

Analysing The Data Extracted On **Not Showing Up for The Medical Appointment**

By
Farhein Akmal (17977491)

Overview:

The data was taken from the website Kaggle . The data set included 110527 records which is relevant to healthcare operations. After examining the dataset, I realized my interest towards health policies. The relationship between medical institutions and common people is an interesting topic to study and analyses. The data which was collected has a lot of errors which is very common in the real world and has its own value to do its examination.

The aim of the analysis is to examine the pattern by number of patients who skipped their appointment and see if it is possible to predict if a patient will show up for the appointment or not.

Understanding the Data:

The data set included the following variables:

- Patient ID
- Appointment ID
- Age
- Gender
- Neighborhood
- Appointment Date
- Scheduled Date
- SMS Received
- Disability
- Scholarship

I will analyze the no show data with Age, Gender, Neighborhood, Difference in days between Appointment Date and Scheduled Date, SMS Received and Scholarship. I will be leaving the disability variables as there are different kind of disabilities included in the data which would be needing more explanations.

Book1 - Excel Farhein Akmal

File Home Insert Page Layout Formulas Data Review View Tell me what you want to do

Clipboard Font Alignment Number Styles Cells Editing

Calibri 11 A A Wrap Text General \$ % , . 0 00 0.0 Conditional Formatting Format as Table Cell Styles Insert Delete Format AutoSum Fill Sort & Find & Filter Select

A1 PatientID

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|----|-------------|---------------|--------|----------------------|----------------------|-----|-------------------|-------------|--------------|----------|------------|---------|--------------|---------|
| | PatientID | AppointmentID | Gender | ScheduledDay | AppointmentDay | Age | Neighbourhood | Scholarship | Hipertension | Diabetes | Alcoholism | Handcap | SMS_received | No-show |
| 1 | 2.98725E+13 | 5642903 | F | 2016-04-29T18:38:08Z | 2016-04-29T00:00:00Z | 62 | JARDIM DA PENHA | 0 | 1 | 0 | 0 | 0 | 0 | No |
| 2 | 5.58998E+14 | 5642503 | M | 2016-04-29T16:08:27Z | 2016-04-29T00:00:00Z | 56 | JARDIM DA PENHA | 0 | 0 | 0 | 0 | 0 | 0 | No |
| 3 | 4.26296E+12 | 5642549 | F | 2016-04-29T16:19:04Z | 2016-04-29T00:00:00Z | 62 | MATA DA PRAIA | 0 | 0 | 0 | 0 | 0 | 0 | No |
| 4 | 8.67951E+11 | 5642828 | F | 2016-04-29T17:29:31Z | 2016-04-29T00:00:00Z | 8 | PONTAL DE CAMBURI | 0 | 0 | 0 | 0 | 0 | 0 | No |
| 5 | 8.84119E+12 | 5642494 | F | 2016-04-29T16:07:23Z | 2016-04-29T00:00:00Z | 56 | JARDIM DA PENHA | 0 | 1 | 1 | 0 | 0 | 0 | No |
| 6 | 9.59851E+13 | 5626772 | F | 2016-04-27T08:36:51Z | 2016-04-29T00:00:00Z | 76 | REPÚBLICA | 0 | 1 | 0 | 0 | 0 | 0 | No |
| 7 | 7.33688E+14 | 5630279 | F | 2016-04-27T15:05:12Z | 2016-04-29T00:00:00Z | 23 | GOIABEIRAS | 0 | 0 | 0 | 0 | 0 | 0 | Yes |
| 8 | 3.44983E+12 | 5630575 | F | 2016-04-27T15:39:58Z | 2016-04-29T00:00:00Z | 39 | GOIABEIRAS | 0 | 0 | 0 | 0 | 0 | 0 | Yes |
| 9 | 5.63947E+13 | 5638447 | F | 2016-04-29T08:02:16Z | 2016-04-29T00:00:00Z | 21 | ANDORINHAS | 0 | 0 | 0 | 0 | 0 | 0 | No |
| 10 | 7.81246E+13 | 5629123 | F | 2016-04-27T12:48:25Z | 2016-04-29T00:00:00Z | 19 | CONQUISTA | 0 | 0 | 0 | 0 | 0 | 0 | No |
| 11 | 7.34536E+14 | 5630213 | F | 2016-04-27T14:58:11Z | 2016-04-29T00:00:00Z | 30 | NOVA PALESTINA | 0 | 0 | 0 | 0 | 0 | 0 | No |
| 12 | 7.54295E+12 | 5620163 | M | 2016-04-26T08:44:12Z | 2016-04-29T00:00:00Z | 29 | NOVA PALESTINA | 0 | 0 | 0 | 0 | 0 | 1 | Yes |
| 13 | 5.66655E+14 | 5634718 | F | 2016-04-28T11:33:51Z | 2016-04-29T00:00:00Z | 22 | NOVA PALESTINA | 1 | 0 | 0 | 0 | 0 | 0 | No |
| 14 | 9.11395E+14 | 5636249 | M | 2016-04-28T14:52:07Z | 2016-04-29T00:00:00Z | 28 | NOVA PALESTINA | 0 | 0 | 0 | 0 | 0 | 0 | No |
| 15 | 9.98847E+13 | 5633951 | F | 2016-04-28T10:06:24Z | 2016-04-29T00:00:00Z | 54 | NOVA PALESTINA | 0 | 0 | 0 | 0 | 0 | 0 | No |
| 16 | 99948393975 | 5620206 | F | 2016-04-26T08:47:27Z | 2016-04-29T00:00:00Z | 15 | NOVA PALESTINA | 0 | 0 | 0 | 0 | 0 | 1 | No |
| 17 | 8.45744E+13 | 5633121 | M | 2016-04-28T08:51:47Z | 2016-04-29T00:00:00Z | 50 | NOVA PALESTINA | 0 | 0 | 0 | 0 | 0 | 0 | No |
| 18 | 1.4795E+13 | 5633460 | F | 2016-04-28T09:28:57Z | 2016-04-29T00:00:00Z | 40 | CONQUISTA | 1 | 0 | 0 | 0 | 0 | 0 | Yes |
| 19 | 1.71354E+13 | 5621836 | F | 2016-04-26T10:54:18Z | 2016-04-29T00:00:00Z | 30 | NOVA PALESTINA | 1 | 0 | 0 | 0 | 0 | 1 | No |
| 20 | 7.22329E+12 | 5640433 | F | 2016-04-29T10:43:14Z | 2016-04-29T00:00:00Z | 46 | DA PENHA | 0 | 0 | 0 | 0 | 0 | 0 | No |
| 21 | 6.22257E+14 | 5626083 | F | 2016-04-27T07:51:14Z | 2016-04-29T00:00:00Z | 30 | NOVA PALESTINA | 0 | 0 | 0 | 0 | 0 | 0 | Yes |
| 22 | 1.21548E+13 | 5628338 | F | 2016-04-27T10:50:45Z | 2016-04-29T00:00:00Z | 4 | CONQUISTA | 0 | 0 | 0 | 0 | 0 | 0 | Yes |
| 23 | 8.6323E+14 | 5616091 | M | 2016-04-25T13:29:16Z | 2016-04-29T00:00:00Z | 13 | CONQUISTA | 0 | 0 | 0 | 0 | 0 | 0 | Yes |

