In [1]:
```
import numpy as np
import pandas as pd
```

In [8]:
```
df = pd.read_csv(r'C:\Users\a2z\Downloads\Medical Appointment.csv.zip')
```

In [9]:
```
df
```

Out[9]:

| | PatientId | AppointmentID | Gender | ScheduledDay | AppointmentDay | Age | Neighbourhood | Scholarship | Hipertension | Diabetes | Alcoholism | Handcap | SMS_received | No-show |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2.987250e+13 | 5642903 | F | 2016-04-29T18:38:08Z | 2016-04-29T00:00:00Z | 62 | JARDIM DA PENHA | 0 | 1 | 0 | 0 | 0 | 0 | No |
| **1** | 5.589978e+14 | 5642503 | M | 2016-04-29T16:08:27Z | 2016-04-29T00:00:00Z | 56 | JARDIM DA PENHA | 0 | 0 | 0 | 0 | 0 | 0 | No |
| **2** | 4.262962e+12 | 5642549 | F | 2016-04-29T16:19:04Z | 2016-04-29T00:00:00Z | 62 | MATA DA PRAIA | 0 | 0 | 0 | 0 | 0 | 0 | No |
| **3** | 8.679512e+11 | 5642828 | F | 2016-04-29T17:29:31Z | 2016-04-29T00:00:00Z | 8 | PONTAL DE CAMBURI | 0 | 0 | 0 | 0 | 0 | 0 | No |
| **4** | 8.841186e+12 | 5642494 | F | 2016-04-29T16:07:23Z | 2016-04-29T00:00:00Z | 56 | JARDIM DA PENHA | 0 | 1 | 1 | 0 | 0 | 0 | No |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **110522** | 2.572134e+12 | 5651768 | F | 2016-05-03T09:15:35Z | 2016-06-07T00:00:00Z | 56 | MARIA ORTIZ | 0 | 0 | 0 | 0 | 0 | 1 | No |
| **110523** | 3.596266e+12 | 5650093 | F | 2016-05-03T07:27:33Z | 2016-06-07T00:00:00Z | 51 | MARIA ORTIZ | 0 | 0 | 0 | 0 | 0 | 1 | No |
| **110524** | 1.557663e+13 | 5630692 | F | 2016-04-27T16:03:52Z | 2016-06-07T00:00:00Z | 21 | MARIA ORTIZ | 0 | 0 | 0 | 0 | 0 | 1 | No |
| **110525** | 9.213493e+13 | 5630323 | F | 2016-04-27T15:09:23Z | 2016-06-07T00:00:00Z | 38 | MARIA ORTIZ | 0 | 0 | 0 | 0 | 0 | 1 | No |
| **110526** | 3.775115e+14 | 5629448 | F | 2016-04-27T13:30:56Z | 2016-06-07T00:00:00Z | 54 | MARIA ORTIZ | 0 | 0 | 0 | 0 | 0 | 1 | No |

110527 rows × 14 columns

In [11]:
```
#Identify and handle missing values using .isnull() in Python or filters in Excel.
df.isnull()
```

Out[11]:
```
PatientId         110527
AppointmentID     110527
Gender            110527
ScheduledDay      110527
AppointmentDay    110527
Age               110527
Neighbourhood     110527
Scholarship       110527
Hipertension      110527
Diabetes          110527
Alcoholism        110527
Handcap           110527
SMS_received      110527
No-show           110527
dtype: int64
```

In [12]:
```
df.isnull().sum()
```

Out[12]:
```
PatientId         0
AppointmentID     0
Gender            0
ScheduledDay      0
AppointmentDay    0
Age               0
Neighbourhood     0
Scholarship       0
Hipertension      0
Diabetes          0
Alcoholism        0
Handcap           0
SMS_received      0
No-show           0
dtype: int64
```

In [13]:
```
df.isnull().count()
```

Out[13]:
```
PatientId         110527
AppointmentID     110527
Gender            110527
ScheduledDay      110527
AppointmentDay    110527
Age               110527
Neighbourhood     110527
Scholarship       110527
Hipertension      110527
Diabetes          110527
Alcoholism        110527
Handcap           110527
SMS_received      110527
No-show           110527
dtype: int64
```

