

THE MEDICAL APPOINTMENT NO-SHOW

Project Report: Data Visualization, Fall 2019

Charith Musku
Bhargav Kandlagunta
Bhagya Reddy

1 Abstract

The problem of patients not showing up for their appointments has been a critical problem for most of the clinical practices across the nation. One study estimates the healthcare industry loses about \$150 billion to patient appointment no shows each year | about \$200 in co-pays, reimbursements, and overhead for every hour-long time slot that goes unfilled. [8] There are many studies trying to understand the main reasons and effective measure to reduce the no-show rate. We want to explore the problem using an open-source dataset for medical appointments no show data available on kaggle [3]. The objective of the project is to determine the most common factors causing no-show of patients using exploratory data analysis and try to model the data with these features to predict if a patient shows up or not for their respective appointment using regression.

2 Introduction

No-show at hospitals is an important national problem; the public healthcare sector is trying to cope with. A no-show in the healthcare sector is an appointment, where the patient or client did not show-up or try to call the hospital to cancel the appointment or reschedule the appointment. Missed appointments are associated with poorer patient outcomes and cost the hospitals dearly. Delayed testing potentially puts patients in danger. Missed screening or patient no-show may result in delayed disease detection. Reducing no-show rates can diminish cost and improve quality of health care delivery. Thus, no-shows reduce scheduling capacity, contribute to inefficiency, lower the quality of care, and negatively affect the working environment for providers and staff.

2.1 Motivation

No-show at hospitals is an important worldwide problem the public healthcare sector is trying to cope with. Lin, Muthuraman, and Lawley (2011) reported patient no-show rates from 22\% to more than 50\%, especially prevalent in mental health, paediatrics and dentistry. It comes as no small surprise that reducing the rate of no-shows has become a priority in the United States and around the world. The first step to solving the problem of missed appointments is identifying why a patient skips a scheduled visit in the first place. What trends are there among patients with higher absence rates? Are there demographic indicators or perhaps time-variant relationships hiding in the data? Ultimately, these are the questions that drove us to take up this project for exploratory data analysis [4]. This research attempts to understand the reason behind patient absenteeism by studying patients' demographic factors, their environmental factors, their behaviour and the relation between these factors that lead to no-show.

With the existing financial challenges to deliver quality health care, it is crucial to find solution to this problem to provide better access to care [6]. One system level approach

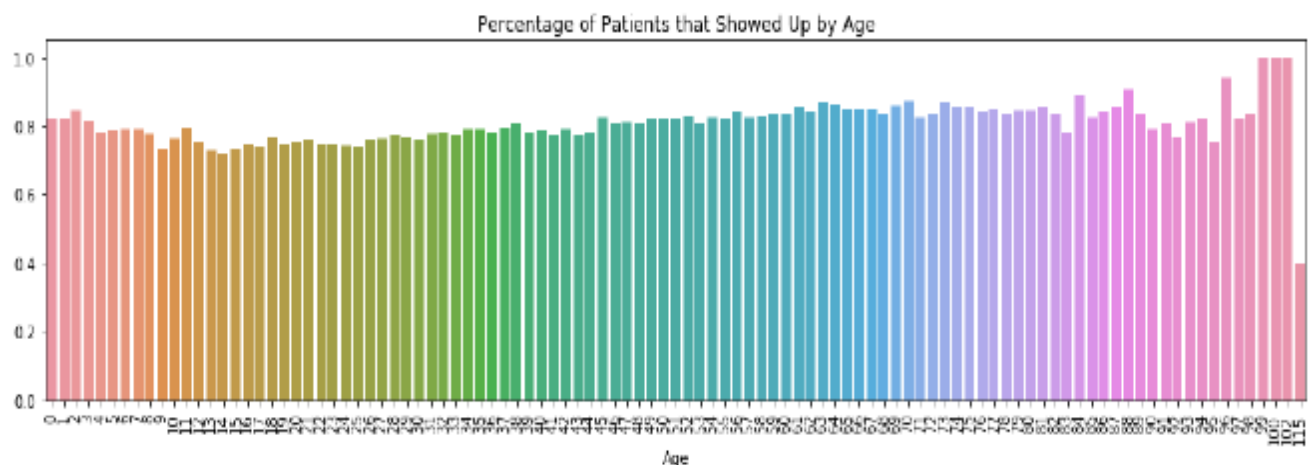
proposed to address the needs of access is to optimize the scheduling systems [7]. For example, reducing the number of missed appointments is crucial as when appointment slots go unused it effectively reduces access to others in need of an appointment. Also there will be a huge impact on the waiting times and delays for other patients because of no-shows along with increased health care costs and decreased access to care.

2.2 Background and Relevant work

The problem of patients missing the appointments has been a very serious problem even though it seems like a mundane thing. A recent survey shows that 42\% of people miss their appointments in general [5]. There have been many studies to figure out the factors that lead to this problem and also to try building solutions to tackle it effectively. From the findings it is clear that the reasons are more complex than expected [10]. Financial, transportation, wait times [2], poor medical literacy, language barriers are few of the reasons to mention. Many of these have limited the number of features for consideration of modelling the data which will not help in understanding the broader context of the reasons for the actual problem. Different datasets, features and volumes have been studied in trying to understand the cause of no-shows in varying health care settings.

For example, a large-scale no-show modelling of a Veterans Affairs (VA) outpatient clinic included 555,183 patients, which scheduled 25,050,479 appointments; however, the study only considered a few variables such as the patient gender, the date of the appointment, and new versus established patients [1]. Another study concentrated on limited number of factors and built a model to observe that age of the patient and the number of days between the scheduling and actual appointment mattered the most using a dataset of 30,000 samples and random forest model. Regression modelling was used by the most to build an effective predictive model. Some studies have incorporated neural networks, naive bayes approaches to try and build more complex models taking in large feature set.

Since the dataset used here is from kaggle, there are a lot of interesting EDA techniques and visualization kernels corresponding to this particular dataset. One particular kernel had a variety of data cleaning and visualization methods considering each feature and comparing the correlations with rest of them using python. A lot of graph types were used to show how the data structures and insights. Here is a sample usage of colored vertical bar graph to represent the percentage of patient no-show data by age [9].



3 Objectives

Using exploratory data analysis methods on the Medical Appointment No Show dataset (from Kaggle), we aim to answer the following research questions:

- What factors are most likely to determine whether a patient shows up to their scheduled doctor's appointment?
- How is the absence/no-show to a scheduled appointment dependent on the general characteristics and behavior patterns of the patient?
- Can we predict whether a patient would show up or not by taking the aforementioned variables as explanatory variables into consideration?

4 Dataset

Dataset Source: Kaggle Competition - Medical Appointment No Shows [3]

This dataset is drawn from 300,000 primary physician visits in Brazil across 2014 and 2015. The information about the appointment was labelled such as, when the patient scheduled the appointment and then the patient has either attended or not. The information about the appointment included demographic data, time data, and conditions concerning the reason for the visit.

We included a total of 15 variables from the original data. The variables and the description of the values are as follows:

- Age: Integer age of patient
- Gender: M or F
- Appointment Reservation Date: Date and time for which the appointment was made
- Appointment Date: Date of appointment without time
- DayOfTheWeek: Day of the week of appointment
- Status: Show-up or No-Show
- Diabetes: 0 or 1 for condition
- Alcoholism: 0 or 1 for condition
- Handicap: 0 or 1 for condition
- Tuberculosis: 0 or 1 for condition
- Smoker: 0 or 1 for condition / non-smoker
- Scholarship: 0 or 1 indicating whether the family of the patient takes part in the Bolsa Familia Program, an initiative that provides families with small cash transfers in exchange for keeping children in school and completing health care visits.
- SMS Reminder: 0 ,1 ,2 for number of text message reminders sent to patient about appointment
- Waiting Time: Integer number of days between when the appointment was made and when the appointment took place.

5. Data Exploration

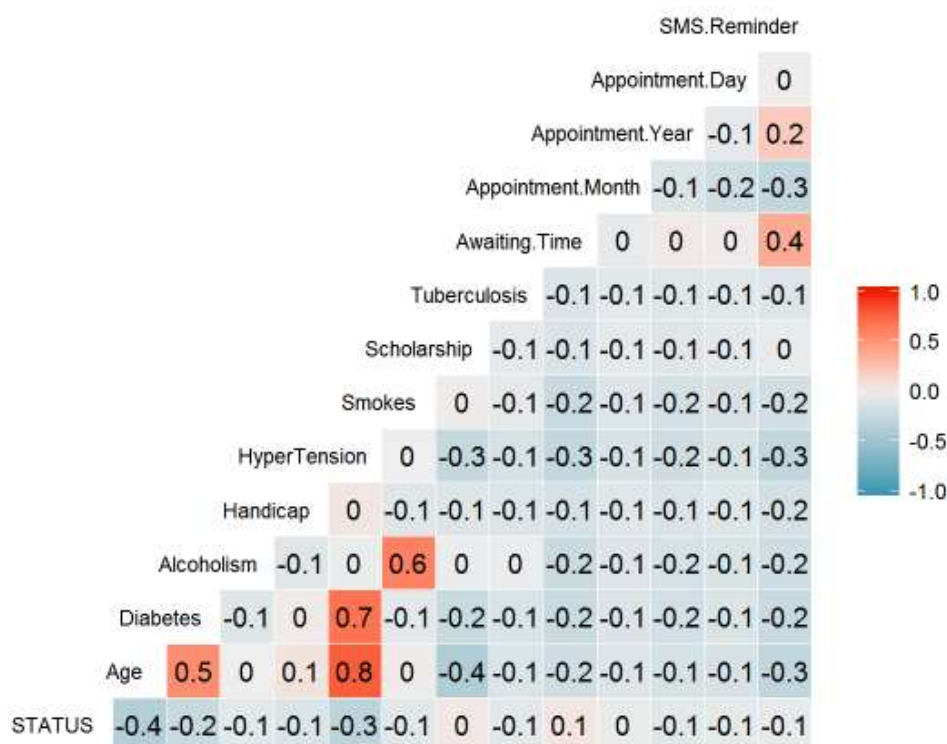
5.1 What factors are most likely to determine whether a patient shows up to their scheduled doctor's appointment?

