Investigate_a_Dataset

April 3, 2019

1 Inverstigating a dataset - No Show Appointments - kagglev2 , may/2016

1.0.1 Introductions :

Questions we are trying to answer :

- what is the overall appointment show-up vs. no show-up rate
- what are the most feature that matter the most of ('Age', 'being alcoholic','Having an SMS', 'Gender', 'Scholarship'), to make the patient make it to his appiontment?

1.0.2 importing the libraries that we gonna use in our inverstigation

```
In [22]: import numpy as np
         import pandas as pd
         import seaborn as sns
         import matplotlib.pyplot as plt
         % matplotlib inline
```

Data Wrangling

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

1.0.3 General Properties:

```
In [23]: # Load the data.
         # df = Data Frame.
         df = pd.read_csv('noshowappointments-kagglev2-may-2016.csv')
```

1.0.4 Show summary of the data:

```
In [24]: df.describe()
```

```
Out[24]:           PatientId  AppointmentID            Age    Scholarship  \
         count  1.105270e+05   1.105270e+05  110527.000000  110527.000000
         mean   1.474963e+14   5.675305e+06      37.088874       0.098266
```

```
std    2.560949e+14   7.129575e+04      23.110205       0.297675
min    3.921784e+04   5.030230e+06      -1.000000       0.000000
25%    4.172614e+12   5.640286e+06      18.000000       0.000000
50%    3.173184e+13   5.680573e+06      37.000000       0.000000
75%    9.439172e+13   5.725524e+06      55.000000       0.000000
max    9.999816e+14   5.790484e+06     115.000000       1.000000

        Hipertension       Diabetes     Alcoholism        Handcap  \
count  110527.000000  110527.000000  110527.000000  110527.000000
mean        0.197246       0.071865       0.030400       0.022248
std         0.397921       0.258265       0.171686       0.161543
min         0.000000       0.000000       0.000000       0.000000
25%         0.000000       0.000000       0.000000       0.000000
50%         0.000000       0.000000       0.000000       0.000000
75%         0.000000       0.000000       0.000000       0.000000
max         1.000000       1.000000       1.000000       4.000000

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

Check null values

```
In [25]: pd.isna(df).sum()

Out[25]: 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
```

1.0.5 _*There no null values_

1.0.6 Show some of the data:

In [26]:
```
df.head()
```

Out[26]:

	PatientId	AppointmentID	Gender	ScheduledDay	\
0	2.987250e+13	5642903	F	2016-04-29T18:38:08Z	
1	5.589978e+14	5642503	M	2016-04-29T16:08:27Z	
2	4.262962e+12	5642549	F	2016-04-29T16:19:04Z	
3	8.679512e+11	5642828	F	2016-04-29T17:29:31Z	
4	8.841186e+12	5642494	F	2016-04-29T16:07:23Z	

	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension	\
0	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1	
1	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0	
2	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0	
3	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0	
4	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	1	

	Diabetes	Alcoholism	Handcap	SMS_received	No-show
0	0	0	0	0	No
1	0	0	0	0	No
2	0	0	0	0	No
3	0	0	0	0	No
4	1	0	0	0	No

1.0.7 Looking for Data need to be cleaned

In [27]:
```
df.info()

