Documentación del Proyecto ETL de Customer Churn

Descripción del Problema

El abandono de clientes (churn) es uno de los desafíos más significativos que enfrentan las empresas basadas en suscripciones, como las compañías de telecomunicaciones. El churn ocurre cuando los clientes cancelan sus servicios, lo que lleva a una pérdida de ingresos recurrentes. Comprender los factores que contribuyen al churn puede ayudar a las empresas a tomar medidas proactivas para mejorar la retención de clientes.

Objetivo:

Este proyecto tiene como objetivo construir un pipeline ETL (Extract, Transform, Load) para analizar datos de clientes, identificar los factores que contribuyen al churn y extraer información valiosa para reducir el abandono. El resultado final incluirá un panel de control para visualizar métricas clave de churn y perspectivas sobre los clientes.

Contexto

Las compañías de telecomunicaciones suelen ofrecer múltiples servicios, como teléfono, internet y suscripciones de TV. Factores como la duración del contrato, la calidad del servicio, los precios y el soporte al cliente pueden influir en la decisión de un cliente de quedarse o abandonar el servicio. Identificar señales tempranas de churn puede ayudar a las empresas a implementar intervenciones específicas para mejorar la satisfacción y la retención de clientes.

Descripción del Conjunto de Datos

El conjunto de datos principal para este proyecto es el **Telco Customer Churn Dataset** obtenido de Kaggle. Este conjunto de datos contiene información detallada sobre los clientes, incluidos los tipos de servicio, detalles del contrato, métodos de pago y un indicador de churn.

Principales Columnas del Dataset:

- **CustomerID: Identificador único del cliente.**
- **Gender: Género del cliente (Male/Female).**
- **SeniorCitizen: Indica si el cliente es un adulto mayor (1) o no (0).**

- **Partner: Si el cliente tiene pareja (Yes/No).**
- **Dependents: Si el cliente tiene dependientes (Yes/No).**
- **Tenure: Tiempo en meses que el cliente ha estado con la empresa.**
- **PhoneService: Si el cliente tiene servicio de teléfono (Yes/No).**
- **MultipleLines: Si el cliente tiene múltiples líneas telefónicas (Yes/No/No phone service).**
- **InternetService: Tipo de conexión a internet (DSL/Fiber optic/No).**
- **OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies: Servicios adicionales contratados (Yes/No/No internet service).**
- **Contract: Tipo de contrato (Month-to-month/One year/Two year).**
- **PaperlessBilling: Si el cliente tiene facturación sin papel (Yes/No).**
- **PaymentMethod: Método de pago (Electronic check, Mailed check, Bank transfer, Credit card).**
- **MonthlyCharges: Monto mensual que paga el cliente.**
- **TotalCharges: Monto total pagado por el cliente.**
- **Churn: Variable objetivo que indica si el cliente ha abandonado el servicio (Yes/No).**

Tamaño del Conjunto de Datos:

- Filas: 7,043
- Columnas: 21

Proceso

1. Fuentes de Datos

- **Fuente:** Kaggle (archivo CSV)

2. Extracción de Datos

- **Configuración del Ambiente de Trabajo**
  Para garantizar la reproducibilidad y estabilidad del proceso, se configuró un ambiente virtual en Python donde se instalaron las librerías necesarias. Entre estas se encuentran pandas para la manipulación de datos, kaggle

para la descarga del dataset, mysql-connector-python para la conexión con MySQL y python-dotenv para la gestión segura de credenciales.

- **Descarga de Datos desde Kaggle**
  El dataset *Telco Customer Churn*, disponible en Kaggle, fue descargado mediante la API de la plataforma. Para ello, se configuró la autenticación con una clave de acceso, lo que permitió la descarga automatizada del archivo CSV con los datos de los clientes.
- **Carga y Validación de los Datos**
  Una vez descargado, el dataset fue cargado en un DataFrame de pandas, donde se realizó una primera inspección para verificar su integridad. Se revisaron aspectos como la cantidad de registros y columnas, los tipos de datos y la presencia de valores nulos o inconsistentes.
- **Limpieza y Preparación Inicial**
  Se aplicaron algunas transformaciones básicas para garantizar la calidad de los datos antes de su almacenamiento. Entre ellas, la conversión de ciertos valores al formato numérico, la eliminación de espacios en blanco y la gestión de valores vacíos.
- **Completar registros**: Debido a que el dataset tiene menos registros que los solicitados se completo con datos aleatorios generados, se garantizo que se mantengan los **CustomerID diferentes para cada registro**
- **Almacenamiento en MySQL**
  Para facilitar su acceso y procesamiento posterior, los datos fueron almacenados en una base de datos MySQL. Se creó una estructura de tabla acorde a las características del dataset, asegurando que cada columna tuviera un tipo de dato apropiado. Luego, se insertaron los registros extraídos, verificando que la carga se realizara correctamente.
- **Validación Final**
  Finalmente, se realizaron consultas sobre la base de datos para verificar que los datos fueron cargados correctamente. Se revisó la cantidad de registros almacenados y se confirmó que las estructuras de datos coincidieran con la fuente original.

