In [91]:

```
# importing libraries and all the library
```

In [159…

```
import pandas as pd
```

In [160…

```
import numpy as np
```

In [161…

```
dt = pd.read_csv(r"D:\DATA ANALYST INTERNSHIP\all datasets\KaggleV2-May-2016.csv")
```

In [162…

```
dt
```

Out[162…

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Ne
0	2.987250e+13	5642903	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	
1	5.589978e+14	5642503	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	
2	4.262962e+12	5642549	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	M,
3	8.679512e+11	5642828	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	
4	8.841186e+12	5642494	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	
...	...	...	...	...	...	...	
110522	2.572134e+12	5651768	F	2016-05-03T09:15:35Z	2016-06-07T00:00:00Z	56	
110523	3.596266e+12	5650093	F	2016-05-03T07:27:33Z	2016-06-07T00:00:00Z	51	
110524	1.557663e+13	5630692	F	2016-04-27T16:03:52Z	2016-06-07T00:00:00Z	21	
110525	9.213493e+13	5630323	F	2016-04-27T15:09:23Z	2016-06-07T00:00:00Z	38	
110526	3.775115e+14	5629448	F	2016-04-27T13:30:56Z	2016-06-07T00:00:00Z	54	

110527 rows × 14 columns

In [163…

```
dt.head(10)
```

Out[163...

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbou
0	2.987250e+13	5642903	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARI
1	5.589978e+14	5642503	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARI
2	4.262962e+12	5642549	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA
3	8.679512e+11	5642828	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PON CA
4	8.841186e+12	5642494	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARI
5	9.598513e+13	5626772	F	2016-04-27T08:36:51Z	2016-04-29T00:00:00Z	76	REP
6	7.336882e+14	5630279	F	2016-04-27T15:05:12Z	2016-04-29T00:00:00Z	23	GOIA
7	3.449833e+12	5630575	F	2016-04-27T15:39:58Z	2016-04-29T00:00:00Z	39	GOIA
8	5.639473e+13	5638447	F	2016-04-29T08:02:16Z	2016-04-29T00:00:00Z	21	ANDOI
9	7.812456e+13	5629123	F	2016-04-27T12:48:25Z	2016-04-29T00:00:00Z	19	CON

In [164...

```
dt.head(5)
```

Out[164...

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbou
0	2.987250e+13	5642903	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARI
1	5.589978e+14	5642503	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARI
2	4.262962e+12	5642549	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA
3	8.679512e+11	5642828	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PON CA
4	8.841186e+12	5642494	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARI

In [165...

```
# checking the shape of the data set
```

In [166…

```
dt.shape
```

Out[166…

```
(110527, 14)
```

In [167…

```
dt.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 110527 entries, 0 to 110526
Data columns (total 14 columns):
 #   Column          Non-Null Count   Dtype  
---  ------          --------------   -----  
 0   PatientId       110527 non-null  float64
 1   AppointmentID   110527 non-null  int64  
 2   Gender          110527 non-null  object 
 3   ScheduledDay    110527 non-null  object 
 4   AppointmentDay  110527 non-null  object 
 5   Age             110527 non-null  int64  
 6   Neighbourhood   110527 non-null  object 
 7   Scholarship     110527 non-null  int64  
 8   Hipertension    110527 non-null  int64  
 9   Diabetes        110527 non-null  int64  
 10  Alcoholism      110527 non-null  int64  
 11  Handcap         110527 non-null  int64  
 12  SMS_received    110527 non-null  int64  
 13  No-show         110527 non-null  object 
dtypes: float64(1), int64(8), object(5)
memory usage: 11.8+ MB
```