The first step is to find the correlations between all the columns to see if any variable stands out as particularly compelling. The correlation matrix shows the statistical measure that indicates the extent to which 2 or more variables fluctuate together. A positive correlation indicates the extent to which those variables increase or decrease in parallel, whereas a negative correlation indicates the extent to which one variable increases as the other decreases. The Pearson correlation coefficient (plotted here), is a measure of how linearly dependent two variables are on each other. The coefficient value is between -1 and +1, and two variables that are perfectly linearly correlated will have a value of +1.

From these correlation values, there are no standouts that are strong linear predictors of whether a patient will miss a visit (given by the status column). It appears that there are other strong and moderate relationships between age and hypertension, Diabetes and Hypertension, Diabetes and Age, Age and Smokes and Awaiting Time and SMS Reminder. However, there are no strong correlations between any variable with the target variable, 'Status'.

Correlation Matrix

Correlation of Factors w.r.t "Status" variable

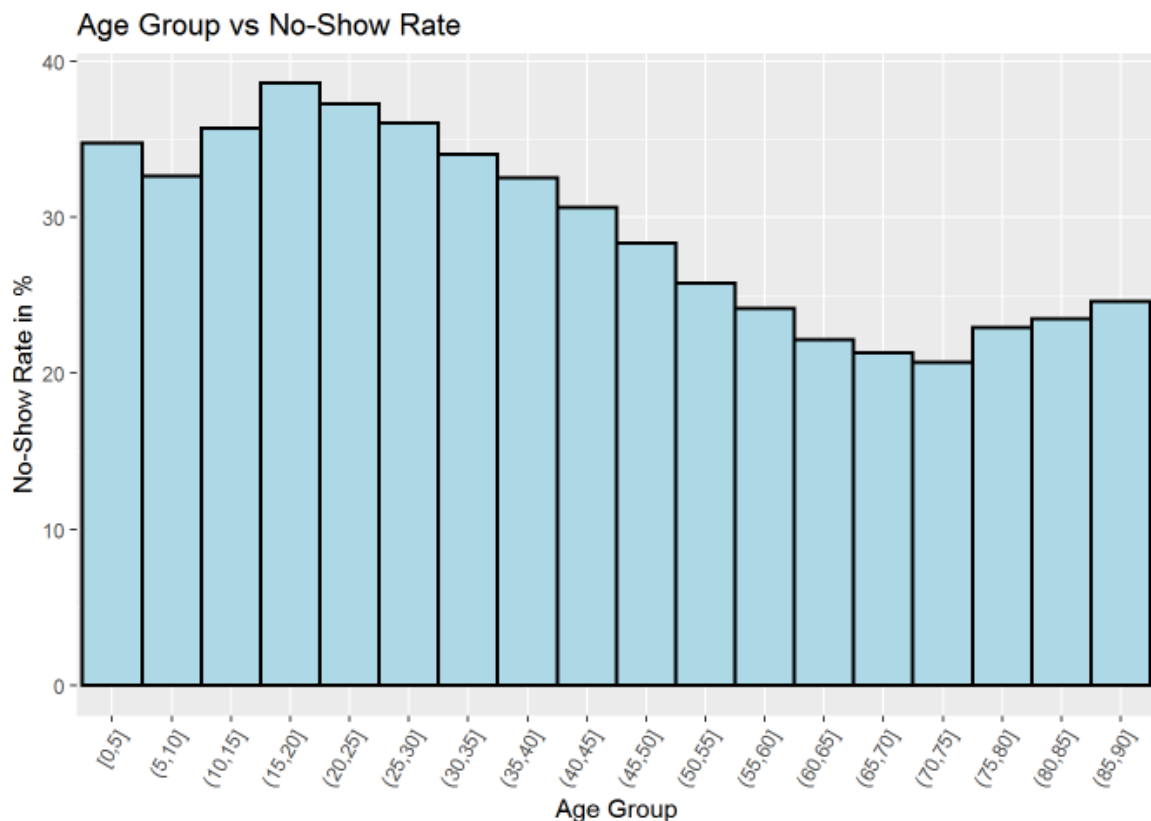


Nevertheless, there could be some meaningful relationships to extract from the data by plotting them individually against each other in the form of univariate or bivariate graphs. Our approach will be to group the data by various fields and then determine if the average absence rate shows a relationship with the fields.

Age of patient:

We were interested in whether age is correlated with absence rate. The data was binned in five years increments to highlight the years with the highest missed appointment rate. We got a bimodal graph with a major peak at 15-20 years (almost 40%) and another minor peak at 85-90 years (about 25%). It exhibits a dip at 70-75 years (around 20%). Thus, the worst group in terms of absence rate is 15–20 years old patients and the best group for attendance is 70–75 years old patients. The correlation between ages and absence rate is strongly negative

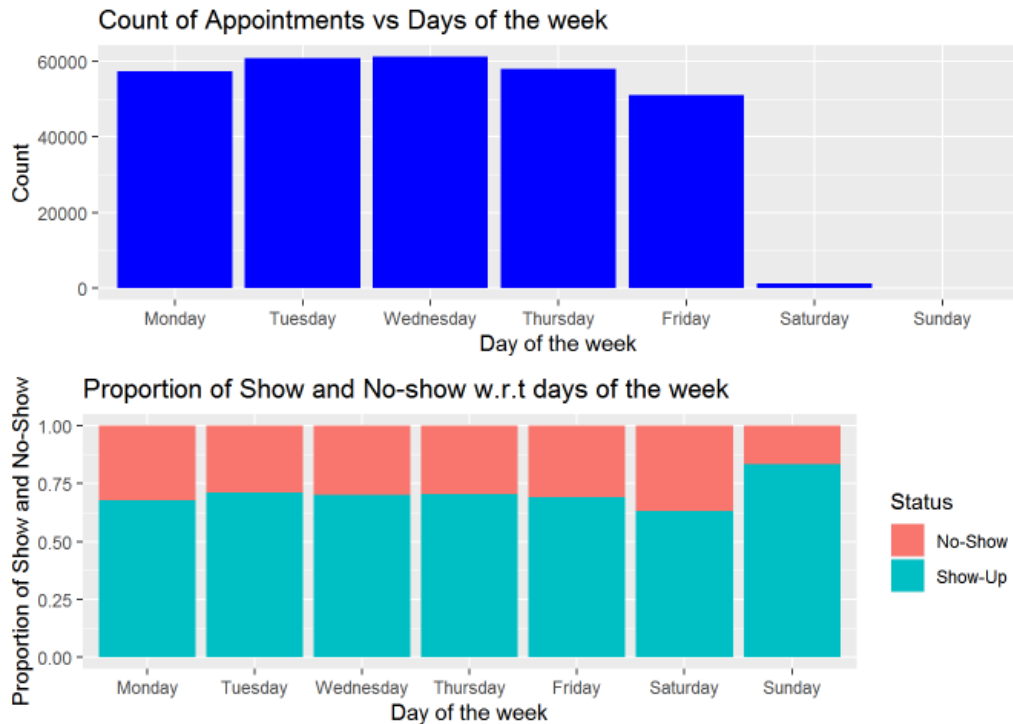
and indicates that as the age of the patient increases, statistically, that patient is less likely to miss a scheduled doctor's appointment.



As expected, the youngest and oldest patients tend to attend their appointments since they pay more attention to their health because of their vulnerability to falling sick. In contrast, patients in their youth or middle age are generally healthier and tend to neglect their health owing to their busy schedules. If we look at very old people in the bracket 75-90 years, we observe a rise in no show rate probably because they need someone to accompany them for their appointments.

