Extracción, Transformación y Carga

Retail Data Analytics: ETL Project

Ramón Morillo Barrera
José Carlos Monescillo Calzado

Link al repositorio de github:
https://github.com/ramonmorillx/ETL-s

1. Objetivos del Proyecto

Este proyecto tiene como propósito principal realizar un proceso de ETL (Extracción, Transformación y Carga) aplicado a un conjunto de datos titulado 'Retail Data Analitycs'. El conjunto de datos a tratar incluye información acerca de datos históricos de ventas de 45 tiendas ubicadas en diferentes regiones.

El objetivo real de este proyecto es convertir estos datos en un valioso recurso analítico que facilite el estudio de tendencias y patrones subyacentes, con la finalidad de apoyar y hacer ver a los pequeños comercios que la recopilación, tratamiento y análisis de datos es una gran herramienta para potenciar las ventas y el beneficio.

Se garantizará que los datos mantengan su consistencia, estén limpios y preparados para poder realizar un análisis posterior en herramientas de business inteligence (inteligencia de negocio) y creación de modelos predictivos.

Objetivos específicos:

1. **Extracción:** Obtención de información de fuentes relacionadas con el tema a tratar, en nuestro caso, el dataset se encontraba en Kaggle por lo que nos facilita el proceso de extracción de los datos.
   Tendremos que traspasar los datos de nuestros archivos .csv a DataFrames de pandas en Python y verificar si los valores de cada fila y columna de nuestros DataFrames se corresponden con el número de valores en cada fila y columna de cada archivo .csv

2. **Transformación:** Este paso en nuestro caso, es el más importante ya que incluye la estandarización de formato, manejo de datos faltantes y reducción de inconsistencias, además del cálculo de las métricas establecidas. En este proceso transformaremos los datos con Python con el objetivo de dejarlos preparados para posteriores análisis.

3. **Carga:** Depositar los datos transformados y procesados en un sistema de datos estructurado. En este caso lo cargaremos como una base de datos relacional que estará optimizada para recibir consultas acerca de los datos.

Justificación de la importancia de realizar un proceso ETL

Un proyecto de ETL es primordial para asegurar tanto la calidad como la accesibilidad y la utilidad de un conjunto de datos. Si los datos no están correctamente transformados y limpios, tras su análisis en un proyecto real, las conclusiones podrían estar sesgadas o ser completamente erróneas.

Es por ello que el enfoque de la ETL asegura que nuestro dataset sea transformado y suponga un valioso recurso para todo aquel que quiera investigar patrones de compra o predecir ventas futuras, etc.

2. Conjunto de datos

El conjunto de datos utilizado fue recopilado en Kaggle, una plataforma web que reúne la comunidad Data Science más grande del mundo, con más de 536 mil miembros. Es una plataforma en la que cualquier persona tiene acceso a datos con licencia pública. Podemos encontrar en ella datos de diversas categorías, desde datos económicos a datos sobre características de películas, entre otros.

Link al conjunto de datos:
https://www.kaggle.com/datasets/manjeetsingh/retaildataset?select=Features+data+set.csv

Diccionario de datos
Nuestro dataset se compone de 3 ficheros .csv (comma separated values), por tanto, el diccionario de datos se compone de 3 tablas, cada una asociada a su correspondiente fichero.

- Features data

Name of column	Description	Type
Store	The store number	int
Date	The week	date
Temperature	Average temperature in the region	float
Fuel_Price	Cost of fuel in the region	float
MarkDown1	Anonymized data related to promotional markdowns. MarkDown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA	float
MarkDown2	Anonymized data related to promotional markdowns. MarkDown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA	float
MarkDown3	Anonymized data related to promotional markdowns. MarkDown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA	float
MarkDown4	Anonymized data related to promotional markdowns. MarkDown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA	float
MarkDown5	Anonymized data related to promotional markdowns. MarkDown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA	float

CPI	The consumer price index	float
Unemployment	The unemployment rate	float
IsHoliday	Whether the week is a special holiday week	bool

