Report on Health Care Appointment Data Set

Name: Milan Sood

Reg No: 12007271

Roll No: RK20RUB50

Medical Appointment Case Study

Exploring and getting insights of the data set having data of appointments of patients having different health related problems such as diabetes, hypertension etc.

Introduction

The recent raise in health related issues has made it fundamental that everyone one has easy access to health care but its been discovered that despite provision of access patient may or may not show up for their appointments leading to hospital incurring losses on operating cost as salaries will be paid whether or not patient visit the hospitals as expected. This report tries to analyze the reason why patients don't show up for their appointments after scheduling one and providing possible solutions the hospitals can consider in improving their revenue.

Dataset used for this project was collected from Kaggle. The dataset contains 110k medical appointments in Brazil, collected in May/June 2016 and is focused on characteristics of patients as well as their presence or absence from scheduled medical appointments. The dataset is made up of 14 unique variables.

Overview

Variables in this dataset includes (PatientId, AppointmentID, Gender, ScheduledDay, AppointmentDay, Age, Hipertension, Diabetes, Alcoholism, Handicap, SMS_received, Neighbourhood, Scholarship and No-show). This analysis will be done using the No-show variable as our dependent variable while other variables will be used as independent variables.

Identifying factors responsible for no-show will definitely help hospitals better determine types of scheduled appointments to plan for and provide incentives that enable patients to show up for their appointments.

- PatientId - Identification of a patient
- AppointmentID - Identification of each appointment
- Gender - Male or Female .
- AppointmentDay - The day of the actual appointment, when they have to visit the doctor.
- SchedulingDay - The day someone called or registered the appointment, this is before the appointment.

- Age - How old is the patient.
- Neighborhood - Where the patient is from.
- Scholarship - True of False .
- Hipertension - True or False
- Diabetes - True or False
- Alcoholism - True or False
- Handicap - True or False
- SMS_received - 1 or more messages sent to the patient.
- No-show - True or False.

This Project will be providing insights into the following questions:

- What percentage of people showed up compared to those that didn't?
- Why Are People Not Showing Up?
- Does scholarship affect a patient 's ability to show up?
- Does SMS alert impact patient availability for their appointment?
- Is there any relationship between distance from the patient 's neighborhood to the hospital and their showing up for appointments?
- What's the time difference between a patient's schedule date and his appointment date? What impact does this have on the patient's ability to show up?
- Is there any relationship between patient age and their ability to show up?
- Does gender play any role in patients missing their appointment?

Data Cleaning

Corrected Column names which are misspelled

```
#changing the name of some cloumns
base_data= base_data.rename(columns={'Hipertension': 'Hypertension', 'Handcap': 'Handicap'})
```

✓ 0.6s

The datatype of the column "ScheduledDay" and "AppointmentDay" was not absolute so changed its format to Date time format.

The values of "Scholarship", "Hypertension", "Diabetes", "Alcoholism", "Handicap" were 0 or 1 so I changed it to True and False.

```
# Converting Scholarships to SMS Columns values from 0 ,1 to Boolean
for each in ["Scholarship", "Hypertension", "Diabetes", "Alcoholism", "Handicap"]:
    base_data[each] = base_data[each].astype(bool)
```

Null Values Check

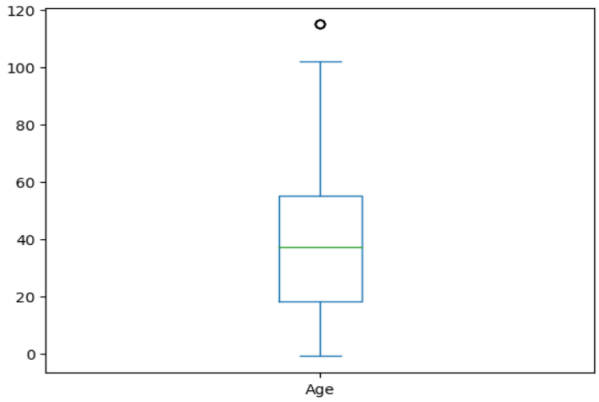
```
base_data.isnull().sum()
```

```
PatientId          0
AppointmentID      0
Gender             0
ScheduledDay       0
AppointmentDay     0
Age                0
Neighbourhood      0
Scholarship        0
Hypertension       0
Diabetes           0
Alcoholism         0
Handicap           0
SMS_received       0
No-show            0
dtype: int64
```

Hence, there are no null values in our data set.

