# Project 2. No Show Appointments

Medical appointments are the services that helps patients to book their appointments with the doctors for their health check. However, not all patients show up for their appointment as their bookings.

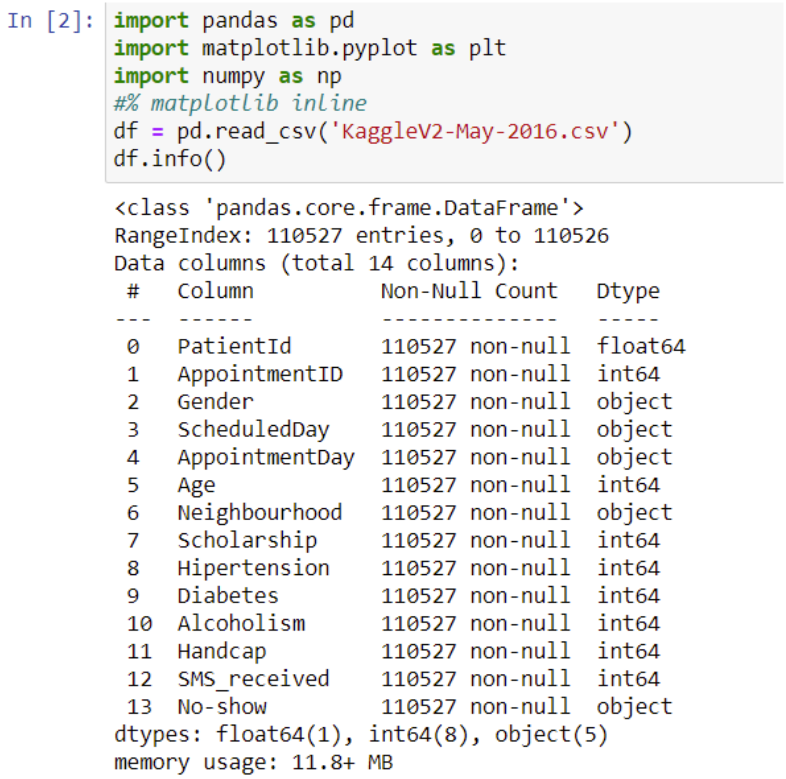
The project is about the investigation of the factors that affecting the medical appointmenst in Brasil.

There are many factors that involve in the appointments postpones. Here, the project is focusing on questions as below:

1. What is the ages in patients involving in the appointments?
2. Is there any difference in the gender affecting the appointment show up?
3. What is the difference in percentage between the patients joining the wellfare program and the patients that not joining the wellfare program for their appointment show up?

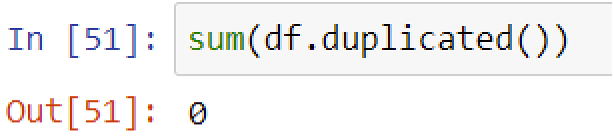
## 1.Data wrangling

Data is store in csv file with the name ‘KaggleV2-May-2016.csv’.First, we need to load the file into the Jupyter Notebook for the analysis.



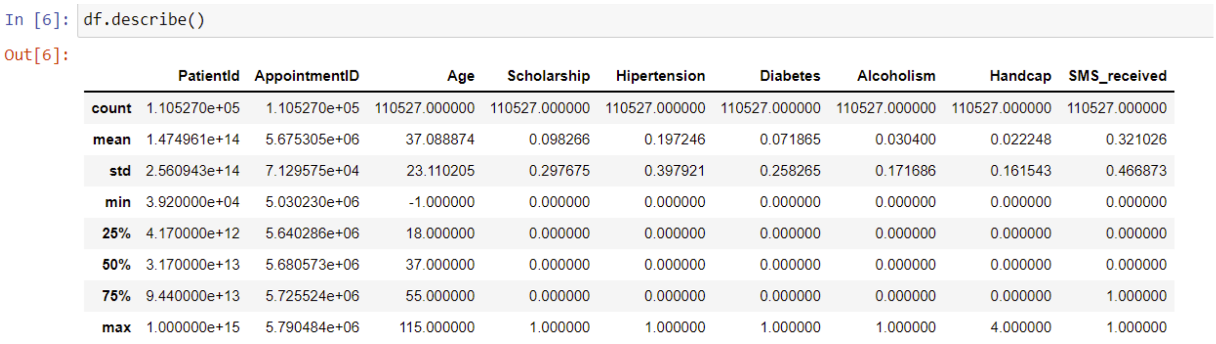
```
In [2]: import pandas as pd
        import matplotlib.pyplot as plt
        import numpy as np
        #% matplotlib inline
        df = pd.read_csv('KaggleV2-May-2016.csv')
        df.info()
```

```
<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
```



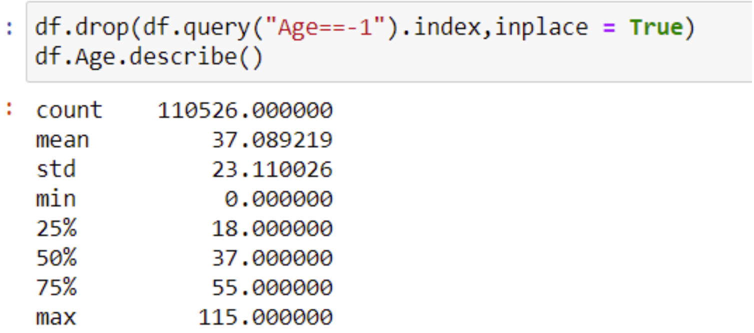
```
In [51]: sum(df.duplicated())

Out[51]: 0
```

There is 110,527 rows and 14 columns in the dataset. There is no null values and duplicates in the dataset. However, there is 1 age group that has the value of -1 that we need to remove it from the dataset as following:

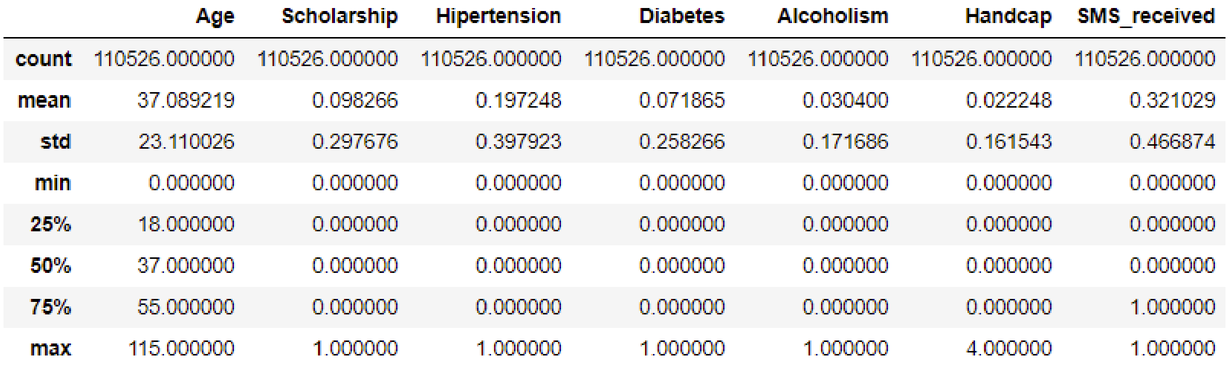


```
In [6]: df.describe()

Out[6]:
```

| | PatientId | AppointmentID | Age | Scholarship | Hipertension | Diabetes | Alcoholism | Handcap | SMS_received |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 1.105270e+05 | 1.105270e+05 | 110527.000000 | 110527.000000 | 110527.000000 | 110527.000000 | 110527.000000 | 110527.000000 | 110527.000000 |
| **mean** | 1.474961e+14 | 5.675305e+06 | 37.088874 | 0.098266 | 0.197246 | 0.071865 | 0.030400 | 0.022248 | 0.321026 |
| **std** | 2.560943e+14 | 7.129575e+04 | 23.110205 | 0.297675 | 0.397921 | 0.258265 | 0.171686 | 0.161543 | 0.466873 |
| **min** | 3.920000e+04 | 5.030230e+06 | -1.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| **25%** | 4.170000e+12 | 5.640286e+06 | 18.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| **50%** | 3.170000e+13 | 5.680573e+06 | 37.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| **75%** | 9.440000e+13 | 5.725524e+06 | 55.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 1.000000 |
| **max** | 1.000000e+15 | 5.790484e+06 | 115.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 4.000000 | 1.000000 |



```
: df.drop(df.query("Age==-1").index,inplace = True)
  df.Age.describe()

: count    110526.000000
  mean         37.089219
  std          23.110026
  min           0.000000
  25%          18.000000
  50%          37.000000
  75%          55.000000
  max         115.000000
```

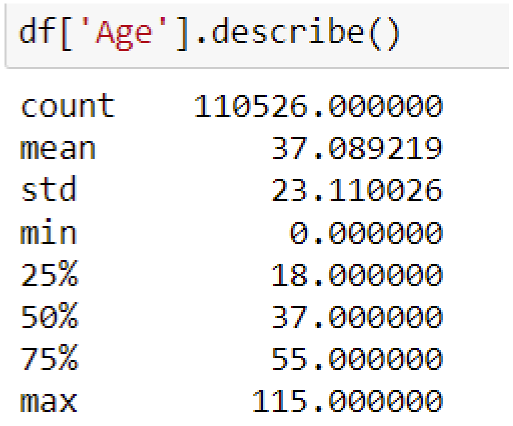


| | Age | Scholarship | Hipertension | Diabetes | Alcoholism | Handcap | SMS_received |
|---|---|---|---|---|---|---|---|
| **count** | 110526.000000 | 110526.000000 | 110526.000000 | 110526.000000 | 110526.000000 | 110526.000000 | 110526.000000 |
| **mean** | 37.089219 | 0.098266 | 0.197248 | 0.071865 | 0.030400 | 0.022248 | 0.321029 |
| **std** | 23.110026 | 0.297676 | 0.397923 | 0.258266 | 0.171686 | 0.161543 | 0.466874 |
| **min** | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| **25%** | 18.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| **50%** | 37.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| **75%** | 55.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 1.000000 |
| **max** | 115.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 4.000000 | 1.000000 |

The dataset is now cleaned with 110,526 rows and 14 columns.

## 2. Analysis

### 2.1 Identify the group age.



```
df['Age'].describe()
```

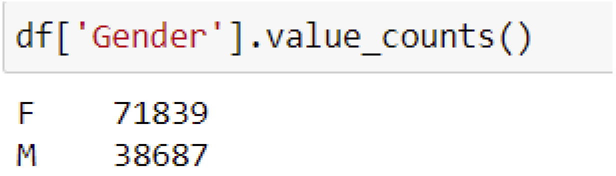
```
count    110526.000000
mean         37.089219
std          23.110026
min           0.000000
25%          18.000000
50%          37.000000
75%          55.000000
max         115.000000
```

- The average age of the sample was 37 years old, in which the minimum age is 0 (less than 1 years old) and the highest was 115 years old.

- 50% of the samples aged from 0 to 37 years old.

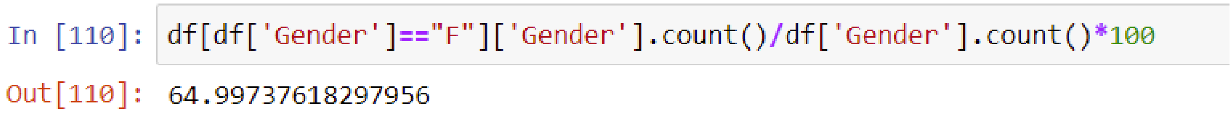
- 75% of the samples age from 0 to 55 years old.

### 2.2 Identify the gender

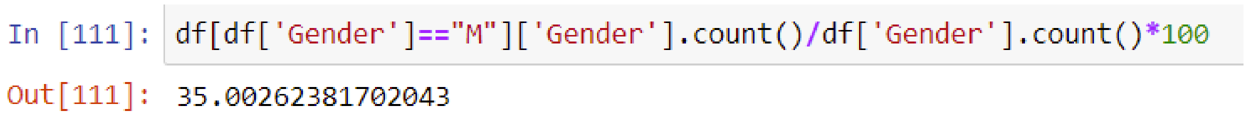


```
df['Gender'].value_counts()
```

```
F    71839
M    38687
```



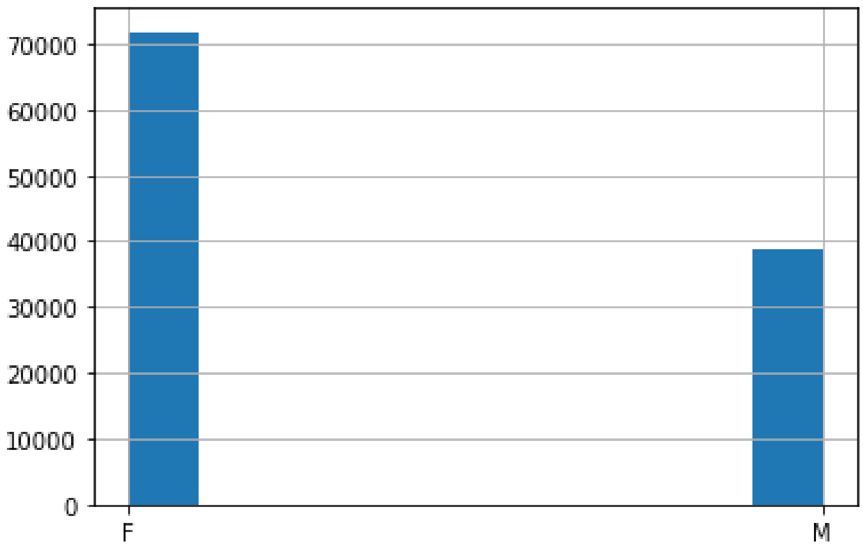
```
In [110]: df[df['Gender']=="F"]['Gender'].count()/df['Gender'].count()*100

Out[110]: 64.99737618297956
```



```
In [111]: df[df['Gender']=="M"]['Gender'].count()/df['Gender'].count()*100

Out[111]: 35.00262381702043
```

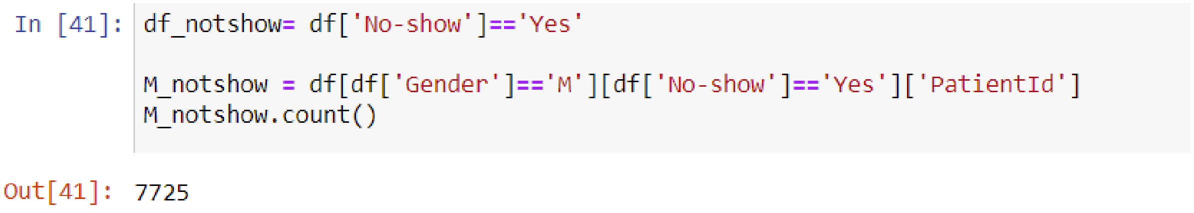
The female number was 71,839 which represent 65% the sample.

The male number was 38,687 which represent 35% the sample.

