PURPOSE

This report provides a detailed explanation of data cleaning and analysis of transactional data of a UK-based and registered non-store online retail occurring between 01/12/2010 and 09/12/2011. This report is intended for:

- Exploring the Datasets.
- Understanding the data.
- Checking Data Issues in Terms of Data Entry and Otherwise
- Cleaning the Data
- Identify Sales Patterns
- Make Projections of Sales

Data Description

This Online Retail II data set contains all the transactions occurring for a UK-based and registered, non-store online retail between 01/12/2009 and 09/12/2011.The company mainly sells unique all-occasion giftware. Many customers of the company are wholesalers.

InvoiceNo	Invoice number. A 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation.	Nominal
StockCode	Product (item) code. A 5-digit integral number uniquely assigned to each distinct product.	Nominal
Description	Product (item) name	Nominal
Quantity	The quantities of each product (item) per transaction.	Numeric
InvoiceDate	Invoice date and time. The day and time when a transaction was generated	Numeric
UnitPrice	Unit price. Product price per unit in sterling (Â£)	Numeric
CustomerID	Customer number. A 5-digit integral number uniquely assigned to each customer	Nominal
Country	Country name. The name of the country where a customer resides	Categorical Nominal

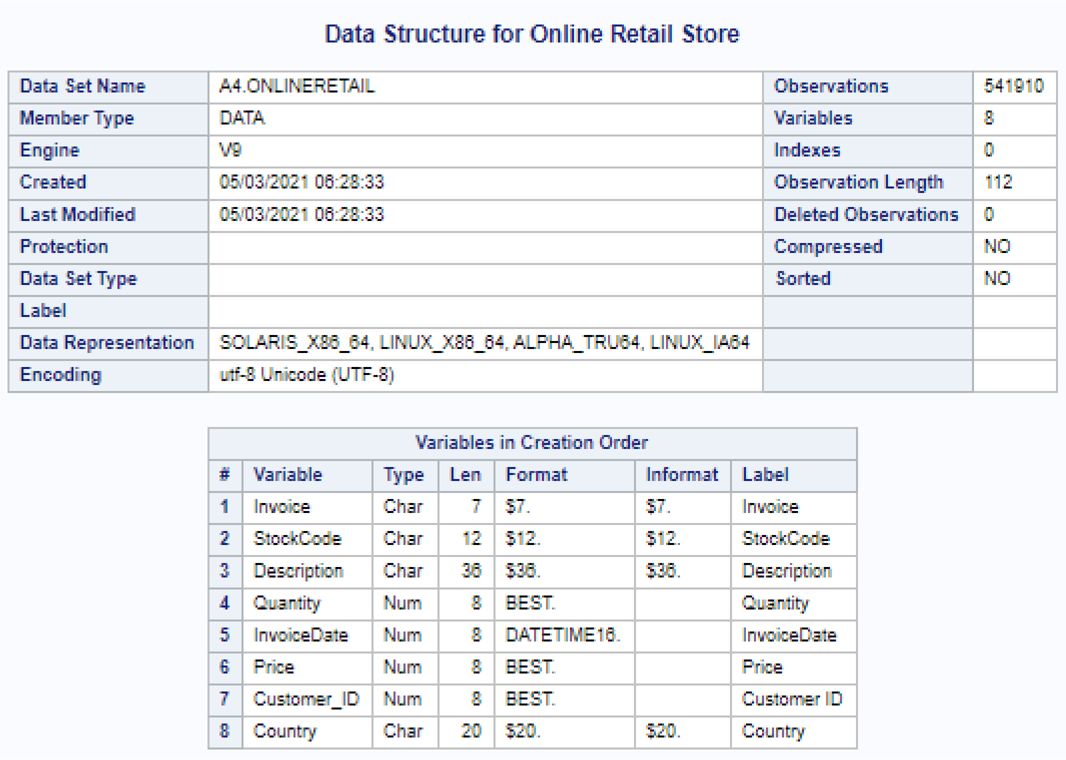
Overview

The analysis starts with exploring the dataset. The number and the type of variables along with the number of observations are explored to get an overview of the dataset. Next, the dataset is cleaned and preprocessed by exploring the numerical and categorical descriptive statistics and handling missing as well as invalid values.

The analysis of dataset is performed after the pre-processing of dataset. The data is explored to gain insights regarding sales of the store. Various research questions are answered such as unusual patterns, time lag, monetary sales monthly and per product. Finally, Sales projections are made to project sales numbers by changing the quantity of certain prominent products

Analyzing Data Set

Table Attributes and Data



Data Structure for Online Retail Store

Data Set Name	A4.ONLINERETAIL	Observations	541910
Member Type	DATA	Variables	8
Engine	V9	Indexes	0
Created	05/03/2021 06:28:33	Observation Length	112
Last Modified	05/03/2021 06:28:33	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