Age

Checking for Outliers in column Age.
As seen from the above box plot there are some outliers in our data set.
After checking for quartiles we can see that there is one patient whose age is -1 and after inspecting more it is found that there are 4 patients whose age is 115.

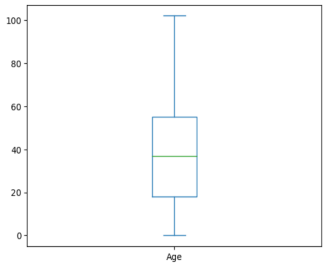



```
1.00    115.0
0.99     86.0
0.75     55.0
0.50     37.0
0.25     18.0
0.00     -1.0
Name: Age, dtype: float64
```

Hence, dropping these rows is the best option as the number of outliers is very less.

```
base_data = base_data[~((base_data.Age == -1) | (base_data.Age == 115))]
```

After removing the Outliers
Now we can see that there are no outliers in the Age column.

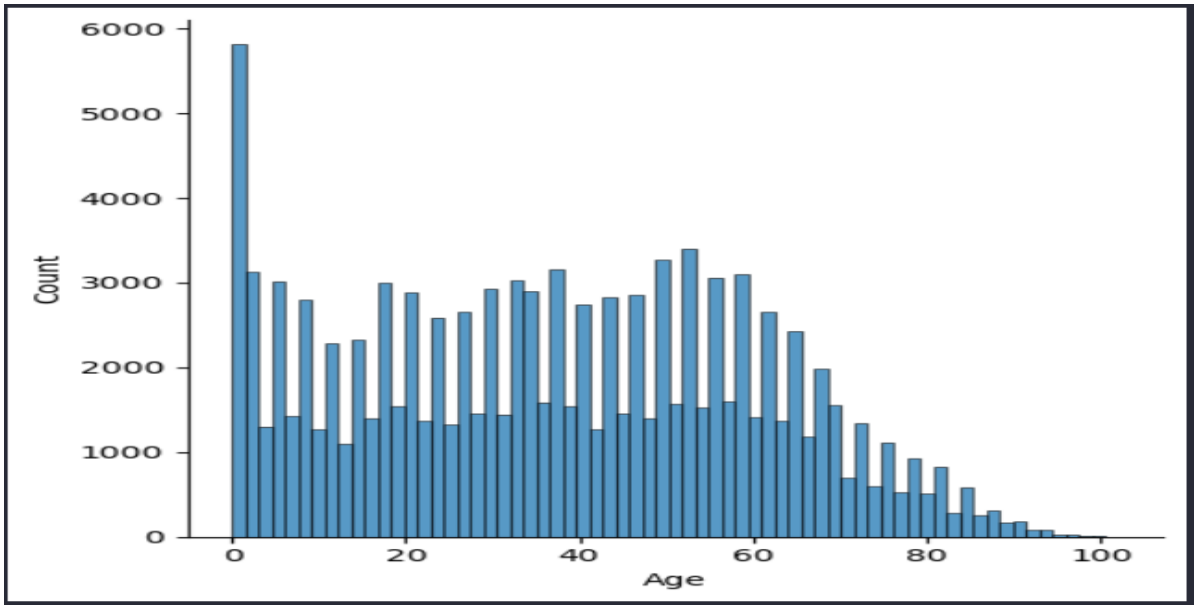


1.00	102.0
0.99	86.0
0.75	55.0
0.50	37.0
0.25	18.0
0.00	0.0

- The assumption that patients with 0 age as toddlers can be considered appropriate, as they received no scholarships and are not suffering from any diseases.
- There are 3539 toddler appointments in the data.