3. Transformación de Datos

Objetivo de la Transformación

La fase de transformación tiene como objetivo limpiar, modificar y enriquecer los datos extraídos para hacerlos más útiles en el análisis del churn. Se aplicaron diversas técnicas para mejorar la calidad de los datos y facilitar su interpretación en el dashboard final.

3.1 Limpieza de Datos

Conversión de *TotalCharges* a numérico

- **Motivo:** La columna ***TotalCharges*** contenía valores en formato de texto, lo que impedía realizar cálculos con ella.
- **Acción realizada**:
  - Se convirtió ***TotalCharges*** a tipo numérico (float).
  - Se eliminaron las filas con valores nulos en ***TotalCharges***, ya que representaban una cantidad mínima de registros.

3.2 Conversión de variables categóricas

- **Motivo:** Las variables categóricas no pueden ser utilizadas directamente en cálculos y modelos analíticos, por lo que se aplicó one-hot encoding para convertirlas en variables numéricas.
- **Acción realizada:**
  - Se aplicó one-hot encoding a ***Contract, PaymentMethod, InternetService, MultipleLines, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, gender, Partner, Dependents, PhoneService, PaperlessBilling,*** y ***Churn.***

3.3 Transformación de *SeniorCitizen*

- **Motivo:** ***SeniorCitizen*** es una variable numérica binaria (0 o 1), pero al tratarse de una condición, es más intuitivo expresarla como una variable categórica (Yes / No).
- **Acción realizada:** Se mapearon los valores 0 → "No" y 1 → "Yes".

3.4 Creación de Nuevas Variables

3.4.1 Agrupación de *tenure* en rangos

Motivo: ***tenure*** (tiempo en meses con la empresa) es una variable numérica con un amplio rango de valores. Agruparla en intervalos facilita el análisis visual y la segmentación de clientes.

Acción realizada:

- Se agruparon los clientes en 5 categorías de permanencia (0-12, 13-24, etc.).
- Se ajustaron los límites para evitar duplicados en los bins.

3.4.2 Creación de *AvgMonthlySpend* (Gasto Promedio Mensual)

Motivo: ***AvgMonthlySpend*** ayuda a identificar patrones en los clientes con altos o bajos gastos y su relación con el churn.

Acción realizada:

- Se creó la columna **AvgMonthlySpend = TotalCharges / tenure.**
- Se evitaron divisiones por 0 imputando valores faltantes con **MonthlyCharges.**

3.4.3 Creación de Indicadores de Clientes

Motivo: Se generaron métricas adicionales para segmentar a los clientes según su comportamiento.

Indicadores creados:

- Clientes con contrato a largo plazo (***LongTermContract***)
- Clientes nuevos (***isNewCustomer***)
- Número de servicios contratados (***MultipleServices***)
- Clientes con bajo gasto mensual (***LowSpender***)

