S670 Exploratory Data Analysis Final Project Report

DOCTOR'S APPOINTMENT NO-SHOW

Group Name: Team No-Show

Team members: *Chhavi Sharma, Prahasan Gadugu, Supriya Ayalur Balasubramanian*

1. Project 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, pediatrics and dentistry. 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 due to administrative costs and other related costs. Each No-show costs the US healthcare system a tremendous loss of nearly $200. Moreover, delayed diagnosis potentially puts patients in danger. No-show patients deprive themselves of professional services, disrupt client-care provider relationships and reduce the opportunity for other patients to receive timely care.

Therefore, 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. This research attempts to understand the reason behind patient absenteeism by studying patients' demographic factors, their environmental factors, their behavior and the relation between these factors that lead to no-show.

2. Relevance

Hospitals, clinics and other medical care centers globally are trying to reduce their costs as much as possible due to wastage of resources caused by missed appointments. This analysis benefits the healthcare sector by providing an insight into achieving this.

2.1 Scientific contribution

It is interesting for the researchers to see how information about patients' demographic and environmental factors and patients' behavior with regard to no-show can be combined, and further developed to create a method to support the healthcare sector. This analysis creates the source of information to perform future research.

2.2 Societal contribution

The societal contributions of this analysis include the results and therefore the method that could be developed keeping the inferences from this analysis in mind that may serve to reduce no-show patients in the healthcare sector. It may help hospitals, clinics and other medical care centers to understand the factors associated with no-show. Ultimately, hospitals can transfer knowledge to their staff members regarding the reduction of the number of no-show patients. This knowledge helps to improve the communication between the staff and their patients and creates a better understanding of the patients.

3. Research Questionnaire

In order to answer the main question, a drill-down is required. This drilldown of the problem statement resulted in the following research questions that we aim to answer using exploratory data analysis methods on the Medical Appointment No-Show dataset (from Kaggle):

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

4. Dataset Description:

Dataset Source: Kaggle
(https://www.kaggle.com/joniarroba/noshowappointments/version/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 attended, it is marked as a 'Show' in the target variable; when the patient schedules an appointment and does not show up, it is marked as a 'No-Show' in the target variable. The information about the appointment included demographic data (factors relating to an individual's personal characteristics that can influence the individual to undertake a certain action), environment or time data (focus on the nature of people's transactions with their physical and sociocultural surroundings), and patient markers (habits and present health conditions or ailments).
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; gender of patient.
- **WaitingTime**: integer; number of days elapsed between when the appointment was made and when the appointment took place.
- **Scholarship**: 0 or 1; indicates 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.
- **AppointmentReservationDate**: date and time for which the appointment was made.
- **AppointmentDate**: date of appointment without time.
- **DayOfTheWeek**: day of the week of appointment.
- **SMSReminder**: 0 ,1 ,2; values for number of text message reminders sent to patient about appointment.
- **Diabetes**: 0 or 1 for condition (1 means patient was scheduled to treat condition).
- **Alcoholism**: 0 or 1 for condition.
- **Handcap**: 0, 1, 2, 3 or 4 for condition.
- **Tuberculosis:** 0 or 1 for condition.
- **Smoker**: 0 or 1 for smoker / non-smoker.
- **Tuberculosis**: 1 or 0 for condition.

Target Variable:

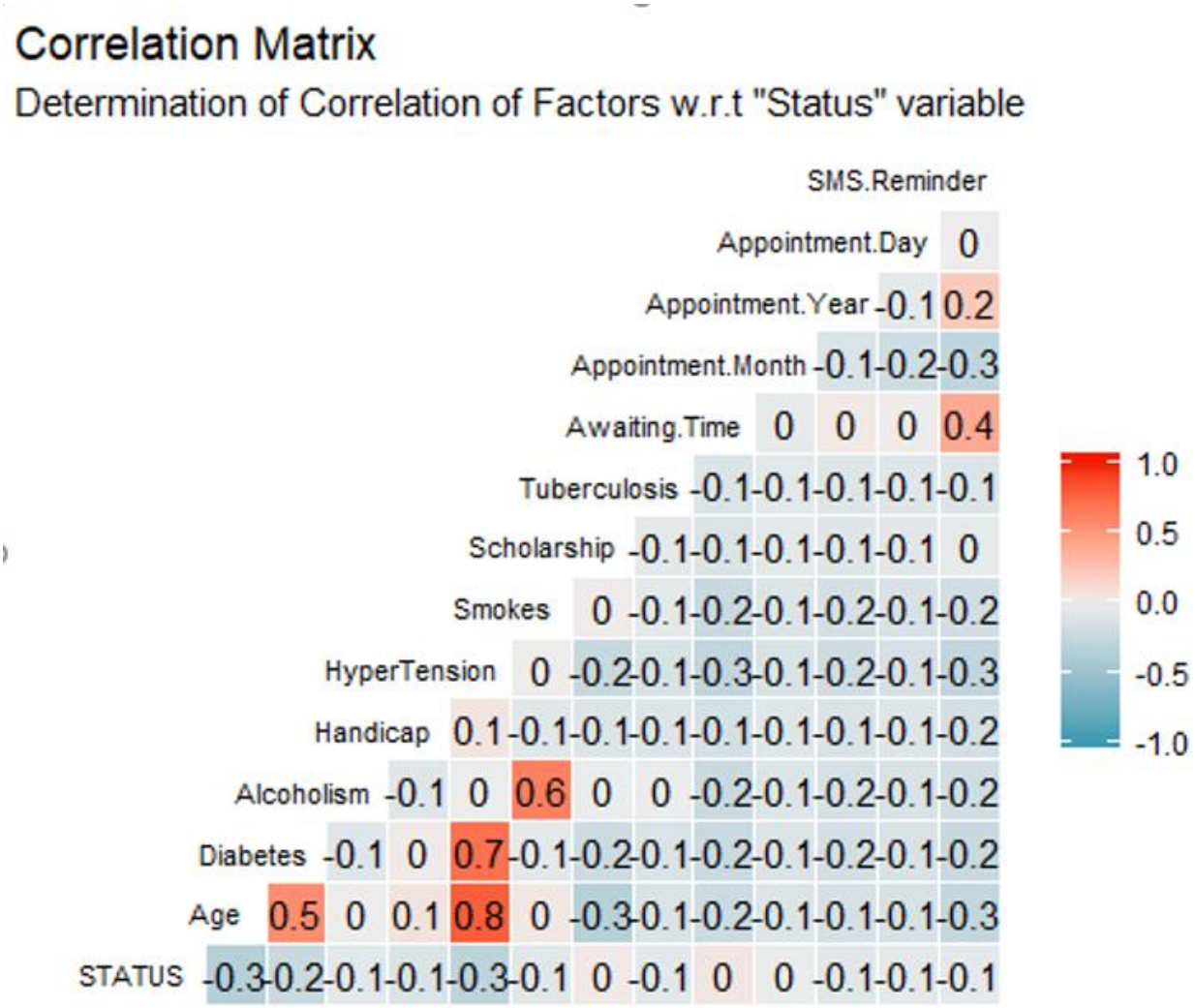
- **Status**: Show-up or No-Show

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.

