APPLICATION OF DATA MINING IN CUSTOMER RELATIONSHIP MANAGEMENT
MARKET BASKET ANALYSIS IN ONLINE RETAIL DATASET

ŞENİZ AKBULUT | AHMET ATA SUHINDOL

DOKUZ EYLÜL ÜNİVERSİTESİ
İZMİR-1982

DEUCENG
Dokuz Eylül University
Dept of Computer Engineering

ABSTRACT

CRM is an iterative process that turns customer data into customer loyalty.

In the analysis of this data, data mining techniques are essentially used.

Association rules are one of the most frequently used methods which are the special application areas of the data mining.

Association rules are the rules that include which items commonly occur together in the same transactions.

The Apriori algorithm is the most popular association rule algorithm which discovers all frequent itemsets in large database of transactions.

This algorithm uses iterative approach to count the frequent itemsets.

The aim of this study is to propose a base for the customer relationship management activities by using data mining tools and applications for a firm in retail sector.

1

BASIC CONCEPTS

Explanations of some basic concepts

Customer relationship management (CRM) is the combination of practices, strategies and technologies that companies use to manage and analyze customer interactions and data throughout the customer lifecycle.It tries to understand the customer profile of the company and to communicate according to these profiles. Moreover CRM targets to gain new customers other than existing customers.

What is CRM?

CRM is defined as customer relationship management and is a software system that helps business owners nurture their relationships with their clientele. A CRM also assists with organization, efficiency, time management, and impressing clients every step of the way.

What is CRM?

CRM software has been around since the mid-1990s, but has come into its own over the last decade. CRM platforms are powerful systems that connect all the data from your sales leads and customers all in one place. A CRM records and analyzes all calls, emails and meetings, helping improve customer service, drive sales, and increase revenue.

2

DATA SET TO BE USED

Basic features of the data set to be used

UCI ONLINE RETAIL II DATASET

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 gift-ware. Many customers of the company are wholesalers.

SUMMARY OF DATASET

Data Set Characteristics:	Multivariate, Sequential, Time-Series
Attribute Characteristics:	Integer, Real
Associated Tasks:	Classification, Clustering

Number of Instances:	541909	**Area:**	Business
Number of Attributes:	8	**Date Donated**	2015-11-06
Missing Values?	N/A	**Number of Web Hits:**	676029

ATTRIBUTE INFORMATION

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

3

BUSINESS PROBLEM

Aim of the work

AIM OF THE PROJECT

There are different algorithms to be used for Association Rules Mining. One of them is the Apriori algorithm. In this project, product association analysis will be handled with “Apriori Algorithm” and the the purchasing tendency of the customer will be revealed who is in the sales process, using the sales data of an e-commerce company.

1

IMPLEMENTATION

Solution produced to the business problem

TOP 5 DATA IN DATASET

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	1/12/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	1/12/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	1/12/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	1/12/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	1/12/2010 8:26	3.39	17850.0	United Kingdom

DATA DESCRIBING

	Quantity	UnitPrice	CustomerID
count	532619.000000	532619.000000	397924.000000
mean	10.240010	3.847635	15294.315171
std	159.573967	41.758101	1713.169877
min	-9600.000000	-11062.060000	12346.000000
25%	1.000000	1.250000	13969.000000
50%	3.000000	2.080000	15159.000000
75%	10.000000	4.130000	16795.000000
max	80995.000000	13541.330000	18287.000000

