Bolsa Familia, Age, and Neighborhood Impact on No-Show Medical Appointments

A look into data from May 2016 of No-Show Medical Appointments for Brazil

Abstract:

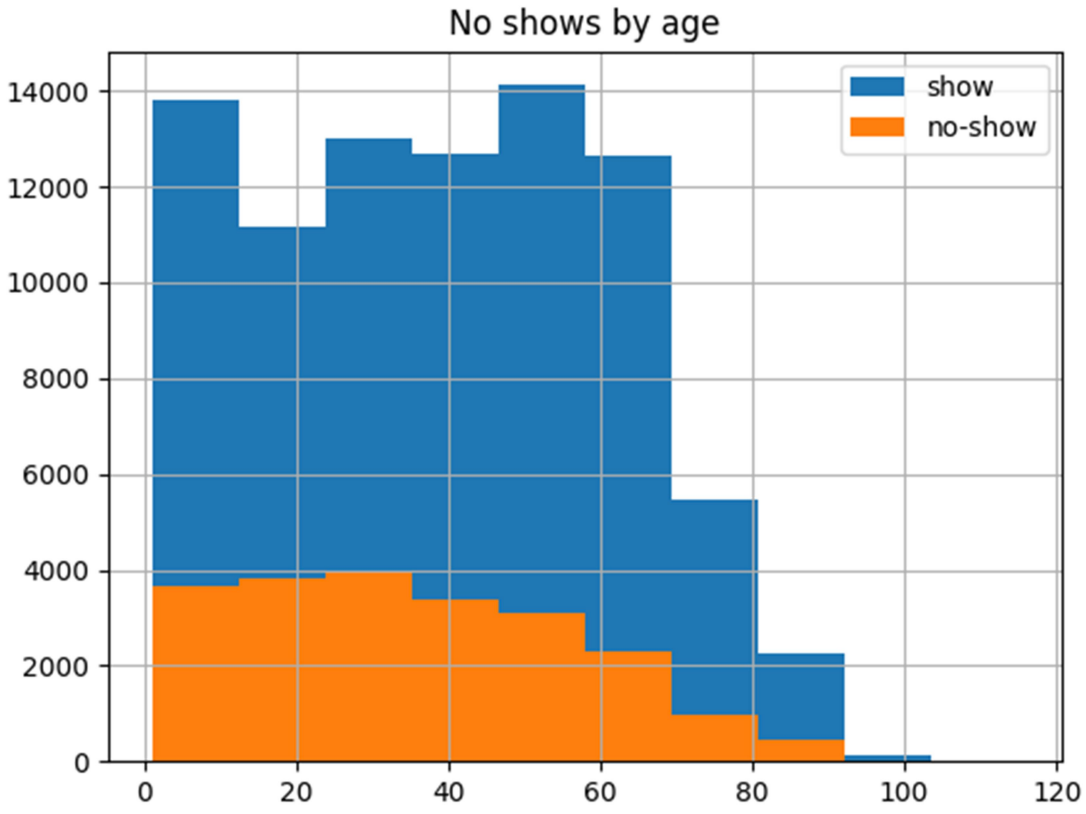
We start with the data provided from noshowappointments-kagglev2-may-2016.csv. After reviewing the initial data we have determined to find what factors affect a no show for a medical appointment. The main questions addressed here are days of the week, gender, age (or even age categories????), neighborhood, if they are, or are not, on the system 'Scholarship' (Bolsa Familia), or the status of certain ailments: Hypertension, Diabetes, Alcoholism, Handicap. While no particular influence by gender, ailment, or weekday were discovered, Age, Neighborhood, and Scholarship all seemed to have some impact. We can clearly see that the patients under the age of 25 and not on the Bolsa Familia system are more likely to be no-shows than other groups. In addition we are able to identify 7 neighborhoods, Itarare, Jesus De Nazareth, Caratorira, Ilha Do Principe, Santos Dumont, Santa Clara, and Santa Cecilia, as having significantly higher no-show rates than other neighborhoods.