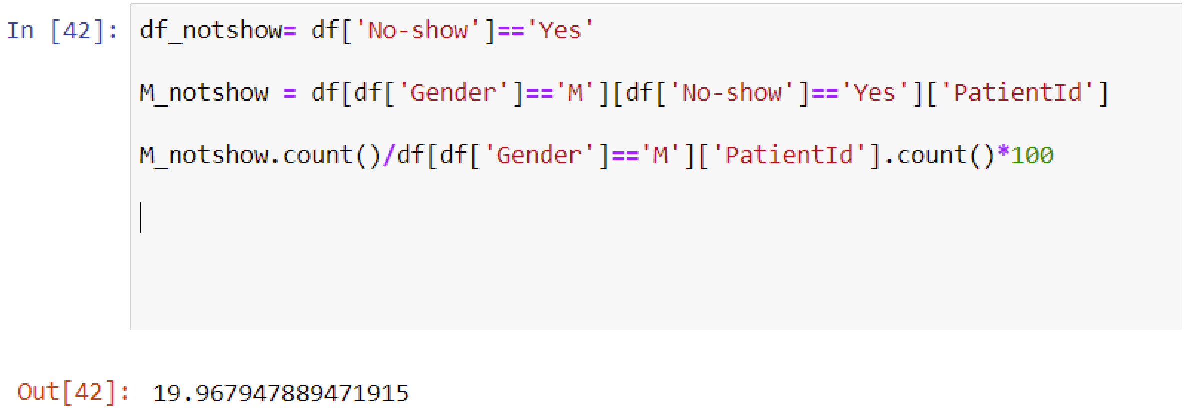


### 2.3 Finding the relationship between the gender and the No-show results

### 2.3.1 Male number that not showing at the appointment



```
In [41]: df_notshow= df['No-show']=='Yes'

         M_notshow = df[df['Gender']=='M'][df['No-show']=='Yes']['PatientId']
         M_notshow.count()

Out[41]: 7725
```

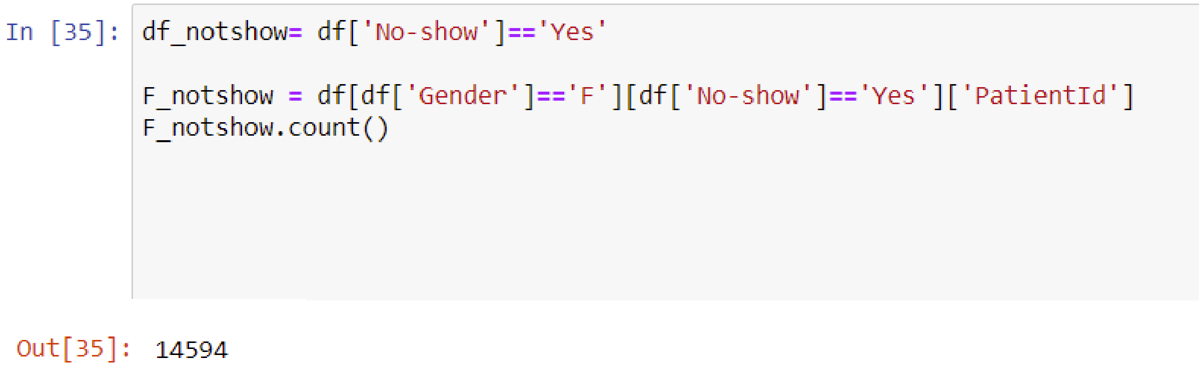


```
In [42]: df_notshow= df['No-show']=='Yes'

         M_notshow = df[df['Gender']=='M'][df['No-show']=='Yes']['PatientId']

         M_notshow.count()/df[df['Gender']=='M']['PatientId'].count()*100

Out[42]: 19.967947889471915
```

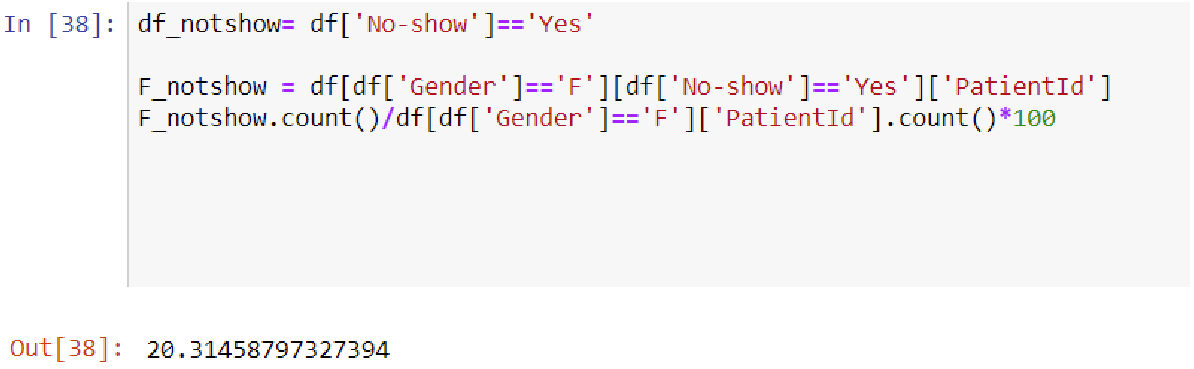
There was 7,725 male out of 38,687  that not appearing in the appointments. The percentage was 19.97%

