First Project Report

Dataset: noshowappointments-kagglev2-may-2016.csv

Questions: 1 – Did most of the patients miss the appointment or not?

2 – What is the percentage of the patients that have Hypertension, Diabetes, Alcoholism or Handicap?

3 – Is the SMS messages help the patients to remember the appointment?

4 – Is the scholarship has effect on their appointment or not?

5 – Is there a relationship between Gender and missing their appointment?

6 – Is there any correlation between the “show” column (The dependent variable) and the other columns (“The independent columns”)?

7 – If there is correlation between the dependent column and independent column, Is it strong correlation or weak correlation?

I made visualizations of data using pandas,seaborn and matplotlib library to answer these questions

I used histograms to know the frequency of rows

Using boxplots to find five summary numbers of age column and detecting the outliers

Using bar plot for each numeric column to find which patients have scholarship

Using bar plots to find if most of the patients have Diabetes,Handicap,Hypertensions

Using heat maps to find the correlation between the dependent column and independent column

Using bar plots between every independent column and dependent column

Data Wrangling: Gathering

Reading the noshowappointments-kagglev2-may-2016.csv file

using pandas to manipulate the data frame

Assesing:

Using methods like Describe, Info, Sample, shape, value_counts to assess the data visually and programmatically

Cleaning:

1 - converting the type of patientid column to integer

2 - removing PatientId column because all rows have the same value after converting to integer so it is useless

3 - removing the rows whose age is less than 1

4 - the handcap col should be 0 or 1 only

5 - removing "T00:00:00Z" from AppointmentDay column

6 - removing characters from ScheduledDay column

7 - ScheduledDay and AppointmentDay columns should be datetime instead of string

8 - making date and time of ScheduledDay column in seperate columns
9 - dropping ScheduledDay column

10 - The name of Hypertension and Handcap column should be Hipertension and Handicap

Conclusions :

1 - Being enrolled in the Scholarship program does not seem to make people more likely to show up the appointment

2 - SMS_received and Scholarship columns have weak correlation between them and Show column

3 - The distribution of Age between the patients is not big

4 - Patients that have any disease like diabetes,alcoholism,handicap or hypertension most of them didn't miss the appointment

5 - Handicap patients are more likely to show up in the appointment compared to people who are not handicap

Limitations :

1 - Some rows has the same date in the Appontmentday and ScheduledDay I think it is entry problem

2 - I used only Descriptive analysis and didnot use inferences or hypotheses to our data

3 - The data showed that SMS_receivers didont have strong correlation with show column although It should be the patients who received messages are more likely to attend the appointment