**Exploratory Data Analysis using SAS Programming**

**What is EDA –**

The Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends or to check assumptions with the help of statistical summary and graphical representations.

The main goal of EDA is to get a full understanding of the data and draw attention to its most important features. It helps in detecting anomalies in data and reveal missing or duplicate values.

**Objective**

In this project, I tried to analyze why would some patient not show up for his/her medical appointment and whether there are reasons for that using the data at hand. I tried to find some correlation between different attributes the data has and if they have any effect on whether the patient shows up or not. The dataset I used contains 110527 medical appointments and its 14 associated variables ( PatientId, AppointmentID, Gender, ScheduledDay, AppointmentDay, Age, Neighbourhood, Scholarship, Hypertension, Diabetes, Alcoholism, Handcap, SMS\_received, No-show ).

**Data Dictionary**

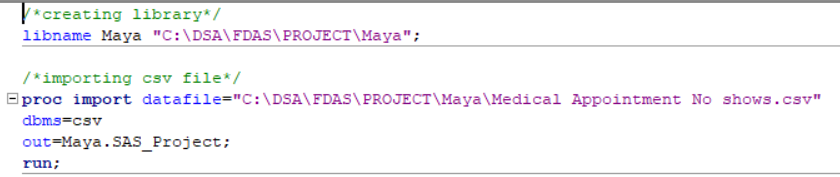
1. PatientId — Identification of a patient
2. AppointmentID — Identification of each appointment
3. Gender — Male or Female . Female is the greater proportion, woman takes way more care of they health in comparison to man.
4. Appointmentday — The day of the actual appointment, when they have to visit the doctor.
5. Sheduledday — The day someone called or registered the appointment, this is before appointment of course.
6. Age — How old is the patient.
7. Neighbourhood — Where the appointment takes place.
8. Scholarship — True of False (1, 0)
9. Hipertension — True or False (1, 0)
10. Diabetes — True or False (1, 0)
11. Alcoholism — True or False (1, 0)
12. Handcap — True or False (1, 0)
13. SMS\_received — 1 or more messages sent to the patient.
14. No-show — True or False (1, 0)

**Questions I tried to answer -**

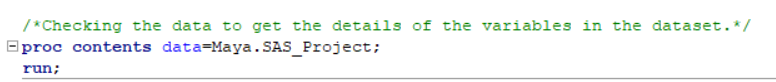
1. What is the percentage of no-shows?
2. What are the factors which may affect if a patient will show up for their scheduled appointment or not?
3. Is there any effect of Gender on if a patient will show up for the booked appointment or not?
4. Are patients with scholarship more likely to miss their appointment?
5. Are patients who don’t receive sms more likely to miss their appointment?
6. Is the time difference between the scheduling and appointment has any impact on whether a patient will show up for their appointment or not?
7. Does patient’s age affect whether a patient will show up or not?
8. What are five neighborhood with highest number of patients missing their appointments from all the neighborhoods.?
9. On which weekdays people don’t show up most often?

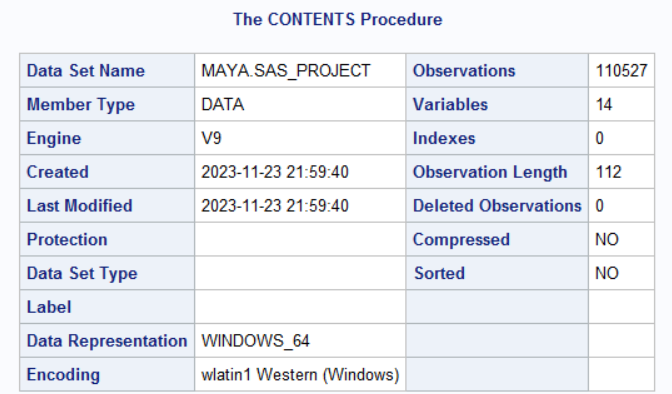
**Setup**

1. I created a SAS library in my name ‘Maya’. Then I used PROC IMPORT procedure to read the .csv file into the SAS environment.



1. I used the PROC CONTENTS to get the detailed information about data including the number of observations = 110527, Variables = 14 along with length and types of variables in the data set.



 A screenshot of a computer screen

Description automatically generated

1. I used PROC MEANS procedure to show the mean values for all the numeric variables in the data by default.

A zero value for the following variables means -

Scholarship = 0 means no scholarship

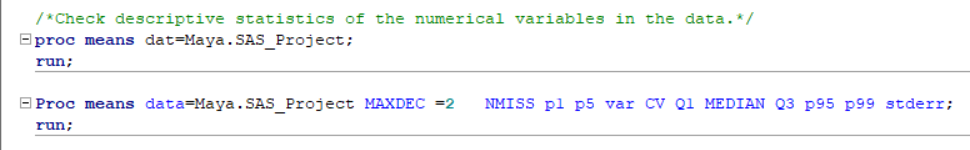
Hipertension = 0 No Hipertension (as spelt in the original dataset)

Diabetes = 0 No Diabetes

Alcoholism = 0 No Alcoholism

Handicap = 0 No Handicap

SMS\_received means = 0 - No sms received 1 — sms received



1. Check missing values in the data set



A screenshot of a computer

Description automatically generated

There are no Missing values in the data as evident from the PROC MEANS output.

**Notes On Data Exploration**

It is evident from the info we got from the initial analysis that there are some columns that need to have their type corrected like dates. Another finding is that our data doesn’t have any missing values. Most of the duplicate values are for the same person for a future appointment.

