Investigate_a_Dataset

August 13, 2021

1 Project: Investigate a Dataset of Medical Appointment No Shows

1.1 Table of Contents

Introduction

This dataset collects information from 100k medical appointments in Brazil and is focused on the question of whether or not patients show up for their appointment. A number of characteristics about the patient are included in each row. 'ScheduledDay' tells us on what day the patient set up their appointment. 'Neighborhood' indicates the location of the hospital. 'Scholarship' indicates whether or not the patient is enrolled in Brasilian welfare program Bolsa Família.

The goal of this investigation is to study the trends among people who don't show up at their scheduled appointments. To do so we will focus on answering some questions like: 1) What is the Overall Show Vs No-Show Percentage? 2) What is the correlation between Age and No-show columns? 3) What is the correlation between Gender and No-show columns? 4) Does receiving an SMS affect commitment to the appointments? 4) Does having a scolarship affect commitment to the appointments?

```
In [46]: # Import statements for all of the packages
         import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt
         import seaborn as sns

         %matplotlib inline
```

Data Wrangling

In this section of the report, we will load in the data, check for cleanliness, and then trim and clean our dataset for analysis.

1.1.1 Data Summary

```
In [47]: # Loading data and printing out a few lines.
         df = pd.read_csv('KaggleV2-May-2016.csv')
         df.head()
```

Out[47]:

	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 [48]:
```
# Taking a look at the shape to know the exact number of patients data collected
df.shape
```

Out[48]: (110527, 14)

In [49]:
```
# Getting some insight from a summary statistic
df.describe()
```

Out[49]:

	PatientId	AppointmentID	Age	Scholarship	Hipertension	Diabetes	Alcoholism	Handcap
count	1.105270e+05	1.105270e+05	110527.000000	110527.000000	110527.000000	110527.000000	110527.000000	110527.000000
mean	1.474963e+14	5.675305e+06	37.088874	0.098266	0.197246	0.071865	0.030400	0.022248
std	2.560949e+14	7.129575e+04	23.110205	0.297675	0.397921	0.258265	0.171686	0.161543
min	3.921784e+04	5.030230e+06	-1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	4.172614e+12	5.640286e+06	18.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	3.173184e+13	5.680573e+06	37.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	9.439172e+13	5.725524e+06	55.000000	0.000000	0.000000	0.000000	0.000000	0.000000
max	9.999816e+14	5.790484e+06	115.000000	1.000000	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
```

1.1.2 From this summary we can see that about 75% of patients don't suffer from Hipertension, Diabetes, Alcoholism or Handcap. And that most of the patients don't have a Scholarship. And about only 32% of the patients received an SMS.

```
In [50]: # Checking data types and null values.
         df.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
```

1.1.3 Luckily, we don't have any missing values in our dataset

1.2 Data Cleaning

1- Checking for and removing Duplicates. 2- Dropping Columns that we won't need in our study (such as Patient ID and Appointment ID). 3- Checking for and removing any illogical data (such as zero or negative age).

```
In [51]: # Checking for Duplicate Rows and dropping them if found.
         df.duplicated().sum()
```

```
Out[51]: 0
```

Luckily, our data has no duplicates.

```
In [53]: df.drop(['PatientId','AppointmentID'], axis = 1, inplace = True)
```

```
In [8]: df.head()
```

```
Out[8]:   Gender          ScheduledDay        AppointmentDay  Age      Neighbourhood  \
        0      F  2016-04-29T18:38:08Z  2016-04-29T00:00:00Z   62    JARDIM DA PENHA
        1      M  2016-04-29T16:08:27Z  2016-04-29T00:00:00Z   56    JARDIM DA PENHA
        2      F  2016-04-29T16:19:04Z  2016-04-29T00:00:00Z   62      MATA DA PRAIA
        3      F  2016-04-29T17:29:31Z  2016-04-29T00:00:00Z    8  PONTAL DE CAMBURI
        4      F  2016-04-29T16:07:23Z  2016-04-29T00:00:00Z   56    JARDIM DA PENHA

           Scholarship  Hipertension  Diabetes  Alcoholism  Handcap  SMS_received  \
        0            0             1         0           0        0             0
        1            0             0         0           0        0             0
        2            0             0         0           0        0             0
        3            0             0         0           0        0             0
        4            0             1         1           0        0             0

          No-show
        0      No
        1      No
        2      No
        3      No
        4      No
```

```
In [11]: # Checking for Zero or Negative Age.
         df[df["Age"] <= 0]
```

```
Out[11]:        Gender          ScheduledDay        AppointmentDay  Age  \
         59          F  2016-04-29T08:08:43Z  2016-04-29T00:00:00Z    0
         63          M  2016-04-27T10:46:12Z  2016-04-29T00:00:00Z    0
         64          M  2016-04-25T13:28:21Z  2016-04-29T00:00:00Z    0
         65          M  2016-04-27T10:48:50Z  2016-04-29T00:00:00Z    0
         67          F  2016-04-29T08:53:02Z  2016-04-29T00:00:00Z    0
         89          M  2016-04-29T10:37:02Z  2016-04-29T00:00:00Z    0
         101         M  2016-04-29T17:24:19Z  2016-04-29T00:00:00Z    0
         104         F  2016-04-28T08:04:48Z  2016-04-29T00:00:00Z    0
         132         M  2016-04-08T09:29:23Z  2016-04-29T00:00:00Z    0
         150         M  2016-04-29T13:43:34Z  2016-04-29T00:00:00Z    0
         188         M  2016-04-29T11:44:49Z  2016-04-29T00:00:00Z    0
         192         M  2016-04-29T10:57:14Z  2016-04-29T00:00:00Z    0
         193         F  2016-03-31T11:14:57Z  2016-04-29T00:00:00Z    0
         194         M  2016-04-01T15:17:10Z  2016-04-29T00:00:00Z    0
         250         M  2016-04-29T10:13:35Z  2016-04-29T00:00:00Z    0
         256         F  2016-04-27T14:19:02Z  2016-04-29T00:00:00Z    0
         266         M  2016-04-29T08:50:09Z  2016-04-29T00:00:00Z    0
         292         F  2016-04-29T17:06:22Z  2016-04-29T00:00:00Z    0
```

```
305         F  2016-04-29T07:49:54Z  2016-04-29T00:00:00Z    0
306         M  2016-04-29T07:58:15Z  2016-04-29T00:00:00Z    0
310         M  2016-04-29T08:40:58Z  2016-04-29T00:00:00Z    0
358         F  2016-03-31T10:07:18Z  2016-04-29T00:00:00Z    0
359         M  2016-04-06T14:20:30Z  2016-04-29T00:00:00Z    0
366         F  2016-03-31T10:14:54Z  2016-04-29T00:00:00Z    0
377         M  2016-03-29T14:39:32Z  2016-04-29T00:00:00Z    0
434         F  2016-03-31T15:01:12Z  2016-04-29T00:00:00Z    0
524         M  2016-03-18T10:35:28Z  2016-04-29T00:00:00Z    0
525         F  2016-04-01T10:53:45Z  2016-04-29T00:00:00Z    0
526         M  2016-04-29T09:22:07Z  2016-04-29T00:00:00Z    0
565         F  2016-04-15T17:39:38Z  2016-04-29T00:00:00Z    0
...       ...                   ...                   ...  ...
109629      F  2016-05-04T16:30:06Z  2016-06-01T00:00:00Z    0
109633      F  2016-05-04T16:30:06Z  2016-06-01T00:00:00Z    0
109646      M  2016-05-04T13:21:32Z  2016-06-01T00:00:00Z    0
109647      M  2016-05-12T12:35:04Z  2016-06-08T00:00:00Z    0
109649      M  2016-05-04T13:21:32Z  2016-06-01T00:00:00Z    0
109650      M  2016-05-12T12:35:04Z  2016-06-08T00:00:00Z    0
109830      F  2016-05-25T13:07:38Z  2016-06-02T00:00:00Z    0
109847      M  2016-05-11T11:33:48Z  2016-06-02T00:00:00Z    0
109848      M  2016-05-17T09:32:32Z  2016-06-02T00:00:00Z    0
109852      F  2016-05-24T10:14:26Z  2016-06-02T00:00:00Z    0
110231      M  2016-04-28T11:32:00Z  2016-06-01T00:00:00Z    0
110235      M  2016-04-28T10:40:49Z  2016-06-01T00:00:00Z    0
110236      M  2016-05-06T12:31:22Z  2016-06-08T00:00:00Z    0
110299      F  2016-05-30T13:47:40Z  2016-06-07T00:00:00Z    0
110313      F  2016-04-19T10:09:05Z  2016-06-06T00:00:00Z    0
110319      F  2016-04-29T10:28:16Z  2016-06-06T00:00:00Z    0
110320      M  2016-05-09T13:02:20Z  2016-06-06T00:00:00Z    0
110321      M  2016-02-11T16:14:32Z  2016-06-01T00:00:00Z    0
110331      F  2016-02-24T15:33:08Z  2016-06-01T00:00:00Z    0
110334      M  2016-06-01T08:12:55Z  2016-06-01T00:00:00Z    0
110335      M  2016-02-11T16:39:26Z  2016-06-01T00:00:00Z    0
110339      M  2016-04-14T13:01:21Z  2016-06-01T00:00:00Z    0
110341      M  2016-05-18T10:55:00Z  2016-06-01T00:00:00Z    0
110342      M  2016-06-06T11:48:00Z  2016-06-08T00:00:00Z    0
110343      F  2016-05-12T10:43:50Z  2016-06-01T00:00:00Z    0
110345      F  2016-05-16T12:30:58Z  2016-06-01T00:00:00Z    0
110346      M  2016-06-06T14:22:34Z  2016-06-08T00:00:00Z    0
110454      F  2016-06-03T15:18:44Z  2016-06-03T00:00:00Z    0
110460      F  2016-06-03T08:56:51Z  2016-06-03T00:00:00Z    0
110507      F  2016-06-08T09:04:18Z  2016-06-08T00:00:00Z    0

