

Data_Mining_1004334

by Adam Moore

Submission date: 13-May-2022 02:20PM (UTC+0100)

Submission ID: 179928614

File name: 12919_Adam_Moore_Data_Mining_1004334_398307_1376790280.docx (213.51K)

Word count: 3536

Character count: 19540

Student name:	Adam Moore	Student number:	ID 1004334
Programme:	BSc Data and Information Systems		
Module:	Data Warehousing and Data Mining		
Module code:	ACCA6026	Contribution to Overall Module Assessment (%):	50%
Lecturer:	Dr Nik Whitehead	Internal Verifier:	Bev Holland
Assignment Title:	Data Mining	Word count (or equivalent):	2,995 excluding figures and tables
Submission deadline:	13/05/2022	Return date of provisional marks & written feedback:	Unknown

Contents

- Introduction 3
- Dataset detail 4
 - Attributes.....4
 - Familiarisation and Exploration.....6
 - Transformation and processing.....7
- Runs and Results 9
- Reflection 13
- Bibliography 14

Introduction

Substantial cost implications associated with storing large volumes of data is now an historical issue [1]. Businesses and organisations can now maintain large databases in a cost-effective way, progress that has led to so called big data becoming widely available for data mining.

Extracting value from this data requires analysts to firstly turn data in its raw form into information, and then in turn generate insight and knowledge from the information [2]. This insight and knowledge can form a basis for positive decisions that add value to a business or organisation in some way.

Data mining is an analytical process that can identify trends, patterns, correlations, anomalies and relationships within data [3] [4]. These findings are generally utilised to build classification, predictive, estimative and descriptive models that enable analysts to predict future data [5] [6]. Data mining works by leveraging statistics and machine learning, a branch of artificial intelligence [7]. The greater the volume and complexity of any data that undergoes the data mining process, the more accurate predictions and insights derived from data mining will likely be. Optimisation for analytics within database management is fundamental if a significant amount of data, or a significant frequency of retrievals are being undertaken to ensure performance is acceptable and timeframes can be met [8].

In order to undertake data mining, typically an extract, transform and load process (ETL process) will need to be completed prior to commencement. The extract section of the process covers sourcing the data, and may involve developing an understanding of source systems, nuances, convention and more. Generally, some transformation, cleaning or reformatting needs to occur before the data is in a sufficient state to be utilised optimally by subsequent processes. Loading this transformed data into any tools or programs that will then analyse it is the final stage in an ETL process.

Once loaded into a data mining application or program, machine learning algorithms (such as discovery and sorting algorithms) search for patterns, relationships, correlations and more within the data [6]. The result of this processing intensive task is a model that explains or describes these relationships numerically. This model, generated through supervised or unsupervised machine learning, can then be used to generate predictions. At this stage, the model is tested, and its accuracy determined. The results delivered can then be analysed to determine how effectively they can help meet business needs through more effective planning, or through the gaining of additional insight and understanding. This step can be achieved via a business intelligence report generated to clearly demonstrate the findings of the process to business stakeholders.

Insight and predictions generated via data mining can be utilised by businesses in a prescriptive manner; this leads to effective, data-driven business decisions that can mitigate risk, exploit opportunity, avoid oversight and generally add value to the business.

Dataset detail

The dataset utilised is publicly available retail/sales data downloaded from the University of California Irvine machine learning repository (UCI ML Repository). The dataset is described by UCI as *"... a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail."* [9]

The dataset has an 8x541909 structure; 8 columns and 541,909 rows. There are 541,909 instances in this dataset.

