PROJECT 2: INVESTIGATE NO-SHOW APPOINTMENT DATASET

I. INTRODUCTION

The project would like to investigate the data about No-show appointment. This dataset is about more than 100,000 medical appointments in Brazil and would like to find out the insightful information about the presence or absence for a scheduled appointment.

This dataset was extracted from the link: https://s3.amazonaws.com/video.udacity-data.com/topher/2018/July/5b57919a_data-set-options/data-set-options.pdf

The description of this dataset was also explained the above link.

The investigator would like to explore the key information of this datasets and answer some questions as below:

- What factors are important for us to know in order to predict if a patient will show up for their scheduled appointment?
- Which gender has higher percentage of appointment show-up?
- What factors are main reasons for not showing up for a schedule appointment?

II. DATA WRANGLING AND CLEANING

To investigate this dataset, the analyst would like to use Python and Jupyter Notebook for cleaning data and exploring information.

- Pandas, numpy and matplotlib were imported.
- The dataset was imported.
- There are 110527 rows and 14 columns in this dataset.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
```

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

```
df.shape
```

```
(110527, 14)
```

The type of variables are described as picture below:

```
<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 Gender variable, there are two values: F (for Female) and M (for Male)

In each variable Scholarship, Hipertension, Diabetes, Alcoholism, Handcap and SMS_received, there are two values: 0 (standing for No) and 1 (standing for Yes).

Few first rows of the dataset was shown as below:

```
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	

Activate Windo

For data cleaning, the data analyst would like to check whether there are any Null values or duplicated rows or not. As you can see the picture below, there are no null values.

For duplicated rows, there are no duplicates in the data.

```
Entrée [7]: # check for duplicateds in the data
            sum(df.duplicated())

Out[7]: 0
```

III. EXPLORATORY DATA ANALYSIS

1. Overview

Let's take a look at the descriptive information of this dataset.

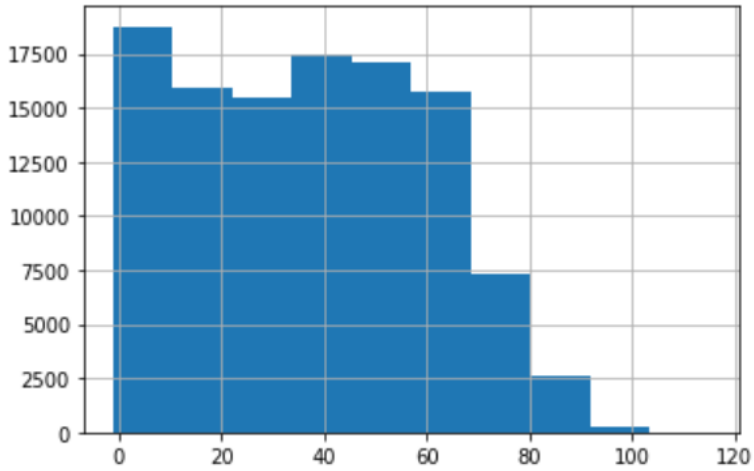
```
df.describe()
```

	PatientId	AppointmentID	Age	Scholarship	Hipertension	Diabetes	Alcoholism	Handcap	SMS_received
count	1.105270e+05	1.105270e+05	110527.000000	110527.000000	110527.000000	110527.000000	110527.000000	110527.000000	110527.000000
mean	1.474963e+14	5.675305e+06	37.088874	0.098266	0.197246	0.071865	0.030400	0.022248	0.321026
std	2.560949e+14	7.129575e+04	23.110205	0.297675	0.397921	0.258265	0.171686	0.161543	0.466873
min	3.921784e+04	5.030230e+06	-1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	4.172614e+12	5.640286e+06	18.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	3.173184e+13	5.680573e+06	37.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	9.439172e+13	5.725524e+06	55.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
max	9.999816e+14	5.790484e+06	115.000000	1.000000	1.000000	1.000000	1.000000	4.000000	1.000000

- Age: the average age is about 37 years old. The youngest is -1 (this should be pregnancy). The oldest is 115 years old. There is 75% of patients older than 55 years old. There are three main groups in this dataset: children age from 0 – 10 years old, group age 34 – 45 and 45 to late 60s.