Full Data Reduced Data Gender Age Neighbourhood Scheduled & App Date Diff

Reason for Choosing this data set:

No show appointment dataset relates to my interest in knowing more about health operations and health policies. This specific dataset aims to identify patients at substantial risk for not showing up for the appointments. The goal is to target the pattern of these no-show patients. I wanted to understand a range of factors contributing behind patients not showing up for the appointments.

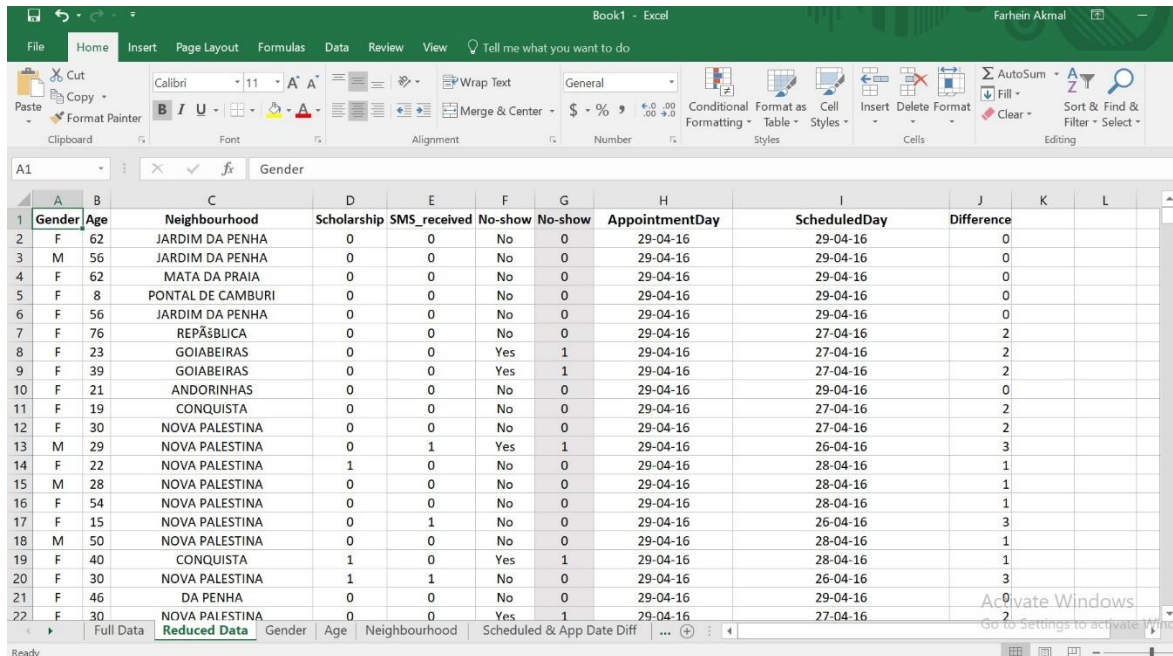
I wish to understand how each variable contribute to the probability of patients missing their appointments.

- Age: Which age group is more likely to miss their appointment?
- Gender: Whether males are more likely to not show up for their appointments or female?
- Neighborhood: If there is/are people from a specific neighborhood/s who results in not showing up for the appointment.
- Difference in Scheduled date and Appointment date: Identify the pattern by analyzing the relationship between patients who skipped their appointment and when did they book for it.
- SMS Received: Is sending patients a reminder text for their appointment make it less likely for them to skip their appointment?
- Scholarship: Is patients with the scholarship have any relation with skipping the appointments?

The analysis was done with the help of Excel to get the basic descriptive analysis as it explains the overview of the issue.

Importing the data:

The data was imported in an excel file in the .csv format from the website Kaggle. The whole data set was copied to a new excel sheet, Book1. The data was reduced and rearranged on the different sheet of the same excel workbook. The data was rearranged according to the variables on which analysis has to be done.