Attributes

There are eight attributes within the dataset:

1. *InvoiceNo*; Six-digit number assigned to each transaction. These numbers are unique to each transaction and serve as an ID/reference. Where numbers are preceded with a C, it indicates an order that has been cancelled after an invoice has been generated. Nominal data type.
2. *StockCode*; Five-digit number assigned to each product. These numbers are uniquely assigned to each product type and each distinct product will share this uniquely assigned stock code. Numeric data type.
3. *Description*; String of text that describes the product. Nominal data type.
4. *Quantity*; The sum of product units associated with each transaction. Numeric data type.
5. *InvoiceDate*; A date-time field that represents the date and time at which each transaction was generated. As an online transaction, this may be interchangeable with the date and time at which orders we placed but we cannot make this assumption. Numeric data type.
6. *UnitPrice*; The individual unit price of each product in Great British pounds (GBP). Numeric data type.
7. *CustomerID*; Five-digit number assigned to each customer. These numbers are uniquely assigned to each customer. Nominal data type.
8. *Country*; The country from which each customer is a resident. Nominal data type.

This dataset has missing values associated with two of the attributes; *Description* and *CustomerID*. There are 1,454 blank records associated with *Description* and 135,080 blank records associated with *CustomerID*. The significant number of blank *CustomerID* records could be explained by usage of a 'guest' checkout system whereby a customer is not forced to create an account, by the so called 'right to erasure' or 'right to be forgotten' within GDPR legislation [10], or most probably a combination of factors such as these.

The table below outlines the number of instances, distinct values and unique values for each attribute. Where applicable, it also indicates the minimum value, maximum value, mean value and standard deviation of each attribute.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
Instances	541909	541909	*540455	541909	541909	541909	*406829	541909
Distinct	25900	4070	4222	722	23260	1630	4372	38
Unique	5481	233	308	308	4242	1104	79	0
Minimum	NA	NA	NA	-80995	NA	-£ 11,062.06	NA	NA
Maximum	NA	NA	NA	80995	NA	£ 38,970.00	NA	NA
Mean	NA	NA	NA	9.552	NA	£ 4.61	NA	NA
StdDev	NA	NA	NA	218.081	NA	96.76	NA	NA

*These values are accurate at the initial analysis phase, as some transformation takes place prior to processing and data mining processes, these values are not representative of the final values. This is also true for some outliers i.e., refund/incorrect orders (negative values).

The distinct column represents the number of values remaining if the attribute had been de-duplicated; that is to say, how many instances remain if we remove every subsequent instance where a value appears and keep only the first instance where a value appears. Distinct gives us a count of how many different values there are.

Unique differs from distinct in that, unique is simply a count of the values that only appear once; if we remove every instance of a value where it appears more than once (including the initial instance) we are left with the total of unique values. Unique gives us a count of the values that only appear once.