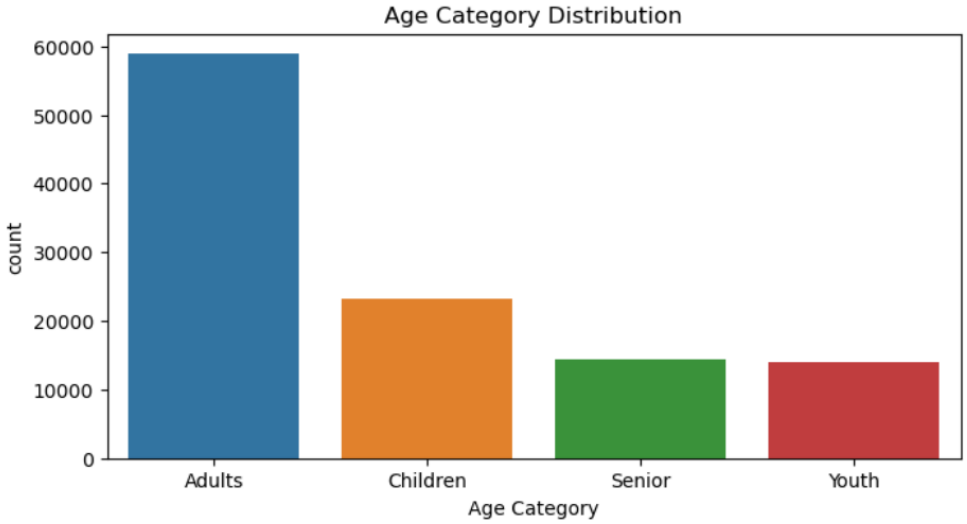
Exploratory Data Analysis

Distribution of Age



From here also we can see that most of the patients are toddlers.
Total count is 110521 where mean Age is 37 and max Age is 102 , 25% of patients are up to 18 years and 75% are about 55 years.

Creating a new column "Age Category" for better understanding the age distribution



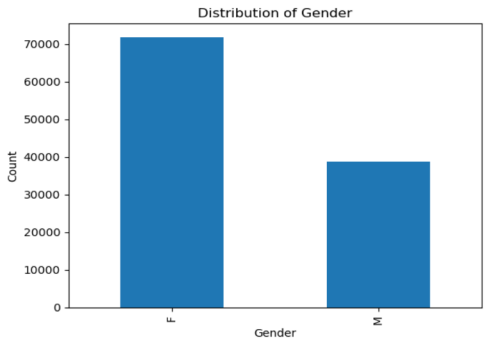
53% of the patients in the data are Adults, 21% of children,13% of Seniors and 12% of Youth.

Unique PatientIds and AppointmentIds:

