Project: Patient Appointment No Shows Factors, EDA

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. The main qeustion we are trying to answer here is why 30% of patients miss their scheduled appointment. We are trying to predict the most important factors that affect the atendance of the patient.

Some questions we can ask to help us explore the data:

1. Does the patient gender has a realation with the atendance?
2. Does the neighborhood play a role in making patients don't show up? "Location of the hospital"
3. Which pateints show up more? Does old age take care of their health more than youth?
4. Does the disease type affect the patient's show up?

In [1]:
```
#import the liberaries that we need for analysis
import pandas as pd                    #for dealing with dataframs
import numpy as np                     #for scientific compution and arraies
import matplotlib.pyplot as plt        #fro visualization
import seaborn as sns                  #fro better visualization
%matplotlib inline

import warnings
warnings.filterwarnings('ignore')
```

Data Wrangling

General Properties

In [2]:
```
base_data=pd.read_csv('KaggleV2-May-2016.csv')
```

In [3]:
```
base_data
```

Out[3]:

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbour
0	2.987250e+13	5642903	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDI PE
1	5.589978e+14	5642503	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDI PE
2	4.262962e+12	5642549	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MAT P
3	8.679512e+11	5642828	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTA CAM
4	8.841186e+12	5642494	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDI PE
...	...	...	...	...	...	...	
110522	2.572134e+12	5651768	F	2016-05-03T09:15:35Z	2016-06-07T00:00:00Z	56	MARIA C
110523	3.596266e+12	5650093	F	2016-05-03T07:27:33Z	2016-06-07T00:00:00Z	51	MARIA C
110524	1.557663e+13	5630692	F	2016-04-27T16:03:52Z	2016-06-07T00:00:00Z	21	MARIA C
110525	9.213493e+13	5630323	F	2016-04-27T15:09:23Z	2016-06-07T00:00:00Z	38	MARIA C
110526	3.775115e+14	5629448	F	2016-04-27T13:30:56Z	2016-06-07T00:00:00Z	54	MARIA C

110527 rows × 14 columns

In [4]:
```
base_data.shape
```

Out[4]:
```
(110527, 14)
```

In [5]:
```
base_data.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
```

In [6]:
```
base_data.duplicated().sum()
```

Out[6]:
```
0
```

In [7]:
```
#modifaying the data
base_data['ScheduledDay']=pd.to_datetime(base_data['ScheduledDay']).dt.date.as
base_data['AppointmentDay']=pd.to_datetime(base_data['AppointmentDay']).dt.dat
```

In [8]:
```
print(base_data)
```

```
           PatientId  AppointmentID Gender ScheduledDay AppointmentDay  Age
\
0       2.987250e+13        5642903      F   2016-04-29     2016-04-29   62
1       5.589978e+14        5642503      M   2016-04-29     2016-04-29   56
2       4.262962e+12        5642549      F   2016-04-29     2016-04-29   62
3       8.679512e+11        5642828      F   2016-04-29     2016-04-29    8
4       8.841186e+12        5642494      F   2016-04-29     2016-04-29   56
...              ...            ...    ...          ...            ...  ...
110522  2.572134e+12        5651768      F   2016-05-03     2016-06-07   56
110523  3.596266e+12        5650093      F   2016-05-03     2016-06-07   51
110524  1.557663e+13        5630692      F   2016-04-27     2016-06-07   21
110525  9.213493e+13        5630323      F   2016-04-27     2016-06-07   38
110526  3.775115e+14        5629448      F   2016-04-27     2016-06-07   54

             Neighbourhood  Scholarship  Hipertension  Diabetes  Alcoholism
\
0          JARDIM DA PENHA            0             1         0           0
1          JARDIM DA PENHA            0             0         0           0
2            MATA DA PRAIA            0             0         0           0
3        PONTAL DE CAMBURI            0             0         0           0
4          JARDIM DA PENHA            0             1         1           0
...                    ...          ...           ...       ...         ...
110522         MARIA ORTIZ            0             0         0           0
110523         MARIA ORTIZ            0             0         0           0
110524         MARIA ORTIZ            0             0         0           0
110525         MARIA ORTIZ            0             0         0           0
110526         MARIA ORTIZ            0             0         0           0

        Handcap  SMS_received No-show
0             0             0      No
1             0             0      No
2             0             0      No
3             0             0      No
4             0             0      No
...         ...           ...     ...
110522        0             1      No
110523        0             1      No
110524        0             1      No
110525        0             1      No
110526        0             1      No

[110527 rows x 14 columns]
```

In [9]:
```
type(base_data)
```

Out[9]:
```
pandas.core.frame.DataFrame
```

In [10]:
```
base_data.columns
```

Out[10]:
```
Index(['PatientId', 'AppointmentID', 'Gender', 'ScheduledDay',
       'AppointmentDay', 'Age', 'Neighbourhood', 'Scholarship', 'Hipertensio
n',
       'Diabetes', 'Alcoholism', 'Handcap', 'SMS_received', 'No-show'],
      dtype='object')
```

In [11]:
```
base_data.dtypes
```

Out[11]:
```
PatientId               float64
AppointmentID             int64
Gender                   object
ScheduledDay      datetime64[ns]
AppointmentDay    datetime64[ns]
Age                       int64
Neighbourhood            object
Scholarship               int64
Hipertension              int64
Diabetes                  int64
Alcoholism                int64
Handcap                   int64
SMS_received              int64
No-show                  object
dtype: object
```