In [16]:
```
#Remove duplicate rows using .drop_duplicates() or Excel's "Remove Duplicates".

df_cleaned = df.drop_duplicates()
```

In [17]:
```
df_cleaned
```

Out[17]:

| | PatientId | AppointmentID | Gender | ScheduledDay | AppointmentDay | Age | Neighbourhood | Scholarship | Hipertension | Diabetes | Alcoholism | Handcap | SMS_received | No-show |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2.987250e+13 | 5642903 | F | 2016-04-29T18:38:08Z | 2016-04-29T00:00:00Z | 62 | JARDIM DA PENHA | 0 | 1 | 0 | 0 | 0 | 0 | No |
| **1** | 5.589978e+14 | 5642503 | M | 2016-04-29T16:08:27Z | 2016-04-29T00:00:00Z | 56 | JARDIM DA PENHA | 0 | 0 | 0 | 0 | 0 | 0 | No |
| **2** | 4.262962e+12 | 5642549 | F | 2016-04-29T16:19:04Z | 2016-04-29T00:00:00Z | 62 | MATA DA PRAIA | 0 | 0 | 0 | 0 | 0 | 0 | No |
| **3** | 8.679512e+11 | 5642828 | F | 2016-04-29T17:29:31Z | 2016-04-29T00:00:00Z | 8 | PONTAL DE CAMBURI | 0 | 0 | 0 | 0 | 0 | 0 | No |
| **4** | 8.841186e+12 | 5642494 | F | 2016-04-29T16:07:23Z | 2016-04-29T00:00:00Z | 56 | JARDIM DA PENHA | 0 | 1 | 1 | 0 | 0 | 0 | No |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **110522** | 2.572134e+12 | 5651768 | F | 2016-05-03T09:15:35Z | 2016-06-07T00:00:00Z | 56 | MARIA ORTIZ | 0 | 0 | 0 | 0 | 0 | 1 | No |
| **110523** | 3.596266e+12 | 5650093 | F | 2016-05-03T07:27:33Z | 2016-06-07T00:00:00Z | 51 | MARIA ORTIZ | 0 | 0 | 0 | 0 | 0 | 1 | No |
| **110524** | 1.557663e+13 | 5630692 | F | 2016-04-27T16:03:52Z | 2016-06-07T00:00:00Z | 21 | MARIA ORTIZ | 0 | 0 | 0 | 0 | 0 | 1 | No |
| **110525** | 9.213493e+13 | 5630323 | F | 2016-04-27T15:09:23Z | 2016-06-07T00:00:00Z | 38 | MARIA ORTIZ | 0 | 0 | 0 | 0 | 0 | 1 | No |
| **110526** | 3.775115e+14 | 5629448 | F | 2016-04-27T13:30:56Z | 2016-06-07T00:00:00Z | 54 | MARIA ORTIZ | 0 | 0 | 0 | 0 | 0 | 1 | No |

110527 rows × 14 columns

In [19]:
```
#Standardize text values like gender,Neighbourhood , etc

df['Gender'] = df['Gender'].str.strip().str.upper().map({'M': 'Male', 'F': 'Female'})

# Optional: Standardize Neighbourhood
df['Neighbourhood'] = df['Neighbourhood'].str.strip().str.title()


# Display first few rows
print(df[['Gender', 'Neighbourhood']].head())
```

```
   Gender      Neighbourhood
0  Female     Jardim Da Penha
1    Male     Jardim Da Penha
2  Female       Mata Da Praia
3  Female  Pontal De Camburi
4  Female     Jardim Da Penha
```

