SJSU INFO 208 — Big Data Technologies  
Mini Project

Analysis on Medical Appointment No-Show data

Amar Ippili

013728975

# **Overview**

The demand for utilizing the public health resources is very high as private health institutions cost a ton. In developing countries like Brazil, where there are many people who belong to lower middle class and poverty class tend to rely solely on public health institutions. As there is a lot of demand and less public health resources it takes lot of time for a patient to get a medical appointment scheduled. Once scheduled, if a person doesn’t show up for an appointment it is a waste of resourceful time for the public health officials, waste of money and a loss of opportunity for someone else who is in dire need of it. Such is the cost of a single appointment in the public health sector in the developing countries.

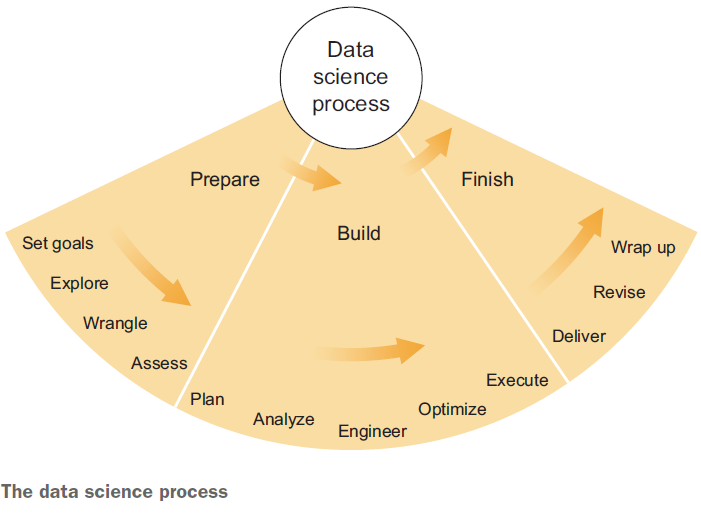
In this mini project I will be doing analysis on a public health institutions medical appointments dataset of more than 110k medical appointments. Here more than 30% of the appointments were a “No-Show”. The public health institutions can send multiple SMS notifications to the patients regarding their scheduled appointments in order to decrease the “No-Shows” percentage. But they wouldn’t want to waste their resources by sending the notifications to everyone, instead want to notify the patients who are most likely to be a “No-Show”. We must find a way to find the risk of a patient not showing up for a medical appointment.

# **My Interests on this Dataset**

I have a longtime interest in health policies and people's experiences with medical institutions, so this dataset naturally interests me. I chose this specific dataset as I want to find how different factors or attributes of a patient can be correlated to the final goal (patient showing up for appointment). My background as an NGO worker who has worked for different social welfare programs gives me an interest in finding insights for this dataset. Also, it was also equally interesting about studying the background of the different attributes in this dataset.

For this purpose I want to follow the Data Science process for the respective dataset to find some insights that can be helpful in dealing with the problem.

According to Brian Godsey’s “Think like a Data Scientist” the Data Science process can be divided in three phases: Prepare, Build and Finish.



# **What are you hoping to achieve?**

Firstly, I need to set a few goals to give me some direction for this project.

Which gender seeks more medical attention?

Which age group seeks more medical attention?

To find the different factors that are having a lot of impact on patients not showing up on the Appointment Day?

Will I be able to predict whether a patient is more likely to show up for an Appointment or not?

# **Dataset Studied**

The dataset I chose is a Medical Appointment data from public hospitals in Brazil. It has roughly around 110k records. It is about 2MB in size. This dataset is a list of medical appointments scheduled in the time period of November 2015 to June 2016.

## **Source (URL or other sourcing location)**

I found the data set in Kaggle.com. You can find it in the below URL.

<https://www.kaggle.com/joniarroba/noshowappointments/home>

# **Description of & Structure of the Data**

This dataset is in CSV format with 110k records and each having 14 different attributes.

* **PatientId** – This is the unique id of a patient.
* **AppointmentID** – This is a unique id given for each appointment made.
* Gender – This attribute defines whether the patient is **male** (**M**) or **female** (**F**).
* **ScheduledDay** – This attribute defines the date on which the appointment was done.
* **AppointmentDay** – This attribute defines the date on which appointment is scheduled and patient needs to show up in the hospital.
* **Age** – This attribute defines the age of the patient.
* **Neighborhood** – This attribute defines the place where appointment takes place.

