Investigating the [Medical Appointment No Shows]

(original source on Kaggle)

Summary of data

Based on my choice, in this project I'll analyze the Medical Appointment No Shows dataset, it's a dataset who holds informations about 110.527 medical appointments of different patients from different neighborhoods in Brazil.

This dataset is built for the purpose of finding out the reasons that made patients not showed up at their scheduled appointment. Thus, it contains one dependent variable called No-Show (which take the 'No' cardinality when the patient shows up at the scheduled appointment and 'Yes' when he don't show up), and to understand the reasons for this behavior, 13 other variables were collected in parallel with the main information, which is the showing up or not at each medical appointment.

Questions

Main question: What are the factors that affect the patient's attendance?

Intermediate questions:

- *What is the overall rate of patients for both (show up / not show up) at the medical appointment?*
- *How many patients are female and are male, and what's the proportion of no-showing up at the appointment for each Gender?*
- *What are the characteristics of patients in terms of age?*
- *How many patients and what's the no-showing rate for each age category?*
- *How many patients have (a scholarship / don't have it), and what's the proportion of no-showing up at the appointment for each situation?*
- *How many patients have(hypertension / don't have it), and what's the proportion of no-showing up at the appointment for each situation?*
- *How many patients are (diabetics / not diabetics), and what's the proportion of no-showing up at the appointment for each situation?*

Describe what you did to investigate these questions

I did a lot to answer these questions. I evaluated the data visually and programmatically, identified all issues related to hygiene, then identified one variable (appointment) and based on it in asking my questions, then compared all the variables that have a relationship to the question that I presented with *the variable that I selected from the beginning, then I set out to explore The data and answer its questions in an illustrated manner using the graph*

Data Wrangling

In this section I'll proceed through three stages:

- ***Load packages & gathering data***
- ***Assessing data***

Cleaning data

In this section I'll go through the steps below:

- ***Renaming to relevant column names***
- ***Converting variables to the correct data types***
- ***Triming data regarding the variables needed in the research questions***

EDA - Exploratory Data Analysis

In this section: I'll explore the data in order to address may research questions.

Finding patterns.

Visualizing relationships.

Building intuition.

Summary statistics and plots communicating your final results

After analyzing these 6 variables, we reach the results below:

- ***The overall no-showing rate is 20.2%, less than the overall show-up rate (79.8%).***
- ***The percentage of females is greater than males _ 65% > 35%.***
- ***From the age 0 to 17, the number of females is almost equal to the number of males,***
- ***From the age 18 to 115, the number of females is almost twice the number of males.***
- ***There is no significant effect of gender on no-showing rate (20% for each gender).***
- ***The most frequent age categories are [3 - 17]) and ([18 - 39] which they mark the highest rates of no-showing 23.5% for both.***

- *There is an effect of age on the patients no-showing rate.*
- *The percentage of having a scholarship is much less than not having a scholarship _ 9.8% < 90.2%.*
- *The percentage of having an hypertension is much less than not having an hypertension _ 19.7% < 80.3%.*
- *The percentage of the diabetic patients is much less than the non-diabetic patients _ 7.2% < 92.8% are s*
- *The data provided is not sufficient to confirm the answer to the research question, and this study remains descriptive because it is limited to only 6 variables, so we need an inference statistic or a machine learning model.*
- *The volume of sample is not good for making a good judgment about research questions*

Conclusions

As mentioned in the introduction, this study remains descriptive and limited to only 4 variables, and in no way provides real and precise explanations for The reasons why patients do not showed up for their scheduled medical appointments because there are any inferential statistics included, however it indicates some factors that may cause this behavior.

References

- https://pandas.pydata.org/
- https://numpy.org/
- https://matplotlib.org/
- https://seaborn.pydata.org/
- https://www.kaggle.com/
- https://classroom.udacity.com/nanodegrees/
- https://stackoverflow.com/
- https://stackexchange.com/
- https://daringfireball.net/projects/markdown/syntax#list
- https://en.wikipedia.org/wiki/Bolsa_Fam%C3%ADlia