No-show medical appointments analysis

Dataset used for analysis:

No-show-medical appointments dataset is used in this project for analysis. This dataset under investigation is a collection of information from 100k medical appointments scheduled in Brazil. The main objective of this data analysis is to find whether patients who scheduled appointment actually showed up or not and if there are more number of no-shows, what are the factors affecting them.

With a given characteristics like gender, age, scholarship and health profile of every patient such as hypertension, alcoholism, diabetes, handicap, below are the questions that were investigated to find what are the characteristics associated with no-shows.

Questions posed:

- Which gender type has the highest number of no-shows?
- What is the patient age distribution of no-shows versus shows?
- What are the trends associated with a patient health profile and shows versus no-shows?
- Which neighborhoods have the highest number of no-shows?
- Do no-shows have a larger time gap between scheduled date and appointment date?
- Are people receiving SMS are more likely to show up?

Data Wrangling in the project:

Data wrangling is mainly done to clean and unify the messy and complex data sets for easier access and analysis. Basically, data wrangling process contains three main activities - data acquisition, data cleaning and data joining.

- The data is acquired from the source in data acquisition step. In this project, no-show medical appointment is loaded from Kaggle.

- Data cleaning process includes dropping off the duplicates, cleaning the erroneous data, renaming or correcting the column names, changing the data types of columns if required and finally removing the outliers.
    - In the dataset investigated, no duplicates were found so no rows had been dropped off.
    - The column names were renamed so maintain the uniformity in them
    - Data type of patient id was changed string type to int and scheduled & appointment days were trimmed to date value and changed to datetime object from string
    - Unique values of every column were calculated to remove the erroneous data
    - Finally outliers of every column are removed
- Data joining process was not required as per the data given

How is the data investigated?

The data investigation is followed by the process of data wrangling. It is considered for every dataset to bring uniformity in data and easiness to the analysis process. Once the data is cleaned and trimmed, it is ready for analysis.

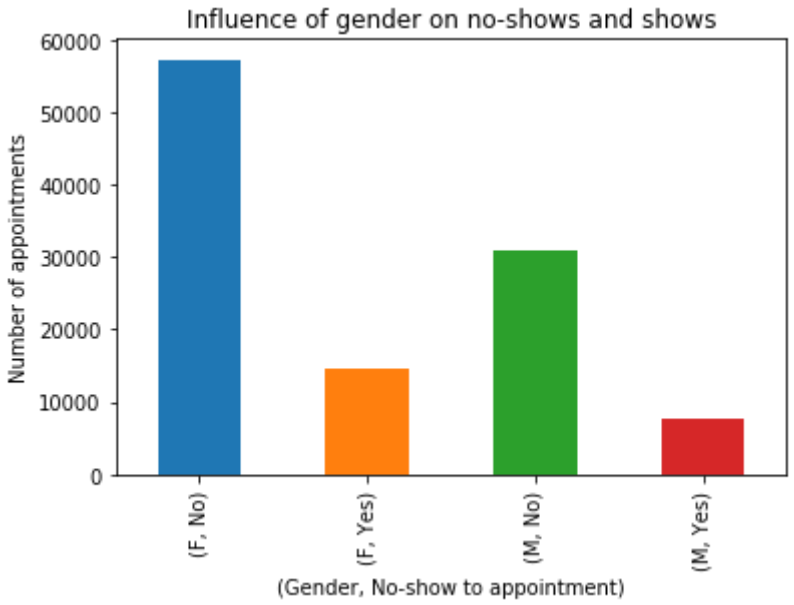
In this project, one question at a time is taken with it's related variables filtered out for analysis. With the given data, the proportion of shows and no-shows were calculated for each question using Pandas and Numpy libraries. The statistics of each question were computed and visualizations were shown using Matplotlib to communicate the results in the best way possible.

Summary:

The main question was What are the most important factors in determining the likelihood of a no-show? The most important factors were:

The major factors would be the neighbourhood they live and the time gap between scheduled day and appointment day. However, there are some other small factors that are affecting too.

Question 1: Which gender type has the highest number of no-shows?



