Project: Investigate a Dataset - [No-show Appointments]

The goal of this project is to investigate a dataset of appointment records for Brasil public hospitals. The data includes some attributes of patients and state if the patients showed up to appointments. The analysis should be focused on finding trends influencing patients to show or not show up to appointments. Using descriptive statistics the following question should be answered: What factors are important for us to know in order to predict if a patient will show up for their scheduled appointment? Predictive analytics is out of scope of this project.

Dataset Description

I have looked into the dataset and managed a few problems like unifying names, removing wrong data, adding new features based on existing data. I have also investigated most of independent variables in the dataset and made a few observations comparing them to each other as well as to the dependent one (no_show). As this was only an exploratory analysis, many potential correlations may remain uncovered. The data should be investigated further with more advanced statistical analysis to potentially reveal new insights and correlations.

['Gender', 'Neighbourhood', 'Scholarship', 'Hipertension', 'Diabetes', 'Alcoholism', 'Handcap', 'SMS_received', 'Age_Stage', 'AttendanceDay', 'AttendanceMonth', 'ScheduledMonth', 'ScheduledDay']

Question(s) for Analysis

1- What is the most frequented gender? and, What is the common case(Show or Not)?

2- What is the most neighbourhood tends to be show?

3- Is having scholarship important for cases to show?

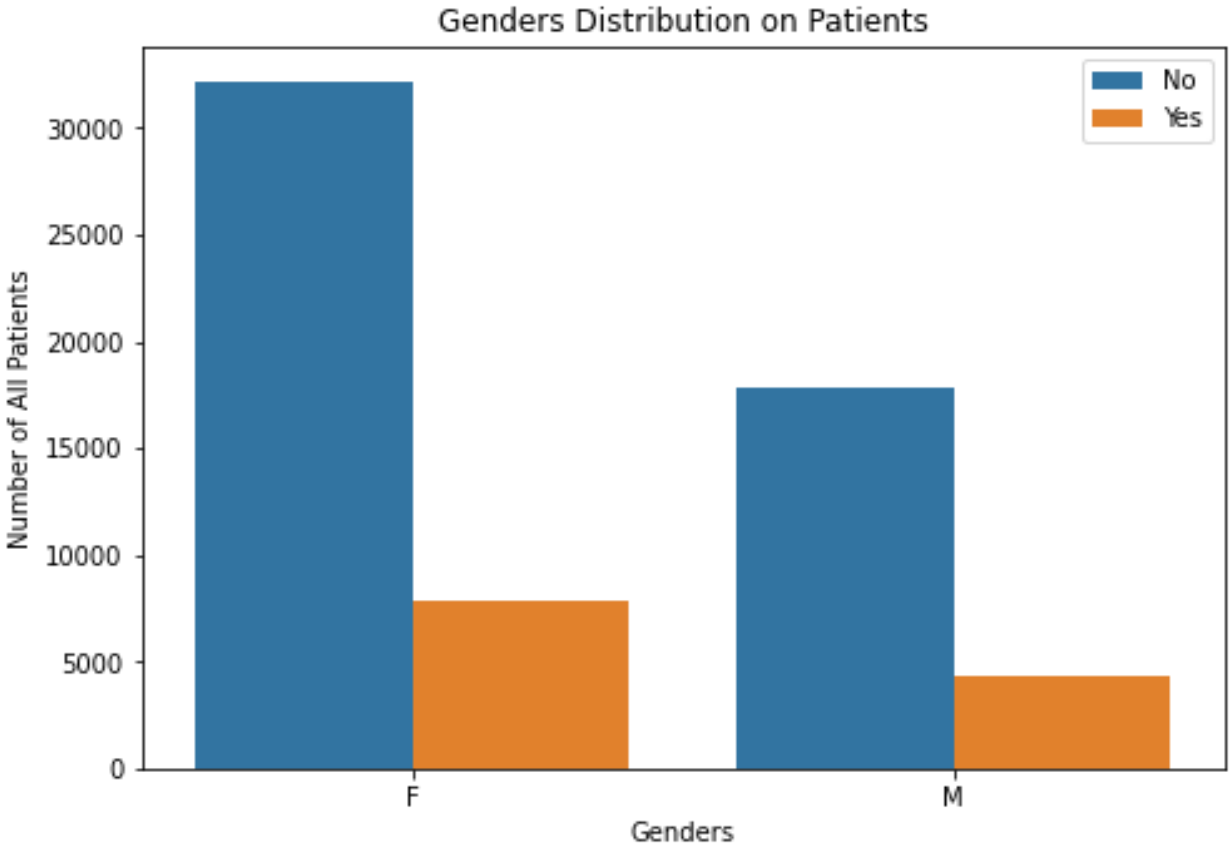
4- Is having an adjective from these list ['Hipertension', 'Diabetes', 'Alcoholism', 'Handcap'] playes an important role in showing?

