Proyecto Final

Profesora:

Natalia Herrera

Estudiantes:

Gary Carballo Barrantes

Marco Fonseca León

André Montoya Fernández

LEAD University

2023

Introducción

En el ámbito actual del analisis de datos, la calidad de los datos es un componente crucial para la toma de decisiones acertadas e informadas. Este proyecto se centra en proporcionar una sólida base en términos de limpieza de datos, abordando los problemas más comunes que pueden surgir al trabajar con conjuntos de datos diversos. La misión en este caso es la detección de esos problemas en con las diferentes funciones aplicadas para saber a qué nos enfrentamos.

Desarrollo

Tenemos dos archivos que contienen los códigos para efectuar la detección de los problemas en cuanto a calidad de los datos. El archivo **Main.py** es el archivo de principal que nos proporciona la información necesaria para cargar y analizar un archivo CSV. Lo que hace el código dentro de este archivo es lo siguiente:

1. El programa solicita al usuario el nombre y la ruta del archivo CSV, almacenando la información en un diccionario.

2. Utilizamos la función de **cargar archivo** para tomar el archivo CSV y convertir las columnas con fechas al formato adecuado.

3. Se realiza un análisis detallado del archivo cargado, incluyendo la detección de caracteres especiales, análisis de datos, verificación de columnas de fecha y evaluación de la calidad del archivo.

4. Se realizan llamadas a diversas funciones del módulo **funciones** para realizar análisis específicos, proporcionando una visión integral del contenido y la calidad del archivo.

Ahora en el archivo de **funciones.py** tenemos todas las funciones aplicadas para poder realizar el análisis del archivo y por ende la detección de esos errores que nos podrían complicar nuestro trabajo más adelante.

El archivo trabaja en este orden:

1. Cargamos el CSV y trata de convertir columnas con fechas al formato adecuado, proporcionando un DataFrame listo para el análisis.

2. Cuenta el número de mayúsculas y letras especiales en un texto, facilitando el análisis de patrones de caracteres en cadenas de texto.

3. Detecta y cuenta el número total de mayúsculas y letras especiales en columnas, ofreciendo información sobre la diversidad de caracteres presentes.

4. Realiza un análisis detallado de los datos, incluyendo el conteo de nulos, duplicados y muestra algunos datos sin espacios, proporcionando una visión general de la calidad de los datos.

5. Verifica las columnas que contienen fechas y trata de convertirlas a objetos de fecha, gestionando posibles errores de formato y rango.

6. Evalúa la calidad del archivo basándose en factores como nulos y duplicados, proporcionando un puntaje global y una evaluación cualitativa de la calidad del archivo.

Resultados

```
Introduce el nombre del archivo: proveedores.csv
Introduce la ruta del archivo proveedores.csv: C:\Users\Marco Fonseca\OneDrive\Escritorio\ultimo cuatri\jueves\proveedores.csv

Información de proveedores.csv:

Mayúsculas totales: 1881124
Letras especiales totales: 4692593

Análisis de datos para proveedores.csv:

c:\Users\Marco Fonseca\OneDrive\Escritorio\ultimo cuatri\jueves\funciones.py:81: FutureWarning: DataFrame.applymap has been deprecated. Use DataFrame.map instead.
  datos_sin_espacios = df.applymap(lambda x: x.strip() if isinstance(x, str) else x)
Nulos por columna:
cedula_proveedor          0
nombre_proveedor          0
tipo_proveedor            0
tamano_proveedor          0
fecha_constitucion    21359
fecha_expiracion      21740
direccion                 6
codigo_postal          1022
provincia              1066
canton                 1103
distrito               1145
dtype: int64

Duplicados en el conjunto de datos: 0

   cedula_proveedor                              nombre_proveedor     tipo_proveedor tamano_proveedor  ... codigo_postal provincia         canton       distrito
0        3002667850  ASOCIACION COSTARRICENCE PARA EL ESTUDIO E INT...  Nacional Jurídico          Grande  ...         11501  San Jose  Montes de Oca      San Pedro
1        3101654585                  GRUPO ALFA NOVA SOCIEDAD ANONIMA  Nacional Jurídico          Grande  ...         20101  Alajuela       Alajuela       Alajuela
2         304260976                      TOMAS ADOLFO BRENES MADRIGAL    Nacional Físico         Pequeña  ...         30501   Cartago      Turrialba      Turrialba
3        3102615222                    THE TRAVELSHOP CENTER LIMITADA  Nacional Jurídico         Pequeña  ...         10203  San Jose         Escazu     San Rafael
4        3101215441     DEPOSITO IRAZU LOS HEREDIANOS SOCIEDAD ANONIMA  Nacional Jurídico          Grande  ...         40103   Heredia        Heredia  San Francisco

[5 rows x 11 columns]

Resultados finales:
                  Porcentaje de Nulos  Duplicados  Porcentaje de Duplicados
cedula_proveedor             0.000000           0                       0.0
nombre_proveedor             0.000000           0                       0.0
tipo_proveedor               0.000000           0                       0.0
tamano_proveedor             0.000000           0                       0.0
fecha_constitucion          51.487320           0                       0.0
fecha_expiracion            52.405747           0                       0.0
direccion                    0.014463           0                       0.0
codigo_postal                2.463600           0                       0.0
provincia                    2.569665           0                       0.0
canton                       2.658856           0                       0.0
distrito                     2.760100           0                       0.0
```

```
Verificación de columnas de fecha:

Columnas de fecha detectadas:
['fecha_constitucion']

No se encontraron errores en las columnas de fecha.

Evaluación de la calidad del archivo:

Calidad Global del Archivo: 4.80
¡Archivo necesita limpieza urgente!
```

Conclusión

Después de aplicar con éxito las funciones diseñadas para la detección de factores nulos, duplicados, mayúsculas y letras especiales, logramos obtener un análisis completo de la calidad del archivo. Este proyecto no se limitó únicamente a la selección y limpieza de un conjunto de datos, sino que se enfocó en la identificación de errores.

Durante el desarrollo, nos enfrentamos a restricciones similares a las que podríamos encontrar en un entorno empresarial real. Abordar un problema concreto y trabajar dentro de ciertos límites nos permitió no solo aplicar técnicas de limpieza de datos, sino también comprender cómo gestionar desafíos y limitaciones en un contexto más práctico.

Con esta experiencia, hemos fortalecido nuestras habilidades en la detección y resolución de problemas específicos en conjuntos de datos, preparándonos para enfrentar desafíos similares en futuros proyectos. Este enfoque práctico ha enriquecido nuestro conocimiento y capacidad para abordar problemas del mundo real de manera efectiva.