Udacity Data Analyst Nanodegree Program

Investigate a Dataset project

Dataset analyzed in this project:

No-show appointments from Kaggle (noshowappointments-kagglev2-may-2016.csv)

Questions I pose for this dataset:

- is there a relationship between the time gap between the appointment day and the scheduled day, and the patient won't show up for the appointment?
- Is there a relationship between the patient receiving an SMS and him/her not showing up for the appointment?
- Is there a relationship between the of the age patients and him/her not showing up for the appointment?

A description of how I investigate the answers for those questions:

- Q1)

  Step1)

  I made a new field in the data frame called Days_Between_Ad_Sd and made it equal to df['AppointmentDay'] - df['ScheduledDay'].

  Step2)

  I made a scatter plot between df["Days_Between_Ad_Sd"] and df['No_show']) but the resulting plot was not very effective at showing any relationship.

  Step3)

  I used the binning technique show in "Data Analysis Process - Case Study 1" to make a bar plot with the following cutoffs -7,0,4,15,179, which split the data into four groups.

- Q2)

  I made 6 counts for every possible outcome patience can have, for example, receiving an SMS & not showing up or not receiving an SMS & showing up, etc... and made 4 bar plots showing the relationship between them.

- Q3)

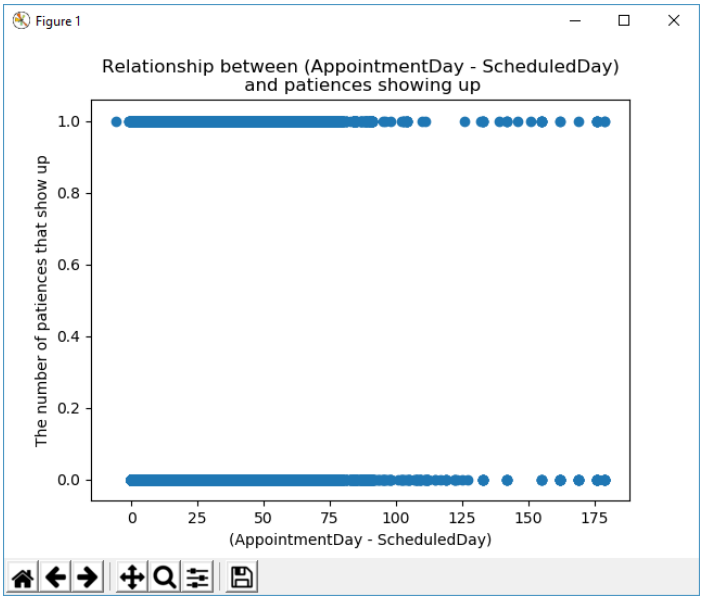
  The same as Q1 step 3 with the cutoffs being -∞,18,37,55,155.

Data wrangling:

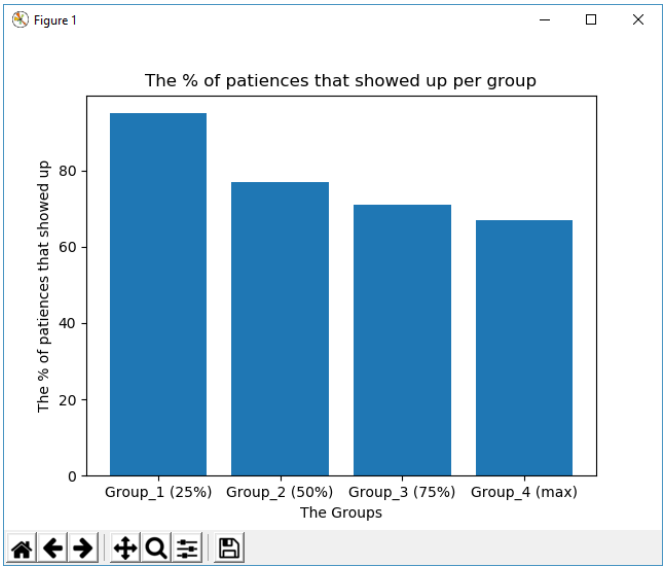
- Dropping all nulls or duplicates
- Changing any negative age into the average age in the date frame
- Change the following fields into int64
    - PatientId
- Change the following fields into datetime64
    - ScheduledDay
    - AppointmentDay
- Change the following fields into Boolean
    - Scholarship
    - Hipertension
    - Diabetes
    - Alcoholism
    - Handcap
    - SMS_received

Results:

- Q1)



As you see the scatter plot as trouble showing the relationship however when using the binning technique shown in “Data Analysis Process - Case Study 1” we can make this plot



Which show that longer that gap between appointment day and the scheduled day the more likely the patient won't show up for the appointment.

Note:

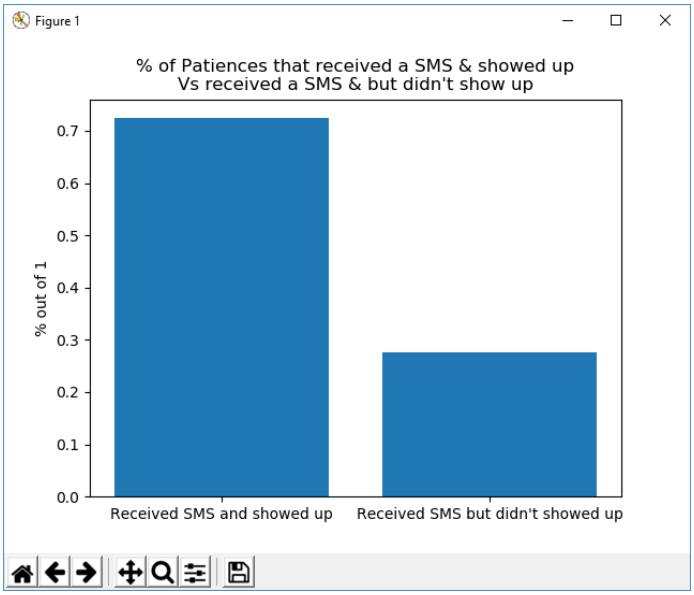
Group_1 = [-6,0]

Group_2 = (0,4]

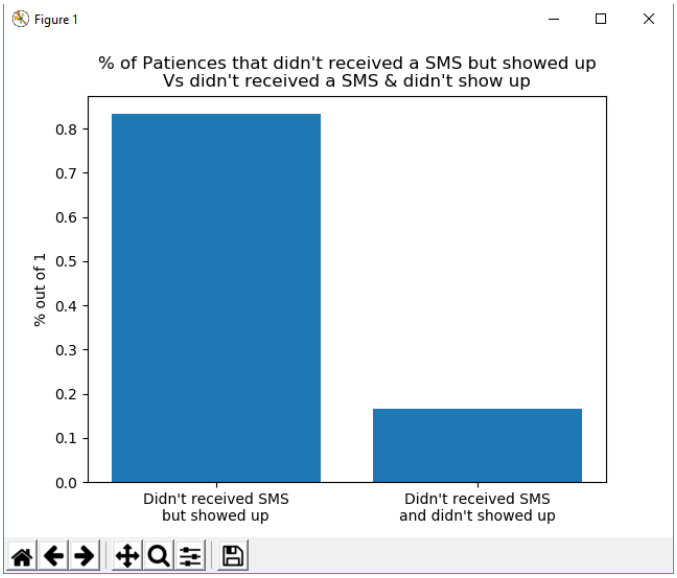
Group_3 = (4,15]

Group_4 = (15,179]

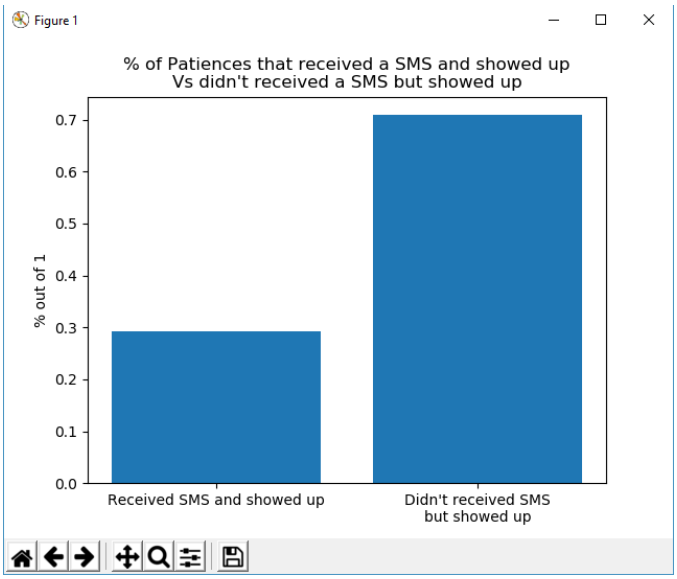
- Q2)



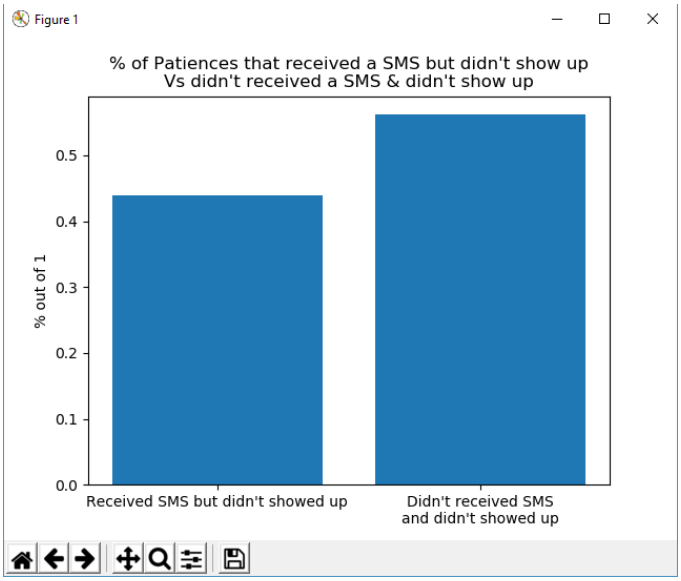
No surprise here, given that a patient receives an SMS they are likely to show up to their appointment about 70% of the time.



From this plot, we can theorize that most people either didn't know that they can receive an SMS for their appointment or didn't bother asking for the service because given that a patient doesn't receive an SMS more than 80% of patients will show up for their appointment.



This plot reinforces the theory stated above, form all the people who showed up for an appointment about %70 did receive an SMS.

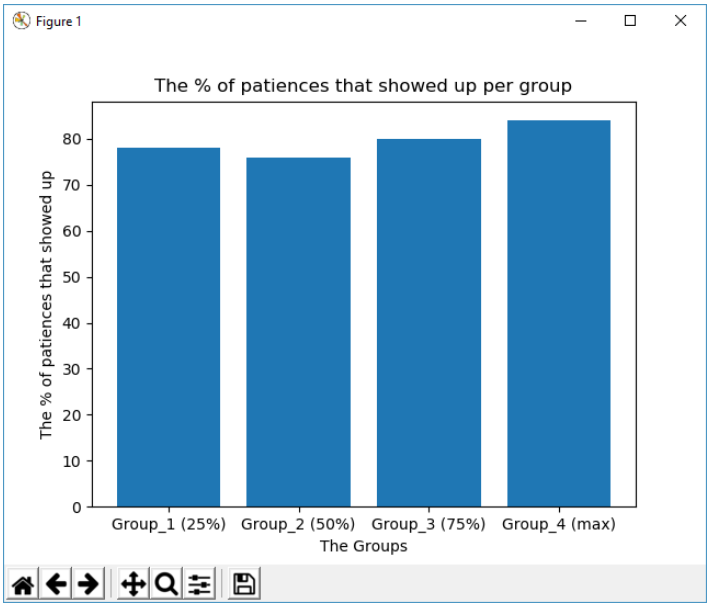


This plot is interesting, form all who didn't show up for their appointment about 45% of them were form patient that receive a SMS and as we theorize before the number of people who don't receive an SMS are the biggest group in this dataset which means that if you did receive an SMS you are more likely to not show up.

Conclusion:

While there is a relationship between receiving an SMS and showing up for an appointment it is a weak one and I think this dataset just shows that most patients don’t ask for an SMS message than showing the relationship between SMS and showing up.

- Q3)



The plot shows there is no relationship between age and showing up for an appointment.

Note:
Group_1 = [0,18]
Group_2 = (18,37]
Group_3 = (37,55]
Group_4 = (55,115]