5- What is the most frequented age and what is the most frequent option to it?

6- How each element in this list ['AttendanceDay', 'AttendanceMonth', 'ScheduledMonth', 'ScheduledDay'] effect the showing or not?

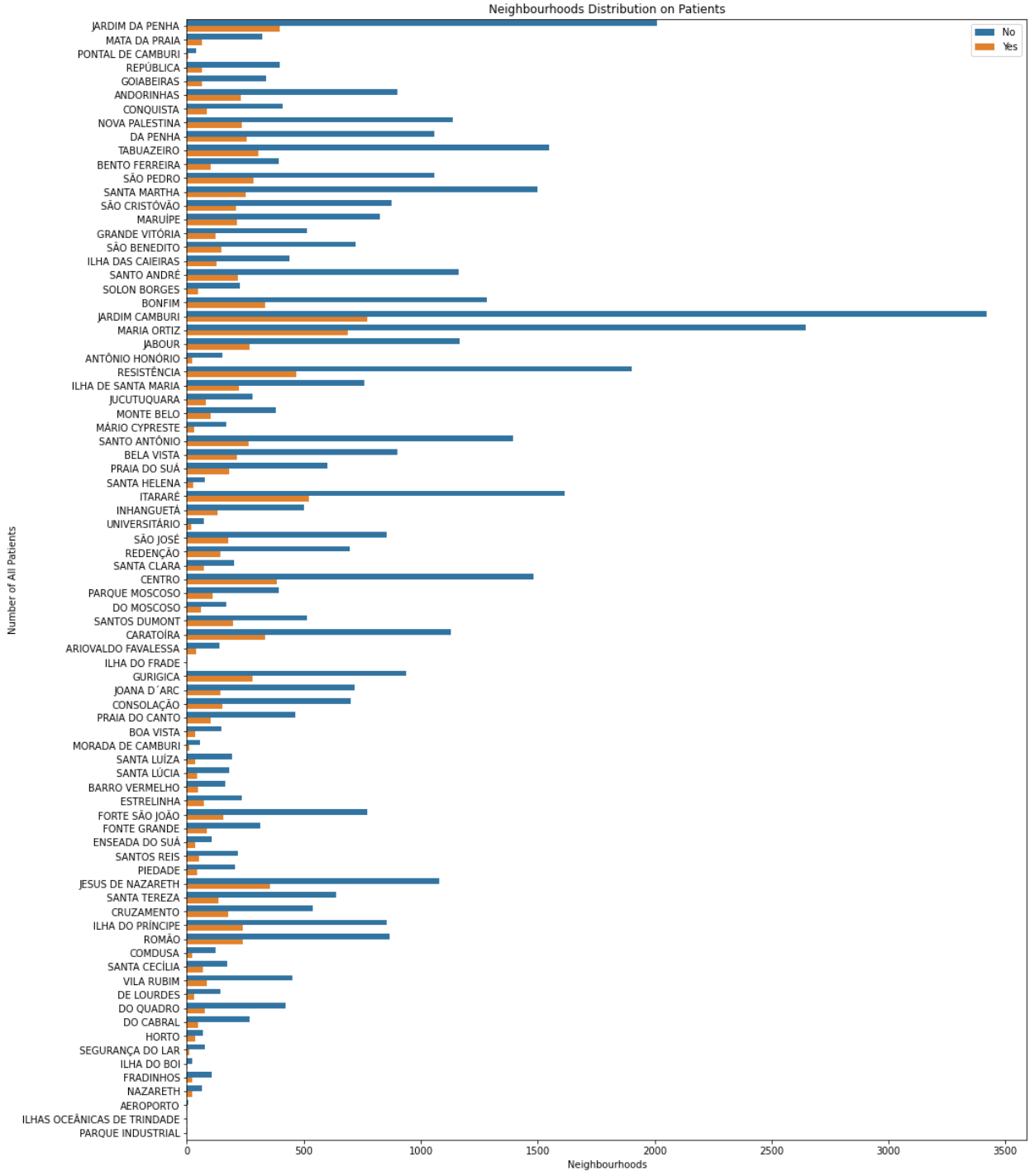
Conclusions

Finding 1- The most frequented gender is females and the common case for females and males is showing for both of them.
