No-show appointments

Source kagglev2-may-2016

Data Wrangling

Data gathering :

- all data are gathered in 1 CSV file (noshowappointments-kagglev2-may-2016.csv)

This data-set 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.

- ‘Scheduled Day’ 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.
- **Be careful about 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.**

Data Assessing :

After exploring data using Jupyter Notebook and checking data types ,value_counts, non-null values of data , loading samples of data , data descriptions and summaries
searching for duplicates and messy data to be tidied .

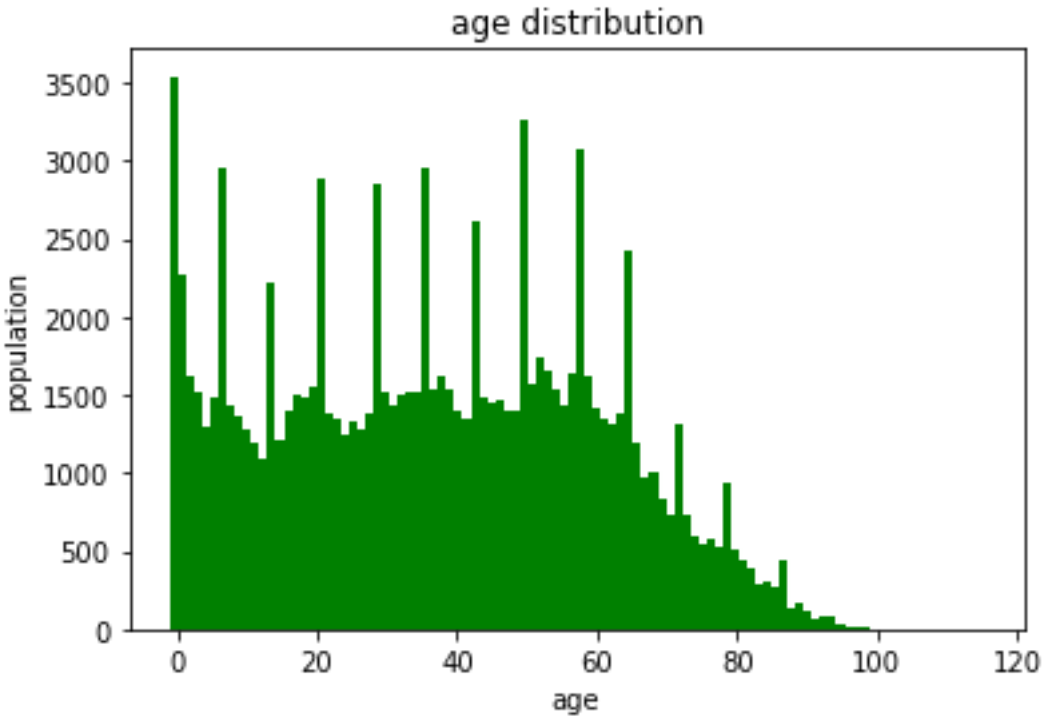
- **Duplication :** same patient may have many appointments with different appointment id so there's no duplication to remove.

Data Cleaning :

- **Changing Datatype**
    - patient-id to int
    - Scheduled-Day to date-time
    - Appointment-Day to date-time

- **Removing outliers**

Age have outliers -1 and above 90 as shown in the below graph



- Changing Gender and No-show from categorical data to numerical for easier investigation .

Data insights

Question 1

what's the percentage of people who attend/missed their appointments ?

Ans.

percentage of people who attend there appointments 80.0 %
percentage of people who miss there appointments 20.0 %

Question 2

What reasons behind the missed appointments ?
Ans.

Exploring reason for the 20% not showing to their appointments assuming and testing assumption for some variables effecting showing/not-showing to the appointments such as

assumption 1:

Assuming Age of people effect the attendance of appointments .