              Neighbourhood  Scholarship  Hipertension  Diabetes  Alcoholism  \
59                CONQUISTA            0             0         0           0
63              SÃO BENEDITO           0             0         0           0
64         ILHA DAS CAIEIRAS           0             0         0           0
```

65	CONQUISTA	0	0	0	0
67	NOVA PALESTINA	0	0	0	0
89	MONTE BELO	0	0	0	0
101	BONFIM	0	0	0	0
104	SANTO ANTÔNIO	0	0	0	0
132	PRAIA DO SUÁ	0	0	0	0
150	ITARARÉ	0	0	0	0
188	NOVA PALESTINA	0	0	0	0
192	CONQUISTA	0	0	0	0
193	NOVA PALESTINA	0	0	0	0
194	REDENÇÃO	0	0	0	0
250	CARATOÍRA	0	0	0	0
256	ARIOVALDO FAVALESSA	0	0	0	0
266	MARIA ORTIZ	0	0	0	0
292	GURIGICA	0	0	0	0
305	JOANA D'ARC	0	0	0	0
306	SANTA MARTHA	0	0	0	0
310	SANTA MARTHA	0	0	0	0
358	CONSOLAÇÃO	0	0	0	0
359	CONSOLAÇÃO	0	0	0	0
366	CONSOLAÇÃO	0	0	0	0
377	DA PENHA	0	0	0	0
434	MORADA DE CAMBURI	0	0	0	0
524	ITARARÉ	0	0	0	0
525	ITARARÉ	0	0	0	0
526	ITARARÉ	0	0	0	0
565	SANTA LUÍZA	0	0	0	0
...	...	...	...	...	...
109629	ROMÃO	0	0	0	0
109633	ROMÃO	0	0	0	0
109646	FORTE SÃO JOÃO	0	0	0	0
109647	FORTE SÃO JOÃO	0	0	0	0
109649	FORTE SÃO JOÃO	0	0	0	0
109650	FORTE SÃO JOÃO	0	0	0	0
109830	SÃO BENEDITO	0	0	0	0
109847	NOVA PALESTINA	0	0	0	0
109848	RESISTÊNCIA	0	0	0	0
109852	RESISTÊNCIA	0	0	0	0
110231	RESISTÊNCIA	0	0	0	0
110235	RESISTÊNCIA	0	0	0	0
110236	RESISTÊNCIA	0	0	0	0
110299	RESISTÊNCIA	0	0	0	0
110313	RESISTÊNCIA	0	0	0	0
110319	RESISTÊNCIA	0	0	0	0
110320	RESISTÊNCIA	0	0	0	0
110321	RESISTÊNCIA	0	0	0	0
110331	RESISTÊNCIA	0	0	0	0
110334	RESISTÊNCIA	0	0	0	0

110335	RESISTÊNCIA	0	0	0	0
110339	RESISTÊNCIA	0	0	0	0
110341	RESISTÊNCIA	0	0	0	0
110342	RESISTÊNCIA	0	0	0	0
110343	RESISTÊNCIA	0	0	0	0
110345	RESISTÊNCIA	0	0	0	0
110346	RESISTÊNCIA	0	0	0	0
110454	RESISTÊNCIA	0	0	0	0
110460	RESISTÊNCIA	0	0	0	0
110507	MARIA ORTIZ	0	0	0	0

	Handcap	SMS_received	No-show
59	0	0	No
63	0	0	No
64	0	1	No
65	0	0	No
67	0	0	No
89	0	0	No
101	0	0	No
104	0	0	Yes
132	0	1	Yes
150	0	0	No
188	0	0	No
192	0	0	No
193	0	1	No
194	0	0	No
250	0	0	Yes
256	0	0	Yes
266	0	0	No
292	0	0	No
305	0	0	No
306	0	0	No
310	0	0	No
358	0	0	Yes
359	0	0	No
366	0	1	Yes
377	0	1	No
434	0	1	Yes
524	0	1	No
525	0	1	Yes
526	0	0	No
565	0	1	No
...	...	...	...
109629	0	0	Yes
109633	0	1	Yes
109646	0	0	No
109647	0	0	Yes
109649	0	0	No

```
109650        0             0     Yes
109830        0             0     Yes
109847        0             1     Yes
109848        0             0      No
109852        0             0      No
110231        0             0      No
110235        0             0     Yes
110236        0             1      No
110299        0             1     Yes
110313        0             1      No
110319        0             1      No
110320        0             1      No
110321        0             1      No
110331        0             1     Yes
110334        0             0      No
110335        0             1      No
110339        0             1     Yes
110341        0             1      No
110342        0             0      No
110343        0             1      No
110345        0             0      No
110346        0             0      No
110454        0             0      No
110460        0             0      No
110507        0             0      No

[3540 rows x 12 columns]
```

1.2.1 We found 3,540 patients with zero or negative ages. We need to fix that by replacing those numbers with 1 to make more sense.

```
In [62]: df.loc[df['Age'] <= 0, 'Age'] = 1
```

We found some Values in "Handcap" coulmns which were greater than 1. We will change those values to 1 to make more sense.

```
In [85]: df[df["Handcap"] > 1]
```

```
Out[85]:        Gender          ScheduledDay        AppointmentDay  Age  \
          946         M  2016-04-14T09:26:08Z  2016-04-29T00:00:00Z   94
          1665        M  2016-03-30T09:16:41Z  2016-04-29T00:00:00Z   64
          1666        M  2016-03-30T09:16:41Z  2016-04-29T00:00:00Z   64
          2071        M  2016-04-29T10:08:48Z  2016-04-29T00:00:00Z   64
          2091        F  2016-04-29T08:13:59Z  2016-04-29T00:00:00Z   11
          2213        F  2016-04-29T11:22:50Z  2016-04-29T00:00:00Z   29
          2214        M  2016-04-29T11:22:20Z  2016-04-29T00:00:00Z   55
          2673        M  2016-04-15T13:06:05Z  2016-04-29T00:00:00Z   17
          5424        M  2016-05-02T13:24:36Z  2016-05-04T00:00:00Z   65
```

5467	F	2016-05-12T09:10:28Z	2016-05-16T00:00:00Z	10
5475	F	2016-05-12T09:10:50Z	2016-05-16T00:00:00Z	34
5485	F	2016-05-05T10:23:26Z	2016-05-09T00:00:00Z	16
5510	F	2016-05-04T11:06:48Z	2016-05-05T00:00:00Z	42
6067	F	2016-05-17T07:41:55Z	2016-05-19T00:00:00Z	10
6156	F	2016-05-13T07:39:38Z	2016-05-17T00:00:00Z	29
6401	F	2016-05-11T07:32:16Z	2016-05-13T00:00:00Z	19
11227	M	2016-04-15T08:49:43Z	2016-05-19T00:00:00Z	18
11230	M	2016-03-18T15:02:00Z	2016-05-19T00:00:00Z	18
14847	M	2016-04-05T09:04:49Z	2016-05-05T00:00:00Z	17
15839	F	2016-05-02T09:46:27Z	2016-05-12T00:00:00Z	52
15845	F	2016-05-02T09:46:40Z	2016-05-12T00:00:00Z	52
16632	F	2016-05-02T08:04:06Z	2016-05-09T00:00:00Z	32
16634	F	2016-04-27T12:03:16Z	2016-05-02T00:00:00Z	32
16635	F	2016-05-02T08:04:06Z	2016-05-09T00:00:00Z	32
16636	F	2016-04-27T12:01:01Z	2016-05-02T00:00:00Z	32
17862	F	2016-05-06T07:59:24Z	2016-05-06T00:00:00Z	37
18241	M	2016-04-28T11:58:14Z	2016-05-04T00:00:00Z	4
19264	F	2016-04-26T09:39:37Z	2016-05-10T00:00:00Z	89
19915	M	2016-04-27T08:25:25Z	2016-05-11T00:00:00Z	16
19981	M	2016-05-09T15:27:02Z	2016-05-12T00:00:00Z	89
...	...	...	...	...
96173	M	2016-05-10T10:57:48Z	2016-06-07T00:00:00Z	82
96457	M	2016-06-06T15:02:34Z	2016-06-07T00:00:00Z	29
96831	F	2016-05-19T14:08:28Z	2016-06-02T00:00:00Z	97
97998	F	2016-06-02T09:21:08Z	2016-06-02T00:00:00Z	8
98045	F	2016-06-07T14:56:08Z	2016-06-07T00:00:00Z	52
98538	M	2016-06-01T15:11:25Z	2016-06-03T00:00:00Z	19
99455	F	2016-06-02T11:06:13Z	2016-06-02T00:00:00Z	40
100538	M	2016-06-07T09:27:29Z	2016-06-07T00:00:00Z	11
101074	F	2016-05-09T09:00:57Z	2016-06-07T00:00:00Z	33
101911	M	2016-06-07T11:10:08Z	2016-06-07T00:00:00Z	84
101912	M	2016-06-07T11:09:45Z	2016-06-07T00:00:00Z	84
101913	M	2016-06-07T11:10:22Z	2016-06-07T00:00:00Z	84
101949	M	2016-06-06T15:14:05Z	2016-06-08T00:00:00Z	12
102276	M	2016-06-06T14:02:28Z	2016-06-06T00:00:00Z	89
102360	F	2016-06-01T09:00:14Z	2016-06-01T00:00:00Z	44
102955	F	2016-06-06T10:45:58Z	2016-06-06T00:00:00Z	15
104268	F	2016-05-13T15:01:17Z	2016-06-02T00:00:00Z	9
104927	M	2016-05-24T16:35:44Z	2016-06-01T00:00:00Z	70
104931	M	2016-05-24T16:35:44Z	2016-06-01T00:00:00Z	70
104932	M	2016-05-24T16:35:44Z	2016-06-01T00:00:00Z	70
105008	F	2016-06-07T09:03:12Z	2016-06-07T00:00:00Z	37
105073	F	2016-05-16T07:20:46Z	2016-06-01T00:00:00Z	35
105322	M	2016-06-01T07:23:50Z	2016-06-01T00:00:00Z	6
105753	M	2016-05-31T14:41:50Z	2016-06-02T00:00:00Z	32
108335	M	2016-06-03T09:11:46Z	2016-06-08T00:00:00Z	45
108376	F	2016-06-01T08:48:28Z	2016-06-07T00:00:00Z	44

