Investigate_a_Dataset

September 12, 2023

1 Project: Investigate a Dataset - [noshowappointments-kagglev2-may-2016.csv]

1.1 Table of Contents

Introduction
    Data Wrangling
    Exploratory Data Analysis
    Conclusions
    ## 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.

1.1.1 Dataset Description

we Have a file contains the data we are going to analyze

1.1.2 Question(s) for Analysis

What factors are important for us to know in order to predict if a patient will show up for their scheduled appointment?

```
In [3]: #import statmentes for all of the packeges that we plan to use
        import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
        #Rember to include a 'Magic word' so that your visualzation are plotted
        #inline with the notebook.see this page for more:
        #http://ipython.readhedocs.io/en/stuble/interactive/magic.html
        %matplotlib inline
```

```
In [9]: # Upgrade pandas to use dataframe.explode() function.
        !pip install --upgrade pandas==0.25.0
```

```
Collecting pandas==0.25.0
  Downloading https://files.pythonhosted.org/packages/1d/9a/7eb9952f4b4d73fbd75ad1d5d6112f407e69
    100% || 10.5MB 2.3MB/s eta 0:00:01  5% |                                | 593kB 62.4MB/s eta 0
```

```
Requirement already satisfied, skipping upgrade: python-dateutil>=2.6.1 in /opt/conda/lib/python
Requirement already satisfied, skipping upgrade: pytz>=2017.2 in /opt/conda/lib/python3.6/site-p
Collecting numpy>=1.13.3 (from pandas==0.25.0)
  Downloading https://files.pythonhosted.org/packages/45/b2/6c7545bb7a38754d63048c7696804a0d9473
    100% || 13.4MB 3.6MB/s eta 0:00:01
Requirement already satisfied, skipping upgrade: six>=1.5 in /opt/conda/lib/python3.6/site-packa
tensorflow 1.3.0 requires tensorflow-tensorboard<0.2.0,>=0.1.0, which is not installed.
Installing collected packages: numpy, pandas
  Found existing installation: numpy 1.12.1
    Uninstalling numpy-1.12.1:
      Successfully uninstalled numpy-1.12.1
  Found existing installation: pandas 0.23.3
    Uninstalling pandas-0.23.3:
      Successfully uninstalled pandas-0.23.3
Successfully installed numpy-1.19.5 pandas-0.25.0
```

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

1.1.3 General Properties

```
In [19]: import pandas as pd

         # Load your data from 'noshowappointments-kagglev2-may-2016.csv' and print out the firs
         df = pd.read_csv('noshowappointments-kagglev2-may-2016.csv')
         df.head()
```

Out[19]:

	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 [20]: #exploring the data shape
         df.shape
```

```
Out[20]: (110527, 14)
```

The data have 110527rows and 14 columns

```
In [21]: #check foe duplication
         df.duplicated().sum()
```

```
Out[21]: 0
```

There is 0 duplication

```
In [22]: # Check the number of unique values in the id
         df['PatientId'].nunique()
```

```
Out[22]: 62299
```

There is 62299 is unique values

```
In [23]: #check for duplications id
         df['PatientId'].duplicated().sum()
```

```
Out[23]: 48228
```

There is 48228 duplicated PatientId

1.1.4 Data Cleaning

```
In [24]: #getting some information about your data
         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
```

In [25]:
```
#identify the row index -1 for the Age
mask=df.query('Age=="-1"')
mask
```

Out[25]:
```
          PatientId  AppointmentID Gender          ScheduledDay  \
99832  4.659432e+14        5775010      F  2016-06-06T08:58:13Z

              AppointmentDay  Age Neighbourhood  Scholarship  Hipertension  \
99832  2016-06-06T00:00:00Z   -1         ROMÃO            0             0

