Material de apoyo – Planteamiento 10

Sesión #10 Limpieza de datos Python

- **Cargar Datos**
  df = pd.read_csv('datos.csv')

- **Visualizar los Datos**
  print(df.head())

- **Valores atípicos (outliers)**
  outliers = df[(df[columna] < limite_inferior) | (df[columna] > limite_superior)]
  print(f"\nValores atípicos en '{columna}':\n", outliers)
  df_sin_outliers = df[(df[columna] >= limite_inferior) & (df[columna] <= limite_superior)]
  return df_sin_outliers

- **Inconsistencias**: D
  df['Genero'].replace({'M': 'Masculino', 'H': 'Masculino', 'F': 'Femenino'}, inplace=True)

- **Detección de Errores y Anomalías**: Identificar valores atípicos, duplicados, datos faltantes y otros tipos de errores que puedan afectar la calidad de los datos (Kimball & Caserta, 2011).

- **Tipos de datos incorrectos**: A veces los datos no tienen el tipo correcto (números almacenados como texto). Es necesario convertirlos al tipo correcto.

- **Eliminar Valores Faltantes**
  df = df.dropna() # Elimina filas con valores NaN
  df.fillna(df.mean(), inplace=True)

- **Filtrar Datos**
  df_filtrado = df[df['edad'] > 18]

- **Corrección de Errores**:
  df['edad'] = pd.to_numeric(df['edad'], errors='coerce')

  df['nombre'] = df['nombre'].str.lower()

- **Eliminar Columnas Innecesarias**:
  df = df.drop(columns=['columna_innecesaria'])

- **Duplicados**:
  df = df.drop_duplicates()

- **Renombrar Columnas**
  df.rename(columns={'viejo_nombre': 'nuevo_nombre'}, inplace=True)

- **Exportar Datos Limpios**
  df.to_csv('datos_limpios.csv', index=False)

- **Ejemplo**

```
import pandas as pd
import numpy as np

data = {
   'Nombre': ['Juan', 'Maria', 'Pedro', 'Maria', None, 'Ana', 'Carlos', 'Juan'],
   'Edad': [25, 30, np.nan, 30, 22, 28, np.nan, 25],
   'Genero': ['M', 'F', 'M', 'F', None, 'F', 'M', 'M'],
   'Email': ['juan@mail.com', 'maria@mail.com', 'pedro@mail.com', 'maria@mail.com',
'ana@mail.com', 'ana@mail.com', None, 'juan@mail.com'],
   'Pais': ['Colombia', 'Colombia', 'Mexico', 'Colombia', 'Argentina', 'Argentina', 'Brasil',
'Colombia']
}

df = pd.DataFrame(data)

print("Data original:")
print(df)

df['Edad'].fillna(df['Edad'].mean(), inplace=True)
df['Genero'].fillna(df['Genero'].mode()[0], inplace=True)
df.dropna(subset=['Email'], inplace=True)
print("\nDespués de manejar valores faltantes:")
print(df)
df.drop_duplicates(inplace=True)
print("\nDespués de eliminar duplicados:")
print(df)
df['Genero'].replace({'M': 'Masculino', 'F': 'Femenino'}, inplace=True)
print("\nDespués de estandarizar los valores:")
print(df)
df.reset_index(drop=True, inplace=True)
print("\nDespués de restablecer el índice:")
print(df)
df.to_csv('DatosLimpiosPrincipiantes.csv', index=False)
```

- **Código de Ejemplo:**

```
import pandas as pd
import numpy as np

df = pd.read_csv('DatosPersonales.csv')

# Reemplazar valores vacíos o incorrectos ('None', 'nan', '-') por NaN
df.replace([None, 'nan', '-'], np.nan, inplace=True)

# Reemplazar 'M', 'H' por 'Masculino' y 'F', 'Mujer' por 'Femenino' para estandarizar
df['Genero'].replace({'M': 'Masculino', 'H': 'Masculino', 'F': 'Femenino', 'Mujer':
'Femenino'}, inplace=True)

df['Edad'].fillna(df['Edad'].median(), inplace=True)
df['Ingresos'].fillna(df['Ingresos'].median(), inplace=True)

df['Pais'].fillna(df['Pais'].mode()[0], inplace=True)

df['Visitas'] = df['Visitas'].astype(float)
df['Visitas'].fillna(int(df['Visitas'].mean()), inplace=True)
df.dropna(subset=['Email'], inplace=True)
df.drop_duplicates(inplace=True)
df.reset_index(drop=True, inplace=True)
df.to_csv('DatosPersonalesLimpios.csv', index=False)
df
```

- **Otro ejemplo:**

```
import pandas as pd
import numpy as np

df = pd.read_csv('Datosucios.csv')  # This line loads the CSV file
print("Data original:")
print(df)
df.replace([None, 'nan'], np.nan, inplace=True)
print("Data reemplazo:")
print(df)

df['Edad'].fillna(df['Edad'].median(), inplace=True)
df['Ingresos'].fillna(df['Ingresos'].median(), inplace=True)
df['Visitas'].fillna(int(df['Visitas'].mean()), inplace=True)
print("Data llena sin valores faltantes:")
print(df)
df.dropna(subset=['Email'], inplace=True)
print("Data elimnada del Email:")
print(df)
```

```
df.reset_index(drop=True, inplace=True)
cleaned_data_path = 'DatosLimpios.csv'
df.to_csv(cleaned_data_path, index=False)
print("Data guradada:")
print(df)

cleaned_data_path
print("Data para descargar:")
print(df)
```

- **Tareas repetitivas**

```
import pandas as pd
import numpy as np

data = {
   'Nombre': ['Juan', 'Maria', 'Pedro', None, 'Ana', 'Carlos', 'Juan'],
   'Edad': [25, 30, np.nan, 22, 28, np.nan, 25],
   'Genero': ['M', 'F', 'M', None, 'F', 'M', 'M'],
   'Email': ['juan@mail.com', 'maria@mail.com', 'pedro@mail.com', 'ana@mail.com',
'ana@mail.com', None, 'juan@mail.com'],
   'Pais': ['Colombia', None, 'Mexico', 'Argentina', 'Argentina', 'Brasil', 'Colombia']
}

df = pd.DataFrame(data)

print("Valores faltantes:\n", df.isnull().sum())

df['Edad'].fillna(df['Edad'].median(), inplace=True)  # Usar la mediana para 'Edad'
df['Genero'].fillna('Desconocido', inplace=True)      # Usar un valor específico para
'Genero'
df['Pais'].fillna(df['Pais'].mode()[0], inplace=True)  # Usar la moda para 'Pais'

df.drop_duplicates(inplace=True)

df['Genero'].replace({'M': 'Masculino', 'F': 'Femenino'}, inplace=True)
df.dropna(subset=['Email'], inplace=True)
df.reset_index(drop=True, inplace=True)
print("\nData limpia:")
print(df).
```