Investigating a dataset

Introduction

The aim of this project is to examine a dataset containing 110,459 medical appointments in Brazil to determine what factors may contribute to a patient missing their appointments. The dataset includes the following patient characteristics,

- PatientId: patient identification number
- AppointmentID: patient appointment number
- Gender: patient gender (M/F)
- ScheduledDay: date and time of the appointment scheduling
- AppointmentDay: date (no timestamp) of the appointment
- Age: patient's age
- Neighborhood: hospital location
- Scholarship: patient enrollment in a scholarship program giving financial aid
- Hipertension: patient diagnosis of hypertension
- Diabetes: patient diagnosis of diabetes
- Alcoholism: patient diagnosis of alcoholism
- Handcap: patient diagnosis of ableism
- SMS_received: number from a SMS text reminder service
- Show-up: whether the patient made the appointment

Based on the specified parameters, we can explore the following questions:

1. Is there a temporal relationship between the date (scheduled or appointment) and attendance?
2. Are there any direct relationships between the patient's medical conditions (such as hypertension, diabetes, etc.), their use of SMS services, or their gender with their appointment attendance?
3. Is there a particular age group that is more likely to miss their appointment?
4. Is there a specific neighborhood where patients are more likely to miss their appointment?

Cleaning Data: `(1_clean_data.py)`

The initial stage of data cleaning involves examining for null and duplicate entries, which found none. Although approximately 4% of the data comprises duplicate patient identification numbers, no identical rows or appointment numbers were found. As every appointment number is unique and probably generated systematically by each hospital, we can discard this column.

To enhance our analysis workflow, we will rectify misspellings and naming inconsistencies by renaming columns. Furthermore, we will modify the data types of the following columns.

- PatientId: string -> integer
- Gender: sting -> bool (True if Male)
- ScheduledDay: sting -> datetime (note: the goal of changing to datetime before
- AppointmentDay: sting -> datetime saving to csv was to ensure uniform formatting)
- Scholarship: integer -> bool (True if 1)
- Hipertension: integer -> bool (True if 1)
- Diabetes: integer -> bool (True if 1)
- Alcoholism: integer -> bool (True if 1)
- Handcap: integer -> bool (True if 1)
- SMS_received: integer -> bool (True if 1)

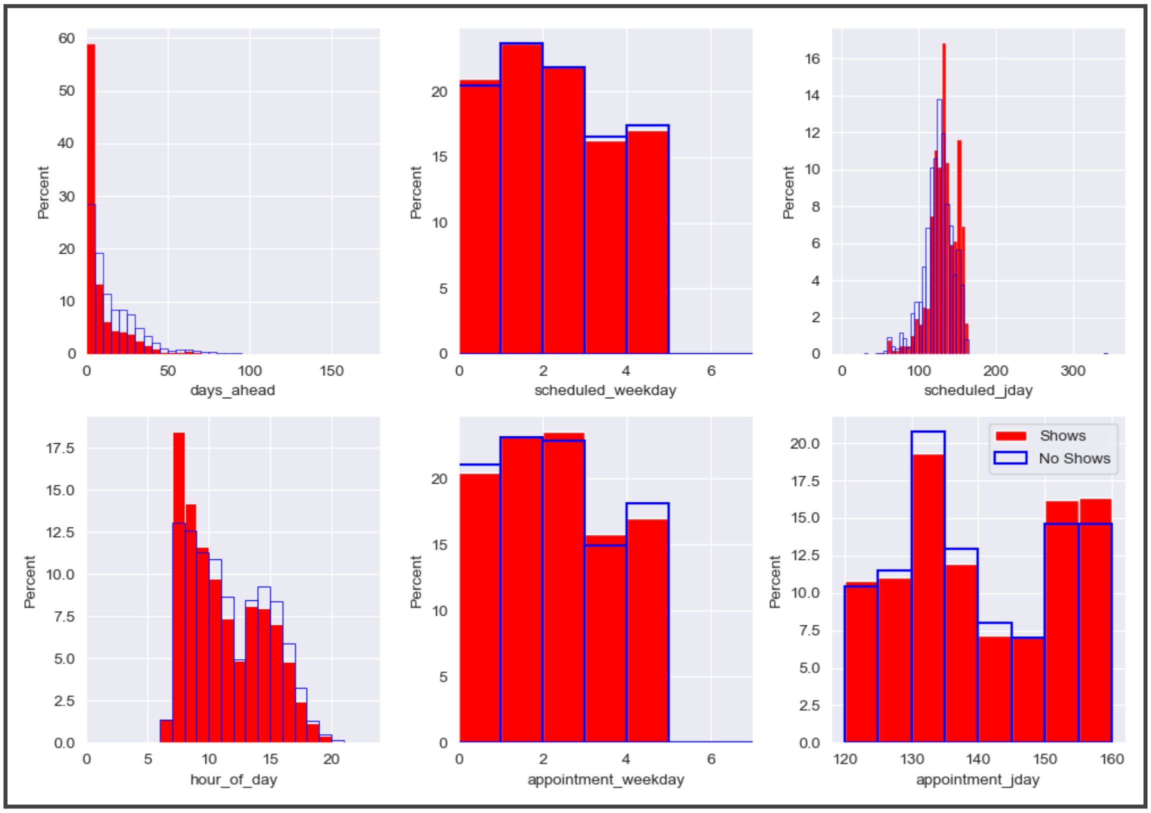
Finally, the dataset has been divided into two subsets - one consisting of patients who attended their appointment (“Shows”) and the other comprising patients who did not (“No Shows’).

Exploring Data `(2_explore_data.py)`

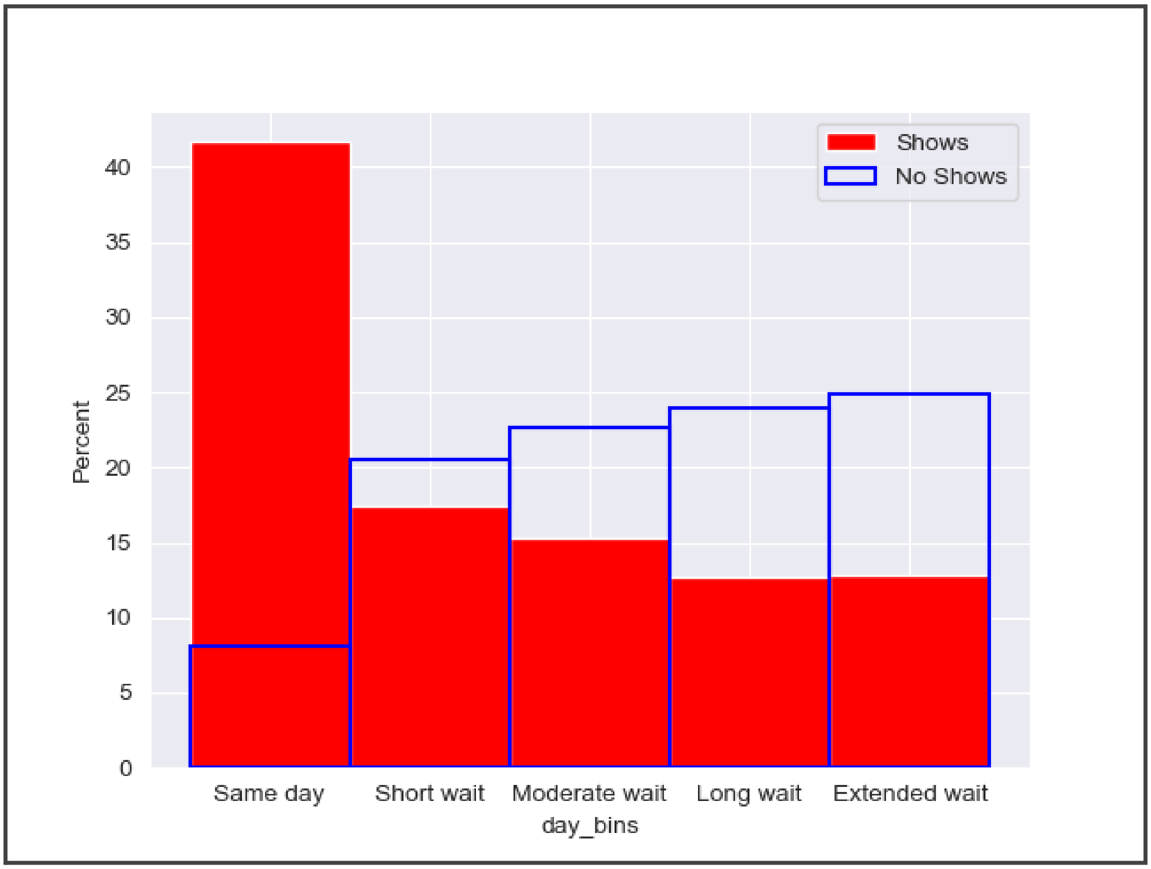
With the data separated into “Shows” and “No Shows’, we can see that the “No Shows” constitute 25% of the original, combined dataset. Next we will revisit the four proposed questions to find potential relationships in the dataset.

