Investigate_a_Dataset

January 20, 2023

1 Project: Investigate a Dataset - [No-show Appointments]

1.1 Table of Contents

1.1.1 Dataset Description

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.

In this notebook we will be analysing this dataset for people how are above 16 years old only.

1.1.2 Data Dictionary

The order of columns listed here is different from the original dataset order. * * refers to newly added columns during the wrangling process. * ** refers to columns that were dropped during the wrangling process.

- **Gender** : Male or Female
- **Age** : How old is the patient
- **Neighbourhood** : Where the appointment takes place, "Where the hospital is within the city of Vitória"
- **AppointmentDay** : The day of the actuall appointment, when they have to visit the doctor.
- * **WaitingDays**
  - Number of days a person have waited before the Appointment
  - `AppointmentDay - ScheduledDay`
- **Scholarship** : Indicates whether or not the patient is enrolled in Brasilian welfare program Bolsa Família
- **Hipertension** : True or False
- **Diabetes** : True or False

- **Alcoholism** : True or False
- **Handcap** : the number of desabilites a person has.
- **SMS_received**
  - True or False
  - indecates that 1 or more messages sent to the patient.
- **No_how**
  - 'No' if the patient showed up to their appointment
  - 'Yes' if they did not show up.
- ** **PatientId** : Identification of a patient
- ** **AppointmentID** : Identification of each appointment
- ** **ScheduledDay** : The date in which someone registered the appointment

1.1.3 Question of Analysis

1. What is the absence ratio generally, and for each gender?
2. What is the effect of each disease on the Absence Ratio?
3. What is Absence Ratio for people who received SMS ant those who didn't?
4. What is Absence Ratio for people with specefic characteristics?
   1. Patients with Hypertension, Diabetes and Handicap.
   2. Patients received SMS and having scholarship.
   3. Patients waited more than one day.
   4. Patients who have their appointments at the same day.
5. How does the absence rate change over the course of the week?
6. What is the correlation between Absence Ratio and the following factors?
   - Age
   - Waiting Days
7. For neighbourhoods hosted most of the appointments, what is the relation between Absence Ratio and number of hosted appointments?

Data Wrangling

Lets have a broad look to our dataset to see if it needs any trimming or cleaning.

Loading and Assessing > By loading our dataset and looking into it, > we will discover some problems that need to be solved before starting our analysis, > then we will solve those problems in the cleaning process.

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt

        pd.options.display.max_rows = 99
        %matplotlib inline
```

Calling the necessary packages that includes some useful functions we will use in the next sections. * **NumPy:** adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. * **Pandas:** providing fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive. * **Matplotlib:** generating plots, histograms, bar charts, and other types of charts with just a few lines of code.

```
In [2]: df= pd.read_csv('Database_No_show_appointments/noshowappointments-kagglev2-may-2016.csv'
        df.rename(columns={'No-show':'No_show' },inplace=True)
        df.head()
```

Out[2]:

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension	Diabetes	Alcoholism	Handcap	SMS_received	No_show
0	2.987250e+13	5642903	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1	0	0	0	0	No
1	5.589978e+14	5642503	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0	0	0	0	0	No
2	4.262962e+12	5642549	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0	0	0	0	0	No
3	8.679512e+11	5642828	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0	0	0	0	0	No
4	8.841186e+12	5642494	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	1	1	0	0	0	No

Reading the content of our data scource using Pandas and assigning it to a Pandas DataFrame, and viewing its first 5 entries to check if it's loaded correctly.

Observation - some columns names need a little change - *ScheduledDay* provides informations about time in seconds, while *AppointmentDay* provides the date only without the time of the appointment.

```
In [3]: def more_about(df):
            """Makes a table of some useful informations for a given Dataframe"""

            columns = [df.dtypes, df.count(), df.isnull().sum(), df.nunique()]
            names   = ["data_type", "n_values",      "n_NaN",      "n_unique"]
            infos   = pd.concat(columns, axis=1, keys=names)

            return infos

        more_about(df)
```

Out[3]:

	data_type	n_values	n_NaN	n_unique
PatientId	float64	110527	0	62299
AppointmentID	int64	110527	0	110527
Gender	object	110527	0	2
ScheduledDay	object	110527	0	103549
AppointmentDay	object	110527	0	27
Age	int64	110527	0	104
Neighbourhood	object	110527	0	81
Scholarship	int64	110527	0	2
Hipertension	int64	110527	0	2
Diabetes	int64	110527	0	2
Alcoholism	int64	110527	0	2
Handcap	int64	110527	0	5
SMS_received	int64	110527	0	2
No_show	object	110527	0	2

Extracting general useful informations about data type, number of null values and number of unique values for each column.

Observation - The dataset contains zero null values. - *ScheduledDay* and *AppointmentDay* need to be Date type. - all of *AppointmentId* values are unique so this column isn't needed in our investigation.

```
In [4]: df.duplicated().sum()
```

```
Out[4]: 0
```

Observation The dataset contains no duplicates.

```
In [5]: df.describe()
```

Out[5]:

	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
```

Observation The min age is -1.

```
In [6]: print(np.sort(df.Age.unique()))
```

```
[ -1   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16
  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34
  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52
  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70
  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88
  89  90  91  92  93  94  95  96  97  98  99 100 102 115]
```

```
In [7]: df[df.Age == -1]
```

```
Out[7]:          PatientId  AppointmentID Gender          ScheduledDay  \
        99832  4.659432e+14        5775010      F  2016-06-06T08:58:13Z

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

               Diabetes  Alcoholism  Handcap  SMS_received No_show
        99832         0           0        0             0      No
```

Observation There is only one entry with age of -1

For the ages in my opinion, a baby or a child is not the one who decides whether to go for the appointment or not, in most cases for a person under 16 years old showing up for the appointment is related to other persons as well (mostly his parents), so we will analyse our data for adults above 16 years old only

Data Cleaning >Now let's take actions for the observations above.

Renaming Columns **Action** Renaming misspelled columns to explain data correctly and unify all the names under one formatting.

```
In [8]: df.rename(columns={'Hipertension':'Hypertension',
                          'Handcap': 'Handicap',
                          'No-show':'No_show',},inplace=True)
```