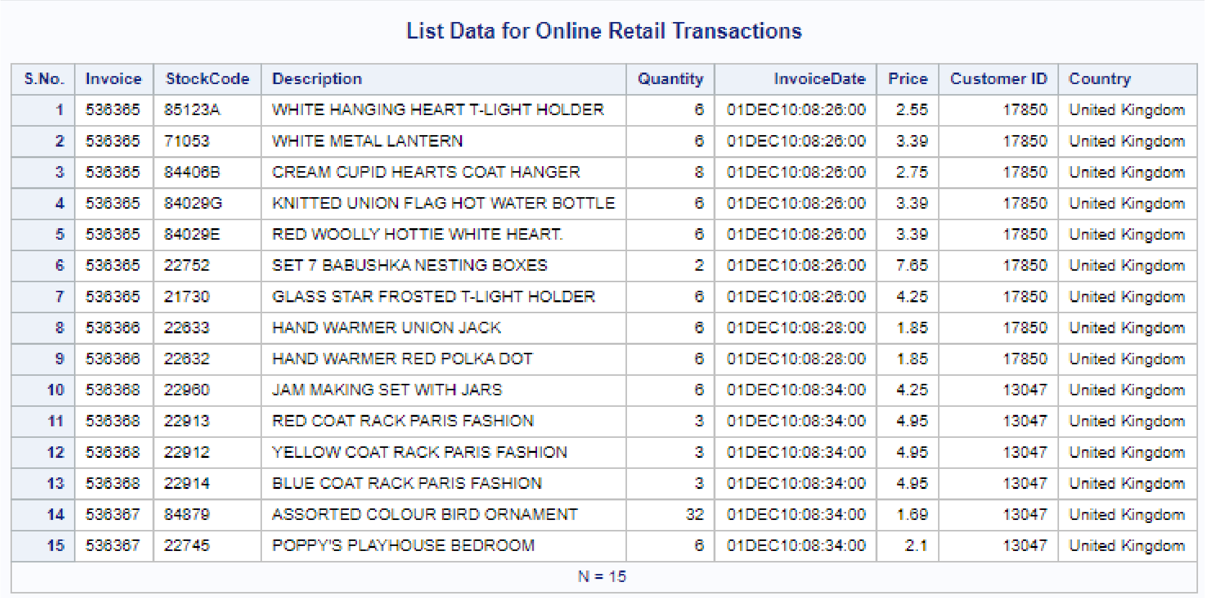
Variables in Creation Order

#	Variable	Type	Len	Format	Informat	Label
1	Invoice	Char	7	$7.	$7.	Invoice
2	StockCode	Char	12	$12.	$12.	StockCode
3	Description	Char	36	$36.	$36.	Description
4	Quantity	Num	8	BEST.		Quantity
5	InvoiceDate	Num	8	DATETIME16.		InvoiceDate
6	Price	Num	8	BEST.		Price
7	Customer_ID	Num	8	BEST.		Customer ID
8	Country	Char	20	$20.	$20.	Country

From the above tables about data structure, it can be observed that:

- Number of Variables: 8
- Number of Observations: 541,910
- Categorical Variables: 5 {Invoice, StockCode, Description, Customer_ID, Country}

- Numerical Variables: 3 {Quantity, Price, InvoiceDate}
- Continuous Numerical Variables: Price
- Discreate Numerical Variable: Quntity



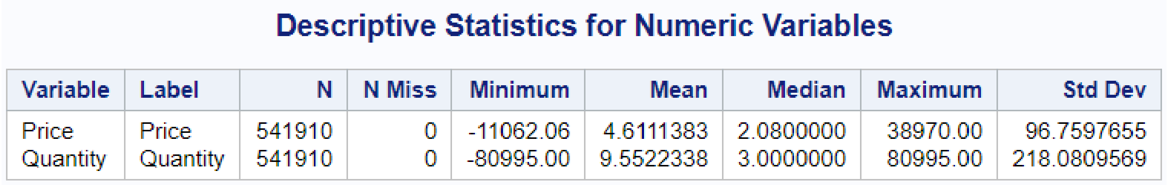
List Data for Online Retail Transactions

S.No.	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
1	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01DEC10:08:26:00	2.55	17850	United Kingdom
2	536365	71053	WHITE METAL LANTERN	6	01DEC10:08:26:00	3.39	17850	United Kingdom
3	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01DEC10:08:26:00	2.75	17850	United Kingdom
4	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01DEC10:08:26:00	3.39	17850	United Kingdom
5	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01DEC10:08:26:00	3.39	17850	United Kingdom
6	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	01DEC10:08:26:00	7.65	17850	United Kingdom
7	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	01DEC10:08:26:00	4.25	17850	United Kingdom
8	536366	22633	HAND WARMER UNION JACK	6	01DEC10:08:28:00	1.85	17850	United Kingdom
9	536366	22632	HAND WARMER RED POLKA DOT	6	01DEC10:08:28:00	1.85	17850	United Kingdom
10	536368	22960	JAM MAKING SET WITH JARS	6	01DEC10:08:34:00	4.25	13047	United Kingdom
11	536368	22913	RED COAT RACK PARIS FASHION	3	01DEC10:08:34:00	4.95	13047	United Kingdom
12	536368	22912	YELLOW COAT RACK PARIS FASHION	3	01DEC10:08:34:00	4.95	13047	United Kingdom
13	536368	22914	BLUE COAT RACK PARIS FASHION	3	01DEC10:08:34:00	4.95	13047	United Kingdom
14	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	01DEC10:08:34:00	1.69	13047	United Kingdom
15	536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	01DEC10:08:34:00	2.1	13047	United Kingdom

N = 15

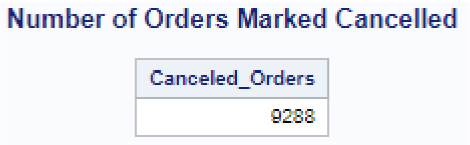
This provides an overview of data and its structure. To dive deeper into the data and issues related, further analysis is done for categorical and numerical variables.

Analyzing Numerical and Categorical Variables



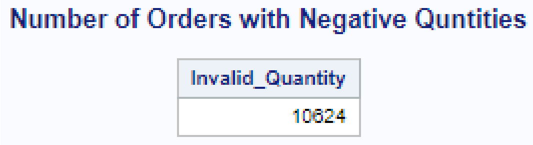
Descriptive Statistics for Numeric Variables

Variable	Label	N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
Price	Price	541910	0	-11062.06	4.6111383	2.0800000	38970.00	96.7597655
Quantity	Quantity	541910	0	-80995.00	9.5522338	3.0000000	80995.00	218.0809569



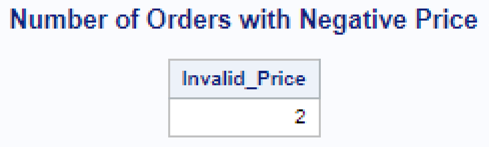
Number of Orders Marked Cancelled

Canceled_Orders
9288



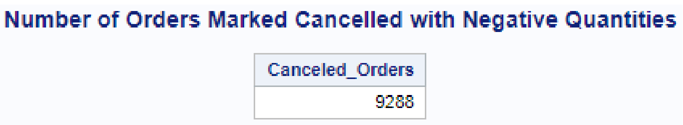
Number of Orders with Negative Quntities