In [12]:
```
base_data.isnull()
```

Out[12]:

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhoo
0	False	False	False	False	False	False	Fals
1	False	False	False	False	False	False	Fals
2	False	False	False	False	False	False	Fals
3	False	False	False	False	False	False	Fals
4	False	False	False	False	False	False	Fals
...	...	...	...	...	...	...	
110522	False	False	False	False	False	False	Fals
110523	False	False	False	False	False	False	Fals
110524	False	False	False	False	False	False	Fals
110525	False	False	False	False	False	False	Fals
110526	False	False	False	False	False	False	Fals

110527 rows × 14 columns

In [13]:
```
base_data.isnull().sum()
```

Out[13]:
```
PatientId         0
AppointmentID     0
Gender            0
ScheduledDay      0
AppointmentDay    0
Age               0
Neighbourhood     0
Scholarship       0
Hipertension      0
Diabetes          0
Alcoholism        0
Handcap           0
SMS_received      0
No-show           0
dtype: int64
```

In [14]:
```
base_data.nunique()
```

Out[14]:
```
PatientId          62299
AppointmentID     110527
Gender                 2
ScheduledDay         111
AppointmentDay        27
Age                  104
Neighbourhood         81
Scholarship            2
Hipertension           2
Diabetes               2
Alcoholism             2
Handcap                5
SMS_received           2
No-show                2
dtype: int64
```

In [15]:
```
base_data.head(12)
```

Out[15]:

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood
0	2.987250e+13	5642903	F	2016-04-29	2016-04-29	62	JARDIM DA PENHA
1	5.589978e+14	5642503	M	2016-04-29	2016-04-29	56	JARDIM DA PENHA
2	4.262962e+12	5642549	F	2016-04-29	2016-04-29	62	MATA DA PRAIA
3	8.679512e+11	5642828	F	2016-04-29	2016-04-29	8	PONTAL DE CAMBUR
4	8.841186e+12	5642494	F	2016-04-29	2016-04-29	56	JARDIM DA PENHA
5	9.598513e+13	5626772	F	2016-04-27	2016-04-29	76	REPÚBLICA
6	7.336882e+14	5630279	F	2016-04-27	2016-04-29	23	GOIABEIRAS
7	3.449833e+12	5630575	F	2016-04-27	2016-04-29	39	GOIABEIRAS
8	5.639473e+13	5638447	F	2016-04-29	2016-04-29	21	ANDORINHAS
9	7.812456e+13	5629123	F	2016-04-27	2016-04-29	19	CONQUISTA
10	7.345362e+14	5630213	F	2016-04-27	2016-04-29	30	NOVA PALESTINA
11	7.542951e+12	5620163	M	2016-04-26	2016-04-29	29	NOVA PALESTINA

In [16]:
```
base_data.tail(12)
```

Out[16]:

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbour
110515	6.456342e+14	5778621	M	2016-06-06	2016-06-08	33	MARIA C
110516	6.923772e+13	5780205	F	2016-06-07	2016-06-08	37	MARIA C
110517	5.574942e+12	5780122	F	2016-06-07	2016-06-07	19	MARIA C
110518	7.263315e+13	5630375	F	2016-04-27	2016-06-07	50	MARIA C
110519	6.542388e+13	5630447	F	2016-04-27	2016-06-07	22	MARIA C
110520	9.969977e+14	5650534	F	2016-05-03	2016-06-07	42	MARIA C
110521	3.635534e+13	5651072	F	2016-05-03	2016-06-07	53	MARIA C
110522	2.572134e+12	5651768	F	2016-05-03	2016-06-07	56	MARIA C
110523	3.596266e+12	5650093	F	2016-05-03	2016-06-07	51	MARIA C
110524	1.557663e+13	5630692	F	2016-04-27	2016-06-07	21	MARIA C
110525	9.213493e+13	5630323	F	2016-04-27	2016-06-07	38	MARIA C
110526	3.775115e+14	5629448	F	2016-04-27	2016-06-07	54	MARIA C

In [17]:
```
base_data['Gender'].value_counts()
```

Out[17]:
```
F    71840
M    38687
Name: Gender, dtype: int64
```

In [19]:
```
base_data['Age'].value_counts()
```

Out[19]:
```
 0      3539
 1      2273
 52     1746
 49     1652
 53     1651
        ...
 115       5
 100       4
 102       2
 99        1
-1         1
Name: Age, Length: 104, dtype: int64
```

Age can't be negative so we will drop the -1

In [20]:
```
base_data = base_data[base_data['Age'] >= 0] # drop negative age
```

In [21]:
```
base_data['Age'].value_counts() # make sure there is no negative age
```

Out[21]:
```
0      3539
1      2273
52     1746
49     1652
53     1651
       ...
98        6
115       5
100       4
102       2
99        1
Name: Age, Length: 103, dtype: int64
```

In [22]:
```
base_data.describe()      #summary statistics
```

Out[22]:

	PatientId	AppointmentID	Age	Scholarship	Hipertension	Diabete
count	1.105260e+05	1.105260e+05	110526.000000	110526.000000	110526.000000	110526.00000
mean	1.474934e+14	5.675304e+06	37.089219	0.098266	0.197248	0.07186
std	2.560943e+14	7.129544e+04	23.110026	0.297676	0.397923	0.25826
min	3.921784e+04	5.030230e+06	0.000000	0.000000	0.000000	0.00000
25%	4.172536e+12	5.640285e+06	18.000000	0.000000	0.000000	0.00000
50%	3.173184e+13	5.680572e+06	37.000000	0.000000	0.000000	0.00000
75%	9.438963e+13	5.725523e+06	55.000000	0.000000	0.000000	0.00000
max	9.999816e+14	5.790484e+06	115.000000	1.000000	1.000000	1.00000

In [23]:
```
#changing the name of some columns
base_data=base_data.rename(columns={'Hipertension':'hypertension','Handcap':'H
```

In [24]:
```
base_data.columns
```

Out[24]:
```
Index(['PatientId', 'AppointmentID', 'Gender', 'ScheduledDay',
       'AppointmentDay', 'Age', 'Neighbourhood', 'Scholarship', 'hypertensio
n',
       'Diabetes', 'Alcoholism', 'Handicap', 'SMSReceived', 'Noshow'],
      dtype='object')
```

In [25]:
```
#dropping some columns
base_data.drop(['PatientId','AppointmentID'], axis=1 , inplace= True)
```

In [26]:
```
base_data
```

Out[26]:

	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	hypertensio
0	F	2016-04-29	2016-04-29	62	JARDIM DA PENHA	0	
1	M	2016-04-29	2016-04-29	56	JARDIM DA PENHA	0	
2	F	2016-04-29	2016-04-29	62	MATA DA PRAIA	0	
3	F	2016-04-29	2016-04-29	8	PONTAL DE CAMBURI	0	
4	F	2016-04-29	2016-04-29	56	JARDIM DA PENHA	0	
...	...	...	...	...	...	...	
110522	F	2016-05-03	2016-06-07	56	MARIA ORTIZ	0	
110523	F	2016-05-03	2016-06-07	51	MARIA ORTIZ	0	
110524	F	2016-04-27	2016-06-07	21	MARIA ORTIZ	0	
110525	F	2016-04-27	2016-06-07	38	MARIA ORTIZ	0	
110526	F	2016-04-27	2016-06-07	54	MARIA ORTIZ	0	

110526 rows × 12 columns

In [27]:
```
base_data.columns
```

Out[27]:
```
Index(['Gender', 'ScheduledDay', 'AppointmentDay', 'Age', 'Neighbourhood',
       'Scholarship', 'hypertension', 'Diabetes', 'Alcoholism', 'Handicap',
       'SMSReceived', 'Noshow'],
      dtype='object')
```

let's go more into the data and check the columns

In [28]:
```
base_data['SMSReceived'].value_counts()
```

Out[28]:
```
0    75044
1    35482
Name: SMSReceived, dtype: int64
```

we see that most of them didn't recieve SMS

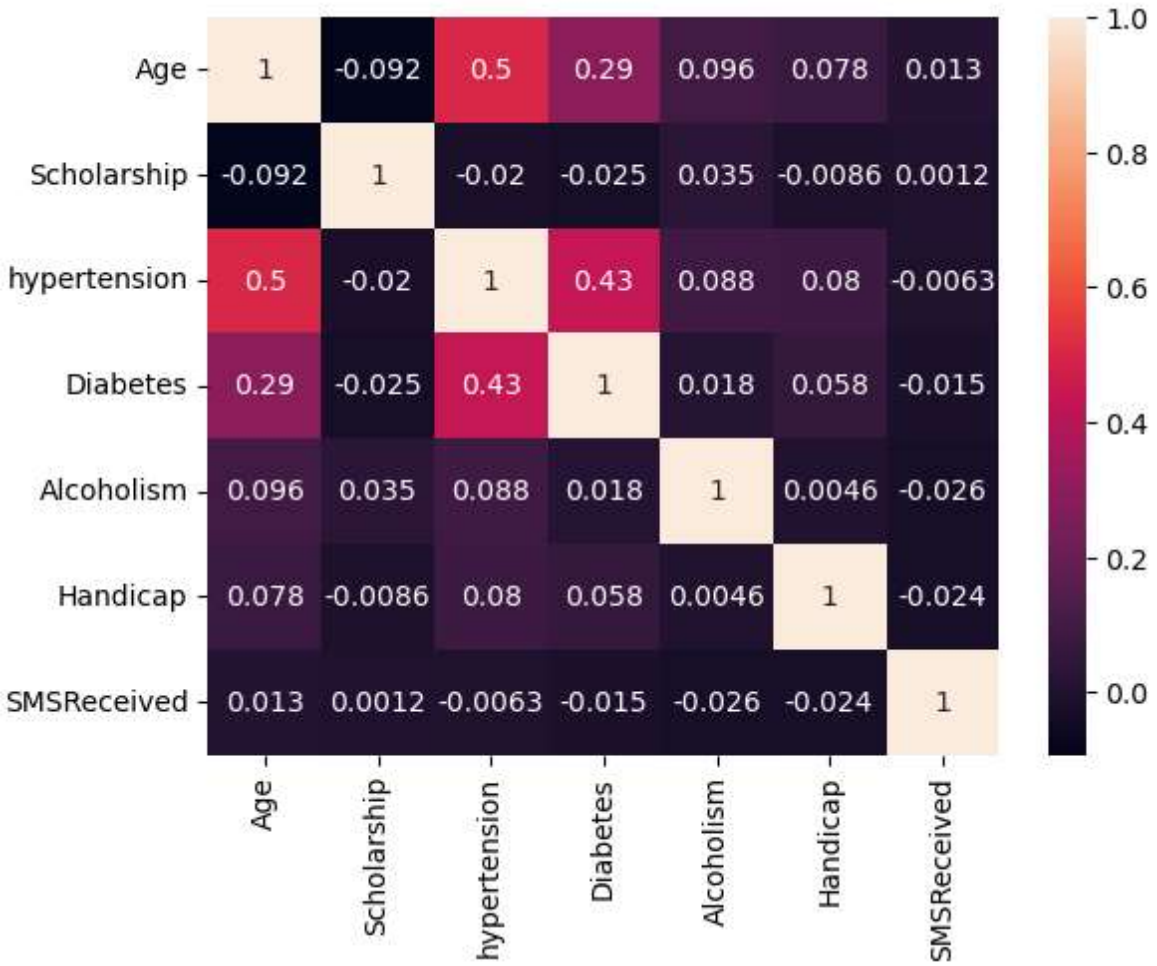
In [29]:
```
base_data['Scholarship'].value_counts()
```