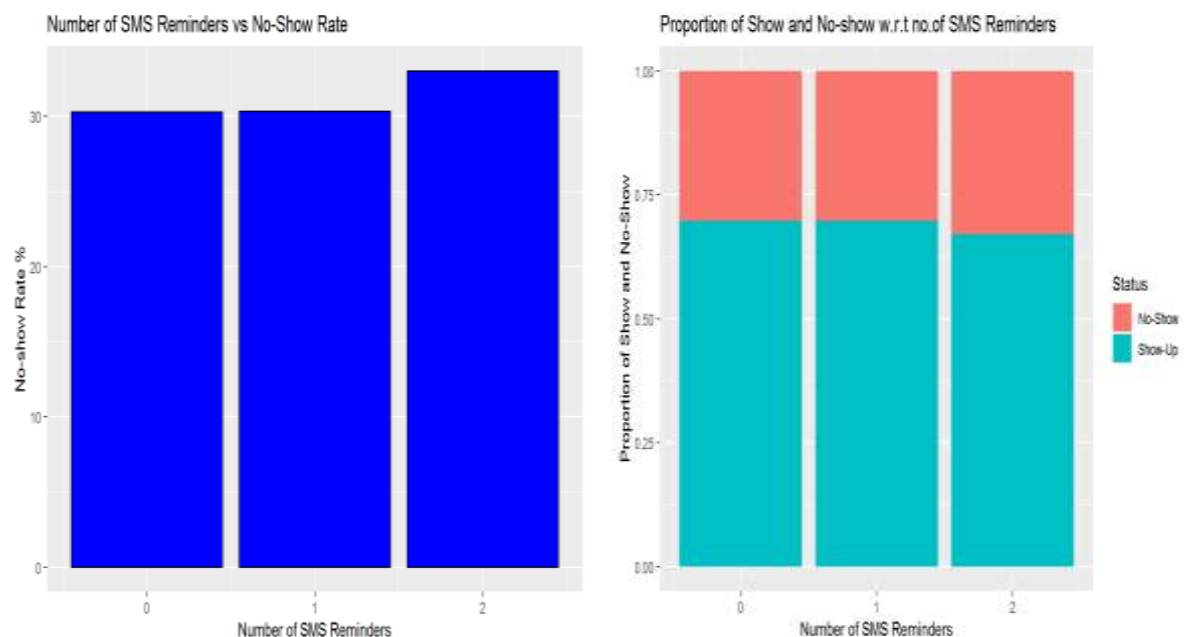
Day of Week:

Finally, the data was broken by day of the week. However, we cannot clearly conclude anything for the weekend owing to their small sample sizes (as can be seen by appointment days of the week count). Therefore, the proportion of Shows and No-shows were plotted for each weekday. As can be seen from the graph, Monday, Friday and Saturday exhibit highest no-show rate, whereas Sunday and Tuesday exhibit the lowest absence rate. Monday should have a high no-show rate since it is the beginning of the week and people are more likely to be busy in their work. However, Friday being the end of week, individuals would either have some end-of-week deliverables at work or some partying planned, hence miss their appointments. Since we cannot conclude anything concrete for Saturday and Sunday, a safe bet to schedule an appointment that would not be skipped by the patients would be the middle three days of the week.



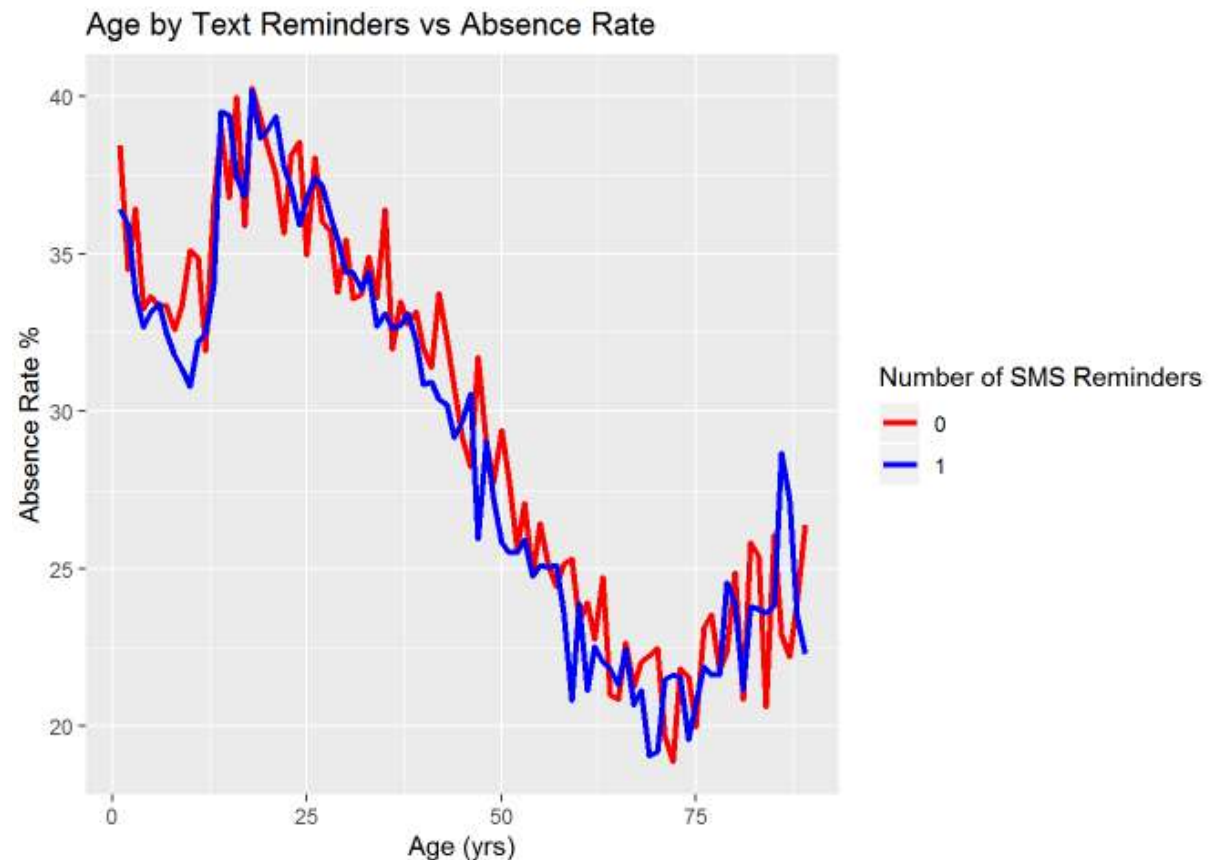
SMS Reminders:

One of the most interesting aspects of the dataset is SMS Reminders that are sent to the patients. We expected that patients receiving text reminders are more likely to attend an appointment. However, as per the data, one text reminder seems to decrease the no-show rate whereas two SMS reminders seem to significantly increase the no-show rate. This could be attributed to imbalanced data. Therefore, we decided to plot the proportion of patients showing/not showing up for the appointment versus the number of SMS reminders sent to the patient. We still observed a similar relation. Thus, we decided to delve deeper into understanding this variable.



We know that younger patients are more likely to sign up for text reminders rather than older patients, but young individuals are also more likely to not show up for their appointments.

Therefore, patients less likely to attend an appointment receive more reminders. To understand this, we segmented the SMS reminders per appointment data by age and created separate curves for number of messages received. Also, we removed the data for 2 SMS reminders since we cannot make any concrete inferences regarding it owing to its low sample size.



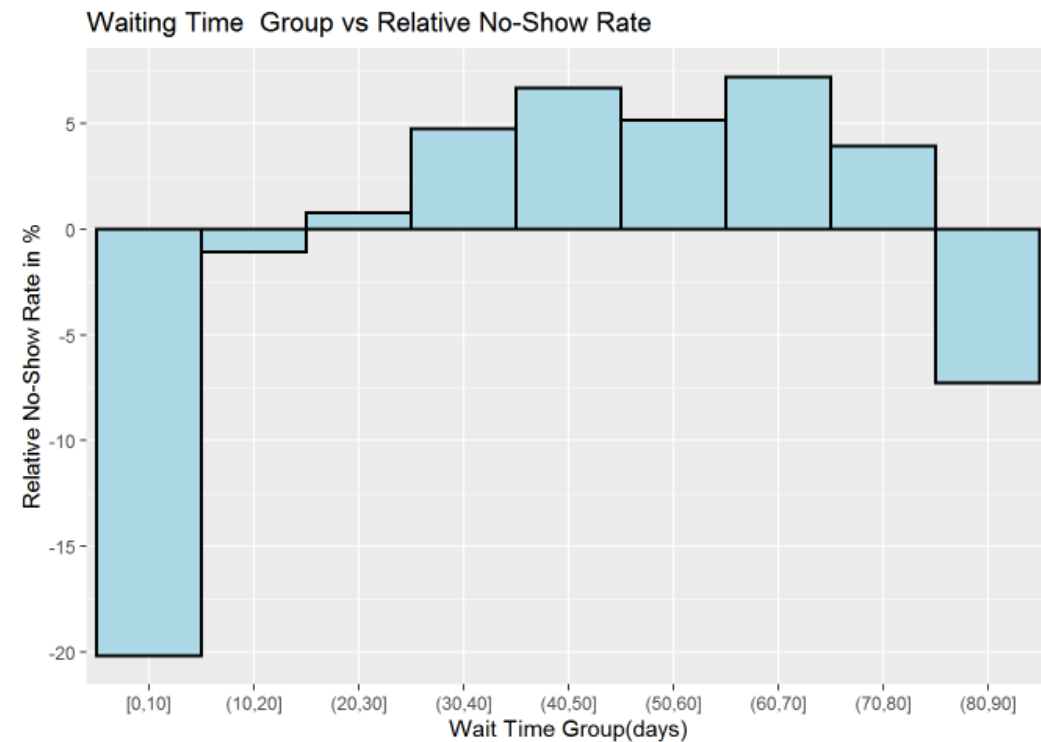
As expected, the average number of text messages sent for an appointment is greater among younger aged patients. A strong negative correlation is observed between the number of text reminders and age. The absence rates for 0 or 1 reminders look almost the same. However, in general, it does seem like for a given age, SMS reminders decrease the absence rate slightly. Thus, SMS reminders could decrease the non-attendance rate of patients.

Waiting time:

We proceeded to search for a possible trend in absence rate by waiting time, or the time between when an appointment was scheduled, and when the appointment took place. We grouped the data into 10-day segments for better analysis. We plotted the relative absence rate versus waiting time graph. The graph seems to demonstrate a lack of any relationship. The incidence of broken appointments increases when appointments are scheduled more than three weeks in advance. Patients who schedule less than 10 days out are 20% less likely to miss an appointment than those scheduling further out.

Increase in patient no-show when waiting time is more than 10 days can be attributed to three basic reasons. Firstly, patients with a longer waiting time will miss the appointment because their condition has more time to change in the interim period. Moreover, patients with a shorter waiting time are more likely to need urgent care or have a problem they deem to be pressing, and it would certainly be in their best interest to show up at the appointment. Also, patients may get an appointment in another hospital sooner resulting in them missing their

scheduled appointment. Lastly, since people have a fast-paced life these days, some other life event may get in the way leading to the patient forgetting/skipping their appointment.

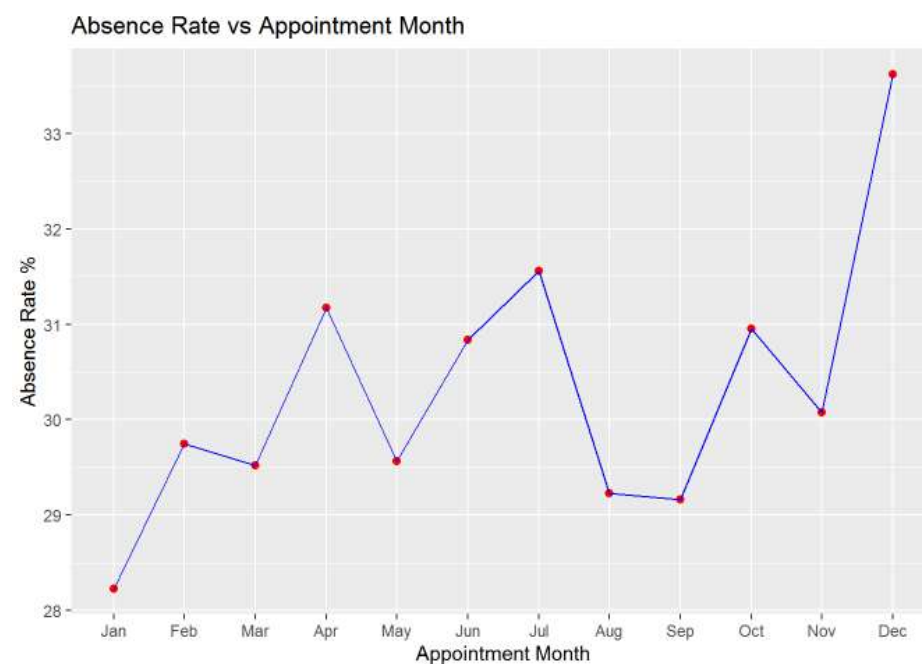


Date of Appointment:

We began the time-series analysis based on the Date of Appointment variable.

Appointment Month:

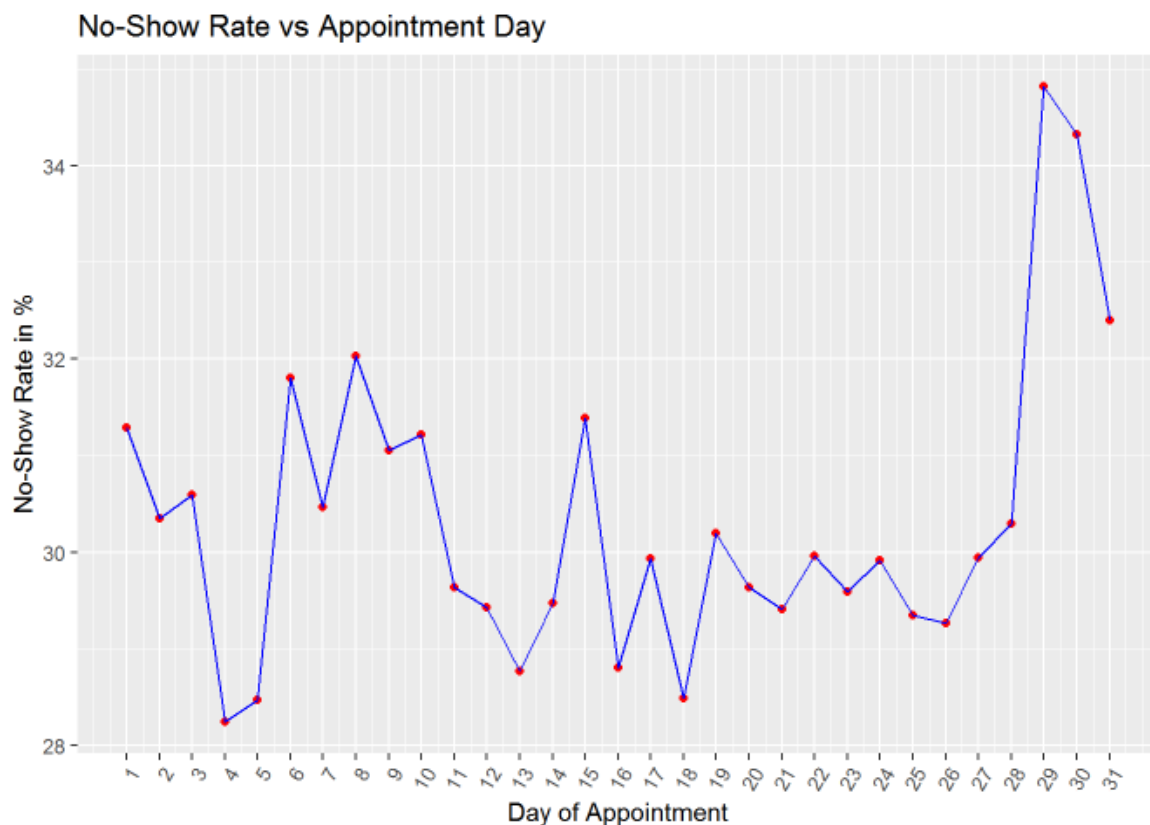
From the date, we extracted and grouped the appointments by month of year and plotted the average absence rate for each month. As per the graph, appointment month and no-show rate are loosely correlated. As the year progresses, an increase in the number of missed appointments is observed with the lowest rate in January (around 28%) and the highest in December (33.5%).



This could possibly be because people tend to get busier towards the end of the year. Also, people may plan their vacations or prefer spending time with their families in the holiday month of December leading to them missing out on their appointments. Therefore, people are more prone to scheduling their appointments in January after their vacations. Moreover, people may make some new year resolutions to be more health conscious thus, leading to a higher percentage of appointments met.

Appointment Day:

Next, we created a similar no-show rate over time graph, but this time by day of the month. Day of the month and absence rate also exhibit a weak positive correlation with no noteworthy trend. The only point to be noticed is that the no-show rate is much higher (almost 35%) towards the last three days of the month.



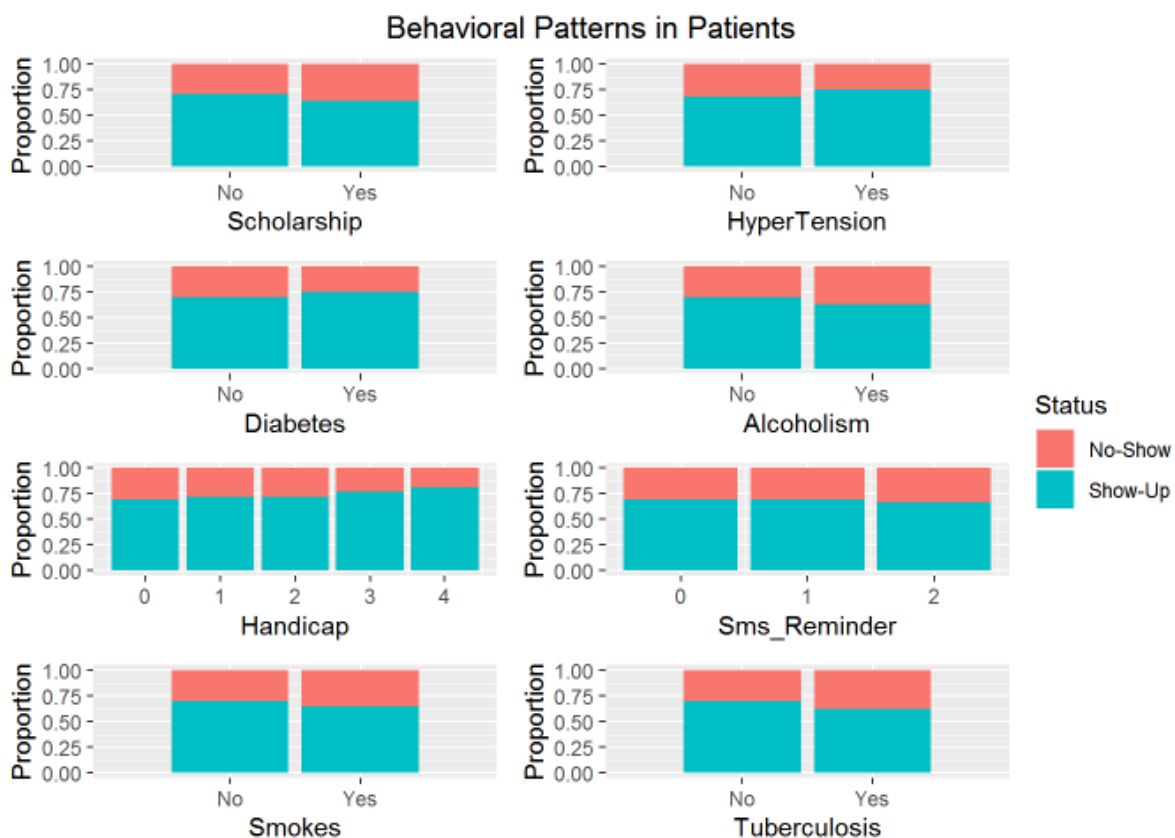
Individuals may miss their appointment towards the end of the month owing to some end of month deliverables at work. Moreover, since healthcare is very expensive in most parts of the world, patients may have some financial constraints towards the end of the month leading to them not showing up for their medical appointments.

