

1809781- Data Mining Assignment

by Jessica Marfell

Submission date: 12-May-2022 04:42PM (UTC+0100)

Submission ID: 179748029

File name: 12281_Jessica_Marfell_1809781-_Data_Mining_Assignment_398307_2126116617.docx (7.05M)

Word count: 3709

Character count: 20609

ACCA6026

Data Warehousing and Data Mining

Data Mining Assignment

**Author: Jessica Marfell
Student Number: 1809781
Date: 12/05/2022**

TABLE OF CONTENTS

1	INTRODUCTION.....	3
2	PURPOSE AND PRACTICE OF DATA MINING	4
3	DATA SET DESCRIPTION.....	6
3.1	InvoiceNo-	8
3.2	StockCode-.....	8
3.3	Description-	10
3.4	Quantity-.....	10
3.5	Invoice Date	11
3.6	UnitPrice-.....	12
3.7	CustomerID-.....	14
3.8	Country-.....	15
3.9	Summary.....	17
4	PRE-PROCESSING AND POST-PROCESSED DATA.....	18
4.1	Orders and Cancellations.....	18
4.2	Best Sellers	20
4.3	Customers.....	22
5	SAMPLE RUNS AND RESULTS	23
5.1	Orders and Cancellations.....	23
5.2	Best Sellers	26
5.3	Customer.....	27
5.4	Summary-.....	30
6	BUSINESS RECOMMENDATIONS	31
7	REFLECTION	32
8	REFERENCES.....	33

1 Introduction

The purpose of this report is to document the work carried out for the data mining assignment. The assignment questions have been answered within the sub sections of the report and analysis has been conducted on the Online Retail Data Set [1]. Tableau and PowerBI have been used for the data processing activities.

2 Purpose and Practice of Data Mining

Data mining is a term used within business intelligence activities, it provides companies and individuals meaningful information from their dataset. Data mining extracts information and provides analysis to determine patterns in large data sets [2]. Data mining has been around since the 1990's and is also referred to as knowledge discovery in databases. Its process is a combination of statistical work, artificial intelligence, and machine learning. These three factors contribute to the identification of patterns within data and can lead to data predictions being processed [3].

The purpose of data mining is to process a set of data that can be turned into meaningful information which will lead to informed decisions being made. It's the process of understanding a set of data by identifying patterns and trends from a large amount of data to make informed decisions and potentially predict outcomes through the use of mathematical and statistical analytics which processes the data into a readable form [4].

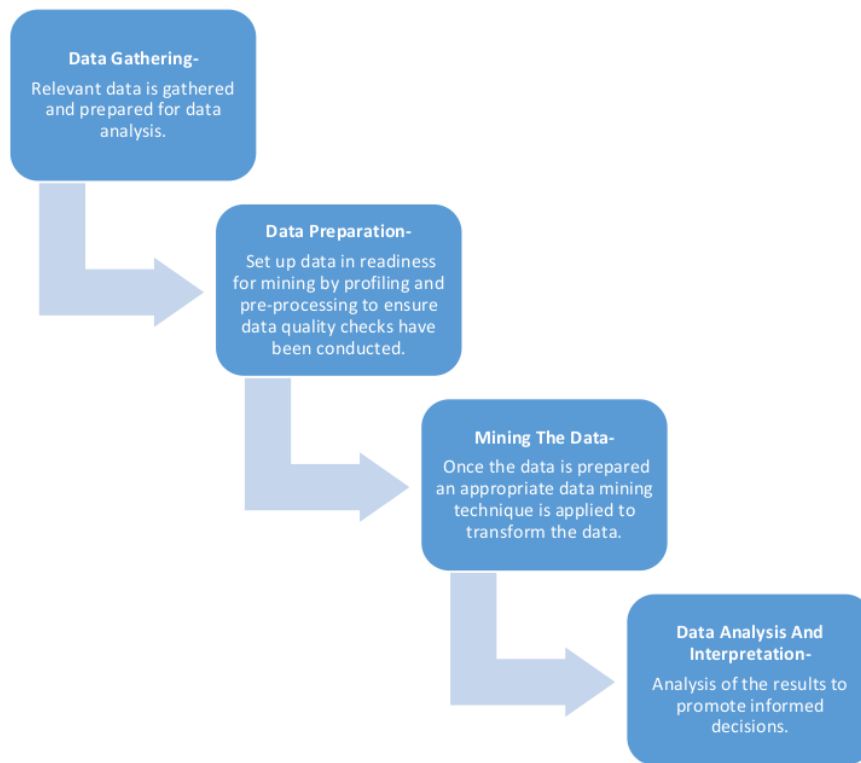
Data mining has advanced during the last decade due to the increase of data availability and automation of data processing. Many firms are using data mining to identify trends within their business. For example, data mining can improve customer services by providing trends on how the target audience interacts with the company and enabling sale forecasts and marketing strategies. Patterns within the data can provide the company with information on when, how much and where certain target audiences purchase promotional discounts.

Data mining is utilised within a variety of industries with the goal to better understand customers and businesses. Some examples are service providers, retail, supermarkets, and crime agencies. Matillion [5] highlights the reasons why data mining works within these organisations. Service providers use processes which analyse the customers data usage to report on trends such as the likelihood they may leave the company. This in turn allows them to target their audience with appropriate marketing materials. This is similar for retail and supermarkets, they too analysis trends to promote their sales by offering their customers bespoke offers and loyalty rewards. This process is commonly done within supermarkets as customers shopping trends can identify their lifestyle trends. For example, a shopper starting to purchase baby essentials may be expecting a child. With the use of data mining tools these trends can be identified which leads to customer satisfaction as they will receive customised experiences.

Data mining is not used solely for financial gain, it is extremely crucial within crime and health agencies also. Highlighting trends such as busier times and locations where incidents happen can allow the deployment of police manpower. It can also be utilised within the health industry to promote a healthier population as trends in illnesses can lead to a pro-active approach to healthcare where patients are advised of risks and benefits.

These real-life business examples use the association rules. Pattern mining can identify patterns quickly however deeper data mining searches can offer calculated statistical outcomes from the dataset. Depending on the user requirements will depend on the depth of search required [6].

The data mining process can be broken down into four manageable stages [7]–



There are various data mining techniques that can be applied. With some examples provided below [7]:



Association rule mining- If-then statements are utilised to identify relationships



Classification- defining the dataset into specific categories such as decision trees and naive bayes classifiers.



Clustering- clustering the data into specific groups for example, through the use of k-means clustering.



Regression- calculating the data to predicted values based on the set of variables.



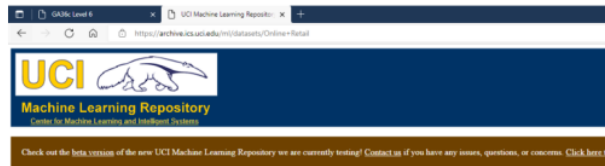
Neural Networks- uses algorithms that stimulate the activity of the human brain such as machine learning.

There are many tools which can support data mining and the identification of patterns within the data. These range from Weka, Rapid Miner to Microsoft products [8].

In summary, data mining is a crucial part for many industries and isn't limited to financial gains. Business requirements will determine the process and techniques used to identify their organisations patterns and trends. Data mining has proven benefits in many industries and is an important factor for business intelligence and making informed decisions.

3 Data Set Description

The dataset used for analysis has been taken from the UCI Machine Learning Repository. The Online Retail Dataset has been downloaded into Tableau and PowerBI. The information displays transactions between 01/12/2010 to 09/12/2011 for a UK-based online retail store which consist of all occasion gifts.



Abstract: This is a transactional 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.			
Data Set Characteristics:	Multivariate, Sequential, Time-Series	Number of Instances:	541909
Attribute Characteristics:	Integer, Real	Number of Attributes:	8
Associated Tasks:	Classification, Clustering	Missing Values?	N/A
		Date Created:	2015-11-08
		Number of Web Hits:	713866

Source:

Dr Daqing Chen, Director, Public Analytics group, chen_d@lboro.ac.uk, School of Engineering, Loughborough University, Loughborough LE11 3AA, UK

Data Set Information:

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536383	85123A	WHITE HANGING H	6	01/12/2010 08:26	2.25	17820	United Kingdom
536383	71053	WHITE METAL LANT	6	01/12/2010 08:26	3.39	17820	United Kingdom
536383	844068	CREAM CUPID HEAL	8	01/12/2010 08:26	2.75	17820	United Kingdom
536383	840296	KNITTED UNION FL	6	01/12/2010 08:26	3.39	17820	United Kingdom
536383	840296	RED WOOLLY HOTT	6	01/12/2010 08:26	3.39	17820	United Kingdom
536383	22752	SET 7 BABUSHKA N	2	01/12/2010 08:26	7.65	17820	United Kingdom
536383	21750	GLASS STAR FROST	6	01/12/2010 08:26	4.25	17820	United Kingdom
536386	22633	HAND WARMER UP	6	01/12/2010 08:26	1.85	17820	United Kingdom
536386	22632	HAND WARMER RE	6	01/12/2010 08:26	1.85	17820	United Kingdom
536387	84879	ASSORTED COLOUR	32	01/12/2010 08:34	1.89	13047	United Kingdom
536387	22745	POPPY'S PLATNOUT	6	01/12/2010 08:34	2.1	13047	United Kingdom

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
541696	23220	SET/4 BADGES DOGS	4	20/03/2011 18:08:00	0.8300	null	United Kingdom
541696	23221	SET/4 BADGES CUTE CREAT...	1	20/03/2011 18:08:00	0.8300	null	United Kingdom
541696	23222	SET/4 BADGES BEETLES	3	20/03/2011 18:08:00	0.8300	null	United Kingdom
541696	23224	SET/4 SKULL BADGES	1	20/03/2011 18:08:00	0.8300	null	United Kingdom
541696	23234	SMALL GLASS HEART TRINK...	1	20/03/2011 18:08:00	4.1300	null	United Kingdom
541696	23236	SMALL CHUNKY GLASS RO...	2	20/03/2011 18:08:00	5.7900	null	United Kingdom
541696	23263	HOME SMALL WOOD LETTE...	1	20/03/2011 18:08:00	10.7900	null	United Kingdom
541696	23263	PACK OF 12 STICKY BUNNIES	2	20/03/2011 18:08:00	1.6300	null	United Kingdom
541696	23285	IVORY HANGING DECORATI...	2	20/03/2011 18:08:00	1.6300	null	United Kingdom
541696	23402	RED EGG SPOON	2	20/03/2011 18:08:00	0.8300	null	United Kingdom
541696	23403	BLUE EGG SPOON	2	20/03/2011 18:08:00	0.8300	null	United Kingdom
541696	23447	12 IVORY ROSE PEG PLACE S...	1	20/03/2011 18:08:00	2.4600	null	United Kingdom
541696	23448	12 DAISY PEGS IN WOOD BOX	1	20/03/2011 18:08:00	3.2900	null	United Kingdom
541696	23452	TOADSTOOL MONEY BOX	2	20/03/2011 18:08:00	5.7900	null	United Kingdom
541696	23454	CHICK GREY HOT WATER BO...	1	20/03/2011 18:08:00	7.6600	null	United Kingdom
541696	23494	ROTATING LEAVES T LIGHT...	7	20/03/2011 18:08:00	1.6300	null	United Kingdom
541696	23504	SKULLS GREETING CARD	1	20/03/2011 18:08:00	0.8300	null	United Kingdom
541696	23506	FANCY FONT BIRTHDAY CARD	1	20/03/2011 18:08:00	0.8300	null	United Kingdom
541696	23537	RED RETROSPOT PUDDING...	1	20/03/2011 18:08:00	8.2900	null	United Kingdom
541696	23555	CERAMIC STRAWBERRY THL...	1	20/03/2011 18:08:00	2.4600	null	United Kingdom
541696	23556	CERAMIC STRAWBERRY MO...	1	20/03/2011 18:08:00	2.4600	null	United Kingdom

Attribute	Information
InvoiceNo	6 digit number that is unique to each transaction. If the letter 'C' appears before the InvoiceNo it indicates a cancellation.
StockCode	5 digit internal number which associates to a distinct product.
Description	Product description.
Quantity	Quantity of each product associated to the transaction.
InvoiceDate	Numeric field generated with the date and time.
UnitPrice	Product price per unit in sterling.
CustomerID	5-digit number uniquely linked to each customer.
Country	Name of country where the customer resides.