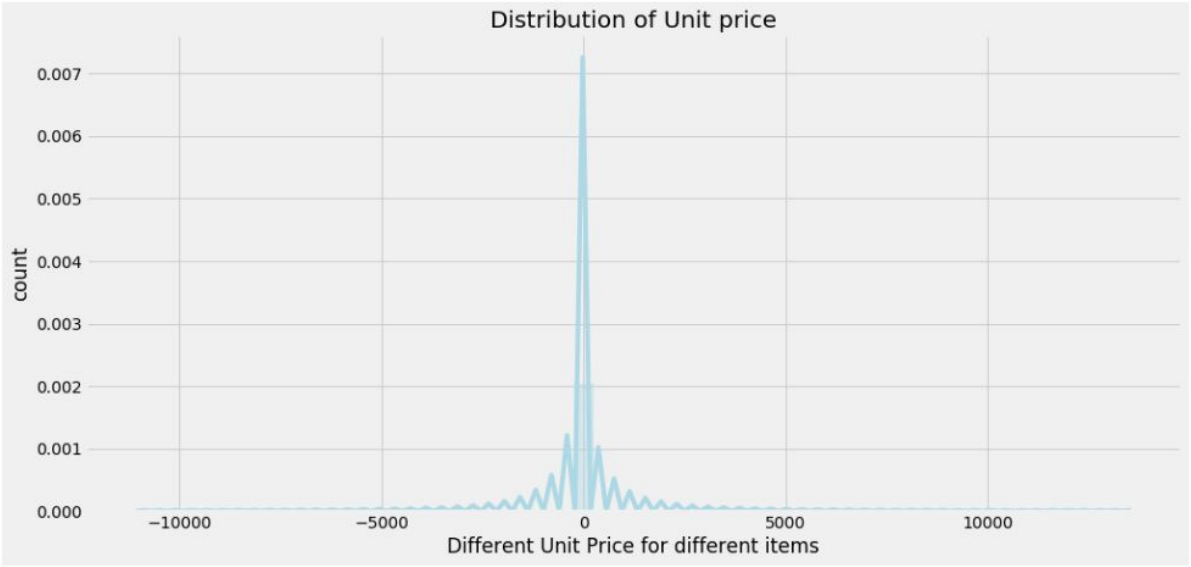
INFORMATION ABOUT DATA

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 532619 entries, 0 to 532618
Data columns (total 8 columns):
InvoiceNo      532619 non-null object
StockCode      532619 non-null object
Description    531165 non-null object
Quantity       532619 non-null int64
InvoiceDate    532619 non-null object
UnitPrice      532619 non-null float64
CustomerID     397924 non-null float64
Country        532619 non-null object
dtypes: float64(2), int64(1), object(5)
memory usage: 32.5+ MB
```

ATTRIBUTES AND DATA TYPES

```
InvoiceNo      object
StockCode      object
Description    object
Quantity        int64
InvoiceDate    object
UnitPrice     float64
CustomerID    float64
Country        object
dtype: object
```

