Project: Investigate a Dataset (No-show Appointments!)

Table of Contents

Introduction

Purpose To perform a Data analysis on a sample Dataset of No-show Appointments

This Dataset contains the records of the patients with various types of diseases who booked appointments and did not showed up on their appointment Day.

```
# Use this cell to set up import statements for all of the packages that you
#   plan to use.
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
sns.set(style="whitegrid")
# 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
```

Data Wrangling

```
# 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()
```

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension	Diabetes	Alcoholism	Handcap	SM
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	
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	
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	
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	
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	

Data Cleaning ()

From the data description and questions to answer, I've determined that some of the dataset columns are not necessary for the analysis process and will therefore be removed. This will help to process the Data Analysis Faster i'll take a 3 step approach to data cleanup

1-Identify and remove duplicate entries

2-Remove unnecessary columns

3-Fix missing and data format issues

```
df.info()
```

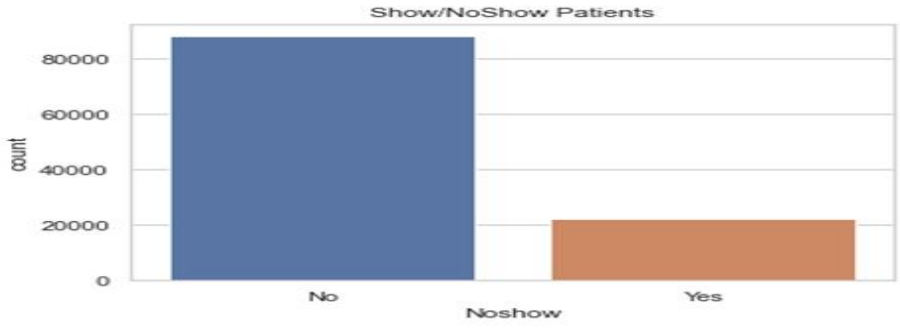
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 110527 entries, 0 to 110526
Data columns (total 14 columns):
 #   Column          Non-Null Count   Dtype  
---  ------          --------------   -----  
 0   PatientId       110527 non-null  float64
 1   AppointmentID   110527 non-null  int64  
 2   Gender          110527 non-null  object 
 3   ScheduledDay    110527 non-null  object 
 4   AppointmentDay  110527 non-null  object 
 5   Age             110527 non-null  int64  
 6   Neighbourhood   110527 non-null  object 
 7   Scholarship     110527 non-null  int64  
 8   Hipertension    110527 non-null  int64  
 9   Diabetes        110527 non-null  int64  
 10  Alcoholism      110527 non-null  int64  
 11  Handcap         110527 non-null  int64  
 12  SMS_received    110527 non-null  int64  
 13  No-show         110527 non-null  object 
```

Exploratory Data Analysis

Research Question 1 (What factors are important for us to know in order to predict if a patient will show up for their scheduled appointment?)

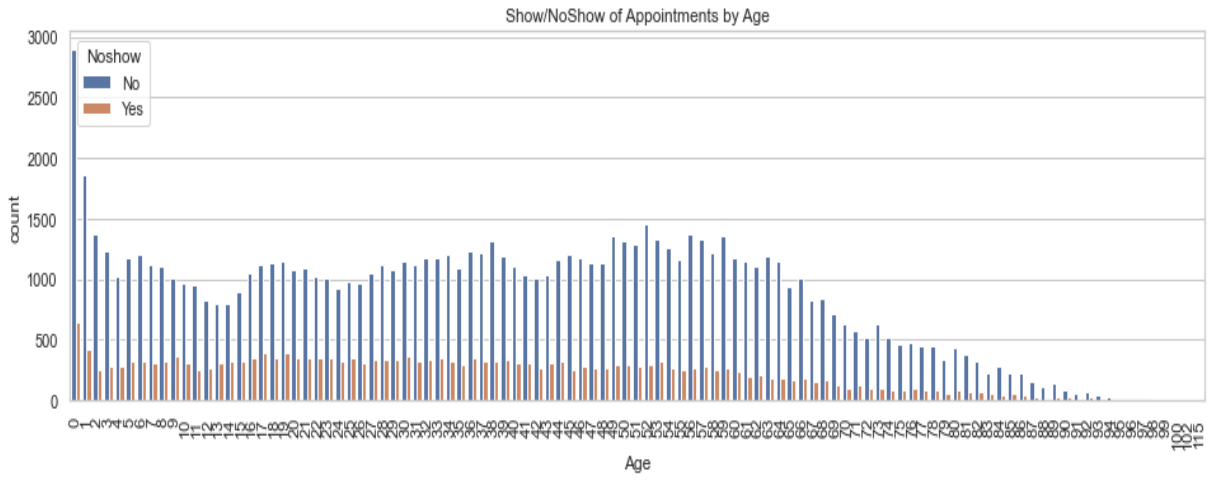
we can see that out of 110519 patients around 88,000 of them have turned up and that's around 80%