The initial observation of the dataset is there's a large amount of information available which has been categorised below. Filters and rules have been utilised within Microsoft Excel, Tableau and PowerBI to examine the information to make the following observations.

3.1 InvoiceNo-

There's a large amount of invoice numbers available with each row associated with an invoice number. The invoice number links up the description and details of the items purchased within the transaction. There are multiple rows displayed which associates to one invoice number.

From the below example, InvoiceNo 536365 is the purchase number and shows the customer purchased 40 items in total between 7 different stock codes at 08.26 on the 01/12/2010. The information stores the price, location and customer id which can be linked to other transactions they may have made. There are no blank entries within this field.

A	B	C	D	E	F	G	H	I
InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01/12/2010 08:26	2.55	17850	United Kingdom	
536365	71053	WHITE METAL LANTERN	6	01/12/2010 08:26	3.39	17850	United Kingdom	
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01/12/2010 08:26	2.75	17850	United Kingdom	
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01/12/2010 08:26	3.39	17850	United Kingdom	
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01/12/2010 08:26	3.39	17850	United Kingdom	
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	01/12/2010 08:26	7.65	17850	United Kingdom	
536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	01/12/2010 08:26	4.25	17850	United Kingdom	

3.2 StockCode-

The StockCode vary from numbers to letters. It is observed that many of the companies Stockcodes are referenced in a similar format to- 10133. All Stockcodes with a similar format have information

linked. For example, StockCode 10133 displays a brown colouring pencil with the associated invoices and customer IDs. It also displays there were 82 items damaged, logged on the 16/11/2011.

A	B	C	D	E	F	G	H
InvoiceN	StockCo	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
576843	10133	damaged	-82	16/11/2011 15:32	£0.00		United Kingdom
539240	10133	COLOURING PENCILS BROWN TUBE	20	16/12/2010 13:15	£0.42 15194		United Kingdom
539252	10133	COLOURING PENCILS BROWN TUBE	20	16/12/2010 14:17	£0.42 17744		United Kingdom
539353	10133	COLOURING PENCILS BROWN TUBE	20	17/12/2010 11:30	£0.42 12782		Portugal
540595	10133	COLOURING PENCILS BROWN TUBE	20	10/01/2011 11:35	£0.42 14321		United Kingdom
541594	10133	COLOURING PENCILS BROWN TUBE	20	19/01/2011 15:30	£0.42 14916		United Kingdom
541699	10133	COLOURING PENCILS BROWN TUBE	40	21/01/2011 09:03	£0.42 17744		United Kingdom
555558	10133	COLOURING PENCILS BROWN TUBE	10	05/06/2011 14:06	£0.42 14159		United Kingdom
555741	10133	COLOURING PENCILS BROWN TUBE	10	07/06/2011 10:05	£0.42 16817		United Kingdom
555820	10133	COLOURING PENCILS BROWN TUBE	20	07/06/2011 12:01	£0.42 16133		United Kingdom
555845	10133	COLOURING PENCILS BROWN TUBE	30	07/06/2011 13:19	£0.42 17015		United Kingdom
555853	10133	COLOURING PENCILS BROWN TUBE	2	07/06/2011 13:44	£0.42 14375		United Kingdom
555859	10133	COLOURING PENCILS BROWN TUBE	20	07/06/2011 14:22	£0.42 17967		United Kingdom
555866	10133	COLOURING PENCILS BROWN TUBE	20	07/06/2011 14:38	£0.42 16841		United Kingdom
556038	10133	COLOURING PENCILS BROWN TUBE	10	08/06/2011 12:38	£0.42 13184		United Kingdom
556103	10133	COLOURING PENCILS BROWN TUBE	1	08/06/2011 16:18	£0.42 16011		United Kingdom
556193	10133	COLOURING PENCILS BROWN TUBE	20	09/06/2011 12:04	£0.42 14132		United Kingdom
556214	10133	COLOURING PENCILS BROWN TUBE	20	09/06/2011 14:14	£0.42 14913		United Kingdom
556486	10133	COLOURING PENCILS BROWN TUBE	20	12/06/2011 13:22	£0.42 17422		United Kingdom
556504	10133	COLOURING PENCILS BROWN TUBE	20	13/06/2011 08:47	£0.42 13936		United Kingdom
556536	10133	COLOURING PENCILS BROWN TUBE	10	13/06/2011 12:36	£0.42 17795		United Kingdom
556580	10133	COLOURING PENCILS BROWN TUBE	20	13/06/2011 14:20	£0.42 17719		United Kingdom
556876	10133	COLOURING PENCILS BROWN TUBE	4	15/06/2011 11:56	£0.42 16241		United Kingdom
556910	10133	COLOURING PENCILS BROWN TUBE	10	15/06/2011 13:25	£0.42 12748		United Kingdom
557009	10133	COLOURING PENCILS BROWN TUBE	20	16/06/2011 11:38	£0.42 14321		United Kingdom
557026	10133	COLOURING PENCILS BROWN TUBE	20	16/06/2011 13:14	£0.42 16571		United Kingdom
557289	10133	COLOURING PENCILS BROWN TUBE	12	19/06/2011 12:58	£0.42 16880		United Kingdom
557612	10133	COLOURING PENCILS BROWN TUBE	49	21/06/2011 12:40	£0.42 14460		United Kingdom
557770	10133	COLOURING PENCILS BROWN TUBE	20	22/06/2011 14:41	£0.42 14643		United Kingdom
557787	10133	COLOURING PENCILS BROWN TUBE	2	22/06/2011 16:16	£0.42 15041		United Kingdom
558049	10133	COLOURING PENCILS BROWN TUBE	1	24/06/2011 14:16	£0.42 15850		United Kingdom

There's a number of outliers within the dataset. There appears to be abbreviations for example M which means Manual and D for discount. When filtered there are many other Stockcodes which are gift tagged and missing descriptions. The inconsistencies within the StockCode format makes it harder to link the transactions to the company's stock.

A	B	C	D	E	F	G	H
InvoiceN	StockCo	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
A563186	B	Adjust bad debt	1	12/08/2011 14:51	-£11,062.06		United Kingdom
A563187	B	Adjust bad debt	1	12/08/2011 14:52	-£11,062.06		United Kingdom
540699	POST		1000	11/01/2011 09:32	£0.00		United Kingdom
542529	DCGS0055		-1	28/01/2011 13:08	£0.00		United Kingdom
542531	DCGS0072		-1	28/01/2011 13:08	£0.00		United Kingdom
542532	DCGS0074		-1	28/01/2011 13:09	£0.00		United Kingdom
542582	DCGS0057		-6	28/01/2011 15:01	£0.00		United Kingdom
547966	DOT		1000	28/03/2011 15:49	£0.00		United Kingdom
552230	DOT	DOTCOM POSTAGE	1	06/05/2011 15:43	£0.00		United Kingdom
554857	POST		800	27/05/2011 10:08	£0.00		United Kingdom
556955	gift_0001_20	to push order through as stock was	10	16/06/2011 09:04	£0.00		United Kingdom
561249	DCGS0073	ebay	-4	26/07/2011 11:51	£0.00		United Kingdom
561250	DCGS0071		-2	26/07/2011 11:51	£0.00		United Kingdom
561251	DCGS0070		-8	26/07/2011 11:51	£0.00		United Kingdom
561252	DCGS0069	ebay	-5	26/07/2011 11:52	£0.00		United Kingdom
561253	DCGS0068	ebay	-10	26/07/2011 11:52	£0.00		United Kingdom
561254	DCGS0067	ebay	-11	26/07/2011 11:52	£0.00		United Kingdom
561255	DCGS0066P		-3	26/07/2011 11:52	£0.00		United Kingdom
561256	DCGS0003	ebay	-7	26/07/2011 11:53	£0.00		United Kingdom
561916	M	Manual	1	01/08/2011 11:44	£0.00 15581		United Kingdom
564761	gift_0001_30		30	30/08/2011 10:48	£0.00		United Kingdom
564762	gift_0001_10		30	30/08/2011 10:48	£0.00		United Kingdom
565556	POST		750	05/09/2011 12:14	£0.00		United Kingdom
568158	PADS	PADS TO MATCH ALL CUSHIONS	1	25/09/2011 12:22	£0.00 16133		United Kingdom
568384	M	Manual	1	27/09/2011 09:46	£0.00 12748		United Kingdom
571035	M	Manual	1	13/10/2011 12:50	£0.00 12446		RSA
574252	M	Manual	1	03/11/2011 13:24	£0.00 12437		France
575505	POST		800	10/11/2011 10:29	£0.00		United Kingdom
575506	C2		150	10/11/2011 10:30	£0.00		United Kingdom
577168	M	Manual	1	18/11/2011 10:42	£0.00 12603		Germany
577696	M	Manual	1	21/11/2011 11:57	£0.00 16406		United Kingdom
580366	DOT	DOTCOM POSTAGE	1	02/12/2011 16:38	£0.00		United Kingdom
550193	PADS	PADS TO MATCH ALL CUSHIONS	1	15/04/2011 09:27	£0.00 13952		United Kingdom
561226	PADS	PADS TO MATCH ALL CUSHIONS	1	26/07/2011 10:13	£0.00 15618		United Kingdom
568200	PADS	PADS TO MATCH ALL CUSHIONS	1	25/09/2011 14:58	£0.00 16198		United Kingdom

3.3 Description-

A variety of descriptions are used within the online retail company. Many have meaningful information, for example- 50's Christmas Gift Bag Large. This information provides the right amount of data within a limited space.

564972	23437	50'S CHRISTMAS GIFT BAG LARGE	12	31/08/2011 15:30	£1.25	12621	Germany
565088	23437	50'S CHRISTMAS GIFT BAG LARGE	12	01/09/2011 10:11	£1.25	16145	United Kingdom
565648	23437	50'S CHRISTMAS GIFT BAG LARGE	3	05/09/2011 16:22	£1.25	18196	United Kingdom

However, there appears to be a variety of incomplete information which range from '?', numbers, incomplete titles, and many more poor examples.

572891	23343		20713	-400	26/10/2011 14:14	£0.00	United Kingdom
537032	21275	?		-30	03/12/2010 16:50	£0.00	United Kingdom

In addition, there are 1454 rows of blank item descriptions for a range of Stockcodes.

A	B	C	D	E	F	G	H
InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536414	22139		56	01/12/2010 11:52	£0.00		United Kingdom
536545	21134		1	01/12/2010 14:32	£0.00		United Kingdom
536546	22145		1	01/12/2010 14:33	£0.00		United Kingdom
536547	37509		1	01/12/2010 14:33	£0.00		United Kingdom
536549	85226A		1	01/12/2010 14:34	£0.00		United Kingdom
536550	85044		1	01/12/2010 14:34	£0.00		United Kingdom
536552	20950		1	01/12/2010 14:34	£0.00		United Kingdom
536553	37461		3	01/12/2010 14:35	£0.00		United Kingdom
536554	84670		23	01/12/2010 14:35	£0.00		United Kingdom
536589	21777		-10	01/12/2010 16:50	£0.00		United Kingdom
536764	84952C		-38	02/12/2010 14:42	£0.00		United Kingdom
536765	84952C		19	02/12/2010 14:43	£0.00		United Kingdom
536995	35951		57	03/12/2010 15:29	£0.00		United Kingdom
536996	22712		-20	03/12/2010 15:30	£0.00		United Kingdom
536997	22028		-20	03/12/2010 15:30	£0.00		United Kingdom
536998	85067		-6	03/12/2010 15:30	£0.00		United Kingdom
536999	21421		110	03/12/2010 15:32	£0.00		United Kingdom
537000	21414		-22	03/12/2010 15:32	£0.00		United Kingdom
537001	21653		-6	03/12/2010 15:33	£0.00		United Kingdom
537002	21651		70	03/12/2010 15:33	£0.00		United Kingdom
537003	85126		-2	03/12/2010 15:33	£0.00		United Kingdom
537004	21814		-30	03/12/2010 15:34	£0.00		United Kingdom
537005	21692		-70	03/12/2010 15:35	£0.00		United Kingdom
537006	21648		-130	03/12/2010 15:36	£0.00		United Kingdom
537007	21172		-80	03/12/2010 15:36	£0.00		United Kingdom
537008	21161		-120	03/12/2010 15:37	£0.00		United Kingdom
537009	84534B		-80	03/12/2010 15:38	£0.00		United Kingdom
537010	22162		-40	03/12/2010 15:38	£0.00		United Kingdom
537011	22139		-5	03/12/2010 15:38	£0.00		United Kingdom
537012	21535		-12	03/12/2010 15:39	£0.00		United Kingdom
537013	35965		-25	03/12/2010 15:40	£0.00		United Kingdom
537014	21040		-20	03/12/2010 15:40	£0.00		United Kingdom

The overall observation for the company's description information is poor. This is down to a high number of blank entries and poor data quality on many other items.

