Project 1: Investigate a Dataset (No-show Appointments)

Introduction

Sourced from the Kaggle website and situated in Brazil, the No-show appointments dataset consists of over 100,000 observations. Each observation starts with a patient's ID and ends with whether they made or missed their medical appointments. The 14 columns are listed below with a short explanation.

- PatientID: Unique ID for each patient in the dataset.
- ScheduledDay: The day and time an appointment date was set.
- AppointmentDay: The day the patient is supposed to make their hospital visit.
- Age: The age, in years, of the patient.
- Neighbourhood: The place where the appointment is supposed to take place.
- Scholarship: Whether the patient is a beneficiary of Bolsa Familia.
- Diabetes: Condition of the patient.
- Alcoholism: Condition of the patient.
- Handicap: Condition of the patient.
- SMS_received: Did the patient receive at least one SMS about the appointment.

Research Questions

In exploring the dataset, these six (6) questions were asked and answered:

1. How many people showed up for their medical appointments?
2. What is the age distribution of the people who showed or missed their appointments?
3. Did one gender show up over the over?
4. What condition is mostly present if a patient showed up?
5. Wait times and no-show. How correlated is one with the other?
6. What are the 10 neighbourhoods with the highest show up attendance.

Data Wrangling

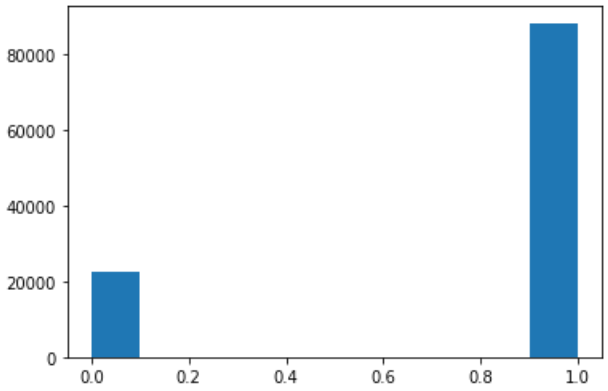
In the data wrangling part of this project, the dataset was loaded into a DataFrame, df. Column names, type, and count of the non-null values of the dataset was gotten using different pandas functions.

The unique values of the Neighbourhood and No-show columns were counted. 'Jardim Camburi' and 'No' had the largest count of 7717 and 88208 respectively.

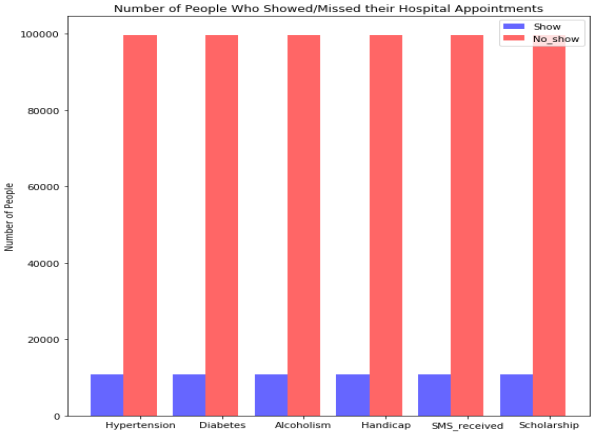
Eight (8) of the fourteen (14) columns were checked, to ensure they had only two unique values. The Handicap column had more than 2 unique values and was noted for adjustments in the cleaning session of the notebook. The age column was also queried for invalid values (values below 0).

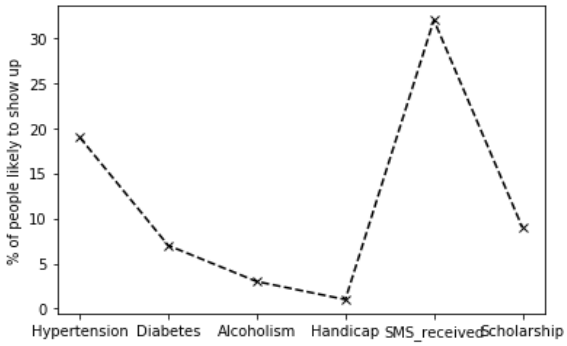
Summary of Results

1. 20% of the observed patients showed up for their appointments.