5.2 How is the absence/no-show to a scheduled appointment dependent on the general characteristics and behavior patterns of the patient?

The next step in this analysis is to investigate how the absence rate of a patient is dependent on his general characteristics, behavioral patterns and health conditions. We grouped by these markers, sub-grouped into each category of the marker (like No/Yes or 0-4) and looked at the proportion of patients who showed up/did not show up for each category. The proportions

were taken instead of absence rate since the sample size in some categories is very small (Refer Appendix).

Broadly, none of the markers seem to considerably influence the no-show. However, on careful analysis, we observe some significant insights like Diabetes and Hypertension patients being least likely to skip an appointment. This is expected since these conditions need constant care and attention. Surprisingly, Tuberculosis and Alcoholic patients are most likely to not show up for their appointments. However, this could be attributed to their low sample sizes (refer appendix). Patients having Bolsa Familia Welfare Scholarship also strangely, attend their appointments less often, which could be because they do not want to spare time for the appointment because of the opportunity cost of not working.



5.3 Can we predict whether a patient would show up or not by taking the aforementioned variables as explanatory variables into consideration?

a) Feature Engineering:

As appointment date and appointment registration date did not give much information towards the status of show/no-show, we did not include them to predict the 'Status' variable. Even though, the behavioral markers of patients did not show any significant trend, they were included to find which feature is more significant when modeled. Since age and waiting time are significant as per our previous analysis, we further broke down age into bins and introduced them into the model as features.

b) Predictive Modelling using Logistic Regression:

Model 1:

Considering all the above-mentioned features and partitioning the data into train set (75%) and test set (25%), we used logistic regression model to predict no-show. It gave us an accuracy of 68.8% with the below confusion matrix:

Confusion Matrix for Model 1:

	No-Show	Show-up
No-Show	2471	20156
Show-up	3165	48959

Test accuracy: 68.8 %

The model established the following features to be contributing the most towards the prediction:

- Day of the week: Monday and Saturday
Without considering the sample sizes, the absolute no-show rate is high on Monday and Saturday, which is why they significantly contribute towards the prediction.
- Gender: Female
The female sample size is more than the male sample size (refer Appendix). Also, the female absence rate is higher as compared to males, thereby dominating the male predictor.
- Not an Alcoholic, smoker or Scholarship holder
Non-alcoholics, non-smokers and patients without a scholarship are much higher in number than their counterparts. Moreover, the proportion of attendance rate is higher for these groups leading to them contributing well in predicting show-up rate.
- Waiting Time
As per previous analysis, waiting time gave a trend by concluding that shorter waiting time decreases no-show rate. Therefore, as expected, it contributes well towards predictions.
- Age Groups 0-45 and 70-80 years
This supports our previous analysis of age group versus absence rate graph. The peak between 0-45 years predicts no-show better whereas the dip between 70-80 years predicts show-up rate better.

Model 2:

Later, we used these features to implement Model 2, which gave us an almost similar accuracy of 68.84% with the following confusion matrix:

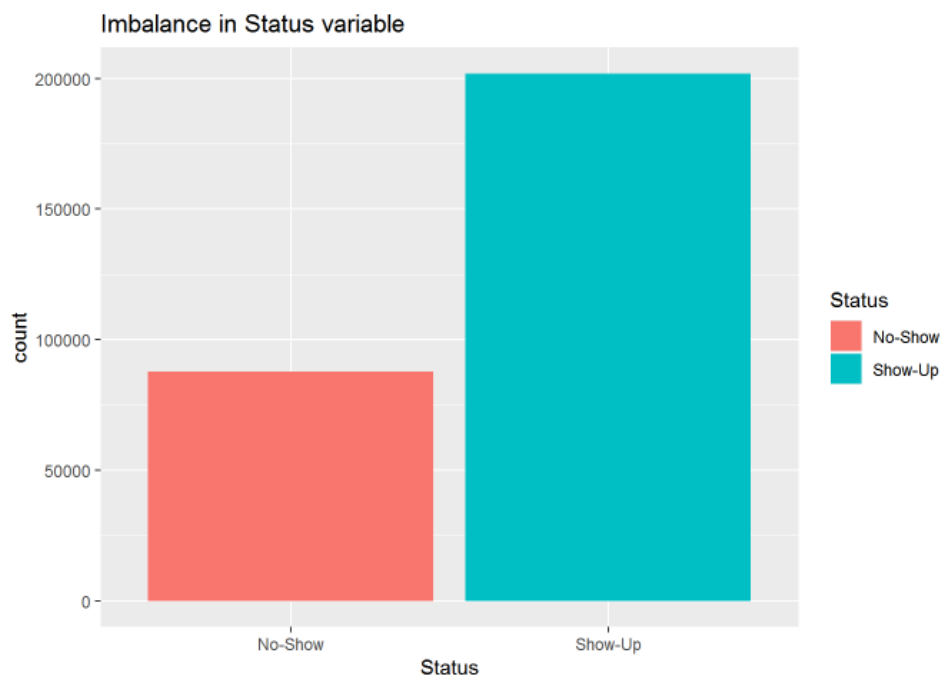
Confusion Matrix for Model 2:

	No-Show	Show-up
No-Show	2317	20310
Show-up	2981	49143

Test accuracy: 68.84 %

Thus, on comparing the confusion matrices, we can conclude that Model 2 predicts Show-up better than Model 1.

However, these conclusions do not take high imbalance in data (graph below):



Model 3:

Later, in Model 3 we balanced the data by down sampling (refer Appendix) the show-up data but achieved a much lower accuracy of 50% (Confusion matrix below):

Confusion Matrix for Model 3:

	No-Show	Show-up
No-Show	20917	1710
Show-up	22597	3465

Test accuracy: 50.08 %

Therefore, since **Model 2 gives the best accuracy (68.84%)** by taking all significant features into account, it can be concluded as the best for predicting the show-up and no-show rates (Refer Appendix for ROC Curve Comparison).

6. Conclusion

Though it appeared at the beginning that there are no significant correlations that can be made from the data, we were able to find out several key relationships on further analysis. Considering that this dataset may not represent all countries and hospitals exactly and some of the groupings result in small sample sizes, following were the most notable discoveries:

- December has the highest absence rate and January has the lowest but there is no trend in between.

- ii. Excluding the weekends, the day of the week with the highest absence rate is Monday followed by Friday. Tuesday had the lowest rate of absences with appointments on Tuesday more likely to be attended than those on the other days of the week.
- iii. As the age of the patient increased, the likelihood that they would not show up to their appointment decreased. The youngest patients had a relatively low rate of missed appointments, then the absence rate rose and peaked in the teens, before gradually declining until age 70 where it increased by a small amount into the upper end of the range.
- iv. Patients who made an appointment less than 10 days in advance were 20% more likely to attend the appointment than those who made the appointment further ahead of time.
- v. At a specified age, patients who received 1 or 2 text messages were around 0.5% less likely to miss an appointment than those receiving no text messages.
- vi. Patients with hypertension or diabetes were more likely to attend a scheduled appointment than the average patient.
- vii. Patients with families in the Bolsa Familia program were more likely to miss an appointment than the average patient.
- viii. We found that for registration hour between 12:00-5:00 am, none of the patients show up. However, later we realized that there was no data for this particular category (refer Appendix).

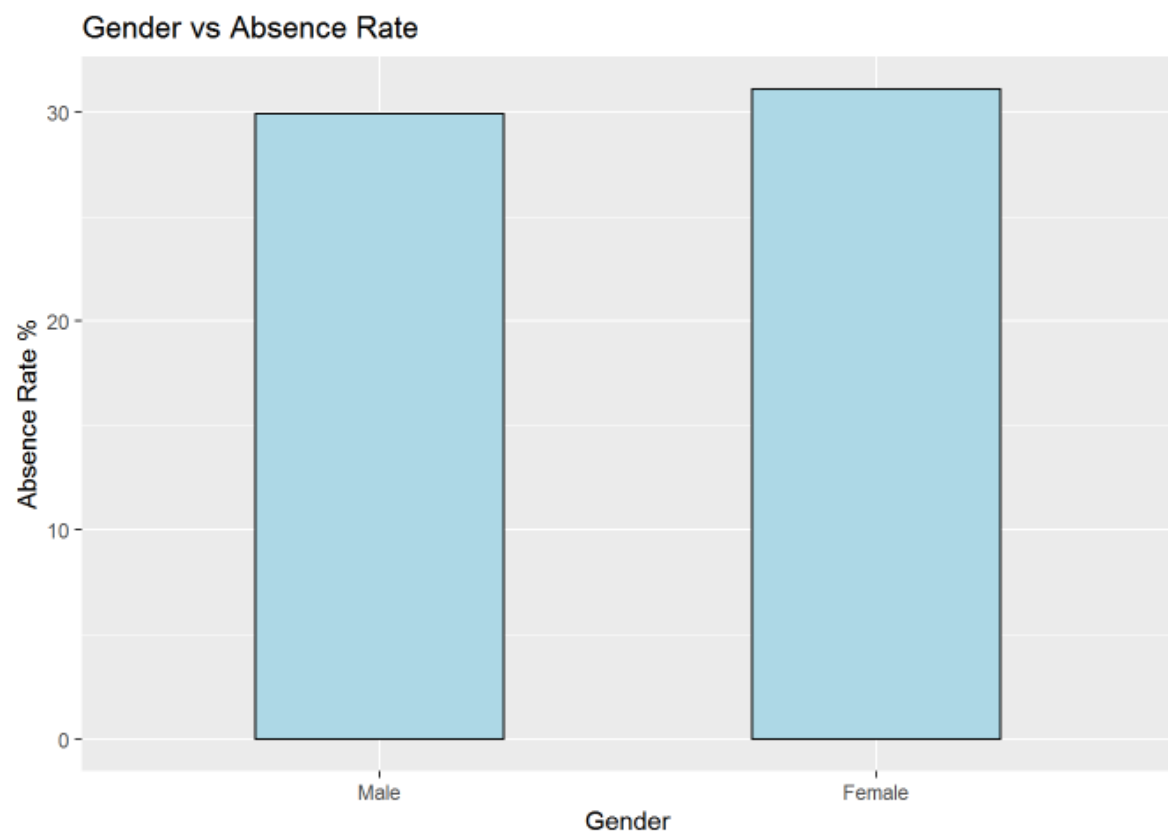
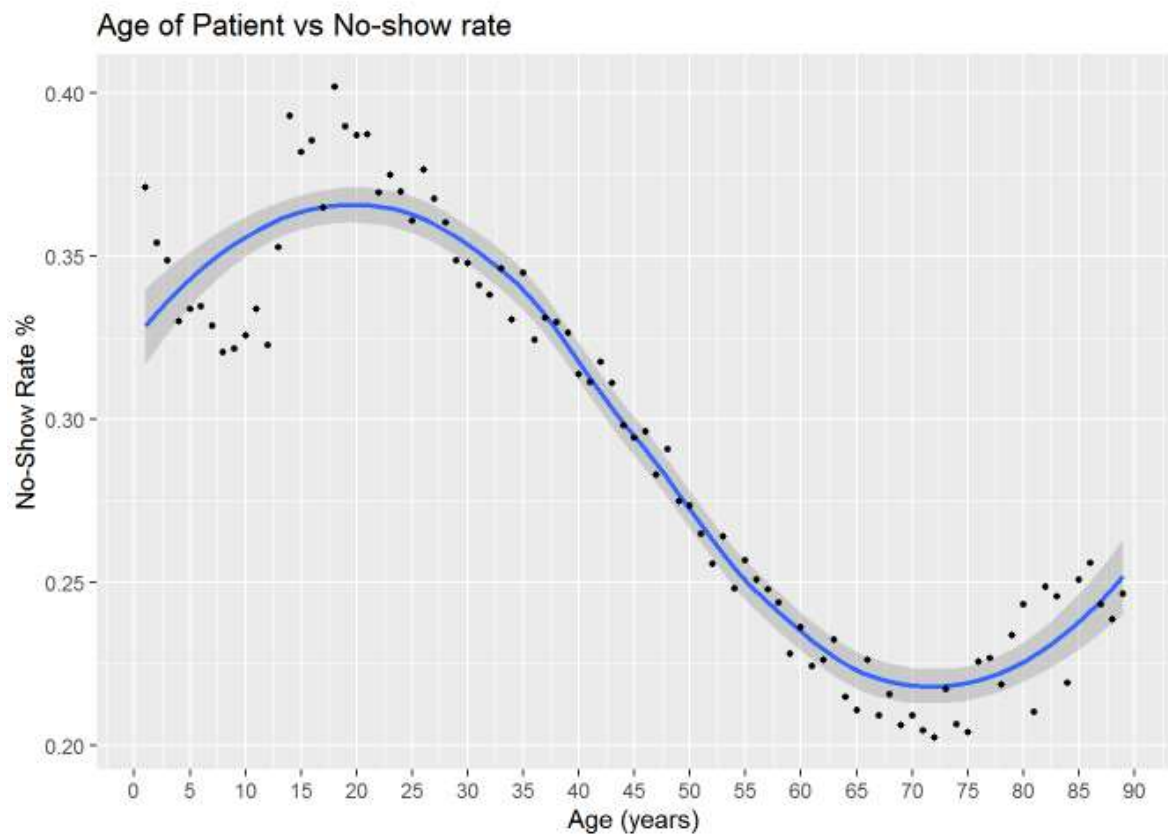
7. Future Scope/Limitations:

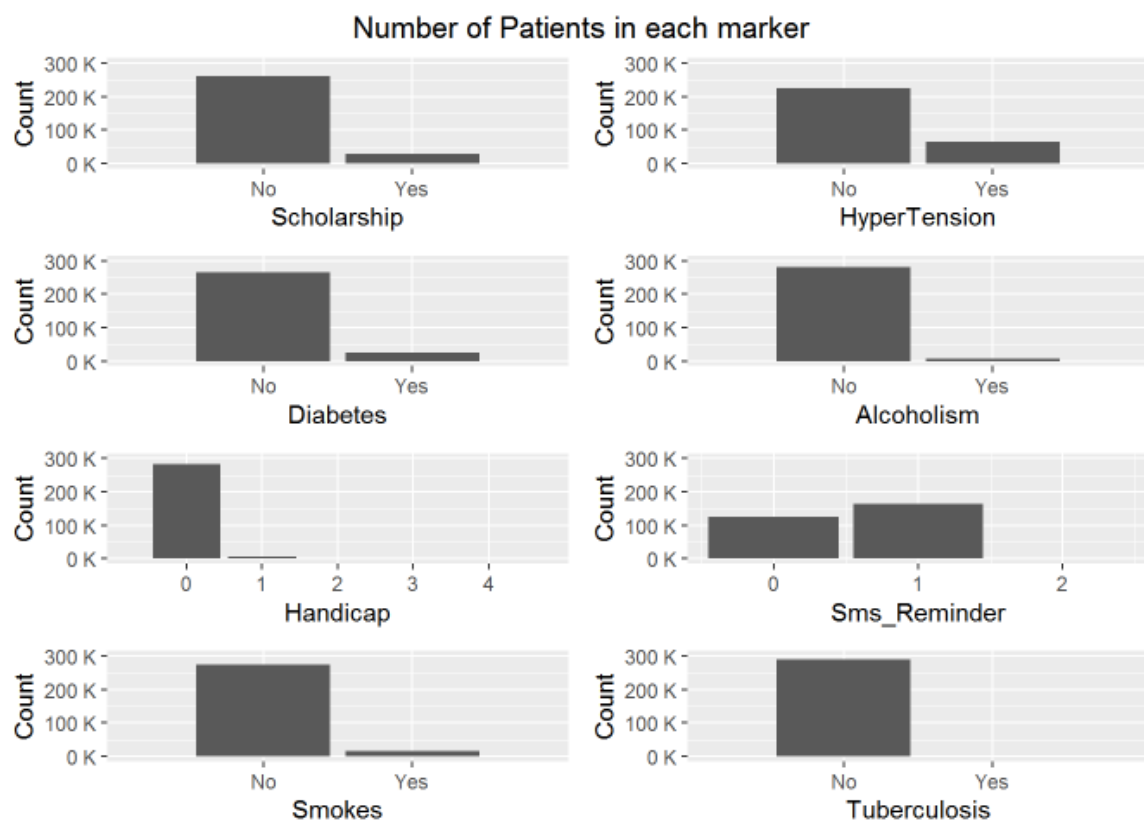
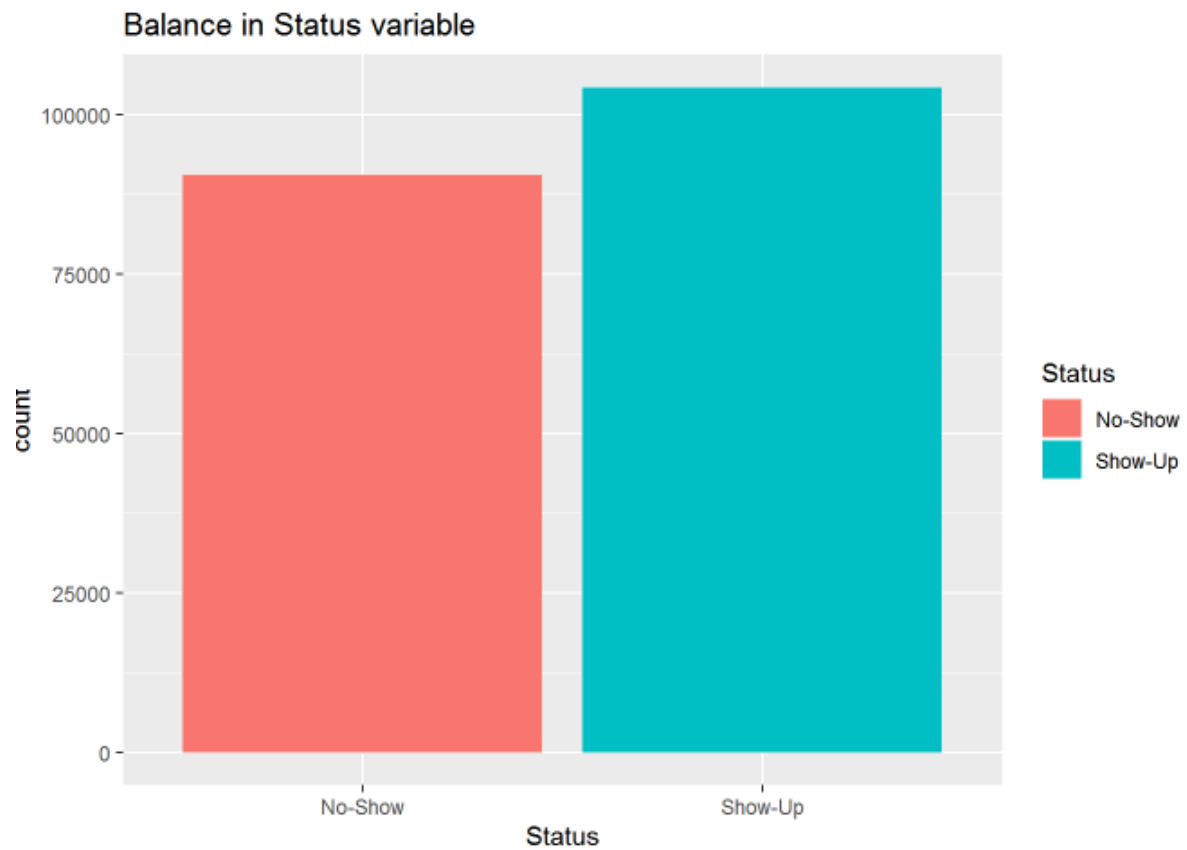
Data is analyzed holistically rather than differentiating on gender. Thus, we could have inferred better insight if segmented on gender. Moreover, age had some negative values, so we removed them and the ages < 90 have been considered for better visualization. However, it could have some meaning like infant data which we did not investigate. Similarly, waiting time only less than 90 days (3 months) is considered. Also, since the data is huge (300,000 instances), due to computation and memory constraints, we were unable to use random forest or boosting algorithms for better modeling. Thus, the accuracy could have been improved with better resources.