In [168…

```
  # for checkig correspondig data types of columns
```

In [169…

```
dt.dtypes
```

Out[169…

```
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
```

In [170…

```
# checking the null values
```

In [171…

```
dt.isnull().sum()
```

Out[171…

```
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 [172…

```
for i in dt.columns:
    print(i,':',sum((dt[i])=='?'))
```

```
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
```

In [173…

```
for i  in dt.columns:
    print(i,':','\n',dt[i].unique())
```

```
PatientId :
 [2.98724998e+13 5.58997777e+14 4.26296230e+12 ... 7.26331493e+13
 9.96997666e+14 1.55766317e+13]
AppointmentID :
 [5642903 5642503 5642549 ... 5630692 5630323 5629448]
Gender :
 ['F' 'M']
ScheduledDay :
 ['2016-04-29T18:38:08Z' '2016-04-29T16:08:27Z' '2016-04-29T16:19:04Z' ...
 '2016-04-27T16:03:52Z' '2016-04-27T15:09:23Z' '2016-04-27T13:30:56Z']
AppointmentDay :
 ['2016-04-29T00:00:00Z' '2016-05-03T00:00:00Z' '2016-05-10T00:00:00Z'
 '2016-05-17T00:00:00Z' '2016-05-24T00:00:00Z' '2016-05-31T00:00:00Z'
 '2016-05-02T00:00:00Z' '2016-05-30T00:00:00Z' '2016-05-16T00:00:00Z'
 '2016-05-04T00:00:00Z' '2016-05-19T00:00:00Z' '2016-05-12T00:00:00Z'
 '2016-05-06T00:00:00Z' '2016-05-20T00:00:00Z' '2016-05-05T00:00:00Z'
 '2016-05-13T00:00:00Z' '2016-05-09T00:00:00Z' '2016-05-25T00:00:00Z'
 '2016-05-11T00:00:00Z' '2016-05-18T00:00:00Z' '2016-05-14T00:00:00Z'
 '2016-06-02T00:00:00Z' '2016-06-03T00:00:00Z' '2016-06-06T00:00:00Z'
 '2016-06-07T00:00:00Z' '2016-06-01T00:00:00Z' '2016-06-08T00:00:00Z']
Age :
 [ 62  56   8  76  23  39  21  19  30  29  22  28  54  15  50  40  46   4
  13  65  45  51  32  12  61  38  79  18  63  64  85  59  55  71  49  78
  31  58  27   6   2  11   7   0   3   1  69  68  60  67  36  10  35  20
  26  34  33  16  42   5  47  17  41  44  37  24  66  77  81  70  53  75
  73  52  74  43  89  57  14   9  48  83  72  25  80  87  88  84  82  90
  94  86  91  98  92  96  93  95  97 102 115 100  99  -1]
Neighbourhood :
 ['JARDIM DA PENHA' 'MATA DA PRAIA' 'PONTAL DE CAMBURI' 'REPÚBLICA'
 'GOIABEIRAS' 'ANDORINHAS' 'CONQUISTA' 'NOVA PALESTINA' 'DA PENHA'
 'TABUAZEIRO' 'BENTO FERREIRA' 'SÃO PEDRO' 'SANTA MARTHA' 'SÃO CRISTÓVÃO'
 'MARUÍPE' 'GRANDE VITÓRIA' 'SÃO BENEDITO' 'ILHA DAS CAIEIRAS'
 'SANTO ANDRÉ' 'SOLON BORGES' 'BONFIM' 'JARDIM CAMBURI' 'MARIA ORTIZ'
 'JABOUR' 'ANTÔNIO HONÓRIO' 'RESISTÊNCIA' 'ILHA DE SANTA MARIA'
 'JUCUTUQUARA' 'MONTE BELO' 'MÁRIO CYPRESTE' 'SANTO ANTÔNIO' 'BELA VISTA'
 'PRAIA DO SUÁ' 'SANTA HELENA' 'ITARARÉ' 'INHANGUETÁ' 'UNIVERSITÁRIO'
 'SÃO JOSÉ' 'REDENÇÃO' 'SANTA CLARA' 'CENTRO' 'PARQUE MOSCOSO'
 'DO MOSCOSO' 'SANTOS DUMONT' 'CARATOÍRA' 'ARIOVALDO FAVALESSA'
 'ILHA DO FRADE' 'GURIGICA' 'JOANA D´ARC' 'CONSOLAÇÃO' 'PRAIA DO CANTO'
 'BOA VISTA' 'MORADA DE CAMBURI' 'SANTA LUÍZA' 'SANTA LÚCIA'
 'BARRO VERMELHO' 'ESTRELINHA' 'FORTE SÃO JOÃO' 'FONTE GRANDE'
 'ENSEADA DO SUÁ' 'SANTOS REIS' 'PIEDADE' 'JESUS DE NAZARETH'
 'SANTA TEREZA' 'CRUZAMENTO' 'ILHA DO PRÍNCIPE' 'ROMÃO' 'COMDUSA'
 'SANTA CECÍLIA' 'VILA RUBIM' 'DE LOURDES' 'DO QUADRO' 'DO CABRAL' 'HORTO'
 'SEGURANÇA DO LAR' 'ILHA DO BOI' 'FRADINHOS' 'NAZARETH' 'AEROPORTO'
 'ILHAS OCEÂNICAS DE TRINDADE' 'PARQUE INDUSTRIAL']
Scholarship :
 [0 1]
Hipertension :
 [1 0]
Diabetes :
 [0 1]
Alcoholism :
 [0 1]
Handcap :
 [0 1 2 3 4]
```