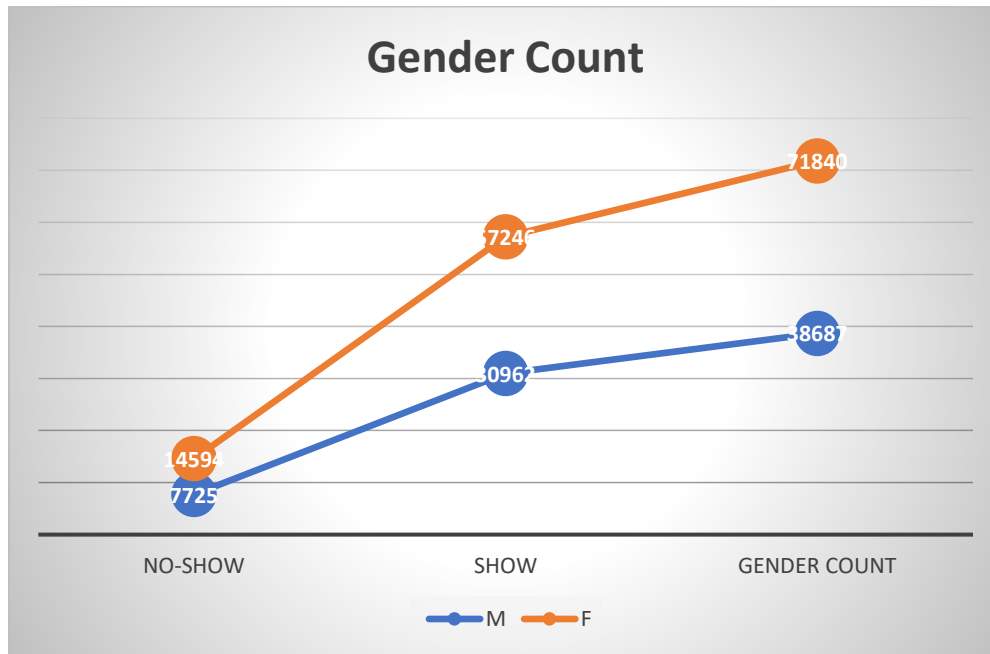
| Gender | Age | Neighbourhood | Scholarship | SMS_received | No-show | AppointmentDay | ScheduledDay | Difference |
|--------|-----|-------------------|-------------|--------------|---------|----------------|--------------|------------|
| F | 62 | JARDIM DA PENHA | 0 | 0 | No | 29-04-16 | 29-04-16 | 0 |
| M | 56 | JARDIM DA PENHA | 0 | 0 | No | 29-04-16 | 29-04-16 | 0 |
| F | 62 | MATA DA PRAIA | 0 | 0 | No | 29-04-16 | 29-04-16 | 0 |
| F | 8 | PONTAL DE CAMBURI | 0 | 0 | No | 29-04-16 | 29-04-16 | 0 |
| F | 56 | JARDIM DA PENHA | 0 | 0 | No | 29-04-16 | 29-04-16 | 0 |
| F | 76 | REPÚBLICA | 0 | 0 | No | 29-04-16 | 27-04-16 | 2 |
| F | 23 | GOIABEIRAS | 0 | 0 | Yes | 29-04-16 | 27-04-16 | 2 |
| F | 39 | GOIABEIRAS | 0 | 0 | Yes | 29-04-16 | 27-04-16 | 2 |
| F | 21 | ANDORINHAS | 0 | 0 | No | 29-04-16 | 29-04-16 | 0 |
| F | 19 | CONQUISTA | 0 | 0 | No | 29-04-16 | 27-04-16 | 2 |
| F | 30 | NOVA PALESTINA | 0 | 0 | No | 29-04-16 | 27-04-16 | 2 |
| M | 29 | NOVA PALESTINA | 0 | 1 | Yes | 29-04-16 | 26-04-16 | 3 |
| F | 22 | NOVA PALESTINA | 1 | 0 | No | 29-04-16 | 28-04-16 | 1 |
| M | 28 | NOVA PALESTINA | 0 | 0 | No | 29-04-16 | 28-04-16 | 1 |
| F | 54 | NOVA PALESTINA | 0 | 0 | No | 29-04-16 | 28-04-16 | 1 |
| F | 15 | NOVA PALESTINA | 0 | 1 | No | 29-04-16 | 26-04-16 | 3 |
| M | 50 | NOVA PALESTINA | 0 | 0 | No | 29-04-16 | 28-04-16 | 1 |
| F | 40 | CONQUISTA | 1 | 0 | Yes | 29-04-16 | 28-04-16 | 1 |
| F | 30 | NOVA PALESTINA | 1 | 1 | No | 29-04-16 | 26-04-16 | 3 |
| F | 46 | DA PENHA | 0 | 0 | No | 29-04-16 | 29-04-16 | 0 |
| F | 30 | NOVA PALESTINA | 0 | 0 | Yes | 29-04-16 | 27-04-16 | 2 |

Cleaning the Data and Making the Data Compatible:

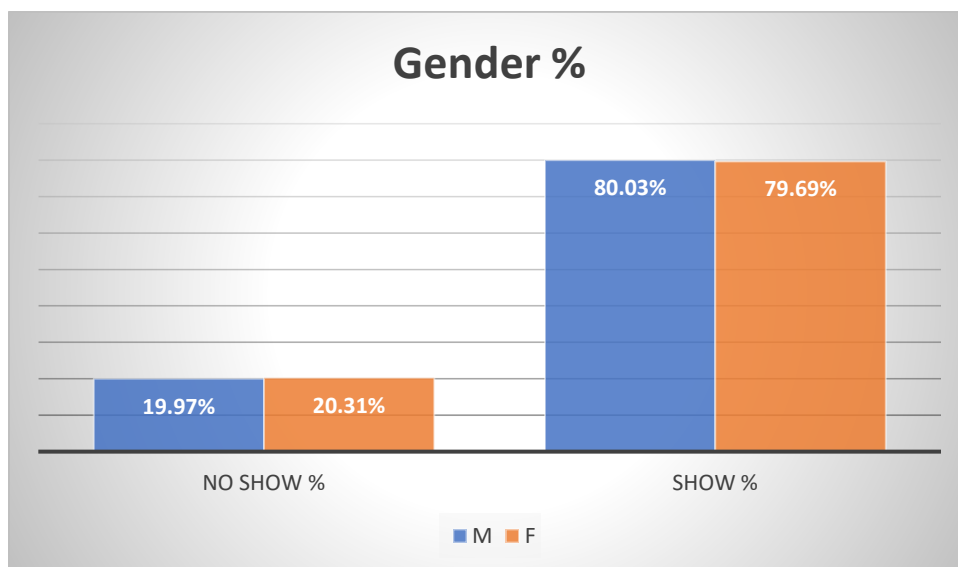
- Age: There was one record for which the age was in negative, it was converted into positive value to get the absolute value for the age of the patients, assuming that it must be a typing error.
- Schedule Date and Appointment Date: There were a few records for which the scheduled date was bigger than the appointment date. Considering the data was filled in the wrong cell, the greater date was then corrected to appointment date and the smaller date was filled under scheduled date for each wrong record. The difference of the appointment date and scheduled date were calculated in the “Reduced Data” sheet of the excel sheet.
- No Show: A new column was created in which the value for not showing up was assigned “1” and the value for showing up was assigned “0”
- We group the main columns for every sheet in the workbook so that, when we filter our records using the filter option on Excel, the data we see adjacent to the cell, doesn’t mix up with other records.

Gender Analysis:

The total number of males and females booked for the appointment were 38687 and 71839 respectively. 14594 out of 71839 females didn't show up for the appointment whereas 7725 out of 38687 males didn't show up for the appointment. If we see just numbers, females have missed more appointments compare to male.



It's important that we see the ratio of no-show males and no-show females to the total males and females. After calculating the percentage, we realized the ration is almost the same. 19.97% of males and 20.31% of females didn't show up for the appointment.

