REPORT ON NOSHOW MEDICAL APPOINTMENTS

MOSTAFA FATHY

1.Introduction

Patients have appointment, we want to know which patient will attend or miss.

Problems We are Trying to Solve

- Know the reason of Why do 30% of patients miss their scheduled appointments

Main Goal

- Create an analytical framework to understand

About Dataset

This dataset collects information from 100k medical appointments in Brazil and is focused on the question of whether or not patients show up for their appointment. A number of characteristics about the patient are included in each row.

"PatientId": indicates the patient ID; duplication is possible due to cases where the same patient booked more than one appointment.

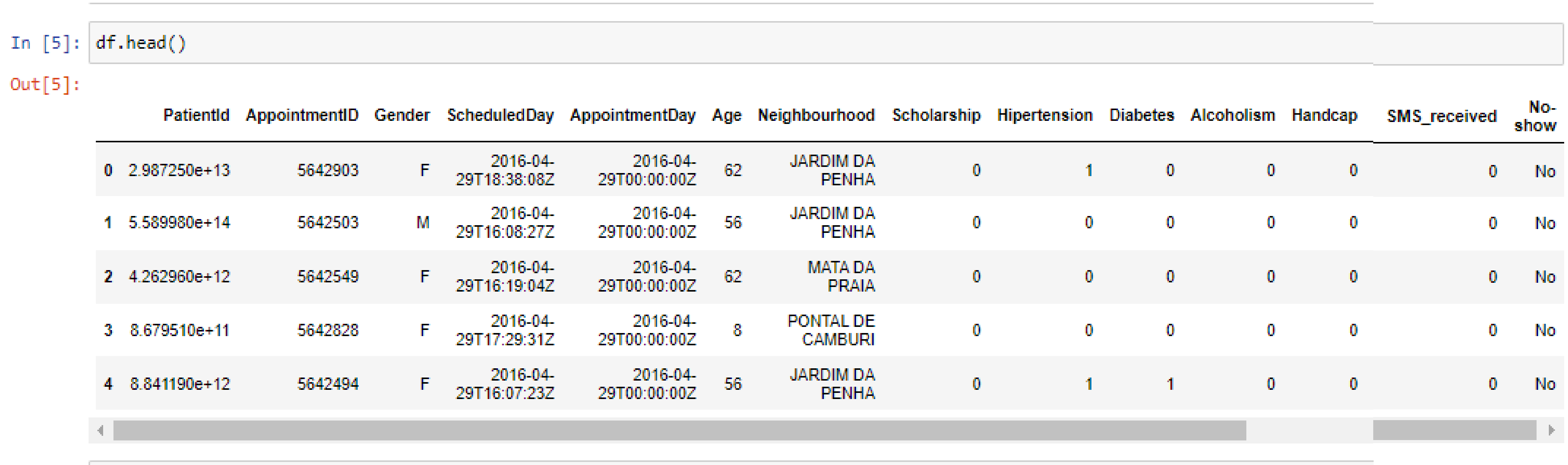
- "AppointmentID": indicates appoint ID, this field should be unique
- "Gender": indicates the patient's gender (M/F)
- "AppointmentDay": indicates the date/time the patient called to book their appointment.
- "Age": indicates the patient's age.

- "ScheduledDay" tells us on what day the patient set up their appointment.
- "Neighborhood" indicates the location of the hospital.
- "Scholarship" indicates whether or not the patient is enrolled in Brasilian welfare program Bolsa Família.
- "Hipertension": indicates whether or not the patient is experiencing Hypertension.
- "Diabetes": indicates whether or not the patient is experiencing Diabetes.
- "Alcoholism": indicates whether or not the patient is experiencing Alcoholism.
- "Handcap": indicates whether or not the patient is with special needs.
- "SMS_received": indicates whether or not the patient has received a reminder text message.
- "No-show" ‘No’ if the patient showed up to their appointment, and ‘Yes’ if they did not show up.

Questions for Analysis

1-What is the overall appointment show vs no show rate?

2- What factors are important to know in order to predict if a patient will show up for their scheduled appointment?