In [20]:
```
df.head()
```

Out[20]:

| | PatientId | AppointmentID | Gender | ScheduledDay | AppointmentDay | Age | Neighbourhood | Scholarship | Hipertension | Diabetes | Alcoholism | Handcap | SMS_received | No-show |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2.987250e+13 | 5642903 | Female | 2016-04-29T18:38:08Z | 2016-04-29T00:00:00Z | 62 | Jardim Da Penha | 0 | 1 | 0 | 0 | 0 | 0 | No |
| **1** | 5.589978e+14 | 5642503 | Male | 2016-04-29T16:08:27Z | 2016-04-29T00:00:00Z | 56 | Jardim Da Penha | 0 | 0 | 0 | 0 | 0 | 0 | No |
| **2** | 4.262962e+12 | 5642549 | Female | 2016-04-29T16:19:04Z | 2016-04-29T00:00:00Z | 62 | Mata Da Praia | 0 | 0 | 0 | 0 | 0 | 0 | No |
| **3** | 8.679512e+11 | 5642828 | Female | 2016-04-29T17:29:31Z | 2016-04-29T00:00:00Z | 8 | Pontal De Camburi | 0 | 0 | 0 | 0 | 0 | 0 | No |
| **4** | 8.841186e+12 | 5642494 | Female | 2016-04-29T16:07:23Z | 2016-04-29T00:00:00Z | 56 | Jardim Da Penha | 0 | 1 | 1 | 0 | 0 | 0 | No |

In [38]:
```
# Clean column names and check them
df.columns = df.columns.str.strip()
print(df.columns)  # Show column names

# Try to convert and split dates if the columns exist
try:
    df['ScheduledDay'] = pd.to_datetime(df['ScheduledDay'])
    df['AppointmentDay'] = pd.to_datetime(df['AppointmentDay'])

    df['ScheduledDate'] = df['ScheduledDay'].dt.date
    df['ScheduledTime'] = df['ScheduledDay'].dt.time
    df['AppointmentDate'] = df['AppointmentDay'].dt.date
    df['AppointmentTime'] = df['AppointmentDay'].dt.time

    print(df[['ScheduledDate', 'ScheduledTime', 'AppointmentDate', 'AppointmentTime']])
except KeyError as e:
    print(f"Column not found: {e}")
```

```
Index(['patientid', 'appointmentid', 'gender', 'scheduleday',
       'appointmentday', 'age', 'neighbourhood', 'scholarship', 'hipertension',
       'diabetes', 'alcoholism', 'handcap', 'sms_received', 'no-show'],
      dtype='object')
Column not found: 'ScheduledDay'
```

In [39]:
```
df.head()
```

Out[39]:

| | patientid | appointmentid | gender | scheduleday | appointmentday | age | neighbourhood | scholarship | hipertension | diabetes | alcoholism | handcap | sms_received | no-show |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | -2.147483648 | 5642903 | Female | 2016-04-29 | 2016-04-29 | 62 | Jardim Da Penha | 0 | 1 | 0 | 0 | 0 | 0 | No |
| **1** | -2.147483648 | 5642503 | Male | 2016-04-29 | 2016-04-29 | 56 | Jardim Da Penha | 0 | 0 | 0 | 0 | 0 | 0 | No |
| **2** | -2.147483648 | 5642549 | Female | 2016-04-29 | 2016-04-29 | 62 | Mata Da Praia | 0 | 0 | 0 | 0 | 0 | 0 | No |
| **3** | -2.147483648 | 5642828 | Female | 2016-04-29 | 2016-04-29 | 8 | Pontal De Camburi | 0 | 0 | 0 | 0 | 0 | 0 | No |
| **4** | -2.147483648 | 5642494 | Female | 2016-04-29 | 2016-04-29 | 56 | Jardim Da Penha | 0 | 1 | 1 | 0 | 0 | 0 | No |

In [23]:
```
df.columns = df.columns.str.lower().str.replace(' ', '_')
```

In [29]:
```
df.dtypes
```

Out[29]:
```
patientid              float64
appointmentid            int64
gender                  object
scheduleday     datetime64[ns]
appointmentday  datetime64[ns]
age                      int32
neighbourhood           object
scholarship              int64
hipertension             int64
diabetes                 int64
alcoholism               int64
handcap                  int64
sms_received             int64
no-show                 object
dtype: object
```