Invalid_Quantity
10624



Number of Orders with Negative Price

Invalid_Price
2



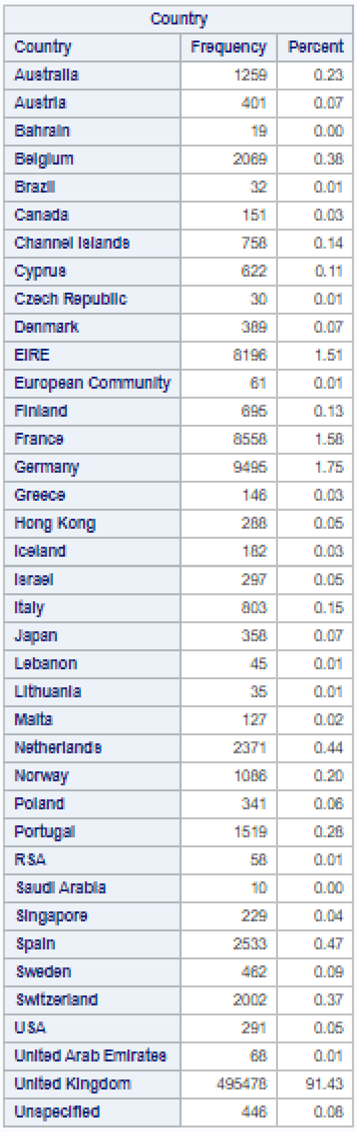
Number of Orders Marked Cancelled with Negative Quantities

Canceled_Orders
9288

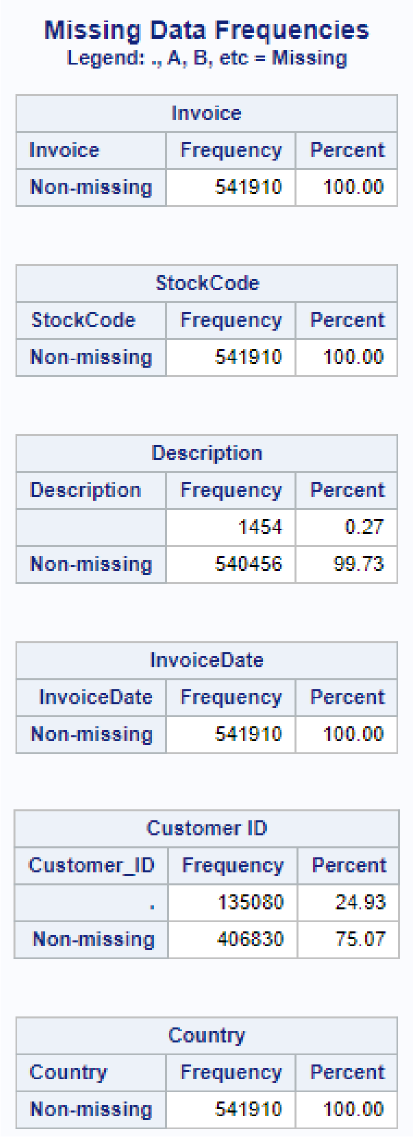
The descriptive statistics of the Numerical Variables has been calculated using the 'proc means' command in SAS. The Invalid values and their frequencies are evaluated using PROC SQL. The above tables provide the following information:

1. All the numerical variables wiz., Price and Quantity have invalid minimum value which is negative. This is an outlier as it falls out of the ranges of valid values for these variables.
2. As clear from tables specifying counts of negative values, there are 2 records with Invalid(negative) Prices and 10624 Invalid Quantities (Negative). Since Quantities cannot be negative, and constitutes just 2% of total observations, these can be dropped from the dataset.
3. There are 9288 orders marked as cancelled and signify that these orders do not constitute towards sales of the store. Thus, these can be dropped from the dataset. Also, all these orders have negative values of quantities.

Frequencies for Categorical Variable Country



Country		
Country	Frequency	Percent
Australia	1259	0.23
Austria	401	0.07
Bahrain	19	0.00
Belgium	2069	0.38
Brazil	32	0.01
Canada	151	0.03
Channel Islands	758	0.14
Cyprus	622	0.11
Czech Republic	30	0.01
Denmark	389	0.07
EIRE	8196	1.51
European Community	61	0.01
Finland	695	0.13
France	8558	1.58
Germany	9495	1.75
Greece	146	0.03
Hong Kong	288	0.05
Iceland	182	0.03
Israel	297	0.05
Italy	803	0.15
Japan	358	0.07
Lebanon	45	0.01
Lithuania	35	0.01
Malta	127	0.02
Netherlands	2371	0.44
Norway	1086	0.20
Poland	341	0.06
Portugal	1519	0.28
RSA	58	0.01
Saudi Arabia	10	0.00
Singapore	229	0.04
Spain	2533	0.47
Sweden	462	0.09
Switzerland	2002	0.37
USA	291	0.05
United Arab Emirates	68	0.01
United Kingdom	495478	91.43
Unspecified	446	0.08



Missing Data Frequencies

Legend: ., A, B, etc = Missing

Invoice		
Invoice	Frequency	Percent
Non-missing	541910	100.00

StockCode		
StockCode	Frequency	Percent
Non-missing	541910	100.00

Description		
Description	Frequency	Percent
	1454	0.27
Non-missing	540456	99.73

InvoiceDate		
InvoiceDate	Frequency	Percent
Non-missing	541910	100.00

Customer ID		
Customer_ID	Frequency	Percent
.	135080	24.93
Non-missing	406830	75.07

Country		
Country	Frequency	Percent
Non-missing	541910	100.00

The Frequency analysis of Categorical variables is performed in SAS using 'PROC FREQ' procedure. Since, except 'Country', all other nominal categorical variables have a large number of values(categories), thus, only the missing values (if present) are evaluated. The above tables reveal the following:

1. There are no missing values in country column. There are 91% of observations that belong to United Kingdom specifying that majority of orders were made in UK from the online store.
2. There are no missing values in Invoice, Stock Code, Invoice Date and Country.
3. There are 0.27% Null values in Description column and almost 25% null values in Customer_ID column. Since these columns just describe the order and explain nothing about the sales itself, these values can be replaced with 'N/A' and 'UNREGISTERED' respectively. Since, the transactions without customer id are for those customers who are not registered with the store.

