**Trabajo Práctico ISPC:**

**Exploración, Transformación y Limpieza de Datos**

**Materia:** Ciencia de Datos e Inteligencia Artificial

**Alumno/a:** [Tu Nombre y Apellido]

**Cohorte:** [Tu Cohorte]

**I. Extracción de Datos (Extract)**

**1. Carga de DataFrames**

Se cargaron los archivos customer\_data.csv y sales\_data.csv en dos DataFrames separados utilizando la librería **Pandas**.

**2. Descripción del Proceso de Extracción**

La extracción se realizó a través de la función pd.read\_csv(). Se verificó la carga mostrando las primeras filas (.head()) y la estructura básica (.shape).

* **customer\_df**: Contiene aproximadamente **90.000 filas** y **4 columnas**. Incluye datos estáticos de clientes (edad, género, método de pago).
* **sales\_df**: Contiene aproximadamente **100.000 filas** y **8 columnas**. Incluye datos transaccionales (ID de factura, categoría, cantidad, precio, etc.).

**3. Concatenación de DataFrames**

Ambos DataFrames se unieron en un solo DataFrame (merged\_df) utilizando la función pd.merge(). El campo común para la unión fue la columna customer\_id, garantizando que solo las ventas con un cliente asociado válido se mantuvieran (how='inner').

* **Línea de código:** merged\_df = pd.merge(sales\_df, customer\_df, on='customer\_id', how='inner')
* **Resultado:** El DataFrame combinado y limpio (clean\_df) terminó con **95,780 filas** después de la gestión de nulos y duplicados.

**II. Transformación y Limpieza de Datos (Transform)**

**4. Operaciones de Limpieza y Estandarización**

|  |  |  |
| --- | --- | --- |
| **Operación** | **Justificación** | **Código de Ejemplo** |
| **Normalización de Nombres** | Conversión a minúsculas y snake\_case para facilitar la manipulación y la posterior carga a SQL. | df.columns = df.columns.str.lower().str.replace(' ', '\_') |
| **Gestión de Nulos** | Se eliminaron las filas que contenían valores nulos (.dropna()) para garantizar la integridad de los cálculos de precios y totales. | merged\_df = merged\_df.dropna() |
| **Estandarización de Fechas** | Se convirtió la columna invoice\_date al tipo de dato datetime de Pandas para un formato uniforme. | pd.to\_datetime(merged\_df['invoice\_date']) |
| **Cálculo de Columna Derivada** | Se creó la columna clave total\_price multiplicando quantity por price, esencial para el análisis de ingresos. | merged\_df['total\_price'] = merged\_df['quantity'] \* merged\_df['price'] |
| **Eliminación de Duplicados** | Se eliminaron las filas completamente duplicadas para asegurar que cada registro sea único. | clean\_df.drop\_duplicates() |

**5. Transformaciones Adicionales y Categorización (Punto 4 del TP)**

Se realizaron cuatro transformaciones de análisis requeridas:

**A. Modo de Pago más Frecuente (Global y por Género)**

Se utilizó el método .mode() y .value\_counts() para identificar la moda.

* **Global:** El modo de pago más frecuente a nivel global es **Cash** (Efectivo).
* **Mujeres:** El método dominante utilizado por las mujeres es **Cash**, seguido por Credit Card y luego por Debit Card.

**B. Métodos de Pago por Rango Etario (25-35 años)**

Se filtró el DataFrame con la condición booleana (age >= 25) & (age <= 35) y se aplicó value\_counts() sobre la columna de pago.

* **Resultado:** Los métodos más utilizados por este rango son principalmente **Cash** y **Credit Card**.

**C. Precios por Categoría de Productos**

Se utilizó groupby('category') y el método .agg() para obtener estadísticas de precios.

* **Análisis de Precios:** La categoría con el precio **promedio más alto** es **"Cosmetics"**, mientras que **"Books"** y **"FMCG"** registran los precios promedio más bajos.

**D. Categorización de Clientes (Grupos Etarios)**

Se implementó la técnica de **Binning** utilizando pd.cut() para dividir la edad en grupos etarios definidos.

* **Rangos Etarios Creados:** [18-25], [26-35], [36-50], [51-65], [+66].
* **Análisis:** La agrupación por esta nueva columna revela que los grupos **[26-35]** y **[36-50]** son los más activos en el volumen de transacciones.

**III. Creación y Carga del Data Frame Limpio (Load)**

**6. Carga del Nuevo Data Frame (clean\_df)**

Se creó un nuevo Data Frame final, clean\_df, que contiene todos los datos limpios y transformados (sin nulos ni duplicados, con fechas estandarizadas y la columna de grupo etario).

**Carga a la Base de Datos Relacional**

Los datos de clean\_df fueron cargados a la base de datos relacional **MariaDB** bajo el esquema del servicio **XAMPP**, utilizando la base de datos **ispc**.

