Analitica de transacciones en linea relacionadas con el conjunto de datos Online Retail

Raul Alejandro Buitrago Castellanos
Universidad Distrital
Francisco Jose de Caldas
Bogota, Colombia
Email: raulhabits@gmail.com

***Resumen*—En este articulo se muestra el analisis realizado al conjunto de datos "Online Retail" perteneciente al sitio web UCI Machine Learning Repository. Este analisis es realizado para la obtencion de conocimiento a partir de la informacion suministrada en los datos correspondiente a las transacciones en linea realizadas en un periodo de tiempo.**

***Keywords*—Conjunto de datos, mineria de datos, cientifico de datos, bigdata, inteligencia de negocios.**

I. Introducción

El análisis de datos siempre ha jugado un papel de vital importancia en la historia de la humanidad ya sea para comprender la naturaleza, mejorar la calidad de vida, el desarrollo de la economia, entre otras.
Además, la evolución de la tecnología ha representado un aumento considerable en cuanto a la capacidad de almacenamiento y procesamiento de información; lo cual permite el uso y tratamiento de grandes volumenes de datos.
La aplicacion del analisis de datos es infinita, puesto que todo aquello que puede ser clasificado y medido se puede analizar, por ejemplo el valor de la moneda frente a otros mercados, las visitas a un sitio web, el uso de alguna herramienta, la inteligencia de negocios, el analisis de ADN, etc. Entre las diversas tecnicas para dicho analisis se pueden destacar la estadistica, el calculo de probabilidades, la mineria de datos, el big data, entre otros.
Por ello se utilizaron conceptos relacionados con estadistica y "Bigdata" para obtener informacion relevante sobre el conjunto de datos.

II. Objetivos

Realizar un analisis estadistico, identificar las variables de mayor influencia, establecer comportamientos y/o patrones para predecir el comportamiento del mercado relacionado a las ventas en linea registradas en el conjunto de datos, utilizando tecnicas de mineria de datos y bigdata.

III. Marco teorico

Para comprender en su totalidad los terminos que se mencionan en este documento es necesario conocer los siguientes terminos.

III-A. Estadistica

III-B. Dataset o conjunto de datos

III-C. Business intelligence o inteligencia de negocios

III-D. Dataware house o bodega de datos

III-E. Data mining o Mineria de datos

III-F. Bigdata

III-G. Machine learning

IV. Estado del arte

V. METODOLOGIA

Para el desarrollo de este documento se plantearon las siguientes tareas

- Reconocimiento de la informacion
  - Identificar el dominio
  - Identificar un problema
  - Objetivo SMART
    - Specified
    - Measurable
    - Attainable
    - Relevant
    - Time able
- Preguntas de investigacion
  - Descriptivas
  - Exploratorias
  - Inferenciales
  - Predictivas
- Analisis exploratorio
  - Frecuencias
  - Medidas de tendencia central
  - Variabilidad
  - Desviacion estandar y varianzas
- Analisis multivariado
  - Analisis de correlacion de variables

VI. Preguntas de investigacion

El exito de una investigacion radica en la oportuna definicion de lo que se va a realizar, por ello las preguntas de investigacion se clasifican en varios tipos de acorde a la intencion del analisis.

- Caracter descriptivo. Las preguntas de caracter descriptivo sirven para identificar y conocer las caracteristicas del conjunto de datos.
    - ¿Cual es el rango de fechas en la medicion?
    - ¿Cual es el valor promedio por unidad?
    - ¿Cual es el producto mas vendido?
    - ¿Cual es el pais con mayor numero de transacciones?

- Caracter exploratorio. Las preguntas de caracter exploratorio consisten en la busqueda de patrones o relaciones que soporten una pregunta de investigacion
    - ¿Cual fue el país que compro la mayor cantidad de productos el mes de enero del año 2012?
    - ¿Cual fue el mes que mayor valor registro en las transacciones?
    - ¿Se registraron transacciones por mayor valor que 100000?
    - ¿Cual es el cliente que menos gasto dinero?

- Caracter inferencial. Las preguntas de caracter inferencial consisten en el planteamiento de una hipotesis que podria ser resuelta con el analisis respectivo de la informacion
    - ¿Fue Francia el país que gasto mas dinero?

