Informe de Análisis

Análisis Exploratorio de Datos (EDA) con Python y Dashboard en Power BI

Introducción

En este informe, se presentan los resultados del análisis realizado a un conjunto de datos ficticios de una cadena de tiendas de comestibles mediante un Análisis Exploratorio de Datos (EDA) utilizando Python y la creación de un dashboard en Power BI para analizar las Ventas, los Clientes y los Empleados.

Análisis Exploratorio de Datos (EDA)

El EDA es una fase crucial en el análisis de datos, ya que permite comprender la estructura de los datos y detectar patrones, anomalías y relaciones significativas. Para llevar a cabo este análisis, se utilizaron los archivos de datos originales y se emplearon diversas librerías de Python como por ejemplo pandas y datetime entre otras.

Proceso de Limpieza de Datos

Se analizó cada archivo individualmente para ver qué transformaciones eran necesarias.

```
Análisis inicial del archivo "categories":

1. Tipo de datos a cambiar: no es necesario hacer ningún cambio

2. Categorías a cambiar:

    - Cambiar el nombre de las columnas:

        - **CategoryId** -> Category_Id
        - **CategoryName** -> Category_Name

3. Columnas con valores nulos: no hay

4. Columnas a eliminar: no es necesario eliminar ninguna columna
```

```
Análisis inicial del archivo "cities":

1. Tipo de datos a cambiar: no es necesario hacer ningún cambio

2. Categorías a cambiar:

    - Cambiar nombre de las columnas:

        - **CityID** -> City_Id
```

```
        - **CityName** -> City_Name
        - **CountryID** -> Country_Id

3. Columnas con valores nulos: no hay

4. Columnas a eliminar: no es necesario eliminar ninguna columna
```

```
Análisis inicial del archivo "countries":

1. Tipo de datos a cambiar: no es necesario hacer ningún cambio

2. Categorías a cambiar:

    - Cambiar nombre de las columnas:

        - **CountryID** -> Country_Id
        - **CountryName** -> Country_Name
        - **CountryCode** -> Country_Code

3. Columnas con valores nulos:

    - **CountryCode**: 1 valor nulo

4. Columnas a eliminar: **CountryCode** no aporta valor a nuestro
análisis
```

```
Análisis inicial del archivo "customers":

1. Tipo de datos a cambiar: no es necesario hacer ningún cambio

2. Categorías a cambiar:

    - Cambiar nombre de las columnas:

        - **CustomerID** -> Customer_Id
        - **CityID** -> City_Id

3. Columnas con valores nulos:

    - **MiddleInitial**: 977 valores nulos

4. Columnas a eliminar:
    - Eliminar 'MiddleInitial' ya que no es necesaria para nuestro
análisis
    - Concatenar las columnas 'FirstName' y 'LastName' en una sola,
'Full_Name' y eliminar las otras 2
    - Eliminar 'Address' ya que no es necesaria para nuestro análisis
```

```
Análisis inicial del archivo "employees":

1. Tipo de datos a cambiar:

    - **BirthDate**: object -> datetime
    - **HireDate**: object -> datetime

2. Categorías a cambiar:

    - **BirthDate**: dejar sólo la fecha
    - **Gender**: cambiar 'M' por Male y 'F' por Female
    - **HireDate**: dejar sólo la fecha
    - Cambiar nombre de las columnas:

        - **EmployeeID** -> Employee_Id
        - **BirthDate** -> Birth_Date
        - **CityID** -> City_Id
        - **HireDate** -> Hire_Date

    - Crear columna **Age** (edad del empleado) a partir de Birth_Date y
**Seniority** (antigüedad del empleado en la empresa) a partir de
Hire_Date

3. Columnas con valores nulos: no hay

4. Columnas a eliminar:
    - Eliminar 'MiddleInitial' ya que no es necesaria para nuestro
análisis
    - Concatenar las columnas 'FirstName' y 'LastName' en una sola,
'Full_Name' y eliminar las otras 2
```

```
Análisis inicial del archivo "products":

1. Tipo de datos a cambiar:

    - **ModifyDate**: object -> datetime
    - **VitalityDays**: float -> int

2. Categorías a cambiar:

    - Cambiar nombre de las columnas:

        - **ProductID** -> Product_Id
        - **ProductName** -> Product_Name
        - **CategoryID** -> Category_Id
        - **ModifyDate** -> Modify_Date
        - **IsAllergic** -> Is_Allergic
        - **VitalityDays** -> Vitality_Days
```

```
    - **ModifyDate**: dejar sólo la fecha
    - **Is_Allergic**: cambiar True por Allergic, False por Not Allergic

3. Columnas con valores nulos: no hay

4. Columnas a eliminar: no son de interés para nuestro análisis
```

```
Análisis inicial del archivo "sales":

1. Tipo de datos a cambiar:

    - **SalesDate**: object -> datetime

2. Categorías a cambiar:

    - Cambiar nombre de las columnas:

        - **SalesID** -> Sales_Id
        - **SalesPersonID** -> Sales_Person_Id
        - **CustomerID** -> Customer_Id
        - **ProductID** -> Product_Id
        - **SalesDate** -> Sales_Date

    - **Sales_Date**: dejar sólo la fecha y crear una nueva columna sólo
con la hora (no los minutos) para ver las ventas por horas

    - Crear nueva columna Total_Price obteniendo el precio del producto
del archivo 'products' y aplicando el descuento.

3. Columnas con valores nulos:

    - **SalesDate**. 67526 valores nulos

4. Columnas a eliminar:

    - **TransactionNumber**: no aporta valor a nuestro análisis
    - **TotalPrice**: todos los valores son ceros
```

A modo de resumen, tras finalizar el EDA, se identificaron y corrigieron problemas de calidad de datos como:

- Homogeneizar el nombre de las columnas
- Eliminar columnas innecesarias que no aportaban valor al análisis
- Gestión de valores nulos y faltantes
- Transformaciones varias de fechas
- Creación de nuevas columnas con información

Los datos limpios se guardaron en otra carpeta añadiendo a cada archivo la terminación "_clean" para diferenciarlos de los archivos originales. Este proceso aseguró que los datos fueran precisos y adecuados para el análisis posterior.

