Context:

A person makes a doctor's appointment, receives all the instructions, and no-show. Who to blame?

300k medical appointments and its 15 variables (characteristics) of each. The most important one is if the patient show-up or no-show the appointment. Variable names are self-explanatory

Problem Statement:

Predict someone to no-show an appointment.

Dataset Description:

PatientId - Identification of a patient AppointmentID - Identification of each appointment
Gender - Male or Female. Female is the greater proportion; woman takes way more care of their health in comparison to man.
DataMarcacaoConsulta - The day of the actual appointment, when they have to visit the doctor.
DataAgendamento - The day someone called or registered the appointment, this is before appointment of course.
Age - How old is the patient.
Neighbourhood - Where the appointment takes place.
Scholarship - Ture or False observation. This is a broad topic, consider reading this article https://en.wikipedia.org/wiki/Bolsa_Fam%C3%ADlia
Hipertension - True or False
Diabetes - True or False
Alcoholism - True or False
Handcap - True or False
SMS_received - 1 or more messages sent to the patient
No-show - True or False

Approach:

Following pointers will be helpful to structure your findings.

1. Try and explore the data to check for missing values/erroneous entries and also comment on redundant features and add additional ones, if needed.

2. It is immediately apparent that some of the column names have typos, so let us clear them up before continuing further, so that we don't have to use alternate spellings every time we need a variable.

3. For convenience, convert the AppointmentRegistration and Appointment columns into datetime64 format and the AwaitingTime column into absolute values.

4. Create a new feature called HourOfTheDay, which will indicate the hour of the day at which the appointment was booked.

5. Identify and remove outliers from Age. Explain using an appropriate plot.

6. Analyse the probability of showing up with respect to different features. Create scatter plot and trend lines to analyse the relation between probability of showing up with respect to age/Houroftheday/awaitingtime. Describe your finding.

7. Create a bar graph to depict probability of showing up for diabetes, alcoholism, hypertension, TB, smokes, scholarship.

8. Create separate bar graphs to show the probability of showing up for male and female, day of the week and sms reminder. Describe your interpretation.

9. Predict the Show-Up/No-Show status based on the features which show the most variation in probability of showing up. They are:

- Age
- Diabetes
- Alchoholism
- Hypertension
- Smokes
- Scholarship
- Tuberculosis

10. Create a dashboard in tableau by choosing appropriate chart types and metrics useful for the business.