In [26]:
```
df['age'] = df['age'].fillna(df['age'].median()).astype('int')
```

In [28]:
```
df['scheduleday'] = pd.to_datetime(df['scheduleday'], errors='coerce')
df['appointmentday'] = pd.to_datetime(df['appointmentday'], errors='coerce')
```

```
C:\Users\a2z\AppData\Local\Temp\ipykernel_4488\248930977O.py:1: UserWarning: Parsing dates in %d-%m-%Y format when dayfirst=False (the default) was specified. Pass `dayfirst=True` or specify a format to silence this warning.
  df['scheduleday'] = pd.to_datetime(df['scheduleday'], errors='coerce')
C:\Users\a2z\AppData\Local\Temp\ipykernel_4488\248930977O.py:2: UserWarning: Parsing dates in %d-%m-%Y format when dayfirst=False (the default) was specified. Pass `dayfirst=True` or specify a format to silence this warning.
  df['appointmentday'] = pd.to_datetime(df['appointmentday'], errors='coerce')
```

In [30]:
```
# Convert 'patientid' column from float to int and remove any '+' signs (if they exist)
df['patientid'] = df['patientid'].astype(int)

# If there are '+' signs in patientid (assuming it is in string format)
# You can remove the '+' signs like this:
df['patientid'] = df['patientid'].astype(str).str.replace('+', '')

# Convert back to int after removing '+' sign
df['patientid'] = df['patientid'].astype(int)

# Check the result
print(df['patientid'].head())
```

```
0   -2147483648
1   -2147483648
2   -2147483648
3   -2147483648
4   -2147483648
Name: patientid, dtype: int32
```

In [31]:
```

```

Out[31]:

| | patientid | appointmentid | gender | scheduleday | appointmentday | age | neighbourhood | scholarship | hipertension | diabetes | alcoholism | handcap | sms_received | no-show |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | -2147483648 | 5642903 | Female | 2016-04-29 | 2016-04-29 | 62 | Jardim Da Penha | 0 | 1 | 0 | 0 | 0 | 0 | No |
| **1** | -2147483648 | 5642503 | Male | 2016-04-29 | 2016-04-29 | 56 | Jardim Da Penha | 0 | 0 | 0 | 0 | 0 | 0 | No |
| **2** | -2147483648 | 5642549 | Female | 2016-04-29 | 2016-04-29 | 62 | Mata Da Praia | 0 | 0 | 0 | 0 | 0 | 0 | No |
| **3** | -2147483648 | 5642828 | Female | 2016-04-29 | 2016-04-29 | 8 | Pontal De Camburi | 0 | 0 | 0 | 0 | 0 | 0 | No |
| **4** | -2147483648 | 5642494 | Female | 2016-04-29 | 2016-04-29 | 56 | Jardim Da Penha | 0 | 1 | 1 | 0 | 0 | 0 | No |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **110522** | -2147483648 | 5651768 | Female | 2016-05-03 | 2016-06-07 | 56 | Maria Ortiz | 0 | 0 | 0 | 0 | 0 | 1 | No |
| **110523** | -2147483648 | 5650093 | Female | 2016-05-03 | 2016-06-07 | 51 | Maria Ortiz | 0 | 0 | 0 | 0 | 0 | 1 | No |
| **110524** | -2147483648 | 5630692 | Female | 2016-04-27 | 2016-06-07 | 21 | Maria Ortiz | 0 | 0 | 0 | 0 | 0 | 1 | No |
| **110525** | -2147483648 | 5630323 | Female | 2016-04-27 | 2016-06-07 | 38 | Maria Ortiz | 0 | 0 | 0 | 0 | 0 | 1 | No |
| **110526** | -2147483648 | 5629448 | Female | 2016-04-27 | 2016-06-07 | 54 | Maria Ortiz | 0 | 0 | 0 | 0 | 0 | 1 | No |

110527 rows × 14 columns

In [ ]:
```

```

In [ ]:
```

```

In [ ]:
```

```