

Anton Hall 1809131 Data Warehousing - ACCA6026

by Anton Hall

Submission date: 12-May-2022 11:46PM (UTC+0100)

Submission ID: 179806209

File name: 12316_Anton_Hall_Anton_Hall_1809131_Data_Warehousing_-_ACCA6026_398307_692334926.docx (911.17K)

Word count: 2654

Character count: 13714

University of Wales, Trinity Saint David

Data Warehousing and Data Mining - ACCA6026

Anton Hall

Contents

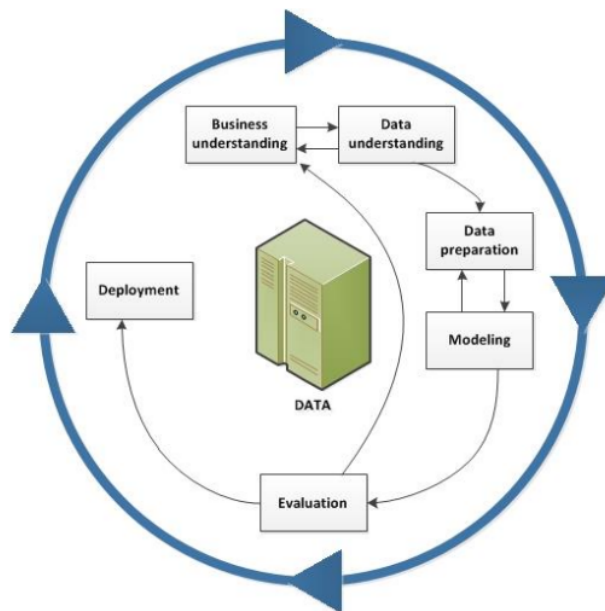
1. Data Mining.....	1
2. Data Set Description & Familiarisation	2
2.1 Data Set Description	2
2.2 Data Familiarisation	6
3. Data Processing.....	8
3.1 Primary Pre-Processing	8
3.1 Missing Data Handling	8
3.2 Numerical Transformation	8
3.3 Data Filtering.....	8
3.4 Additional Attributes.....	8
3.5 .arff File Creation	9
3.6 Weka Visualisation.....	9
4. Sample Runs.....	10
5. Business Recommendations	13
6. Conclusion	14
References	15
Bibliography	Error! Bookmark not defined.

1. Data Mining

The amount of data that we collect is ever increasing and the cost of storing and maintaining this data is ever decreasing. This makes storing data for any type of business easier, however, making valuable use of that data is not always achieved. A technique referred to as Data Mining is often used to discover patterns and valuable information from data. Data mining is also sometimes referred to as Knowledge discovery in databases (KDD).

“Data mining helps you understand how the elements of your business relate to one another. It provides clues about actions that you can take to make your business run more smoothly and generate more revenue.” [1]

CRISP-DM, which stands for Cross-Industry Standard Process for Data Mining is a good place to start to understand the methodology of data mining. It sets out the typical phases and tasks of a typical data mining project and provides an overview of the data mining lifecycle – the figure below illustrates this: [2]



CRISP-DM is broken down into 6 logical steps: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment.

The first step is Business Understanding, and its focus is to understand the goals that are to be achieved with the data. It is carried out so that the developer has a clear understanding of the context of the project before they start working on the data. An example of a goal could be to understand if there are any seasonal patterns in water consumption for a water company.

Next is Data Understanding, this is carried out to work out what can be achieved from the data. At this stage the analyst will check for completeness, quality, any data governance issues etc. It is important at this stage to discuss with any business stakeholders any issues that may be present in the data which could potentially affect the goal.

The third phase is Data Preparation, this is where any pre-processing on the data is carried out. Traditionally this phase of the project can take the longest. It includes any processing on the data necessary to carry out the analysis on the data.

The fourth step is modelling, this is the core of any data mining project. This step is where the goal of the project should be achieved, for example you should be able to gain the answer to the water consumption seasonality question. Some potential algorithms that may be used at this stage include decision trees, regression, and k-means.

Next is Evaluation, where the results of the modelling need to be verified. This is where the results are interrogated to see if we can answer the goal of the project.

Whether the answer to the goal is yes or not, the results need to be presented and fed back to the business. This is the deployment stage. The modelling results should be presented in a clear and understandable manner for less data minded business members to interpret. [3]

2. Data Set Description & Familiarisation

2.1 Data Set Description

The dataset that will be used as part of this data mining report is the Online Retail Data Set from the UCI Machine Learning repository: <https://archive.ics.uci.edu/ml/datasets/Online+Retail>. It describes the data as transactional which contains all the transactions occurring between 01/12/2010 and 09/12/2011. The data is for a UK based business that is a registered non-store online retail company. The company sells unique all occasion gifts and many of the customers are wholesalers.

The data contains eight attributes: [4]

Attribute No.	Attribute Header	Attribute Description
1	InvoiceNo	Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
2	StockCode	Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
3	Description	Product (item) name. Nominal.
4	Quantity	The quantities of each product (item) per transaction. Numeric.
5	InvoiceDate	Invoice Date and time. Numeric, the day and time when each transaction was generated.
6	UnitPrice	Unit price. Numeric, Product price per unit in sterling.
7	CustomerID	Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
8	Country	Country name. Nominal, the name of the country where each customer resides.

The number of instances in the data: 541910

Power BI has been utilised to investigate the quality and distribution of each attribute in the dataset.

InvoiceNo:

Count	541909
Error	0
Empty	0
Distinct	25900
Unique	5841
Empty string	0
Min	536365
Max	C581569

The invoice number column contains a value for every row and is text due the 'c' that gets added if an order is a cancelation.

StockCode:

Count	541909
Error	0
Empty	0
Distinct	4070
Unique	233
Empty string	0
Min	10002
Max	m

The stock code column also contains no empty rows and is also a text column.

Description:

Count	541909
Error	0
Empty	1454
Distinct	4224
Unique	308
Empty string	0
Min	4 PURPLE FLOCK DINNER CANDLES
Max	wrongly sold sets

The description column contains 1454 empty rows that will need to be addressed. It also contains 4224 distinct values that may be difficult to categorise into groups.

Quantity:

Count	541909
Error	0
Empty	0
Distinct	722
Unique	308
NaN	0
Zero	0
Min	-80995
Max	80995
Average	9.55224955
Standard deviation	218.081157850234
Even	324839
Odd	217070

The quantity column is fully populated. There is max of 80995 and a min of -80995 which may indicate an error in the data as the average is 9.55.

InvoiceDate:

Count	541909
Error	0
Empty	0
Distinct	23260
Unique	4242
Min	01/12/2010 08:26:00
Max	09/12/2011 12:50:00
Average	04/07/2011 13:34:57

The invoice data attribute is a date column with no null entries. The min is the 01/12/2010 and the max is the 09/12/2011. The date will need to be carefully filtered if the analysis covers data on a month by month or yearly comparison.

UnitPrice:

Count	541909
Error	0
Empty	0
Distinct	1630
Unique	1104
NaN	0
Zero	2515
Min	-11062.06
Max	38970
Average	4.611113626
Standard deviation	96.759853061178

The unit price column contains no empties, however it does have a large range with a max of 38970 and a min of -11062.02 which when compared with the average of 4.61 seems suspicious.

CustomerID:

Count	541909
Error	0
Empty	135080
Distinct	4373
Unique	79
NaN	0
Zero	0
Min	12346
Max	18287
Average	15287.69057
Standard deviation	1713.60030332162
Even	194630
Odd	212199

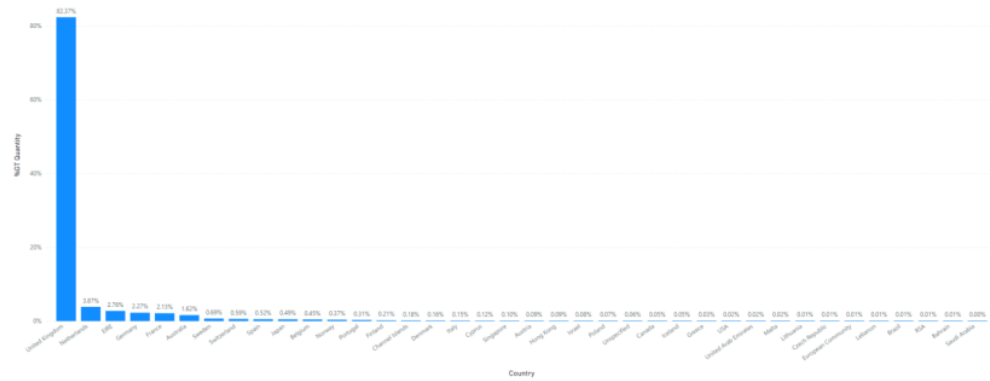
The customer ID column contains 135080 missing entries', this will need to be investigated and addressed.

Country:

Count	541909
Error	0
Empty	0
Distinct	38
Unique	0
Empty string	0
Min	Australia
Max	Unspecified

The country column contains 38 distinct entries.

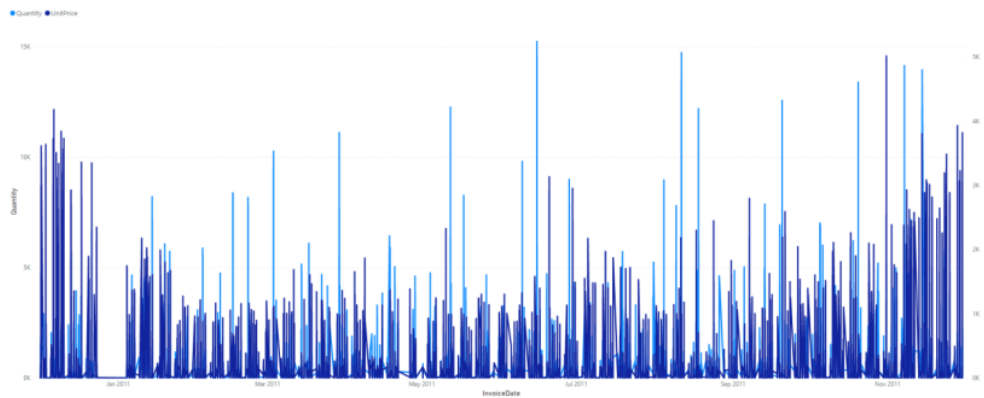
Power BI has also been utilised to visualise the data and investigate any suspicious data found in the column distribution.



The first attribute to considered is the country as there are 38 values it would be good to see any correlation with this being a UK based company with overseas sales, however, we can obviously see that most sales are made in the UK.



As shown by the tree map above comparing quantity ordered to description there is many different categories that could be considered in the item type. It may prove difficult to carry out analysis by item description because of this.



Visualising a comparison between unit price and quantity over time highlights some increased activity in the months of December and November. It also highlights many gaps with little activity, the most obvious is zero activity over the Christmas period. The data has been filtered to show a full year of data to ensure it is accurate, given that there is more data for December across two years in this representation.

