Data Set Description

1)Data Set Description:

Describe the nature and context of the data set.

- This data set is about medical appointments of the patients in Brazil which records the information of them showing up or not in the day of the appointment. Likewise, a background information about the patient is also given e.g. age, gender, address and their diseases etc. This data set was extracted from Kaggle website and the volume of the dataset is really high, as it has 110527 rows and a lot of relevant columns to work with.
  Details about the columns:
  - **01 - PatientId**
  - It is the patient ID to identify the patient.
  - **02 - AppointmentID**
  - A unique ID to identify each appointment.
  - **03 - Gender**
  - Identifies whether patient is male or female.
  - **04 - Scheduled Day**
  - The date of appointment
  - **05 - Appointment Day**
  - The day when a patient made the appointment to visit the doctor.
  - **06 - Age**
  - what is the age of the patient.
  - **07 - Neighborhood**
  - Where the patient made the appointment.
  - **08 - Scholarship**
  - It tells about whether patient does have the scholarship or not. The name of this scholarship is Bolsa Familia. Bolsa Família provided financial aid to poor Brazilian families.
  - **09 - Hipertension**
  - Does the patient have hypertension or not?
  - **10 - Diabetes**
  - Is the patient suffering from diabetes or not?
  - **11- Alcoholism**
  - Is the patient a alcoholic or not?
  - **12-Handcap**
  - Does the patient have any handicap or not?
  - **13.-SMS_received**
  - Whether the patient has received the SMS or not and if yes how many?
  - **14-No-show**
  - Did the patient show up during the time of appointment or not?

- **b) Identify and discuss potential outcomes from analysis of the data set.**
  There is a lot of potential analysis that can be done with this dataset. For example, we can see if there is a relationship between the gender and their diseases like diabetes or hypertension. Moreover, we can also see the potential relationship between their locations and the diseases they are carrying. Moreover, we can determine and see the outcome of whether the patient will show up and the most correlation it has with the other factors of the patient, whether it has a dependent variable or not.
- **Identify initial data quality, data integrity, and/or data ethics issues related to the data set?**
  The data quality is high, likewise very organized. Moreover, here is the dataset information:
    - **01 – PatientId**
    - It is a unique id for every row to differentiate the patient. However, it is not relevant to analysis.
    - **02 – Appointment ID**
    - It is also a unique id for every row to differentiate the patient. However, it is not relevant to analysis.
    - **03 - Gender**
    - Specifically mentions two genders male and female with Strings “F” and “M”
    - **04 - Scheduled Day**
    - The date of appointment in format
    - **05 - Appointment Day**
    - The appointment information in datetime format
    - **06 - Age**
    - An integer number providing the information.
    - **07 - Neighborhood**
    - Where the patient neighborhood visited, and there are 80 unique locations.
    - **08 - Scholarship**
    - 0 and 1 value stating true or false.
    - **09 - Hipertension**
    - 0 and 1 value stating true or false.
    - **10 - Diabetes**
    - 0 and 1 value stating true or false.
    - **11- Alcoholism**
    - 0 and 1 value stating true or false.
    - **12-Handcap**
    - 0 and 1 value stating true or false.
    - **13.-SMS_received**
    - 0 and 1 value stating true or false.
    - **14-No-show**
    - No-show string value “No” and “Yes” whether the patient showed up or not

  Also, the data integrity is high, as there are no mistaken values. Likewise, there is no missing values. Some of the values might be anomaly, however, they can be fixed.