In [178…
```
import pandas as pd
```

In [179…
```
df = pd.read_csv('Medical_Appointment_No_Shows.csv')
```

In [180…
```
df
```

Out[180…

	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 [181…
```
df.isnull().sum()
```

Out[181…
```
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 [182…
```
#Since there are no null values, we need not drop any null value
```

In [183…
```
df.duplicated().sum()
```

Out[183… 0

In [184…
```
#There are no duplicate values in the dataset
```

In [185…
```
df.head()
```

Out[185…

	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

In [186…
```
df.rename(columns= {'Handcap':'Handicap'},inplace = True)
```

In [187…
```
df.rename(columns= {'No-show':'No_Show'},inplace = True)
```

In [188…
```
df.keys()
```

Out[188…
```
Index(['PatientId', 'AppointmentID', 'Gender', 'ScheduledDay',
       'AppointmentDay', 'Age', 'Neighbourhood', 'Scholarship', 'Hipertension',
       'Diabetes', 'Alcoholism', 'Handicap', 'SMS_received', 'No_Show'],
      dtype='object')
```

In [189…
```
df['Gender'] = df['Gender'].replace('F', 'female')
df['Gender'] = df['Gender'].replace('M', 'male')
```

In [190…
```
df.head()
```

Out[190…

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension	Diabetes	Alcoholism	Handicap	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 [191…
```
df['Neighbourhood']=df['Neighbourhood'].str.lower()
```

In [192…
```
df['Neighbourhood']
```

Out[192…
```
0           jardim da penha
1           jardim da penha
2             mata da praia
3         pontal de camburi
4           jardim da penha
                ...
110522          maria ortiz
110523          maria ortiz
110524          maria ortiz
110525          maria ortiz
110526          maria ortiz
Name: Neighbourhood, Length: 110527, dtype: object
```

In [193…
```
df['ScheduledDay'] = pd.to_datetime(df['ScheduledDay'])
```

In [194…
```
df['ScheduledDay']
```

Out[194…
```
0       2016-04-29 18:38:08+00:00
1       2016-04-29 16:08:27+00:00
2       2016-04-29 16:19:04+00:00
3       2016-04-29 17:29:31+00:00
4       2016-04-29 16:07:23+00:00
                   ...
110522  2016-05-03 09:15:35+00:00
110523  2016-05-03 07:27:33+00:00
110524  2016-04-27 16:03:52+00:00
110525  2016-04-27 15:09:23+00:00
110526  2016-04-27 13:30:56+00:00
Name: ScheduledDay, Length: 110527, dtype: datetime64[ns, UTC]
```

In [195…
```
df['AppointmentDay'] = pd.to_datetime(df['AppointmentDay'])
```

In [196…
```
df['AppointmentDay']
```

Out[196…
```
0       2016-04-29 00:00:00+00:00
1       2016-04-29 00:00:00+00:00
2       2016-04-29 00:00:00+00:00
3       2016-04-29 00:00:00+00:00
4       2016-04-29 00:00:00+00:00
                   ...
110522  2016-06-07 00:00:00+00:00
110523  2016-06-07 00:00:00+00:00
110524  2016-06-07 00:00:00+00:00
110525  2016-06-07 00:00:00+00:00
110526  2016-06-07 00:00:00+00:00
Name: AppointmentDay, Length: 110527, dtype: datetime64[ns, UTC]
```

In [197…
```
df['Age']
```

Out[197…
```
0         62
1         56
2         62
3          8
4         56
          ..
110522    56
110523    51
110524    21
110525    38
110526    54
Name: Age, Length: 110527, dtype: int64
```

In [198…
```
df.dtypes
```

Out[198…
```
PatientId                     float64
AppointmentID                   int64
Gender                         object
ScheduledDay      datetime64[ns, UTC]
AppointmentDay    datetime64[ns, UTC]
Age                             int64
Neighbourhood                  object
Scholarship                     int64
Hipertension                    int64
Diabetes                        int64
Alcoholism                      int64
Handicap                        int64
SMS_received                    int64
No_Show                        object
dtype: object
```

In [199…
```
df.columns = df.columns.str.lower()
```

In [200…
```
df.columns
```

Out[200…
```
Index(['patientid', 'appointmentid', 'gender', 'scheduledday',
       'appointmentday', 'age', 'neighbourhood', 'scholarship', 'hipertension',
       'diabetes', 'alcoholism', 'handicap', 'sms_received', 'no_show'],
      dtype='object')
```

In [201…
```
df['age'].value_counts()
```

Out[201…
```
0      3539
1      2273
52     1746
49     1652
53     1651
       ...
115       5
100       4
102       2
99        1
-1        1
Name: age, Length: 104, dtype: int64
```

In [202…
```
df['age'] = df['age'].replace(0, np.nan)
```

In [203…
```
df['age'].isna().sum()
```

Out[203… 3539

In [204…
```
df['age'].dropna(inplace=True)
```

In [205…
```
df['age'].value_counts()
```

Out[205…
```
 1.0      2273
 52.0     1746
 49.0     1652
 53.0     1651
 56.0     1635
          ...
 115.0       5
 100.0       4
 102.0       2
 99.0        1
-1.0         1
Name: age, Length: 103, dtype: int64
```

In [212…
```
idx=df[df['age']==-1].index
```

In [215…
```
print(idx)
```