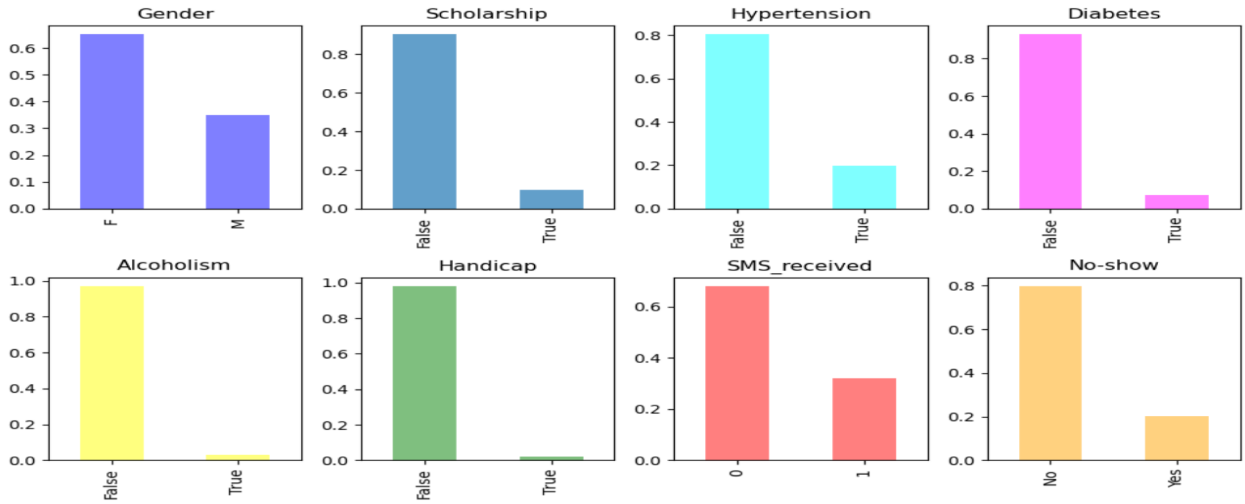
There are only 62299 unique patient ids while the number of unique appointment ids are 110527. This shows that 43% of the appointments are made by the recurring patients.

Distribution of Gender

There are 110527 appointments in total. 65% of the appointments are reserved by females and 35% are reserved by Male. This indicates that the probability of ill health is distributed disproportionately among the Male and Female.

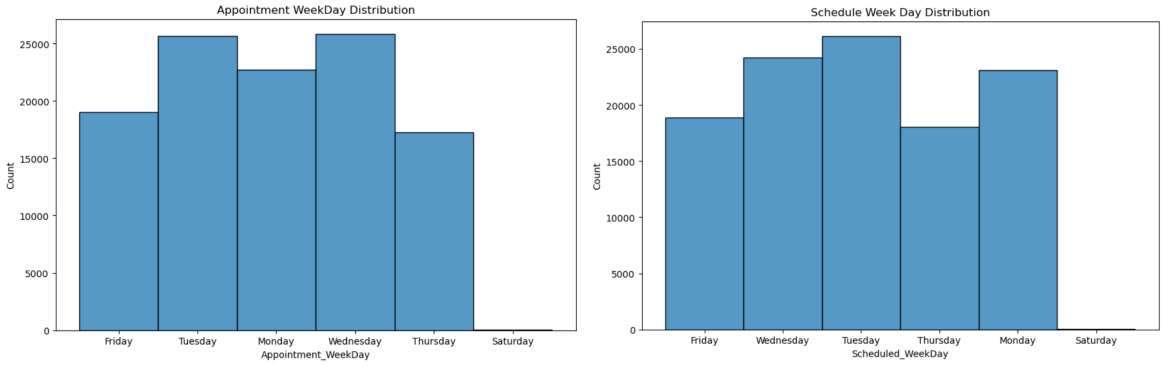


Categorical Variables



1. **Scholarship** : 85% of people who reserved appointments have no scholarship.
2. **Hypertension**, **Alcoholism** & **Diabetes** : 20% of the people are suffering from Hypertension. Less than 10% of people are suffering from Diabetes and Alcoholism.
3. **SMS_received** : Approximately 70% of the people did not receive any SMS alerts about the appointments.
4. **No-Show:** Only 20% of the appointments made turned out to be a no-show.

