Investigate a Dataset

The chosen data:

No show appointments: 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 Brazilian welfare program Bolsa Família.
- 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.

Questions posed:

What are the factors that affects the "No Show Appointment" the most?

- We will investigate the relation between the "now show" and each independent variable: AppointmentDay, Age, Neighbourhood, Scholarship, Hypertension, Diabetes, Alcoholism, Handicap, SMS_received

Is the difference between ScheduledDay and AppointmentDay affects "No Show Appointment"?

Steps of investigating the questions:

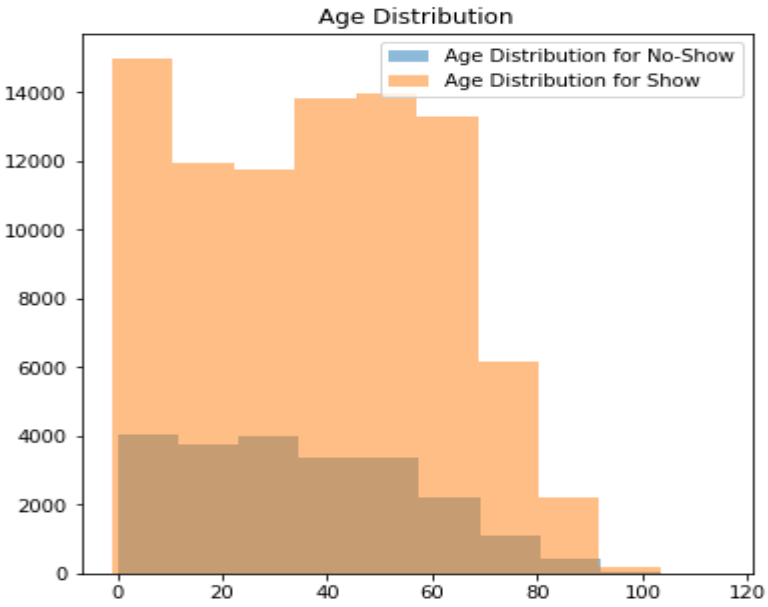
- Splitting the data set into two: one for show appointments and the other for no show
- Calculating the age distribution and mean age for each dataset
- Calculate the percentage of no show for each neighbourhood (no show per neighbourhood / total appointment for the same neighbourhood)
- Calculate the percentage of no show per day of week
- For other features, calculate the percentage of no show appointments with the feature / total number of appointments with the feature

Data Wrangling:

- The data does not contain null values or duplicates
- Adjusting ScheduledDay and AppointmentDay to datatime type.
  To do that we adjusted the content from "2016-04-29T18:38:08Z" to "2016-04-29 18:38:08"
- editing columns names: {'Hipertension' : 'Hypertension', 'Handcap' : 'Handicap', 'No-show' : 'No_show'}
- Changing "No_show" column form Yes & No to 1 & 0

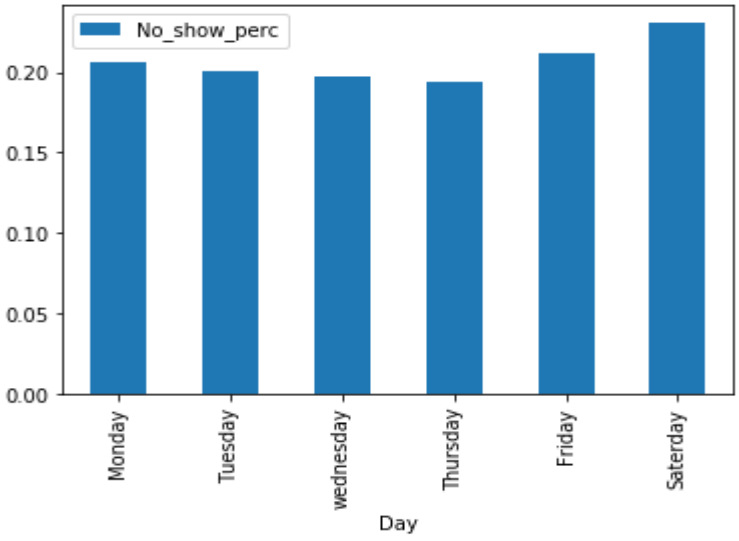
Find Result:

- The age does not highly affect the appointment Shwo/No_show because the age distribution is similar.



- Neighborhoods: SANTOS DUMONT, SANTA CECÍLIA, SANTA CLARA, and ITARARÉ has a percentage of No_show over 25%
  Note: ILHAS OCEÂNICAS DE TRINDADE is considered outlier.

- Saturday and Friday have the highest No_show value. However, the difference is not very high.



-

- For each feature, the value of No_show with the feature / the total value with appoints with the feature is calculated.

	Feature	No_Show_Perc_with_feature
0	Scholarship	0.237363
1	Hypertension	0.173020
2	Diabetes	0.180033
3	Alcoholism	0.201488
4	Handicap	0.179236
5	SMS_received	0.275745

27% of appointments with SMS received do not show.
23% of appointments with scholarship do not show.
20% of appointment with alcoholism do not show.

Links used in investigating the dataset

- https://www.w3schools.com/python/python_try_except.asp
- https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DatetimeIndex.dayofweek.html
- https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.nlargest.html