Cleaning Data Set

The datasets contain certain inaccuracies and missing values that must be fixed before proceeding with any kind of analysis. These issues have been described above and below are the solutions applied to solve these issues.

Fixing Invalid Values (Data Inaccuracy)

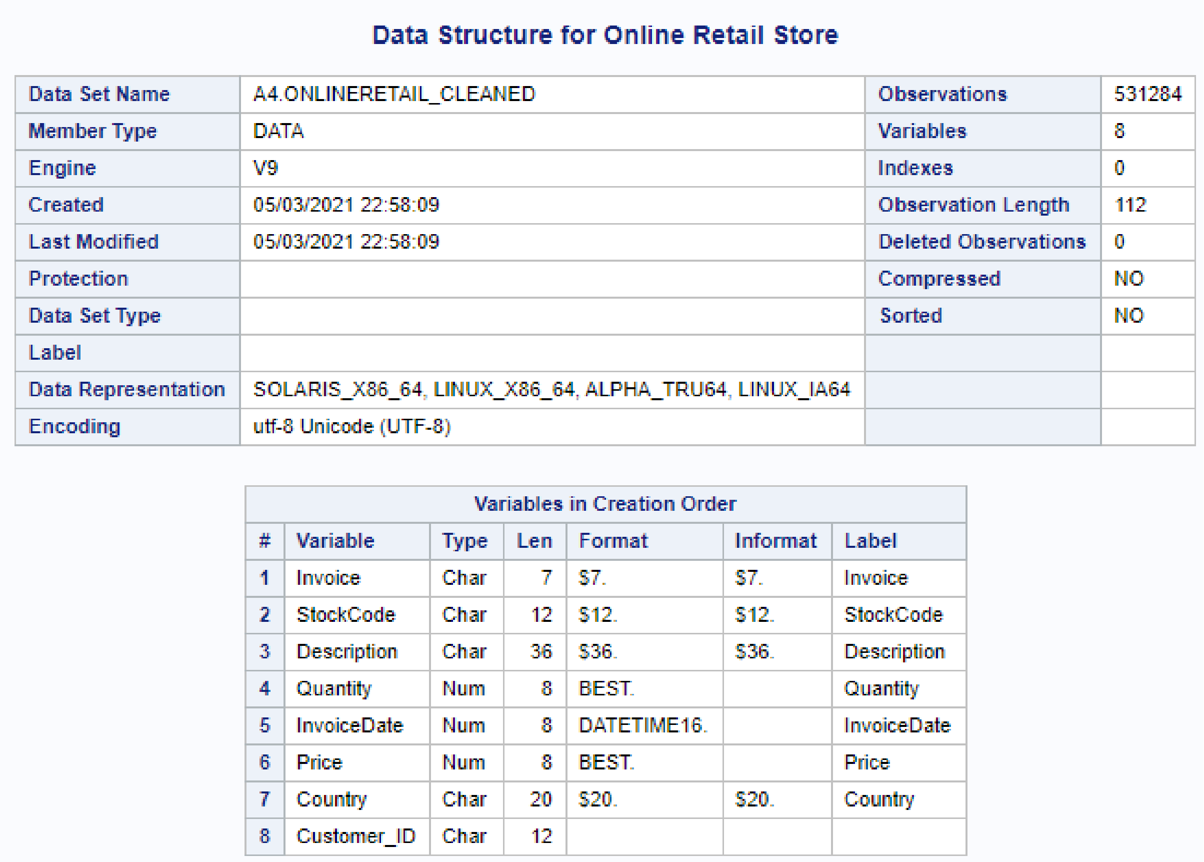
- All the orders marked as cancelled i.e., the Invoice Number starts with 'C' are removed from the dataset using PROC DATA step.
- All orders with Negative values of Quantity and Price are removed from the dataset. Since, there is no information to evaluate these variables, the orders are removed from the dataset using PROC DATA Step.

Fixing Missing Values

- All the orders with empty Description are filled with description 'N/A' since we do not have any information as to what that particular product represents.

- All the orders with null values of Customer_ID are filled with ‘UNREGISTERED’ since these customers are not registered with the store. To achieve this, the format of Customer_ID is changed from Integer to Character.

Datasets After Cleaning

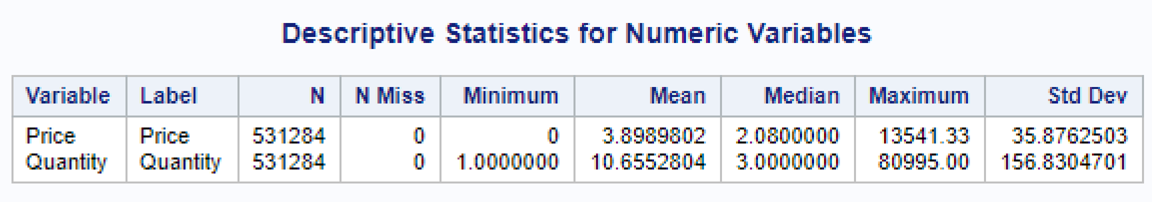


Data Structure for Online Retail Store

Data Set Name	A4.ONLINERETAIL_CLEANED	Observations	531284
Member Type	DATA	Variables	8
Engine	V9	Indexes	0
Created	05/03/2021 22:58:09	Observation Length	112
Last Modified	05/03/2021 22:58:09	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Variables in Creation Order

#	Variable	Type	Len	Format	Informat	Label
1	Invoice	Char	7	$7.	$7.	Invoice
2	StockCode	Char	12	$12.	$12.	StockCode
3	Description	Char	36	$36.	$36.	Description
4	Quantity	Num	8	BEST.		Quantity
5	InvoiceDate	Num	8	DATETIME16.		InvoiceDate
6	Price	Num	8	BEST.		Price
7	Country	Char	20	$20.	$20.	Country
8	Customer_ID	Char	12			