VISUALIZING THE UNIT PRICE



DATA COUNTS BY COUNTRY

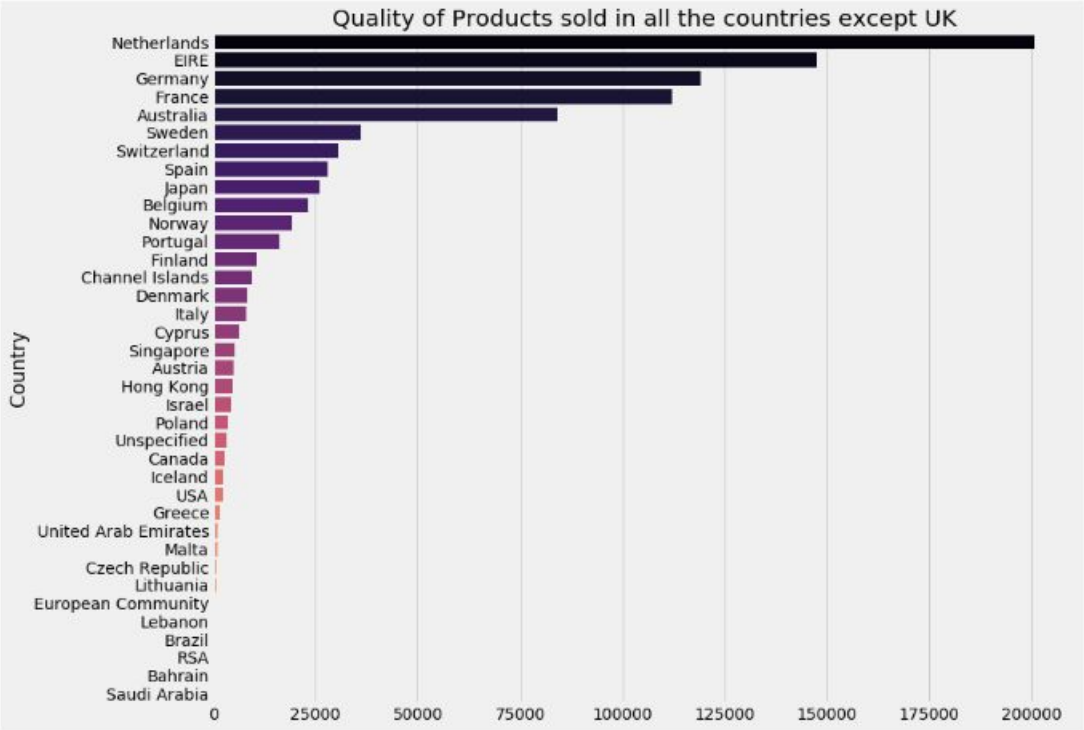
```
United Kingdom     487620
Germany              9042
France               8408
EIRE                 7894
Spain                2485
Netherlands          2363
Belgium              2031
Switzerland          1967
Portugal             1501
Australia            1185
Norway               1072
Italy                 758
Channel Islands       748
Finland               685
Cyprus                614
Sweden                451
Unspecified           446
Austria               398
Denmark               380
Poland                330
Name: Country, dtype: int64
```

VISUALIZATION OF DIFFERENT VALUE COUNTS FOR COUNTRY IN DATASET

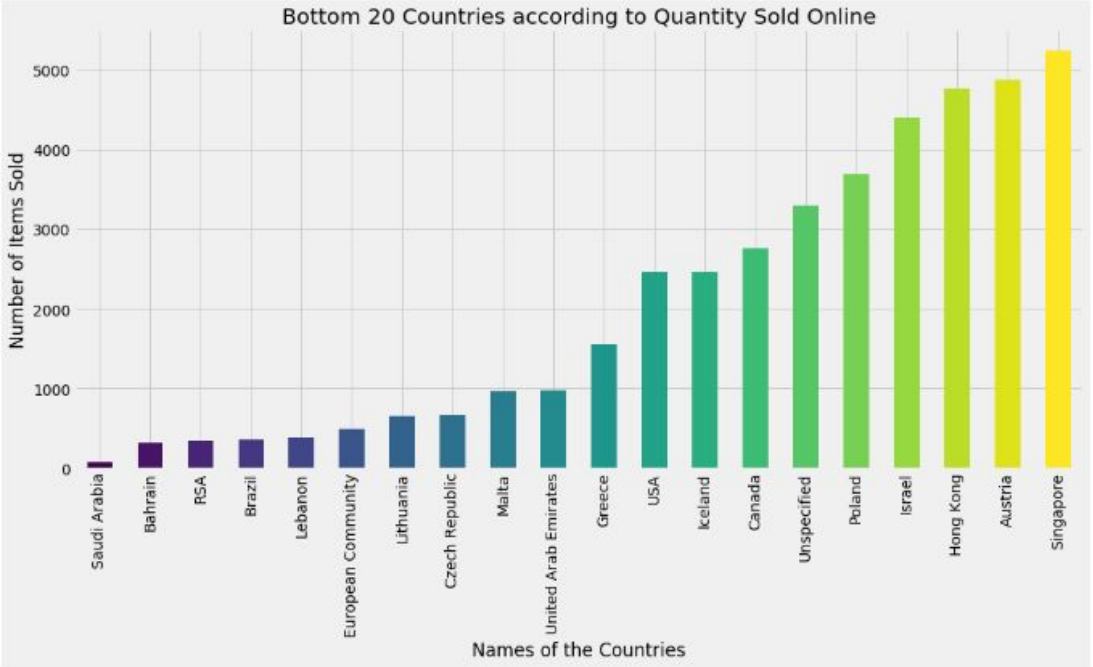
DATA COUNTS BY COUNTRY