        # check if dplicates in the data
        print("Num of dublicated : ", + sum(df.duplicated()))

        #check the age if it is '- number' or = 0
        df[df["Age"] <= 0]
```

```
<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
Num of dublicated :  0
```

Out[27]:

	PatientId	AppointmentID	Gender	ScheduledDay	\
59	7.184428e+13	5638545	F	2016-04-29T08:08:43Z	
63	2.366233e+14	5628286	M	2016-04-27T10:46:12Z	
64	1.885174e+14	5616082	M	2016-04-25T13:28:21Z	
65	2.718818e+14	5628321	M	2016-04-27T10:48:50Z	
67	8.647128e+13	5639264	F	2016-04-29T08:53:02Z	
89	4.314932e+14	5640380	M	2016-04-29T10:37:02Z	
101	6.822847e+13	5642813	M	2016-04-29T17:24:19Z	
104	3.326985e+12	5632495	F	2016-04-28T08:04:48Z	
132	8.468587e+14	5560742	M	2016-04-08T09:29:23Z	
150	5.392419e+10	5641476	M	2016-04-29T13:43:34Z	
188	1.883366e+14	5640933	M	2016-04-29T11:44:49Z	
192	1.165338e+12	5640561	M	2016-04-29T10:57:14Z	
193	2.379698e+11	5532452	F	2016-03-31T11:14:57Z	
194	4.442538e+11	5539566	M	2016-04-01T15:17:10Z	
250	2.942723e+14	5640181	M	2016-04-29T10:13:35Z	
256	8.599713e+10	5629906	F	2016-04-27T14:19:02Z	
266	1.198173e+13	5639237	M	2016-04-29T08:50:09Z	
292	4.645238e+13	5642766	F	2016-04-29T17:06:22Z	
305	9.794819e+11	5638226	F	2016-04-29T07:49:54Z	
306	2.648746e+11	5638370	M	2016-04-29T07:58:15Z	
310	3.732371e+14	5639056	M	2016-04-29T08:40:58Z	
358	7.349270e+14	5531831	F	2016-03-31T10:07:18Z	
359	7.975283e+11	5551583	M	2016-04-06T14:20:30Z	
366	3.499167e+14	5531933	F	2016-03-31T10:14:54Z	
377	9.123627e+11	5522585	M	2016-03-29T14:39:32Z	
434	9.625417e+14	5533985	F	2016-03-31T15:01:12Z	
524	5.815365e+13	5488633	M	2016-03-18T10:35:28Z	
525	4.239228e+13	5537936	F	2016-04-01T10:53:45Z	
526	8.574376e+12	5639659	M	2016-04-29T09:22:07Z	
565	5.569896e+14	5591488	F	2016-04-15T17:39:38Z	
...	...	...	...	...	
109629	3.689222e+13	5660428	F	2016-05-04T16:30:06Z	
109633	3.689222e+13	5660427	F	2016-05-04T16:30:06Z	
109646	5.941811e+13	5659194	M	2016-05-04T13:21:32Z	
109647	5.635258e+13	5691391	M	2016-05-12T12:35:04Z	
109649	5.941811e+13	5659193	M	2016-05-04T13:21:32Z	
109650	5.635258e+13	5691390	M	2016-05-12T12:35:04Z	

109830	6.931429e+11	5739239	F	2016-05-25T13:07:38Z
109847	9.174483e+11	5685977	M	2016-05-11T11:33:48Z
109848	5.853540e+12	5707361	M	2016-05-17T09:32:32Z
109852	1.663316e+13	5731578	F	2016-05-24T10:14:26Z
110231	1.731548e+14	5634706	M	2016-04-28T11:32:00Z
110235	3.279373e+14	5634274	M	2016-04-28T10:40:49Z
110236	1.174717e+13	5669996	M	2016-05-06T12:31:22Z
110299	5.944513e+10	5746323	F	2016-05-30T13:47:40Z
110313	8.535734e+12	5600209	F	2016-04-19T10:09:05Z
110319	4.524190e+11	5640301	F	2016-04-29T10:28:16Z
110320	3.459268e+13	5675103	M	2016-05-09T13:02:20Z
110321	1.388630e+11	5338643	M	2016-02-11T16:14:32Z
110331	2.739486e+13	5392089	F	2016-02-24T15:33:08Z
110334	1.174717e+13	5756343	M	2016-06-01T08:12:55Z
110335	3.747953e+11	5338761	M	2016-02-11T16:39:26Z
110339	3.731746e+12	5584042	M	2016-04-14T13:01:21Z
110341	4.696832e+12	5713894	M	2016-05-18T10:55:00Z
110342	3.147163e+11	5776749	M	2016-06-06T11:48:00Z
110343	1.759439e+12	5690822	F	2016-05-12T10:43:50Z
110345	1.473952e+14	5702537	F	2016-05-16T12:30:58Z
110346	5.577525e+12	5777724	M	2016-06-06T14:22:34Z
110454	6.142460e+11	5772400	F	2016-06-03T15:18:44Z
110460	4.321846e+13	5769545	F	2016-06-03T08:56:51Z
110507	4.769462e+14	5786918	F	2016-06-08T09:04:18Z

	AppointmentDay	Age	Neighbourhood	Scholarship	\
59	2016-04-29T00:00:00Z	0	CONQUISTA	0	
63	2016-04-29T00:00:00Z	0	SÃO BENEDITO	0	
64	2016-04-29T00:00:00Z	0	ILHA DAS CAIEIRAS	0	
65	2016-04-29T00:00:00Z	0	CONQUISTA	0	
67	2016-04-29T00:00:00Z	0	NOVA PALESTINA	0	
89	2016-04-29T00:00:00Z	0	MONTE BELO	0	
101	2016-04-29T00:00:00Z	0	BONFIM	0	
104	2016-04-29T00:00:00Z	0	SANTO ANTÔNIO	0	
132	2016-04-29T00:00:00Z	0	PRAIA DO SUÁ	0	
150	2016-04-29T00:00:00Z	0	ITARARÉ	0	
188	2016-04-29T00:00:00Z	0	NOVA PALESTINA	0	
192	2016-04-29T00:00:00Z	0	CONQUISTA	0	
193	2016-04-29T00:00:00Z	0	NOVA PALESTINA	0	
194	2016-04-29T00:00:00Z	0	REDENÇÃO	0	
250	2016-04-29T00:00:00Z	0	CARATOÍRA	0	
256	2016-04-29T00:00:00Z	0	ARIOVALDO FAVALESSA	0	
266	2016-04-29T00:00:00Z	0	MARIA ORTIZ	0	
292	2016-04-29T00:00:00Z	0	GURIGICA	0	
305	2016-04-29T00:00:00Z	0	JOANA D‘ARC	0	
306	2016-04-29T00:00:00Z	0	SANTA MARTHA	0	
310	2016-04-29T00:00:00Z	0	SANTA MARTHA	0	
358	2016-04-29T00:00:00Z	0	CONSOLAÇÃO	0	

359	2016-04-29T00:00:00Z	0	CONSOLAÇÃO	0		
366	2016-04-29T00:00:00Z	0	CONSOLAÇÃO	0		
377	2016-04-29T00:00:00Z	0	DA PENHA	0		
434	2016-04-29T00:00:00Z	0	MORADA DE CAMBURI	0		
524	2016-04-29T00:00:00Z	0	ITARARÉ	0		
525	2016-04-29T00:00:00Z	0	ITARARÉ	0		
526	2016-04-29T00:00:00Z	0	ITARARÉ	0		
565	2016-04-29T00:00:00Z	0	SANTA LUÍZA	0		
...	...	...	...	...		
109629	2016-06-01T00:00:00Z	0	ROMÃO	0		
109633	2016-06-01T00:00:00Z	0	ROMÃO	0		
109646	2016-06-01T00:00:00Z	0	FORTE SÃO JOÃO	0		
109647	2016-06-08T00:00:00Z	0	FORTE SÃO JOÃO	0		
109649	2016-06-01T00:00:00Z	0	FORTE SÃO JOÃO	0		
109650	2016-06-08T00:00:00Z	0	FORTE SÃO JOÃO	0		
109830	2016-06-02T00:00:00Z	0	SÃO BENEDITO	0		
109847	2016-06-02T00:00:00Z	0	NOVA PALESTINA	0		
109848	2016-06-02T00:00:00Z	0	RESISTÊNCIA	0		
109852	2016-06-02T00:00:00Z	0	RESISTÊNCIA	0		
110231	2016-06-01T00:00:00Z	0	RESISTÊNCIA	0		
110235	2016-06-01T00:00:00Z	0	RESISTÊNCIA	0		
110236	2016-06-08T00:00:00Z	0	RESISTÊNCIA	0		
110299	2016-06-07T00:00:00Z	0	RESISTÊNCIA	0		
110313	2016-06-06T00:00:00Z	0	RESISTÊNCIA	0		
110319	2016-06-06T00:00:00Z	0	RESISTÊNCIA	0		
110320	2016-06-06T00:00:00Z	0	RESISTÊNCIA	0		
110321	2016-06-01T00:00:00Z	0	RESISTÊNCIA	0		
110331	2016-06-01T00:00:00Z	0	RESISTÊNCIA	0		
110334	2016-06-01T00:00:00Z	0	RESISTÊNCIA	0		
110335	2016-06-01T00:00:00Z	0	RESISTÊNCIA	0		
110339	2016-06-01T00:00:00Z	0	RESISTÊNCIA	0		
110341	2016-06-01T00:00:00Z	0	RESISTÊNCIA	0		
110342	2016-06-08T00:00:00Z	0	RESISTÊNCIA	0		
110343	2016-06-01T00:00:00Z	0	RESISTÊNCIA	0		
110345	2016-06-01T00:00:00Z	0	RESISTÊNCIA	0		
110346	2016-06-08T00:00:00Z	0	RESISTÊNCIA	0		
110454	2016-06-03T00:00:00Z	0	RESISTÊNCIA	0		
110460	2016-06-03T00:00:00Z	0	RESISTÊNCIA	0		
110507	2016-06-08T00:00:00Z	0	MARIA ORTIZ	0		

	Hipertension	Diabetes	Alcoholism	Handcap	SMS_received	No-show
59	0	0	0	0	0	No
63	0	0	0	0	0	No
64	0	0	0	0	1	No
65	0	0	0	0	0	No
67	0	0	0	0	0	No
89	0	0	0	0	0	No
101	0	0	0	0	0	No

104	0	0	0	0	0	Yes
132	0	0	0	0	1	Yes
150	0	0	0	0	0	No
188	0	0	0	0	0	No
192	0	0	0	0	0	No
193	0	0	0	0	1	No
194	0	0	0	0	0	No
250	0	0	0	0	0	Yes
256	0	0	0	0	0	Yes
266	0	0	0	0	0	No
292	0	0	0	0	0	No
305	0	0	0	0	0	No
306	0	0	0	0	0	No
310	0	0	0	0	0	No
358	0	0	0	0	0	Yes
359	0	0	0	0	0	No
366	0	0	0	0	1	Yes
377	0	0	0	0	1	No
434	0	0	0	0	1	Yes
524	0	0	0	0	1	No
525	0	0	0	0	1	Yes
526	0	0	0	0	0	No
565	0	0	0	0	1	No
...	...	...	...	...	...	...
109629	0	0	0	0	0	Yes
109633	0	0	0	0	1	Yes
109646	0	0	0	0	0	No
109647	0	0	0	0	0	Yes
109649	0	0	0	0	0	No
109650	0	0	0	0	0	Yes
109830	0	0	0	0	0	Yes
109847	0	0	0	0	1	Yes
109848	0	0	0	0	0	No
109852	0	0	0	0	0	No
110231	0	0	0	0	0	No
110235	0	0	0	0	0	Yes
110236	0	0	0	0	1	No
110299	0	0	0	0	1	Yes
110313	0	0	0	0	1	No
110319	0	0	0	0	1	No
110320	0	0	0	0	1	No
110321	0	0	0	0	1	No
110331	0	0	0	0	1	Yes
110334	0	0	0	0	0	No
110335	0	0	0	0	1	No
110339	0	0	0	0	1	Yes
110341	0	0	0	0	1	No
110342	0	0	0	0	0	No

```
110343            0         0            0         0              1      No
110345            0         0            0         0              0      No
110346            0         0            0         0              0      No
110454            0         0            0         0              0      No
110460            0         0            0         0              0      No
110507            0         0            0         0              0      No

[3540 rows x 14 columns]
```

1.0.8 Data Cleaning!

Fixing typos in data columns

```
In [28]: # rename = Rename function
         df.rename(columns = {'Hipertension': 'Hypertension',
                              'Handcap': 'Handicap','No-show':'No_show'}, inplace = True)
```

converting some columns that has date to a datetime datatype

```
In [29]: # converting datatype to datetime
         df['ScheduledDay'] = pd.to_datetime(df['ScheduledDay'])
         df['AppointmentDay'] = pd.to_datetime(df['AppointmentDay'])
```

fixing the ages column that <0 by taking the mean of all ages.

```
In [35]: # we fix any age has 0 or less values because it's not make sense to have those values
         # in the data (there is no patient has the exactly the age 0 or less)
         mean_ages = df['Age'].mean()
         df[df['Age'] <= 0] = mean_ages
```

1.0.9 convert no show data to 0 and 1 instead of 'Yes' and 'No'

```
In [36]: df.No_show[df['No_show'] == 'Yes'] = '1'
         df.No_show[df['No_show'] == 'No'] = '0'
         df['No_show'] = pd.to_numeric(df['No_show'])
```

```
/opt/conda/lib/python3.6/site-packages/ipykernel_launcher.py:1: 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#
  """Entry point for launching an IPython kernel.
/opt/conda/lib/python3.6/site-packages/ipykernel_launcher.py:2: 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#
```

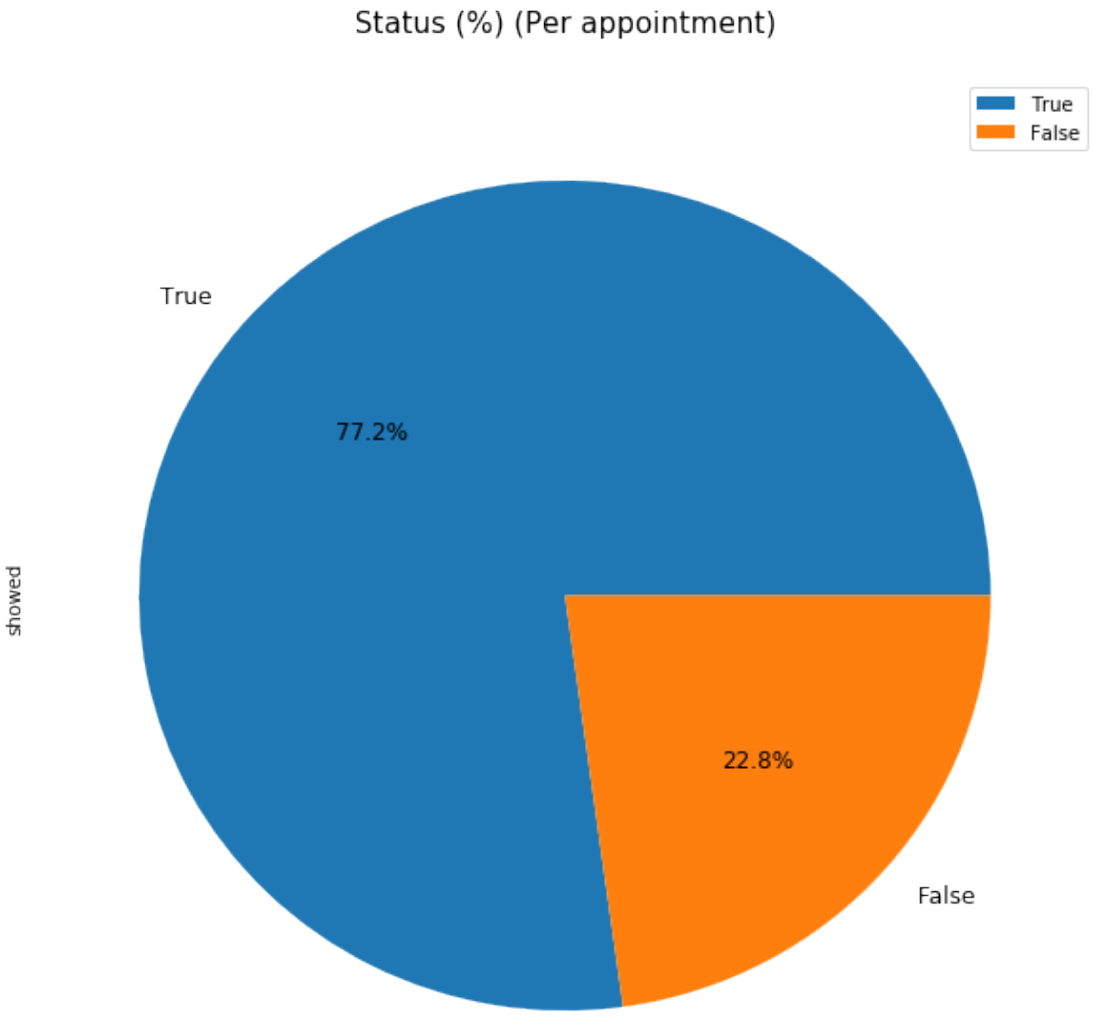
```
In [37]: # create a mask for people who came
         showed = df['No_show'] == 0
         not_showed = df['No_show'] == 1
         df['showed'] = showed
         df['not_showed'] = not_showed
```

Exploratory Data Analysis

1.0.10 1- (what is the overall appointment show-up vs. no show-up rate?)

```
In [38]: allP = df['showed'].value_counts()
         print(allP[1] / allP.sum() * 100)
         pieChart = allP.plot.pie(figsize=(10,10), autopct='%1.1f%%', fontsize = 12);
         pieChart.set_title("Status" + ' (%) (Per appointment)\n', fontsize = 15);
         plt.legend();
```

```
77.1820460159
```



1.0.11 2- (what are the most feature that matter the most of ('Age', 'being alcoholic','Having an SMS', 'Gender', 'Scholarship'), to make the patient make it to his appiontment?)

```
In [39]: df.head()
```

```
Out[39]:       PatientId  AppointmentID Gender        ScheduledDay  \
          0  2.987250e+13      5642903.0      F  2016-04-29 18:38:08
          1  5.589978e+14      5642503.0      M  2016-04-29 16:08:27
          2  4.262962e+12      5642549.0      F  2016-04-29 16:19:04
          3  8.679512e+11      5642828.0      F  2016-04-29 17:29:31
          4  8.841186e+12      5642494.0      F  2016-04-29 16:07:23

                 AppointmentDay   Age      Neighbourhood  Scholarship  Hypertension  \
```

0	2016-04-29 00:00:00	62.0	JARDIM DA PENHA	0.0	1.0
1	2016-04-29 00:00:00	56.0	JARDIM DA PENHA	0.0	0.0
2	2016-04-29 00:00:00	62.0	MATA DA PRAIA	0.0	0.0
3	2016-04-29 00:00:00	8.0	PONTAL DE CAMBURI	0.0	0.0
4	2016-04-29 00:00:00	56.0	JARDIM DA PENHA	0.0	1.0

	Diabetes	Alcoholism	Handicap	SMS_received	No_show	showed	not_showed
0	0.0	0.0	0.0	0.0	0.0	True	False
1	0.0	0.0	0.0	0.0	0.0	True	False
2	0.0	0.0	0.0	0.0	0.0	True	False
3	0.0	0.0	0.0	0.0	0.0	True	False
4	1.0	0.0	0.0	0.0	0.0	True	False

print the average age of people who show up.

```
In [40]: df.Age[showed].mean()
```

```
Out[40]: 39.075187264820002
```

print the average age of people not show up

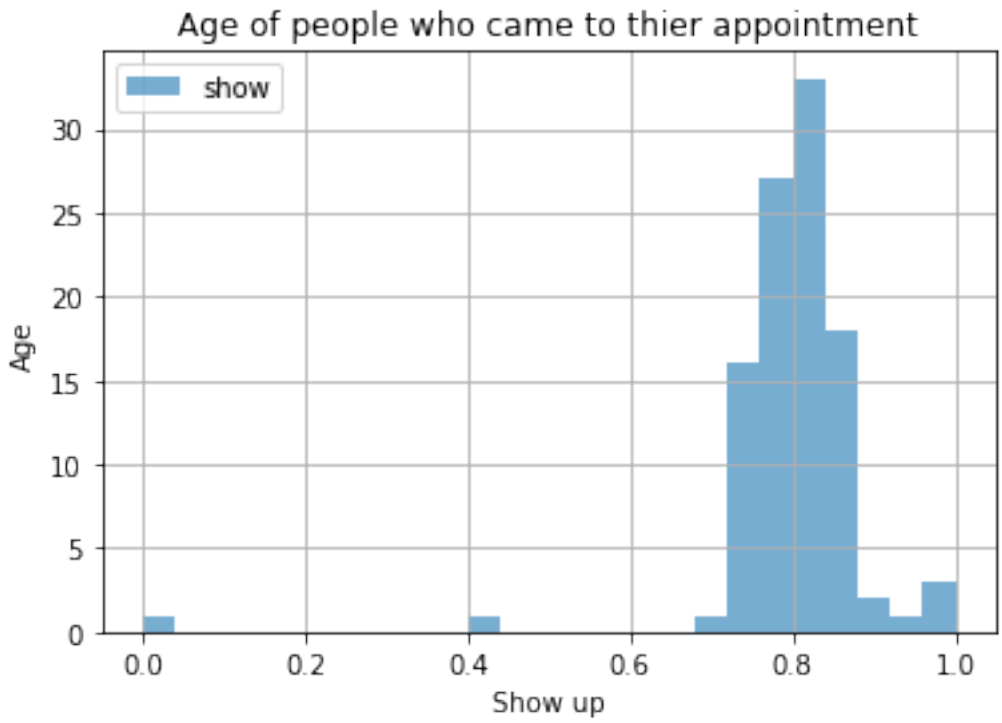
```
In [41]: df.Age[not_showed].mean()
```

```
Out[41]: 35.329151291512915
```

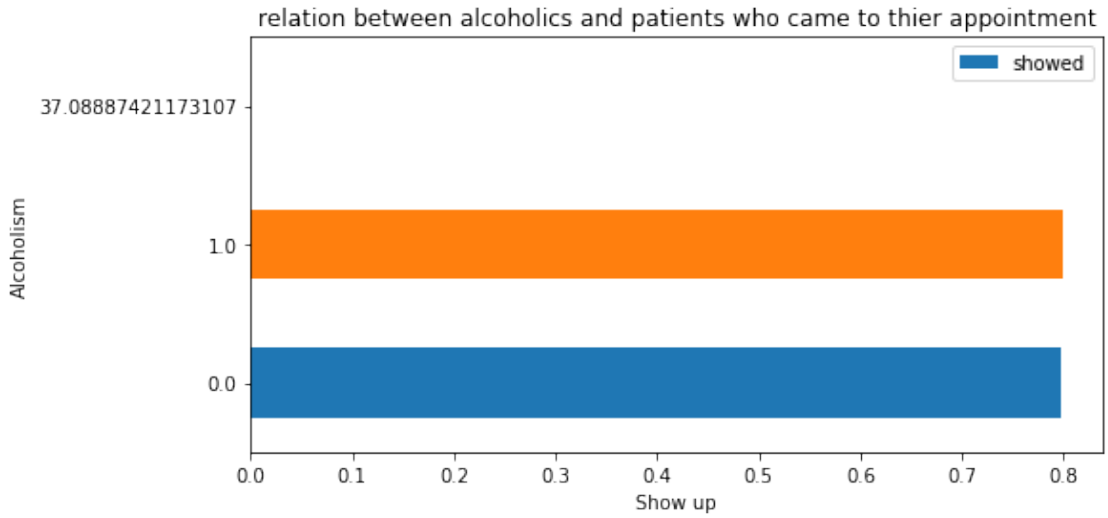
histogram to show the realtion between the (age,alcoholic people, SMS recievers, Gender, Scholarship) and the people who show up.

```
In [42]: df.groupby('Age')['showed'].mean().hist(alpha=0.6,bins=25,label='show');
         plt.xlabel("Show up")
         plt.ylabel("Age")
         plt.title("Age of people who came to thier appointment")
         plt.legend()
```

```
Out[42]: <matplotlib.legend.Legend at 0x7fe08b2e36d8>
```



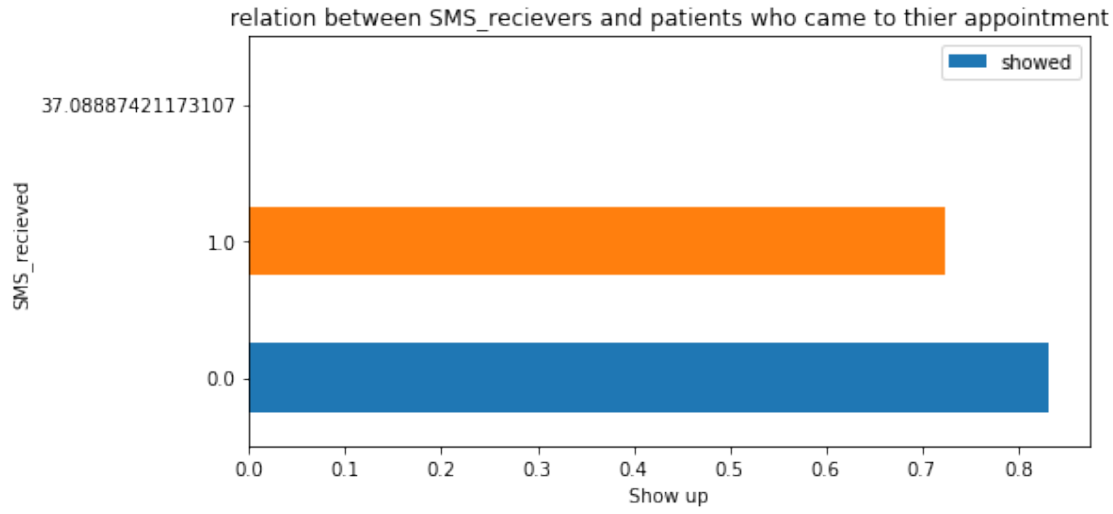

```
In [44]: print(df.groupby('Alcoholism')['showed'].mean())
         df.groupby('Alcoholism')['showed'].mean().plot(kind='barh',figsize=(8,4));
         plt.xlabel("Show up")
         plt.ylabel("Alcoholism")
         plt.title("relation between alcoholics and patients who came to thier appointment")
         plt.legend()

Alcoholism
0.000000     0.797321
1.000000     0.798512
37.088874    0.000000
Name: showed, dtype: float64


Out[44]: <matplotlib.legend.Legend at 0x7fe08b17c860>
```



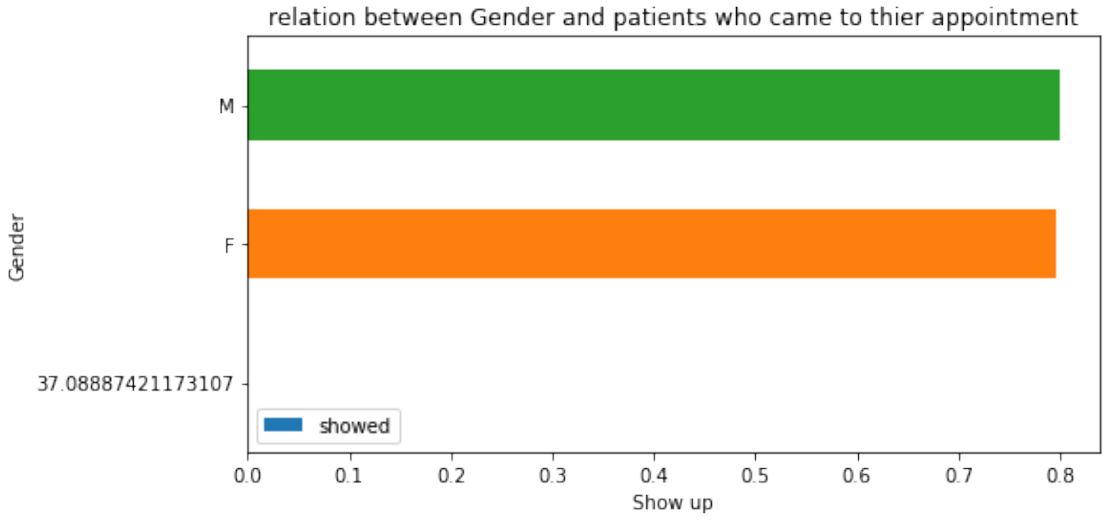

```
In [45]: print(df.groupby('SMS_received')['showed'].mean())
         df.groupby('SMS_received')['showed'].mean().plot(kind='barh',figsize=(8,4));
         plt.xlabel("Show up")
         plt.ylabel("SMS_recieved")
         plt.title("relation between SMS_recievers and patients who came to thier appointment")
         plt.legend()
```

```
SMS_received
0.000000     0.832712
1.000000     0.723348
37.088874    0.000000
Name: showed, dtype: float64
```

```
Out[45]: <matplotlib.legend.Legend at 0x7fe08b123b38>
```



```
In [46]: print(df.groupby('Gender')['showed'].mean())
         df.groupby('Gender')['showed'].mean().plot(kind='barh',figsize=(8,4));
         plt.xlabel("Show up")
         plt.ylabel("Gender")
         plt.title("relation between Gender and patients who came to thier appointment")
         plt.legend()

Gender
37.08887421173107    0.000000
F                    0.796415
M                    0.799154
Name: showed, dtype: float64


Out[46]: <matplotlib.legend.Legend at 0x7fe08b098550>
```

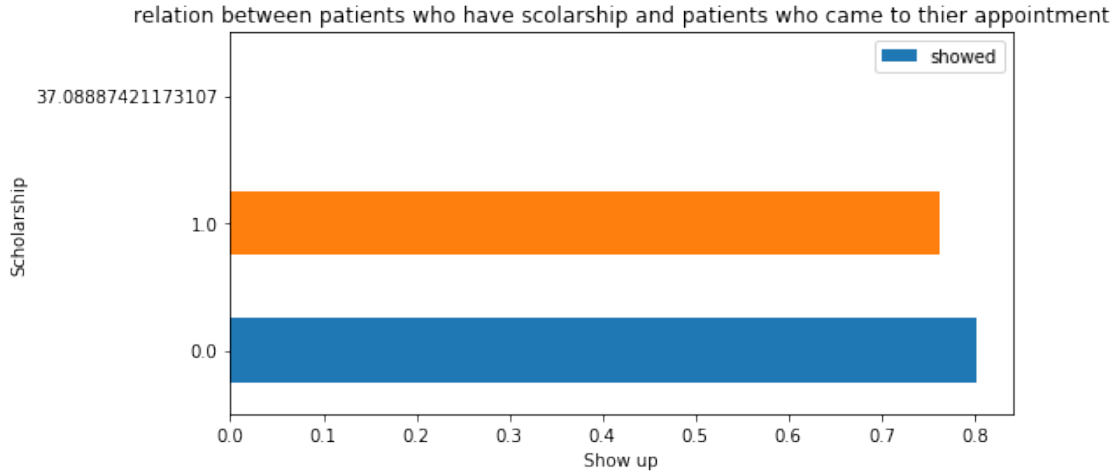


completing the visualiztion to see what is the feature that affect patients to come.

```
In [47]: print(df.groupby('Scholarship')['showed'].mean())
         df.groupby('Scholarship')['showed'].mean().plot(kind = 'barh',figsize=(8,4))
         plt.xlabel("Show up")
         plt.ylabel("Scholarship")
         plt.title("relation between patients who have scolarship and patients who came to thier
         plt.legend()

Scholarship
0.000000     0.801316
1.000000     0.762143
```

```
37.088874    0.000000
Name: showed, dtype: float64
```

```
Out[47]: <matplotlib.legend.Legend at 0x7fe08af6e518>
```



As we can see Age is the most factor that affect people who will attend or not!
Conclusions

what we noticed: Sending a text message is not necessary to confirm the patient's will presence.

The age is the most important factor that decides if a patient would come or not the average of age for people who will be most likely to show up is 39.07518726482 , and the average ages for people who are not show up is 35.329151291512915.

About 22.8% of people that schedule an appointment did not make it to thier appointment

Most of people who has Scholarship are not show up with a percentage of 76.2% of showing and patients who don't have a scholarship have the percentage 80.1%

*No any relation between gender and alcoholic that decide if the person would come to his appointment or not!

```
In [48]: from subprocess import call
         call(['python', '-m', 'nbconvert', 'Investigate_a_Dataset.ipynb'])
```

```
Out[48]: 0
```

```
In [ ]:
```