

Missed Appointment Data Analysis

In [1]:

```
#import required modules
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

Loading Dataset

In [2]:

```
df1=pd.read_csv('c:\\users\\manu\\desktop\\appoint2.csv')
```

In [3]:

```
df1
```

Out[3]:

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipert
0	2.987250e+13	5642903	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1
1	5.589978e+14	5642503	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0
2	4.262962e+12	5642549	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0
3	8.679512e+11	5642828	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0
4	8.841186e+12	5642494	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	1
5	9.598513e+13	5626772	F	2016-04-27T08:36:51Z	2016-04-29T00:00:00Z	76	REPÚBLICA	0	1
6	7.336882e+14	5630279	F	2016-04-27T15:05:12Z	2016-04-29T00:00:00Z	23	GOIABEIRAS	0	0
7	3.449833e+12	5630575	F	2016-04-27T15:39:58Z	2016-04-29T00:00:00Z	39	GOIABEIRAS	0	0
8	5.639473e+13	5638447	F	2016-04-29T08:02:16Z	2016-04-29T00:00:00Z	21	ANDORINHAS	0	0
9	7.812456e+13	5629123	F	2016-04-27T12:48:25Z	2016-04-29T00:00:00Z	19	CONQUISTA	0	0
10	7.345362e+14	5630213	F	2016-04-27T14:58:11Z	2016-04-29T00:00:00Z	30	NOVA PALESTINA	0	0
11	7.542951e+12	5620163	M	2016-04-26T08:44:12Z	2016-04-29T00:00:00Z	29	NOVA PALESTINA	0	0
12	5.666548e+14	5634718	F	2016-04-28T11:33:51Z	2016-04-29T00:00:00Z	22	NOVA PALESTINA	1	0
13	9.113946e+14	5636249	M	2016-04-28T14:52:07Z	2016-04-29T00:00:00Z	28	NOVA PALESTINA	0	0
14	9.988472e+13	5633951	F	2016-04-28T10:06:24Z	2016-04-29T00:00:00Z	54	NOVA PALESTINA	0	0

15	9.994839e+10	5620206	F	2016-04-28T08:51:47Z	2016-04-29T00:00:00Z	15	NOVA PALESTINA	0	0
16	8.457439e+13	5633121	M	2016-04-28T08:51:47Z	2016-04-29T00:00:00Z	50	NOVA PALESTINA	0	0
17	1.479497e+13	5633460	F	2016-04-28T09:28:57Z	2016-04-29T00:00:00Z	40	CONQUISTA	1	0
18	1.713538e+13	5621836	F	2016-04-26T10:54:18Z	2016-04-29T00:00:00Z	30	NOVA PALESTINA	1	0
19	7.223289e+12	5640433	F	2016-04-29T10:43:14Z	2016-04-29T00:00:00Z	46	DA PENHA	0	0
20	6.222575e+14	5626083	F	2016-04-27T07:51:14Z	2016-04-29T00:00:00Z	30	NOVA PALESTINA	0	0
21	1.215484e+13	5628338	F	2016-04-27T10:50:45Z	2016-04-29T00:00:00Z	4	CONQUISTA	0	0
22	8.632298e+14	5616091	M	2016-04-25T13:29:16Z	2016-04-29T00:00:00Z	13	CONQUISTA	0	0
23	2.137540e+14	5634142	F	2016-04-28T10:27:05Z	2016-04-29T00:00:00Z	46	CONQUISTA	0	0
24	8.734858e+12	5641780	F	2016-04-29T14:19:19Z	2016-04-29T00:00:00Z	65	TABUAZEIRO	0	0
25	5.819370e+12	5624020	M	2016-04-26T15:04:17Z	2016-04-29T00:00:00Z	46	CONQUISTA	0	1
26	2.578785e+10	5641781	F	2016-04-29T14:19:42Z	2016-04-29T00:00:00Z	45	BENTO FERREIRA	0	1
27	1.215484e+13	5628345	F	2016-04-27T10:51:45Z	2016-04-29T00:00:00Z	4	CONQUISTA	0	0
28	5.926172e+12	5642400	M	2016-04-29T15:48:02Z	2016-04-29T00:00:00Z	51	SÃO PEDRO	0	0
29	1.225776e+12	5642186	F	2016-04-29T15:16:29Z	2016-04-29T00:00:00Z	32	SANTA MARTHA	0	0
...
110497	7.935892e+14	5757745	M	2016-06-01T09:46:33Z	2016-06-01T00:00:00Z	76	MARIA ORTIZ	0	0
110498	9.433654e+13	5787655	F	2016-06-08T10:21:14Z	2016-06-08T00:00:00Z	59	MARIA ORTIZ	0	0
110499	8.219692e+14	5757697	F	2016-06-01T09:42:56Z	2016-06-01T00:00:00Z	66	MARIA ORTIZ	0	1
110500	4.434384e+14	5787233	F	2016-06-08T09:35:13Z	2016-06-08T00:00:00Z	59	MARIA ORTIZ	0	0
110501	4.544252e+11	5758133	M	2016-06-01T10:19:12Z	2016-06-01T00:00:00Z	44	MARIA ORTIZ	0	0
110502	7.316229e+14	5787937	F	2016-06-08T10:50:42Z	2016-06-08T00:00:00Z	22	GOIABEIRAS	0	0
110503	2.362182e+13	5759473	F	2016-06-01T13:00:36Z	2016-06-01T00:00:00Z	64	OLON BORGES	0	0
110504	9.947983e+12	5788052	F	2016-06-08T11:06:21Z	2016-06-08T00:00:00Z	4	MARIA ORTIZ	0	0
110505	5.667344e+13	5758455	F	2016-06-01T10:45:50Z	2016-06-01T00:00:00Z	55	MARIA ORTIZ	0	0
110506	8.973883e+11	5758779	M	2016-06-01T11:09:20Z	2016-06-01T00:00:00Z	5	MARIA ORTIZ	0	0
110507	4.769462e+14	5786918	F	2016-06-08T09:04:18Z	2016-06-08T00:00:00Z	0	MARIA ORTIZ	0	0

110508	9.433654e+13	5757656	F	2016-06-01T09:41:00Z	2016-06-01T00:00:00Z	59	MARIA ORTIZ	0	0
	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipert
110509	4.952968e+14	5786750	M	2016-06-08T08:50:51Z	2016-06-08T00:00:00Z	33	MARIA ORTIZ	0	0
110510	2.362182e+13	5757587	F	2016-06-01T09:35:48Z	2016-06-01T00:00:00Z	64	SOLON BORGES	0	0
110511	8.235996e+11	5786742	F	2016-06-08T08:50:20Z	2016-06-08T00:00:00Z	14	MARIA ORTIZ	0	0
110512	9.876246e+13	5786368	F	2016-06-08T08:20:01Z	2016-06-08T00:00:00Z	41	MARIA ORTIZ	0	0
110513	8.674778e+13	5785964	M	2016-06-08T07:52:55Z	2016-06-08T00:00:00Z	2	ANTÔNIO HONÓRIO	0	0
110514	2.695685e+12	5786567	F	2016-06-08T08:35:31Z	2016-06-08T00:00:00Z	58	MARIA ORTIZ	0	0
110515	6.456342e+14	5778621	M	2016-06-06T15:58:05Z	2016-06-08T00:00:00Z	33	MARIA ORTIZ	0	1
110516	6.923772e+13	5780205	F	2016-06-07T07:45:16Z	2016-06-08T00:00:00Z	37	MARIA ORTIZ	0	0
110517	5.574942e+12	5780122	F	2016-06-07T07:38:34Z	2016-06-07T00:00:00Z	19	MARIA ORTIZ	0	0
110518	7.263315e+13	5630375	F	2016-04-27T15:15:06Z	2016-06-07T00:00:00Z	50	MARIA ORTIZ	0	0
110519	6.542388e+13	5630447	F	2016-04-27T15:23:14Z	2016-06-07T00:00:00Z	22	MARIA ORTIZ	0	0
110520	9.969977e+14	5650534	F	2016-05-03T07:51:47Z	2016-06-07T00:00:00Z	42	MARIA ORTIZ	0	0
110521	3.635534e+13	5651072	F	2016-05-03T08:23:40Z	2016-06-07T00:00:00Z	53	MARIA ORTIZ	0	0
110522	2.572134e+12	5651768	F	2016-05-03T09:15:35Z	2016-06-07T00:00:00Z	56	MARIA ORTIZ	0	0
110523	3.596266e+12	5650093	F	2016-05-03T07:27:33Z	2016-06-07T00:00:00Z	51	MARIA ORTIZ	0	0
110524	1.557663e+13	5630692	F	2016-04-27T16:03:52Z	2016-06-07T00:00:00Z	21	MARIA ORTIZ	0	0
110525	9.213493e+13	5630323	F	2016-04-27T15:09:23Z	2016-06-07T00:00:00Z	38	MARIA ORTIZ	0	0
110526	3.775115e+14	5629448	F	2016-04-27T13:30:56Z	2016-06-07T00:00:00Z	54	MARIA ORTIZ	0	0

110527 rows × 14 columns



In [4]:

```
type(df1)
```

Out[4]:

```
pandas.core.frame.DataFrame
```

Data:

- 'ScheduledDay' tells us on what day the patient set up their appointment.
- 'Neighbourhood' indicates the location of the hospital.
- 'Scholarship' indicates whether or not the patient is enrolled in the Brazilian scholarship program.
- 'No-show' says 'No' if the patient showed up to their appointment, and 'Yes' if they did not show up.

Expioratory Analysis

In [5]:

```
df1.head() # top 5 records
```

Out[5]:

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

In [6]:

```
df1.head(2)
```

Out[6]:

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertensor
0	2.987250e+13	5642903	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1
1	5.589978e+14	5642503	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0

In [7]:

```
df1.tail() #
```

Out[7]:

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipert
110522	2.572134e+12	5651768	F	2016-05-03T09:15:35Z	2016-06-07T00:00:00Z	56	MARIA ORTIZ	0	0
110523	3.596266e+12	5650093	F	2016-05-03T07:27:33Z	2016-06-07T00:00:00Z	51	MARIA ORTIZ	0	0
110524	1.557663e+13	5630692	F	2016-04-27T16:03:52Z	2016-06-07T00:00:00Z	21	MARIA ORTIZ	0	0
110525	9.213493e+13	5630323	F	2016-04-27T15:09:23Z	2016-06-07T00:00:00Z	38	MARIA ORTIZ	0	0
110526	3.775115e+14	5629448	F	2016-04-27T13:30:56Z	2016-06-07T00:00:00Z	54	MARIA ORTIZ	0	0

In [8]:

```
# row wise selection
df1[100:300]
```

Out [8]:

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertensi
100	1.663388e+14	5574764	F	2016-04-12T15:47:45Z	2016-04-29T00:00:00Z	20	BONFIM	0	0
101	6.822847e+13	5642813	M	2016-04-29T17:24:19Z	2016-04-29T00:00:00Z	0	BONFIM	0	0
102	4.225492e+13	5634093	F	2016-04-28T10:21:39Z	2016-04-29T00:00:00Z	54	MÁRIO CYPRESTE	0	1
103	3.715425e+11	5639379	F	2016-04-29T09:01:27Z	2016-04-29T00:00:00Z	50	SANTO ANTÔNIO	0	0
104	3.326985e+12	5632495	F	2016-04-28T08:04:48Z	2016-04-29T00:00:00Z	0	SANTO ANTÔNIO	0	0
105	2.921558e+12	5639399	M	2016-04-29T09:02:40Z	2016-04-29T00:00:00Z	51	BELA VISTA	0	1
106	4.439514e+14	5639773	M	2016-04-29T09:34:00Z	2016-04-29T00:00:00Z	54	SANTO ANTÔNIO	0	1
107	7.954396e+12	5608249	F	2016-04-20T13:30:12Z	2016-04-29T00:00:00Z	26	SANTO ANTÔNIO	0	0
108	4.337498e+13	5638829	F	2016-04-29T08:28:16Z	2016-04-29T00:00:00Z	38	BELA VISTA	0	0
109	1.295988e+14	5572081	F	2016-04-12T10:01:23Z	2016-04-29T00:00:00Z	34	SANTO ANTÔNIO	0	0
110	3.737430e+12	5574527	F	2016-04-12T15:10:45Z	2016-04-29T00:00:00Z	34	SANTO ANTÔNIO	0	0
111	3.772422e+11	5574534	F	2016-04-12T15:11:24Z	2016-04-29T00:00:00Z	56	SANTO ANTÔNIO	0	1
112	9.538868e+12	5594665	F	2016-04-18T11:05:04Z	2016-04-29T00:00:00Z	59	SANTO ANTÔNIO	0	1
113	5.999817e+13	5600591	M	2016-04-19T10:55:18Z	2016-04-29T00:00:00Z	8	NOVA PALESTINA	0	0
114	3.585621e+13	5591384	F	2016-04-15T16:58:02Z	2016-04-29T00:00:00Z	15	BELA VISTA	0	0
115	8.472380e+11	5641363	F	2016-04-29T13:24:28Z	2016-04-29T00:00:00Z	54	SANTO ANTÔNIO	0	1
116	4.673335e+12	5574680	F	2016-04-12T15:35:30Z	2016-04-29T00:00:00Z	18	SANTO ANTÔNIO	0	0
117	7.444632e+08	5539219	F	2016-04-01T14:27:30Z	2016-04-29T00:00:00Z	34	SANTO ANTÔNIO	1	0
118	9.711616e+12	5539858	F	2016-04-01T16:01:28Z	2016-04-29T00:00:00Z	35	SANTO ANTÔNIO	0	0
119	3.293153e+11	5546448	F	2016-04-05T16:14:11Z	2016-04-29T00:00:00Z	33	MÁRIO CYPRESTE	0	0
120	6.297229e+12	5590755	F	2016-04-15T14:54:29Z	2016-04-29T00:00:00Z	16	MÁRIO CYPRESTE	0	0
121	4.564312e+14	5540024	F	2016-04-01T16:40:41Z	2016-04-29T00:00:00Z	18	MÁRIO CYPRESTE	0	0
122	4.599699e+12	5642643	M	2016-04-29T16:34:59Z	2016-04-29T00:00:00Z	46	BELA VISTA	0	1
123	3.673278e+13	5621388	F	2016-04-26T10:12:46Z	2016-04-29T00:00:00Z	31	BELA VISTA	1	0
124	5.475453e+11	5642319	F	2016-04-	2016-04-	7	SANTO	0	0

				29T15:35:28Z	29T00:00:00Z		ANTONIO		
	PatientID	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertensi
125	9.997213e+13	5542592	F	2016-04-05T09:09:15Z	2016-04-29T00:00:00Z	36	MARIO CYPRESTE	0	1
126	9.447582e+14	5633576	F	2016-04-28T09:37:22Z	2016-04-29T00:00:00Z	67	PRAIA DO SUÁ	0	0
127	4.224166e+12	5561194	M	2016-04-08T10:14:36Z	2016-04-29T00:00:00Z	42	PRAIA DO SUÁ	0	1
128	9.488697e+14	5571842	F	2016-04-12T09:41:01Z	2016-04-29T00:00:00Z	34	PRAIA DO SUÁ	1	0
