Udacity Data Analyst Nanodegree

Investigating a Dataset

Introduction

For this project, we were tasked with choosing a dataset from a curated list and going through the complete data analysis process, including accessing the data, cleaning and wrangling, analyzing it to answer research questions, and creating visualizations to illustrate or draw conclusions. For my investigation, I chose a dataset, originally provided by Kaggle, with information about no-shows to medical appointments in Brazil. I chose this dataset because my main data interests are healthcare, child health, and disability. I thought that disability status might have an impact on appointment attendance, and was also curious about how age would interact with attendance. However, once I dug into the data, I was surprised by the results I ultimately found.

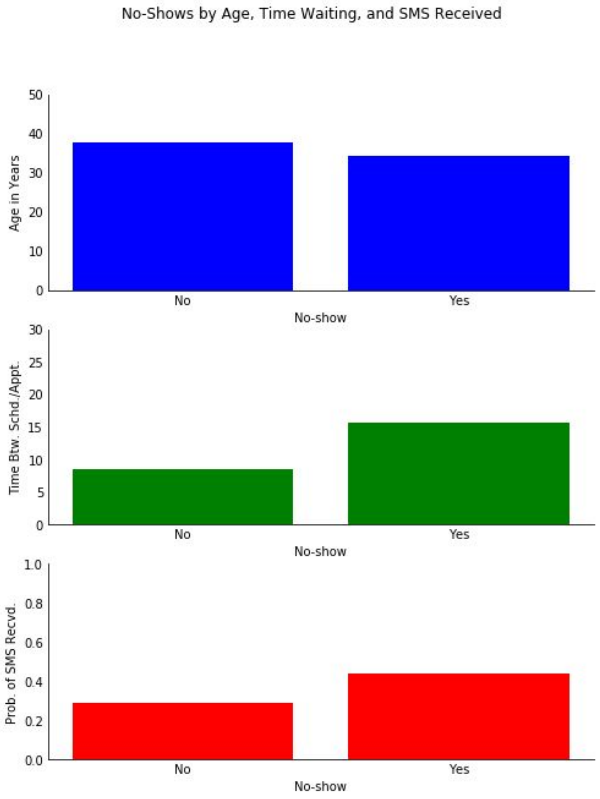
Data Cleaning/Wrangling

After reading in the .csv file using Pandas and viewing the data, I made several modifications. First, I renamed two columns to correct misspelled names. Then, I converted the datetime-like string objects in the ScheduledDay and AppointmentDay columns to datetime objects, and because AppointmentDay contains no unique times, selected the date attribute of both to use in my dataset. Because those columns were mostly interesting because of what they implied about the distance between them (i.e., the time between the appointment scheduling and the actual appointment date), I created a new column, TimeBetween, to hold the difference as timedelta objects. Then, after looking through the data, I realized there were some strange values in the Age and new TimeBetween columns, so I restricted the parameters of those to reduce extreme outliers. Finally, I dropped several unnecessary columns so I could focus on the ones I wanted to analyze.

Data Analysis

Question 1: What features are associated with attendance?

As mentioned above, I chose this project thinking that disability status might have an impact, so once the data was clean I grouped by the No-Show column and viewed the means of each feature to see what features seemed most associated with attendance. I was surprised to realize that most of the features in the dataset did not appear to have much, if any, correlation with attendance. That is, the means of most features (including disability status) were roughly the same for both No-Show and Show. The features that ultimately appeared most closely associated with attendance were Age and the created TimeBetween, as well as a very slight difference in the column SMS_Received (for whether or not a patient received an SMS reminder of their appointment).

