Investigate a Dataset(NoShow appointments)

Introduction :

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. The main qeustion we are trying to answer here is why 30% of patients miss their scheduled appointment. We are trying to predict the most important factors that affect the atendance of the patient.

Data Wrangling :

Correct Incosistencies in Data

Below we will correct some of the inconsistencies in the data:

- PatientId is an Integer and AppointmentID is Float , but both don't have any numerical values they should be string but i will igonre them and drop them .
- Data Type of ScheduledDay and AppointmentDay will be changed to DateTime.
- Typo's in the Column names will be corrected
- As the AppointmentDay has 00:00:00 in it's TimeStamp, we will ignore it.
- As we removed the Time from AppointmentDay's TimeStamp we will do a similar thing for ScheduledDay also. (Ideally the Time in AppointmentDay column will help us better rather than in the ScheduledDay)
- there's row with Age = -1 which not make sense .

EDA :

1. Does the Age has an affect the Appointment ?

- the median is a 37 and the box plot is between 18 to 55
- age 0 and 1 is the most values in the dataset and they affect on the show that 0 and 1 most of them attend the Appointment but looking to other values we can't say that show affect by age
- all Age bins almost have same number of show , bin 13-29 have the greatest number in not attend

2. Does the gender has an affect the Appointment ?

- Females are the most appointments in dataset than males
- the distrubtion for both F & M vs noshow seem to be the same in the ratio
- When it comes to show up, there is no Significant difference between males and females regardless of: Agg , Being diabetic or not,Receiving SMS or not
- but we can say that gender affect the noshow

3. Does the Scholarship has an affect the Appointment ?

- only 9.8 % have Scholarship and 90.2 don't have
- most of patients with no Scholarship attend show almost 75% of them , so we can't say that Scholarship affect the show
- But when we investigated it further, we concluded that those who were diabetic or hipertension and received the scholaship had higher show up rates

4. Does the Hypertension has an affect the Appointment ?

- we can see that there are around 80% of patients without Hypertension and out of them around 70000 have come for the visit.
- Out of the 20% of patients with Hypertension and most of them have come for the visit.
- So, Hypertension feature could help us in determining if a patient will turn up for the visit after an appointment.

5. Does the Diabetes has an affect the Appointment ?

- we can see that there are around 92% of patients without Diabetes and out of them around 80000 have come for the visit.
- Out of the 10% of patients with Diabetes and most of them have come for the visit.
- So, Diabetes feature could help us in determining if a patient will turn up for the visit after an appointment.

6. Does the Alcoholism has an affect the Appointment ?

- we can see that there are around 99% of patients without Alcoholism and out of them around 80000 have come for the visit.
- Out of the 1% of patients with Alcoholism and most of them have come for the visit.
- As the percentage of visits for patients with and without Alcoholism is the same it may not help us in determining if a patient will come for a visit.

7. Does the Handicap has an affect the Appointment ?

- we can see that there are around 110,000 patients without Handicap and out of them around 80% have come for the visit.
- As we can see a clear distinction between different Handicap levels this feature will help us in determining if a patient will turn up for the visit after taking an appointment.

8. Does the Neighbourhood has an affect the Appointment ?

- It’s looks like the ratio of Show to NoShow is almost the same for all Neighbourhood's
- We see that some neighborhood have more people show up for their appointment and this indicates that this area have increase in disease

9. Does the Time has an affect the Appointment ?

- we can see that most of the patients are booking their appointments on the same day. The next highest waiting times are 2days, 4 days and 1 day.
- Looks like that takes the appointments doesn't work over the weekends as we do not see any appointments taken on Saturday and Sunday.

Conclusions

Now we can see the factors that affect the absence of the patients more clearly.

- The gender and age are the most important factor as we saw earlier that female and youth show up for their appointment more than male and old people.
- Neighbohood and hypertension come after gender and age as there are some neighborhoods that the diseases are spread and patients with hypertension tend to show up if they have it or not.
- So we need to search for more factors to help patient remmenber their appointments and show up.
- No Limitations in the dataset that null values & missing data and duplicated data are 0