- correlation between Age and No-show =0.0608

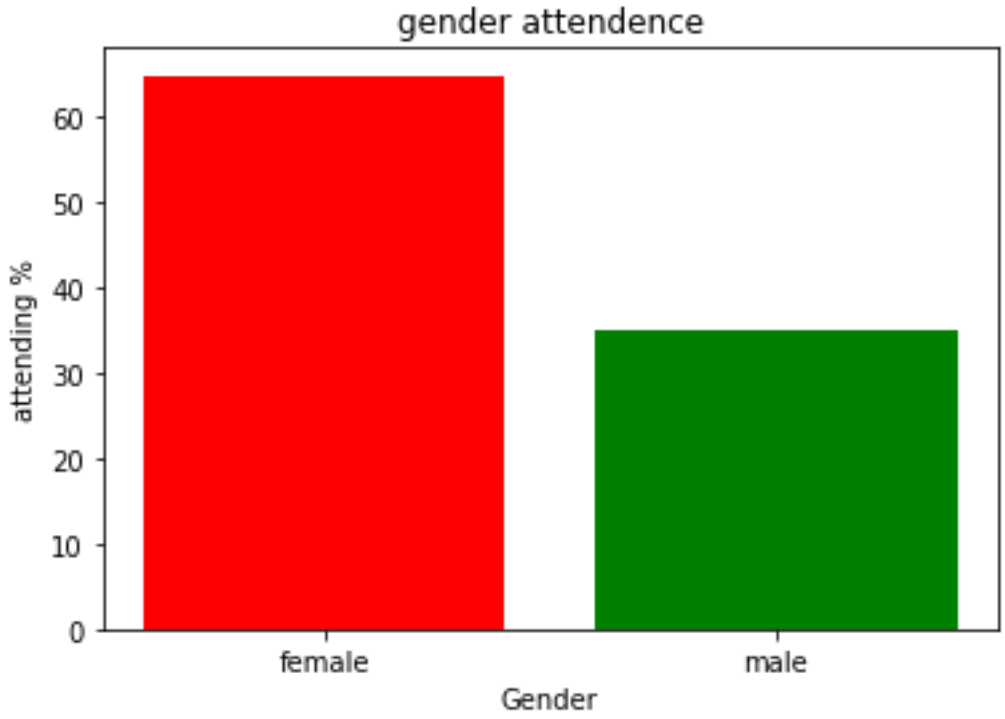
very weak correlation

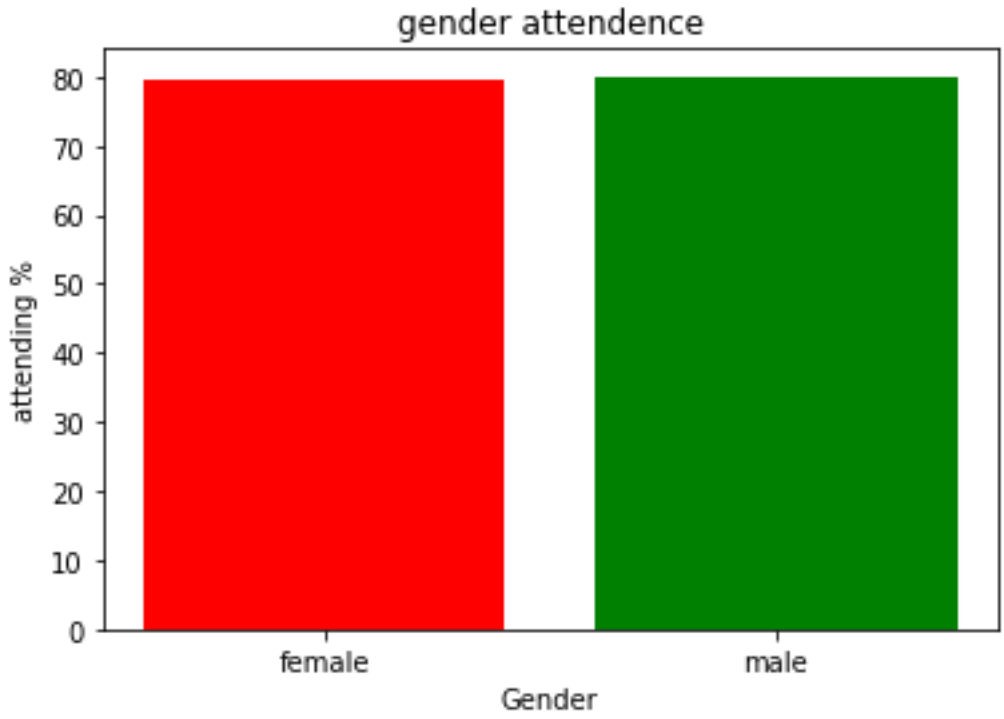
- **Conclusion** people's Age has no effect on attending their appointments

assumption 2:

Assuming Gender of people effect the attendance of appointments .

- percentage of females who attended out of the total attendance 64.89%
- percentage of males who attended out of the total attendance 35.1 %





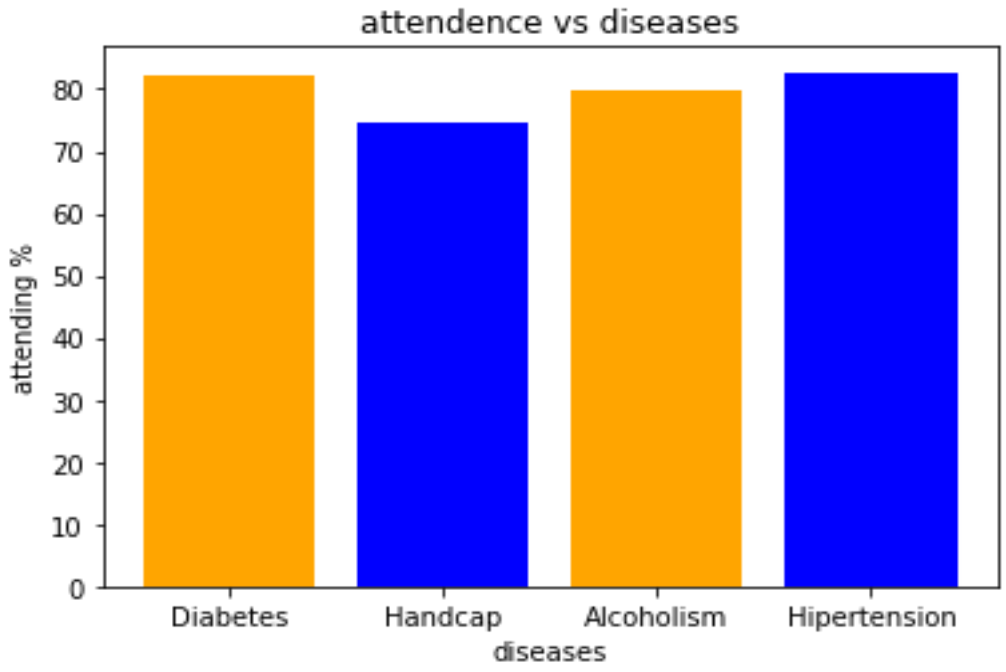
- percentage of females who attend out of the total female signed for appointment 79.68 %
- percentage of males who attend out of the total male signed for appointment 80.03%

- 79.6 % of the female patients attend their appointment and 20.4% missed their appointment
- 80% of the male patients attend their appointment and 20% missed their appointment
- **Conclusion** both male and female having the same percentage attending and missing appointments so gender doesn't effect No-show

Assumption 3:

Assuming type of disease effect the attendance of appointments

- calculating how type of disease may effect attendance of appointments as shown in the below figure.