8. References

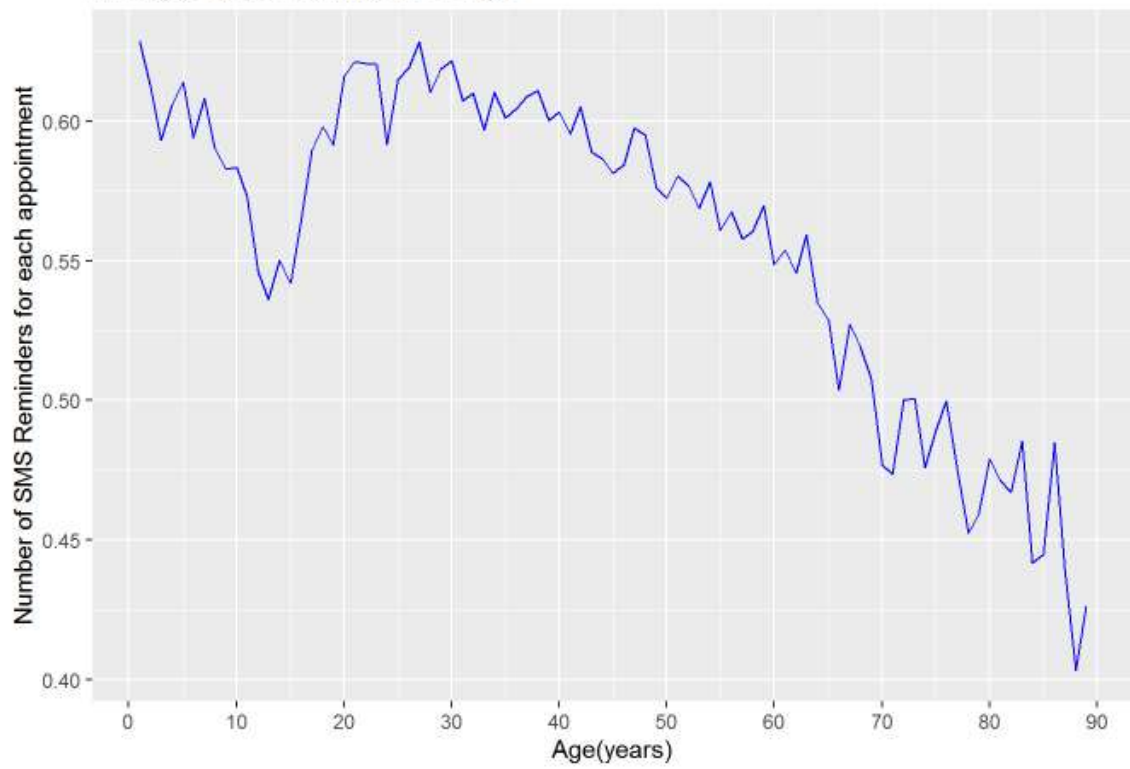
- [1] M. L. Davies, R. M. Go_man, J. H. May, R. J. Monte, K. L. Rodriguez, Y. C. Tjader, and D. L. Vargas. Large-scale no-show patterns and distributions for clinic operational research, Feb 2016.
- [2] J. Gold. In cities, the average doctor wait-time is 18.5 days, Apr 2019.
- [3] JoniHoppen. Medical appointment no shows, Aug 2017.
- [4] W. Koehrsen. Exploratory data analysis with r, Aug 2017.
- [5] N. L. Lacy, A. Paulman, and M. D. Reuter. Nao
- [6] I. Mohammadi, H. Wu, A. Turkcan, T. Toscos, and B. N. Doebbeling. Data analytics and modeling for appointment no-show in community health centers, 2018.
- [7] N. A. of Engineering (US), I. of Medicine (US) Committee on Engineering, and the Health Care System. Building a better delivery system: A new engineering/health care partnership, 2005.
- [8] PatientPop. 5 reasons patients no-show for healthcare appointments, Jun 2019.
- [9] Samratp. Predict show/noshow - eda visualization model, May 2019.
- [10] K. T. The effect of no-shows on patients and hospitals, Jun 2017.
- 5mi l. lacy, Nov 2004.