```
df['Age'].hist()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x27b1125
```



- Scholarship: just only 9.8% of the total patients are enrolled in Brasilian welfare program. This means around 90% of the total patients that didn't have scholarship.
- Hipertension: nearly 20% in total 110527 patients get high blood pressure.
- Diabetes: around 7% people in the dataset get diabetes.
- Alcoholism: only 3% people have alcohol abuse
- Handcap: only 2% of the total patients are handicapped
- SMS_received: 32% of the patients received the text message about the appointment reminder.
- No-show: There are 88,208 patients that didn't show up to the appointment. Only 22319 out of more than 110k cases appeared in the appointment.

```
# count values in No-show column
df['No-show'].value_counts()
```

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

Gender: There are about 72,000 female and around 39,000 male in this dataset.

```
df['Gender'].value_counts()
```

```
F    71840
M    38687
Name: Gender, dtype: int64
```

2. Findings

Let's go deep into the data explorations.

```
NoShow = df['No-show']
counts = NoShow.value_counts()
percent = NoShow.value_counts(normalize=True)
pd.DataFrame({'counts': counts, 'percent' : percent})
```

	counts	percent
No	88208	0.798067
Yes	22319	0.201933

There is up to 80% of the patients that didn't show up to the scheduled appointment. Therefore, this project will pay special attentions to find out which are the reasons behind not coming up as scheduled appointment.

The analyst would like to find out the relationship between No-show and other variables.

Relationship between Gender and No-Show

```
F = df.Gender
counts = F.value_counts()
percent = F.value_counts(normalize=True)
pd.DataFrame({'counts': counts, 'percent' : percent})
```

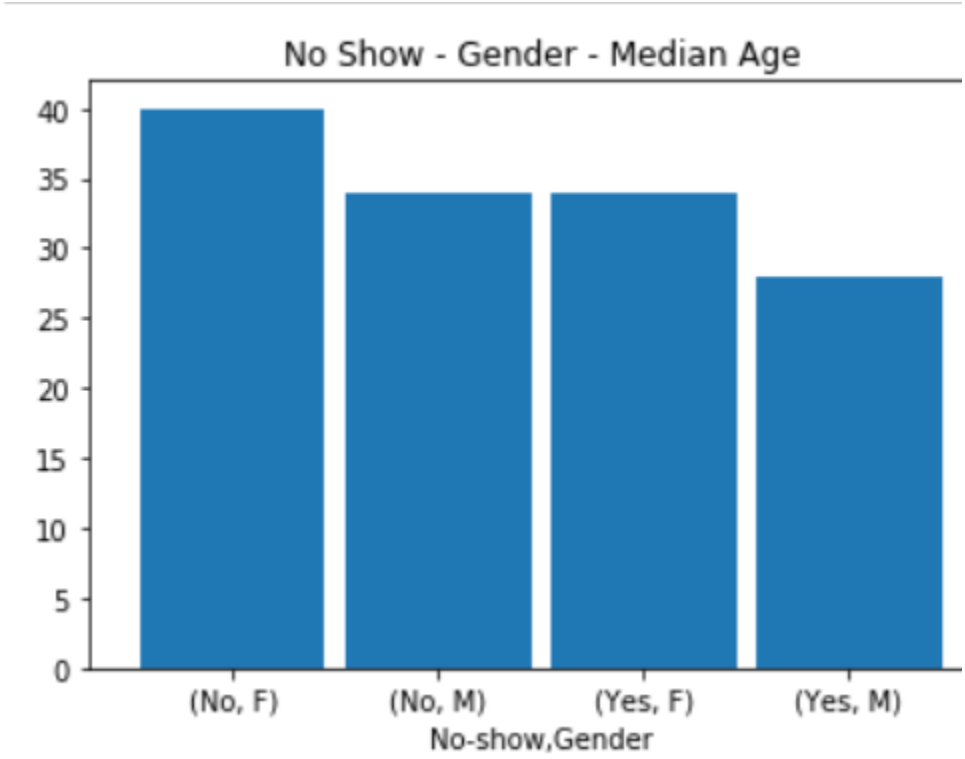
	counts	percent
F	71840	0.649977
M	38687	0.350023