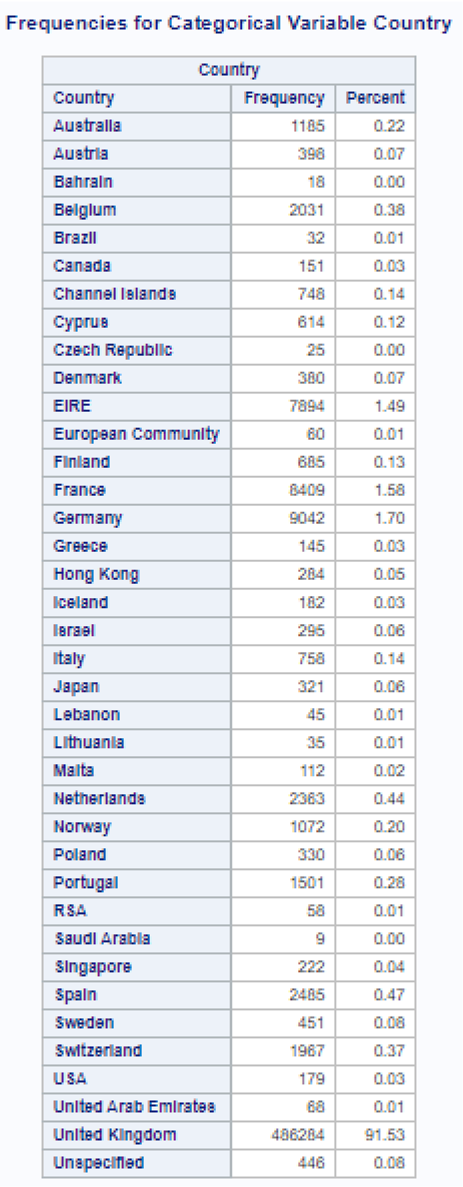


Descriptive Statistics for Numeric Variables

Variable	Label	N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
Price Quantity	Price Quantity	531284 531284	0 0	0 1.0000000	3.8989802 10.6552804	2.0800000 3.0000000	13541.33 80995.00	35.8762503 156.8304701

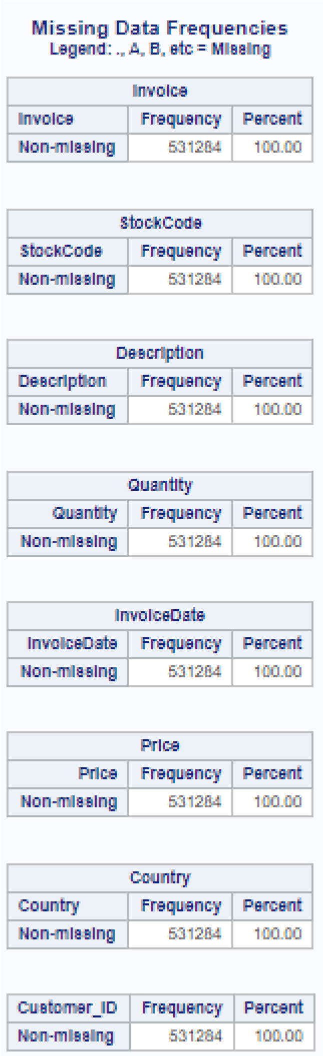
After the dataset has been cleaned, following observations can be made:

- Number of Orders in Cleaned Dataset: 531,284
- Minimum Price is 0.0 and Minimum Quantity is 1
- There are no missing values anymore in any of the variable.
- The Invalid values of both Description and Customer_ID columns are handled. There are 133,359 unregistered customers and 592 products without description.



Frequencies for Categorical Variable Country

Country		
Country	Frequency	Percent
Australia	1185	0.22
Austria	398	0.07
Bahrain	18	0.00
Belgium	2031	0.38
Brazil	32	0.01
Canada	151	0.03
Channel Islands	748	0.14
Cyprus	614	0.12
Czech Republic	25	0.00
Denmark	380	0.07
EIRE	7894	1.49
European Community	60	0.01
Finland	685	0.13
France	8409	1.58
Germany	9042	1.70
Greece	145	0.03
Hong Kong	284	0.05
Iceland	182	0.03
Israel	295	0.06
Italy	758	0.14
Japan	321	0.06
Lebanon	45	0.01
Lithuania	35	0.01
Malta	112	0.02
Netherlands	2363	0.44
Norway	1072	0.20
Poland	330	0.06
Portugal	1501	0.28
RSA	58	0.01
Saudi Arabia	9	0.00
Singapore	222	0.04
Spain	2485	0.47
Sweden	451	0.08
Switzerland	1967	0.37
USA	179	0.03
United Arab Emirates	68	0.01
United Kingdom	486284	91.53
Unspecified	446	0.08



Missing Data Frequencies

Legend: ., A, B, etc = Missing

Invoice		
Invoice	Frequency	Percent
Non-missing	531284	100.00

StockCode		
StockCode	Frequency	Percent
Non-missing	531284	100.00

Description		
Description	Frequency	Percent
Non-missing	531284	100.00

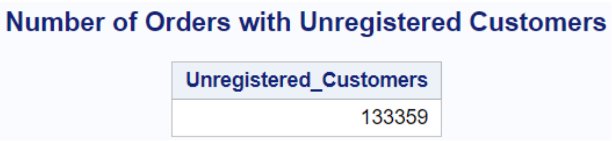
Quantity		
Quantity	Frequency	Percent
Non-missing	531284	100.00

InvoiceDate		
InvoiceDate	Frequency	Percent
Non-missing	531284	100.00

Price		
Price	Frequency	Percent
Non-missing	531284	100.00

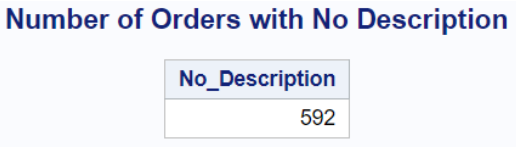
Country		
Country	Frequency	Percent
Non-missing	531284	100.00