3.4 Quantity-

The Quantity attribute range from -80995 to 80995. There are many minus numbers identified within the dataset with informal notes typed into the description and missing data. For example, quality -1479 has a note that states- mystery! Only ever imported 1800. This information is not useful to the user or decision maker as it is incomplete.

A	B	C	D	E	F	G	H
InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
C581484	23843	PAPER CRAFT , LITTLE BIRDIE	-80995	09/12/2011 09:27	£2.08	16446	United Kingdom
C541433	23166	MEDIUM CERAMIC TOP STORAGE JAR	-74215	18/01/2011 10:17	£1.04	12346	United Kingdom
556690	23005	printing smudges/thrown away	-9600	14/06/2011 10:37	£0.00		United Kingdom
556691	23005	printing smudges/thrown away	-9600	14/06/2011 10:37	£0.00		United Kingdom
C536757	84347	ROTATING SILVER ANGELS T-LIGHT HLDR	-9360	02/12/2010 14:23	£0.03	15838	United Kingdom
556687	23003	Printing smudges/thrown away	-9058	14/06/2011 10:36	£0.00		United Kingdom
546152	72140F	throw away	-5368	09/03/2011 17:25	£0.00		United Kingdom
573596	79323W	Unsaleable, destroyed.	-4830	31/10/2011 15:17	£0.00		United Kingdom
566768	16045		-3667	14/09/2011 17:53	£0.00		United Kingdom
565304	16259		-3167	02/09/2011 12:18	£0.00		United Kingdom
C550456	21108	FAIRY CAKE FLANNEL ASSORTED COLOUR	-3114	18/04/2011 13:08	£2.10	15749	United Kingdom
560039	20713	wrongly marked. 23343 in box	-3100	14/07/2011 14:27	£0.00		United Kingdom
545990	84598	check	-3000	08/03/2011 13:07	£0.00		United Kingdom
569466	23270	incorrect stock entry.	-2880	04/10/2011 11:42	£0.00		United Kingdom
560600	18007		-2834	19/07/2011 17:04	£0.00		United Kingdom
573593	79323LP	Unsaleable, destroyed.	-2618	31/10/2011 15:16	£0.00		United Kingdom
540564	22617	mouldy, thrown away.	-2600	10/01/2011 10:36	£0.00		United Kingdom
547560	72732	thrown away-can't sell	-2472	23/03/2011 17:28	£0.00		United Kingdom
552733	23059	Thrown away-rusty	-2376	11/05/2011 10:48	£0.00		United Kingdom
573594	79323P	Unsaleable, destroyed.	-2003	31/10/2011 15:16	£0.00		United Kingdom
C550456	21175	GIN + TONIC DIET METAL SIGN	-2000	18/04/2011 13:08	£1.85	15749	United Kingdom
C550456	85123A	WHITE HANGING HEART T-LIGHT HOLDER	-1930	18/04/2011 13:08	£2.55	15749	United Kingdom
565310	21829	?missing	-1897	02/09/2011 12:26	£0.00		United Kingdom
558379	22618	mouldy, unsaleable.	-1681	28/06/2011 16:34	£0.00		United Kingdom
573591	79323B	Unsaleable, destroyed.	-1671	31/10/2011 15:15	£0.00		United Kingdom
574816	85036C	damages wax	-1560	07/11/2011 10:37	£0.00		United Kingdom
C556522	22920	HERB MARKER BASIL	-1515	13/06/2011 11:21	£0.55	16938	United Kingdom
551837	79164	damages/credits from ASOS.	-1512	04/05/2011 13:23	£0.00		United Kingdom
576764	16008	check	-1510	16/11/2011 13:10	£0.00		United Kingdom
549527	21620	mystery! Only ever imported 1800	-1479	08/04/2011 16:03	£0.00		United Kingdom
557108	23113	Damaged	-1440	16/06/2011 15:58	£0.00		United Kingdom
557596	23114	Damaged	-1440	21/06/2011 11:33	£0.00		United Kingdom
557597	23115	Damaged	-1440	21/06/2011 11:33	£0.00		United Kingdom
557599	23116	Damaged	-1440	21/06/2011 11:34	£0.00		United Kingdom
557600	23117	Damaged	-1440	21/06/2011 11:35	£0.00		United Kingdom
569464	23268	Incorrect stock entry.	-1440	04/10/2011 11:40	£0.00		United Kingdom

The highest quantity noted is 80995. There are many other Stockcodes with a high quantity value with some association to CustomerIDs. It is assumed that the quantity field is the amount of stock the customer purchased, for example the assumption is made that under invoice no- 581483, customer id 16446 from the UK purchased 80995 paper bags under StockCode 23843 for £2.08 each on the 9th of December 2011.

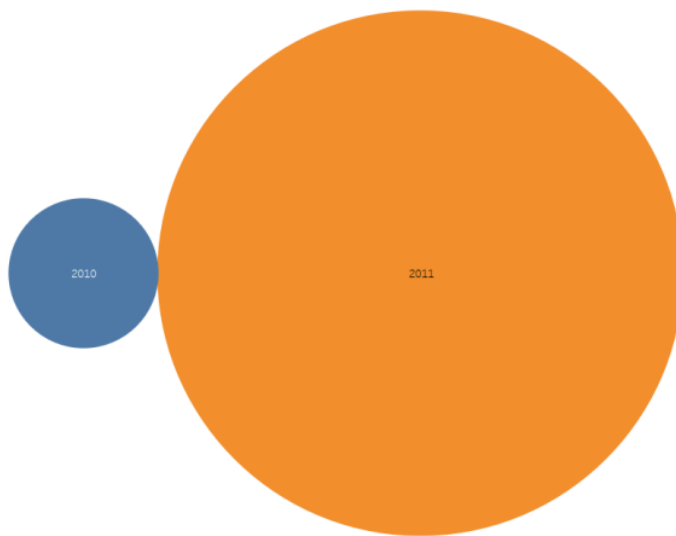
A	B	C	D	E	F	G	H	I
InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	
581483	23843	PAPER CRAFT , LITTLE BIRDIE	80995	09/12/2011 09:15	£2.08	16446	United Kingdom	

On further investigation for InvoiceNo- C581484 the customer made a return of all items as the datasets displays a value of -80995 for the same item.

The outliers in this dataset are the very high and low quantity values linked to the Stockcodes. InvoiceNo with a C in front of is a cancellation. This matches up some of the dataset and answer some of the large quantities and minuses.

3.5 Invoice Date

Data has been captures for 2010 and 2011. The data displayed more invoices for 2011 than 2010 which is expected as the data started to be collated from 01/12/2010.



3.6 UnitPrice-

The Companies unit prices range from -11062.06 to 38970. There is a number of missing fields and poor information displayed in this column. There is assumption made that the information is displayed in pennies, however this would leave outliers within the dataset as some unit prices are displayed as 0.001. If the information was configured to display as currency unit prices this will make the information easier to categorises.

Reviewing the lower unit price there are a number of debt payments and missing stock descriptions for unit prices of 0.

A	B	C	D	E	F	G	H
InvoiceNo	StockCo	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
A563186		B Adjust bad debt	1	12/08/2011 14:51	-11062.06		United Kingdom
A563187		B Adjust bad debt	1	12/08/2011 14:52	-11062.06		United Kingdom
536414	22139		56	01/12/2010 11:52	0		United Kingdom
536545	21134		1	01/12/2010 14:32	0		United Kingdom
536546	22145		1	01/12/2010 14:33	0		United Kingdom
536547	37509		1	01/12/2010 14:33	0		United Kingdom
536549	85226A		1	01/12/2010 14:34	0		United Kingdom
536550	85044		1	01/12/2010 14:34	0		United Kingdom
536552	20950		1	01/12/2010 14:34	0		United Kingdom
536553	37461		3	01/12/2010 14:35	0		United Kingdom
536554	84670		23	01/12/2010 14:35	0		United Kingdom
536589	21777		-10	01/12/2010 16:50	0		United Kingdom
536764	84952C		-38	02/12/2010 14:42	0		United Kingdom
536765	84952C		19	02/12/2010 14:43	0		United Kingdom
536941	22734 amazon		20	03/12/2010 12:08	0		United Kingdom
536942	22139 amazon		15	03/12/2010 12:08	0		United Kingdom
536995	35951		57	03/12/2010 15:29	0		United Kingdom
536996	22712		-20	03/12/2010 15:30	0		United Kingdom
536997	22028		-20	03/12/2010 15:30	0		United Kingdom
536998	85067		-6	03/12/2010 15:30	0		United Kingdom
536999	21421		110	03/12/2010 15:32	0		United Kingdom
537000	21414		-22	03/12/2010 15:32	0		United Kingdom
537001	21653		-6	03/12/2010 15:33	0		United Kingdom
537002	21651		70	03/12/2010 15:33	0		United Kingdom
537003	85126		-2	03/12/2010 15:33	0		United Kingdom
537004	21814		-30	03/12/2010 15:34	0		United Kingdom
537005	21692		-70	03/12/2010 15:35	0		United Kingdom
537006	21648		-130	03/12/2010 15:36	0		United Kingdom
537007	21172		-80	03/12/2010 15:36	0		United Kingdom
537008	21161		-120	03/12/2010 15:37	0		United Kingdom
537009	84534B		-80	03/12/2010 15:38	0		United Kingdom
537010	22162		-40	03/12/2010 15:38	0		United Kingdom
537011	22139		-5	03/12/2010 15:38	0		United Kingdom
537012	21535		-12	03/12/2010 15:39	0		United Kingdom
537013	35965		-25	03/12/2010 15:40	0		United Kingdom
537014	21040		-20	03/12/2010 15:40	0		United Kingdom
537015	37464		-14	03/12/2010 15:41	0		United Kingdom

For the higher unit prices, many items are displayed as a minus quantity although have descriptions associated to the item.

