**Data Cleaning & Processing Summary**

**1. Initial Cleaning**

**-** Removed duplicates using: df.drop\_duplicates(inplace=True)

- Checked null values using: df.isnull().sum()

- Stripped whitespace from column names: df.columns = df.columns.str.strip()

- Converted column headers to string: df.columns = df.columns.astype(str)

**2. Date Formatting**

**-** Converted ScheduledDay & AppointmentDay from ISO format to 'dd-mm-yyyy':

df['ScheduledDay'] = pd.to\_datetime(df['ScheduledDay']).dt.strftime('%d-%m-%Y')

df['AppointmentDay'] = pd.to\_ datetime(df['AppointmentDay']).dt.strftime('%d-%m-%Y')

**3. Feature Engineering**

- Created 'WaitingDays' column:

df['WaitingDays'] = (pd.to\_datetime(df['AppointmentDay']) - pd.to\_datetime(df['ScheduledDay'])).dt.days

- Created 'AgeGroup' Using a custom function with define and if and elif statements:

def categorize\_age(age):

if age < 0:

return 'Invalid'

elif age <= 12:

return 'Child'

elif age <= 18:

return 'Teen'

elif age <= 35:

return 'Young Adult'

elif age <= 60:

return 'Adult'

else:

return 'Senior’

**4. Column Reordering**

- Reordered columns for better readability:

new\_order = [...desired order...]

df = df[new\_order]

**5. Exporting Cleaned Data**

- Saved the final DataFrame to Excel:

file\_path = r'C:/.../Task\_1\_medical\_appointments\_data.xlsx'

df.to\_excel(file\_path, index=False)

* **Summary:**
* Cleaned raw data
* Transformed and engineered new features
* Aggregated key columns
* Reorganized structure
* Exported the final cleaned data