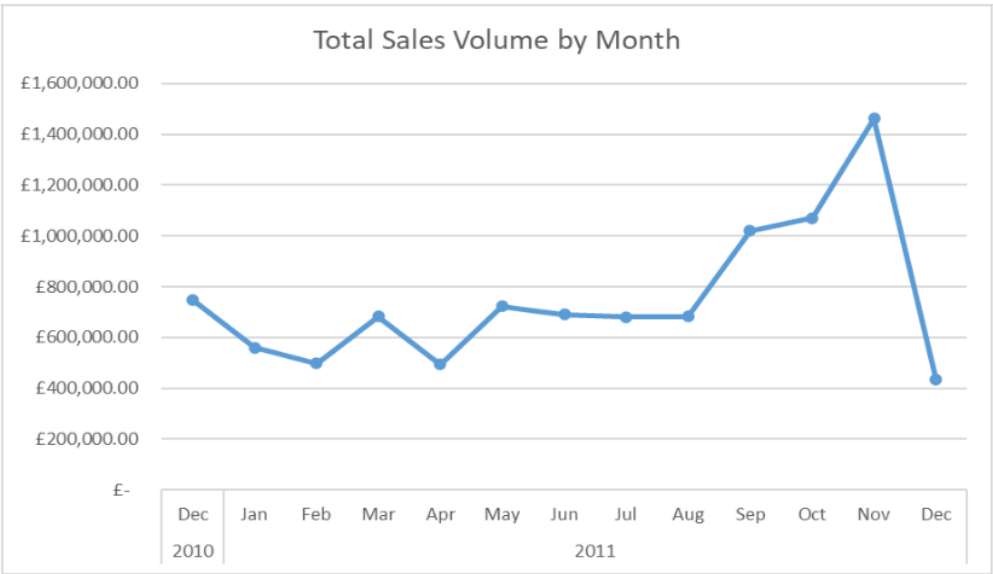
Familiarisation and Exploration

During exploration, the hypothesis that a viable field for clustering that was a boolean/logical value which represented whether the date-time was between 7am and 7pm was generated. This was achieved in excel by duplicating the date-time field, converting it to numeric format, and rounding it to zero decimal places. This gave a numeric date with no decimal values. In another column, this numeric date with no decimal value was subtracted from the date-time value which outputted just the time in a numeric format. Using the following nested logical test, the Boolean value was generated to denote a true/false as (1/0) as to whether the *InvoiceTime* was between 7am and 7pm:

=IF(AND(InvoiceTime>=0. 291666666666667, InvoiceTime<=0. 791666666666667),1,0)

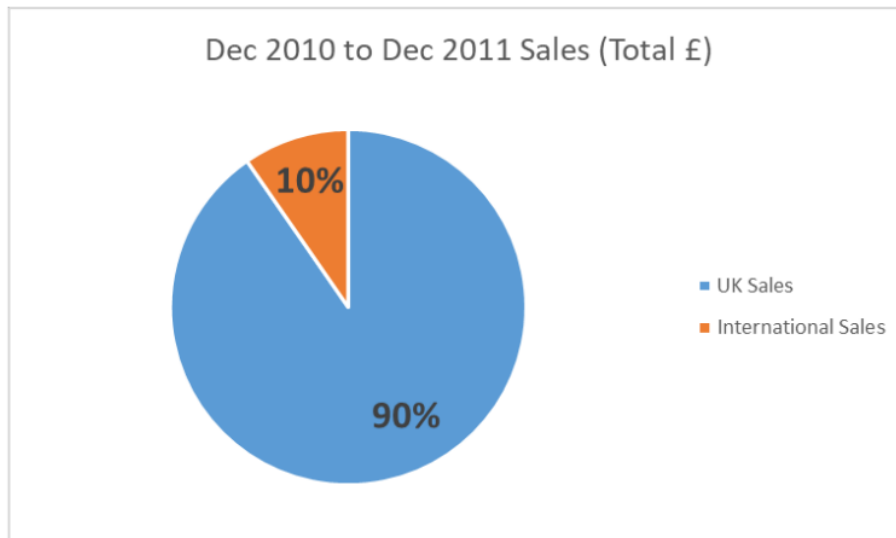
After wrangling data to generate this field, it was noted that a very small volume of instances that were not generated between 7am and 7pm, and these were all generated between 6am and 7am. Due to the 24-hour nature of online retail, this analysis (although not the intention) enables us to assume the invoice date time is indeed the time the invoice is generated for a transaction, and not the time the order was placed. This considerably limits the value of the time component of the date-time field as it is not linked to the customer interaction with the online retailer, but rather the system or staff generating invoices on behalf of the retailer.

At this stage, further visual exploration of the data was undertaken to prevent further invalid assumptions being made leading to redundant undertakings such as the *InvoiceTime* field. The attribute *TotalUnitsCost* was generated, by multiplying *Quantity* by *UnitPrice*. An aggregate pivot table was generated, to further explore the data. The first visual created to provide some insight into sales volume over time can be seen below:



This is an interesting visual; despite what appears to be an excellent year with an upward trajectory, there is a steep drop from Nov-11 to Dec-11. Also, the only year-on-year month that can be directly compared is December, and 2011 has considerably worse sales volume than 2010 despite the apparent general upward trajectory. It is reasonably difficult to draw any conclusions from this.

Most revenue from sales is generated by sales made to UK based customers. The large number of countries present in the list also adds some noise; an *InternationalSale* field is added to address this, a simple visualisation demonstrating the split between UK sales and international sales (by sum of total sales) is below:



This clearly demonstrates that most sales are made to UK-based customers. This isn't surprising insight for a UK-based firm, but it does help to build the overall picture of the business and its transactions.