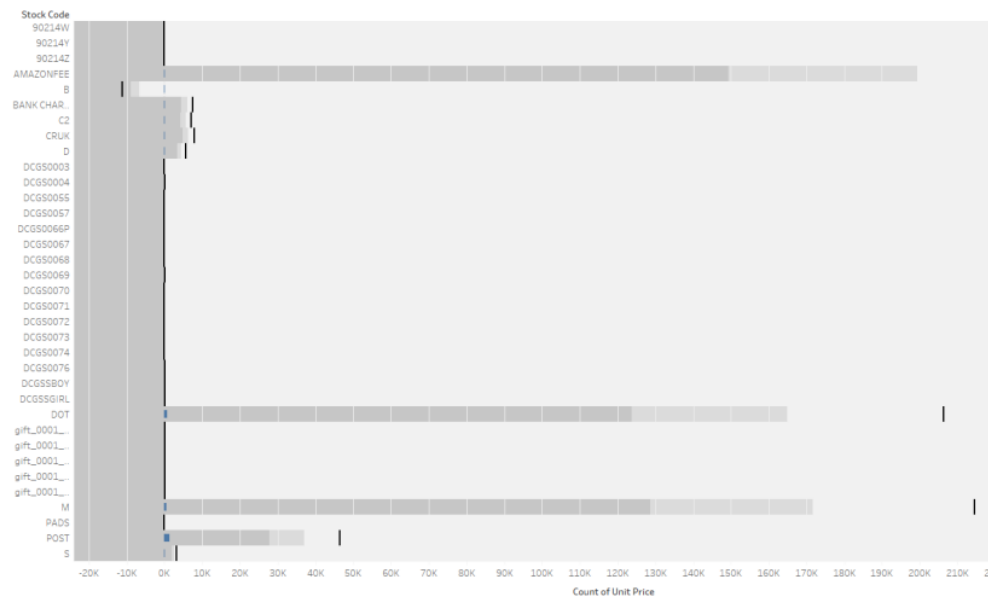
A	B	C	D	E	F	G	H	I
C564340	AMAZONFEE	AMAZON FEE	-1	24/08/2011 14:50	4527.65		United Kingdom	
C558037	AMAZONFEE	AMAZON FEE	-1	24/06/2011 12:33	4534.24		United Kingdom	
C562062	AMAZONFEE	AMAZON FEE	-1	02/08/2011 12:17	4575.64		United Kingdom	
C546989	AMAZONFEE	AMAZON FEE	-1	18/03/2011 12:59	5225.03		United Kingdom	
C544589	AMAZONFEE	AMAZON FEE	-1	21/02/2011 15:11	5258.77		United Kingdom	
C537647	AMAZONFEE	AMAZON FEE	-1	07/12/2010 15:41	5519.25		United Kingdom	
C566889	AMAZONFEE	AMAZON FEE	-1	15/09/2011 13:50	5522.14		United Kingdom	
C544587	AMAZONFEE	AMAZON FEE	-1	21/02/2011 15:07	5575.28		United Kingdom	
C546987	AMAZONFEE	AMAZON FEE	-1	18/03/2011 12:56	5693.05		United Kingdom	
C558036	AMAZONFEE	AMAZON FEE	-1	24/06/2011 12:31	5791.18		United Kingdom	
C553354	AMAZONFEE	AMAZON FEE	-1	16/05/2011 13:54	5876.4		United Kingdom	
C574897	AMAZONFEE	AMAZON FEE	-1	07/11/2011 15:03	5877.18		United Kingdom	
C570025	AMAZONFEE	AMAZON FEE	-1	07/10/2011 10:29	5942.57		United Kingdom	
C573549	AMAZONFEE	AMAZON FEE	-1	31/10/2011 13:23	5942.57		United Kingdom	
C559917	AMAZONFEE	AMAZON FEE	-1	13/07/2011 15:21	6497.47		United Kingdom	
C564341	AMAZONFEE	AMAZON FEE	-1	24/08/2011 14:53	6662.51		United Kingdom	
C537652	AMAZONFEE	AMAZON FEE	-1	07/12/2010 15:51	6706.71		United Kingdom	
C562086	AMAZONFEE	AMAZON FEE	-1	02/08/2011 12:27	6721.37		United Kingdom	
C551699	M Manual		-1	03/05/2011 14:12	6930	16029	United Kingdom	
C553355	AMAZONFEE	AMAZON FEE	-1	16/05/2011 13:58	7006.83		United Kingdom	
C566899	AMAZONFEE	AMAZON FEE	-1	15/09/2011 13:53	7427.97		United Kingdom	
C551685	POST POSTAGE		-1	03/05/2011 12:51	8142.75	16029	United Kingdom	
C551697	POST POSTAGE		1	03/05/2011 13:46	8142.75	16029	United Kingdom	
C574902	AMAZONFEE	AMAZON FEE	-1	07/11/2011 15:21	8286.22		United Kingdom	
A563185	B Adjust bad debt		1	12/08/2011 14:50	11062.06		United Kingdom	
C580604	AMAZONFEE	AMAZON FEE	-1	05/12/2011 11:35	11586.5		United Kingdom	
C537644	AMAZONFEE	AMAZON FEE	-1	07/12/2010 15:34	13474.79		United Kingdom	
C537630	AMAZONFEE	AMAZON FEE	-1	07/12/2010 15:04	13541.33		United Kingdom	
C537632	AMAZONFEE	AMAZON FEE	1	07/12/2010 15:08	13541.33		United Kingdom	
C537651	AMAZONFEE	AMAZON FEE	-1	07/12/2010 15:49	13541.33		United Kingdom	
C540118	AMAZONFEE	AMAZON FEE	-1	05/01/2011 09:57	16453.71		United Kingdom	
C540117	AMAZONFEE	AMAZON FEE	-1	05/01/2011 09:55	16888.02		United Kingdom	
C580605	AMAZONFEE	AMAZON FEE	-1	05/12/2011 11:36	17836.46		United Kingdom	
C556445	M Manual		-1	10/06/2011 15:31	38970	15098	United Kingdom	

Filtering the data, it appears the companies most expensive item is a rustic sideboard for £165 and the cheapest is difficult to determine as the unit prices are set to 0 for many items.

541199	57403	22827 RUSTIC SEVENTEEN DRAWER SIDEBOARD	1 08/11/2011 14:08	165 14535	United Kingdom
541199	577130	22827 RUSTIC SEVENTEEN DRAWER SIDEBOARD	1 17/11/2011 19:58	165 15861	United Kingdom

537466	21488		-50	07/12/2010 10:33	0	United Kingdom
537534	85064	CREAM SWEETHEART LETTER RACK	1	07/12/2010 11:48	0	United Kingdom
537534	94832	ZINC WILLIE WINKIE CANDLE STICK	1	07/12/2010 11:48	0	United Kingdom
537534	94952	BOX OF 24 COCKTAIL PARASOLS	2	07/12/2010 11:48	0	United Kingdom
537534	48184	DOORMAT ENGLISH ROSE	3	07/12/2010 11:48	0	United Kingdom
537534	48111	DOORMAT 3 SMILEY CATS	1	07/12/2010 11:48	0	United Kingdom
537534	22697	GREEN REGENCY TEACUP AND SAUCER	1	07/12/2010 11:48	0	United Kingdom
537534	22682	FRENCH BLUE METAL DOOR SIGN 7	1	07/12/2010 11:48	0	United Kingdom
537534	22680	FRENCH BLUE METAL DOOR SIGN 5	1	07/12/2010 11:48	0	United Kingdom
537534	22681	FRENCH BLUE METAL DOOR SIGN 6	2	07/12/2010 11:48	0	United Kingdom
537534	22679	FRENCH BLUE METAL DOOR SIGN 4	5	07/12/2010 11:48	0	United Kingdom
537534	22686	FRENCH BLUE METAL DOOR SIGN NO	2	07/12/2010 11:48	0	United Kingdom
537534	22683	FRENCH BLUE METAL DOOR SIGN 8	3	07/12/2010 11:48	0	United Kingdom
537534	22676	FRENCH BLUE METAL DOOR SIGN 1	3	07/12/2010 11:48	0	United Kingdom
537534	22625	RED KITCHEN SCALES	1	07/12/2010 11:48	0	United Kingdom
537534	22624	IVORY KITCHEN SCALES	3	07/12/2010 11:48	0	United Kingdom
537534	22619	SET OF 6 SOLDIER SKITTLES	1	07/12/2010 11:48	0	United Kingdom
537534	22520	CHILDS GARDEN TROWEL BLUE	1	07/12/2010 11:48	0	United Kingdom
537534	22524	CHILDS GARDEN GLOVES BLUE	3	07/12/2010 11:48	0	United Kingdom
537534	22502	PCMC BASKET WICKER SMALL	29	07/12/2010 11:48	0	United Kingdom
537534	22501	PCMC BASKET WICKER LARGE	2	07/12/2010 11:48	0	United Kingdom
537534	22476	EMPIRE UNION JACK TV DINNER TRAY	1	07/12/2010 11:48	0	United Kingdom
537534	22473	TV DINNER TRAY VINTAGE FAISLEY	1	07/12/2010 11:48	0	United Kingdom
537534	22474	SPACEBOY TV DINNER TRAY	1	07/12/2010 11:48	0	United Kingdom
537534	22472	TV DINNER TRAY DOLLY GIRL	1	07/12/2010 11:48	0	United Kingdom
537534	22520	CHILDS GARDEN TROWEL BLUE	1	07/12/2010 11:48	0	United Kingdom
537534	22514	CHILDS GARDEN SPADE BLUE	1	07/12/2010 11:48	0	United Kingdom
537534	22516	CHILDS GARDEN RAKE BLUE	1	07/12/2010 11:48	0	United Kingdom
537534	22432	WATERING CAN PINK BUNNY	1	07/12/2010 11:48	0	United Kingdom
537534	22428	ENAMEL FIRE BUCKET CREAM	3	07/12/2010 11:48	0	United Kingdom
537534	22427	ENAMEL FLOWER JUG CREAM	1	07/12/2010 11:48	0	United Kingdom
537534	22376	AIRLINE BAG VINTAGE JET SET WHITE	1	07/12/2010 11:48	0	United Kingdom
537534	22375	AIRLINE BAG VINTAGE JET SET BROWN	1	07/12/2010 11:48	0	United Kingdom
537534	22377	AIRLINE BAG VINTAGE JET SET CHAMPION	1	07/12/2010 11:48	0	United Kingdom

Data processing in Tableau displays many of the higher unit prices are associated with AmazonFees.



3.7 CustomerID-

Using the same InvoiceNo as the example above- 536365 that links to customer 17850. As the CustomerID links to the invoice numbers they have completed with the company it can be seen that customer 17850 has 35 separate transactions completed over a period of two years.

However, on the dataset there are 135080 items which don't have an associated customer ID (blank fields). This could be they are not registered with the company and may have checked out as a guest

on their website. These missing links means the transactions could have been one offs or returning customers. This information is critical in associating customers to their invoice numbers.

A	B	C	D	E	F	G	H
InvoiceNo	StockCo	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536385	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01/12/2010 08:26	2.55	17800	United Kingdom
536385	71053	WHITE METAL LANTERN	6	01/12/2010 08:26	3.39	17800	United Kingdom
536385	844068	CREAM CUPID HEARTS COAT HANGER	8	01/12/2010 08:26	2.75	17800	United Kingdom
536385	840296	KNITTED UNION FLAG HOT WATER BOTTLE	6	01/12/2010 08:26	3.39	17800	United Kingdom
536385	840296	RED WOOLLY HOTTIE WHITE HEART.	6	01/12/2010 08:26	3.39	17800	United Kingdom
536385	22752	SET 7 BABUSHKA NESTING BOXES	2	01/12/2010 08:26	7.65	17800	United Kingdom
536385	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	01/12/2010 08:26	4.25	17800	United Kingdom
536386	22633	HAND WARMER UNION JACK	6	01/12/2010 08:28	1.85	17800	United Kingdom
536386	22632	HAND WARMER RED POLKA DOT	6	01/12/2010 08:28	1.85	17800	United Kingdom
536372	22632	HAND WARMER RED POLKA DOT	6	01/12/2010 09:01	1.85	17800	United Kingdom
536372	22633	HAND WARMER UNION JACK	6	01/12/2010 09:01	1.85	17800	United Kingdom
536373	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01/12/2010 09:02	2.55	17800	United Kingdom
536373	71053	WHITE METAL LANTERN	6	01/12/2010 09:02	3.39	17800	United Kingdom
536373	844068	CREAM CUPID HEARTS COAT HANGER	8	01/12/2010 09:02	2.75	17800	United Kingdom
536373	20679	EDWARDIAN PARASOL RED	6	01/12/2010 09:02	4.95	17800	United Kingdom
536373	37370	RETRO COFFEE MUGS ASSORTED	6	01/12/2010 09:02	1.06	17800	United Kingdom
536373	21871	SAVE THE PLANET MUG	6	01/12/2010 09:02	1.06	17800	United Kingdom
536373	21871	VINTAGE BILLBOARD DRINK ME MUG	6	01/12/2010 09:02	1.06	17800	United Kingdom
536373	21068	VINTAGE BILLBOARD LOVE/HATE MUG	6	01/12/2010 09:02	1.06	17800	United Kingdom
536373	82483	WOOD 2 DRAWER CABINET WHITE FINISH	2	01/12/2010 09:02	4.95	17800	United Kingdom
536373	82488	WOOD 5/3 CABINET ANT WHITE FINISH	4	01/12/2010 09:02	6.95	17800	United Kingdom
536373	82482	WOODEN PICTURE FRAME WHITE FINISH	6	01/12/2010 09:02	2.1	17800	United Kingdom
536373	82494	WOODEN FRAME ANTIQUE WHITE	6	01/12/2010 09:02	2.55	17800	United Kingdom
536373	840296	KNITTED UNION FLAG HOT WATER BOTTLE	6	01/12/2010 09:02	3.39	17800	United Kingdom
536373	840296	RED WOOLLY HOTTIE WHITE HEART.	6	01/12/2010 09:02	3.39	17800	United Kingdom
536373	22752	SET 7 BABUSHKA NESTING BOXES	2	01/12/2010 09:02	7.65	17800	United Kingdom
536373	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	01/12/2010 09:02	4.25	17800	United Kingdom
536375	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01/12/2010 09:32	2.55	17800	United Kingdom
536375	71053	WHITE METAL LANTERN	6	01/12/2010 09:32	3.39	17800	United Kingdom
536375	844068	CREAM CUPID HEARTS COAT HANGER	8	01/12/2010 09:32	2.75	17800	United Kingdom
536375	20679	EDWARDIAN PARASOL RED	6	01/12/2010 09:32	4.95	17800	United Kingdom
536375	37370	RETRO COFFEE MUGS ASSORTED	6	01/12/2010 09:32	1.06	17800	United Kingdom
536375	21871	SAVE THE PLANET MUG	6	01/12/2010 09:32	1.06	17800	United Kingdom
536375	21871	VINTAGE BILLBOARD DRINK ME MUG	6	01/12/2010 09:32	1.06	17800	United Kingdom
536375	21068	VINTAGE BILLBOARD LOVE/HATE MUG	6	01/12/2010 09:32	1.06	17800	United Kingdom
536375	82483	WOOD 2 DRAWER CABINET WHITE FINISH	2	01/12/2010 09:32	4.95	17800	United Kingdom
536375	82488	WOOD 5/3 CABINET ANT WHITE FINISH	4	01/12/2010 09:32	6.95	17800	United Kingdom