```
Int64Index([99832], dtype='int64')
```

In [218…
```
df = df.drop(99832)
```

In [219…
```
df
```

Out[219…

	patientid	appointmentid	gender	scheduledday	appointmentday	age	neighbourhood	scholarship	hipertension	diabetes	alcoholism	handicap	sms_received	no_show
0	2.987250e+13	5642903	female	2016-04-29 18:38:08+00:00	2016-04-29 00:00:00+00:00	62.0	jardim da penha	0	1	0	0	0	0	No
1	5.589978e+14	5642503	male	2016-04-29 16:08:27+00:00	2016-04-29 00:00:00+00:00	56.0	jardim da penha	0	0	0	0	0	0	No
2	4.262962e+12	5642549	female	2016-04-29 16:19:04+00:00	2016-04-29 00:00:00+00:00	62.0	mata da praia	0	0	0	0	0	0	No
3	8.679512e+11	5642828	female	2016-04-29 17:29:31+00:00	2016-04-29 00:00:00+00:00	8.0	pontal de camburi	0	0	0	0	0	0	No
4	8.841186e+12	5642494	female	2016-04-29 16:07:23+00:00	2016-04-29 00:00:00+00:00	56.0	jardim da penha	0	1	1	0	0	0	No
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
110522	2.572134e+12	5651768	female	2016-05-03 09:15:35+00:00	2016-06-07 00:00:00+00:00	56.0	maria ortiz	0	0	0	0	0	1	No
110523	3.596266e+12	5650093	female	2016-05-03 07:27:33+00:00	2016-06-07 00:00:00+00:00	51.0	maria ortiz	0	0	0	0	0	1	No
110524	1.557663e+13	5630692	female	2016-04-27 16:03:52+00:00	2016-06-07 00:00:00+00:00	21.0	maria ortiz	0	0	0	0	0	1	No
110525	9.213493e+13	5630323	female	2016-04-27 15:09:23+00:00	2016-06-07 00:00:00+00:00	38.0	maria ortiz	0	0	0	0	0	1	No
110526	3.775115e+14	5629448	female	2016-04-27 13:30:56+00:00	2016-06-07 00:00:00+00:00	54.0	maria ortiz	0	0	0	0	0	1	No

110526 rows × 14 columns

In [221…

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 110526 entries, 0 to 110526
Data columns (total 14 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   patientid       110526 non-null  float64
 1   appointmentid   110526 non-null  int64
 2   gender          110526 non-null  object
 3   scheduledday    110526 non-null  datetime64[ns, UTC]
 4   appointmentday  110526 non-null  datetime64[ns, UTC]
 5   age             106987 non-null  float64
 6   neighbourhood   110526 non-null  object
 7   scholarship     110526 non-null  int64
 8   hipertension    110526 non-null  int64
 9   diabetes        110526 non-null  int64
 10  alcoholism      110526 non-null  int64
 11  handicap        110526 non-null  int64
 12  sms_received    110526 non-null  int64
 13  no_show         110526 non-null  object
dtypes: datetime64[ns, UTC](2), float64(2), int64(7), object(3)
memory usage: 12.6+ MB
```

In [222…

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

Out[222…
```
jardim camburi                 7717
maria ortiz                    5805
resistência                    4431
jardim da penha                3877
itararé                        3514
                               ...
ilha do boi                      35
ilha do frade                    10
aeroporto                         8
ilhas oceânicas de trindade       2
parque industrial                 1
Name: neighbourhood, Length: 81, dtype: int64
```

In [ ]: