REPORT

Based on the dataset „No-show appointments” from 100k medical appointments in Brazil (focused on the question of whether or not patients show up for their appointment) I have answered the questions below:

1. Which single variables effect patient's show up to a visit the most (base question)?
2. How receiving the SMS and not receiving the SMS before the visit affects the no-show when scheduling day is different then appointment day?
3. How the time interval between the visit and scheduling day affect not showing up?

Data wrangling steps:

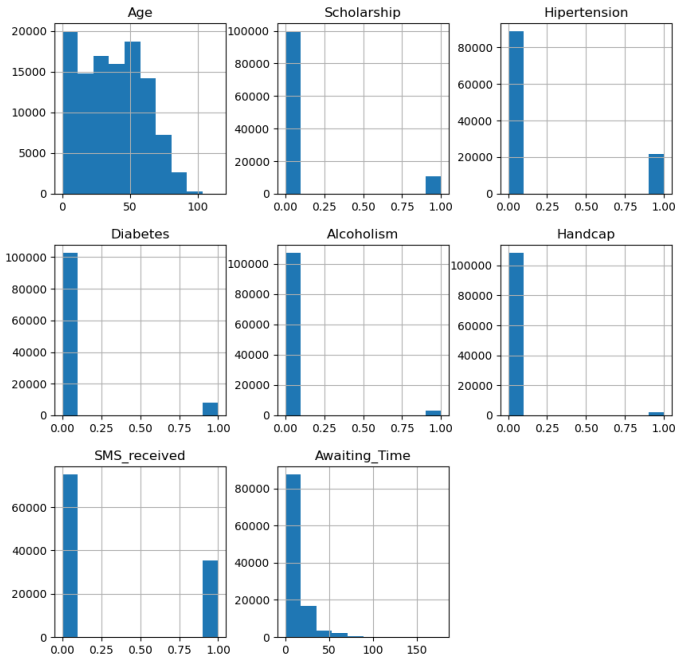
```
- first look
- checking data types
- dropping duplicated rows
- dropping rows with N/A values (we want to work with complete data,
  predicting boolean values is not recommended)
- checking unique values of appropriate columns
```

```
outcomes:
- raw dataset dimensions: 14 columns x 110527 data rows (header excluded)
- dataset dimensions after de-dup and N/A cleanup: 14 columns x 110527 data rows (header excluded)
- duplicated rows count: 0
- rows with N/A count: 0
- cleaning data from corrupted values:
```

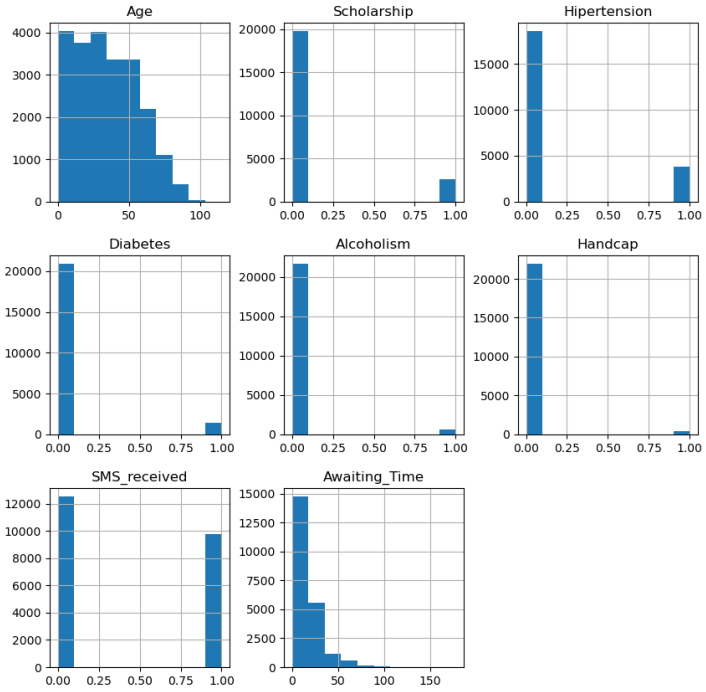
```
- changing handcap level values to boolean parameter (to treat all levels of handcap as a handcap in general)
- setting up AppointmentID as index
- changing dates datatype to datetime YYYY-MM-DD
- adding column Awaiting_Time (number of days from 0 to x, how long patient waits for the visit)
- adding column Weekday_Of_Appointment
- final dataset dimensions and columns after data wrangling
```

Plots:

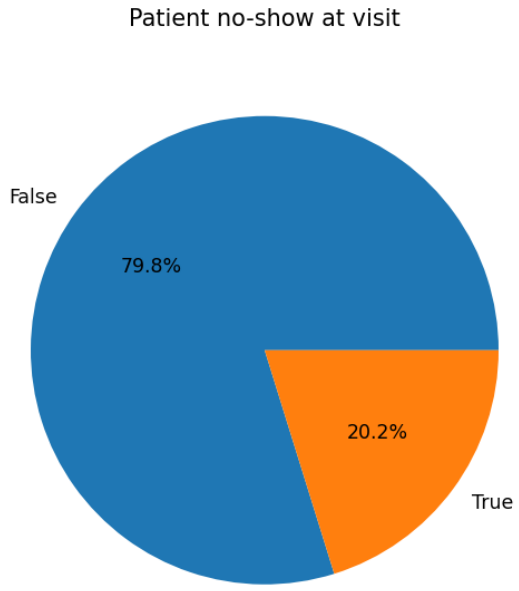
1. Histogram for overview on counts



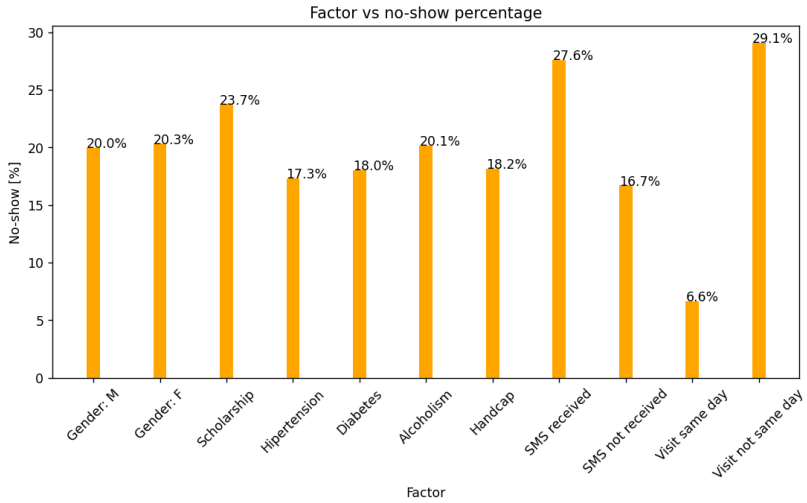
2. No-show only histogram



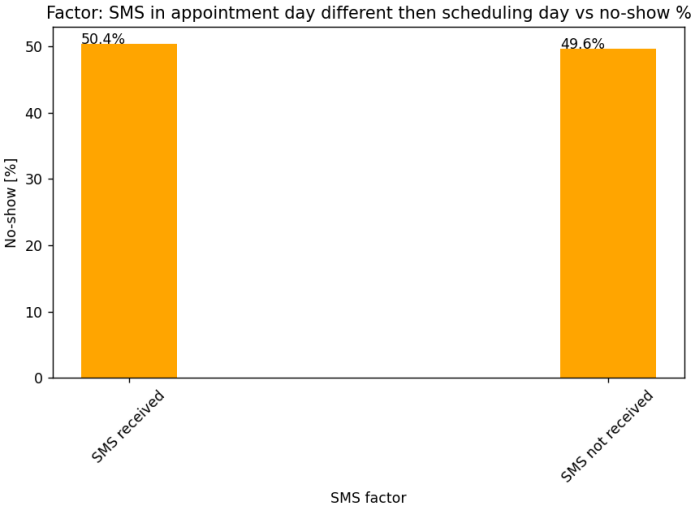
3. Pie chart : no-show vs show in percents



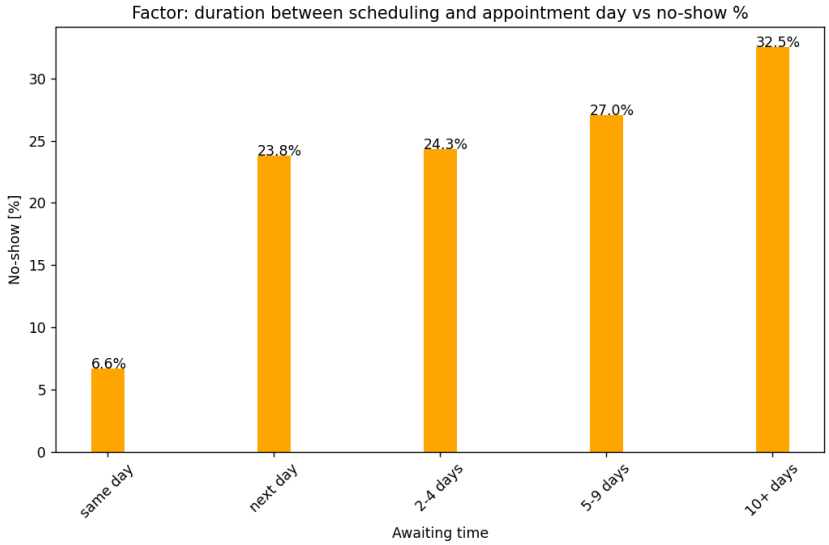
4. Plot : percentage of the factor in no-show



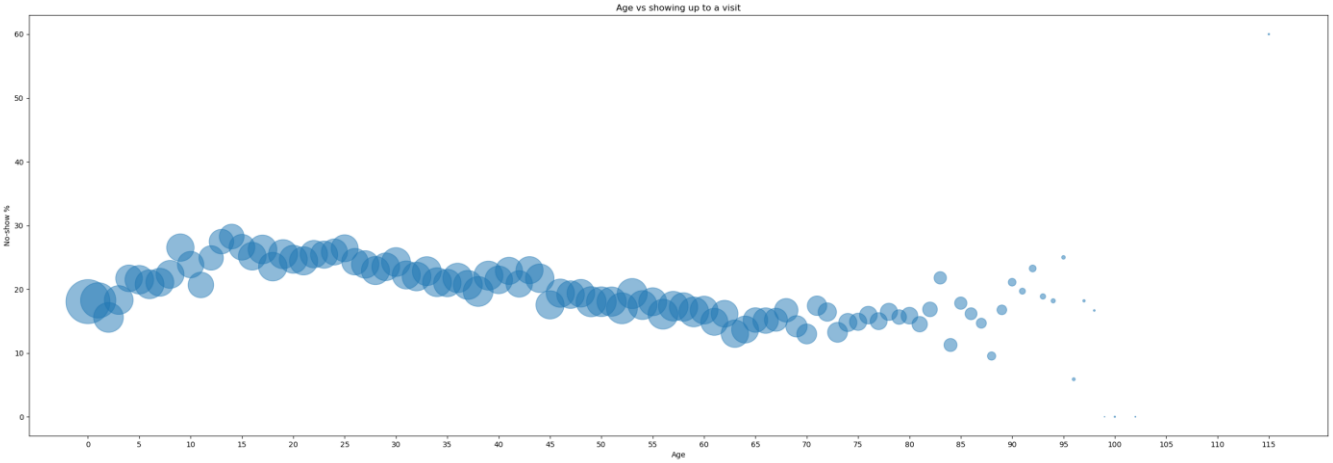
5. Plot : SMS in appointment day different then scheduling day impact on no-show



