8fxava9cp

April 21, 2025

[2]:

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

[3]:

```
import numpy as np
import pandas as pd
```

[4]:

```
df=pd.read_csv('/content/drive/MyDrive/DATASETS/Appointments_Dset.csv')
```

[5]:

```
df.head(10)
```

[5]:

```
      PatientId  AppointmentID Gender         ScheduledDay  \
0  2.987250e+13        5642903      F  2016-04-29T18:38:08Z
1  5.589978e+14        5642503      M  2016-04-29T16:08:27Z
2  4.262962e+12        5642549      F  2016-04-29T16:19:04Z
3  8.679512e+11        5642828      F  2016-04-29T17:29:31Z
4  8.841186e+12        5642494      F  2016-04-29T16:07:23Z
5  9.598513e+13        5626772      F  2016-04-27T08:36:51Z
6  7.336882e+14        5630279      F  2016-04-27T15:05:12Z
7  3.449833e+12        5630575      F  2016-04-27T15:39:58Z
8  5.639473e+13        5638447      F  2016-04-29T08:02:16Z
9  7.812456e+13        5629123      F  2016-04-27T12:48:25Z

         AppointmentDay  Age      Neighbourhood  Scholarship  Hipertension  \
0  2016-04-29T00:00:00Z   62    JARDIM DA PENHA            0             1
1  2016-04-29T00:00:00Z   56    JARDIM DA PENHA            0             0
2  2016-04-29T00:00:00Z   62      MATA DA PRAIA            0             0
3  2016-04-29T00:00:00Z    8  PONTAL DE CAMBURI            0             0
4  2016-04-29T00:00:00Z   56    JARDIM DA PENHA            0             1
5  2016-04-29T00:00:00Z   76          REPÚBLICA            0             1
6  2016-04-29T00:00:00Z   23         GOIABEIRAS            0             0
7  2016-04-29T00:00:00Z   39         GOIABEIRAS            0             0
8  2016-04-29T00:00:00Z   21         ANDORINHAS            0             0
9  2016-04-29T00:00:00Z   19          CONQUISTA            0             0

   Diabetes  Alcoholism  Handcap  SMS_received No-show
0         0           0        0             0      No
```

```
1         0           0        0             0      No
2         0           0        0             0      No
3         0           0        0             0      No
4         1           0        0             0      No
5         0           0        0             0      No
6         0           0        0             0     Yes
7         0           0        0             0     Yes
8         0           0        0             0      No
9         0           0        0             0      No
```

[7]:

```
df.columns = [col.strip().lower().replace('-', '_') for col in df.columns]

print("Cleaned Column Names:\n", df.columns.tolist())
```

```
Cleaned Column Names:
 ['patientid', 'appointmentid', 'gender', 'scheduledday', 'appointmentday',
'age', 'neighbourhood', 'scholarship', 'hipertension', 'diabetes', 'alcoholism',
'handcap', 'sms_received', 'no_show']
```

[8]:

```
df.shape
```

[8]:

```
(110527, 14)
```

[9]:

```
df = df.drop_duplicates()

print("Shape after removing duplicates:", df.shape)
```

```
Shape after removing duplicates: (110527, 14)
```

#There are no Duplicates

[10]:

```
df.isnull().sum()
```

[10]:

```
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
```

No Missing Values

[11]:
```
df.dtypes
```

[11]:
```
patientid         float64
appointmentid       int64
gender             object
scheduledday       object
appointmentday     object
age                 int64
neighbourhood      object
scholarship         int64
hipertension        int64
diabetes            int64
alcoholism          int64
handcap             int64
sms_received        int64
no_show            object
dtype: object
```

[12]:
```
df['scheduledday'] = pd.to_datetime(df['scheduledday'])
df['appointmentday'] = pd.to_datetime(df['appointmentday'])
```

[14]:
```
print("ScheduledDay Type:", df['scheduledday'].dtype)
print("AppointmentDay Type:", df['appointmentday'].dtype)
```

```
ScheduledDay Type: datetime64[ns, UTC]
AppointmentDay Type: datetime64[ns, UTC]
```

[15]:
```
df.describe()
```

[15]:
```
          patientid  appointmentid            age    scholarship  \
count  1.105270e+05   1.105270e+05  110527.000000  110527.000000
mean   1.474963e+14   5.675305e+06      37.088874       0.098266
std    2.560949e+14   7.129575e+04      23.110205       0.297675
min    3.921784e+04   5.030230e+06      -1.000000       0.000000
25%    4.172614e+12   5.640286e+06      18.000000       0.000000
50%    3.173184e+13   5.680573e+06      37.000000       0.000000
75%    9.439172e+13   5.725524e+06      55.000000       0.000000
max    9.999816e+14   5.790484e+06     115.000000       1.000000

        hipertension       diabetes     alcoholism        handcap  \
count  110527.000000  110527.000000  110527.000000  110527.000000
mean        0.197246       0.071865       0.030400       0.022248
std         0.397921       0.258265       0.171686       0.161543
min         0.000000       0.000000       0.000000       0.000000
25%         0.000000       0.000000       0.000000       0.000000
50%         0.000000       0.000000       0.000000       0.000000
```

```
75%          0.000000       0.000000       0.000000       0.000000
max          1.000000       1.000000       1.000000       4.000000

        sms_received
count  110527.000000
mean        0.321026
std         0.466873
min         0.000000
25%         0.000000
50%         0.000000
75%         1.000000
max         1.000000
```

Above we found that age is min -1 so it is invalid we have to change that

[16]:
```
df = df[df['age'] >= 0]
```

[18]:
```
df['age'].min()  #Checking whether it is updated or not
```

[18]:
```
0
```

1 Clean Categorical Columns

[20]:
```
df['gender'].unique()
```

[20]:
```
array(['F', 'M'], dtype=object)
```

[23]:
```
df['no_show'].unique()
```

[23]:
```
array(['No', 'Yes'], dtype=object)
```

[25]:
```
df['handcap'].value_counts()
```

[25]:
```
handcap
0    108285
1      2042
2       183
3        13
4         3
Name: count, dtype: int64
```

2 Value Meaning (likely) Count

0 No disability 108,285

1 Some disability

But values 2, 3, 4 are super rare (only 199 total — less than 0.2%). This might be data entry error or simply unnecessary granularity.

[28]:

```
df['handcap'] = df['handcap'].apply(lambda x: 1 if x > 0 else 0)
```

[29]:

```
df['handcap'].value_counts()
```

[29]:

```
handcap
0    108285
1      2241
Name: count, dtype: int64
```

#Let's check for duplicate rows in the dataset.

[30]:

```
df.duplicated().sum()
```

[30]:

```
np.int64(0)
```

Here's a short summary of the changes made to the dataset:

2.0.1 Summary of Changes:

1. **Handled missing values**: No missing values were found, so no further action was needed.
2. **Removed duplicate rows**: The dataset had no duplicates, so nothing was removed.
3. **Standardized text values**:
   - The `Gender` column values were consistent (`F` for female, `M` for male).
   - The `Handcap` column values were confirmed as numerical and standardized.
4. **Converted date formats**:
   - The `ScheduledDay` and `AppointmentDay` columns were converted to `datetime` type for consistency.
5. **Renamed columns**:
   - All column names were cleaned to be lowercase with no spaces (e.g., `No-show` became `no_show`).
6. **Checked and fixed data types**:
   - Ensured that `Age` is an integer, and both `ScheduledDay` and `AppointmentDay` are `datetime`.

The dataset is now cleaned and ready for further analysis.

[ ]:

```

```