For the below attribute there are only two possible values. **‘1’** means **True** and **‘0’** means **False**.

* **Scholarship** – This attribute defines whether the patient is eligible for the scholarship program.
* **Hypertension** – This attribute defines whether the patient is suffering from Hypertension.
* **Diabetes** - This attribute defines whether the patient is suffering from Diabetes.
* **Alcoholism** - This attribute defines whether the patient is suffering from Alcoholism.
* **Handicap** - This attribute defines whether the patient is having any disabilities. According to the data I find that a patient can have 0 to 4 disabilities.
* **SMS\_recieved** - This attribute defines whether 1 or more SMS notifications about the appointment are sent or not.
* **No-show** - This attribute defines whether the patient showed up for the appointment or not. If ‘Yes’ then the patient didn’t show up and if ‘No’ then the patient showed up.

The **Scholarship progr**am is a **social welfare program** in **Brazil** called **Bolsa Família**. Bolsa Família provides financial aid to poor Brazilian families; and if they have children, families must ensure that the children attend school and are [vaccinated](https://en.wikipedia.org/wiki/Vaccination). If they exceed the total of permitted school absences, they are dropped from the program and their funds are suspended. The program attempts to both reduce short-term [poverty](https://en.wikipedia.org/wiki/Poverty) by direct [cash transfers](https://en.wikipedia.org/wiki/Welfare_benefit) and fight long-term poverty by increasing [human capital](https://en.wikipedia.org/wiki/Human_capital) among the [poor](https://en.wikipedia.org/wiki/Poverty) through [conditional cash transfers](https://en.wikipedia.org/wiki/Conditional_Cash_Transfer).

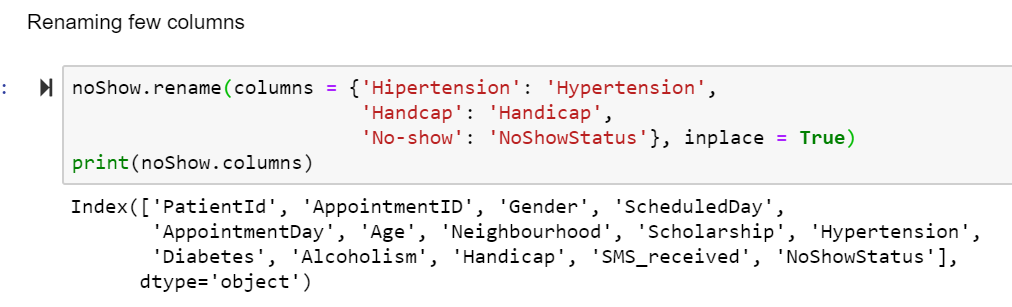
Bolsa Família currently gives families with [per-capita monthly income](https://en.wikipedia.org/wiki/Per-capita_income) below $140 BRL (poverty line, ~$56 USD) a monthly stipend of $32 BRL (~$13 USD) per vaccinated child (< 16 years old) attending school (up to 5), and $38 BRL (~$15 USD) per youth (16 or 17 years old) attending school (up to 2). Furthermore, to families whose per-capita monthly income below $70 BRL (extreme poverty line, ~$28 USD), the program gives the Basic Benefit $70 BRL per month.

So, if the data value in Scholarship section is ‘1’ then the patient is eligible for the Bolsa Família program. If the value is ‘0’ the patient is not eligible for the Bolsa Família program anymore.

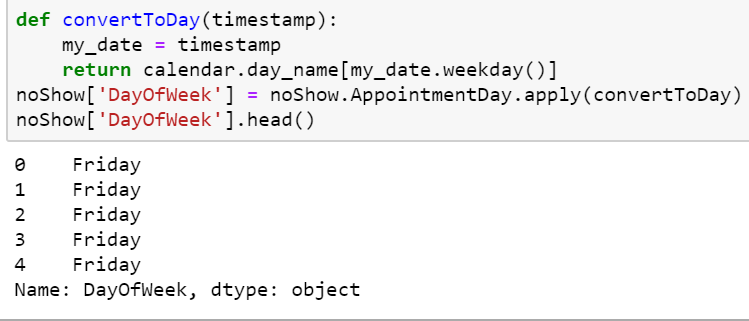
# **Wrangling of Dataset**

Before starting any exploratory analysis on the dataset, I need to do some Data wrangling, sometimes referred to as data munging which is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics. Here as the data is already in CSV format and is around 110k records, it can be easily handled by python pandas library. So I don’t need to change any format though I need to clean the data.