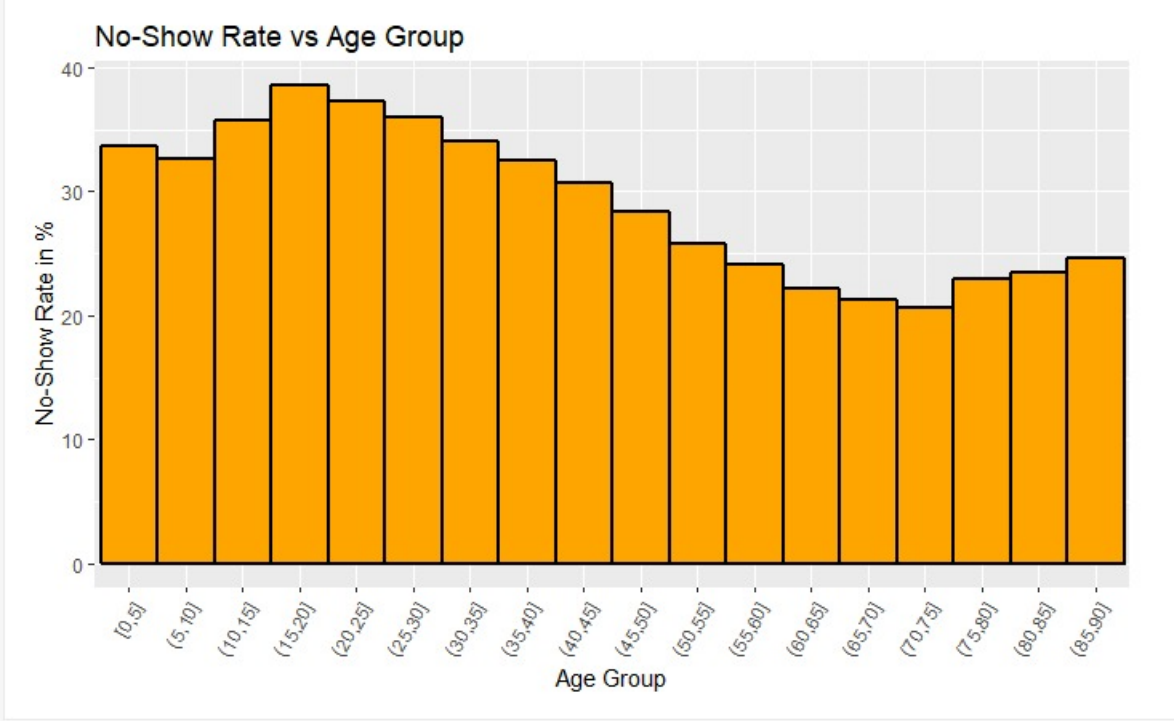
From these correlation values, there are no standouts that are strong linear predictors of whether a patient will miss a visit (given by the status column). It appears that there are other strong and moderate relationships between age and hypertension, Diabetes and Hypertension, Diabetes and Age, Age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



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.

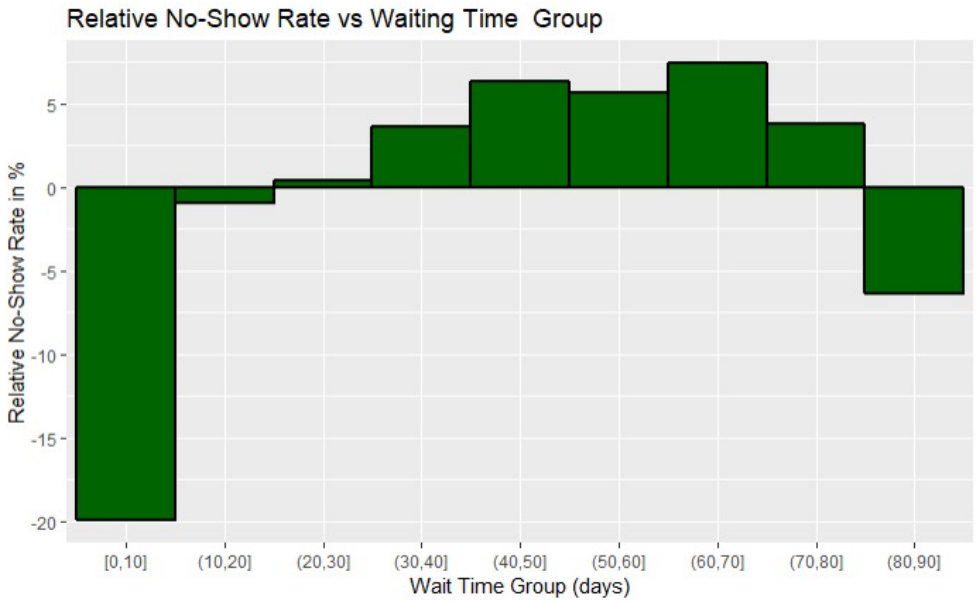


As expected, the youngest and oldest patients tend to attend their appointments since they pay more attention to their health because of their vulnerability to falling sick. In contrast, patients in their youth or middle age are generally healthier and tend to neglect their health owing to their busy 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.

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.

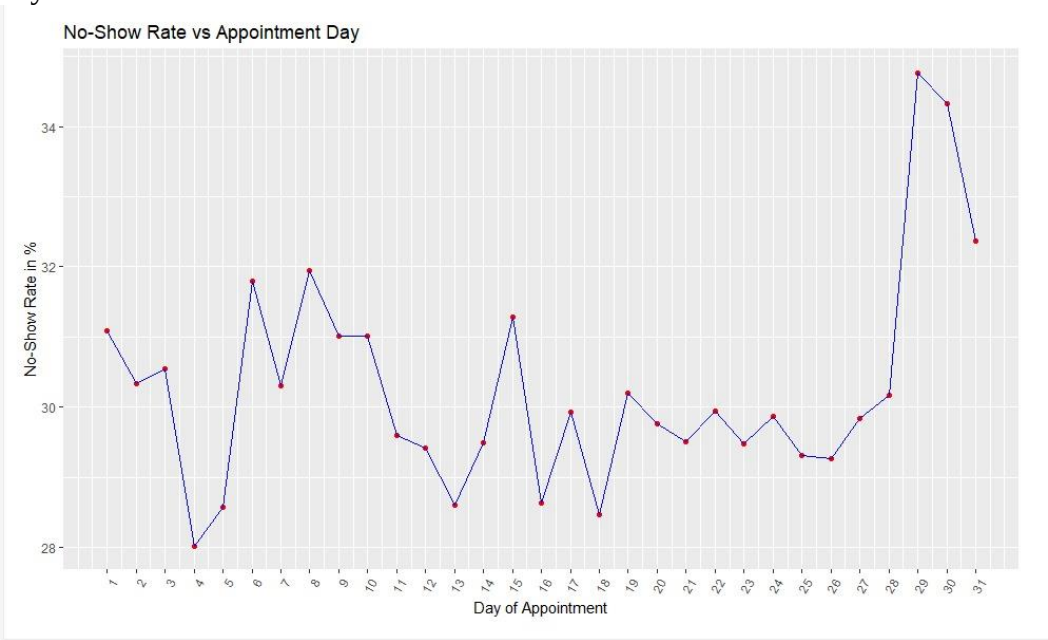
Appointment Month:
From the date, we extracted and grouped the appointments by month of year and plotted the average absence rate for each month. As per the graph, appointment month and no-show rate are loosely correlated. As the year progresses, an increase in the number of missed appointments is observed 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

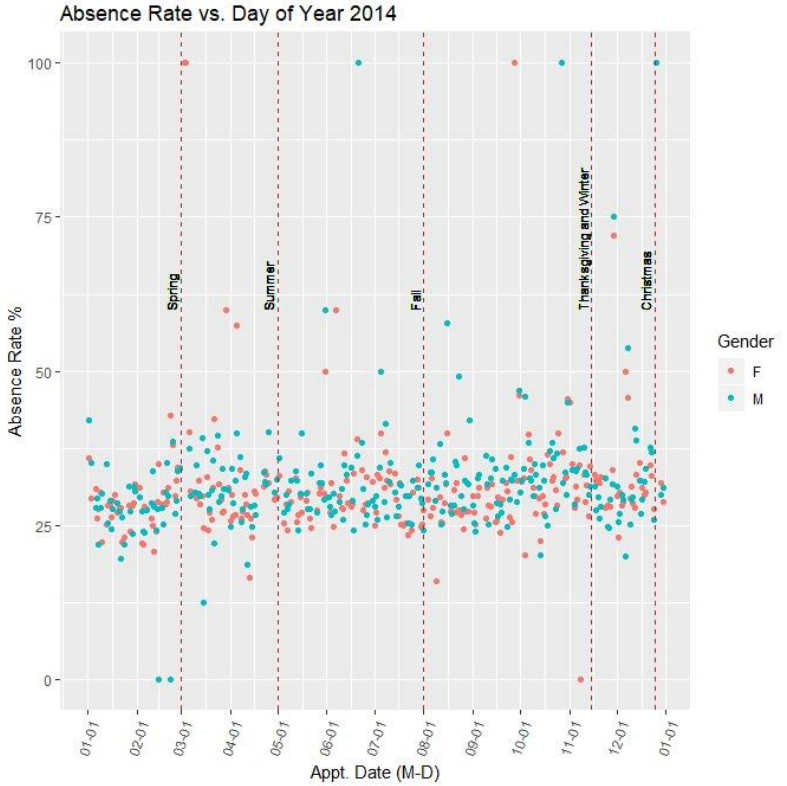
new year resolutions to be more health conscious thus, leading to a higher percentage of appointments met.

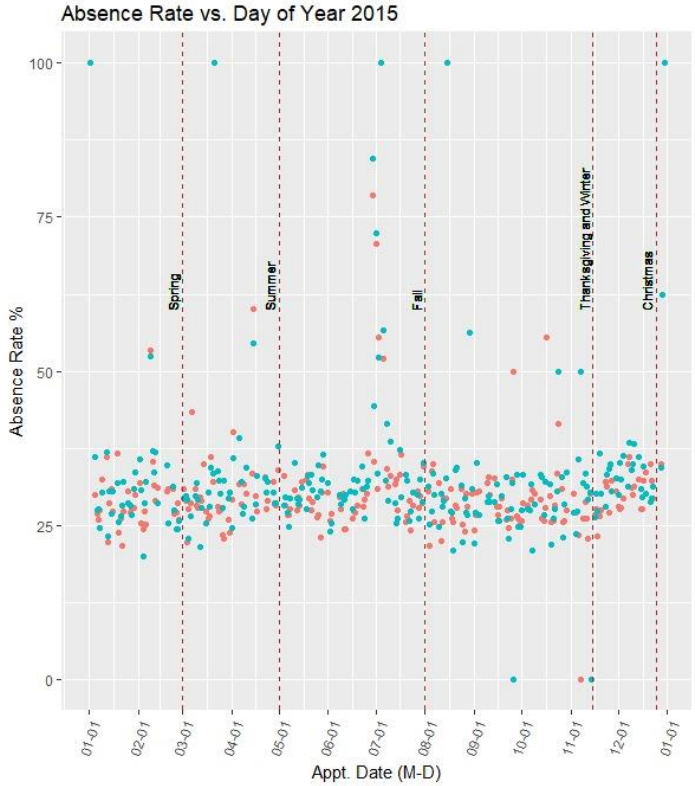
Appointment Day:

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.





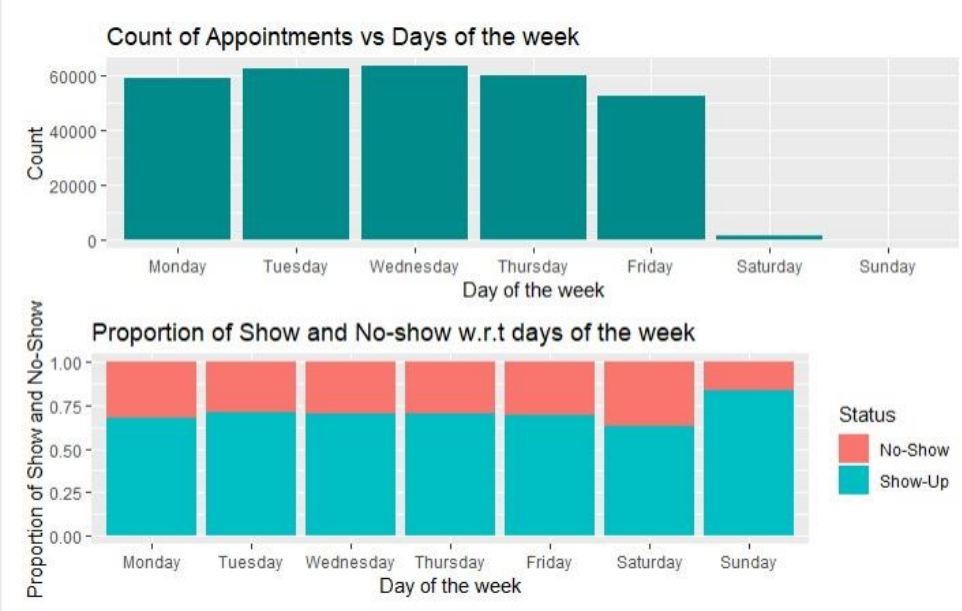
For better understanding, we plotted the absence rate with respect to the Date for years 2014 and 2015 differentiating on gender. This is done for season-wise and holiday analysis of absence rate. Holistically, the no-show rate was observed to be constant for the years 2014 and 2015 with a few outliers. The outliers are higher during summers indicating that the absence rate is slightly higher in summers. In 2015, out of all the outliers, the absence rate of males is much higher as compared to females.

There are several days for which the absence rate is 100% which should be investigated. These points might correspond to public holidays. However, these days are not consistent across the two years, so there could be other major events that correspond to an increase in absences. Thus, we overlaid the holidays to see if it reveals a relationship. No significant trend was observed.