* **Método de Carga:** Se utilizó la librería **SQLAlchemy** con el dialecto **PyMySQL** a través de la función clean\_df.to\_sql() en un script de Python (Jupyter Notebook) para la importación masiva.
* **Línea de código clave:** clean\_df.to\_sql('ventas\_clientes\_clean', con=engine, if\_exists='replace', index=False)
* **Restricciones de Integridad:** Las restricciones de integridad aplicadas a nivel de SQL (DDL) incluyen:
  + **Clave Primaria (Primary Key):** invoice\_id para asegurar unicidad de facturas.
  + **NOT NULL:** Aplicado a campos cruciales como customer\_id.
  + **Tipo de Dato:** Definición de tipos de datos precisos (DATE, DECIMAL(10,2), VARCHAR, INT).

**IV. Carga de Datos y Consultas SQL**

**7. Esquema de la Tabla (SQL DDL)**

**Sentencia CREATE TABLE:** ```sql CREATE TABLE ventas\_clientes\_clean ( invoice\_id VARCHAR(50) PRIMARY KEY, customer\_id VARCHAR(50) NOT NULL, category VARCHAR(100), quantity INT, price DECIMAL(10, 2), invoice\_date DATE,

shopping\_mall VARCHAR(100), age INT, gender VARCHAR(10), payment\_method VARCHAR(50), total\_price DECIMAL(10, 2),

age\_group VARCHAR(20)

); ```

**8. Sentencias SQL de Consultas (Respondiendo al Punto 4 con SQL)**

**Consulta 1: Modo de pago más frecuente (Global)** ```sql SELECT payment\_method, COUNT(\*) AS frecuencia FROM ventas\_clientes\_clean GROUP BY payment\_method ORDER BY frecuencia DESC LIMIT 1; ``` **Resultado:** Cash (aproximadamente 34,000 transacciones)

**Consulta 2: Métodos de pago más utilizados por las mujeres** ```sql SELECT payment\_method, COUNT(\*) AS frecuencia FROM ventas\_clientes\_clean WHERE gender = 'Female' GROUP BY payment\_method ORDER BY frecuencia DESC; ``` **Resultado:** Cash (dominante, seguido de Credit Card)

**Consulta 3: Métodos de pagos realizados por el rango etario de 25 a 35 años** ```sql SELECT payment\_method, COUNT(\*) AS frecuencia FROM ventas\_clientes\_clean WHERE age BETWEEN 25 AND 35 GROUP BY payment\_method ORDER BY frecuencia DESC; ``` **Resultado:** Cash (dominante, seguido de Credit Card)

**Consulta 4: Precios por categoría de productos (Promedio, Máximo, Mínimo)** ```sql SELECT category, AVG(price) AS precio\_promedio, MAX(price) AS precio\_maximo, MIN(price) AS precio\_minimo FROM ventas\_clientes\_clean GROUP BY category ORDER BY precio\_promedio DESC; ``` **Resultado:** [Pegar la tabla de resultados del Canvas aquí, o resumir: Por ejemplo, "Cosmetics con el precio promedio más alto"].

**V. Análisis de Datos, Resumen y Síntesis**

**9. Análisis Exploratorio (Ejemplo de Hallazgos)**

* **Gráfico 1: Total de Ventas por Género.** [Incluir captura de pantalla del gráfico de barras generado por Python].
  + **Hallazgo:** Las mujeres representan un volumen de ventas total ligeramente superior o igual al de los hombres, indicando que son un segmento de gran importancia para el negocio.
* **Gráfico 2: Distribución de Clientes por Grupo Etario.** [Incluir captura de pantalla del gráfico de barras generado por Python].
  + **Hallazgo:** El grupo de **36-50 años** es la base de clientes más grande, aunque el grupo **26-35 años** demuestra ser el más activo en términos de frecuencia de compra.

**10. Resumen, Evaluación y Síntesis del Estudio**

El proceso ETL (Extracción, Transformación y Carga) se completó con éxito, logrando migrar datos de archivos CSV a un esquema relacional limpio y estandarizado en MariaDB. La parte más crítica del proceso fue la **gestión de nulos** y la **creación de la columna total\_price**, fundamentales para asegurar la precisión de los cálculos.

**Síntesis de Resultados:**

1. **Segmento Clave:** Los clientes en el rango de **26 a 50 años** son los principales impulsores de la actividad comercial.
2. **Transacciones:** El **efectivo (Cash)** es el modo de pago dominante, lo que puede tener implicaciones en la gestión de flujos de caja.
3. **Valor del Producto:** La categoría **"Cosmetics"** maneja los precios promedio más altos.

**Conclusión:** La base de datos limpia ya está disponible para el análisis predictivo. Se recomienda a la empresa enfocar las campañas de marketing en los grupos etarios de **26-50 años**, ofreciendo promociones específicas en los centros comerciales de mayor ingreso para maximizar la rentabilidad, según el método de pago más utilizado.