Conclusion 1- We can make offers to males to encarage them to attend.

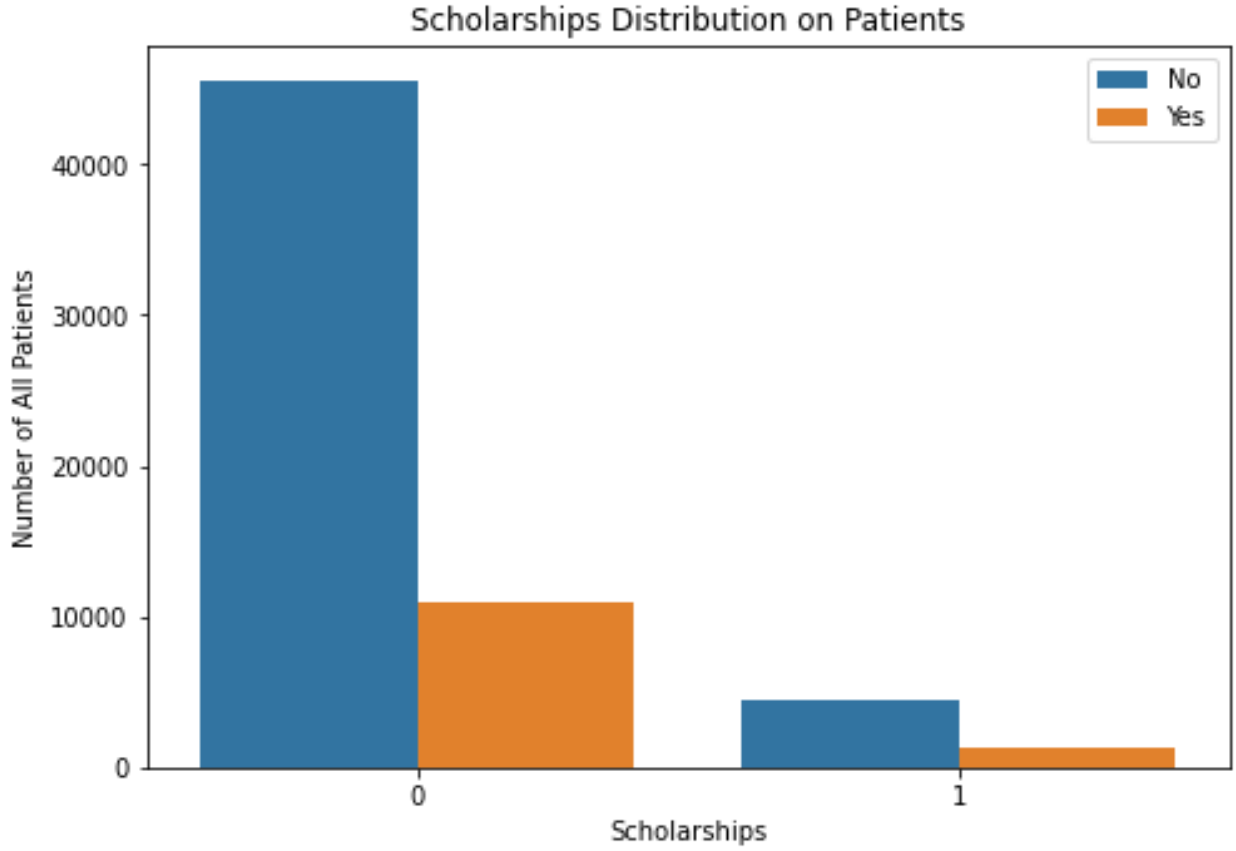


Finding 2-The most frequented nieghbourhood is JARDIM CAMBURI and most of cases come from this tends to be show, also the same investigation for the most frequented 5 neighbourhoods.

