Makarem-Investigate-A-Dataset-NoShowAppoientment

December 24, 2018

1 Investigate a Dataset (No show Appointment)

Done by: Makarem Al-Salman

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.

```
<li>ScheduledDay tells us on what day the patient set up their appointment.</li>
```

'Neighborhood' indicates the location of the hospital.

'Scholarship' indicates whether or not the patient is enrolled in Brasilian welfare program Bolsa Família.

the last column: says 'No' if the patient showed up to their appointment, and 'Yes' if they did not show up.

## Posted Questions

Do the female patient care more about their appointments?

Does having a Diabetes affect on the patient commitment of their appointments?

Which age range is more commitment of their appointments?

```
In [195]: # Use this cell to set up import statements for all of the packages that you
          #   plan to use.
          import pandas as pd
          import numpy as np
          % matplotlib inline
```

## Data Wrangling

```
In [196]: # Loading data and Perform operations to inspect data
          # types and look for instances of missing or possibly errant data.
          df_appointments=pd.read_csv('No_show_appointment.csv')
          df_appointments.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 [197]: df_appointments.head(20)
```

```
Out[197]:        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

                    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
5   2016-04-29T00:00:00Z   76          REPÚBLICA            0             1
6   2016-04-29T00:00:00Z   23         GOIABEIRAS            0             0
7   2016-04-29T00:00:00Z   39         GOIABEIRAS            0             0
8   2016-04-29T00:00:00Z   21         ANDORINHAS            0             0
9   2016-04-29T00:00:00Z   19          CONQUISTA            0             0
10  2016-04-29T00:00:00Z   30     NOVA PALESTINA            0             0
11  2016-04-29T00:00:00Z   29     NOVA PALESTINA            0             0
12  2016-04-29T00:00:00Z   22     NOVA PALESTINA            1             0
13  2016-04-29T00:00:00Z   28     NOVA PALESTINA            0             0
14  2016-04-29T00:00:00Z   54     NOVA PALESTINA            0             0
15  2016-04-29T00:00:00Z   15     NOVA PALESTINA            0             0
16  2016-04-29T00:00:00Z   50     NOVA PALESTINA            0             0
17  2016-04-29T00:00:00Z   40          CONQUISTA            1             0
18  2016-04-29T00:00:00Z   30     NOVA PALESTINA            1             0
19  2016-04-29T00:00:00Z   46           DA PENHA            0             0

    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
5          0           0        0             0      No
6          0           0        0             0     Yes
7          0           0        0             0     Yes
8          0           0        0             0      No
9          0           0        0             0      No
10         0           0        0             0      No
11         0           0        0             1     Yes
12         0           0        0             0      No
13         0           0        0             0      No
14         0           0        0             0      No
15         0           0        0             1      No
16         0           0        0             0      No
17         0           0        0             0     Yes
18         0           0        0             1      No
19         0           0        0             0      No
```

```
In [198]: # Create a list of unique values in handcap column
          list(df_appointments['Handcap'].unique())

Out[198]: [0, 1, 2, 3, 4]

In [199]: # this returns a tuple of the dimensions of the dataframe
          df_appointments.shape
```

```
Out[199]: (110527, 14)
```

```
In [200]: # although the datatype for AppointmentDay, ScheduledDay appears to be object, further
          # investigation shows it's a string
          print('appointment Day data type:',type(df_appointments['AppointmentDay'][0]))
          print('Scheduled Day data type:',type(df_appointments['ScheduledDay'][0]))
```

```
appointment Day data type: <class 'str'>
Scheduled Day data type: <class 'str'>
```

```
In [201]: sum(df_appointments.duplicated())
```

```
Out[201]: 0
```

```
In [202]: # this returns useful descriptive statistics for each column of data
          df_appointments.describe()
```

```
Out[202]:           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
```

1.1.1 Data Cleaning

After assessing and exploring the data I found that the dataset has no missing data & no duplicate data!! WOW. But it has the following problems:
icorrect data type for 'AppointmentDay' , 'ScheduledDay' ,and 'PatientId' columns.
outlier values in 'Age' column like (-1, 115).
Misspelled names in 'Hipertension' , 'No-show' , and 'Handcap' columns.

1. Columns Remaing: I will start with renaming 'Hipertension' , 'No-show' , and 'Handcap' columns.

```
In [203]: #renaming  Hipertension , and Handcap columns
          df_appointments.rename(columns={'Hipertension':'Hypertension','Handcap':'handicap','No
          #check the result
          df_appointments.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
Hypertension      110527 non-null int64
Diabetes          110527 non-null int64
Alcoholism        110527 non-null int64
handicap          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
```

2.changing datatypes: second, i will change the datatype for:
'AppointmentDay' from string to date & time
'ScheduledDay' from string to date & time
'PatientId' from float to string

```
In [204]: #changing datatype of PatientId column by using numpy functions
          df_appointments['PatientId']= (df_appointments['PatientId']).astype(str)
          #changing datatype by using pandas function
          df_appointments['AppointmentDay']=pd.to_datetime(df_appointments['AppointmentDay'])
          df_appointments['ScheduledDay']=pd.to_datetime(df_appointments['ScheduledDay'])
          #check the result
```

```
df_appointments.info()
df_appointments.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 110527 entries, 0 to 110526
Data columns (total 14 columns):
PatientId         110527 non-null object
AppointmentID     110527 non-null int64
Gender            110527 non-null object
ScheduledDay      110527 non-null datetime64[ns]
AppointmentDay    110527 non-null datetime64[ns]
Age               110527 non-null int64
Neighbourhood     110527 non-null object
Scholarship       110527 non-null int64
Hypertension      110527 non-null int64
Diabetes          110527 non-null int64
Alcoholism        110527 non-null int64
handicap          110527 non-null int64
SMS_received      110527 non-null int64
No_show           110527 non-null object
dtypes: datetime64[ns](2), int64(8), object(4)
memory usage: 11.8+ MB
```

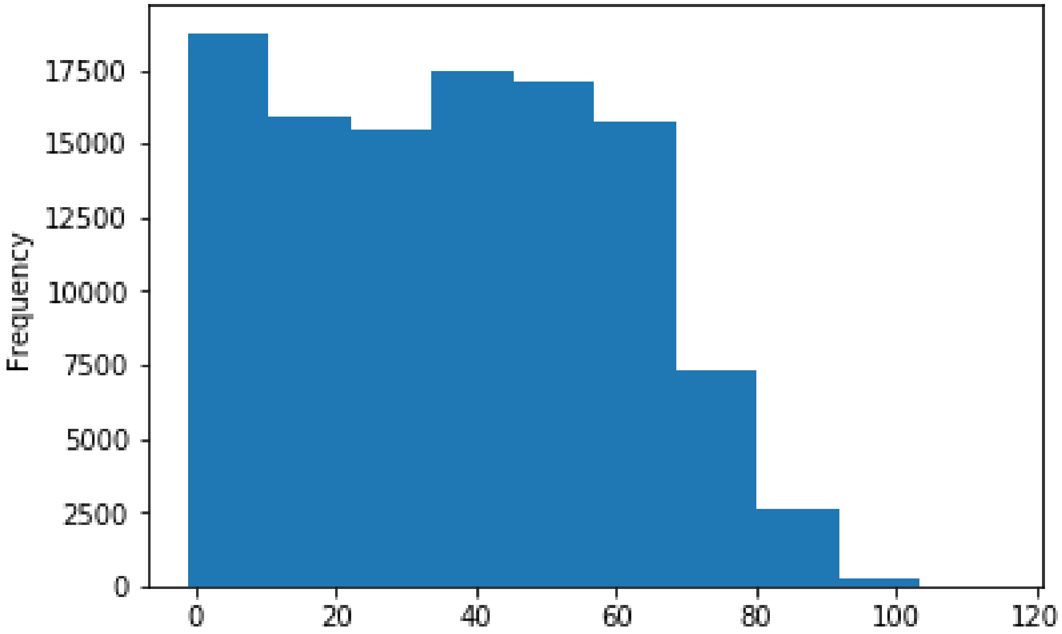
```
Out[204]:            PatientId  AppointmentID Gender        ScheduledDay AppointmentDay  \
          0  2.98724998243e+13        5642903      F 2016-04-29 18:38:08     2016-04-29
          1  5.58997776694e+14        5642503      M 2016-04-29 16:08:27     2016-04-29
          2  4.26296229995e+12        5642549      F 2016-04-29 16:19:04     2016-04-29
          3     867951213174.0        5642828      F 2016-04-29 17:29:31     2016-04-29
          4  8.84118644818e+12        5642494      F 2016-04-29 16:07:23     2016-04-29

             Age      Neighbourhood  Scholarship  Hypertension  Diabetes  Alcoholism  \
          0   62    JARDIM DA PENHA            0             1         0           0
          1   56    JARDIM DA PENHA            0             0         0           0
          2   62      MATA DA PRAIA            0             0         0           0
          3    8  PONTAL DE CAMBURI            0             0         0           0
          4   56    JARDIM DA PENHA            0             1         1           0

             handicap  SMS_received No_show
          0         0             0      No
          1         0             0      No
          2         0             0      No
          3         0             0      No
          4         0             0      No
```

3.Removing outlier data: finally I will remove the unrealistic data in 'Age' column in the following steps:

```
<li>detacte the outliler values.</li>
```

<li>define the boundries.</li>
<li>replace the outliers value with the mean of patients ages.</li>

```
In [205]: #detacting the outliler values
          # plot Age to draw histogram
          df_appointments['Age'].plot(kind='hist');
```



the concentration of the values is very high in range of 0 years to 90 years. therefore, i will consider the values less than 0 and above 90 as outliers.

```
In [206]: #define the boundries
          Lower_bound = 0.0
          Upper_bound = 0.997
          result = df_appointments['Age'].quantile([Lower_bound,Upper_bound])
          result
```

```
Out[206]: 0.000    -1.0
          0.997    90.0
          Name: Age, dtype: float64
```

any value is more than -1 is acceptable value and any value is more than 90 is an outlier value

case 1 : -1.0 < -1 < 90.0 - False

case 2 : -1.0 < 20 < 90.0 - True

```
In [207]: #definin the values accepted range
          true_value =(result.loc[Lower_bound] < df_appointments['Age'].values) & (df_appointmen
          #check the result
          true_value
```

```
Out[207]: array([ True,  True,  True, ...,  True,  True,  True], dtype=bool)
```

```
In [208]: #use this line to check the result
          #df_appointments.Age[true_value]
```

```
In [209]: #obtaining the outlier values by reversing the true values
          false_value = ˜true_value
          false_value
```

```
Out[209]: array([False, False, False, ..., False, False, False], dtype=bool)
```

```
In [210]: #calculating the mean of the patients ages
          mean_value = np.mean(df_appointments.Age[true_value])
          mean_value = int(mean_value)
          mean_value
```

```
Out[210]: 36
```

```
In [211]: #replace the ouliers values with the mean value
          df_appointments.Age[false_value].fillna(mean_value,inplace=True)
          #check the result
          df_appointments
```

```
Out[211]:                 PatientId  AppointmentID Gender        ScheduledDay  \
          0       2.98724998243e+13        5642903      F 2016-04-29 18:38:08
          1       5.58997776694e+14        5642503      M 2016-04-29 16:08:27
          2       4.26296229995e+12        5642549      F 2016-04-29 16:19:04
          3          867951213174.0        5642828      F 2016-04-29 17:29:31
          4       8.84118644818e+12        5642494      F 2016-04-29 16:07:23
          5       9.59851332313e+13        5626772      F 2016-04-27 08:36:51
          6       7.33688164477e+14        5630279      F 2016-04-27 15:05:12
          7       3.44983339412e+12        5630575      F 2016-04-27 15:39:58
          8         5.639472995e+13        5638447      F 2016-04-29 08:02:16
          9       7.81245643693e+13        5629123      F 2016-04-27 12:48:25
          10      7.34536231958e+14        5630213      F 2016-04-27 14:58:11
          11      7.54295136844e+12        5620163      M 2016-04-26 08:44:12
          12      5.66654781423e+14        5634718      F 2016-04-28 11:33:51
          13      9.11394617216e+14        5636249      M 2016-04-28 14:52:07
          14      9.98847233349e+13        5633951      F 2016-04-28 10:06:24
          15          99948393975.0        5620206      F 2016-04-26 08:47:27
          16      8.45743929428e+13        5633121      M 2016-04-28 08:51:47
          17      1.47949661912e+13        5633460      F 2016-04-28 09:28:57
          18      1.71353782452e+13        5621836      F 2016-04-26 10:54:18
          19      7.22328918422e+12        5640433      F 2016-04-29 10:43:14
```

```
20      6.22257462899e+14        5626083      F 2016-04-27 07:51:14
21      1.21548437528e+13        5628338      F 2016-04-27 10:50:45
22      8.63229818888e+14        5616091      M 2016-04-25 13:29:16
23      2.13753979426e+14        5634142      F 2016-04-28 10:27:05
24      8.73485799688e+12        5641780      F 2016-04-29 14:19:19
25       5.8193699788e+12        5624020      M 2016-04-26 15:04:17
26          25787851512.0        5641781      F 2016-04-29 14:19:42
27      1.21548437528e+13        5628345      F 2016-04-27 10:51:45
28      5.92617169253e+12        5642400      M 2016-04-29 15:48:02
29      1.22577616366e+12        5642186      F 2016-04-29 15:16:29
...                   ...            ...    ...                 ...
110497  7.93589177751e+14        5757745      M 2016-06-01 09:46:33
110498  9.43365361457e+13        5787655      F 2016-06-08 10:21:14
110499  8.21969177626e+14        5757697      F 2016-06-01 09:42:56
110500  4.43438443335e+14        5787233      F 2016-06-08 09:35:13
110501     454425189389.0        5758133      M 2016-06-01 10:19:12
110502  7.31622885365e+14        5787937      F 2016-06-08 10:50:42
110503  2.36218168228e+13        5759473      F 2016-06-01 13:00:36
110504  9.94798255557e+12        5788052      F 2016-06-08 11:06:21
110505   5.6673438856e+13        5758455      F 2016-06-01 10:45:50
110506     897388334326.0        5758779      M 2016-06-01 11:09:20
110507  4.76946211847e+14        5786918      F 2016-06-08 09:04:18
110508  9.43365361457e+13        5757656      F 2016-06-01 09:41:00
110509  4.95296829376e+14        5786750      M 2016-06-08 08:50:51
110510  2.36218168228e+13        5757587      F 2016-06-01 09:35:48
110511     823599626588.0        5786742      F 2016-06-08 08:50:20
110512  9.87624564474e+13        5786368      F 2016-06-08 08:20:01
110513  8.67477849953e+13        5785964      M 2016-06-08 07:52:55
110514  2.69568517714e+12        5786567      F 2016-06-08 08:35:31
110515  6.45634214296e+14        5778621      M 2016-06-06 15:58:05
110516  6.92377244368e+13        5780205      F 2016-06-07 07:45:16
110517  5.57494241893e+12        5780122      F 2016-06-07 07:38:34
110518  7.26331492534e+13        5630375      F 2016-04-27 15:15:06
110519  6.54238778939e+13        5630447      F 2016-04-27 15:23:14
110520  9.96997666246e+14        5650534      F 2016-05-03 07:51:47
110521  3.63553377464e+13        5651072      F 2016-05-03 08:23:40
110522  2.57213436929e+12        5651768      F 2016-05-03 09:15:35
110523  3.59626632874e+12        5650093      F 2016-05-03 07:27:33
110524  1.55766317299e+13        5630692      F 2016-04-27 16:03:52
110525  9.21349314356e+13        5630323      F 2016-04-27 15:09:23
110526  3.77511518121e+14        5629448      F 2016-04-27 13:30:56

       AppointmentDay  Age      Neighbourhood  Scholarship  Hypertension  \
0          2016-04-29   62    JARDIM DA PENHA            0             1
1          2016-04-29   56    JARDIM DA PENHA            0             0
2          2016-04-29   62      MATA DA PRAIA            0             0
3          2016-04-29    8  PONTAL DE CAMBURI            0             0
4          2016-04-29   56    JARDIM DA PENHA            0             1
```

```
5         2016-04-29  76        REPÚBLICA            0             1
6         2016-04-29  23       GOIABEIRAS            0             0
7         2016-04-29  39       GOIABEIRAS            0             0
8         2016-04-29  21       ANDORINHAS            0             0
9         2016-04-29  19        CONQUISTA            0             0
10        2016-04-29  30   NOVA PALESTINA            0             0
11        2016-04-29  29   NOVA PALESTINA            0             0
12        2016-04-29  22   NOVA PALESTINA            1             0
13        2016-04-29  28   NOVA PALESTINA            0             0
14        2016-04-29  54   NOVA PALESTINA            0             0
15        2016-04-29  15   NOVA PALESTINA            0             0
16        2016-04-29  50   NOVA PALESTINA            0             0
17        2016-04-29  40        CONQUISTA            1             0
18        2016-04-29  30   NOVA PALESTINA            1             0
19        2016-04-29  46         DA PENHA            0             0
20        2016-04-29  30   NOVA PALESTINA            0             0
21        2016-04-29   4        CONQUISTA            0             0
22        2016-04-29  13        CONQUISTA            0             0
23        2016-04-29  46        CONQUISTA            0             0
24        2016-04-29  65       TABUAZEIRO            0             0
25        2016-04-29  46        CONQUISTA            0             1
26        2016-04-29  45   BENTO FERREIRA            0             1
27        2016-04-29   4        CONQUISTA            0             0
28        2016-04-29  51        SÃO PEDRO            0             0
29        2016-04-29  32     SANTA MARTHA            0             0
...              ... ...              ...          ...           ...
110497    2016-06-01  76      MARIA ORTIZ            0             0
110498    2016-06-08  59      MARIA ORTIZ            0             0
110499    2016-06-01  66      MARIA ORTIZ            0             1
110500    2016-06-08  59      MARIA ORTIZ            0             0
110501    2016-06-01  44      MARIA ORTIZ            0             0
110502    2016-06-08  22       GOIABEIRAS            0             0
110503    2016-06-01  64     SOLON BORGES            0             0
110504    2016-06-08   4      MARIA ORTIZ            0             0
110505    2016-06-01  55      MARIA ORTIZ            0             0
110506    2016-06-01   5      MARIA ORTIZ            0             0
110507    2016-06-08   0      MARIA ORTIZ            0             0
110508    2016-06-01  59      MARIA ORTIZ            0             0
110509    2016-06-08  33      MARIA ORTIZ            0             0
110510    2016-06-01  64     SOLON BORGES            0             0
110511    2016-06-08  14      MARIA ORTIZ            0             0
110512    2016-06-08  41      MARIA ORTIZ            0             0
110513    2016-06-08   2  ANTÔNIO HONÓRIO            0             0
110514    2016-06-08  58      MARIA ORTIZ            0             0
110515    2016-06-08  33      MARIA ORTIZ            0             1
110516    2016-06-08  37      MARIA ORTIZ            0             0
110517    2016-06-07  19      MARIA ORTIZ            0             0
110518    2016-06-07  50      MARIA ORTIZ            0             0
```

```
110519     2016-06-07   22        MARIA ORTIZ            0             0
110520     2016-06-07   42        MARIA ORTIZ            0             0
110521     2016-06-07   53        MARIA ORTIZ            0             0
110522     2016-06-07   56        MARIA ORTIZ            0             0
110523     2016-06-07   51        MARIA ORTIZ            0             0
110524     2016-06-07   21        MARIA ORTIZ            0             0
110525     2016-06-07   38        MARIA ORTIZ            0             0
110526     2016-06-07   54        MARIA ORTIZ            0             0

        Diabetes  Alcoholism  handicap  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
5              0           0         0             0      No
6              0           0         0             0     Yes
7              0           0         0             0     Yes
8              0           0         0             0      No
9              0           0         0             0      No
10             0           0         0             0      No
11             0           0         0             1     Yes
12             0           0         0             0      No
13             0           0         0             0      No
14             0           0         0             0      No
15             0           0         0             1      No
16             0           0         0             0      No
17             0           0         0             0     Yes
18             0           0         0             1      No
19             0           0         0             0      No
20             0           0         0             0     Yes
21             0           0         0             0     Yes
22             0           0         0             1     Yes
23             0           0         0             0      No
24             0           0         0             0      No
25             0           0         0             1      No
26             0           0         0             0      No
27             0           0         0             0      No
28             0           0         0             0      No
29             0           0         0             0      No
...          ...         ...       ...           ...     ...
110497         0           0         0             0      No
110498         0           0         0             0      No
110499         1           0         0             0      No
110500         0           0         0             0      No
110501         0           0         0             0      No
110502         0           0         0             0      No
110503         0           0         0             0      No
```

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

[110527 rows x 14 columns]
```

```
In [212]: #use this line to check the result
          #df_appointments.Age[false_value]
```

```
In [213]: # use this line to check the result
          #df_appointments['Age']
```

## Exploratory Data Analysis Now I have done with cleaning my data It is clean and clear. I will move on to exploration and compute statistics to answer the question.

1.1.2 Research Question 1: Do the female patient care more about their appointments?

In this question I will find whather the female patients are more committed to their appointments

```
In [214]: #findinging the number of female and male patients
          df_appointments['Gender'].value_counts()
```

```
Out[214]: F    71840
          M    38687
          Name: Gender, dtype: int64
```

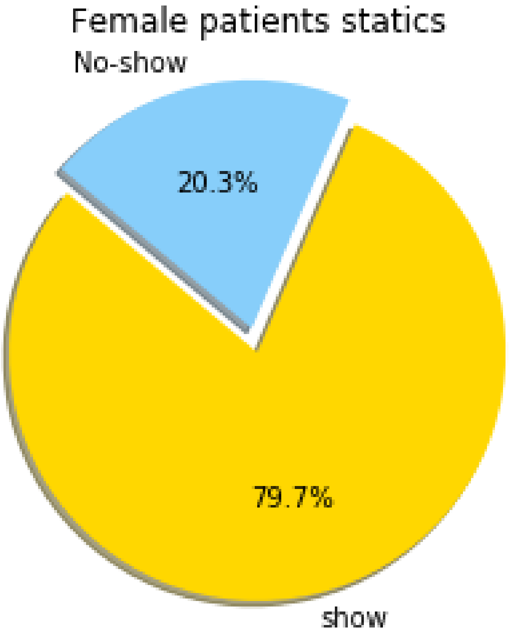
```
In [215]: #calculting the number of female patients who came to their appointment
          df_F_show = df_appointments.loc[(df_appointments['Gender'] == "F") & (df_appointments[
          F_show = df_F_show['PatientId'].count()
          F_show
```

```
Out[215]: 57246
```

```
In [216]: #calculting the number of female patients who skip to their appointment
          df_F_Noshow = df_appointments.loc[(df_appointments['Gender'] == "F") & (df_appointment
          F_Noshow = df_F_Noshow['PatientId'].count()
          F_Noshow
```

```
Out[216]: 14594
```

```
In [217]: #draw a pie chart to illustrate the result

          #import matplotlib
          import matplotlib.pyplot as plt

          # Data to plot
          labels = 'show', 'No-show'
          sizes = [F_show, F_Noshow]
          colors = ['gold', 'lightskyblue']
          explode = (0.1, 0)  # explode 1st slice

          # Plot
          plt.pie(sizes, explode=explode, labels=labels, colors=colors,
                  autopct='%1.1f%%', shadow=True, startangle=140)
          plt.title('Female patients statics')
          plt.axis('equal')
          plt.show()
```

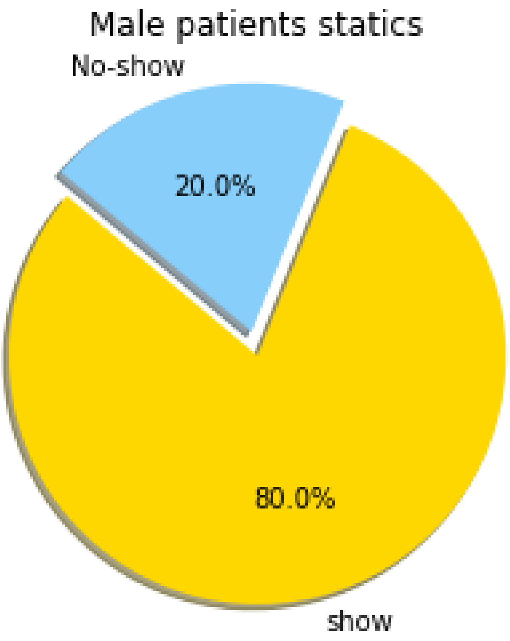


this graph show us that Almost 80% of female patients tend to attend thier appointments

```
In [218]: #calculting the number of male patients who came to their appointment
          df_M_show = df_appointments.loc[(df_appointments['Gender'] == "M") & (df_appointments[
          M_show = df_M_show['PatientId'].count()
          M_show

Out[218]: 30962
```

```
In [219]: #calculting the number of male patients who skip their appointment
          df_M_Noshow = df_appointments.loc[(df_appointments['Gender'] == "M") & (df_appointment
          M_Noshow = df_M_Noshow['PatientId'].count()
          M_Noshow

Out[219]: 7725
```

```
In [220]: #draw a pie chart to illustrate the result
          # Data to plot
          labels = 'show', 'No-show'
          sizes = [M_show, M_Noshow]
          colors = ['gold', 'lightskyblue']
          explode = (0.1, 0)  # explode 1st slice

          # Plot
          plt.pie(sizes, explode=explode, labels=labels, colors=colors,
                  autopct='%1.1f%%', shadow=True, startangle=140)

          plt.title('Male patients statics')
          plt.axis('equal')
          plt.show()
```



this graph show us that 80% of male patients tend to attend thier appointments

1.1.3 therefore,

gender does not have an effect on patients commitment toward their appointments

1.1.4 Research Question 2: Does having a Diabetes affect on the patient commitment of their appointments?

In this question I will find wether or not diabetics tend to skip thier appoientments

```
In [221]: #findinging the number of diabetics
          df_appointments['Diabetes'].value_counts()
```

```
Out[221]: 0    102584
          1      7943
          Name: Diabetes, dtype: int64
```

```
In [222]: #calculting the number of diabetics who came to their appointment
          df_D_show = df_appointments.loc[(df_appointments['Diabetes'] == 1) & (df_appointments[
          D_show_count = df_D_show['PatientId'].count()
          D_show_count
```

```
Out[222]: 6513
```

```
In [223]: #calculting the number of diabetics who skip their appointment
          df_D_Noshow = df_appointments.loc[(df_appointments['Diabetes'] == 1) & (df_appointment
          D_Noshow_count = df_D_Noshow['PatientId'].count()
          D_Noshow_count
```

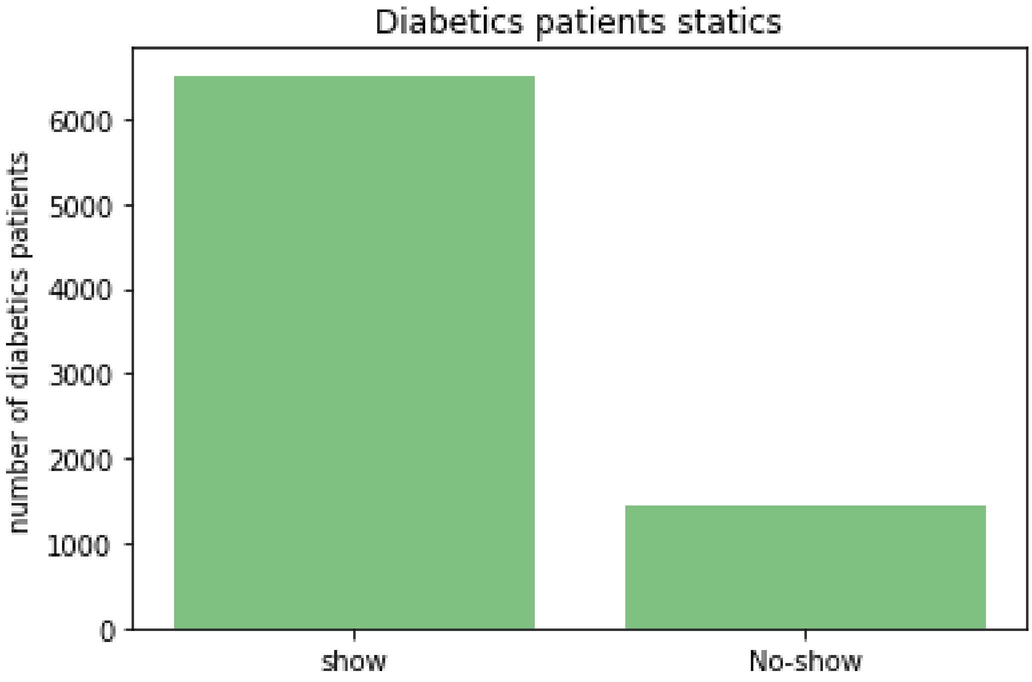
```
Out[223]: 1430
```

```
In [224]: #total diabetics
          D_total =  7943
          #calculting the percentage for both diabetics who attend their appointment and who's n
          show_count_Percentage = int((D_show_count/D_total)*100)
          print('Percentage of diabetics who comes to their appointment: ',show_count_Percentage
          Noshow_count_Percentage = int((D_Noshow_count/D_total)*100)
          print('Percentage of diabetics who skip their appointment:'  ,Noshow_count_Percentage
```

```
Percentage of diabetics who comes to their appointment:  81 %
Percentage of diabetics who skip their appointment: 18 %
```

```
In [225]: #draw a Bar chart to illustrate the result
          objects = ('show', 'No-show')
          y_pos = np.arange(len(objects))
          values = [D_show_count,D_Noshow_count]
```

```
plt.bar( y_pos, values, align='center', alpha=0.5, color='g')
plt.xticks(y_pos, objects)

plt.ylabel('number of diabetics patients')
plt.title('Diabetics patients statics')

plt.show()
```



1.1.5 As result :

Diabetes does not have an effect on patients commitment toward their appointments

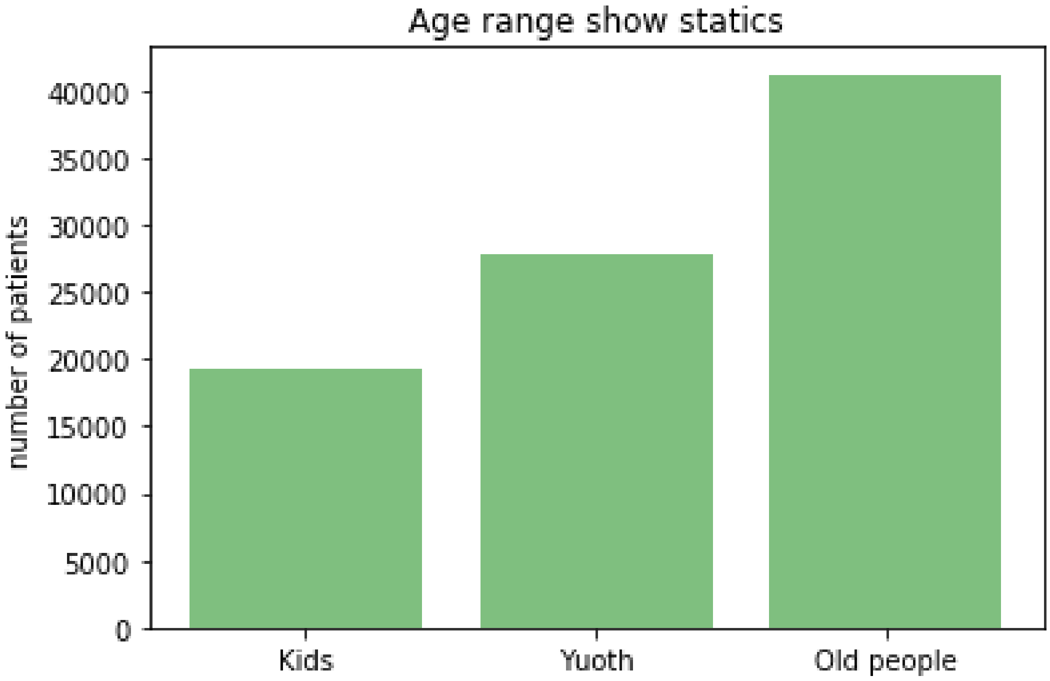
1.1.6 Research Question 3: Which age range is more commitment of their appointments?

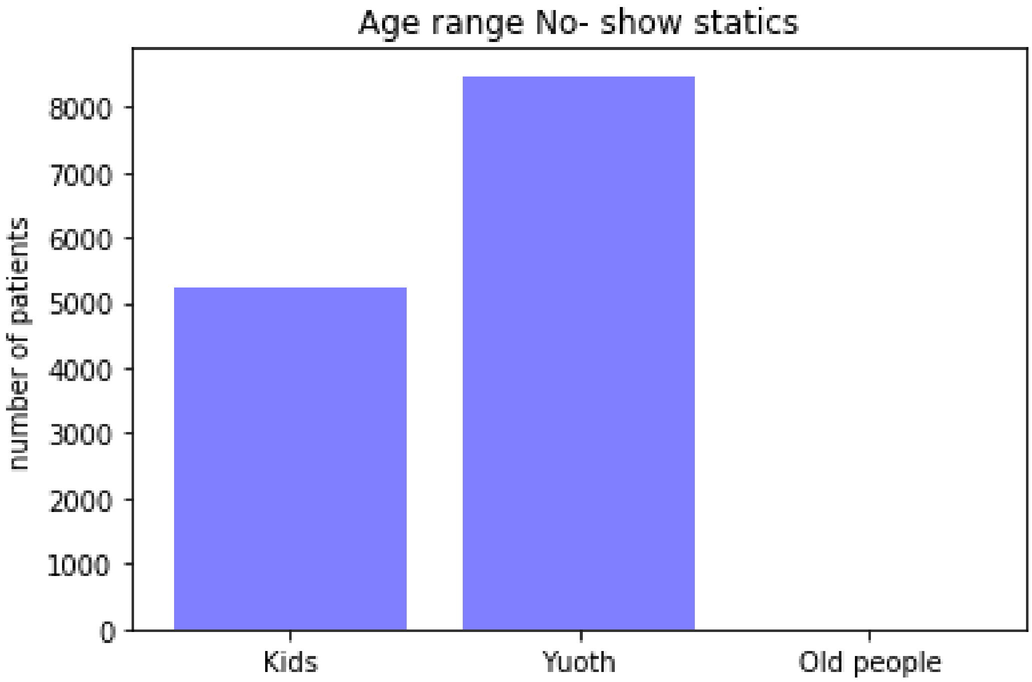
In this Question I will find out which age range (children , youth, older people) tend to attend thier appointments.

```
In [226]: #calculating the number of kids who came to their appointment
          df_kids_show = df_appointments.loc[(df_appointments['Age']> -1) & (df_appointments['Ag
          kids_show_count = df_kids_show['PatientId'].count()
          kids_show_count

Out[226]: 19220
```

```
In [227]: #calculating the number of kids who skip their appointment
          df_kids_Noshow = df_appointments.loc[(df_appointments['Age']> -1) & (df_appointments['
          kids_Noshow_count = df_kids_Noshow['PatientId'].count()
          kids_Noshow_count

Out[227]: 5248

In [228]: #calculating the number of youth who came to their appointment
          df_youth_show = df_appointments.loc[(df_appointments['Age']> 15) & (df_appointments['A
          youth_show_count = df_youth_show['PatientId'].count()
          youth_show_count

Out[228]: 27741

In [229]: #calculating the number of youth who came to their appointment
          df_youth_Noshow = df_appointments.loc[(df_appointments['Age']> 15) & (df_appointments[
          youth_Noshow_count = df_youth_Noshow['PatientId'].count()
          youth_Noshow_count

Out[229]: 8474

In [230]: #calculating the number of old people who came to their appointment
          df_old_show = df_appointments.loc[(df_appointments['Age']> 40) & (df_appointments['Age
          old_show_count = df_old_show['PatientId'].count()
          old_show_count

Out[230]: 41238

In [231]: #calculating the number of old people who came to their appointment
          df_old_Noshow = df_appointments.loc[(df_appointments['Age']> 40) & (df_appointments['A
          old_Noshow_count = df_old_Noshow['PatientId'].count()
          old_Noshow_count

Out[231]: 0

In [232]: #draw a Bar chart to illustrate the result of show
          objects = ('Kids', 'Yuoth','Old people')
          y_pos = np.arange(len(objects))
          values = [kids_show_count, youth_show_count, old_show_count]

          plt.bar( y_pos, values, align='center', alpha=0.5, color='g')
          plt.xticks(y_pos, objects)

          plt.ylabel('number of patients')
          plt.title('Age range show statics')

          plt.show()

          #draw a Bar chart to illustrate the result ob no show
```

```
objects = ('Kids', 'Yuoth','Old people')
y_pos = np.arange(len(objects))
values = [kids_Noshow_count, youth_Noshow_count, old_Noshow_count]

plt.bar( y_pos, values, align='center', alpha=0.5, color='b')
plt.xticks(y_pos, objects)

plt.ylabel('number of patients')
plt.title('Age range No- show statics')

plt.show()
```





1.1.7 As result :

Even though the number of kids and youth patient who attend thier appointment is very high. But all older people came to thier appointments which means they have a higher commitment to their appointments

## Conclusions

the main focus of this report is looking at relationships between patient variables and his commitment towards his appointment. I chose to investigate the relationship between (patient age, diabetes, patent gender) with the patient appointments attendance. I have found that these variables do not affect the patient commitment to his appointments except the age. older people tend to have a higher commitment. I think this relationship is strong and direct since usually the older the human get, they suffer from more diseases

limitations: handicap have anonymous 4 unique values. thier meaning is not clear. therefore, I could not use them in investigation.

I need the appointment location neighbourhood along with the patient neighbourhood to find out if the desteance affact on the patient appointments attendance.

```
In [ ]:
```