1. Temporal relationship

The range of the appointment dates span April 29, 2016 to June 8 2016, while the scheduled dates range November 11, 2015 to June 8 2016. Five entries include a scheduling date listed after the appointment date, which will be dropped from the following analysis. Some potential relationships to consider are the schedule/appointment weekday (i.e., scheduled_weekday and appointment_weekday), schedule/appointment julian day (i.e., scheduled_jday and appointment_jday), scheduled hour (i.e., hour_of_day), and days between scheduling and appointment (i.e., days_ahead). Below we plot the relative distributions of each subset for these six parameters.

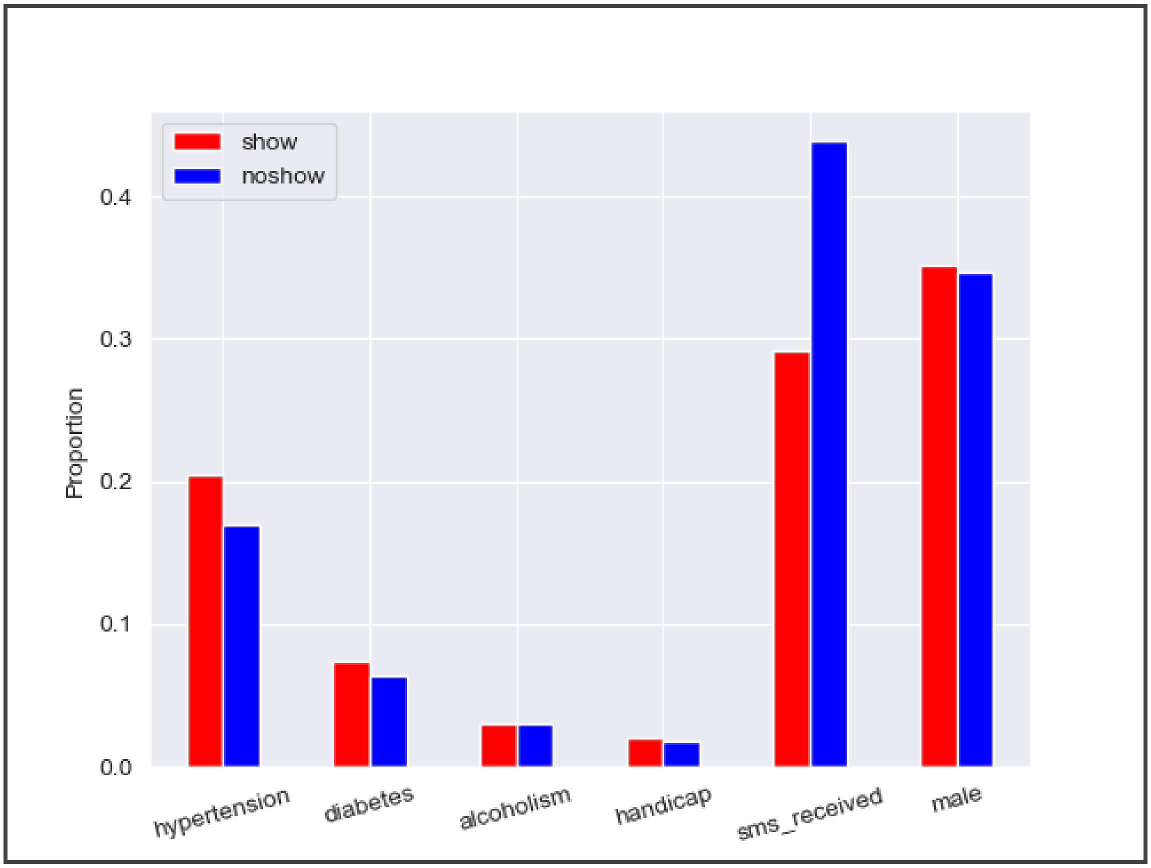


From the two rightmost graphs above, we can observe that a higher proportion of "No Shows" were scheduled in April, and slightly higher percent of appointment dates missed in May. The middle two graphs show that the proportions of scheduling weekdays are nearly identical in the both separated subsets, while appointment weekdays indicated a possible increase of missed appointments on Mondays and Fridays, relative to the "Shows" subset. In the lower left, the scheduling hour of the day indicates an increased possibility for missed appointments if scheduled later in the day. Lastly in the upper left, we observe that nearly 60% of the "Shows" subset have appointments scheduled the same day, whereas the "No Shows" appear to increase as days increase. When rebinning these days of separation based on the quartiles of "No Shows", plotted below, we see a positive correlation of missed appointments with increased wait times as well as a negative correlation in the "Shows" data.