Extracting Day of the Week for AppointmentDay and ScheduleDay

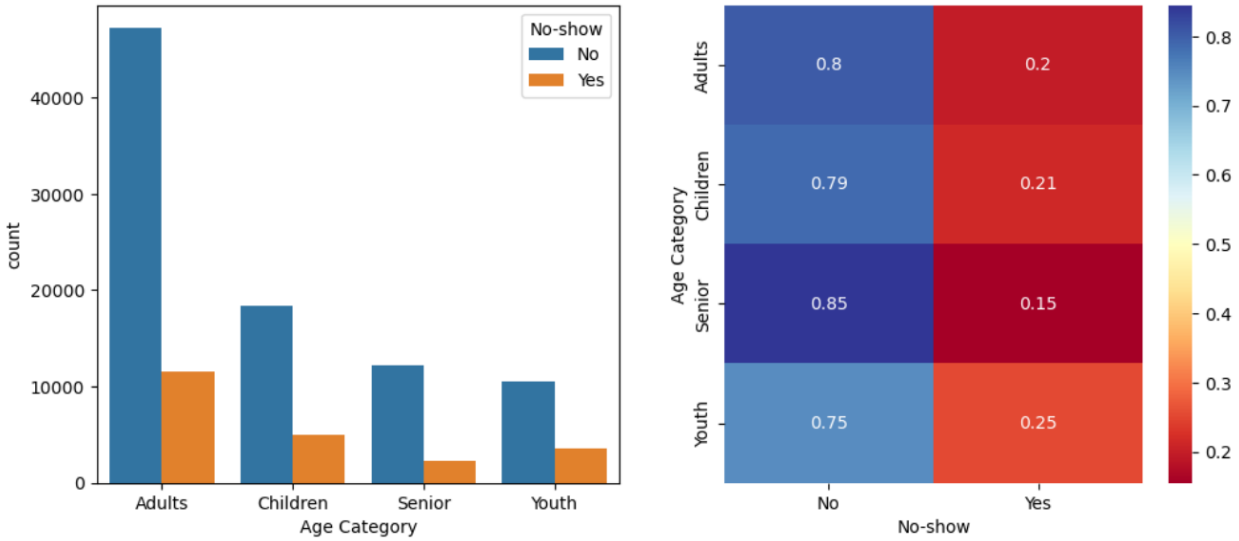


- The above visual shows that 46% of the appointments are reserved on Tuesday and Wednesday.
- The next favorable days for the appointments as observed above are Monday, Friday and Thursday.
- The least number of appointments are observed during Saturday, which is obvious as it is a weekend day.

Similar kind of trend is followed in Scheduled day as seen in Appointment Day. Highest number of appointments are scheduled on Tuesdays followed by Wednesdays and Mondays.