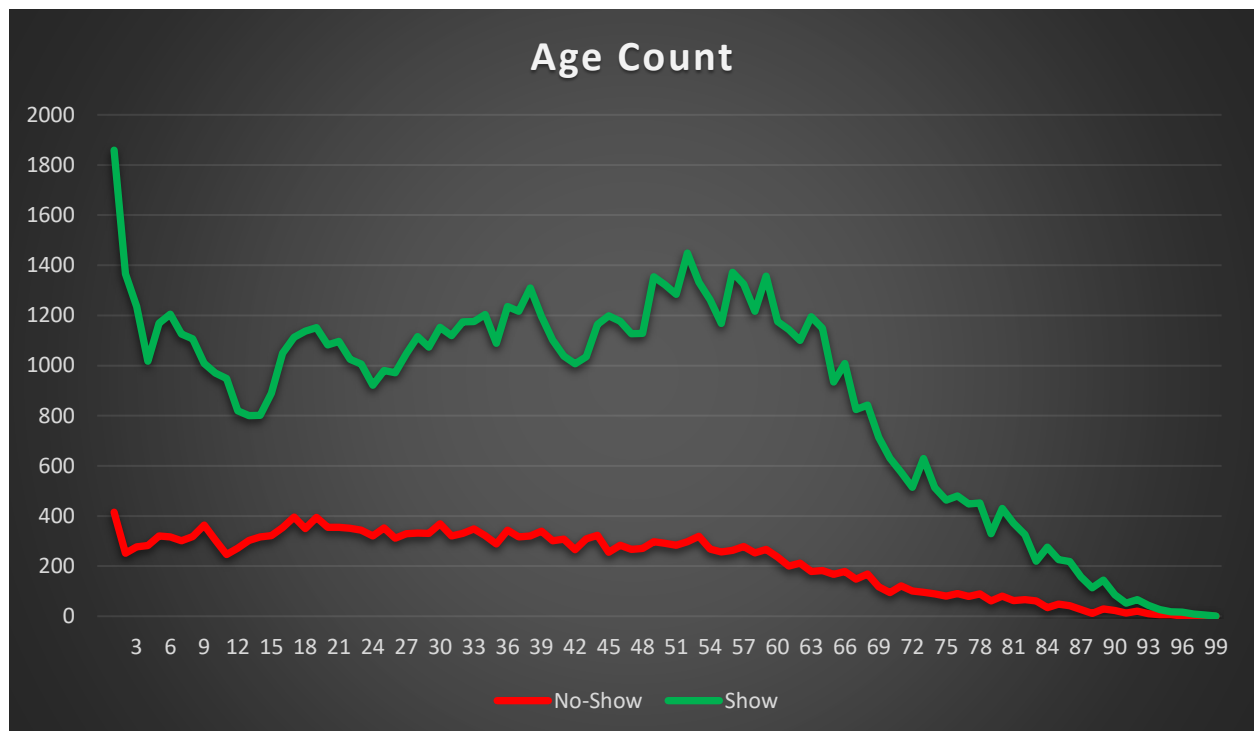


This shows that we should not analyze a data looking at its value, we should always concentrate on the ratio to compare the result as it is more valuable. Hence, looking at the gender data in regards to the no-show appointment data, we infer that the gender of the patients does not affect the appointment as there is no substantial difference between the gender's data. Therefore, gender can be ignored while framing the weighted role of the gender attribute to the probability for not showing up for an appointment.

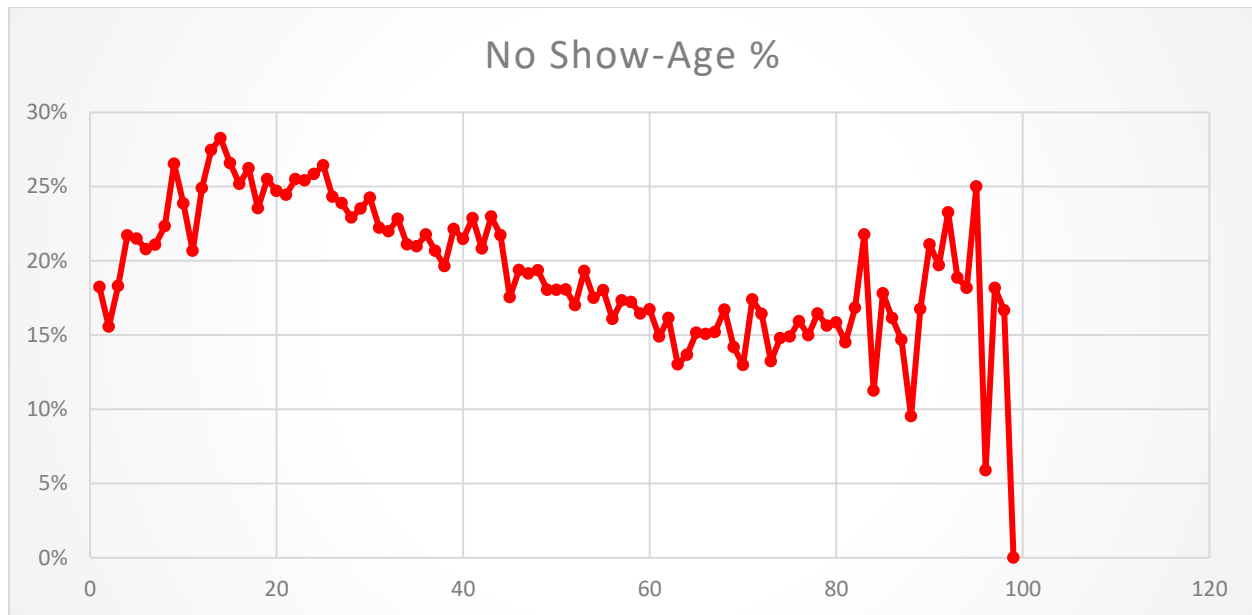
Age Analysis:

The variable age can help us notice the bracket of age of patients, who misses their appointments compare to other group of ages. I initially started with splitting the age into different age bins, but later realized the analysis with splitting the age into different bins will make the analysis weaker. It would be better if we treat each age separately. For the given data the age ranged from 0 to 115, whereas considering the practicality of life, I decided to work on age range from 1 to 99 for the largest realist bracket.

The graph below shows the number of patients showed up and number of patients which didn't show up for the appointment plotted according to the age (from 1 to 99).