- Caracter predictivo. Las preguntas de caracter predictivo permiten analizar el comportamiento de la informacion a traves del tiempo, para de esta forma descubrir, proyectar, o realizar hipotesis sobre estados futuros.
    - Pendiente

- Informacion sobre la cantidad de unidades involucradas en las transacciones.

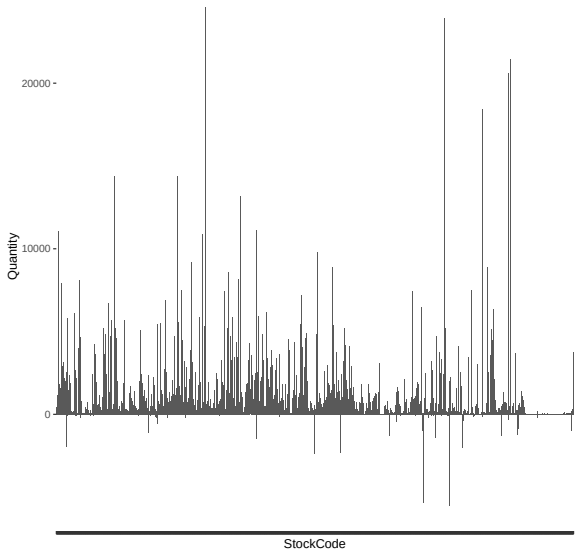
Tabla I

Statistic	N	Mean	St. Dev.	Min	Max
cantidad	232,959	9.410	242.464	−80,995	80,995

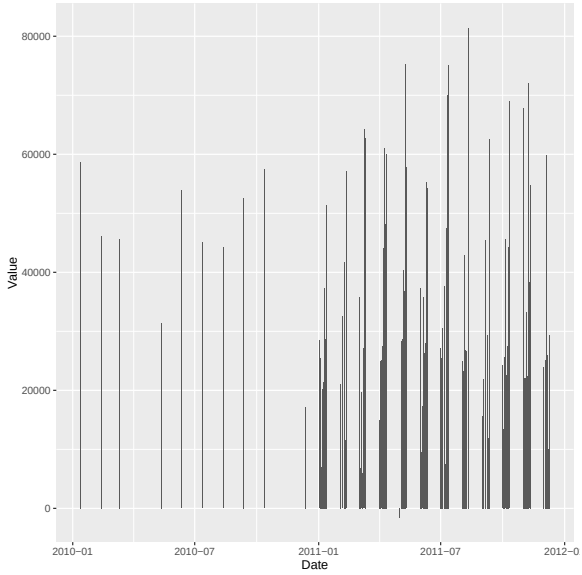
- Informacion correspondiente al periodo de tiempo de la medicion [1] ”2010-01-122011-12-10”Min. 1st Qu. Median Mean 3rd Qu. ”2010-01-122011-03-042011-06-092011-05-132011-09-06”Max. ”2011-12-10”

- Información correspondiente a los productos.

- Para conocer el producto mas vendido, se procede con una grafica que contiene la cantidad de productos por unidad vendidos



- Para conocer los registros de ventas de las transacciones se grafica la informacion de las transacciones mensuales



REFERENCIAS

[1] B. Klaus and P. Horn, Robot Vision. Cambridge, MA: MIT Press, 1986.

[2] L. Stein, “Random patterns,” in Computers and You, J. S. Brake, Ed. New York: Wiley, 1994, pp. 55-70.

[3] R. L. Myer, “Parametric oscillators and nonlinear materials,” in Nonlinear Optics, vol. 4, P. G. Harper and B. S.

[4] Wherret, Eds. San Francisco, CA: Academic, 1977, pp. 47-160.

[5] E. F. Moore, “Gedanken-experiments on sequential machines,” in Automata Studies (Ann. of Mathematical

[6] Studies, no. 1), C. E. Shannon and J. McCarthy, Eds. Princeton, NJ: Princeton Univ. Press, 1965, pp. 129-153.

[7] Westinghouse Electric Corporation (Staff of Technology and Science, Aerospace Div.), Integrated Electronic

[8] Systems. Englewood Cliffs, NJ: Prentice-Hall, 1970.