hand, focuses more on data integration and organization. It will combine data analysis to help managers make operational-tactical or strategic business decisions.

Data mining can be used to aid the objectives of the business intelligence system. Business Intelligence could be an idea of applying a group of technologies to convert information into meaning data. Data processing applied math analysis yet as information visual image. Giant amounts of knowledge| of information originating completely different in several numerous formats and from different sources may be consolidated and regenerate to key business knowledge.

Presents a general read on however information square measure remodeled to business intelligence. The method involves each business consultants and technical consultants. It converts an outsized scale of information to meaning outcomes therefore on offer decision making support to finish users.

Business intelligence (BI) has two basic different meanings related to the use of the term intelligence. The primary less frequently is the human intelligence capacity applied in business affairs/activities. The intelligence of Business is a new field of investigation of the application of human cognitive faculties and artificial intelligence technologies to management and decision support in different business problems.

The second relates to intelligence as information valued for its currency and relevance. It is expert information, knowledge, and technologies efficient in the management of the organizational and individual business. Therefore, in this sense business intelligence is a broad category of applications and technologies for gathering providing access to and analyzing data to help enterprise users make better business decisions.

The term implies having a comprehensive knowledge of all of the factors

that affect the business. It is imperative that firms have in-depth

knowledge about factors such as the customer's competitor's business

partners economic environment and internal operations to make effective

and good quality

1.5 |The Tools technologies used

1.5.1 | SPSS Modeler (IBM.inc Programmer)

is leading visual data science and machine-learning solution It helps enterprises accelerate time to value and achieve desired outcomes by speeding up operational tasks for data scientists. Leading organizations worldwide rely on IBM for data preparation and discovery, predictive analytics, model management and deployment, and machine learning to monetize data assets. SPSS Modeler empowers organizations to tap into data assets and modern applications, with complete algorithms and models that are ready for immediate use. It's suited for hybrid environments to meet robust governance and security requirements and is available in IBM Watson® Studio. SPSS Modeler helps you:

- Take advantage of open source-based innovation, including R or Python
- Empower data scientists of all skills — programmatic and visual
- Exploit a hybrid approach — on-premises and in the public or private cloud
- Start small and scale to an enterprise-wide, governed approach

1.5.2 | Classification (datamining Technique)

This analysis is used to retrieve important and relevant information about data, and metadata. This data mining method helps to classify data in different classes.

1.5.3 | Clustering (datamining Technique)

Clustering analysis is a data mining technique to identify data that are like each other. This process helps to understand the differences and similarities between the data.

1.5.4 | Association Rules (datamining Technique)

This data mining technique helps to find the association between two or more Items. It discovers a hidden pattern in the data set

1.6| What is market basket analysis?

Market basket analysis is a data mining technique that allows us to discover relationships and associations in our data. This technique is commonly used to analysis transactional data sets where we aim to find associations between products purchased together. By performing this method of analysis, we are able to derive associations between products and these associations are called association rules.

1.7| REQUIREMENT

1.7.1| Software

- **Microsoft Excel**
- **IBM SPSS Modeler**

1.7.2| Hardware

- **CPU Core i7**
- **RAM 8 GB**

Chapter 2

2| DATA PREPARATION AND PREPROCESSING

2.1| Business Understanding

In the first phase of a data-mining project, before you approach data or tools, you define what you're out to accomplish and define the reasons for wanting to achieve this goal. The business understanding phase includes four tasks (primary activities, each of which may involve several smaller parts). These are

- ✓ Identifying your business goals
- ✓ Assessing your situation
- ✓ Defining your data-mining goals
- ✓ *Producing your project plan*

2.2|Data understanding

In the first phase of a data-mining project, a data must be is checked very well and the identification and understanding of all components and elements of data as well as understanding structure of data and determine whether this data is good for the work of mining it or there are many problems and the search for other data

2.2.1 | Data Understanding: [Quantity](#)

Number of instances (records)

Rule of thumb: 5,000 or more desired

if less, results are less reliable;

Number of attributes (fields)

Rule of thumb: for each field, 10 or more instances

If more fields, use feature reduction and selection

Number of targets

Rule of thumb: >100 for each class

if very unbalanced, use stratified sampling

2.3| What is Data?

- Collection of data objects “and their” attributes
- An attribute is a property or characteristic of an object
- Examples: eye color of “a Person” and temperature
- An attribute is also known “as variable” field characteristic, or feature
- A collection of attributes describes an object
- An object is also known as a record, point, case, sample, entity, or

Attributes

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

instance.

2.4.1 | Data preparation

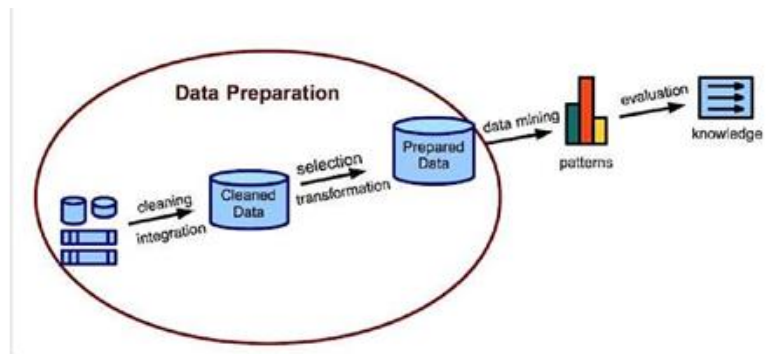


Figure 2. data preparation

- The data preparation normally consumes about 90% of the time.
- The time factor is usually dependent on
 - dimensionality
 - a high number of instances.

2.4.2 | Data Preparation methods

2.4.3 | Data Normalization:

This type of scaling transforms the data into a specific range.

For sequential/temporal data

2.4.4 | Removing outliers:

Outliers are those data points that are inconsistent with the majority of the data points.

	A	B	C	D	E	F
50	93402	32.122	CASH	F	37	whole milk,yogurt,processed cheese,pickled vegetables,soda
51	16869	43.315	CHEQUE	M	35	whole milk,curd,yogurt,pastry
52	12478		CARD	M	34	packaged fruit/vegetables,brown bread,canned beer
53	42015	39.238	CASH	M	25	rolls/buns,oil,bottled water,chewing gum,chocolate marshmallow,hygiene articles,napkins
54	30247	10.62	CARD	F	35	ham,beef,whipped/sour cream,ice cream,rolls/buns,cat food
55	75366	-10.54	CASH	M	47	rolls/buns,pastry,sugar
56	69334	49.178	CASH	F	17	other vegetables,whole milk,frozen vegetables,canned fish,salty snack,seasonal products,detergent
57	3362	-48.4	CARD	Male	29	sausage,pastry
58	88382	15.716	CARD	F	24	sausage,beef,whole milk
59	27995	26.251	CHEQUE	M	22	frankfurter,tropical fruit,rolls/buns,brown bread,sugar
60	28582	45.796	CARD	F	25	rolls/buns,pastry,soda
61	43835		CARD	Female	46	curd cheese,coffee
62	13816	26.348	CARD	F	31	red/blush wine,newspapers
63	1193		CASH	M	42	sausage,whole milk,curd
64	12032	30.426	CARD	M	17	tropical fruit,pip fruit,berries,whole milk,frozen potato products,rolls/buns,pickled vegetables,chocolate
65	11652	42.713	CASH	F	40	red/blush wine
66	76336	28.534	CASH	M	35	whole milk,butter,margarine,specialty fat,specialty chocolate,candles,flower (seeds)
67	19976	-31.2	CASH	M	49	frankfurter,citrus fruit,whole milk,domestic eggs,oil,sparkling wine,specialty chocolate,newspapers
68	83968	46.27	CARD	Male	39	whole milk,meat spreads

2.5 | Data Preprocessing

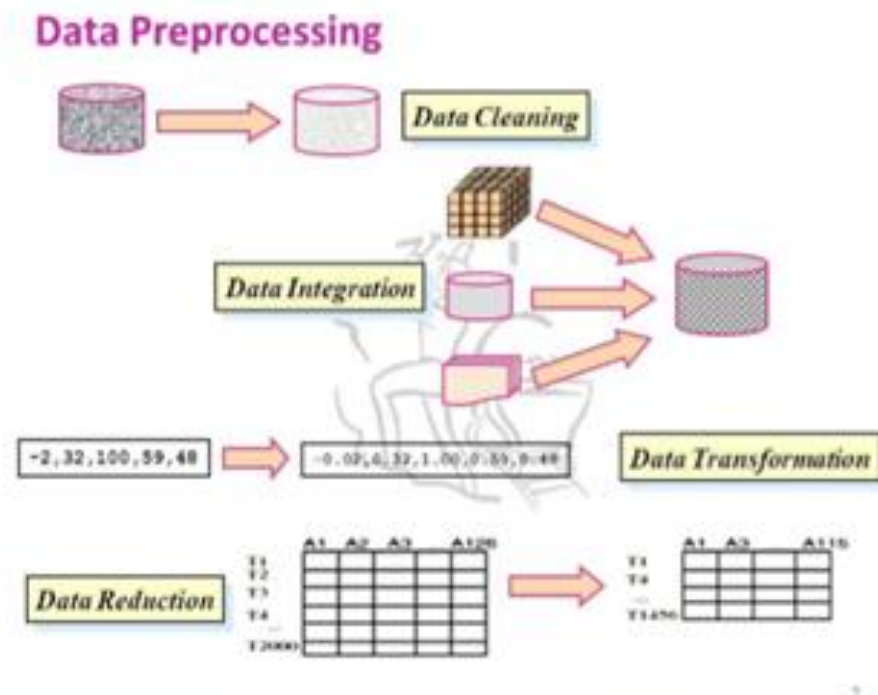


Figure 3 data preprocessing

2.5.1| Why Preprocess the Data?

- Data in the real world is dirty and **incomplete** lacking attribute values, lacking certain attributes of interest, or containing only aggregate data and missing
- Data preprocessing can be separated into two major tasks:
 - Data Cleaning
 - Data Reduction

2.5.2| Why Is Data Preprocessing Important?

- No quality data, no quality mining results
- Quality decisions must be based on quality data duplicate or missing data may cause incorrect or even misleading statistics.

- Data warehouse needs consistent integration of quality data
 - Data extraction, cleaning, and transformation comprises the majority of the work of building a data mining system

2.6 | Multi-Dimensional Measure of Data Quality

- Completeness
- Consistency
- Timeliness
- Value added
- Interpretability
- Accessibility

2.7 | Data Cleaning Steps

Data Cleaning consists with dealing with the following:

- Missing values
- Noisy Data

2.7.1 | Missing Data.

A missing value (M v) is an empty cell in the table that represents a dataset.

	A	B	C	D	E	F
50	93402	32.122	CASH	F	37	whole milk,yogurt,processed cheese,pickled vegetables,soda
51	16869	43.315	CHEQUE	M	35	whole milk,curd,yogurt,pastry
52	12478		CARD	M	34	packaged fruit/vegetables,brown bread,canned beer
53	42015	39.238	CASH	M	25	rolls/buns,oil,bottled water,chewing gum,chocolate marshmallow,hygiene articles,napkins
54	30247	10.62	CARD	F	35	ham,beef,whipped/sour cream,ice cream,rolls/buns,cat food
55	75366	-10.54	CASH	M	47	rolls/buns,pastry,sugar
56	69334	49.178	CASH	F	17	other vegetables,whole milk,frozen vegetables,canned fish,salty snack,seasonal products,detergent
57	3362	-48.4	CARD	Male	29	sausage,pastry
58	88382	15.716	CARD	F	24	sausage,beef,whole milk
59	27995	26.251	CHEQUE	M	22	frankfurter,tropical fruit,rolls/buns,brown bread,sugar
60	28582	45.796	CARD	F	25	rolls/buns,pastry,soda
61	43835		CARD	Female	46	curd cheese,coffee
62	13816	26.348	CARD	F	31	red/blush wine,newspapers
63	1193		CASH	M	42	sausage,whole milk,curd
64	12032	30.426	CARD	M	17	tropical fruit,pip fruit,berries,whole milk,frozen potato products,rolls/buns,pickled vegetables,chocolate
65	11652	42.713	CASH	F	40	red/blush wine
66	76336	28.534	CASH	M	35	whole milk,butter,margarine,specialty fat,specialty chocolate,candles,flower (seeds)
67	19976	-31.2	CASH	M	49	frankfurter,citrus fruit,whole milk,domestic eggs,oil,sparkling wine,specialty chocolate,newspapers
68	83968	46.27	CARD	Male	39	whole milk,meat spreads

2.7.2 | Dealing with missing values

A. Ignore records with missing values:

This is usually done when the class label is missing.

This method is not effective, unless the record contains several attributes with missing values.

B. Fill in the missing value manually:

In general, this approach is time-consuming and maybe not feeble given a large data set with many missing values.

C. Fill in the missing value manually:

Replace all missing values by the same constant such as “unknown”. Although this method is simple but it is not recommended because results with “unknown values are not interesting

2.7.3 | Noisy Data

- Noise is a random error in **measured variable**.
- Noisy data is meaningless data.
- Any data that has been received, stored or changed in such a manner that it **cannot be read or used** by the program that originally created it can be described as noisy.

2.7.4 | Source of Noisy data:

1. Data entry problem.
2. Faulty data collection instruments.
3. Data transmission

2.7.5 | How to handle noisy data?

- Binning method
- Clustering
- Combined computer and human inspections
- Regression

2.8 | Inconsistent Data

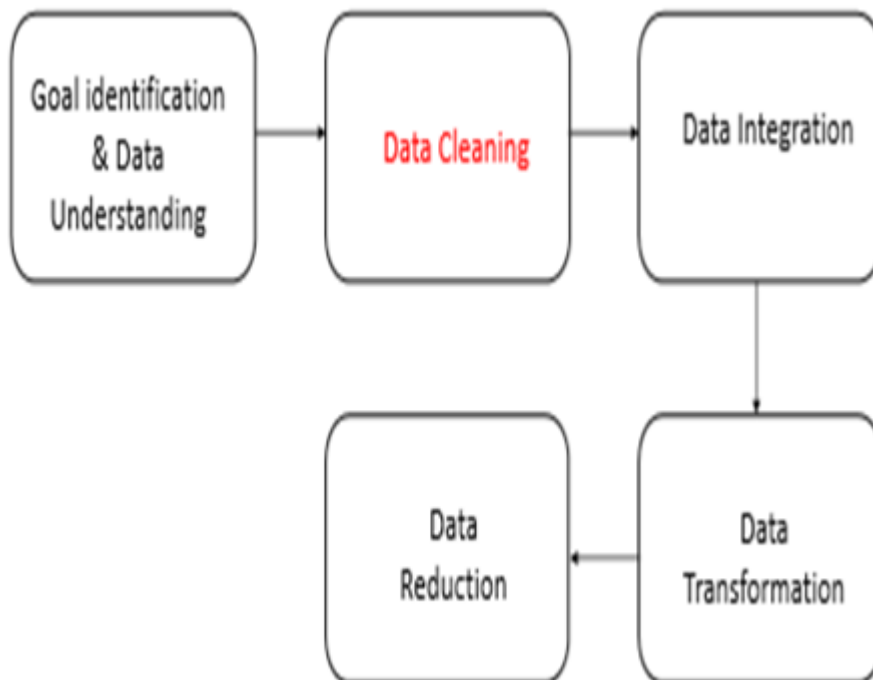


Figure 4 Inconsistent Data

-Data which is inconsistent with our models, should be dealt with.

-Common sense can also be used to detect such kind of inconsistency:

- The same name occurring differently in an application.
- Different names can appear to be the same (Dennis Vs Denis)
- Inappropriate values (Males being pregnant, or having a negative age) Was rating “1,2,3”, now rating “A, B, C”
- Difference between duplicate records
- We want to transform all dates to the same format internally
- Some systems accept dates in many formats
- e.g. “Sep 24, 2003”, 9/24/03, 24.09.03, etc.
- dates are transformed internally to a standard value
- Frequently, just the year (YYYY) is sufficient
- For more details, we may need the month, the day, the hour, etc.
- Representing date as YYYYMM or YYYYMMDD can be OK

Inconsistent Data

A	B	C	D	E	F	G	H	I	J	K	L
Customer	price	Payment	Gender	age	citrus fruit	tropical fru	whole milk	pip fruit	other vege	rolls/buns	potted plan
27039	42.712	CHEQUE	M	46	T	F	F	F	F	F	F
25011	25.357	CASH	F	28	F	T	F	F	F	F	F
94024	20.618	CASH	M	36	F	F	T	F	F	F	F
73966	23.688	CARD	F	26	F	F	F	T	F	F	F
32653	18.813	CARD	M	24	F	F	T	F	T	F	F
28663	46.487	CARD	F	35	F	F	T	F	F	F	F
46674	14.047	CASH	F	30	F	F	F	F	F	T	F
12687	22.203	CASH	M	22	F	F	F	F	T	T	F
89009	22.975	CHEQUE	F	46	F	F	F	F	F	F	T
65017	14.569	CASH	M	22	F	F	T	F	F	F	F
67670	10.328	CASH	F	18	F	T	F	F	T	F	F
28353	13.78	CASH	F	48	T	T	T	F	F	F	F
29082	36.509	CARD	M	43	F	F	F	F	F	F	F
33133	10.201	CHEQUE	F	43	F	F	F	F	F	T	F
19310	10.374	CASH	F	24	F	T	F	F	F	F	F
91867	34.822	CHEQUE	F	19	F	F	F	F	F	F	F
14612	42.248	CARD	M	31	F	F	F	F	F	F	F
96120	18.169	CASH	F	29	F	F	F	F	F	F	F
2657	10.753	CASH	F	26	F	F	F	F	F	F	F
51200	32.318	CARD	F	38	F	F	F	F	F	F	F
91382	31.72	CASH	M	38	F	F	F	F	T	F	F
23677	36.833	CASH	F	43	F	F	F	F	F	F	F
35389	31.179	CHEQUE	F	41	F	T	F	F	F	F	F
66030	21.681	CASH	M	48	F	T	F	F	T	T	F
45384	29.854	CASH	M	31	F	F	F	F	F	F	F
87577	15.27	CARD	F	23	F	F	F	F	F	F	F
14848	32.232	CHEQUE	F	32	F	F	F	F	F	T	F
45482	42.567	CARD	M	34	F	F	F	F	F	F	F
655	44.504	CASH	F	22	F	F	F	F	F	F	F

2.9 | Data Integration:

- Combines data from multiple sources into a coherent store.
- Increasingly data a mining projects require data from more than one data source.
- Such as multiple databases, data warehouse, flat files and historical data.

2.9.1 | Data Integration

- ✚ Data Warehouse: is a structure that links information from two or more databases.

- + Data warehouse brings data from different data sources into a central repository.
- + It performs some data integration, clean-up, and summarization, and distribute the information data marts.