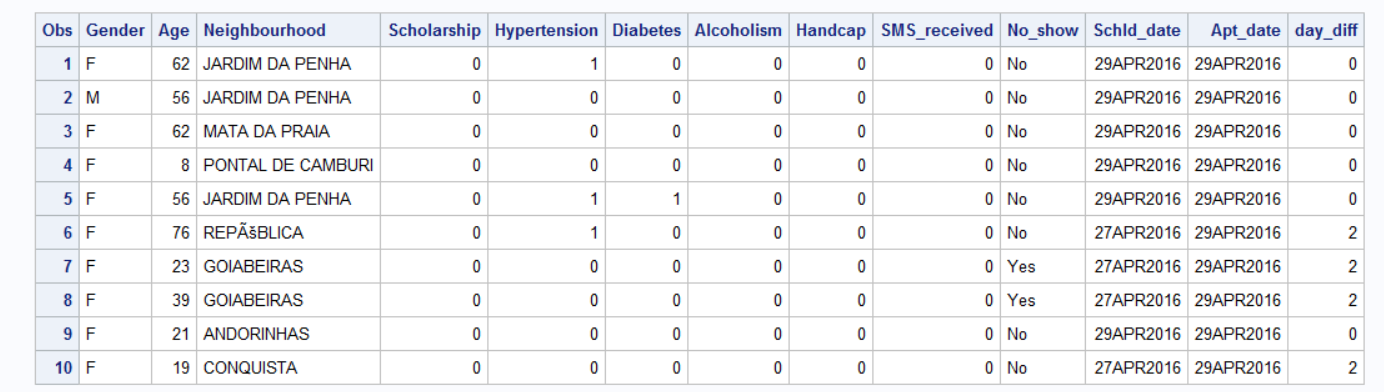
1. I tried to make the following changes to the data set
   * Getting the date only part from datetime values
   * Rename hipertension column to Hypertension
   * Create a new column day\_diff for difference between scheduling and appointment days
   * Drop patientid and appointmentid columns

A computer code with blue and white text

Description automatically generated

1. Checking the updated data

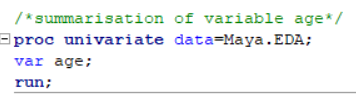




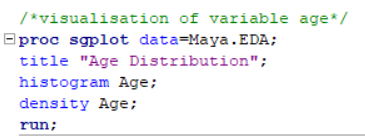
**Univariate Analysis**

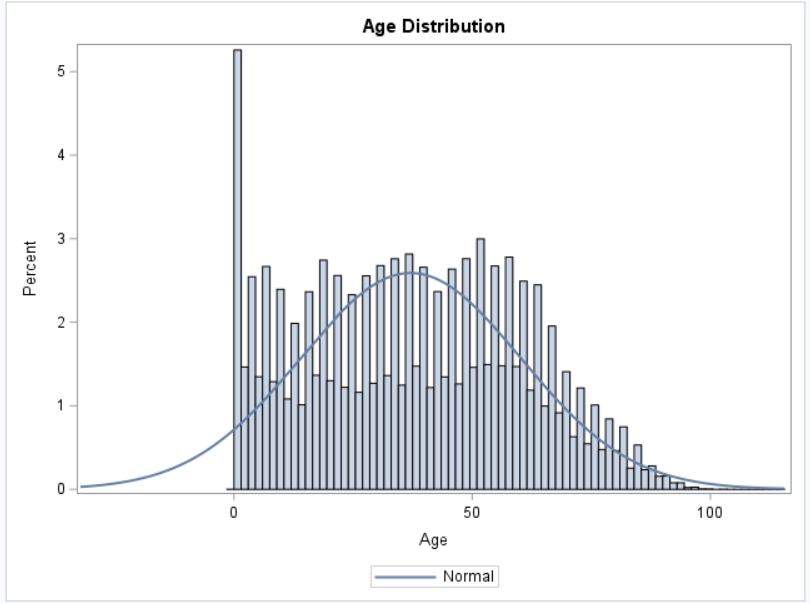
1.Numeric variables

Age

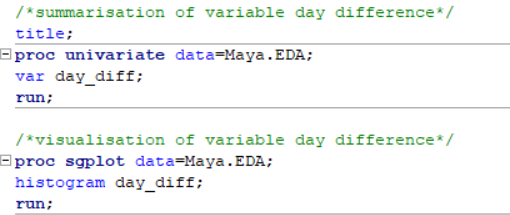


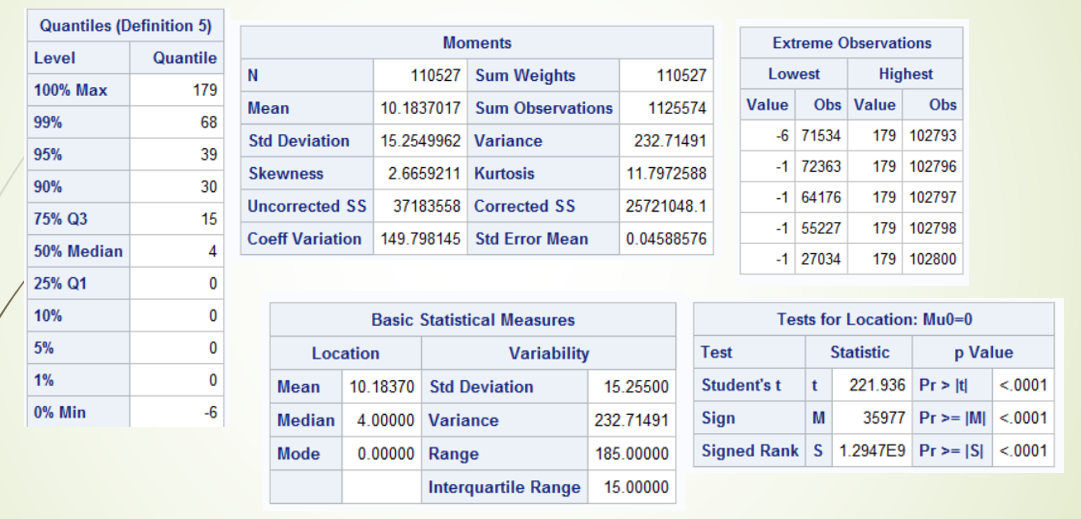


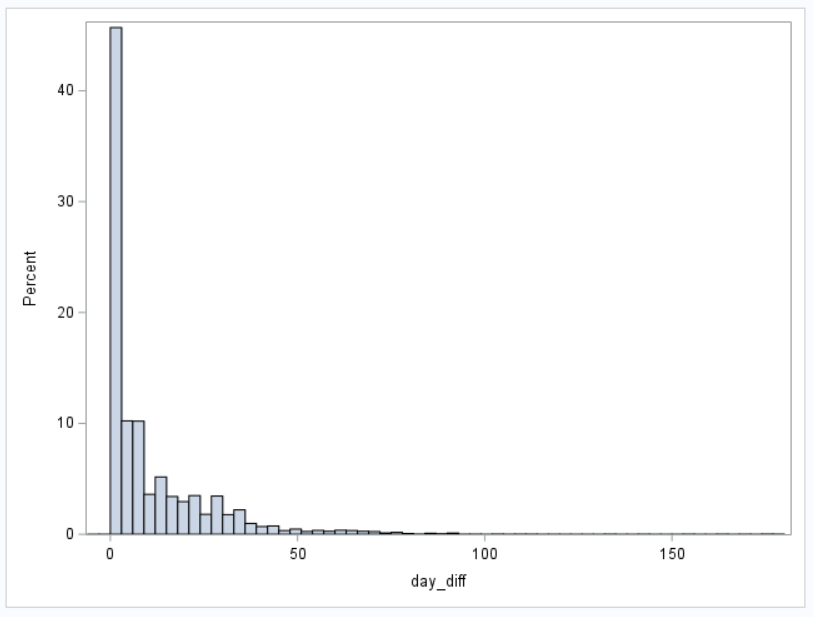




Day Difference



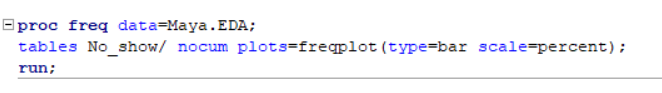


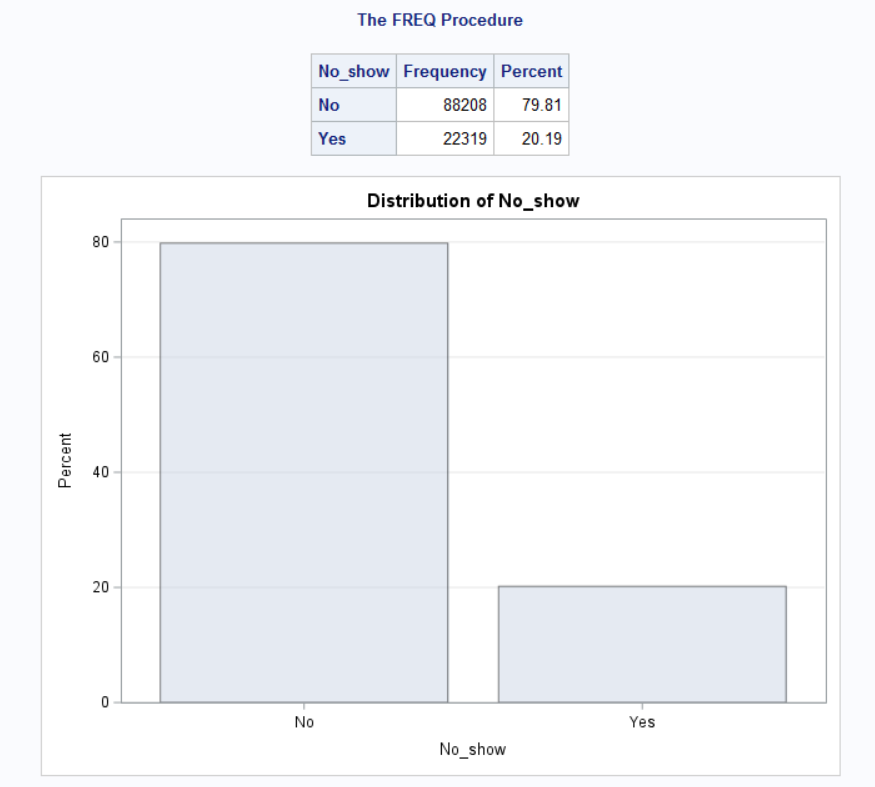


**Bivariate Analysis**

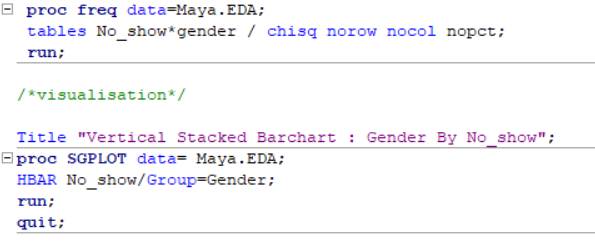
1. What is the percentage of no-shows?