```
109484      M  2016-05-31T11:45:57Z  2016-06-02T00:00:00Z   64
109733      F  2016-06-03T16:11:00Z  2016-06-07T00:00:00Z   34
109975      M  2016-06-02T16:07:36Z  2016-06-06T00:00:00Z   39
110107      F  2016-06-02T06:44:00Z  2016-06-06T00:00:00Z   44

          Neighbourhood  Scholarship  Hipertension  Diabetes  Alcoholism  \
946          BELA VISTA            0             1         1           0
1665       SANTA MARTHA            0             1         0           1
1666       SANTA MARTHA            0             1         0           1
2071       SANTA MARTHA            0             1         0           1
2091         ANDORINHAS            0             0         0           0
2213         VILA RUBIM            0             0         0           0
2214          DO QUADRO            0             0         0           0
2673       SANTA TEREZA            0             0         0           0
5424        SANTO ANDRÉ            0             1         1           1
5467     BENTO FERREIRA            0             0         0           0
5475        JUCUTUQUARA            0             0         0           0
5485       SANTA TEREZA            0             0         0           0
5510          SÃO PEDRO            0             1         0           0
6067        RESISTÊNCIA            0             0         0           0
6156      SANTO ANTÔNIO            0             0         0           0
6401         VILA RUBIM            0             0         0           0
11227       RESISTÊNCIA            0             0         0           0
11230       RESISTÊNCIA            0             0         0           0
14847      SANTA TEREZA            0             0         0           0
15839        DE LOURDES            0             1         0           0
15845        DE LOURDES            0             1         0           0
16632      SANTA MARTHA            0             0         0           0
16634      SANTA MARTHA            0             0         0           0
16635      SANTA MARTHA            0             0         0           0
16636      SANTA MARTHA            0             0         0           0
17862       SANTO ANDRÉ            0             0         0           0
18241      PRAIA DO SUÁ            0             0         0           0
19264  ILHA DO PRÍNCIPE            0             1         0           0
19915        BELA VISTA            0             0         0           0
19981        BELA VISTA            0             1         1           0
...                 ...          ...           ...       ...         ...
96173     MATA DA PRAIA            0             1         0           0
96457   JARDIM DA PENHA            0             0         0           0
96831           MARUÍPE            0             1         0           0
97998        TABUAZEIRO            0             0         0           0
98045        DE LOURDES            0             1         0           0
98538         SÃO PEDRO            0             0         0           0
99455          DO QUADRO            0             0         0           0
100538       GOIABEIRAS            1             0         0           0
101074     SANTA MARTHA            0             0         0           0
101911      RESISTÊNCIA            0             1         0           0
101912      RESISTÊNCIA            0             1         0           0
```

101913	RESISTÊNCIA	0	1	0	0
101949	ESTRELINHA	0	0	0	0
102276	BELA VISTA	0	1	1	0
102360	ROMÃO	0	1	1	0
102955	JOANA D'ARC	1	0	0	0
104268	ITARARÉ	0	0	0	0
104927	ANDORINHAS	0	0	0	1
104931	ANDORINHAS	0	0	0	1
104932	ANDORINHAS	0	0	0	1
105008	SANTO ANDRÉ	0	0	0	0
105073	BELA VISTA	1	0	1	0
105322	SANTA CLARA	0	0	0	0
105753	REDENÇÃO	0	0	0	0
108335	ROMÃO	0	0	0	0
108376	ROMÃO	0	1	1	0
109484	DA PENHA	0	1	1	0
109733	JUCUTUQUARA	0	0	0	0
109975	PRAIA DO SUÁ	1	0	0	0
110107	RESISTÊNCIA	0	0	0	0

	Handcap	SMS_received	No-show
946	2	1	No
1665	2	1	No
1666	2	0	No
2071	2	0	No
2091	2	0	No
2213	2	0	No
2214	3	0	No
2673	2	1	No
5424	2	0	Yes
5467	2	0	No
5475	2	0	Yes
5485	2	0	No
5510	2	0	No
6067	2	0	Yes
6156	2	0	No
6401	2	0	No
11227	2	0	No
11230	2	0	No
14847	2	1	No
15839	2	1	No
15845	2	0	No
16632	2	0	No
16634	2	0	Yes
16635	2	0	No
16636	2	0	Yes
17862	2	0	No
18241	2	1	No

```
19264          2              1      Yes
19915          2              1       No
19981          2              0       No
...          ...            ...      ...
96173          2              0       No
96457          2              0       No
96831          2              0       No
97998          2              0       No
98045          2              0       No
98538          4              0       No
99455          2              0       No
100538         2              0       No
101074         2              0       No
101911         2              0      Yes
101912         2              0       No
101913         2              0      Yes
101949         2              0      Yes
102276         2              0       No
102360         2              0       No
102955         2              0       No
104268         4              1      Yes
104927         3              0       No
104931         3              0       No
104932         3              0       No
105008         2              0       No
105073         2              1       No
105322         2              0       No
105753         2              0       No
108335         2              1       No
108376         2              1       No
109484         2              0       No
109733         2              1       No
109975         2              1       No
110107         2              1       No

[199 rows x 12 columns]
```

```
In [91]: df.loc[df['Handcap'] > 1, 'Handcap'] = 1
```

1.2.2 To deal better with "No-show" column we need to convert its values from "yes" and "no" to "1" and "0".

```
In [92]: dummy = pd.get_dummies(df["No-show"])
         dummy.head()
```

```
Out[92]:    No  Yes
         0   1    0
         1   1    0
```

```
2   1    0
3   1    0
4   1    0
```

In [93]:
```
df2 = pd.concat((df,dummy), axis = 1)
df2.head()
```

Out[93]:

	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension	Diabetes	Alcoholism	Handcap	SMS_received	No-show	No	Yes
0	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1	0	0	0	0	No	1	0
1	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0	0	0	0	0	No	1	0
2	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0	0	0	0	0	No	1	0
3	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0	0	0	0	0	No	1	0
4	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	1	1	0	0	0	No	1	0

In [94]:
```
df2.rename(columns = {"No": "show_numeric"} , inplace = True )
df2.head()
```

Out[94]:

	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension	Diabetes	Alcoholism	Handcap	SMS_received	No-show	show_numeric	Yes
0	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1	0	0	0	0	No	1	0
1	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0	0	0	0	0	No	1	0
2	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0	0	0	0	0	No	1	0
3	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0	0	0	0	0	No	1	0
4	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	1	1	0	0	0	No	1	0

```
         3      No             1    0
         4      No             1    0
```

```
In [95]: df2.drop(["Yes"], axis =1, inplace = True)
```

```
In [96]: df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 110527 entries, 0 to 110526
Data columns (total 13 columns):
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
show_numeric      110527 non-null uint8
dtypes: int64(7), object(5), uint8(1)
memory usage: 10.2+ MB
```