129	5.537367e+13	5560735	F	2016-04-08T09:28:50Z	2016-04-29T00:00:00Z	22	PRAIA DO SUÁ	0	0
...
270	8.617228e+12	5620528	F	2016-04-26T09:08:53Z	2016-04-29T00:00:00Z	45	CARATOÍRA	1	1
271	1.981635e+14	5612133	F	2016-04-25T07:58:09Z	2016-04-29T00:00:00Z	12	CARATOÍRA	0	0
272	5.119616e+12	5594085	F	2016-04-18T10:12:57Z	2016-04-29T00:00:00Z	51	NOVA PALESTINA	0	1
273	8.442143e+12	5612157	F	2016-04-25T07:59:21Z	2016-04-29T00:00:00Z	44	SANTO ANTÔNIO	0	0
274	3.995366e+12	5641070	F	2016-04-29T12:16:28Z	2016-04-29T00:00:00Z	41	MARIA ORTIZ	0	0
275	5.878159e+13	5641067	M	2016-04-29T12:15:33Z	2016-04-29T00:00:00Z	7	MARIA ORTIZ	0	0
276	1.192393e+12	5641191	M	2016-04-29T12:47:35Z	2016-04-29T00:00:00Z	48	MARIA ORTIZ	0	0
277	8.587449e+13	5641281	F	2016-04-29T13:11:41Z	2016-04-29T00:00:00Z	23	MARIA ORTIZ	0	0
278	3.784661e+13	5641065	M	2016-04-29T12:14:41Z	2016-04-29T00:00:00Z	12	MARIA ORTIZ	0	0
279	5.673474e+12	5638397	F	2016-04-29T07:59:32Z	2016-04-29T00:00:00Z	28	DA PENHA	1	0
280	2.164857e+12	5640108	M	2016-04-29T10:08:02Z	2016-04-29T00:00:00Z	3	ANDORINHAS	0	0
281	3.198485e+11	5640366	M	2016-04-29T10:35:49Z	2016-04-29T00:00:00Z	79	ILHA DO FRADE	0	0
282	9.762242e+12	5541090	F	2016-04-05T07:39:33Z	2016-04-29T00:00:00Z	3	ANDORINHAS	0	0
283	5.567160e+12	5512177	F	2016-03-28T08:37:23Z	2016-04-29T00:00:00Z	2	ANDORINHAS	0	0
284	4.761172e+14	5520021	F	2016-03-29T09:28:01Z	2016-04-29T00:00:00Z	40	ANDORINHAS	1	0
285	6.342844e+10	5430959	F	2016-03-04T09:51:43Z	2016-04-29T00:00:00Z	20	SANTA MARTHA	0	0
286	8.386226e+12	5471716	F	2016-03-15T13:11:49Z	2016-04-29T00:00:00Z	3	ANDORINHAS	0	0
287	2.246214e+13	5532909	M	2016-03-31T12:39:06Z	2016-04-29T00:00:00Z	43	CARATOÍRA	0	0
288	2.246214e+13	5532908	M	2016-03-31T12:39:06Z	2016-04-29T00:00:00Z	43	CARATOÍRA	0	0
289	7.222383e+12	5566072	M	2016-04-11T09:50:19Z	2016-04-29T00:00:00Z	7	CARATOÍRA	0	0
290	7.222383e+12	5566070	M	2016-04-11T09:50:19Z	2016-04-29T00:00:00Z	7	CARATOÍRA	0	0

				11T09:30:18Z	29T00:00:00Z				
	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertensi
291	5.264573e+14	5642545	F	2016-04-29T16:17:16Z	2016-04-29T00:00:00Z	17	GURIGICA	0	0
292	4.645238e+13	5642766	F	2016-04-29T17:06:22Z	2016-04-29T00:00:00Z	0	GURIGICA	0	0
293	9.936998e+12	5637808	M	2016-04-29T07:24:33Z	2016-04-29T00:00:00Z	17	ANDORINHAS	0	0
294	6.492948e+12	5464289	M	2016-03-14T11:18:07Z	2016-04-29T00:00:00Z	60	ANDORINHAS	0	1
295	2.811753e+13	5470709	M	2016-03-15T10:57:15Z	2016-04-29T00:00:00Z	54	ANDORINHAS	0	1
296	9.984591e+13	5640119	F	2016-04-29T10:08:47Z	2016-04-29T00:00:00Z	83	ILHA DE SANTA MARIA	0	0
297	1.412746e+13	5616693	F	2016-04-25T14:39:05Z	2016-04-29T00:00:00Z	16	ANDORINHAS	0	0
298	4.434592e+13	5473380	F	2016-03-15T16:51:57Z	2016-04-29T00:00:00Z	22	ANDORINHAS	0	0
299	9.532898e+13	5618957	M	2016-04-26T07:34:07Z	2016-04-29T00:00:00Z	1	ANDORINHAS	0	0

200 rows × 14 columns



In [9]:

```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 110527 entries, 0 to 110526
Data columns (total 14 columns):
PatientId      110527 non-null float64
AppointmentID  110527 non-null int64
Gender         110527 non-null object
ScheduledDay   110527 non-null object
AppointmentDay 110527 non-null object
Age           110527 non-null int64
Neighbourhood  110527 non-null object
Scholarship    110527 non-null int64
Hipertension   110527 non-null int64
Diabetes       110527 non-null int64
Alcoholism     110527 non-null int64
Handcap       110527 non-null int64
SMS_received   110527 non-null int64
No-show        110527 non-null object
dtypes: float64(1), int64(8), object(5)
memory usage: 11.8+ MB
```

In [10]:

```
df1.columns
```

Out[10]:

```
Index(['PatientId', 'AppointmentID', 'Gender', 'ScheduledDay',
       'AppointmentDay', 'Age', 'Neighbourhood', 'Scholarship', 'Hipertension',
       'Diabetes', 'Alcoholism', 'Handcap', 'SMS_received', 'No-show'],
      dtype='object')
```

In [11]:

```
df1.dtypes
```

Out[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
```

Column Wise Selection

```
In [12]:
```

```
df1['Gender'].head()
```

```
Out[12]:
```

```
0    F
1    M
2    F
3    F
4    F
Name: Gender, dtype: object
```

```
In [13]:
```

```
df1[['Age', 'No-show', 'Diabetes']].head()
```

```
Out[13]:
```

	Age	No-show	Diabetes
0	62	No	0
1	56	No	0
2	62	No	0
3	8	No	0
4	56	No	1

Overall Patient Attendance

```
In [14]:
```

```
df1['No-show'].unique()
```

```
Out[14]:
```

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

```
In [15]:
```

```
df1['No-show'].value_counts()
```

```
Out[15]:
```

```
No      88208
Yes     22319
Name: No-show, dtype: int64
```

```
In [16]:
```



```
df1['No-show'].value_counts()['No']/(df1['No-show'].value_counts()['Yes']+df1['No-show'].value_counts()['No'])*100
```

Out[16]:

79.8067440534892

Here, 79.8% patients have attended the appointment.