```
ax = sns.countplot(x=df.Noshow, data=df)
ax.set_title("Show/NoShow Patients")
plt.show()
```



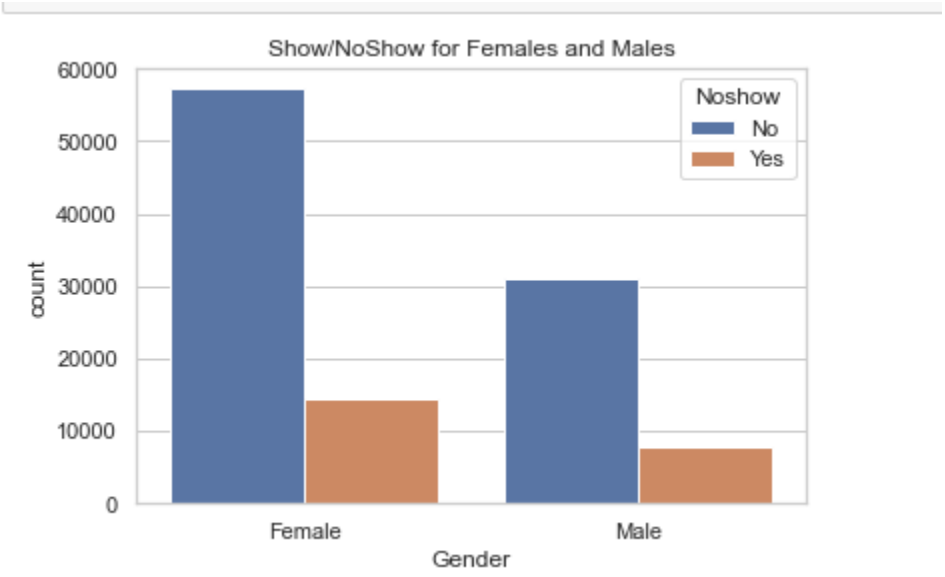
Research Question 2 (Does age affect)

the ratio of Show to NoShow is nearly the same for all ages except for "Age 0" and "Age 1" We will get better clarity on the ratio of Show to NoShow for all ages. so age does not affect the commitment to visit



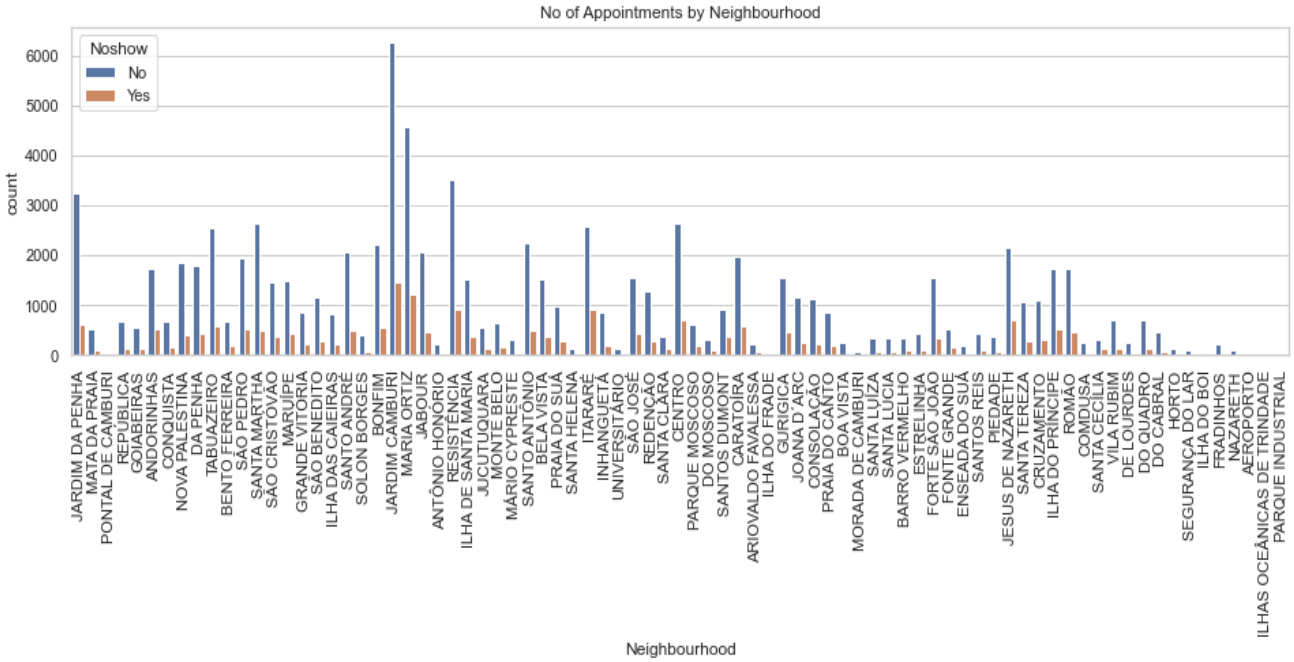
Question 3 Can gender affect visit ?

We can see that of the 88,000 patients that appeared, about 57,000 were female and 31,000 were male. Of the 22,500 patients who did not come for a visit, about 15,000 were females and 7,500 were males The ratio of females to males who attended appears to be the same as that which did not come to visit, and therefore gender affect



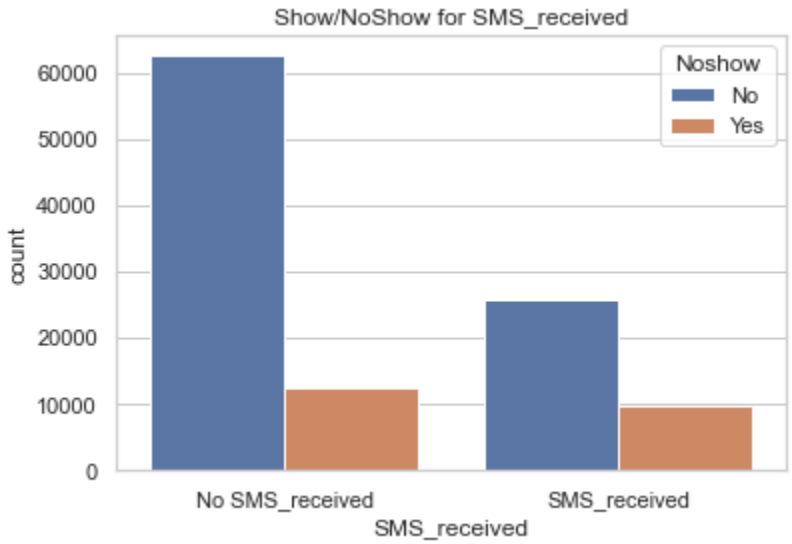