Process

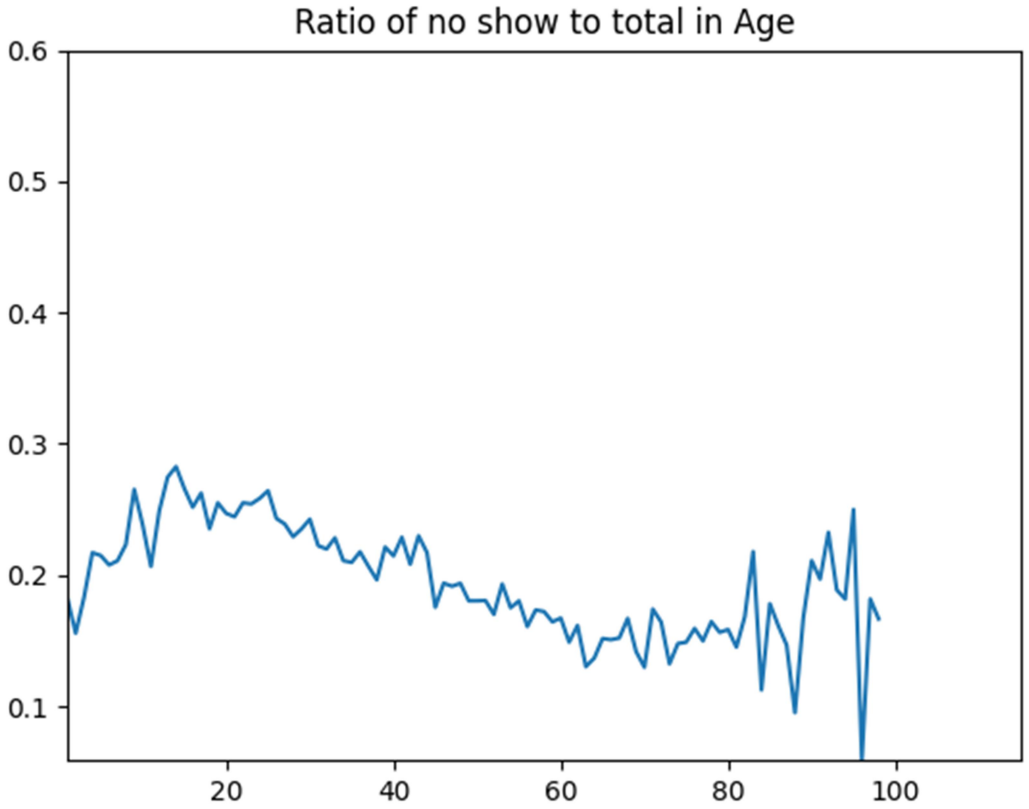
During an early check of the data I discovered that there was a row with an age of -1. As this isn't a valid age number, I purged that particular row and ignored its data. All dates were fixed as well in order to check them against the day of the week they represented, this was done by conversion from sting to pandas datetiime type objects. Other than that the data was left as is, as no major issues were found. There were no duplicated rows, and no other irregularities found at any point.

Age and No Show Appointments

Next we looked into the effect of age on the no-shows. The histogram provided by the raw data indicates a possibly strong relationship between age and not showing up for appointments.



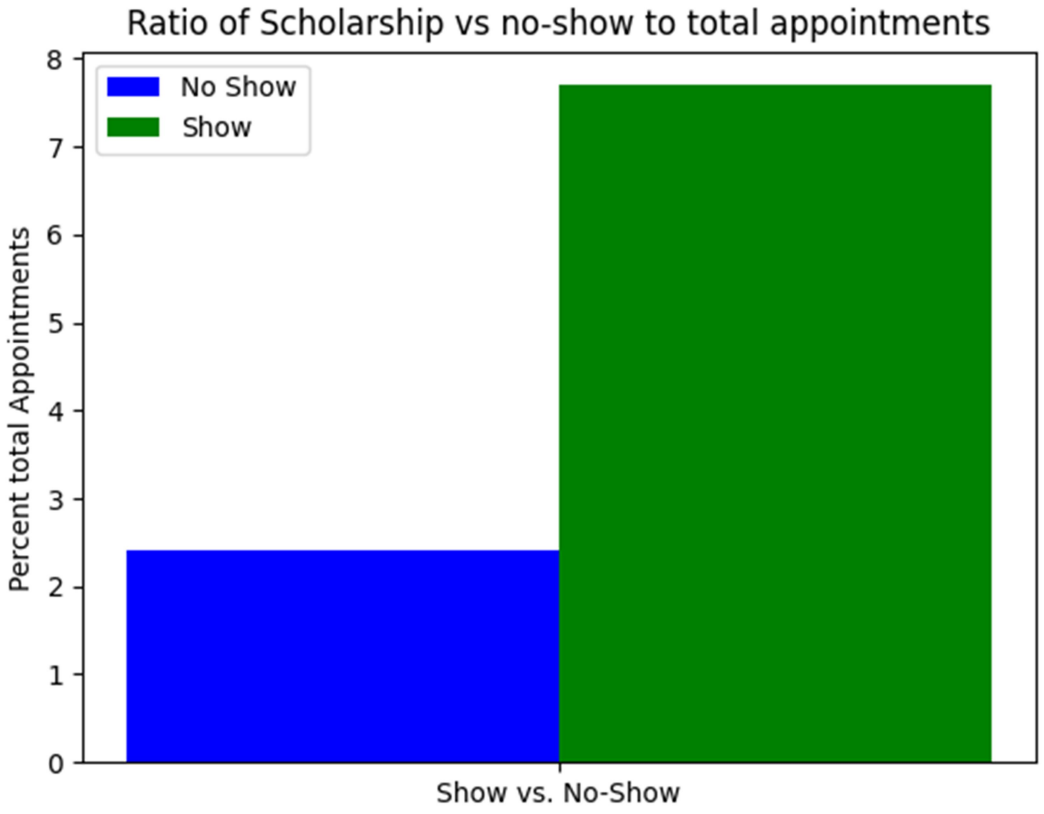
In order to further explore this possibility I took a look at the ratio between the total number of patients and the number that are no shows for their appointments.