Bi-variate Analysis

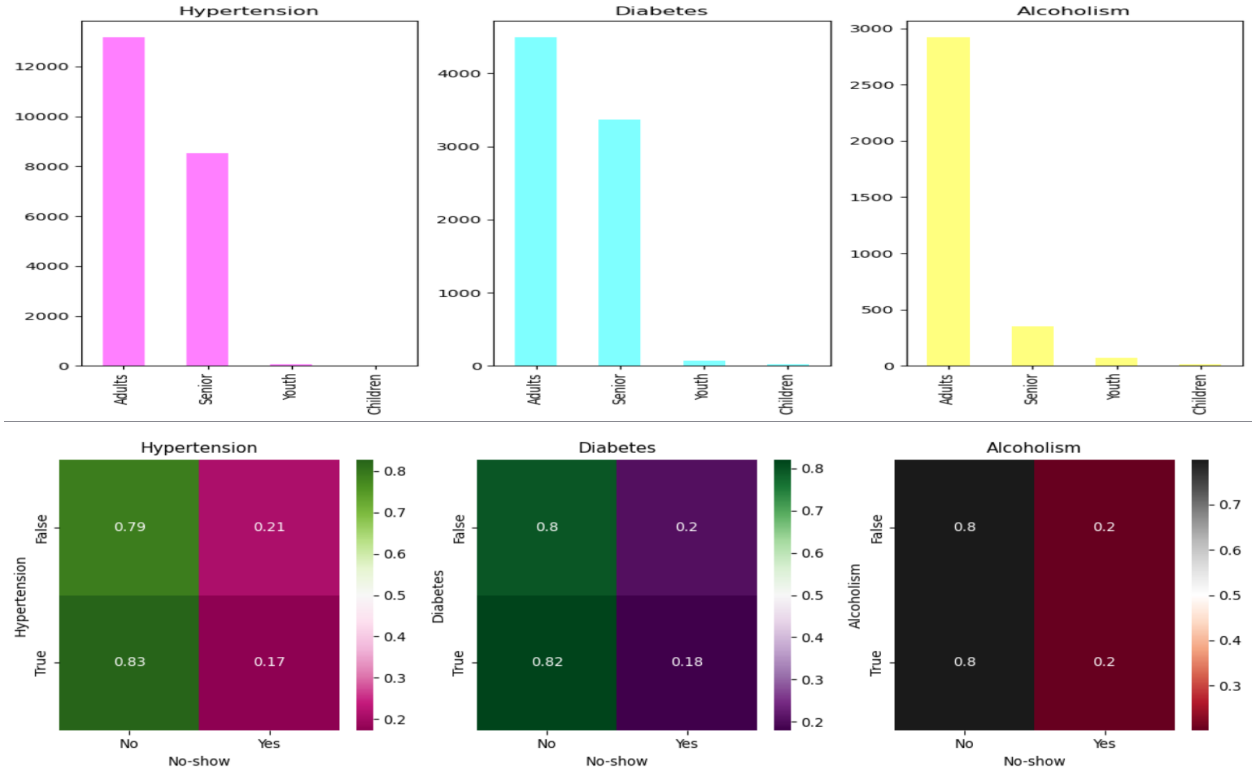
Age Category Vs No-Show



On observing the bar plot, we may infer that the highest number of No-shows are observed in Adults. But the percentage values on the left gives a clear idea of the situation. Out of 100 appointments booked by the Youth, 25 of them turn out to be NO-shows which is the highest in the age categories. Seniors are

among the age group with less percentage of No-shows. This proves that Age can be an important factor in determining the probability of a patient showing/not showing up for his/her appointment.

(Hypertension, Diabetes, Alcoholism) Vs No-show



Hypertension

- There are 88,000 patients not suffering from hypertension and 79% of the people are showing up for their appointment.
- Out of 22,500 patients with Hypertension, 83% of them show up for appointments.

Diabetes

- There are 102,000 patients not suffering from hypertension and 80% of the people are showing up for their appointment.
- Out of 8,500 patients with Hypertension, 82% of them show up for appointments.