Fixing date problems

Actions - Changing *ScheduledDay* and *AppointmentDay* to be Date type and to co-tain informations about the date only without the time. - Making new column to hold informations about number of waiting days ***WaitingDays*** = (AppointmentDay - ScheduledDay) - Changing the informations in *AppointmentDay* to the name of the day instead of the date.

```
In [9]: df['ScheduledDay']   = pd.to_datetime(pd.to_datetime(df['ScheduledDay']).dt.date)
        df['AppointmentDay'] = pd.to_datetime(df['AppointmentDay'])

        df["WaitingDays"]    = (df['AppointmentDay'] - df['ScheduledDay']).dt.days

        df['AppointmentDay'] = df['AppointmentDay'].dt.day_name()
```

```
In [10]: df[df.WaitingDays < 0]
```

```
Out[10]:           PatientId  AppointmentID Gender ScheduledDay AppointmentDay  Age  \
         27033  7.839273e+12        5679978      M   2016-05-10         Monday   38
         55226  7.896294e+12        5715660      F   2016-05-18        Tuesday   19
         64175  2.425226e+13        5664962      F   2016-05-05      Wednesday   22
         71533  9.982316e+14        5686628      F   2016-05-11       Thursday   81
         72362  3.787482e+12        5655637      M   2016-05-04        Tuesday    7

                Neighbourhood  Scholarship  Hypertension  Diabetes  Alcoholism  \
         27033    RESISTÊNCIA            0             0         0           0
         55226  SANTO ANTÔNIO            0             0         0           0
         64175     CONSOLAÇÃO            0             0         0           0
         71533  SANTO ANTÔNIO            0             0         0           0
         72362     TABUAZEIRO            0             0         0           0

                Handicap  SMS_received No_show  WaitingDays
         27033         1             0     Yes           -1
         55226         1             0     Yes           -1
         64175         0             0     Yes           -1
         71533         0             0     Yes           -6
         72362         0             0     Yes           -1
```

Observation Some persons waited a negative number of days.

Dropping

Actions - drop unnecessary columns and order the other columns. - drop observed rows that contain typos. - trim rows according to age to be for people above 16 years old.

```
In [11]: df= df[['Gender','Age','Neighbourhood','AppointmentDay','WaitingDays',
                 'Scholarship','Hypertension','Diabetes','Alcoholism','Handicap',
                 'SMS_received', 'No_show']]
```

```
df= df.query('WaitingDays >= 0 and Age >= 16')

df.head()
```

Out[11]:

	Gender	Age	Neighbourhood	AppointmentDay	WaitingDays	Scholarship	\
0	F	62	JARDIM DA PENHA	Friday	0	0	
1	M	56	JARDIM DA PENHA	Friday	0	0	
2	F	62	MATA DA PRAIA	Friday	0	0	
4	F	56	JARDIM DA PENHA	Friday	0	0	
5	F	76	REPÚBLICA	Friday	2	0	

	Hypertension	Diabetes	Alcoholism	Handicap	SMS_received	No_show
0	1	0	0	0	0	No
1	0	0	0	0	0	No
2	0	0	0	0	0	No
4	1	1	0	0	0	No
5	1	0	0	0	0	No

```
In [12]: more_about(df)
```

Out[12]:

	data_type	n_values	n_NaN	n_unique
Gender	object	86054	0	2
Age	int64	86054	0	87
Neighbourhood	object	86054	0	81
AppointmentDay	object	86054	0	6
WaitingDays	int64	86054	0	127
Scholarship	int64	86054	0	2
Hypertension	int64	86054	0	2
Diabetes	int64	86054	0	2
Alcoholism	int64	86054	0	2
Handicap	int64	86054	0	5
SMS_received	int64	86054	0	2
No_show	object	86054	0	2

Exploratory Data Analysis

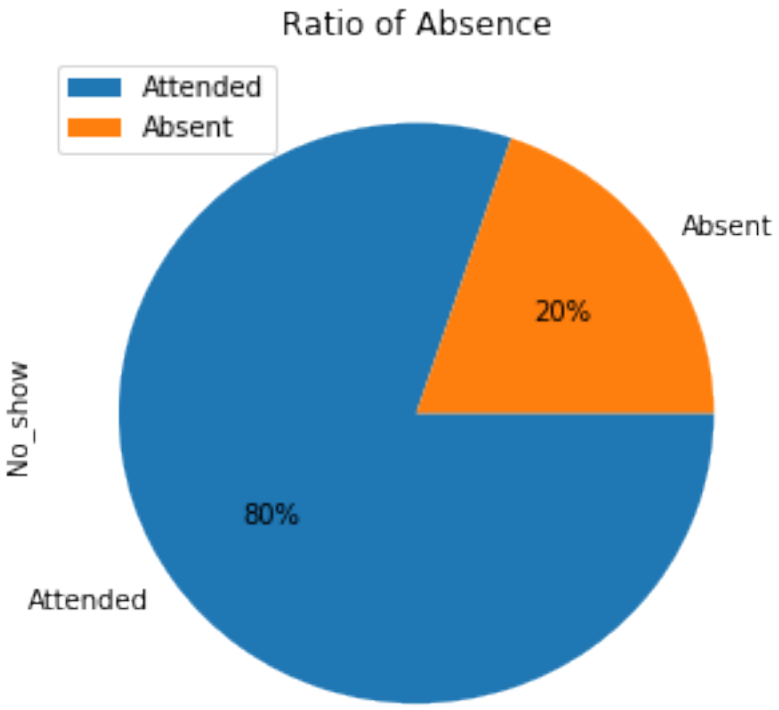
1.1.4 What factors are more correlated with the patient not showing up for their appointment?

Here we will take every single factor and see how much its affecting the attendence of the patients, there will be a repeatitive code blocks so its better to handle some of them with custom functions.

Research Question 1 (What is the absence ratio generally, and for each gender?)

```
In [13]: df.No_show.value_counts().plot(kind= 'pie',  labels = ['Attended', 'Absent'],
                                        figsize=(5, 5), counterclock = False, autopct='%1.0f%%')

         plt.title('Ratio of Absence')
```

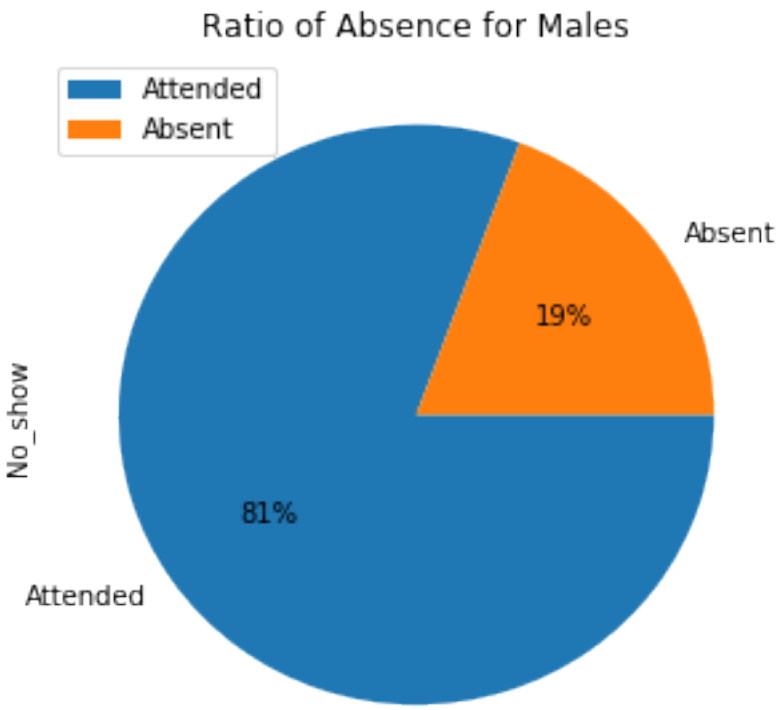
```
plt.legend()
plt.show()
df.No_show.value_counts()
```




```
Out[13]: No      68987
         Yes     17067
         Name: No_show, dtype: int64
```

Observation General Absence Ratio is 20%

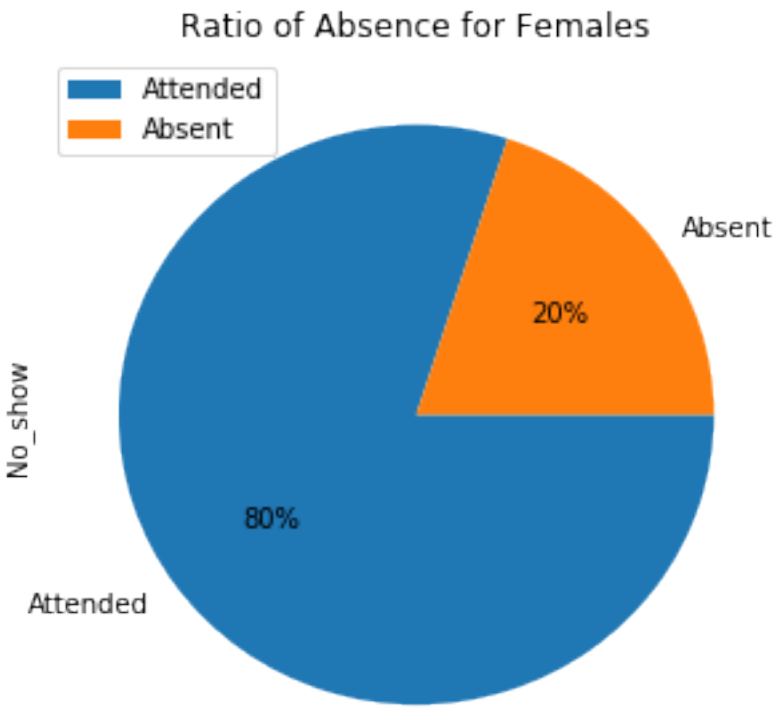
```
In [14]: values=df[df.Gender == "M"].No_show.value_counts()

         values.plot(kind= 'pie',  labels = ['Attended', 'Absent'],
                     figsize=(5, 5), counterclock = False, autopct='%1.0f%%')
         plt.title('Ratio of Absence for Males')
         plt.legend()
         plt.show()
         values
```




```
Out[14]: No      21112
         Yes      5068
         Name: No_show, dtype: int64
```

Observation Males Absence Ratio is 19%

```
In [15]: values=df[df.Gender == "F"].No_show.value_counts()

         values.plot(kind= 'pie',  labels = ['Attended', 'Absent'],
                     figsize=(5, 5), counterclock = False, autopct='%1.0f%%')
         plt.title('Ratio of Absence for Females')
         plt.legend()
         plt.show()

         values
```



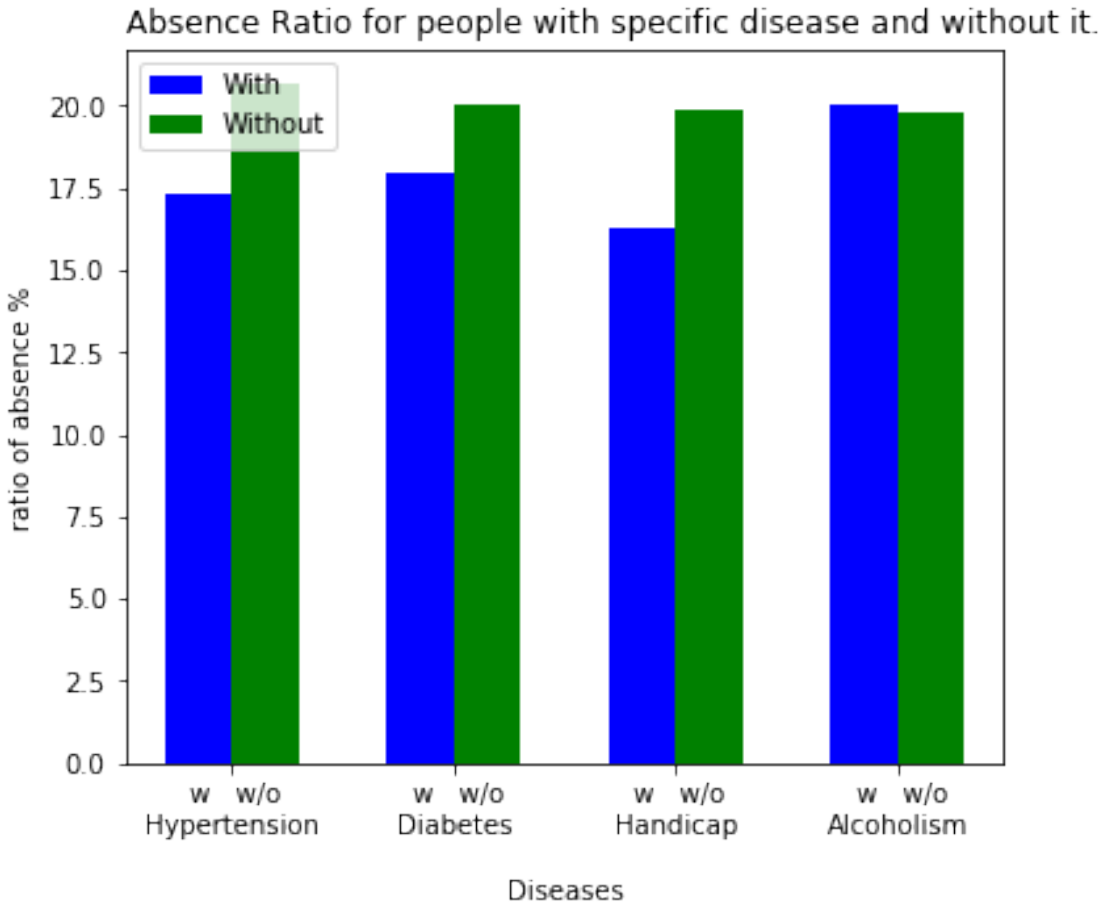

```
Out[15]: No      47875
         Yes     11999
         Name: No_show, dtype: int64
```

Observation Females Absence Ratio is 20%

Research Question 2 (What is the effect of each disease on the Absence Ratio?)

```
In [16]: ratio = pd.DataFrame()
         diseases = ['Hypertension','Diabetes','Handicap','Alcoholism']
         for disease in diseases:
             total  = df[disease].value_counts()
             absent = df.query('No_show == "Yes"')[disease].value_counts()
             ratio = ratio.append({'disease': disease,
                                   'with': (absent*100/total).loc[1],
                                   'without': (absent*100/total).loc[0]}, ignore_index=True)

         i = np.arange(len(ratio))
         b = 0.3
         plt.subplots(figsize=(i[-1]+3,5))
         plt.bar(i  , ratio['with']   , color = 'b', width = b, label='With')
         plt.bar(i+b, ratio['without'], color = 'g', width = b, label='Without')
```

```
plt.xticks(i+b/2,' w   w/o\n'+ratio.disease)

plt.title("Absence Ratio for people with specific disease and without it.",loc='left')
plt.xlabel("\nDiseases")
plt.ylabel("ratio of absence %")
plt.legend()
plt.show()
ratio
```



Out[16]:

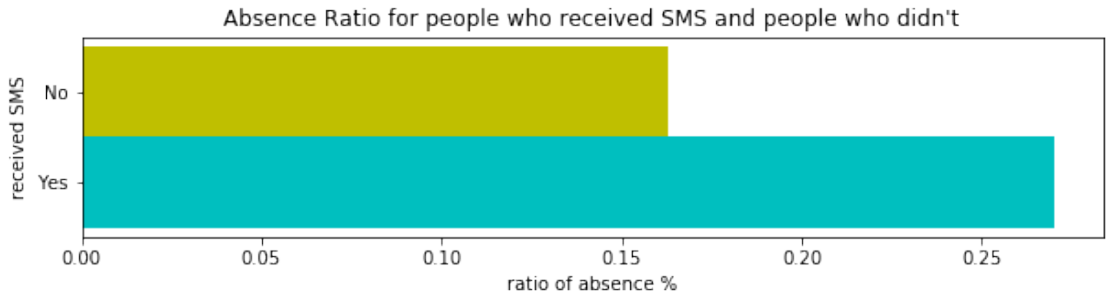
	disease	with	without
0	Hypertension	17.284291	20.696361
1	Diabetes	17.952279	20.023550
2	Handicap	16.281513	19.912363
3	Alcoholism	20.011965	19.825658

Observation - Absence Ratio for patients with Hypertension is 17.3%, and 20.7% for those without it - Absence Ratio for patients with Diabetes is 18%, and 20% for those without it - Absence Ratio for patients with Handicap is 16.3%, and 19.9% for those without it

Research Question 3 (What is Absence Ratio for people who received SMS ant those who didn't?)

```
In [17]: total  = df.SMS_received.value_counts()
         absent = df.query('No_show == "Yes"').SMS_received.value_counts()
         ratio  = absent/total
         print(ratio)

         b = 0.2
         plt.subplots(figsize=(10,2))
         plt.barh(0, ratio.loc[1], color = 'c', height=b)
         plt.barh(b, ratio.loc[0], color = 'y', height=b)
         plt.yticks([0,b],['Yes','No'])
         plt.title("Absence Ratio for people who received SMS and people who didn't")
         plt.xlabel("ratio of absence %")
         plt.ylabel("received SMS")
         plt.show()
```

```
0    0.163149
1    0.270157
Name: SMS_received, dtype: float64
```



Observation - Absence Ratio for patients who received an SMS is 27%, and 16.3% for those who didn't receive any.

Research Question 4 (What is Absence Ratio for patients with specefic characteristics?)

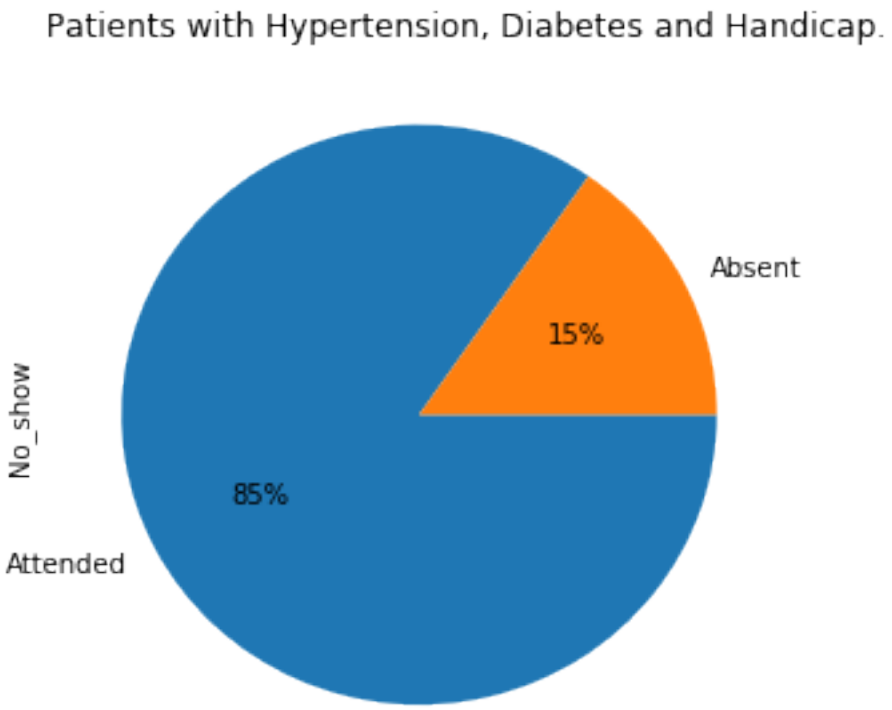
```
>
A. Patients with Hypertension, Diabetes and Handicap.
B. Patients received SMS and having scholarship.
C. Patients waited more than one day.
D. Patients who have their appointments at the same day.
```

```
In [18]: def combo_effect(factors):
             """
```

```
            For patients with certain characteristics combined
            how much percent showed up and how much absent?

            **REQUIRED Packages** pandas and matplotlib

            Arguments: list with one or more than one factor as strings.

            Outputs: a pie showing their absence rate
            """
            q_string = ''
            for factor in factors:
                q_string += factor + ' >= 1 and '
            q_string = q_string[:-4]

            matchs = df.query(q_string).No_show.value_counts()
            matchs.plot(kind= 'pie', labels = ['Attended', 'Absent'],
                        figsize=(5, 5), counterclock = False, autopct='%1.0f%%')

            msg = str(matchs.sum()) +' patients meet this characteristics\n'
            print(msg,matchs)
```

```
In [19]: combo_effect(['Hypertension','Diabetes','Handicap'])
         plt.title("Patients with Hypertension, Diabetes and Handicap.", loc='left');
```

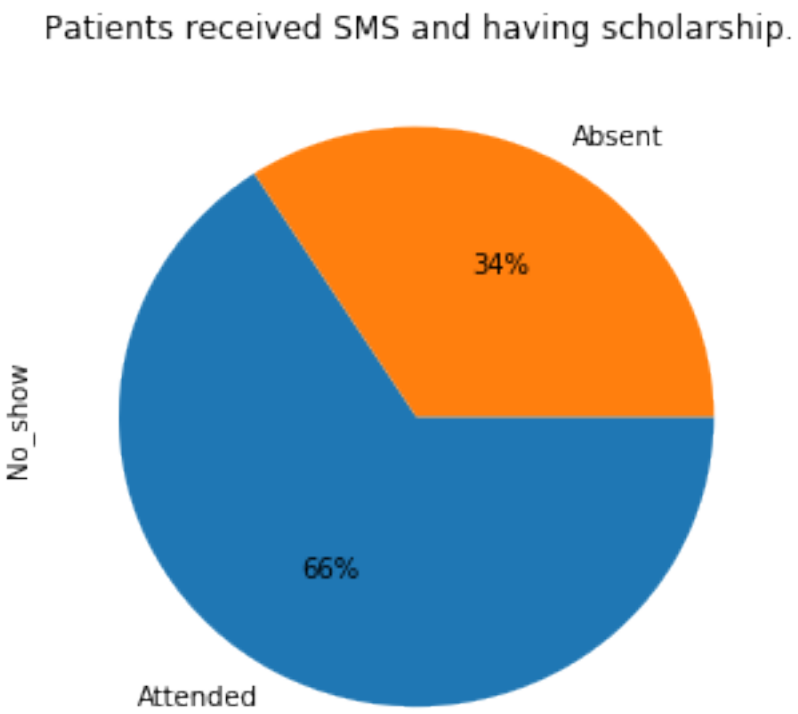
```
358 patients meet this characteristics
 No     303
Yes     55
Name: No_show, dtype: int64
```



Observation - Absence Ratio for patients with Hypertension, Diabetes and Handicap is 15%

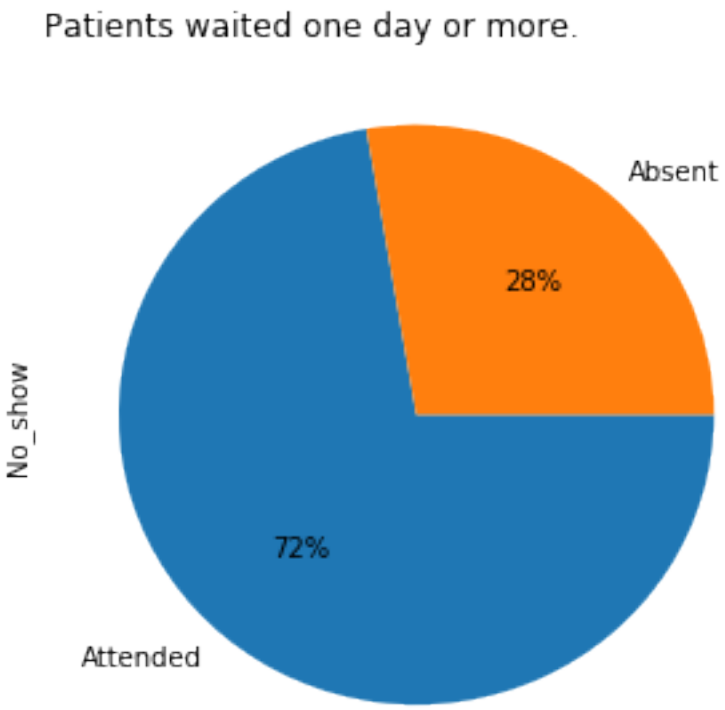
```
In [20]: combo_effect(['SMS_received','Scholarship'])
         plt.title("Patients received SMS and having scholarship.", loc='left');
```

```
2740 patients meet this characteristics
 No     1804
Yes     936
Name: No_show, dtype: int64
```



Observation - Absence Ratio for patients received SMS and having scholarship is 34%

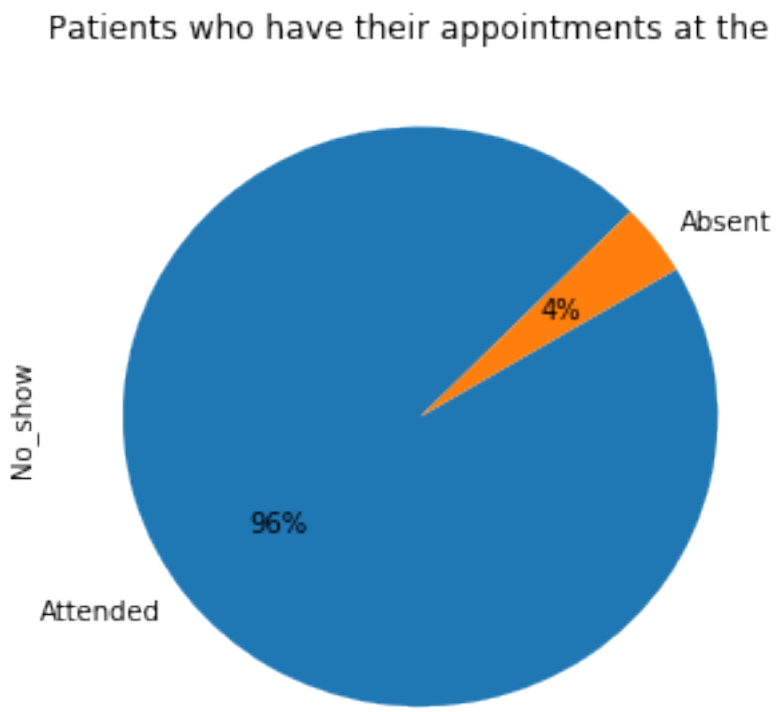
```
In [21]: combo_effect(['WaitingDays'])
         plt.title("Patients waited one day or more.", loc='left');

57528 patients meet this characteristics
 No     41602
Yes    15926
Name: No_show, dtype: int64
```



Observation - Absence Ratio for patients waited one day or more is 28%

```
In [22]: matchs = df.query('WaitingDays == 0').No_show.value_counts()
         matchs.plot(kind= 'pie', labels = ['Attended', 'Absent'],
                     figsize=(5, 5), startangle = 45, autopct='%1.0f%%')
         plt.title("Patients who have their appointments at the same day", loc='left')

         msg = str(matchs.sum()) +' patients meet this characteristics\n'
         print(msg,matchs)
```

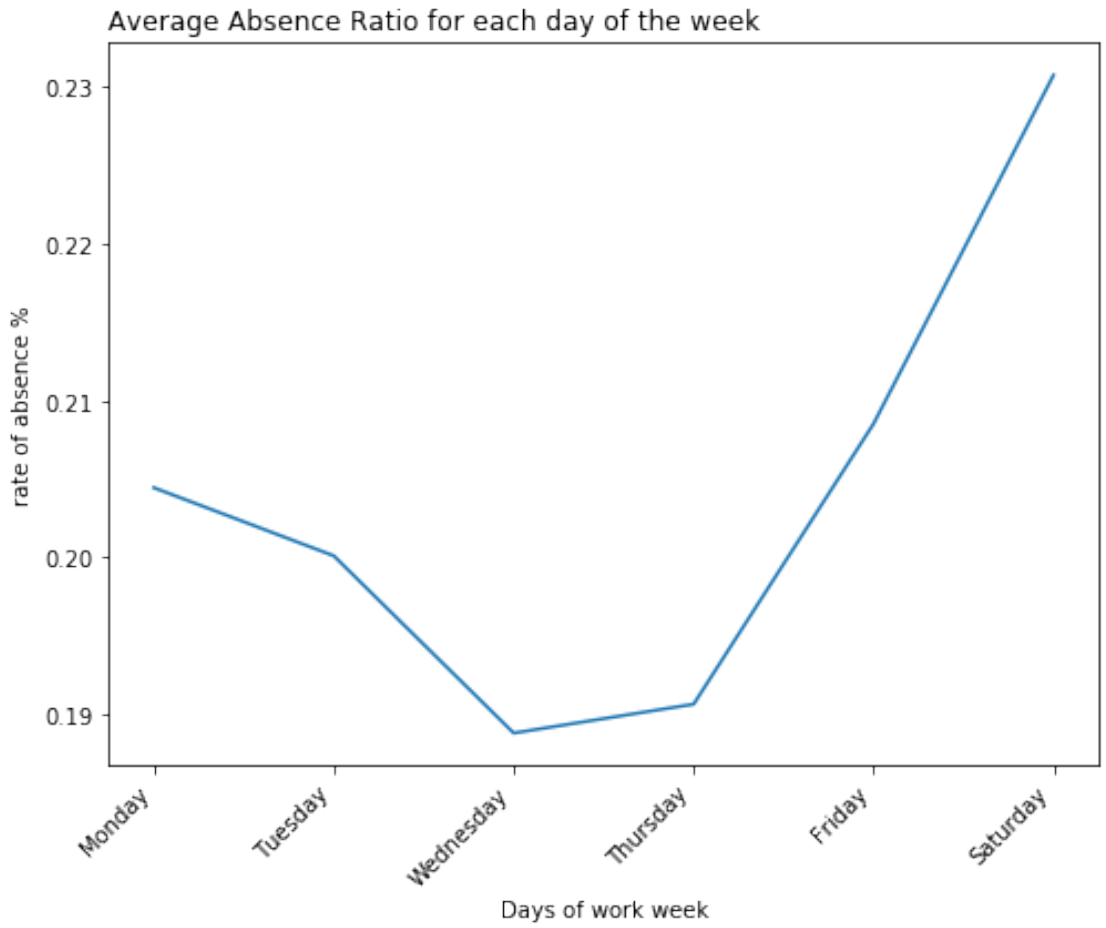
```
28526 patients meet this characteristics
 No     27385
Yes     1141
Name: No_show, dtype: int64
```



Observation - Absence Ratio for patients who have their appointments at the same day is 4%

Research Question 5 (How does the absence rate change over the course of the week?)

```
In [23]: days=['Monday', 'Tuesday', 'Wednesday', 'Thursday','Friday','Saturday']
         days_not = df.query('No_show == "Yes"').AppointmentDay.value_counts()
         days_tot = df.AppointmentDay.value_counts()
         proportion = days_not/days_tot

         x= np.arange(len(days))
         plt.figure(figsize= (8,6))
         plt.plot(x, proportion[days])
         plt.xticks(x, days, rotation=45, horizontalalignment='right')

         plt.title("Average Absence Ratio for each day of the week",loc='left')
         plt.ylabel("rate of absence %")
         plt.xlabel("Days of work week")
         plt.show()
         proportion[days]
```




```
Out[23]: Monday       0.204453
         Tuesday      0.200109
         Wednesday    0.188801
         Thursday     0.190642
         Friday       0.208533
         Saturday     0.230769
         Name: AppointmentDay, dtype: float64
```

Observation - Absence Ratio is lower in the middle of the week. - lowest average Absence Ratio is 18.9% in Wednesdays. - Highest average Absence Ratio is 23.1% in Saturdays.

Research Question 6 (What is the correlation between Absence Ratio and the following factors?) - Age - Number of waiting days

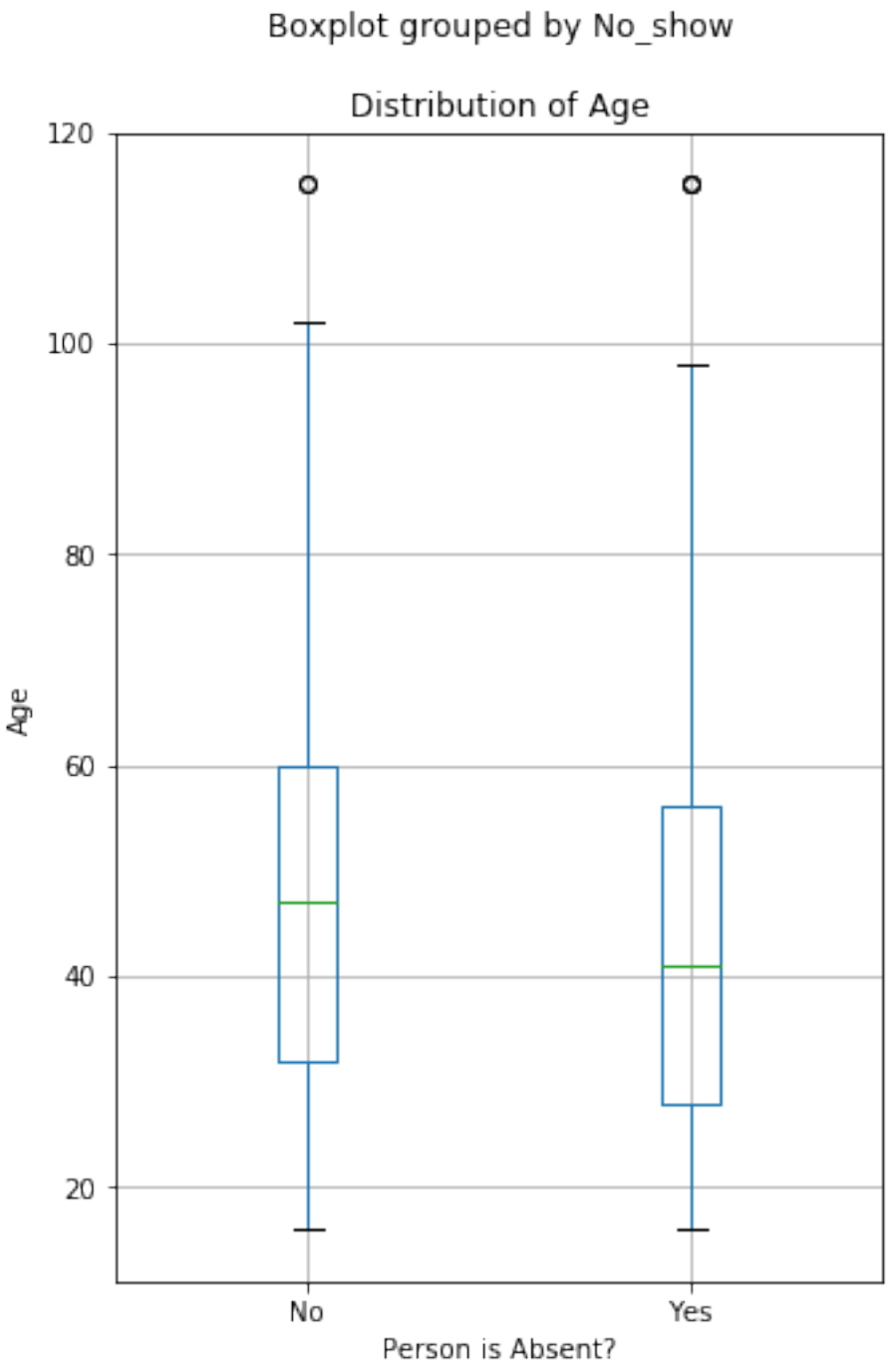
```
In [24]: def factor_box(factor):
             """
             box plot showing the distribution of the given factor according to No_show column.
```

```
    """
    df.boxplot(column = factor, by= 'No_show', rot=0,figsize=(5,8))
    plt.title('Distribution of '+factor)
    plt.xlabel('Person is Absent?')
    plt.ylabel(factor)


def factor_scatter(factor):
    """
    scatter plot to show the correlation between the given factor and Absence Ratio
    """
    ratio = df.query('No_show == "Yes"')[factor].value_counts()/df[factor].value_counts
    i = np.array(ratio.index)
    plt.subplots(figsize=(10,5))
    plt.scatter(x = i , y = ratio.loc[i], color='darkorange')
    plt.xticks(range(0,len(i)+50,10))
    plt.title("Correlation between Absence Ratio and "+factor)
    plt.ylabel("Absence Ratio")
    plt.xlabel(factor)
    plt.show()
```