```
Country
Netherlands            200937
EIRE                   147447
Germany                119263
France                 112104
Australia               84209
Sweden                  36083
Switzerland             30630
Spain                   27951
Japan                   26016
Belgium                 23237
Norway                  19338
Portugal                16258
Finland                 10704
Channel Islands          9491
Denmark                  8235
Italy                    8112
Cyprus                   6361
Singapore                5241
Austria                  4881
Hong Kong                4773
Israel                   4409
Poland                   3684
Unspecified              3300
Canada                   2763
Iceland                  2458
USA                      2458
Greece                   1557
United Arab Emirates      982
Malta                     970
Czech Republic            671
Lithuania                 652
European Community        499
Lebanon                   386
Brazil                    356
RSA                       352
Bahrain                   314
Saudi Arabia               80
Name: Quantity, dtype: int64
```

HOW MANY QUANTITY OF PRODUCTS HAVE BEEN SOLD ONLINE FOR EACH COUNTRY



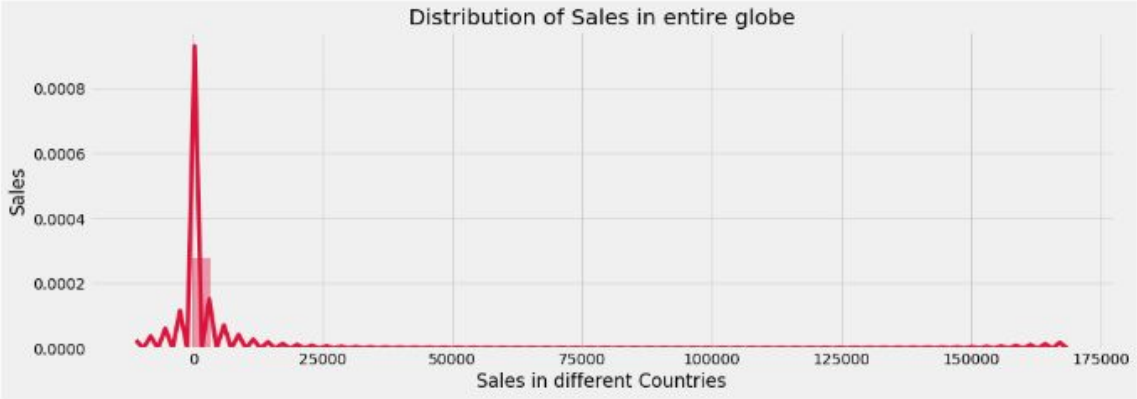
LAST 20 COUNTRY



WORLD CLOUD


Most Occuring word in the Description list
LIGHT
CUTLERY
SET
HANGING
Description
dtype
RETROSPOT
HOT
METAL
HEART
HOTTIE
HANGER
CUPID
BOTTLE
COAT
CIRCUS
CHILDREN
FLAG
WATER
Name
KNITTED
UNION
RED
WHITE
APRON
PACK
SPACEBOY
LANTERN
WOOLLY
HOLDER
CREAM
DOLLY
GIRL
BAKING
object
NAPKINS
PIECE
PARADE
Length

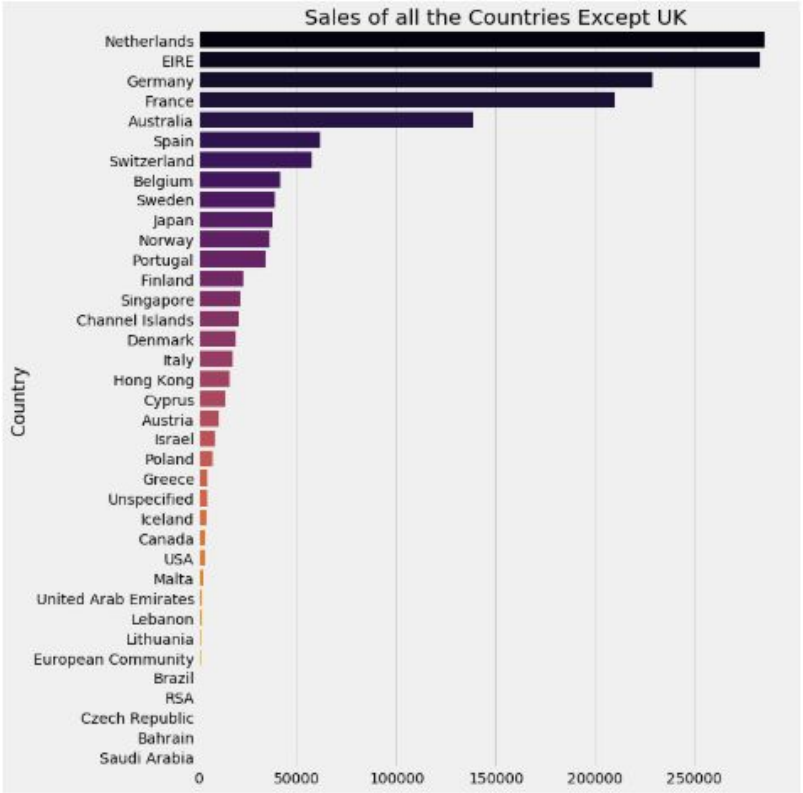
VISUALIZING THE SALES IN THE ENTIRE GLOBE



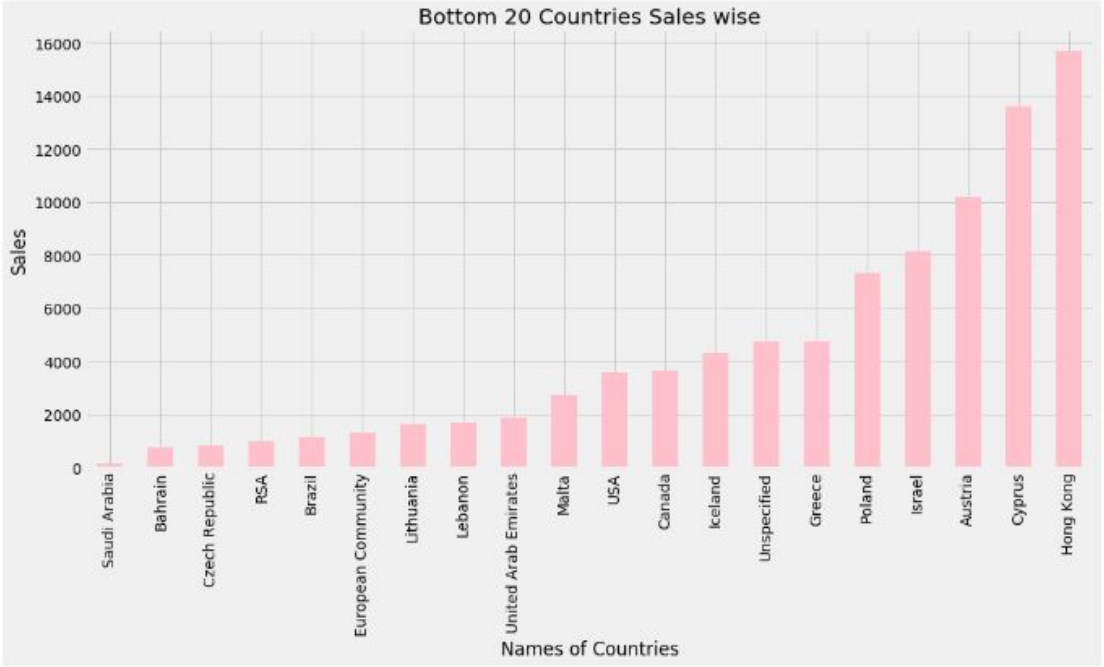
COUNTRY SALES