```
SMS_received :
 [0 1]
No-show :
 ['No' 'Yes']
```

In [174…

```
dt.describe(include='all')
```

Out[174…

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	
count	1.105270e+05	1.105270e+05	110527	110527	110527	110527.00
unique	NaN	NaN	2	103549	27	
top	NaN	NaN	F	2016-05-06T07:09:54Z	2016-06-06T00:00:00Z	
freq	NaN	NaN	71840	24	4692	
mean	1.474963e+14	5.675305e+06	NaN	NaN	NaN	37.08
std	2.560949e+14	7.129575e+04	NaN	NaN	NaN	23.11
min	3.921784e+04	5.030230e+06	NaN	NaN	NaN	-1.00
25%	4.172614e+12	5.640286e+06	NaN	NaN	NaN	18.00
50%	3.173184e+13	5.680573e+06	NaN	NaN	NaN	37.00
75%	9.439172e+13	5.725524e+06	NaN	NaN	NaN	55.00
max	9.999816e+14	5.790484e+06	NaN	NaN	NaN	115.00

In [ ]:

In [175…

```
# Clean and extract from AppointmentDay
```

In [176…

```
dt['AppointmentDay'] = pd.to_datetime(dt['AppointmentDay'], errors='coerce', utc=Tr
```

In [177…

```
# extract date without time
```

In [178…

```
dt['Appointment_Date'] = dt['AppointmentDay'].dt.date
```

In [179…

```
# extract day no
```

In [180…

```
dt['Appointment_DayName'] = dt['AppointmentDay'].dt.day_name()
```

In [181…

```
dt['Appointment_DayNum'] = dt['AppointmentDay'].dt.dayofweek
```

In [182…

```
print(dt[['AppointmentDay', 'Appointment_Date', 'Appointment_DayName']].head())
```

```
              AppointmentDay Appointment_Date Appointment_DayName
0 2016-04-29 00:00:00+00:00       2016-04-29              Friday
1 2016-04-29 00:00:00+00:00       2016-04-29              Friday
2 2016-04-29 00:00:00+00:00       2016-04-29              Friday
3 2016-04-29 00:00:00+00:00       2016-04-29              Friday
4 2016-04-29 00:00:00+00:00       2016-04-29              Friday
```

In [183…

```
# drop original
```

In [184…

```
dt = dt.drop(columns=['AppointmentDay'])
```

In [185…

```
dt.head(5)
```

Out[185…

	PatientId	AppointmentID	Gender	ScheduledDay	Age	Neighbourhood	Scholarshi
0	2.987250e+13	5642903	F	2016-04-29T18:38:08Z	62	JARDIM DA PENHA	
1	5.589978e+14	5642503	M	2016-04-29T16:08:27Z	56	JARDIM DA PENHA	
2	4.262962e+12	5642549	F	2016-04-29T16:19:04Z	62	MATA DA PRAIA	
3	8.679512e+11	5642828	F	2016-04-29T17:29:31Z	8	PONTAL DE CAMBURI	
4	8.841186e+12	5642494	F	2016-04-29T16:07:23Z	56	JARDIM DA PENHA	

In [186…

```
#droping the appoint day num as it not nessary
```

In [187…

```
dt = dt.drop(columns=['Appointment_DayNum'])
```

In [188…

```
dt.head(5)
```

Out[188…

	PatientId	AppointmentID	Gender	ScheduledDay	Age	Neighbourhood	Scholarshi
0	2.987250e+13	5642903	F	2016-04-29T18:38:08Z	62	JARDIM DA PENHA	
1	5.589978e+14	5642503	M	2016-04-29T16:08:27Z	56	JARDIM DA PENHA	
2	4.262962e+12	5642549	F	2016-04-29T16:19:04Z	62	MATA DA PRAIA	
3	8.679512e+11	5642828	F	2016-04-29T17:29:31Z	8	PONTAL DE CAMBURI	
4	8.841186e+12	5642494	F	2016-04-29T16:07:23Z	56	JARDIM DA PENHA	

In [189…

```python
# now cleane and extract the scheduleday column
```

In [190…

```python
dt['ScheduledDay'] = pd.to_datetime(dt['ScheduledDay'], errors='coerce', utc=True)
```

In [191…