- Sales data

Name of column	Description	Type
Store	The store number	int
Dept	The department number	int
Date	The week	date
Weekly_Sales	Sales for the given department in the given store	float
IsHoliday	Whether the week is a special holiday week	bool

- Store data

Name of column	Description	Type
Store	Number of the store	int
Type	Type of the store	str
Size	Size on square feet of the store	int

Frecuencia de actualización de los datos.
Según la fuente de nuestros datos (Kaggle), la frecuencia de actualización de los mismos es 'nunca', es decir, se publicaron por primera vez en el año 2017 y no se han vuelto a actualizar. Observamos que no hay frecuencia de actualización para ninguno de nuestros datos, por tanto, no lo tendremos en cuenta en las métricas de data quality ya que no lo podemos medir y además evitaremos distorsionar los resultados de calidad del dato.

3. Características de los datos

Descripción de los tipos de datos manejados

Nuestros datos son datos estructurados, pues son datos organizados en un formato rígido, en filas y columnas, distribuidos en varias tablas que se ajustan bien a una base de datos relacional (SQL).

Los datos están organizados en 3 ficheros .csv, es un formato simple de archivo de texto donde los valores están separados por comas. Las ventajas que esto supone son que este tipo de formato es ampliamente utilizado, sencillo de manipular y compatible con casi todas las herramientas.

Nos encontramos con 1 tipo de dato que se repite en las 3 tablas, de los cuáles podremos beneficiarnos para realizar nuestro modelo de datos e interconectar las tablas entre sí.

- Store: que indica el identificador de cada tienda, esto nos facilita crear el modelo entidad relación y su modelo relacional en tablas

Es de tipo 'int', por lo que contiene el departamento de cada tienda como número entero. En nuestro caso, el departamento (Store), será nuestra columna óptima para relacionar las tablas.

Las diferentes tablas del proyecto contienen información relevante para nuestro proyecto de ETL, de las que podremos calcular diversas métricas que nos proporcionen información valiosa para el negocio.

- **Features data set:** Contiene datos adicionales relacionados con la tienda, el departamento y la actividad regional para las fechas indicadas.
- **Sales data set:** Datos históricos de ventas, que abarcan desde el 5 de febrero de 2010 hasta el 1 de noviembre de 2012.
- **Stores data set:** Contiene datos adicionales relacionados con la tienda, el tipo de tienda y el tamaño de las mismas para las fechas indicadas.