Day of Week:

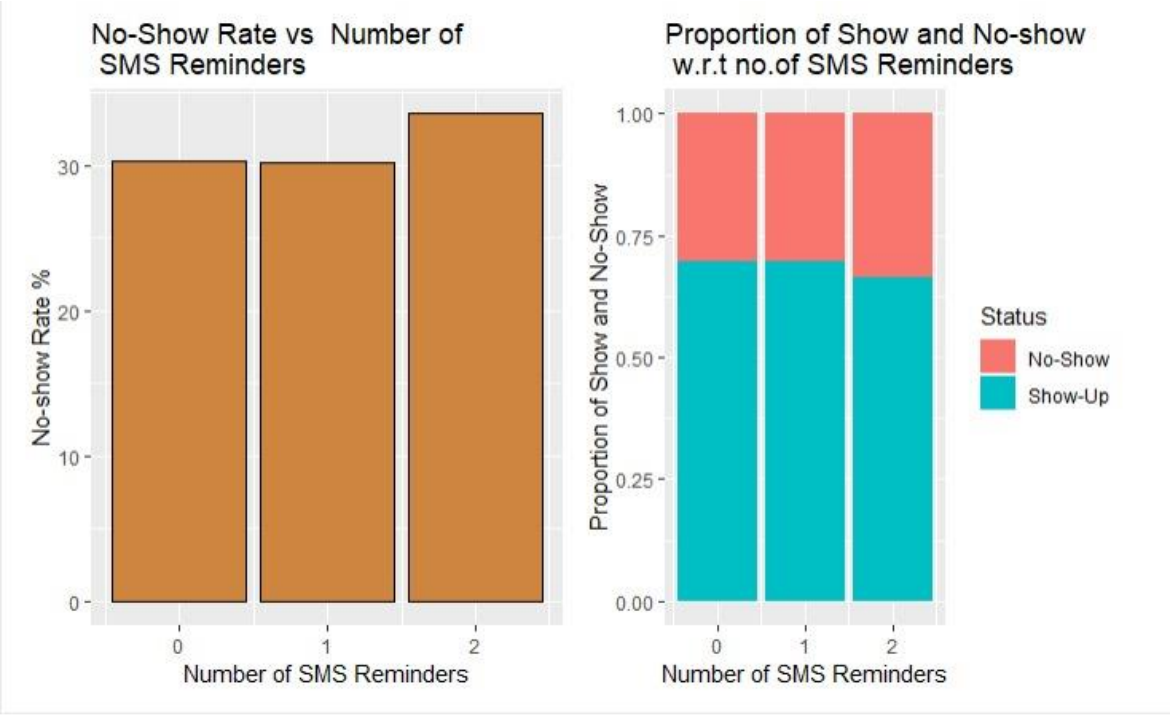
Finally, the data was broken by day of the week. However, we cannot clearly conclude anything for the weekend owing to their small sample sizes (as can be seen by appointment days of the week count). Therefore, the proportion of Shows and No-shows were plotted for each weekday. As can be seen from the 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.

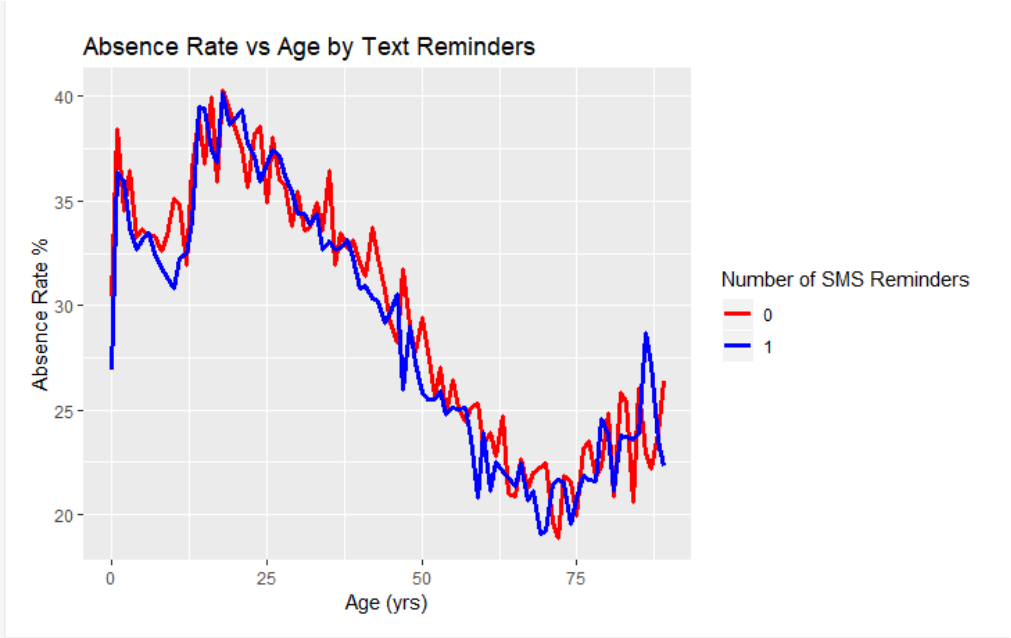


SMS Reminders:

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.

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.

Finally, more information for these categories is required to draw any concrete inferences.

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:

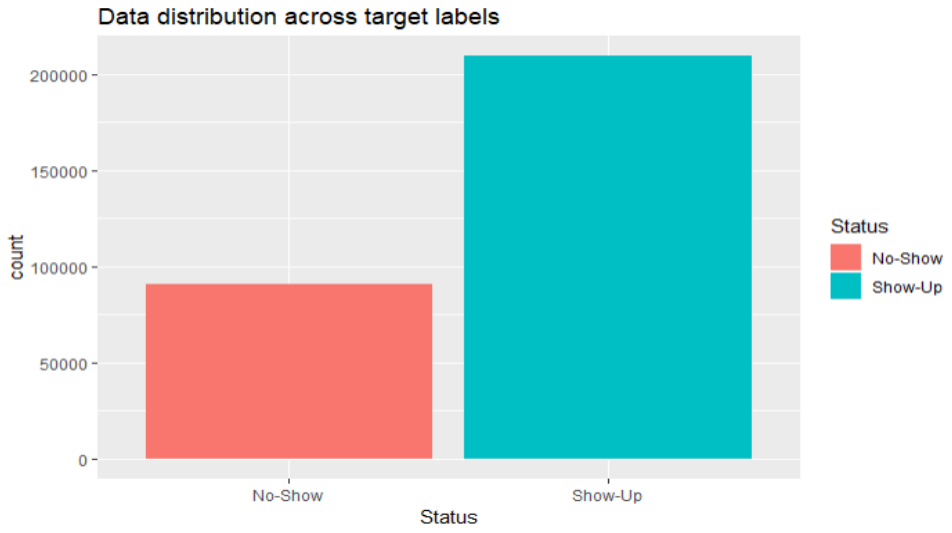
Confusion Matrix for Model 2:

	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:

i. 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).

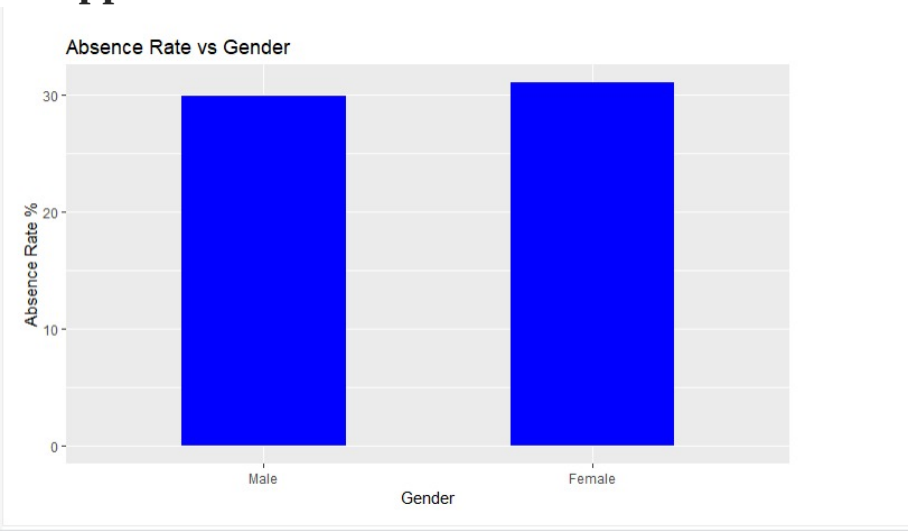
7. Future Scope/Limitations:

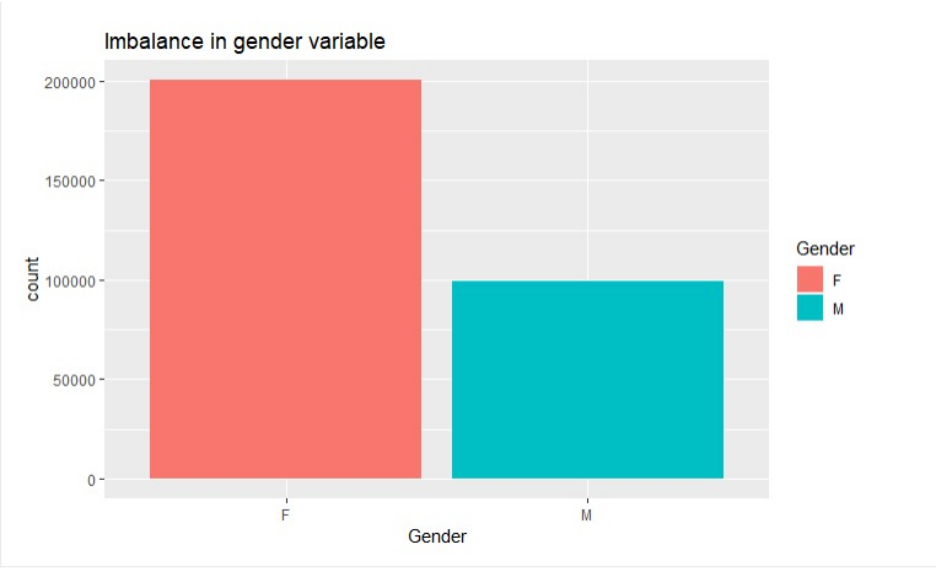
Data is analyzed holistically rather than differentiating on gender. Thus, we could have inferred better insight if segmented on gender. Moreover, age had some negative values, so we removed them and the ages < 90 have been considered for better visualization. However, it could have some meaning like 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.

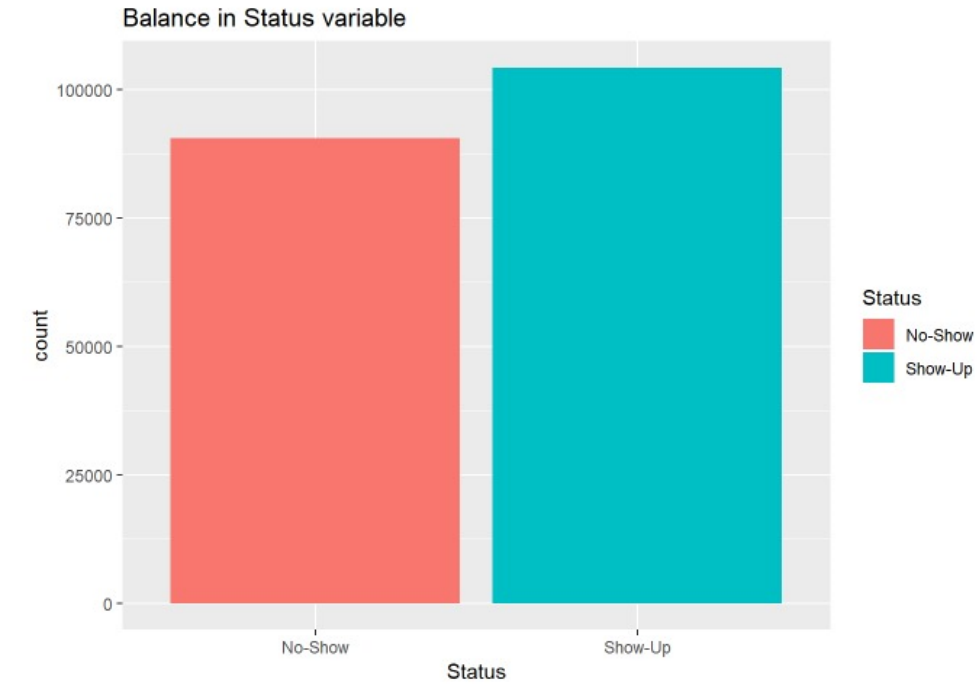
8. References

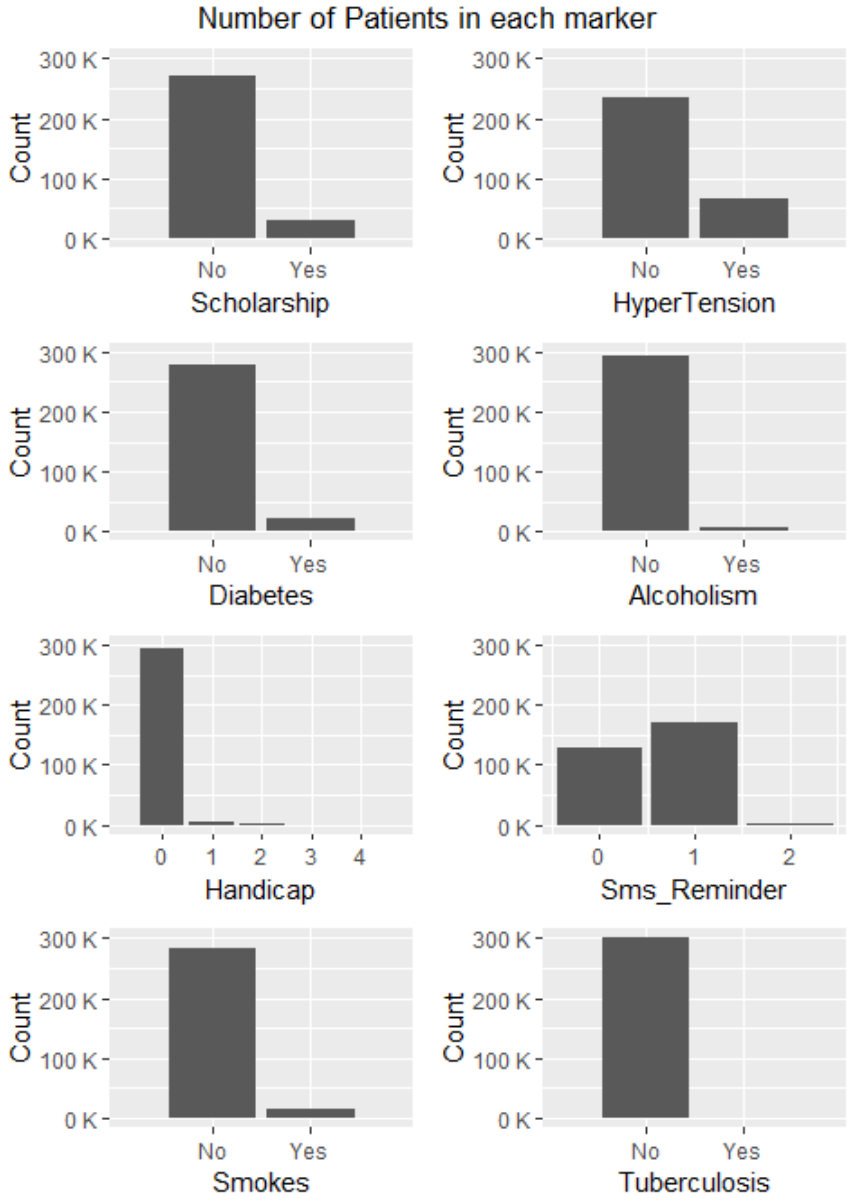
a. https://medium.com/@williamkoehrsen/exploratory-data-analysis-with-r-f9d3a4eb6b16
b. https://www.kaggle.com/yousuf28/medical-appointment-no-show-in-r/report
c. https://www.kaggle.com/jph84562/data-exploration-and-visualization/report
d. https://www.kaggle.com/cnjn22/exploratory-analysis-for-medical-appointment
e. https://www.kaggle.com/skirmer/excavating-insights-in-medical-no-shows
f. https://www.kaggle.com/kchandrasekhar/predict-with-logistic-regression-algorithm
g. The Healthcare No-Show Reduction Method- Darrel Farro_Thesis – Utrecht University