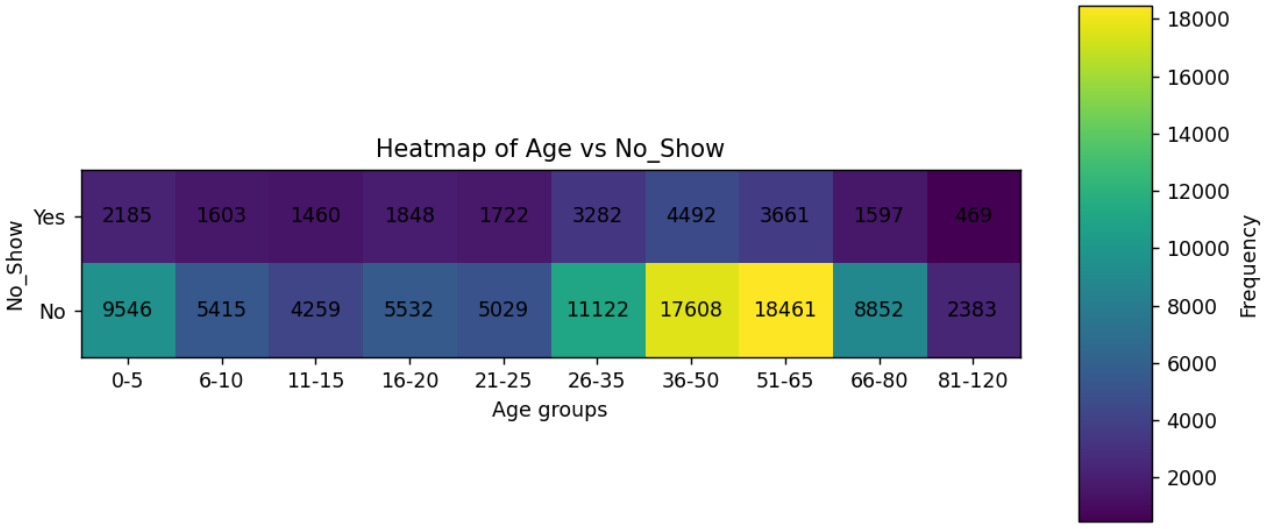
6. Plot : time interval between the visit and scheduling day affect on not showing up:



7. Scatter plot: how does age affect on no-show percentage, with visualising size of age set size



8. Heatmap showing counts of show and no show for different age groups



Conclusions:

1. Based on the analysis, age, taking scholarship, receiving SMS and having to wait for the visit seems to be the 4 factors which have greater share in no-show. Receiving SMS as a factor which increases chances of not showing seems to be strange. After taking a deeper look at this factor, we can spot that when scheduling day is the same as appointment day, patients are not receiving SMS. It means that we should take SMS factor under consideration only for the visits which are not in the scheduling day.

   ```
   Count of SMS received if Awaiting_Time = 0: 0
   ```

   Additionally, females are slightly more prone to no show at the visit than men, but the difference is very small.
2. Analysis on SMS and awaiting time shows, that the percentage difference between SMS impact on no-showing to the visit in the day different than visit's scheduling day is very similar. We can say that not receiving SMS is not affecting not showing up to a visit.
3. Based on further analysis, we can spot that increase of awaiting time affects on the no-show percentage – not showing up percentage is increasing.
4. Moreover, patients between age ~15 and 25 seems to more likely skip the visit.
5. Additional research can be done on Neighbourhood factor, and some combined factors, e.g. „How does age affect on alcoholics showing up to a visit?", or „How does alcoholism affect people with scholarship showing up to a visit?".

Limitations:

1. Important limitation in the dataset is information about timing of receiving the SMS and appointment date. Receiving the SMS 1 or 2 days earlier is different then receiving SMS on the same day.