```
In [97]: df2.head()
```

Out[97]:

	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	\
0	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	
1	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	
2	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	
3	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	
4	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	

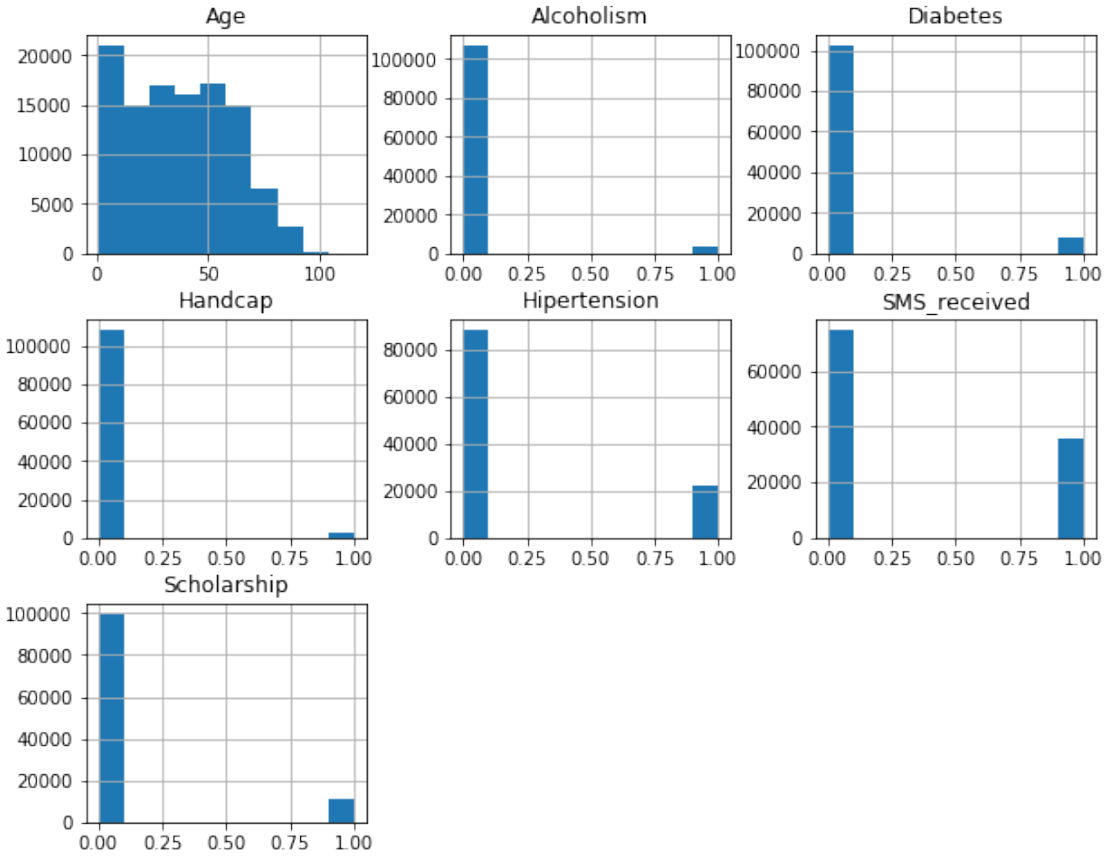
	Scholarship	Hipertension	Diabetes	Alcoholism	Handcap	SMS_received	\
0	0	1	0	0	0	0	
1	0	0	0	0	0	0	
2	0	0	0	0	0	0	
3	0	0	0	0	0	0	
4	0	1	1	0	0	0	

	No-show	show_numeric
0	No	1
1	No	1
2	No	1
3	No	1
4	No	1

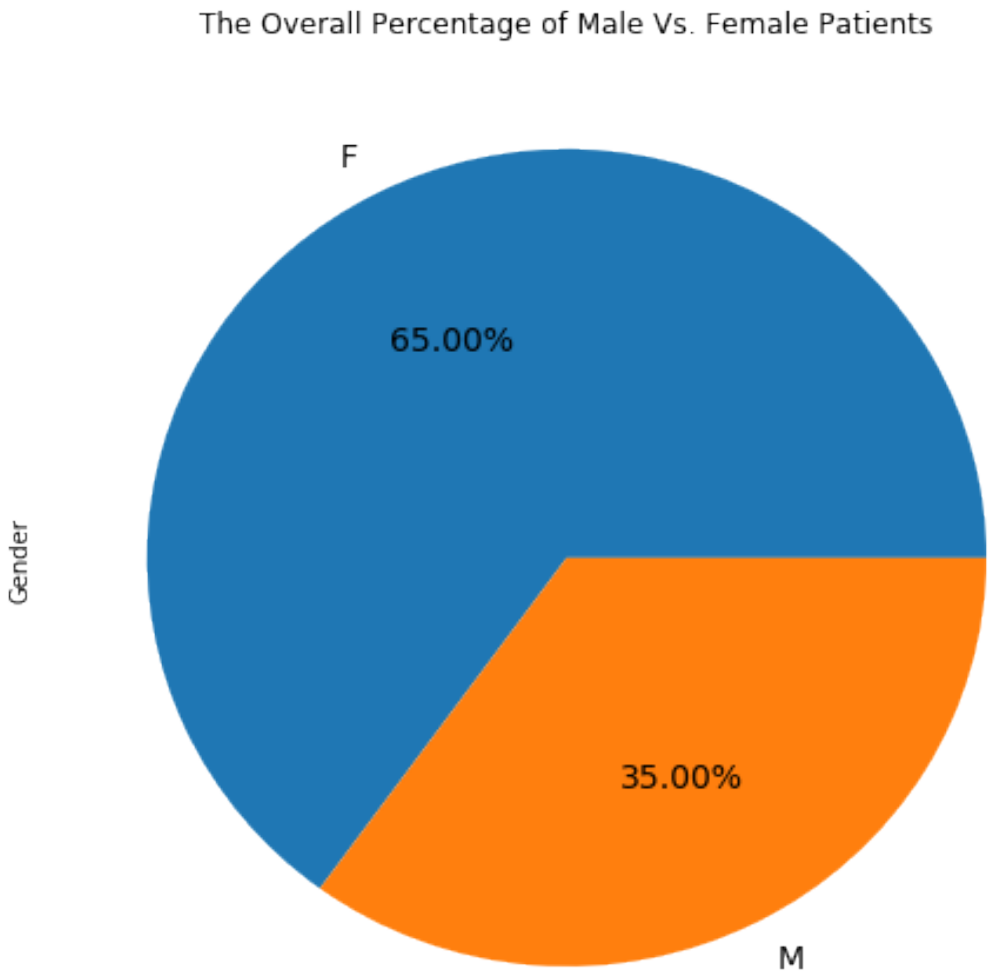
Exploratory Data Analysis

Now, we will look deeper into the data with visual plots to gain insights about the different types of patients data gathered and their percentages. Then we will take a look at the correlation between each parameter and no show rate.

```
In [98]: # Taking a general look at the histogram of the whole data.
         df.hist(figsize = (10,8));
```

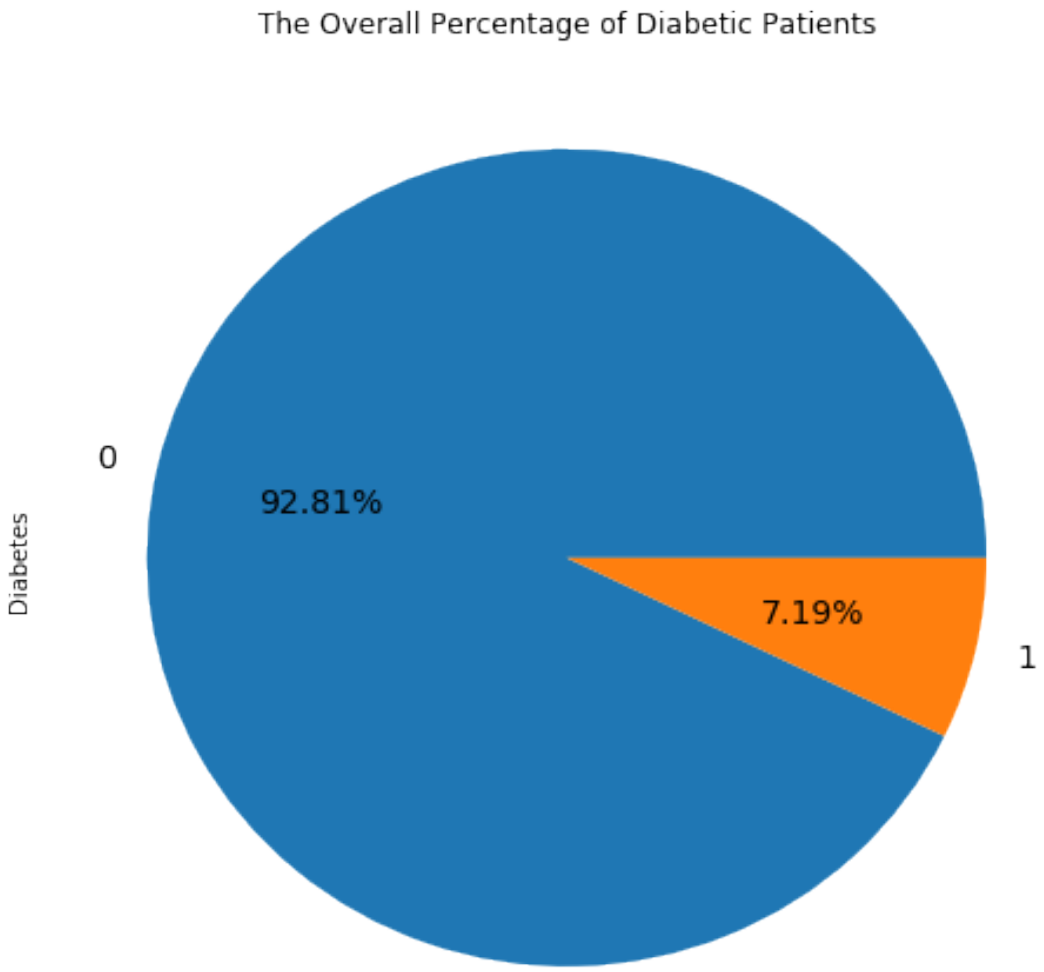



