Project 2 – Investigate dataset

Table of Content:

1. Introduction

The Dataset I choose for my project is “No-Show Appointment” from Kaggle. The data within this set is showing information for about 100,000 patients, reviewing some of their characteristics in separate columns like health status, appointments details, gender, age…etc. ant its relationship with showing up to their appointments.

We are going to analysis this provided data trying to find answers to our questions which are:

1- What is the characteristics affecting the status of showing up at the appointments?
2- Does receiving SMS related with showing up status?

2. Data Wrangling

I’m going to use Jupyter Notebook to analyst this Dataset.

```
#Importing the tools to be used within this project
%matplotlib inline

import matplotlib
import numpy as np
import matplotlib.pyplot as plt
```

```
#importing the study case data
import pandas as dp
nonshown_data = dp.read_csv(r'C:\Users\coffe\Documents\DND\Project2\noshow
appointments-kagglev2-may-2016.csv')
#Print out few lines to view and look for missing and useless data
nonshown_data.head()
```

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholar
0	2.987250e+13	5642903	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	
1	5.589978e+14	5642503	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	
2	4.262962e+12	5642549	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	
3	8.679512e+11	5642828	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	
4	8.841186e+12	5642494	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	

3. Cleaning Data

```
#Cleaning the Data
#choocing the unnecessary colums to delete
del_col=[ 'PatientId',  'AppointmentID', 'Scholarship', 'Hipertension', 'D
iabetes', 'Alcoholism']

#deleting the columns
nonshown_data= nonshown_data.drop(del_col,1)

#previewing the new dataset
nonshown_data.head()
```

```
#Checking Data types
nonshown_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 110527 entries, 0 to 110526
Data columns (total 8 columns):
 #   Column          Non-Null Count   Dtype 
---  ------          --------------   ----- 
 0   Gender          110527 non-null  object
 1   ScheduledDay    110527 non-null  object
 2   AppointmentDay  110527 non-null  object
 3   Age             110527 non-null  int64 
 4   Neighbourhood   110527 non-null  object
 5   Handcap         110527 non-null  int64 
 6   SMS_received    110527 non-null  int64 
 7   No-show         110527 non-null  object
dtypes: int64(3), object(5)
memory usage: 6.7+ MB
```

```
#Checking the length of data set
len(nonshown_data)
```

```
#Checking the null values in our data set
nonshown_data.isna().sum()
```

```
#Checking the null values in our data set
nonshown_data.isna().sum()
```

```
Gender          0
ScheduledDay    0
AppointmentDay  0
Age             0
Neighbourhood   0
Handcap         0
SMS_received    0
No-show         0
dtype: int64
```

```
# check if there is a dplicated  data
print("Num of dublicated : ", + sum(nonshown_data.duplicated()))
```

```
Num of dublicated :  642
```

```
#Dropping duplicated data
nonshown_data.drop_duplicates(inplace=True)
```

```
# check if there is a dplicated  data after dropping to make sure there is no more duplicated data
print("Num of dublicated : ", + sum(nonshown_data.duplicated()))
```

```
Num of dublicated :  0
```

```
#printing to check if duplication deleted successfully
print(nonshown_data.duplicated().sum())
```

```
0
```

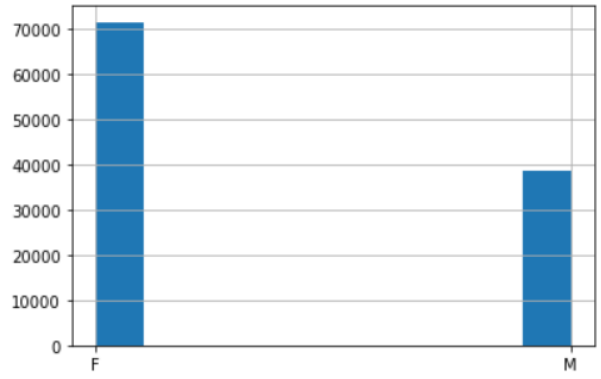
4. Data Analysis and Exploration

Exploring Charactaristics:

```
# Find the sex distribution within the dataset
gndr_dstrb = nonshown_data['Gender'].value_counts().reset_index()
gndr_dstrb.columns = ['Gender' , '']
gndr_dstrb
```

	Gender	
0	F	71840
1	M	38687

```
# plot distribution of Gender in this data set
nonshown_data['Gender'].hist();
```

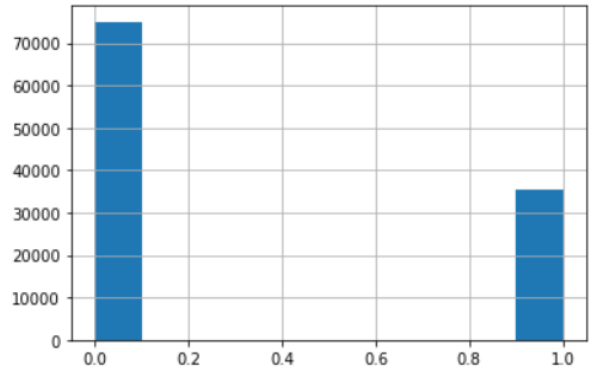



```
# Statistics of received SMS
sms_rcvd = nonshown_data['SMS_received'].value_counts().reset_index()
sms_rcvd.columns = ['SMS_received' , '']
print(nonshown_data.SMS_received.mean() * 100)
sms_rcvd
```

```
32.10256317460892
```

	SMS_received	
0	0	75045
1	1	35482

```
nonshown_data['SMS_received'].hist();
```




```
#Average Age for the patients within the dataset
nonshown_data['Age'].mean()
```

```
37.08887421173107
```

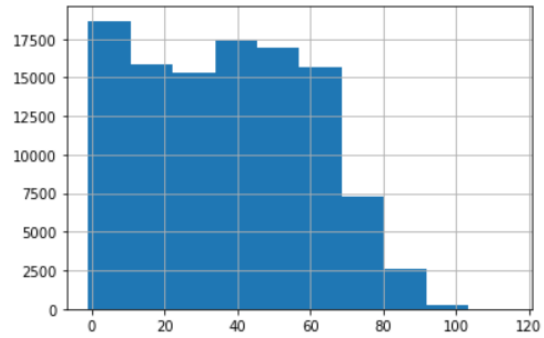
```
#MAX age within the data set
print(nonshown_data['Age'].max())
```

```
115
```

```
#MIN age within the data set
print(nonshown_data['Age'].min())
```

```
-1
```

```
# plot distribution of Age in this data set
nonshown_data['Age'].hist();
```




```
# view dimensions of dataset
nonshown_data.shape
```

```
(109885, 8)
```

My observations so far can be summarized as follows:

- The number of women in the dataset is almost twice the number of men
- The number of patients who did not received SMS is almost twice the number of men
- Age average is about 37 Years
- The highest registered age group is 115 and the lowest is under one year

Analyzing the data:

```
#Counting No-show
nonshown_data['No-show'].value_counts()
```

```
No     88208
Yes    22319
Name: No-show, dtype: int64
```

```
#Splitting the patients acoording to attendance status into two groups to count percentage of absence
showed = nonshown_data['No-show'] == 'No'
not_showed = nonshown_data['No-show'] == 'Yes'
#Adding the new coulnms to our dataset
nonshown_data['showed'] = showed
nonshown_data['not_showed'] = not_showed
nonshown_data.head()
```