In [5]:
```
df.head()
```

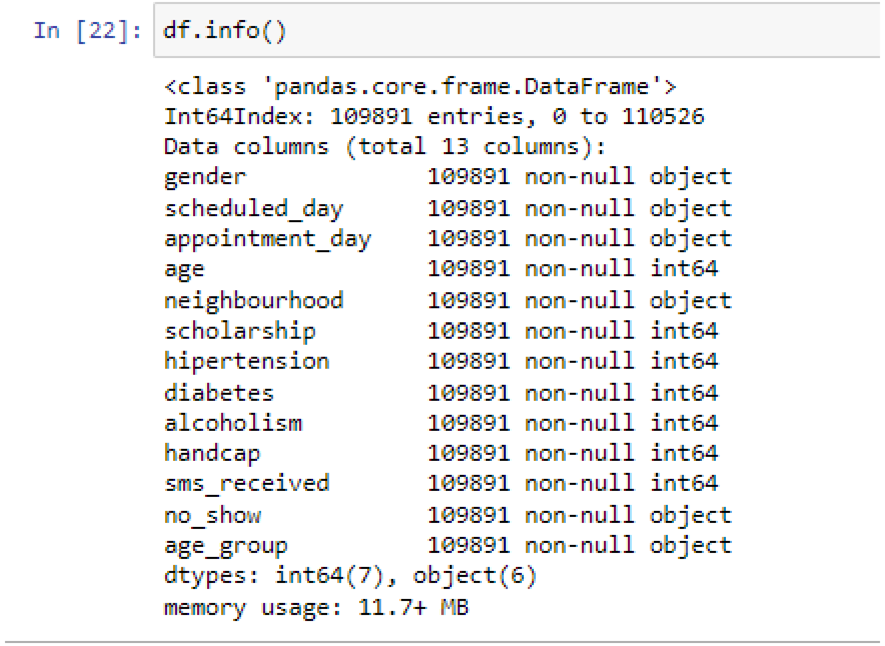
Out[5]:

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension	Diabetes	Alcoholism	Handcap	SMS_received	No-show
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	0	No
1	5.589980e+14	5642503	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0	0	0	0	0	No
2	4.262960e+12	5642549	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0	0	0	0	0	No
3	8.679510e+11	5642828	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0	0	0	0	0	No
4	8.841190e+12	5642494	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	1	1	0	0	0	No

2.Data Cleaning & Pre-processing

Data Cleaning is an important phase in any data science project, if our data is clean then only we can provide it to our machine learning model. Uncleaned Data can further lead our model with low accuracy. And, If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset.

- drop duplicated rows
- drop invalid value of age
- drop null values

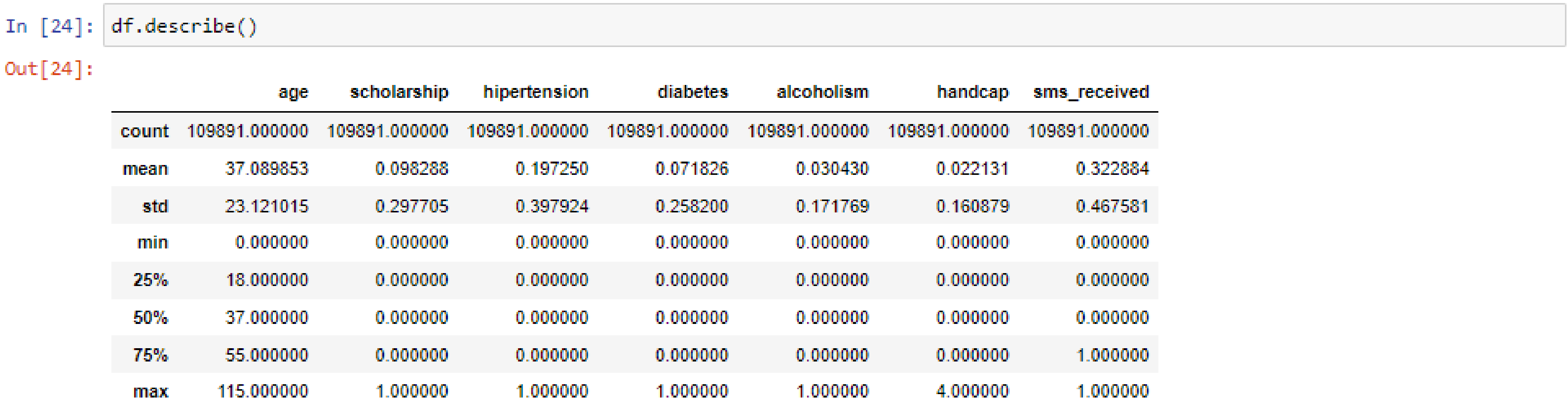


```
In [22]: df.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 109891 entries, 0 to 110526
Data columns (total 13 columns):
gender             109891 non-null object
scheduled_day      109891 non-null object
appointment_day    109891 non-null object
age                109891 non-null int64
neighbourhood      109891 non-null object
scholarship        109891 non-null int64
hipertension       109891 non-null int64
diabetes           109891 non-null int64
alcoholism         109891 non-null int64
handcap            109891 non-null int64
sms_received       109891 non-null int64
no_show            109891 non-null object
age_group          109891 non-null object
dtypes: int64(7), object(6)
memory usage: 11.7+ MB
```

3. EDA & Business Implication

EDA stands for exploratory data analysis where we explore our data and grab insights from it. EDA helps us in getting knowledge in form of various plots and diagrams where



```
In [24]: df.describe()
```

