Project 1:
Medical Appointment No-Show Analysis

Group 1: Eddie Pellissier, Eric Lidiak, Ross Boersma, Stephanie Abegg

The Data

The dataset used in this study has over 110,000 rows of data regarding medical appointments for an unnamed medical care company in Brazil, from April-June 2016. The dataset had fourteen columns of data, including unique identifiers (Patient ID and Appointment ID), scheduling data (Scheduled Day, Appointment Day, SMS Received), demographic data (gender, age, neighborhood, Bolsa Família), health conditions (Hypertension, Diabetes, Alcoholism, Handicap), and appointment attendance (No show = Yes or No).

The dataset was found on Kaggle (https://www.kaggle.com/datasets/joniarroba/noshowappointments).

Before analysis, the dataset was cleaned. This involved:

- Correcting typos in the column names (e.g. “Hipertension” was corrected to “Hypertension” and “Handcap” was corrected to “Handicap”);
- Making column names consistent with separate words being capitalized and separated with underscores (e.g. “No-show” became “No_Show”, “SMS_received” became “SMS_Received”, “AppointmentID” became “Appointment_ID”);
- Checking for duplicate entries (there were none);
- Removing outliers (e.g. a patient with age of -1 was eliminated); and
- Making the conditions entries consistently defined as 0 or 1. (In the original dataset the conditions Hypertension, Diabetes, and Alcoholism were identified as 0 or 1 while Handicap was 0, 1, 2, 3, 4; so for consistency all of the non-zero Handicap were replaced with 1).

In addition, some new columns were added. These included

- Breaking down Scheduled Date into Date, Time, Day of Week, and Month;
- Breaking down Appointment Date into Date, Day of Week, and Month (no time was provided);
- Finding days between Scheduled Date and Appointment Date, and creating a Days Between column;
- Adding a No Show Boolean (which is a bit more versatile than the Yes/No in the original dataset);
- Finding Latitude and Longitude coordinates for the Neighborhoods (this used the Geoapify API).

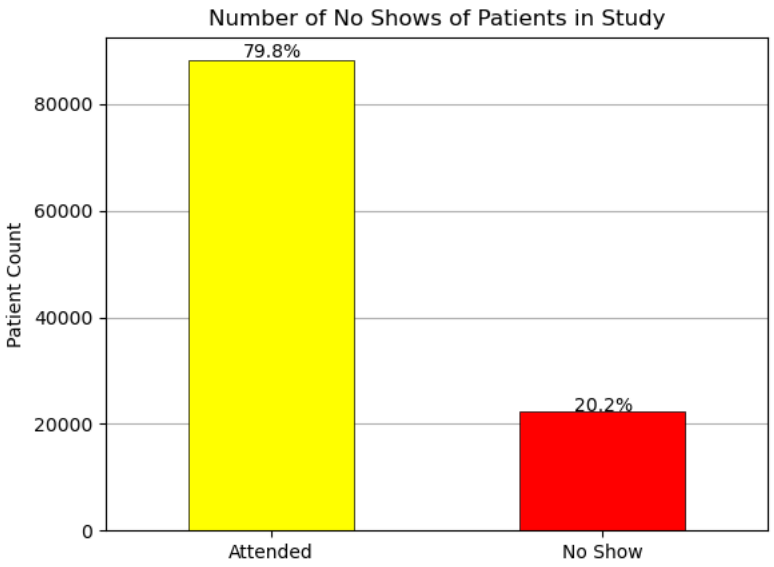
We had a couple of key assumptions:

- Cancelled appointments are not listed, i.e. the no shows are those who just did not show up.
- Each row in the dataset represents an independent datapoint. This means that appointments scheduled by a patient already in the dataset (in other words, patients who scheduled multiple appointments) were independent of other appointments scheduled by the same patient. We did confirm that the majority of patients in the dataset had only one or two appointments, so this assumption should not affect the results much.

Number of No Shows

What is the overall proportion of No Shows?

Over the three-month time span represented by the dataset, 79.8% of appointments were attended and 20.2% of the appointments were no show. This is visually shown the graph below.

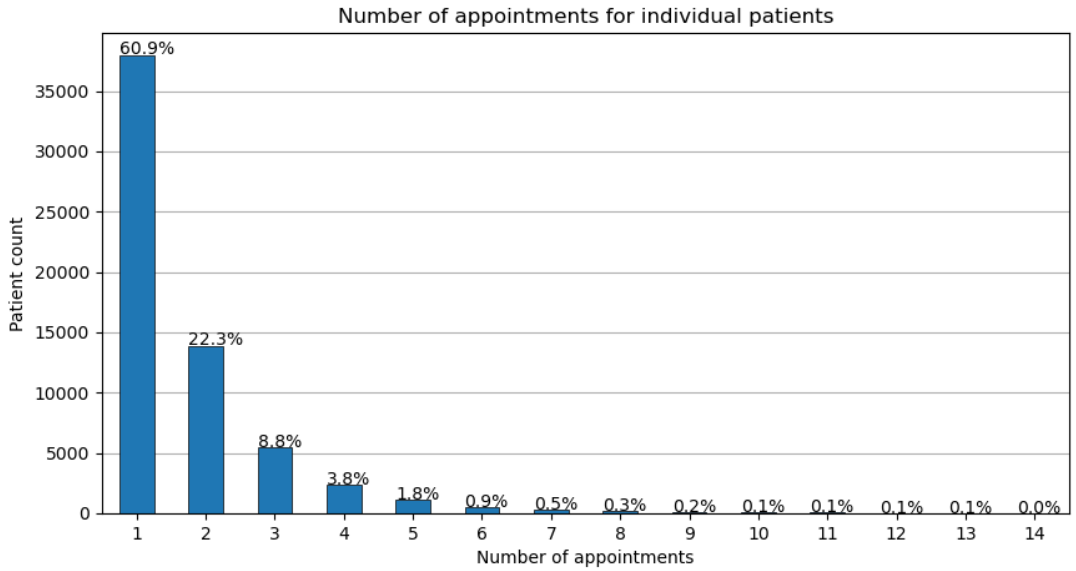


Multiple Appointments

How many patients scheduled multiple appointments?

In this analysis, we treated each entry as an independent record. To confirm this was a good simplifying assumption, we checked how many times the same person appeared in the dataset (i.e. had multiple appointments).

The bar graph below shows that during the months of April-June 2016, over 61% of the patients in the dataset had only one appointment, 83% had one or two appointments, and 99% had seven or fewer appointments.



What are the outliers?

Thirteen different patients had more than 50 appointments during the three-month timeframe of the dataset, and the greatest number of appointments a single patient had during the three-month timeframe of the dataset 88 appointments. Many of these appointments were scheduled on the same day as other appointments, suggesting that seeing different doctors or areas of the clinic on the same day constitutes different appointments. The table to the right shows the outliers.

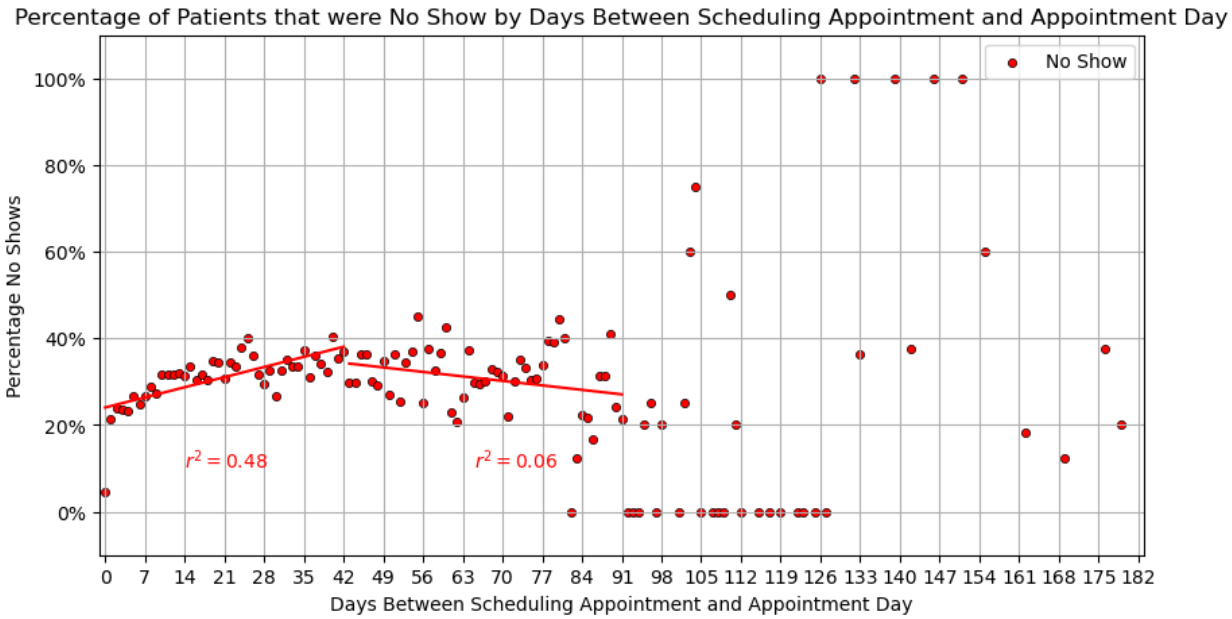
#_of_Visits	Patient_count
50	1
51	1
54	1
55	1
57	1
62	4
65	1
70	1
84	1
88	1

The overall conclusion is that it is a valid assumption to treat each appointment individually. While it is possible that an individual patient who has multiple appointments can bias the data towards their demographic and behavior, the number of such patients is small compared to the size of the dataset, and not likely to impact averages in a meaningful way.

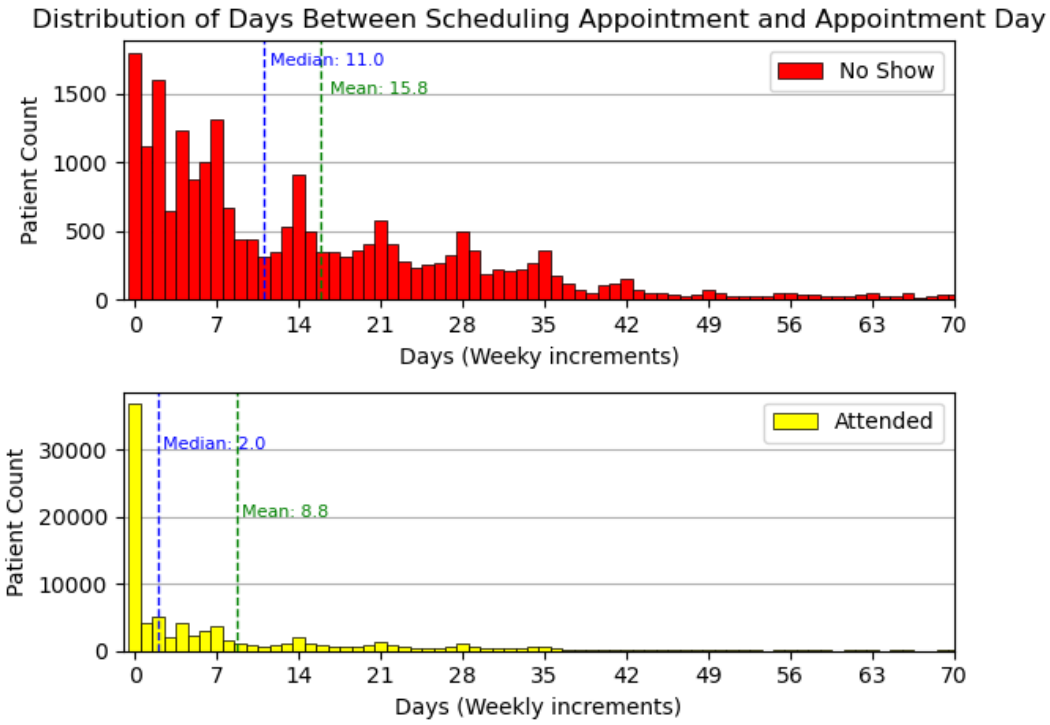
Appointment Scheduling

Does the time between scheduling an appointment and appointment date impact appointment attendance?

One question we investigated was how the time between scheduling an appointment and the appointment data impacts appointment attendance. The following scatterplot shows the percentage of no shows vs. days between scheduling appointment and appointment day. It is no surprise to discover that same-day appointments have a very low rate of no shows, whereas the proportion of no shows increases as the days between scheduling and appointment increases. After about 6 weeks, the proportion of no shows levels out.



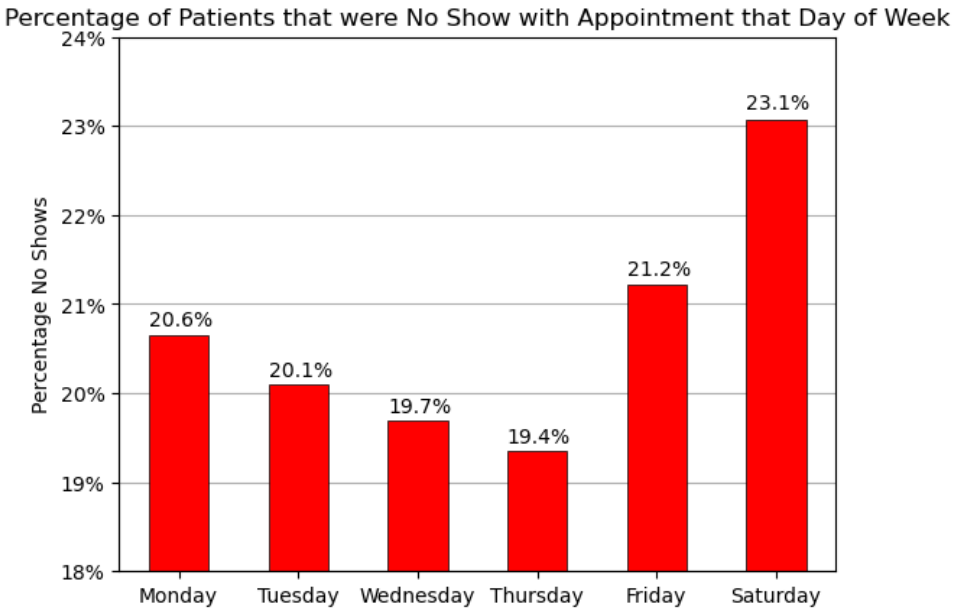
The histograms below show the distribution of days between scheduling appointment and appointment day, separated for no show and attended appointments. The large spike on the attended appointment distribution at 0 days indicates the high likelihood of appointment attendance for same-day appointments. On average, someone who misses a scheduled appointment scheduled it 15.8 days out. On average, someone who attended an appointment scheduled it 8.8 days out.



The overall general conclusion is that scheduling appointments closer to the appointment date results in a higher chance of attendance.

Does the day of the week of the appointment impact no-shows?

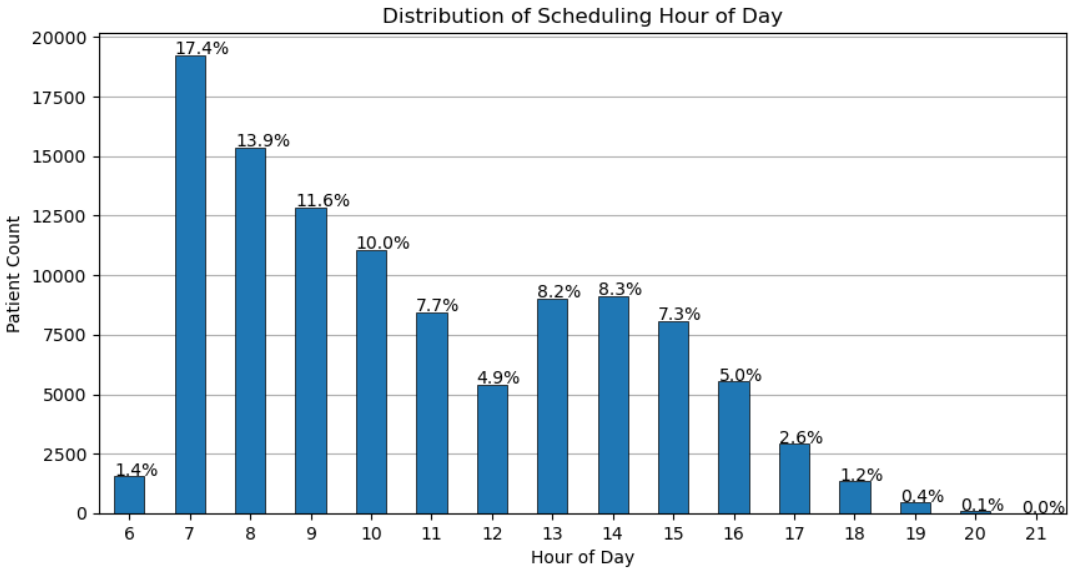
The bar chart below shows the proportion of no shows by day of week. Friday and Saturday have the highest percent of no-shows. Appointments on Monday-Thursday had a much higher chance of being attended, with the no show rate decreasing up until Thursday.



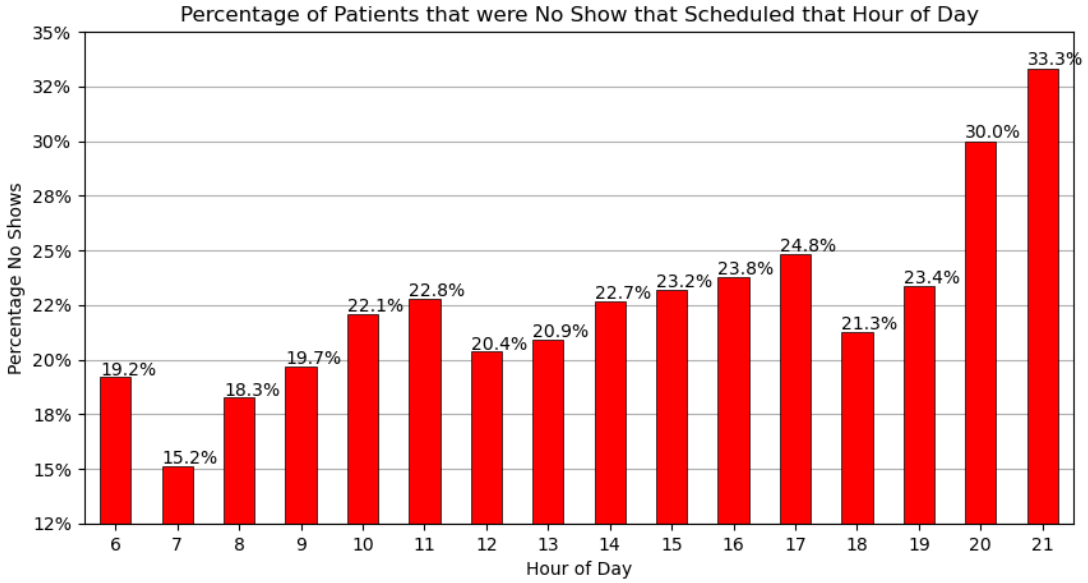
Does the time of day that an appointment is scheduled impact the likelihood of a no-show?

The dataset had the time of day that the appointment was scheduled, but unfortunately not the time of day of the actual appointment. So, we investigated the former and only wished we could investigate the latter.

The histogram below shows the distribution of scheduling times. Morning from 7-11 am is the most common time for people to schedule appointments, with a dip in scheduling at lunch hour, rising back up in early afternoon, and then fewer and fewer scheduling appointments towards later afternoon and evening.



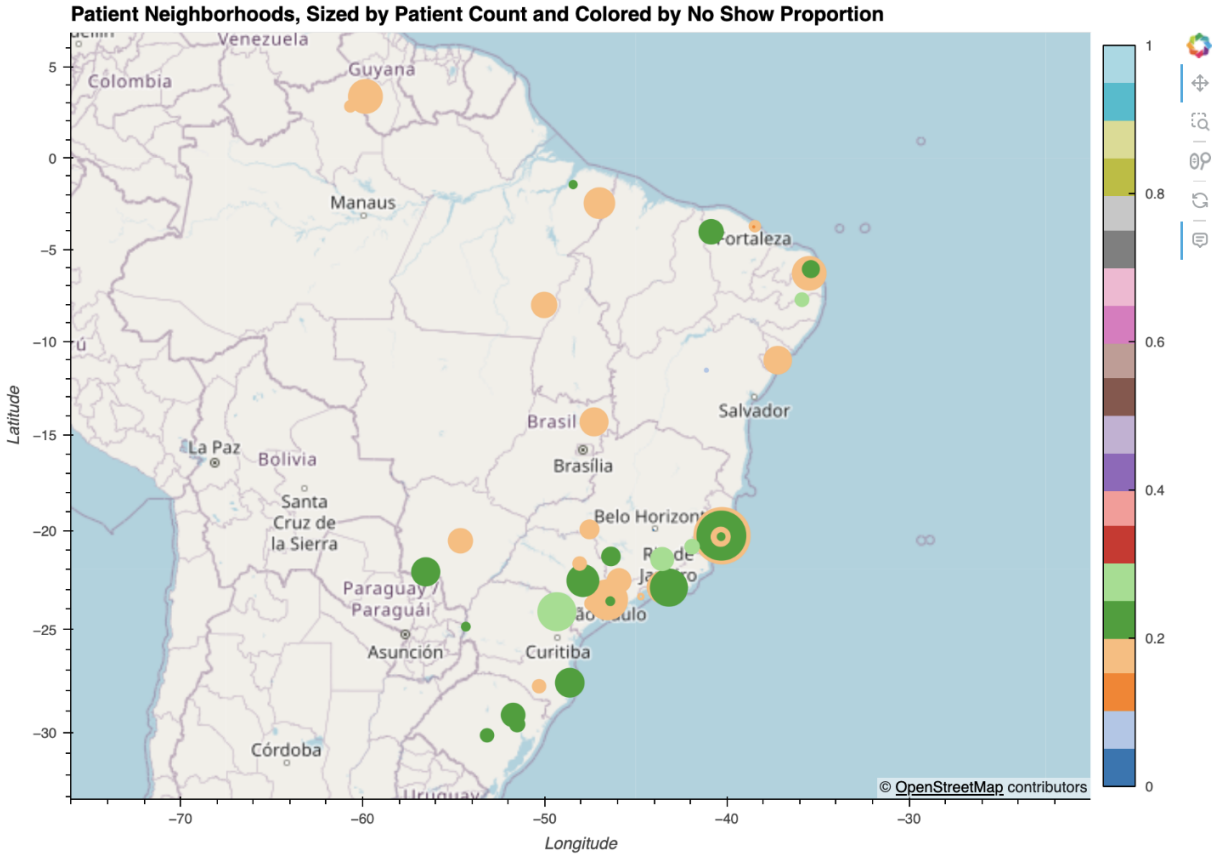
Next we looked at the proportion of no shows for each hour of the day. This graph is shown below and although it has the same x-axis as the distribution of scheduling by hour of day above, it look quite different. We see that scheduling an appointment earlier in the day is more likely to result in an attended appointment. People who scheduled between 6am and 9am attended 81.9% of their appointments (no show 18.1%), while people who scheduled after 7pm attended only 68.3% of the time (no show 31.7%).



Location

Does location have an impact the likelihood of a no-show?

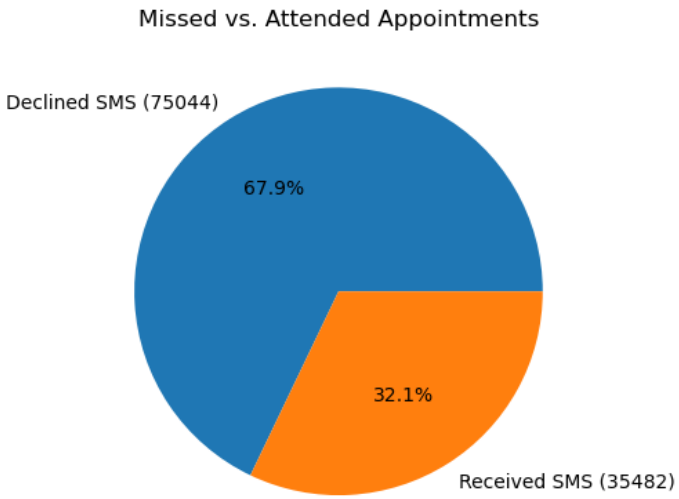
The dataset provided the neighborhood of each appointment. The Geoapify API was used to get the Latitude and Longitude of each neighborhood, and therefore gain the ability to plot no show statistics on a map. The map below shows that urban areas provided the largest sample size (as expected). The proportion of no-shows is greater in more concentrated urban areas.



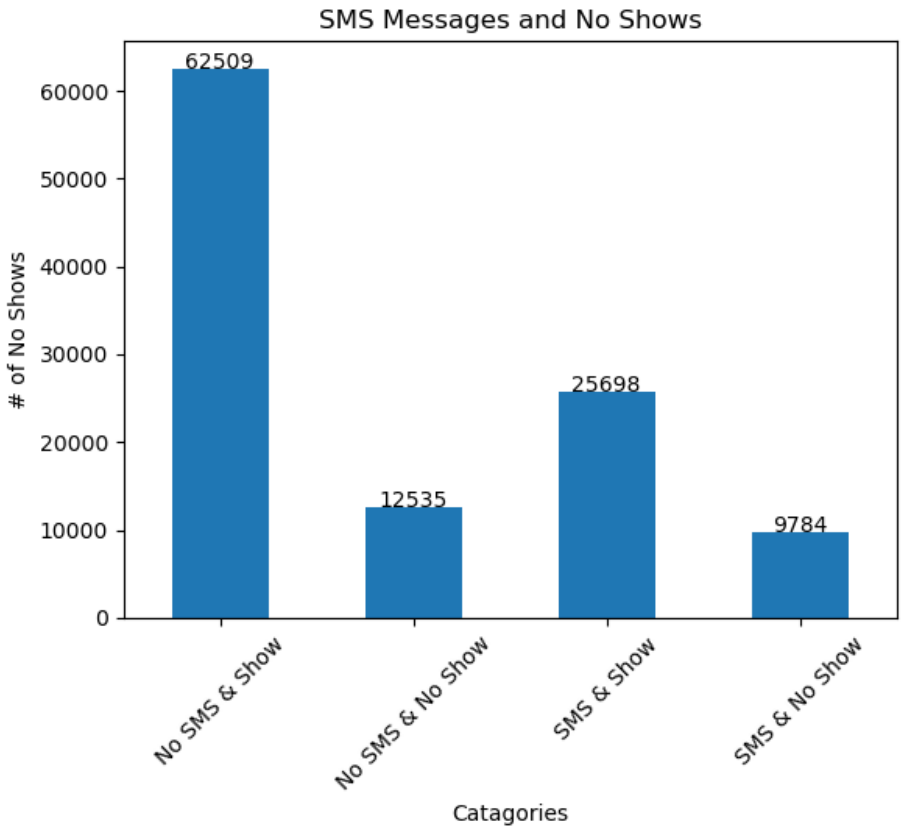
SMS Messaging

Can text messaging affect appointment attendance?

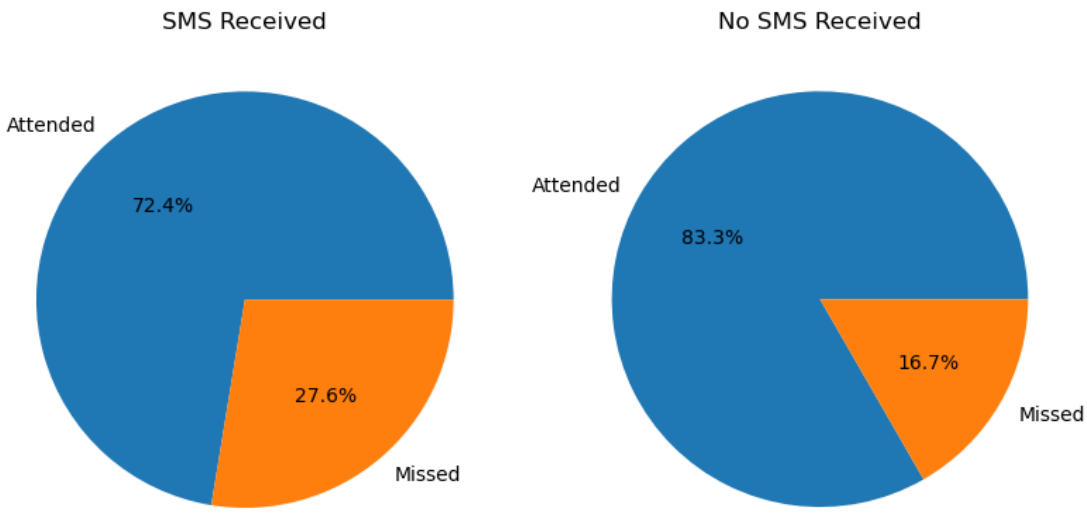
Only 32% opted in and received SMS text message appointment reminders, while 68% opted out of SMS text messaging. The pie graph below gives a nice visual.



The largest category of people were those who opted out of SMS text messaging and still showed up. About 12,500 patients who did not receive an SMS reminder were no show, while about 9,800 patients who received an SMS reminder were no show. The bar chart below gives a nice visual.



Note that the bar graph above does not directly indicate the affect of the SMS message. In order to determine if sending the SMS message had a positive affect on appointment attendance, we must look at the proportions. The pie charts below give the proportions on missed and attended appointments for patients who received an SMS and patients who opted out of an SMS.

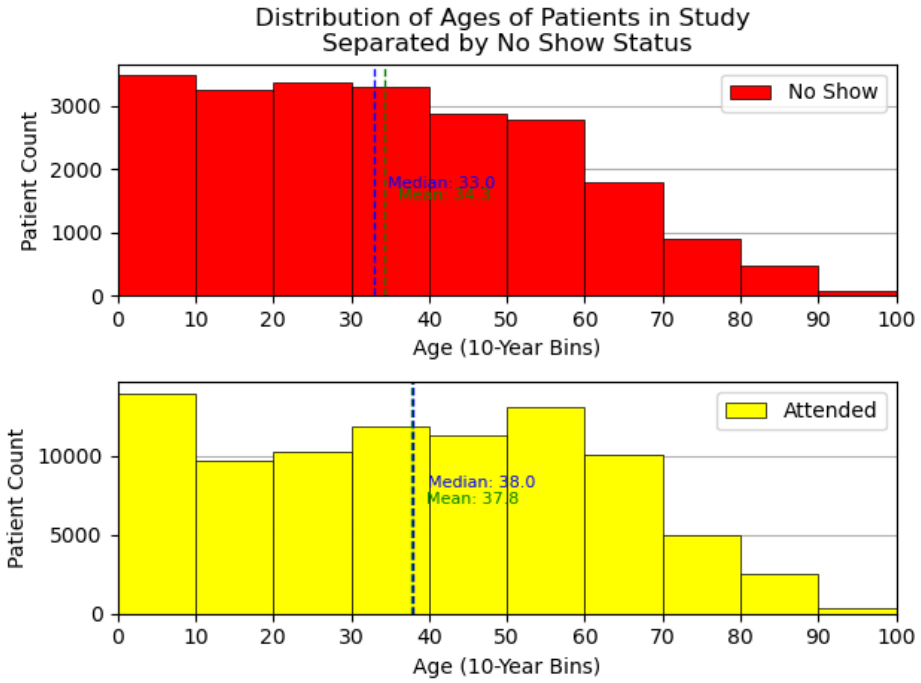


The pie charts indicate that 27.6% of people who received a SMS message still did not attend their appointments, whereas 16.7% of people who did not receive a text message did not attend their appointments. This means that people who received an SMS message were 11% *more* likely to miss their scheduled appointments. This could be that the type of person who opts into receive SMS messages is inherently more likely to miss an appointment.

Demographics

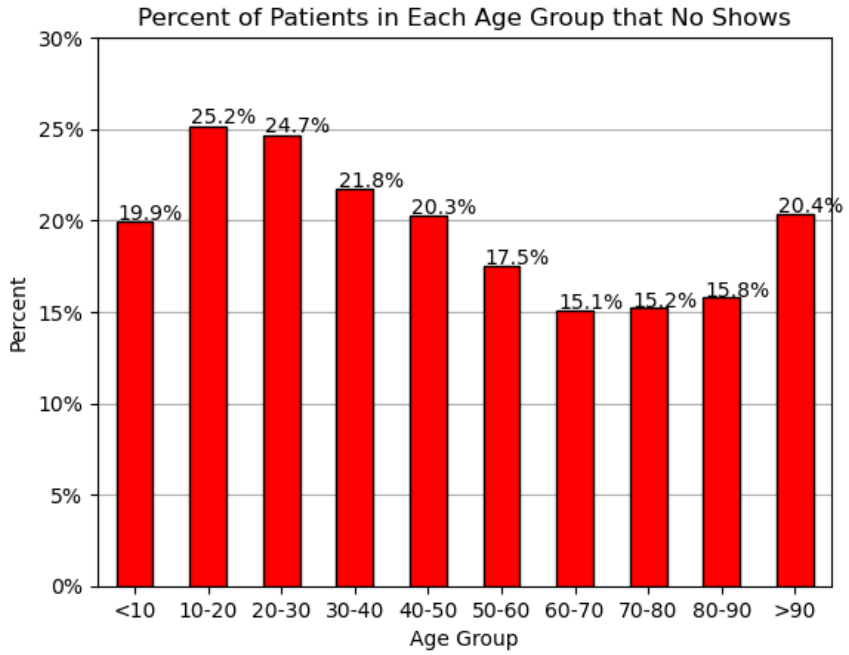
What impact does age have on attendance?

The following histogram shows the distribution of ages of people who attended appointments and were no show to appointments.



The distributions show that there was a good distribution of ages in the study and that people of all ages missed and attended appointments, and that children (ages 0-10) have the highest volume of appointments of any 10-year age group. The main difference between the distributions for the no shows and attended patients is the median and mean ages: the median age of someone who misses an appointment is 33 years old, which is five years younger than the median age of people who attend appointments (38 years old); similarly, the average age is 3.5 years younger for no shows.

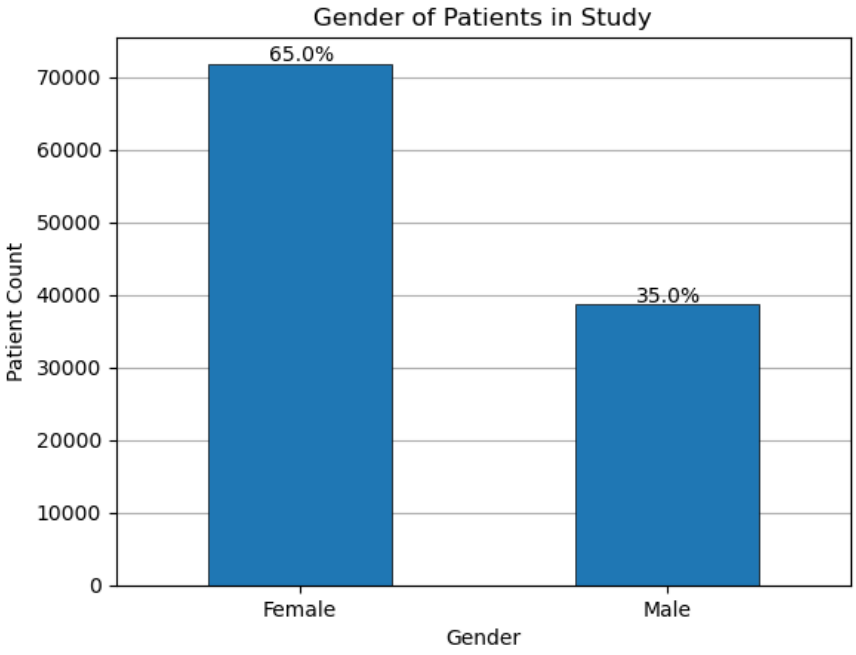
The lower median and mean ages for no shows suggests that no shows tend to be of a slightly younger age group and that older age groups tend to attend appointments. To investigate this further, we crated the following bar chart, which shows the proportion of no-shows in each age group, on 10-year age increments.



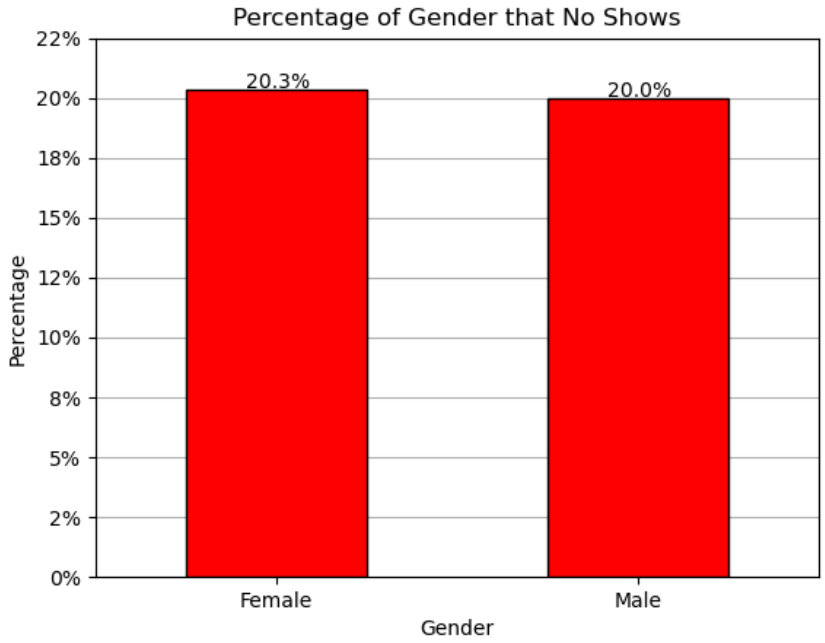
The bar chart shows that people between the ages of 10-30 are the most likely age group to miss scheduled appointments. In general, children and people aged 60-90 are the least likely to miss appointments. For the elderly (over 90 years old), the no show rate increases again, which could have some more morbid reasons.

Are men or women more likely to attend their appointments?

The following histogram shows the distribution of genders of patients in the study. A majority (65%) of the patients who scheduled appointments for this particular medical care company in Brazil were female.

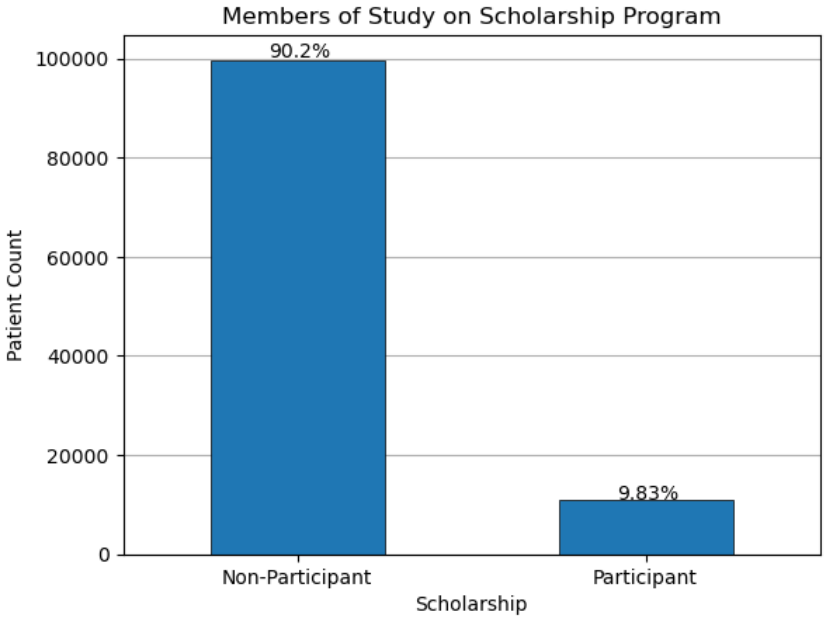


However, despite the dichotomy in the gender of patients in the study, the difference in no shows between the two genders was negligible.

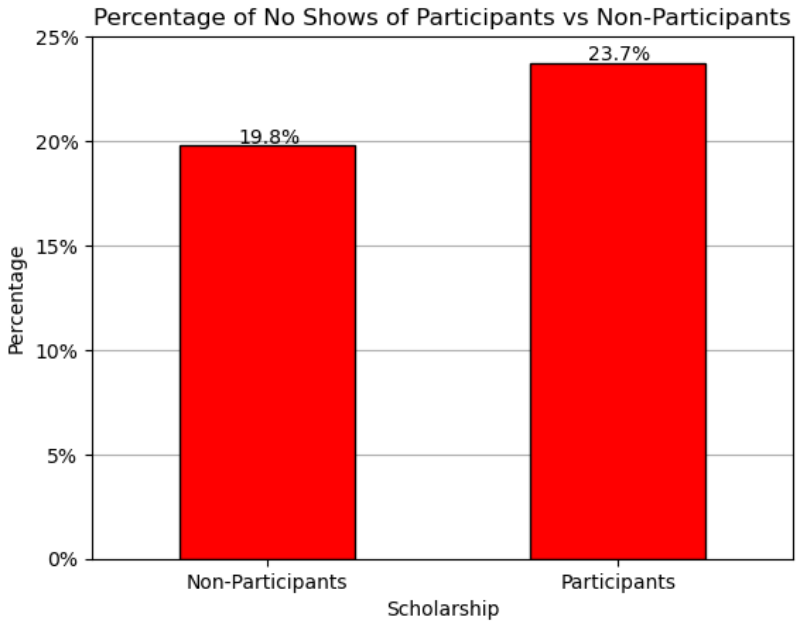


Do patients on Brazil's Bolsa Família Program attend appointments at a higher rate?

The Bolsa Família program is a Brazilian social welfare program launched in 2003. Bolsa Família has been credited with significantly reducing poverty rates and improving educational and health outcomes for children in Brazil. Roughly 1 in 10 people who had scheduled appointments were participants in the program, as shown in the bar chart below.



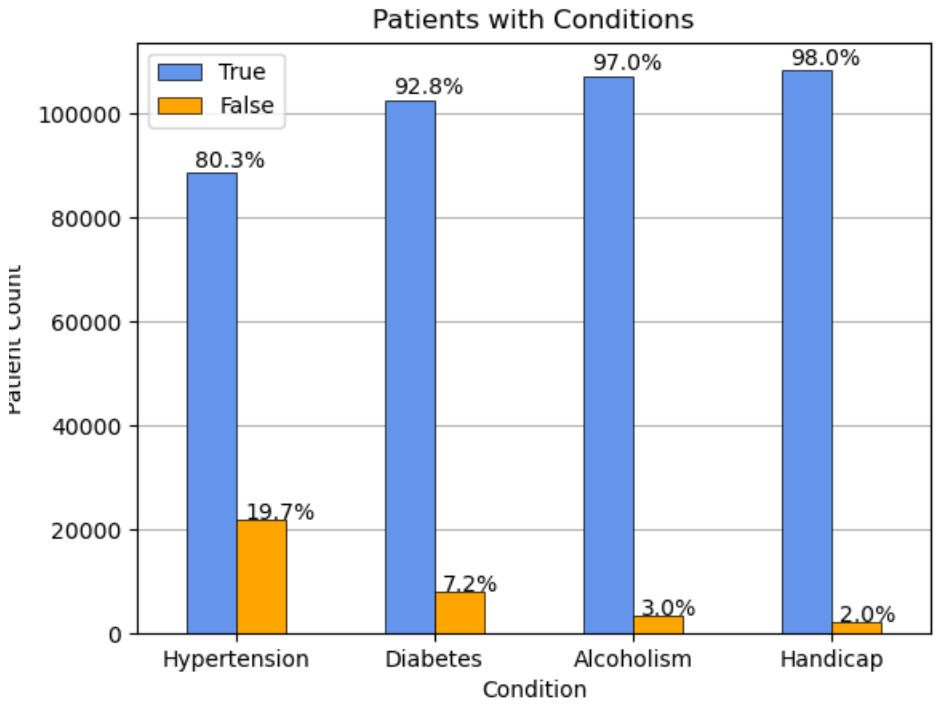
Though incentivized to attend, participants in the Bolsa Família program were more likely to miss appointments, as shown in the bar chart below, which gives the percentage of patients that were no shows, broken down by non-participants and participants in the program. This likely highlights the additional barriers that low income individuals experience when trying to get to an appointment.



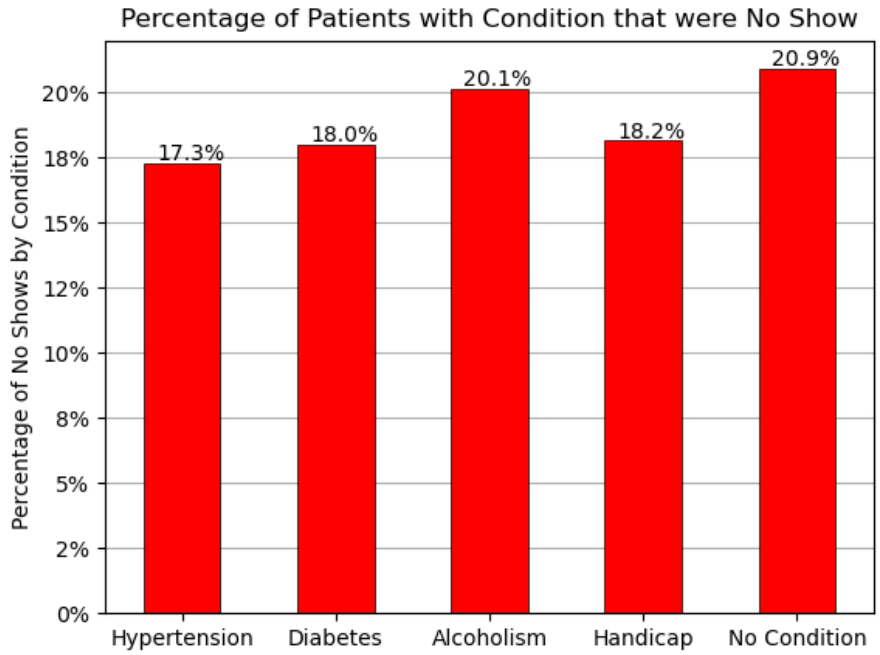
Health Conditions

How do special health conditions impact attendance?

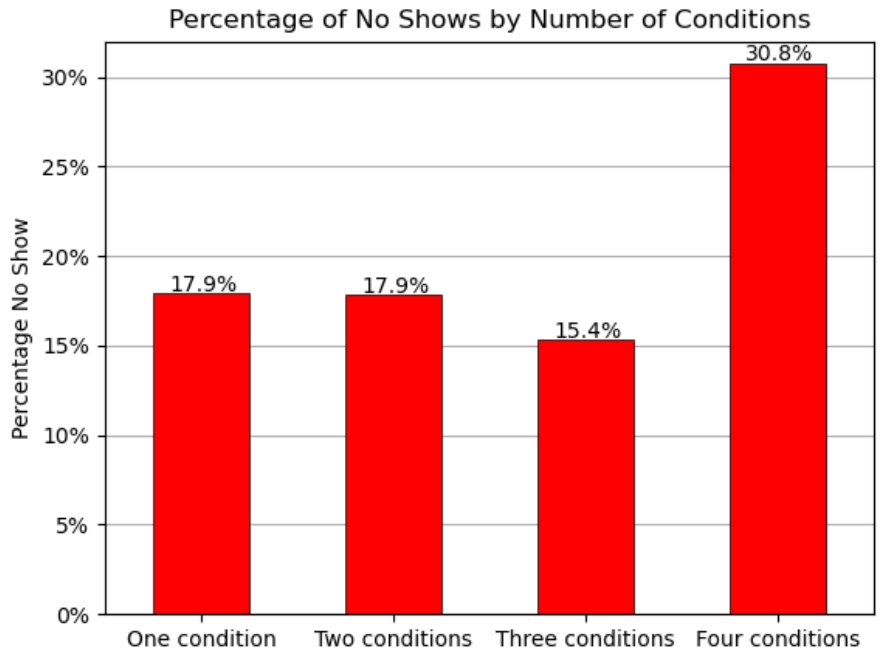
The dataset had four special health conditions: Hypertension, Diabetes, Alcoholism, and Handicap. The following bar chart gives a breakdown of the number of patients with and without the specified conditions. Hypertension impacted about 20% of the patients, while only about 2% of all patients had what the data considered to be a handicap.



As the following bar chart shows, people with special conditions were more likely to attend appointments than those who did not have special conditions. Of those with special conditions, people with Hypertension had the highest chance of showing up, while people with Alcoholism showed up at about the same rate as people with no special conditions.



We also investigated the effect on attendance of having multiple conditions. As shown in the following bar chart, people with three conditions were most likely to attend their appointments, which could be because they were in greater need of medical help. Interestingly, people with four conditions were significantly more likely to miss their appointments than people who had three or less special conditions. This could be because these people tend to be "box checkers" (i.e. checking all boxes on intake forms) and inherently the type to is appointments, or they could be genuinely quite medically compromised and have difficulty attending appointments. It is important to note that there were only 5 unique patients with four conditions, whereas there were anywhere from 330 to 10047 unique patients with one, two, or three conditions, which makes the four-condition data less reliable than the data for fewer conditions.



Further Study

What other information would we have liked to have?

There are several variables that, if added to the dataset, could enhance our analysis. Here are some examples:

- Have a larger timespan of data, like for a few years instead of just three months of 2016. This would allow inclusion of a seasonal analysis of no shows.
- Whereas we had the time of day at which each appointment was scheduled, we did not have the appointment time of day. If we did, we could have investigated how the time of day plays a role in appointment attendance.
- Have more demographic data, such as income, access to transportation, employment information, children, marriage status, etc. These would provide further interesting analyses on what influences appointment attendance.
- Give the type of appointments people had (e.g. one-year checkup, medical procedure, trauma, etc.).

- Give the clinic name, to answer the question of if individual clinics have different appointment attendance rates (and why?).

What other questions could we answer with more time?

With limited time (i.e. two weeks and working full-time jobs on the side) to do this project, we had to narrow our focus to doing an in-depth analysis of one particular variable of the data (i.e. no shows) rather than a superficial analysis of several variables. Here are some other questions we could answer with more time:

- Analyze the combined effect of different variables. This would involve a multi-linear regression. For example, there is likely a correlation between age and opting into the SMS program and presence of conditions, so looking at these variables together and the effect on appointment attendance would be more instructive than looking at them individually.
- We could investigate the question of why people in urban areas were more likely to miss their appointments. For example, do urban areas tend to have a different age distribution?
- Most of our analysis focused on finding total counts and proportions of no shows across various metrics (e.g. different ages, genders, locations, conditions, appointment scheduling time spans, etc.). Given more time, we would do more statistical analyses (hypothesis testing, multilinear regression, t-test, chi square, etc.)
- Investigate the effect of removing from the dataset the patients who scheduled multiple appointments. Also, it might also be interesting to look at appointment frequency as another indicator of missed appointments (i.e. is a patient who has more appointments less likely to be a no show?).