Investigate_a_Dataset

October 14, 2022

0.1

1 Projet: Examiner un ensemble de données - No show appointments

1.1 Table of Contents

1.2

1.3 Introduction

1.3.1 Dataset Description

PatientId : identification d'une patient sous traitement médical.

AppointmentId : identification rendez-vous.

Gender : la construction socioculturelle des rôles masculins et féminins et des rapports entre les hommes et les femmes

ScheduledId : journée aménagée ou planifiée selon un programme, emploi du temps du patient.

AppointmenDay : jour rendez-vous fixé à une heure précise ou date du patient.

Age : Durée écoulée entre la naissance d'une personne; moment de la vie correspondant à cette durée

Neighbourhood : un quartier où vivent les patients.

Scholarship : études ou réalisations académiques ; apprendre à un haut niveau.

Hipertension : L'hypertension est une élévation de la pression artérielle, en particulier de la pression diastolique.

Diabetes : l'un de plusieurs troubles caractérisés par une augmentation de la production d'urine.

Alcoholism : addiction la consommation de boisson alcoolisée.

Handcap : patient avec un handicap physique ou mental.

SMS_received : Patient ayant un message.

No-show : patient arrive ou qui n'arrive à leur rendez-vous.

1.3.2 Questions d'analyse

Son état de santé est-il une problèmes pour venir aux rendez-vous prévus?

[1]:
```
# Importer paquets

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

[2]:
```
# mise à jour pandas pour utilisé dataframe.explode() function.
#!pip --upgrade install pandas==0.25.0
```

1.4

1.5 Data Wrangling

1.5.1 Les propriétés générales

Combien de ligne et colonne l'ensemble de donné ?

Combien de valeur manquante l'ensemble de donné ?

Est-ce qu'il y a de patient avoir trois maladie ?

Combien des patients aucune des trois maladies ?

[3]:
```
# importation donnée
df = pd.read_csv('noshowappointments-kagglev2-may-2016.csv')
```

[4]:
```
# Copie le donnée
show_df = df.copy()
```

[5]:
```
# Affiche 5 premier ligne
show_df.head()
```

[5]:

	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. Nombre de ligne et colonne

[6]:
```
show_df.shape
```

[6]:
```
(110527, 14)
```

[7]:
```
# type ensembble de donné
show_df.dtypes
```

[7]:
```
PatientId         float64
AppointmentID       int64
Gender             object
ScheduledDay       object
AppointmentDay     object
Age                 int64
Neighbourhood      object
Scholarship         int64
Hipertension        int64
Diabetes            int64
Alcoholism          int64
Handcap             int64
SMS_received        int64
No-show            object
dtype: object
```

2. Nombre manquante

[8]:
```
# Information sur l'ensemble de donnée
show_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 110527 entries, 0 to 110526
Data columns (total 14 columns):
 #   Column          Non-Null Count   Dtype
```

```
---  ------          --------------   -----
 0   PatientId       110527 non-null  float64
 1   AppointmentID   110527 non-null  int64
 2   Gender          110527 non-null  object
 3   ScheduledDay    110527 non-null  object
 4   AppointmentDay  110527 non-null  object
 5   Age             110527 non-null  int64
 6   Neighbourhood   110527 non-null  object
 7   Scholarship     110527 non-null  int64
 8   Hipertension    110527 non-null  int64
 9   Diabetes        110527 non-null  int64
 10  Alcoholism      110527 non-null  int64
 11  Handcap         110527 non-null  int64
 12  SMS_received    110527 non-null  int64
 13  No-show         110527 non-null  object
dtypes: float64(1), int64(8), object(5)
memory usage: 11.8+ MB
```

3. Voir les nombres de patient présentent 3 maladies

[9]:
```
# patient porte 3 maladies et handicap
df_porte = show_df.query('Hipertension == 1 & Diabetes == 1 & Alcoholism == 1 &␣
 ↪Handcap == 1')

# Nombre de ligne c'est le nombre de patient porte 3 maladie et handicap
df_porte.shape
```

[9]:
```
(12, 14)
```

4. Nombre d'une patient ne présente leur maladie

[10]:
```
# nombre d'une patient aucune de cette maladie et non pas handicap
df_auc = show_df.query('Hipertension == 0 & Diabetes == 0 & Alcoholism == 0')
# Nombre de ligne c'est le nombre de patient aucune 3 maladie et non pas␣
 ↪handicap
df_auc.shape
```

[10]:
```
(85312, 14)
```

5. Nombre de patient avoir au moins une maladie

[11]:
```
# Patient avoir une ou plusieur maladie
df_avoir = show_df.query('Hipertension == 1 or Diabetes == 1 or Alcoholism ==␣
 ↪1')
df_avoir
```

[11]:
```
           PatientId  AppointmentID Gender          ScheduledDay  \
0       2.987250e+13        5642903      F  2016-04-29T18:38:08Z
4       8.841186e+12        5642494      F  2016-04-29T16:07:23Z
```

```
5       9.598513e+13        5626772      F  2016-04-27T08:36:51Z
25      5.819370e+12        5624020      M  2016-04-26T15:04:17Z
26      2.578785e+10        5641781      F  2016-04-29T14:19:42Z
…              …            …    …                   …
110483  1.642781e+12        5769404      F  2016-06-03T08:47:58Z
110492  6.456342e+14        5786741      M  2016-06-08T08:50:19Z
110496  8.544295e+13        5779046      F  2016-06-06T17:35:38Z
110499  8.219692e+14        5757697      F  2016-06-01T09:42:56Z
110515  6.456342e+14        5778621      M  2016-06-06T15:58:05Z

              AppointmentDay  Age    Neighbourhood  Scholarship  Hipertension  \
0       2016-04-29T00:00:00Z   62  JARDIM DA PENHA            0             1
4       2016-04-29T00:00:00Z   56  JARDIM DA PENHA            0             1
5       2016-04-29T00:00:00Z   76        REPÚBLICA            0             1
25      2016-04-29T00:00:00Z   46        CONQUISTA            0             1
26      2016-04-29T00:00:00Z   45   BENTO FERREIRA            0             1
…                       …  …            …          …           …
110483  2016-06-03T00:00:00Z   60   PRAIA DO CANTO            0             1
110492  2016-06-08T00:00:00Z   33      MARIA ORTIZ            0             1
110496  2016-06-08T00:00:00Z   37      MARIA ORTIZ            0             1
110499  2016-06-01T00:00:00Z   66      MARIA ORTIZ            0             1
110515  2016-06-08T00:00:00Z   33      MARIA ORTIZ            0             1

        Diabetes  Alcoholism  Handcap  SMS_received No-show
0              0           0        0             0      No
4              1           0        0             0      No
5              0           0        0             0      No
25             0           0        0             1      No
26             0           0        0             0      No
…          …         …      …           …     …
110483         0           0        0             0      No
110492         0           0        0             0     Yes
110496         0           0        0             0     Yes
110499         1           0        0             0      No
110515         0           0        0             0     Yes

[25215 rows x 14 columns]
```

1.6 Nettoyage des données

Renommer colonne

Convertir header columns to lowercase

Convertir le type de l'ensemble de donnée

1. Renommer les colonnes suivant : Hipertension,Handcap, SMS_received, No-show

[12]:
```
# Renommer colonne
show_df = show_df.rename(columns = {'Hipertension': 'Hypertension', 'Handcap':
 ↪'Handicap','SMS_received':'Smsreceived', 'No-show': 'Noshow'})
# confirme modification
show_df.head(1)
```

[12]:
```
      PatientId  AppointmentID Gender          ScheduledDay  \
0  2.987250e+13        5642903      F  2016-04-29T18:38:08Z

          AppointmentDay  Age    Neighbourhood  Scholarship  Hypertension  \
0  2016-04-29T00:00:00Z   62  JARDIM DA PENHA            0             1

   Diabetes  Alcoholism  Handicap  Smsreceived Noshow
0         0           0         0            0     No
```

2. Convertir l' entête des colonne en miniscule

[13]:
```
# Convertir colonne en miniscule
show_df.rename(columns=lambda x: x.strip().lower(), inplace=True)
# Confirme la modification
show_df.head(1)
```

[13]:
```
      patientid  appointmentid gender          scheduledday  \
0  2.987250e+13        5642903      F  2016-04-29T18:38:08Z

          appointmentday  age    neighbourhood  scholarship  hypertension  \
0  2016-04-29T00:00:00Z   62  JARDIM DA PENHA            0             1

   diabetes  alcoholism  handicap  smsreceived noshow
0         0           0         0            0     No
```

Je fais convertir le colonne en miniscule

3. Convertir la type de quelques ensemble de donnée

[14]:
```
# convertir type ensemble de donnée
show_df.scheduledday = pd.to_datetime(show_df.scheduledday).dt.date.
 ↪astype('datetime64[ns]')
# confirme
show_df.head(1)
```

[14]:
```
      patientid  appointmentid gender scheduledday        appointmentday  age  \
0  2.987250e+13        5642903      F   2016-04-29  2016-04-29T00:00:00Z   62

     neighbourhood  scholarship  hypertension  diabetes  alcoholism  handicap  \
0  JARDIM DA PENHA            0             1         0           0         0

   smsreceived noshow
0            0     No
```

[15]:
```
# Convertir le type appointmentday vers datetime
show_df.appointmentday = show_df.appointmentday.astype('datetime64[ns]')
# confirme
show_df.head(1)
```

[15]:
```
      patientid  appointmentid gender scheduledday appointmentday  age  \
0  2.987250e+13        5642903      F   2016-04-29     2016-04-29   62

     neighbourhood  scholarship  hypertension  diabetes  alcoholism  handicap  \
0  JARDIM DA PENHA            0             1         0           0         0

   smsreceived noshow
0              0     No
```

On cherche l'age minimum

[16]:
```
show_df.age.min()
```

[16]:
```
-1
```

J'utilise query pour chrercher l'âge inférieur Zéro

[17]:
```
show_df.query('age < 0')
```

[17]:
```
          patientid  appointmentid gender scheduledday appointmentday  age  \
99832  4.659432e+14        5775010      F   2016-06-06     2016-06-06   -1

      neighbourhood  scholarship  hypertension  diabetes  alcoholism  \
99832         ROMÃO            0             0         0           0

       handicap  smsreceived noshow
99832         0            0     No
```

Supprime l'âge inférieur zéro car n'est valable

[18]:
```
show_df.drop(show_df[show_df.age < 0].index,inplace=True)
show_df.age.min()
```

[18]:
```
0
```

[19]:
```
# Confirmé
show_df.age.min()
```

[19]:
```
0
```

1.7 L'analyse exploratoire des données

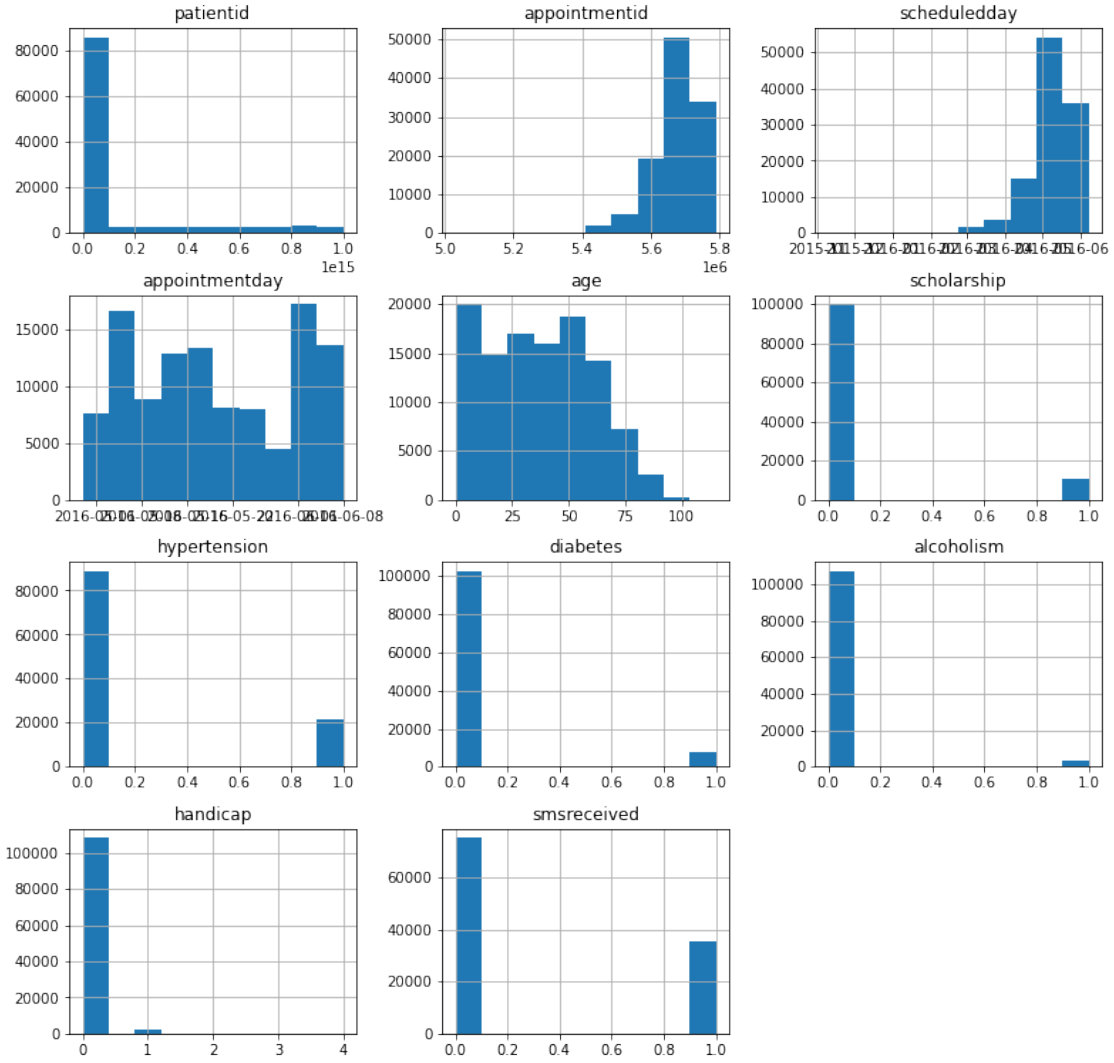
Histogram des ensembles de données

Diagramme circulaire pour les trois maladies

1.7.1 Quelle est la maladie qui a infecté beaucoup de patients ?

Histogramme des ensembles de données

[20]:
```
#figsize = (width = 13, height = 10)
show_df.hist(figsize = (13,13));
```



1.8 Fonction pour diagramme circulaire

On fait le diagramme circulaire pour voir le pourcentage des patients porte ou aucune de maladie
>pnd : patient n'a pas de cette maladie;

pilm : patient il y a maladie;

nbr1 : nombre d'une patient qui n'a pas de maladie;

nbr2 : patient il y a de maladie;

titre : Titre maladie de patient.

[21]:
```
def circle(pnd,pilm,nbr1,nbr2,titre):
    maladie = [pnd,pilm]
    nombre = [nbr1,nbr2]
    figsize = (13,10)
    plt.title(titre, color='b', fontsize=25)
    plt.pie(x=nombre, labels= maladie, autopct ='%.0f%%')
    plt.show()
```

1.8.1 Cherche le nombre de patient il n'y a pas de maladie et le patient il y a de maladie avec value_counts() pour faire le diagramme circulaire

1.Hypertation

[22]:
```
# nombre patients porte et aucune de cette maladie
show_df.hypertension.value_counts()
```

[22]:
```
0    88725
1    21801
Name: hypertension, dtype: int64
```

2. Diabetes

[23]:
```
# nombre patients porte et aucune de cette maladie
show_df.diabetes.value_counts()
```

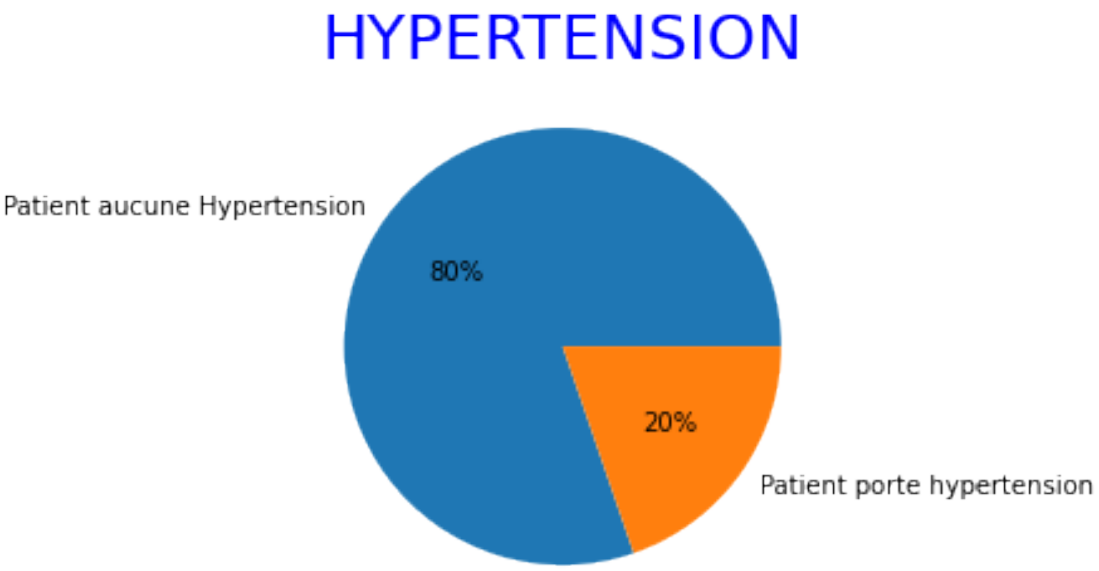
[23]:
```
0    102583
1      7943
Name: diabetes, dtype: int64
```

3. Alcoholism

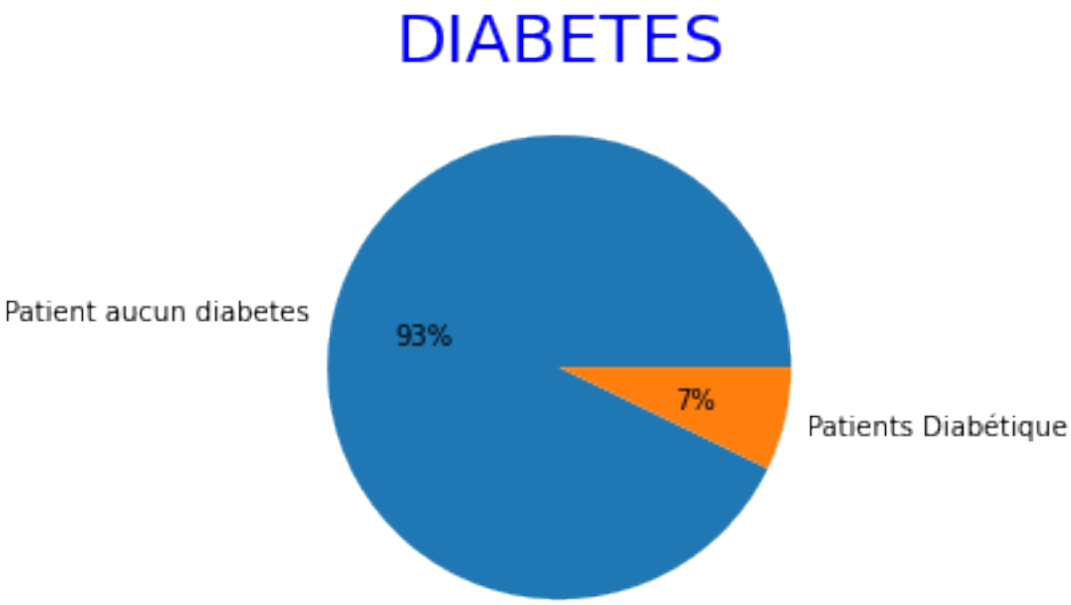
[24]:
```
# nombre patients porte et aucune de cette maladie
show_df.alcoholism.value_counts()
```

[24]:
```
0    107166
1      3360
Name: alcoholism, dtype: int64
```

[25]:
```
circle('Patient aucune Hypertension','Patient porte hypertension',88725,21801,␣
↪'HYPERTENSION')
```



[26]:
```
circle('Patient aucun diabetes','Patients Diabétique',102584,7943, 'DIABETES')
```



[27]:
```
circle('Patient aucune alcoholism','alcoholism',107167,3360, 'ALCOHOLISM')
```

1.9 Voir le nombre de patient manque leur rendez-vous

1. 21801 nombre de patient hypertention

[28]:
```
show_df.query('hypertension == 1 & noshow=="No"')
```

[28]:

	patientid	appointmentid	gender	scheduledday	appointmentday	age	\
0	2.987250e+13	5642903	F	2016-04-29	2016-04-29	62	
4	8.841186e+12	5642494	F	2016-04-29	2016-04-29	56	
5	9.598513e+13	5626772	F	2016-04-27	2016-04-29	76	
25	5.819370e+12	5624020	M	2016-04-26	2016-04-29	46	
26	2.578785e+10	5641781	F	2016-04-29	2016-04-29	45	
…	…	…	…	…	…	…	
110471	3.187119e+14	5781360	F	2016-06-07	2016-06-07	84	
110475	2.123885e+14	5779726	F	2016-06-07	2016-06-07	54	
110476	9.278752e+12	5678369	F	2016-05-10	2016-06-06	80	
110483	1.642781e+12	5769404	F	2016-06-03	2016-06-03	60	
110499	8.219692e+14	5757697	F	2016-06-01	2016-06-01	66	

	neighbourhood	scholarship	hypertension	diabetes	alcoholism	\
0	JARDIM DA PENHA	0	1	0	0	
4	JARDIM DA PENHA	0	1	1	0	
5	REPÚBLICA	0	1	0	0	
25	CONQUISTA	0	1	0	0	
26	BENTO FERREIRA	0	1	0	0	

```
…                 …           …           …       …         …
110471      RESISTÊNCIA           0           1         0           0
110475      RESISTÊNCIA           0           1         0           0
110476      RESISTÊNCIA           0           1         0           0
110483   PRAIA DO CANTO           0           1         0           0
110499      MARIA ORTIZ           0           1         1           0

        handicap  smsreceived noshow
0              0            0     No
4              0            0     No
5              0            0     No
25             0            1     No
26             0            0     No
…            …          …     …
110471         0            0     No
110475         0            0     No
110476         0            1     No
110483         0            0     No
110499         0            0     No

[18029 rows x 14 columns]
```

2. 7943 nombre de patient diabetiques

[29]:
```
show_df.query('diabetes == 1 & noshow=="No"')
```

[29]:
```
           patientid  appointmentid gender scheduledday appointmentday  age  \
4       8.841186e+12        5642494      F   2016-04-29     2016-04-29   56
36      7.298459e+13        5637975      F   2016-04-29     2016-04-29   63
37      1.578132e+12        5637986      F   2016-04-29     2016-04-29   64
41      8.224325e+14        5633339      F   2016-04-28     2016-04-29   71
47      5.894585e+11        5633116      F   2016-04-28     2016-04-29   39
…                …           …    …          …              …  …
110448  2.629184e+14        5756082      F   2016-06-01     2016-06-06   40
110456  4.994742e+12        5772107      F   2016-06-03     2016-06-03   79
110468  3.227475e+11        5763322      F   2016-06-02     2016-06-07   76
110477  2.798494e+13        5673472      F   2016-05-09     2016-06-06   67
110499  8.219692e+14        5757697      F   2016-06-01     2016-06-01   66

          neighbourhood  scholarship  hypertension  diabetes  alcoholism  \
4       JARDIM DA PENHA            0             1         1           0
36        SÃO CRISTÓVÃO            0             1         1           0
37           TABUAZEIRO            1             1         1           0
41              MARUÍPE            0             0         1           0
47              MARUÍPE            0             1         1           0
…                 …          …           …       …         …
110448      RESISTÊNCIA            0             1         1           0
110456      RESISTÊNCIA            0             1         1           0
```

```
110468      RESISTÊNCIA            0             1         1           0
110477      RESISTÊNCIA            0             0         1           0
110499      MARIA ORTIZ            0             1         1           0

        handicap  smsreceived noshow
4              0            0     No
36             0            0     No
37             0            0     No
41             0            0     No
47             0            0     No
…            …            …     …
110448         0            1     No
110456         0            0     No
110468         0            1     No
110477         0            1     No
110499         0            0     No

[6513 rows x 14 columns]
```

3. 3360 nombre de patient alcoholism

[30]:
```
show_df.query('alcoholism == 1 & noshow=="No"')
```

[30]:
```
           patientid  appointmentid gender scheduledday appointmentday  age  \
46      1.379437e+11        5615608      M   2016-04-25     2016-04-29   58
133     3.587186e+12        5580520      M   2016-04-14     2016-04-29   69
186     7.329661e+13        5587737      M   2016-04-15     2016-04-29   66
207     6.359796e+13        5642700      M   2016-04-29     2016-04-29   46
331     8.218631e+11        5639649      F   2016-04-29     2016-04-29   23
…                …              …    …          …              …    …
109912  3.486968e+12        5755218      M   2016-06-01     2016-06-03   56
109947  8.126729e+13        5748881      M   2016-05-31     2016-06-02   62
110071  9.648483e+13        5758772      M   2016-06-01     2016-06-06   54
110167  9.733160e+11        5756807      M   2016-06-01     2016-06-03   64
110174  7.942549e+12        5741957      M   2016-05-30     2016-06-02   59

        neighbourhood  scholarship  hypertension  diabetes  alcoholism  \
46      SÃO CRISTÓVÃO            0             1         0           1
133      PRAIA DO SUÁ            0             0         1           1
186          REDENÇÃO            0             1         0           1
207           MARUÍPE            0             0         0           1
331     SÃO CRISTÓVÃO            1             0         0           1
…                 …          …           …       …         …
109912         BONFIM            0             1         1           1
109947    RESISTÊNCIA            0             0         1           1
110071         BONFIM            0             1         0           1
110167      BOA VISTA            0             0         0           1
110174   SÃO BENEDITO            0             1         0           1
```

```
        handicap  smsreceived noshow
46             0            1     No
133            0            0     No
186            0            0     No
207            0            0     No
331            0            0     No
…            …          …    …
109912         0            0     No
109947         0            0     No
110071         0            1     No
110167         0            0     No
110174         0            1     No

[2683 rows x 14 columns]
```

1.9.1 Quel est l'âge moyenne d'une patient qui n'arrive au rendez-vous?

On va chercher d'abord combien de patient arrive au rendez-vous

[31]:
```
show_df = show_df.query('hypertension == 1 or diabetes == 1 or alcoholism == 1')
```

[32]:
```
absent = show_df.query('noshow=="No"')
absent
```

[32]:
```
           patientid  appointmentid gender scheduledday appointmentday  age  \
0       2.987250e+13        5642903      F   2016-04-29     2016-04-29   62
4       8.841186e+12        5642494      F   2016-04-29     2016-04-29   56
5       9.598513e+13        5626772      F   2016-04-27     2016-04-29   76
25      5.819370e+12        5624020      M   2016-04-26     2016-04-29   46
26      2.578785e+10        5641781      F   2016-04-29     2016-04-29   45
…              …            …    …          …              …  …
110475  2.123885e+14        5779726      F   2016-06-07     2016-06-07   54
110476  9.278752e+12        5678369      F   2016-05-10     2016-06-06   80
110477  2.798494e+13        5673472      F   2016-05-09     2016-06-06   67
110483  1.642781e+12        5769404      F   2016-06-03     2016-06-03   60
110499  8.219692e+14        5757697      F   2016-06-01     2016-06-01   66

          neighbourhood  scholarship  hypertension  diabetes  alcoholism  \
0       JARDIM DA PENHA            0             1         0           0
4       JARDIM DA PENHA            0             1         1           0
5             REPÚBLICA            0             1         0           0
25            CONQUISTA            0             1         0           0
26       BENTO FERREIRA            0             1         0           0
…                   …          …           …       …         …
110475      RESISTÊNCIA            0             1         0           0
110476      RESISTÊNCIA            0             1         0           0
110477      RESISTÊNCIA            0             0         1           0
```

```
110483   PRAIA DO CANTO            0             1         0           0
110499      MARIA ORTIZ            0             1         1           0

        handicap  smsreceived noshow
0              0            0     No
4              0            0     No
5              0            0     No
25             0            1     No
26             0            0     No
…            …            …     …
110475         0            0     No
110476         0            1     No
110477         0            1     No
110483         0            0     No
110499         0            0     No

[20734 rows x 14 columns]
```

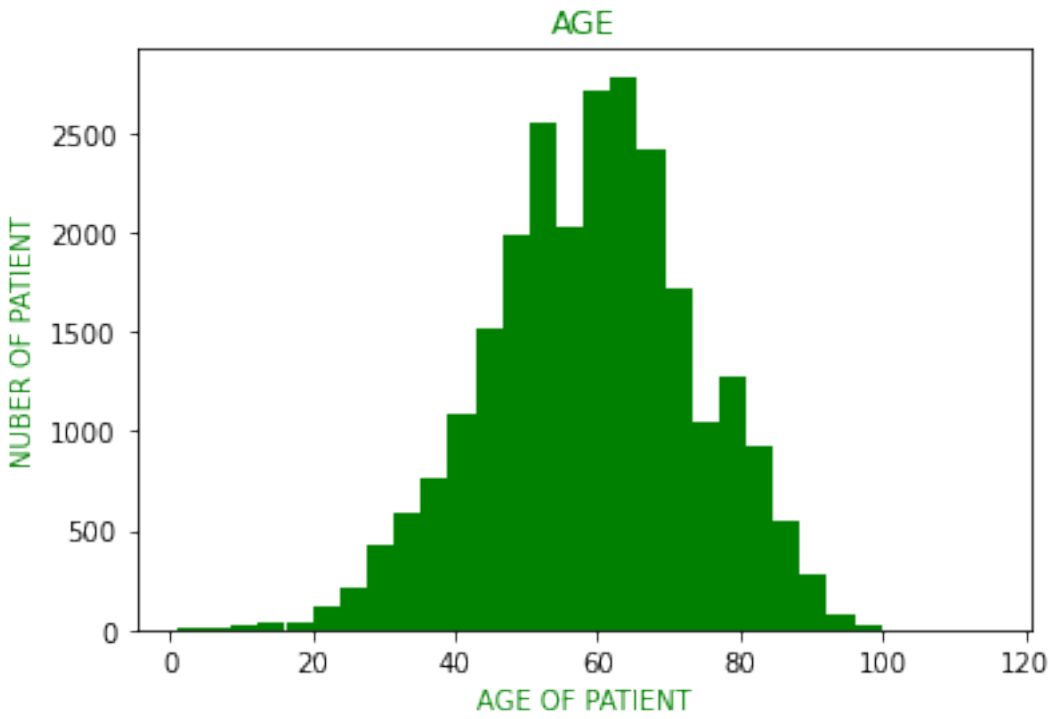
[33]:
```
absent.age.mean()
```

[33]:
```
59.352995080544034
```

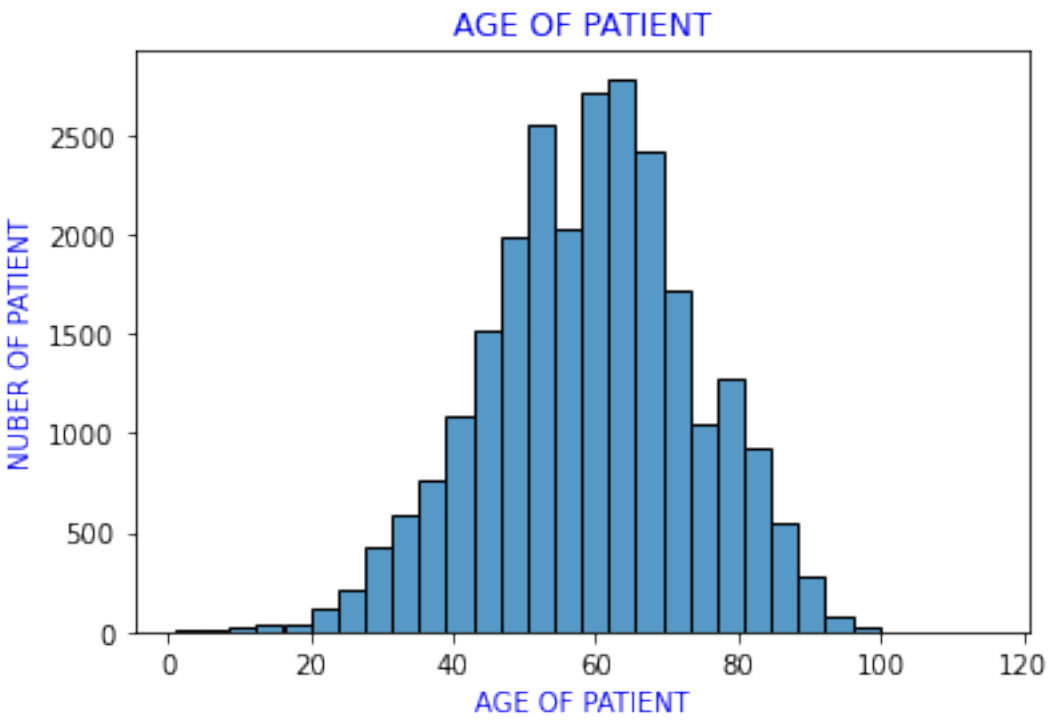
1.10 Répresentation histogramme Age

[34]:
```
# Histogram for age using Matplotlib
def age_patient() :
    plt.figure(figsize=(6,4))
    plt.hist(show_df.age, color='g', bins=30)
    plt.title('AGE', color='g')
    plt.xlabel('AGE OF PATIENT', color='g')
    plt.ylabel('NUBER OF PATIENT', color='g')
    plt.show()
```

[35]:
```
age_patient()
```



[36]:
```
# Histogram for age using Seaborn
plt.figure(figsize=(6,4))
sns.histplot(show_df.age, bins=30)
plt.title('AGE OF PATIENT', color='b')
plt.xlabel('AGE OF PATIENT', color='b')
plt.ylabel('NUBER OF PATIENT', color='b')
plt.show()
```



1.11 Diagramme d'une âge des patients présent et absent sur rendez-vous

[37]:
```
# absent = show_df.query('noshow=="No"') : nombre des patient absent au
  ↪rendez-vous
absent
```

[37]:

	patientid	appointmentid	gender	scheduledday	appointmentday	age	\
0	2.987250e+13	5642903	F	2016-04-29	2016-04-29	62	
4	8.841186e+12	5642494	F	2016-04-29	2016-04-29	56	
5	9.598513e+13	5626772	F	2016-04-27	2016-04-29	76	
25	5.819370e+12	5624020	M	2016-04-26	2016-04-29	46	
26	2.578785e+10	5641781	F	2016-04-29	2016-04-29	45	
…	…	…	…	…	…	…	
110475	2.123885e+14	5779726	F	2016-06-07	2016-06-07	54	
110476	9.278752e+12	5678369	F	2016-05-10	2016-06-06	80	
110477	2.798494e+13	5673472	F	2016-05-09	2016-06-06	67	
110483	1.642781e+12	5769404	F	2016-06-03	2016-06-03	60	
110499	8.219692e+14	5757697	F	2016-06-01	2016-06-01	66	

	neighbourhood	scholarship	hypertension	diabetes	alcoholism	\
0	JARDIM DA PENHA	0	1	0	0	
4	JARDIM DA PENHA	0	1	1	0	
5	REPÚBLICA	0	1	0	0	

```
25            CONQUISTA            0             1         0           0
26       BENTO FERREIRA            0             1         0           0
…                  …           …           …       …         …
110475      RESISTÊNCIA            0             1         0           0
110476      RESISTÊNCIA            0             1         0           0
110477      RESISTÊNCIA            0             0         1           0
110483   PRAIA DO CANTO            0             1         0           0
110499      MARIA ORTIZ            0             1         1           0

        handicap  smsreceived noshow
0              0            0     No
4              0            0     No
5              0            0     No
25             0            1     No
26             0            0     No
…            …          …     …
110475         0            0     No
110476         0            1     No
110477         0            1     No
110483         0            0     No
110499         0            0     No

[20734 rows x 14 columns]
```

[38]:
```
# Les patient present lors d'une rendez-vous
present = show_df.query('noshow == "Yes"')
present
```

[38]:
```
           patientid  appointmentid gender scheduledday appointmentday  age  \
44      2.741649e+11        5635414      F   2016-04-28     2016-04-29   78
126     9.447582e+14        5633576      F   2016-04-28     2016-04-29   67
131     4.755938e+13        5637150      M   2016-04-28     2016-04-29   29
212     4.266984e+14        5642059      M   2016-04-29     2016-04-29   62
270     8.617228e+12        5620528      F   2016-04-26     2016-04-29   45
…                …              …    …           …              …  …
110386  2.957279e+12        5582576      F   2016-04-14     2016-06-01   48
110399  9.437123e+13        5692938      F   2016-05-12     2016-06-07   17
110492  6.456342e+14        5786741      M   2016-06-08     2016-06-08   33
110496  8.544295e+13        5779046      F   2016-06-06     2016-06-08   37
110515  6.456342e+14        5778621      M   2016-06-06     2016-06-08   33

        neighbourhood  scholarship  hypertension  diabetes  alcoholism  \
44      SÃO CRISTÓVÃO            0             1         1           0
126      PRAIA DO SUÁ            0             0         1           0
131      PRAIA DO SUÁ            0             0         0           1
212     SANTOS DUMONT            0             1         1           0
270         CARATOÍRA            1             1         0           0
```