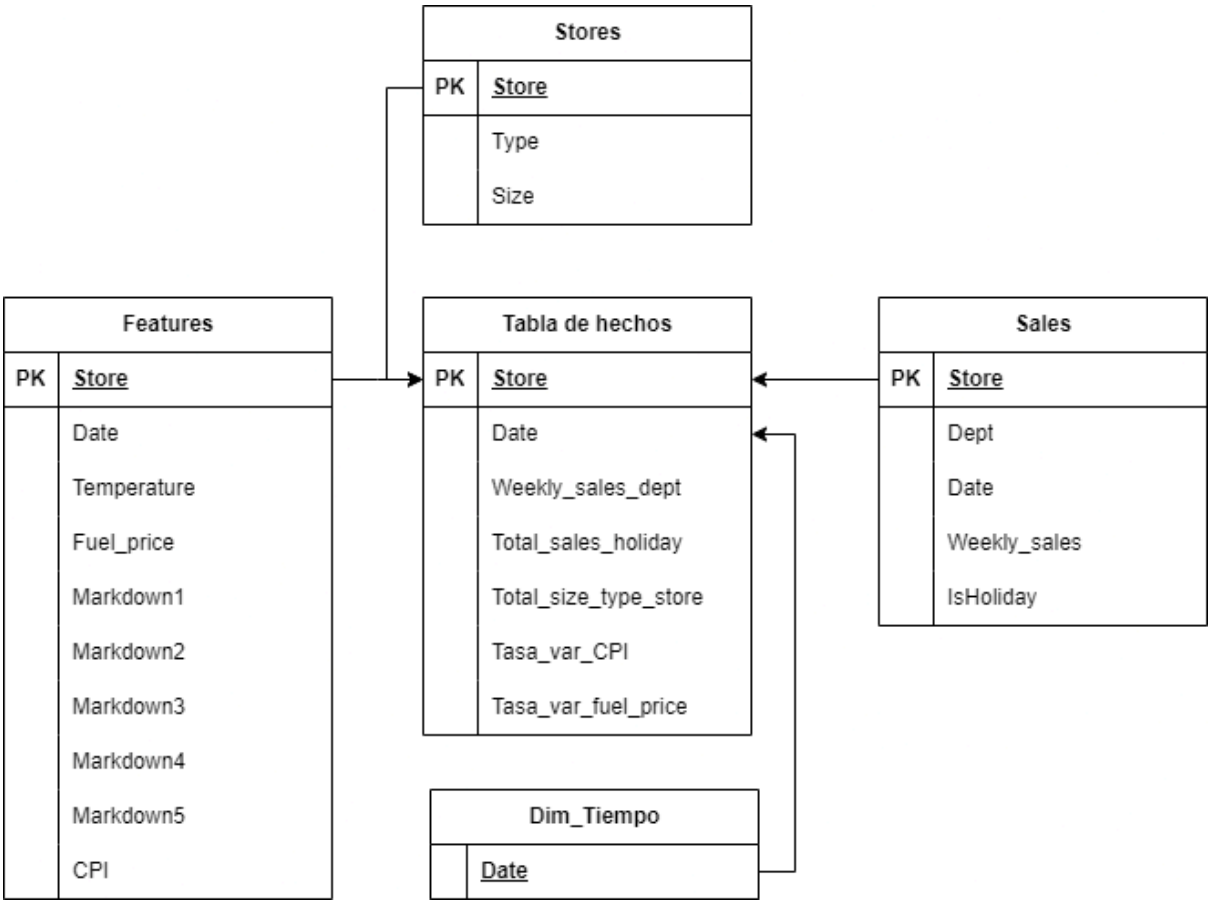
Definición del Modelo de Datos

Hemos optado por definir un modelo relacional de estrella en el que las diferentes tablas se unen mediante el id de la tienda (Store), ya que el id de la tienda es un registro común en todas las tablas de los archivos disponibles, siendo un registro único e identificable. El id será la clave primaria en todas las tablas que se unen a nuestra tabla de hechos, pues no puede haber dos registros con el mismo id ni registros nulos, pues no sabríamos a que tienda pertenece y tendríamos que eliminarlo.

Añadimos la tabla de dimensión de tiempo ya que en nuestros datos tenemos 2 tablas que incluyen la columna 'Date' que contiene fechas que van desde el 05/02/2010 hasta el 01/11/2012.

Las métricas que calcularemos son las siguientes:

- Ventas semanales totales por departamento
- Total de ventas por tienda en días festivos
- Tamaño total de tiendas según su tipo
- Tasa de variación media del 'CPI' a lo largo de los 3 años
- Tasa de variación media del 'Fuel_price' a lo largo de los 3 años



4. Calidad de los Datos

En cuanto a la calidad de los datos, solo se medirán aquellas métricas de las que tengamos suficiente información para calcular, debido a que, si damos por hecho valores para métricas que no podemos calcular, distorsionará nuestro valor real de calidad del dato.

Vamos a realizar un estudio de calidad de los datos, teniendo en cuenta las siguientes métricas de calidad del dato:

- **Precisión:** Evalúa la exactitud en la representación de los valores registrados, especialmente en términos numéricos y de decimales. En nuestro caso, que tengan el mismo número de decimales en cada tabla.
- **Estructura:** Mide si los datos siguen un formato o esquema estructural definido, como el tipo de dato y la longitud esperada.

- **Completitud**: Evalúa el porcentaje de datos presentes frente al total de datos esperados. En nuestro caso que no hayan datos faltantes.
- **Razonabilidad:** Evalúa si los datos tienen sentido y cumplen con ciertas reglas lógicas o umbrales, basándose en el contexto.
- **Identificabilidad:** Evalúa si cada dato puede ser identificado de manera única. En nuestro caso que los id no se repitan en una misma tabla, ya que debería ser clave primaria.

En este caso no mediremos:

- **Linaje**: No sabemos qué transformaciones previas han recibido nuestros datos. Necesitaríamos la información sobre el origen de los datos, qué transformaciones se realizaron antes de que el dataset fuese publicado y un historial de flujo de los datos desde su recolección hasta su publicación.
- **Puntualidad**: No tenemos una frecuencia de actualización disponible. Únicamente sabemos que se publicaron en 2017.
- **Integridad**: No tenemos datos que referencian directamente a otros datos.
- **Semántica:** La semántica requiere evaluar si los datos son fieles a su contexto. Sin conocer el propósito del dataset o las restricciones del dominio, no se puede determinar si los valores cumplen con su intención. Por ejemplo, en la columna 'Type' donde aparecen los valores 'A', 'B' y 'C', no sabemos a qué tipo de tienda corresponde exactamente cada uno.

Evaluación de calidad de datos

En nuestro caso, las métricas de calidad calculadas son:

- Precisión: 60.92%
- Completitud: 83.90%
- Estructura: 81.92%
- Razonabilidad: 81.83%
- Identificabilidad: 0.031%

La calidad total de nuestros datos es de: 61.71%

5. Limpieza de los Datos

En la limpieza y preparación de datos realizaremos los siguientes cambios para estandarizar los datos y facilitar la interpretabilidad y el análisis de los mismos.

En primer lugar, como no sabemos el valor de los datos NaN, vamos a imputar dichos valores por 0 en caso de datos numéricos y por el valor 'Desconocido' en el caso de datos textuales. Si contamos con algún id de alguna tienda que presente un valor nulo, directamente eliminaremos dicha fila, ya que la información de dicho registro carece de sentido.

En segundo lugar, estandarizamos todas las columnas numéricas, añadiendo 1 o 3 decimales, dependiendo de la mayoría de registros en cada tabla. Así tendremos todas las columnas con la misma longitud de decimales.

En tercer lugar, ajustar las columnas 'Date' para que presenten formato tipo fecha, comprobar que la columna 'IsHoliday' solo presenta valores booleanos [True, False], y corregir los registros que estén mal.

6. Problemas y Próximos Pasos

Como problemas encontrados, la mayoría de ellos surgen a la hora de calcular las métricas de calidad de los datos, hubieron métricas que no se pueden calcular debido a la falta de información. Pues no tenemos actualización de los datos ni tenemos conocimiento de que hayan recibido transformaciones previas, por tanto, descartamos el linaje y la puntualidad. Tanto la semántica como la integridad no han sido posibles de calcular debido al desconocimiento del contexto de los datos. En este caso, al ser un dataset de Kaggle, no tenemos un contexto completo de los datos, solo lo que se nos proporciona en la descripción.

Para solucionar este tipo de problemáticas en un ámbito real donde se mantiene un contacto directo con el cliente, deberíamos pedirle las indicaciones necesarias al mismo cliente para adaptar dicho contexto a nuestros datos. En el caso de que el cliente no nos proporcione la información necesaria, tendríamos que reunirnos con él para establecer una serie de directrices o indicaciones con el objetivo de facilitar el proceso de ETL.

Una gran mejora a futuro sería incluir un registro de actualizaciones de los datos, con los que poder calcular métricas y obtener valores de periodos semanales o anuales recientes. Estos datos facilitan la obtención de conclusiones que se ajusten al tiempo y a la realidad, promoviendo la aplicación de medidas correctivas o de promociones que potencien las ventas del negocio.

Como propuesta para futuros pasos me gustaría recomendar el uso de estos datos para la realización de estudios acerca del impacto de los diferentes factores que pueden influir en las ventas de un negocio, factores como el precio de la gasolina, el índice de precios al consumidor. Incluso analizar las ventas en periodos estacionales como pudieran ser meses concretos en los que las personas suelen tener vacaciones. Estos datos pueden proporcionar una clara idea sobre patrones de compra que se ajustan a ciertos comportamientos de la economía.