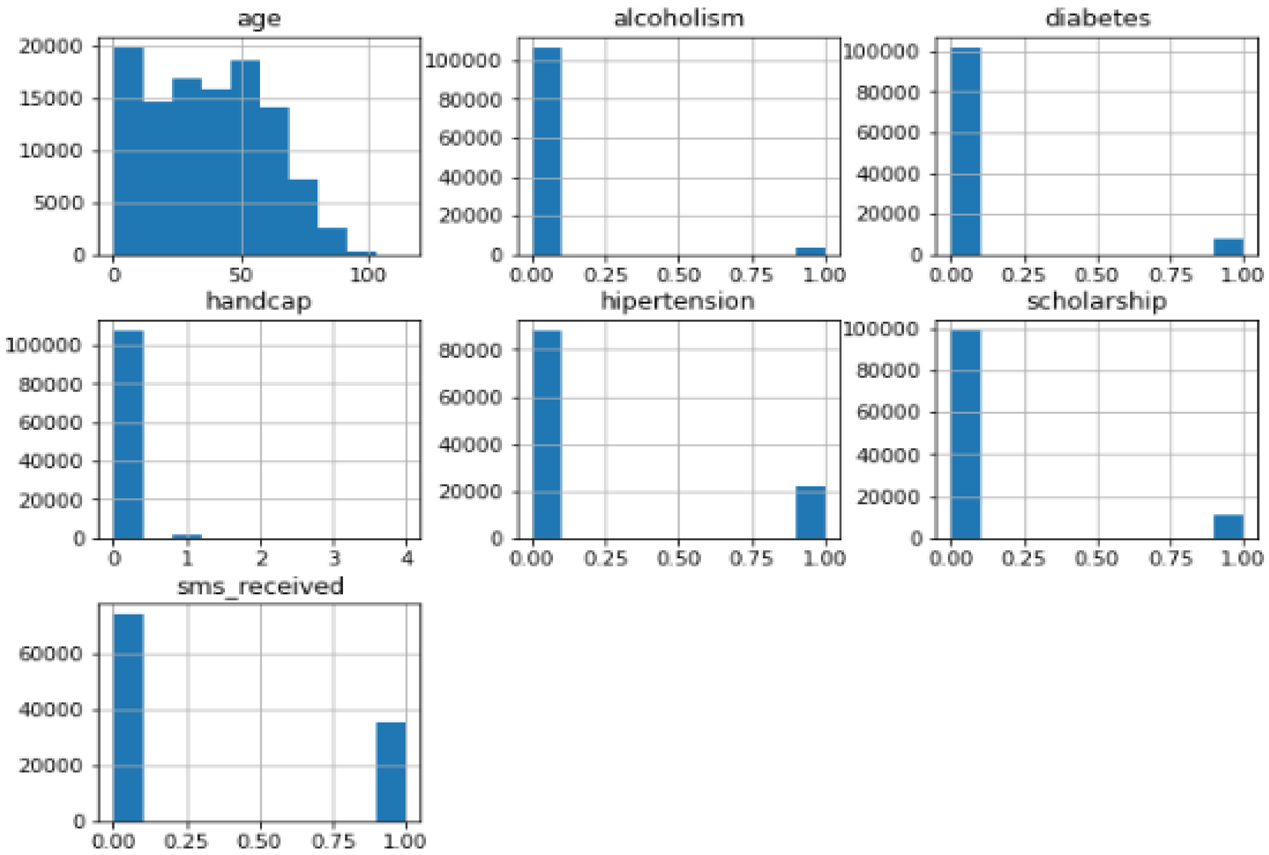
Out[24]:

	age	scholarship	hipertension	diabetes	alcoholism	handcap	sms_received
count	109891.000000	109891.000000	109891.000000	109891.000000	109891.000000	109891.000000	109891.000000
mean	37.089853	0.098288	0.197250	0.071826	0.030430	0.022131	0.322884
std	23.121015	0.297705	0.397924	0.258200	0.171769	0.160879	0.467581
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	18.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	37.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	55.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
max	115.000000	1.000000	1.000000	1.000000	1.000000	4.000000	1.000000

Observation:

Age: ages between 0 -115 years. Mean is 37 years

Uni-variate Analysis - By Histogram



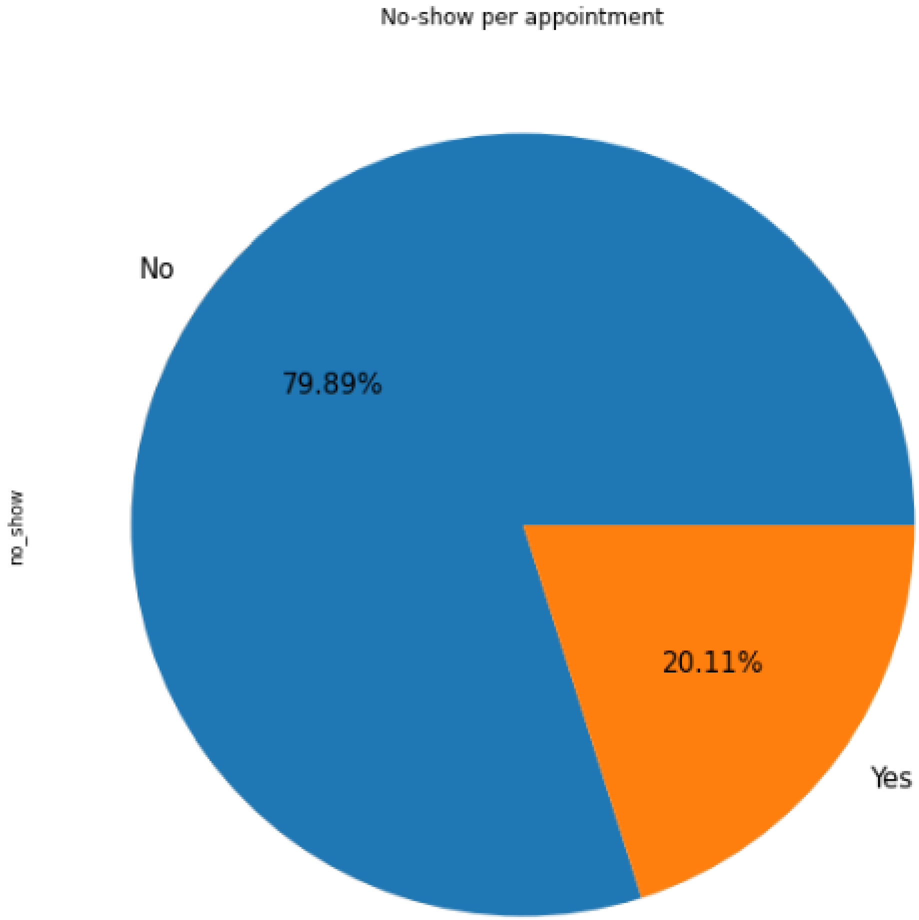
Question 1: what is the overall appointment show vs no show rate ?

Observation:

the overall show rate is 79.89%.

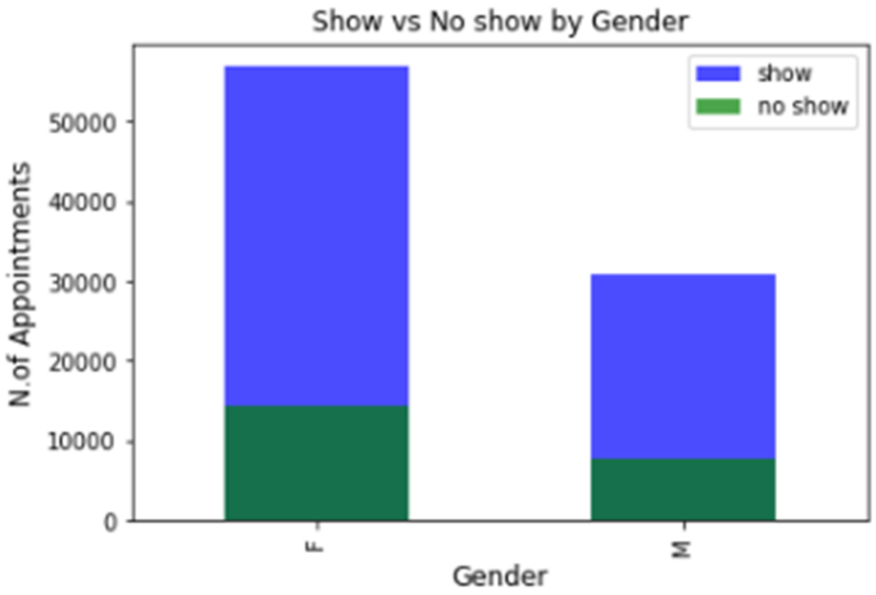
the overall no show rate is 20.11%.