```
Country
Netherlands          285446.34
EIRE                 283453.96
Germany              228867.14
France               209715.11
Australia            138521.31
Spain                 61577.11
Switzerland           57089.90
Belgium               41196.34
Sweden                38378.33
Japan                 37416.37
Norway                36165.44
Portugal              33747.10
Finland               22546.08
Singapore             21279.29
Channel Islands       20450.44
Denmark               18955.34
Italy                 17483.24
Hong Kong             15691.80
Cyprus                13590.38
Austria               10198.68
Israel                 8135.26
Poland                 7334.65
Greece                 4760.52
Unspecified            4749.79
Iceland                4310.00
Canada                 3666.38
USA                    3580.39
Malta                  2725.59
United Arab Emirates   1902.28
Lebanon                1693.88
Lithuania              1661.06
European Community     1300.25
Brazil                 1143.60
RSA                    1002.31
Czech Republic          826.74
Bahrain                 754.14
Saudi Arabia            145.92
Name: Sales, dtype: float64
```

SALES OF COUNTRIES



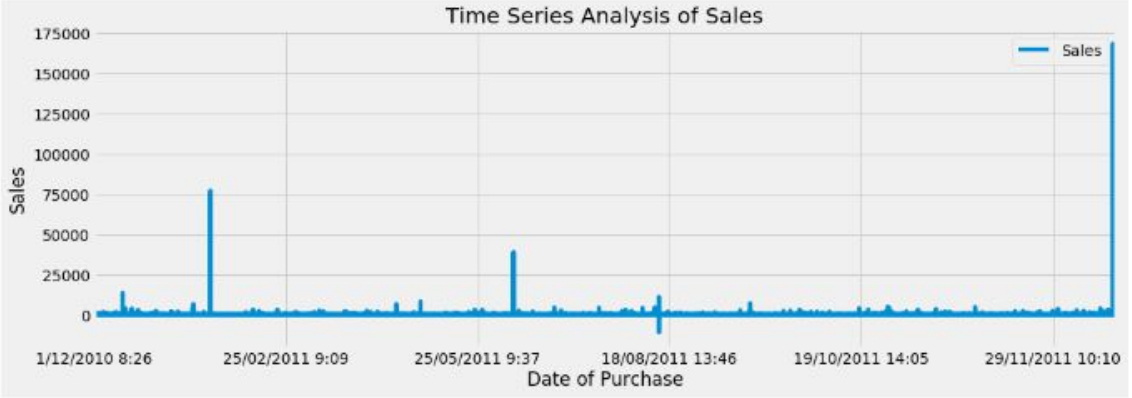
LAST 20 SALES WISE COUNTRIES

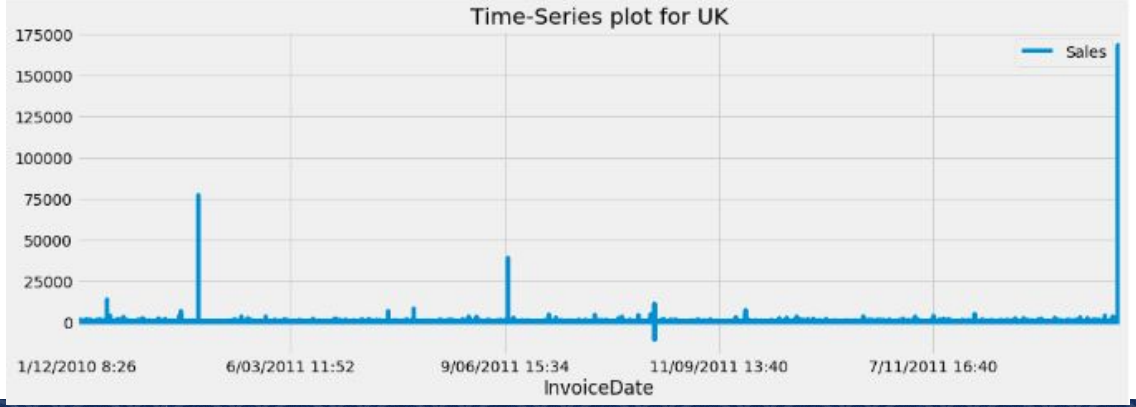


What is Time Series Analysis?

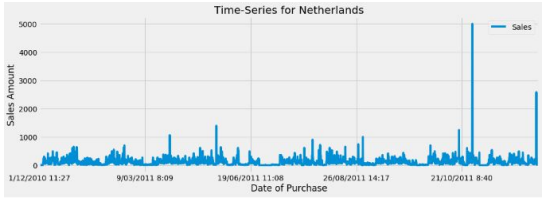
A time series analysis is a process of analyzing an observation of data points collected over a period of time, i.e time series data. In time series analysis, data analysts record data observations in constant intervals for a set of time periods instead of recording data observations randomly. The rate of observation (time interval) can be from milliseconds to several years.

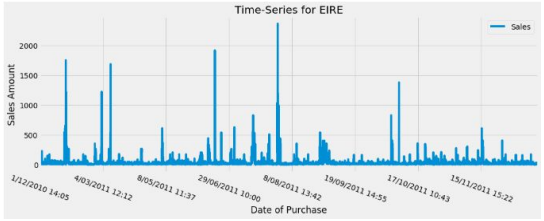
SALES AND INVOICE DATE TIME SERIES ANALYSIS

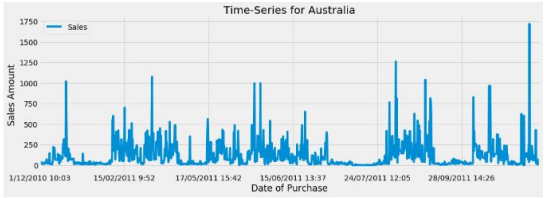


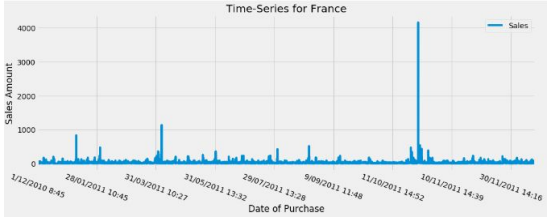


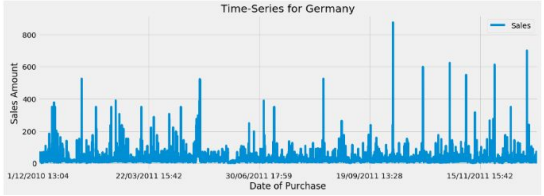
TIME SERIES FOR COUNTRIES

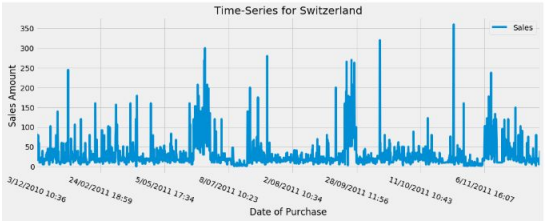




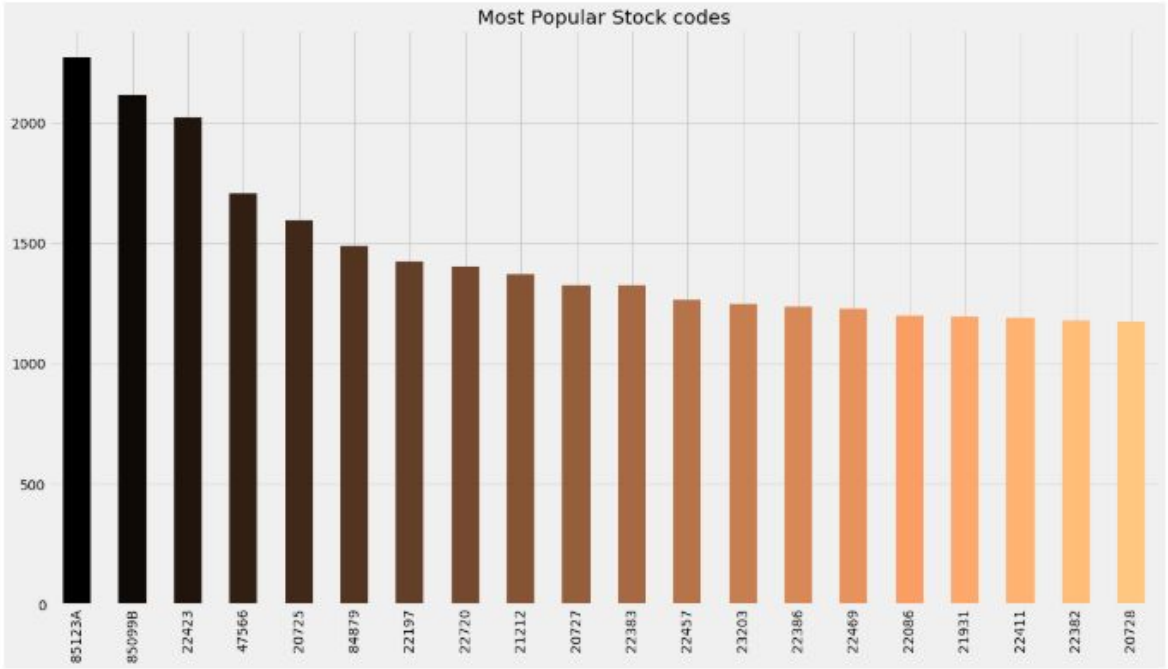








GROUP BY STOCK CODES



SORTED DATASET BY SALES AMOUNT

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Sales
294522	A563186	B	Adjust bad debt	1	12/08/2011 14:51	-11062.06	NaN	United Kingdom	-11062.06
294523	A563187	B	Adjust bad debt	1	12/08/2011 14:52	-11062.06	NaN	United Kingdom	-11062.06