Question 4 what about the neighborhood ?

we can see that the number of patients for few Neighbourhood's is very high.



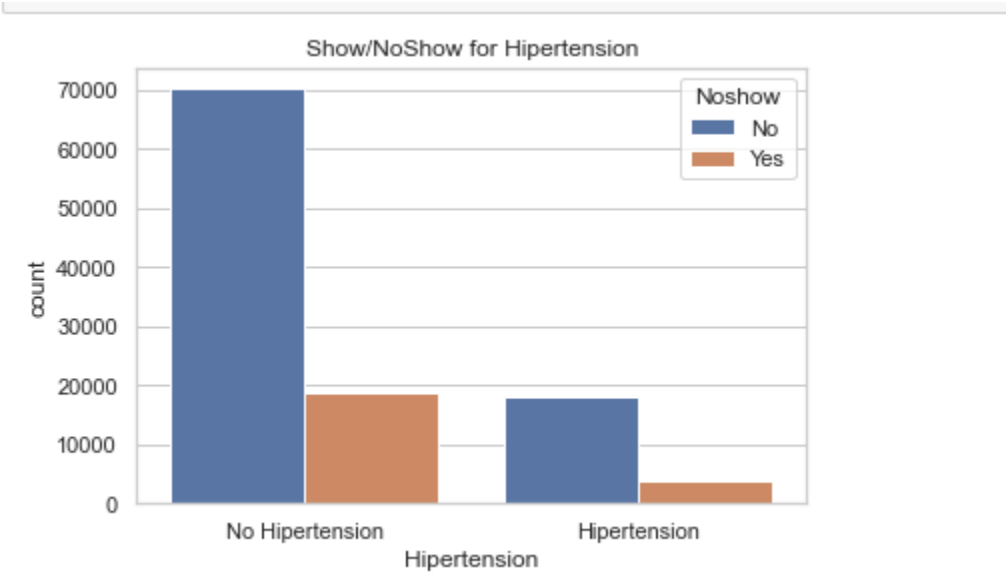
Question 5 what about the SMSReceived ?

we can see that there are about 75,000 patients who did not receive text messages, and about 84% of them attended the visit. Of the 35,500 patients who received text messages, about 72% attended the visit



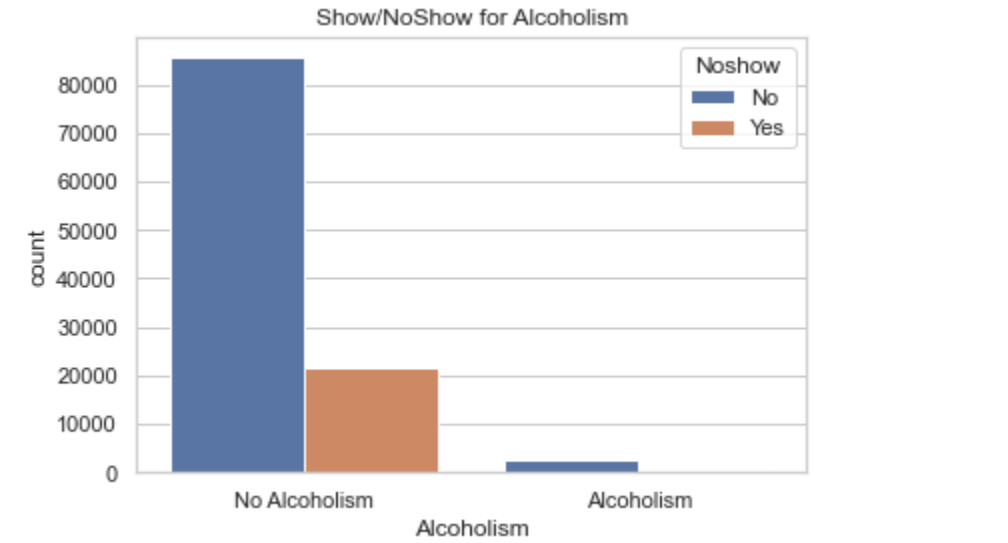
Question 6 what about the Hypertension ?

we can see that there are about 88,000 patients suffering from high blood pressure and about 78% of them attended the visit. Of 22,500 patients with high blood pressure, about 85% came to visit. Therefore, the high blood pressure feature can help us determine whether a patient will show up on a post-appointment visit



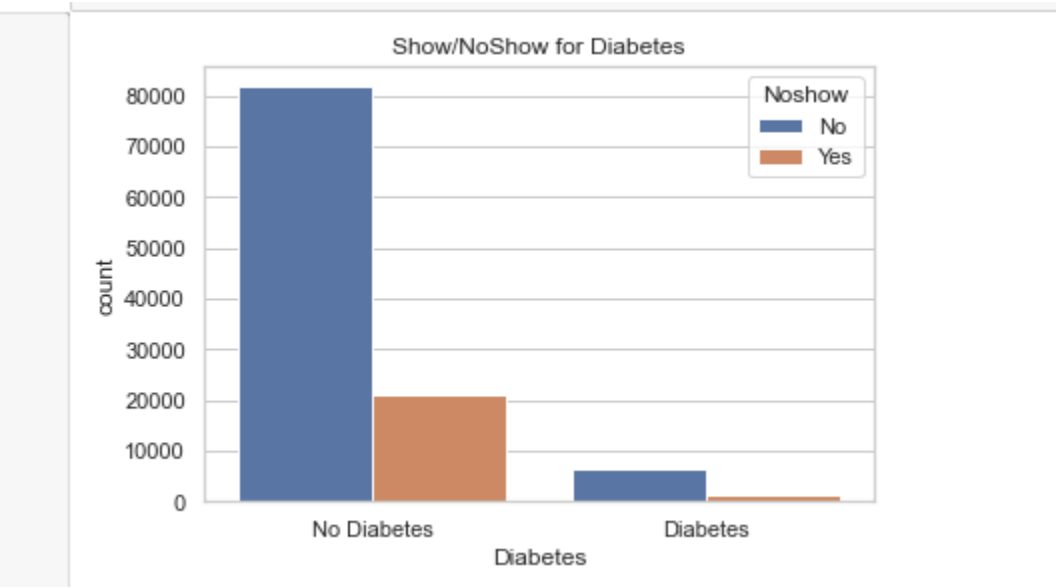