Creación del Dashboard en Power BI

Una vez limpiados los datos, se importaron en Power BI para elaborar un dashboard interactivo que permitiera visualizar y analizar diferentes aspectos del negocio. El dashboard incluye varias visualizaciones clave enfocadas en:

Ventas

- Distribución de ventas por productos y categorías.
- Tendencias de ventas a lo largo del tiempo.
- Comparación de ventas por horas.

Clientes

- Análisis demográfico de los clientes.
- Segmentación de clientes según comportamiento de compra.
- Identificación de los principales clientes.

Empleados

- Rendimiento individual.
- Análisis demográfico de los empleados.
- Análisis de la edad y la antigüedad.

Modelado y transformación de los datos

Combinación de tablas

Con el fin de conseguir un modelo en estrella, se realizan las siguientes combinaciones de tablas:

- Se combina la tabla “cities_clean” con “countries_clean”.
- Se combina la tabla “customers_clean” con “cities_clean”, y “employees_clean” con “cities_clean”.
- Se combina la tabla “products_clean” con “categories_clean”.

Agrupación de tablas

Con el fin de obtener la categoría de los clientes en base al valor de “Total_Price” de la tabla “sales_clean”, se duplica la tabla “sales_clean” y se agrupan la suma de “Total_Price” por “Customer_Id”. Posteriormente, se combina con “customers_clean” y se crea una columna donde se aplica la lógica para categorizar a los clientes.

Hallazgos y consideraciones

Hay **dos hallazgos clave** que caben ser resaltados antes de comentar las visualizaciones y análisis del punto anterior.

El primer hallazgo clave es que **hay productos cuya categoría no concuerda con la descripción del producto**. Por ejemplo, dentro de la categoría "Meat" (carne) hay productos como plátanos, agua carbonatada o vino. En el caso de que fueran datos reales, sería conveniente advertir a la empresa que genera los datos que hay productos mal categorizados.

El segundo hallazgo clave es que, aunque en el archivo "countries" aparecen países de todo el mundo, el archivo "cities" sólo contiene datos de ventas en Estados Unidos (país con identificador número 32). En consecuencia, **el archivo "sales" sólo contiene datos de ventas en Estados Unidos.**

Una consideración a tener en cuenta también. Para el análisis de este conjunto de datos, asumo que la ciudad de la tabla "customers" es la ciudad de procedencia del cliente. Del mismo modo asumo que, la ciudad de la tabla "employees", es la ciudad en la que trabaja el empleado y, en consecuencia, la ciudad en la que hay tienda física y donde analizaremos la facturación por ciudades.

Conclusiones

Estas son las conclusiones tras el análisis.

Ventas

- El producto que más factura es "Bread – Calabrese Baguette" con $18,87 millones seguido de "Shrimp – 31/40" con $18,72 millones.
- Por categorías, la categoría con mayor facturación es "Confections" con casi $557 millones. Hay que destacar aquí que, dado el problema de productos mal categorizados, este dato no es fiable. De todas formas, se deja en el dashboard ya que, si en el futuro se corrige este error, aparecería el análisis correctamente sólo con actualizar el informe.
- Como se puede observar en el gráfico de "Ventas por Semana", las ventas son muy estables semana a semana en el periodo del que disponemos datos. En semana 1 y semana 19, la facturación es menor debido a que se abrieron menos días.
- En el periodo dado, la hora con más facturación es entre las 16 h y las 17 h, y la que menos entre las 3 h y las 4 h.

Clientes

- Tucson es la ciudad de la que mayor número de clientes proceden, 1.104.
- El cliente que más ha gastado es Wayne Chan de la ciudad de Washington.
- Segmentación de clientes según importe de compra:
    - VIP (Compras >= $80.000): suponen el 17,67% del total
    - Medio (Compras entre $23.000 y $80.000): suponen el 75,02%
    - Bajo (Compras < $23.000): suponen el 7,3%

- En el gráfico de Ventas por Hora se puede analizar individualmente para cada cliente sus hábitos de consumo, a qué horas compras más y a qué horas menos.

Empleados

- El empleado con mayor facturación es Devon Brewer en la ciudad de Baltimore con $190 millones.
- La ciudad con mayor número de empleados es Lubbock con 3 empleados, seguida con 2 por Baltimore, Columbus y New Orleans.
- Los empleados se distribuyen de la siguiente forma: 8 mujeres (34,78%) y 15 hombres (65,22%).
- La distribución de las edades por género es la siguiente:
    - Mujeres:
        - Media de edad 59,25 años
        - Edad mínima: 44 años
        - Edad máxima: 74
    - Hombres:
        - Media de edad: 56,47
        - Edad mínima: 36
        - Edad máxima: 74
- La antigüedad media en la empresa de los empleados es de 12,38 años para las mujeres y 11,6 para los hombres.

Recomendaciones

Basado en los hallazgos del análisis, se recomienda:

- Corregir la tabla “products”, en concreto, asignar a cada producto su categoría correcta.
- Sería conveniente añadir una tabla “stores” donde se indique la localización correcta de cada tienda.
- Implementar estrategias de marketing personalizadas para las distintas categorías de clientes identificadas.
- La edad media de los trabajadores (tanto mujeres como hombres) es bastante elevada. Sería conveniente planificar a medio plazo el relevo de los empleados con gente más joven.