       Diabetes  Alcoholism  Handcap  SMS_received No-show
99832         0           0        0             0      No
```

In [26]:
```
#remove the -1 value
df.drop(index=99832)
```

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
5       9.598513e+13        5626772      F  2016-04-27T08:36:51Z
6       7.336882e+14        5630279      F  2016-04-27T15:05:12Z
7       3.449833e+12        5630575      F  2016-04-27T15:39:58Z
8       5.639473e+13        5638447      F  2016-04-29T08:02:16Z
9       7.812456e+13        5629123      F  2016-04-27T12:48:25Z
10      7.345362e+14        5630213      F  2016-04-27T14:58:11Z
11      7.542951e+12        5620163      M  2016-04-26T08:44:12Z
12      5.666548e+14        5634718      F  2016-04-28T11:33:51Z
13      9.113946e+14        5636249      M  2016-04-28T14:52:07Z
14      9.988472e+13        5633951      F  2016-04-28T10:06:24Z
15      9.994839e+10        5620206      F  2016-04-26T08:47:27Z
16      8.457439e+13        5633121      M  2016-04-28T08:51:47Z
17      1.479497e+13        5633460      F  2016-04-28T09:28:57Z
18      1.713538e+13        5621836      F  2016-04-26T10:54:18Z
```

```
19      7.223289e+12        5640433      F  2016-04-29T10:43:14Z
20      6.222575e+14        5626083      F  2016-04-27T07:51:14Z
21      1.215484e+13        5628338      F  2016-04-27T10:50:45Z
22      8.632298e+14        5616091      M  2016-04-25T13:29:16Z
23      2.137540e+14        5634142      F  2016-04-28T10:27:05Z
24      8.734858e+12        5641780      F  2016-04-29T14:19:19Z
25      5.819370e+12        5624020      M  2016-04-26T15:04:17Z
26      2.578785e+10        5641781      F  2016-04-29T14:19:42Z
27      1.215484e+13        5628345      F  2016-04-27T10:51:45Z
28      5.926172e+12        5642400      M  2016-04-29T15:48:02Z
29      1.225776e+12        5642186      F  2016-04-29T15:16:29Z
...              ...            ...    ...                   ...
110497  7.935892e+14        5757745      M  2016-06-01T09:46:33Z
110498  9.433654e+13        5787655      F  2016-06-08T10:21:14Z
110499  8.219692e+14        5757697      F  2016-06-01T09:42:56Z
110500  4.434384e+14        5787233      F  2016-06-08T09:35:13Z
110501  4.544252e+11        5758133      M  2016-06-01T10:19:12Z
110502  7.316229e+14        5787937      F  2016-06-08T10:50:42Z
110503  2.362182e+13        5759473      F  2016-06-01T13:00:36Z
110504  9.947983e+12        5788052      F  2016-06-08T11:06:21Z
110505  5.667344e+13        5758455      F  2016-06-01T10:45:50Z
110506  8.973883e+11        5758779      M  2016-06-01T11:09:20Z
110507  4.769462e+14        5786918      F  2016-06-08T09:04:18Z
110508  9.433654e+13        5757656      F  2016-06-01T09:41:00Z
110509  4.952968e+14        5786750      M  2016-06-08T08:50:51Z
110510  2.362182e+13        5757587      F  2016-06-01T09:35:48Z
110511  8.235996e+11        5786742      F  2016-06-08T08:50:20Z
110512  9.876246e+13        5786368      F  2016-06-08T08:20:01Z
110513  8.674778e+13        5785964      M  2016-06-08T07:52:55Z
110514  2.695685e+12        5786567      F  2016-06-08T08:35:31Z
110515  6.456342e+14        5778621      M  2016-06-06T15:58:05Z
110516  6.923772e+13        5780205      F  2016-06-07T07:45:16Z
110517  5.574942e+12        5780122      F  2016-06-07T07:38:34Z
110518  7.263315e+13        5630375      F  2016-04-27T15:15:06Z
110519  6.542388e+13        5630447      F  2016-04-27T15:23:14Z
110520  9.969977e+14        5650534      F  2016-05-03T07:51:47Z
110521  3.635534e+13        5651072      F  2016-05-03T08:23:40Z
110522  2.572134e+12        5651768      F  2016-05-03T09:15:35Z
110523  3.596266e+12        5650093      F  2016-05-03T07:27:33Z
110524  1.557663e+13        5630692      F  2016-04-27T16:03:52Z
110525  9.213493e+13        5630323      F  2016-04-27T15:09:23Z
110526  3.775115e+14        5629448      F  2016-04-27T13:30:56Z

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

4	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0
5	2016-04-29T00:00:00Z	76	REPÚBLICA	0
6	2016-04-29T00:00:00Z	23	GOIABEIRAS	0
7	2016-04-29T00:00:00Z	39	GOIABEIRAS	0
8	2016-04-29T00:00:00Z	21	ANDORINHAS	0
9	2016-04-29T00:00:00Z	19	CONQUISTA	0
10	2016-04-29T00:00:00Z	30	NOVA PALESTINA	0
11	2016-04-29T00:00:00Z	29	NOVA PALESTINA	0
12	2016-04-29T00:00:00Z	22	NOVA PALESTINA	1
13	2016-04-29T00:00:00Z	28	NOVA PALESTINA	0
14	2016-04-29T00:00:00Z	54	NOVA PALESTINA	0
15	2016-04-29T00:00:00Z	15	NOVA PALESTINA	0
16	2016-04-29T00:00:00Z	50	NOVA PALESTINA	0
17	2016-04-29T00:00:00Z	40	CONQUISTA	1
18	2016-04-29T00:00:00Z	30	NOVA PALESTINA	1
19	2016-04-29T00:00:00Z	46	DA PENHA	0
20	2016-04-29T00:00:00Z	30	NOVA PALESTINA	0
21	2016-04-29T00:00:00Z	4	CONQUISTA	0
22	2016-04-29T00:00:00Z	13	CONQUISTA	0
23	2016-04-29T00:00:00Z	46	CONQUISTA	0
24	2016-04-29T00:00:00Z	65	TABUAZEIRO	0
25	2016-04-29T00:00:00Z	46	CONQUISTA	0
26	2016-04-29T00:00:00Z	45	BENTO FERREIRA	0
27	2016-04-29T00:00:00Z	4	CONQUISTA	0
28	2016-04-29T00:00:00Z	51	SÃO PEDRO	0
29	2016-04-29T00:00:00Z	32	SANTA MARTHA	0
...	...	...	...	...
110497	2016-06-01T00:00:00Z	76	MARIA ORTIZ	0
110498	2016-06-08T00:00:00Z	59	MARIA ORTIZ	0
110499	2016-06-01T00:00:00Z	66	MARIA ORTIZ	0
110500	2016-06-08T00:00:00Z	59	MARIA ORTIZ	0
110501	2016-06-01T00:00:00Z	44	MARIA ORTIZ	0
110502	2016-06-08T00:00:00Z	22	GOIABEIRAS	0
110503	2016-06-01T00:00:00Z	64	SOLON BORGES	0
110504	2016-06-08T00:00:00Z	4	MARIA ORTIZ	0
110505	2016-06-01T00:00:00Z	55	MARIA ORTIZ	0
110506	2016-06-01T00:00:00Z	5	MARIA ORTIZ	0
110507	2016-06-08T00:00:00Z	0	MARIA ORTIZ	0
110508	2016-06-01T00:00:00Z	59	MARIA ORTIZ	0
110509	2016-06-08T00:00:00Z	33	MARIA ORTIZ	0
110510	2016-06-01T00:00:00Z	64	SOLON BORGES	0
110511	2016-06-08T00:00:00Z	14	MARIA ORTIZ	0
110512	2016-06-08T00:00:00Z	41	MARIA ORTIZ	0
110513	2016-06-08T00:00:00Z	2	ANTÔNIO HONÓRIO	0
110514	2016-06-08T00:00:00Z	58	MARIA ORTIZ	0
110515	2016-06-08T00:00:00Z	33	MARIA ORTIZ	0
110516	2016-06-08T00:00:00Z	37	MARIA ORTIZ	0
110517	2016-06-07T00:00:00Z	19	MARIA ORTIZ	0

110518	2016-06-07T00:00:00Z	50	MARIA ORTIZ	0				
110519	2016-06-07T00:00:00Z	22	MARIA ORTIZ	0				
110520	2016-06-07T00:00:00Z	42	MARIA ORTIZ	0				
110521	2016-06-07T00:00:00Z	53	MARIA ORTIZ	0				
110522	2016-06-07T00:00:00Z	56	MARIA ORTIZ	0				
110523	2016-06-07T00:00:00Z	51	MARIA ORTIZ	0				
110524	2016-06-07T00:00:00Z	21	MARIA ORTIZ	0				
110525	2016-06-07T00:00:00Z	38	MARIA ORTIZ	0				
110526	2016-06-07T00:00:00Z	54	MARIA ORTIZ	0				

	Hipertension	Diabetes	Alcoholism	Handcap	SMS_received	No-show
0	1	0	0	0	0	No
1	0	0	0	0	0	No
2	0	0	0	0	0	No
3	0	0	0	0	0	No
4	1	1	0	0	0	No
5	1	0	0	0	0	No
6	0	0	0	0	0	Yes
7	0	0	0	0	0	Yes
8	0	0	0	0	0	No
9	0	0	0	0	0	No
10	0	0	0	0	0	No
11	0	0	0	0	1	Yes
12	0	0	0	0	0	No
13	0	0	0	0	0	No
14	0	0	0	0	0	No
15	0	0	0	0	1	No
16	0	0	0	0	0	No
17	0	0	0	0	0	Yes
18	0	0	0	0	1	No
19	0	0	0	0	0	No
20	0	0	0	0	0	Yes
21	0	0	0	0	0	Yes
22	0	0	0	0	1	Yes
23	0	0	0	0	0	No
24	0	0	0	0	0	No
25	1	0	0	0	1	No
26	1	0	0	0	0	No
27	0	0	0	0	0	No
28	0	0	0	0	0	No
29	0	0	0	0	0	No
...	...	...	...	...	...	...
110497	0	0	0	0	0	No
110498	0	0	0	0	0	No
110499	1	1	0	0	0	No
110500	0	0	0	0	0	No
110501	0	0	0	0	0	No
110502	0	0	0	0	0	No