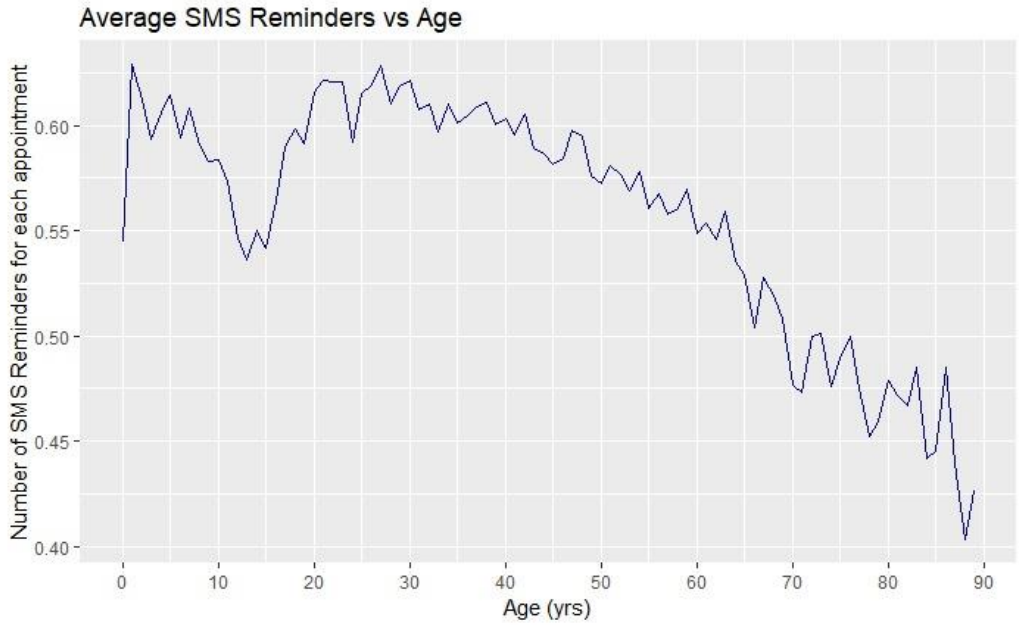
9. Appendix

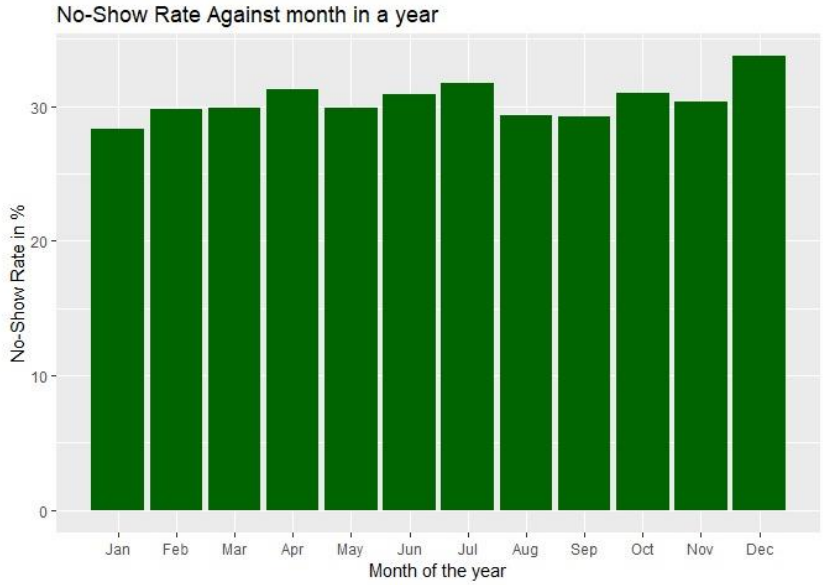

Absence Rate vs Gender
Absence Rate %
30
20
10
0
Male
Female
Gender

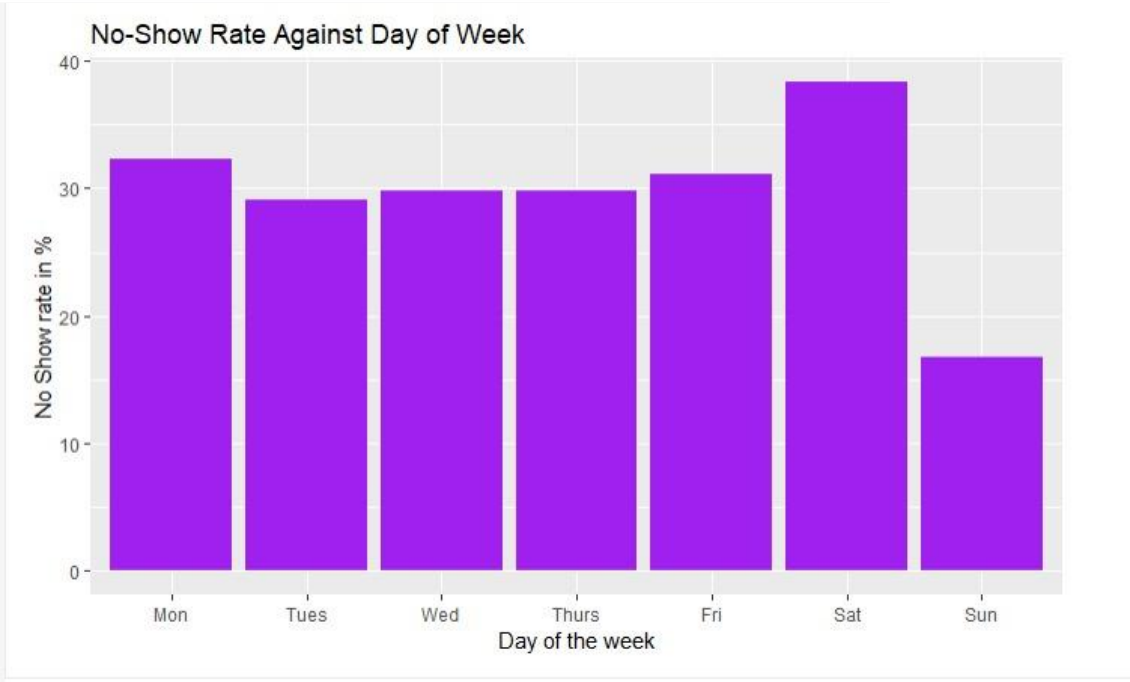

Imbalance in gender variable
count
200000
150000
100000
50000
0
F
M
Gender
Gender
F
M