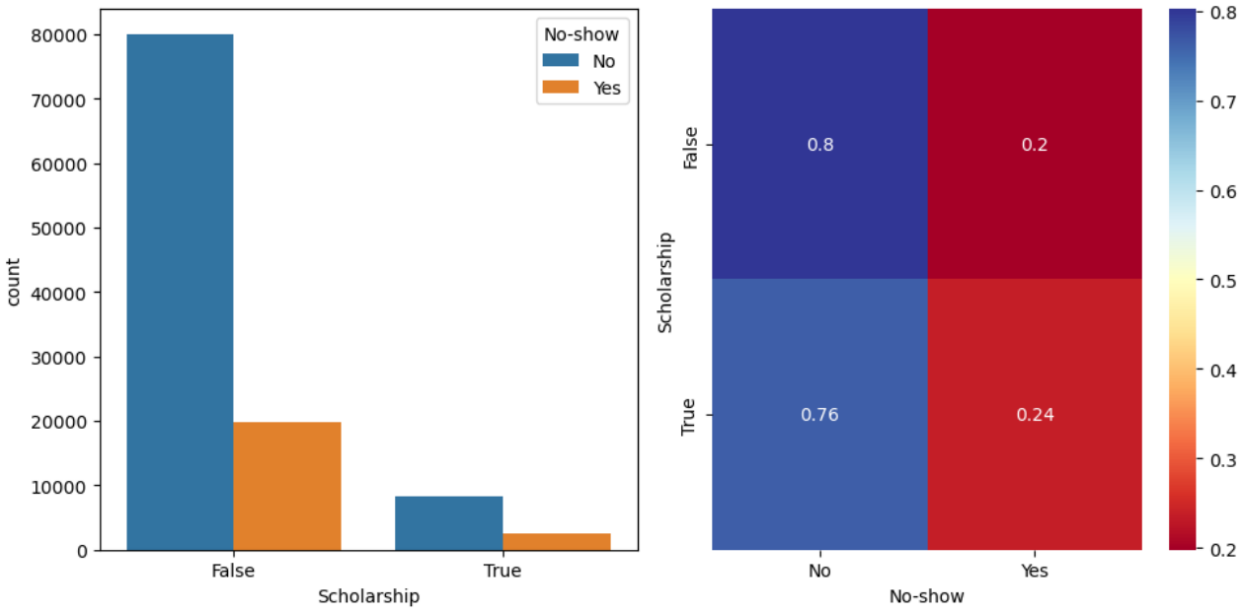
Alcoholism

- 80% of both the alcoholics and non alcoholics are showing up for their appointments.

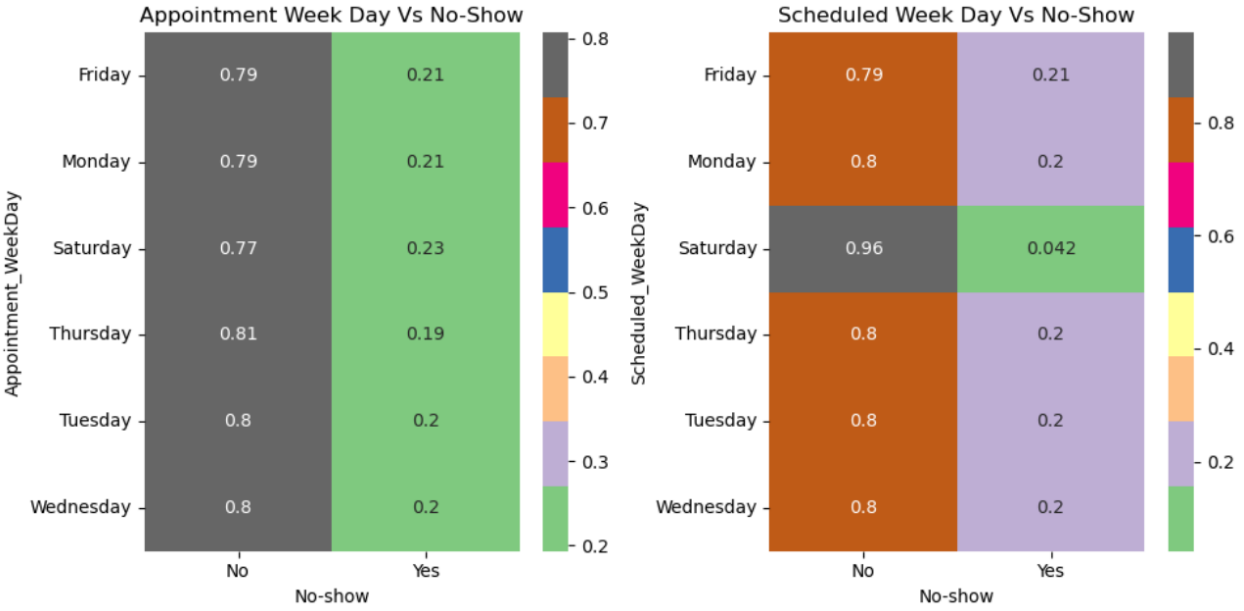
The above graphs clearly convey that the probability of suffering/not suffering from Hypertension or Diabetes has a significant effect on a patient's probability of showing up for their appointment. Whereas, Alcoholics or Non alcoholics have the same probability.