Online Retail 312 of 547809 records found

A	B	C	D	E	F	G	H
InvoiceNo	StockCo	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536414	22139		56	01/12/2010 11:52	0		United Kingdom
536544	21773	DECORATIVE ROSE BATHROOM BOTTLE	1	01/12/2010 14:32	2.51		United Kingdom
536544	21774	DECORATIVE CATS BATHROOM BOTTLE	2	01/12/2010 14:32	2.51		United Kingdom
536544	21786	POLKA DOT RAIN HAT	4	01/12/2010 14:32	0.85		United Kingdom
536544	21787	RAIN PONCHO RETROSPOT	2	01/12/2010 14:32	1.66		United Kingdom
536544	21790	VINTAGE SNAP CARDS	9	01/12/2010 14:32	1.66		United Kingdom
536544	21791	VINTAGE HEADS AND TAILS CARD GAME	2	01/12/2010 14:32	2.51		United Kingdom
536544	21801	CHRISTMAS TREE DECORATION WITH BELL	10	01/12/2010 14:32	0.43		United Kingdom
536544	21802	CHRISTMAS TREE HEART DECORATION	9	01/12/2010 14:32	0.43		United Kingdom
536544	21803	CHRISTMAS TREE STAR DECORATION	11	01/12/2010 14:32	0.43		United Kingdom
536544	21809	CHRISTMAS HANGING TREE WITH BELL	1	01/12/2010 14:32	2.51		United Kingdom
536544	21810	CHRISTMAS HANGING STAR WITH BELL	3	01/12/2010 14:32	2.51		United Kingdom
536544	21811	CHRISTMAS HANGING HEART WITH BELL	1	01/12/2010 14:32	2.51		United Kingdom
536544	21821	GLITTER STAR GARLAND WITH BELLS	6	01/12/2010 14:32	7.62		United Kingdom
536544	21822	GLITTER CHRISTMAS TREE WITH BELLS	1	01/12/2010 14:32	4.21		United Kingdom
536544	21823	PAINTED METAL HEART WITH HOLLY BELL	2	01/12/2010 14:32	2.98		United Kingdom
536544	21844	RED RETROSPOT MUG	2	01/12/2010 14:32	5.91		United Kingdom
536544	21851	LILAC DIAMANTE PEN IN GIFT BOX	1	01/12/2010 14:32	4.21		United Kingdom
536544	21870	I CAN ONLY PLEASE ONE PERSON MUG	1	01/12/2010 14:32	3.36		United Kingdom
536544	21871	SAVE THE PLANET MUG	5	01/12/2010 14:32	3.36		United Kingdom
536544	21874	GIN AND TONIC MUG	1	01/12/2010 14:32	3.36		United Kingdom
536544	21879	HEARTS GIFT TAPE	1	01/12/2010 14:32	1.66		United Kingdom
536544	21884	CAKES AND BOWS GIFT TAPE	1	01/12/2010 14:32	1.66		United Kingdom
536544	21888	BINGO SET	1	01/12/2010 14:32	7.62		United Kingdom
536544	21889	WOODEN BOX OF DOMINOES	2	01/12/2010 14:32	2.51		United Kingdom
536544	21892	TRADITIONAL WOODEN CATCH CLIP GAME	3	01/12/2010 14:32	2.51		United Kingdom
536544	21894	POTTING SHED SEED ENVELOPES	1	01/12/2010 14:32	2.51		United Kingdom
536544	21911	GARDEN METAL SIGN	1	01/12/2010 14:32	3.36		United Kingdom
536544	21912	VINTAGE SNAKES & LADDERS	3	01/12/2010 14:32	7.62		United Kingdom
536544	21913	VINTAGE SEASIDE JIGSAW PUZZLES	1	01/12/2010 14:32	7.62		United Kingdom
536544	21914	BLUE HARMONICA IN BOX	3	01/12/2010 14:32	2.51		United Kingdom
536544	21915	RED HARMONICA IN BOX	2	01/12/2010 14:32	2.51		United Kingdom
536544	21922	UNION STRIPE WITH FRINGE HAMMOCK	2	01/12/2010 14:32	16.98		United Kingdom
536544	21928	JUMBO BAG SCANDINAVIAN FAISLEY	1	01/12/2010 14:32	4.21		United Kingdom
536544	21935	SUKI SHOULDER BAG	1	01/12/2010 14:32	3.36		United Kingdom
536544	21942	SKULLS DESIGN FLANNEL	2	01/12/2010 14:32	1.66		United Kingdom
536544	21944	KITTENS DESIGN FLANNEL	1	01/12/2010 14:32	1.66		United Kingdom

Online Retail 135080 of 547809 records found

3.8 Country-

Transactions have been carried out in 36 different countries with 13 invoices not having a specified Country. These invoices total 446 different stock items purchased by 4 customers with allocated customer IDs and 5 invoices from unknown customers.

There are no blank fields displayed for Country.

A	B	C	D	E	F	G	H
InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
553857	23090	VINTAGE GLASS T-LIGHT HOLDER	12	19/05/2011 13:30	0.83		Unspecified
553857	47021G	SET/6 BEAD COASTERS GAUZE BAG GOLD	48	19/05/2011 13:30	0.39		Unspecified
553857	79030D	TUMBLER BAROQUE	24	19/05/2011 13:30	0.39		Unspecified
553857	84877A	PINK ROUND COMPACT MIRROR	24	19/05/2011 13:30	1.25		Unspecified
553857	22178	VICTORIAN GLASS HANGING T-LIGHT	12	19/05/2011 13:30	1.25		Unspecified
553857	84913B	MINT GREEN ROSE TOWEL	8	19/05/2011 13:30	4.65		Unspecified
553857	85125	SMALL ROUND CUT GLASS CANDLESTICK	3	19/05/2011 13:30	4.95		Unspecified
553857	21380	WOODEN HAPPY BIRTHDAY GARLAND	6	19/05/2011 13:30	2.95		Unspecified
553857	47590A	BLUE HAPPY BIRTHDAY BUNTING	3	19/05/2011 13:30	5.45		Unspecified
553857	47590B	PINK HAPPY BIRTHDAY BUNTING	3	19/05/2011 13:30	5.45		Unspecified
553857	47369A	PINK GREEN EMBROIDERY COSMETIC BAG	6	19/05/2011 13:30	2.95		Unspecified
559521	21634	ASSORTED MINI MADRAS NOTEBOOK	24	08/07/2011 16:26	0.39		Unspecified
559521	21212	PACK OF 72 RETROSPOT CAKE CASES	13	08/07/2011 16:26	0.55		Unspecified
559521	22964	3 PIECE SPACEBOY COOKIE CUTTER SET	5	08/07/2011 16:26	2.1		Unspecified
559521	23254	CHILDRENS CUTLERY DOLLY GIRL	5	08/07/2011 16:26	4.15		Unspecified
559521	23256	CHILDRENS CUTLERY SPACEBOY	5	08/07/2011 16:26	4.15		Unspecified
559521	84378	SET OF 3 HEART COOKIE CUTTERS	7	08/07/2011 16:26	1.25		Unspecified
559521	20983	12 PENCILS TALL TUBE RED RETROSPOT	7	08/07/2011 16:26	0.85		Unspecified
559521	20982	12 PENCILS TALL TUBE SKULLS	7	08/07/2011 16:26	0.85		Unspecified
559521	22367	CHILDRENS APRON SPACEBOY DESIGN	6	08/07/2011 16:26	1.95		Unspecified
559521	20981	12 PENCILS TALL TUBE WOODLAND	7	08/07/2011 16:26	0.85		Unspecified
559521	22991	GIRAFFE WOODEN RULER	3	08/07/2011 16:26	1.95		Unspecified
559521	22992	REVOLVER WOODEN RULER	5	08/07/2011 16:26	1.95		Unspecified
559521	21380	WOODEN HAPPY BIRTHDAY GARLAND	1	08/07/2011 16:26	2.95		Unspecified

Tableau has highlighted the area's orders have been processed for and highlights the UK as the main source on the heat map.





3.9 Summary

The above information has been observations made on the attributes detailed on the online retail dataset. Excel, Tableau and PowerBI have been the data processing tools to identify the data characteristics which is displayed below.

Attribute	Overall observation
InvoiceNo	All entries have an allocated InvoiceNo.
StockCode	There is no consistency with the format of Stockcodes.
Description	The description information is poor due to missing values and poor data available in many others.
Quantity	There is a vast amount of minus and plus values for the quantity within the dataset.
InvoiceDate	Invoices have been captured for 2010 and 2011. 2011 is displayed as the busier year.
Unit Price	The unit price is not currency correct as some are displayed as 0.001
CustomerID	135080 items are not linked to a CustomerID
Country	Processed orders from 36 different countries with the heatmap identifying the UK as the main source of orders.

4 Pre-processing and post-processed data

This section will look at the available data within the dataset and the processing requirements to analyse the following areas-

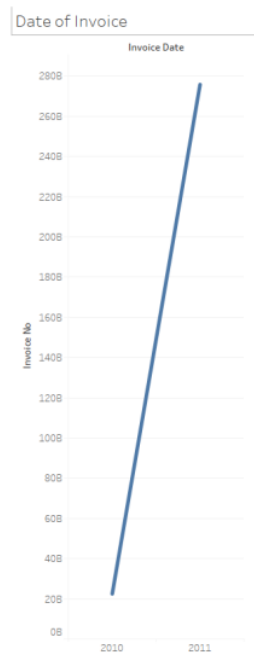


4.1 Orders and Cancellations

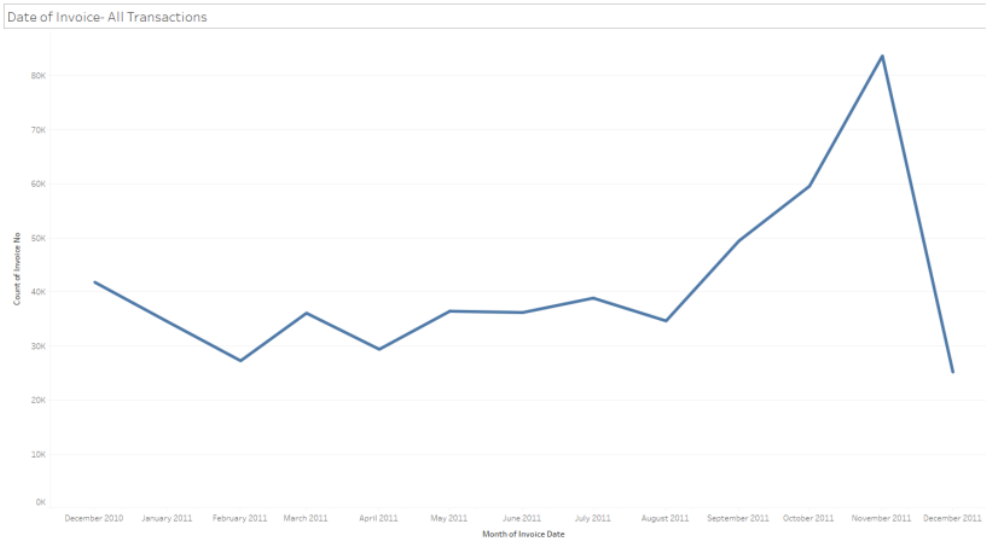
There is a total of 541,909 transactions stored within the dataset, reviewing the attribute information it highlights, invoices beginning with a 'C' are cancellations.