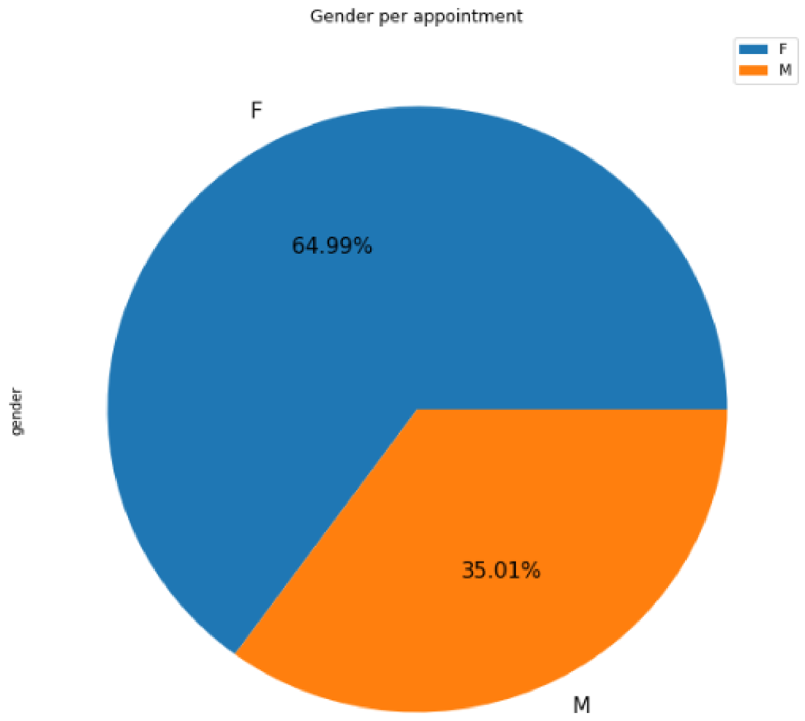
We can see the overall show rate is higher.



Question 2: What factors are important to know in order to predict if a patient will show up for their scheduled appointment?

1.Gender:





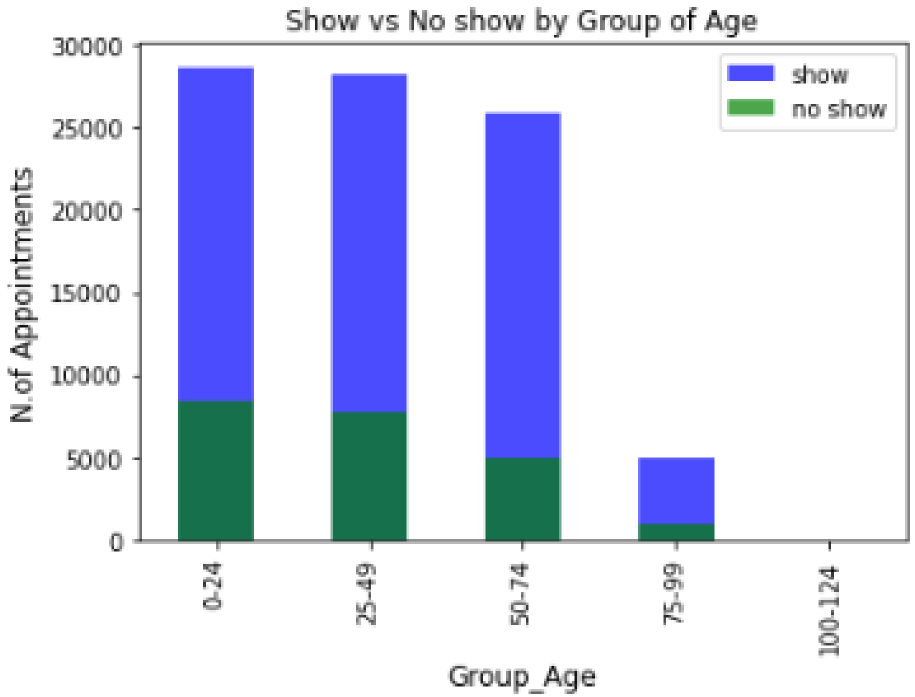
Observation:

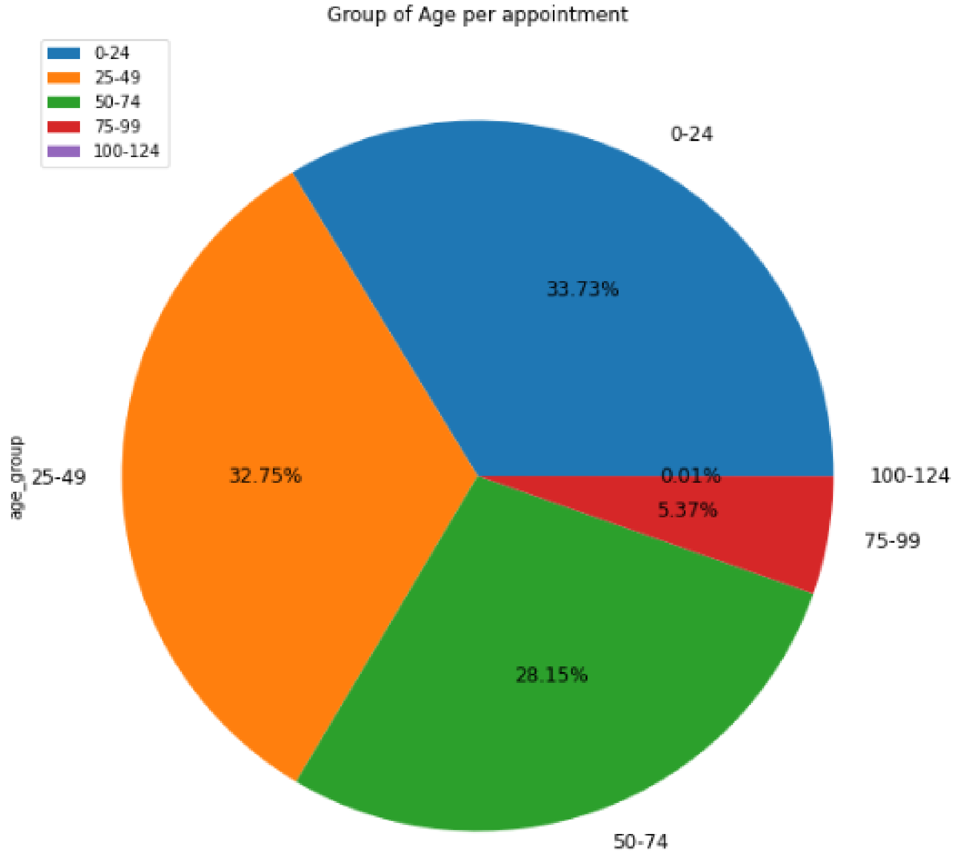
The proportion of appointments where patients who are females is 64.99% .