40798	539856	21116	OWL DOORSTOP	1	22/12/2010 14:41	0.00	NaN	United Kingdom	0.00
325066	565938	23066	NaN	-13	8/09/2011 10:54	0.00	NaN	United Kingdom	-0.00
140178	548646	21472	NaN	-140	1/04/2011 13:15	0.00	NaN	United Kingdom	-0.00
...	...	...	...	...	...	...	...	...	...
294521	A563185	B	Adjust bad debt	1	12/08/2011 14:50	11062.06	NaN	United Kingdom	11062.06
14839	537632	AMAZONFEE	AMAZON FEE	1	7/12/2010 15:08	13541.33	NaN	United Kingdom	13541.33
218664	556444	22502	PICNIC BASKET WICKER 60 PIECES	60	10/06/2011 15:28	649.50	15098.0	United Kingdom	38970.00
60580	541431	23166	MEDIUM CERAMIC TOP STORAGE JAR	74215	18/01/2011 10:01	1.04	12346.0	United Kingdom	77183.60
531138	581483	23843	PAPER CRAFT , LITTLE BIRDIE	80995	9/12/2011 9:15	2.08	16446.0	United Kingdom	168469.60

SORTED DATASET BY UNIT PRICE

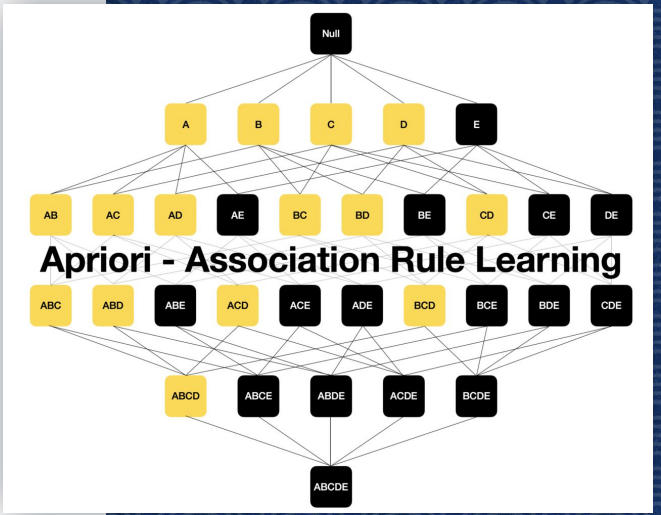
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Sales
14839	537632	AMAZONFEE	AMAZON FEE	1	7/12/2010 15:08	13541.33	NaN	United Kingdom	13541.33
294521	A563185	B	Adjust bad debt	1	12/08/2011 14:50	11062.06	NaN	United Kingdom	11062.06
170180	551697	POST	POSTAGE	1	3/05/2011 13:46	8142.75	16029.0	United Kingdom	8142.75
292286	562955	DOT	DOTCOM POSTAGE	1	11/08/2011 10:14	4505.17	NaN	United Kingdom	4505.17
263194	560373	M	Manual	1	18/07/2011 12:30	4287.63	NaN	United Kingdom	4287.63
414584	573080	M	Manual	1	27/10/2011 14:20	4161.06	12536.0	France	4161.06
414560	573077	M	Manual	1	27/10/2011 14:13	4161.06	12536.0	France	4161.06
398897	571751	M	Manual	1	19/10/2011 11:18	3949.32	12744.0	Singapore	3949.32
367877	569382	M	Manual	1	3/10/2011 16:44	3155.95	15502.0	United Kingdom	3155.95
341714	567353	M	Manual	1	19/09/2011 16:14	2653.95	NaN	Hong Kong	2653.95
117455	546558	M	Manual	1	15/03/2011 9:50	2583.76	NaN	Hong Kong	2583.76
292031	562946	M	Manual	1	11/08/2011 9:38	2500.00	15581.0	United Kingdom	2500.00
142170	548813	M	Manual	1	4/04/2011 13:03	2382.92	12744.0	Singapore	2382.92
484322	578149	DOT	DOTCOM POSTAGE	1	23/11/2011 11:11	2275.54	NaN	United Kingdom	2275.54
515827	580610	DOT	DOTCOM POSTAGE	1	5/12/2011 11:48	2196.67	NaN	United Kingdom	2196.67
398898	571751	M	Manual	1	19/10/2011 11:18	2118.74	12744.0	Singapore	2118.74
516069	580612	DOT	DOTCOM POSTAGE	1	5/12/2011 11:58	2114.00	NaN	United Kingdom	2114.00
142173	548820	M	Manual	1	4/04/2011 13:04	2053.07	12744.0	Singapore	2053.07
336841	566927	M	Manual	1	15/09/2011 15:20	2033.10	17846.0	United Kingdom	2033.10