Conclusion: From the visualization it is pretty much clear that this distribution is skewed towards women. Though women are likely to show up to their appointment, the no-shows percent of both male and women doesn't vary too much. Apparently, gender has no influence on proportion of no-shows

Question 2: What is the patient age distribution of no-shows versus shows?

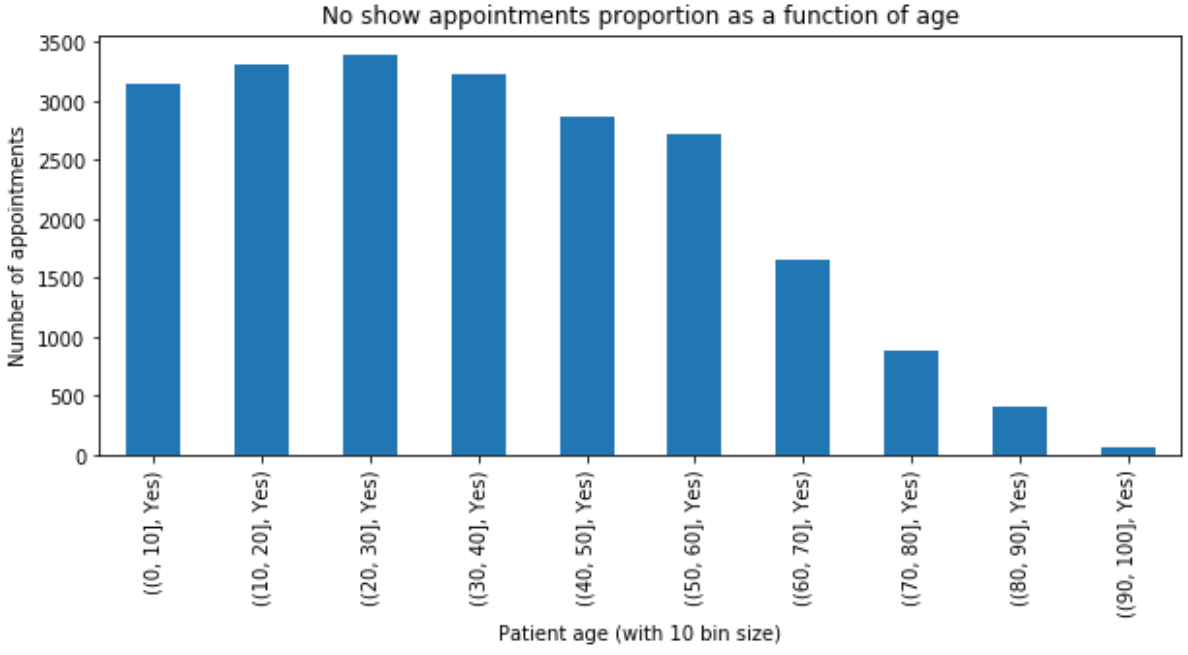


Fig: No-show appointments stats

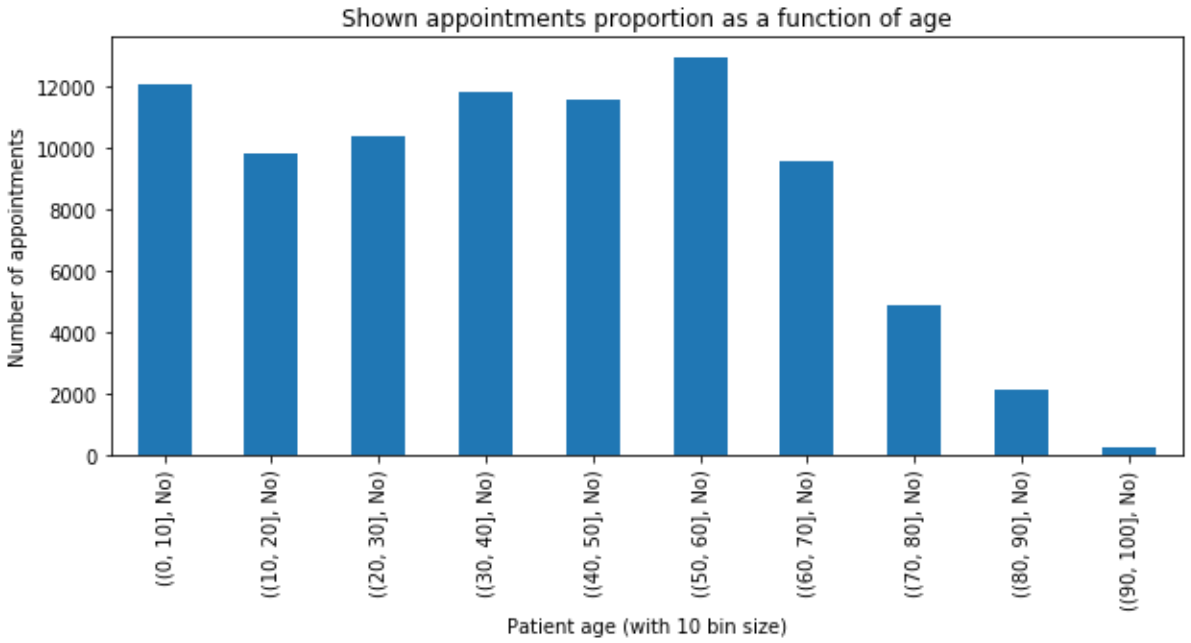


Fig: Shown appointments stats

Conclusion: The number of no-shows were increasing till the age bin - (20,30] and then they had started decreasing. This doesn't clearly depicts how the age was influencing the no-shows proportion. Similarly with the shown appointments, there doesn't seem to be any trends associated.

Question 3: What are the trends associated with a patient health profile and shows versus no-shows?

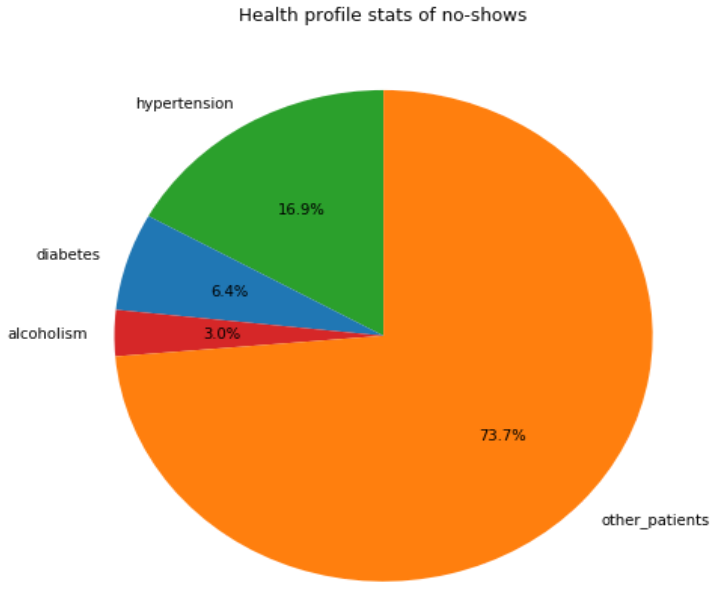


Fig:Health profile stats of no-shows

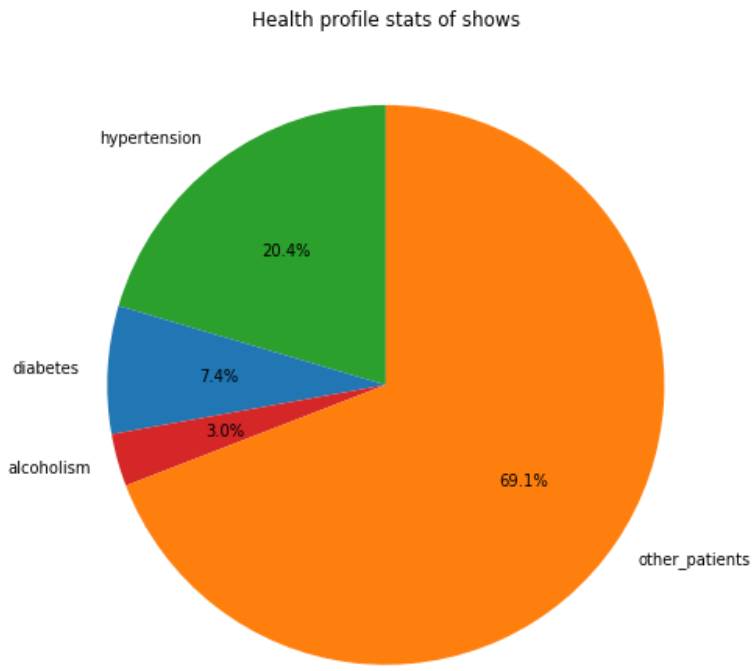


Fig: Health profile stats of shows

Conclusion: The health profile of patients doesn't seem to have much influence on showing up to appointment or not. Patients with hypertension have 16.9% no-show rate and 20% show rate which is not a significant difference. Similarly with alcoholism and diabetes, their proportions doesn't seem to vary significantly.

But when the combined stats are analysed, it seems like patients with hypertension are more likely to not show up than diabetic patients and diabetic patients are more likely to not show up than alcoholic patients.

Question 4: Which neighbourhoods have the highest number of no-shows?

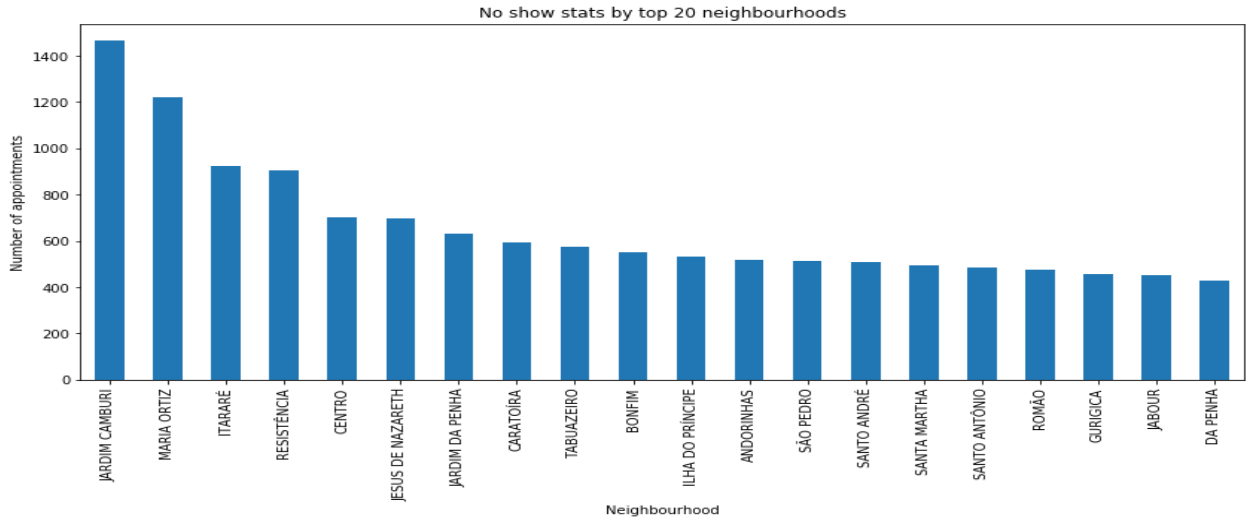


Fig: Top 20 neighbourhoods of no-shows

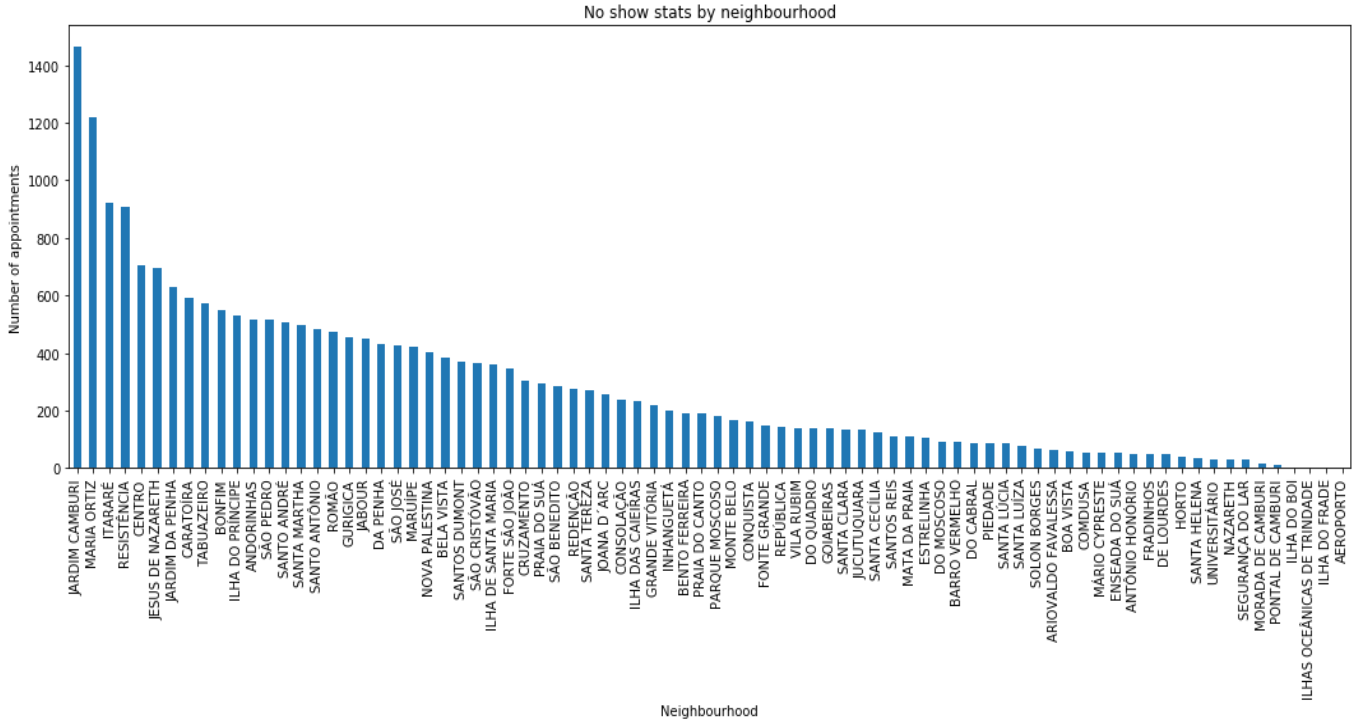


Fig: Neighbourhood stats of no-shows

Conclusion: The graph clearly depicts that few neighbourhoods are more likely to escape from their appointments than others. Obviously, this counts as a major factor that influence no-shows

Question 5: Do no-shows have a larger time gap between scheduled date and appointment date?

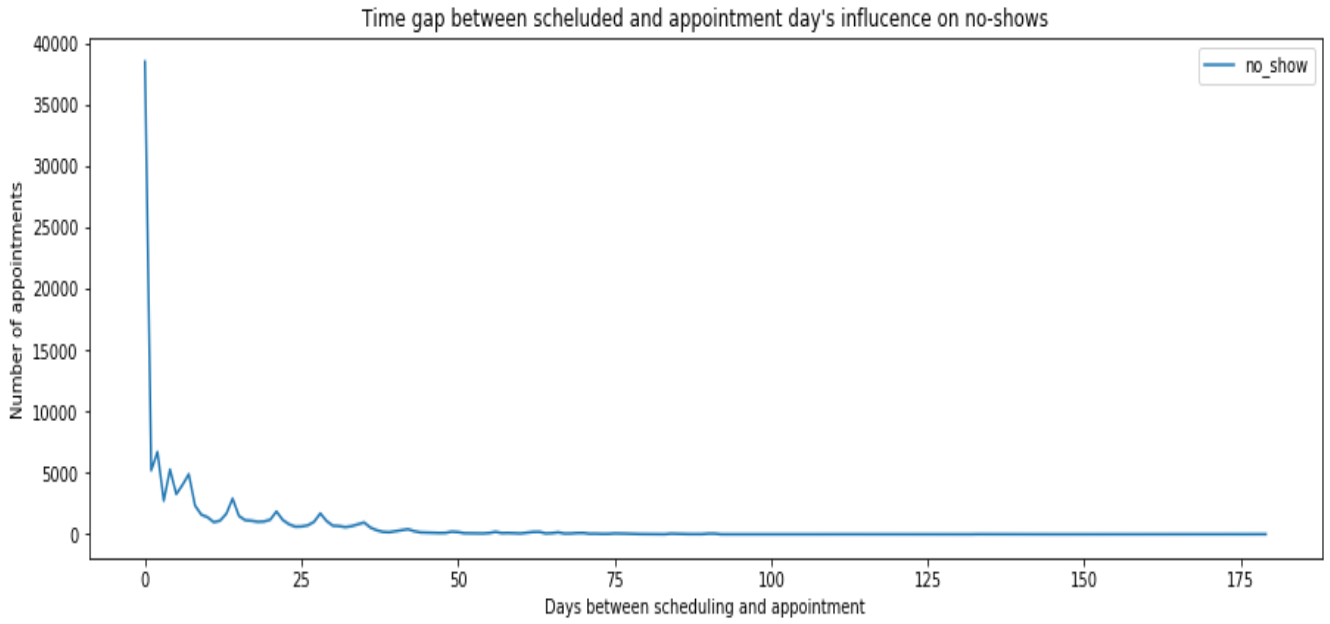


Fig: Difference in days between scheduled and appointment days trends

Conclusion: It seems that patients are a lot more likely to show up when the appointment is scheduled on the same day. However, other than for the same day appointments, the graph seems to be going down with few distractions in between them

Question 6: Are people receiving SMS are more likely to show up?

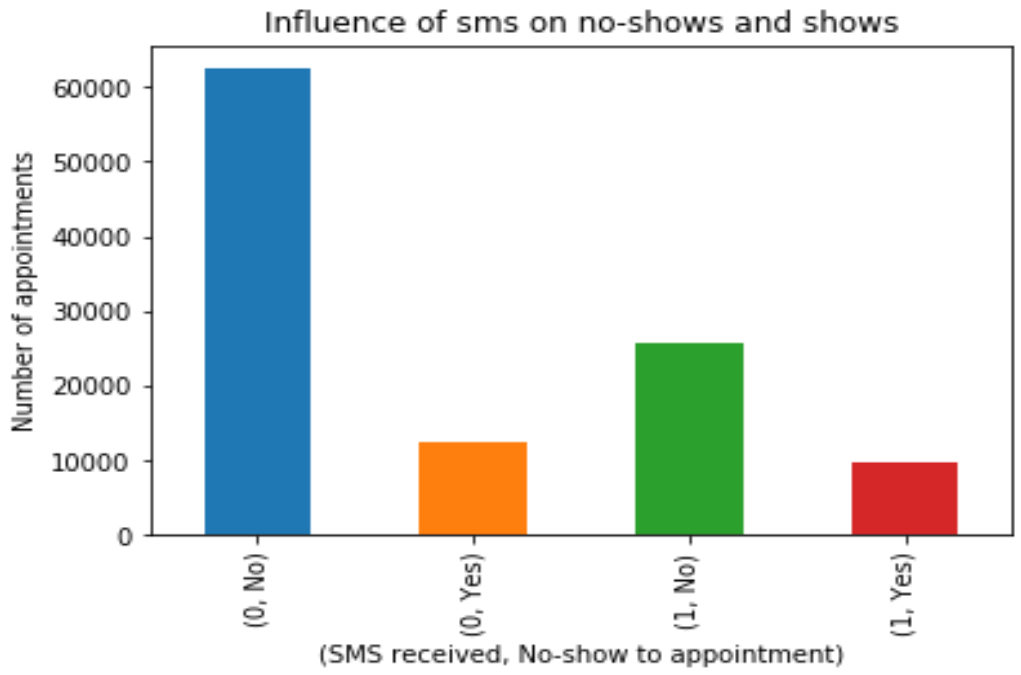


Fig: SMS influence on no-shows vs shows

Conclusion: Apparently, SMS is not a major factor that is influencing the percentage of no-shows because patients without receiving SMS showed up to the appointment in a great proportion. Sending SMS doesn't really help patients to show up. Therefore, SMS hardly has an effect on no-shows.