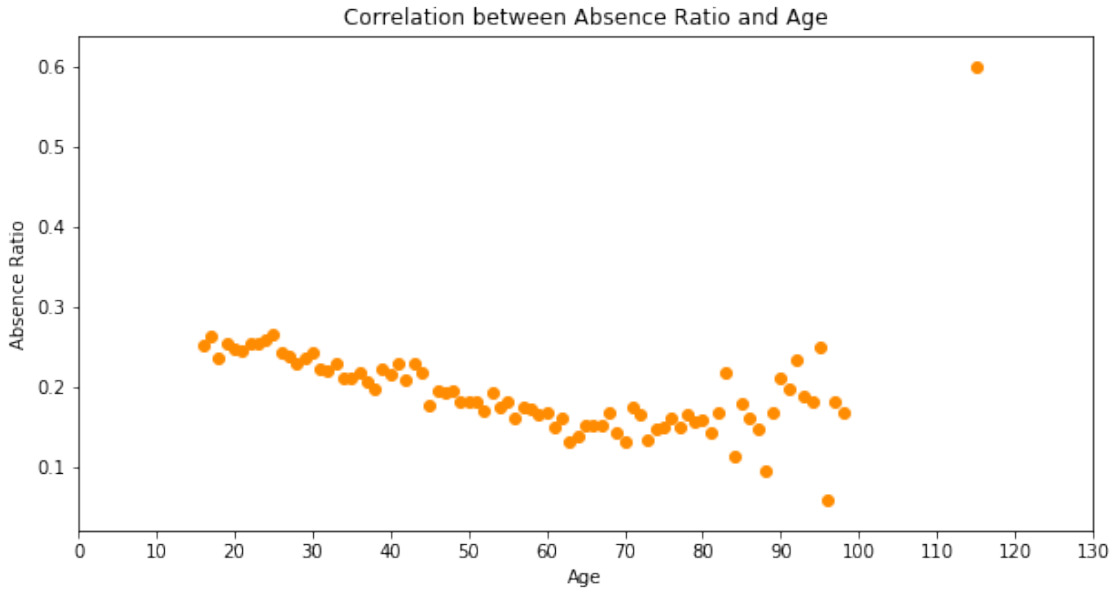
In [25]:
```
factor_box("Age")
pd.DataFrame(df.groupby(['No_show']).Age.describe())
```

Out[25]:

No_show	count	mean	std	min	25%	50%	75%	max
No	68987.0	46.635062	18.197023	16.0	32.0	47.0	60.0	115.0
Yes	17067.0	42.729302	17.962758	16.0	28.0	41.0	56.0	115.0




```
In [26]: factor_scatter("Age")
         pd.DataFrame(df.Age.describe()).transpose()
```

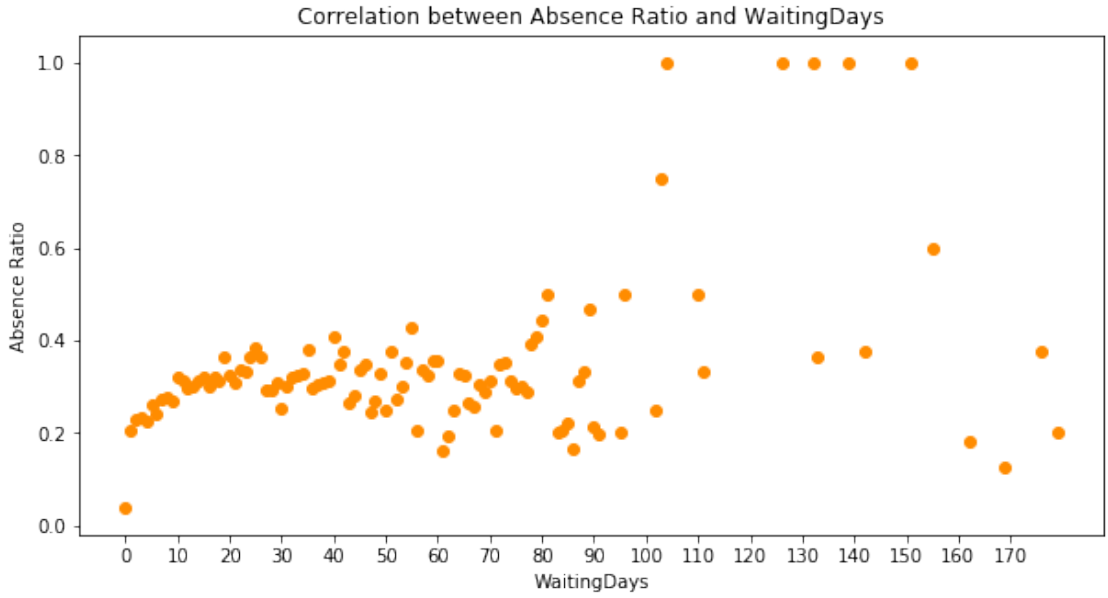


Out[26]:

	count	mean	std	min	25%	50%	75%	max
Age	86054.0	45.860436	18.217391	16.0	31.0	45.0	59.0	115.0

Observation - There is a negative correlation between Absence Ratio and Age.

In [27]:
```
factor_scatter("WaitingDays")
pd.DataFrame(df.WaitingDays.describe()).transpose()
```



Out[27]:

	count	mean	std	min	25%	50%	75%	max
WaitingDays	86054.0	10.320159	15.64358	0.0	0.0	4.0	15.0	179.0

Observation - There is a positive correlation between Absence Ratio and WaitingDays for waiting days from 1 to 30 days, and for Waiting days more than 30 the correlation is hard to figure out.

Research Question 6 (For neighbourhoods hosted most of the appointments, what is the relation between Absence Ratio and number of hosted appointments?)