Out[29]:
```
0    99665
1    10861
Name: Scholarship, dtype: int64
```

In [30]:
```
base_data['Neighbourhood'].value_counts()
```

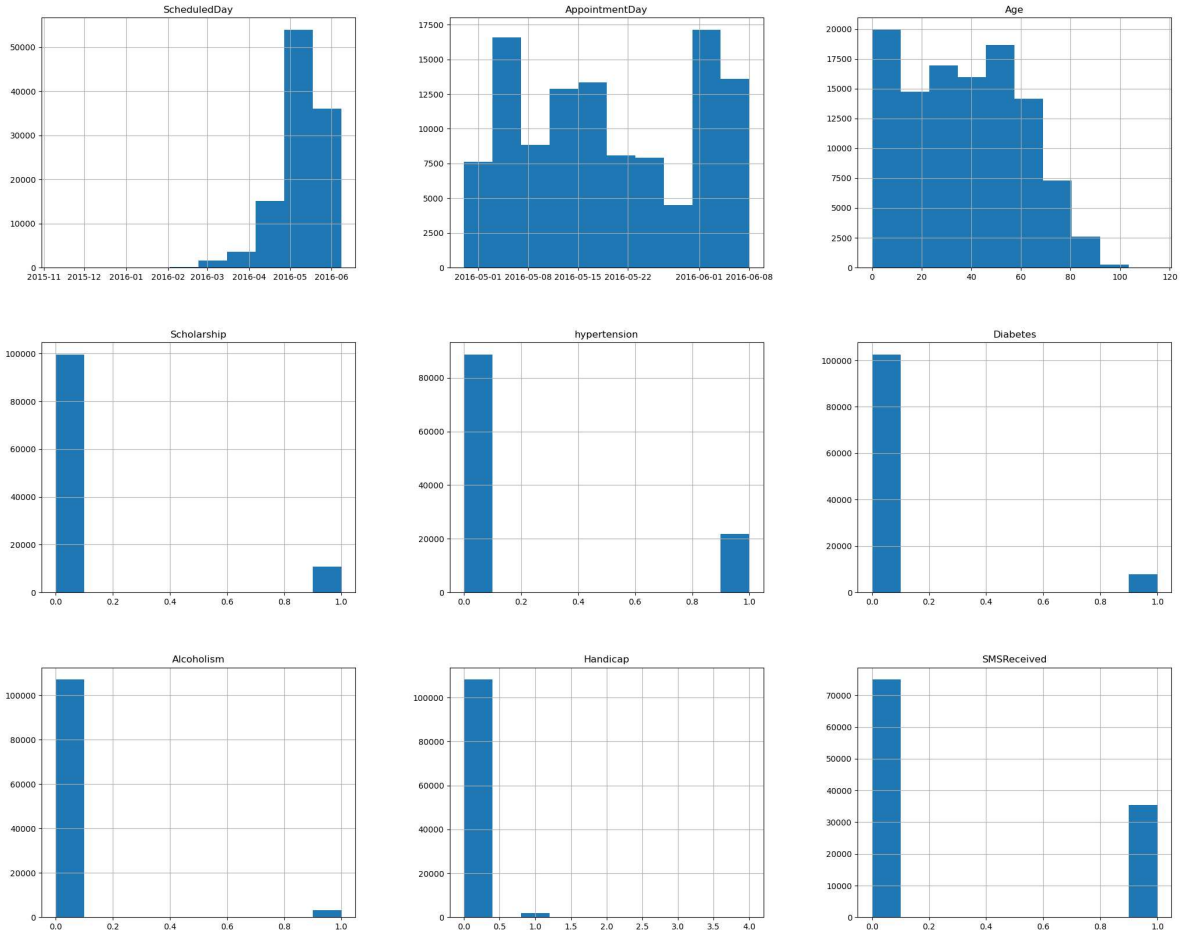
Out[30]:
```
JARDIM CAMBURI                 7717
MARIA ORTIZ                    5805
RESISTÊNCIA                    4431
JARDIM DA PENHA                3877
ITARARÉ                        3514
                               ...
ILHA DO BOI                      35
ILHA DO FRADE                    10
AEROPORTO                         8
ILHAS OCEÂNICAS DE TRINDADE       2
PARQUE INDUSTRIAL                 1
Name: Neighbourhood, Length: 81, dtype: int64
```

In [31]:
```
sns.heatmap(base_data.corr(),annot=True)
plt.show()
```



let's make it more clear through visualization

In [31]:
```
base_data.hist(figsize=(25,20));
```



In [32]:
```
# Rename incorrect columns names
base_data = base_data.rename(columns={'Handcap':'Handicap', 'Hipertension':'Hy
```

In [33]:
```
base_data.columns
```

Out[33]:
```
Index(['Gender', 'ScheduledDay', 'AppointmentDay', 'Age', 'Neighbourhood',
       'Scholarship', 'hypertension', 'Diabetes', 'Alcoholism', 'Handicap',
       'SMSReceived', 'Noshow'],
      dtype='object')
```

In [34]:
```
base_data['Noshow'].value_counts()
```

Out[34]:
```
No     88207
Yes    22319
Name: Noshow, dtype: int64
```

In [35]:
```
# rename the No-show column to avoid misleading

base_data = base_data.rename(columns={'Noshow':'Absent'})
```

In [36]:
```
base_data.columns
```

Out[36]:
```
Index(['Gender', 'ScheduledDay', 'AppointmentDay', 'Age', 'Neighbourhood',
       'Scholarship', 'hypertension', 'Diabetes', 'Alcoholism', 'Handicap',
       'SMSReceived', 'Absent'],
      dtype='object')
```

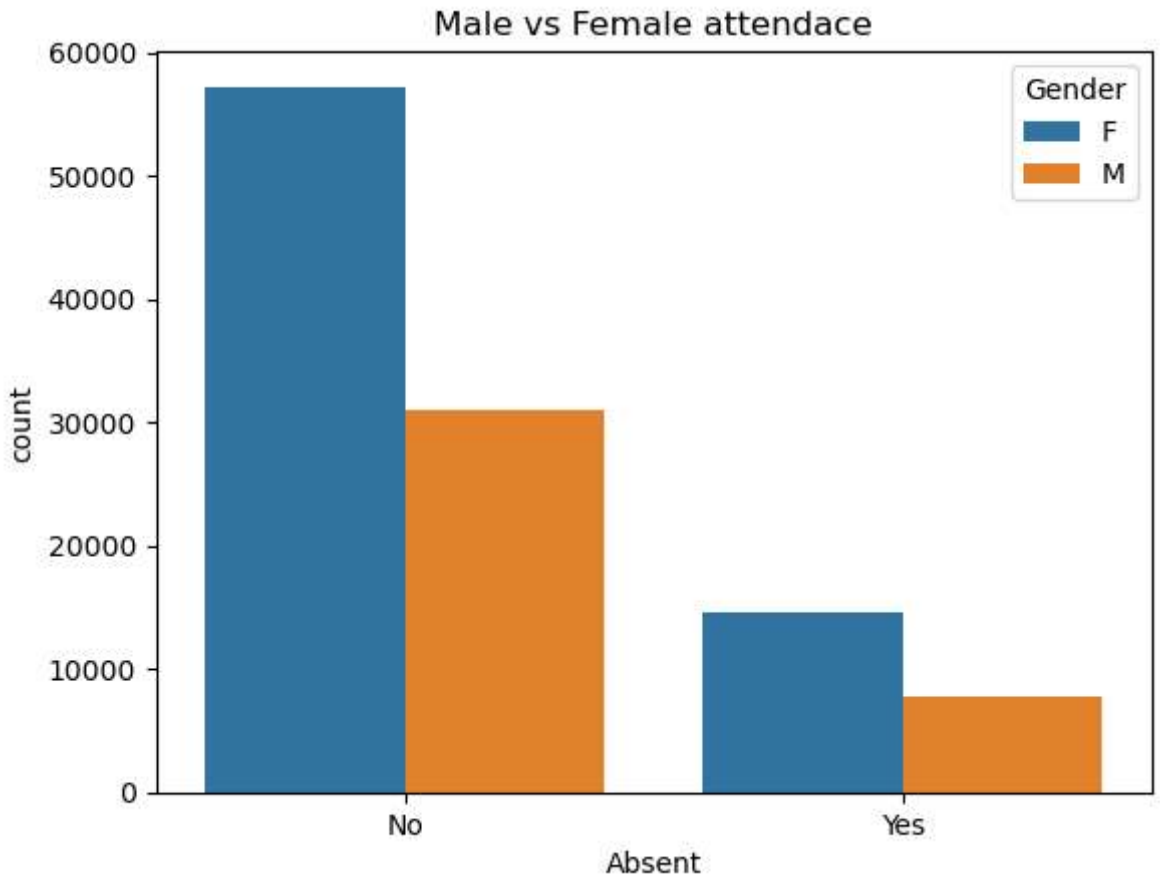
Exploratory Data Analysis

1. Does the patient gender has a realation with the atendance?

In [37]:
```
base_data['Gender'].value_counts()
```

Out[37]:
```
F    71839
M    38687
Name: Gender, dtype: int64
```

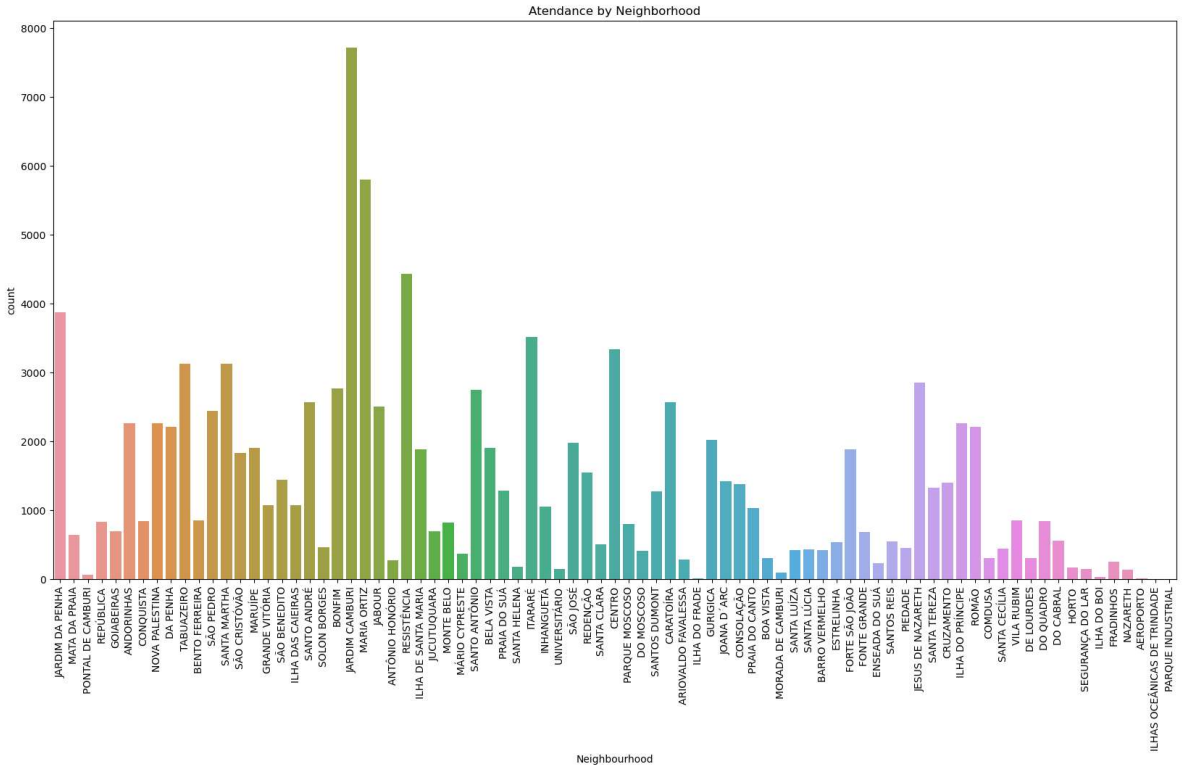
In [38]:
```
sns.countplot(x=base_data['Absent'], hue=base_data['Gender']);
plt.title('Male vs Female attendace');
```



The number of females show up is greater than the males. May be because we have more data of females but that also show that they vist hospitals more in general.

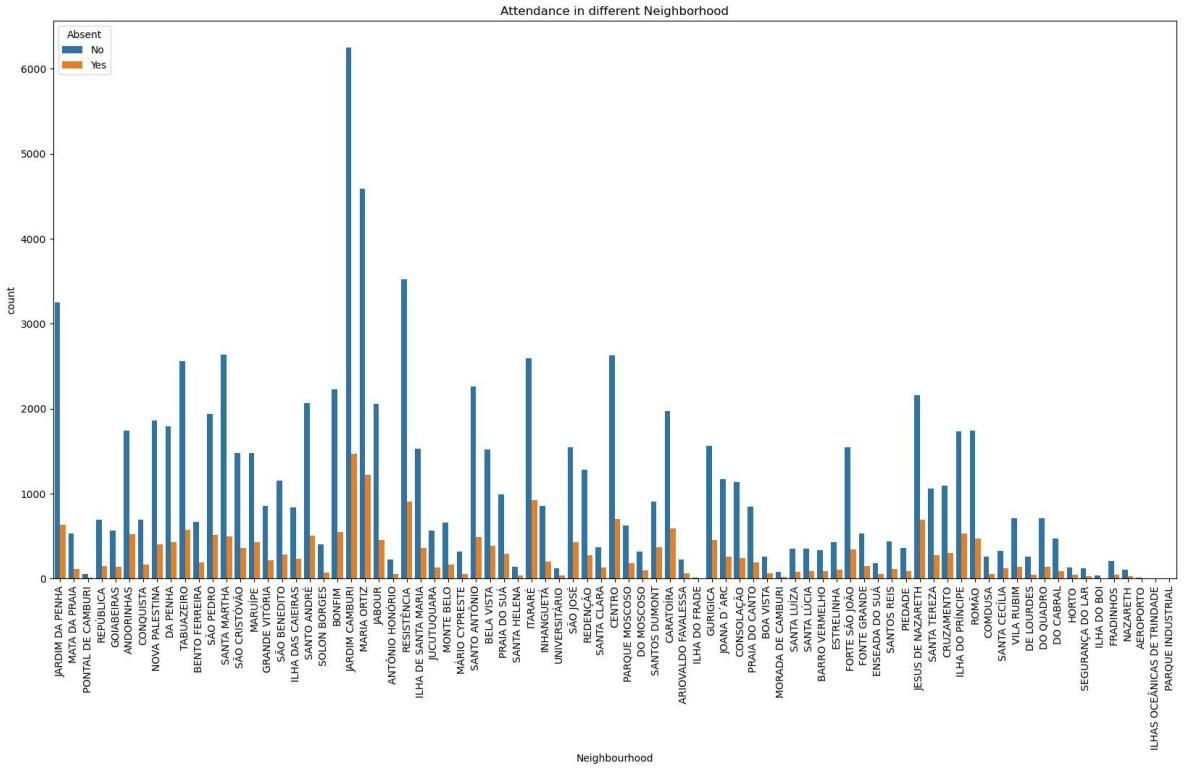
2. Does the neighborhood play a role in making patients don't show up? "Location of the hospital"