4. Carga de Datos

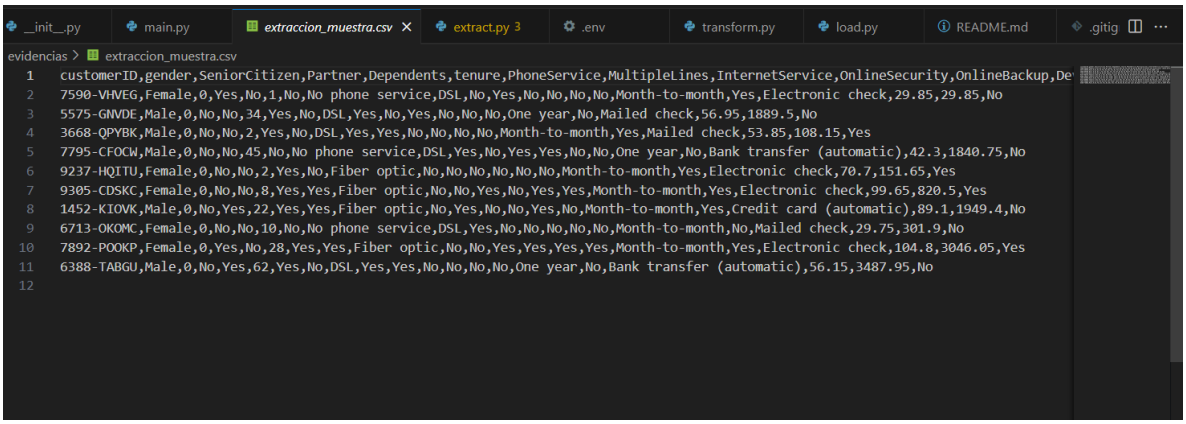
- Cargar los datos transformados en una base de datos final para su análisis.
- Asegurar la integridad de los datos e indexación para una recuperación optimizada.

5. Creación de Dashboard

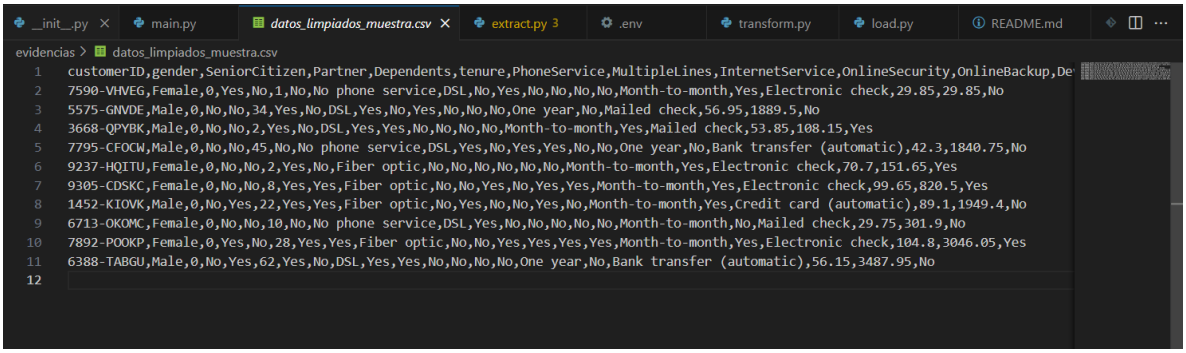
- Usar una herramienta de visualización (por ejemplo, Power BI o Looker Studio) para crear paneles interactivos que muestren:
  - Tasas de churn por demografía y tipo de servicio.
  - Tendencias de churn a lo largo del tiempo.
  - Factores clave que contribuyen al churn.

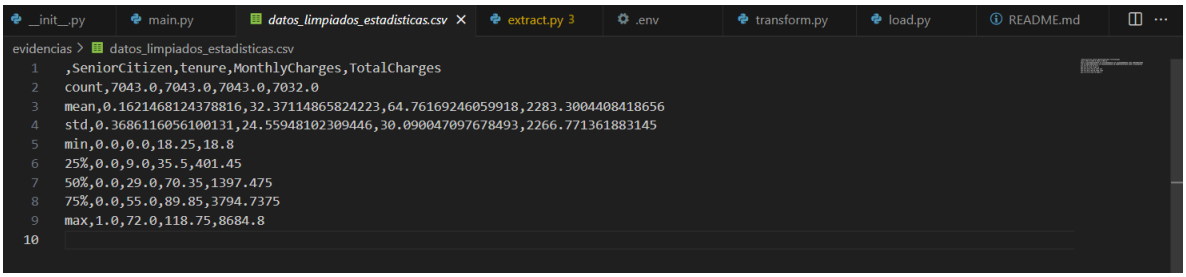
Evidencias del Proceso ETL

1. **Extracción de Datos**
   - **Extraccion_muestra.csv: Se guarda una muestra de los primeros 10 registros extraídos del dataset, para verificar que los datos fueron descargados correctamente.**



- **Datos_limpiados_valores_nulos.csv, datos_limpiados_muestra.csv y datos_limpiados_estadisticas.csv: Proceso de validación y limpieza de datos.**