Transformation and processing

Additional columns were created in excel for *InvoiceDateOnly*, *Month*, *Season*, *IsSpring*, *IsSummer*, *IsAutumn* and *IsWinter*. *InvoiceDateOnly* is a date field that does not include time. *Month* is a mmm-yy formatted field representing the month and year of each instance. *Season* classifies each record as being in a different season; December to February is Winter, March to May is Spring, June to August is summer and September to November is Autumn. The remaining columns are Boolean attributes that denote in binary terms, whether each record belongs to a season or not. These binary fields can later be used in clustering analysis.

Many blank cells are present in the data, which much be addressed so the data can later be converted to *arff* format and imported to WEKA. In the *description* attribute, 1,454 blank instances were converted to "Not Available". A further 135,080 blank instances were present in the *CustomerID* attribute. These were converted to a value of 0, as opposed to null (which should be default for blank values). The alter special function in excel (F5 > Special > Blanks and CTRL+Enter) was used to address these at scale efficiently.

Symbols are present in the data that are not readable by WEKA; these symbols are in the *Description* attribute and are " and '. The find and replace functionality in excel was used to remove these symbols.

There are 11,806 records in the data that relate to transactions with a sum value of £0 or less. Upon investigation, 78.7% of these records (9,288) relate to cancelled transactions. Where transactions

aren't cancelled, there is often no description or a vague description referencing damage, lost parcels, mixed up orders and other indications of problematic orders. These are potentially useful; theoretically, using data mining, analysis could be undertaken to detect patterns between customers or countries and problematic orders – results could highlight unfavourable customers (also potential fraud) and the business could take action to deny such customers their services. In this analysis, however, these records will be removed on the basis that they are outliers; further action will be taken to remove the top 0.5% of records as outliers. Power Query within excel was used to achieve this as it handles big data transformations far easier than excel can in normal tables, as a result of the efficient M programming language that power query is built on. All column names were validated to ensure they contained no spaces also.

The data was then aggregated into a per-order basis, once again using power query. To achieve this, much detail must be sacrificed such as description, stock code and unit price – this is due to the nature of aggregation and loss of granularity. However, upon analysis it was noted that the objective is to analyse different items and item types over time, not different orders based on the scale of their cost. This table therefore is not best designed to address the business question, and the aggregation was re-worked to be on a *StockCode* basis rather than an *InvoiceNo* basis. A more favourable table, for this task, would be to sacrifice other attributes and aggregate StockCode quantities sold, and revenue generated by season. The resultant table has 20,261 rows. Each row represents a *StockCode* sold in a season and whether it was a national or international sale, the aggregate value columns are unit quantities sold and total units' cost. This *.xlsx* file was the converted to a csv file, imported to WEKA and further converted to a *.arff* file and saved.

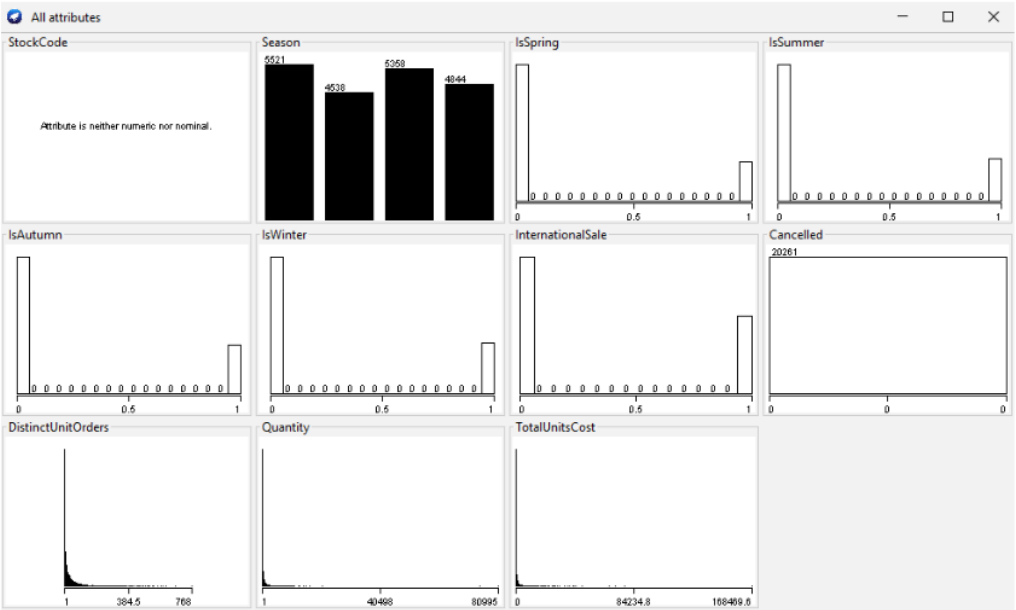
The *.arff* header is as follows:

```
@relation Retail_Aggregate

@attribute StockCode string
@attribute Season {Winter, Spring, Autumn, Summer}
@attribute IsSpring numeric
@attribute IsSummer numeric
@attribute IsAutumn numeric
@attribute IsWinter numeric
@attribute InternationalSale numeric
@attribute Cancelled numeric
@attribute DistinctUnitOrders numeric
@attribute Quantity numeric
@attribute TotalUnitsCost numeric

@data
```