Question 7 what about the Alcoholism ?

we can see that there are about 105,000 patients who do not suffer from alcoholism and about 80% of them attended the visit. Of the 5,500 patients with alcohol addiction, about 80% attended the visit. Since the rate of visits for non-alcoholic patients is the same, this may not help us determine whether or not the patient is coming for a visit



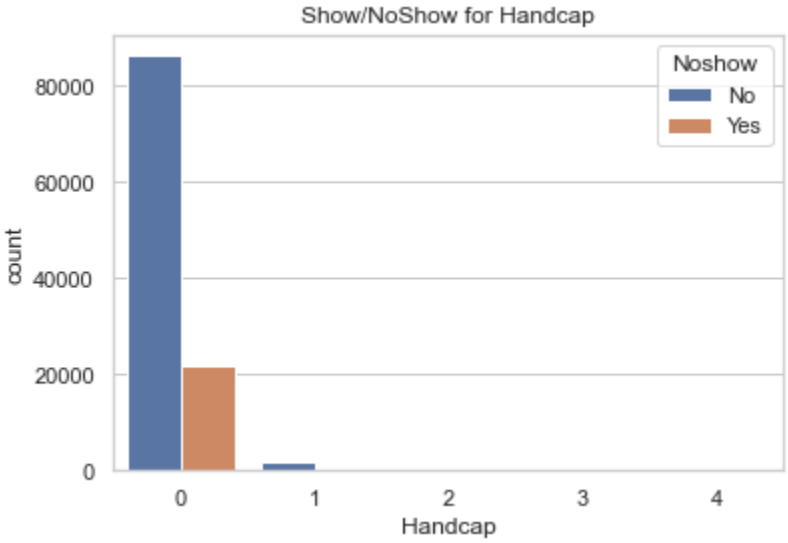
Question 8 what about the Diabetes ?

we can see that there are about 102,000 diabetics and about 80% of them attended the visit. Of the 8,500 diabetic patients, about 83% came to visit. Therefore, the diabetes feature can help us determine whether a patient will attend the post-appointment visit



Question 9 what about Handcap ?

we can see that there are about 110,000 unobstructed patients and about 80% of them have come for a visit. Since we see a clear distinction between different levels of disability, this feature will help us determine if a patient will come for a visit after making an appointment.



Conclusions

By appling everything we had learned in the class and used most of the functions explained By analyzing and tracking the results

1-The Age values included a negative value which created problem in Analysing the Dataset so the negative value was changed into positive value Age did not seem to be a major factor

2-Gender is importent factor

3-Neighborhood and high blood pressure come after gender as there are some neighborhoods where diseases are common and high blood pressure patients tend to appear if they have it or not

4- receive sms does not effect to attend

5- one Patient can have more than one appointment