Investigate a dataset
By Emil Delvaux

For this project, I analysed the 'No-show appointments' dataset.

Question: Which features are important in determining whether a patient showed up to their appointment or not?

Data wrangling

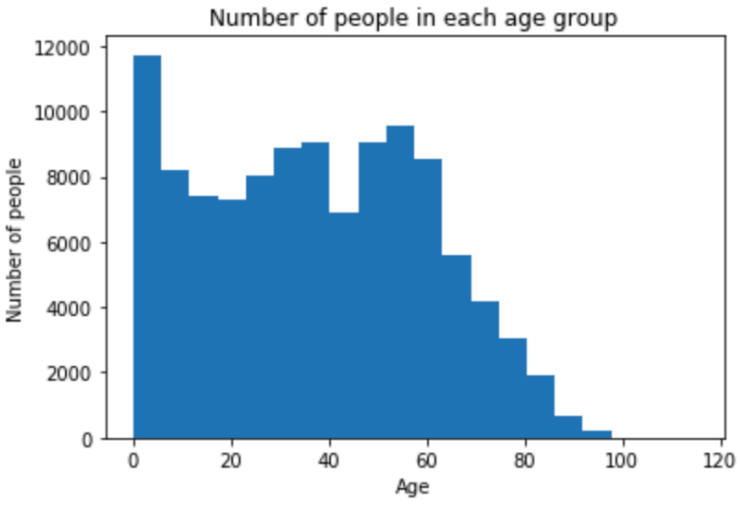
My first step in data wrangling was checking for missing values and duplicated values in the data, of which there were none, so I did not have to remove any rows. Next I used the describe function on the dataset to give me a brief overview of the data. There were a few things that stood out to me when doing this, firstly, I noticed that there was an individual with an age of -1, which is not possible, so I removed that patient from the dataset. I also noticed that the max age was quite high at 115, however since this is technically possible I did not remove this value.

I also noticed that the Handcap column had a max value of 4, when it should have been 1. I knew the max had to be 1 as Kaggle provided a breakdown of the data and said the Handcap column was a boolean, therefore the only two possible values would be 0 (False) or 1 (True). There were also some rows with a Handcap of 2 and 3, so I removed these from the dataset.

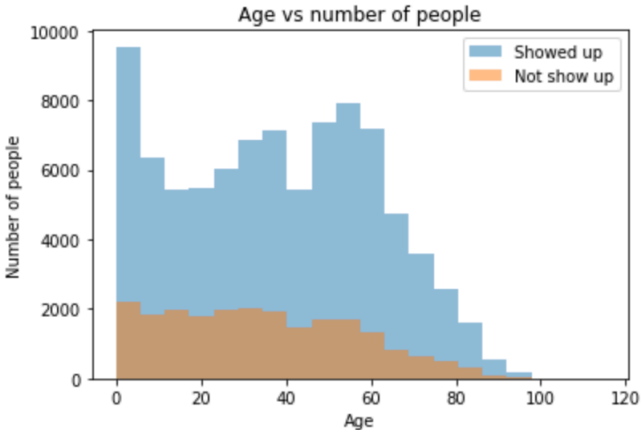
Next, I checked the datatypes of the data and noticed that ScheduledDay and AppointmentDay were saved as Objects, so I had to convert them into DateTime. Because the values were saved in a yyyy-mm-ddThh:mm:ssZ format, I was not sure how to convert them into DateTime without the time being present, so I used the split method to split the object on the 'T' chacter, and saved the date into one column and the time into another column. I then dropped the time column as I did not need it.

Lastly, I changed the 'No-show' column as it was confusing. I renamed the column to 'Show_up' and replaced 'No' with 1, and 'Yes' with 0. This way, the column represented a 1 if the patient showed up, and a 0 if the patient did not show up.