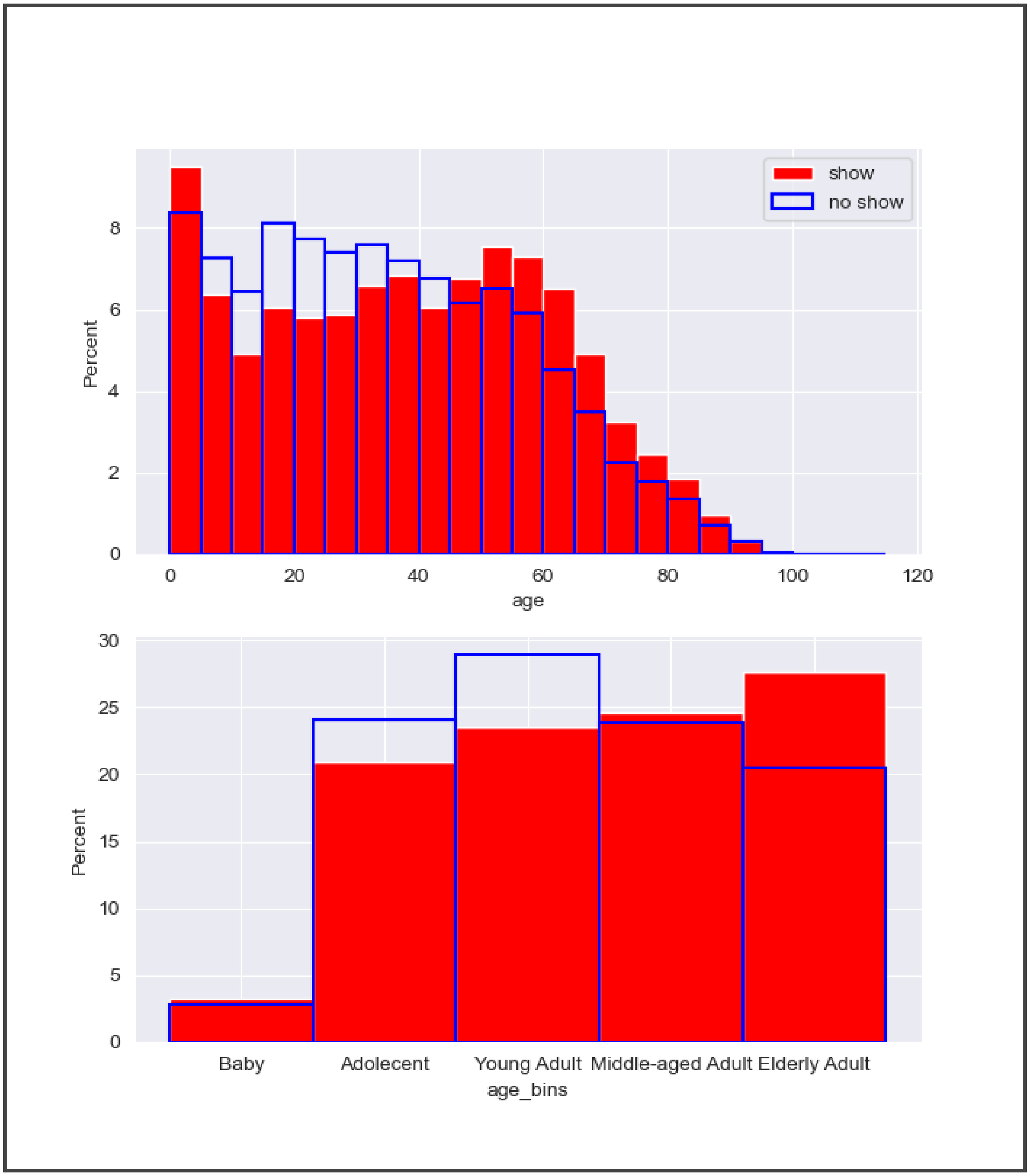
2. Patient attributes

Next we consider the proportions of patient attributes in both datasets, plotted below. The only attribute that appears to be associated with the "No Shows" data is paradoxically the SMS texting system.



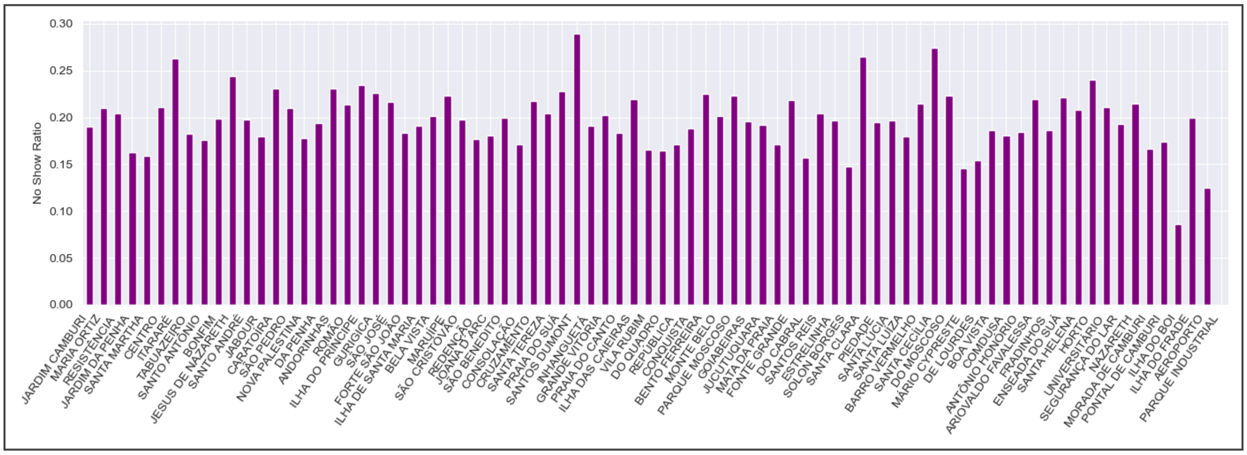
3. Age groups

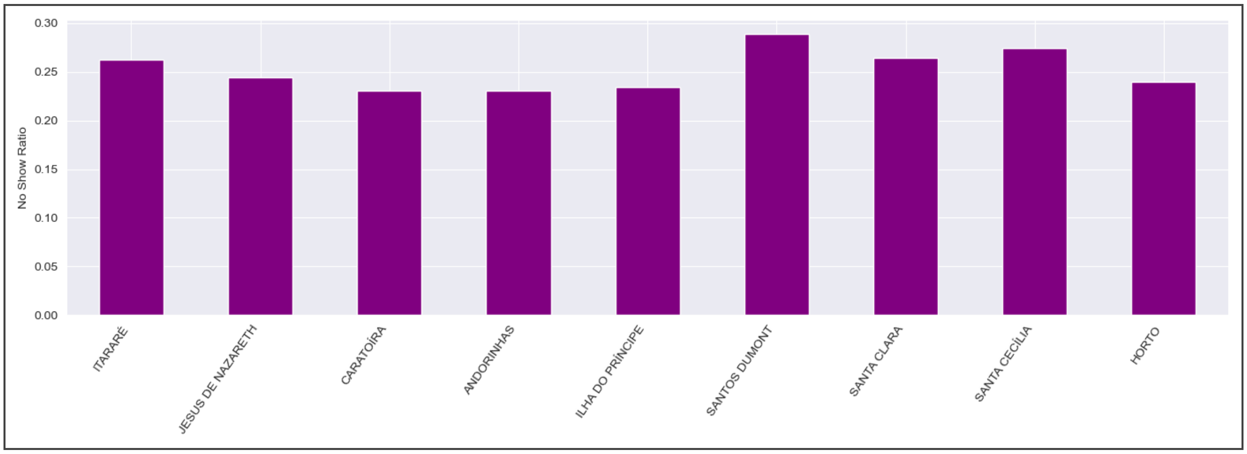
Looking at distribution of patient's ages, plotted below, we observe a higher portion of missed appointments from younger patients. After binning the age data into the "No Shows" quartiles, plus babies, we see that adolescence and young adults miss appointments more than their older counterparts.



4. Neighborhoods

In the neighborhood data, we see the average ratio of missed appointments lowers to 20%, with nine neighborhoods having ratios greater than one standard deviation about the mean (ratio > 19.80 + 0.03). Those high-ratio neighborhoods are also plotted below.





Drawing Conclusions

In conclusion, the raw dataset contained no null or duplicate data but during the data exploration some entries listed scheduling dates which followed the appointment date, thus were dropped. Some data hygiene was performed to make columns more uniform and easier to access.

In the exploration, we set out to explore the following questions,

1. Is there a temporal relationship?
2. Are there any direct relationships with the patient's attributes?
3. Is there a particular age group associated with missed appointments?
4. Is there a specific neighborhood associated with missed appointments?

Our analysis found a direct correlation between the rate of missed appointments with increasing days separating scheduling and appointment dates. Additionally, patients that received a SMS text reminder had missed appointments at a higher rate than those who had not received a reminder; however, this is likely indirectly related to the previous observation. Younger patients appear to miss their appointments more often than older patients. And lastly, there were nine neighborhoods associated with higher ratios of missed appointments.