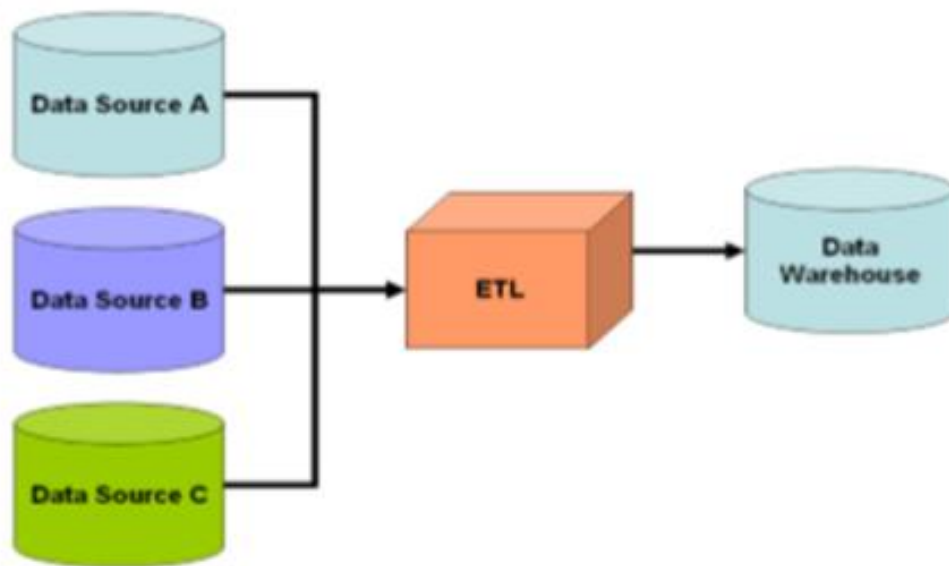
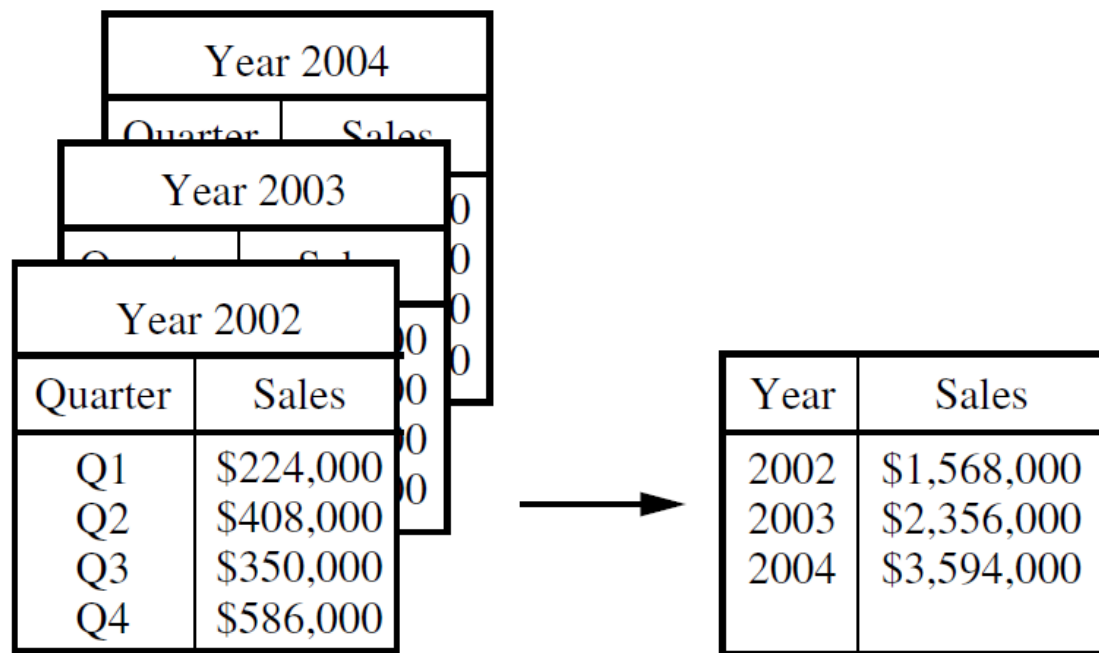


Figure 5 Data Integration

2.10| Data Reduction

Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results



Chapter 3

3| Modeling

3.1| Modeling

3.1| Modeling

This is the most important part of the data mining process because it searches for new patterns in the data.

The modeling phase includes four tasks. These are

- ✓ Selecting modeling techniques

Specify the technique(s) that will use in project

- ✓ Designing test(s)

The test in this task is the test that will use to determine how well model works. It may be as simple as splitting data into a group of cases for model training and another group for model testing. Training data is used

to fit mathematical forms to the data model

- ✓ Building model(s)

In this process the model and structure of the technique used in the data mining process is constructed

✓ Assessing model(s)

In the process the model is being evaluated correctly or you create a new model

3.2 | Clustering Algorithms

3.2.1 | What is Clustering?

- A cluster of data objects can be treated as one group.
- While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.
- The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

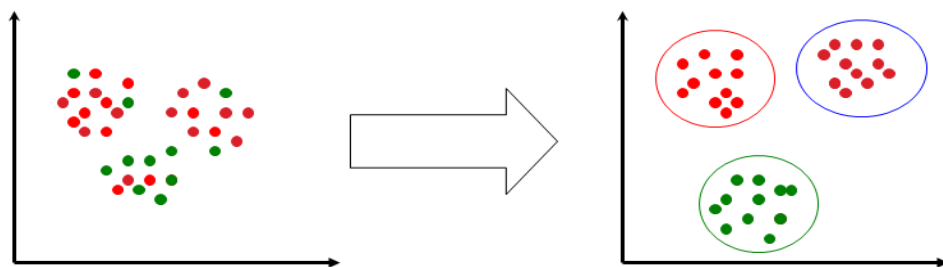


Figure 6 Clustering Algorithms

3.2.2|The Goal of clustering

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data.

3.2.3|Requirements

The main requirements that a clustering algorithm should satisfy are:

- **Scalability - Data must be scalable otherwise we may get the wrong result. Fig II shows simple graphical example where we may get the wrong result.**
- Clustering algorithm must be able to deal with different types of attributes.
- Clustering algorithm must be able to find clustered data with the arbitrary shape.
- Clustering algorithm must be insensitive to noise and outliers.
- Interpret-ability and Usability - Result obtained must be interpretable and usable so that maximum knowledge about.
- Insensitivity to order of input records.
- High dimensionality.
- Interpretability and usability.

3.4|The classification algorithm used in the project

3.4.1|K-means clustering Algorithm

K-Means algorithms are extremely easy to implement and very efficient computationally speaking. Those are the main reasons that explain why they are so popular. But they are not very good to identify classes when dealing with in groups that do not have a spherical distribution shape.

The K-Means algorithm aims to find and group into classes the data points that have high similarity between them. In the terms of the algorithm, this similarity is understood as the opposite of the distance between datapoints. The closer the data points are, the more similar and more likely to belong to the same cluster they will be.

3.4.2| goals of this algorithm

is to find groups in the data, with the number of groups represented by the variable K . The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K -means clustering algorithm are:

The centroids of the K clusters, which can be used to label new data.

Labels for the training data (each data point is assigned to a single cluster)

3.4.3|Key Concepts

- Squared Euclidean Distance

The most commonly used distance in K-Means is the squared Euclidean distance. An example of this distance between two points x and y in m -dimensional space is:

$$d(x, y)^2 = \sum_{j=1}^m (x_j - y_j)^2 = \|x - y\|_2^2$$

Here, j is the *the* dimension (or feature column) of the sample points x and y .

3.4.4|Cluster Inertia

Cluster inertia is the name given to the Sum of Squared Errors within the clustering context, and is represented as follows:

$$SSE = \sum_{i=1}^n \sum_{j=1}^k w^{(i,j)} \left\| \mathbf{x}^{(i)} - \boldsymbol{\mu}^{(j)} \right\|_2^2$$

Where $\mu(j)$ is the centroid for cluster j , and $w(i, j)$ is 1 if the sample $x(i)$ is in cluster j and 0 otherwise.

k-Means can be understood as an algorithm that will try to minimize the cluster inertia factor.

3.5| Algorithm Steps

Here is step by step k-means clustering algorithm

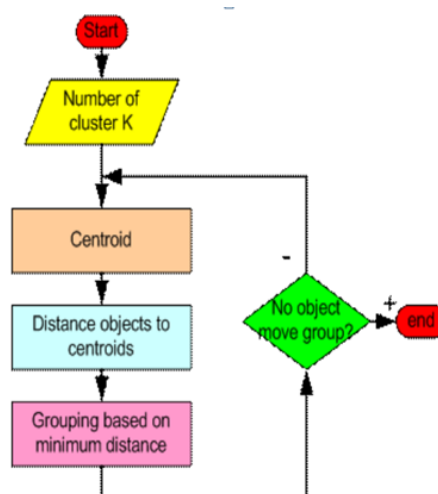


Figure 7 K -means clustering algorithm flowchart

1. First, we need to choose k , the number of clusters that we want to be found.
2. Then, the algorithm will select randomly the the centroids of each cluster.
3. It will be assigned each datapoint to the closest centroid (using Euclidean distance).
4. It will be computed the cluster inertia.
5. The new centroids will be calculated as the mean of the points that belong to the centroid of the previous step. In other words, by calculating the minimum quadratic error of the datapoints to the center of each cluster, moving the center towards that point
6. Back to step 3

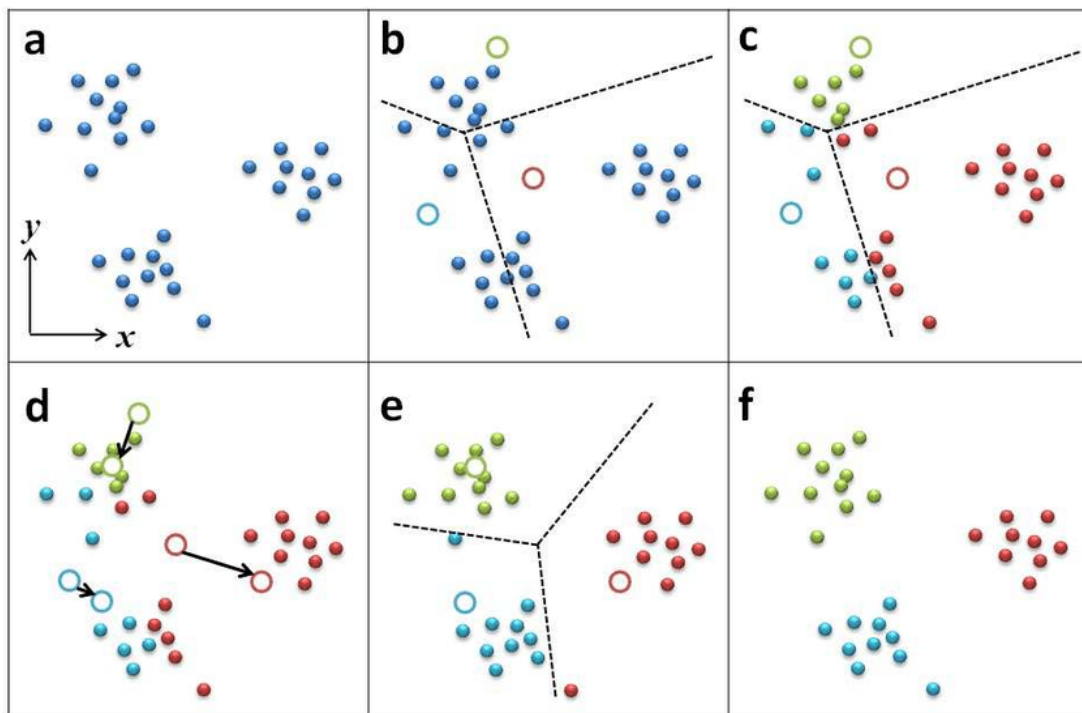


Figure 8 K-Means Hyperparameters

3.6|K-Means Hyperparameters

- Number of clusters: The number of clusters and centroids to generate.
- Maximum iterations: Of the algorithm for a single run.
- Number initial: The number of times the algorithm will be run with different centroid seeds. The final result will be the best output of the number defined of consecutives runs, in terms of inertia.