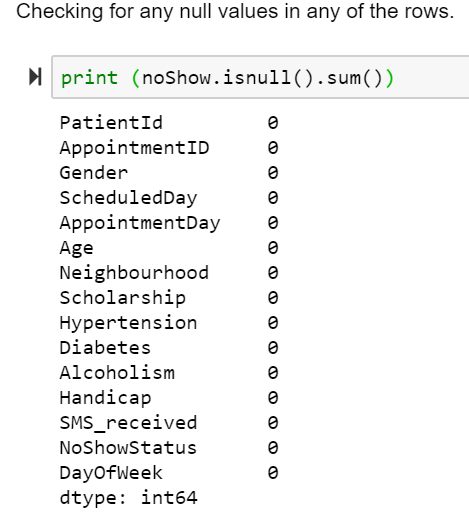
On some exploration of the dataset I found that the column names translated from Spanish language to English. I have corrected those column names for easy understanding.

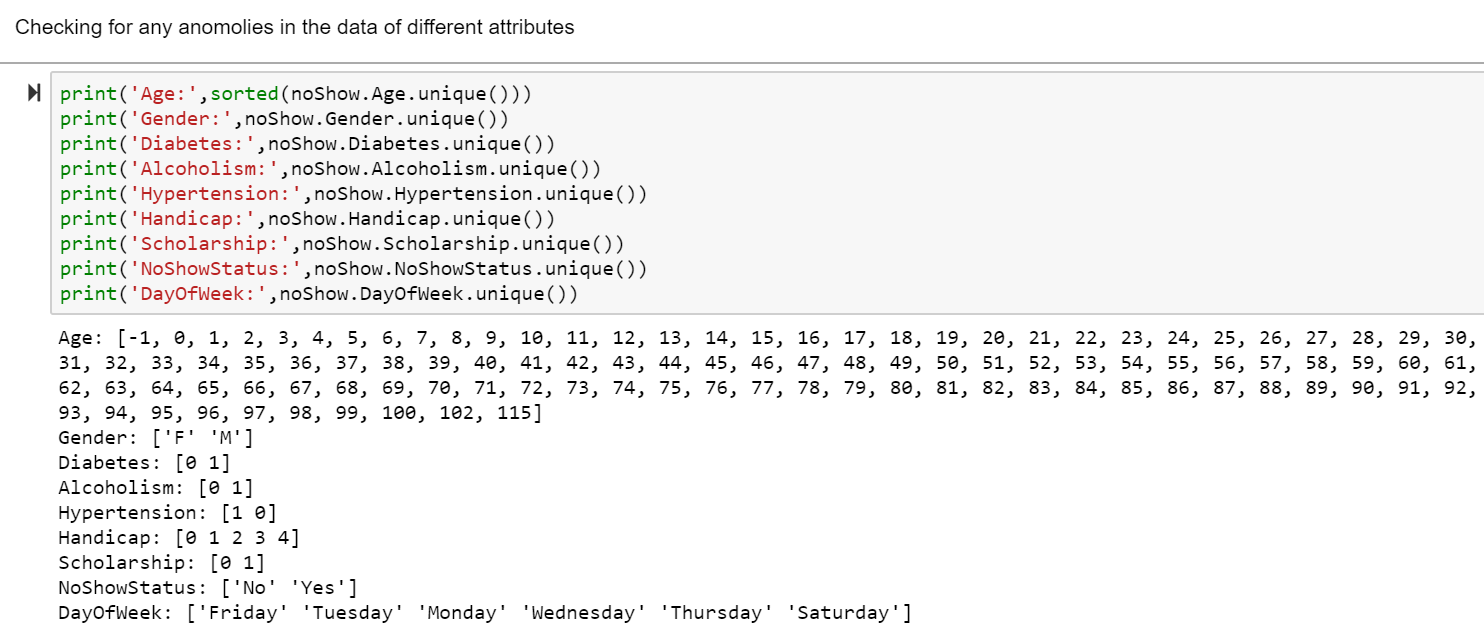


Also, for the convenience of operations on datetime I have converted the ScheduledDay and AppointmentDay columns into datetime64 format.

