Posgrado de Ingeniería de Sistemas

Preprocesamiento de datos

Alberto Benavides
24 de febrero de 2020

1. Sobre los datos

Se cuenta con un conjunto de datos consistente en históricos de ventas y producción mensuales de entre los años 2014 a 2019 (ambos incluidos) de artículos de la empresa Fritos Totis [2]. Dicho conjunto de datos se encuentra en formato `XLSX`, propio de los archivos generados por Microsoft Excel [1]. El acomodo de estos datos está dado en bloques de siete renglones, uno por cada año, por cada artículo definido por un código, con 24 columnas por cada bloque que corresponden, por pares, a los doce meses del año. De cada par de columnas por renglón por artículo, la primera corresponde a la producción y la segunda a la venta. Un ejemplo de esto puede verse en la figura 1.1 (p. 2).

2. Valores separados por comas

Para poder hacer un análisis de datos es necesario tener un conjunto de los mismos en un formato que facilite su preprocesamiento [3]. Por este motivo, se procede a convertirlos en formato `CSV` [4] en renglones que informen, por columnas, el código del artículo, las producciones, ventas y la fecha en que fueron reportadas. Para este fin, se utilizó la librería `datetime` [5] con la que se convirtieron los meses y años a formato de fecha compatible para futuro preprocesamiento de la información. Los datos un ejemplo de los datos procesados se muestra en el cuadro 2.1 (p. 2). En tal cuadro puede verse que existen valores reportados con el símbolo de `-` por lo que falta aclarar qué significa dicho símbolo para poder analizar los datos.

		Enero		Febrero		Marzo	
AÑO	Codigo	Prod.	Venta	Prod.	Venta	Prod.	Venta
2014	A1	97,964	-	10,746	-	91,992	-
2015		54,005	-	78,098	-	124,682	-
2016		87,423	93,608	91,517	88,144	68,425	66,808
2017		84,142	83,304	60,306	79,328	46,342	74,032
2018		48,138	87,432	113,050	102,816	94,201	87,728
2019		92,202	89,414	102,440	112,046	133,470	99,705
2020		-	-	-	-	-	-

Figura 1.1: Extracto del artículo con código A1 obtenido del documento original en formato `XLSX`.

Tabla 2.1: Ejemplo de datos limpios tras procesamiento.

Código	Producción	Ventas	Fecha
A1	97, 964	-	2014-01-01
A1	10, 746	-	2014-02-01
A1	91, 992	-	2014-03-01
A1	50, 939	-	2014-04-01
A1	26, 826	-	2014-05-01

Referencias

[1] File formats that are supported in Excel. `https://support.office.com/en-us/article/file-formats-that-are-supported-in-excel-0943ff2c-6014-4e8d-aaea-b83d51d46247`. Accedido el 24-02-2020.

[2] Fritos Totis. `http://www.totis.com.mx/`. Accedido el 24-02-2020.

[3] Elisa Schaeffer. Práctica 1: Preparación de los datos con `bash`. `https://elisa.dyndns-web.com/teaching/comp/datasci/p1.html`. Accedido el 24-02-2020.

[4] Gerrit J.J. van den Burg. Handling Messy `CSV` Files. `https://towardsdatascience.com/handling-messy-csv-files-2ef829aa441d`. Accedido el 24-02-2020.

[5] Python Software Foundation. `datetime` – Basic date and time types. `https://docs.python.org/3.8/library/datetime.html`. Accedido el 24-02-2020.