In [39]:
```
plt.figure(figsize=(20,10))
sns.countplot(x=base_data.Neighbourhood);
plt.title('Atendance by Neighborhood')
plt.xticks(rotation=90);
```



We see that some neighborhood have more people show up for their appointment and this indicates that this area have increase in diseases

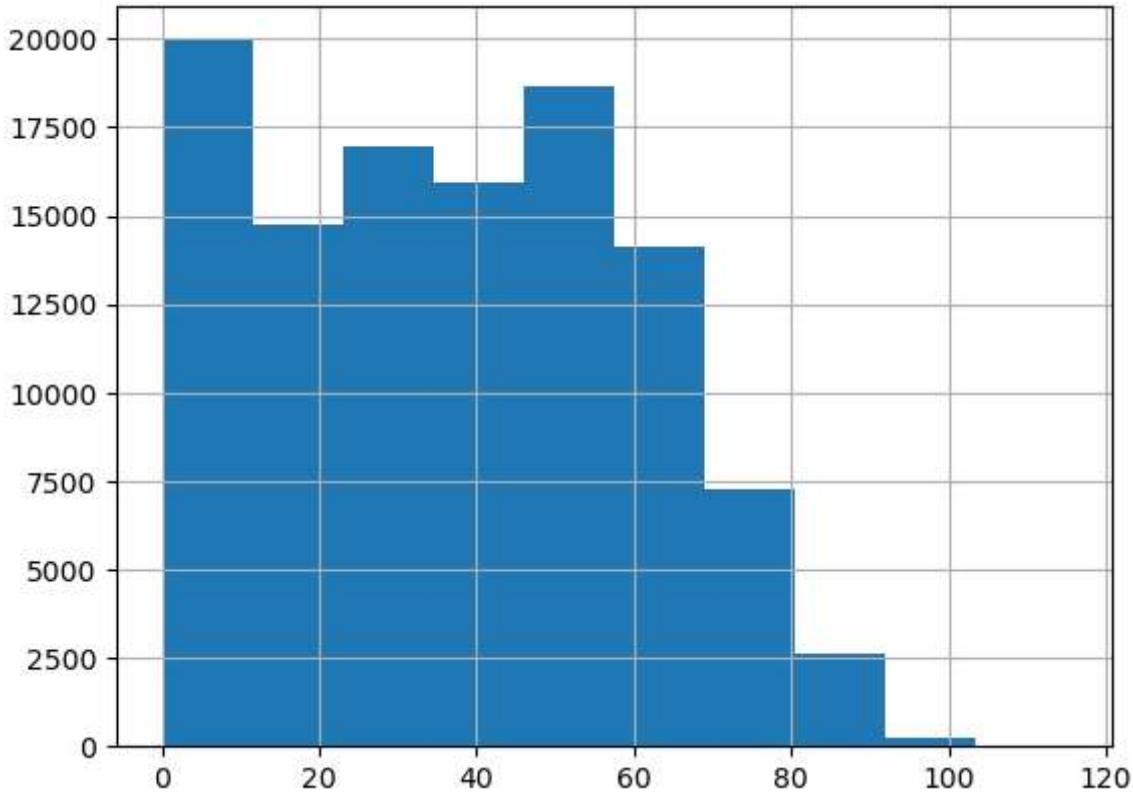
In [40]:
```
plt.figure(figsize=(20, 10))
sns.countplot(x=base_data['Neighbourhood'], hue=base_data['Absent']);
plt.xticks(rotation=90);
plt.title('Attendance in different Neighborhood');
```



In most neighborhoods patients attend more in the more the area where there are more disease

3. Which pateints show up more? Does old age take care of their health more than youth?

In [41]:
```
base_data['Age'].hist(bins=10);
```



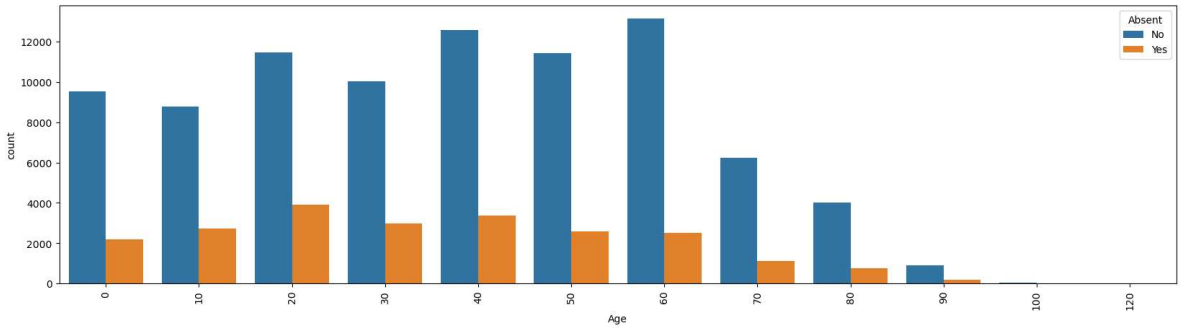
We see that most of the patients in the data are youth