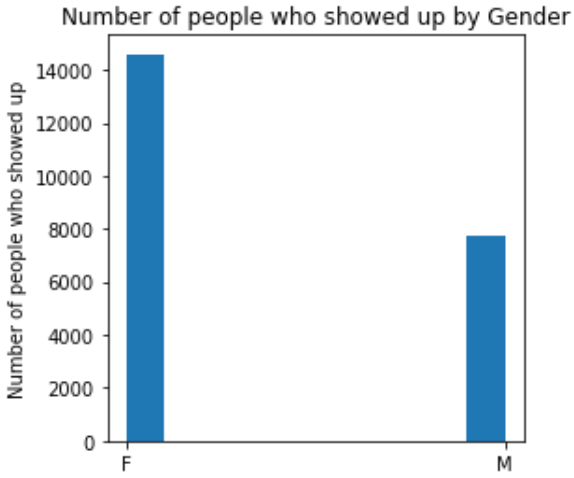


2. If grouped by observed variables, the percentage of people who showed up for their appointments is highest at 32% for people who received at least one SMS, and lowest for people who are handicapped. Of all the conditions observed, the people who are hypertensive had a higher show up rate.

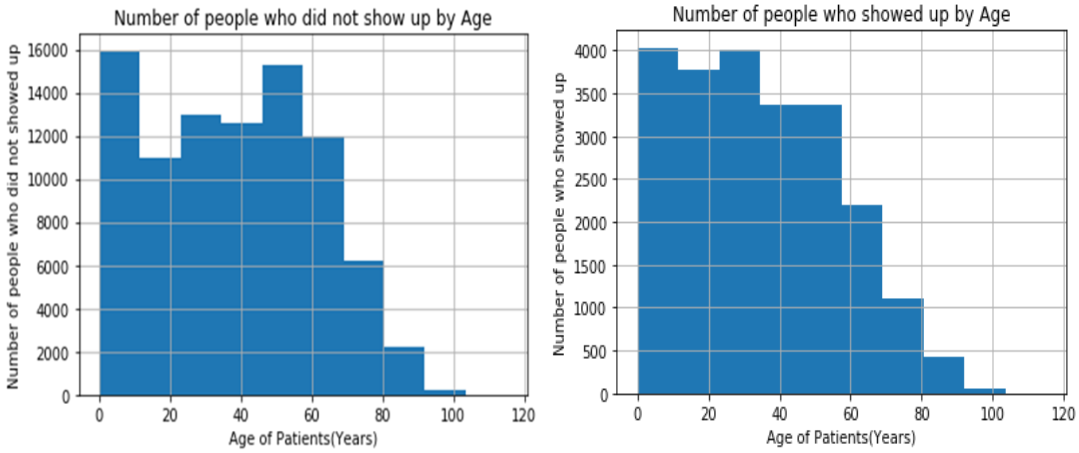




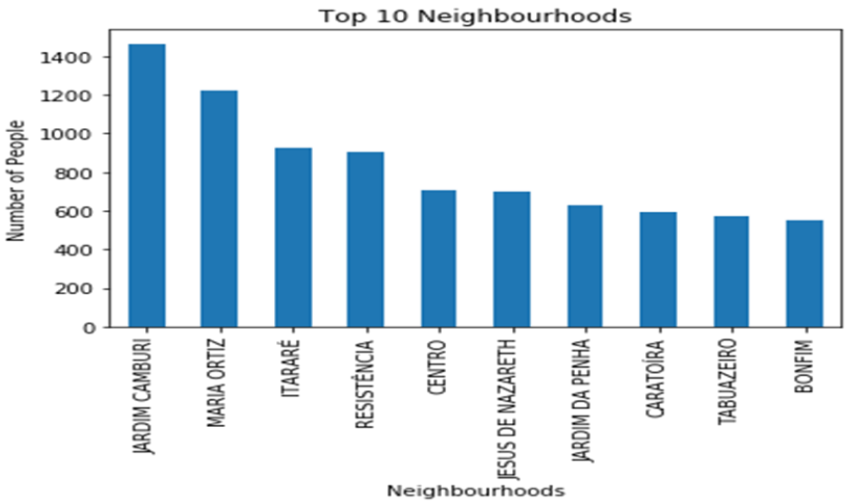
3. 13% of the female respondents went for their appointments compared to the 6% for males.



4. The wait time (the number of days between the day of the actual appointment and the day it was scheduled) was **8 days for the people who missed and 15 days who made** their appointments.
5. The age distribution of the people who showed up is more right skewed than for those who missed their appointment.



6. The bar plot below shows the top 10 neighbourhoods with the highest show-up numbers.


Top 10 Neighbourhoods
Number of People
0
200
400
600
800
1000
1200
1400
JARDIM CAMBURI
MARIA ORTIZ
ITARARÉ
RESISTÊNCIA
CENTRO
JESUS DE NAZARETH
JARDIM DA PENHA
CARATOÍRA
TABUAZEIRO
BONFIM
Neighbourhoods