Investigate a Dataset (No Show appointments)

Introduction:

This dataset collects information from 100k medical appointments in Brazil and is focused on the question of whether patients show up for their appointment. The main question 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 attendance of the patient.

Data Wrangling:

Correct Inconsistencies in Data

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

- Patient is an Integer and Appointment is Float, but both don't have any numerical values they should be string, but I will ignore them and drop them.
- Data Type of Scheduled Day and Appointment Day will be changed to Date Time.
- Typos in the Column names will be corrected
- As the Appointment Day has 00:00:00 in its Time Stamp, we will ignore it.
- As we removed the Time from Appointment Days' Time Stamp, we will do a similar thing for Scheduled Day also. (Ideally the Time in Appointment Day column will help us better rather than in the Scheduled Day)
- 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 effect 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 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.

3. 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.
- Neighborhood 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 remember their appointments and show up.
- No Limitations in the dataset that null values & missing data and duplicated data are 0