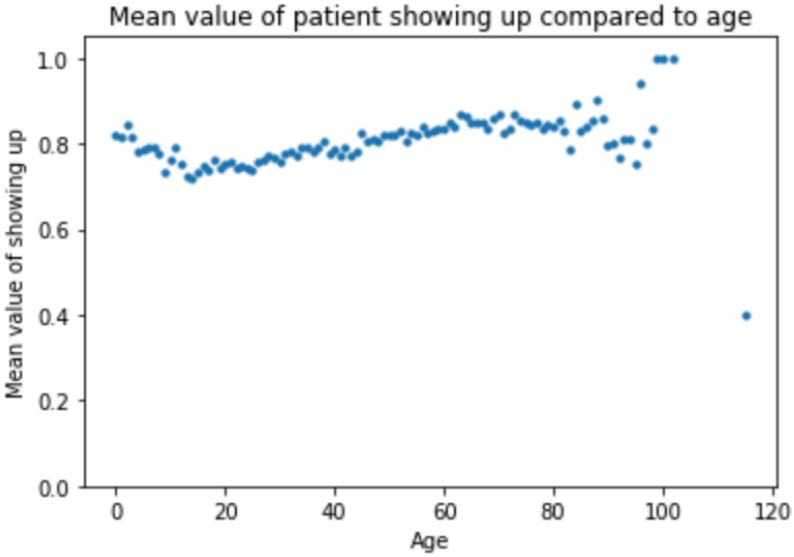
Data exploration



My first piece of data exploration was a simple histogram showing the age distribution of all people involved in the dataset. This showed us that there were a large amount of younger individuals in the 0-5 age range, and also a large amount of people in the 50-60 age range, which makes sense as this is when people normally have medical complications.



My second piece of data exploration was comparing the effect age had on whether or not people showed up to their appointments. As this graph shows, more people definitely showed up to their appointments compared to people who didn't. One could also be mistaken for thinking that as age increases past 60, the number of people showing up to their appointments decreases, however this is simply because fewer people are going to the doctors in general past that age, likely due to like expectancy and people passing away after that age. A better representation would be to group the data by peoples age, and then find the mean value of people who showed up by each age group. This is represented in the graph below.



What we can see here is that between the ages 20 to 80, it seems that people are less likely to miss their appointments, and then after 80 years old the data becomes a bit scattered, likely due to the low number of samples representing these age groups, so they are more susceptible to be affected by extreme values. We can see that apart from the anomalous sample of 115 years old, the age group most likely to miss an appointment is between the age range of 15-18.

I then calculated the mean number of people who showed up to their appointment and the mean number of people who did not show up to their appointment, and received the results below.

```
Percent of people who did not show up: 20.19%
Percent of people who did show up: 79.81%
```

Once I had calculated the mean value of not showing up to an appointment, I decided to calculate the mean value of not showing up to an appointment if an individual did not have a scholarship vs if they did have a scholarship, and received the results below.

```
Percent of people who did not have a scholarship and did not show up: 19.80%
Percent of people who had a scholarship and did not show up: 23.77%
```

This shows us that people who have a scholarship are roughly 4% more likely to miss an appointment than people who do not have a scholarship.

Lastly, I calculated the mean value of not showing up to an appointment if an individual received an SMS vs if they did not receive an SMS, and got the results below.

```
Percent of people who received an SMS and did not show up: 27.59%
Percent of people who did not receive an SMS and did not show up: 16.69%
```

What this tells us is that people who received an SMS were actually 10% more likely to not show up to their appointment compared to people who did not receive an SMS.

Conclusion

In conclusion, this dataset seems to show that around 1 in 5 people will miss their appointment, and that age seems to play some sort of role in this, as between the ages of 20-80, there seems to be a positive correlation between age and showing up to an appointment. Whether you receive an SMS or not also seems to have an effect on your likelihood to show up, with people received an SMS being 10% less likely to show up to their appointment. However, these results are not conclusive as I have not performed any statistical tests on them, therefore I do not know if these results are statistically significant or not.

This dataset was relatively easy to work with, as there were no duplicate or missing values, however, there were a few erroneous data points such as a negative age value that needed removing. The last column of 'No-show' was quite confusing so I renamed it and changed the values as explained above, to make the dataset easier to comprehend. I also had to change the ScheduleDay and AppointmentDay datatypes to DateTime in a rather unorthodox manner which involved me splitting the values into two, one with the date, and the other with the time, and then dropping the time value. I'm sure there exists a way that is more conventional that I could have used, however in the end, the outcome was the same.