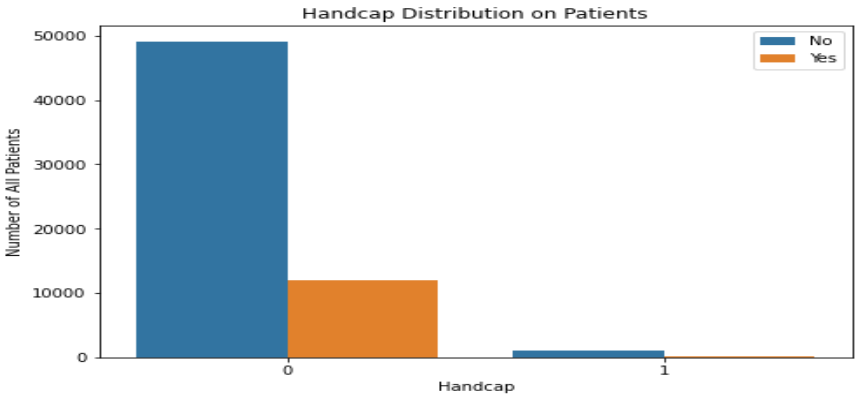
Conclusion 2-It is important to take pledge for advance booking form patients those come from nieghbourhoods other those five to ensure they will attend and will not waste the opportunity on others.

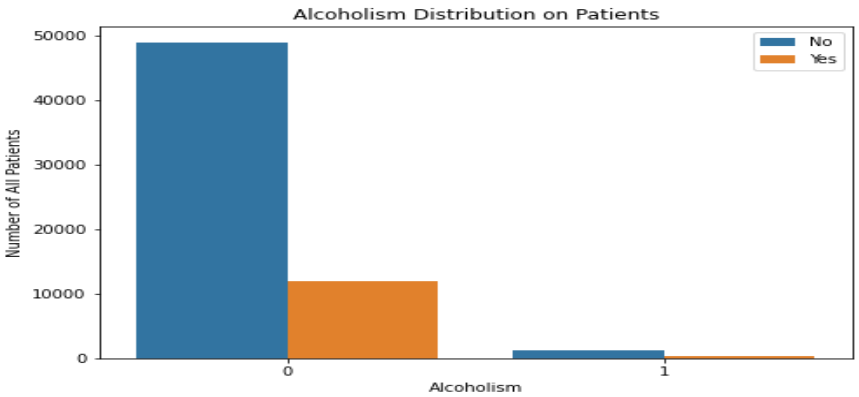


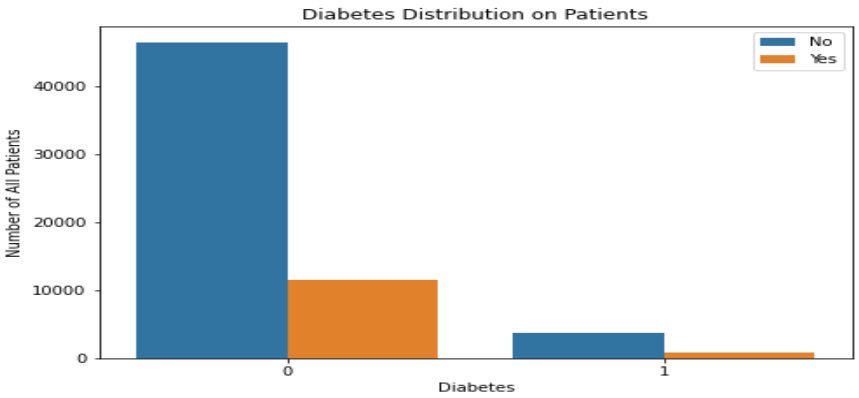
Finding 3-Most of patients were not had a scholarship at the same time most of them are showing.
Conclusion 3-So, it is not important to offer a scholarship for patients cause it will have a high cost to the owner.