Customer_ID	Frequency	Percent
Non-missing	531284	100.00



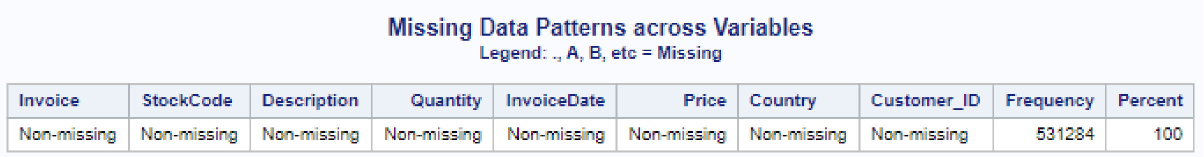
Number of Orders with Unregistered Customers

Unregistered_Customers
133359



Number of Orders with No Description

No_Description
592



Missing Data Patterns across Variables

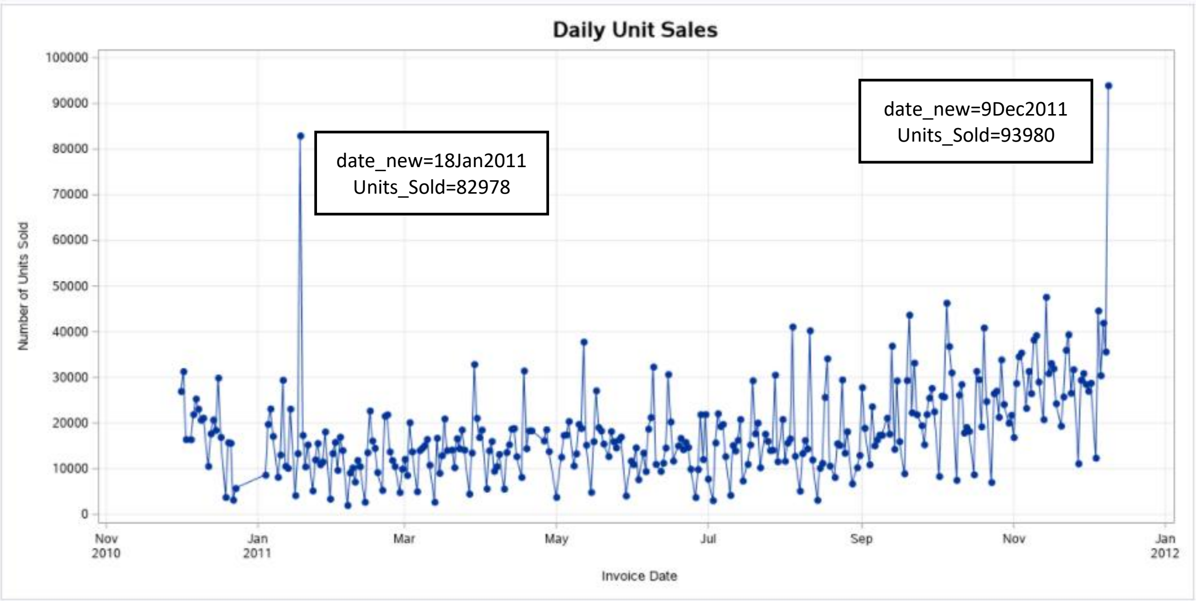
Legend: ., A, B, etc = Missing

Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Country	Customer_ID	Frequency	Percent
Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	Non-missing	531284	100

Exploring Data Set

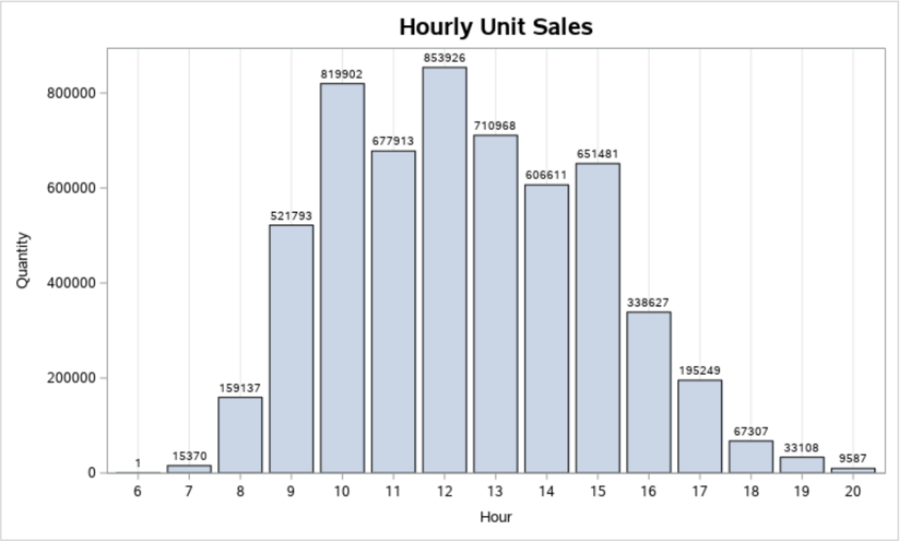
Below is a series of analysis performed on cleaned dataset in order to answer some of the research questions. These analyses are complemented by charts and graphs along with interpretations.