```
…                …           …           …       …          …
110386    RESISTÊNCIA           0           1       0          0
110399    RESISTÊNCIA           0           1       0          0
110492    MARIA ORTIZ           0           1       0          0
110496    MARIA ORTIZ           0           1       0          0
110515    MARIA ORTIZ           0           1       0          0

        handicap  smsreceived noshow
44             0            0    Yes
126            0            0    Yes
131            0            0    Yes
212            0            0    Yes
270            0            1    Yes
…            …          …     …
110386         0            1    Yes
110399         0            1    Yes
110492         0            0    Yes
110496         0            0    Yes
110515         0            0    Yes

[4481 rows x 14 columns]
```

Statistiques descriptives

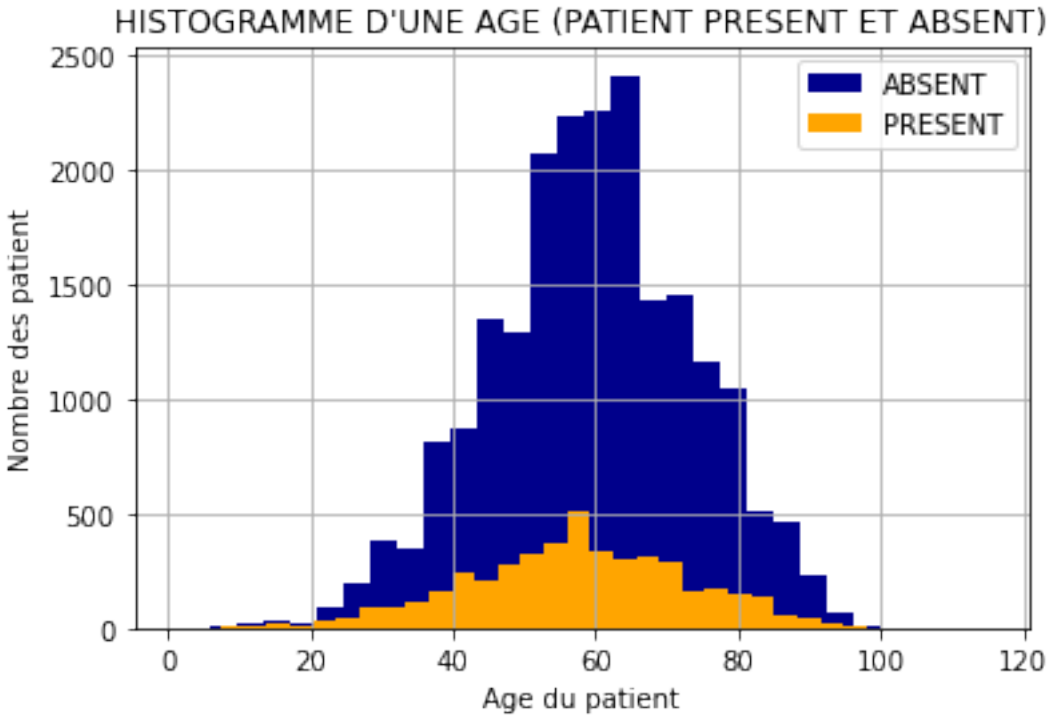
[39]:
```
absent.age.describe()
```

[39]:
```
count    20734.000000
mean        59.352995
std         14.367270
min          2.000000
25%         50.000000
50%         60.000000
75%         69.000000
max        115.000000
Name: age, dtype: float64
```

[40]:
```
present.age.describe()
```

[40]:
```
count    4481.000000
mean       57.364204
std        15.478612
min         1.000000
25%        48.000000
50%        57.000000
75%        68.000000
max        98.000000
Name: age, dtype: float64
```

[41]:
```
# Histogramme d'âge de patient
absent.age.hist(color = 'darkblue', bins=30, label="ABSENT")
present.age.hist(color = 'orange',bins=30,  label="PRESENT")
plt.legend();
plt.title("HISTOGRAMME D'UNE AGE (PATIENT PRESENT ET ABSENT)")
plt.xlabel("Age du patient")
plt.ylabel("Nombre des patient");
plt.savefig('noshow.png')
plt.show()
```



La plupart des patient ne sont pas présent au rendez-vous. L'âge moyenne d'une patient présent est 57 ans et les absent est 59 ans

1.12

1.13 Conclusions

La santé de patient à l'origine de manque rendez-vous, peut-être il souffre beaucoup à date de rendez-vous. Il y a 12 patients porte trois maladie et Handicap, 84115 patients aucune de trois maladie et n'ont pas handicap

59 ans l'âge moyenne de patient qui n'arrive au rendez-vous, il peut oublié la date ou heure.

1.14 Limite

84115 personnes enreigistres sans maladie et ils n'ont pas handicap donc il y a d'erreur sur cette donnée.

[42]:
```
# Convert to html
#from subprocess import call
#call(['python', '-m', 'nbconvert', 'Investigate_a_Dataset.ipynb'])
```