```
pd.crosstab(df['Gender'],df['No-show']).apply(lambda r: r/r.sum(), axis=1)
```

No-show	No	Yes
Gender		
F	0.796854	0.203146
M	0.800321	0.199679

There are 65% female patients and 35% male ones in the dataset. However, both gender show the same percentage of not showing up to the appointment. 80% of female patients didn't show up to the appointment and neither did male. This seems that the difference in gender didn't play any important roles in predicting who will not appear as scheduled appointment.

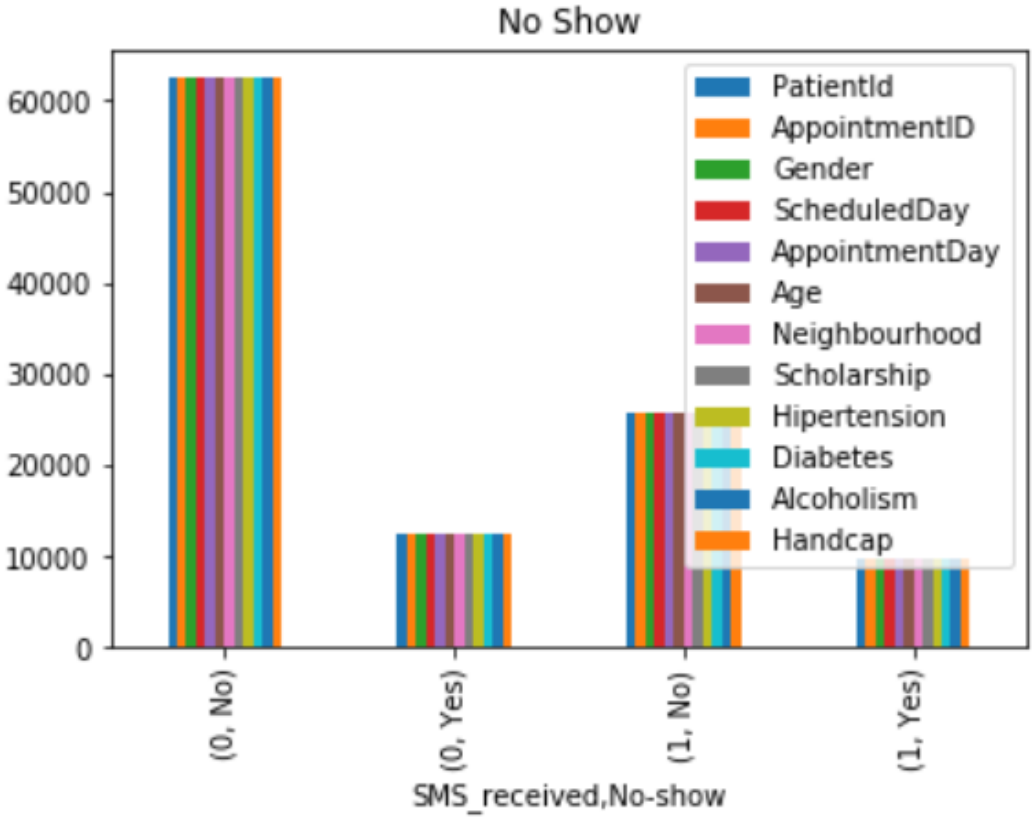
Relationship between No-show – Gender – Median age



The median age of female who didn't show up to the appointment is about 40 years old

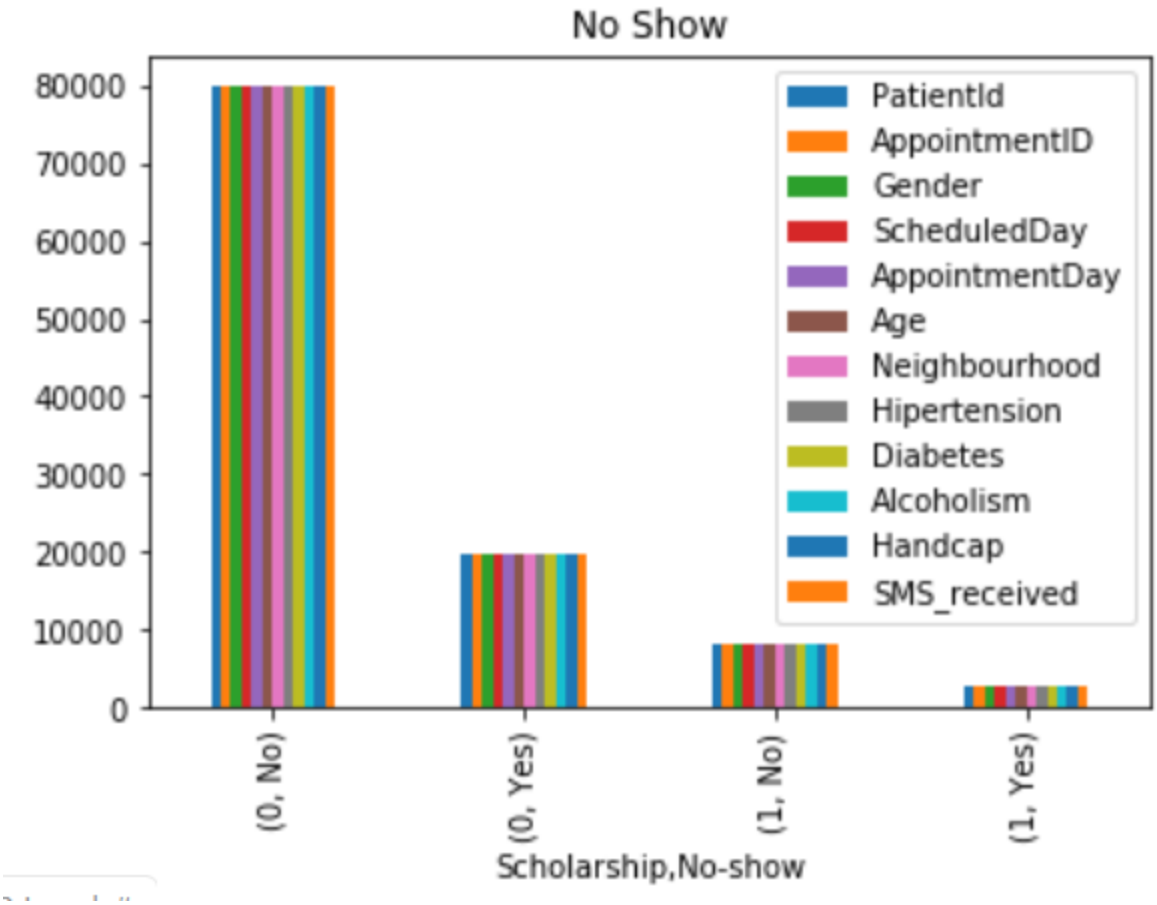
The median of male who didn't show up to the appointment is about 35 years old