The resultant visualisation of the inputs can be seen below:



Runs and Results

The StockCode attribute was converted to nominal using the unsupervised 'Strings to nominal' filter in WEKA. The simplest clustering methodology, k-means, was utilised to then cluster results; the results (4 clusters specified) can be seen below:

Final cluster centroids:					
Attribute	Full Data (20261.0)	Cluster#			
		0 (5518.0)	1 (4540.0)	2 (4846.0)	3 (5357.0)
StockCode	10125	10002	23000	22746	22865
Season	Winter	Winter	Spring	Summer	Autumn
DistinctUnitOrders	26.1638	23.3938	22.2837	22.5922	35.5361
Quantity	275.8194	245.6403	237.8344	250.7893	361.7398
TotalUnitsCost	526.4639	489.4582	445.6599	463.4025	690.1082

These results add no value and generate no additional insight; due to the way pre-processing was handled to this point, the resultant clusters are simply split into their seasonal components as that is a field provided to the algorithm. The full, disaggregated version of the data was reverted to at this stage.

Following a review of the disaggregated data, further outliers were removed including a record cited as 'Amazon fee'. Discretisation of attributes and conversion of strings to nominals was utilised in WEKA to further pre-process this data. Attempts to classify the description field using J48 and string to word vector were unsuccessful due to hardware constraints or improper methodology.

String to word vector was used to generate additional attributes that denoted the presence of common descriptive words in a row; these were used in the clustering process, seen below:

Final cluster centroids:						
Attribute	Cluster#					
	Full Data (530095.0)	0 (97307.0)	1 (94753.0)	2 (276124.0)	3 (44816.0)	4 (17095.0)
=====						
Quantity	10.2349	10.727	11.3243	10.0745	8.5117	8.503
IsSpring	0.1909	1	0	0	0	0.2287
IsSummer	0.2064	0	1	0	0.2541	0.1908
IsAutumn	0.3588	0	0	0.5952	0.4783	0.2592
IsWinter	0.2439	0	0	0.4048	0.2676	0.3213
UnitPrice	3.8447	4.079	3.8339	3.9251	3.4346	2.3456
InternationalSale	0.0849	0.0829	0.0991	0.0791	0.1128	0.0362
BAG	0.0966	0.1098	0.1497	0.0926	0.0165	0
BLUE	0.0358	0.041	0.0395	0.0301	0.0043	0.1606
BOTTLE	0.0218	0.0069	0.0114	0.0356	0	0
BOX	0.0451	0.0446	0.0464	0.0544	0.0023	0.001
CAKE	0.0395	0.0406	0.0314	0.0338	0.1044	0
CARD	0.0267	0.0311	0.033	0.0288	0.0006	0
CERAMIC	0.0167	0.0183	0.0187	0.0153	0.0243	0
CHRISTMAS	0.0469	0.0025	0.0164	0.0779	0.0322	0.0048
DECORATION	0.0243	0.0242	0.022	0.0284	0.0076	0.0144
DESIGN	0.0557	0.0642	0.0618	0.0503	0.0794	0
GLASS	0.0225	0.0261	0.0262	0.0249	0.0003	0
GREEN	0.0168	0.0192	0.0219	0.0176	0.0025	0.0014
HANGING	0.0314	0.0414	0.0322	0.0327	0.0077	0.0119
HEART	0.073	0.0646	0.0697	0.084	0.0499	0.0212
HOLDER	0.0299	0.0344	0.0322	0.0338	0	0.0048
HOME	0.0103	0.0082	0.0076	0.0128	0.003	0.0168
LUNCH	0.0344	0.0417	0.0557	0.0322	0	0
METAL	0.0393	0.0117	0.0107	0.0071	0	0.9788
PACK	0.0293	0.0331	0.0342	0.0328	0	0
PAISLEY	0.0174	0.0197	0.0209	0.017	0.0149	0
PANTRY	0.0179	0.023	0.0094	0.008	0.0927	0
PAPER	0.0264	0.0362	0.0237	0.0226	0.0448	0
PINK	0.055	0.0688	0.0696	0.0547	0.0158	0.0025
POLKADOT	0.0232	0.0304	0.0332	0.021	0.0094	0
RED	0.0794	0.0884	0.0944	0.0875	0.0075	0.0018
RETROSPOT	0.0649	0.0804	0.0707	0.0615	0.0645	0
ROSE	0.0174	0.0184	0.0194	0.0174	0.0181	0.0002
SET	0.1009	0.0861	0.0031	0	1	0
SIGN	0.0321	0.0041	0.0033	0.0018	0	0.926
SMALL	0.0266	0.0302	0.0295	0.0303	0	0.0008
SPACEBOY	0.0201	0.0197	0.0203	0.0192	0.0338	0
T-LIGHT	0.0267	0.0323	0.0298	0.0296	0.0003	0.0011
TEA	0.0227	0.0236	0.0182	0.0175	0.0657	0.0135
VINTAGE	0.0627	0.0442	0.067	0.0694	0.0764	0
WATER	0.0187	0.0049	0.0076	0.0315	0	0
WHITE	0.0384	0.0413	0.0429	0.0423	0.0122	0.0053
WOODEN	0.0237	0.022	0.0157	0.023	0.0575	0

Clearly, there is a huge amount of noise and a great deal of complexity that make these results difficult to interpret, perhaps meaningless. Additional runs were utilised, including a version substantially cut down with isolated attributes that theoretically should align with seasonal demand:

```

Wrapped clusterer:
kMeans
=====

Number of iterations: 4
Within cluster sum of squared errors: 99084.44538508315

Initial starting points (random):

Cluster 0: {0 6,1 Spring,2 1.25}
Cluster 1: {0 3,1 Summer,2 0.83}
Cluster 2: {0 1,2 12.75}
Cluster 3: {0 1,1 Summer,2 4.13}
Cluster 4: {0 25,1 Summer,2 8.29}

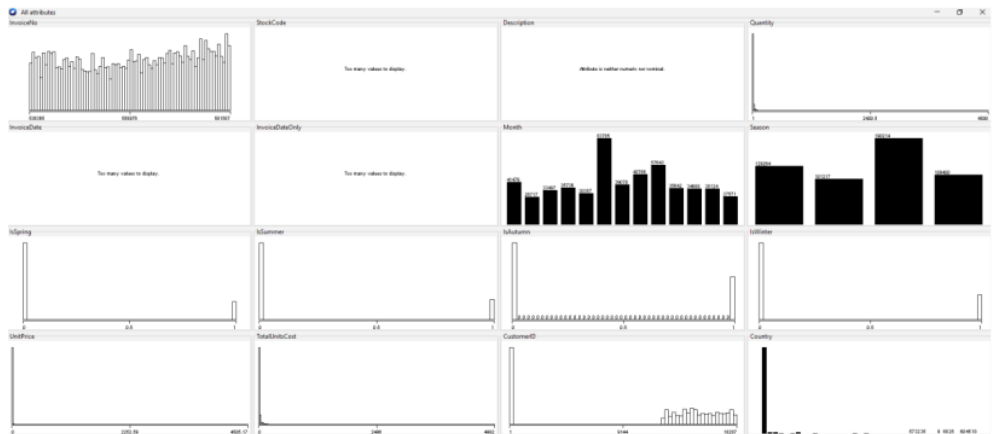
Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute          Cluster#
                   Full Data      0      1      2      3      4
                   (530095.0) (101211.0) (109192.0) (129099.0) (157542.0) (33051.0)
=====
Quantity           10.2349    10.6349    11.0284    9.2636    8.9966    16.0851
Season             Autumn     Spring     Summer     Winter     Autumn     Autumn
UnitPrice          3.8447    4.0066    3.7651    4.1196    3.6534    3.4497
InternationalSale  0.0849    0.0809    0.0999    0.0713    0        0.5049
CARD               0.0267    0.0299    0.0286    0.0246    0.0257    0.0228
CHRISTMAS          0.0469    0.0024    0.0135    0.0409    0        0.5399
HEART              0.073     0.0621    0.0659    0.0819    0.0839    0.0432

```

This doesn't particularly appear to be the case; perhaps because wholesalers may purchase out of season. The full visualisation of the original attributes can be seen below:



Clustering, an unsupervised machine learning technique, seeks to find natural grouping or potential relationships in a dataset [11]. Clustering generally groups unlabelled data, the labelled variant of machine learning is classification [11], which was not used in this analysis, although certainly could and perhaps should have been. There are several clustering techniques, of which k-means has been utilised in this analysis. K-means is an example of partitioning clustering [12] [11] [13]. K represents the number of clusters and can be defined by the user in WEKA.

This was selected for its unsupervised nature, flexibility and high efficiency as the proof-of-concept clustering methodology. Were any well-defined and interesting clusters detected, more complex methodologies would have been considered for further analysis.

No meaningful and useable insight and knowledge was garnered from this undertaking and a significant shift in methodology and execution is fundamentally required to yield significant results that can generate useful business recommendations.

Were clear relationships between types of items and seasons identified, such recommendations could have been put forward regarding the timings of stock acquisition. For example, if Christmas related items were generally purchased in Autumn, stocking them in late Summer could be a viable way to ensure stock is readily available when needed. If Christmas related items saw low order volume throughout Spring and Summer, then they would needlessly consume warehouse space and so efforts to shift them; actions could be taken to mitigate this inefficiency by the business.

If there was a particular period of increased frequency of orders, staffing levels could be altered to improve efficiency by ensuring increased staffing levels were in place during periods of high order volume, and lower staffing levels (if possible) were in place during periods of lower order volume.

It was expected that this type of knowledge and insight would be acquired from undertaking this analysis, but as previously cited, a substantial change in methodology is required to perform sufficient analysis that conclusions such as these could be drawn from.

Reflection

Disappointingly, the data mining undertaken must be considered unsuccessful. No significant clear findings were made, and so astute business advice could not be prescribed in this instance. Utilising output measures such as root mean squared error, or relative error would have been structured ways of measuring accuracy and efficiency and enabled an evaluative framework for the models built.

Unclean data with erroneous entries, colloquial references, blank cells, and other data quality issues leads to a substantial pre-processing workload to be undertaken [14]. This human driven element of pre-processing is complex and can lead to human-based error which can invalidate results. Communicating and presenting results in such an environment where business stakeholders are not necessarily mathematically minded is clearly a crucial challenge to overcome, if it can be overcome, though, substantial knowledge can be derived from raw data and converted into business improvement. The scale at which unsupervised machine learning algorithm's function is vast [15], and they perform tasks that are beyond human capability. The ability to detect patterns, relationships and explain them is unparalleled when leveraged correctly.

In the next cycle, clearer hypotheses would be generated, and methodology clearly established. In addition to this, ascertaining how cancelled orders affect the business could enable data mining that provides knowledge. If there is sufficient evidence that cancelled orders affect the business negatively, and a process could be established to mitigate or eradicate this negative consequence if information could be available ahead of time, I would build a model (regression or classification, regression might be best) that attempted to predict the likelihood that orders that would be cancelled or classify them as "will be cancelled" or "won't be cancelled". Regression would likely be more viable as the value would allow the business to manage this supposed risk as they see fit, or to vary processes according to the value returned.

Natural language processing (NLP) or alternate ML methods could be utilised to classify products into groups based on their text descriptions [16]. These groups could then be used in cluster/classification analysis to build clustering and regression models to support planning, targeted marketing and more within the business.

Bibliography

- [1] M. e. a. Strohbach, "Big data storage," in *New horizons for a data-driven economy*, Springer, 2016, pp. 119-141.
- [2] Y. e. a. Cheng, "Data and knowledge mining with big data towards smart production," *Journal of Industrial Information Integration*, vol. 9, pp. 1-13, 2018.
- [3] M. J. H. Mughal, "Data mining: Web data mining techniques, tools and algorithms: An overview," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 6, 2018.
- [4] W. Raghupathi, "Data mining in healthcare," *Healthcare Informatics: Improving Efficiency through Technology, Analytics, and Management*, pp. 353-372, 2016.
- [5] A. S. Osman, "Data mining techniques," 2019.
- [6] R. S. King, *Cluster analysis and data mining: An introduction*, Stylus Publishing, 2015.
- [7] M. J. a. W. M. J. Zaki, *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*, Cambridge University Press, 2020.
- [8] C. M. e. a. Keet, "The data mining optimization ontology," *Journal of web semantics*, vol. 32, pp. 43-53, 2015.
- [9] UCI, "Online Retail Data Set," UCI, 27 08 2012. [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Online+Retail#>. [Accessed 24 04 2022].
- [10] "Regulation (EU) 2016/679 of the European Parliament and of the Council," UK Government, 2016. [Online]. Available: <https://www.legislation.gov.uk/eur/2016/679/article/17>. [Accessed 07 05 2022].
- [11] Google, "What is Clustering?," Google inc, 02 10 2020. [Online]. Available: <https://developers.google.com/machine-learning/clustering/overview>. [Accessed 26 12 2021].
- [12] L. X. L. a. G. W. R Li, "Global Optimal and Minimal Solutions to K-means Cluster Analysis," in *Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp)*, 2015.
- [13] L. R. a. O. Maimon, "Clustering Methods," *Data Mining and Knowledge Discovery Handbook*, pp. 321-352, 2005.
- [14] J. e. a. Liu, "Rethinking big data: A review on the data quality and usage issues," *ISPRS journal of photogrammetry and remote sensing*, vol. 115, pp. 134-142, 2016.
- [15] A. e. a. Roohi, "Unsupervised machine learning in pathology: the next frontier," *Surgical Pathology Clinics*, vol. 13, no. 2, pp. 349-358, 2020.
- [16] C. R. X. a. J. Z. Zong, "Text Data Mining," Springer, 2021.

FINAL GRADE

68/100

GENERAL COMMENTS

Instructor

Adam David Moore	
1 engagement with literature comment	available
excellent range of references and cited well throughout the report	20
2 Knowledge and Understanding comment	available
Quite a long introduction, some of the information you provide here would be better suited to the main body of the report, however, the understanding of the material is sound and backed up with citations throughout.	
3. Cognitive and Intellectual skills comment	available
discussion on preprocessing is particularly good, results gained (and their flaws) are covered and justified	25
4. Practical Application Skills comment	available
Main tools used - excel for data cleaning and then imported into a weka ARFF file - adding additional fields for classes may have benefitted you rather than having so many numeric fields	35
5. Transferable Skills for life and Professional practice comment	available
very well structured report no major issues, some usages of informal language, no clear conclusion	15
Total:	68

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14