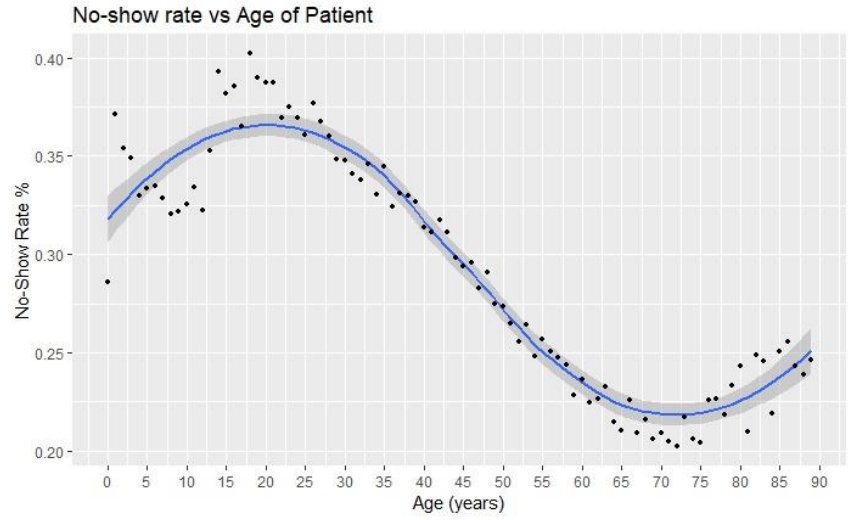

Balance in Status variable
count
100000
75000
50000
25000
0
No-Show
Show-Up
Status
Status
No-Show
Show-Up

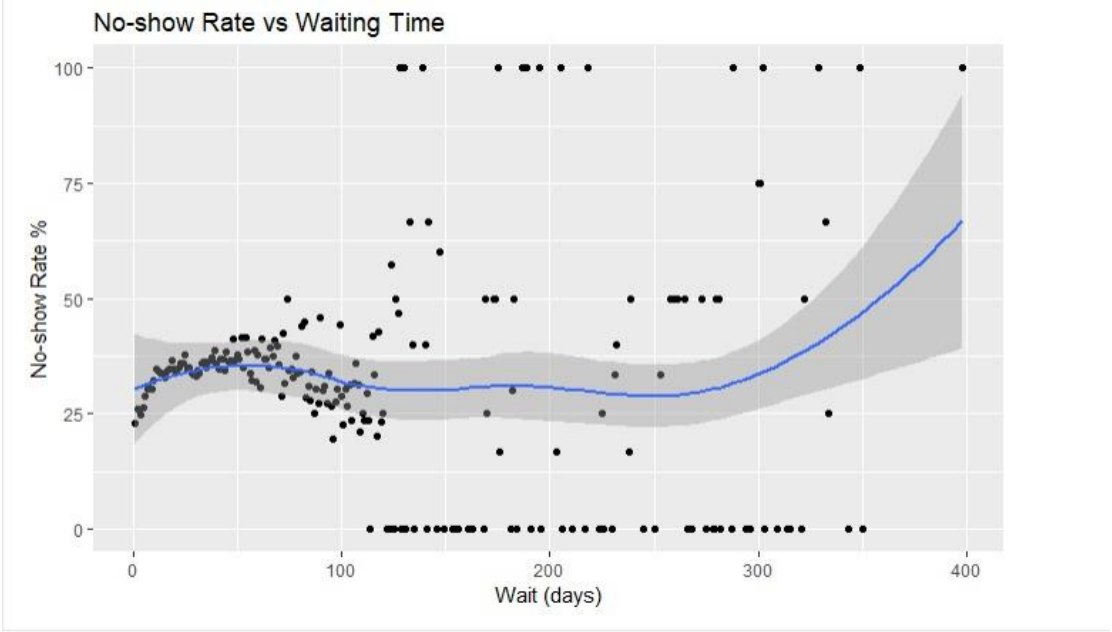

Number of Patients in each marker
Count
300 K
200 K
100 K
0 K
No
Yes
Scholarship
HyperTension
Diabetes
Alcoholism
0
1
2
3
4
Handicap
Sms_Reminder
Smokes
Tuberculosis

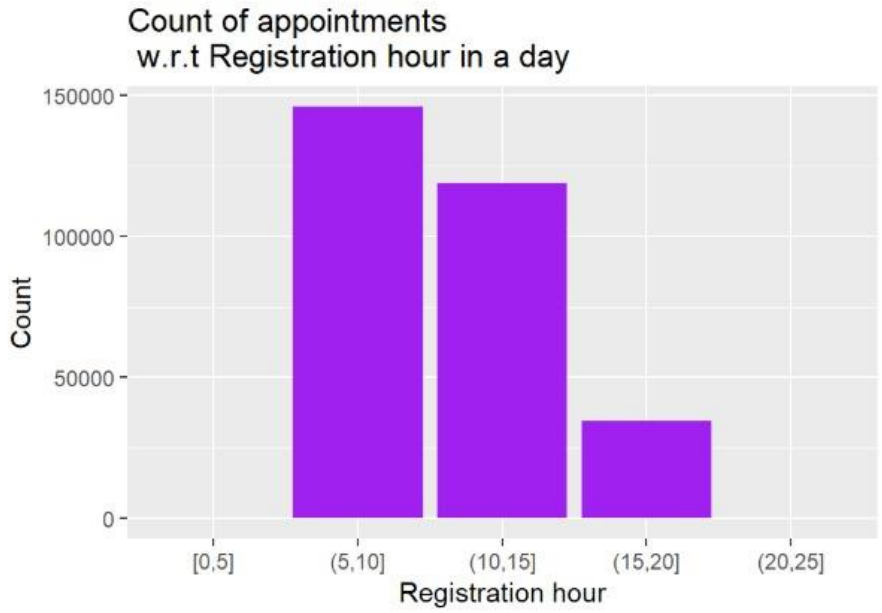

Average SMS Reminders vs Age
Number of SMS Reminders for each appointment
0.60
0.55
0.50
0.45
0.40
0
10
20
30
40
50
60
70
80
90
Age (yrs)

