Project: What factors influence a patient to show up for their scheduled appointment?

1. Dataset Analysed

For this project the dataset analysed was the **No-show appointment** available in this link.

Transcribing the dataset description:

" *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. A number of characteristics about the patient are included in each row.*

- *''ScheduledDay' tells us on what day the patient set up their appointment.*
- *'Neighborhood' indicates the location of the hospital.*
- *'Scholarship' indicates whether or not the patient is enrolled in Brasilian welfare program Bolsa Família.*
- *Be careful about the encoding of the last column: it says 'No' if the patient showed up to their appointment, and 'Yes' if they did not show up*."

2. Questions about the dataset

The main question raised about the dataset is:

- **What are the factors that influence a patient to not show up in a medical appointment?**

Other questions can be raised, such as:

- **What hospitals have a higher number of patients not showing up?**
- **Is the SMS communication effective to decrease the no-show rate?**

3. Walkthrough and Data Wrangling

To answer the questions, first, it was necessary to understand if all the required information was available in the dataset and with quality.

Therefore, an extensive data quality analysis was done to identify the data type of each field and discard impossible and null values.

Some fields, such as day of the week and number of days between the appointment and the scheduled date were created from their original fields already available in the database.

After all the quality treatment, it was checked the amount of unique records in the class fields (hospitals, gender, etc.) and the statistics of the numeric fields,

because depending on the size of the content, some adjustments would be necessary or a different approach would be required.

One example is the **handicap** field. Checking its content we see that there are 4 values, but there is no explanation of what each value means, thus, we replaced the records different of 0 to 1, turning this field into a Boolean one.

After all the treatment, the dataset was divided in two groups: “Show” and “No Show” and from there all the features were analysed.

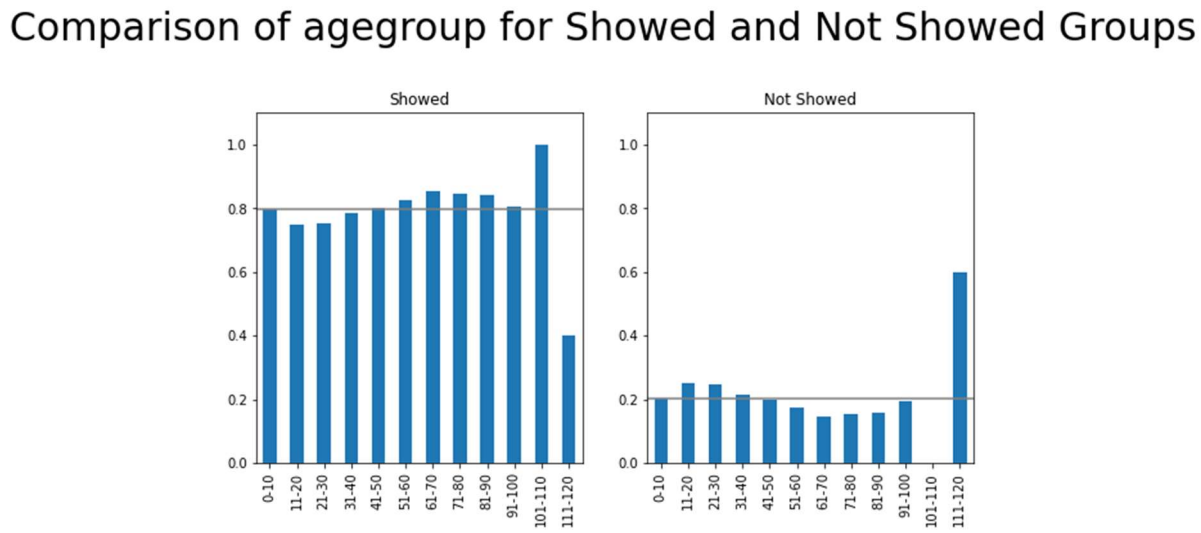
In order to have a reference of what is a normal no show rate, it was calculated the proportion of now show patients to the total, and recorded this variable as the reference value for all the class features. Anything above this reference value is a potential feature to influence the no show.

As the proportion of both groups is not the same, we plotted the proportional value in relation to the total dataset. This is well defined and described in the functions created for this purpose.

For the numeric fields, it was used a similar approach, overlapping both groups in a histogram chart to identify potential influences.

4. Final Results

All the results were plotted and explained along the **.ipynb** file, but among all the features analysed, we can see that the age group, the neighbourhood and the appointment day are the most relevant ones, following the above the rate logic. See the chart below:



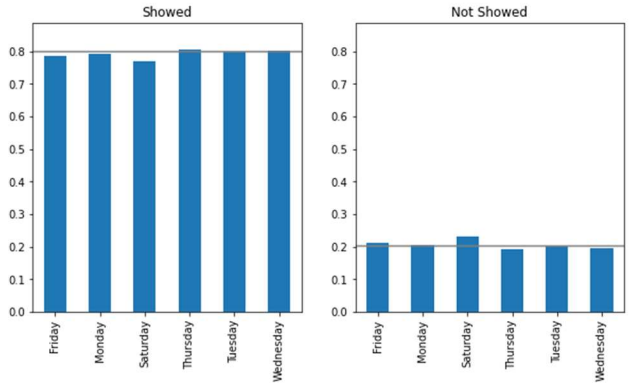
The horizontal line represents the no-show rate of the entire dataset, so the age group between 11-40 years old are among the groups that have not showed the most.

It makes sense if we consider that these people are in the labour age, so, an additional question should be included in the dataset: “Full time employed?”.

The group above 100 years old should be considered as an outlier because they are not representative in the dataset.

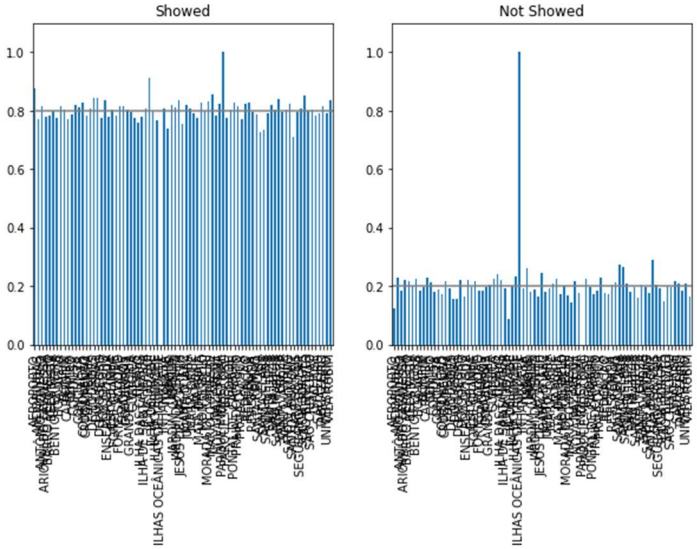
Considering the day of the week, Fridays and Saturdays are the days that are above the rate as seen in the plot below:

Comparison of appointment_weekday for Showed and Not Showed Groups



In total, there are 81 hospitals to be analysed, and very few information is provided related to their distances to the patient houses or the kind of services in the neighbourhood. This could be a reason to affect some hospitals, as seen:

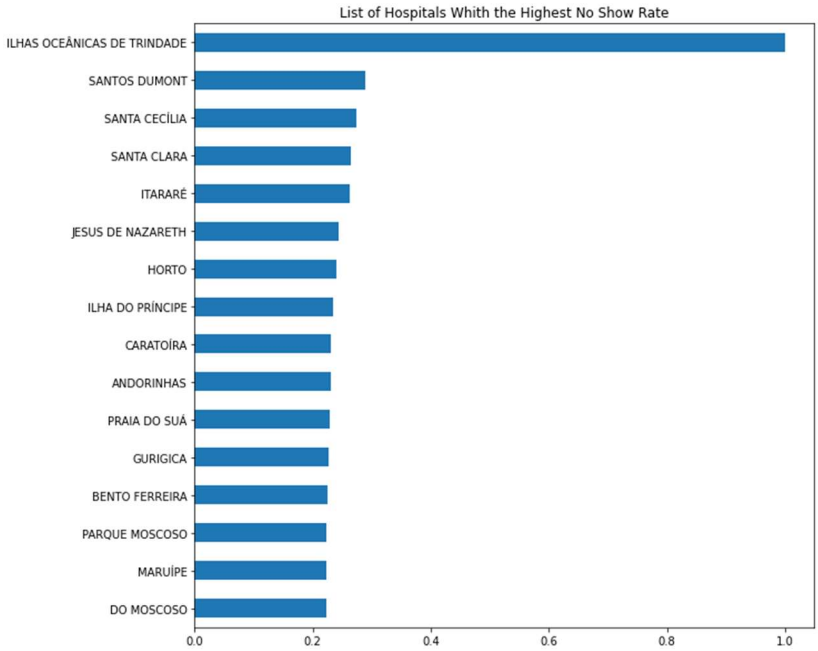
Comparison of neighbourhood for Showed and Not Showed Groups



There are a couple of hospitals that are above the no-show rate, and it is impossible to read the chart, so I selected only those hospitals that are above the rate in 10% to narrow the results to only the worse places in terms of no-show rate.

If there is a possibility to communicate the authorities, a closer look would be recommended to these institutions to find the root cause of the no-show. And for that more information should be collected

The institutions are listed below:



To finish up, the sms system seems to be doing the opposite of help, if analysed individually, the group that didn't show up is bigger with the people that received the sms compared to those that did not receive it.

Other questions should be raised here: How are they checking it? Is the message clear enough? With what anticipation it arrives to the people?