In [17]:

```
df1['No-show'].value_counts()['Yes']/(df1['No-show'].value_counts()['Yes']+df1['No-show'].value_counts()['No'])*100
```

Out[17]:

20.193255946510806

Here, 20.19% patients have not attended the appointment.

Gender-wise analysis

In [18]:

```
df1['Gender'].value_counts()
```

Out[18]:

```
F    71840
M    38687
Name: Gender, dtype: int64
```

In [19]:

```
female=df1[df1.Gender=='F']['No-show'].value_counts()
```

In [20]:

```
female['No']/(female['Yes']+female['No'])*100
```

Out[20]:

79.68541202672606

Here, 79.68% females have attended the appointment.

In [21]:

```
male=df1[df1.Gender=='M']['No-show'].value_counts()
male['No']/(male['Yes']+male['No'])*100
```

Out[21]:

80.03205211052808

Here, 80.03% males have attended the appointment.

From above results, we can say that there is no impact of gender on showing up for appointment.

Diabetes wise analysis

In [22]:

```
diabetes=df1['Diabetes'].value_counts()
```

In [23]:

```
diabetes
```

Out[23]:

```
0    102584
1      7943
Name: Diabetes, dtype: int64
```

In [24]:

```
diabetes[1]/(diabetes[0]+diabetes[1])*100
```

Out[24]:

```
7.186479321794674
```

In [25]:

```
df1.columns
```

Out[25]:

```
Index(['PatientId', 'AppointmentID', 'Gender', 'ScheduledDay',
       'AppointmentDay', 'Age', 'Neighbourhood', 'Scholarship', 'Hipertension',
       'Diabetes', 'Alcoholism', 'Handcap', 'SMS_received', 'No-show'],
      dtype='object')
```

In [26]:

```
diabetes_no_show=df1[df1.Diabetes==1]['No-show'].value_counts()
```

In [27]:

```
diabetes_no_show['Yes']/(diabetes_no_show['No']+diabetes_no_show['Yes'])*100
```

Out[27]:

```
18.00327332242226
```

Hipertension wise analysis

In [28]:

```
diabetes=df1['Hipertension'].value_counts()
diabetes[1]/(diabetes[0]+diabetes[1])*100
```

Out[28]:

```
19.72459218109602
```

In [29]:

```
diabetes_no_show=df1[df1.Hipertension==1]['No-show'].value_counts()
diabetes_no_show['Yes']/(diabetes_no_show['No']+diabetes_no_show['Yes'])*100
```

Out[29]:

```
17.301958625751112
```

In [30]:

```
df1.columns
```

Out[30]:

```
Index(['PatientId', 'AppointmentID', 'Gender', 'ScheduledDay',  
      'AppointmentDay', 'Age', 'Neighbourhood', 'Scholarship', 'Hipertension',  
      'Diabetes', 'Alcoholism', 'Handcap', 'SMS_received', 'No-show'],  
      dtype='object')
```

Alcoholism wise analysis

In [31]:

```
diabetes=df1['Alcoholism'].value_counts()  
diabetes[1]/(diabetes[0]+diabetes[1])*100
```

Out[31]:

3.039981181068879

In [32]:

```
diabetes_no_show=df1[df1.Alcoholism==1]['No-show'].value_counts()  
diabetes_no_show['Yes']/(diabetes_no_show['No']+diabetes_no_show['Yes'])*100
```

Out[32]:

20.148809523809526

Handcap wise analysis

In [33]:

```
diabetes=df1['Handcap'].value_counts()  
diabetes[1]/(diabetes[0]+diabetes[1])*100
```

Out[33]:

1.8508447538249584

In [34]:

```
diabetes_no_show=df1[df1.Handcap==1]['No-show'].value_counts()  
diabetes_no_show['Yes']/(diabetes_no_show['No']+diabetes_no_show['Yes'])*100
```

Out[34]:

17.92360430950049

SMS_received

In [35]:

```
df1['SMS_received'].value_counts()
```

Out[35]:

```
0    75045  
1    35482  
Name: SMS_received, dtype: int64
```

In [36]:

```
diabetes=df1['SMS_received'].value_counts()  
diabetes[0]/(diabetes[0]+diabetes[1])*100
```

Out[36]:

67.89743682539108

In [37]:

```
diabetes_no_show=df1[df1.SMS_received==1]['No-show'].value_counts()  
diabetes_no_show['Yes']/(diabetes_no_show['No']+diabetes_no_show['Yes'])*100
```

Out[37]:

27.574544839637

Here, 27% patients have not attended the appoointment even after getting the SMS.