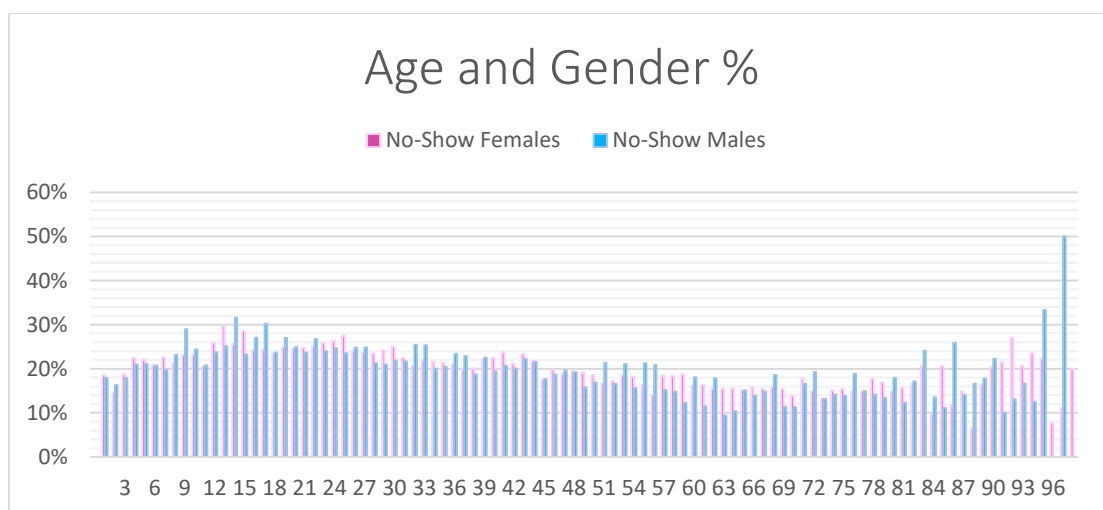


As we discussed earlier, it is important to see the percentage of the attribute, therefore, we calculated the percentage of the number of patients who didn't show up for the appointment upon the total number of patients of that specific age. The resulted graph is below:



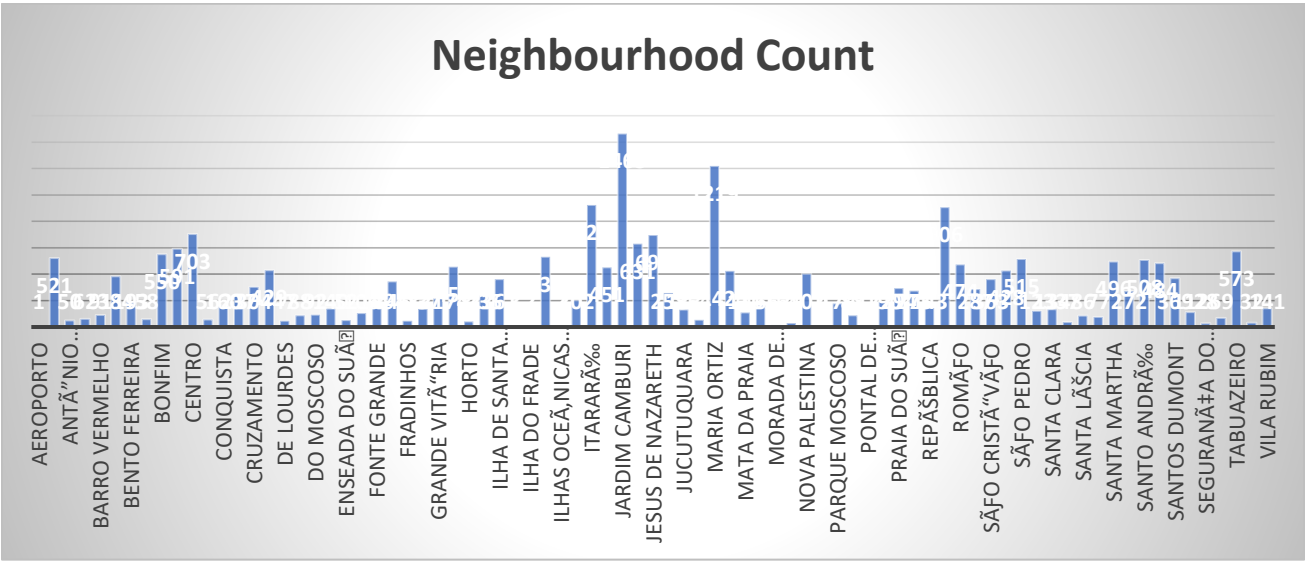
Looking at the graph, we can infer that the variable age does contribute a lot on the rate of the no-show appointments. The age range from 13 to 25 years old patients has the maximum cancellations. The rate then decreases till the age 80 and then again increases. Considering the number of people in the later age group has a small sample size, we can explain why there is a sudden increase in the graph at the end.

We then try to observe males and females separately with the age, below is the graph for the same, which shows there is no significant difference between the gender.

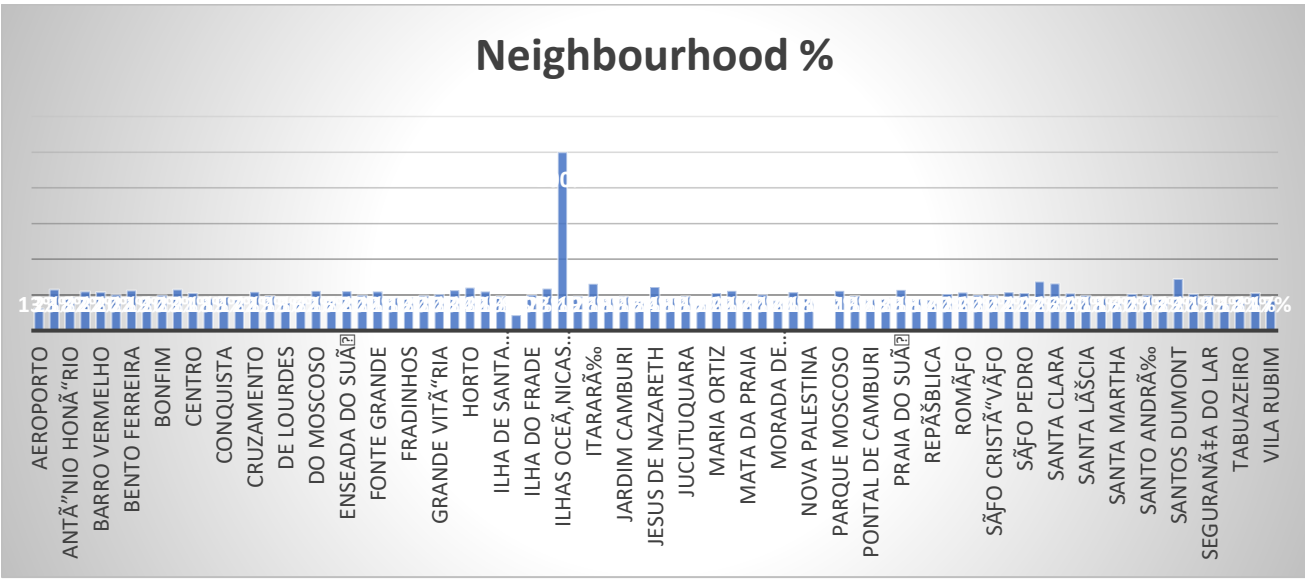


Neighborhood Analysis:

In this analysis, we want to extract information on which suburbs have the highest probability for the patients not showing up for their appointments. There are in total 81 unique suburbs in the records. According to the count of no show of each suburb the line graph looks like this.



Whereas if we have a look on the ratio of number of no-show up is to total number of appointments from the suburbs, the graph looks like this:



These graphs show a vital information as in this case, the number of no shows and percentage, both holds a very important role in analysis. We can't just rely on the percentage of no show as it might be misleading, for example a suburb has two no-shows and the total number of appointment from that suburb was two as well, this will result in 100% no show, in our data "ILHAS OCEÂNICAS DE TRINDADE" is that suburb. It also means that ILHAS OCEÂNICAS DE

TRINDADE brings the low revenue and has non-reliable patients in terms of showing up, compare to other suburbs. Therefore leaving this specific suburb, we focus on other top 4 suburbs with highest percentage of no-show, which are:

| Neighbourhood | No Show % |
|---------------|-----------|
| ITARARÃ | 26% |
| SANTA CECÍLIA | 27% |
| SANTA CLARA | 26% |
| SANTOS DUMONT | 29% |

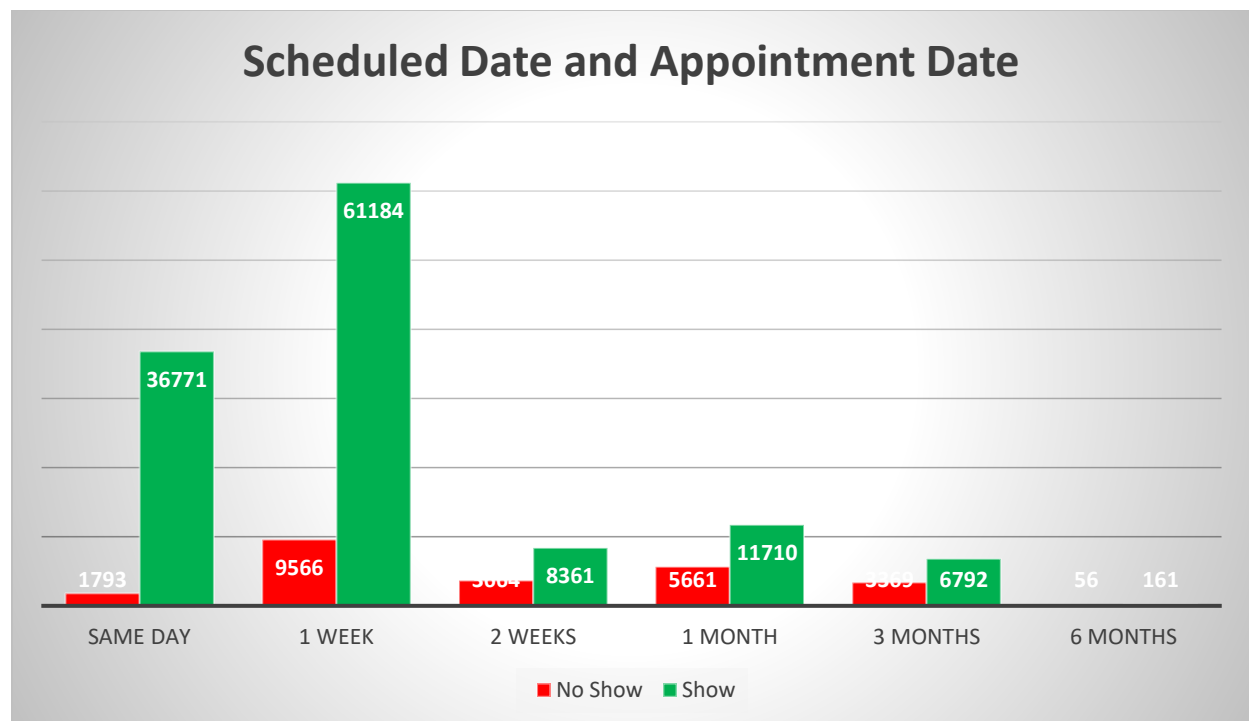
Scheduled Date and Appointment Date Analysis:

This variable where we study the difference in the two dates ie the date when the appointment was book and the date for which the appointment was book; gives an idea that how likely it is to skip an appointment with n number of duration in between. It let us find the relationship between the difference in dates with the no-show appointment.

For our data, I have grouped the difference in dates in following categories:

- Same Day: which means the appointment was booked for the same day
- 1 Week: which means the there is a difference of 0 to 7 days
- 2 Weeks: difference of 8 to 14 days
- 1 Month: difference of 15 to 30 days
- 3 Months: difference of 31 to 90 days
- 6 Months: difference of 91 to 180 days

There was no data for more than 180 days.



According to this graph, we can say the most amount of no-show was seen when the difference in the scheduled date and appointment date was of 1 week i.e. 1 to 7 days. It came as a surprise to me to see that were a sizable number of no-show for the same day appointments as well, whereas according to me the same day appointment is less likely to be missed. It would be

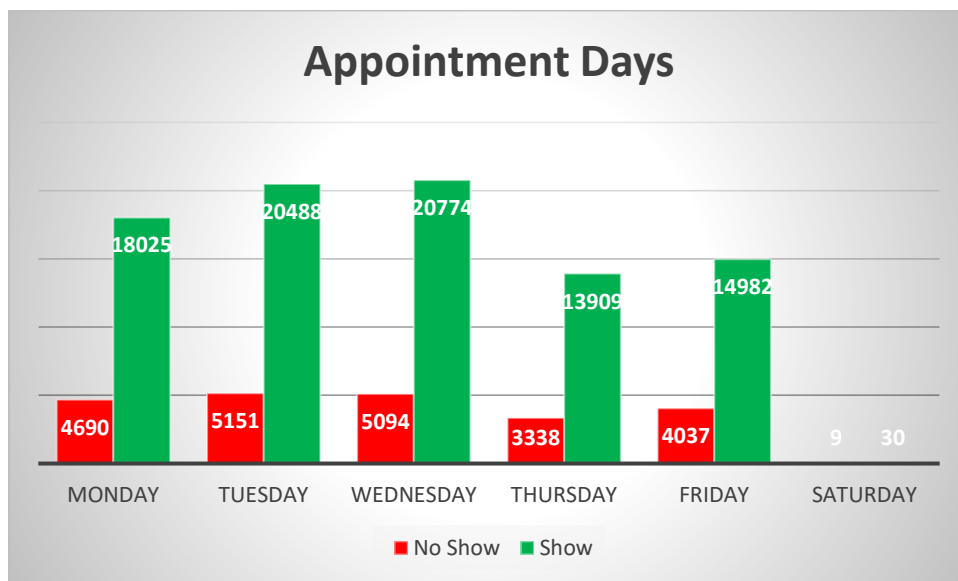
interesting we find out the suburbs of these same day no-show appointments, as I assume either patient lives very near to the clinic or very far away from the clinic.

Appointment Day Analysis:

The total number of appointments booked on the specific day of the week are:

| Days | Total Booked |
|-----------|--------------|
| Monday | 22715 |
| Tuesday | 25639 |
| Wednesday | 25868 |
| Thursday | 17247 |
| Friday | 19019 |
| Saturday | 39 |
| Sunday | 0 |

We plotted a graph of no-show and show with respect to the days of the week. According to this graph, the most appointments were booked for Tuesday and Wednesday; and the most no shows were also for Tuesday and Wednesday.

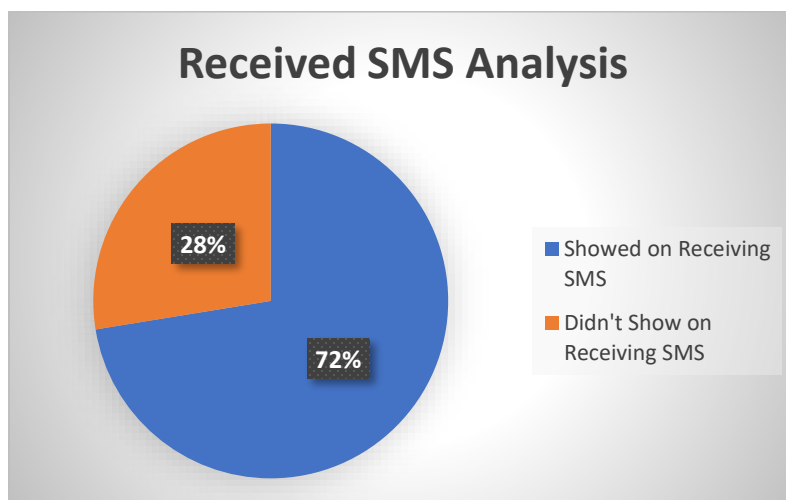
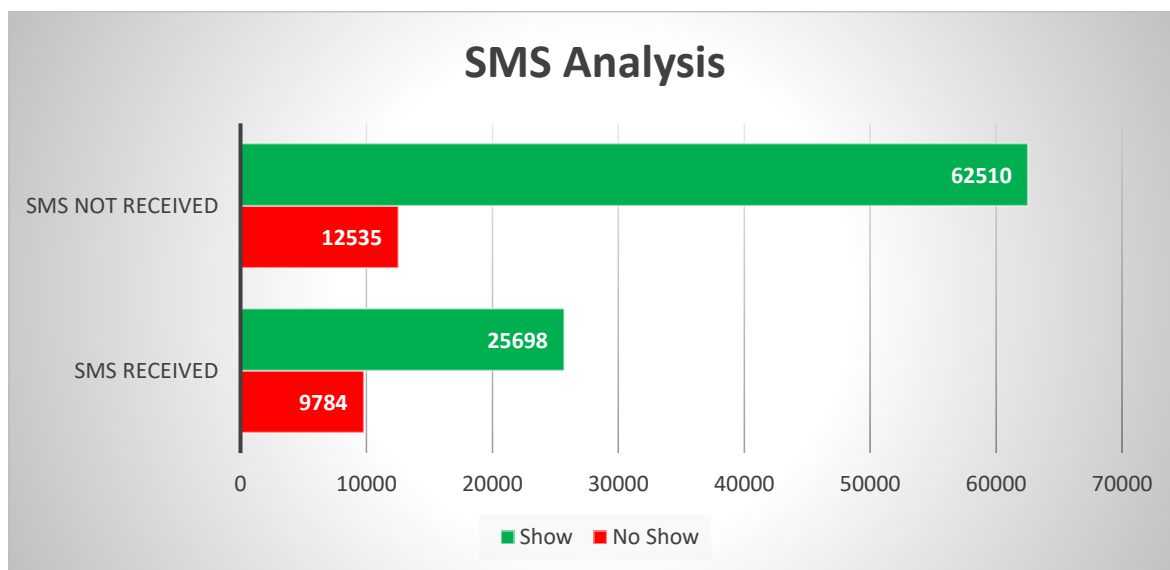


SMS Sent/Receive Analysis:

The SMS were sent to the patients as a reminder for their appointments, the total number of SMS sent were 35482 out of 110527 records, which means 75045 patients didn't receive the SMS. We calculated the number of no shows and shows for the appointment in respect to the SMS received or not.

| | No Show | Show |
|------------------|---------|-------|
| SMS Received | 9784 | 25698 |
| SMS Not Received | 12535 | 62510 |

Later we represented these figures on the bar graph.



| | |
|------------------------------|-----|
| Showed on Receiving SMS | 72% |
| Didn't Show on Receiving SMS | 28% |

Usually, sending a reminder can decrease the no-show appointments, still there were still 28% of patients who didn't show up for the appointment even after receiving the SMS. An interesting factor to study would be at what time these messages were sent.

We performed the chi-square test for this variable. A chi-square test is a way to determine the relationship between two categorical variables. (Stephanie, n.d.)

$$\chi^2_c = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where,

c: is the degree of freedom

O: is the observed value

E: is the expected value

For our data,

| | SMS Received | SMS Not Received | Total |
|----------------|---------------------|-------------------------|--------------|
| No Show | 9784 | 12535 | 22319 |
| Show | 25698 | 62510 | 88208 |
| Total | 35482 | 75045 | |

Expected = 22319 / 88208 = 0.253

Observed = 9784 / 25698 = 0.380

Chi-square value = ((0.380-0.253)^2)/0.253

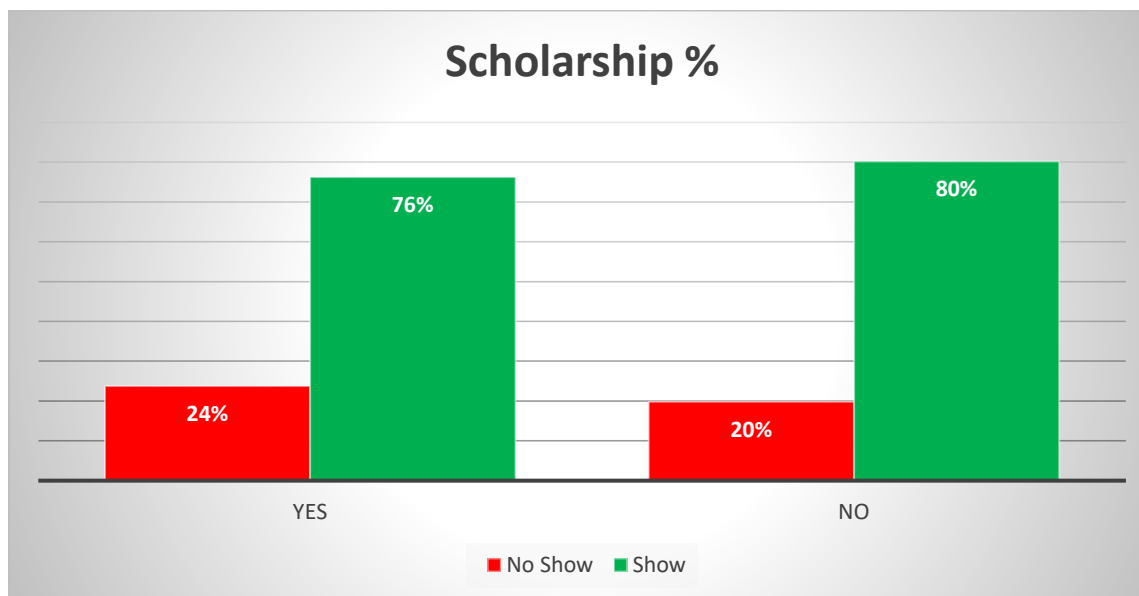
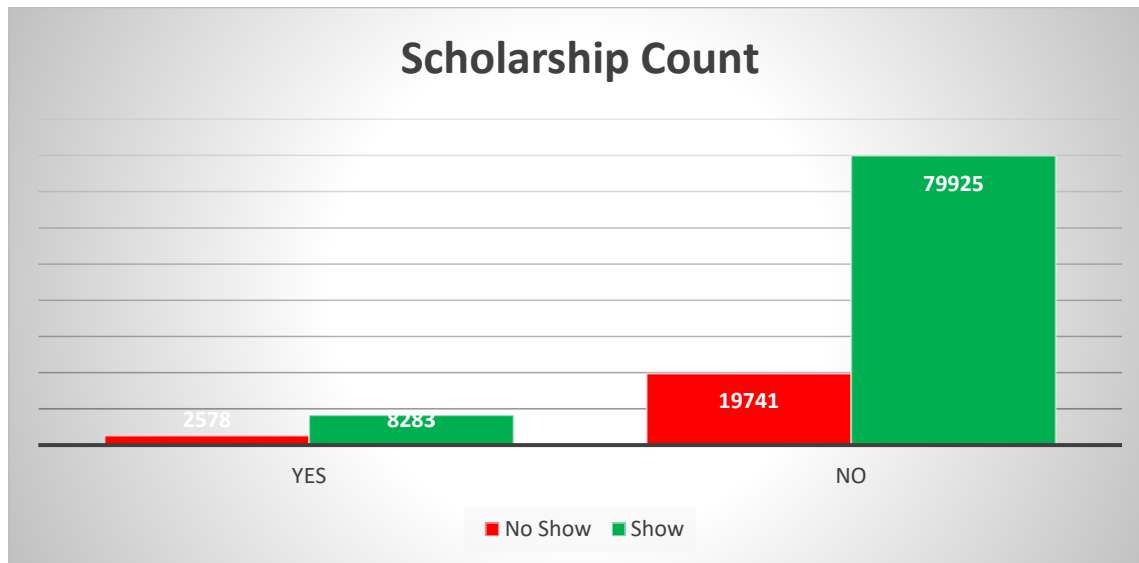
= 0.0637

With a degree of freedom 1 and chi-square value 0.0637, the cumulative probability i.e. $P(\chi^2 \leq CV)$ is 0.2

This shows that receiving a SMS does have a relationship with the patients' presence for their appointment.

Scholarship Analysis:

The variable informs whether the patient has received a scholarship or not. 10861 out of 110527 patients have received the scholarship. 2578 patients out of those 10861 with scholarship patients didn't show up for the appointment.



The difference is not huge but still comparatively the patients with scholarship missed their appointments more.

We performed the chi-square test for this variable as well.

| | Scholarship | No Scholarship | Total |
|---------|-------------|----------------|-------|
| No Show | 2578 | 19741 | 22319 |
| Show | 8283 | 79925 | 88208 |
| Total | 10861 | 99666 | |

Expected = $22319 / 88208 = 0.253$

Observed = $2578 / 19741 = 0.130$

Chi-square value = $((0.130 - 0.253)^2) / 0.253$

= 0.972

With a degree of freedom 1 and chi-square value 0.972, the cumulative probability i.e. $P(X^2 \leq CV)$ is 0.68

This shows that having a scholarship does have a relationship with the patients' presence for their appointment.

Conclusion:

The gender attribute doesn't contribute much to the prediction of patients not showing up for the appointment whereas when it comes to the age, the age group 13 to 25 years had the maximum number of no-show appointments. The top 4 neighborhoods which showed the missing of the appointment by 26-29% were ITARARÃ%, SANTA CECÃLIA, SANTA CLARA, SANTOS DUMONT. Tuesday and Wednesday were the two days when there were the maximum no-show appointments and majority of these missed appointments were booked within the 7 days. The clinic sent SMS reminders to the patients still there were 28% of patients who didn't show up for the appointment even after receiving a SMS. The scholarship attribute didn't give a huge result for the prediction but still patients with the scholarship missed their appointment 4% more compare to the patients without scholarship.

Future Work:

The analysis can further be continued for the disability variable. Also, the dataset can be upgraded by adding variables like the cause of the appointment, was it a follow-up appointment or a new appointment, the time of the SMS sent to the patients and more information about the neighborhood location.

It would be interesting if we could also gather the information on why did the patients didn't show up for the appointment. Although it would be difficult to get the response from the patients who haven't co-operated with the appointment system to come forward and participate in research which questions them the reason of their no-show for the appointment, but it's not impossible. (George & Rubin, 2003)

"Interventions to reduce non-attendance need robust evaluation within a broad-based context that includes patient, organizational, quality and health economic perspectives." (George & Rubin, 2003)

References:

George, A., & Rubin, G. (2003). Non-attendance in general practice: a systematic review and its implications for access to primary health care. *Family Practice*, 20(2), 178-184.

Stephanie. (n.d.). Chi-Square Statistic: How to Calculate It / Distribution - Statistics How To. Retrieved from <http://www.statisticshowto.com/probability-and-statistics/chi-square/>