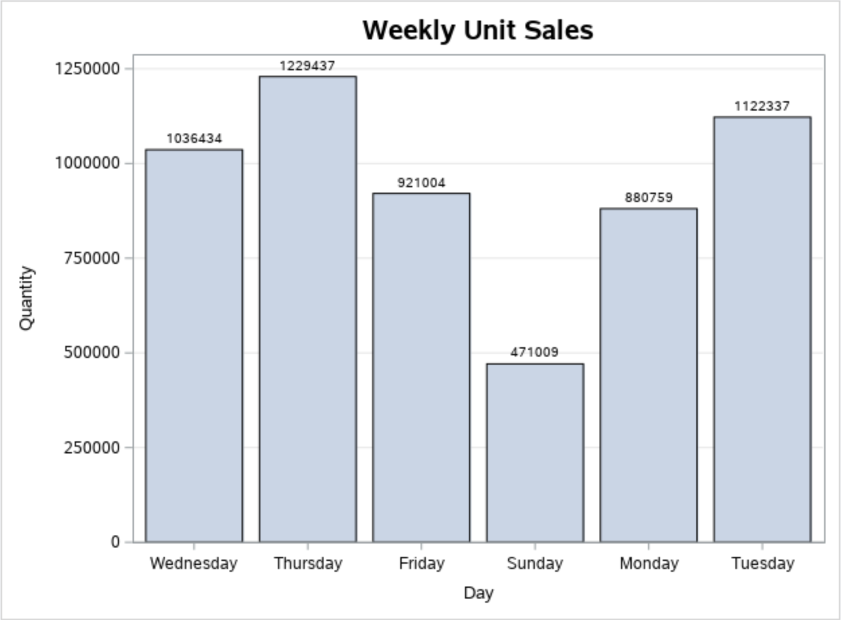
Unusual Unit Sales



- The above plot is developed by grouping together orders based on dates and aggregating the Quantity sold by adding all the quantities by day. A Series Plot is then developed with Invoice Date on x-axis and Number of Units sold on y-axis.
- As observed from the plot, the Units Sold were abnormally high on two specific dates January 18, 2011 and December 9, 2011.
- Although, the Unit sales vary 0 to 40,000 for most of the time, but for the above-mentioned dates, the number of units sold were 82978 and 93980 which are unusually high.

Time Lag in Sales

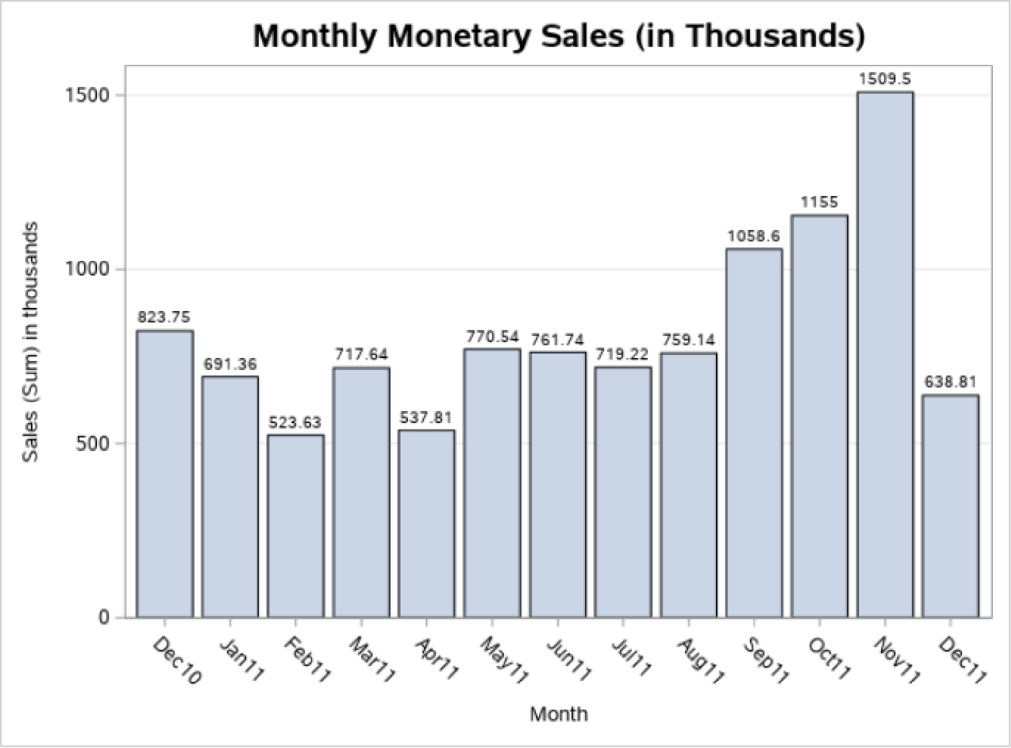




- The number of quantities sold on hourly and weekly are plotted above.
- From the plot of Hourly Unit Sales, it can be observed that there is no order from 2100 hrs. to 0600 hrs. i.e., there is a time lag of 9 hours in the Sales even though the orders are placed online. This suggest that no one places orders at nighttime.
- Apart from hourly distribution, the Weekly Unit Sales plot shows that no order is placed on Saturdays. This also signifies a time lag wherein after Friday the store receives orders from Sunay onwards.

Visualizing and Aggregating Monetary Sales

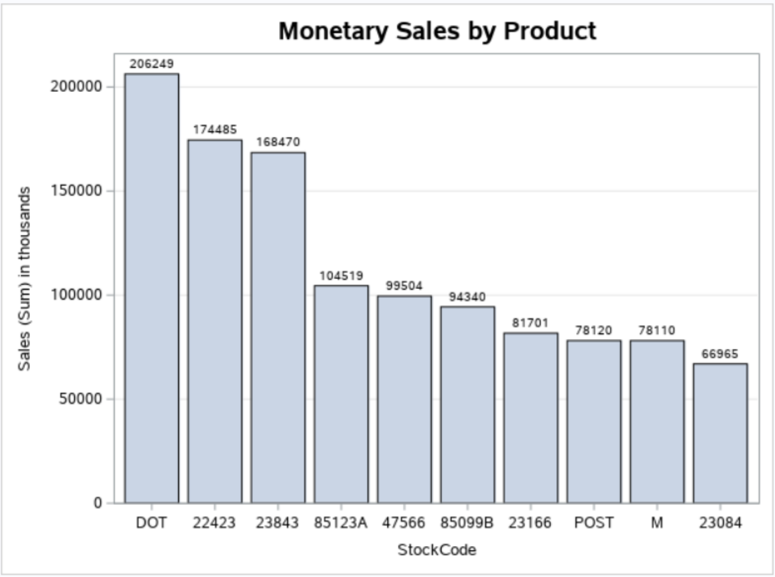
Monthly Sales

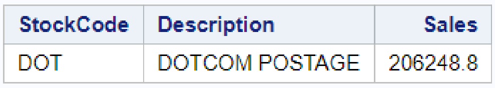


The above bar graph shows the monthly distribution of monetary sales (in thousands). The followings observations can be made:

- November 2011 was the highest selling month making a total sales value of 1509.5K while February 2011 was the lowest selling month with a sale of 523.63K only
- There was a steady sale from May to August 2011 with a value of around 760K.
- The year 2011 saw a dip in the sales for first few months as compared to ending of year 2010 with a sales of 823.75K in December 2010 while a sale of 523.63 in February 2011.
- There was a steep rise in the sales from September to November 2011 followed by a sharp decrease in December 2011 with a sale of just 638.81K

Highest Earning Product





StockCode	Description	Sales
DOT	DOTCOM POSTAGE	206248.8

The above plot displays the distribution of Monetary sales by Products. For convenience, the plot has been constructed by taking only 10 products in descending order of sales. The following observations can be made from the plot and table:

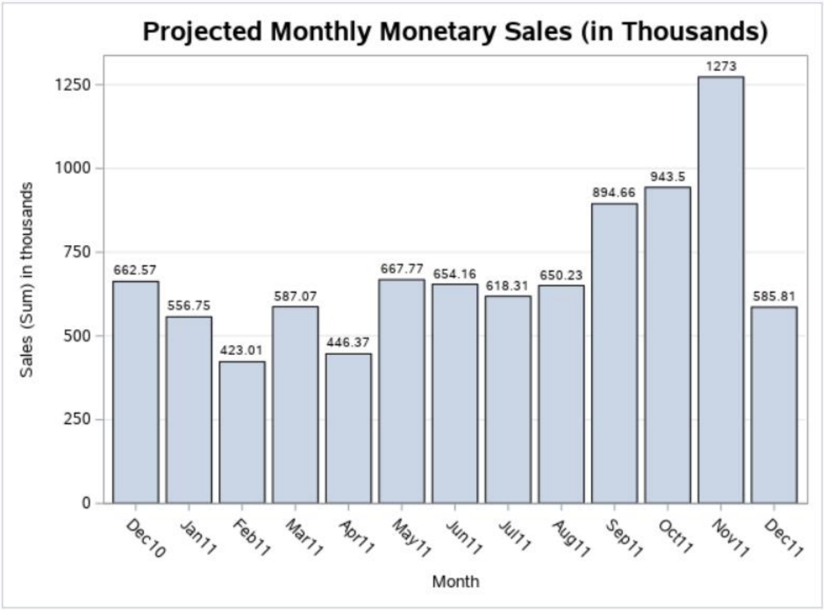
- The Product DOTCOM POSTAGE is the highest selling product with a net sale of 206,248.8 pound sterling. It is identified by StockCode ‘DOT’.
- DOT is followed by STockCodes 22423 and 22843 with almost similar sales of 1744,85 and 168,470-pound sterling respectively.

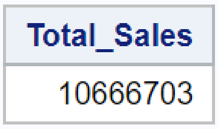
Sales Projections

The Sales estimations are performed for year 2012, by making the following modifications to the existing dataset

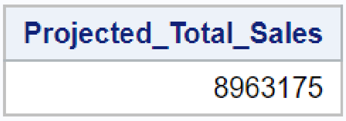
- The products that contributed to the bottom 20% of the monetary sales i.e., 21,33,340.6-pound sterling are removed from the dataset.
- The number of units of top 10 products based on amount of monetary sale are increased by 10%.

After doing the modifications, the sales for 2012 is projected by analyzing the modified dataset.





Total_Sales
10666703



Projected_Total_Sales
8963175

From the above bar chart and total sales, the following observations can be made:

- The Total Sales after making the modifications got reduced instead of increasing. The initial total sales of 10,666,703x-pound sterling got reduced to just 84% of value at 8963175-pound sterling. Thus, these modifications should not be performed.
- The Month-wise distribution of projected sales also displays a decrease in per month sales value when compared to actual distribution explained earlier. The highest earning month sales got reduced to 1273K pound sterling from 1509.5K pound sterling. The same can be observed for rest of the months.

Conclusions

Based on the analysis and interpretations, the following are the conclusions:

- The dataset consists of 531,284 transactions with an overall sale of 10,666,703 Pound sterling.
- Although, the Unit sales varies 0 to 40,000 for most of the time, but there are two days with unusually high spike in number of units sold that is January 18 and December 9th of 2011 with unit sales of 82978 and 93980 respectively.
- The highest number of units are sold on Thursdays while lowest on Sundays. There were no orders made on Saturdays.
- The orders are placed for specific time of the day starting from 6 am to 8pm. There are no orders between 9pm to 5am, since its nighttime.
- The highest sales were recorded in November 2011 with 1509.5K pound sterling sales, while the lowest sales were recorded in February 2011 with 523.63K pound sterling. The sales initially dropped at starting of year 2011 but gradually increased thereafter with a steep rise in November and a dramatic fall in December 2011.

- DOTCOM POSTAGE was the highest earning product with a total sale of 206.249K pound sterling.
- The projected sales for year 2012 obtained by removing products contributing to bottom 20% sales and increasing stock of top 10 products is demotivating as it gets reduced to 84% of the original total sales.

Summary

In this report, the datasets describing transactional data of a UK-based and registered non-store online retail occurring between 01/12/2010 and 09/12/2011 are processed. The processing, analysis and interpretation of data involved the following steps.

- Understanding Variables
- Accuracy Check
- Missing Values Check
- Fixing Accuracy and Missing Value Issues
- Exploration of Datasets to answer research questions:
    - Unusual Unit Sales of products
    - Time Lag in Sales and Order Placement
- Visualizing Transactional dataset
- Evaluating Sales on various parameters
    - Highest Monthly Sales
    - Top earning Product
- Projecting Sales
    - Removing Products that brought less sales
    - Increasing number of Units of top 10 highest selling products
    - Comparing projected increase in Sales with sales of 2011