9. Appendix

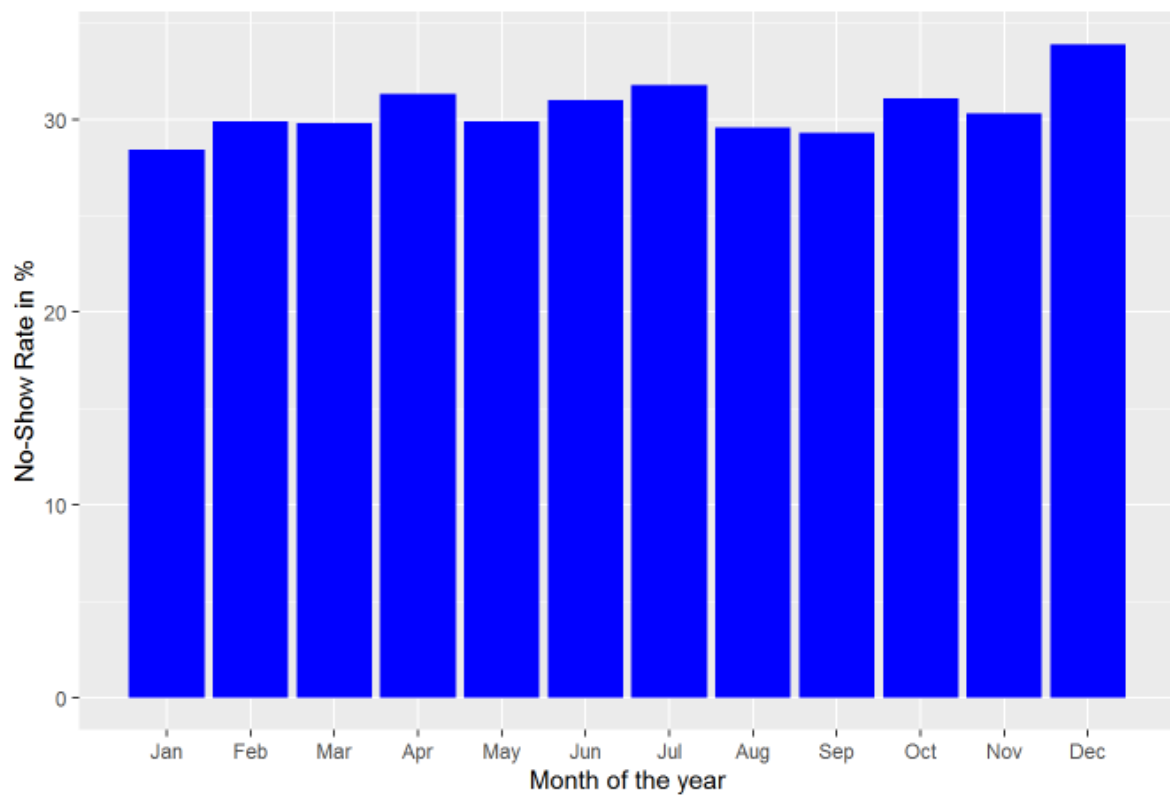


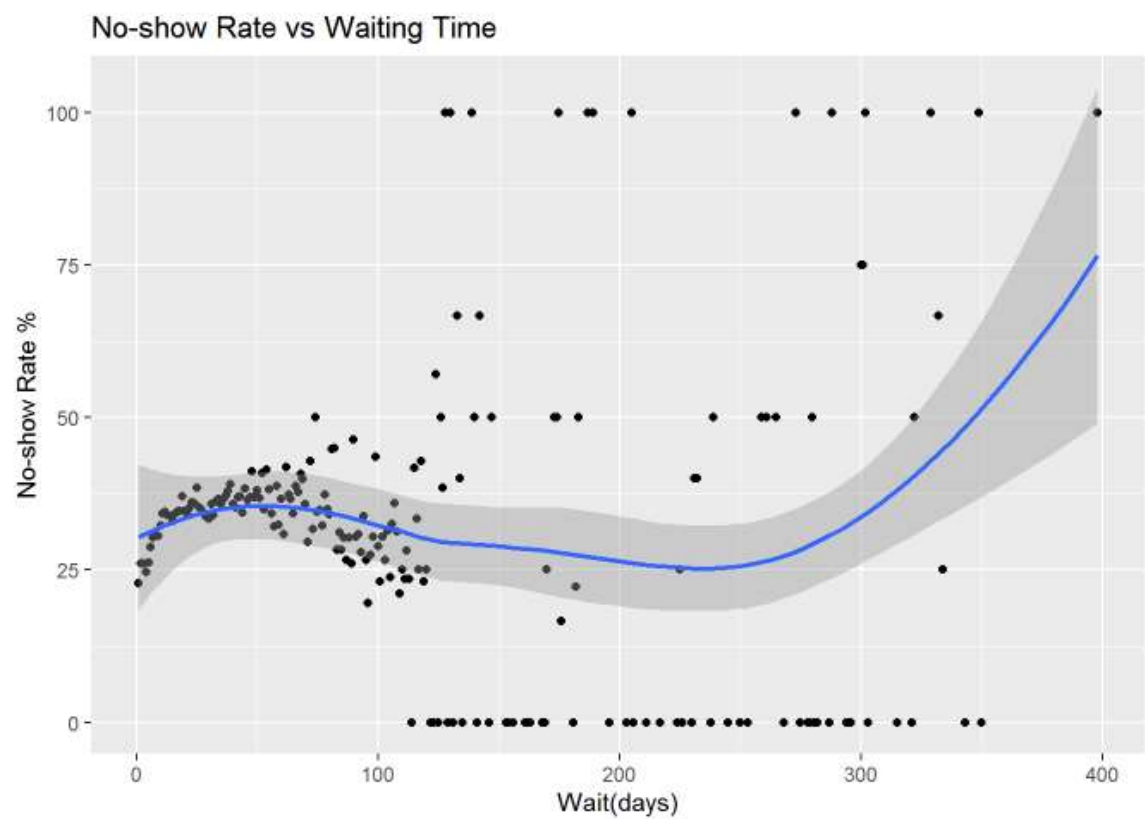
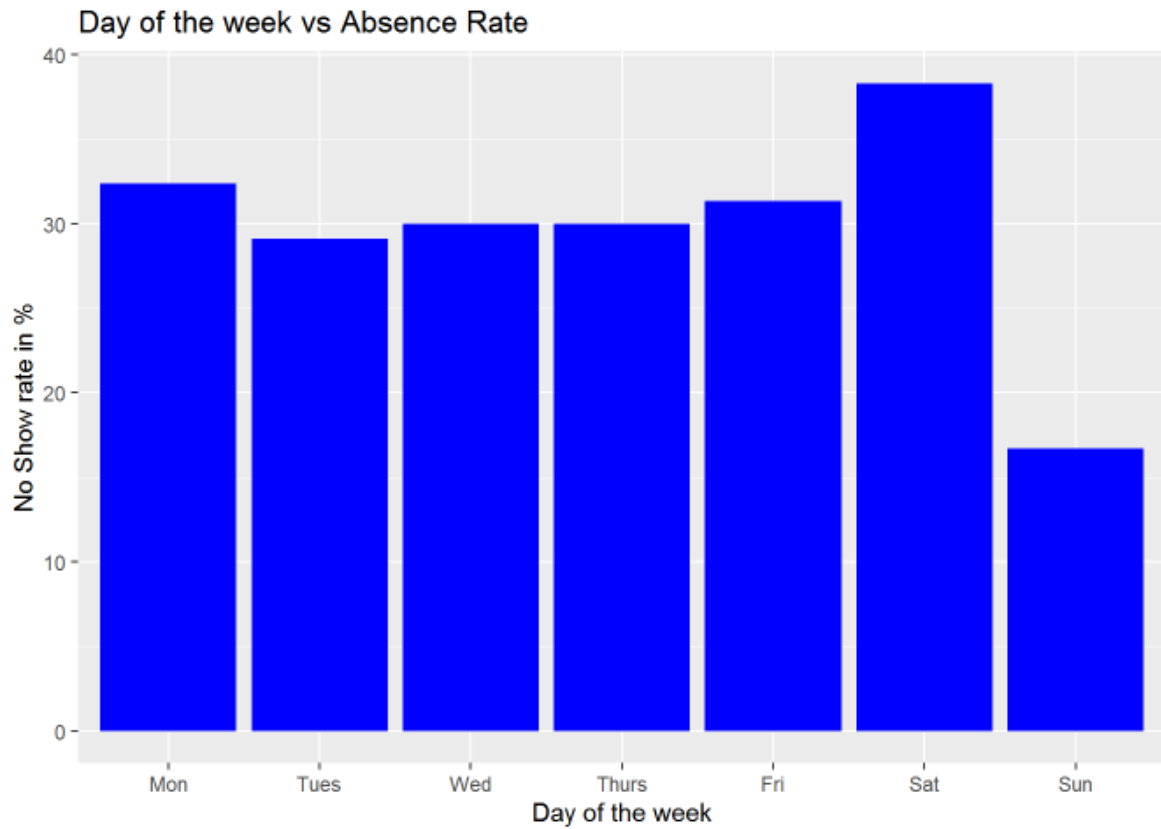


Average SMS Reminders vs Age

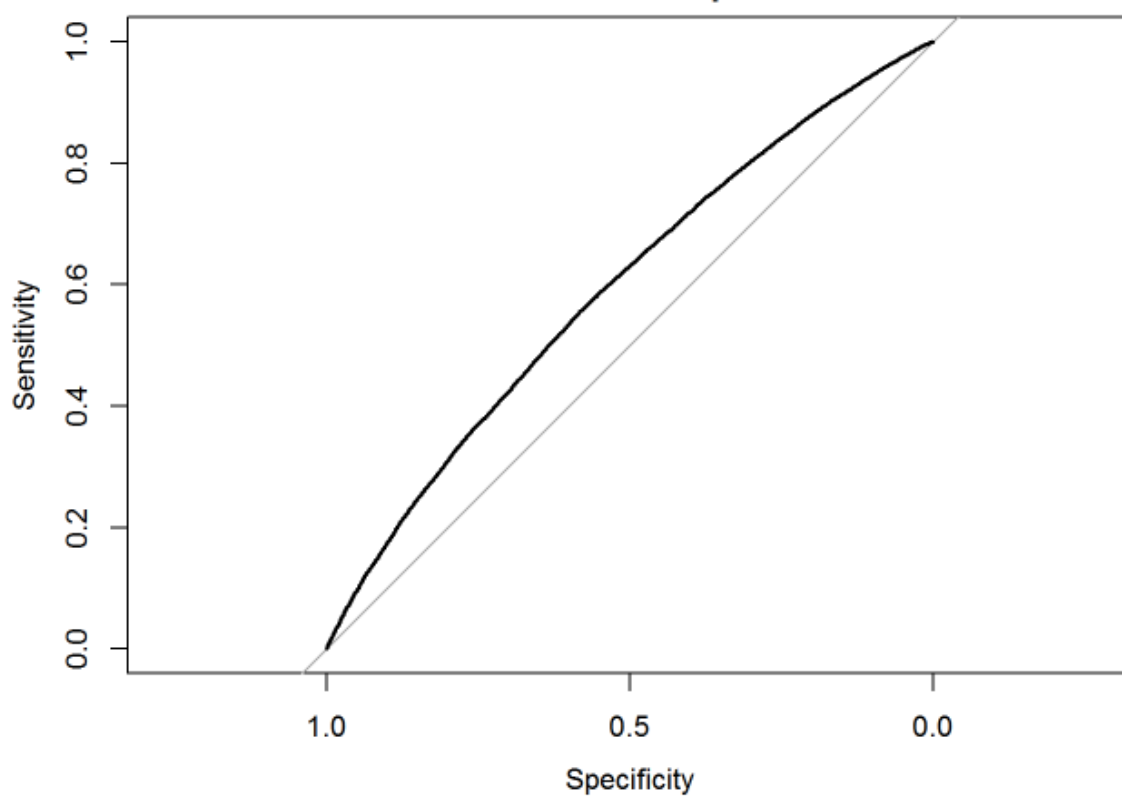


No-Show Rate Against month in a year





Model 1 ROC plot



Model 2 ROC plot

