Investigate_a_Dataset

May 28, 2022

> **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. Once you complete this project, remove these **Tip** sections from your report before submission. 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: patient Data Analysis

1.1 Table of Contents

## Introduction

1.1.1 Dataset Description

> **Tip**: In this section of the report, provide a brief introduction to the dataset you've selected/downloaded for analysis. Read through the description available on the homepage-links present here. List all column names in each table, and their significance. In case of multiple tables, describe the relationship between tables.

1.1.2 Question(s) for Analysis

> **Tip**: Clearly state one or more questions that you plan on exploring over the course of the report. You will address these questions in the **data analysis** and **conclusion** sections. Try to build your report around the analysis of at least one dependent variable and three independent variables. If you're not sure what questions to ask, then make sure you familiarize yourself with the dataset, its variables and the dataset context for ideas of what to explore.

> **Tip**: Once you start coding, use NumPy arrays, Pandas Series, and DataFrames where appropriate rather than Python lists and dictionaries. Also, **use good coding practices**, such as, define and use functions to avoid repetitive code. Use appropriate comments within the code cells, explanation in the mark-down cells, and meaningful variable names.

```
In [1]: # Use this cell to set up import statements for all of the packages that you
        #   plan to use.
        import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
        % matplotlib inline
        # 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
```

```
In [6]: # Upgrade pandas to use dataframe.explode() function.
        !pip install --upgrade pandas==0.25.0
```

```
Requirement already up-to-date: pandas==0.25.0 in /opt/conda/lib/python3.6/site-packages (0.25.0
Requirement already satisfied, skipping upgrade: numpy>=1.13.3 in /opt/conda/lib/python3.6/site-
Requirement already satisfied, skipping upgrade: python-dateutil>=2.6.1 in /opt/conda/lib/python
Requirement already satisfied, skipping upgrade: pytz>=2017.2 in /opt/conda/lib/python3.6/site-p
Requirement already satisfied, skipping upgrade: six>=1.5 in /opt/conda/lib/python3.6/site-packa
```

1.1.3 General

```
In [2]: # 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 = pd.read_csv("noshowappointments-kagglev2-may-2016.csv")
        df.head()
```

```
Out[2]:       PatientId  AppointmentID Gender          ScheduledDay  \
        0  2.987250e+13        5642903      F  2016-04-29T18:38:08Z
        1  5.589978e+14        5642503      M  2016-04-29T16:08:27Z
        2  4.262962e+12        5642549      F  2016-04-29T16:19:04Z
        3  8.679512e+11        5642828      F  2016-04-29T17:29:31Z
        4  8.841186e+12        5642494      F  2016-04-29T16:07:23Z

                 AppointmentDay  Age      Neighbourhood  Scholarship  Hipertension  \
        0  2016-04-29T00:00:00Z   62    JARDIM DA PENHA            0             1
        1  2016-04-29T00:00:00Z   56    JARDIM DA PENHA            0             0
        2  2016-04-29T00:00:00Z   62      MATA DA PRAIA            0             0
        3  2016-04-29T00:00:00Z    8  PONTAL DE CAMBURI            0             0
        4  2016-04-29T00:00:00Z   56    JARDIM DA PENHA            0             1

           Diabetes  Alcoholism  Handcap  SMS_received No-show
        0         0           0        0             0      No
        1         0           0        0             0      No
        2         0           0        0             0      No
        3         0           0        0             0      No
        4         1           0        0             0      No
```

```
In [25]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 110527 entries, 0 to 110526
Data columns (total 14 columns):
PatientId         110527 non-null float64
AppointmentID     110527 non-null int64
Gender            110527 non-null object
ScheduledDay      110527 non-null object
AppointmentDay    110527 non-null object
Age               110527 non-null int64
Neighbourhood     110527 non-null object
Scholarship       110527 non-null int64
Hipertension      110527 non-null int64
Diabetes          110527 non-null int64
Alcoholism        110527 non-null int64
Handcap           110527 non-null int64
SMS_received      110527 non-null int64
No-show           110527 non-null object
dtypes: float64(1), int64(8), object(5)
memory usage: 11.8+ MB
```

```
In [12]: df.duplicated()
```

```
Out[12]: 0         False
         1         False
         2         False
         3         False
         4         False
                   ...
         110522    False
         110523    False
         110524    False
         110525    False
         110526    False
         Length: 110527, dtype: bool
```

```
In [6]: df.drop(['AppointmentDay','ScheduledDay'],axis=1 , inplace=True)
```

```
In [7]: df.head()
```

```
Out[7]:       PatientId  AppointmentID Gender  Age      Neighbourhood  Scholarship  \
        0  2.987250e+13        5642903      F   62    JARDIM DA PENHA            0
        1  5.589978e+14        5642503      M   56    JARDIM DA PENHA            0
        2  4.262962e+12        5642549      F   62      MATA DA PRAIA            0
        3  8.679512e+11        5642828      F    8  PONTAL DE CAMBURI            0
        4  8.841186e+12        5642494      F   56    JARDIM DA PENHA            0
```

```
   Hipertension  Diabetes  Alcoholism  Handcap  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
3             0         0           0        0             0      No
4             1         1           0        0             0      No
```

```
In [8]: df.drop(['AppointmentID','PatientId'],axis=1 , inplace=True)
```

```
In [9]: df.head()
```

```
Out[9]:   Gender  Age      Neighbourhood  Scholarship  Hipertension  Diabetes  \
        0      F   62    JARDIM DA PENHA            0             1         0
        1      M   56    JARDIM DA PENHA            0             0         0
        2      F   62      MATA DA PRAIA            0             0         0
        3      F    8  PONTAL DE CAMBURI            0             0         0
        4      F   56    JARDIM DA PENHA            0             1         1

           Alcoholism  Handcap  SMS_received No-show
        0           0        0             0      No
        1           0        0             0      No
        2           0        0             0      No
        3           0        0             0      No
        4           0        0             0      No
```

```
In [10]: df.describe()
```

```
Out[10]:                  Age    Scholarship   Hipertension       Diabetes  \
         count  110527.000000  110527.000000  110527.000000  110527.000000
         mean       37.088874       0.098266       0.197246       0.071865
         std        23.110205       0.297675       0.397921       0.258265
         min        -1.000000       0.000000       0.000000       0.000000
         25%        18.000000       0.000000       0.000000       0.000000
         50%        37.000000       0.000000       0.000000       0.000000
         75%        55.000000       0.000000       0.000000       0.000000
         max       115.000000       1.000000       1.000000       1.000000

                   Alcoholism        Handcap   SMS_received
         count  110527.000000  110527.000000  110527.000000
         mean        0.030400       0.022248       0.321026
         std         0.171686       0.161543       0.466873
         min         0.000000       0.000000       0.000000
         25%         0.000000       0.000000       0.000000
         50%         0.000000       0.000000       0.000000
         75%         0.000000       0.000000       1.000000
         max         1.000000       4.000000       1.000000
```

```
In [9]: age=df.query('Age=="-1"')
        age
```

```
Out[9]:           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
```

1.1.4 Data Cleaning

> **Tip**: Make sure that you keep your reader informed on the steps that you are taking in your investigation. Follow every code cell, or every set of related code cells, with a markdown cell to describe to the reader what was found in the preceding cell(s). Try to make it so that the reader can then understand what they will be seeing in the following cell(s).

```
In [28]: df.drop(index=99832, inplace=True)
```

```
In [13]: df.head()
```

```
Out[13]:       PatientId  AppointmentID Gender          ScheduledDay  \
         0  2.987250e+13        5642903      F  2016-04-29T18:38:08Z
         1  5.589978e+14        5642503      M  2016-04-29T16:08:27Z
         2  4.262962e+12        5642549      F  2016-04-29T16:19:04Z
         3  8.679512e+11        5642828      F  2016-04-29T17:29:31Z
         4  8.841186e+12        5642494      F  2016-04-29T16:07:23Z

                  AppointmentDay  Age      Neighbourhood  Scholarship  Hipertension  \
         0  2016-04-29T00:00:00Z   62    JARDIM DA PENHA            0             1
         1  2016-04-29T00:00:00Z   56    JARDIM DA PENHA            0             0
         2  2016-04-29T00:00:00Z   62      MATA DA PRAIA            0             0
         3  2016-04-29T00:00:00Z    8  PONTAL DE CAMBURI            0             0
         4  2016-04-29T00:00:00Z   56    JARDIM DA PENHA            0             1

            Diabetes  Alcoholism  Handcap  SMS_received No-show
         0         0           0        0             0      No
         1         0           0        0             0      No
         2         0           0        0             0      No
         3         0           0        0             0      No
         4         1           0        0             0      No
```

```
In [11]: df['Noshow'].value_counts(normalize=True)
```

```
Out[11]: No     0.798067
         Yes    0.201933
         Name: Noshow, dtype: float64
```

```
In [4]: df.rename(columns={'No-show':'Noshow'},inplace=True)
In [8]: df.head()
```

```
Out[8]:       PatientId  AppointmentID Gender          ScheduledDay  \
        0  2.987250e+13        5642903      F  2016-04-29T18:38:08Z
        1  5.589978e+14        5642503      M  2016-04-29T16:08:27Z
        2  4.262962e+12        5642549      F  2016-04-29T16:19:04Z
        3  8.679512e+11        5642828      F  2016-04-29T17:29:31Z
        4  8.841186e+12        5642494      F  2016-04-29T16:07:23Z

                 AppointmentDay  Age      Neighbourhood  Scholarship  Hipertension  \
        0  2016-04-29T00:00:00Z   62    JARDIM DA PENHA            0             1
        1  2016-04-29T00:00:00Z   56    JARDIM DA PENHA            0             0
        2  2016-04-29T00:00:00Z   62      MATA DA PRAIA            0             0
        3  2016-04-29T00:00:00Z    8  PONTAL DE CAMBURI            0             0
        4  2016-04-29T00:00:00Z   56    JARDIM DA PENHA            0             1

           Diabetes  Alcoholism  Handcap  SMS_received Noshow
        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
```

```
In [15]: dfshow=df[df['Noshow']=='No']
```

```
In [17]: dfshow.head()
```

```
Out[17]:       PatientId  AppointmentID Gender          ScheduledDay  \
         0  2.987250e+13        5642903      F  2016-04-29T18:38:08Z
         1  5.589978e+14        5642503      M  2016-04-29T16:08:27Z
         2  4.262962e+12        5642549      F  2016-04-29T16:19:04Z
         3  8.679512e+11        5642828      F  2016-04-29T17:29:31Z
         4  8.841186e+12        5642494      F  2016-04-29T16:07:23Z

                  AppointmentDay  Age      Neighbourhood  Scholarship  Hipertension  \
         0  2016-04-29T00:00:00Z   62    JARDIM DA PENHA            0             1
         1  2016-04-29T00:00:00Z   56    JARDIM DA PENHA            0             0
         2  2016-04-29T00:00:00Z   62      MATA DA PRAIA            0             0
         3  2016-04-29T00:00:00Z    8  PONTAL DE CAMBURI            0             0
         4  2016-04-29T00:00:00Z   56    JARDIM DA PENHA            0             1

            Diabetes  Alcoholism  Handcap  SMS_received Noshow
         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
```

```
In [16]: dfNoshow=df[df['Noshow']!='No']
In [18]: dfNoshow.head()

Out[18]:        PatientId  AppointmentID Gender          ScheduledDay  \
        6   7.336882e+14        5630279      F  2016-04-27T15:05:12Z
        7   3.449833e+12        5630575      F  2016-04-27T15:39:58Z
        11  7.542951e+12        5620163      M  2016-04-26T08:44:12Z
        17  1.479497e+13        5633460      F  2016-04-28T09:28:57Z
        20  6.222575e+14        5626083      F  2016-04-27T07:51:14Z

                  AppointmentDay  Age   Neighbourhood  Scholarship  Hipertension  \
        6   2016-04-29T00:00:00Z   23      GOIABEIRAS            0             0
        7   2016-04-29T00:00:00Z   39      GOIABEIRAS            0             0
        11  2016-04-29T00:00:00Z   29  NOVA PALESTINA            0             0
        17  2016-04-29T00:00:00Z   40       CONQUISTA            1             0
        20  2016-04-29T00:00:00Z   30  NOVA PALESTINA            0             0

            Diabetes  Alcoholism  Handcap  SMS_received Noshow
        6          0           0        0             0    Yes
        7          0           0        0             0    Yes
        11         0           0        0             1    Yes
        17         0           0        0             0    Yes
        20         0           0        0             0    Yes
```

## Exploratory Data Analysis

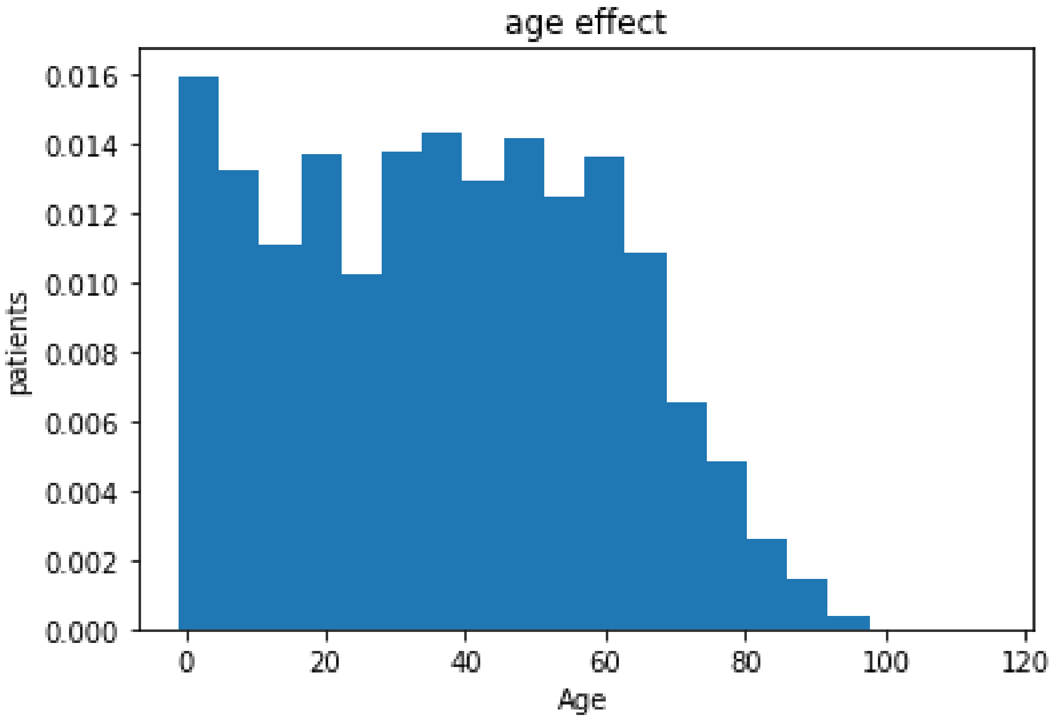
> **Tip**: Now that you've trimmed and cleaned your data, you're ready to move on to exploration. **Compute statistics** and **create visualizations** with the goal of addressing the research questions that you posed in the Introduction section. You should compute the relevant statistics throughout the analysis when an inference is made about the data. Note that at least two or more kinds of plots should be created as part of the exploration, and you must compare and show trends in the varied visualizations.

> **Tip**: - Investigate the stated question(s) from multiple angles. 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. You should explore at least three variables in relation to the primary question. This can be an exploratory relationship between three variables of interest, or looking at how two independent variables relate to a single dependent variable of interest. Lastly, you should perform both single-variable (1d) and multiple-variable (2d) explorations.

1.1.5 Question 1

```
In [9]: # Use this, and more code cells, to explore your data. Don't forget to add
        #   Markdown cells to document your observations and findings.
        plt.hist(df['Age'],bins=20,normed=True)

        plt.xlabel('Age')
        plt.ylabel('patients')
        plt.title('age effect')
        plt.legend()
```

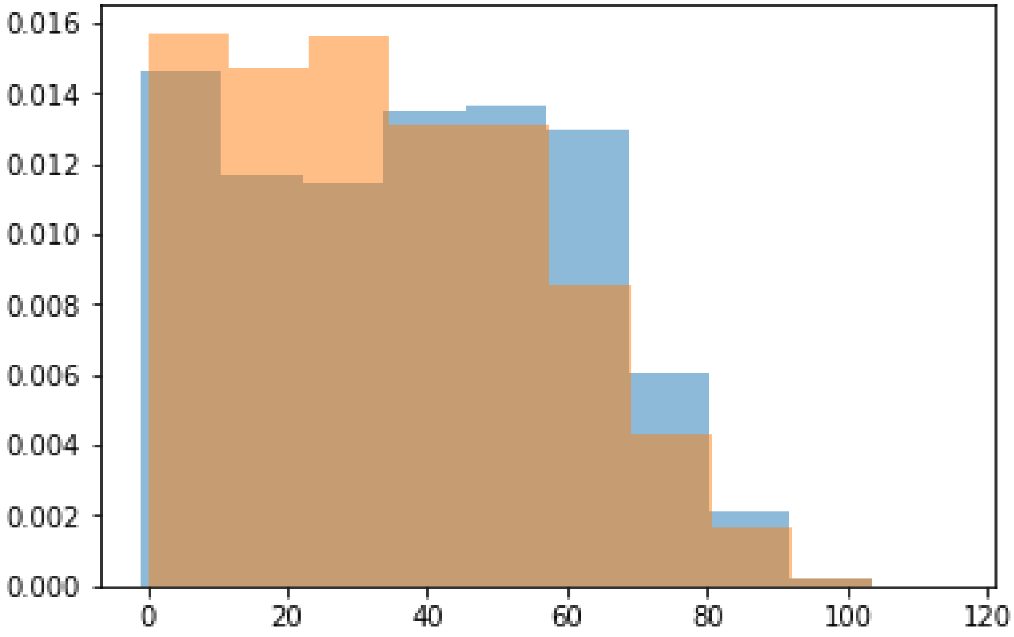


```
In [31]: df['Age'].describe()

Out[31]: count    110526.000000
         mean         37.089219
         std          23.110026
         min           0.000000
         25%          18.000000
         50%          37.000000
         75%          55.000000
         max         115.000000
         Name: Age, dtype: float64
```

From the plan, we see that the age group that goes to the medical examination is children, and it gradually decreases to the age of 20, then increases again to the age of 35, then declines

```
In [22]: plt.hist(dfshow['Age'],normed=True,alpha=0.5)
         plt.hist(dfNoshow['Age'],normed=True,alpha=0.5);
```

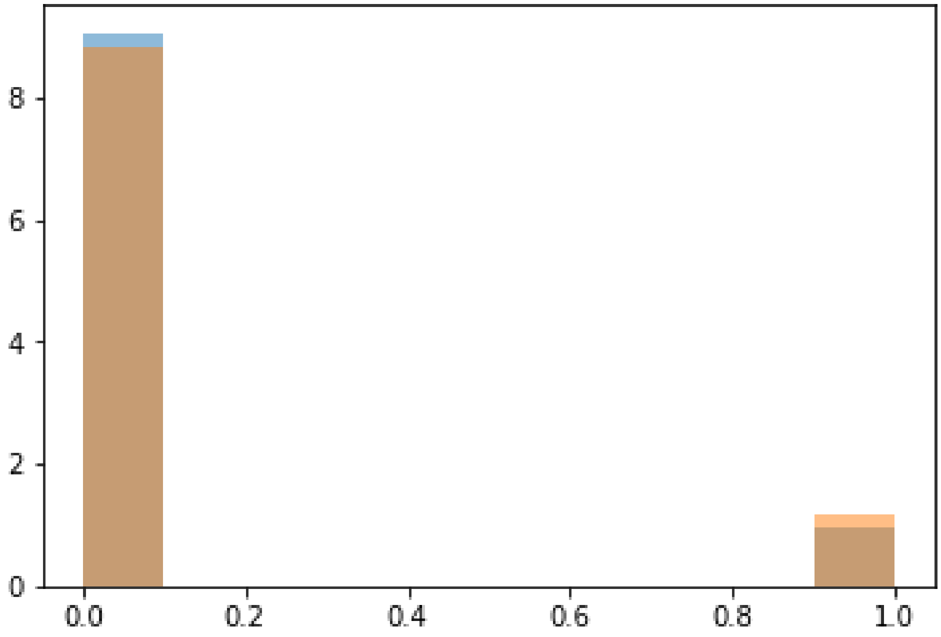


Question 2

```
In [12]: # Continue to explore the data to address your additional research
         #   questions. Add more headers as needed if you have more questions to
         #   investigate.
         df['Scholarship'].value_counts()

Out[12]: 0    99666
         1    10861
         Name: Scholarship, dtype: int64

In [23]: plt.hist(dfshow['Scholarship'],normed=True,alpha=0.5)
         plt.hist(dfNoshow['Scholarship'],normed=True,alpha=0.5);
```



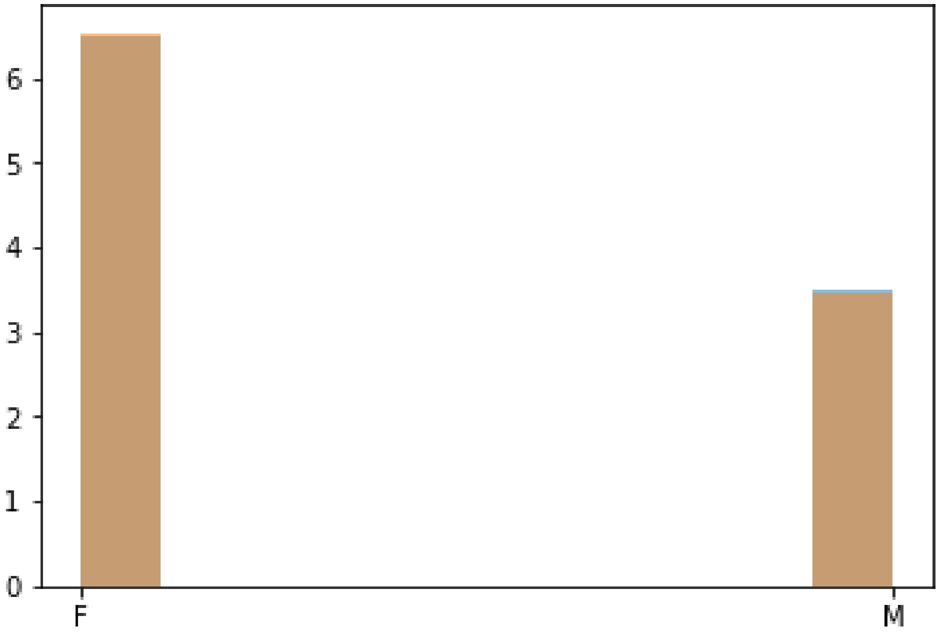
The attendance rate of people who attended through scholarships is almost equal to the people who do not have scholarships

Question 3

```
In [3]: df['Gender'].value_counts()

Out[3]: F    71840
        M    38687
        Name: Gender, dtype: int64

In [24]: Gender
```

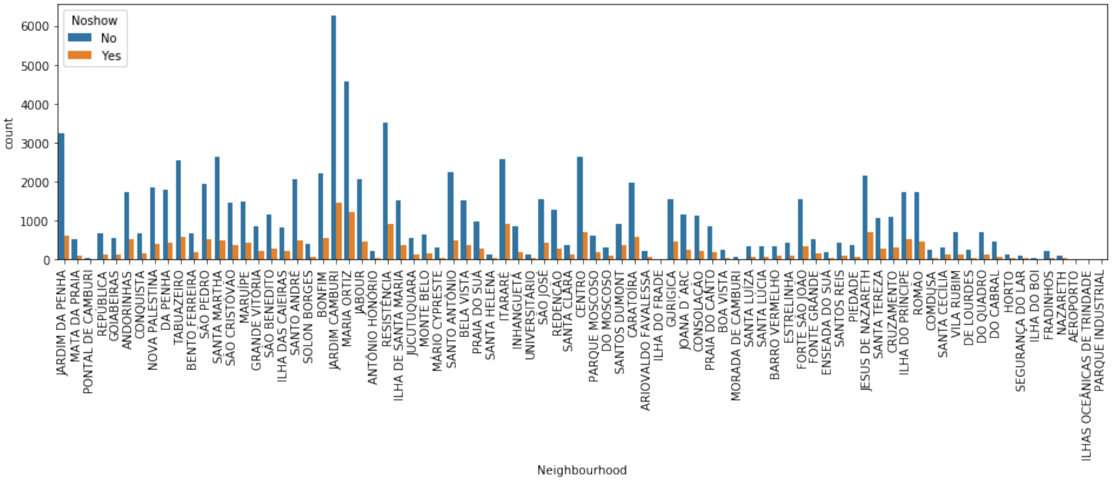


Here we see that the percentage of women who go to the medical examination is equal to the number of women who do not go to the medical examination and the number of men who go to the medical examination is almost equal to the number of men who do not go to the medical examination

Question 4

```
In [13]: plt.figure(figsize=(16,4))
         plt.xticks(rotation=90)
         ax=sns.countplot(x=df.Neighbourhood,hue=df.Noshow)

         plt.show()
```



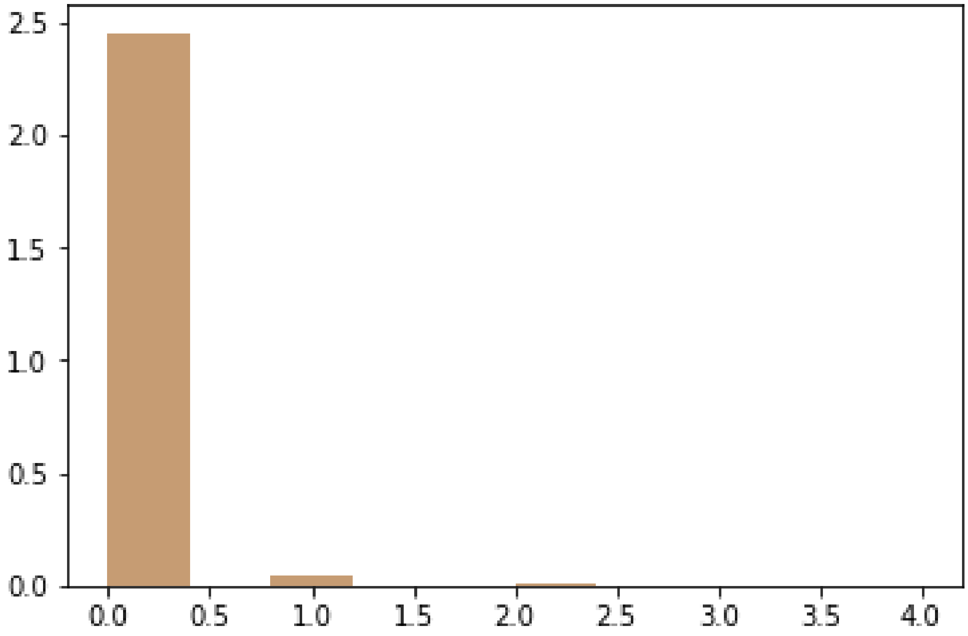
Here we notice the presence of a larger number of residents of certain neighborhoods than other neighborhoods and this is because the population ratio differs in each neighborhood from the other

Question 5

```
In [11]: df['Handcap'].value_counts()
```

```
Out[11]: 0    108286
         1      2042
         2       183
         3        13
         4         3
         Name: Handcap, dtype: int64
```

```
In [25]: plt.hist(dfshow['Handcap'],normed=True,alpha=0.5)
         plt.hist(dfNoshow['Handcap'],normed=True,alpha=0.5);
```



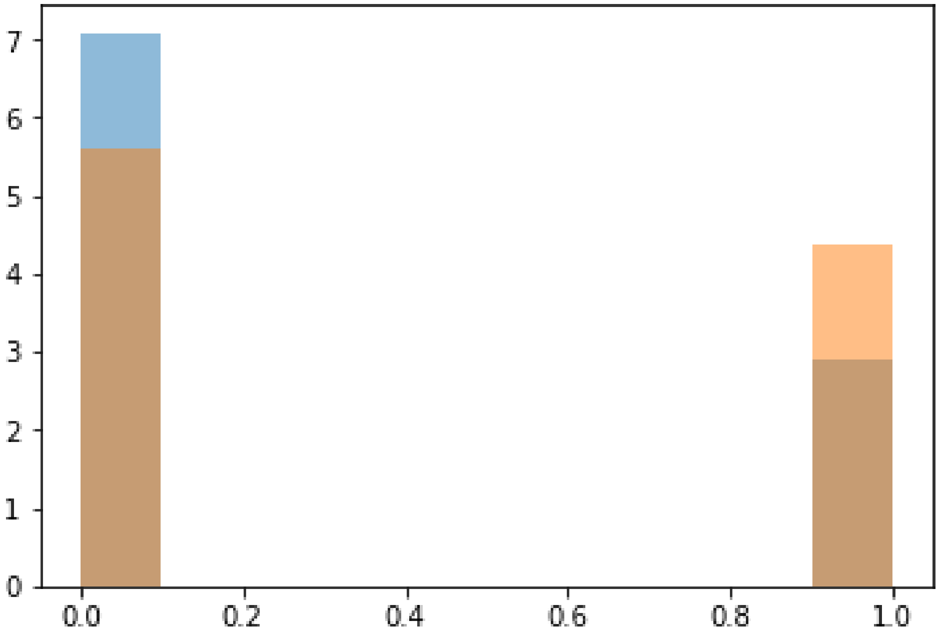
The percentage of the number of people who have a handicap and go for a medical examination is equal to the number of people who have a handicap and do not go

Question 6

```
In [13]: df['SMS_received'].value_counts()
```

```
Out[13]: 0    75045
         1    35482
         Name: SMS_received, dtype: int64
```

```
In [26]: plt.hist(dfshow['SMS_received'],normed=True,alpha=0.5)
         plt.hist(dfNoshow['SMS_received'],normed=True,alpha=0.5);
```



```
In [ ]: People who did not send a text message had a greater attendance rate than the number of
        And here we see that there is a problem in sending messages where the wrong data is sent
```

Conclusions From what I reviewed before you, we see that the most important values in which we are good is age, and we have proven age that mainly affects the percentage of attendance, absence and text messages. We will show you the messages and we have seen that it is important that the messages are sent correctly even Each patient receives the correct return in the medical examination

Limitations We see that neighborhood values differ from one neighborhood to another, and this is a numerical difference resulting from the different population density in each neighborhood, so it would have been better to have data on the population density in each neighborhood

```
In [30]: from subprocess import call
         call(['python', '-m', 'nbconvert', 'Investigate_a_Dataset.ipynb'])

Out[30]: 0

In [ ]:
```