Data analysis of the Online Retails DataSet

Raul Alejandro Buitrago Castellanos
Ingeniero de Sistemas
Universidad Distrital Francisco José de Caldas
Bogotá D.C., Colombia
Email: raulhabits@gmail.com

***Abstract*—This document contains a data analysis of the "Online Retail" dataset provided by the UCI Machine Learning Repository web site, this analysis is made to get knowledge from the transactions described in the dataset through a time line.**

***Keywords*—Dataset, Data Mining, Big Data Analytics, Business intelligence.**

I. Introduction

The data analysis has an important role in the human history. It could be to understand the nature, increase the life quality, economics, and others.
With the technology evolution, the information storage and processing of it has been increased allowing the use and treatment of big data sets.
The application of the data analysis is infinite because every things that you can establish a measure may be analyzed, an example could be the analysis of the behavior of the money value, the visits to a web site, the business intelligence, the ADN analysis.
Between all the techniques to the data analysis are commonly used the statistics, calculating probabilities, bigdata methodologies, data mining, and others.
For it this document is made using some statistics techniques, bigdata tasks, and data mining to analyze the dataset.

II. Objectives

Make a data analysis to the dataset using statistic techniques to identify the influence between the variables establishing behaviors or patterns to understanding and modeling the Online Retails information, using bigdata and data mining techniques.

III. Theoretical Framework

To a complete understanding of this document the reader must take some minutes to check the following terms.

A. Statistics

Is the studying of random phenomena, to obtain some conclusions from a test dataset.

B. Dataset

Is part of the statistic process, because it's a collection of records used to describe the population.

C. Data mining

Is the use of statistic tools, and computing science approach to find patterns in a dataset.

D. Machine learning

Is a set of techniques used in the data mining to establish relationships and predict the behavior of a dataset using some IA algorithms.

E. Big Data Analytics

Is the combination between business intelligence and analytics techniques, those are commonly used in data mining and statistical analysis. Some of the techniques mentioned are the K-Means, Naive Bayes, K-Nearest neighbor, clustering, and regression.

IV. State of the art

There are several ways to use the data science applied in marketing, and the use of data mining techniques, in this case for this dataset some authors made a representative work based in clustering and decision trees, obtaining good results establishing some relationships between some fields obtaining patterns.
Some of the comments that other authors provide us in their researches are that the data mining process takes a long time, because those tasks are too complex

- Data preparation
- Model interpretation
- Evaluation

And another considerations are related with the correlation, because correlation in some cases doesn't implies causation.

V. Methodology

The work and data analysis of this research as suggest one of the big data methodologies has the next tasks.

- Information recognizing
  - Identifying the domain
  - Identifying a problem
  - SMART Objective
    - Specified
    - Measurable
    - Attainable
    - Relevant
    - Time able
- Research Questions
  - Descriptive
  - Exploratory

    - Inferential
    - Predictive
- Exploratory analysis
    - Frequencies
    - Central tendency measures
- Multivariate analysis
    - Correlation Analysis between variables
- Finding patterns
    - Linear regression
    - Clustering
- Conclusions

VI. Information Recognizing

In this section will be described the data domain, the dataset, and the variables recognizing, and the SMART objectives.

A. The domain of the data.

One of the common uses of the business intelligence is oriented to the marketing, because all the organizations needs to know the current state of them, the history of their behavior, the capacity to satisfy the client requests and to increase their competitive.
In this data analysis the dataset used is provided by UCI on the official web site and it contains some records related with the transactions of a store through a time period.

B. The Dataset

Like the UCI web page says, the dataset contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers, it's made off 541909 records of online retails and is composed by a the next fields.

- 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. Invoice 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.

C. Smart Objectives

Check the behavior of the online retails from the past, to establish patterns, tendencies, and behaviors of the market. This objective must be the of high quality because the stakeholder may use it to take decisions.