493256	578833	DOT	DOTCOM POSTAGE	1	25/11/2011 15:23	2028.25	NaN	United Kingdom	2028.25

1

ASSOCIATION RULE MINING

Using Apriori Rule

Apriori is an algorithm for frequent item set mining and association rule learning over relational databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis.



FOUR PARAMETER OF APRIORI ALGORITHM

> min_support: support refers to the popularity of item and can be calculated by finding the number of transactions containing a particular item divided by the total number of transactions.

Support(diaper) = (Transactions containing (diaper))/(Total Transactions) Support(diaper) = 150 / 1000 = 15 %

> min_confidence: Confidence refers to the likelihood that an item B is also bought if item A is bought. It can be calculated by finding the number of transactions where A and B are bought together, divided by the total number of transactions where A is bought. Mathematically, it can be represented as:

Confidence(A → B) = (Transactions containing both (A and B))/(Transactions containing A)

The confidence of likelihood of purchasing a diaper if a customer purchase milk. Confidence(milk → diaper) = (Transactions containing both (milk and diaper))/(Transactions containing milk) Confidence(milk → daiper) =30 / 120 = 25 %

Confidence is similar to Naive Based Algorithm.

> min_lift: Lift refers to the increase in the ratio of the sale of B when A is sold. Lift(A → B) can be calculated by dividing Confidence(A → B) divided by Support(B). Mathematically it can be represented as: Lift(A→B) = (Confidence (A→B))/(Support (B))

Lift(milk → diaper) = (Confidence (milk → diaper))/(Support (diaper)) Lift(milk → diaper) = 25 / 15 = 1.66

So by Lift theory, there is 1.66 times more chance of buying milk and diaper together then just buying diaper alone.

> min_length: How many Items do we want to associate in our rules.

RULES FOR APRIORY ALGORITHM

1. Set a minimum value for support and confidence. This means that we are only interested in finding rules for the items that have certain default existence (e.g. support) and have a minimum value for co-occurrence with other items (e.g. confidence).

2. Extract all the subsets having a higher value of support than a minimum threshold.

3. Select all the rules from the subsets with confidence value higher than the minimum threshold.

4. Order the rules by descending order of Lift.

SPLITTING DATA ACCORDING TO THE REGION TRANSACTIONS

```
# Transactions done in France
basket_France = (data[data['Country'] =="France"]
          .groupby(['InvoiceNo', 'Description'])['Quantity']
          .sum().unstack().reset_index().fillna(0)
          .set_index('InvoiceNo'))

# Transactions done in the United Kingdom
basket_UK = (data[data['Country'] =="United Kingdom"]
          .groupby(['InvoiceNo', 'Description'])['Quantity']
          .sum().unstack().reset_index().fillna(0)
          .set_index('InvoiceNo'))

# Transactions done in Portugal
basket_Por = (data[data['Country'] =="Portugal"]
          .groupby(['InvoiceNo', 'Description'])['Quantity']
          .sum().unstack().reset_index().fillna(0)
          .set_index('InvoiceNo'))

basket_Sweden = (data[data['Country'] =="Sweden"]
          .groupby(['InvoiceNo', 'Description'])['Quantity']
          .sum().unstack().reset_index().fillna(0)
          .set_index('InvoiceNo'))
```

BUILDING MODEL

```
def hot_encode(x):
    if(x<= 0):
        return 0
    if(x>= 1):
        return 1

# Encoding the datasets
basket_encoded = basket_France.applymap(hot_encode)
basket_France = basket_encoded

basket_encoded = basket_UK.applymap(hot_encode)
basket_UK = basket_encoded

basket_encoded = basket_Por.applymap(hot_encode)
basket_Por = basket_encoded

basket_encoded = basket_Sweden.applymap(hot_encode)
basket_Sweden = basket_encoded
```

BUILDING MODEL

```
# Building the model
frq_items = apriori(basket_France, min_support = 0.05, use_colnames = True)

# Collecting the inferred rules in a dataframe
rules = association_rules(frq_items, metric ="lift", min_threshold = 1)
rules = rules.sort_values(['confidence', 'lift'], ascending =[False, False])
rules.head()
```

RULES

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
45	(JUMBO BAG WOODLAND ANIMALS)	(POSTAGE)	0.076531	0.765306	0.076531	1.000	1.306667	0.017961	inf
258	(PLASTERS IN TIN CIRCUS PARADE, RED TOADSTOOL ...	(POSTAGE)	0.051020	0.765306	0.051020	1.000	1.306667	0.011974	inf
270	(RED TOADSTOOL LED NIGHT LIGHT, PLASTERS IN TI...	(POSTAGE)	0.053571	0.765306	0.053571	1.000	1.306667	0.012573	inf
302	(SET/20 RED RETROSPOT PAPER NAPKINS, SET/6 RED...	(SET/6 RED SPOTTY PAPER PLATES)	0.102041	0.127551	0.099490	0.975	7.644000	0.086474	34.897959
300	(SET/6 RED SPOTTY PAPER PLATES, SET/20 RED RET...	(SET/6 RED SPOTTY PAPER CUPS)	0.102041	0.137755	0.099490	0.975	7.077778	0.085433	34.489796

THANKS!

Any questions?