Relationship between No-show – SMS_received



According to the graph above, among the patient who didn't receive the SMS, almost of them didn't come up as scheduled appointment. The number of these patients also accounts for the highest proportion in comparisons to who received SMS but didn't show up or who showed up without SMS notification. Therefore, not receiving SMS can be considered one of the main reasons that the patient didn't come up.

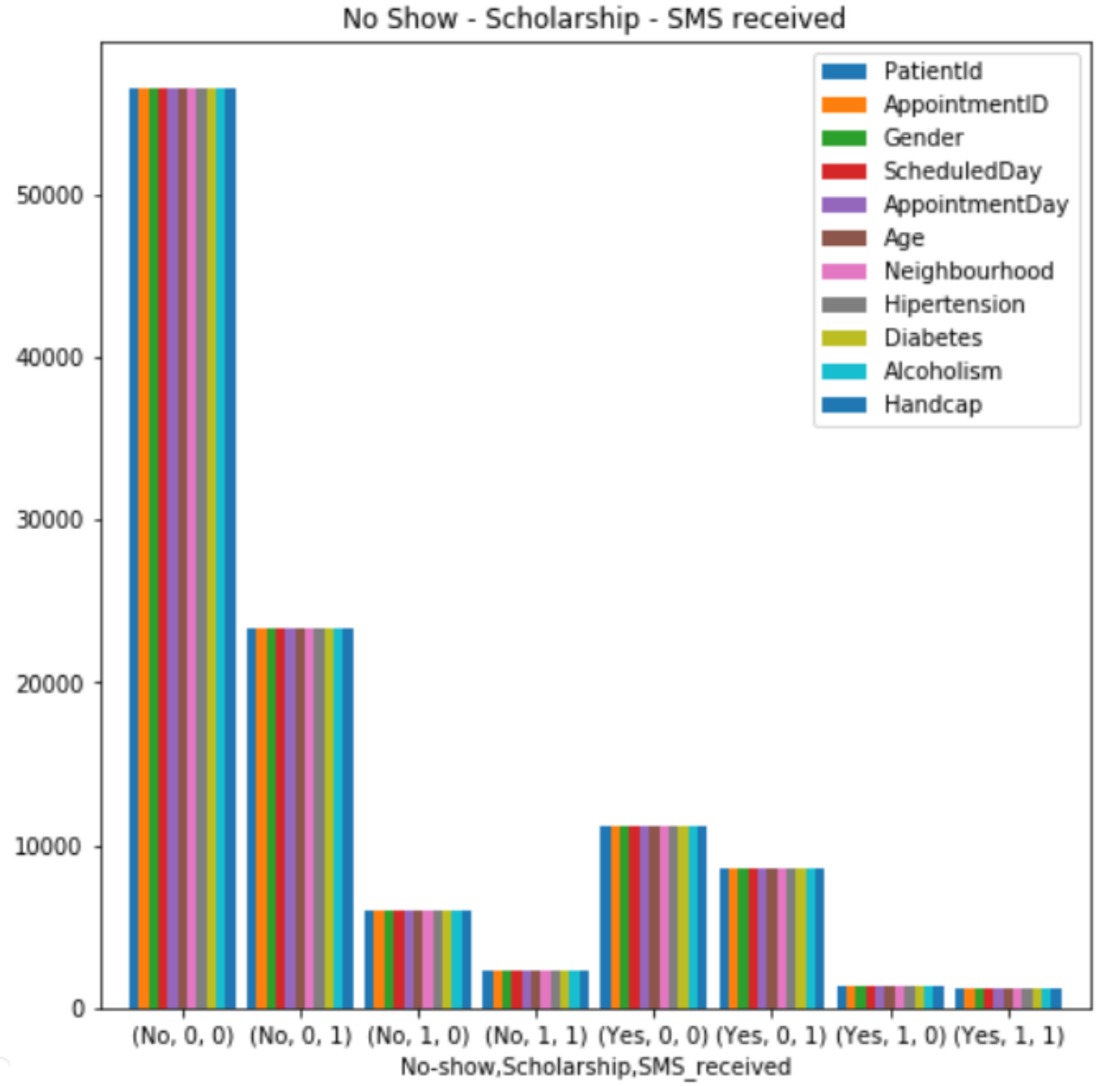
Relationship between No-show and Scholarship



Scholarship	0	1
No-show		
No	0.906097	0.093903
Yes	0.884493	0.115507

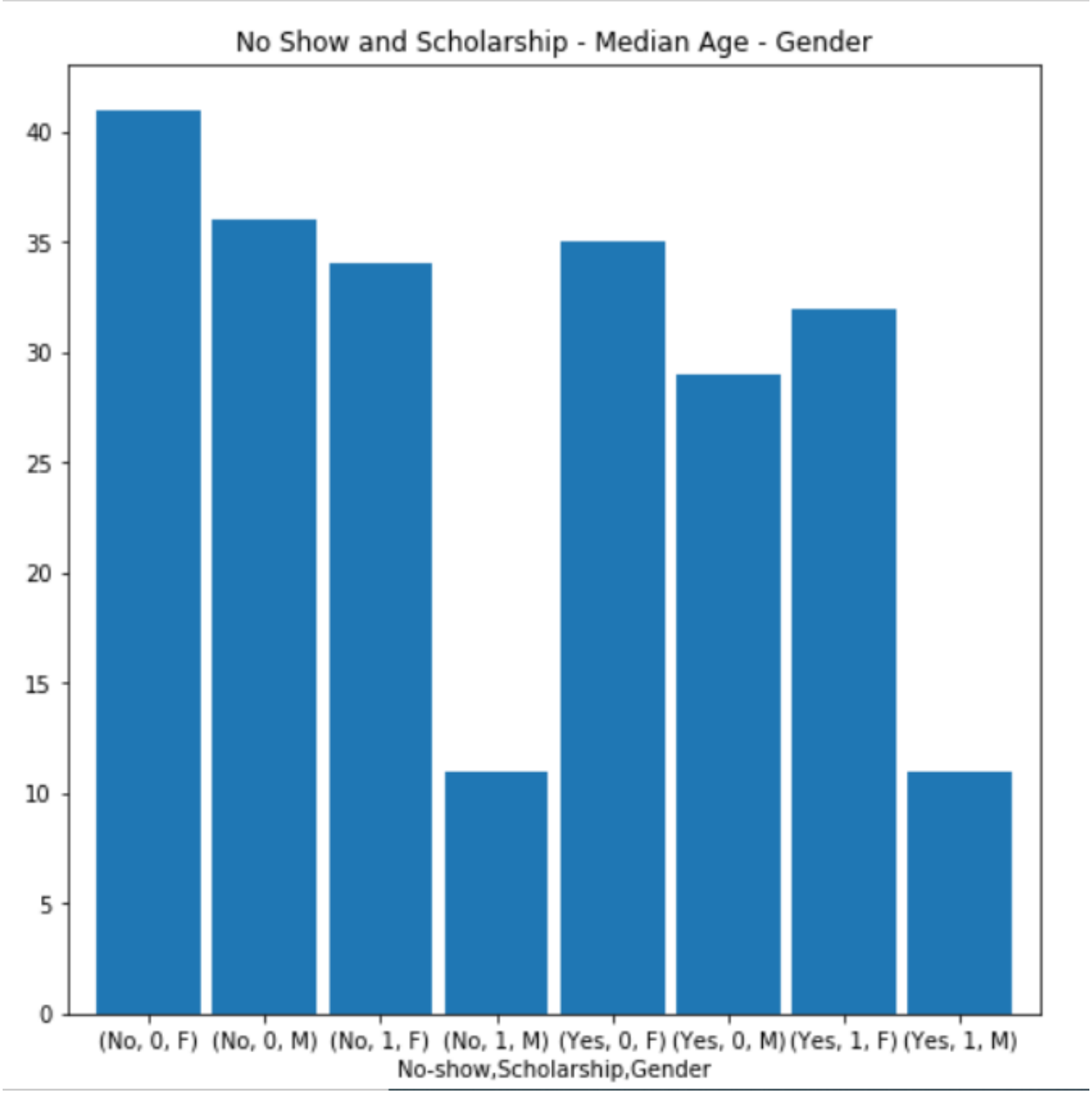
Among 88,208 patients (out of total 110k patients) that didn't show up to the appointment, only 9% of them had scholarship. 