```
In [157]: df2.Gender.value_counts().plot(kind= 'pie', autopct='%.2f%%', figsize=(8,8), fontsize
```



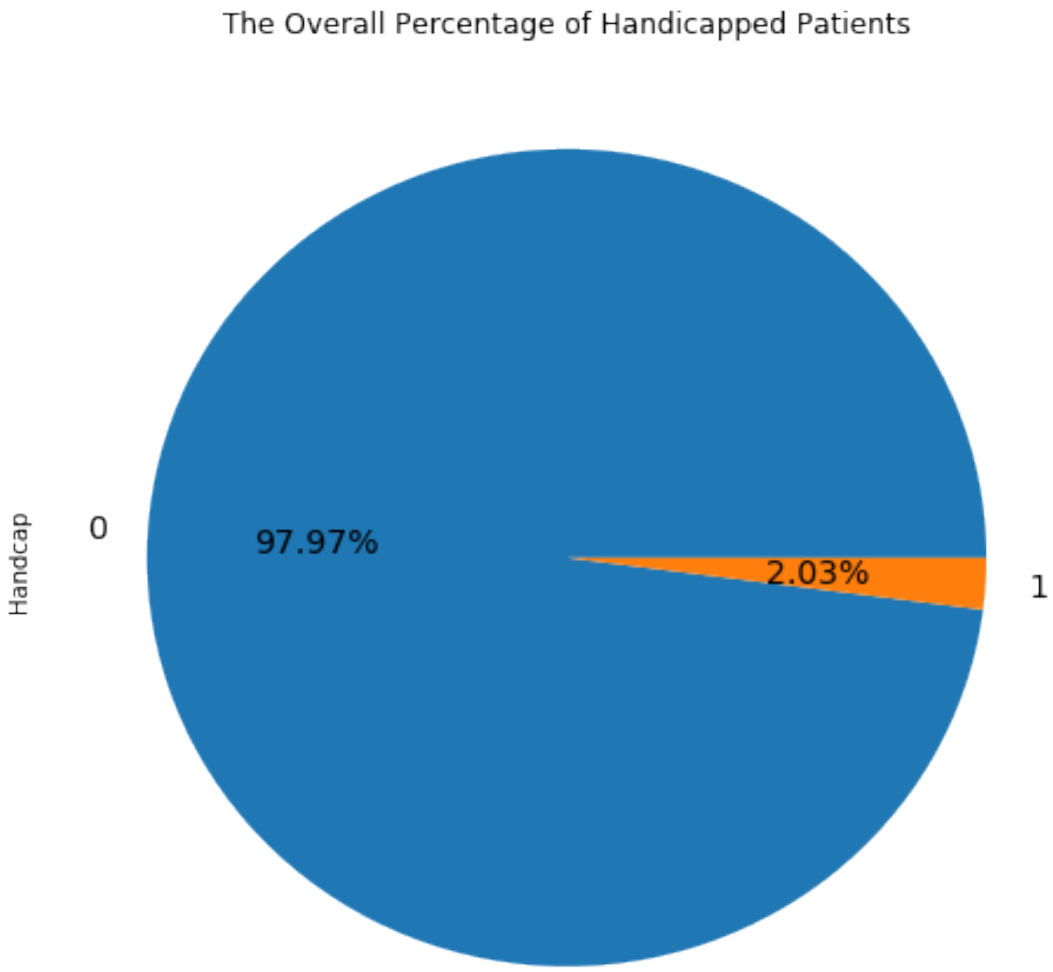
Generally, there are more female patients in the gathered dataset.

```
In [158]: df2.Diabetes.value_counts().plot(kind= 'pie', autopct='%.2f%%', figsize=(8,8), fontsiz
```



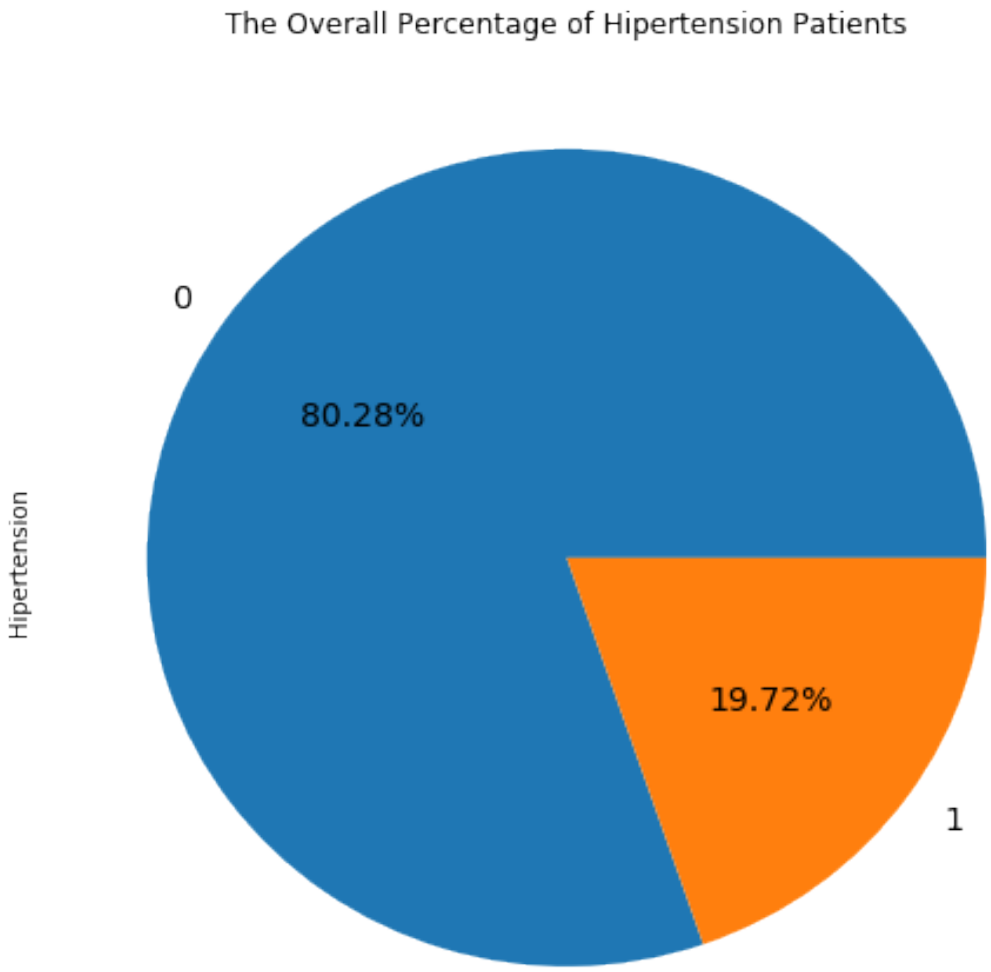
Only 7.19% of the patients are diabetic.

```
In [159]: df2.Handcap.value_counts().plot(kind= 'pie', autopct='%.2f%%', figsize=(8,8), fontsize
```



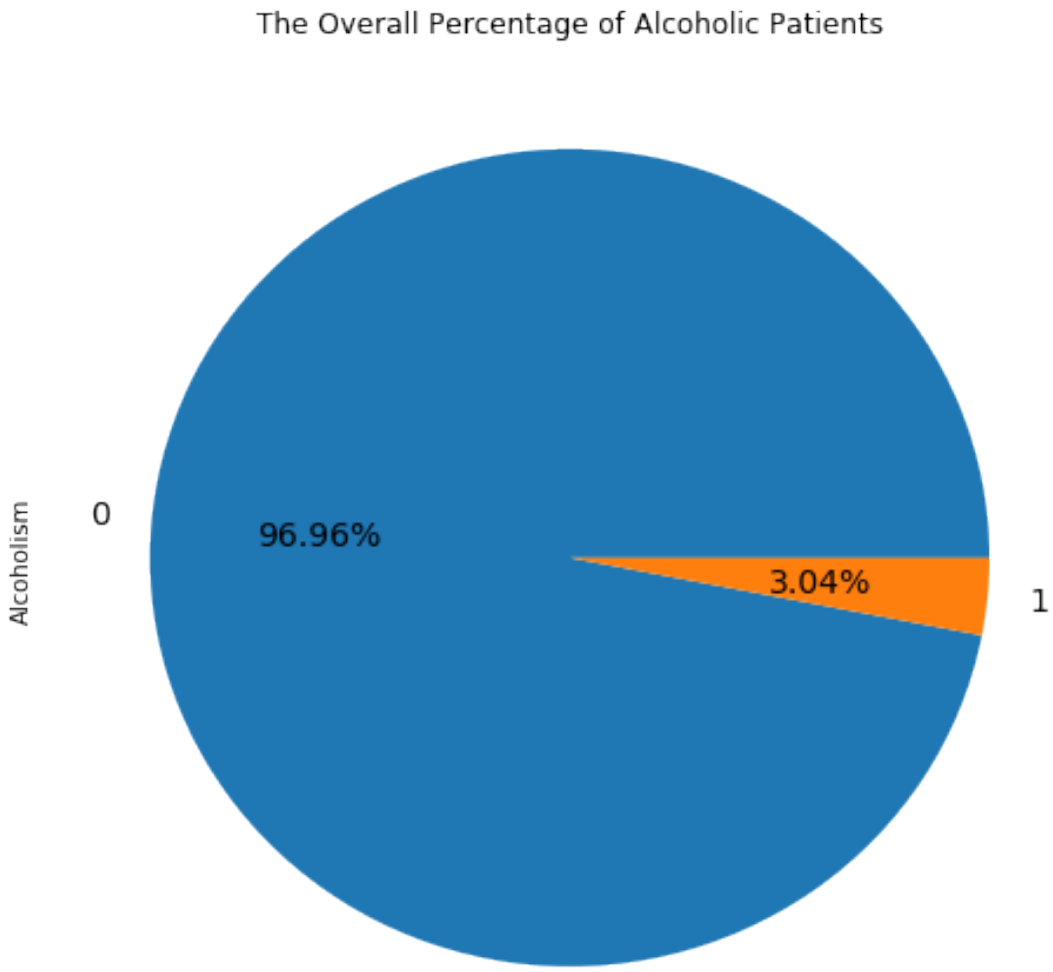
Only 2% of patients are Handicapped.