```python
dt['ScheduledDay'] = dt['ScheduledDay'].dt.date
```

In [193…

```python
print(dt['ScheduledDay'].dtype)
```

```
object
```

In [194…

```python
dt['ScheduledDay'] = pd.to_datetime(dt['ScheduledDay'], errors='coerce', utc=True)
```

In [195…

```python
# extract day and date by its name
```

In [196…

```python
dt['Scheduled_Date'] = dt['ScheduledDay'].dt.date
dt['Scheduled_DayName'] = dt['ScheduledDay'].dt.day_name()
```

In [197…

```python
print(dt[['ScheduledDay', 'Scheduled_Date', 'Scheduled_DayName']].head())
```

```
               ScheduledDay Scheduled_Date Scheduled_DayName
0 2016-04-29 00:00:00+00:00     2016-04-29            Friday
1 2016-04-29 00:00:00+00:00     2016-04-29            Friday
2 2016-04-29 00:00:00+00:00     2016-04-29            Friday
3 2016-04-29 00:00:00+00:00     2016-04-29            Friday
4 2016-04-29 00:00:00+00:00     2016-04-29            Friday
```

In [198…

```python
# dropping the unnessary columns
```

In [199…

```python
dt = dt.drop(columns=['ScheduledDay'])
```

In [200…

```python
dt = dt.drop(columns=['Scheduled_DayName'])
```

In [201…

```python
dt = dt.drop(columns=['Appointment_DayName'])
```

In [202…

```
dt.head(5)
```

Out[202…

	PatientId	AppointmentID	Gender	Age	Neighbourhood	Scholarship	Hipertension
0	2.987250e+13	5642903	F	62	JARDIM DA PENHA	0	1
1	5.589978e+14	5642503	M	56	JARDIM DA PENHA	0	0
2	4.262962e+12	5642549	F	62	MATA DA PRAIA	0	0
3	8.679512e+11	5642828	F	8	PONTAL DE CAMBURI	0	0
4	8.841186e+12	5642494	F	56	JARDIM DA PENHA	0	1

In [203…

```
# now remove scientific notation from patientid columns
```

In [204…

```
print(dt['PatientId'].dtype)
```

```
float64
```

In [205…

```
# Handle missing values and convert properly
```

In [206…

```
# Convert to string safely
dt['PatientId'] = dt['PatientId'].apply(lambda x: str(int(x)) if pd.notnull(x) else
```

In [207…

```
dt['PatientId'] = dt['PatientId'].str.zfill(2)
```

In [208…

```
print(dt['PatientId'].head())
print(dt['PatientId'].dtype)
```

```
0      29872499824296
1     558997776694438
2       4262962299951
3        867951213174
4       8841186448183
Name: PatientId, dtype: object
object
```

In [209…

```
dt['PatientId'] = dt['PatientId'].str.zfill(2)
```

In [210…

```
# Clean PatientId column
if 'PatientId' in dt.columns:
    dt['PatientId'] = dt['PatientId'].apply(lambda x: str(int(x)) if pd.notnull(x)
    print("\n✅ PatientId column cleaned successfully!")
    print(dt['PatientId'].head())
```

```
✅ PatientId column cleaned successfully!
0     29872499824296
1    558997776694438
2      4262962299951
3       867951213174
4      8841186448183
Name: PatientId, dtype: object
```

In [211…

```
dt.head(5)
```

Out[211…

	PatientId	AppointmentID	Gender	Age	Neighbourhood	Scholarship	Hiperten
0	29872499824296	5642903	F	62	JARDIM DA PENHA	0	
1	558997776694438	5642503	M	56	JARDIM DA PENHA	0	
2	4262962299951	5642549	F	62	MATA DA PRAIA	0	
3	867951213174	5642828	F	8	PONTAL DE CAMBURI	0	
4	8841186448183	5642494	F	56	JARDIM DA PENHA	0	

In [212…

```
dt.head(10)
```

Out[212…

	PatientId	AppointmentID	Gender	Age	Neighbourhood	Scholarship	Hiperten
0	29872499824296	5642903	F	62	JARDIM DA PENHA	0	
1	558997776694438	5642503	M	56	JARDIM DA PENHA	0	
2	4262962299951	5642549	F	62	MATA DA PRAIA	0	
3	867951213174	5642828	F	8	PONTAL DE CAMBURI	0	
4	8841186448183	5642494	F	56	JARDIM DA PENHA	0	
5	95985133231274	5626772	F	76	REPÚBLICA	0	
6	733688164476661	5630279	F	23	GOIABEIRAS	0	
7	3449833394123	5630575	F	39	GOIABEIRAS	0	
8	56394729949972	5638447	F	21	ANDORINHAS	0	
9	78124564369297	5629123	F	19	CONQUISTA	0	

In [213…