In [42]:
```
base_data['Age'] = [round(a,-1) for a in base_data['Age']] #this trick makes c

base_data['Age'].value_counts()                          ##it easier visualizing
```

Out[42]:
```
40     15960
60     15628
20     15342
50     14012
30     13026
0      11731
10     11526
70      7365
80      4776
90      1090
100       65
120        5
Name: Age, dtype: int64
```

In [43]:
```
plt.figure(figsize=(20,5))
sns.countplot(x=base_data['Age'], hue=base_data['Absent'])
plt.xticks(rotation=90);
```



This shows that the ratio are close but youth still show up more which the oppisite of what we argued at the beginning

4. Does the disease type affect the patient's show up?

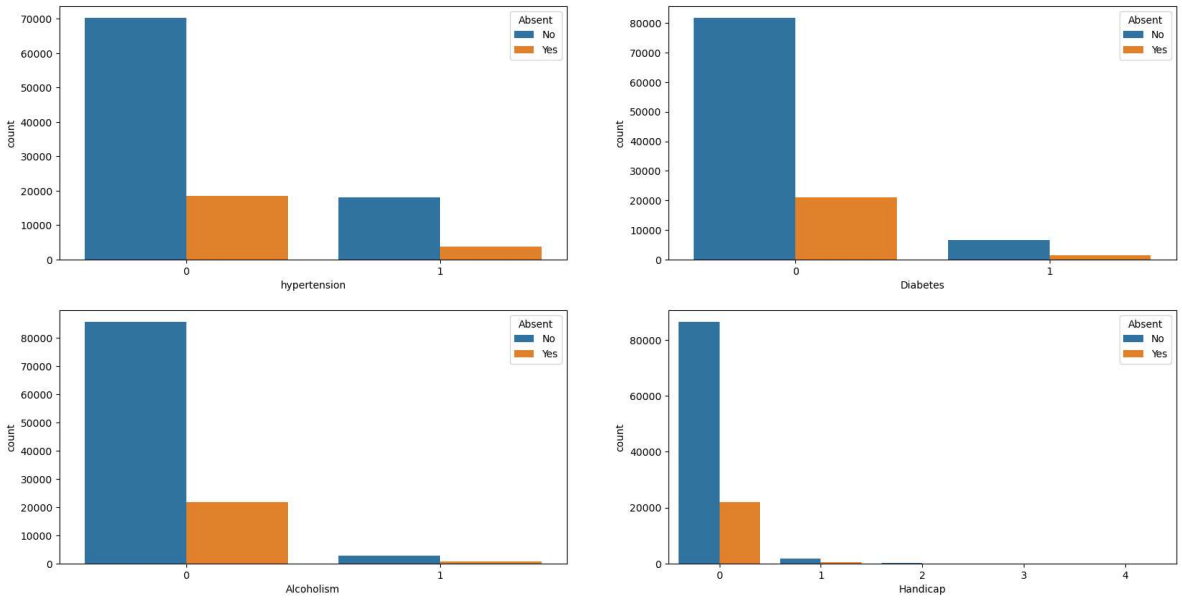
In [44]:
```
base_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 110526 entries, 0 to 110526
Data columns (total 12 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   Gender          110526 non-null  object
 1   ScheduledDay    110526 non-null  datetime64[ns]
 2   AppointmentDay  110526 non-null  datetime64[ns]
 3   Age             110526 non-null  int64
 4   Neighbourhood   110526 non-null  object
 5   Scholarship     110526 non-null  int64
 6   hypertension    110526 non-null  int64
 7   Diabetes        110526 non-null  int64
 8   Alcoholism      110526 non-null  int64
 9   Handicap        110526 non-null  int64
 10  SMSReceived     110526 non-null  int64
 11  Absent          110526 non-null  object
dtypes: datetime64[ns](2), int64(7), object(3)
memory usage: 11.0+ MB
```

In [50]:
```
disease_columns = base_data[['hypertension','Diabetes','Alcoholism','Handicap'
```

In [52]:
```
plt.figure(figsize=(20,10));
plt.subplot(2,2,1)
sns.countplot(disease_columns['hypertension'],hue=base_data['Absent'])
plt.subplot(2,2,2)
sns.countplot(disease_columns['Diabetes'],hue=base_data['Absent'])
plt.subplot(2,2,3)
sns.countplot(disease_columns['Alcoholism'],hue=base_data['Absent'])
plt.subplot(2,2,4)
sns.countplot(disease_columns['Handicap'],hue=base_data['Absent'])
```

Out[52]: <AxesSubplot:xlabel='Handicap', ylabel='count'>



We see that most of them don't have a disease and show up for appointment but we notice that patients of hypertension show up either when they are infected or not which is a mark that hypertension will probably show up more.

Conclusions

Now we can see the factors that affect the absence of the patients more clearly. The gender and age are the most important factor as we saw earlier that female and youth show up for their appointment more than male and old people. Neighbohood and hypertension come after gender and age as there are some neighborhoods that the diseases are spread and patients with hypertension tend to show up if they have it or not. So we need to search for more factors to help patient remmenber their appointments and show up.