Scholarship Vs No-Show

- There are 100,000 patients with no scholarship and 80% of them are showing up for their appointment.
- Out of 10,500 patients with Scholarship, 76% of them show up for appointment.



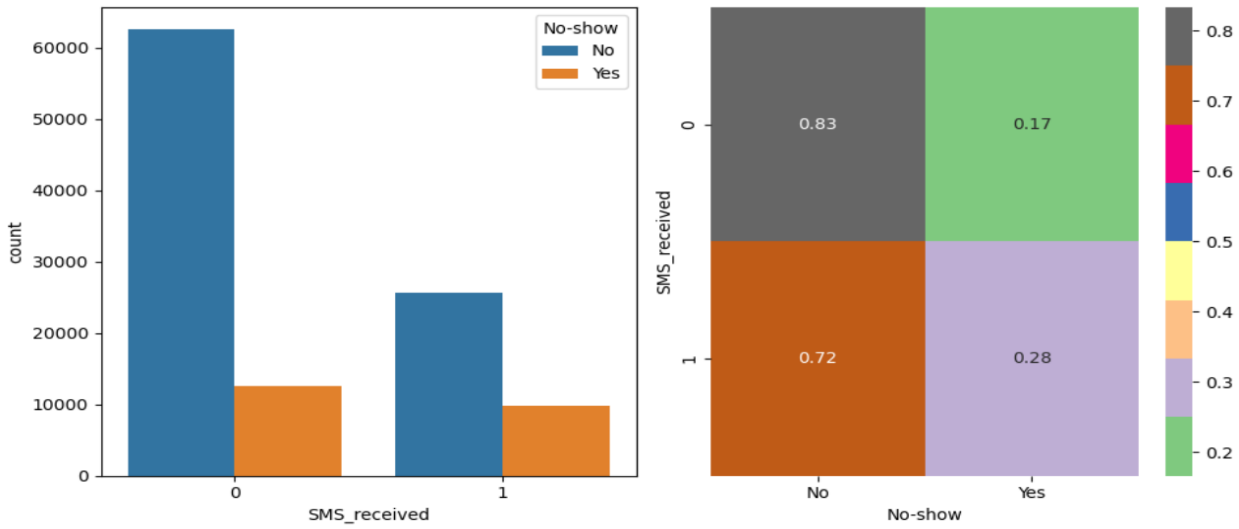
(Appointment Day,Scheduled Day) Vs No-Show



It looks like there is no significant relation between the weekday and the probability of the patient showing/not showing up for the appointment. But we can observe that the percentage of no shows is high during Saturdays. The number of days between Scheduled date and Appointment date could reveal more details.

SMS Received Vs No-Show

- There were 35,500 patients who received SMS and 72% of those patients showed up for their appointment.
- Out of around 75,000 patients who did not receive SMS, 83% of them showed up for appointment



Conclusion

1. From the above analysis, it is clear that Gender, Age, Neighbourhood, Scholarship,Hypertension and Diabetes are the factors that have a notable effect on the probability of No-show/show.
2. The factors that affect the absence of the patients more clearly are Gender and age which are the most important factors as we saw that females and youth show up for their appointment more than male and old people.
3. The patients with hypertension tend to show up if they have it or not. So we need to search for more factors to help patient remember their appointments and show up.
4. Scholarship has no impact on people's ability to show up for their appointments , this can be an indication that medical fees are fairly affordable.
5. Females are more health conscious than Males as they tend to show up on their appointment days.
6. Patient Age has a direct relationship on their ability to show up.
7. The no show frequency on Fridays is a little higher than average, but very comparable to the one on Mondays.
8. sending an SMS seems to be somewhat helpful in reducing no shows, but the correlation is very small.

Milan1508/HealthCare_Appointments_Data_Analysis: Exploring and getting insights of the data set having data of appointments of patients having different health related problems such as diabetes, hypertension etc (github.com)