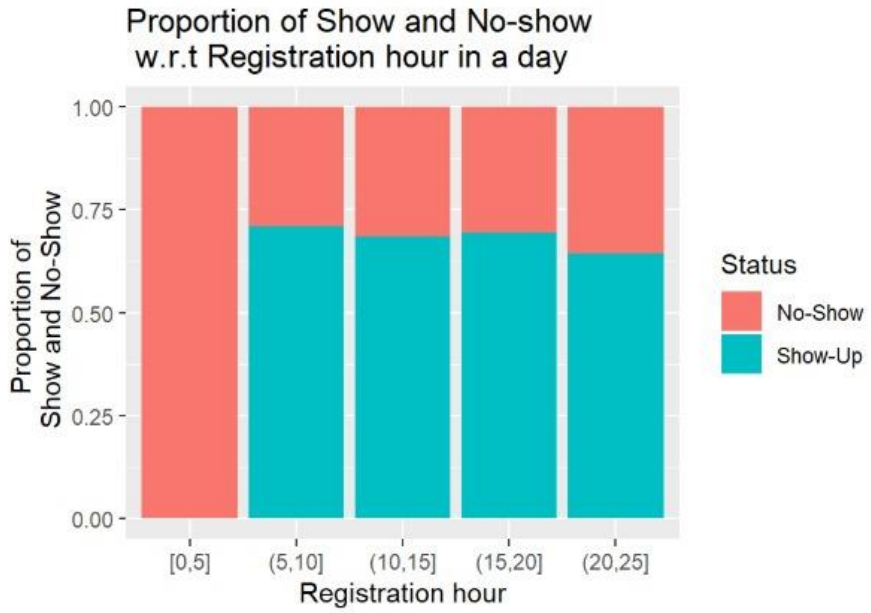

No-Show Rate Against month in a year
No-Show Rate in %
0
10
20
30
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month of the year

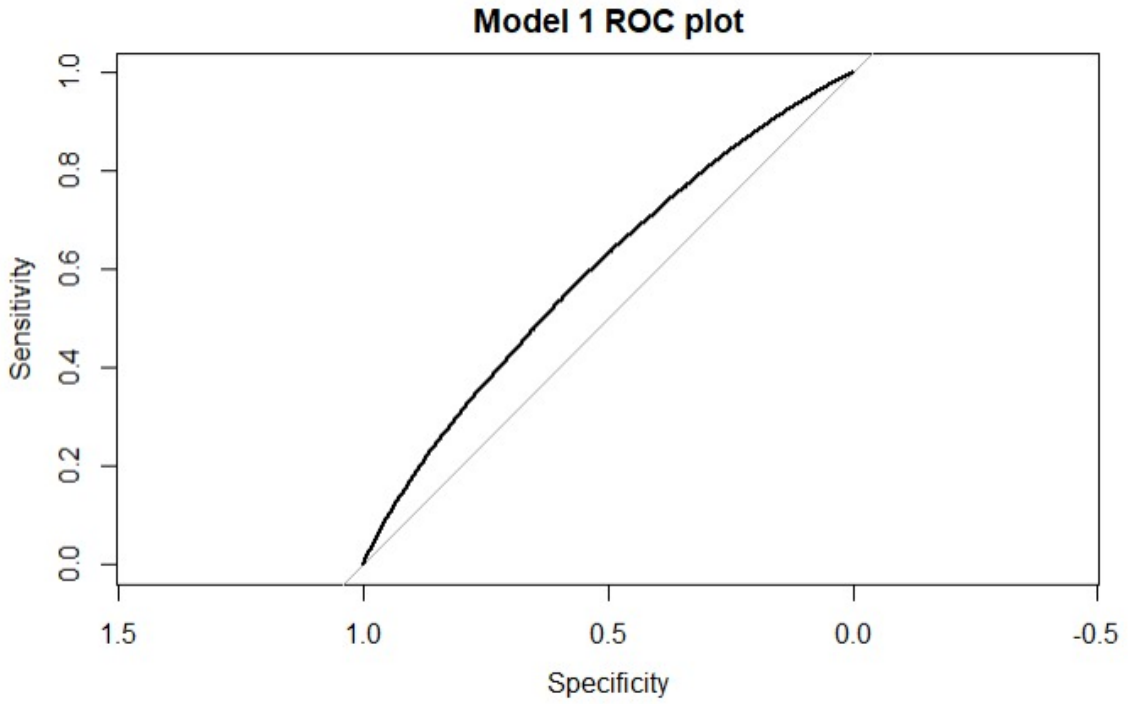

No-Show Rate Against Day of Week
No Show rate in %
0
10
20
30
40
Mon
Tues
Wed
Thurs
Fri
Sat
Sun
Day of the week

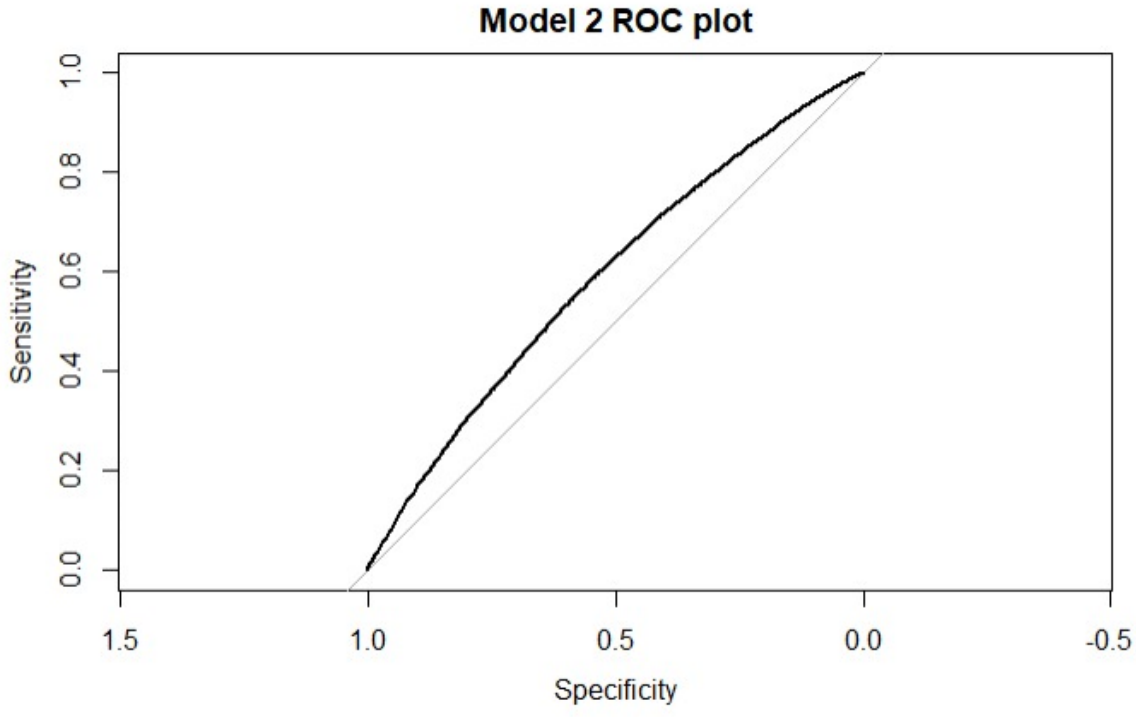

No-show rate vs Age of Patient
No-Show Rate %
0.20
0.25
0.30
0.35
0.40
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
Age (years)


No-show Rate vs Waiting Time
No-show Rate %
100
75
50
25
0
0
100
200
300
400
Wait (days)


Count of appointments
w.r.t Registration hour in a day
Count
150000
100000
50000
0
[0,5]
(5,10]
(10,15]
(15,20]
(20,25]
Registration hour


Proportion of Show and No-show
w.r.t Registration hour in a day
Proportion of
Show and No-Show
1.00
0.75
0.50
0.25
0.00
[0,5]
(5,10]
(10,15]
(15,20]
(20,25]
Registration hour
Status
No-Show
Show-Up


Model 1 ROC plot
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1.5
1.0
0.5
0.0
-0.5
Specificity


Model 2 ROC plot
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1.5
1.0
0.5
0.0
-0.5
Specificity