The proportion of appointments where patients who are males is 35.01% .

The Show up rate of patients who are male is the highest with 80.14%, where the show up rate of patients who are female is 79.75% .

2.Age:


Show vs No show by Group of Age
show
no show
N.of Appointments
30000
25000
20000
15000
10000
5000
0
0-24
25-49
50-74
75-99
100-124
Group_Age


Group of Age per appointment
0-24
25-49
50-74
75-99
100-124
age_group
0-24
33.73%
25-49
32.75%
50-74
28.15%
75-99
5.37%
100-124
0.01%

Observation:

The proportion of appointments where patients who are group1 (from 0 to 24) is 33.73%

The proportion of appointments where patients who are group2 (from 25 to 49) is 32.75%

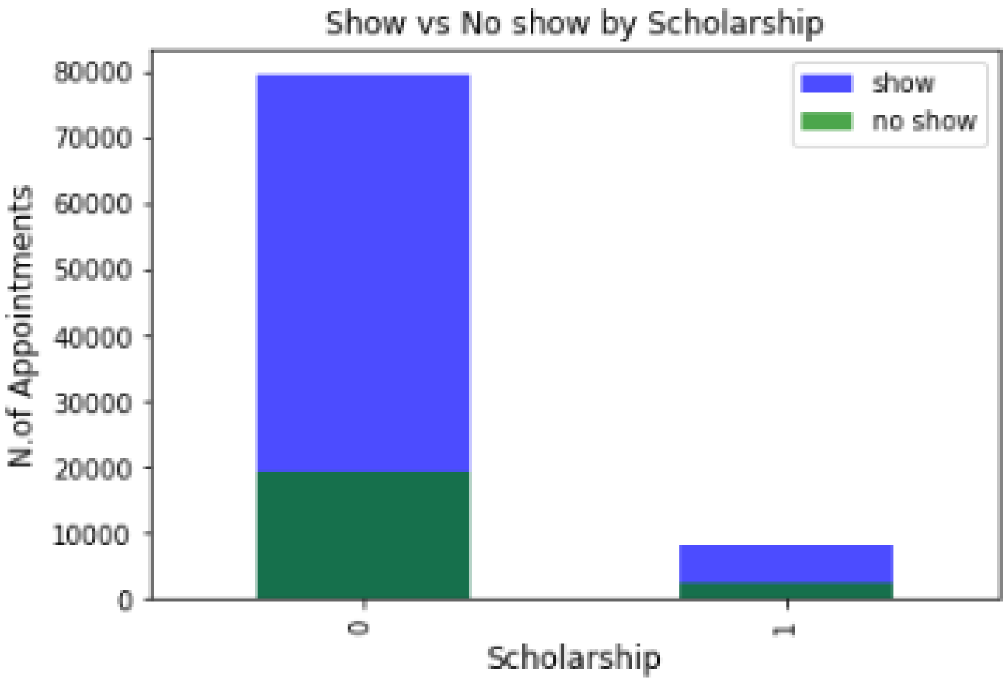
The proportion of appointments where patients who are group3 (from 50 to 74) is 28.15%