```
In [160]: df2.Hipertension.value_counts().plot(kind= 'pie', autopct='%.2f%%', figsize=(8,8), fon
```



About 19.7% of patients suffer from Hypertension.

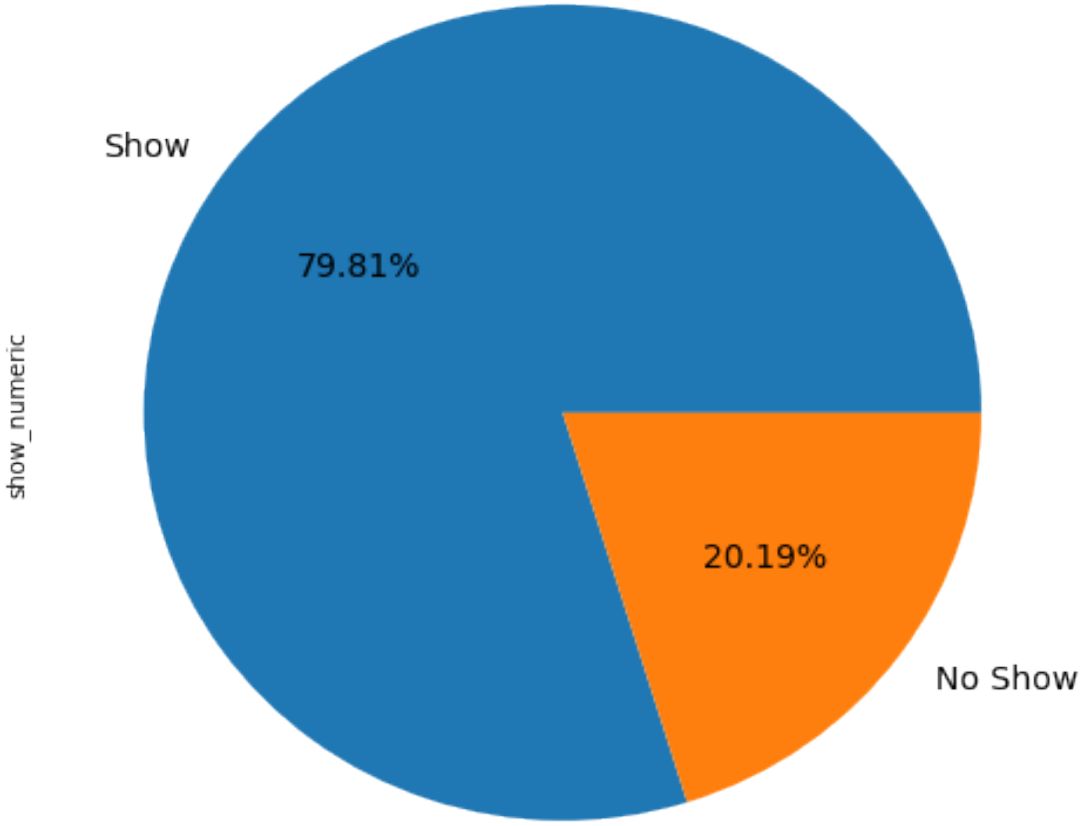
```
In [161]: df2.Alcoholism.value_counts().plot(kind= 'pie', autopct='%.2f%%', figsize=(8,8), fonts
```



And Finally, only 3% are alcoholic.

1.2.3 The Overall Show Vs No-Show Percentage.

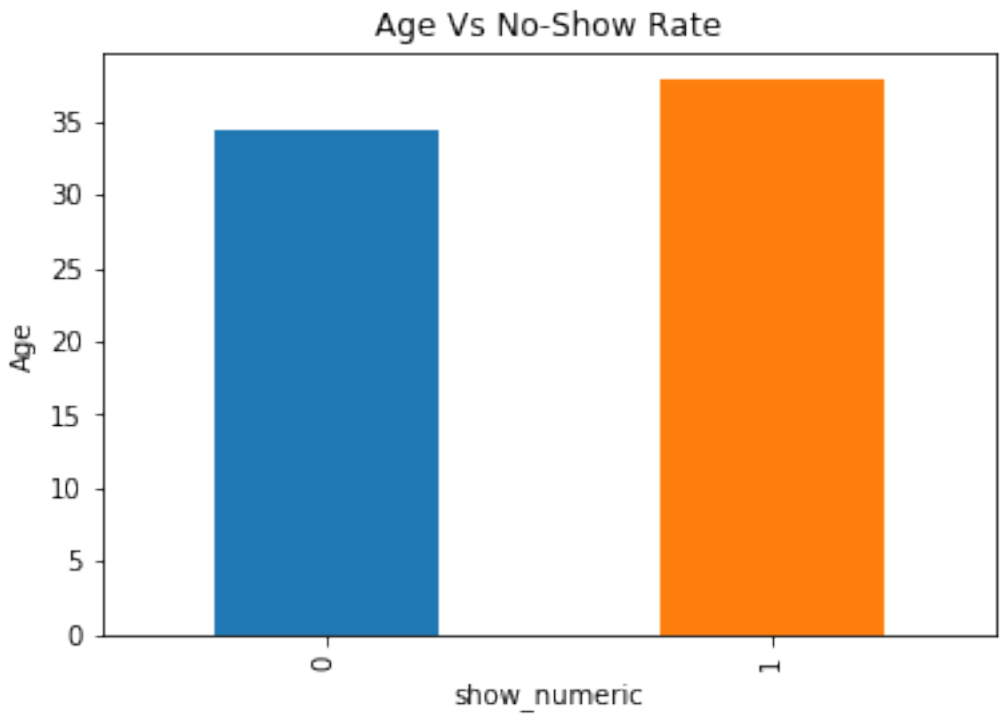
```
In [162]: df2.show_numeric.value_counts().plot(kind= 'pie', autopct='%.2f%%', labels = ["Show",
```



About 20% of patients don't show up at their scheduled appointments.

1.2.4 The Correlation between Age and No-Show rate.

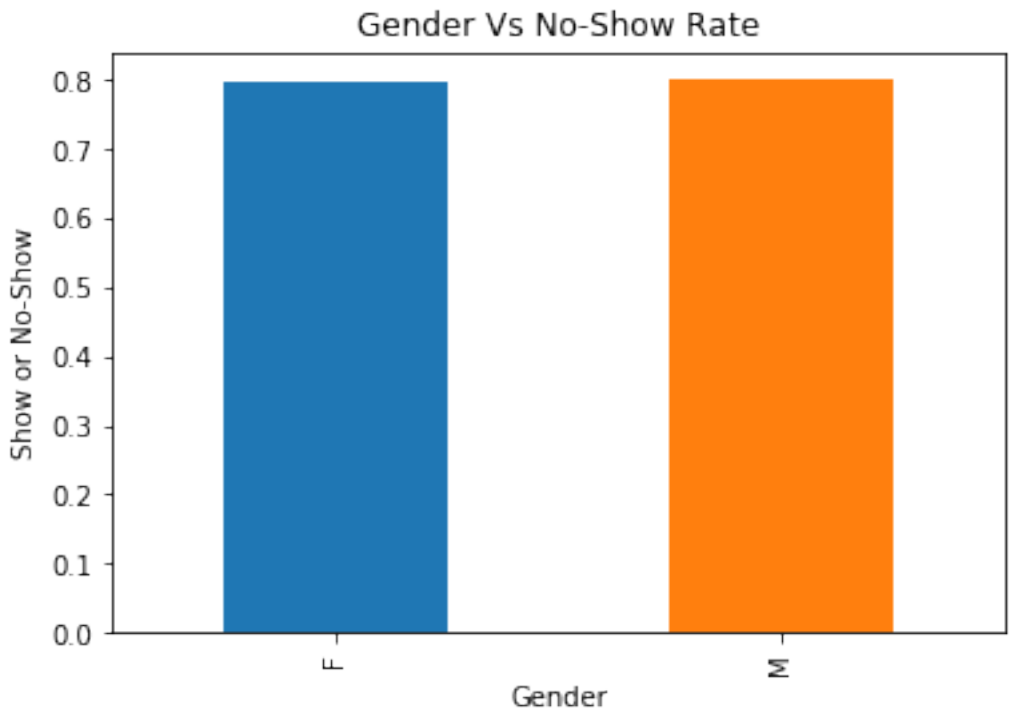
```
In [164]: df2.groupby("show_numeric")["Age"].mean().plot(kind = "bar", title = "Age Vs No-Show R
          plt.ylabel("Age");
```



Obviously, the average age in both show and no show cases is between 35 and 40. which means that age has no significant effect on commitment to scheduled appointments.

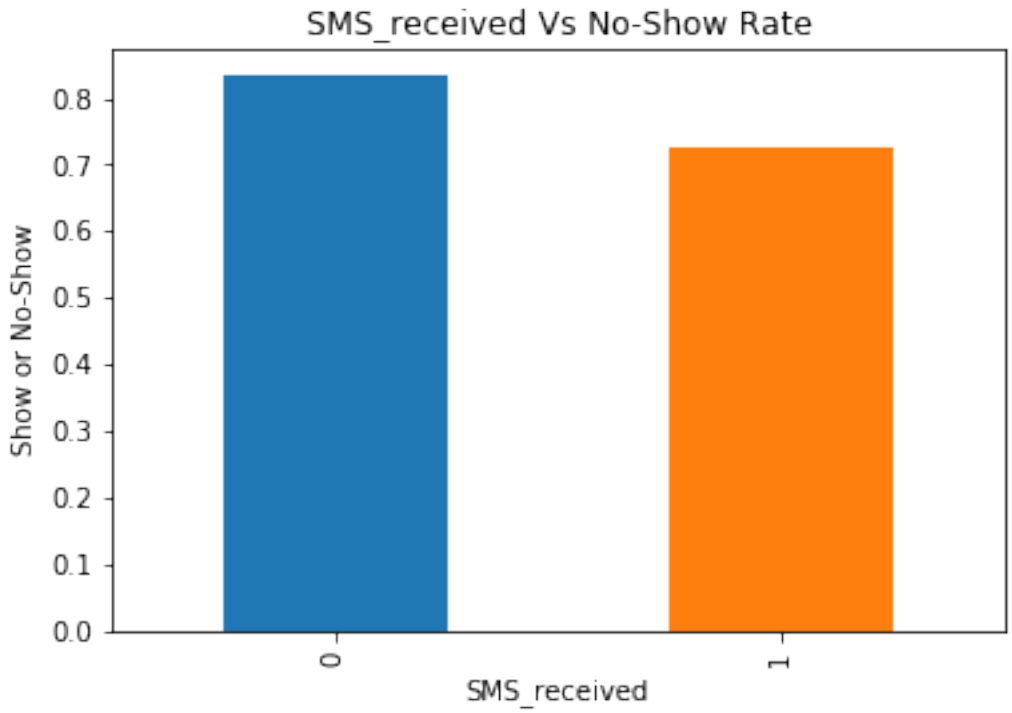
1.2.5 The Correlation between Gender and No-Show rate.

```
In [165]: df2.groupby("Gender")["show_numeric"].mean().plot(kind = "bar", title = "Gender Vs No-
          plt.ylabel("Show or No-Show");
```



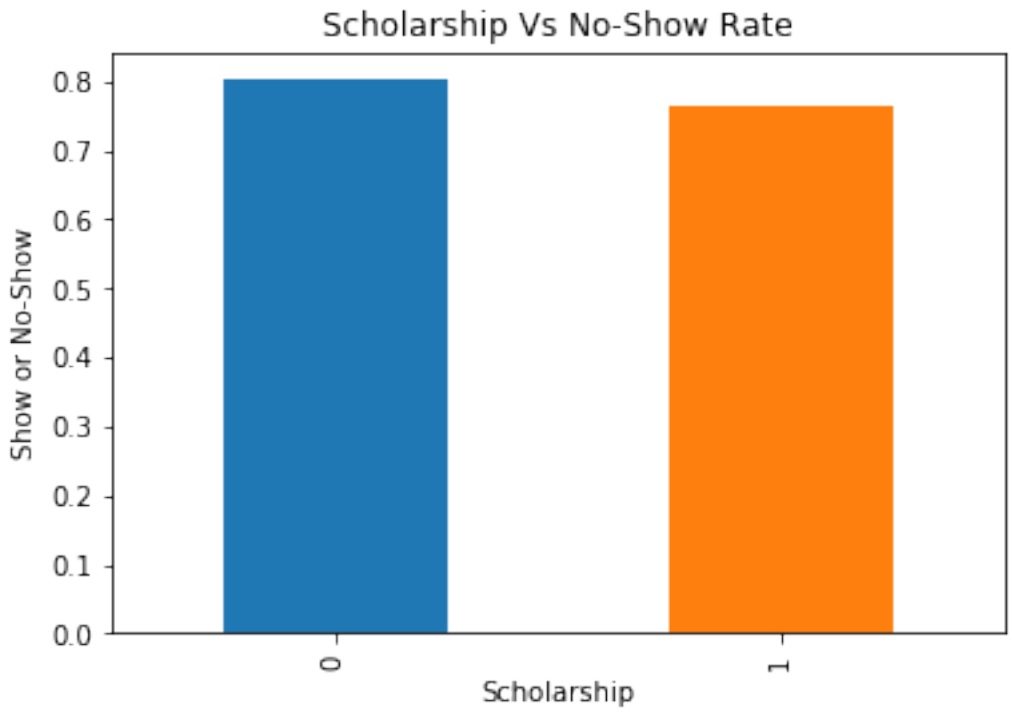
Also patient's gender doesn't seem to have any effect on commitment to scheduled appointments.

```
In [167]: df2.groupby("SMS_received")["show_numeric"].mean().plot(kind = "bar", title = "SMS_rec
          plt.ylabel("Show or No-Show");
```



Here, we would suspect that receiving an SMS should have a higher impact on no show rates. On the contrary, a higher percentage of people who did not receive an SMS showed up at their scheduled appointments more than those who received an SMS.

```
In [168]: df2.groupby("Scholarship")["show_numeric"].mean().plot(kind = "bar", title = "Scholars
          plt.ylabel("Show or No-Show");
```

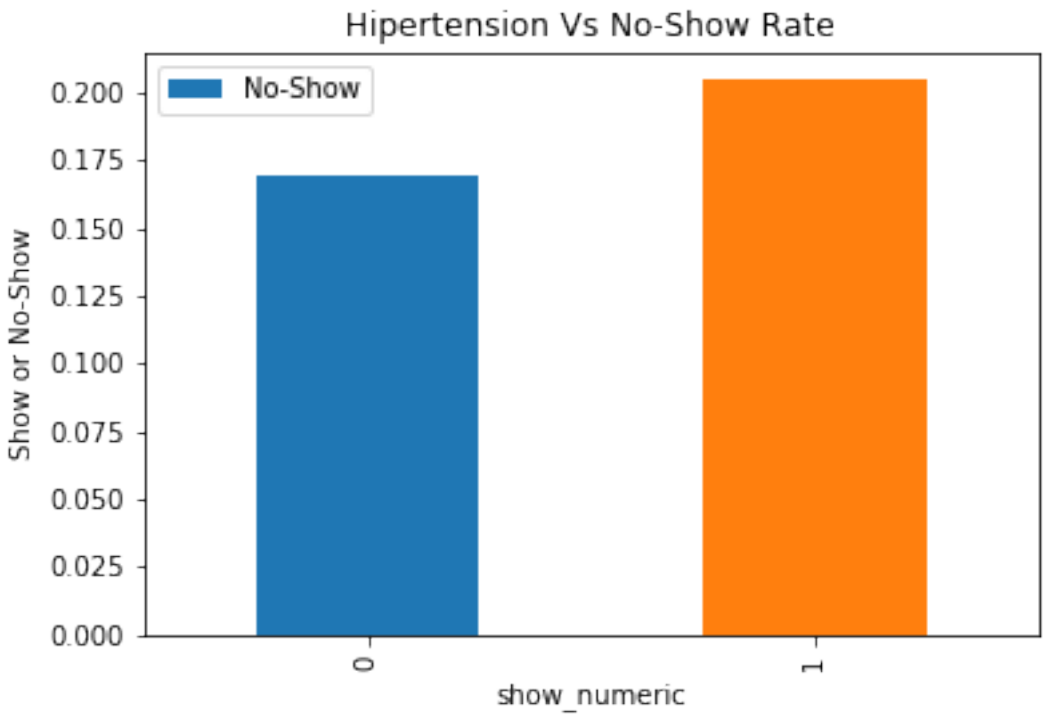


Also, we would suspect that having a Scholarship should have a higher impact on no show rates. On the contrary, a higher percentage of people who do not have a Scholarship showed up at their scheduled appointments more than those who don't.

1.2.6 Now, we will investigate the correlation between having a chronic condition and no-show rate.

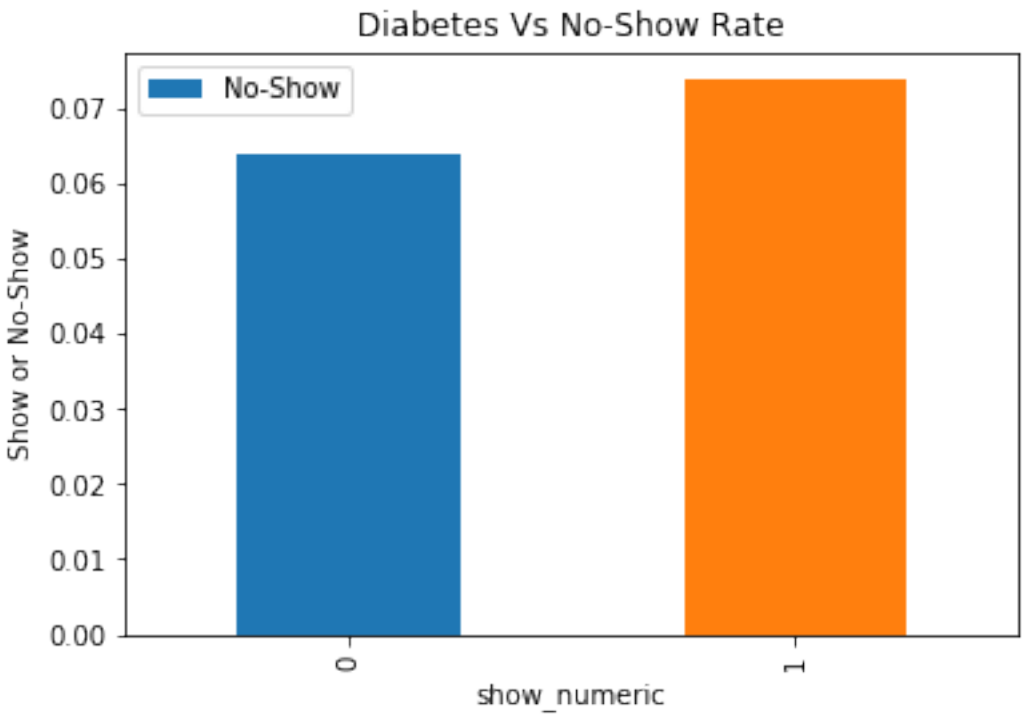
```
In [202]: def chronic_correlation(df, chronic_disease, plot_title):
              df2.groupby("show_numeric")[chronic_disease].mean().plot(kind = "bar", title = plo
              plt.legend(["No-Show"])
              plt.ylabel("Show or No-Show")