	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Handcap	SMS_received	No-show	showed	not_showed
0	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	0	No	True	False
1	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0	No	True	False
2	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0	No	True	False
3	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0	No	True	False
4	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0	No	True	False

```
#defining Percentage Function

#The percentage of patients who did  attended their appointments
#The percentage of patients who did not attended their appointments
def percentage(part,whole):
    return 100 * float(part) / float(whole)

showed =percentage (nonshown_data[nonshown_data['No-show']=='No'].shape[0], nonshown_data['No-show'].shape[0] )
showed
```

```
79.80534513354872
```

```
not_showed =percentage (nonshown_data[nonshown_data['No-show']=='Yes'].shape[0], nonshown_data['No-show'].shape[0] )
not_showed
```

```
20.10465486645129
```

```
#Male Patient who didnt showed up for their appoitnments
male_not_attended = percentage(nonshown_data[(nonshown_data['No-show']=='Yes') & (nonshown_data['Gender']=='M')].shape[0],nonsho
wn_data[nonshown_data['Gender']=='M'].shape[0])
male_not_attended
```

```
19.864847303443796
```

```
#Male Patient who  showed up for their appoitnments
male_attended = percentage(nonshown_data[(nonshown_data['No-show']=='No') & (nonshown_data['Gender']=='M')].shape[0],nonshown_da
ta[nonshown_data['Gender']=='M'].shape[0])
male_attended
```

```
80.1351526965562
```

```
#Female Patient who didnt showed up for their appoitnments
Fmale_not_attended = percentage(nonshown_data[(nonshown_data['No-show']=='Yes') & (nonshown_data['Gender']=='F')].shape[0],nonsh
own_data[nonshown_data['Gender']=='F'].shape[0])
Fmale_not_attended
```

```
20.23386080380899
```

```
#Female Patient who showed up for their appoitnments
Fmale_attended = percentage(nonshown_data[(nonshown_data['No-show']=='No') & (nonshown_data['Gender']=='F')].shape[0],nonshown_d
ata[nonshown_data['Gender']=='F'].shape[0])
Fmale_attended
```

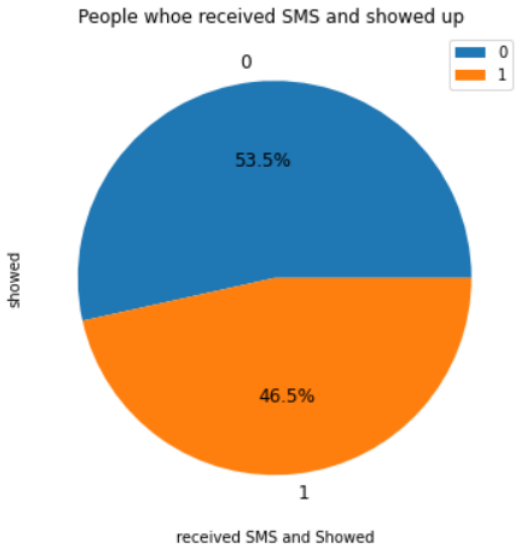
```
79.766139196191
```

```
# Percentage of people who did/didn't received SMS and showed up
rcvd_showed = nonshown_data.groupby('SMS_received')['showed'].mean() * 100
print(rcvd_showed)
```

```
SMS_received
0    83.296689
1    72.425455
Name: showed, dtype: float64
```

```
#Plotting the Percentage of people who received SMS and showed up and who didnt receive and showed up
plt.xlabel("received SMS and Showed")
plt.ylabel("didn't receive and showed up")
pltchart = rcvd_showed.plot.pie(figsize=(6,6), autopct='%1.1f%%', fontsize = 12);
plt.title("People whoe did/didn't received SMS and showed up ")
plt.legend()
```

```
<matplotlib.legend.Legend at 0x127f54c97c0>
```