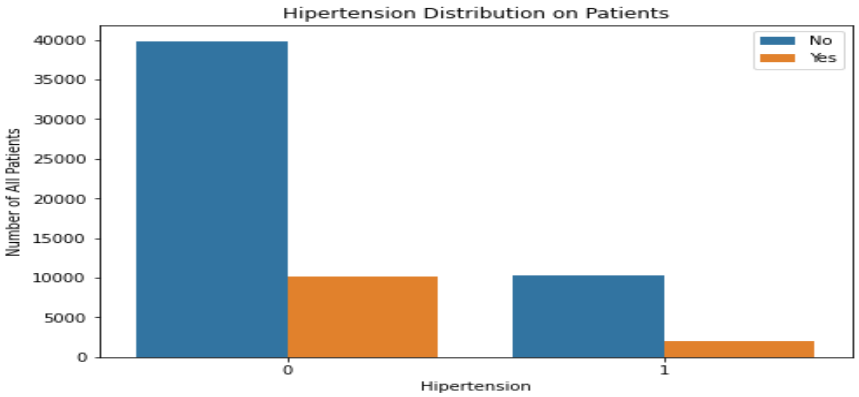


Finding 4-Most of cases do not have any adjective and at the same time most of them are showing.
Conclusion 4-It is not important to marke them adjectives or disabilities to ensure that, they will attend.