```
110503             0         0          0         0             0      No
110504             0         0          0         0             0      No
110505             0         0          0         0             0      No
110506             0         0          0         0             0      No
110507             0         0          0         0             0      No
110508             0         0          0         0             0      No
110509             0         0          0         0             0      No
110510             0         0          0         0             0      No
110511             0         0          0         0             0      No
110512             0         0          0         0             0      No
110513             0         0          0         0             0      No
110514             0         0          0         0             0      No
110515             1         0          0         0             0     Yes
110516             0         0          0         0             0     Yes
110517             0         0          0         0             0      No
110518             0         0          0         0             1      No
110519             0         0          0         0             1      No
110520             0         0          0         0             1      No
110521             0         0          0         0             1      No
110522             0         0          0         0             1      No
110523             0         0          0         0             1      No
110524             0         0          0         0             1      No
110525             0         0          0         0             1      No
110526             0         0          0         0             1      No

[110526 rows x 14 columns]
```

In [20]:
```
# inspection for missing 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
```

In [ ]: No missing values

In [53]:
```
#correction the coulmens name
df.rename(columns={'Hipertension': 'Hypertension'})
df.rename(columns={'No-show': 'No_show'})
```

Out[53]:

	Gender	Age	Neighbourhood	Scholarship	Hypertension	Diabetes	\
0	F	62	JARDIM DA PENHA	0	1	0	
1	M	56	JARDIM DA PENHA	0	0	0	
2	F	62	MATA DA PRAIA	0	0	0	
3	F	8	PONTAL DE CAMBURI	0	0	0	
4	F	56	JARDIM DA PENHA	0	1	1	
5	F	76	REPÚBLICA	0	1	0	
6	F	23	GOIABEIRAS	0	0	0	
7	F	39	GOIABEIRAS	0	0	0	
8	F	21	ANDORINHAS	0	0	0	
9	F	19	CONQUISTA	0	0	0	
10	F	30	NOVA PALESTINA	0	0	0	
11	M	29	NOVA PALESTINA	0	0	0	
12	F	22	NOVA PALESTINA	1	0	0	
13	M	28	NOVA PALESTINA	0	0	0	
14	F	54	NOVA PALESTINA	0	0	0	
15	F	15	NOVA PALESTINA	0	0	0	
16	M	50	NOVA PALESTINA	0	0	0	
17	F	40	CONQUISTA	1	0	0	
18	F	30	NOVA PALESTINA	1	0	0	
19	F	46	DA PENHA	0	0	0	
20	F	30	NOVA PALESTINA	0	0	0	
21	F	4	CONQUISTA	0	0	0	
22	M	13	CONQUISTA	0	0	0	
23	F	46	CONQUISTA	0	0	0	
24	F	65	TABUAZEIRO	0	0	0	
25	M	46	CONQUISTA	0	1	0	
26	F	45	BENTO FERREIRA	0	1	0	
27	F	4	CONQUISTA	0	0	0	
28	M	51	SÃO PEDRO	0	0	0	
29	F	32	SANTA MARTHA	0	0	0	
...	...	...	...	...	...	...	
110497	M	76	MARIA ORTIZ	0	0	0	
110498	F	59	MARIA ORTIZ	0	0	0	
110499	F	66	MARIA ORTIZ	0	1	1	
110500	F	59	MARIA ORTIZ	0	0	0	
110501	M	44	MARIA ORTIZ	0	0	0	
110502	F	22	GOIABEIRAS	0	0	0	
110503	F	64	SOLON BORGES	0	0	0	

110504	F	4	MARIA ORTIZ	0	0	0
110505	F	55	MARIA ORTIZ	0	0	0
110506	M	5	MARIA ORTIZ	0	0	0
110507	F	0	MARIA ORTIZ	0	0	0
110508	F	59	MARIA ORTIZ	0	0	0
110509	M	33	MARIA ORTIZ	0	0	0
110510	F	64	SOLON BORGES	0	0	0
110511	F	14	MARIA ORTIZ	0	0	0
110512	F	41	MARIA ORTIZ	0	0	0
110513	M	2	ANTÔNIO HONÓRIO	0	0	0
110514	F	58	MARIA ORTIZ	0	0	0
110515	M	33	MARIA ORTIZ	0	1	0
110516	F	37	MARIA ORTIZ	0	0	0
110517	F	19	MARIA ORTIZ	0	0	0
110518	F	50	MARIA ORTIZ	0	0	0
110519	F	22	MARIA ORTIZ	0	0	0
110520	F	42	MARIA ORTIZ	0	0	0
110521	F	53	MARIA ORTIZ	0	0	0
110522	F	56	MARIA ORTIZ	0	0	0
110523	F	51	MARIA ORTIZ	0	0	0
110524	F	21	MARIA ORTIZ	0	0	0
110525	F	38	MARIA ORTIZ	0	0	0
110526	F	54	MARIA ORTIZ	0	0	0

	Alcoholism	Handcap	SMS_received	No_show
0	0	0	0	No
1	0	0	0	No
2	0	0	0	No
3	0	0	0	No
4	0	0	0	No
5	0	0	0	No
6	0	0	0	Yes
7	0	0	0	Yes
8	0	0	0	No
9	0	0	0	No
10	0	0	0	No
11	0	0	1	Yes
12	0	0	0	No
13	0	0	0	No
14	0	0	0	No
15	0	0	1	No
16	0	0	0	No
17	0	0	0	Yes
18	0	0	1	No
19	0	0	0	No
20	0	0	0	Yes
21	0	0	0	Yes
22	0	0	1	Yes

```
23              0         0             0      No
24              0         0             0      No
25              0         0             1      No
26              0         0             0      No
27              0         0             0      No
28              0         0             0      No
29              0         0             0      No
...           ...       ...           ...     ...
110497          0         0             0      No
110498          0         0             0      No
110499          0         0             0      No
110500          0         0             0      No
110501          0         0             0      No
110502          0         0             0      No
110503          0         0             0      No
110504          0         0             0      No
110505          0         0             0      No
110506          0         0             0      No
110507          0         0             0      No
110508          0         0             0      No
110509          0         0             0      No
110510          0         0             0      No
110511          0         0             0      No
110512          0         0             0      No
110513          0         0             0      No
110514          0         0             0      No
110515          0         0             0     Yes
110516          0         0             0     Yes
110517          0         0             0      No
110518          0         0             1      No
110519          0         0             1      No
110520          0         0             1      No
110521          0         0             1      No
110522          0         0             1      No
110523          0         0             1      No
110524          0         0             1      No
110525          0         0             1      No
110526          0         0             1      No

[110527 rows x 10 columns]
```

In [43]:
```
df.drop_duplicates(subset=['PatientId', 'No-show'])
df.shape
```

Out[43]:
```
(110527, 14)
```

In [49]:
```
# removing unnecessary data
df.drop(['PatientId','AppointmentID','ScheduledDay','AppointmentDay'], axis=1, inplace=
df.head()
```

```
Out[49]:   Gender  Age      Neighbourhood  Scholarship  Hipertension  Diabetes  \
         0      F   62    JARDIM DA PENHA            0             1         0
         1      M   56    JARDIM DA PENHA            0             0         0
         2      F   62      MATA DA PRAIA            0             0         0
         3      F    8  PONTAL DE CAMBURI            0             0         0
         4      F   56    JARDIM DA PENHA            0             1         1

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

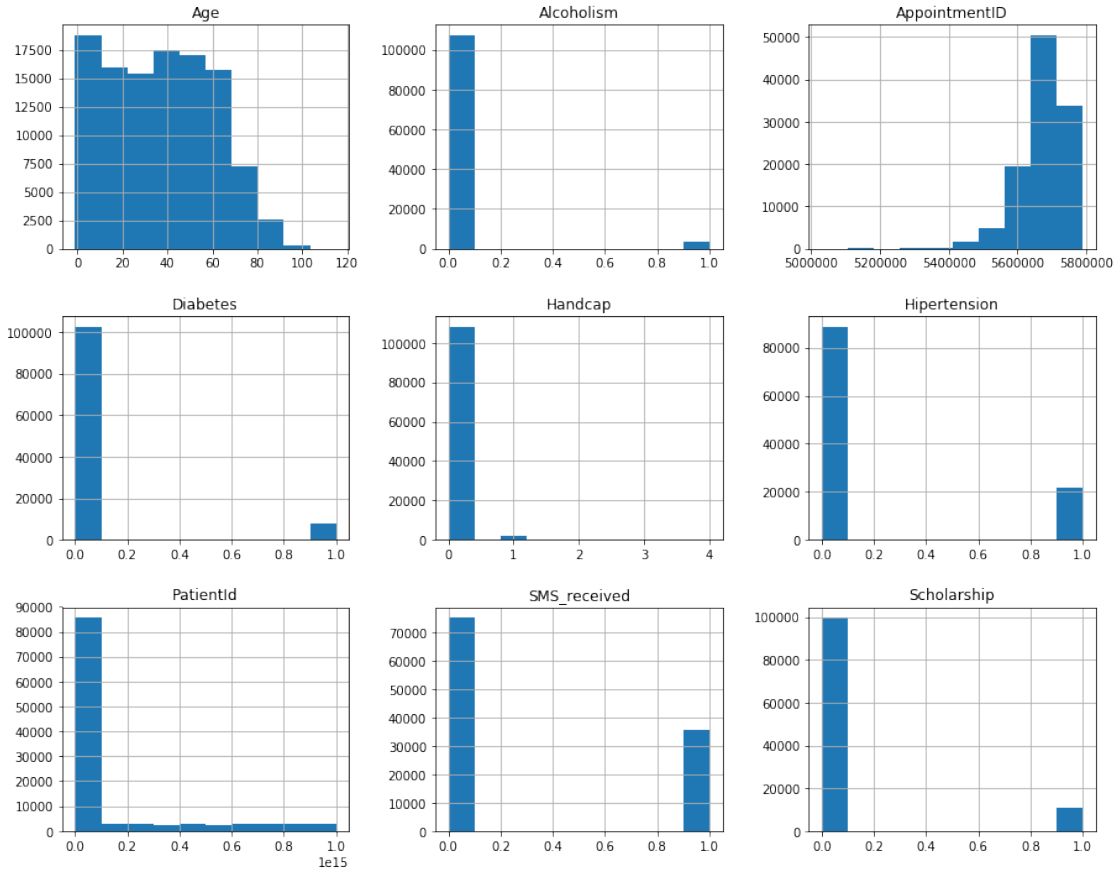
Exploratory Data Analysis

Now that you've prepared and refined your dataset, it's time to delve into exploration. Begin by calculating statistical measures and crafting graphical representations, all aimed at addressing the research inquiries you introduced in the initial section of your analysis

1.1.5 Analysis

```
In [42]: #histogram of whole data
         df.hist(figsize=(15,12)),
```

```
Out[42]: (array([[<matplotlib.axes._subplots.AxesSubplot object at 0x7fb2ffafea58>,
                  <matplotlib.axes._subplots.AxesSubplot object at 0x7fb2ffa0d5f8>,
                  <matplotlib.axes._subplots.AxesSubplot object at 0x7fb2ff9c84e0>],
                 [<matplotlib.axes._subplots.AxesSubplot object at 0x7fb2ff980780>,
                  <matplotlib.axes._subplots.AxesSubplot object at 0x7fb2ff9a19b0>,
                  <matplotlib.axes._subplots.AxesSubplot object at 0x7fb2ff9a15f8>],
                 [<matplotlib.axes._subplots.AxesSubplot object at 0x7fb2ff909588>,
                  <matplotlib.axes._subplots.AxesSubplot object at 0x7fb2ff8c17b8>,
                  <matplotlib.axes._subplots.AxesSubplot object at 0x7fb2ff8de630>]], dtype=obje
```




```
In [44]: # Assigning name of show and no show facilitate recall
         # DataFrame for 'No' and 'Yes' in the 'No-show' column
         show = df['No-show'] == 'No'
         noshow = df['No-show'] == 'Yes'
         df[show].count(),df[noshow].count()
```

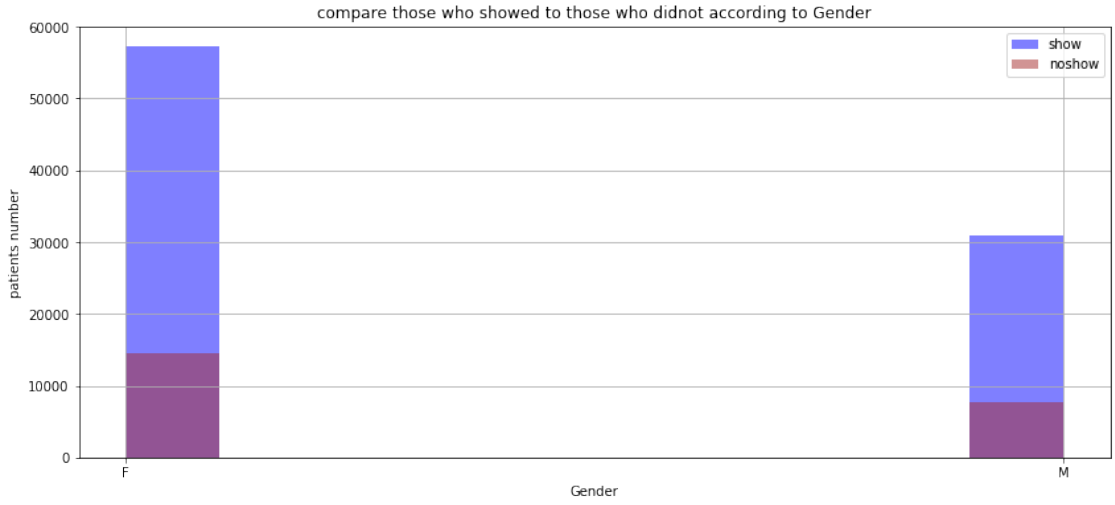
```
Out[44]: (PatientId         88208
          AppointmentID     88208
          Gender            88208
          ScheduledDay      88208
          AppointmentDay    88208
          Age               88208
          Neighbourhood     88208
          Scholarship       88208
          Hipertension      88208
          Diabetes          88208
          Alcoholism        88208
          Handcap           88208
          SMS_received      88208
          No-show           88208
```

```
dtype: int64, PatientId         22319
AppointmentID     22319
Gender            22319
ScheduledDay      22319
AppointmentDay    22319
Age               22319
Neighbourhood     22319
Scholarship       22319
Hipertension      22319
Diabetes          22319
Alcoholism        22319
Handcap           22319
SMS_received      22319
No-show           22319
dtype: int64)
```

The No show is 88208 biger than no-show 22319

1.1.6 Analysing

```
In [68]: import matplotlib.pyplot as plt  # Import the matplotlib library
         # compare those who showed to those who didnot according to Gender

         plt.figure(figsize=(14, 6))
         df.Gender[show].hist(alpha=0.5,color='blue',label='show')
         df.Gender[noshow].hist(alpha=0.5,color='brown',label='noshow')
         plt.legend();
         plt.title('compare those who showed to those who didnot according to Gender')
         plt.xlabel('Gender')
         plt.ylabel('patients number')

Out[68]: Text(0,0.5,'patients number')
```



```
In [69]: # compare those who showed to those who didnot according to Gender

         print(df.Gender[show].value_counts())
         print(df.Gender[noshow].value_counts())
```

```
F    57246
M    30962
Name: Gender, dtype: int64
F    14594
M     7725
Name: Gender, dtype: int64
```

Gender is negligible famale are more than meale

```
In [80]: import matplotlib.pyplot as plt  # Import the matplotlib library
         # compare those who showed to those who didnot according to enrollment in the Brasilan

         plt.figure(figsize=(16, 8))
         df.Gender[show].hist(alpha=0.5,color='green',label='show')
         df.Gender[noshow].hist(alpha=0.5,color='red',label='noshow')
         plt.legend();
         plt.title(' compare those who showed to those who didnot according to enrollment in the
         plt.xlabel('Welfare')
         plt.ylabel('patients number')
```

```
Out[80]: Text(0,0.5,'patients number')
```

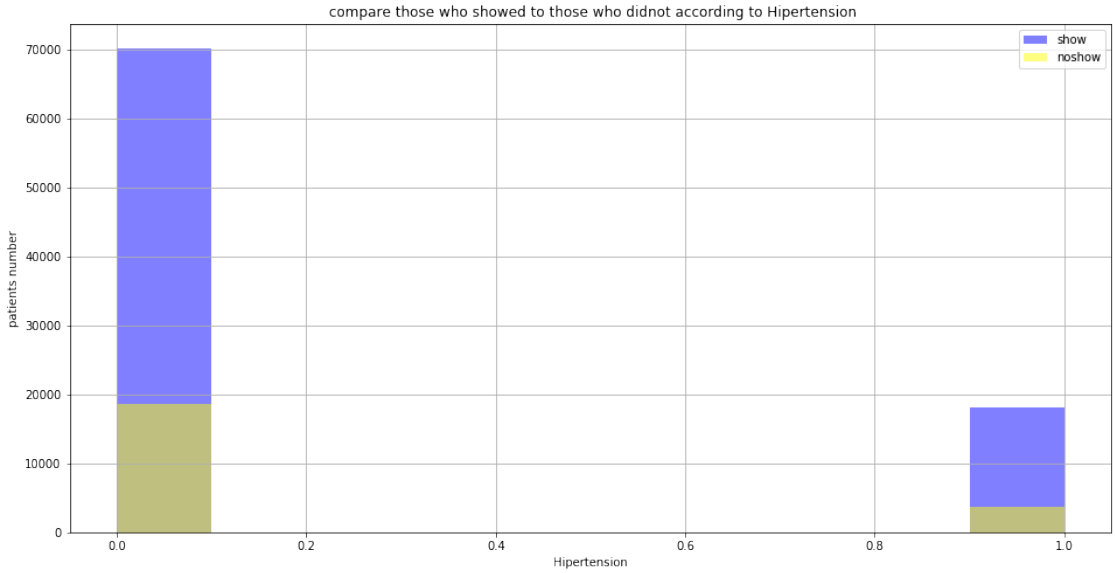
```
In [81]: #compare those who showed to those who didnot according to gender
         print(df.Scholarship[show].value_counts())
         print(df.Scholarship[noshow].value_counts())
```

```
0    79925
1     8283
Name: Scholarship, dtype: int64
0    19741
1     2578
Name: Scholarship, dtype: int64
```

enrollment in the Brasilan welfare program is negligible

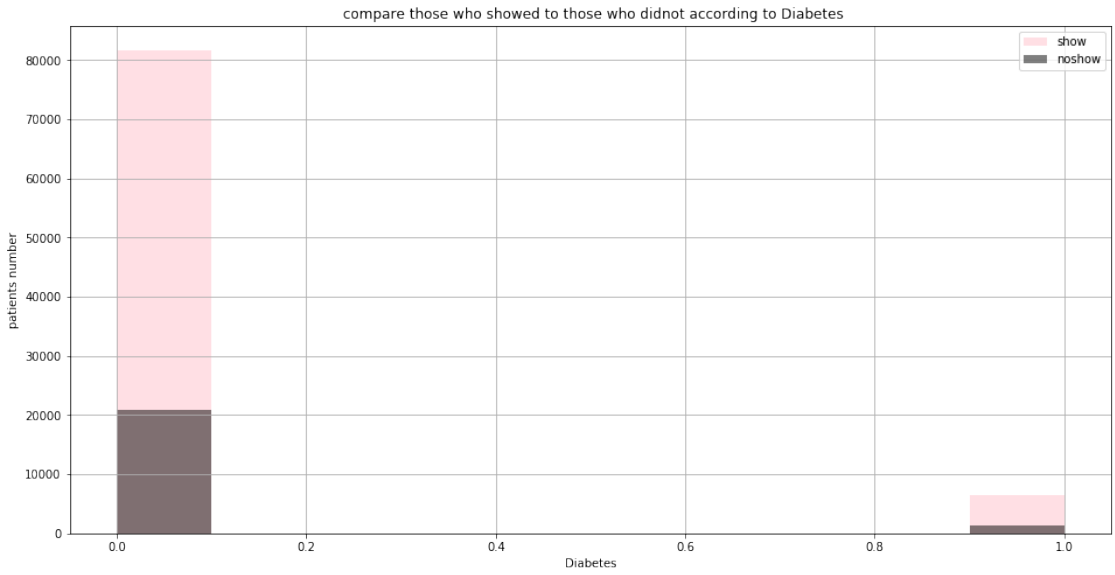
```
In [84]: import matplotlib.pyplot as plt  # Import the matplotlib library
         #compare those who showed to those who didnot according to Hypertension

         plt.figure(figsize=(16, 8))
         df.Hipertension[show].hist(alpha=0.5,color='blue',label='show')
         df.Hipertension[noshow].hist(alpha=0.5,color='yellow',label='noshow')
         plt.legend();
         plt.title(' compare those who showed to those who didnot according to Hipertension')
         plt.xlabel('Hipertension')
         plt.ylabel('patients number')
```

```
Out[84]: Text(0,0.5,'patients number')
```



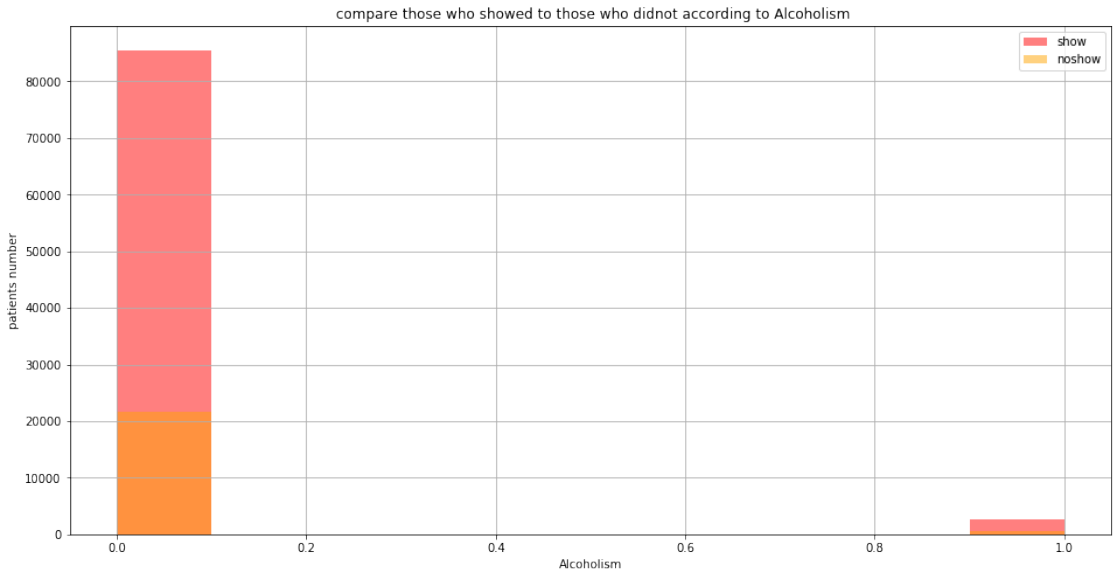
Hypertension is negligible

```
In [85]: import matplotlib.pyplot as plt  # Import the matplotlib library
         #compare those who showed to those who didnot according to Diabetes

         plt.figure(figsize=(16, 8))
         df.Diabetes[show].hist(alpha=0.5,color='pink',label='show')
         df.Diabetes[noshow].hist(alpha=0.5,color='black',label='noshow')
         plt.legend();
         plt.title(' compare those who showed to those who didnot according to Diabetes')
         plt.xlabel('Diabetes')
         plt.ylabel('patients number')

Out[85]: Text(0,0.5,'patients number')
```



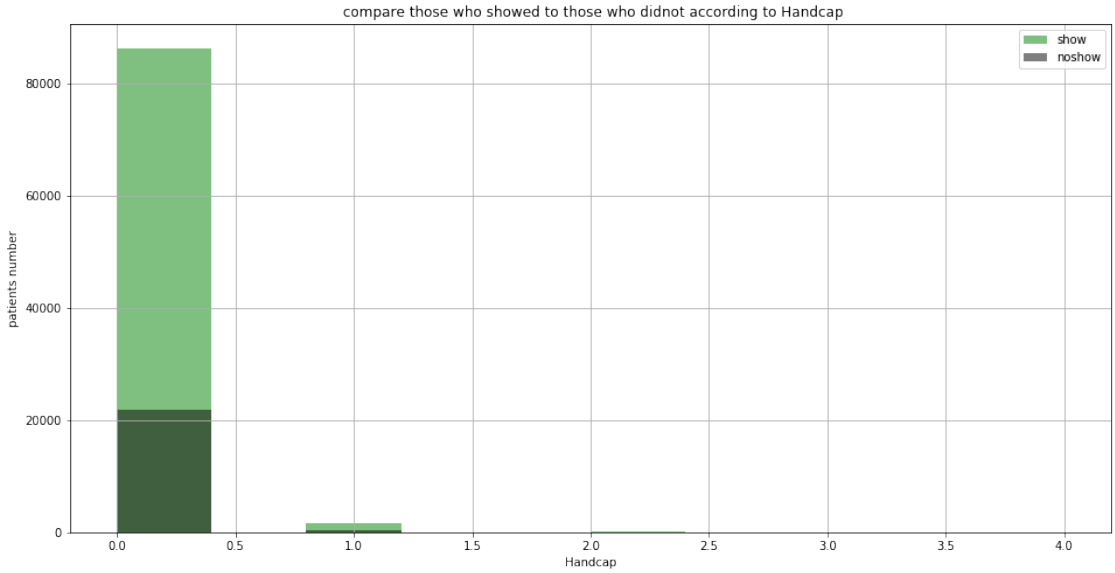
Diabetes is negligible

```
In [86]: import matplotlib.pyplot as plt  # Import the matplotlib library
         #compare those who showed to those who didnot according to Alcoholism

         plt.figure(figsize=(16, 8))
         df.Alcoholism[show].hist(alpha=0.5,color='red',label='show')
         df.Alcoholism[noshow].hist(alpha=0.5,color='orange',label='noshow')
         plt.legend();
         plt.title(' compare those who showed to those who didnot according to Alcoholism')
         plt.xlabel('Alcoholism')
         plt.ylabel('patients number')

Out[86]: Text(0,0.5,'patients number')
```



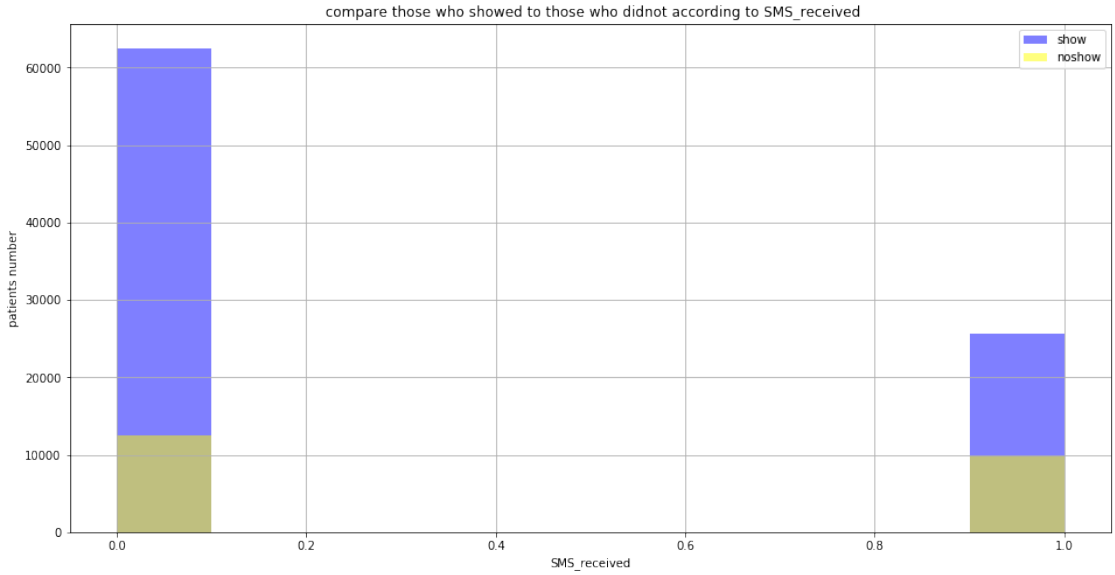
Alcoholism is negligible

```
In [88]: import matplotlib.pyplot as plt  # Import the matplotlib library
         #compare those who showed to those who didnot according to Handcap

         plt.figure(figsize=(16, 8))
         df.Handcap[show].hist(alpha=0.5,color='green',label='show')
         df.Handcap[noshow].hist(alpha=0.5,color='black',label='noshow')
         plt.legend();
         plt.title(' compare those who showed to those who didnot according to Handcap')
         plt.xlabel('Handcap')
         plt.ylabel('patients number')

Out[88]: Text(0,0.5,'patients number')
```