CustomerID	Quantity	UnitPrice	Description	InvoiceDate	CustomerID	Quantity	UnitPrice	Description	InvoiceDate
15098	-1	38,970.00	Manual	2011-06		-1	17,836.46	AMAZON FEE	2011-12
16029	-1	8,142.75	POSTAGE	2011-05		-1	16,888.02	AMAZON FEE	2011-01
16029	1	8,142.75	POSTAGE	2011-05		-1	16,453.71	AMAZON FEE	2011-01
16029	-1	6,930.00	Manual	2011-05		-1	13,541.33	AMAZON FEE	2010-12
12744	-2	6,068.06	Manual	2011-10		1	13,541.33	AMAZON FEE	2010-12
12744	2	6,068.06	Manual	2011-10		-1	13,541.33	AMAZON FEE	2010-12
17448	-1	4,287.63	Manual	2011-07		-1	13,474.79	AMAZON FEE	2010-12
12536	1	4,161.06	Manual	2011-10		-1	11,586.50	AMAZON FEE	2011-12
12536	-2	4,161.06	Manual	2011-10		1	11,062.06	Adjust bad debt	2011-08
12536	1	4,161.06	Manual	2011-10		-1	8,286.22	AMAZON FEE	2011-11
12744	4	4,037.77	Manual	2011-04		-1	7,427.97	AMAZON FEE	2011-09
12744	-4	4,037.77	Manual	2011-04		-1	7,006.83	AMAZON FEE	2011-05
15502	-1	3,155.95	Manual	2011-08		-1	6,721.37	AMAZON FEE	2011-08
15502	1	3,155.95	Manual	2011-10		-1	6,706.71	AMAZON FEE	2010-12
18102	-1	3,060.60	Manual	2011-07		-1	6,662.51	AMAZON FEE	2011-08
15581	1	2,500.00	Manual	2011-08		-1	6,497.47	AMAZON FEE	2011-07
12744	1	2,053.07	Manual	2011-04		-1	5,942.57	AMAZON FEE	2011-10
12744	-1	2,053.07	Manual	2011-04		-1	5,942.57	AMAZON FEE	2011-10
17846	1	2,033.10	Manual	2011-09		-1	5,877.18	AMAZON FEE	2011-11
17940	-1	1,867.86	Discount	2011-04		-1	5,876.40	AMAZON FEE	2011-05
17940	1	1,867.86	Manual	2011-04		-1	5,791.18	AMAZON FEE	2011-06
12748	-1	1,829.84	Manual	2011-09		-1	5,693.05	AMAZON FEE	2011-03
12539	-1	1,715.85	Manual	2011-01		-1	5,575.28	AMAZON FEE	2011-02
14911	-1	1,687.17	Manual	2011-03		-1	5,522.14	AMAZON FEE	2011-09
14911	-1	1,687.17	Manual	2011-03		-1	5,519.25	AMAZON FEE	2010-12
14911	1	1,687.17	Manual	2011-03		-1	5,258.77	AMAZON FEE	2011-02

Analysing both customer IDs with and without values against large transactions highlights large amount of Manual, Postage and Amazon fees. These should be considered when carrying out the analysis.

Given the above familiarisation, the main goal for the analysis will be to look at trends in sales by customer ID. This may highlight to the business their loyal customers or busy periods within the week or year. To do this a total transaction column will be created as well as breaking down the invoice data column into days of the week, week of the year and month. It is evident that most sales take place in the UK, however there could be a link between item type and description but given the variety of the descriptions this may be hard to analyse without further knowledge of item grouping.

3. Data Processing

Pre processing will be carried out in Excel.

3.1 Primary Pre-Processing

3.1 Missing Data Handling

Attribute	No. Of Missing	No. Of Questionable Records	Action
Description	1454	68	Replace with "Missing"
CustomerID	135080	0	Replace with 0

3.2 Numerical Transformation

Attribute	Transformation	Action
UnitPrice	different decimal places	set decimal places to 2

3.3 Data Filtering

Attribute	Filter	Records Removed	Reason
Description	Remove admin transactions	AMAZON FEE Manual DOTCOM POSTAGE Discount CRUK Commission Adjust bad debt Bank Charges	We are looking for trends in item sales not admin fees
Invoice Date	Remove records to get one full year of transactions	Records < 01/12/2011	We are going to analyse trends by month and week

3.4 Additional Attributes

Attribute	Formula
MonthNo	MONTH(Invoice Date)
WeekNo	WEEKNUM(Invoice Date)
DayOfWeek	DAY(Invoice Date)
TotalTransation	Quantity*UnitPrice

Online Retail - Notepad

File Edit View

```
@RELATION onlineretail

@ATTRIBUTE InvoiceNo STRING
@ATTRIBUTE StockCode STRING
@ATTRIBUTE Description STRING
@ATTRIBUTE Quantity NUMERIC
@ATTRIBUTE UnitPrice NUMERIC
@ATTRIBUTE TotalTransaction NUMERIC
@ATTRIBUTE InvoiceDate DATE dd/mm/yyyy
@ATTRIBUTE MonthNo NUMERIC
@ATTRIBUTE WeekNo NUMERIC
@ATTRIBUTE DayOfWeek {Mon,Tue,Wed,Thu,Fri,Sat,Sun}
@ATTRIBUTE CustomerID NUMERIC
@ATTRIBUTE country {United-Kingdom,France,Australia,Netherlands,Germany,Norway,EIRE,Switzerland,Spain,Poland,Portu

@DATA
C579887,84946,"ANTIQUE SILVER T-LIGHT GLASS",-1,1.25,-1.25,30/11/2011,11.00,49.00,Mon,16717,United-Kingdom
C579887,85048,"15CM CHRISTMAS GLASS BALL 20 LIGHTS",-1,7.95,-7.95,30/11/2011,11.00,49.00,Mon,16717,United-Kingdom
C579887,23490,"T-LIGHT HOLDER HANGING LOVE BIRD",-3,3.75,-11.25,30/11/2011,11.00,49.00,Mon,16717,United-Kingdom
```

Ln 1, Col 1 100% Windows (CRLF) UTF-8

“An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the Weka machine learning software.”[5]

Figure 1 displays 16 plots showing the distribution of various attributes in the 'All attributes' dataset. The attributes are arranged in a 4x4 grid. Each plot shows a distribution of values, with some attributes having multiple peaks or outliers.

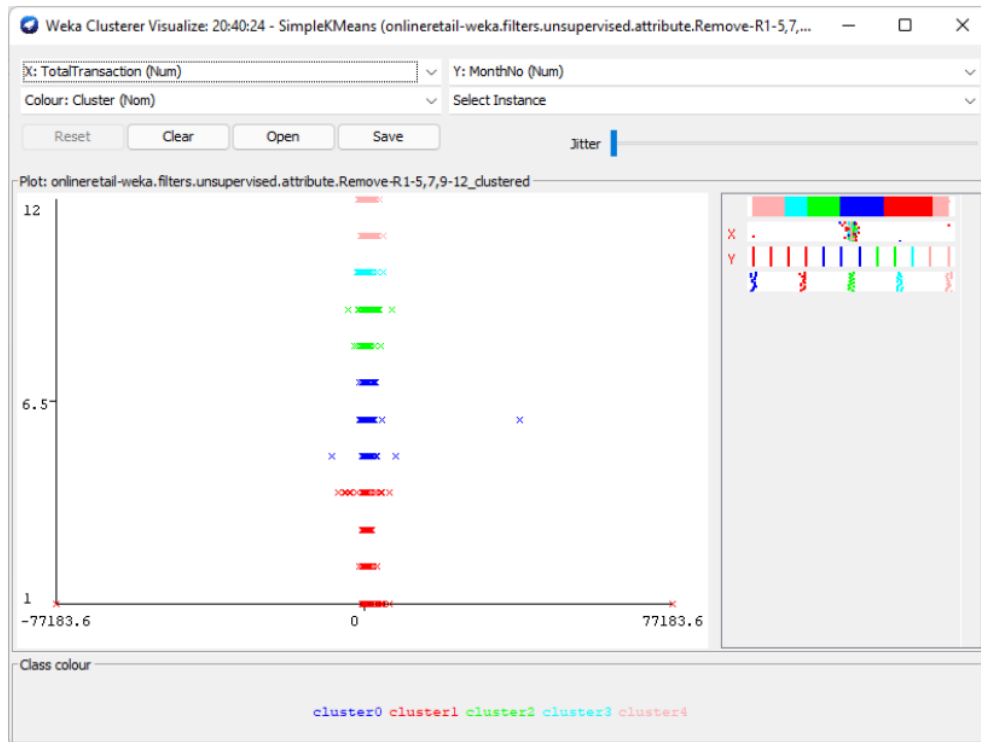
- InvoiceNo:** The distribution is highly skewed, with a very high peak at the maximum value (8142.75) and a long tail extending down to 0.
- StockCode:** The distribution is highly skewed, with a very high peak at the maximum value (77182.6) and a long tail extending down to 0.
- Description:** The distribution is highly skewed, with a very high peak at the maximum value (1252384720000) and a long tail extending down to 0.
- Quantity:** The distribution is highly skewed, with a very high peak at the maximum value (74215) and a long tail extending down to 0.
- UnitPrice:** The distribution is highly skewed, with a very high peak at the maximum value (8142.75) and a long tail extending down to 0.
- TotalTransaction:** The distribution is highly skewed, with a very high peak at the maximum value (77182.6) and a long tail extending down to 0.
- InvoiceDate:** The distribution is highly skewed, with a very high peak at the maximum value (1252384720000) and a long tail extending down to 0.
- MonthNo:** The distribution is highly skewed, with a very high peak at the maximum value (12) and a long tail extending down to 0.
- WeekNo:** The distribution is highly skewed, with a very high peak at the maximum value (52) and a long tail extending down to 0.
- DayOfWeek:** The distribution is highly skewed, with a very high peak at the maximum value (77060) and a long tail extending down to 0.
- CustomerID:** The distribution is highly skewed, with a very high peak at the maximum value (16287) and a long tail extending down to 0.
- country:** The distribution is highly skewed, with a very high peak at the maximum value (16287) and a long tail extending down to 0.