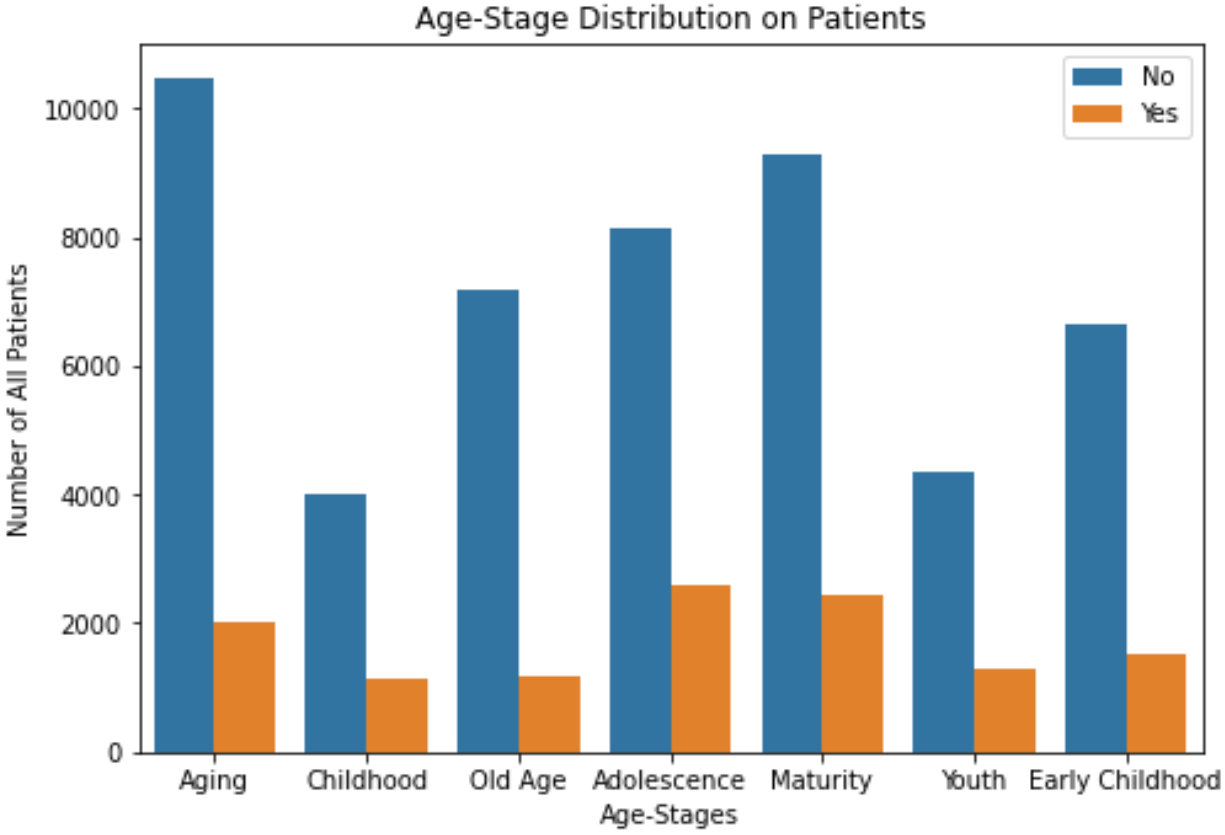

Handcap Distribution on Patients
No
Yes
Number of All Patients
50000
40000
30000
20000
10000
0
0
1
Handcap


Alcoholism Distribution on Patients
No
Yes
Number of All Patients
50000
40000
30000
20000
10000
0
0
1
Alcoholism

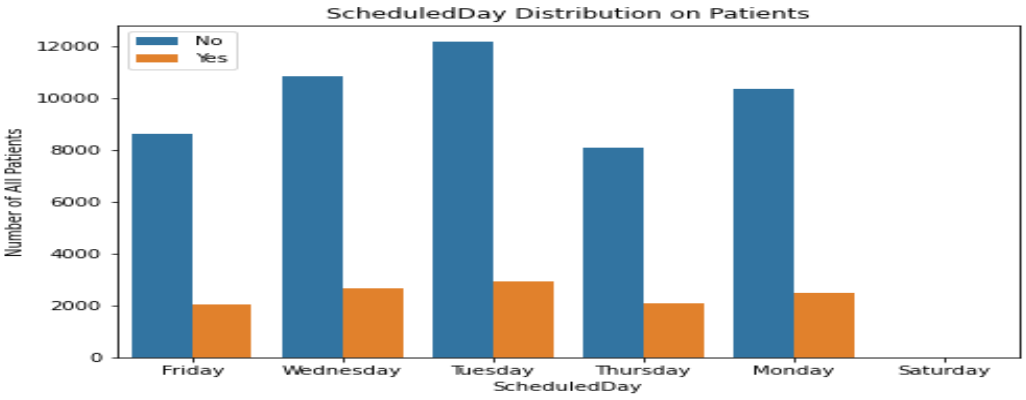

Diabetes Distribution on Patients
No
Yes
Number of All Patients
40000
30000
20000
10000
0
0
1
Diabetes

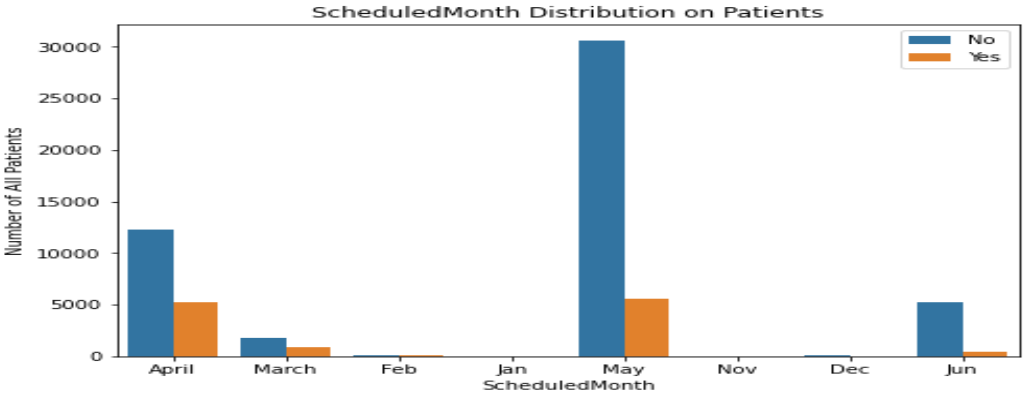

Hipertension Distribution on Patients
No
Yes
Number of All Patients
40000
35000
30000
25000
20000
15000
10000
5000
0
0
1
Hipertension

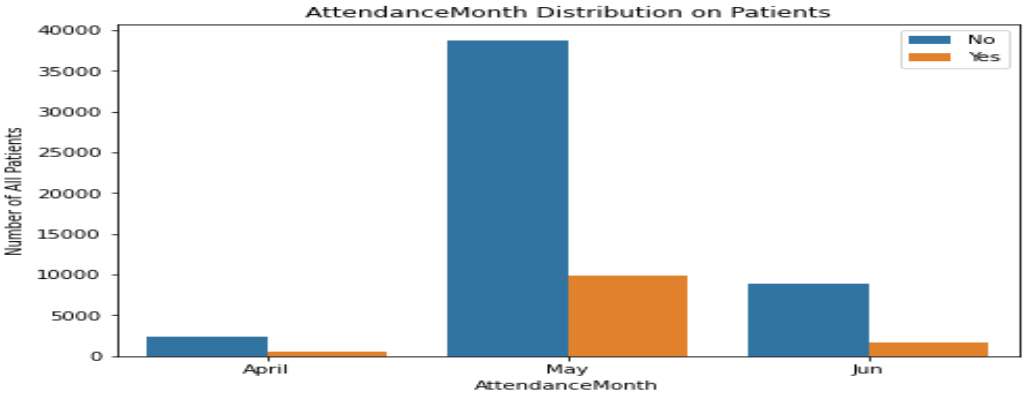
Finding 5-Most of cases are in the stages Aging, Maturity, and Adolescence and also most of them tends to show.
Conclusion 5-It is better to conduct an in-depth study on the fact that the disease is confined to a certain age group, it is possible to spread a specific disease that no one knows about.

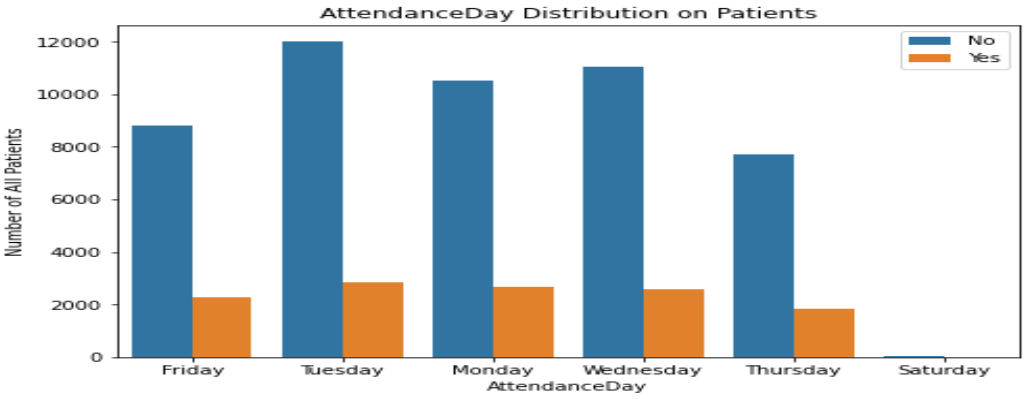


Finding 6-ost of cases prefer to attend and schedule at all days of week except Saturday and most of them are showing. Almost all of cases prefer to attend and schedule at the month of May and almost all of them are showing
Conclusion 6-For a lower cost, please close the days and months when attendance is not sufficient and refer to other nearby places for examination.


ScheduledDay Distribution on Patients
No
Yes
Number of All Patients
0
2000
4000
6000
8000
10000
12000
Friday
Wednesday
Tuesday
Thursday
Monday
Saturday
ScheduledDay


ScheduledMonth Distribution on Patients
No
Yes
Number of All Patients
0
5000
10000
15000
20000
25000
30000
April
March
Feb
Jan
May
Nov
Dec
Jun
ScheduledMonth


AttendanceMonth Distribution on Patients
No
Yes
Number of All Patients
0
5000
10000
15000
20000
25000
30000
35000
40000
April
May
Jun
AttendanceMonth


AttendanceDay Distribution on Patients
No
Yes
Number of All Patients
0
2000
4000
6000
8000
10000
12000
Friday
Tuesday
Monday
Wednesday
Thursday
Saturday
AttendanceDay

Limitations:

The data set includes mistakes data like age with negative value and ap pointment day before scheduled day.
Adding distances between the neighborhoods and the hospitable to the dataset may give good results on the neighborhood's impact .