We can see here that the most people who do not show up for their appointments are younger. With a fairly smooth line down until around 60 years old at which point until 80 year of age we have a fairly consistent number of people who do not show compared to the number of total patients. Younger patients probably skip more appointments due to being in better health, the older patients, 80 and older, seems to be chaotic possibly even indicating a slight increase, but it should be considered that there are 10 times fewer data points total for the older groups.

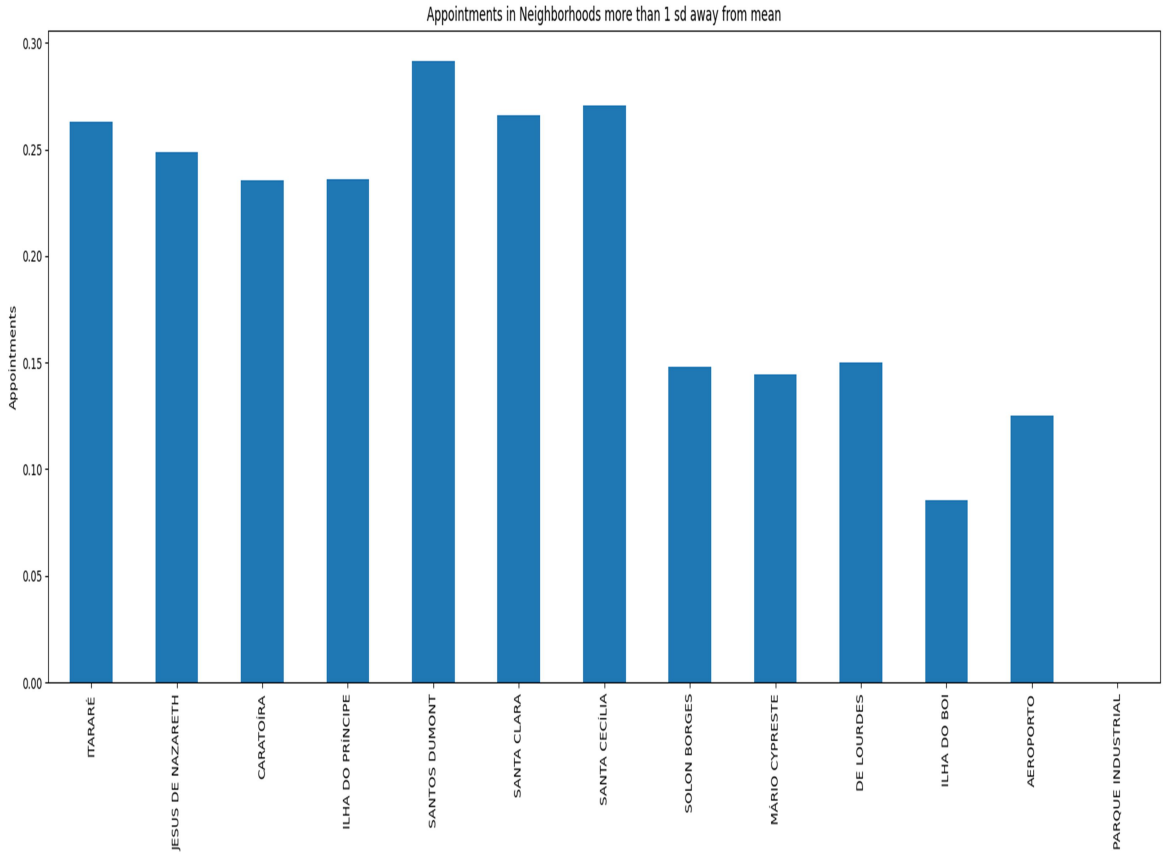
Scholarship and No Show Appointments

Patients with Bolsa Familia are significantly more likely to show up to their appointments overall. While only a small number of patients vs the total are on the state scholarship, nearly 4 times as many show up to appointments as miss it proportionately, 2% no show vs 7% making their appointments.



Neighborhood and No Show Appointments

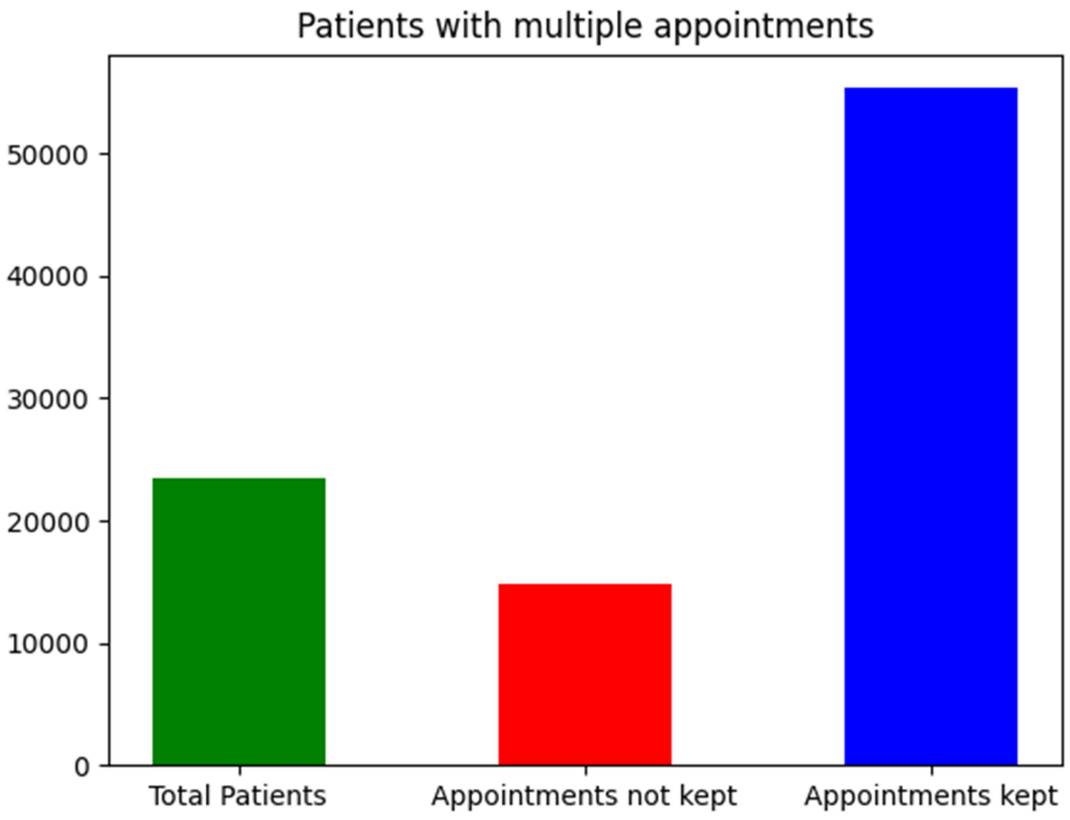
No show appointments vary significantly from neighborhood to neighborhood. With certain neighborhoods showing increased number of no-shows compared to the number of total appointments. The chart below shows highlighted neighborhoods 7 2+ standard deviations higher than the mean for no-shows and 6 2+ below the mean. The following page contains a look at all nieghborhoods.



The seven bars on the left represent the locations with the most no-shows, while the 5 bars and empty column on the right represent locations with the fewest. There was one neighborhood with zero no-shows.

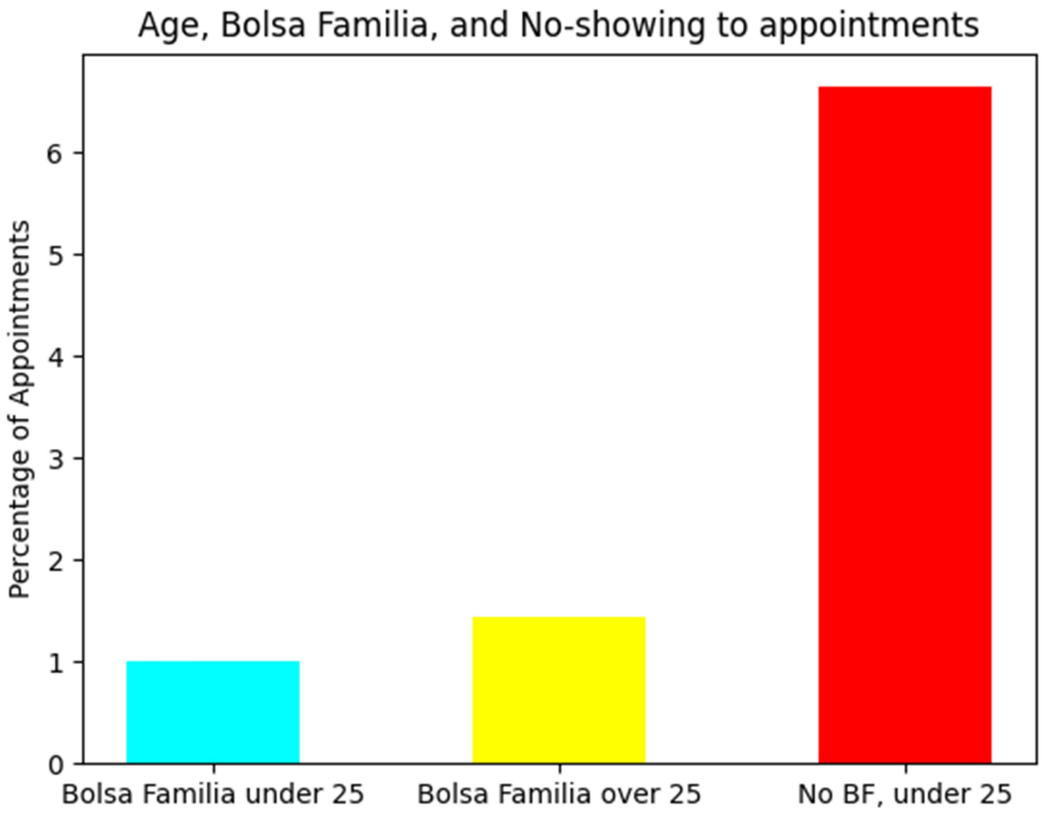
Repeat Cancel Patients and No Show Appointments

Here I look at was the number of patients who were on the list multiple times and their particular outlook. There were 23,421 patients who made 70,138 total appointments. Of these, 22,414 later showed up for 55,361 appointments. Leaving a total 1007 patients who did not show up to later appointments. This number, representing less than 1% of the total dataset shows us that repeat cancellation is not a significant factor in the no-show numbers. The graph below shows us the total no-shows among patients with multiple appointments.



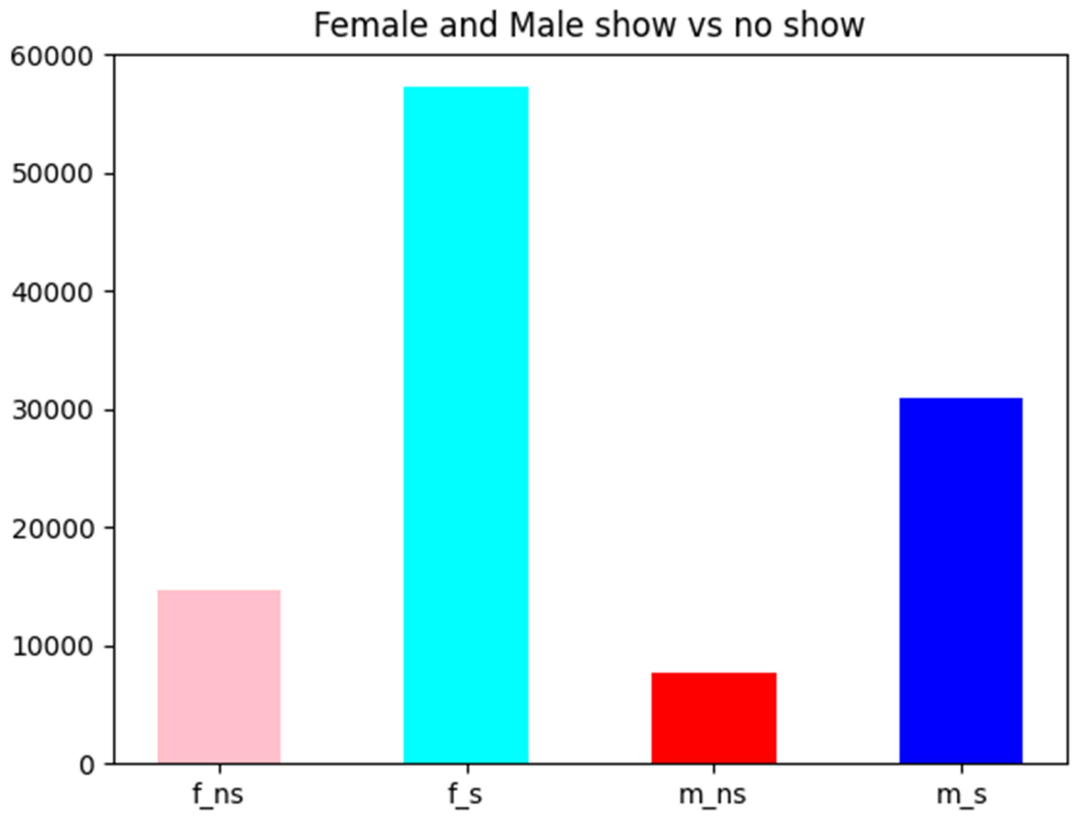
Some combined variable analysis

After combing through the appointment information, we have found four categories that seem to have some impact on the likelihood of no-show appointments, Being in certain neighborhoods increased the chances, being of certain age groups, as well as not having Bolsa Familia. When we look at the combined series of Age and Bolsa Familia we can see a significant impact on the likelihood of not showing up. With nearly 6% of all appointments being missed having people in both categories.



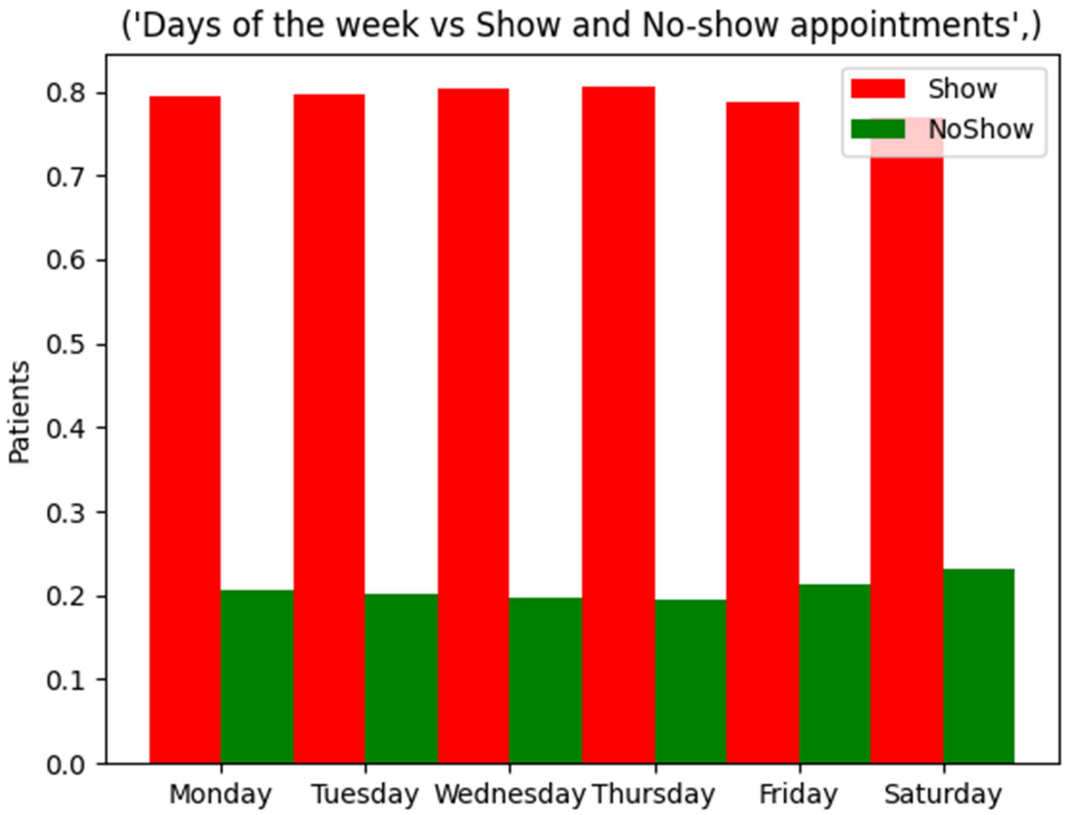
Gender and No Show Appointments

We first looked into the gender category and found that approximately 20% of all persons male or female, 20.3% of women and 19.9% of men, were part of the no-show category. (Fig 1.) The ratio of female to male no shows vs total populations is approximately the same. It is unlikely that this is a contributing factor.



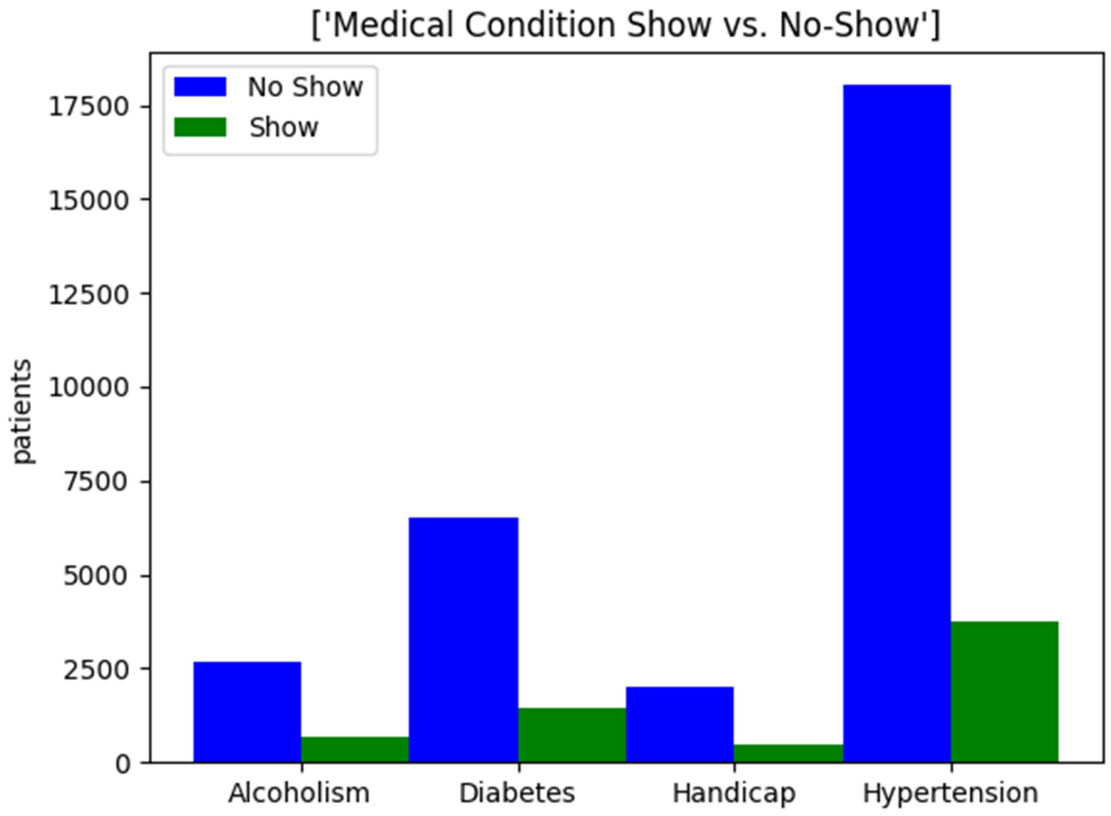
Days of the week and No Show Appointments

Another possibility to consider is that certain days of the week may be a factor in no-show appointments. The graph below shows no-show and show-up appointments compared to total appointments. While Saturday does seem to have a slight uptick in no-show appointments, 23% of Saturday appointments vs 20% on average per day the rest of the week. I am not certain that this is a contributing factor.



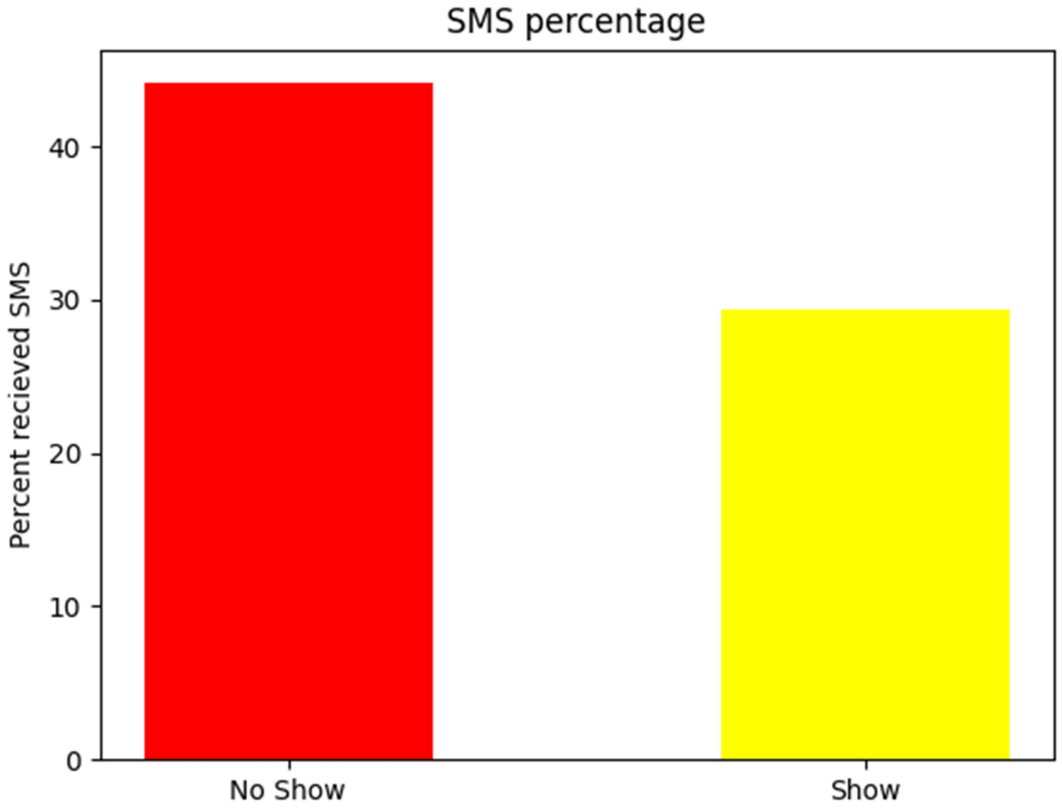
Medical Conditions and No Show Appointments

A review of the medical condition vs. ratio of no-show patients shows almost no difference in Diabetes, or Alcoholism. There was a small reduction in the number of no-shows in patients with Handicap, and a slight increase in patients with Hypertension. None of this amounts to any meaning though as the numbers are within close range of each other, <1% variance in Handicap and <4% in Hypertension.



SMS messages and No Show Appointments

As far as SMS messages showed up those seemed to have potentially negative impact, as 45% of patients who no-showed received SMS messages, as compared to 29% of those who did show up to appointments.



Conclusion

It appears as though missed medical appointments are primarily based on a combined age and scholarship status, with neighborhood, age and scholarship all having some impact on the number of no-show medical appointments. Repeat no-show patients are not a significant factor for determining if a patient will show up or not. The dominant factors in no-shows being if a patient is on Bolsa Familia, above the age of 25, or from certain neighborhoods, while things like gender, day of the week, or particular ailments, had no obvious effect. SMS messages showed a potential negative link, but not significant.

Limitations

This analysis applies only to the data provided here, and should not be taken to represent any larger group, as no verification of the sample has been done. It is possible that some of the variables interact in more complicated ways not explored in this analysis, such as SMS interactions with both age and neighborhood. Factors not included in the dataset that may have altered the outcome can include, but are not limited to, income group of the patients, familial nature of the patient, relationship of multiple patients, in particular for patients below the age of majority and requiring a guardian to care for them. Other medical conditions that may require guardianship, or other types of aid, that could have impacted the ability to show up to appointments. Even the possibility of miscommunication is unable to be verified with just the given data. Something for future research should include multiple month studies, as this sample was just one month's worth of data, and could be an outlier month. This is hinted at by the lack of no-shows from a single neighborhood.

Sources for syntax or concept research:

https://www.geeksforgeeks.org/split-pandas-dataframe-by-rows/

https://stackoverflow.com/questions/24886624/pycharm-does-not-show-plot

https://thispointer.com/pandas-find-duplicate-rows-in-a-dataframe-based-on-all-or-selected-columns-using-dataframe-duplicated-in-python/

https://www.w3resource.com/pandas/series/series-str-count.php