---

2. Carga de Datos en MySQL

- **Consulta SQL: Se ejecuta esta consulta para verificar la creación de la base de datos, la tabla de datos extraídos y cuántos registros han sido insertados en la base de datos.**

```
mysql> Show databases
    -> ;
+--------------------+
| Database           |
+--------------------+
| customer_churn     |
| ejec1_etl          |
| etl_db             |
| information_schema |
| modelo_uao         |
| mysql              |
| performance_schema |
| sys                |
+--------------------+
8 rows in set (0.00 sec)

mysql>
```

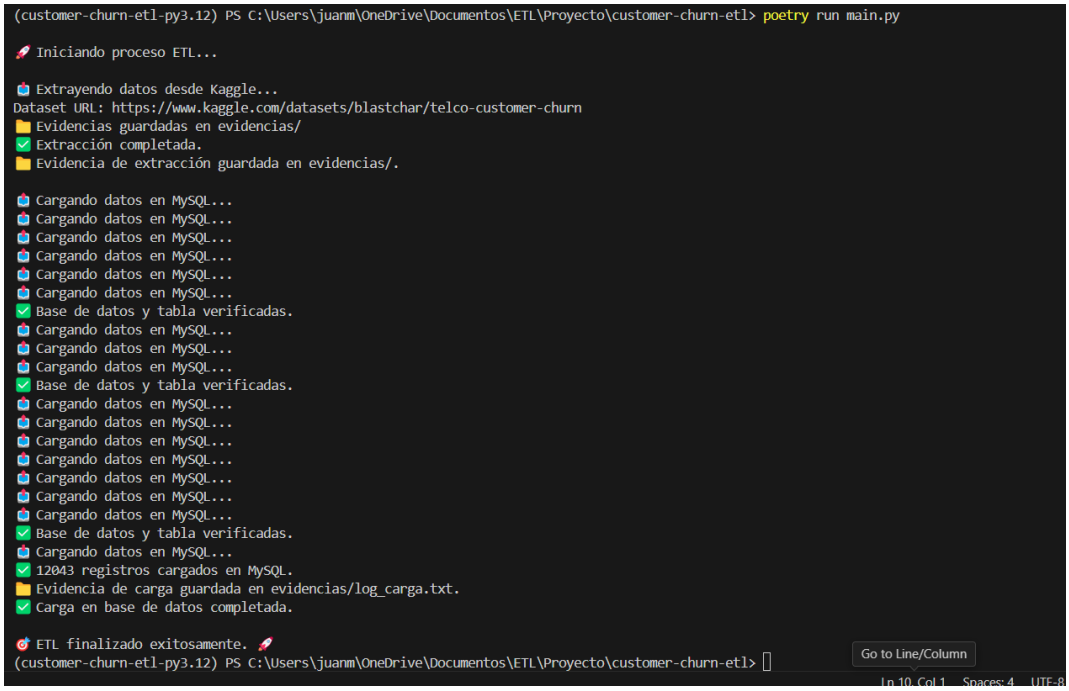
```
mysql> Show tables;
+--------------------------+
| Tables_in_customer_churn |
+--------------------------+
| customer_churn_extract   |
+--------------------------+
1 row in set (0.00 sec)

mysql>
```

```
mysql> SELECT COUNT(*) AS total_registros FROM customer_churn_extract;
+------------------+
| total_registros |
+------------------+
|           12043 |
+------------------+
1 row in set (0.00 sec)

mysql>
```

3. **Respuesta en Terminal: Cadena de respuestas de la terminal a lo largo del proceso de extracción para validar el funcionamiento de la lógica.**



4. **Transformaciones: Código de las funciones con las transformaciones.**

```python
def clean_total_charges(df):
    """Convierte TotalCharges a numérico y elimina valores nulos."""
    df["TotalCharges"] = pd.to_numeric(df["TotalCharges"], errors="coerce")
    df.dropna(subset=["TotalCharges"], inplace=True)
    return df

def encode_categorical(df):
    """Aplica one-hot encoding a variables categóricas."""
    categorical_cols = ["Contract", "PaymentMethod", "InternetService", "MultipleLines",
                        "OnlineSecurity", "OnlineBackup", "DeviceProtection", "TechSupport",
                        "StreamingTV", "StreamingMovies", "gender", "Partner", "Dependents", "PhoneService",
    df = pd.get_dummies(df, columns=categorical_cols, drop_first=True)
    return df

def categorize_senior_citizen(df):
    """Convierte SeniorCitizen en una variable categórica."""
    df["SeniorCitizen"] = df["SeniorCitizen"].map({0: "No", 1: "Yes"})
    return df

def create_tenure_groups(df):
    """Crea una nueva columna para agrupar el tenure en rangos."""
    max_tenure = df["tenure"].max() + 1  # Asegurar que el último bin sea único
    bins = [0, 12, 24, 48, 72, max_tenure]
    labels = ["0-12", "13-24", "25-48", "49-72", "73+"]
    df["tenure_group"] = pd.cut(df["tenure"], bins=bins, labels=labels, include_lowest=True)
    return df

def calculate_avg_monthly_spend(df):
    """Crea la columna AvgMonthlySpend para analizar patrones de gasto."""
    df["AvgMonthlySpend"] = df["TotalCharges"] / df["tenure"]
    df["AvgMonthlySpend"] = df["AvgMonthlySpend"].fillna(df["MonthlyCharges"])  # Evitar divisiones por 0
    return df
```