9

was created using these two attributes. These may need to be removed to produce more accurate results. We can also see the large number of missing CustomerIDs that have been replaced with a 0 clearly. The WeekNo column indicates a large increase in orders towards the later part of the year as well as the DayOfWeek suggesting that there is a particularly quiet day.

4. Sample Runs

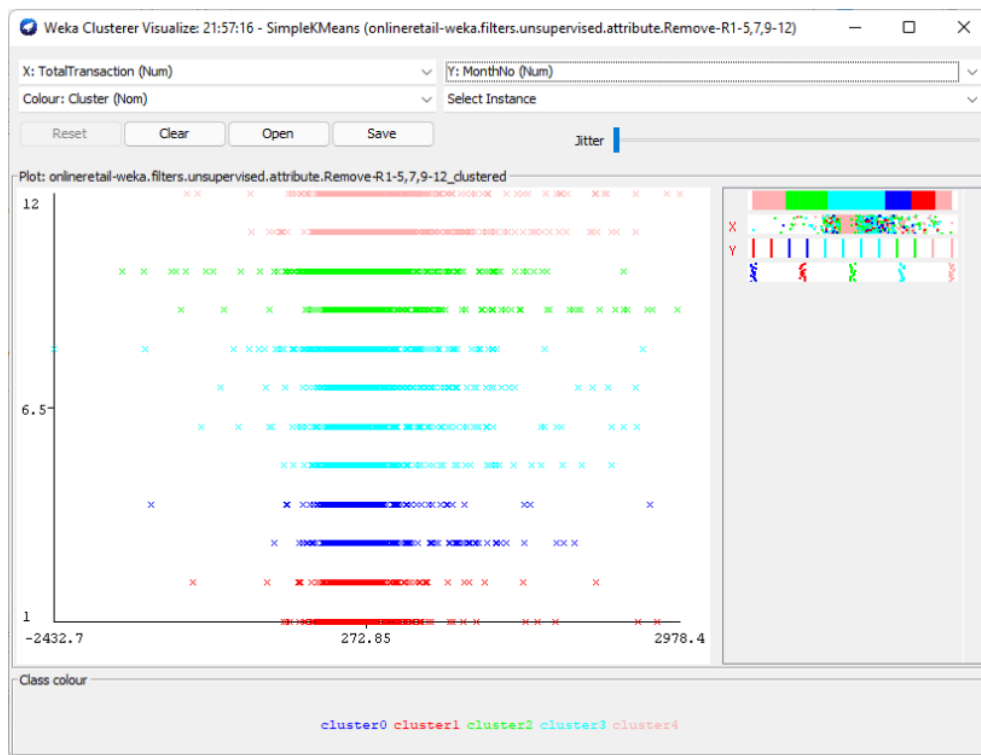
The first analysis method considered was the SimpleKmeans tool. This is an unsupervised clustering machine learning technique. The aim was to check the for any clusters using Total Transaction and Month.



Clustered Instances

- 0 112979 (22%)
- 1 129077 (25%)
- 2 85296 (17%)
- 3 60591 (12%)
- 4 126931 (25%)

Gaining any insight into the results is difficult due to the large outliers in the TransactionTotal, these will need to be removed.



Clustered Instances

- 0 66409 (13%)
- 1 62647 (12%)
- 2 110690 (22%)
- 3 148155 (29%)
- 4 126923 (25%)

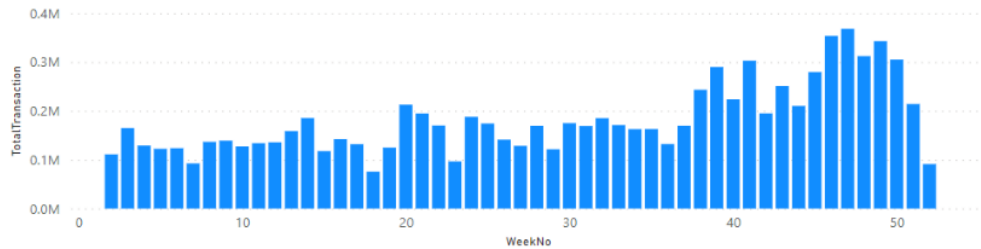
After removing the outliers of anything greater and less than 3000 the results are slightly clearer. The results indicate that in the later months of the year higher total transactions are made. This is evident by the more points in the top right of the chart above from months 8 to 12. We can also see that more transactions are made in the later months of the year due to the number of instances in each cluster. Clusters 2 and 4 cover 2 months each and hold 22% and 25% of all instances respectively whereas cluster 3 holds 29% of instances but covers the mid 4 months of the year.

Carrying out classification data mining in Weka did not yield any results on the dataset. Further analysis on the clean dataset was carried out in PowerBI.

TotalTransaction by Month



TotalTransaction by WeekNo



Comparing both week number and month with total transaction reveals the same insight as the cluster analysis, showing an increase in sales towards the end of the year, possibly indicating an influx on Christmas shopping.

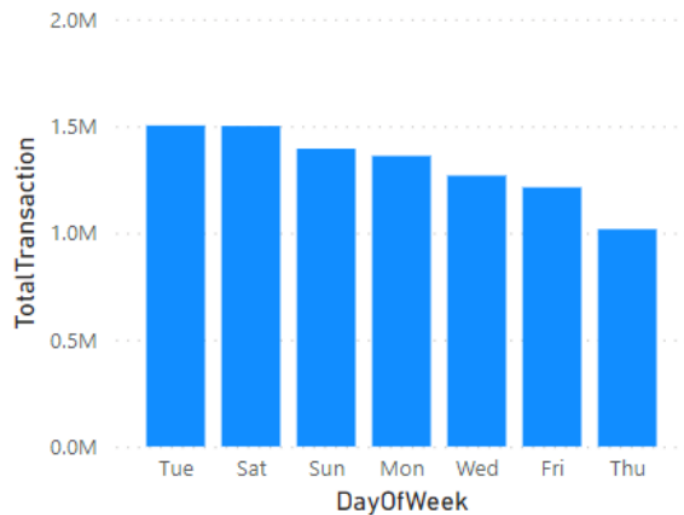
Description	TotalTransaction	%GT TotalTransaction
REGENCY CAKESTAND 3 TIER	158,859.27	1.72%
PARTY BUNTING	97,384.50	1.05%
JUMBO BAG RED RETROSPOT	90,160.33	0.97%
WHITE HANGING HEART T-LIGHT HOLDER	86,385.60	0.93%
POSTAGE	63,505.85	0.69%
ASSORTED COLOUR BIRD ORNAMENT	52,620.99	0.57%
CHILLI LIGHTS	51,134.07	0.55%
PAPER CHAIN KIT 50S CHRISTMAS	48,817.53	0.53%
RABBIT NIGHT LIGHT	45,386.38	0.49%
SPOTTY BUNTING	41,300.46	0.45%
JUMBO BAG PINK POLKADOT	40,693.34	0.44%
BLACK RECORD COVER FRAME	35,014.53	0.38%
DOORMAT KEEP CALM AND COME IN	35,007.17	0.38%
Total	9,254,824.56	100.00%

Given that the item descriptions are so varied, it's difficult to identify any patterns in them. The above shows the items listed and sorted by the total transaction amount with the most popular item being a cake stand making up 1.72% of the total transaction amount.

Description	TotalTransaction	%GT TotalTransaction
PAPER CHAIN KIT 50S CHRISTMAS	36,204.83	1.04%
RABBIT NIGHT LIGHT	35,619.38	1.03%
REGENCY CAKESTAND 3 TIER	34,774.35	1.00%
JUMBO BAG RED RETROSPOT	29,502.77	0.85%
HOT WATER BOTTLE KEEP CALM	23,856.75	0.69%
WHITE HANGING HEART T-LIGHT HOLDER	23,071.37	0.66%
POSTAGE	22,603.42	0.65%
PAPER CHAIN KIT VINTAGE CHRISTMAS	21,958.67	0.63%
HOT WATER BOTTLE TEA AND SYMPATHY	20,410.28	0.59%
DOORMAT KEEP CALM AND COME IN	20,019.21	0.58%
CHILLI LIGHTS	19,830.59	0.57%
JUMBO BAG 50S CHRISTMAS	18,878.01	0.54%
POPCORN HOLDER	18,656.29	0.54%
Total	3,470,700.30	100.00%

The above data is filtered to only show transactions taking place between the months of September and November and we can see that more items relating to Christmas are purchased during this time.

TotalTransaction by DayOfWeek



Viewing total transactions by the day that they were made indicates that Tuesday is the day where most transactions are made throughout the year. Nearly 50% less transactions are made on a Thursday overall which is the day where least sales are made.

5. Business Recommendations

Based on the findings outlined above the first business recommendation would be to cater to the customers buying patterns. It is evident that sales increase in the build up to Christmas so making sure that there are relevant items to purchase during this time would help to increase those sales

even further. It is also evident that most sales are made on a Tuesday so potentially focusing marketing on a different day such as Thursday where least sales are made could boost sales on the worst performing day. As well as increasing marketing, it would be advisable for the business to introduce offers on particularly low performing days of the week. Another business recommendation would be to ensure that there are sufficient staffing levels on the busiest days of trading, therefore based on the data available, the business should be suitably staffed on Tuesdays to keep up with sale demands.

In terms of business recommendations for the data there several improvements that can be made for recording sales. Firstly, a category could be added for each item type which aid in the analysis of pattern spotting by item type. Also, there is large amounts of missing data and some questionable records. As well as this there seems to be huge transactions which relate to business overheads rather than product sales which makes analysis on purely sale patterns difficult.

6. Conclusion

Reflecting on the above, the approach works well as a methodology in approaching data mining projects and is relatively efficient. However, without having background information about the company and the industry they operate in, as well as any prior background to the data to understand the goals of the analysis from the company's point of view, it makes producing accurate results much more difficult.

Approaching analysis in this way produces a well-documented, systematic approach that will speed up the process in the future. It adds additional information to the company allowing direction in improving areas where they may be under performing. Also, in terms of creating machine learning models, they are repeatable and efficient ways of analysing data.

If similar analysis was to be undertaken again, it would be beneficial to learn more background information about the company and its data to get a clear outlook on the goals to be achieved before completing the analysis.

References

- [1] M. S. Brown, "Data mining for dummies."
- [2] "CRISP-DM Help Overview - IBM Documentation." [Online]. Available: <https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=dm-crisp-help-overview>. [Accessed: 03-May-2022].
- [3] "CRISP-DM methodology leader in data mining and big data | by Israel Rodrigues | Towards Data Science." [Online]. Available: <https://towardsdatascience.com/crisp-dm-methodology-leader-in-data-mining-and-big-data-467efd3d3781>. [Accessed: 06-May-2022].
- [4] "UCI Machine Learning Repository: Online Retail Data Set." [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Online+Retail>. [Accessed: 06-May-2022].
- [5] "Attribute-Relation File Format (ARFF)." [Online]. Available: <https://www.cs.waikato.ac.nz/ml/weka/arff.html>. [Accessed: 08-May-2022].

FINAL GRADE

53/100

GENERAL COMMENTS

Instructor

Anton Hall	
1 engagement with literature comment	available
5 references 4 of which are websites and one of which is a book (data mining for dummies) a far wider range of sources and source types is expected at this level - sources are cited within the body	20
2 Knowledge and Understanding comment	available
Data mining well summarised, this section could have been significantly expanded	10
3. Cognitive and Intellectual skills comment	available
Excellent evidence of work done and results gained but discussion of the results and choices made for the investigation is limited in places	25
4. Practical Application Skills comment	available
Powe BI and weka were used for visualisation and data exploration- really good demonstrations of the explorations carried out but the explanation could use significantly expanding	35
5. Transferable Skills for life and Professional pract comment	available
no clear introduction caption and reference all images the report generally well laid out with only minor grammatical errors	15
Total:	53

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17