VII. Research Questions

A correct investigation begins with the right questions, then an optimal definition of what else will be the proposal of the research. The following questions are used in this research, these questions must be classified according with the kind of it.

A. Descriptive.

A descriptive question is used to identify and know the characteristics of the dataset.

- What's the min date of the measurement?

```
## [1] "2010-12-01"
```

- What's the max date of the measurement?

```
## [1] "2011-12-09"
```

- What's the product quantity sold in the measurement?

```
## [1] 5660981
```

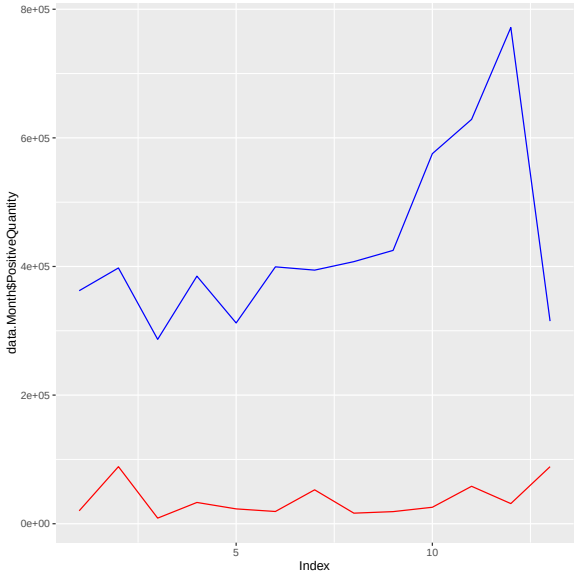
- What's the product quantity returned in the measurement?

```
## [1] 484531
```

B. Exploratory.

An exploratory question consists in the searching of patterns or relations to support an investigation question.

- Which month had major product sales and devolutions?



	Index	YearMonth	PositiveQuantity	NegativeQuantity
1	1	2010 - 12	362316.00	20088.00
42482	2	2011 - 1	397716.00	88750.00
77629	3	2011 - 2	286695.00	8706.00
105336	4	2011 - 3	384950.00	33078.00
142084	5	2011 - 4	312176.00	23078.00
172000	6	2011 - 5	399425.00	19034.00
209030	7	2011 - 6	394337.00	52714.00
245904	8	2011 - 7	407539.00	16423.00
285422	9	2011 - 8	425016.00	18817.00
320706	10	2011 - 9	575416.00	25599.00
370932	11	2011 - 10	628745.00	58213.00
431674	12	2011 - 11	771598.00	31312.00
516385	13	2011 - 12	315052.00	88719.00

Table I: Dataset of product sales and devolution

	Country	PositiveQuantity	NegativeQuantity	Index
44153	Bahrain	0.00	54.00	1
47164	Japan	0.00	45.00	2
50792	Israel	100.00	0.00	3
54381	Cyprus	144.00	0.00	4
72851	Channel Islands	259.00	4.00	5
58992	Poland	288.00	0.00	6
72247	Iceland	315.00	0.00	7
72986	Lebanon	386.00	0.00	8
69008	Greece	526.00	0.00	9
66153	Finland	765.00	0.00	10
43858	Belgium	792.00	9.00	11
70759	Singapore	1091.00	0.00	12
43780	Italy	1121.00	25.00	13
69624	Hong Kong	1121.00	0.00	14
43781	Portugal	2094.00	16.00	15
62999	Switzerland	2993.00	0.00	16
43420	Sweden	3097.00	1.00	17
45623	Spain	3845.00	8.00	18
45512	Australia	5644.00	0.00	19
44295	EIRE	8794.00	106.00	20
44206	Germany	9077.00	171.00	21
43857	France	9199.00	44.00	22
57394	Netherlands	20417.00	0.00	23
42482	United Kingdom	325648.00	88267.00	24

Table II: Information of buys in January of 2011

- Which country bought the major products quantity in January of 2011?
- Which country had major sales and devolutions of products?

C. *Inferential.*

An inferential question consists in the creation of an hypothesis to be solved analyzing the information.

- Was France the country with major sales in January of 2011?
  The Table III suggest that United Kingdom had the major sales in January of 2011

VIII. Exploratory Analysis

In this section was made an analysis in some scenarios to check the information related with the ”frequencies” and the ”central tendency measures”

The following information contains the resume of the exploratory analysis of the variables through the measure period

	Index	Country	PositiveQuantity	NegativeQuantity
100811	1	Saudi Arabia	80.00	5.00
38314	2	Bahrain	314.00	54.00
395473	3	RSA	352.00	0.00
157300	4	Brazil	356.00	0.00
72986	5	Lebanon	386.00	0.00
168150	6	European Community	499.00	2.00
7987	7	Lithuania	652.00	0.00
103599	8	Czech Republic	671.00	79.00
217685	9	Malta	970.00	26.00
89571	10	United Arab Emirates	982.00	0.00
69008	11	Greece	1557.00	1.00
14939	12	Iceland	2458.00	0.00
164465	13	USA	2458.00	1424.00
119192	14	Canada	2763.00	0.00
152713	15	Unspecified	3300.00	0.00
6609	16	Poland	3684.00	31.00
31983	17	Israel	4409.00	56.00
69624	18	Hong Kong	4773.00	4.00
31465	19	Austria	4881.00	54.00
70759	20	Singapore	5241.00	7.00
29733	21	Cyprus	6361.00	44.00
7215	22	Italy	8112.00	113.00
20018	23	Denmark	8235.00	47.00
20001	24	Channel Islands	9491.00	12.00
34084	25	Finland	10704.00	38.00
7135	26	Portugal	16258.00	78.00
1237	27	Norway	19338.00	91.00
7280	28	Belgium	23237.00	85.00
9784	29	Japan	26016.00	798.00
6422	30	Spain	27951.00	1127.00
5321	31	Switzerland	30630.00	305.00
30079	32	Sweden	36083.00	446.00
198	33	Australia	84209.00	556.00
27	34	France	112104.00	1624.00
1110	35	Germany	119263.00	1815.00
1405	36	EIRE	147447.00	4810.00
386	37	Netherlands	200937.00	809.00
1	38	United Kingdom	4733819.00	469990.00

Table III: Global information about the countries consumption

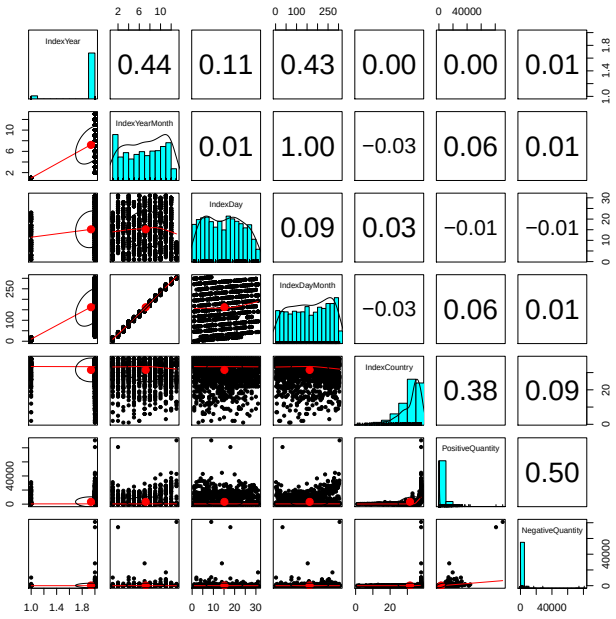
	Variable	mean	min	max	var	sd
1	Year	2010.92	2010.00	2011.00	0.07	0.27
2	Month	7.55	1.00	12.00	12.31	3.51
3	Day	15.02	1.00	31.00	75.07	8.66
4	Sold	10.45	0.00	80995.00	24115.74	155.29
5	Returned	0.89	0.00	80995.00	23424.97	153.05

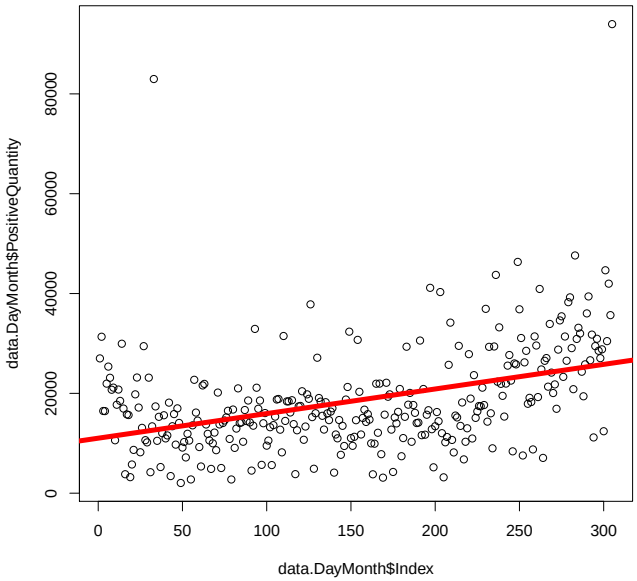
Table IV: Summary of main variables

As you can see in the table, it’s a general overview around the dataset because meeting the exploratory information of the data you can assume you know the data set.

IX. Multivariate Analysis

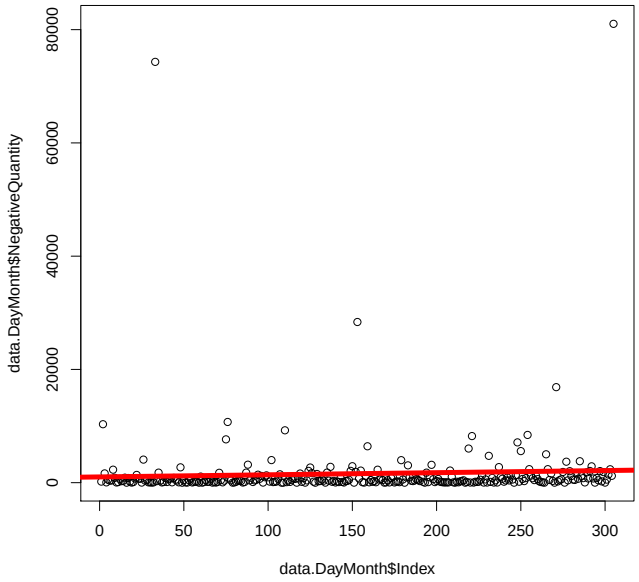
In this section you’ll find the correlation analysis of variables to check the importance of them. This analysis allows checking the influence of an attribute on other, this analysis is commonly used to discard any attribute if it isn’t important for other attributes.





As you can see in the previous chart we can think some conclusions like these

- Applying the Pearson's correlation, we can discard the value related with the Country, because the correlation index is too near than 0
- The year and the month are some of the most important variables of the data analysis of this study.
- The Sales affects directly the Devolutions



X. Finding Patterns

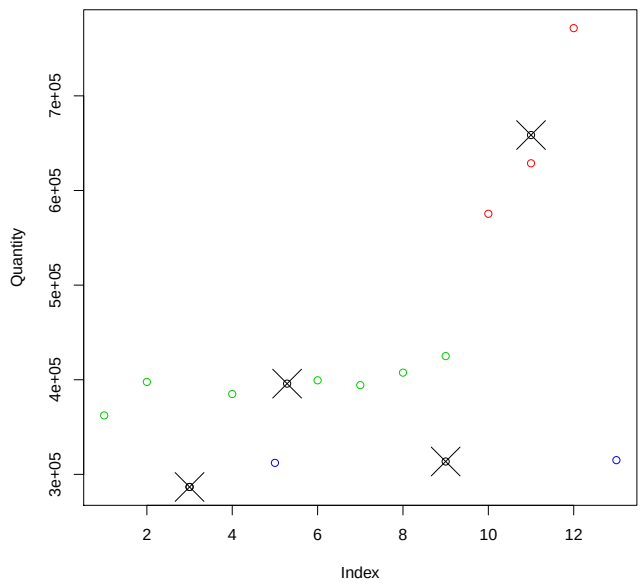
In this section we are going to establish some patterns between the data using the linear regression algorithm and the K-means algorithm.