```
dt.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 110527 entries, 0 to 110526
Data columns (total 14 columns):
 #   Column            Non-Null Count   Dtype 
---  ------            --------------   ----- 
 0   PatientId         110527 non-null  object
 1   AppointmentID     110527 non-null  int64 
 2   Gender            110527 non-null  object
 3   Age               110527 non-null  int64 
 4   Neighbourhood     110527 non-null  object
 5   Scholarship       110527 non-null  int64 
 6   Hipertension      110527 non-null  int64 
 7   Diabetes          110527 non-null  int64 
 8   Alcoholism        110527 non-null  int64 
 9   Handcap           110527 non-null  int64 
 10  SMS_received      110527 non-null  int64 
 11  No-show           110527 non-null  object
 12  Appointment_Date  110527 non-null  object
 13  Scheduled_Date    110527 non-null  object
dtypes: int64(8), object(6)
memory usage: 11.8+ MB
```

In [214…

```
# Convert to numeric (handles scientific notation)
# dt['PatientId'] = pd.to_numeric(df['PatientId'], errors='coerce')

# Drop or handle missing IDs if any
# dt = dt.dropna(subset=['PatientId'])

# Convert to integer type
dt['PatientId'] = dt['PatientId'].astype('int64')
```

In [215…

```
dt.dtypes
```

Out[215…

```
PatientId            int64
AppointmentID        int64
Gender              object
Age                  int64
Neighbourhood       object
Scholarship          int64
Hipertension         int64
Diabetes             int64
Alcoholism           int64
Handcap              int64
SMS_received         int64
No-show             object
Appointment_Date    object
Scheduled_Date      object
dtype: object
```

In [216…

```
dt.head(5)
```

Out[216…

	PatientId	AppointmentID	Gender	Age	Neighbourhood	Scholarship	Hiperten
0	29872499824296	5642903	F	62	JARDIM DA PENHA	0	
1	558997776694438	5642503	M	56	JARDIM DA PENHA	0	
2	4262962299951	5642549	F	62	MATA DA PRAIA	0	
3	867951213174	5642828	F	8	PONTAL DE CAMBURI	0	
4	8841186448183	5642494	F	56	JARDIM DA PENHA	0	

In [220…

```
dt.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 110527 entries, 0 to 110526
Data columns (total 14 columns):
 #   Column            Non-Null Count   Dtype 
---  ------            --------------   ----- 
 0   PatientId         110527 non-null  int64 
 1   AppointmentID     110527 non-null  int64 
 2   Gender            110527 non-null  object
 3   Age               110527 non-null  int64 
 4   Neighbourhood     110527 non-null  object
 5   Scholarship       110527 non-null  int64 
 6   Hipertension      110527 non-null  int64 
 7   Diabetes          110527 non-null  int64 
 8   Alcoholism        110527 non-null  int64 
 9   Handcap           110527 non-null  int64 
 10  SMS_received      110527 non-null  int64 
 11  No-show           110527 non-null  object
 12  Appointment_Date  110527 non-null  object
 13  Scheduled_Date    110527 non-null  object
dtypes: int64(9), object(5)
memory usage: 11.8+ MB
```

In [226…

```
dt['Appointment_Date'] = dt['Appointment_Date'].astype(str)
```

In [227…

```
dt['Scheduled_Date'] = dt['Scheduled_Date'].astype(str)
```

In [228…

```
dt.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 110527 entries, 0 to 110526
Data columns (total 14 columns):
 #   Column            Non-Null Count   Dtype 
---  ------            --------------   ----- 
 0   PatientId         110527 non-null  int64 
 1   AppointmentID     110527 non-null  int64 
 2   Gender            110527 non-null  object
 3   Age               110527 non-null  int64 
 4   Neighbourhood     110527 non-null  object
 5   Scholarship       110527 non-null  int64 
 6   Hipertension      110527 non-null  int64 
 7   Diabetes          110527 non-null  int64 
 8   Alcoholism        110527 non-null  int64 
 9   Handcap           110527 non-null  int64 
 10  SMS_received      110527 non-null  int64 
 11  No-show           110527 non-null  object
 12  Appointment_Date  110527 non-null  object
 13  Scheduled_Date    110527 non-null  object
dtypes: int64(9), object(5)
memory usage: 11.8+ MB
```

In [232…

```
dt.to_csv("Medical Appointment No Shows .csv", index=False)
```

In [ ]: