No Show Appointment Data Analysis

Done by Okonkwo Chukwuebuka Malcom

Dated: 6th August, 2022

I carried out an analysis on the No Show Appointment Dataset using Python which can be gotten from Kaggle. This dataset collects information from over 100,000 medical appointments from 10th November, 2015 to 8th June, 2016 in Brazil and is focused on the question of whether or not patients show up for their appointment. A number of characteristics about the patient are included in each row and details of each row can be found in their respective fields (columns).

To be precise, the dataset consist of 110,527 rows of data and 14 columns. The details of the columns are given below;

01 – PatientId - Identification of a patien

02 – AppointmentID - Identification of each appointment

03 – Gender - Male or Female.

04 – ScheduledDay - The day of the actual appointment, when they have to visit the doctor.

05 – AppointmentDay- The day someone called or registered the appointment, this is before appointment of course.

06 – Age - How old is the patient.

07 – Neighbourhood - Where the appointment takes place.

08 – Scholarship - scholarship variable means this concept - Wikipedia. True of False .

09 – Hipertension - True or False

10 – Diabetes - True or False

11 – **Alcoholism -** True or False

12 – **Handcap -** True or False

13 - **SMS_received -** 1 or more messages sent to the patient.

14 - **No-show -** it says 'No' if the patient showed up to their appointment, and 'Yes' if they did not show up.

I loaded the data on my Jupyter Notebook and I also imported my Numpy, Pandas, and Matplotlib which are the python packages I would use to perform my analysis.

Before I began my analysis, I had to understand the data and draft research questions which I intend to answer form the data. The Research Questions I came up with are:

- Research Question 1: Overall Attendance Rate of the Appointments
- Research Question 2: Proportion Of Gender Amongst Patients and their overall Attendance ratio
- Research Question 3: Trends in Number of Appointments
- Research Question 4: Age Distribution of the Patients
- Research Question 5: Appointments by Neighborhood
- Research Question 6: Does Receiving SMS influence the Attendance Rate?
- Research Question 7: The percentage of Attendance with respect to illness
- Research Question 8: Summary Statistics on the Number of Appointment Days before the Scheduled Day

Before I could start to work on the research question, I performed data wrangling and data manipulation by renaming some of the columns which were misspelt and I also changed the cases of the column names which would them easier to work with. I changed the data types of some of the columns.

The no-show columns had an encoding that could be confusing so I had to recode it from (No, Yes) to (Present, Absent).

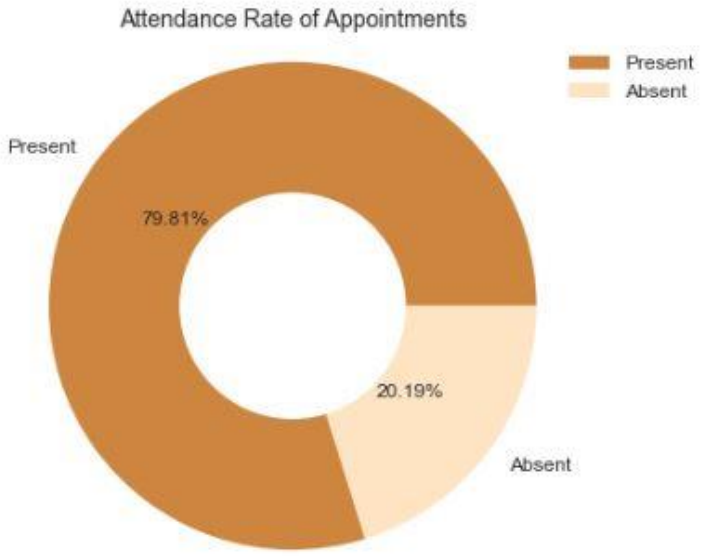
The dataset does not have null values or duplicate rows. From the age column, the minimum was given as -1 which is very unlikely so I dropped row (1 row).

On checking the patientid and appointmentid columns, I figured out that there are 62298 unique patients that have booked 110,526 appointments.

Exploratory Data Analysis / Findings

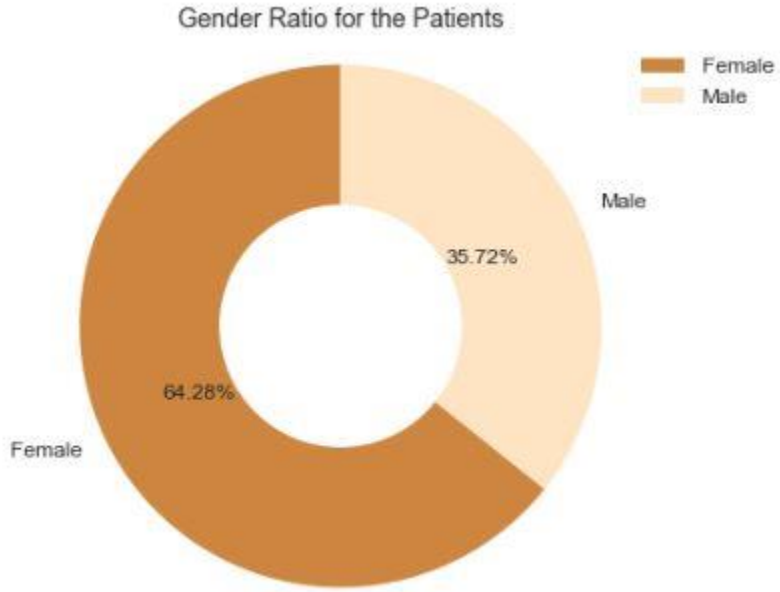
- Research Question 1

To begin my analysis with the first research question, I used a pie chart or visualize it. From my analysis, there was an attendance rate of 79.81% from the patients for their appointments.



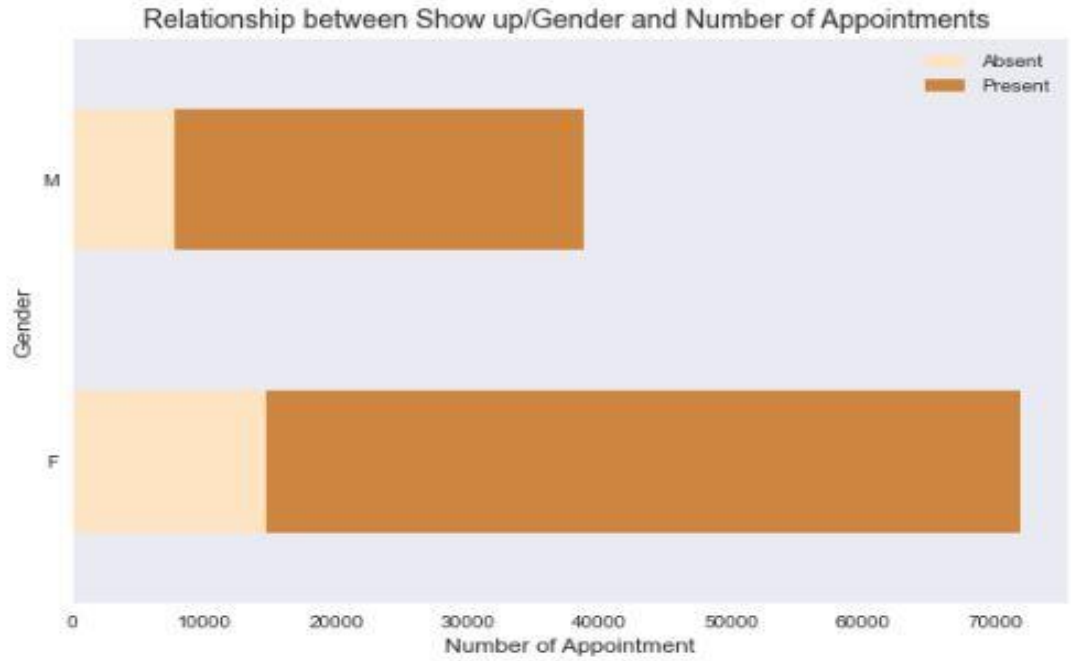
- Research Question 2

To know the proportion of gender among the patients, I represent it with a pie chart. It shows that 64.28% of the patients are female while the rest are male



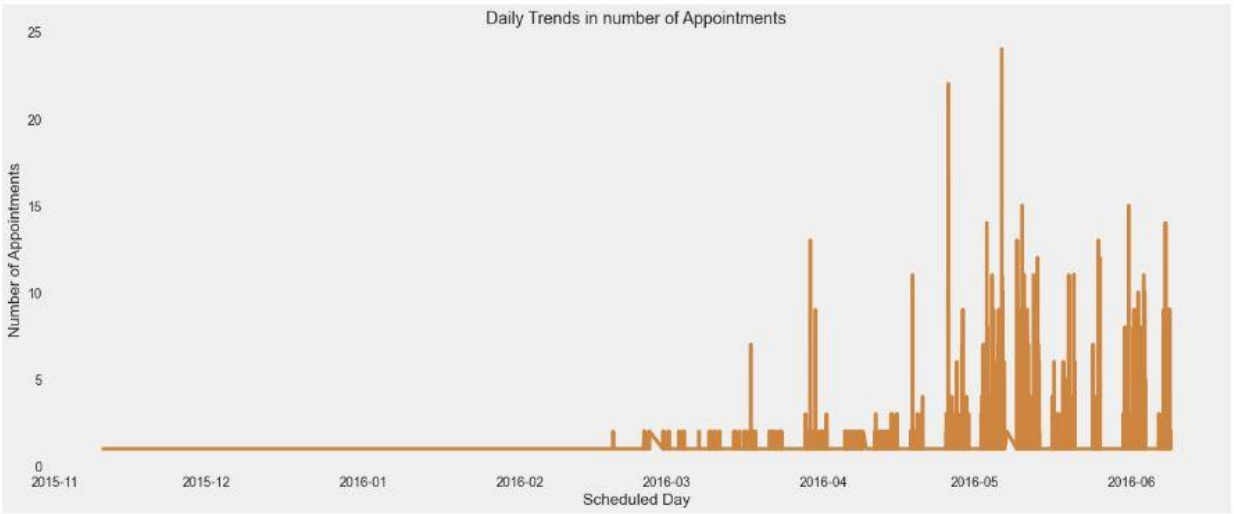
To check the attendace rate of each gender, I used a bar plot to visualize it

no_show	Absent	Present
gender		
F	14594	57245
M	7725	30962



- Research Question 3:

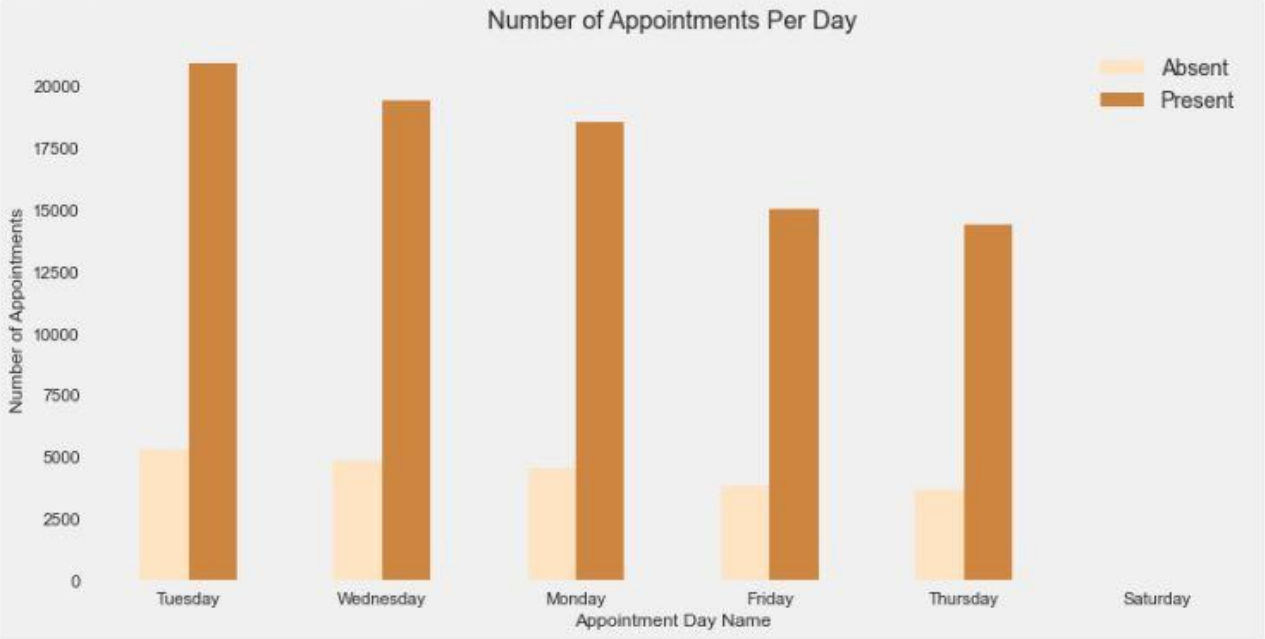
I created a line chart to represent the number of appointment for a day to show the trend of appointments schedule per day.



From the chart, the day with the highest number of appointment is on the 6th of June, 2016.

Further analysis was done to investigate the workdays of the doctor for appointments as well as figure out the busiest days.

no_show appointment_dayname	Absent	Present	total	Present_ratio	Absent_ratio
Tuesday	5291	20877	26168	0.797806	0.202194
Wednesday	4879	19383	24262	0.798904	0.201096
Monday	4561	18523	23084	0.802417	0.197583
Friday	3887	15028	18915	0.794502	0.205498
Thursday	3700	14373	18073	0.795275	0.204725
Saturday	1	23	24	0.958333	0.041667



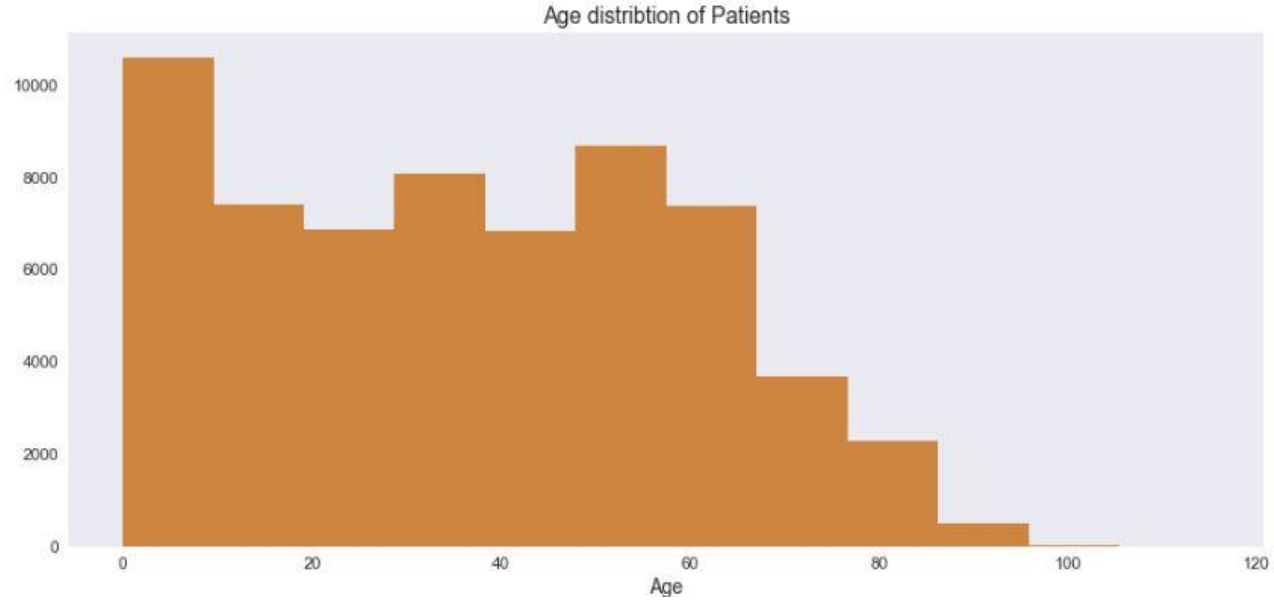
From the chart, we can dedcue that Tuesdays and Wednesdays are the days with the highest appointments. Tuesdays having the most then for weekdays, thursdays have the least appointments on Weekdays. I recommend more appointments be scheduled for thursdays because it also has a low absent ratio.

Looking at the pivot table and chart above, I can deduce that Saturdays are not appointment days but could probably be used in cases of Emergency but this is a theory.

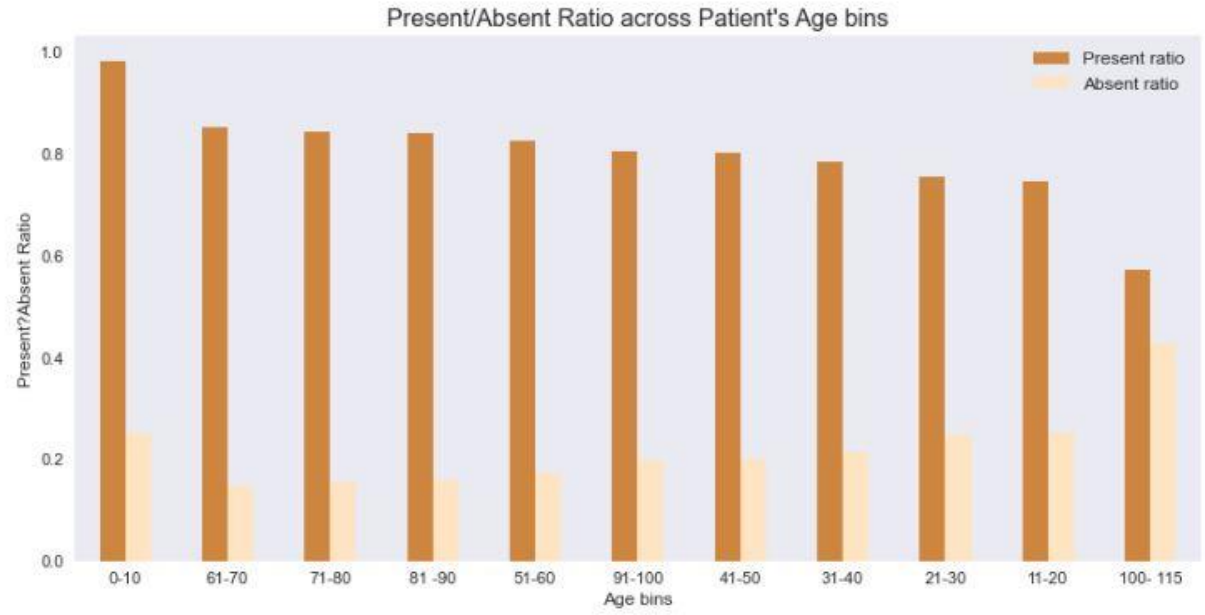
- Research Question 4:

To figure out the age distribution of the patients, I used a histogram to represent the distribution as well as figuring out the minimum age of the patients

```
The minimum age of the patients is 0 years
The maximum age of the patients is 115 years
```



Further analysis was done to find out the attendance rate of the patients within a certain age group



This chart was created using the table below.

Present Ratio can be explained as the number of patients who have been present for an appointment over the total number of appointments

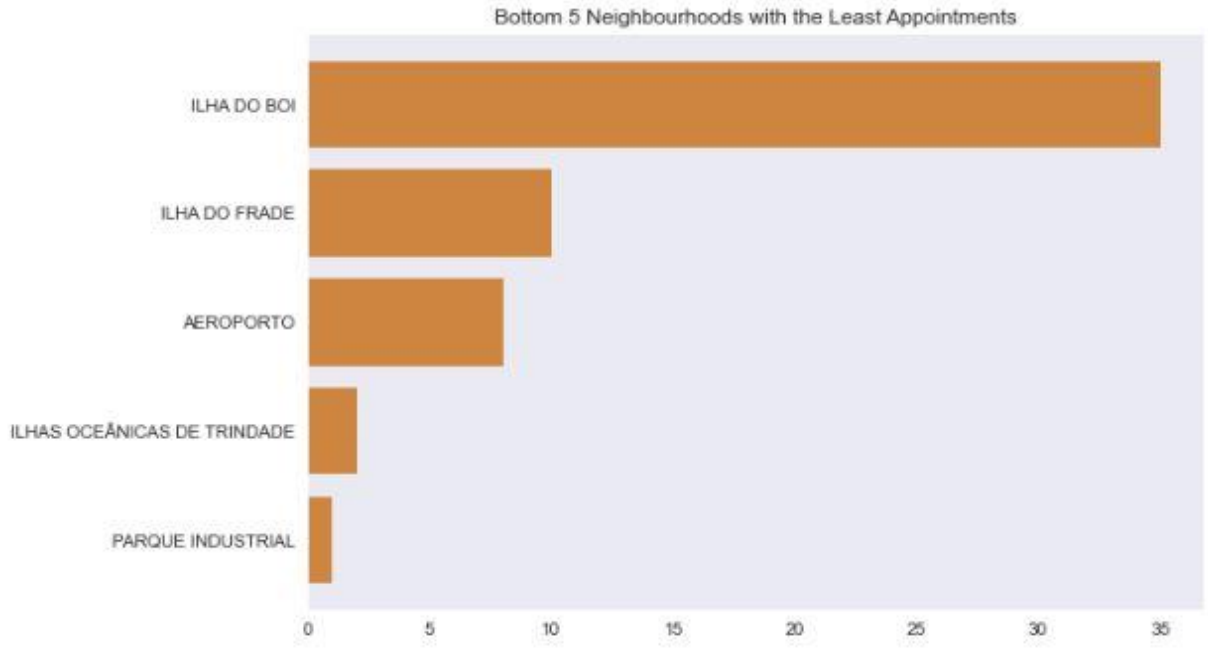
Absent Ratio can be explained as the number of patients who have been absent for an appointment over the total number of appointments

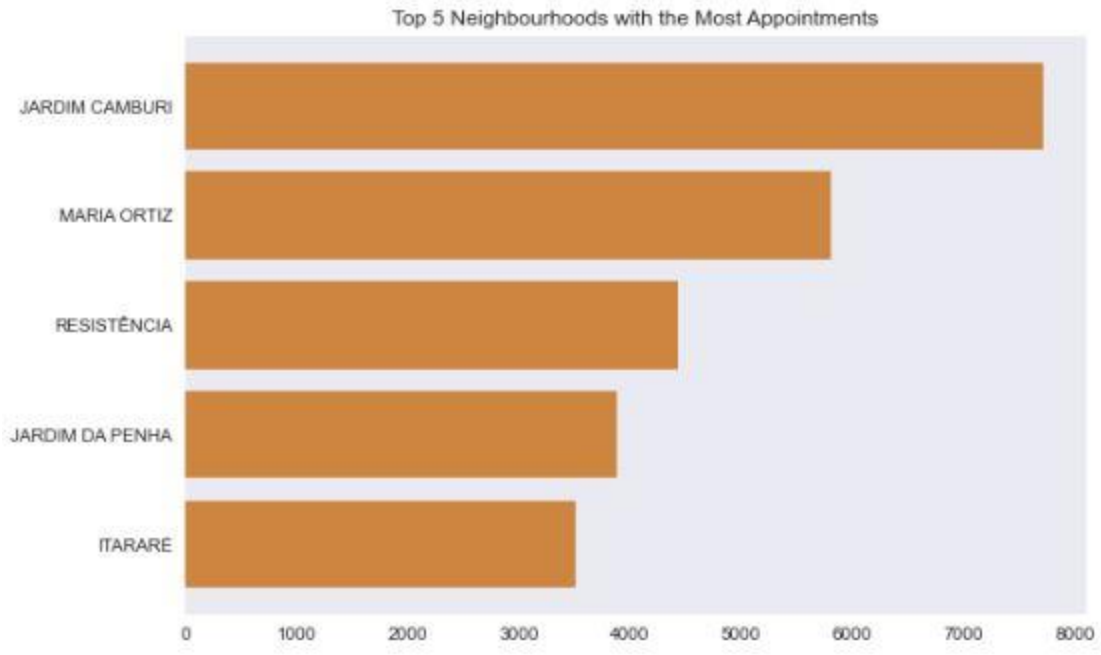
no_show age_bins	Present_ratio	Absent_ratio
0-10	0.983629	0.249047
61-70	0.852712	0.147288
71-80	0.844608	0.155392
81 -90	0.840016	0.159984
51-60	0.826256	0.173744
91-100	0.803987	0.196013
41-50	0.800971	0.199029
31-40	0.785278	0.214722
21-30	0.753755	0.246245
11-20	0.747462	0.252538
100- 115	0.571429	0.428571

Based on attendance rate, patients within the age of 0 - 10 have the most likely chance to show up for appointments.

- Research Question 5:

The Neighborhood column signify the neighborhood where the appointment takes place in Brazil. From my analysis, there are 81 neighborhoods in Brazil which were used as locations for appointments with the doctor



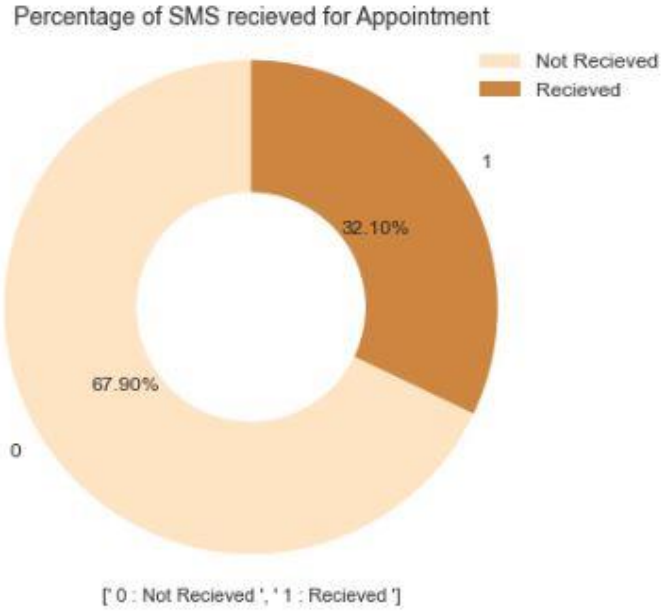


From the analysis above, The neighborhood with the highest appointments are 'ITARARÉ', 'JARDIM DA PENHA', 'RESISTÊNCIA', 'MARIA ORTIZ' and 'JARDIM CAMBURI' while the neighborhood with the least number of appointments are 'PARQUE INDUSTRIAL', 'ILHAS OCEÂNICAS DE TRINDADE', 'AEROPORTO', 'ILHA DO FRADE' and 'ILHA DO BOI'.

- Research Question 6:

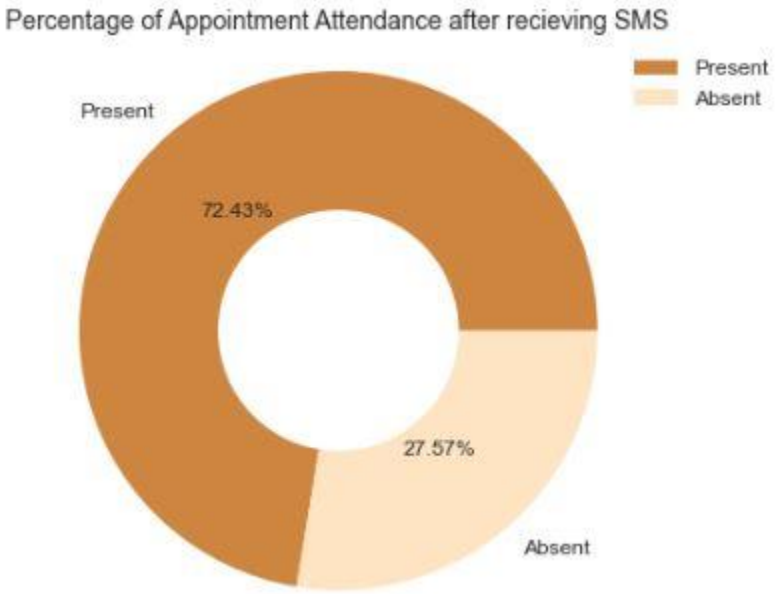
Before I go into that analysis, I checked the proportion of patients who received SMS for their appointments

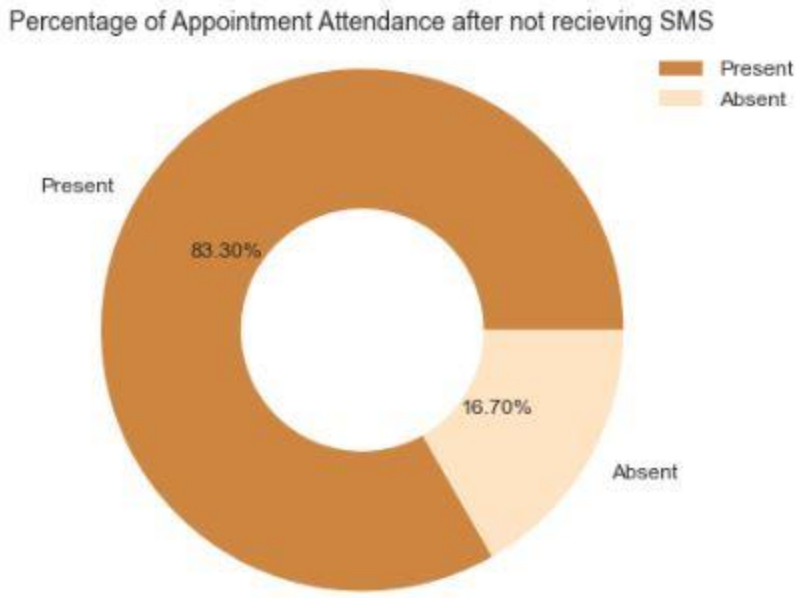
```
0    75044
1    35482
Name: sms_received, dtype: int64
```



This shows that only 32.10% of the appointments received an SMS while the remaining 67.90% did not get an SMS

Further analysis would be done to see if there is a correlation between SMS recieved and rate of Attendance. To do this, I checked the attendance for patients who received SMS to patients who didn't receive SMS





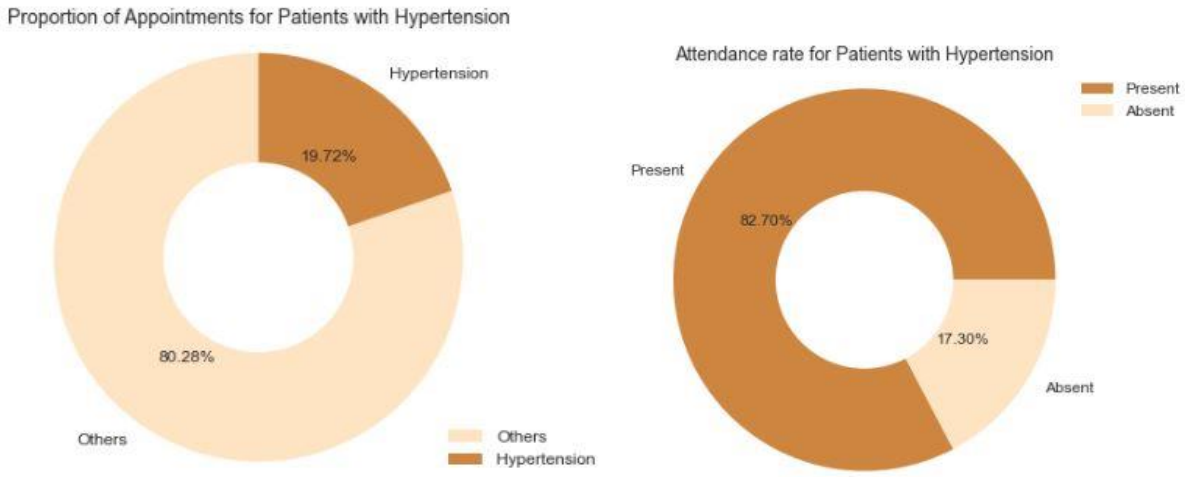
From the chart above, There is a 72.43% attendance after receiving SMS and there is an 83.30% attendance after SMS was not received.

From this analysis above, we can deduce that the SMS received does not influence the Attendance rate.

- Research Question 7:

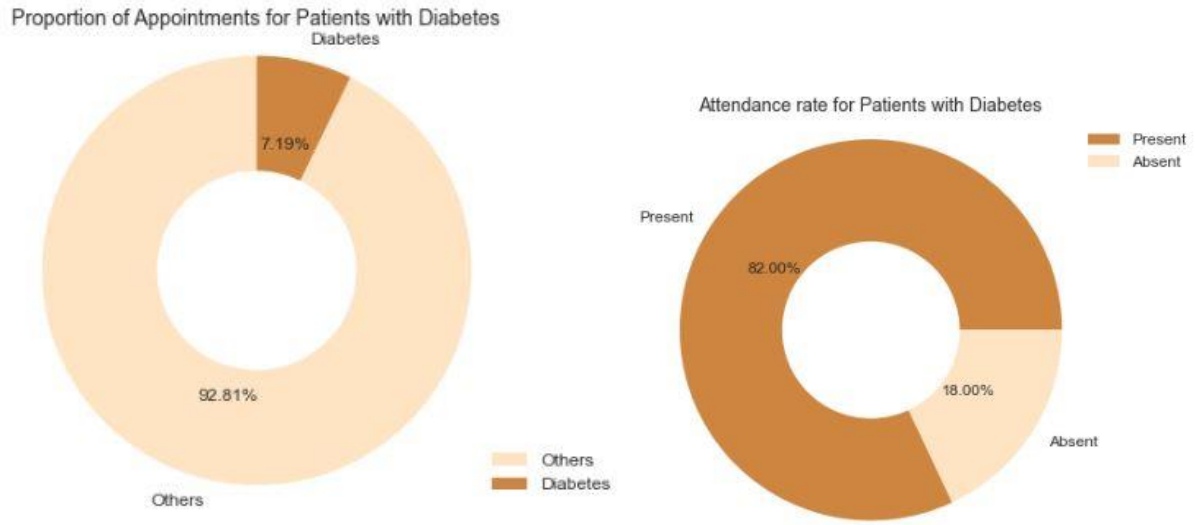
I wanted to investigate the percentage of patients with a certain illness and checking their attendance rate

For Patients with Hypertension:



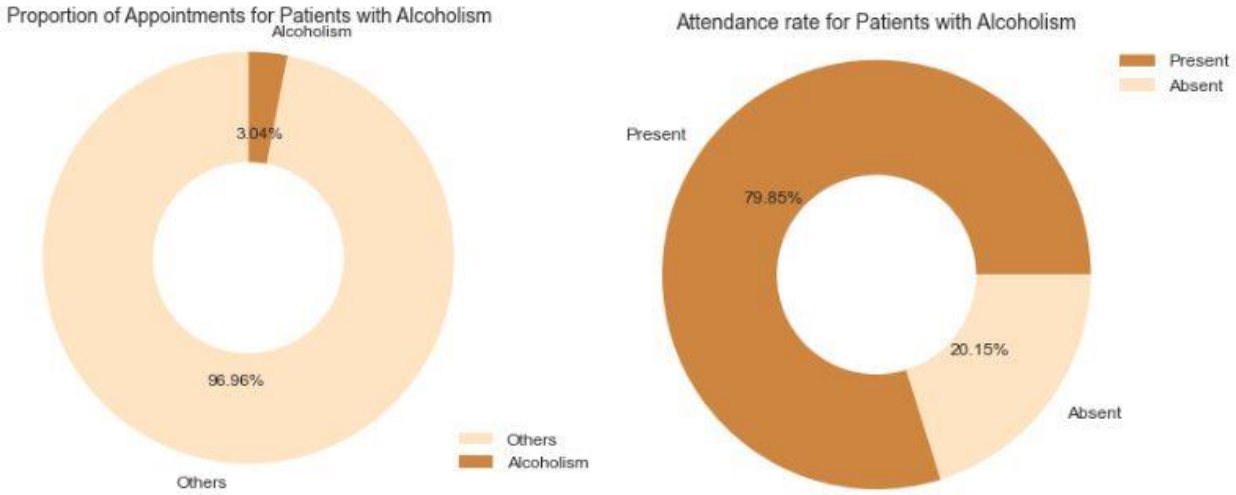
From the chart above, 19.72% of the scheduled appointments are for patients with Hypertension and they have an attendance rate of 82.70%.

For Patients with Diabetes:



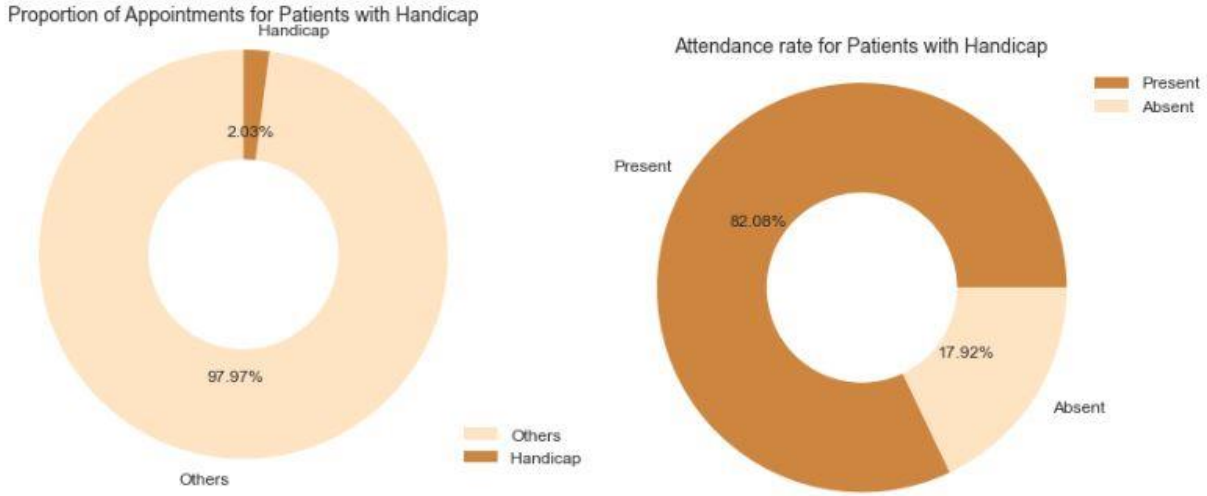
From the chart above, 7.19% of the scheduled appointments are for patients with Diabetes and they have an attendance rate of 82.00%.

For Patients with Alcoholism:



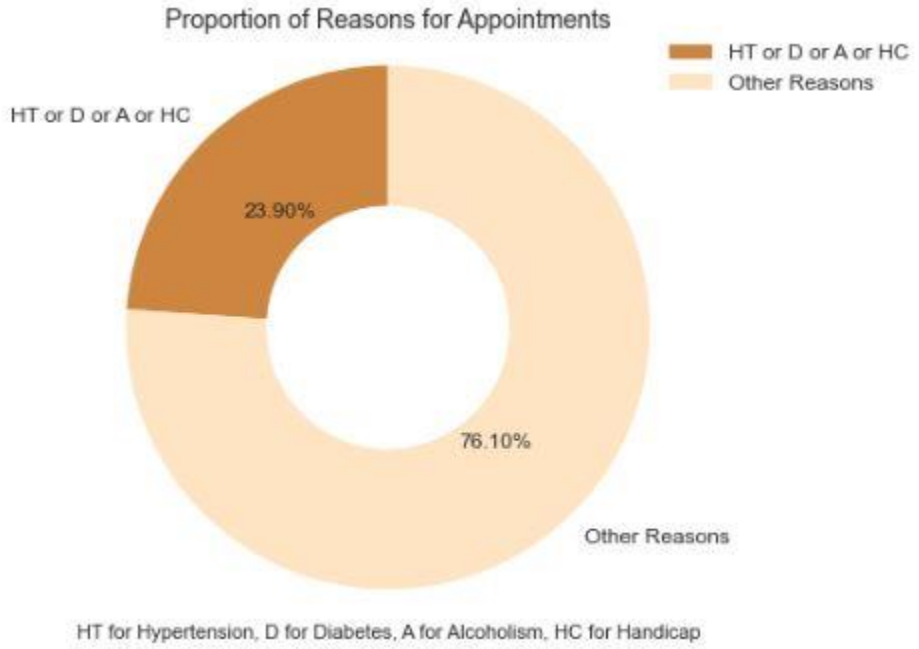
From the chart above, 7.19% of the scheduled appointments are for patients with Alcoholism and they have an attendance rate of 79.85%.

For Patients with Handicap:



From the chart above, 2.03% of the scheduled appointments are for patients with Handicap and they have an attendance rate of 82.08%.

For patients who neither have Hypertension, Diabetes, Alcoholism nor Handicap:



- Research Question 8:

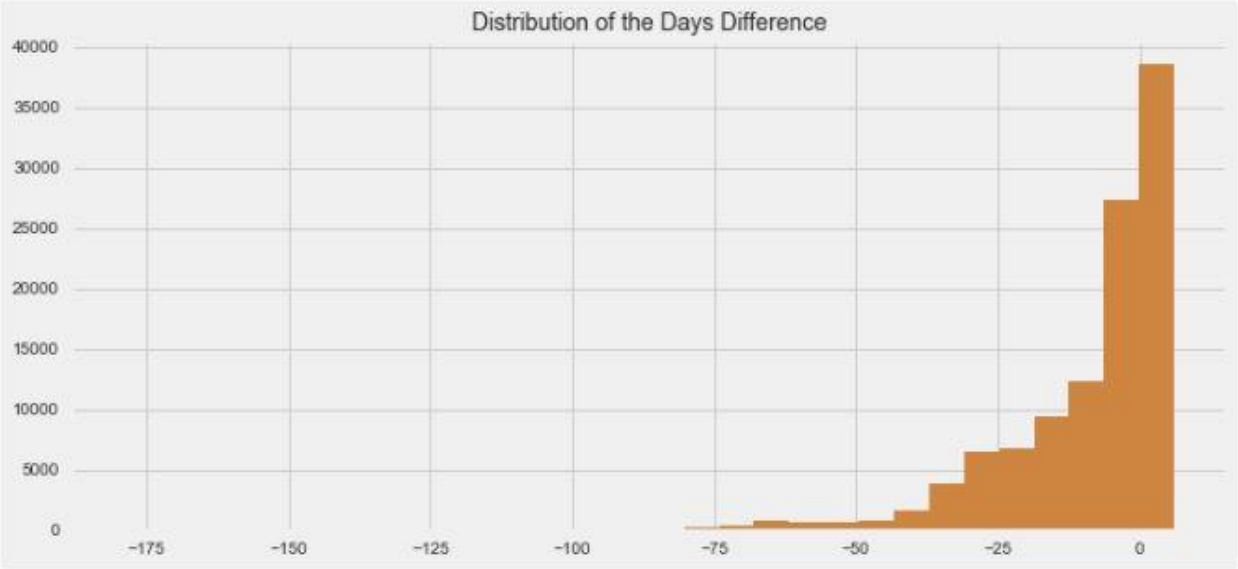
From the documentation,

- Scheduled Day: The day of the actuall appointment, when they have to visit the doctor.
- Appointment Day: The day someone called or registered the appointment, this is before appointment of course

To get the number of days before scheduled day, A difference would be taken between Scheduled days and Appointment days using

number of appointment days before scheduled days = scheduled day - appointment day

After the day difference was calculated, the distribution was plotted on Histogram



The summary statistics of the column was taken

```
count    110526.000000
mean        -10.183794
std          15.255034
min        -179.000000
25%         -15.000000
50%          -4.000000
75%           0.000000
max           6.000000
Name: day_diff, dtype: float64
```

Looking at the chart and the distribution of the values of the day difference, there are a lot of negative values. This is one of the limitations of the dataset. This means that the scheduled day for the appointment had been done then an appointment date was registered after the scheduled day.

Conclusion

There are 110,527 appointments while there are 62,299 patients. There are most likely patients with more than one appointments.

The attendance rate of the patients are relatively high. On checking within the age distribution, ages 0 – 10 years have the highest attendance rate.

I see no relationship between the SMS received and the attendance rate of the patient. More details in the report

Over 70% of the patients do not have their illness accounted for

Appointment days to see the doctor are majorly from Mondays to Fridays. Tuesdays having the most appointments so far. Saturdays were seldomly used. I presume that those are emergency situations. No appointment was done on a Sunday.

The neighborhood with the highest appointments are 'itararé', 'Jardim Da Penha', 'Resistência', 'Maria Ortiz' and 'Jardim Camburi' while the neighborhood with the least number of appointments are 'Parque Industrial', 'Ilhas Oceânicas De Trindade', 'Aeroporto', 'Ilha Do Frade' and 'Ilha Do Boi'

On 6th of June, 2016, the highest number of appointment was recorded

Limitations

Misconception with the no-show columns. Rumor has it that No meant the patient did not attend appointment while yes meant the patient attended the appointment but the documentation stated otherwise.

The scheduled day is meant to come after the appointment day but the day difference were giving negative values which is not supposed to be so

The dataset appeared to have a negative age which cannot be so

76.10% of the patients do not have any of the stated illness, but must have scheduled an appointment for reasons not stated in the data. This makes the data incomplete