In [38]:

```
diabetes_no_show=df1[df1.SMS_received==0]['No-show'].value_counts()  
diabetes_no_show['Yes']/(diabetes_no_show['No']+diabetes_no_show['Yes'])*100
```

Out[38]:

16.703311346525417

Age Wise Analysis

In [39]:

```
df1.Age.describe()
```

Out[39]:

```
count    110527.000000  
mean       37.088874  
std        23.110205  
min        -1.000000  
25%        18.000000  
50%        37.000000  
75%        55.000000  
max       115.000000  
Name: Age, dtype: float64
```

In [40]:

```
df1=df1[(df1.Age>0) & (df1.Age<100)]
```

In [41]:

```
df1.info()
```

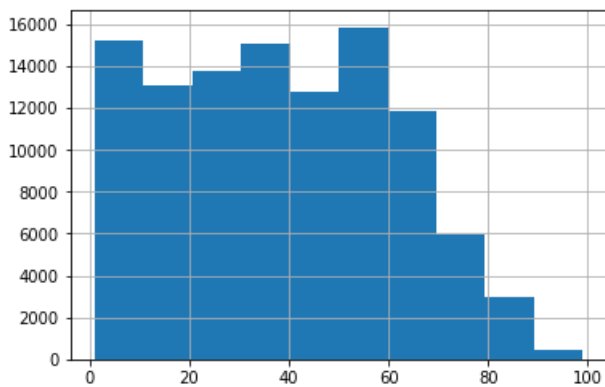
```
<class 'pandas.core.frame.DataFrame'>  
Int64Index: 106976 entries, 0 to 110526  
Data columns (total 14 columns):  
PatientId      106976 non-null float64  
AppointmentID  106976 non-null int64  
Gender         106976 non-null object  
ScheduledDay   106976 non-null object  
AppointmentDay 106976 non-null object  
Age            106976 non-null int64  
Neighbourhood  106976 non-null object  
Scholarship    106976 non-null int64  
Hypertension   106976 non-null int64  
Diabetes       106976 non-null int64  
Alcoholism     106976 non-null int64  
Handcap        106976 non-null int64  
SMS_received   106976 non-null int64  
No-show        106976 non-null object  
dtypes: float64(1), int64(8), object(5)  
memory usage: 12.2+ MB
```

In [42]:

```
df1.Age.hist(bins=10)
```

Out[42]:

<matplotlib.axes._subplots.AxesSubplot at 0x99f75f8>



In [43]:

```
df1.columns
```

Out[43]:

```
Index(['PatientId', 'AppointmentID', 'Gender', 'ScheduledDay',  
      'AppointmentDay', 'Age', 'Neighbourhood', 'Scholarship', 'Hipertension',  
      'Diabetes', 'Alcoholism', 'Handcap', 'SMS_received', 'No-show'],  
      dtype='object')
```

In [44]:

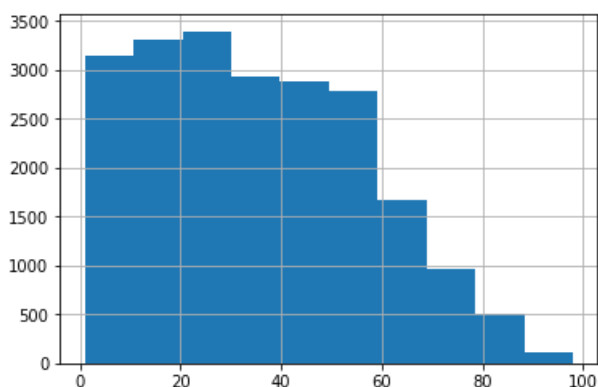
```
df1.columns=['PatientId', 'AppointmentID', 'Gender', 'ScheduledDay',  
            'AppointmentDay', 'Age', 'Neighbourhood', 'Scholarship', 'Hipertension',  
            'Diabetes', 'Alcoholism', 'Handcap', 'SMS_received', 'No_show']
```

In [45]:

```
df1[df1.No_show=='Yes'].Age.hist(bins=10)
```

Out[45]:

<matplotlib.axes._subplots.AxesSubplot at 0x99e3438>



In [46]:

```
df1['Gender'].replace('F',0,inplace=True)  
df1['Gender'].replace('M',1,inplace=True)
```

C:\ProgramData\Anaconda3\lib\site-packages\pandas\core\generic.py:5886: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: <http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy>
self._update_inplace(new_data)

In [47]:

```
df1['No_show'].replace('No',0,inplace=True)  
df1['No_show'].replace('Yes',1,inplace=True)
```

C:\ProgramData\Anaconda3\lib\site-packages\pandas\core\generic.py:5886: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: <http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy>
self._update_inplace(new_data)

In []:

```
df1.corr()
```

In [65]:

```
df1[df.No_show==1].Neighbourhood.value_counts().sort_values(ascending=False)[0:20].plot(kind='bar')
```

```
-----  
NameError                                Traceback (most recent call last)  
<ipython-input-65-b241c79d8aa8> in <module>()  
----> 1 df1[df.No_show==1].Neighbourhood.value_counts().sort_values(ascending=False)[0:20].plot(kind='bar')
```

NameError: name 'df' is not defined

In [50]:

```
df1.head()
```

Out[50]:

	PatientId	AppointmentId	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertensor
0	2.987250e+13	5642903	0	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1
1	5.589978e+14	5642503	1	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0
2	4.262962e+12	5642549	0	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0
3	8.679512e+11	5642828	0	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0
4	8.841186e+12	5642494	0	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	1

In [51]:

```
df1['PatientId']=df1['PatientId'].astype(int)
```

C:\ProgramData\Anaconda3\lib\site-packages\ipykernel_launcher.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: <http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy>
"""Entry point for launching an IPython kernel.

In [52]:

```
df1.head()
```

Out[52]:

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension
0	-2147483648	5642903	0	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1
1	-2147483648	5642503	1	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0
2	-2147483648	5642549	0	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0
3	-2147483648	5642828	0	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0
4	-2147483648	5642494	0	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	1

In [53]:

```
df1.describe()
```

Out[53]:

	PatientId	AppointmentID	Gender	Age	Scholarship	Hipertension	Diabetes	Alc
count	1.069760e+05	1.069760e+05	106976.000000	106976.000000	106976.000000	106976.000000	106976.000000	106976.000000
mean	-2.140303e+09	5.675432e+06	0.344629	38.309004	0.101041	0.203784	0.074250	0.0314
std	1.509123e+08	7.133383e+04	0.475249	22.456395	0.301385	0.402812	0.262179	0.1744
min	-2.147484e+09	5.030230e+06	0.000000	1.000000	0.000000	0.000000	0.000000	0.0000
25%	-2.147484e+09	5.640488e+06	0.000000	19.000000	0.000000	0.000000	0.000000	0.0000
50%	-2.147484e+09	5.680740e+06	0.000000	38.000000	0.000000	0.000000	0.000000	0.0000
75%	-2.147484e+09	5.725632e+06	1.000000	56.000000	0.000000	0.000000	0.000000	0.0000
max	2.141274e+09	5.790484e+06	1.000000	99.000000	1.000000	1.000000	1.000000	1.0000

In [54]:

```
df1.shape
```

Out[54]:

(106976, 14)

In [55]:

```
df1.head()
```

Out[55]:

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension
0	-2147483648	5642903	0	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1
1	-2147483648	5642503	1	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension
2	2147483648	5642549	0	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0
3	2147483648	5642828	0	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0
4	2147483648	5642494	0	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	1

◀		▶
---	--	---

In [56]:

```
type(df1['ScheduledDay'][0])
```

Out[56]:

str

In [63]:

```
# {0:'Mon',1:'Tues',2:'Weds',3:'Thurs',4:'Fri',5:'Sat',6:'Sun'}
pd.to_datetime(df1['ScheduledDay']).dt.dayofweek
```

Out[63]:

```
0      4
1      4
2      4
3      4
4      4
5      2
6      2
7      2
8      4
9      2
10     2
11     1
12     3
13     3
14     3
15     1
16     3
17     3
18     1
19     4
20     2
21     2
22     0
23     3
24     4
25     1
26     4
27     2
28     4
29     4
..
110496  0
110497  2
110498  2
110499  2
110500  2
110501  2
110502  2
110503  2
110504  2
110505  2
110506  2
110508  2
110509  2
110510  2
110511  2
110512  2
110513  2
110514  2
110515  0
```



```
110515    0
110516    1
110517    1
110518    2
110519    2
110520    1
110521    1
110522    1
110523    1
110524    2
110525    2
110526    2
Name: ScheduledDay, Length: 106976, dtype: int64
```

In [64]:

```
df1['week_day'][df1.No_show==1].value_counts().plot(kind='bar')
```

Out[64]:

<matplotlib.axes._subplots.AxesSubplot at 0xae7df28>