3.7|Challenges of K-Means

- The output for any fixed training set won't be always the same, because the initial centroids are set randomly and that will influence the whole algorithm process.
- As stated before, due to the nature of Euclidean distance, it is not a suitable algorithm when dealing with clusters that adopt non-spherical shapes.

3.8|Points to be Considered When Applying K-Means

- Features must be measured on the same scale, so it may be necessary to perform z-score standardization or max-min scaling.
- When dealing with categorical data, we will use the get dummies function.
- Exploratory Data Analysis (EDA) is very helpful to have an overview of the data and determine if K-Means is the most appropriate algorithm.
- The minibatch method is very useful when there is a large number of columns, however, it is less accurate.

3.9|How to Choose the Right K Number

Choosing the right number of clusters is one of the key points of the K-Means algorithm. To find this number there are some methods:

- Field knowledge
- Business decision
- Elbow Method

As being aligned with the motivation and nature of Data Science, the elbow method is the preferred option as it relies on an analytical method backed with data, to make a decision.

3.10|Elbow Method

The elbow method is used for determining the correct number of clusters in a dataset. It works by plotting the ascending values of K versus the total error obtained when using that K.

$$\% \text{ Variance} = \frac{\text{Variance between groups}}{\text{Total variance}}$$

The goal is to find the k that for each cluster will not rise significantly the variance

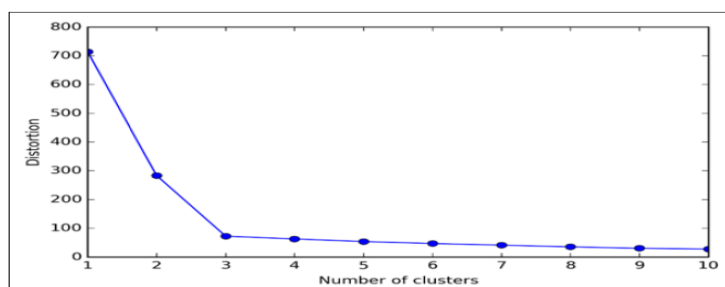


Figure 9 Elbow Method

In this case, we will choose the k=3, where the elbow is located

3.11| K-Means Limitations

Although K-Means is a great clustering algorithm, it is most useful when we know beforehand the exact number of clusters and when we are dealing with spherical-shaped distributions.

The following picture show what we would obtain if we use K-means clustering in each dataset even if we knew the exact number of clusters beforehand:

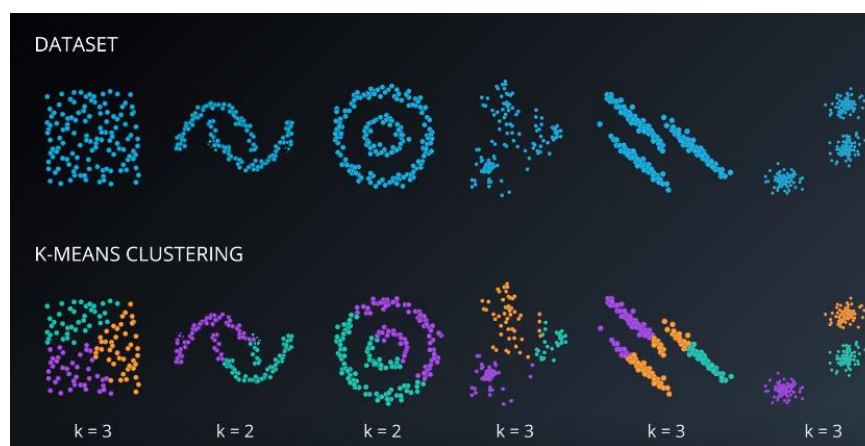


Figure 10 K-Means Limitations

It is quite common to take the K-Means algorithm as a benchmark to evaluate the performance of other clustering methods.

3.12| RESEARCH METHODOLOGY

3.12.1| Association Rule

As the outcome of the market basket analysis, association rule is a useful data mining method for mining frequent patterns, associations,

correlations, or causal structures among sets of items in transaction databases, the main idea of this technique is producing rules on associations between products from a transaction-based dataset.

3.12.2| Definition of Association Rule

is a rule-based Data Mining method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.

rule by the following steps.

Let $I = \{ i_1, i_2, \dots, i_m \}$ be a collection of m items in the market basket data.

Let $T = \{ t_1, t_2, \dots, t_n \}$ be the set of all transactions in the market basket data.

Each transaction t_i contains a subset of items from I .

A transaction t_i is said to contain an X , which is a collection of items, if X is a

subset of t_i ($X \subseteq t_i, t_i \in T$).

An association rule is an implication expression of the form $X \rightarrow Y$.

X and Y are disjoint collection of items ($X \neq \emptyset, Y \neq \emptyset, X \cap Y = \emptyset$).

The main idea of association rule is finding the relationship between purchases of different products.

It can be represented as:

IF {purchase A & B ($A, B \in X$)} THEN {purchase C ($C \in Y$)}

X is an antecedent. Y is a Consequent.

So, we can get:

Antecedent \rightarrow Consequent [support, confidence]

The strength of an association rule is measured by *support* and *confidence*.

Support is the percentage of transactions that include both the antecedent and

the consequent. It determines how often a rule is applicable to the given dataset. If P

means probability, that

$$\text{Support } (X \rightarrow Y) = P(X / Y)$$

Ex1 | Example of Support

ID	Items	Support Calculus
1	A, B, C	Total Support = 5
2	A, B, D	Support {AB} = $2/5 = 40\%$
3	A, C	Support {AC} = $2/5 = 40\%$
4	B, C	Support {BC} = $3/5 = 60\%$
5	B, C, D	Support {ABC} = $1/5 = 20\%$

Confidence is the percentage of antecedent transactions that also have the consequent item collection. It determines how frequently items in consequent (Y)

appear in the transactions, which contain the antecedent (X). If P means probability,

that

$$\text{Confidence } (X \rightarrow Y) = P(Y / X)$$

Ex2| Example of Confidence

ID	Items	Confidence Calculus
1	A, B, C	
2	A, B, D	Confidence $\{A \rightarrow B\} = \{AB\} / \{A\} = 2/3 = 66\%$
3	A, C	Confidence $\{B \rightarrow C\} = \{BC\} / \{B\} = 3/4 = 75\%$
4	B, C	Confidence $\{C \rightarrow D\} = \{CD\} / \{C\} = 1/4 = 25\%$
5	B, C, D	Confidence $\{AB \rightarrow C\} = \{ABC\} / \{AB\} = 1/2 = 50\%$

Both of the support and confidence are very important in association rule. As the measures of interestingness, they respectively reflect the usefulness and certainty of discovered rules (Han and Kamber).

A low support percentage means there is a low probability the chosen items were purchased X and Y together. And a low confidence percentage means there is a low percentage of customers who purchased X will also bought Y. Therefore, both a minimum support threshold and a minimum confidence threshold are necessary for this study. We want to find all rules $X \rightarrow Y$ that

satisfied the following two criteria:

The percentage of X and Y both appear must equal or higher than the percentage of minimum support threshold of all given transactions.

The percentage of Y appears in the given transaction that contain X must equal or higher than the percentage of minimum condition threshold

The performance of an association rule is measured by *lift*.

Lift is one of the correlation measures. The occurrence of the set of items, X,

is independent of the occurrence of Y.

X and Y

are dependent and correlated. That is $\text{lift}(x, y) = \frac{\text{Confidence}(A \rightarrow B)}{\text{Support}(B)} = \frac{P(X \cup Y)}{P(X)P(Y)}$

If the value of lift is greater than 1, it indicates a rule that is useful in finding consequent set of items. In another words, the occurrence of X and Y are positively correlated. If the value of lift is less than 1, it means X is negatively correlated with Y.

If the value of life is equal to 1, it means no correlation between X and Y. Computing the lift is more useful than only selecting transactions randomly.

3.12.3|Process of Association Rule

In general, there is a two-step process to solve the association rule problem.

1. Generating Frequent Set of Items

Setting a minimum support threshold based on the given dataset.

Then generating all sets of items from the given dataset that satisfied the support exceeds the minimum support threshold.

2. Generating Association Rules

Setting a minimum condition threshold based on the given dataset. Then generating all sets of items from the frequent set of items that satisfied the condition exceeds the minimum condition threshold.

3.13| We can list the associations for two algorithms as follows.

3.13.1|Apriori

Being developed Apriori Algorithms provide great benefits to be achieved during the development of the association rules of data mining. Because of this reason, **Apriori** Algorithms became the most popular algorithm in application of Association Rules.

Apriori node extracts rules which contain large information then, picks over set of rules. **Apriori** provides 5 different methods to pick over rules and uses sophisticated induction schematic to process efficient large data sets. **Apriori** is faster than Generalized Rules Induction (Gri) for bigger problems. **Apriori** requires input and output zones to be categorical because it is optimized for this type of data.

In Market Basket Analysis problems to identify the relation between products on sale there 2 criteria used which are support and confidence/trust. Rules Support Criterion identifies in one relation what is the proportion of repetition to every shopping. Rules Trust Criterion identifies what is the probability of someone buying B product who bought A product already

3.13.2|carma

CARMA makes the calculation of small-scale farms online CARMA displays the existing association rules online to the user and allows the user to change the minimal support and minimum trust parameters in any operation of the first scan of the database. CARMA constitutes a set of objects as they pass through the movements. After reading each movement, it first increases the numbers of the objections of the sub-clusters of the movement.

Then, if all of the existing subclasses of the object instance provide the minimum support value, and if they are larger than the read-out of the database, the object instances from the movement are created. An upper bound on the number of objects is computed so that the probability that a object is likely to be extreme can be precisely predicted. This is the sum of the current number and the estimate of its occurrence before the object instance is created. Estimating the probability of occurrence (maximum leaks) is calculated when the node is first created.

The CARMA model extracts a set of rules from the data without having to guess and target fields. In contrast to "Apriori" and "GRI", CARMA node provides only preliminary support instead of structure settings for support rule (support for the premise and consequent). At this point, these rules can be used for a variety of applications in a wider area.

Chapter 4

4| IMPELIMENTATIONS

4.1| Data Preprocessing

Step 1:

Data Cleaning

- Missing values
- Noisy Data

	A	B	C	D	E	F
50	93402	32.122	CASH	F	37	whole milk,yogurt,processed cheese,pickled vegetables,soda
51	16869	43.315	CHEQUE	M	35	whole milk,curd,yogurt,pastry
52	12478		CARD	M	34	packaged fruit/vegetables,brown bread,canned beer
53	42015	39.238	CASH	M	25	rolls/buns,oil,bottled water,chewing gum,chocolate marshmallow,hygiene articles,napkins
54	30247	10.62	CARD	F	35	ham,beef,whipped/sour cream,ice cream,rolls/buns,cat food
55	75366	-10.54	CASH	M	47	rolls/buns,pastry,sugar
56	69334	49.178	CASH	F	17	other vegetables,whole milk,frozen vegetables,canned fish,salty snack,seasonal products,detergent
57	3362	-48.4	CARD	Male	29	sausage,pastry
58	88382	15.716	CARD	F	24	sausage,beef,whole milk
59	27995	26.251	CHEQUE	M	22	frankfurter,tropical fruit,rolls/buns,brown bread,sugar
60	28582	45.796	CARD	F	25	rolls/buns,pastry,soda
61	43835		CARD	Female	46	curd cheese,coffee
62	13816	26.348	CARD	F	31	red/blush wine,newspapers
63	1193		CASH	M	42	sausage,whole milk,curd
64	12032	30.426	CARD	M	17	tropical fruit,pip fruit,berries,whole milk,frozen potato products,rolls/buns,pickled vegetables,chocolate
65	11652	42.713	CASH	F	40	red/blush wine
66	76336	28.534	CASH	M	35	whole milk,butter,margarine,specialty fat,specialty chocolate,candles,flower (seeds)
67	19976	-31.2	CASH	M	49	frankfurter,citrus fruit,whole milk,domestic eggs,oil,sparkling wine,specialty chocolate,newspapers
68	83968	46.27	CARD	Male	39	whole milk,meat spreads

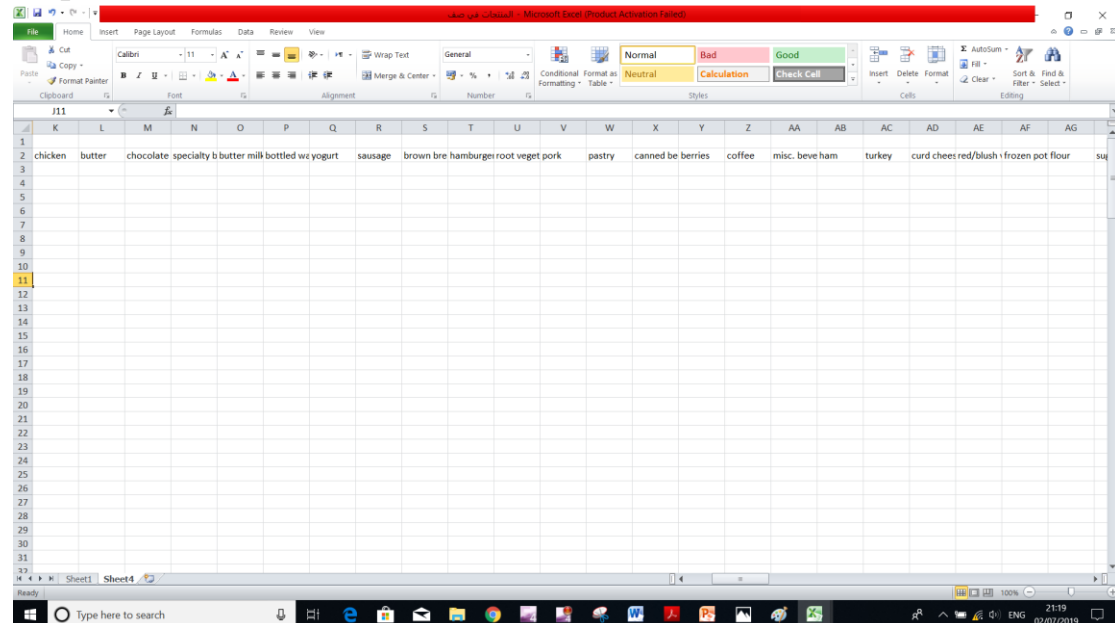
Manually enter the missing value and noisy data, assign a default value depending on the most implicit meaning

Step 2:

Item_Type									
citrus fruit	semi-finish	margarine	ready soups						
tropical fruit	yogurt	coffee							
whole milk									
pip fruit	yogurt	cream che	meat spreads						
other vegetable	whole milk	condensec	long life bakery product						
whole milk	butter	yogurt	rice	abrasive cleaner					
rolls/buns									
other vegetable	UHT-milk	rolls/buns	bottled be	liquor (appetizer)					
potted plants									
whole milk	cereals								
tropical fruit	other vege	white brea	bottled wa	chocolate					
citrus fruit	tropical fru	whole milk	butter	curd	yogurt	flour	bottled wa	dishes	
beef									
frankfurter	rolls/buns	soda							
chicken	tropical fruit								
butter	sugar	fruit/veget	newspapers						
fruit/vegetable juice									
packaged fruit/vegetables									
chocolate									
specialty bar									
other vegetables									
butter milk	pastry								
tropical fruit	cream che	processed	detergent	newspapers					

Manually we separated the products from some and cleared the iteration

Step 3:



Manually is to Transpose products into columns

Step 4:

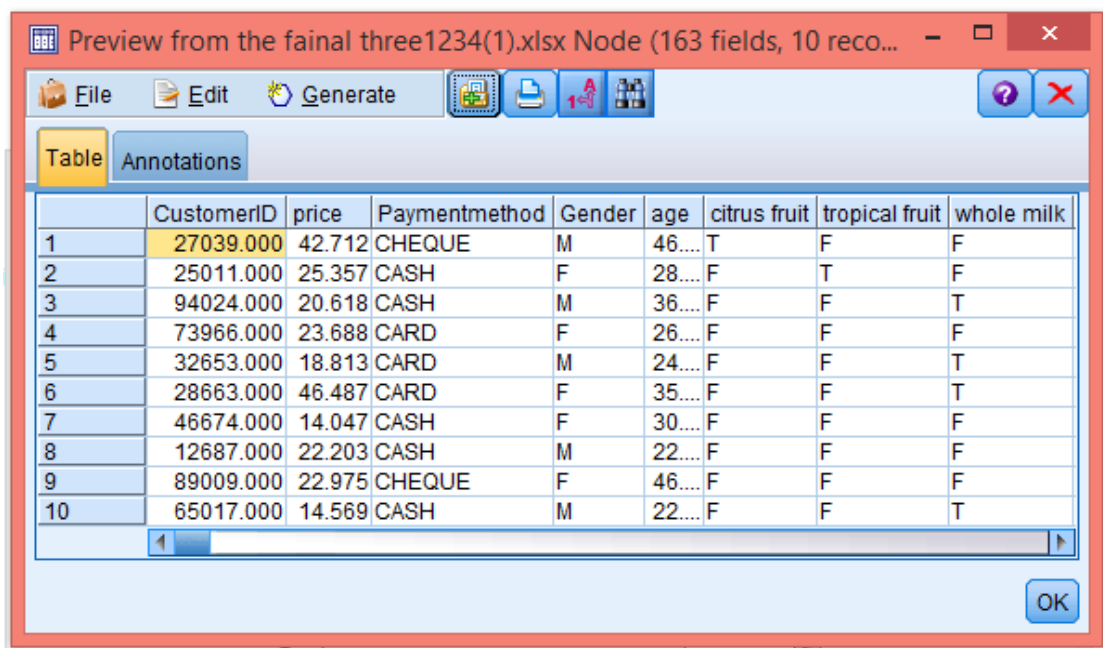
A	B	C	D	E	F	G	H	I
Customer	price	Payment	Gender	age	citrus fruit	tropical fru	whole milk	pip fruit
27039	42.712	CHEQUE	M	46	T	F	F	F
25011	25.357	CASH	F	28	F	T	F	F
94024	20.618	CASH	M	36	F	F	T	F
73966	23.688	CARD	F	26	F	F	F	T
32653	18.813	CARD	M	24	F	F	T	F
28663	46.487	CARD	F	35	F	F	T	F
46674	14.047	CASH	F	30	F	F	F	F
12687	22.203	CASH	M	22	F	F	F	F
89009	22.975	CHEQUE	F	46	F	F	F	F
65017	14.569	CASH	M	22	F	F	T	F
67670	10.328	CASH	F	18	F	T	F	F
28353	13.78	CASH	F	48	T	T	T	F
29082	36.509	CARD	M	43	F	F	F	F
33133	10.201	CHEQUE	F	43	F	F	F	F
19310	10.374	CASH	F	24	F	T	F	F

Manually is putting on true and false products as per customers buy

4.2| Implement k-means algorithm

Step 1:

Load the Excel source.



	CustomerID	price	Paymentmethod	Gender	age	citrus fruit	tropical fruit	whole milk
1	27039.000	42.712	CHEQUE	M	46...	T	F	F
2	25011.000	25.357	CASH	F	28...	F	T	F
3	94024.000	20.618	CASH	M	36...	F	F	T
4	73966.000	23.688	CARD	F	26...	F	F	F
5	32653.000	18.813	CARD	M	24...	F	F	T
6	28663.000	46.487	CARD	F	35...	F	F	T
7	46674.000	14.047	CASH	F	30...	F	F	F
8	12687.000	22.203	CASH	M	22...	F	F	F
9	89009.000	22.975	CHEQUE	F	46...	F	F	F
10	65017.000	14.569	CASH	M	22...	F	F	T

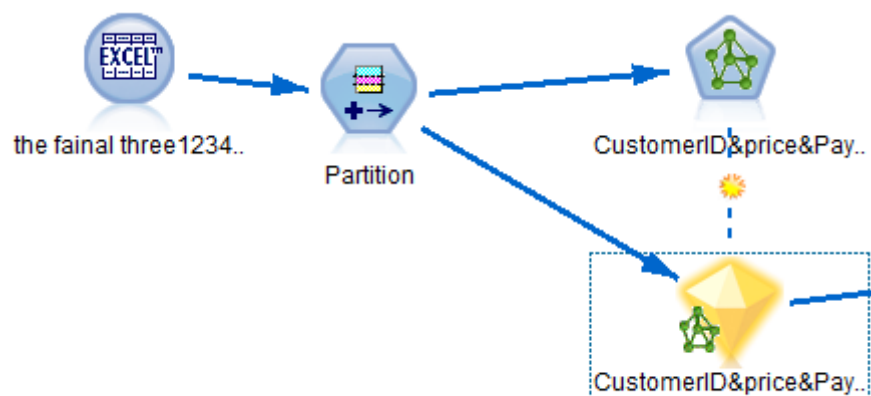
Step 2:

use Partition Node .