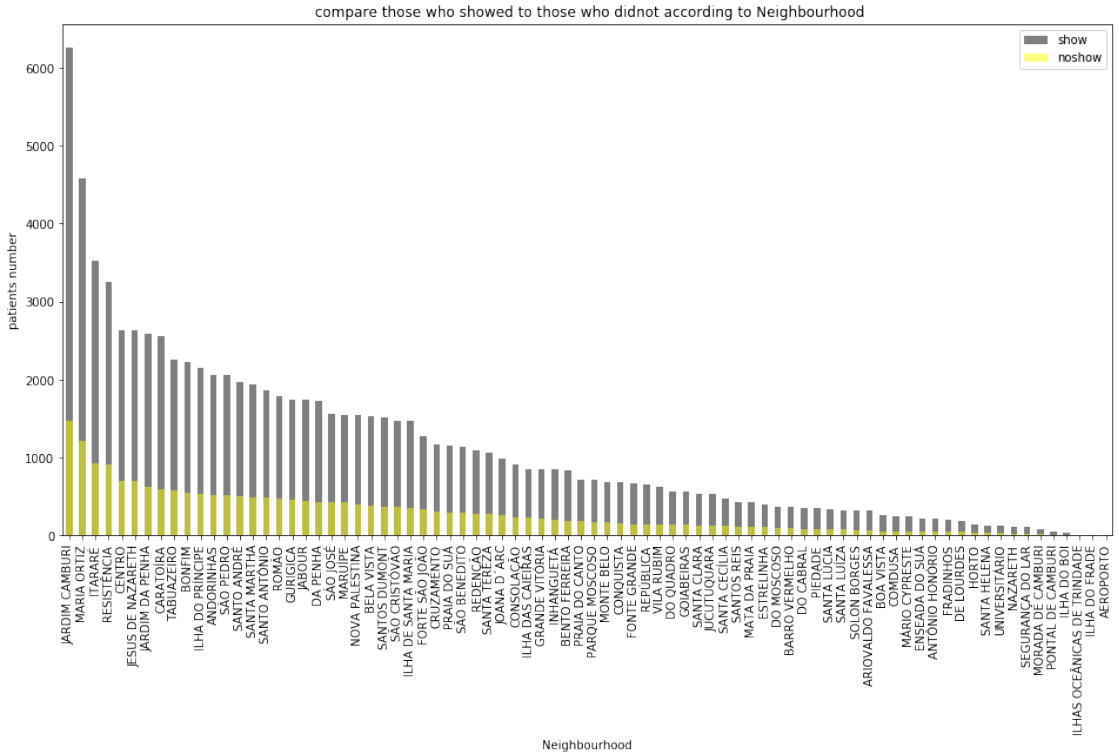
Handcap is negligible

```
In [90]: import matplotlib.pyplot as plt  # Import the matplotlib library
         #compare those who showed to those who didnot according to SMS_received

         plt.figure(figsize=(16, 8))
         df.SMS_received[show].hist(alpha=0.5,color='blue',label='show')
         df.SMS_received[noshow].hist(alpha=0.5,color='yellow',label='noshow')
         plt.legend();
         plt.title(' compare those who showed to those who didnot according to SMS_received')
         plt.xlabel('SMS_received')
         plt.ylabel('patients number')

Out[90]: Text(0,0.5,'patients number')
```



In [94]:
```
import matplotlib.pyplot as plt  # Import the matplotlib library
#compare those who showed to those who didnot according to Neighbourhood

plt.figure(figsize=(16, 8))
df.Neighbourhood[show].value_counts().plot(kind='bar',alpha=0.5,color='black',label='sh
df.Neighbourhood[noshow].value_counts().plot(kind='bar',alpha=0.5,color='yellow',label=
plt.legend();
plt.title(' compare those who showed to those who didnot according to Neighbourhood')
plt.xlabel('Neighbourhood')
plt.ylabel('patients number')
```

Out[94]:
```
Text(0,0.5,'patients number')
```




```
In [ ]: import matplotlib.pyplot as plt  # Import the matplotlib library
        #compare those who showed to those who didnot according to SMS_received

        plt.figure(figsize=(16, 8))
        df.SMS_received[show].hist(alpha=0.5,color='blue',label='show')
        df.SMS_received[noshow].hist(alpha=0.5,color='yellow',label='noshow')
        plt.legend();
        plt.title(' compare those who showed to those who didnot according to SMS_received')
        plt.xlabel('SMS_received')
        plt.ylabel('patients number')
```

I can assert that there is a significant correlation between the patients' choice of neighborhood and their attendance at the clinic.

```
In [92]: import matplotlib.pyplot as plt  # Import the matplotlib library
         #compare those who showed to those who didnot according to Age
         plt.figure(figsize=(16, 8))
         df.Age[show].hist(alpha=0.5,color='red',label='show')
         df.Age[noshow].hist(alpha=0.5,color='pink',label='noshow')
         plt.legend();
         plt.title(' compare those who showed to those who didnot according to Age')
         plt.xlabel('Age')
         plt.ylabel('patients number')
```

Out[92]: Text(0,0.5,'patients number')

"Patients between the ages of 0 to 10 exhibited a higher appointment rate compared to all other age groups. Appointment rates tend to decrease as patients grow older.

Conclusions n conclusion, it is evident that the neighborhood plays a significant role in patients showing up at the clinic. Age also plays a crucial role, with the 0-10 age group showing the highest attendance, followed by the 35-70 age group. Surprisingly, a larger number of patients attended appointments without receiving an SMS reminder.These findings highlight the complex interplay of factors affecting patient attendance and suggest that more research is needed to explore the nuances of this relationship.

1.2 Submitting your Project

Tip: Before you submit your project, you need to create a .html or .pdf version of this notebook in the workspace here. To do that, run the code cell below. If it worked correctly, you should get a return code of 0, and you should see the generated .html file in the workspace directory (click on the orange Jupyter icon in the upper left).

Tip: Alternatively, you can download this report as .html via the **File** > **Download as** submenu, and then manually upload it into the workspace directory by clicking on the orange Jupyter icon in the upper left, then using the Upload button.

Tip: Once you've done this, you can submit your project by clicking on the "Submit Project" button in the lower right here. This will create and submit a zip file with this .ipynb doc and the .html or .pdf version you created. Congratulations!

```
In [1]: from subprocess import call
        call(['python', '-m', 'nbconvert', 'Investigate_a_Dataset.ipynb'])
```

Out[1]: 0

In [ ]: