Investigate a Dataset Project

(No-show appointments)

Abdulrahman Hatem Metawally

Jan 2022

Contents

Introduction

This project developed for Udacity Data Analyst Nanodegree.

The project was about analyzing a specific dataset using the Python libraries NumPy, pandas, and Matplotlib, and then communicating findings about it as it will be shown in this paper.

The dataset which I analyzed is (No-show appointments) from Kaggle,

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. You can know more details from Kaggle.

Steps Details

Here we know that this dataset focused on the question of whether or not patients show up for their appointment, so I tried to find any relation between the characteristics there and this question answer, but before that I had to make sure that the data is ready by cleaning it, so I searched for:

1. Duplicates, and I didn't find any
2. Null values, and I didn't find any
3. Unlogic values, and I found some of them in 'Handcap' and 'Age' as outliers, so I removed them.

Then, I removed unnecessary columns like: 'PatientId' and 'AppointmentID' as they not important in analysis.

After that I divided the main table into two tables:

- **showed**: for patients who showed up their appointment.
- **not_showed**: for patients who didn't showe up their appointment.

Then I started to ask questions about this dataset:

- Can we predict if a patient will show up for his scheduled appointment or not?
- If yes, what are the characteristics correlated with showing up for scheduled appointment?

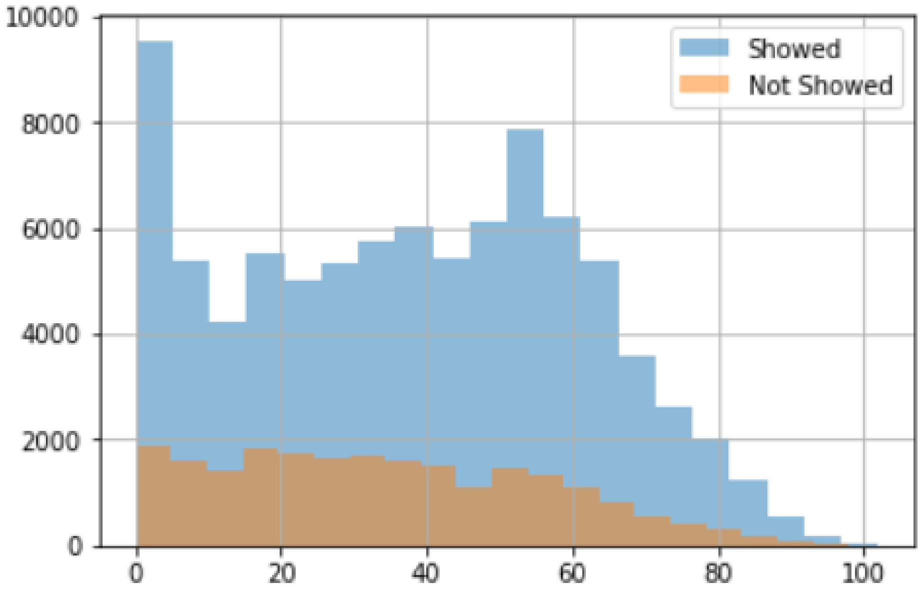
These were my main questions to analyze this dataset.

So, I started to find any relation between the most of characteristics and showing up the appointment. And here are the details:

Age

I plotted the relation between age and number of showed up and not showed up the appointments from patients and I found the following:

- Number of Patients above 60 much less than the rest of patients in this dataset. And the two graphs are approximately seeming to be **right skewed.**
- Children aged 0-5 years and adults aged 50-55 years have the most showing up appointment rate.

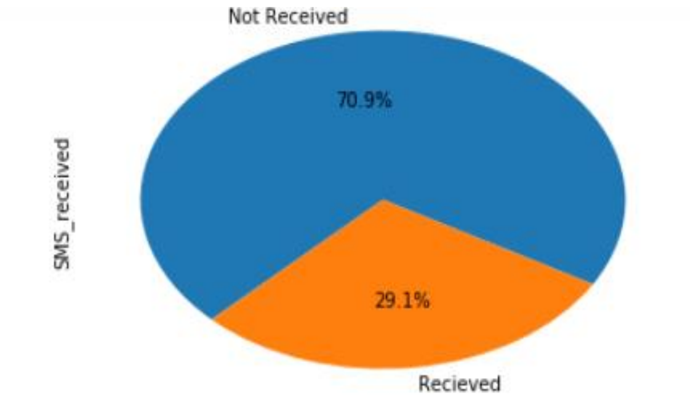


SMS Received

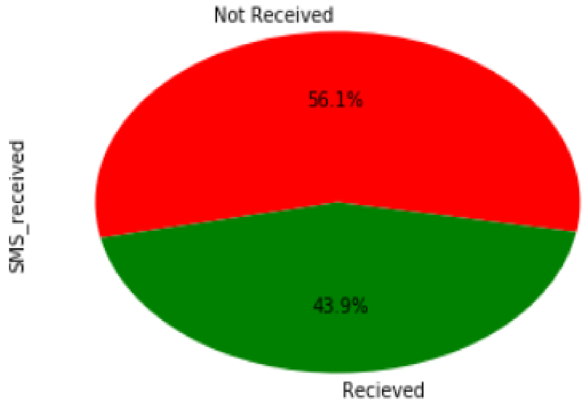
The relation between receiving SMS and number of showed up and not showed up the appointments from patients is as following:

- Altogether, number of not received SMS is **less** than received.
- For the patients who didn't show up their appointment the percentage of SMS received is **higher** than that for patients who showed up their appointment, and this was a bit strange info.

Showed Up



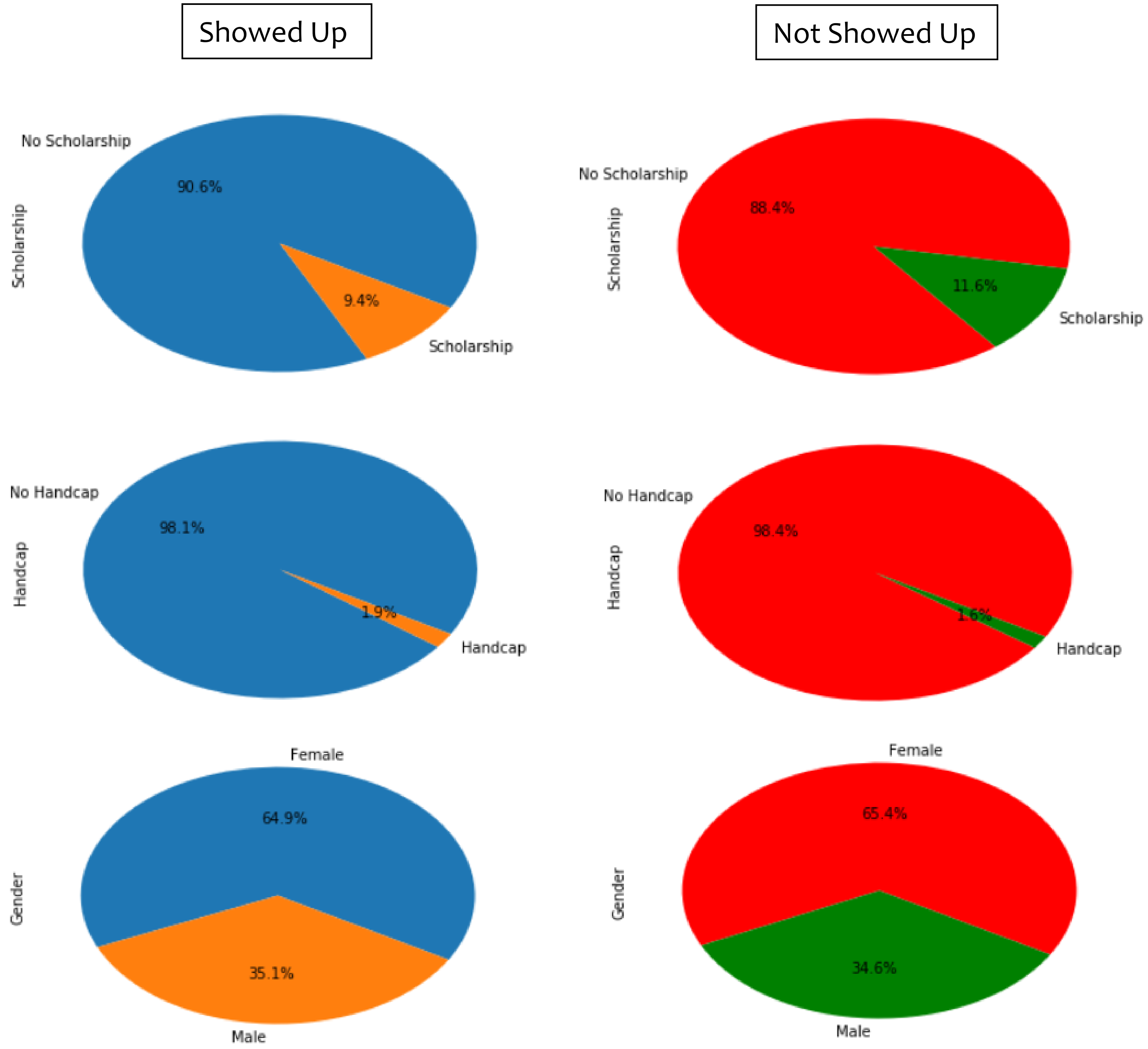
Not Showed Up

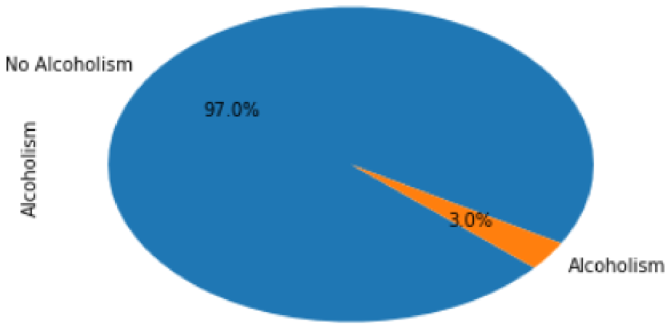


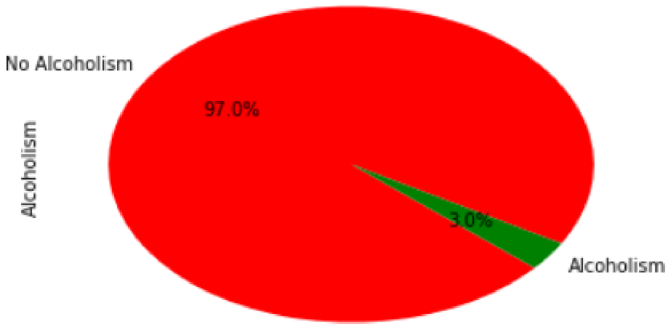
Other Characteristics

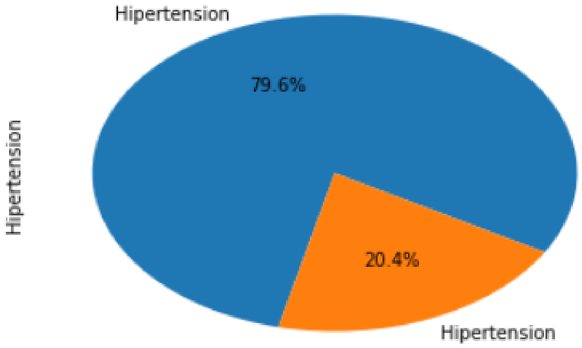
The relation between other characteristics and number of showed up and not showed up the appointments from patients is as following:

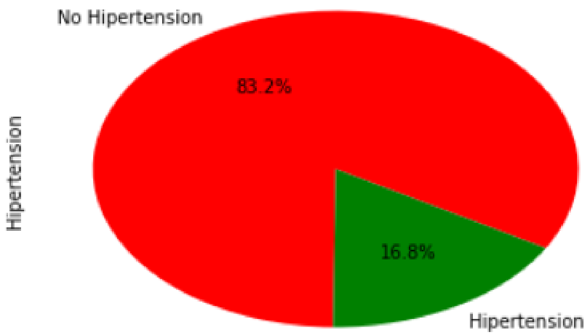
- Altogether, the ratio between the percentage of these characteristics in patients who showed up their appointment and who did not **approximately the same** or very close to each other.

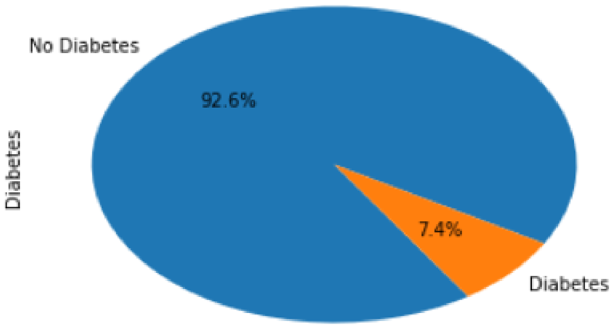


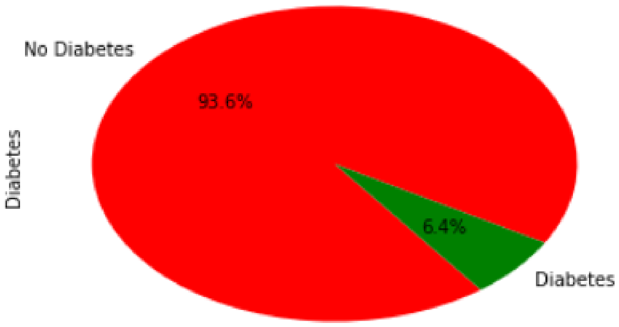

No Alcoholism
97.0%
Alcoholism
3.0%
Alcoholism


No Alcoholism
97.0%
Alcoholism
3.0%
Alcoholism


Hipertension
79.6%
Hipertension
20.4%
Hipertension


No Hipertension
83.2%
Hipertension
16.8%
Hipertension


No Diabetes
92.6%
Diabetes
7.4%
Diabetes


No Diabetes
93.6%
Diabetes
6.4%
Diabetes

Summary and Conclusions

- Altogether, number of patients who showed up their appointment are **much higher** than patients who didn’t show up their appointment.
- ‘**Age**’ and ‘**SMS_received**’ are the most characteristics affecting the appointments showing up.
- For ‘**Age**’, children aged 0-5 years and adults aged 50-55 years have the most showing up appointment rate. So, we can predict that patients at this age are more likely to show up their appointments.
- For ‘**SMS_received’**, for the patients who didn’t show up their appointment the percentage of SMS received is **higher** than that for patients who showed up their appointment, this means that not receiving SMS is better than receiving it as receiving SMS makes the probability of not showing up the appointments higher, and this was a bit strange info, and I think we shouldn’t make prediction based on this parameter as this parameter does not classify and predict the behavior precisely due to the closeness of the ratio.
- Finally, we do not have a specific parameter able to determine if the patient will show up his appointment or not.
- I think that the patient’s attendance for his appointment is more related to the culture related to education and the environment, most likely the prepayment for booking the appointment will be influential, likewise for the patient’s health condition, does he seriously need to visit the doctor or not, and these data were not present in this dataset.