### 2.3.2 Female number that not showing at the appointment

Number of female that didn't show up:



```
In [35]: df_notshow= df['No-show']=='Yes'

         F_notshow = df[df['Gender']=='F'][df['No-show']=='Yes']['PatientId']
         F_notshow.count()

Out[35]: 14594
```

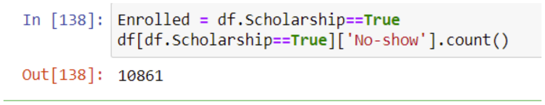


```
In [38]: df_notshow= df['No-show']=='Yes'

         F_notshow = df[df['Gender']=='F'][df['No-show']=='Yes']['PatientId']
         F_notshow.count()/df[df['Gender']=='F']['PatientId'].count()*100

Out[38]: 20.31458797327394
```

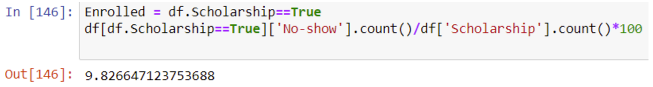
There was 14,594 female out of 71,839 that not appearing in the appointments. The percentage was 20.3%

**Conclusion:** a little bit higher of male rate that not showing at the appointments compared with the female. Therefore, Gender is not the factor that affecting the results of the appoinments.

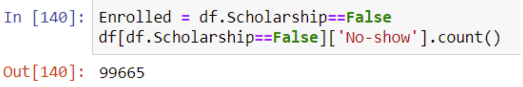
### 2.3 Identify people involving the Brasilian wellfare



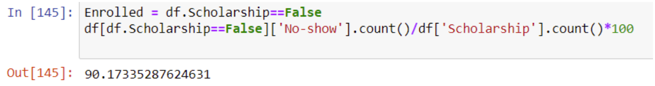
```
In [138]: Enrolled = df.Scholarship==True
          df[df.Scholarship==True]['No-show'].count()

Out[138]: 10861
```



```
In [146]: Enrolled = df.Scholarship==True
          df[df.Scholarship==True]['No-show'].count()/df['Scholarship'].count()*100

Out[146]: 9.826647123753688
```

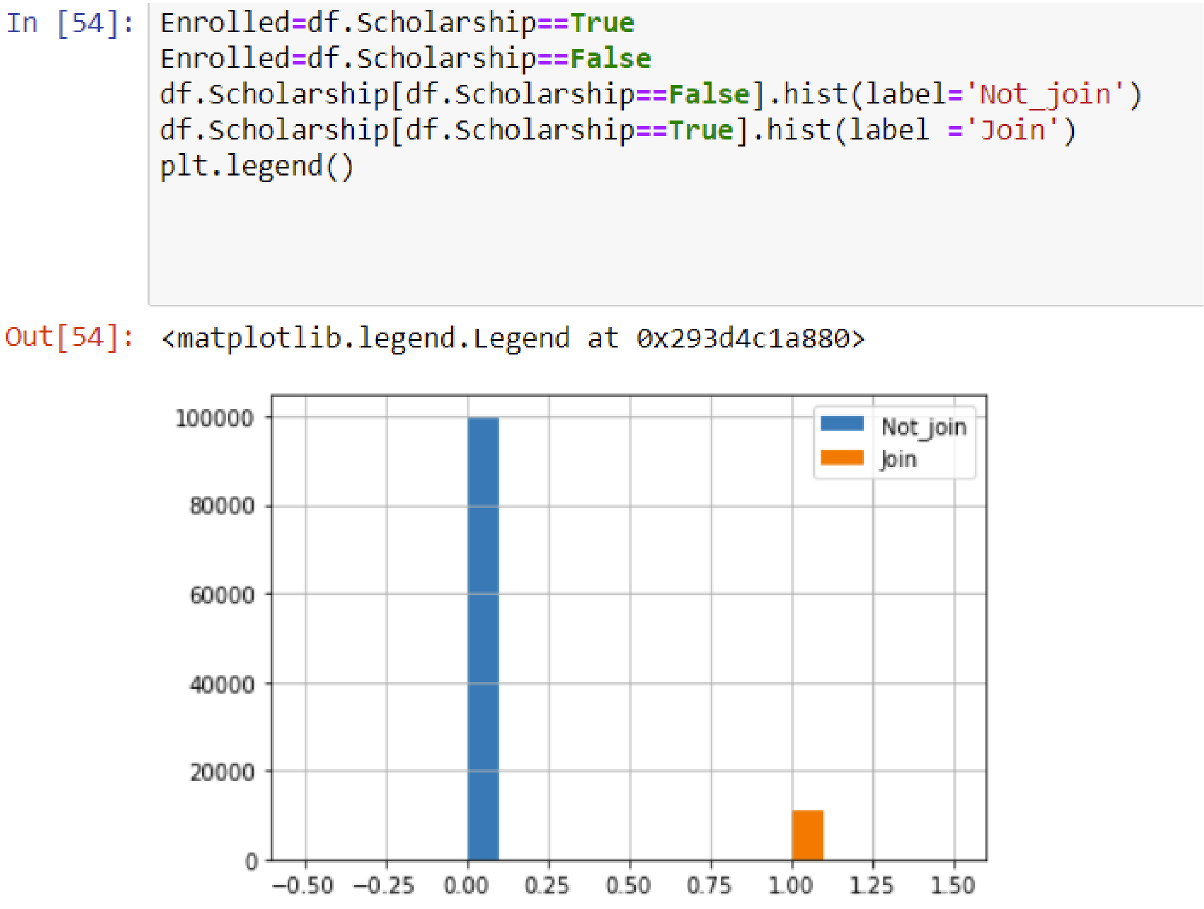


```
In [140]: Enrolled = df.Scholarship==False
          df[df.Scholarship==False]['No-show'].count()

Out[140]: 99665
```



```
In [145]: Enrolled = df.Scholarship==False
          df[df.Scholarship==False]['No-show'].count()/df['Scholarship'].count()*100

Out[145]: 90.17335287624631
```

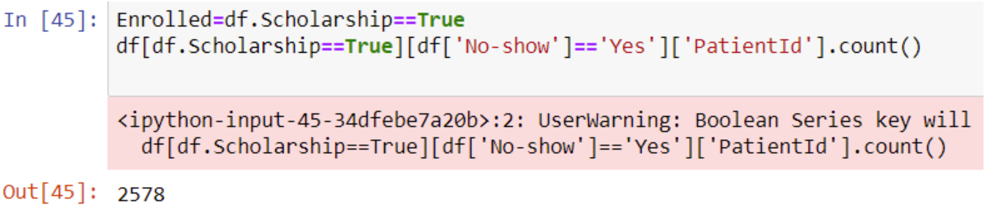
There are 10,861 (9.8%) people joining in the Brasilian and 99,665 (90%) not joining the program.



```
In [54]: Enrolled=df.Scholarship==True
         Enrolled=df.Scholarship==False
         df.Scholarship[df.Scholarship==False].hist(label='Not_join')
         df.Scholarship[df.Scholarship==True].hist(label ='Join')
         plt.legend()

Out[54]: <matplotlib.legend.Legend at 0x293d4c1a880>
```

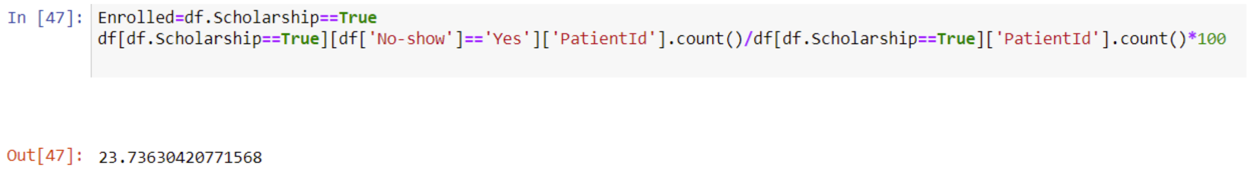
## 2.4 Relationship between the wellfare program with the No-show results

### 2.4.1 Number of people joining the wellfare program that didn't show up at the appointments



```
In [45]: Enrolled=df.Scholarship==True
         df[df.Scholarship==True][df['No-show']=='Yes']['PatientId'].count()

         <ipython-input-45-34dfebe7a20b>:2: UserWarning: Boolean Series key will
           df[df.Scholarship==True][df['No-show']=='Yes']['PatientId'].count()

Out[45]: 2578
```

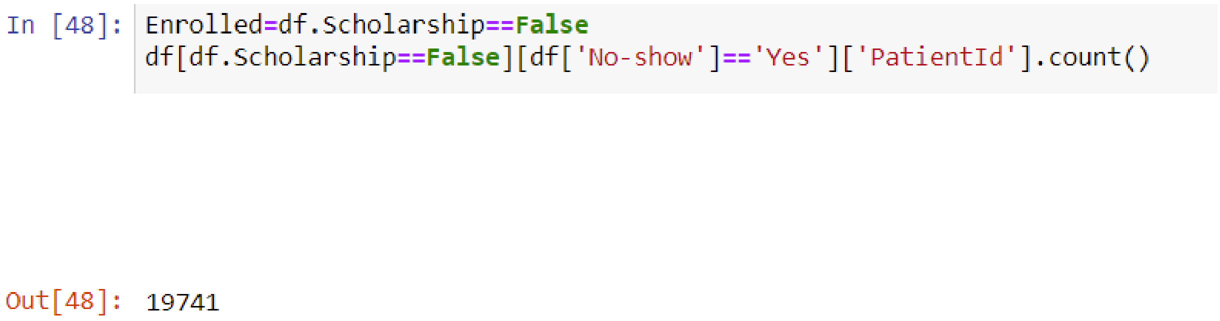
### 2.4.2 **The percentage of people joining the wellfare program that didn’t show up at the appointments**



```
In [47]: Enrolled=df.Scholarship==True
         df[df.Scholarship==True][df['No-show']=='Yes']['PatientId'].count()/df[df.Scholarship==True]['PatientId'].count()*100

Out[47]: 23.73630420771568
```

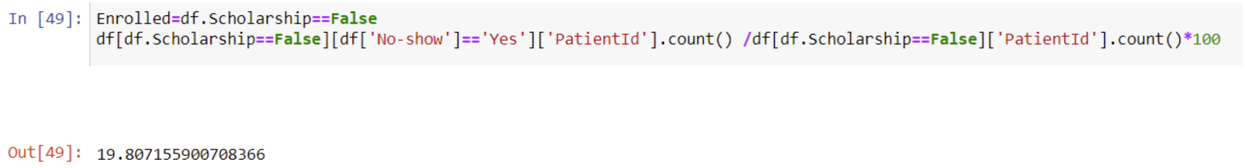
There are 2,578 people **joining the program** didn’t show up at the appointmes. The percentage was 23.74%

### 2.4.3 Number of people not joining the wellfare program that didn’t show up at the appointments



```
In [48]: Enrolled=df.Scholarship==False
         df[df.Scholarship==False][df['No-show']=='Yes']['PatientId'].count()

Out[48]: 19741
```

### 2.4.4 The percentage of people not joining the wellfare program that didn’t show up at the appointments



```
In [49]: Enrolled=df.Scholarship==False
         df[df.Scholarship==False][df['No-show']=='Yes']['PatientId'].count() /df[df.Scholarship==False]['PatientId'].count()*100

Out[49]: 19.807155900708366
```

There are 19,741 out of 99,665 (results in 2.3)  not joining the program didn’t show up at the appointmes. The percentage was 19.8 %

## Conclusion

- The average age of the sample was 37 years old, in which the minimum age is 0 (less than 1 years old) and the highest was 115 years old, 50% of the samples aged from 0 to 37 years old and 75% from 0 to 55 years old.

- There is no difference in between female and male patients in showing up for their appointments.

- The percentage of patients having the wellfare program that didn’t show up for their appointments(23.74%) is higher compared with the patients who don’t have the wellfare program (19.8%)