In [203]: chronic_correlation(df2, "Hipertension", "Hipertension Vs No-Show Rate")
```



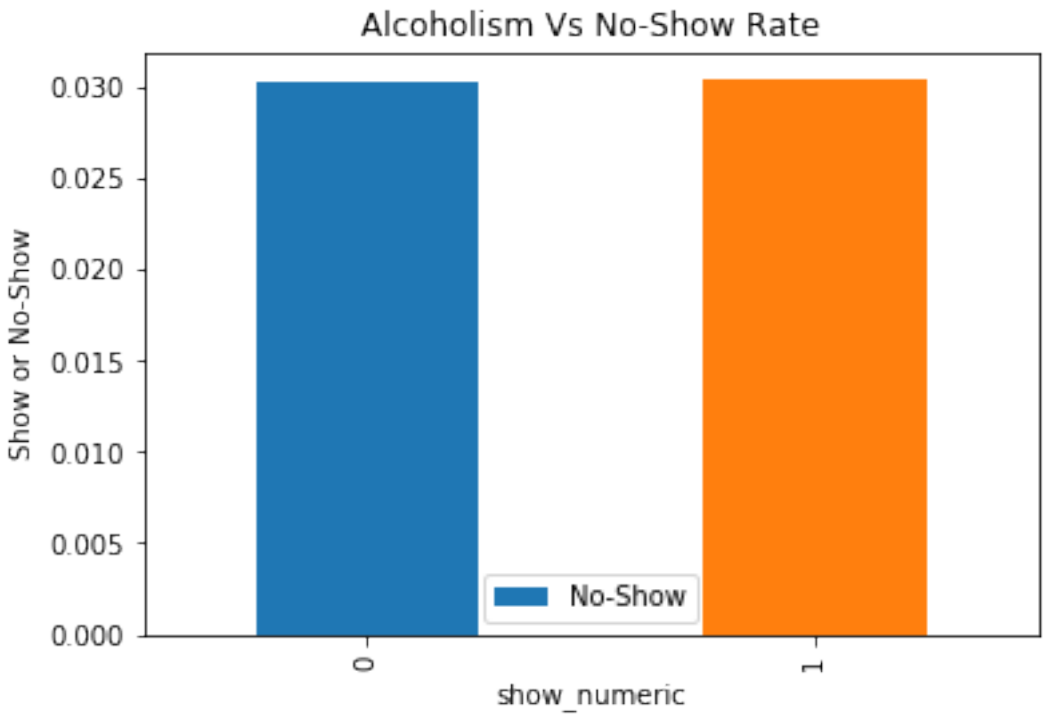
From this graph we can conclude that patients who suffer from Hipertension have a higher tendency to show up on their scheduled appointments.

```
In [206]: chronic_correlation(df2, "Diabetes", "Diabetes Vs No-Show Rate")
```



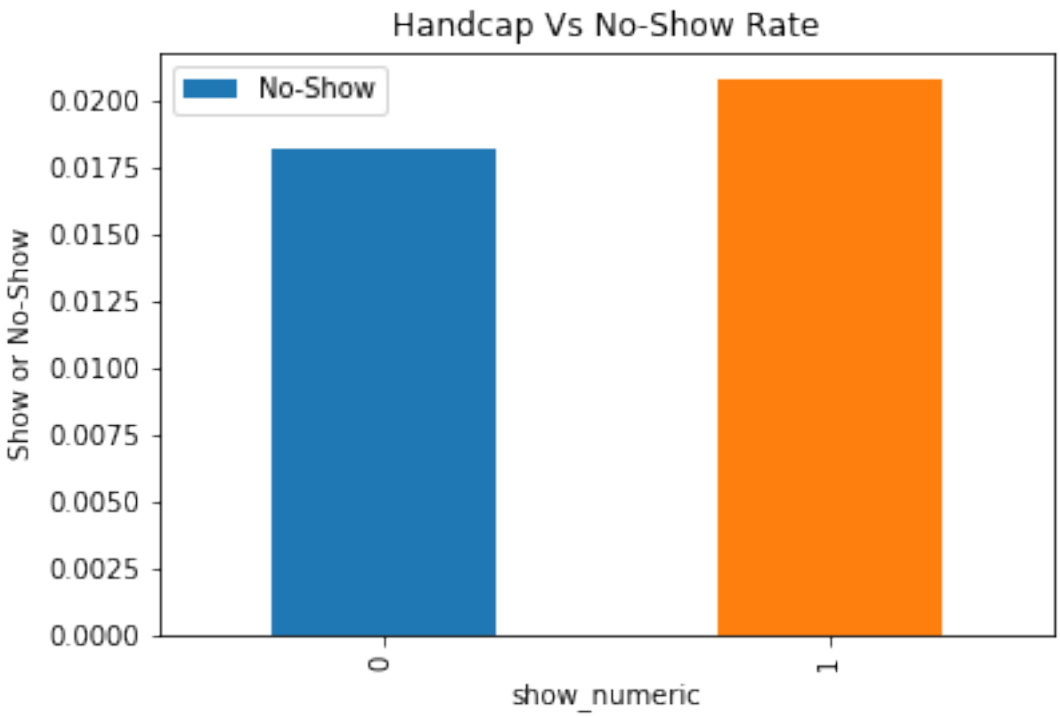
Also, from this graph we can conclude that patients who suffer from Diabetes have a higher tendency to show up on their scheduled appointments.

```
In [207]: chronic_correlation(df2, "Alcoholism", "Alcoholism Vs No-Show Rate")
```



From this graph we can see that alcoholism has no effect on commitment to scheduled appointments.

```
In [209]: chronic_correlation(df2, "Handcap", "Handcap Vs No-Show Rate")
```



Also, from this graph we can conclude that Handicapped patients have a higher tendency to show up on their scheduled appointments.

1.3

1.4 Conclusions

1.4.1 From the above study, we can conclude the following:

1- The overall percentage of show to no-show rate is 80% - 20%.

2- The overall number of female patients is more than male patients. However, gender has no impact on commitment.

3- There's about 7.19% diabetic patients, 3% Alcoholic, 2% Handicapped and 20% suffer from Hypertension.

4- Having a scholarship and receiving an SMS message as a reminder although seem encouraging but don't have a positive impact on commitment.

5- Suffering from chronic illness such as diabetes, handcap or hypertension results in more commitment to scheduled appointments. Howevere, being Alcoholic has no impact at all.

1.4.2 Limitations:

1- The analysis does not state or imply that one change causes another based solely on a correlation.

2- Age Column had -1 and 0 values which I rounded up to be 1.

3- Handcap Column had 2,3 and 4 values which appeared while plotting the correlation between Handcap and no show rate. I replaced them all with 1 to make more sense.

4- The initial non numeric no-show values were hard to deal with inplotting, so I had to change them to numeric.

5- Many of the data was categorical which made it more challenging to be analyzed.

```
In [210]: from subprocess import call
          call(['python', '-m', 'nbconvert', 'Investigate_a_Dataset.ipynb'])
```

```
Out[210]: 0
```

```
In [ ]:
```