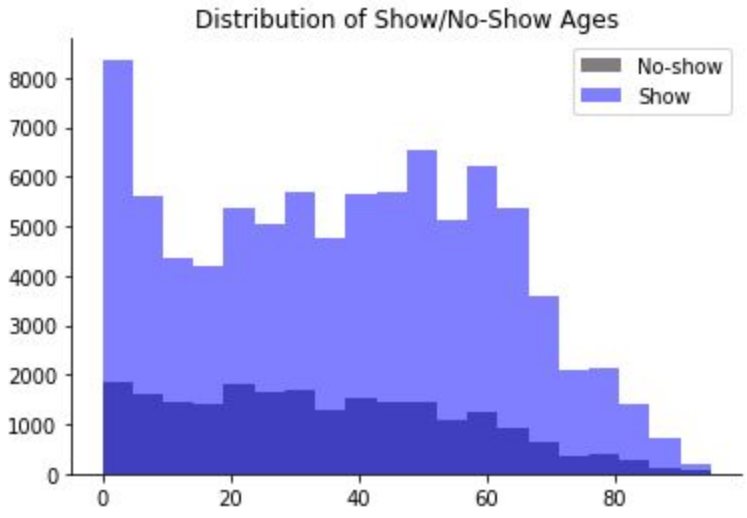


Based on the above, I created bar charts (left) for those three features and looked at the results. Interestingly, patients who did not show appeared slightly *more* likely to have received an SMS reminder, contrary to expectations. Patients who did not show also appeared to trend slightly younger and to have waited slightly longer between scheduling and appointment. There were no very sharp differences that would indicate that any one feature was probably the cause of a no-show. However, the biggest difference was probably in TimeBetween, with no-show patients waiting an average of one week longer than patients who did show.

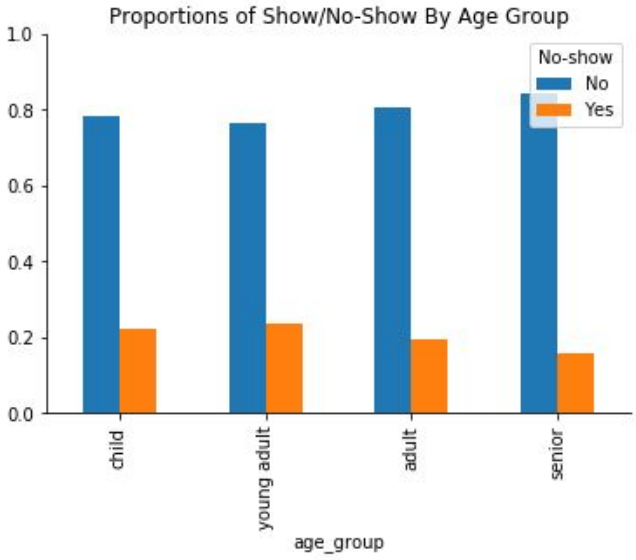
Question 2: How are ages distributed between no-shows and 'shows?'

While there was not a huge difference between ages in the no-show and 'show' groups, I was still curious about the fact that the average of the no-show group appeared to be about three and a half years younger than that of the show group, and wanted to see how the ages would be distributed if analyzed further. I wanted to see if a particular age group would emerge as attending less/missing more appointments. To tackle that question, I first

created a histogram showing the distribution of ages for patients who did and did not show up to their appointments. I found that the distributions were roughly similar, with an overall right skew, although the 'no-show' distribution was much smoother. The child health advocate in me was relieved to see a large spike of 'show' appointments in childhood, although a sizeable proportion of no-shows also appear to be happening in early childhood.



To investigate further, I created four age groups (child, young adult, adult, and senior) and used the Pandas cut function to assign values based on each Patient's age to a new column, age_group. I grouped by age_group and no-show status to aggregate the size of each group (i.e. how many patients in each age group fall into show/no-show) and normalized the results to smooth over differences in age group size. I again plotted the resulting pivot table as a bar chart, looking at the proportions of show/no-show in each age group.



Again, the differences between the age groups are too subtle to be regarded as clearly definitive results. However, the "young adult" group edges the others out in no-shows, with 24% of young adults missing appointments. Disappointingly, 22% of children appear to have missed appointments, while only 20% of adults and 16% of seniors did so.

Conclusions

While I do not feel that this dataset yielded any particularly conclusive results, it did indicate that there may be some association between age and missed appointments, as well as how far in advance that appointment is scheduled. One further direction for research would be to dig further into the age groups and see if it is possible to uncover why some ages are missing more appointments. My particular research interests lead me to want to investigate why children, in particular, are missing appointments, as regular pediatric care is extremely important for the developing child. Do those children's parents face a specific barrier to attendance, is there a cultural factor at work, or is something else happening? Similarly, research could be done on whether the young adult no-shows are due to external barriers or more internal factors, since young adults often tend to be more cavalier about their health in general. Another avenue for research is the scheduling factor: are appointments scheduled further in advance truly more likely to be missed, and if so, can that be ameliorated?