91% of them didn't have scholarship. According to the graph, there is a majority people in total patients (80k out of 110k people) that neither had scholarship nor showed up to the appointment. Therefore, it can be concluded that there is a positive correlation between no scholarship and no-show. Having no scholarship made people less likely to show up in the appointments.

Relationship between No-show – Scholarship – SMS_received



When taking three variable no-show, scholarship and SMS received into consideration, we can see that they have strong relationship with each other. The number of patients who didn't show up because of no SMS receipt and no scholarships accounts for the biggest cluster. There are approximately 60,000 patients like that, take up for around 60% of the total more than 110,000 patients. There is only a small group of patients who still showed up regardless of no scholarship and no SMS notices.

Relationship between No-show – Scholarship – Gender – Median Age



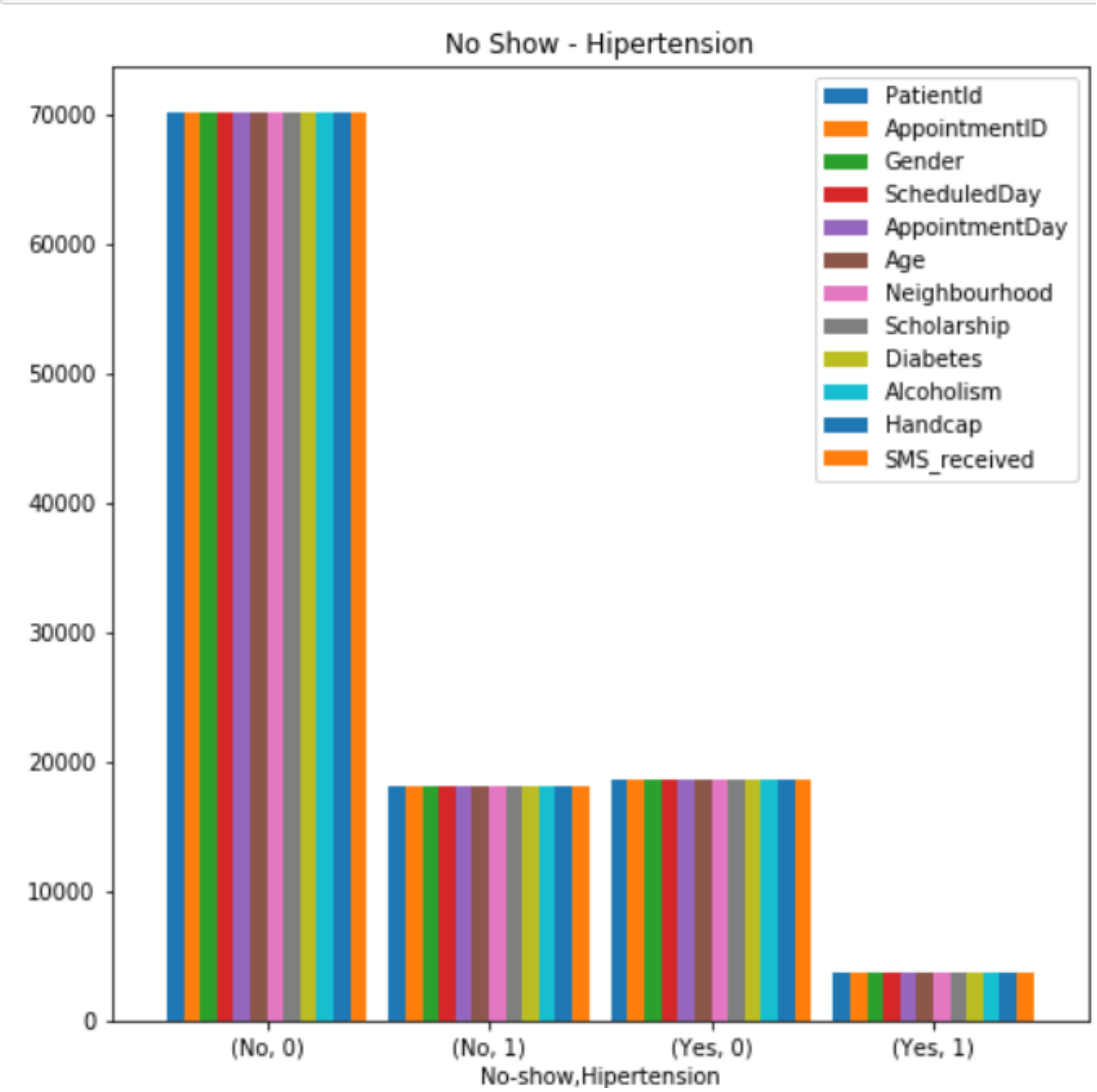
The median age of female patients who didn't had scholarship and show up is around 41 years old.

The median age of male patients who didn't had scholarship and show up is around 36 years old.

Relationship between No show and Hipertension disease

According to the early information statistics, hipertension accounts for the highest percentage of patients in comparisons to other diseases or health conditions. There are approximately 20% of

total patients having high blood pressure. Among them, there are more people who didn't show up to a scheduled appointment than who did.



IV. CONCLUSION

In summary, no scholarship and no SMS receiving are considered the important factors to predict that patients may not show up as scheduled appointment. Patients who at had no scholarship or receive no SMS message or didn't get both were less likely to appear in the appointments.

Difference in gender didn't make any difference in predicting the trends.

Health conditions didn't play significant roles in predicting the trends as the percentage of people having health problems was pretty low. However, among them, hypertension is the most important information. There are more high blood pressure people who didn't show up to a scheduled appointment than who did.

Regarding age, there is a group of female patients who are around 41 years old and didn’t show up to the appointment because of no scholarship. There is also a similar male group like that with the median age of 36 years old.