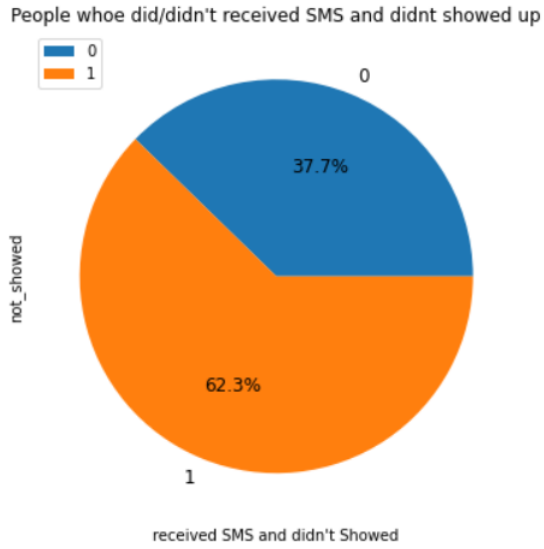



```
#Percentage of people who did/didn treceived SMS and did'nt show up
ntrcvd_ntshowed = nonshown_data.groupby('SMS_received')['not_showed'].mean() * 100
print(ntrcvd_ntshowed)
```

```
SMS_received
0    16.703311
1    27.574545
Name: not_showed, dtype: float64
```

```
#Plotting the Percentage of people who received SMS and showed up and who didnt receive and didnt show up
plt.xlabel("received SMS and didn't Showed")
plt.ylabel("didnt receive and didn't showed up")
pltchart = ntrcvd_ntshowed.plot.pie(figsize=(6,6), autopct='%1.1f%%', fontsize = 12);
plt.title("Status" + ' (%) (Per appointment)\n', fontsize = 15);
plt.title("People whoe did/didn't received SMS and didn't showed up")
plt.legend()
```

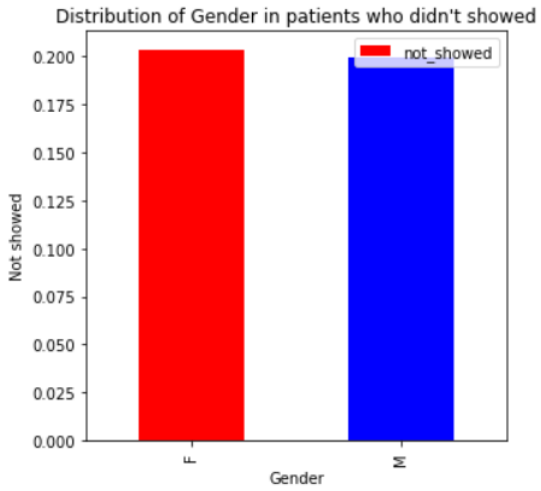
```
<matplotlib.legend.Legend at 0x127f554f7c0>
```



```
# using group by function to find relations between Gender and not attending the appontment
print(nonshown_data.groupby('Gender')['not_showed'].mean())
colors=['red' , 'blue']
nonshown_data.groupby('Gender')['not_showed'].mean().plot(kind='bar',figsize=(5,5) , color= colors );
plt.xlabel("Gender")
plt.ylabel("Not showed")
plt.title("Distribution of Gender in patients who didn't showed")
plt.legend()
```

```
Gender
F    0.203146
M    0.199679
Name: not_showed, dtype: float64
```

```
<matplotlib.legend.Legend at 0x127f5771c40>
```



5. Conclusion

The first thing I noticed in our data set is that about 80% of patients showed up for their appointments, while 20% didn't showed up which I tried to find the possible reasons for their absence.

Overall the gender distribution within our sample is divided to 65% Females and 35 % Males, the median age for the sample is 37 years old, and the range of age is between less than one year and 115 years old.

Around 32% of patients received SMS regarding their appointments, but surprisingly a lot of patient who didn't showed up for their appointments has received SMS about it.

If we took a closer look to the sample, 19% of male hasn't showed up while 80% of them did, on the other side 20% of female patients didn't showed up and 80% did, we can't say that one gender is most likely won't show up, I think gender doesn't affect on the status of showing up.

We can tell that older people are most likely show up for their appointment's vs younger ones , SMS is a good way to courage and remind people to attend their appointments but in my opinion we have to include more characteristics to determine what are the most likely reasons for not showing up.

The Dataset size is quite enough to find facts and conclusions, but I think if we add more characteristics to the data will help us to find more about the most common reasons why patients don't show up for appointments, for example clinic, medical specialty, reason of visit, time period between reservation and booked appointment, but so far we found interesting finding a]in general which answered our questions clearly.