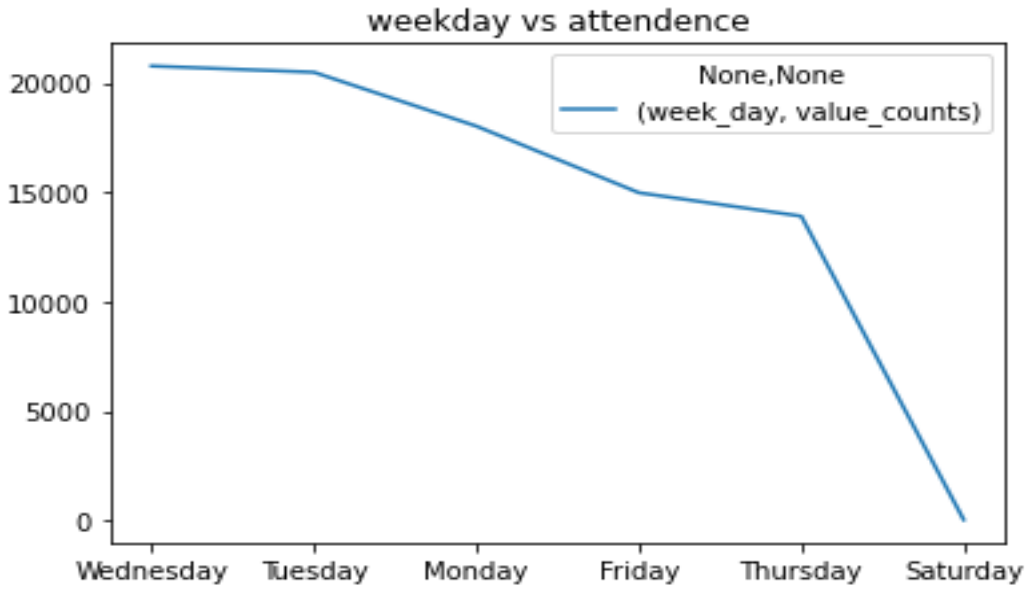


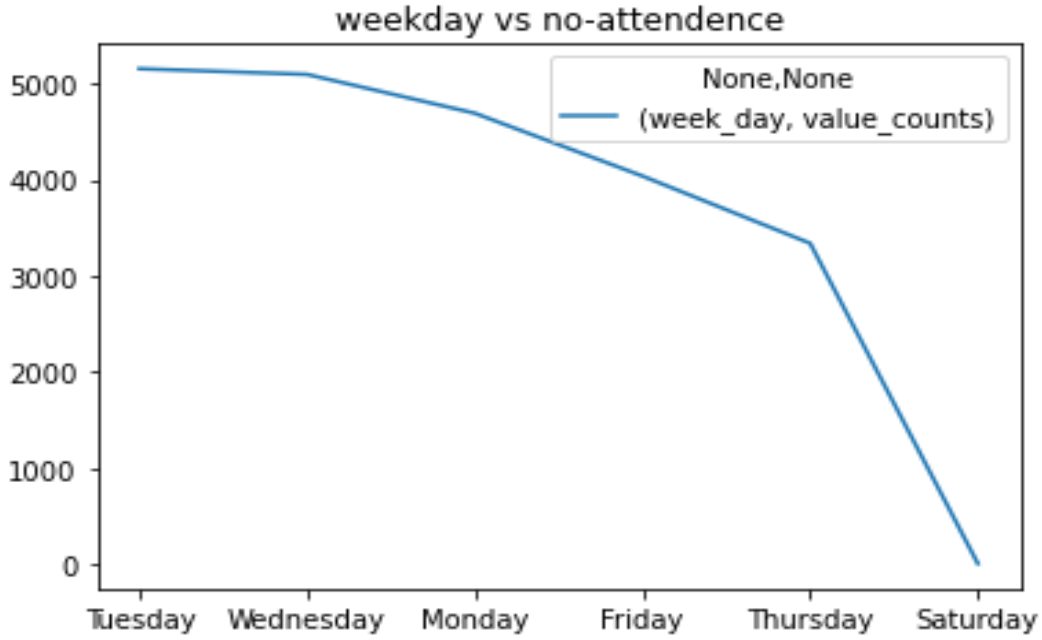
- **Conclusion** People have diabetes and hypertension are more likely to attend their appointments

Assumption 4:

Assuming week_day effect people attendance in case of weekend or middle of the week

- as shown in the below figures Saturday has the lowest appointment.





- **Conclusion** people prefer more to attend their appointments in the middle days of the week more than weekend days (Saturday)

Assumption 5:

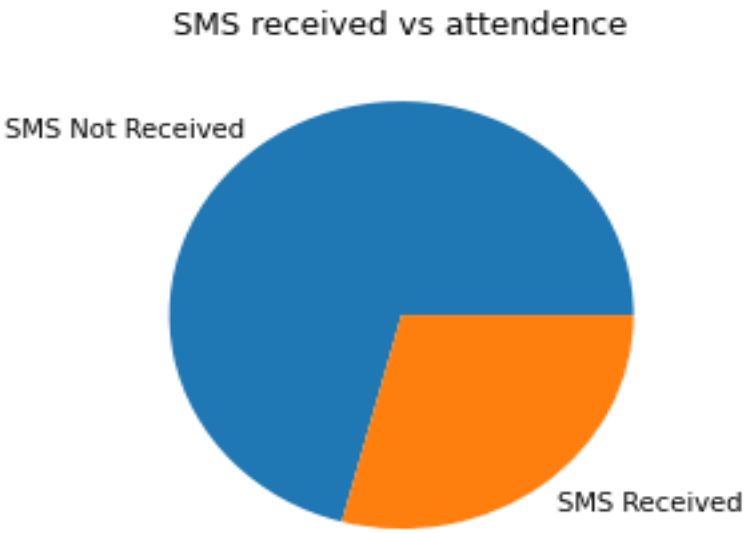
Assuming people having scholarship more likely to attend appointments

- correlation between difference scholarship and no-show = -0.029135
very weak correlation.
- **Conclusion** difference scholarship has no effect on attending their appointments

Assumption 6 :

Assuming receiving/ not-receiving SMS effect attendance.

- 29% only from people how attended their appointments received SMS



- **Conclusion** the majority of the people who attended didn't receive sms so receiving sms has no effect on attendance.

Assumption 7 :

Assuming the difference between appointment day and scheduled day getting longer 'll increase the chance of missing the appointment

- correlation between difference in days and no-show = -0.180074

very weak correlation.

- **Conclusion** difference in days has no effect on attending their appointments.

Final conclusion :

- after investigation and testing multi assumptions to find the reason behind missing appointments non of the assumptions passed and couldn't find an accurate reason .

- the data-set is not sufficient enough to answer the questions may be if the size of the data-set increased 'll give a good judgment about the questions I asked and could find the reason behind missing appointments .

Storing :

cleaned data stored in new CSV file named (No-show_appointment_new.csv)