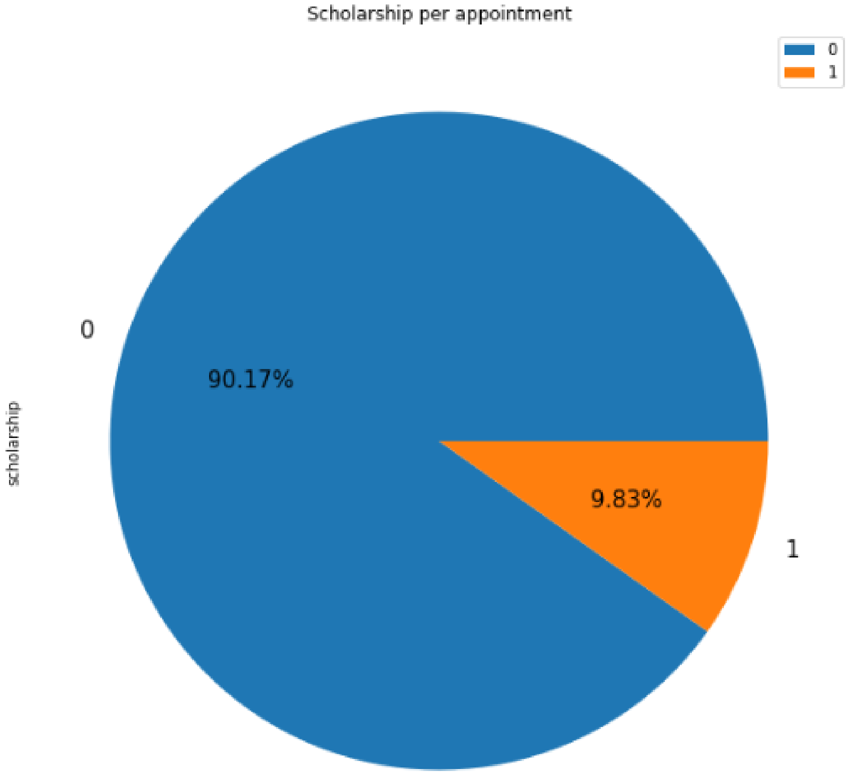
The proportion of appointments where patients who are group4 (from 75 to 99) is 5.37%

The proportion of appointments where patients who are group5 (from 100 to 124) is 0.01%

The Show up rate of patients who are group4 is the highest with 84.19%

3. Scholarship:





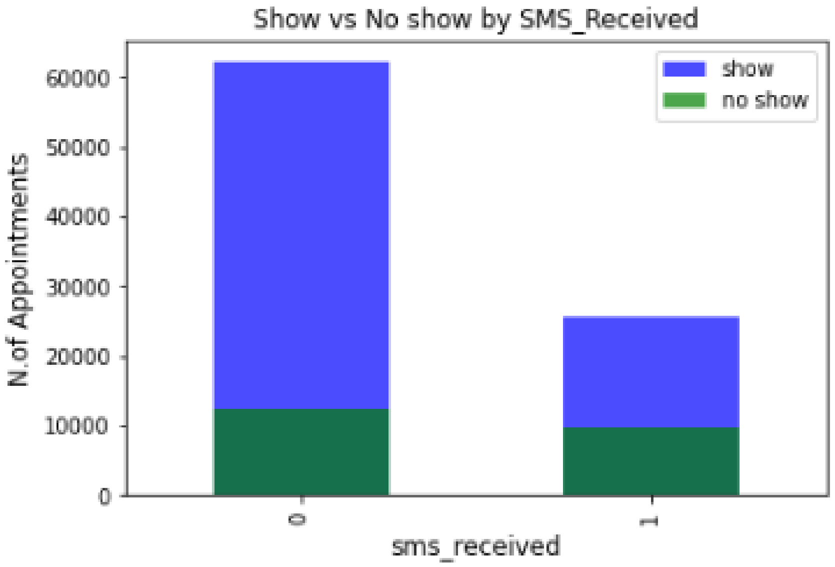
Observation:

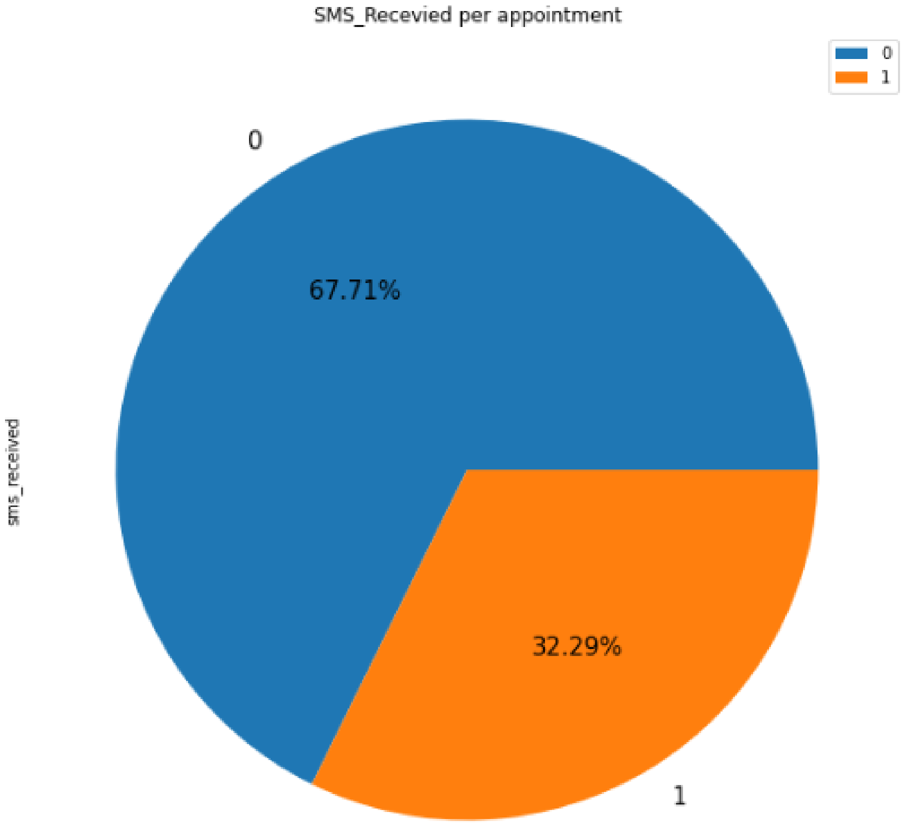
The proportion of appointments where patients who have scholarship is 9.83%