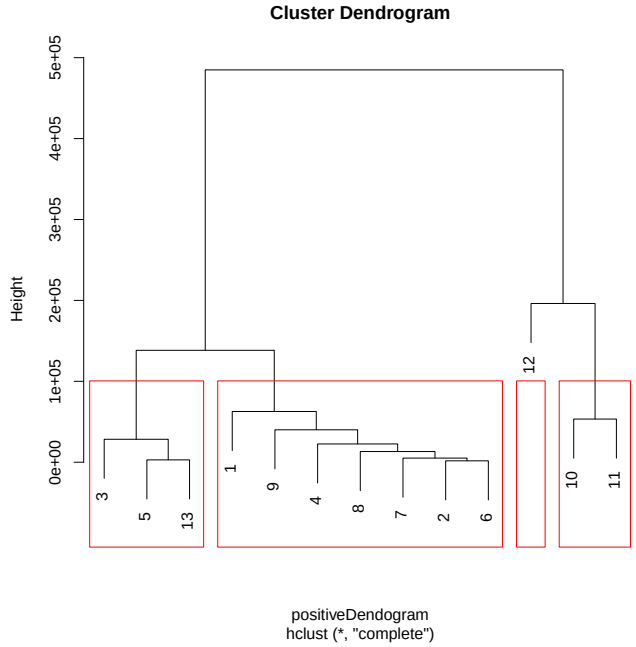
A. Linear Regression

This algorithm is used to build a model to reproduce the information drawing the nearest line to all the points.
For this dataset we'll use this algorithm in the relationship between the month days and the quantity.

B. Clustering

This algorithm is used to build a model to reproduce the information drawing the nearest centroid to all the points of the cluster.
For this dataset we'll use this algorithm in the relationship between the months days and the quantity.





XI. Conclusions

- The data analysis of the dataset allows establishing some relationships and patterns to describe the behavior of the data.
- The critical task of the data analysis consists in understanding the dataset, the data preparation, and the tasks related with the patterns definition.
- The linear regression and the clustering using the K-means algorithm are good options to analyze some behaviors of the marketing.

References

[1] G. Canavos, Probabilidad y estadistica aplicaciones y metodos. Mc-Graw Hill, 1988.

[2] H. Chen, BUSINESS INTELLIGENCE AND ANALYTICS: FROM BIG DATA TO BIG IMPACT, MIS Quarterly Vol. 36 No. 4, pp. 1165-1188/December 2012

[3] Daqing Chen, Sai Liang Sain, and Kun Guo, Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining, Journal of Database Marketing and Customer Strategy Management, Vol. 19, No. 3, pp. 197â€“208, 2012

[4] Ashishkumar Singh, Grace Rumantir, Annie South, Blair Bethwaite, Proceedings of the 2014 International Conference on Big Data Science and Computing.

[5] A decision-making framework for precision marketing, Zhen You, Yain-Whar Si, Defu Zhang, XiangXiang Zeng, Stephen C.H. Leung c, Tao Li, Expert Systems with Applications, 42 (2015) 3357â€“3367.

[6] XIE, Yihui. Dynamic documents with R and knitr. Chapman & Hall. Second edition. 2015.

[7] DE CASTRO KORGI, Rodrigo. El universo LATEX. Facultad de ciencias. Universidad Nacional de Colombia. Segunda edicion. 2003.