5. Resultados: Datos resultantes de las transformaciones realizadas.

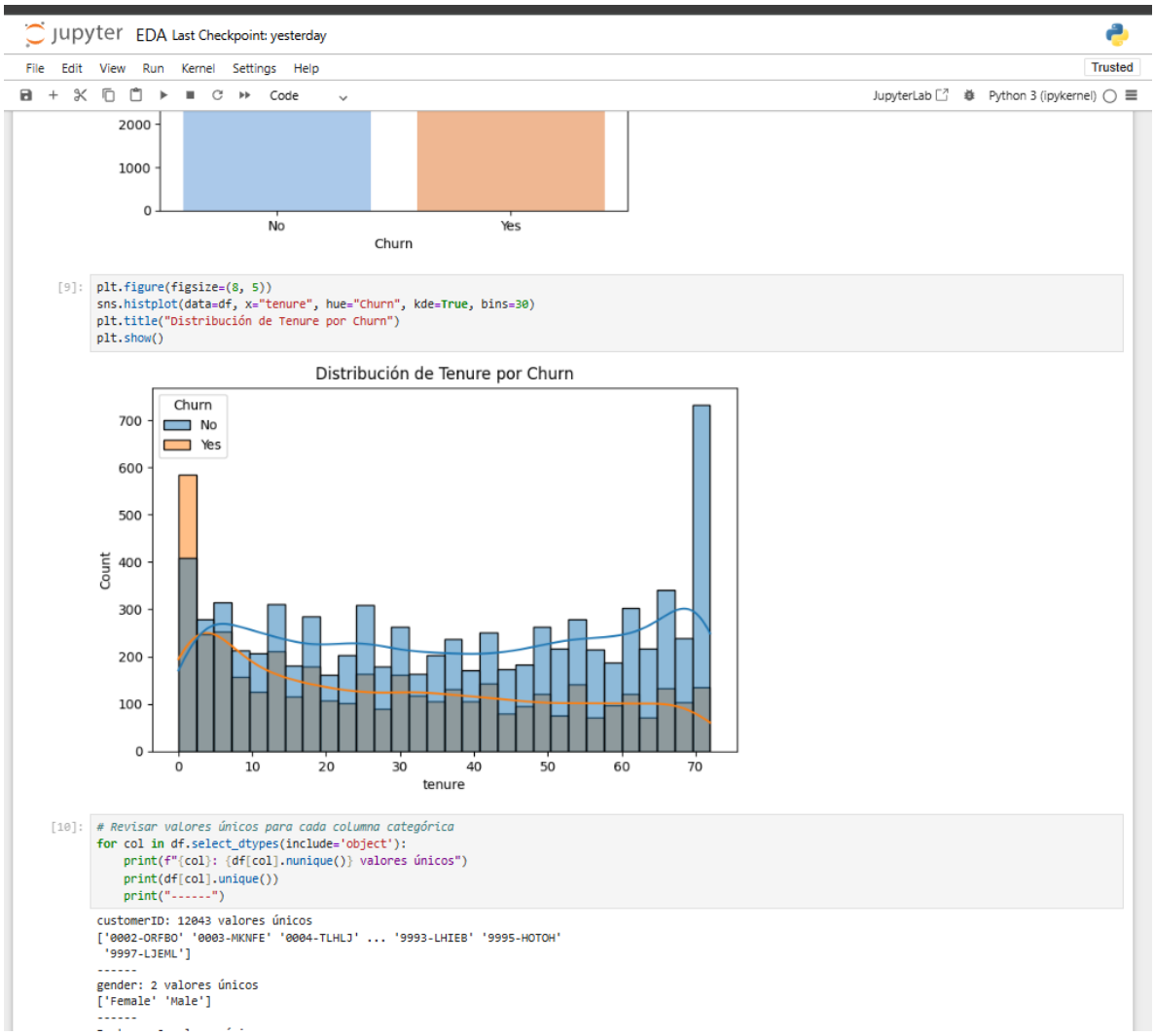
evidencias > extraccion_muestra.csv

```
customerID,gender,SeniorCitizen,Partner,Dependents,tenure,PhoneService,MultipleLines,InternetService,OnlineSecurity,OnlineBackup,Dev
7590-VHVEG,Female,0,Yes,No,1,No,No phone service,DSL,No,Yes,No,No,No,Month-to-month,Yes,Electronic check,29.85,29.85,No
5575-GNVDE,Male,0,No,No,34,Yes,No,DSL,Yes,No,Yes,No,No,No,One year,No,Mailed check,56.95,1889.5,No
3668-QPYBK,Male,0,No,No,2,Yes,No,DSL,Yes,Yes,No,No,No,No,Month-to-month,Yes,Mailed check,53.85,108.15,Yes
7795-CFOCW,Male,0,No,No,45,No,No phone service,DSL,Yes,No,Yes,Yes,No,No,One year,No,Bank transfer (automatic),42.3,1840.75,No
9237-HQITU,Female,0,No,No,2,Yes,No,Fiber optic,No,No,No,No,No,No,Month-to-month,Yes,Electronic check,70.7,151.65,Yes
9305-CDSKC,Female,0,No,No,8,Yes,Yes,Fiber optic,No,No,Yes,No,Yes,Yes,Month-to-month,Yes,Electronic check,99.65,820.5,Yes
1452-KIOVK,Male,0,No,Yes,22,Yes,Yes,Fiber optic,No,Yes,No,No,Yes,No,Month-to-month,Yes,Credit card (automatic),89.1,1949.4,No
6713-OKOMC,Female,0,No,No,10,No,No phone service,DSL,Yes,No,No,No,No,No,Month-to-month,No,Mailed check,29.75,301.9,No
7892-POOKP,Female,0,Yes,No,28,Yes,Yes,Fiber optic,No,No,Yes,Yes,Yes,Yes,Month-to-month,Yes,Electronic check,104.8,3046.05,Yes
6388-TABGU,Male,0,No,Yes,62,Yes,No,DSL,Yes,Yes,No,No,No,No,One year,No,Bank transfer (automatic),56.15,3487.95,No
```