The proportion of appointments where patients who don't have scholarship is 90.17%

The Show up rate of patients who are don't have scholarship is the highest with 80.27%, where the show up rate of patients who have scholarship is 76.33%

4. SMS Received:





Observation:

The proportion of appointments where patients who receive SMS 67.71%

The proportion of patients who don't receive SMS is 32.29%

The Show-up rate of patients who don't receive SMS is the highest with 83.45%, where the show-up rate of patients who receive SMS is 72.42%

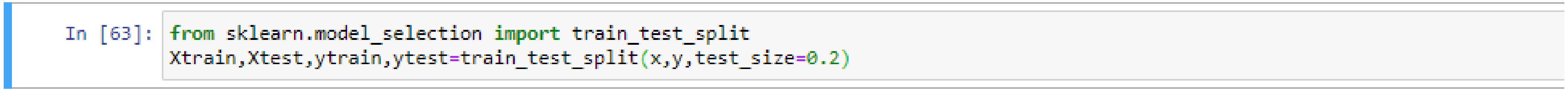
4. Modeling Building

After cleaning and processing the data then comes the modeling part which includes building Machine Learning models.

Then the data needs to be split into 2 sets

1. Training set - This will be the part of the dataset which the model will be using to train itself, the size should be at least 60-70% of the total data we've.

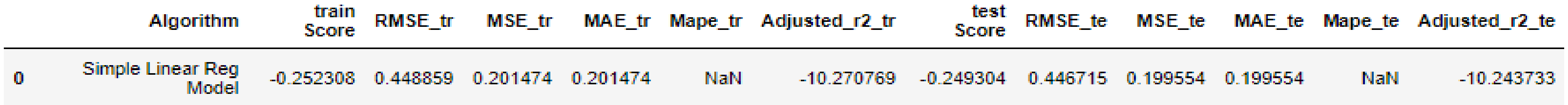
2. Testing set - To evaluate how the model is performing on the unseen data on which the model will be doing future predictions on, test set is used. It helps to understand how much error is there between actual and predicted values.



```
In [63]: from sklearn.model_selection import train_test_split
         Xtrain,Xtest,ytrain,ytest=train_test_split(x,y,test_size=0.2)
```

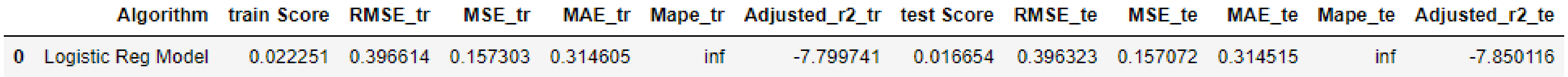
We need to build different regression algorithms:

Linear Regression



	Algorithm	train Score	RMSE_tr	MSE_tr	MAE_tr	Mape_tr	Adjusted_r2_tr	test Score	RMSE_te	MSE_te	MAE_te	Mape_te	Adjusted_r2_te
0	Simple Linear Reg Model	-0.252308	0.448859	0.201474	0.201474	NaN	-10.270769	-0.249304	0.446715	0.199554	0.199554	NaN	-10.243733

Logistic Regression



	Algorithm	train Score	RMSE_tr	MSE_tr	MAE_tr	Mape_tr	Adjusted_r2_tr	test Score	RMSE_te	MSE_te	MAE_te	Mape_te	Adjusted_r2_te
0	Logistic Reg Model	0.022251	0.396614	0.157303	0.314605	inf	-7.799741	0.016654	0.396323	0.157072	0.314515	inf	-7.850116

5.Recommendation

The best performance is given by Logistic Regression model.

THANK YOU