Investigate_a_Dataset

April 12, 2019

> **Tip**: Welcome to the Investigate a Dataset project! You will find tips in quoted sections like this to help organize your approach to your investigation. Before submitting your project, it will be a good idea to go back through your report and remove these sections to make the presentation of your work as tidy as possible. First things first, you might want to double-click this Markdown cell and change the title so that it reflects your dataset and investigation.

1 Project: Investigate a Dataset (No Show Appointments)

1.1 Table of Contents

## Introduction

I have choosed to use the no show appointments, as it relevant to my field of work, and I have been alwasy wondering how can we predict the no show, so here is my analysis for this data set.

Main Questions I tried to answer:

1. Does the time delta (Wait Time) between the Scd_day and app_day is a reason for no show ?
2. what is the impact of the patient age on the show and no show?

Some other question I thought about:

3. Does the patient area has any relevance to the no show ?

```
In [2]: # Use this cell to set up import statements for all of the packages that you
        #   plan to use.

        # Remember to include a 'magic word' so that your visualizations are plotted
        #   inline with the notebook. See this page for more:
        #   http://ipython.readthedocs.io/en/stable/interactive/magics.html
        import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
import re
import time as t
import datetime as dt
from datetime import date

df = pd.read_csv('noshowappointments-kagglev2-may-2016.csv')
#df = pd.read_csv('modified_no_show_dataset.csv')
```

## Data Wrangling & Cleaning

Tip: In this section of the report, you will load in the data, check for cleanliness, and then trim and clean your dataset for analysis. Make sure that you document your steps carefully and justify your cleaning decisions.

1.1.1 General Properties

```
In [30]: # Load your data and print out a few lines. Perform operations to inspect data
         #   types and look for instances of missing or possibly errant data.
         df_pat = df.PatientId.astype(int,inplace=True)
         df['PatientId'] = df_pat
         df.head()
```

```
Out[30]:          PatientId  AppointmentID Gender          ScheduledDay  \
         0   29872499824296        5642903      F  2016-04-29T18:38:08Z
         1  558997776694438        5642503      M  2016-04-29T16:08:27Z
         2    4262962299951        5642549      F  2016-04-29T16:19:04Z
         3     867951213174        5642828      F  2016-04-29T17:29:31Z
         4    8841186448183        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
```

I converted the patient_id to int to get rid of the decimal format

```
In [31]: cols = df.columns
         new_cols = ['patient_id', 'appointment_id', 'gender', 'scheduled_day', 'appointment_day
                     'hipertension', 'diabetes', 'alcoholism', 'handcap', 'sms_received', 'no_sh
         df.columns = new_cols
         df.head()
```

```
Out[31]:         patient_id  appointment_id gender         scheduled_day  \
         0   29872499824296         5642903      F  2016-04-29T18:38:08Z
         1  558997776694438         5642503      M  2016-04-29T16:08:27Z
         2    4262962299951         5642549      F  2016-04-29T16:19:04Z
         3     867951213174         5642828      F  2016-04-29T17:29:31Z
         4    8841186448183         5642494      F  2016-04-29T16:07:23Z

                 appointment_day  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
```

Changed the column names to make sure its all following the same pattern and causes no issues in futher processing

```
In [32]: #changing cols [6:] to boolean
         df.scholarship = df.scholarship.astype(bool,inplace =True)
         df.hipertension = df.hipertension.astype(bool,inplace = True)
         df.diabetes = df.diabetes.astype(bool, inplace = True)
         df.alcoholism = df.alcoholism.astype(bool, inplace = True)
         df.sms_received = df.sms_received.astype(bool,inplace = True)
         df.head()
```

```
Out[32]:         patient_id  appointment_id gender         scheduled_day  \
         0   29872499824296         5642903      F  2016-04-29T18:38:08Z
         1  558997776694438         5642503      M  2016-04-29T16:08:27Z
         2    4262962299951         5642549      F  2016-04-29T16:19:04Z
         3     867951213174         5642828      F  2016-04-29T17:29:31Z
         4    8841186448183         5642494      F  2016-04-29T16:07:23Z

                 appointment_day  age      neighbourhood  scholarship  hipertension  \
         0  2016-04-29T00:00:00Z   62    JARDIM DA PENHA        False          True
         1  2016-04-29T00:00:00Z   56    JARDIM DA PENHA        False         False
         2  2016-04-29T00:00:00Z   62      MATA DA PRAIA        False         False
         3  2016-04-29T00:00:00Z    8  PONTAL DE CAMBURI        False         False
         4  2016-04-29T00:00:00Z   56    JARDIM DA PENHA        False          True

            diabetes  alcoholism  handcap  sms_received no_show
         0     False       False        0         False      No
```

```
1     False       False        0         False      No
2     False       False        0         False      No
3     False       False        0         False      No
4      True       False        0         False      No
```

Changed the columns with '1' and '0' to boolean for more consistency in the DataSet

```
In [33]: df.no_show.replace('Yes', True, inplace= True)

         df.no_show.replace('No', False, inplace= True)
```

Changed the no_show column also to Boolean

```
In [37]: df.shape

Out[37]: (110527, 14)
```

Checking some characteristics of the DataSet

```
In [3]: df.isnull().sum()

Out[3]: Unnamed: 0         0
        patient_id         0
        appointment_id     0
        gender             0
        scheduled_day      0
        appointment_day    0
        age                0
        neighbourhood      0
        scholarship        0
        hipertension       0
        diabetes           0
        alcoholism         0
        handcap            0
        sms_received       0
        no_show            0
        dtype: int64
```

Checking some characteristics of the DataSet

```
In [4]: df.nunique()

Out[4]: Unnamed: 0         110527
        patient_id          62299
        appointment_id     110527
        gender                  2
        scheduled_day      103549
        appointment_day        27
        age                   104
        neighbourhood          81
```

```
scholarship            2
hipertension           2
diabetes               2
alcoholism             2
handcap                5
sms_received           2
no_show                2
dtype: int64
```

Checking some characteristics of the DataSet

```
In [38]: new_schd_day = []
         for day in df.scheduled_day:
             d = dt.datetime.strptime(day, '%Y-%m-%dT%H:%M:%SZ')
             new_schd_day.append(d.strftime('%Y-%m-%d %H:%M:%S'))
         new_schd_day[:5]
         df.scheduled_day = new_schd_day
```

Working on the Date and Time format to be standrized

```
In [39]: new_appt_day = []
         for day in df.appointment_day:
             d = dt.datetime.strptime(day, '%Y-%m-%dT%H:%M:%SZ')
             new_appt_day.append(d.strftime('%Y-%m-%d'))
         new_appt_day[:5]
         df.appointment_day = new_appt_day
```

Working on the Date and Time format to be standrized

```
In [40]: # df = df.drop(df.age == -1, inplace = True)
         df.drop(df[df.age == -1].index, inplace=True)
         df.reset_index(drop=True,inplace=True)
```

Droping any record with -1 as age, eventhoug it should not affect the statistics, but for age groups

```
In [41]: df_new = pd.read_csv('appt_day_less_schd_day.csv')
         to_be_removed = df_new.appointment_id
         appt_id = df.query('appointment_id in @to_be_removed').appointment_id
         df.drop(appt_id.index,inplace=True)
         df.reset_index(drop=True,inplace=True)
         df

Out[41]:             patient_id  appointment_id gender        scheduled_day  \
         0       29872499824296         5642903      F  2016-04-29 18:38:08
         1      558997776694438         5642503      M  2016-04-29 16:08:27
         2        4262962299951         5642549      F  2016-04-29 16:19:04
         3         867951213174         5642828      F  2016-04-29 17:29:31
         4        8841186448183         5642494      F  2016-04-29 16:07:23
```

```
5         95985133231274         5626772      F  2016-04-27 08:36:51
6        733688164476661         5630279      F  2016-04-27 15:05:12
7          3449833394123         5630575      F  2016-04-27 15:39:58
8         56394729949972         5638447      F  2016-04-29 08:02:16
9         78124564369297         5629123      F  2016-04-27 12:48:25
10       734536231958495         5630213      F  2016-04-27 14:58:11
11         7542951368435         5620163      M  2016-04-26 08:44:12
12       566654781423437         5634718      F  2016-04-28 11:33:51
13       911394617215919         5636249      M  2016-04-28 14:52:07
14        99884723334928         5633951      F  2016-04-28 10:06:24
15           99948393975         5620206      F  2016-04-26 08:47:27
16        84574392942817         5633121      M  2016-04-28 08:51:47
17        14794966191172         5633460      F  2016-04-28 09:28:57
18        17135378245248         5621836      F  2016-04-26 10:54:18
19         7223289184215         5640433      F  2016-04-29 10:43:14
20       622257462899397         5626083      F  2016-04-27 07:51:14
21        12154843752835         5628338      F  2016-04-27 10:50:45
22       863229818887631         5616091      M  2016-04-25 13:29:16
23       213753979425692         5634142      F  2016-04-28 10:27:05
24         8734857996885         5641780      F  2016-04-29 14:19:19
25         5819369978796         5624020      M  2016-04-26 15:04:17
26           25787851512         5641781      F  2016-04-29 14:19:42
27        12154843752835         5628345      F  2016-04-27 10:51:45
28         5926171692527         5642400      M  2016-04-29 15:48:02
29         1225776163665         5642186      F  2016-04-29 15:16:29
...                  ...             ...    ...                  ...
110491   793589177751417         5757745      M  2016-06-01 09:46:33
110492    94336536145654         5787655      F  2016-06-08 10:21:14
110493   821969177626116         5757697      F  2016-06-01 09:42:56
110494   443438443334614         5787233      F  2016-06-08 09:35:13
110495      454425189389         5758133      M  2016-06-01 10:19:12
110496   731622885364982         5787937      F  2016-06-08 10:50:42
110497    23621816822757         5759473      F  2016-06-01 13:00:36
110498     9947982555566         5788052      F  2016-06-08 11:06:21
110499    56673438855979         5758455      F  2016-06-01 10:45:50
110500      897388334326         5758779      M  2016-06-01 11:09:20
110501   476946211846992         5786918      F  2016-06-08 09:04:18
110502    94336536145654         5757656      F  2016-06-01 09:41:00
110503   495296829375937         5786750      M  2016-06-08 08:50:51
110504    23621816822757         5757587      F  2016-06-01 09:35:48
110505      823599626588         5786742      F  2016-06-08 08:50:20
110506    98762456447375         5786368      F  2016-06-08 08:20:01
110507    86747784995281         5785964      M  2016-06-08 07:52:55
110508     2695685177138         5786567      F  2016-06-08 08:35:31
110509   645634214296344         5778621      M  2016-06-06 15:58:05
110510    69237724436761         5780205      F  2016-06-07 07:45:16
110511     5574942418928         5780122      F  2016-06-07 07:38:34
110512    72633149253362         5630375      F  2016-04-27 15:15:06
```

```
110513   65423877893936        5630447      F  2016-04-27 15:23:14
110514  996997666245785        5650534      F  2016-05-03 07:51:47
110515   36355337746436        5651072      F  2016-05-03 08:23:40
110516    2572134369293        5651768      F  2016-05-03 09:15:35
110517    3596266328735        5650093      F  2016-05-03 07:27:33
110518   15576631729893        5630692      F  2016-04-27 16:03:52
110519   92134931435557        5630323      F  2016-04-27 15:09:23
110520  377511518121127        5629448      F  2016-04-27 13:30:56

       appointment_day  age      neighbourhood  scholarship  hipertension  \
0           2016-04-29   62    JARDIM DA PENHA        False          True
1           2016-04-29   56    JARDIM DA PENHA        False         False
2           2016-04-29   62      MATA DA PRAIA        False         False
3           2016-04-29    8  PONTAL DE CAMBURI        False         False
4           2016-04-29   56    JARDIM DA PENHA        False          True
5           2016-04-29   76          REPÚBLICA        False          True
6           2016-04-29   23         GOIABEIRAS        False         False
7           2016-04-29   39         GOIABEIRAS        False         False
8           2016-04-29   21         ANDORINHAS        False         False
9           2016-04-29   19          CONQUISTA        False         False
10          2016-04-29   30     NOVA PALESTINA        False         False
11          2016-04-29   29     NOVA PALESTINA        False         False
12          2016-04-29   22     NOVA PALESTINA         True         False
13          2016-04-29   28     NOVA PALESTINA        False         False
14          2016-04-29   54     NOVA PALESTINA        False         False
15          2016-04-29   15     NOVA PALESTINA        False         False
16          2016-04-29   50     NOVA PALESTINA        False         False
17          2016-04-29   40          CONQUISTA         True         False
18          2016-04-29   30     NOVA PALESTINA         True         False
19          2016-04-29   46           DA PENHA        False         False
20          2016-04-29   30     NOVA PALESTINA        False         False
21          2016-04-29    4          CONQUISTA        False         False
22          2016-04-29   13          CONQUISTA        False         False
23          2016-04-29   46          CONQUISTA        False         False
24          2016-04-29   65         TABUAZEIRO        False         False
25          2016-04-29   46          CONQUISTA        False          True
26          2016-04-29   45     BENTO FERREIRA        False          True
27          2016-04-29    4          CONQUISTA        False         False
28          2016-04-29   51          SÃO PEDRO        False         False
29          2016-04-29   32       SANTA MARTHA        False         False
...                ...  ...                ...          ...           ...
110491      2016-06-01   76        MARIA ORTIZ        False         False
110492      2016-06-08   59        MARIA ORTIZ        False         False
110493      2016-06-01   66        MARIA ORTIZ        False          True
110494      2016-06-08   59        MARIA ORTIZ        False         False
110495      2016-06-01   44        MARIA ORTIZ        False         False
110496      2016-06-08   22         GOIABEIRAS        False         False
110497      2016-06-01   64       SOLON BORGES        False         False
```

```
110498      2016-06-08    4        MARIA ORTIZ        False         False
110499      2016-06-01   55        MARIA ORTIZ        False         False
110500      2016-06-01    5        MARIA ORTIZ        False         False
110501      2016-06-08    0        MARIA ORTIZ        False         False
110502      2016-06-01   59        MARIA ORTIZ        False         False
110503      2016-06-08   33        MARIA ORTIZ        False         False
110504      2016-06-01   64       SOLON BORGES        False         False
110505      2016-06-08   14        MARIA ORTIZ        False         False
110506      2016-06-08   41        MARIA ORTIZ        False         False
110507      2016-06-08    2    ANTÔNIO HONÓRIO        False         False
110508      2016-06-08   58        MARIA ORTIZ        False         False
110509      2016-06-08   33        MARIA ORTIZ        False          True
110510      2016-06-08   37        MARIA ORTIZ        False         False
110511      2016-06-07   19        MARIA ORTIZ        False         False
110512      2016-06-07   50        MARIA ORTIZ        False         False
110513      2016-06-07   22        MARIA ORTIZ        False         False
110514      2016-06-07   42        MARIA ORTIZ        False         False
110515      2016-06-07   53        MARIA ORTIZ        False         False
110516      2016-06-07   56        MARIA ORTIZ        False         False
110517      2016-06-07   51        MARIA ORTIZ        False         False
110518      2016-06-07   21        MARIA ORTIZ        False         False
110519      2016-06-07   38        MARIA ORTIZ        False         False
110520      2016-06-07   54        MARIA ORTIZ        False         False

        diabetes  alcoholism  handcap  sms_received  no_show
0          False       False        0         False    False
1          False       False        0         False    False
2          False       False        0         False    False
3          False       False        0         False    False
4           True       False        0         False    False
5          False       False        0         False    False
6          False       False        0         False     True
7          False       False        0         False     True
8          False       False        0         False    False
9          False       False        0         False    False
10         False       False        0         False    False
11         False       False        0          True     True
12         False       False        0         False    False
13         False       False        0         False    False
14         False       False        0         False    False
15         False       False        0          True    False
16         False       False        0         False    False
17         False       False        0         False     True
18         False       False        0          True    False
19         False       False        0         False    False
20         False       False        0         False     True
21         False       False        0         False     True
22         False       False        0          True     True
```

```
23         False       False        0         False    False
24         False       False        0         False    False
25         False       False        0          True    False
26         False       False        0         False    False
27         False       False        0         False    False
28         False       False        0         False    False
29         False       False        0         False    False
...          ...         ...      ...           ...      ...
110491     False       False        0         False    False
110492     False       False        0         False    False
110493      True       False        0         False    False
110494     False       False        0         False    False
110495     False       False        0         False    False
110496     False       False        0         False    False
110497     False       False        0         False    False
110498     False       False        0         False    False
110499     False       False        0         False    False
110500     False       False        0         False    False
110501     False       False        0         False    False
110502     False       False        0         False    False
110503     False       False        0         False    False
110504     False       False        0         False    False
110505     False       False        0         False    False
110506     False       False        0         False    False
110507     False       False        0         False    False
110508     False       False        0         False    False
110509     False       False        0         False     True
110510     False       False        0         False     True
110511     False       False        0         False    False
110512     False       False        0          True    False
110513     False       False        0          True    False
110514     False       False        0          True    False
110515     False       False        0          True    False
110516     False       False        0          True    False
110517     False       False        0          True    False
110518     False       False        0          True    False
110519     False       False        0          True    False
110520     False       False        0          True    False

[110521 rows x 14 columns]
```

Cleaning up records with scheduled_day > appointmet_day

```
In [43]: #adding time delta 'wait_time' column to the Data Set

         appt_day = pd.to_datetime(df.appointment_day)
         schd_day = pd.to_datetime(df.scheduled_day)
```

```
df['wait_time'] = appt_day.dt.date - schd_day.dt.date

df.wait_time = df.wait_time.astype('timedelta64[D]',inplace=True)
df.wait_time = df.wait_time.astype(int)
df.head(100)
```

```
Out[43]:          patient_id  appointment_id gender        scheduled_day  \
         0    29872499824296         5642903      F  2016-04-29 18:38:08
         1   558997776694438         5642503      M  2016-04-29 16:08:27
         2     4262962299951         5642549      F  2016-04-29 16:19:04
         3      867951213174         5642828      F  2016-04-29 17:29:31
         4     8841186448183         5642494      F  2016-04-29 16:07:23
         5    95985133231274         5626772      F  2016-04-27 08:36:51
         6   733688164476661         5630279      F  2016-04-27 15:05:12
         7     3449833394123         5630575      F  2016-04-27 15:39:58
         8    56394729949972         5638447      F  2016-04-29 08:02:16
         9    78124564369297         5629123      F  2016-04-27 12:48:25
         10  734536231958495         5630213      F  2016-04-27 14:58:11
         11    7542951368435         5620163      M  2016-04-26 08:44:12
         12  566654781423437         5634718      F  2016-04-28 11:33:51
         13  911394617215919         5636249      M  2016-04-28 14:52:07
         14   99884723334928         5633951      F  2016-04-28 10:06:24
         15      99948393975         5620206      F  2016-04-26 08:47:27
         16   84574392942817         5633121      M  2016-04-28 08:51:47
         17   14794966191172         5633460      F  2016-04-28 09:28:57
         18   17135378245248         5621836      F  2016-04-26 10:54:18
         19    7223289184215         5640433      F  2016-04-29 10:43:14
         20  622257462899397         5626083      F  2016-04-27 07:51:14
         21   12154843752835         5628338      F  2016-04-27 10:50:45
         22  863229818887631         5616091      M  2016-04-25 13:29:16
         23  213753979425692         5634142      F  2016-04-28 10:27:05
         24    8734857996885         5641780      F  2016-04-29 14:19:19
         25    5819369978796         5624020      M  2016-04-26 15:04:17
         26      25787851512         5641781      F  2016-04-29 14:19:42
         27   12154843752835         5628345      F  2016-04-27 10:51:45
         28    5926171692527         5642400      M  2016-04-29 15:48:02
         29    1225776163665         5642186      F  2016-04-29 15:16:29
         ..              ...             ...    ...                  ...
         70   67144894855774         5552914      M  2016-04-06 17:59:58
         71    1846317738622         5552936      F  2016-04-06 18:12:55
         72    7746485718662         5638014      F  2016-04-29 07:37:37
         73   45421316129453         5552934      F  2016-04-06 18:12:38
         74    9672968175572         5597628      F  2016-04-18 17:29:12
         75     148894173528         5597632      F  2016-04-18 17:32:53
         76    6549277227425         5597643      M  2016-04-18 17:40:18
         77    5753721241256         5642767      F  2016-04-29 17:06:27
         78     625926531749         5597672      M  2016-04-18 17:52:49
         79   99128824246583         5597673      M  2016-04-18 17:53:25
```

```
80    1486714718477         5597685      M  2016-04-18 18:03:12
81   19767951968224         5597689      F  2016-04-18 18:06:35
82     182712485992         5638939      M  2016-04-29 08:36:19
83     227497896765         5637742      M  2016-04-29 07:20:46
84   26879963992389         5637915      F  2016-04-29 07:31:04
85   74727351113223         5623102      F  2016-04-26 13:34:14
86    3376224477447         5595347      M  2016-04-18 12:31:34
87    4143141735632         5595356      M  2016-04-18 12:32:25
88    4448345555999         5595358      M  2016-04-18 12:32:35
89  431493164159576         5640380      M  2016-04-29 10:37:02
90  878252996786747         5595362      M  2016-04-18 12:33:05
91    2294295126913         5598651      F  2016-04-19 07:51:31
92  295467429931514         5638591      M  2016-04-29 08:11:38
93   63225327996426         5639376      F  2016-04-29 09:01:10
94    8192146244379         5640054      M  2016-04-29 10:03:12
95  198624862183842         5640307      M  2016-04-29 10:28:54
96   79376248773989         5623692      M  2016-04-26 14:28:39
97    5253342488842         5565493      F  2016-04-11 09:00:00
98  372596436556933         5571906      F  2016-04-12 09:44:42
99     124621344153         5641893      F  2016-04-29 14:38:28

   appointment_day  age        neighbourhood  scholarship  hipertension  \
0       2016-04-29   62      JARDIM DA PENHA        False          True
1       2016-04-29   56      JARDIM DA PENHA        False         False
2       2016-04-29   62        MATA DA PRAIA        False         False
3       2016-04-29    8    PONTAL DE CAMBURI        False         False
4       2016-04-29   56      JARDIM DA PENHA        False          True
5       2016-04-29   76            REPÚBLICA        False          True
6       2016-04-29   23           GOIABEIRAS        False         False
7       2016-04-29   39           GOIABEIRAS        False         False
8       2016-04-29   21           ANDORINHAS        False         False
9       2016-04-29   19            CONQUISTA        False         False
10      2016-04-29   30       NOVA PALESTINA        False         False
11      2016-04-29   29       NOVA PALESTINA        False         False
12      2016-04-29   22       NOVA PALESTINA         True         False
13      2016-04-29   28       NOVA PALESTINA        False         False
14      2016-04-29   54       NOVA PALESTINA        False         False
15      2016-04-29   15       NOVA PALESTINA        False         False
16      2016-04-29   50       NOVA PALESTINA        False         False
17      2016-04-29   40            CONQUISTA         True         False
18      2016-04-29   30       NOVA PALESTINA         True         False
19      2016-04-29   46             DA PENHA        False         False
20      2016-04-29   30       NOVA PALESTINA        False         False
21      2016-04-29    4            CONQUISTA        False         False
22      2016-04-29   13            CONQUISTA        False         False
23      2016-04-29   46            CONQUISTA        False         False
24      2016-04-29   65           TABUAZEIRO        False         False
25      2016-04-29   46            CONQUISTA        False          True
```

```
26      2016-04-29   45       BENTO FERREIRA        False          True
27      2016-04-29    4            CONQUISTA        False         False
28      2016-04-29   51            SÃO PEDRO        False         False
29      2016-04-29   32         SANTA MARTHA        False         False
..             ...  ...                  ...          ...           ...
70      2016-04-29   62         SOLON BORGES        False         False
71      2016-04-29   30               BONFIM         True         False
72      2016-04-29   61       JARDIM CAMBURI        False         False
73      2016-04-29   68            REPÚBLICA        False          True
74      2016-04-29   64          MARIA ORTIZ        False         False
75      2016-04-29   60               JABOUR        False         False
76      2016-04-29   28      ANTÔNIO HONÓRIO        False         False
77      2016-04-29   27               JABOUR        False         False
78      2016-04-29   21          MARIA ORTIZ        False         False
79      2016-04-29   67          MARIA ORTIZ        False         False
80      2016-04-29   68               JABOUR        False         False
81      2016-04-29   49               JABOUR        False         False
82      2016-04-29   71               JABOUR        False         False
83      2016-04-29   36          RESISTÊNCIA        False         False
84      2016-04-29   29          RESISTÊNCIA        False         False
85      2016-04-29   69          RESISTÊNCIA        False          True
86      2016-04-29   10  ILHA DE SANTA MARIA        False         False
87      2016-04-29    2  ILHA DE SANTA MARIA        False         False
88      2016-04-29    1          JUCUTUQUARA        False         False
89      2016-04-29    0           MONTE BELO        False         False
90      2016-04-29   11          JUCUTUQUARA        False         False
91      2016-04-29   10               BONFIM        False         False
92      2016-04-29    2               BONFIM        False         False
93      2016-04-29    1               BONFIM        False         False
94      2016-04-29   10               BONFIM        False         False
95      2016-04-29    1               BONFIM        False         False
96      2016-04-29    3               BONFIM        False         False
97      2016-04-29   35               BONFIM        False         False
98      2016-04-29   51               BONFIM        False         False
99      2016-04-29    1               BONFIM        False         False

    diabetes  alcoholism  handcap  sms_received  no_show  wait_time
0      False       False        0         False    False          0
1      False       False        0         False    False          0
2      False       False        0         False    False          0
3      False       False        0         False    False          0
4       True       False        0         False    False          0
5      False       False        0         False    False          2
6      False       False        0         False     True          2
7      False       False        0         False     True          2
8      False       False        0         False    False          0
9      False       False        0         False    False          2
10     False       False        0         False    False          2
```

```
11     False       False        0          True     True          3
12     False       False        0         False    False          1
13     False       False        0         False    False          1
14     False       False        0         False    False          1
15     False       False        0          True    False          3
16     False       False        0         False    False          1
17     False       False        0         False     True          1
18     False       False        0          True    False          3
19     False       False        0         False    False          0
20     False       False        0         False     True          2
21     False       False        0         False     True          2
22     False       False        0          True     True          4
23     False       False        0         False    False          1
24     False       False        0         False    False          0
25     False       False        0          True    False          3
26     False       False        0         False    False          0
27     False       False        0         False    False          2
28     False       False        0         False    False          0
29     False       False        0         False    False          0
..       ...         ...      ...           ...      ...        ...
70     False       False        0         False    False         23
71     False       False        0          True    False         23
72     False       False        0         False    False          0
73      True       False        0          True    False         23
74     False       False        0          True    False         11
75     False       False        0         False    False         11
76     False       False        0         False     True         11
77     False       False        0         False    False          0
78     False       False        0          True    False         11
79     False       False        0          True     True         11
80     False       False        0          True    False         11
81     False       False        0          True    False         11
82     False       False        0         False    False          0
83     False       False        0         False    False          0
84     False       False        0         False    False          0
85     False       False        0         False    False          3
86     False       False        0          True    False         11
87     False       False        0         False    False         11
88     False       False        0         False    False         11
89     False       False        0         False    False          0
90     False       False        0          True     True         11
91     False       False        0          True    False         10
92     False       False        0         False    False          0
93     False       False        0         False    False          0
94     False       False        0         False    False          0
95     False       False        0         False    False          0
96     False       False        0          True    False          3
97     False       False        0          True    False         18
```

```
98     False        False        0          True    False          17
99     False        False        0         False    False           0

[100 rows x 15 columns]
```

Adding another column to the dataset ’wait_time

```
In [44]: # Adding another column that has the age groups
         df['age_group'] = (df.age.apply(lambda x: min(int(x / 10) , 9))
             )
         df.age_group.value_counts().sort_index()
         type(df.age[5])
         df.head(100)
```

```
Out[44]:          patient_id  appointment_id gender        scheduled_day  \
         0    29872499824296         5642903      F  2016-04-29 18:38:08
         1   558997776694438         5642503      M  2016-04-29 16:08:27
         2     4262962299951         5642549      F  2016-04-29 16:19:04
         3      867951213174         5642828      F  2016-04-29 17:29:31
         4     8841186448183         5642494      F  2016-04-29 16:07:23
         5    95985133231274         5626772      F  2016-04-27 08:36:51
         6   733688164476661         5630279      F  2016-04-27 15:05:12
         7     3449833394123         5630575      F  2016-04-27 15:39:58
         8    56394729949972         5638447      F  2016-04-29 08:02:16
         9    78124564369297         5629123      F  2016-04-27 12:48:25
         10  734536231958495         5630213      F  2016-04-27 14:58:11
         11    7542951368435         5620163      M  2016-04-26 08:44:12
         12  566654781423437         5634718      F  2016-04-28 11:33:51
         13  911394617215919         5636249      M  2016-04-28 14:52:07
         14   99884723334928         5633951      F  2016-04-28 10:06:24
         15      99948393975         5620206      F  2016-04-26 08:47:27
         16   84574392942817         5633121      M  2016-04-28 08:51:47
         17   14794966191172         5633460      F  2016-04-28 09:28:57
         18   17135378245248         5621836      F  2016-04-26 10:54:18
         19    7223289184215         5640433      F  2016-04-29 10:43:14
         20  622257462899397         5626083      F  2016-04-27 07:51:14
         21   12154843752835         5628338      F  2016-04-27 10:50:45
         22  863229818887631         5616091      M  2016-04-25 13:29:16
         23  213753979425692         5634142      F  2016-04-28 10:27:05
         24    8734857996885         5641780      F  2016-04-29 14:19:19
         25    5819369978796         5624020      M  2016-04-26 15:04:17
         26      25787851512         5641781      F  2016-04-29 14:19:42
         27   12154843752835         5628345      F  2016-04-27 10:51:45
         28    5926171692527         5642400      M  2016-04-29 15:48:02
         29    1225776163665         5642186      F  2016-04-29 15:16:29
         ..              ...             ...    ...                  ...
         70   67144894855774         5552914      M  2016-04-06 17:59:58
         71    1846317738622         5552936      F  2016-04-06 18:12:55
```

```
72    7746485718662         5638014      F  2016-04-29 07:37:37
73   45421316129453         5552934      F  2016-04-06 18:12:38
74    9672968175572         5597628      F  2016-04-18 17:29:12
75     148894173528         5597632      F  2016-04-18 17:32:53
76    6549277227425         5597643      M  2016-04-18 17:40:18
77    5753721241256         5642767      F  2016-04-29 17:06:27
78     625926531749         5597672      M  2016-04-18 17:52:49
79   99128824246583         5597673      M  2016-04-18 17:53:25
80    1486714718477         5597685      M  2016-04-18 18:03:12
81   19767951968224         5597689      F  2016-04-18 18:06:35
82     182712485992         5638939      M  2016-04-29 08:36:19
83     227497896765         5637742      M  2016-04-29 07:20:46
84   26879963992389         5637915      F  2016-04-29 07:31:04
85   74727351113223         5623102      F  2016-04-26 13:34:14
86    3376224477447         5595347      M  2016-04-18 12:31:34
87    4143141735632         5595356      M  2016-04-18 12:32:25
88    4448345555999         5595358      M  2016-04-18 12:32:35
89  431493164159576         5640380      M  2016-04-29 10:37:02
90  878252996786747         5595362      M  2016-04-18 12:33:05
91    2294295126913         5598651      F  2016-04-19 07:51:31
92  295467429931514         5638591      M  2016-04-29 08:11:38
93   63225327996426         5639376      F  2016-04-29 09:01:10
94    8192146244379         5640054      M  2016-04-29 10:03:12
95  198624862183842         5640307      M  2016-04-29 10:28:54
96   79376248773989         5623692      M  2016-04-26 14:28:39
97    5253342488842         5565493      F  2016-04-11 09:00:00
98  372596436556933         5571906      F  2016-04-12 09:44:42
99     124621344153         5641893      F  2016-04-29 14:38:28

   appointment_day  age        neighbourhood  scholarship  hipertension  \
0       2016-04-29   62      JARDIM DA PENHA        False          True
1       2016-04-29   56      JARDIM DA PENHA        False         False
2       2016-04-29   62        MATA DA PRAIA        False         False
3       2016-04-29    8    PONTAL DE CAMBURI        False         False
4       2016-04-29   56      JARDIM DA PENHA        False          True
5       2016-04-29   76            REPÚBLICA        False          True
6       2016-04-29   23           GOIABEIRAS        False         False
7       2016-04-29   39           GOIABEIRAS        False         False
8       2016-04-29   21           ANDORINHAS        False         False
9       2016-04-29   19            CONQUISTA        False         False
10      2016-04-29   30       NOVA PALESTINA        False         False
11      2016-04-29   29       NOVA PALESTINA        False         False
12      2016-04-29   22       NOVA PALESTINA         True         False
13      2016-04-29   28       NOVA PALESTINA        False         False
14      2016-04-29   54       NOVA PALESTINA        False         False
15      2016-04-29   15       NOVA PALESTINA        False         False
16      2016-04-29   50       NOVA PALESTINA        False         False
17      2016-04-29   40            CONQUISTA         True         False
```

```
18      2016-04-29   30        NOVA PALESTINA         True         False
19      2016-04-29   46              DA PENHA        False         False
20      2016-04-29   30        NOVA PALESTINA        False         False
21      2016-04-29    4             CONQUISTA        False         False
22      2016-04-29   13             CONQUISTA        False         False
23      2016-04-29   46             CONQUISTA        False         False
24      2016-04-29   65            TABUAZEIRO        False         False
25      2016-04-29   46             CONQUISTA        False          True
26      2016-04-29   45        BENTO FERREIRA        False          True
27      2016-04-29    4             CONQUISTA        False         False
28      2016-04-29   51             SÃO PEDRO        False         False
29      2016-04-29   32          SANTA MARTHA        False         False
..             ...  ...                   ...          ...           ...
70      2016-04-29   62          SOLON BORGES        False         False
71      2016-04-29   30                BONFIM         True         False
72      2016-04-29   61        JARDIM CAMBURI        False         False
73      2016-04-29   68             REPÚBLICA        False          True
74      2016-04-29   64           MARIA ORTIZ        False         False
75      2016-04-29   60                JABOUR        False         False
76      2016-04-29   28       ANTÔNIO HONÓRIO        False         False
77      2016-04-29   27                JABOUR        False         False
78      2016-04-29   21           MARIA ORTIZ        False         False
79      2016-04-29   67           MARIA ORTIZ        False         False
80      2016-04-29   68                JABOUR        False         False
81      2016-04-29   49                JABOUR        False         False
82      2016-04-29   71                JABOUR        False         False
83      2016-04-29   36           RESISTÊNCIA        False         False
84      2016-04-29   29           RESISTÊNCIA        False         False
85      2016-04-29   69           RESISTÊNCIA        False          True
86      2016-04-29   10   ILHA DE SANTA MARIA        False         False
87      2016-04-29    2   ILHA DE SANTA MARIA        False         False
88      2016-04-29    1           JUCUTUQUARA        False         False
89      2016-04-29    0            MONTE BELO        False         False
90      2016-04-29   11           JUCUTUQUARA        False         False
91      2016-04-29   10                BONFIM        False         False
92      2016-04-29    2                BONFIM        False         False
93      2016-04-29    1                BONFIM        False         False
94      2016-04-29   10                BONFIM        False         False
95      2016-04-29    1                BONFIM        False         False
96      2016-04-29    3                BONFIM        False         False
97      2016-04-29   35                BONFIM        False         False
98      2016-04-29   51                BONFIM        False         False
99      2016-04-29    1                BONFIM        False         False

    diabetes  alcoholism  handcap  sms_received  no_show  wait_time  age_group
0      False       False        0         False    False          0          6
1      False       False        0         False    False          0          5
2      False       False        0         False    False          0          6
```

```
3      False      False        0        False    False         0         0
4       True      False        0        False    False         0         5
5      False      False        0        False    False         2         7
6      False      False        0        False     True         2         2
7      False      False        0        False     True         2         3
8      False      False        0        False    False         0         2
9      False      False        0        False    False         2         1
10     False      False        0        False    False         2         3
11     False      False        0         True     True         3         2
12     False      False        0        False    False         1         2
13     False      False        0        False    False         1         2
14     False      False        0        False    False         1         5
15     False      False        0         True    False         3         1
16     False      False        0        False    False         1         5
17     False      False        0        False     True         1         4
18     False      False        0         True    False         3         3
19     False      False        0        False    False         0         4
20     False      False        0        False     True         2         3
21     False      False        0        False     True         2         0
22     False      False        0         True     True         4         1
23     False      False        0        False    False         1         4
24     False      False        0        False    False         0         6
25     False      False        0         True    False         3         4
26     False      False        0        False    False         0         4
27     False      False        0        False    False         2         0
28     False      False        0        False    False         0         5
29     False      False        0        False    False         0         3
..       ...        ...      ...          ...      ...       ...       ...
70     False      False        0        False    False        23         6
71     False      False        0         True    False        23         3
72     False      False        0        False    False         0         6
73      True      False        0         True    False        23         6
74     False      False        0         True    False        11         6
75     False      False        0        False    False        11         6
76     False      False        0        False     True        11         2
77     False      False        0        False    False         0         2
78     False      False        0         True    False        11         2
79     False      False        0         True     True        11         6
80     False      False        0         True    False        11         6
81     False      False        0         True    False        11         4
82     False      False        0        False    False         0         7
83     False      False        0        False    False         0         3
84     False      False        0        False    False         0         2
85     False      False        0        False    False         3         6
86     False      False        0         True    False        11         1
87     False      False        0        False    False        11         0
88     False      False        0        False    False        11         0
89     False      False        0        False    False         0         0
```

```
90      False        False        0         True     True         11          1
91      False        False        0         True    False         10          1
92      False        False        0        False    False          0          0
93      False        False        0        False    False          0          0
94      False        False        0        False    False          0          1
95      False        False        0        False    False          0          0
96      False        False        0         True    False          3          0
97      False        False        0         True    False         18          3
98      False        False        0         True    False         17          5
99      False        False        0        False    False          0          0

[100 rows x 16 columns]
```

Adding another column for age groups

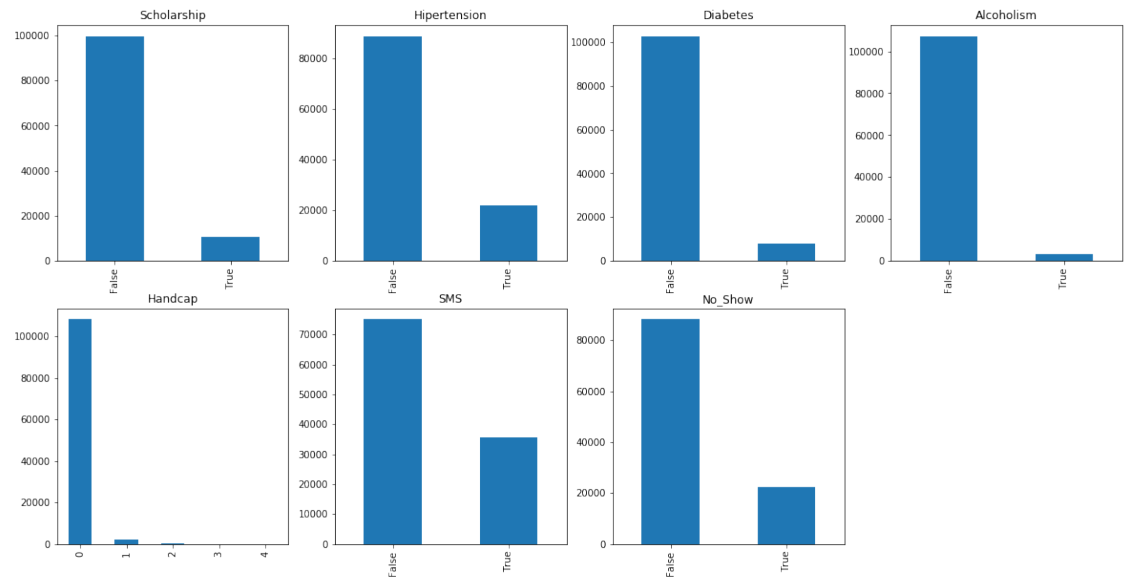
```
In [45]: df.to_csv('modified_no_show_dataset.csv')
```

Saving all of these chnages to a new CS file to avoid editing everytime i run the sheet

1.2 Data Visulization:

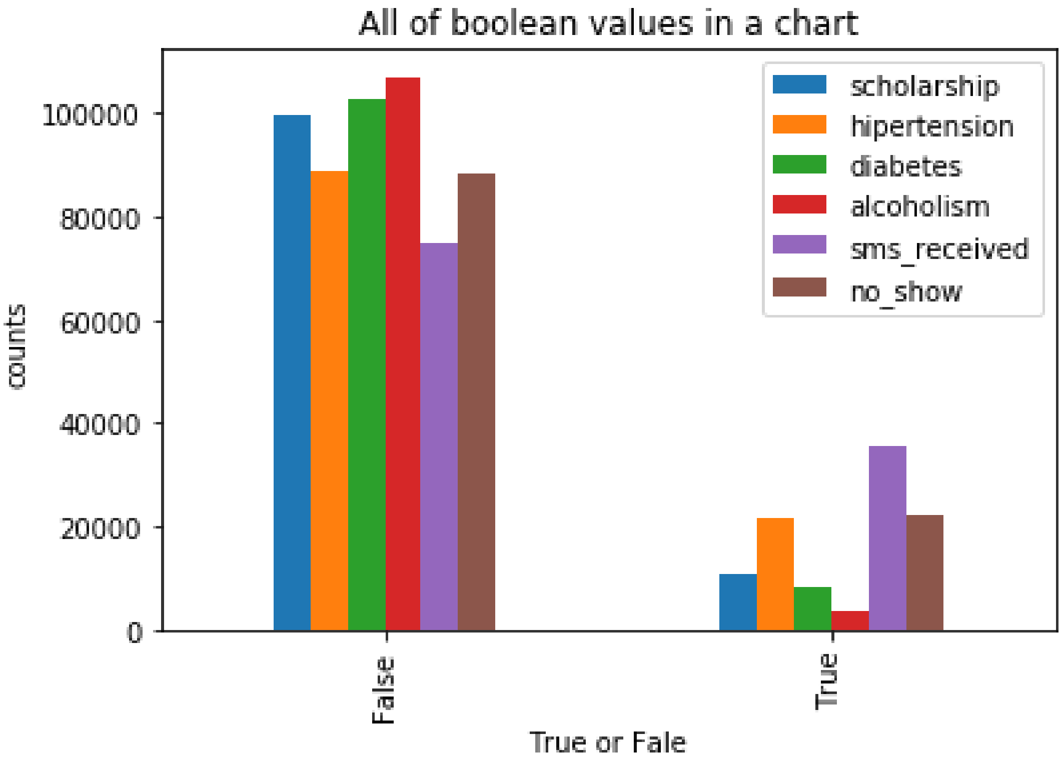
```
In [46]: figure = plt.figure(figsize=(20,10))
         fig_dim = (2,4)

         plt.subplot2grid(fig_dim, (0,0), title='Scholarship')
         df.scholarship.value_counts().plot(kind='bar')

         plt.subplot2grid(fig_dim, (0,1))
         df.hipertension.value_counts().plot(kind='bar', title='Hipertension')

         plt.subplot2grid(fig_dim, (0,2))
         df.diabetes.value_counts().plot(kind='bar', title='Diabetes')

         plt.subplot2grid(fig_dim, (0,3))
         df.alcoholism.value_counts().plot(kind='bar', title='Alcoholism')

         plt.subplot2grid(fig_dim, (1,0))
         df.handcap.value_counts().plot(kind='bar', title='Handcap')

         plt.subplot2grid(fig_dim, (1,1))
         df.sms_received.value_counts().plot(kind='bar', title='SMS')

         plt.subplot2grid(fig_dim, (1,2))
         df.no_show.value_counts().plot(kind='bar', title='No_Show')
```

```
Out[46]: <matplotlib.axes._subplots.AxesSubplot at 0x7fe0171d3c50>
```



Value Counts in bar chart for all of the parameters in the dataset

```
In [15]: df[['scholarship','hipertension', 'diabetes', 'alcoholism', 'sms_received', 'no_show']]
         plt.title('All of boolean values in a chart')
         plt.ylabel('counts')
         plt.xlabel('True or Fale')

Out[15]: Text(0.5,0,'True or Fale')
```



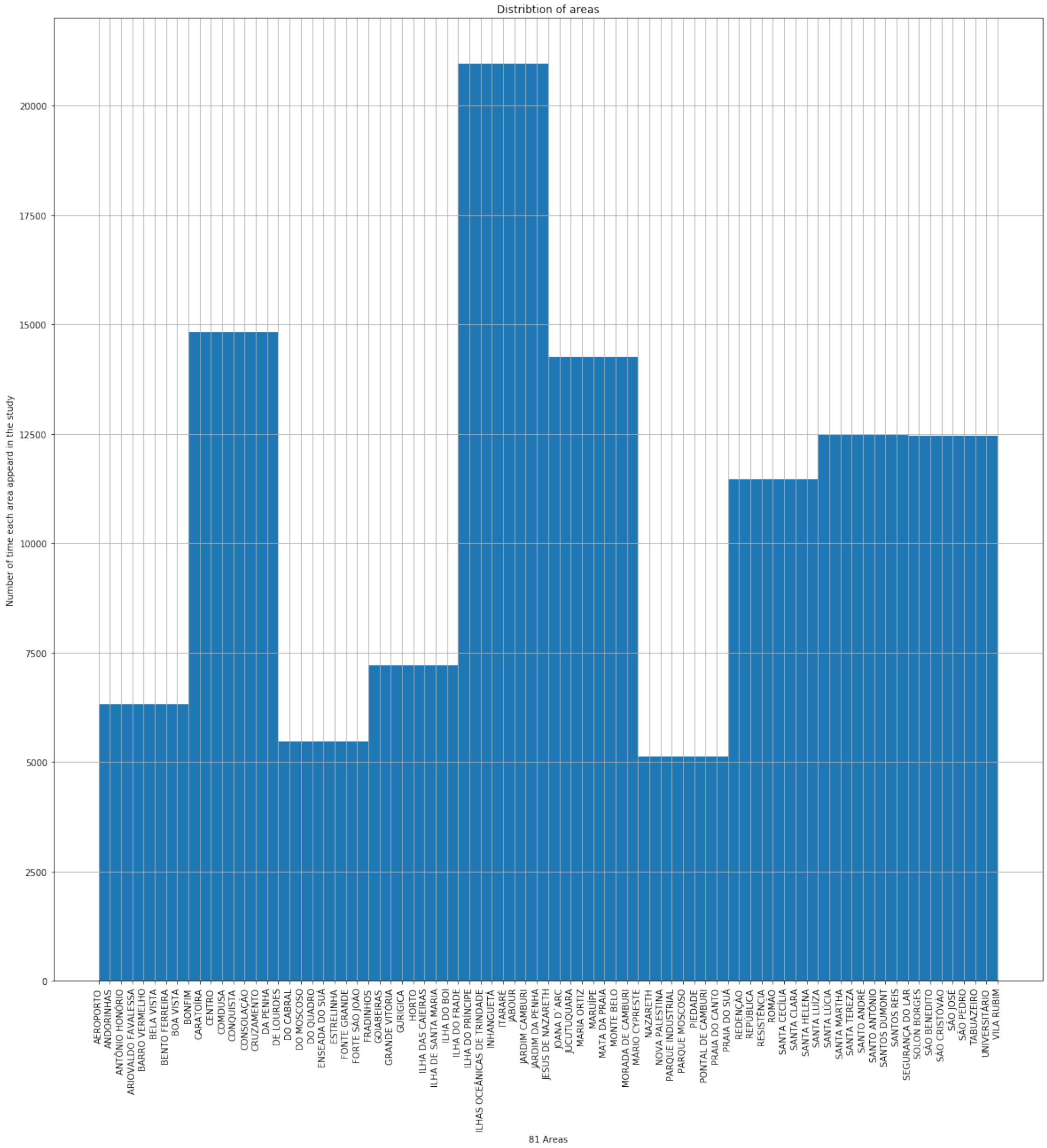
One Bar Chart for all value counts for either True or false

1. So the percentage of no show is around 25%
2. Number of no shows when message received is 9784 (9%)
3. Number of no shows when message wasn't received is 12535 (11.5%)
4. Number of no shows when patient is not hipertensive is 18547 (16.9%)
5. Number of no shows when patient is not on Scholarshop is 19741 (18%)
6. Number of no shows when patient is not Diabetic is 20889 (19%)
7. Number of no shows when patient is not Alcoholic is 21642 (19.7%)
8. Number of no shows when patient is not handicaped is 21912 (19.9%)

Some Statistics to allow me to see which parameter alone impact the no_show

```
In [13]: unique_area = list(df.neighbourhood.unique())
         x_area = unique_area

         df.neighbourhood.hist(figsize=(20,20))
         plt.xticks(rotation='vertical');
         plt.title('Distribtion of areas')
         plt.xlabel('81 Areas')
         plt.ylabel('Number of time each area appeard in the study')
```

```
Out[13]: Text(0,0.5,'Number of time each area appeard in the study')
```



```
In [24]: df.patient_id.value_counts()
```

```
Out[24]: 822145925426128     88
         99637671331         84
         26886125921145      70
         33534783483176      65
         871374938638855     62
         6264198675331       62
         258424392677        62
         75797461494159      62
```

```
66844879846766     57
872278549442       55
89239687393655     54
8435223536         51
853439686798559    50
14479974122994     46
65433599726992     46
9452745294842      42
81894521843749     42
188232341789524    40
2271579924275      38
9496196639835      38
13364929297498     37
1484143378533      35
986162815579582    34
88834999836575     34
712458866975343    33
6128878448536      30
416755661551767    30
81213966782532     29
8634164126317      24
36994987339512     23
                   ..
29739554385665      1
98683352133221      1
5394313945329       1
48689197872217      1
9675119787546       1
763619586595        1
983874124283357     1
737858311826761     1
8169988527774       1
2212945531847       1
961392519656997     1
271517596623238     1
8249496395977       1
137479426839        1
6529316371746       1
3212962263947       1
198193457888        1
1425822256863       1
216133833234618     1
37589497678822      1
31638467315         1
89279955685         1
5883897911366       1
56326578686847      1
869587212288428     1
```

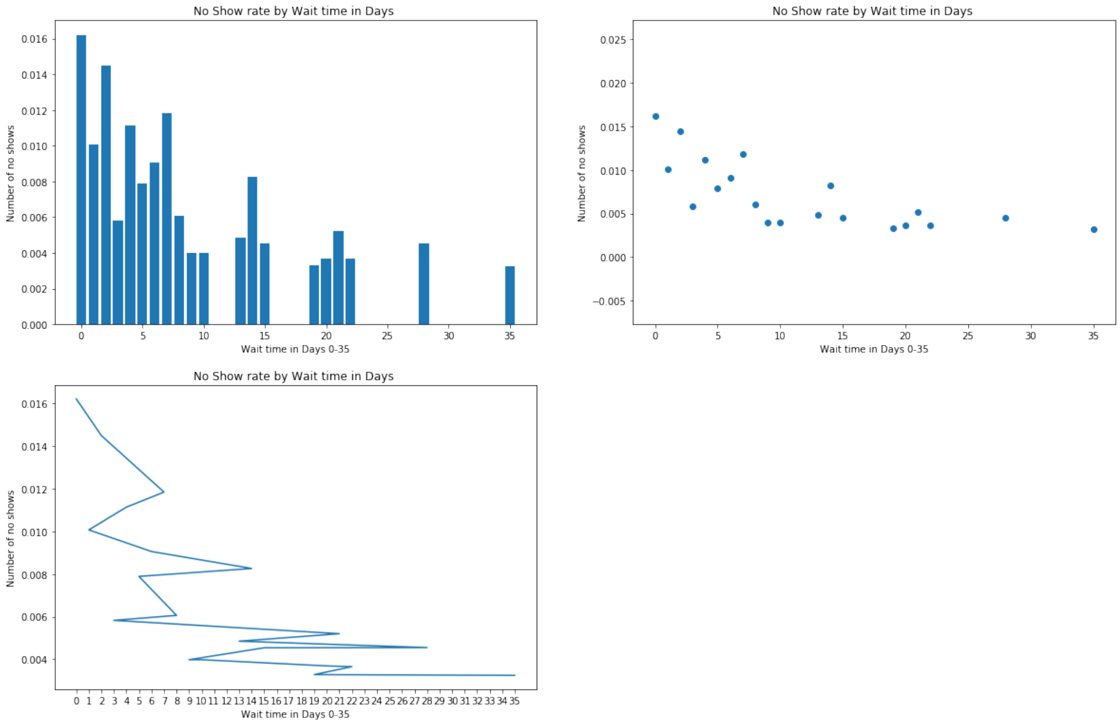
```
        735858598529        1
        2886912523138       1
        68129842443312      1
        99264711372         1
        57863365759569      1
        Name: patient_id, Length: 62298, dtype: int64
```

```
In [25]: df.wait_time.value_counts()
         df.query('no_show').wait_time.value_counts()
```

```
Out[25]: 0      38562
         2       6725
         4       5290
         1       5213
         7       4906
         6       4037
         5       3277
         14      2913
         3       2737
         8       2332
         21      1861
         28      1706
         13      1682
         9       1605
         15      1503
         10      1391
         20      1187
         22      1173
         16      1151
         12      1115
         17      1107
         29      1089
         19      1044
         18      1021
         27      1013
         11       987
         35       963
         23       822
         34       808
         26       731
                ...
         103        5
         109        5
         111        5
         98         5
         95         5
         112        5
         108        5
```

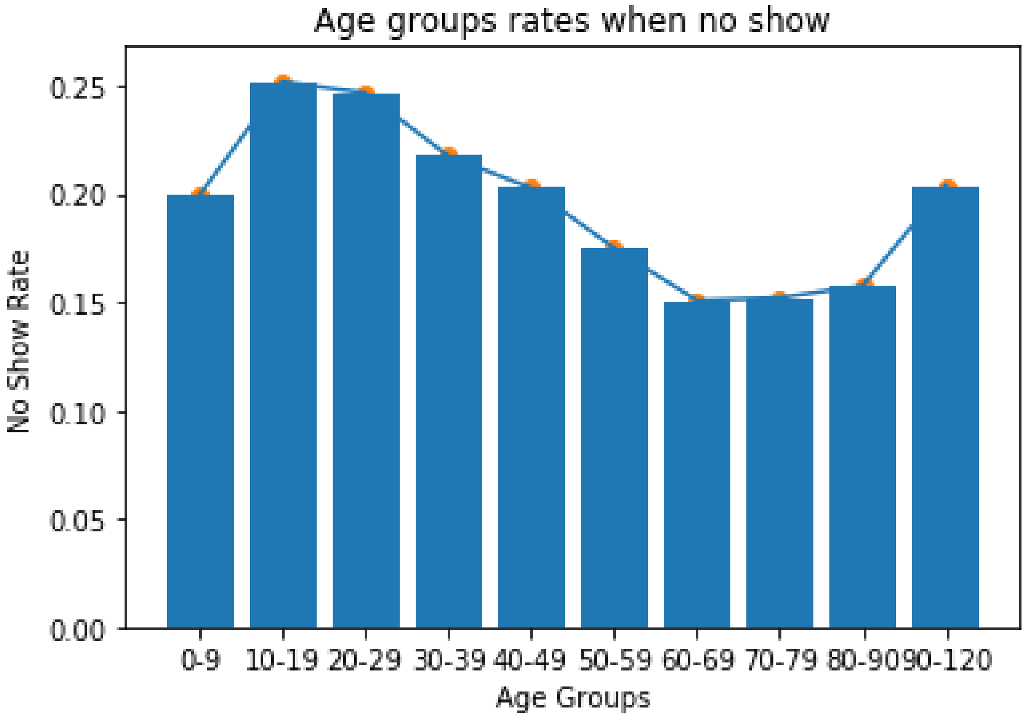
```
96          4
102         4
105         4
119         4
92          3
122         3
107         2
115         2
93          2
94          2
97          2
110         2
126         1
125         1
123         1
117         1
139         1
82          1
101         1
132         1
151         1
146         1
127         1
Name: wait_time, Length: 129, dtype: int64
```

```
In [9]: #calculate the rate of wait time to see how does it affect the no_show

        rate = df.query('no_show').wait_time.value_counts()[0:20] / df.shape[0]
        plt.figure(figsize=(20, 20))
        x=df.query('no_show').wait_time.value_counts().index[0:20]
        #y=df.query('no_show').wait_time.value_counts().values[0:20]
        y=rate
        plt.subplot(321)
        plt.bar(x,y);
        plt.title('No Show rate by Wait time in Days ')
        plt.xlabel('Wait time in Days 0-35')
        plt.ylabel('Rate of no shows')
        plt.subplot(322)
        plt.scatter(x,y);
        plt.title('No Show rate by Wait time in Days ')
        plt.xlabel('Wait time in Days 0-35')
        plt.ylabel('Rate of no shows')
        plt.subplot(323)
        plt.plot(x,y);
        plt.xticks(range(0,36));
        plt.title('No Show rate by Wait time in Days ')
        plt.xlabel('Wait time in Days 0-35')
        plt.ylabel('Rate of no shows')
plt.show()
```



The Above plot show the wait days rate with no show and 0 wait days are the most affecting

```
In [4]: val = []
        for i in range(0,10):
            val.append((df[df.age_group == i].query('no_show').count()/df[df.age_group == i].cou

        plt.bar([0,1,2,3,4,5,6,7,8,9], val)
        plt.scatter([0,1,2,3,4,5,6,7,8,9], val)
        plt.plot([0,1,2,3,4,5,6,7,8,9], val)
        # plt.xticks([0,1,2,3,4,5,6,7,8,9], rotation='vertical')
        labels = ['0-9','10-19','20-29','30-39','40-49','50-59','60-69','70-79','80-90','90-120'
        plt.xticks(range(0,10),labels);
        plt.title('Age groups rates when no show')
        plt.xlabel('Age Groups')
        plt.ylabel('No Show Rate')

Out[4]: Text(0,0.5,'No Show Rate')
```



The above bar,scatter and line plot shows Age Groups showing which age group has the most no_show

## Exploratory Data Analysis

> **Tip**: Now that you've trimmed and cleaned your data, you're ready to move on to exploration. Compute statistics and create visualizations with the goal of addressing the research questions that you posed in the Introduction section. It is recommended that you be systematic with your approach. Look at one variable at a time, and then follow it up by looking at relationships between variables.

1.3 Question that my anlysis answered:

1.3.1 Does the time delta (Wait Time) between the Scd_day and app_day is a reason for no show ?

1.3.2 what is the impact of the patient age on the show and no show?

## Conclusions

> **Tip**: Finally, summarize your findings and the results that have been performed. Make sure that you are clear with regards to the limitations of your exploration. If you haven't done any statistical tests, do not imply any statistical conclusions. And make sure you avoid implying causation from correlation!

> **Tip**: Once you are satisfied with your work here, check over your report to make sure that it is satisfies all the areas of the rubric (found on the project submission page at

the end of the lesson). You should also probably remove all of the "Tips" like this one so that the presentation is as polished as possible.

As a Conculsion there isn't one paramter that causes a no show, but multiple, for example wait time does have an impact and the highest impact when the wait time is 0

Another factor that I looked into which is the age of the patient, the higest numbers of no show is for ages between 10-29

1.4 Submitting your Project

Before you submit your project, you need to create a .html or .pdf version of this notebook in the workspace here. To do that, run the code cell below. If it worked correctly, you should get a return code of 0, and you should see the generated .html file in the workspace directory (click on the orange Jupyter icon in the upper left).

Alternatively, you can download this report as .html via the **File** > **Download as** submenu, and then manually upload it into the workspace directory by clicking on the orange Jupyter icon in the upper left, then using the Upload button.

Once you've done this, you can submit your project by clicking on the "Submit Project" button in the lower right here. This will create and submit a zip file with this .ipynb doc and the .html or .pdf version you created. Congratulations!

```
In [11]: from subprocess import call
         call(['python', '-m', 'nbconvert', 'Investigate_a_Dataset.ipynb'])
```

```
Out[11]: 0
```