After processing the data to remove the cancellation transactions (9288 entries) the dataset is left with 532,621 sales.

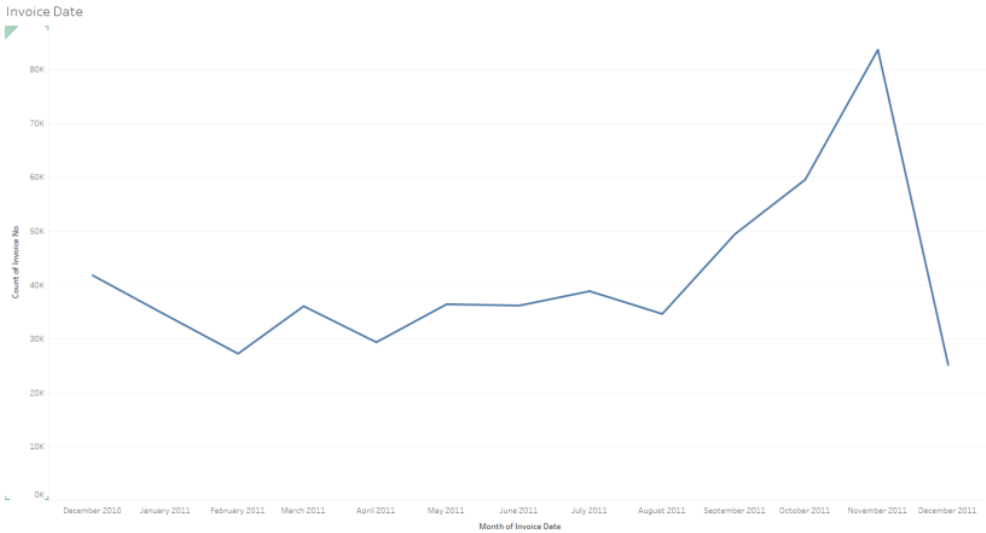
The overall count of invoices displayed between the 2010 and 2011 pre processing the data is displayed below-



Once the information was primary processed to highlight all orders with a count total for the data period more meaningful data is displayed-

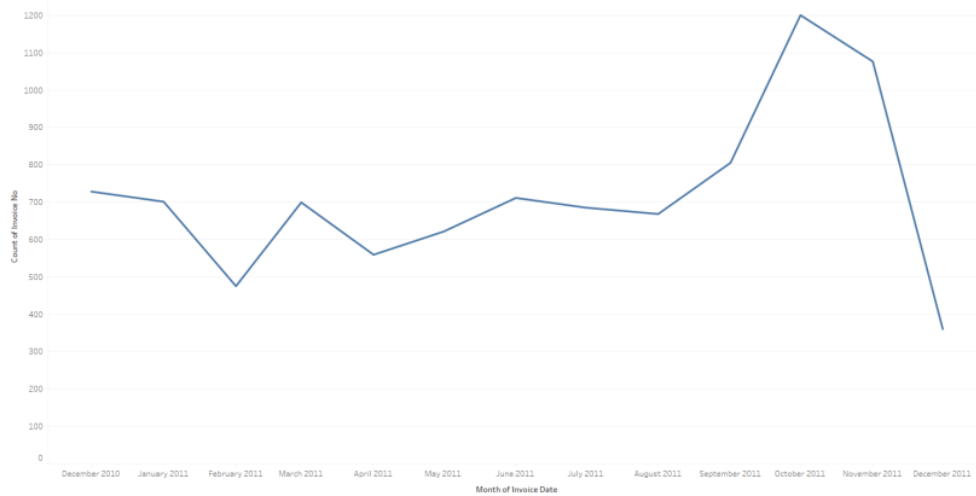


Although this information can identify order trends additional processing is required. Removing the 9288 entries which are cancelled invoices allows the secondary processed data to display accurate information for completed orders.



In order to determine the busier months for the company, incomplete and cancelled orders needed to be removed as they were the data outliers for reporting on the completed orders.

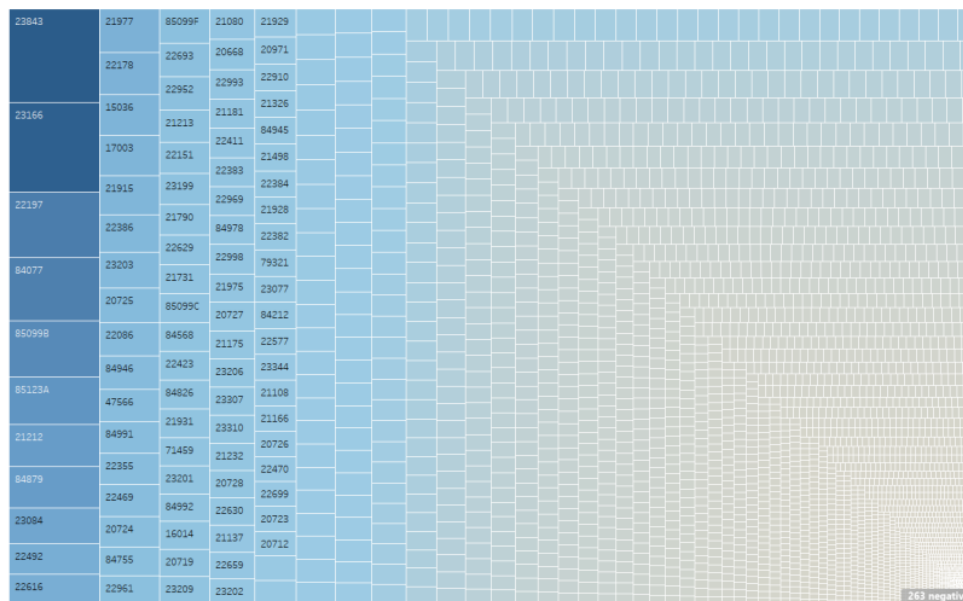
Cancellation Invoice Date



4.2 Best Sellers

The initial data set displays a variety of orders processed and includes cancellations.

Quantity of Stock Ordered-Stock ID Displayed



The data needs to be processed to determine the 263 negative values. Once the data has been processed it displays 1,337 individual entries with 263 different negative values. The stock codes are available however some descriptions identify company waste and blank entries.

23005	20713	72732	85096C	79164	16008	21620	21195	23113	23114	23115	23116	23117	23268
		23059	23269	22740	85099F	22578			23084	84929	35957		
		16259	84611B	85078									
				22351	21622								
23003	23270	85172	37370	23306	21201								
				22130									
		79323P	72807A										
		18007	21700	22127	90057								
72140F	85175	40016	85036B	35954	21391								
		21830	35610A	84997C	21098								
		79323LP		16015	22796								
		22618	79323GR	79323S	85104								
79923W	22617	79323B	85204	79030D									
			79342B	79323G									
16045	47566B	23343		22030									
			35610C	22945									

[illegible]

21 | P a g e

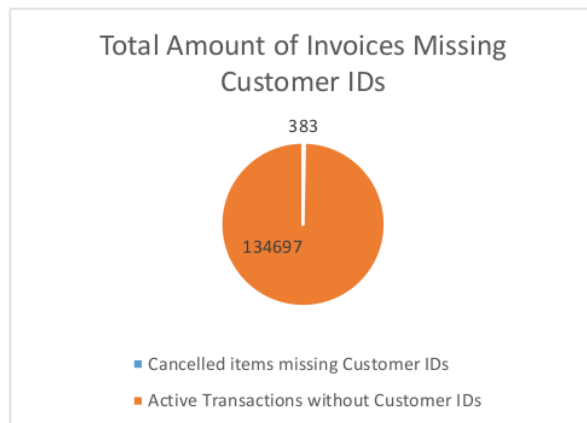
Cancelled Orders Stock Code

23843 Stock Code: 23843 Quantity: -80,995	84347				M	21106	20971	85123A	
	22920	22273	475668	15034	71477	84077	D		
	22161								
	22147								
	22423								
	15036								
	23166								
	23199								
	22102								
	23170								
	22992								
	22991								
	21069								
	21871								
	48185								
	22328								
	22149								

4.3 Customers

There are 135,080 items without associated customer IDs available. This means information gathering on common trends may not be accurate due to lack of customer knowledge.

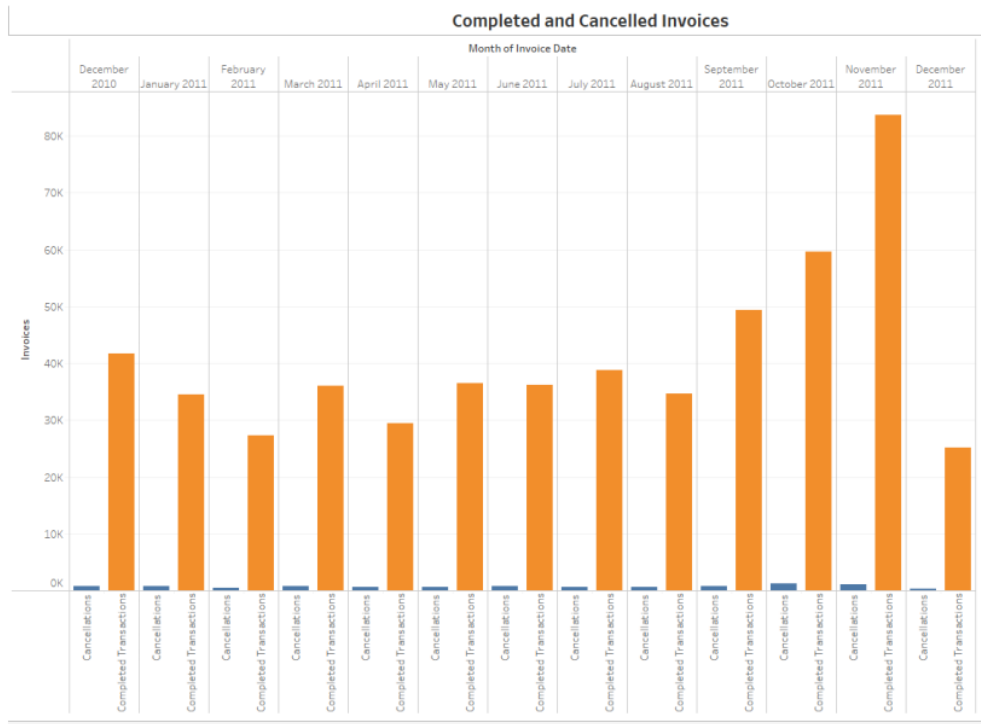
Filtering the dataset to remove 383 cancelled transactions leaves 134,697 active transactions without Customer IDs. Analysis will be carried out on transactions with associated customer IDs to identify returning clients and potential relationships. The second process will allocate the remaining 134,697 transactions with a generic code. This will improve data quality and is advised to be followed within the company as it will easily highlight missing customer IDs without leaving fields blank.



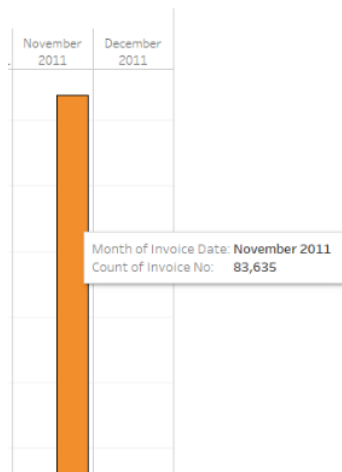
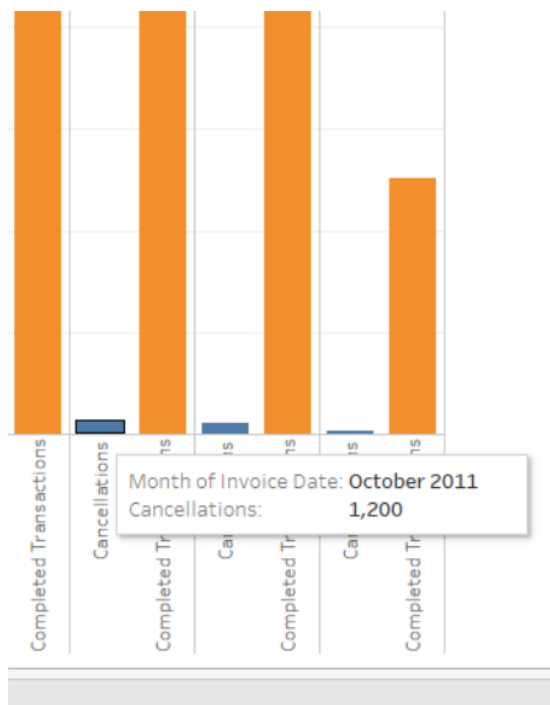
5 Sample Runs and Results

5.1 Orders and Cancellations

In order to compare the cancelled orders, the dataset has been joined together to display the total for cancelled and completed invoices. From this information it can be seen there are significantly more orders completed than cancelled within the company.

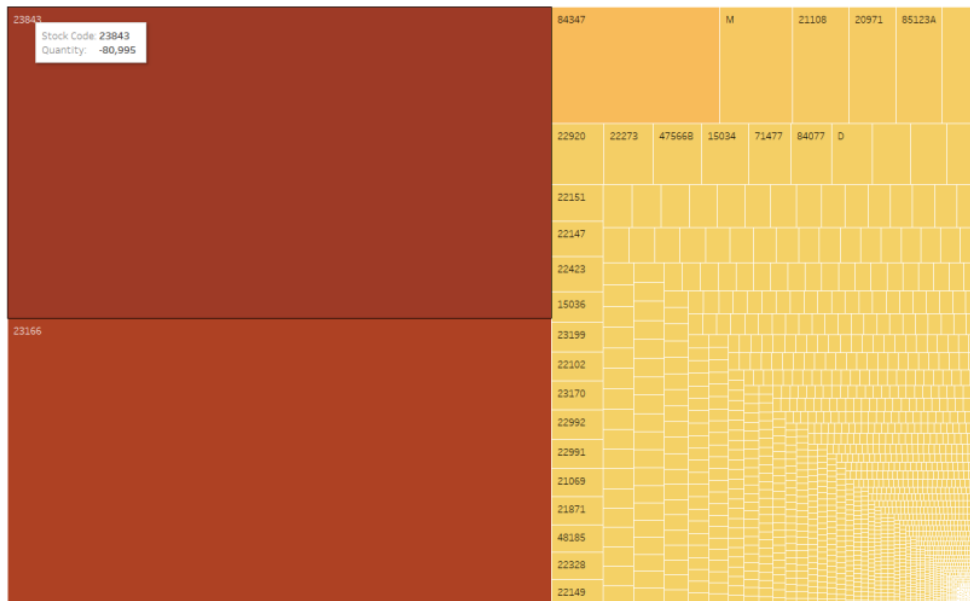


The highest month for cancelled orders was October 2011 with 1,200 cancelled and November was the company's busiest month with 83,635 processed invoices.



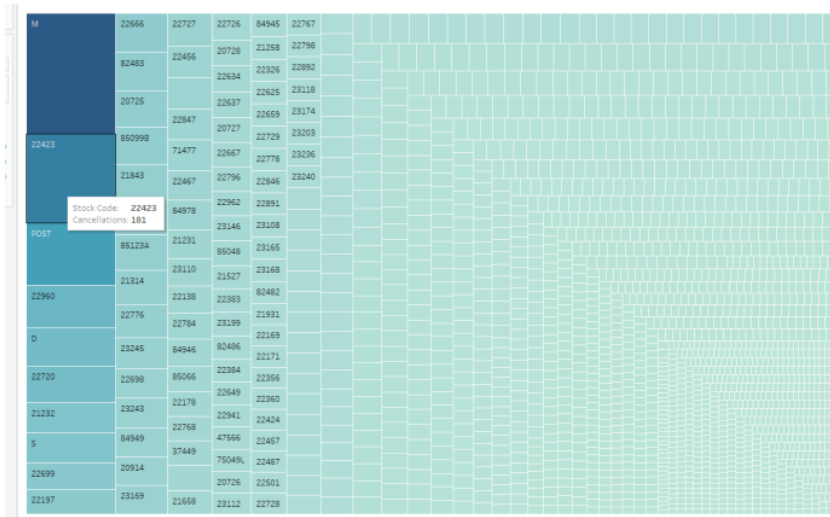
Classifying the data displays, StockCode 23843 was the largest item to be cancelled. These were paper Craft, Little Birdie, however upon reflection this was under one cancelled invoice and not multiple returns.

Cancelled Orders Stock Code

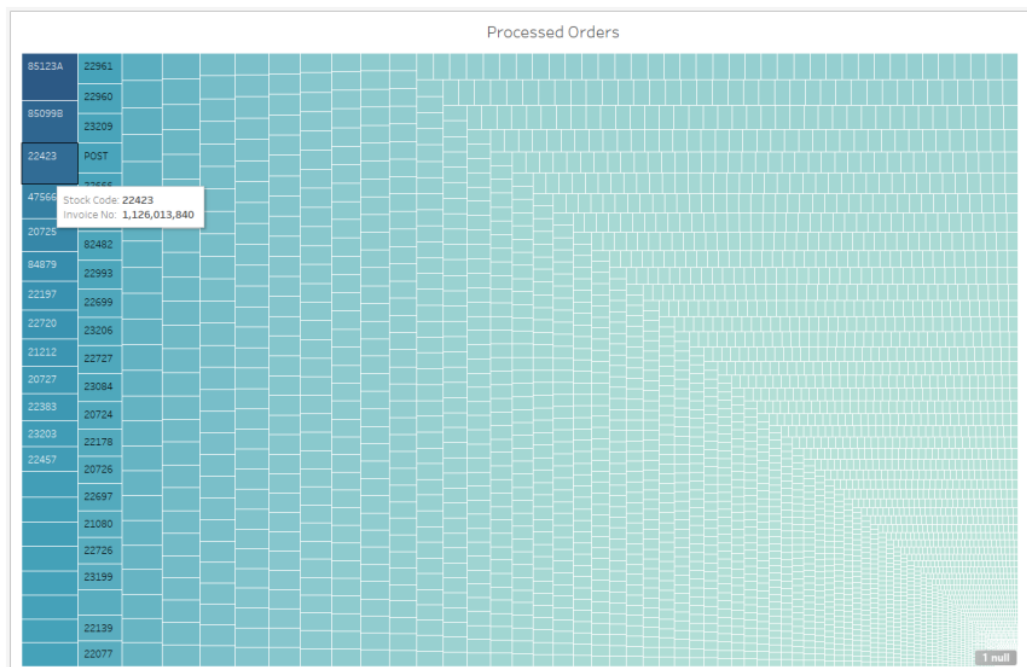


To allow the data to display the highest number of returns on their invoices the data displays all cancelled items and count of InvoiceNo. StockCode M has 244 cancelled invoices against it with the description of manual.

StockCode 22423 is the highest company item returned with 181 cancelled invoices attached to it which is the Regency Cakestand.



In comparison to the orders processed for item 22423 (1,126,013,840) the percentage returned is not a concern for the company.

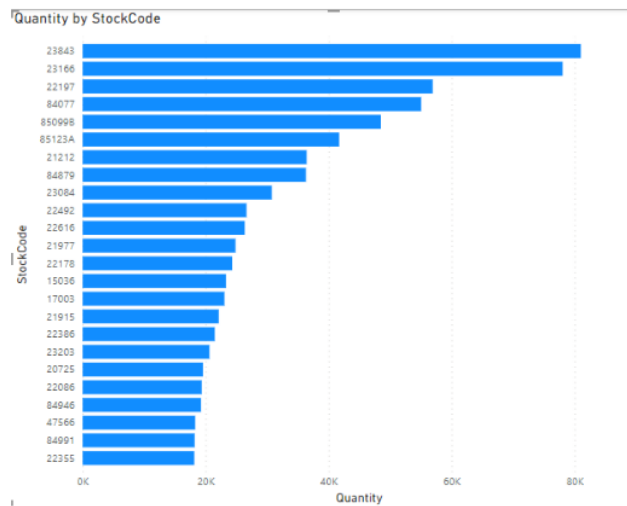


5.2 Best Sellers

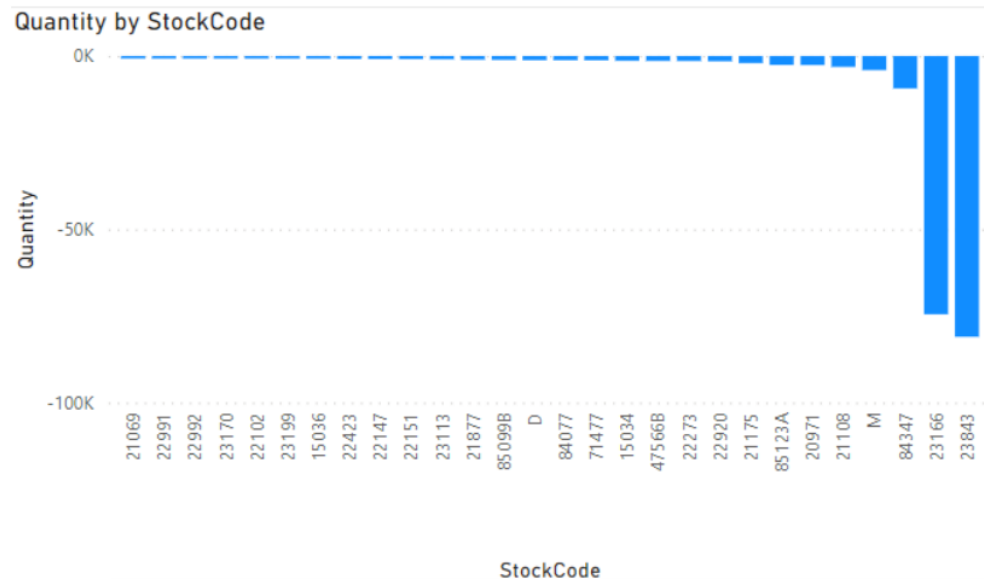
Processed orders and cancellations have identified the active number of orders processed for the company without including their cancellations.

Comparing processed and cancelled orders needs to be identified otherwise there is a risk of incorrect figures being displayed. For example, the below report highlights StockCode 23843 has the highest number of sales however when looking at the companies' cancellations all of these have been cancelled.

Below report displays active Stockcodes-



Below displays cancelled Stockcodes-

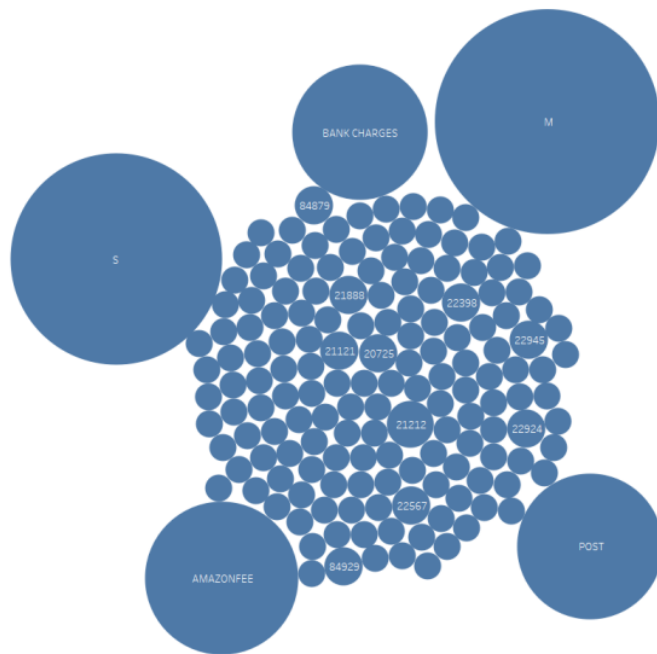


From this information it displays items which have sold more however minus the cancelation of the same item means they are not the companies best seller. StockCode 22197 (Small popcorn holder) has the highest amount sales.

5.3 Customer

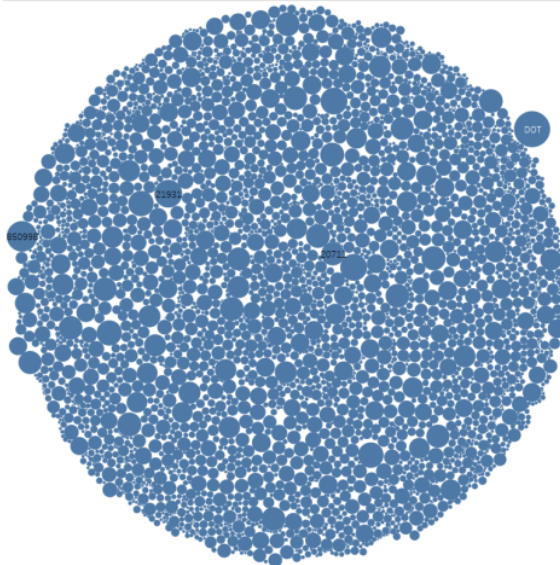
The processed data has been joined together to display 135,080 transactions without a CustomerID. 383 of these have been cancelled transactions for a variety of items. The main outlier for the results is the large number of items that are not stock related, such as AMAZONFEE. The data shows multiple returns for stock items however data quality work is required to ensure the right Stockcodes are used.

Cancelled Transactions Without CustomerID



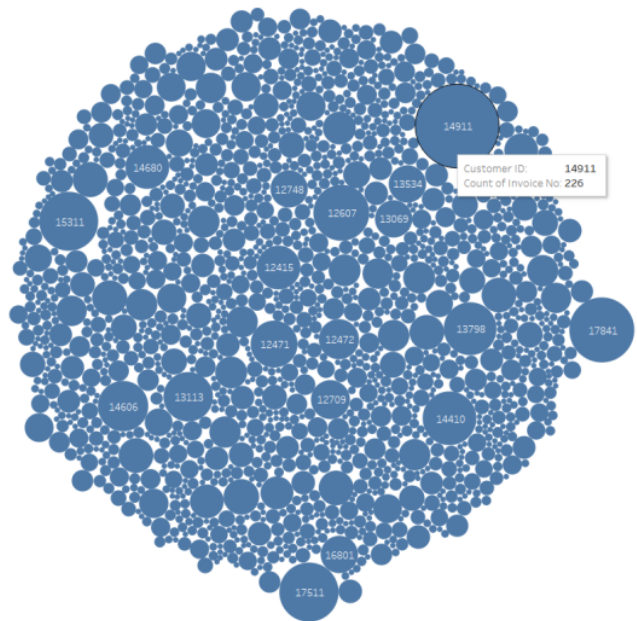
The data shows for example, 497 returns for stock item, 85099B (JUMBO BAG RED RETROSPOT). The below image displays examples and highlights high return rates on specific items. As they are not logged against a CustomerID the relationship can not be made.

Transactions Without a Customer ID



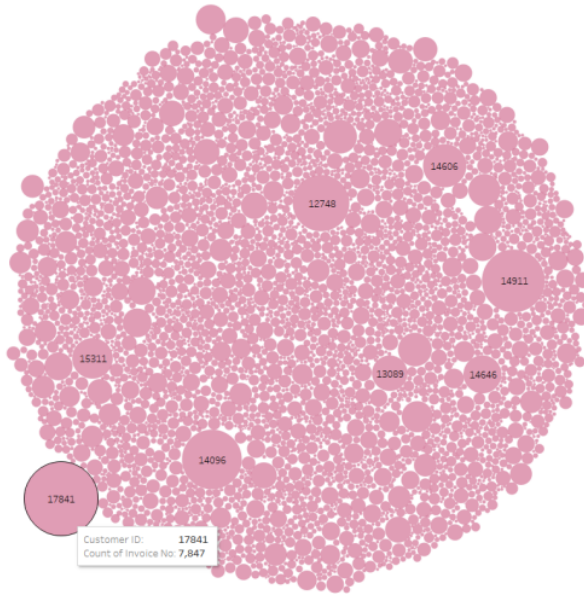
There is a total of 406,829 transactions with CustomerID available. Of those, 8,905 are cancellations. Processing the data displays the top ten customers, with customer 14911 cancelling 226 invoices in total.

Cancelled Items With Associated CustomerID



397,924 active transactions have CustomerID associated to invoices. The data displays many returning customers with one customer returning 7,847 times.

Transactions with CustomerID



5.4 Summary-



-There are low cancellation rates compared to orders.



-November 2011 was the company's busiest month with 83,635 invoices processed.



-There were 135,080 transactions without a CustomerID. Data processes has categorised these customers by validating the field and inserting a generic ID to highlight the trends for all customers.



-Customer ID 17841 returned the highest number of times with 7,847 separate transactions processed with the company.



-The highest cancelled item with 181 cancellations was the Regency Cakestand.



-StockCode 22197 (Small popcorn holder) has the highest amount sales.

6 Business Recommendations

The following business recommendations have been collated after analysing and processing the companies' data-

- The description field should be made mandatory and monitored to improve data quality. This will provide the company with meaningful information on what is associated to what invoices, customers, and stock codes. This field is the descriptive field that joins the information.
- A consistent format of StockCode should be put in place as some codes are numeric and others are text. A recommendation would be to validate the data input by only allowing numeric data.
- When processing the data to determine the best sellers, consideration needs to be given to the completed orders which excludes the cancellations as the information needs to be processed to determine a true reflection on the companies' sales trends.
- The data displays unspecified information for Country with 13 different invoices with 4 of them allocated CustomerID. A suggested improvement would be to ensure when the customer has a linked ID the Country information is populated.
- Validation on the CustomerID information should be made mandatory. This would increase the awareness of customer return rate and a rule could be set up to generate a code if customers do guest checkout a generic number should be inserted into the field which will highlight the number of customers not signed up to the company and using the guest checkout.
- An observation from the findings regarding cancelled invoices would be to introduce a reason for returns code, this field can be non-mandated for invoices not returned but would allow the company to determine if they are continually having faulty items returned or items that are not fit for purpose. From the dataset investigated this information is not present.
- A service improvement for the company would be to update the description field to display more meaningful data as the information doesn't justify the reason there were 244 separate invoices cancelled for StockCode M.
- A review of the cancelled orders in comparison to the active orders will determine if the item is selling and can provide information on the return percentage. It is advised the company reviews a range of cancelled items instead of just the higher proportions as there may be underlining trends for lower purchased items and its cancellation rate.
- Further data validation and comparison of information in different data mining tools may highlight underlining trends. Classification methods have been used throughout the reporting process however the business should investigate other data mining processes such as clustering the data and categorising them into themes. This may highlight different trends and support the ones highlighted above.
- It is recommended that the business captures more information to allow for further data mining. For example, capturing the customers email address will allow promotional offers to be sent to them and relationships between different databases should be considered. The information provided in the data set is one small section of the business and does not link to employee information, address details etc. The company should consider a data warehouse to store their information and to build on the data mining that has been conducted above.

7 Reflection

On reflection the area's investigated have provided sufficient results for the company to review.

One area that needs to be considered for its accuracy is the UnitPrice, it has been processed in sterling to allow for easier categorising however this means the accountancy team will need to be made aware as the pricing may differ depending on exchange rates of the 36 countries it's shipped to. Which may lead to inconsistencies with the data prices. The pricing strategy by the trading government highlights this importance with acknowledgment being made for additional company taxes which may be required when shipping to countries outside of the UK [9].

Introducing data validation rules would improve the efficiency of data being processed. A substantial amount of time was taken to identify missing fields and inserting data to ensure result were displayed consistently. Highlighting area's where data quality improvements were required also took resource, however categorising the results into sub sections to process and display data made the reporting process more efficient as these sections were set out at the start.

Some of the advantages highlighted throughout was the use of data mining tools. Tableau and PowerBI processed the validated dataset to display the results with minimal manual input. This saved time and provided meaningful information that could be reported.

A variety of outputs were used through the use of different graphs and tools. The classification of data made it easier to transfer and display meaningful results.

The processes followed answered the area's to be investigated with supporting data and concludes with the below information-



However, some notable disadvantages have been highlighted. In order to retrieve meaningful information clear stakeholder requirements, need to be identified at the start of the data processing as assumptions were made on what information the company would like to see.

The data required manual processing to validate missing information such as CustomerID and descriptions have been used to categorise results however further data validation and processing of fields may have benefited the way the results were displayed as categorising areas further could have identified more trends in cancelled items for example.

To conclude, if this activity were to be completed again, the companies' data processing requirements would be set out at the start and more investigation would be completed on the relationships between attributes and missing data. Additional reporting tools such as Weka would also be investigated alongside Tableau and PowerBI.

8 References

- [1] U. Learning, "UCI Machine Learning," UCI Learning , 2012. [Online]. Available: <https://archive.ics.uci.edu/ml/machine-learning-databases/00352/>. [Accessed 22 04 2022].
- [2] I. C. Education, "Data Mining," IBM, 15 01 2021. [Online]. Available: <https://www.ibm.com/cloud/learn/data-mining>. [Accessed 12 05 2022].
- [3] SAS, "Data Mining What is it," SAS, [Online]. Available: https://www.sas.com/en_gb/insights/analytics/data-mining.html. [Accessed 22 04 2022].
- [4] S. Valcheva, "7 Real-World Examples Of Data Mining In Business, Marketing, Retail," Intellspot, [Online]. Available: <https://www.intellspot.com/data-mining-examples/>. [Accessed 20 04 2022].
- [5] Matillion, "5 real life applications of Data Mining and Business Intelligence," Matillion, 02 06 2020. [Online]. Available: <https://www.matillion.com/resources/blog/5-real-life-applications-of-data-mining-and-business-intelligence>. [Accessed 30 04 2022].
- [6] Tableau, "An Introduction To Data Mining Techniques," Tableau, [Online]. Available: <https://www.tableau.com/learn/articles/introduction-data-mining-techniques>. [Accessed 15 04 2022].
- [7] C. Stedman, "Data Mining," TechTarget, [Online]. Available: <https://www.techtarget.com/searchbusinessanalytics/definition/data-mining>. [Accessed 01 05 2022].
- [8] Jigsawacademy, "Data Mining Tools," [Online]. Available: <https://www.jigsawacademy.com/blogs/data-science/data-mining-tools/>. [Accessed 03 05 2022].
- [9] International Trade Administration, "Export Pricing Strategy," International Trade Administration, [Online]. Available: <https://www.trade.gov/pricing-strategy>. [Accessed 02 05 2022].

1809781- Data Mining Assignment

GRADEMARK REPORT

FINAL GRADE

71 / 100

GENERAL COMMENTS

Instructor

Jessica Marfell

1 engagement with literature
comment available

9 references used and cited throughout, a
little over reliance on web sources- try to
incorporate books and journals into your
references (especially at level 6) 20

2 Knowledge and Understanding
comment available

good discussion that denotes
understanding of the tools and techniques
demonstrated good supporting images
with citations throughout 10

3. Cognitive and Intellectual skills
comment available

excellent discussion of results and
multiple visualisations of the results
gained with justifications for why the
investigation was made. 25

4. Practical Application Skills
comment available

good discussion and justification of the
use of tableaux and the stages to complete
the experiment - excellent range of
investigations completed and
demonstration of pre processing 35

5. Transferable Skills for life and Professional practice
comment available

Very well structured report summary
should have bullets rather than icons for a
professional business report 15

Total: 71

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

PAGE 18

PAGE 19

PAGE 20

PAGE 21

PAGE 22

PAGE 23

PAGE 24

PAGE 25

PAGE 26

PAGE 27

PAGE 28