In [28]:
```
hosts = df.Neighbourhood.value_counts()
count = hosts.sum()
indices = hosts.index

top_hosts = []
top_count = 0
i = 0

while top_count < count/2 :
    host = indices[i]
    top_hosts.append(host)
    top_count = hosts[top_hosts].sum()
    i +=1

other_hosts = hosts.drop(top_hosts).index

print('top', len(top_hosts), 'hosted', hosts[top_hosts].sum(), 'appointments')
print('all', len(indices), 'neighbourhoods hosted', hosts.sum(), 'appointments')
hosts.head()
```

```
top 16 hosted 43538 appointments
all 81 neighbourhoods hosted 86054 appointments
```

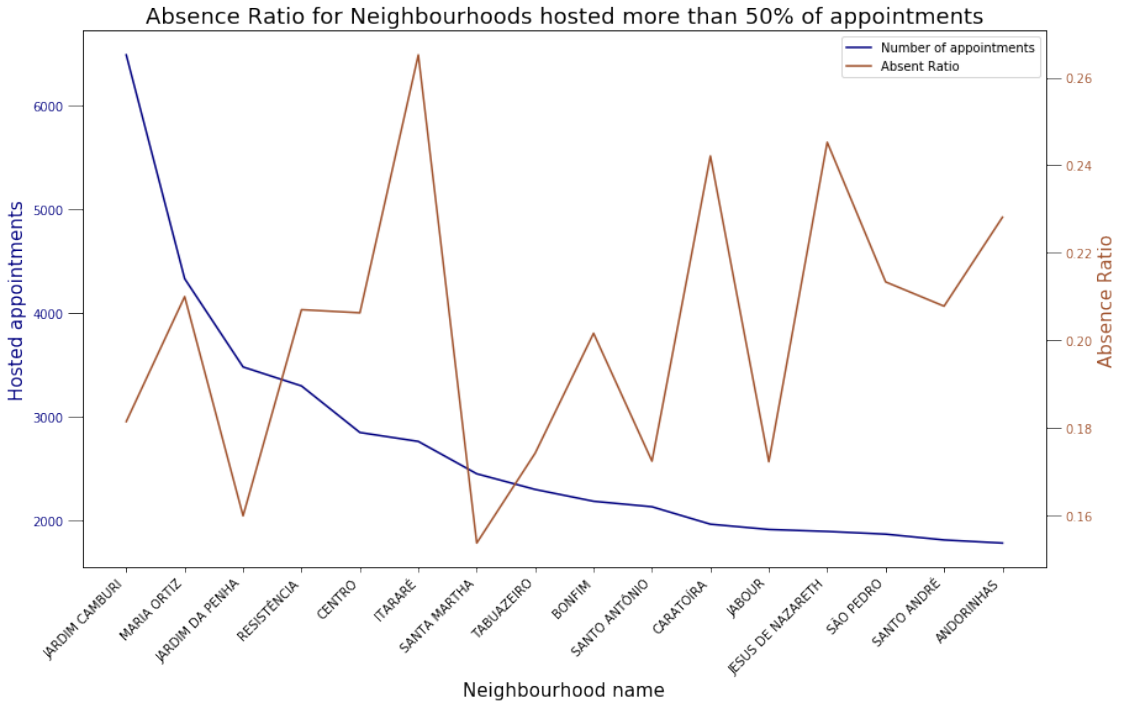
Out[28]:
```
JARDIM CAMBURI     6491
MARIA ORTIZ        4333
JARDIM DA PENHA    3482
RESISTÊNCIA        3299
CENTRO             2850
Name: Neighbourhood, dtype: int64
```

Action aquiring the neighbourhoods that hosted most of the appointments (more than 50%)

Observation 16 ot of 81 neighbourhoods hosted 50,6% of all the appointments

In [29]:
```
hosts_absent = df.query('No_show == "Yes"').Neighbourhood.value_counts()
hosts_ratio = hosts_absent/hosts
x = np.arange(len(top_hosts))
```

```
fig, ax1 = plt.subplots(figsize=(14,8))

ax1.set_xlabel('Neighbourhood name',fontsize = 15)
ax1.set_ylabel('Hosted appointments', color = 'navy', fontsize = 15)
plot_1 = ax1.plot(x , hosts[top_hosts], color = 'navy', label = "Number of appointments
ax1.tick_params(axis ='y', labelcolor = 'navy', size = 12)
plt.xticks(x, top_hosts, rotation=45, horizontalalignment='right')


ax2 = ax1.twinx()

ax2.set_ylabel('Absence Ratio', color = 'sienna', fontsize = 15)
plot_2 = ax2.plot(x , hosts_ratio[top_hosts], color = 'sienna', label = "Absent Ratio")
ax2.tick_params(axis ='y', labelcolor = 'sienna', size = 12)


lns = plot_1 + plot_2
labels = [l.get_label() for l in lns]
plt.legend(lns, labels, loc=0)
plt.title("Absence Ratio for Neighbourhoods hosted more than 50% of appointments",size=
plt.show()
```



Observation - There is no relation between Absence Ratio and number of hosted ap-

pointments.

Conclusions

What is the absence ratio generally, and for each gender? - General Absence Ratio is 20% - Females Absence Ratio is 20% - Males Absence Ratio is 19%

What is the effect of each disease on the Absence Ratio? - Absence Ratio for patients with Hypertension is 17.3%, and 20.7% for those without it - Absence Ratio for patients with Diabetes is 18%, and 20% for those without it - Absence Ratio for patients with Handicap is 16.3%, and 19.9% for those without it

What is Absence Ratio for people who received SMS ant those who didn't? - Absence Ratio for patients who received an SMS is 27%, and 16.3% for those who didn't receive any.

What is Absence Ratio for people with specefic characteristics? - Absence Ratio for patients with Hypertension, Diabetes and Handicap is 15% - Absence Ratio for patients received SMS and having scholarship is 34% - Absence Ratio for patients waited one day or more is 28% - Absence Ratio for patients who have their appointments at the same day is 4%

How does the absence rate change over the course of the week? - Absence Ratio is lower in the middle of the week. - lowest average Absence Ratio is 18.9% in Wednesdays. - Highest average Absence Ratio is 23.1% in Saturdays.

What is the correlation between Absence Ratio and the following factors? - **Age:** there is a negative correlation between Absence Ratio and Age. - **Waiting days:** there is a positive correlation between Absence Ratio and WaitingDays for waiting days from 1 to 30 days, and for Waiting days more than 30 the correlation is hard to figure out.

For neighbourhoods hosted most of the appointments, what is the relation between Absence Ratio and number of hosted appointments? - 16 ot of 81 neighbourhoods hosted 50,6% of all the appointments - There is no relation between Absence Ratio and number of hosted appointments.

1.2 Limitations

Population >Brazil had an estimated population of 215 Million in 2022, with a 0.7% growth rate, according to IBGE (Instituto Brasileiro de Geografia e Estatistica). Brazil is the seventh most populous country in the world.

Vitória, the city from which the data were collected, had an estimated population of 369k according to IBGE, and the city is the capital of the state of Espírito Santo.

This dataset collects informations from 100k medical appointments and after wrangling our dataset the number reduced to be 86k appointments, this number is only 0.04% of Brasil population, and 23% of Vitória population.

Time >The data covers a short time period with olny 3 months, and there is no information about the appointment time during the day.

Locations > Some neighbourhoods have higher absence rate than others, this might depend in other factors such as the distance from the patient's home to the hospital, the avilability of transport, the level of service in that neighbourhood, and all this imformations is missed.