Reporte de calidad de datos

Se realizó el EDA correspondiente con el dataset de origen en el cual teníamos 11 datasets que contienen información sobre Olist, sus ventas, productos, vendedores pymes, entre otra información, del año 2016 año 2018.

El EDA realizado a los datasets de origen evidenció que estos datasets recibidos debian ser normalizado en algunas de las columnas (dimensiones), estos datasets no fueron manejados mediante descargas en la fuente, se decidió darles manejo subiendolos a un repositorio en gitHub y cargandolos directamente en un dataframe de pandas.

Para un informe más detallado se reportará cada dataset por separado para así evidenciar correctamente la calidad de todos los datos de origen:

1. **closed_deals_dataset:**

Cuenta con 14 dimensiones y 872 registros

Tiene la siguiente cantidad de valores nulos en las siguientes dimensiones:

business_segment: 1

lead_type: 6

lead_behaviour_profile: 177

has_company: 779

has_gtin: 778

average_stock: 776

business_type: 10

declared_product_catalog_size: 773

En este dataset tenemos un porcentaje de nulos de: 28%, no tenemos filas duplicadas, ni valores duplicados, se realizó una normalización en el tipo de dato en columna 'won_date', sin afectar el dato y el formato en el que estaba originalmente.

El dataset contaba con los siguientes tipos de datos: Object, float(64) y ocupaba la siguiente cantidad espacio en memoria: 92.2 KiB.

2. **customers_dataset:**

Cuenta con 5 dimensiones y 99441 registros

En este dataset no tenemos valores nulos, no tenemos filas duplicadas, ni valores duplicados, no se generó normalización en ninguna de las columnas.

El dataset contaba con los siguientes tipos de datos: Object, Int(64) y ocupaba la siguiente cantidad espacio en memoria: 3.8 MiB.

3. **geolocation_dataset**

Cuenta con 5 dimensiones y 1000163 registros. En este dataset no tenemos valores nulos, tenemos un total de 261831 valores duplicados y un porcentaje de 12.8% de filas duplicadas, no se generó normalización en ninguna de las columnas.

El dataset contaba con los siguientes tipos de datos: Object, Int(64), float(64) y ocupaba la siguiente cantidad espacio en memoria: 3.8 MiB

4. **marketing_qualified_leads_dataset:**

Cuenta con 4 dimensiones y 8000 registros. Tiene la siguiente cantidad de valores nulos en la siguiente dimensión:

En este dataset tenemos un porcentaje de nulos de: 0.2%, no tenemos filas duplicadas, ni valores duplicados, se realizó una normalización en el tipo de dato en columna 'first_contact_date', sin afectar el dato y el formato en el que estaba originalmente.

El dataset contaba con los siguientes tipos de datos: Object, y ocupaba la siguiente cantidad espacio en memoria: 250.1 KiB

5. **order_items_dataset:**

Cuenta con 7 dimensiones y 112650 registros. En este dataset no tenemos valores nulos, no tenemos filas duplicadas, ni valores duplicados, se realizó una normalización en el tipo de dato en columna 'shipping_limit_date', sin afectar el dato y el formato en el que estaba originalmente.

El dataset contaba con los siguientes tipos de datos: Object, Int(64), float(64) y ocupaba la siguiente cantidad espacio en memoria: 6.0 MiB

6. **order_payments_dataset**

Cuenta con 5 dimensiones y 103886 registros. En este dataset no tenemos valores nulos, no tenemos filas duplicadas, ni valores duplicados, no se generó normalización en ninguna de las columnas.

El dataset contaba con los siguientes tipos de datos: Object, Int(64), float(64) y ocupaba la siguiente cantidad espacio en memoria: 4.0 MiB

7. **order_reviews_dataset:**

Cuenta con 7 dimensiones y 99224 registros. Tiene la siguiente cantidad de valores nulos en las siguientes dimensiones:

review_comment_title: 87656

review_comment_message: 58247

En este dataset tenemos un porcentaje de nulos de: 21.0%, no tenemos filas duplicadas, ni valores duplicados, se realizó una normalización en el tipo de dato en las columnas 'review_creation_date' y 'review_answer_timestamp', sin afectar el dato y el formato en el que estaba originalmente.

El dataset contaba con los siguientes tipos de datos: Object, Int(64), float(64) y ocupaba la siguiente cantidad espacio en memoria: 5.3 MiB

8. **orders_dataset:**

Cuenta con 8 dimensiones y 99441 registros. Tiene la siguiente cantidad de valores nulos en las siguientes dimensiones:

order_approved_at: 160

order_delivered_carrier_date: 1783

order_delivered_customer_date: 2965

En este dataset tenemos un porcentaje de nulos de: 0.6%, no tenemos filas duplicadas, ni valores duplicados, se realizó una normalización en el tipo de dato en las columnas 'order_purchase_timestamp', 'order_approved_at', 'rder_delivered_carrier_date', 'order_delivered_customer_date' y 'order_estimated_delivery_date', sin afectar el dato y el formato en el que estaba originalmente.

El dataset contaba con los siguientes tipos de datos: Object, y ocupaba la siguiente cantidad espacio en memoria: 6.1 MiB

9. **products_dataset:**

Cuenta con 9 dimensiones y 32951 registros. Tiene la siguiente cantidad de valores nulos en las siguientes dimensiones:

product_name_lenght: 610

product_category_name: 610

product_description_lenght: 610

product_photos_qty: 610

product_weight_g: 2

product_length_cm: 2

product_height_cm: 2

product_width_cm: 2

En este dataset tenemos un porcentaje de nulos de: 0.8%, no tenemos filas duplicadas, ni valores duplicados, no se generó normalización en ninguna de las columnas. Este contaba con los siguientes tipos de datos: Object, float(64), y ocupaba la siguiente cantidad espacio en memoria: 2.3 MiB

10. **sellers_dataset:**

Cuenta con 4 dimensiones y 3095 registros. En este dataset no tenemos valores nulos, no tenemos filas duplicadas, ni valores duplicados, no se generó normalización en ninguna de las columnas. El dataset contaba con los siguientes tipos de datos: Object, Int(64), y ocupaba la siguiente cantidad espacio en memoria: 96.8 KiB

11. **category_translat_dataset:**

Cuenta con 2 dimensiones y 71 registros. En este dataset no tenemos valores nulos, no tenemos filas duplicadas, ni valores duplicados, no se generó normalización en ninguna de las columnas. El dataset contaba con los siguientes tipos de datos: Object, y ocupaba la siguiente cantidad espacio en memoria: 1.2 KiB.