Created a new column call “Day of the Week” which can be verified as a factor of influence for the “No-Show” of the appointment.

Null value check and a data anomalies check on the dataframe.

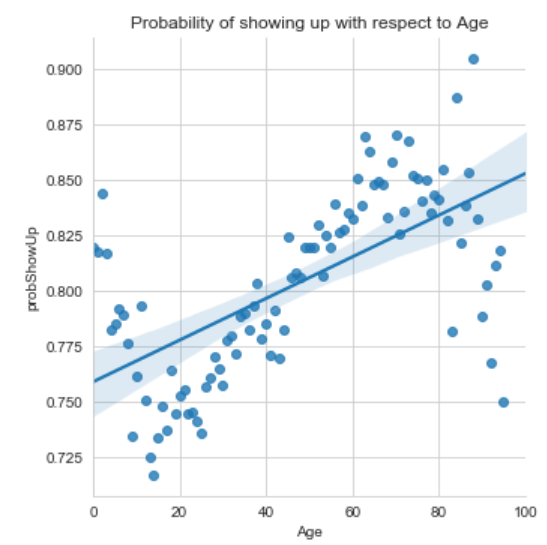


This means there are no null values in any of the columns.

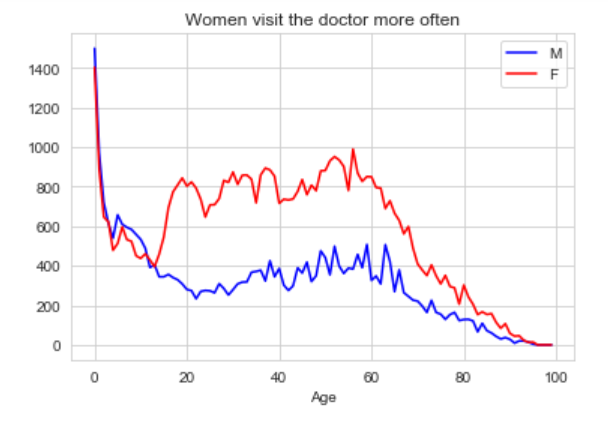
We do have some impossible ages such as -1, and some pretty absurd ages such as 100 and beyond. I do admit that it is possible to live 113 years and celebrate living so long, and some people do live that long, but most people don't. So, I will treat the ages greater than 95 as outliers and remove the absurd data from the data frame.

# **Questions to Ask of the Dataset**

**Which age group seeks more medical attention?**



Though there are few outliers, you can observe the trend. As age increases the probability of a person showing up for an appointment increases. It is a linear dependency.

**Which gender seeks more medical attention?**

Women are twice as likely to visit the doctor. But this effect is true only from a certain age.

Until the age of 14, boys and girls attend the doctor similarly. Starting from around that age, however, there are more than 2.5 woman visits for every man visit.

Possible explanations:

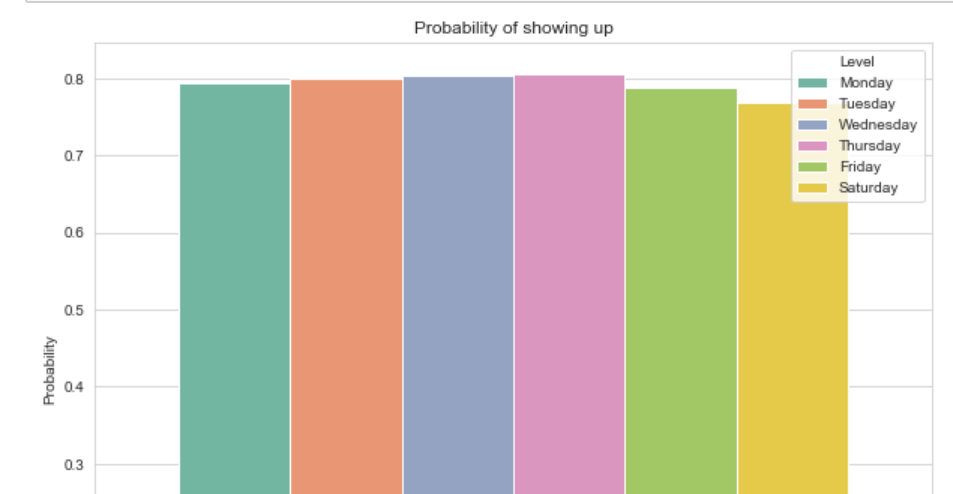
1) Since the difference starts when approaching puberty, it might be due the different care that males and females seek regarding their reproductive system.

2) Men and women might suffer from different medical conditions that would require a different visiting frequency.

3) A difference in lifestyle that can allow women more visits

**To find the different factors that are having a lot of impact on patients not showing up on the Appointment Day?**

Apart from age and Gender, I was not able to find a tangible factor that is having a lot of impact on the patients showing up on Appointment Day. But day of the week can be considered as a factor.



As you can observe from the above plot that the probability of patient showing up for an appointment increases till mid-week and decreases towards weekends. It might imply that people have many plans on the weekends and that might be a reason for the decline in the probability of showing up for an appointment.

**Will I be able to predict whether a patient is more likely to show up for an Appointment or not?**

I guess with the application of a proper mathematical prediction model for the given data we might be able to predict whether a patient is more likely to show up for an Appointment or not. But for the current data as I was not able to deduce many factors that lead to the final result apart from age and gender, the accuracy of the prediction will be less.

# **Summary**

From the acquired dataset I got to learn a lot. While studying the dataset and its attributes I learnt a lot about the Brazilian social welfare programs that are focused to break the transgressional cycle of the poverty. As the data is from the public hospitals of a developing country, the insights obtained from this dataset may not be applicable to other countries as the sample data is biased to the people of middle and poverty class. (An assumption that the rich people in developing countries tend to visit private hospitals than public hospitals). I noticed that 70% of patients were women. This made me curious and I have started analyzing if gender plays a vital role in deciding whether a person shows up for an appointment or not. But I was not able to find much difference with the regard, but I was able to find that women from the age of 14 tend to visit the hospital 2.5 more times than that of men of same age group. Also, I have observed that as the age increases the probability of patients showing up for an appointment increase. Though I was able to deduce so much, due to my lack of knowledge in implementing the right statistical model stopped me from creating a model to predict the likelihood a patient showing up for a medical appointment.

Through this Mini-Project I learnt a lot of stuff about the Data Science process. I always used to think that the important phase of data science project is the “Hypothesis & Modelling” phase or “Build” phase according to “Think like a Data Scientist” book. But I came to know that in real time environment the acquiring the appropriate data is a very difficult task. Even if we acquired the data, it won’t be like the tailored datasets that we find during our coursework. The data obtained may be of different formats like JSON, XML, CSV, database etc. We must be acquainted with all the formats. Also learnt that first we need to explore the data before we proceed with the analysis of it. Finding the null values or outliers, refining the data types are very important. Our initial investment of time in this face will help us to save a lot of time in the next phases of the project. According to a survey, in a data science project 60% of the time and efforts are spent in the acquisition, exploration and wrangling of the data. We must have goals set during the start of a project as that will provide us the direction for exploration and building our project. I felt that given more data or I was able to deduce more attributes from the given data and with a good knowledge of statistical models, I could have created a prediction model to predict the probability of a patient likely to show up for an appointment.

# **References**

1. Godsey, B. (2017). Think like a data scientist: Tackle the data science process step-by-step. Shelter Island: Manning.
2. Information about the maximum recorded age in Brazil: <https://en.wikipedia.org/wiki/List_of_the_oldest_people_by_country>
3. Information regarding Bosla Família**:**

<https://en.wikipedia.org/wiki/Bolsa_Fam%C3%ADlia>

1. Information about the dataset:

<https://www.kaggle.com/joniarroba/noshowappointments/home>