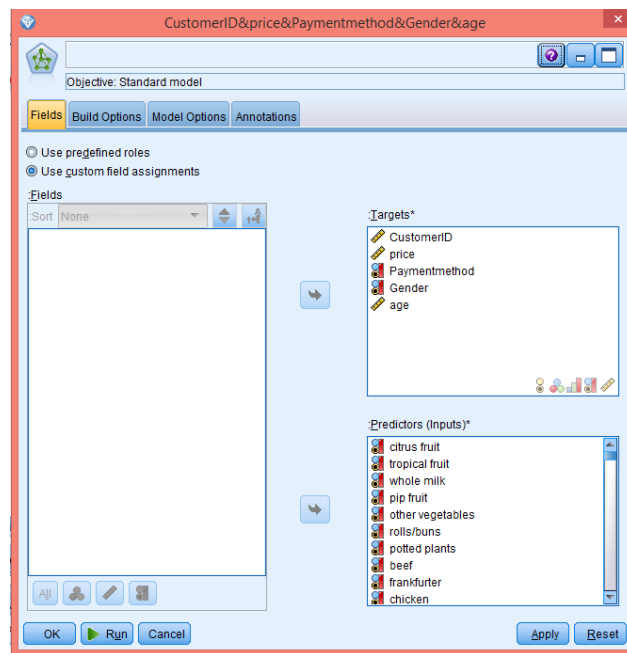
Partition nodes are used to generate a partition field that splits the data into separate subsets or samples for the training, testing, and validation stages of model building

Step 3:
using neural network.



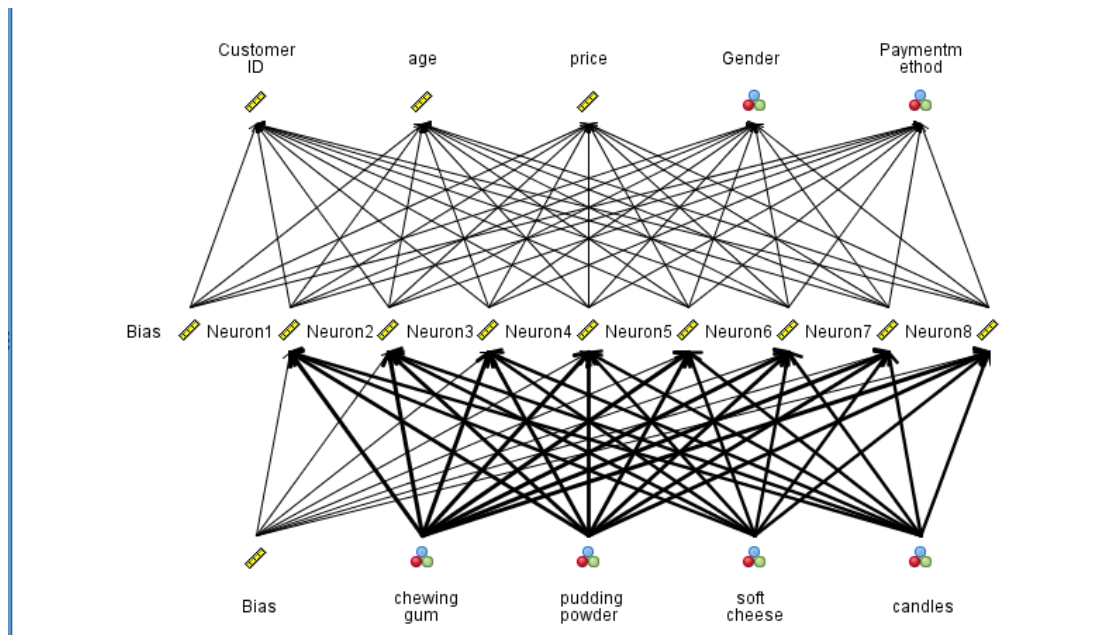
A neural network can approximate a wide range of predictive models with minimal demands on model structure and assumption

Step 4:
select attribute target.



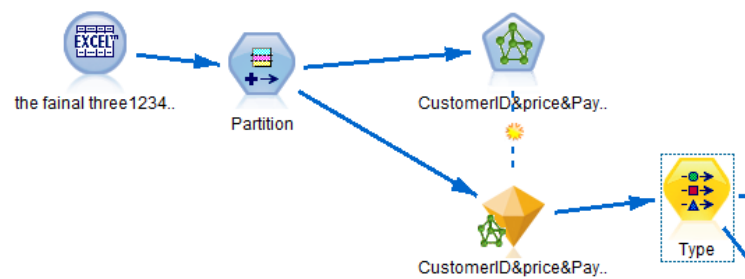
Attribute target is (Customer ID – price – Payment method -Gender -age).

Neural Network result.



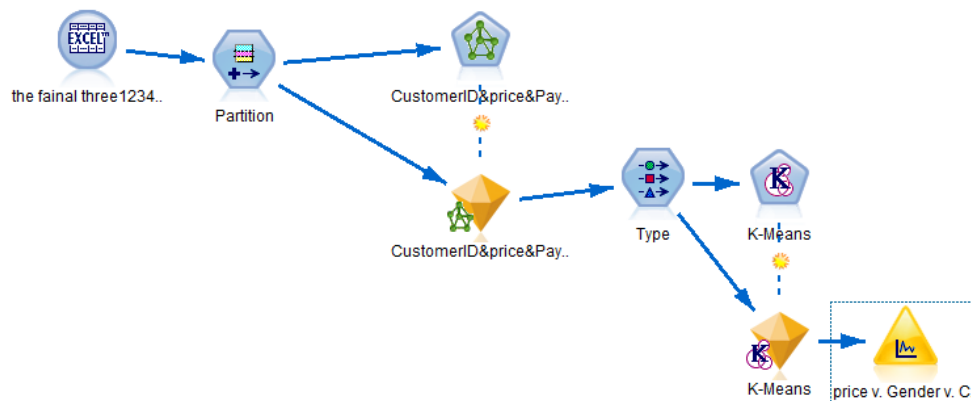
The conclusion from the previous picture is that most of the products have been bought (chewing gum-pudding powder-soft cheese-candles)

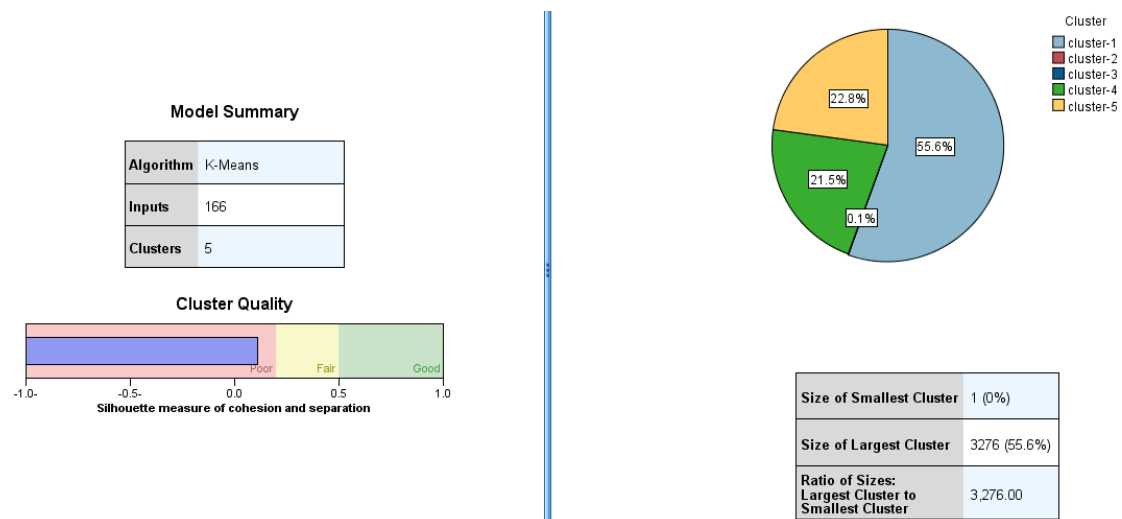
Step 5:
using type node.



Field properties can be specified in a source node or in a separate Type node. The functionality is similar in both nodes

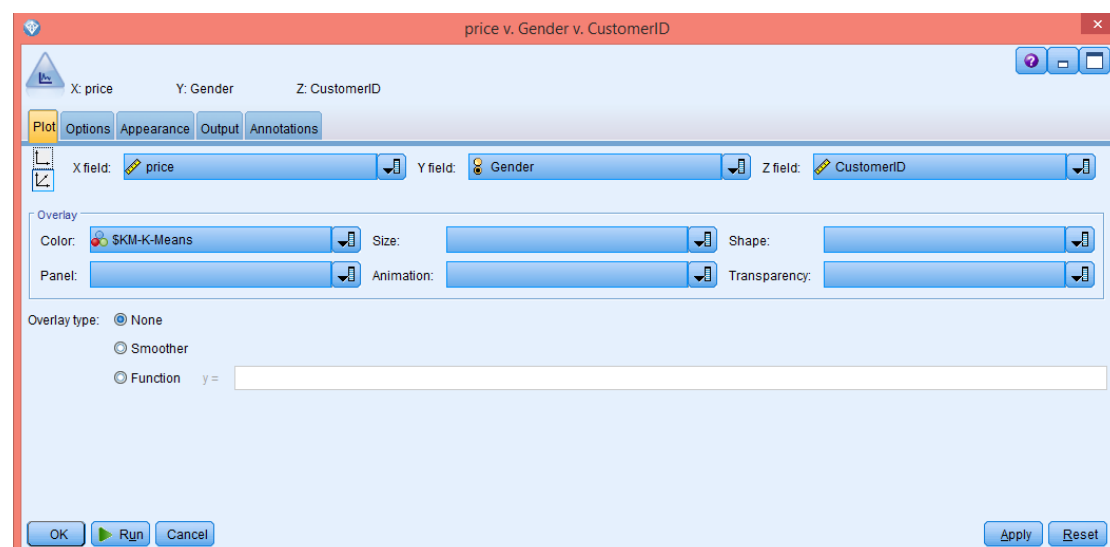
Step 6:
Using k-mean algorithm.



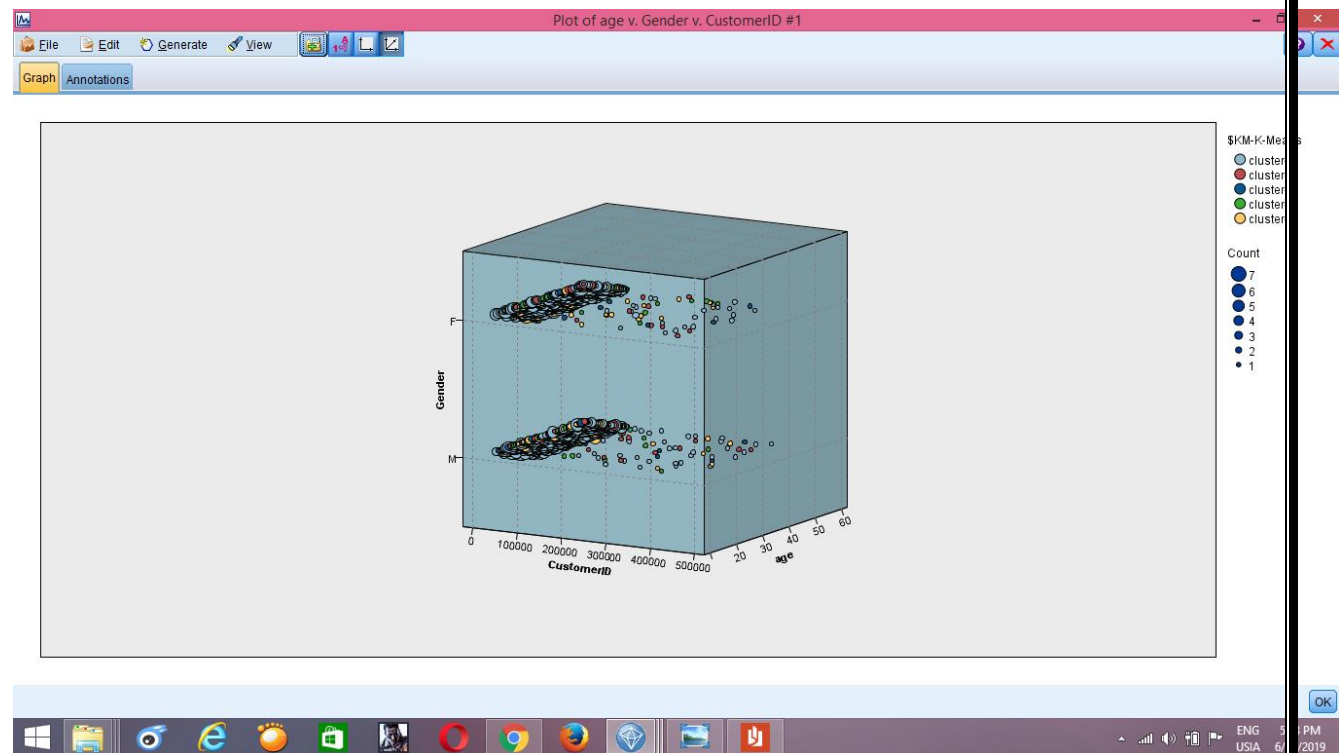


The previous image shows the size of % all cluster , size of smallest cluster and largest cluster

Step 7:
select fields into plot node (x, y, z).



Step 8: K-Mean Result.



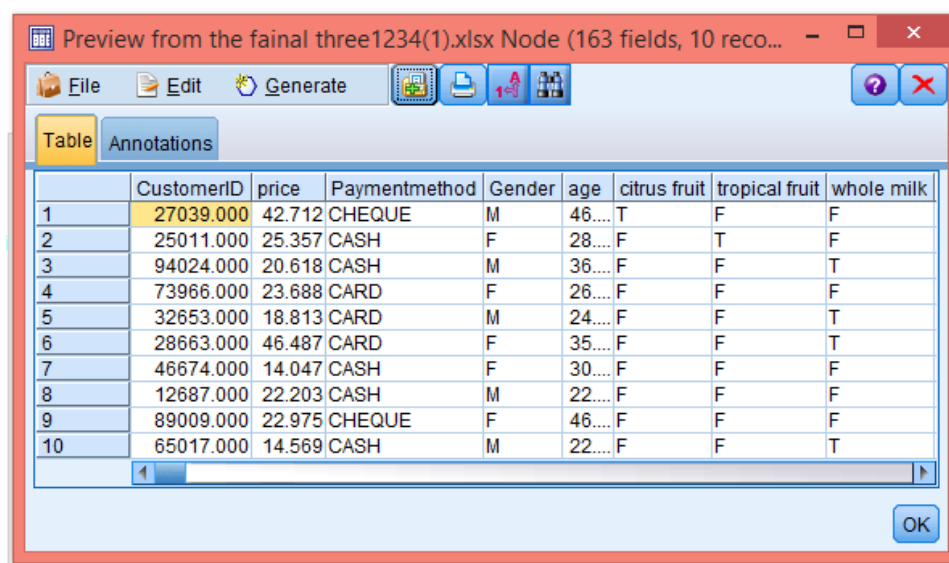
4.3| Data analysis and results

4.3.1 SPSS Modeling Process

The following figures show how to build an association rules algorithm in SPSS. All Four-steps questions mentioned in next image can be answered by loading different classified datasets into this model.

Step 1:

Load the Excel source.

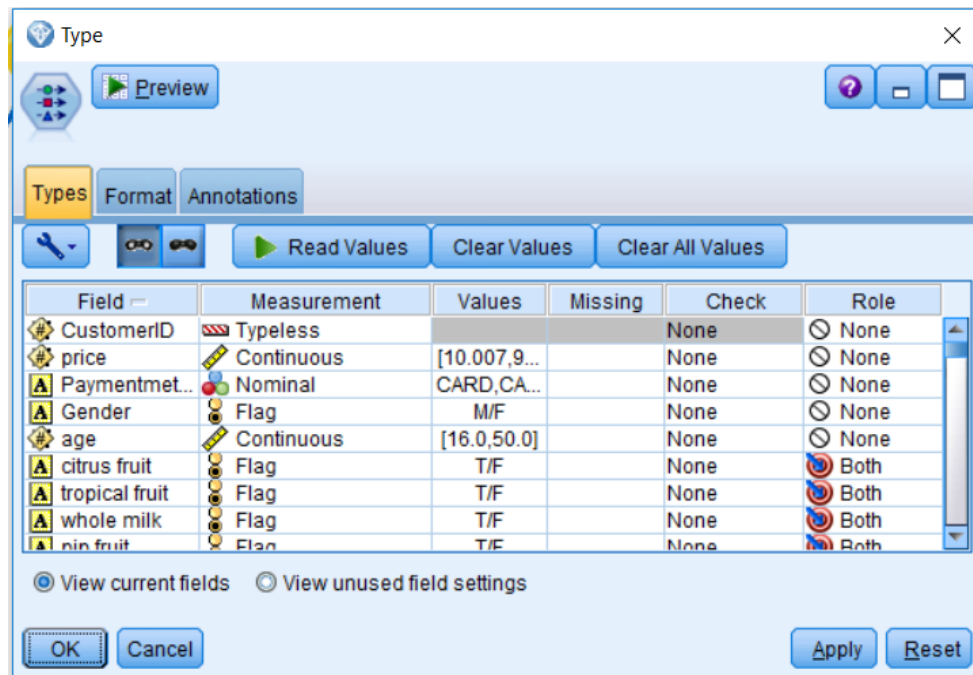


	CustomerID	price	Paymentmethod	Gender	age	citrus fruit	tropical fruit	whole milk
1	27039.000	42.712	CHEQUE	M	46....	T	F	F
2	25011.000	25.357	CASH	F	28....	F	T	F
3	94024.000	20.618	CASH	M	36....	F	F	T
4	73966.000	23.688	CARD	F	26....	F	F	F
5	32653.000	18.813	CARD	M	24....	F	F	T
6	28663.000	46.487	CARD	F	35....	F	F	T
7	46674.000	14.047	CASH	F	30....	F	F	F
8	12687.000	22.203	CASH	M	22....	F	F	F
9	89009.000	22.975	CHEQUE	F	46....	F	F	F
10	65017.000	14.569	CASH	M	22....	F	F	T

Step 2:

Assign a “Type” to the loaded Excel source.

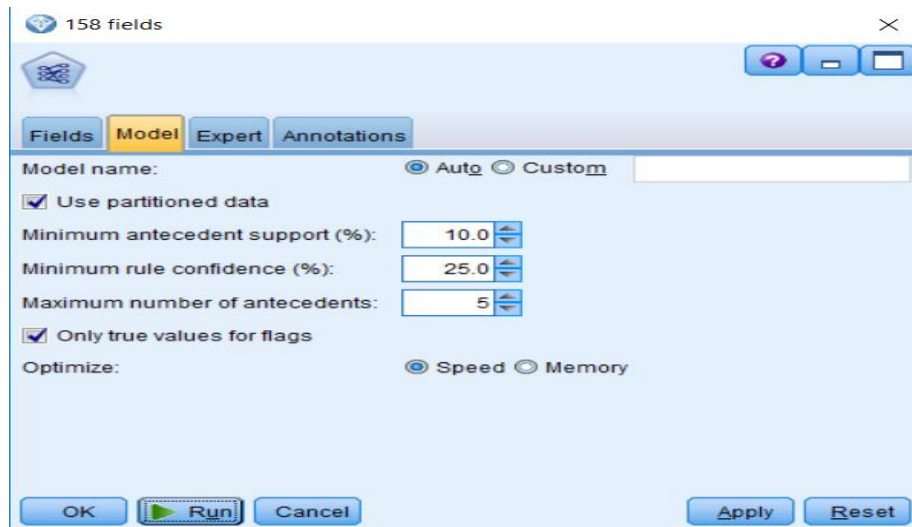
Changing all roles of product to “Both”, which means the roles of the fields are both input (predictor) and target (predicted). Then changing all products’ measurement level to “Flag” since we are using “F” and “T” in data.



Step 3:

Add an “Apriori” model to the Type.

Setting the minimum antecedent support to **10%**, the minimum rule confidence to **25%** and the maximum number of antecedents to 5 (products). The reason why we control the maximum number of antecedents to two is limiting the number of generated rules. If we allow more antecedents, more rules will be generated. That will increase the difficulty level of mining rules

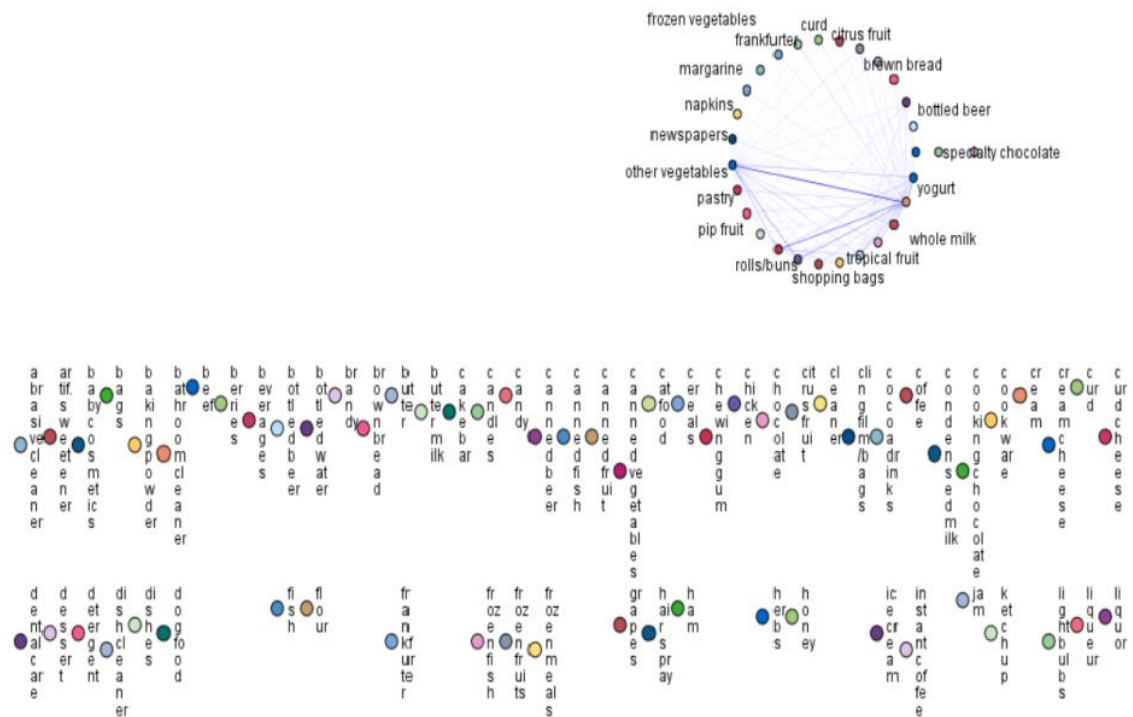


Step 4:

Run the Apriori Node Model to get the results.

The results include all rules satisfied 10% minimum support and 25% maximum confidence. We can sort the list by either support or confidence to do further analysis.

4.3.2| Results

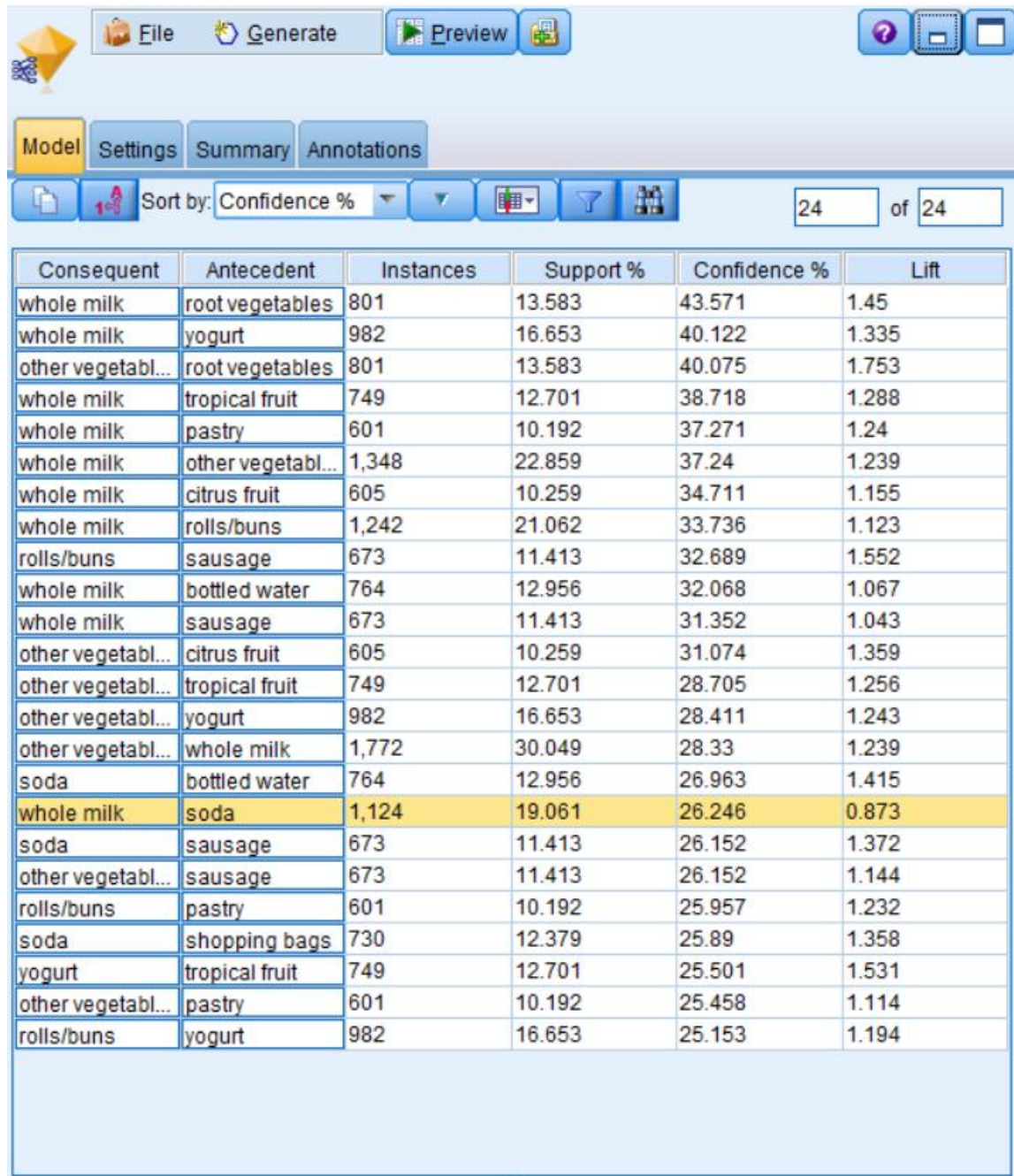


gives us a graphical understanding about levels of products links.

The heavier line it shows, the stronger links. It provides us some one-to-one

relationships between products. Some strong links such as *whole milk* and *root vegetables*

4.3.3| Most Interesting Association Rules



Consequent	Antecedent	Instances	Support %	Confidence %	Lift
whole milk	root vegetables	801	13.583	43.571	1.45
whole milk	yogurt	982	16.653	40.122	1.335
other vegetabl...	root vegetables	801	13.583	40.075	1.753
whole milk	tropical fruit	749	12.701	38.718	1.288
whole milk	pastry	601	10.192	37.271	1.24
whole milk	other vegetabl...	1,348	22.859	37.24	1.239
whole milk	citrus fruit	605	10.259	34.711	1.155
whole milk	rolls/buns	1,242	21.062	33.736	1.123
rolls/buns	sausage	673	11.413	32.689	1.552
whole milk	bottled water	764	12.956	32.068	1.067
whole milk	sausage	673	11.413	31.352	1.043
other vegetabl...	citrus fruit	605	10.259	31.074	1.359
other vegetabl...	tropical fruit	749	12.701	28.705	1.256
other vegetabl...	yogurt	982	16.653	28.411	1.243
other vegetabl...	whole milk	1,772	30.049	28.33	1.239
soda	bottled water	764	12.956	26.963	1.415
whole milk	soda	1,124	19.061	26.246	0.873
soda	sausage	673	11.413	26.152	1.372
other vegetabl...	sausage	673	11.413	26.152	1.144
rolls/buns	pastry	601	10.192	25.957	1.232
soda	shopping bags	730	12.379	25.89	1.358
yogurt	tropical fruit	749	12.701	25.501	1.531
other vegetabl...	pastry	601	10.192	25.458	1.114
rolls/buns	yogurt	982	16.653	25.153	1.194

The table show us in the first row.

If a customer purchases *root vegetable* there is 43.571% probability that the customer will order *whole milk*

The possibility of coexistence of these products in shopping vouchers is 13.583 %.

5 | Appendix

6 | REFERENCE

- Introduction to Data Mining *by (Tan, Steinbach & Kumar)*
- Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management *by (Gordon S. Linoff & Michael J. A. Berry)*
- Data Mining: Concepts and Techniques Second Edition (*Jiawei Han University of Illinois at Urbana-Champaign Micheline Kamber*)
- Data Mining Dummies

ملخص المشروع

الهدف من المشروع هو كيف يمكن تحليل المنتجات وتطبيق استخراج البيانات في ذكاء الأعمال واستخدام أدوات استخراج البيانات لإيجاد بعض العلاقة بين المنتجات أو فرز البيانات إلى مجموعات باستخدام خوارزمية تجميع يمكن استخدامها في عملية التسويق وزيادة نتائج الأرباح لتحقيق يمكن تنفيذ سوق تعادل تنافسي مع هذه التقنيات واستخراج البيانات بشكل فعال في قطاعات الأعمال المختلفة من خلال تصميم منهجي لنمذجة استخراج البيانات وتنفيذها.