evidencias > transformacion_muestra.csv

```
customerID,SeniorCitizen,tenure,MonthlyCharges,TotalCharges,Contract_One year,Contract_Two year,PaymentMethod_Credit card (automatic
7590-VHVEG,No,1,29.85,29.85,False,False,False,True,False,False,False,True,False,False,False,False,True,False,False,False,False,Fals
5575-GNVDE,No,34,56.95,1889.5,True,False,False,False,True,False,False,False,False,False,True,False,False,False,True,False,False,Fal
3668-QPYBK,No,2,53.85,108.15,False,False,False,False,True,False,False,False,False,False,True,False,True,False,False,False,False,Fal
7795-CFOCW,No,45,42.3,1840.75,True,False,False,False,False,False,False,True,False,False,True,False,False,False,True,False,True,Fals
9237-HQITU,No,2,70.7,151.65,False,False,False,True,False,True,False,False,False,False,False,False,False,False,False,False,False,Fal
9305-CDSKC,No,8,99.65,820.5,False,False,False,False,True,False,True,False,False,True,False,False,False,False,False,True,False,False,False
1452-KIOVK,No,22,89.1,1949.4,False,False,True,False,False,True,False,False,True,False,False,True,False,False,False,False,False,Fals
6713-OKOMC,No,10,29.75,301.9,False,False,False,False,True,False,False,True,False,False,True,False,False,False,False,False,False,Fal
7892-POOKP,No,28,104.8,3046.05,False,False,False,True,False,True,False,False,True,False,False,False,False,False,True,False,True,Fal
6388-TABGU,No,62,56.15,3487.95,True,False,False,False,False,False,False,False,False,False,True,False,True,False,False,False,False,F
```

6. EDA: Notebook con el análisis exploratorio de datos.



7. **Respuesta en terminal: Cadena de respuestas de la terminal a lo largo del proceso de ETL para validar el funcionamiento de la lógica.**