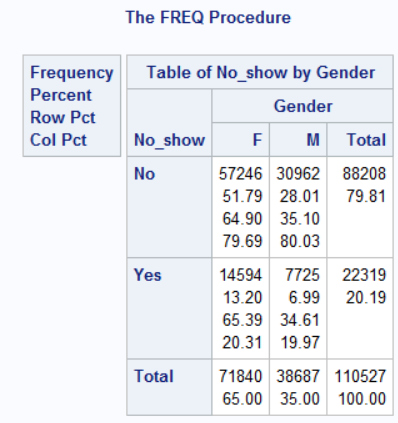
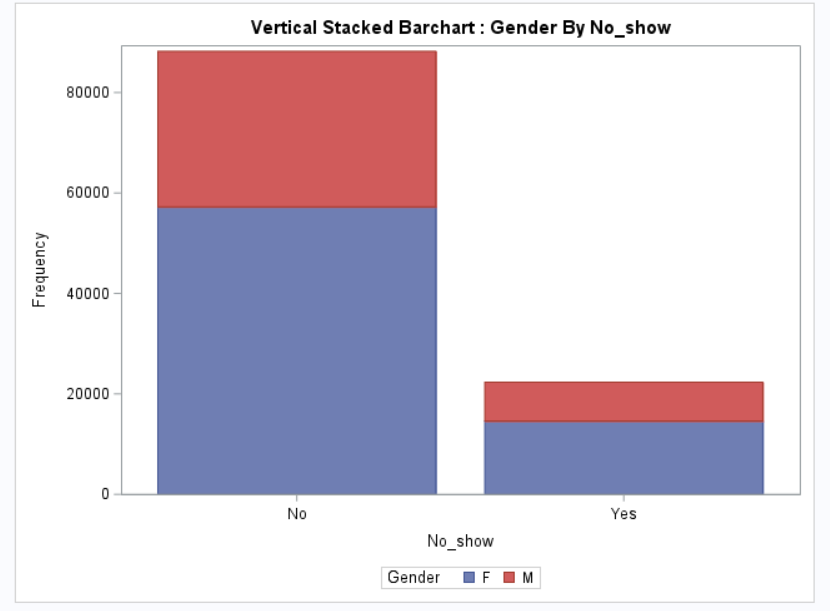
I used PROC FREQ procedure to get the percentage of Show and create a bar graph.





1. Is there any effect of Gender on if a patient will show up for the booked appointment or not?



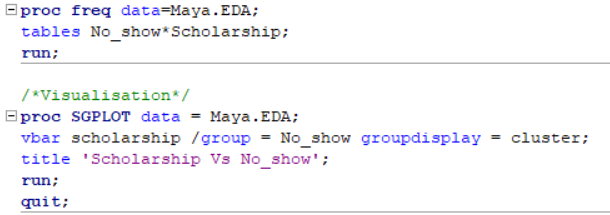
 

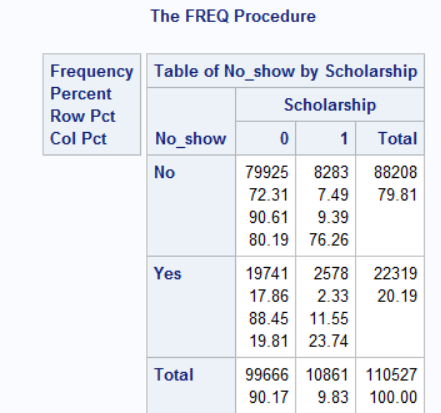
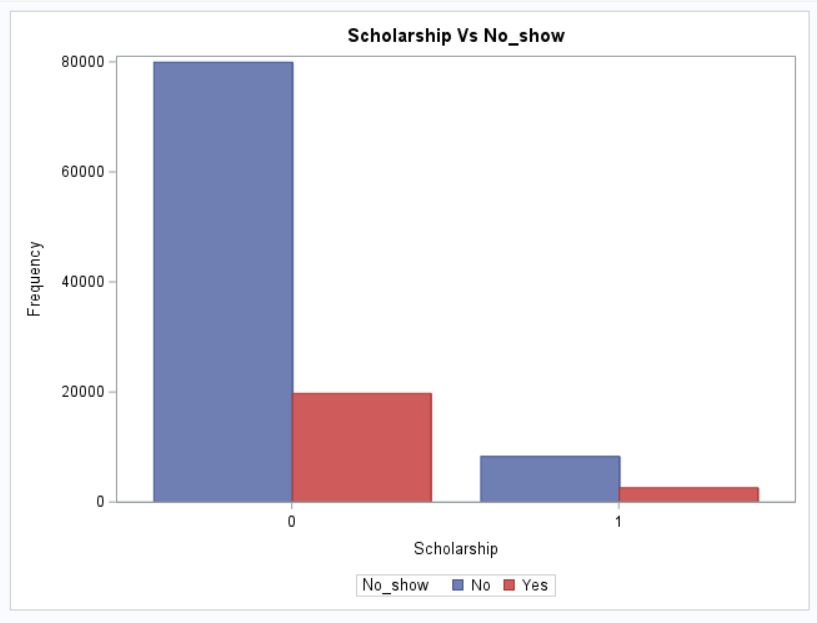
Females who missed their appointment — 14594/110527 = 0.132 \* 100 = 13.2

Males who missed their appointment — 7725/110527 = 0.0699 \* 100 = 6.99

*Findings* — As seen from the results, percentage of Females missing their appointment is double the number of the males missing their appointment. Based on findings we can say that females are more likely to miss their appointment.

1. Are patients with scholarships more likely to miss their appointment?



Patients without scholarship who show up for their appointment — 79925/99666 = 0.8019 \* 100 = 80.19

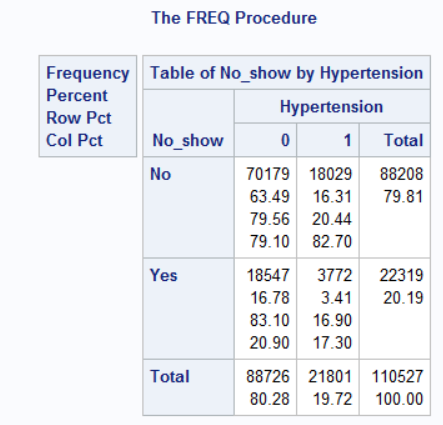
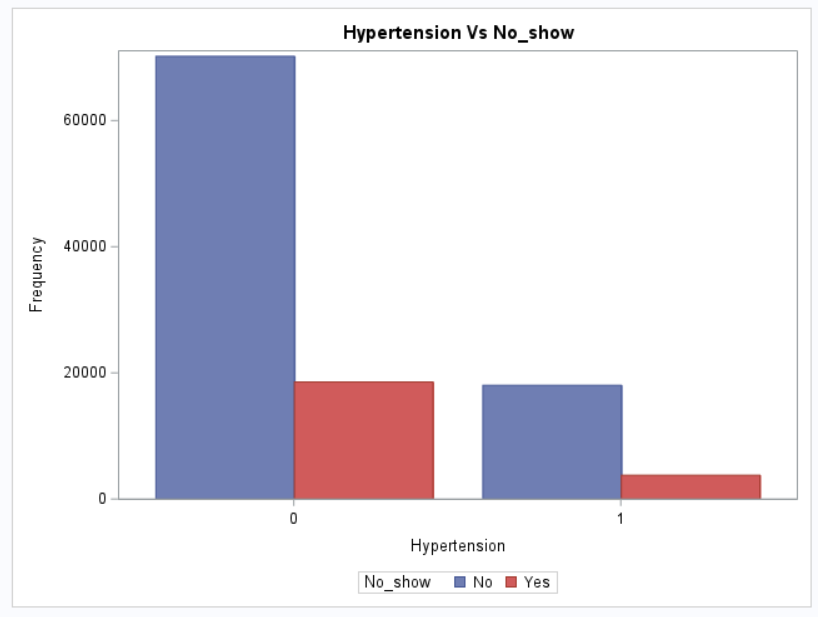
Patients with scholarship who show up for their appointment — 8283/10861 = 0.7626 \* 100 = 76.26

***Finding*** — It seems that patients with scholarships are actually more likely to miss their appointment.

1. Are patients with hypertension more likely to miss their appointment?

A screenshot of a computer program

Description automatically generated

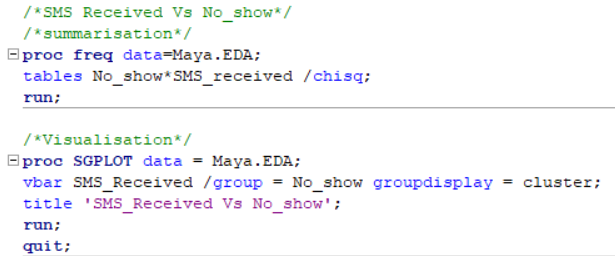
 

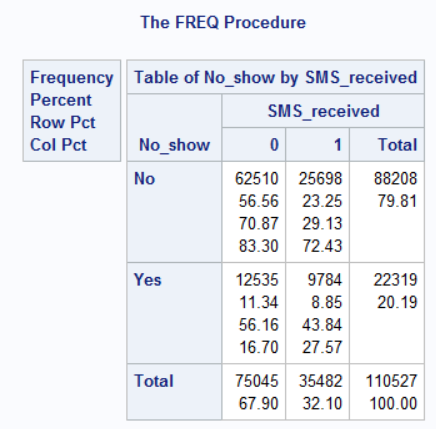
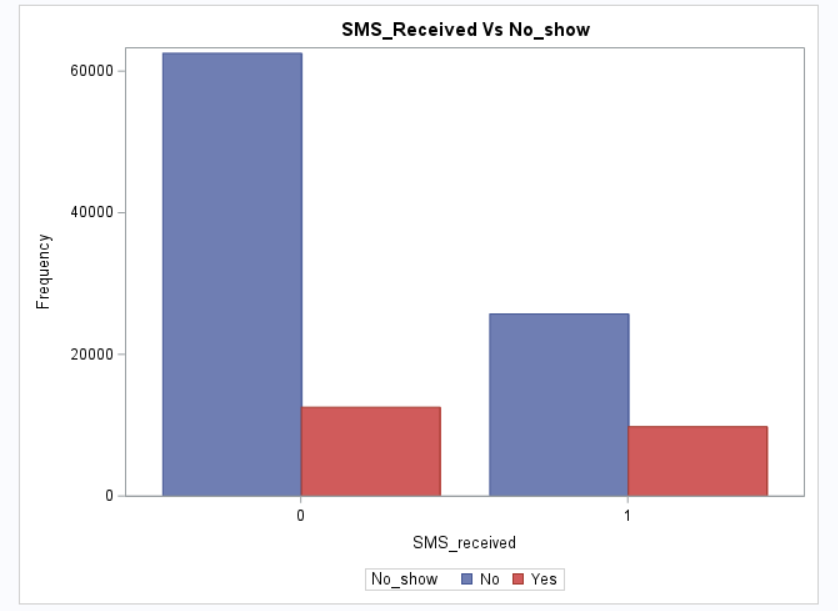
The patients who do not have hypertension and showed up for their appointment — 70179/88208=0.7956

Patients with hypertension who showed up for their appointments — 18029/21801=0.827

***Finding*** — It seems that patients with hypertension are actually more likely to show up for their appointment

1. Are patients who don't receive SMS more likely to miss their appointment?



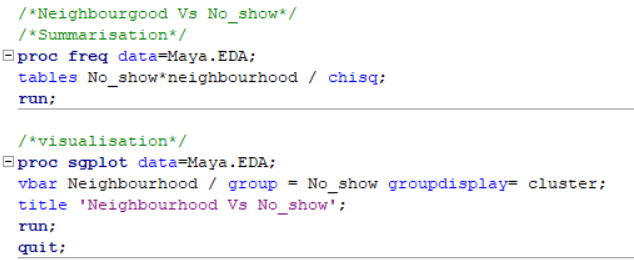
 

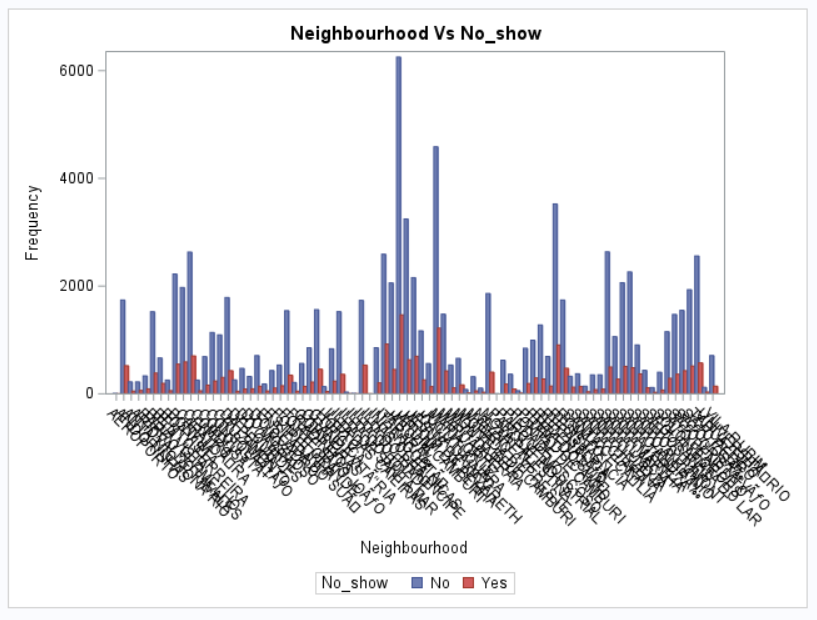
Patients who did not receive sms and showed up for their appointment — 62510/75045=0.833

Pateints who received sms and showed up for their appointment — 25698/35482 = 0.7243

***Finding*** — A strange finding here suggests that patients who received an SMS are more likely to miss their appointment.

1. Which are the neighborhood where patients more than 500 are missing their appointments ?





Since the contingency table is too big because of large number of neighbourhoods, here I am only attaching the grouped bar chart

As per the analysis — The following neighborhoods are the ones where most patients missed their appointment.

Jardim Camburi tops the chart with 1465 missed appointments.

Jardim Camburi — 1465

Maria Ortiz — 1219

ITARARE — 923

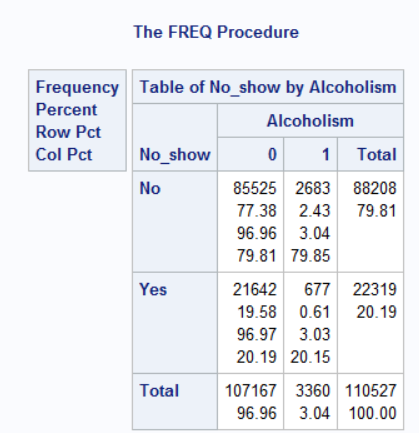
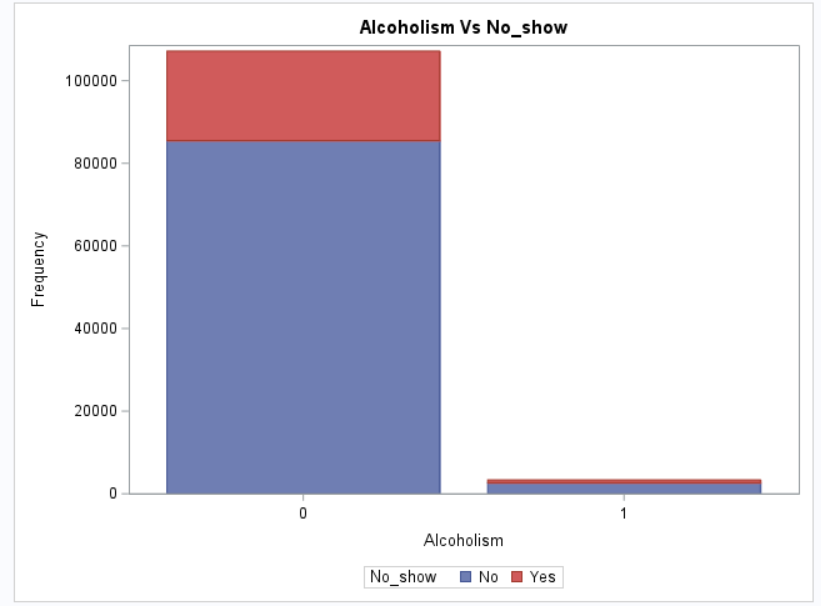
Resistencia — 906

Centro — 703

1. Does alcoholism affect whether a patient will show up or not?

A screenshot of a computer code

Description automatically generated

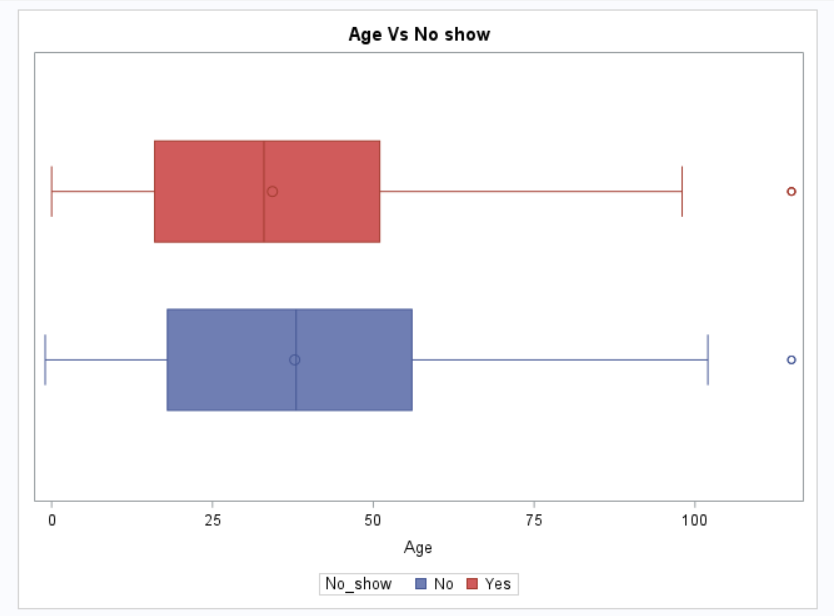
 

***Findings –*** As per the chart, patients who are not alcoholic tends to miss the appointment more.

1. Is there an effect of Age if the patient will turn up for their appointment or not.?

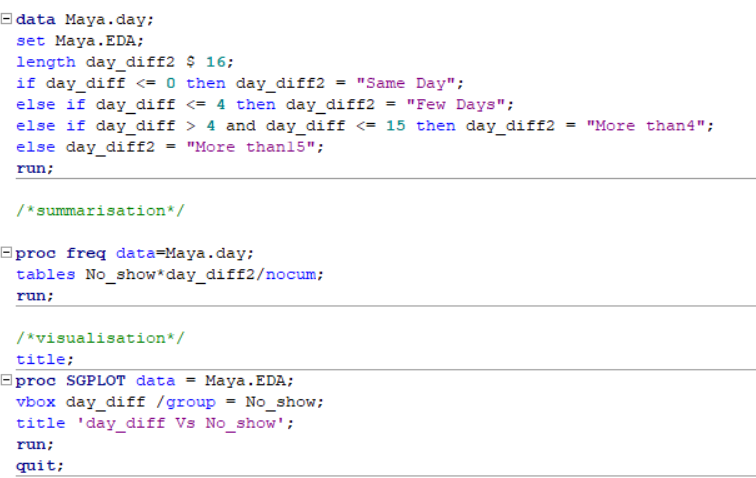
A screenshot of a computer code

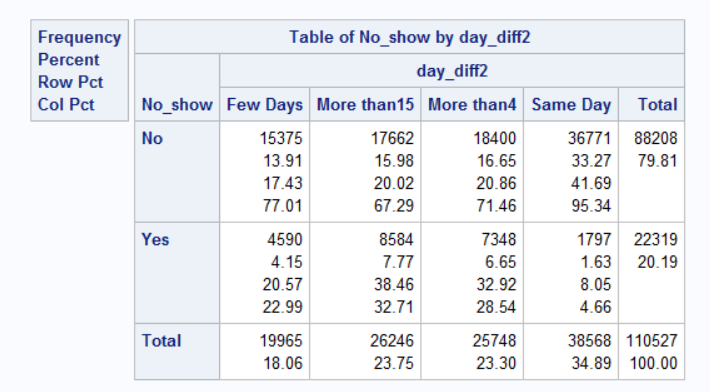
Description automatically generated

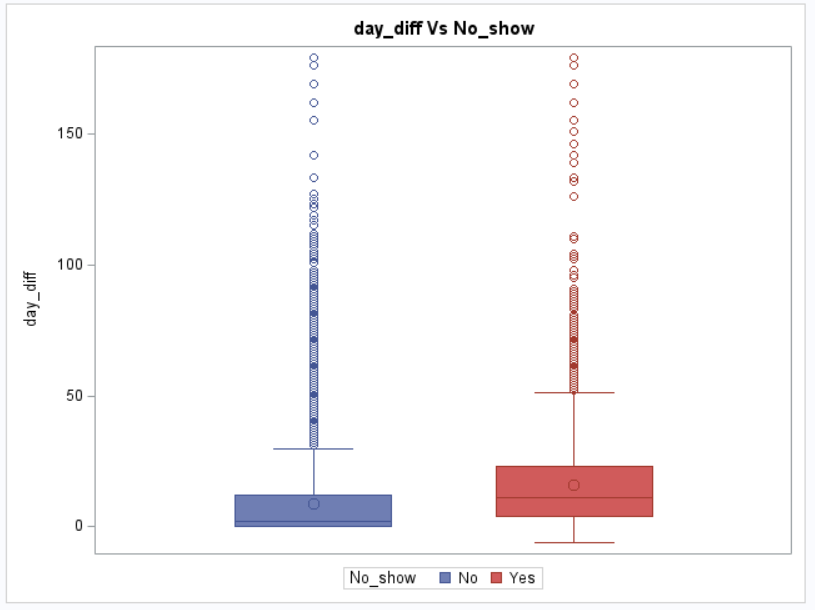


***Finding*** — There is no clear correlation between age and whether the patient shows up for an appointment or not but it is evident from the histogram above that younger patients between the age of 6 years and 42 years are likely to miss their appointments.

1. Is the time difference between the scheduling and appointment related to whether a patient will show up for an appointment or not?







No Show Percentages as per difference in dates of scheduling an appointment and actual appointment day

Few day — 4601/22319 = 0.2061 \*100 = 20.61

More than 15 days — 8579/22319= 0.3844 \*100 = 38.44

More than 4 days = 7333/22319 = 0.3286 \*100 = 32.86

Same day = 1806/22319 = 0.0809 \*100 = 8.09

***Finding*** — It appears that the longer the period between the scheduling date and appointment date the more likely is that the patient won’t show up for their appointment.

**T-test – Day difference Vs No show**

A screen shot of a computer

Description automatically generated



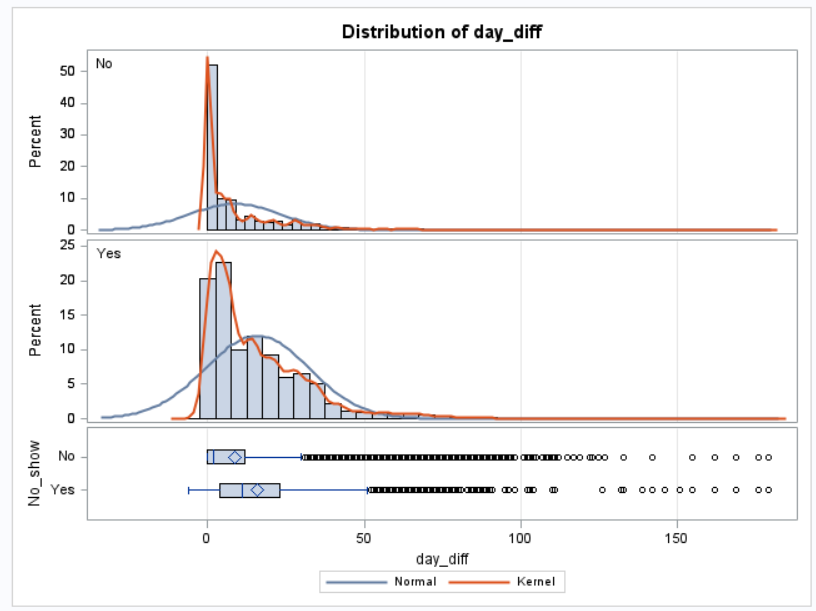
H0 : number of days between scheduled date and appointment date and whether they show upto appointment are independent

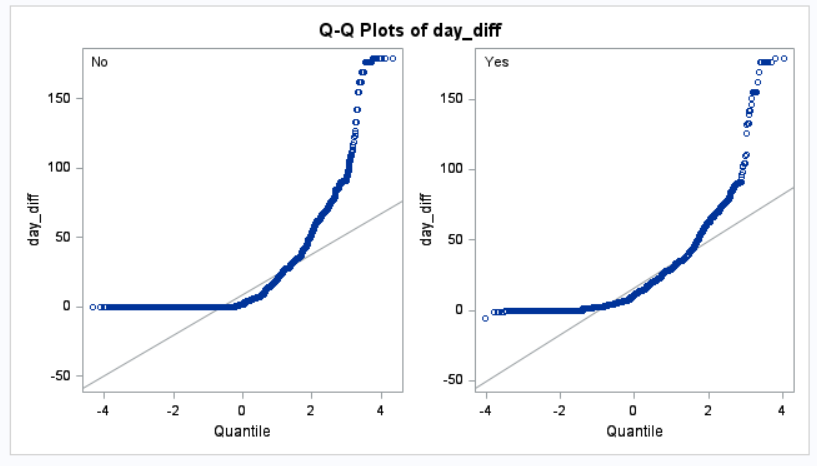
Ha : number of days between scheduled date and appointment date and whether they show upto appointment are not independent

Here we can see 4 tables ,the first one is descriptive statistics of continuous variable between different category of categorical variable.In that table you can see the difference of mean between 2 levels.

Then we will move on to the forth table which is equality of variance. Here as the p-value is less than 5% ,we can reject null hypothesis. In other words the variance of day difference is not equal between no\_show yes and no\_show no.

When we move on to third table the p-value is less than 5% which means the mean day difference is not equal between two levels of no\_show so the difference of 7.07 is statistically significant.

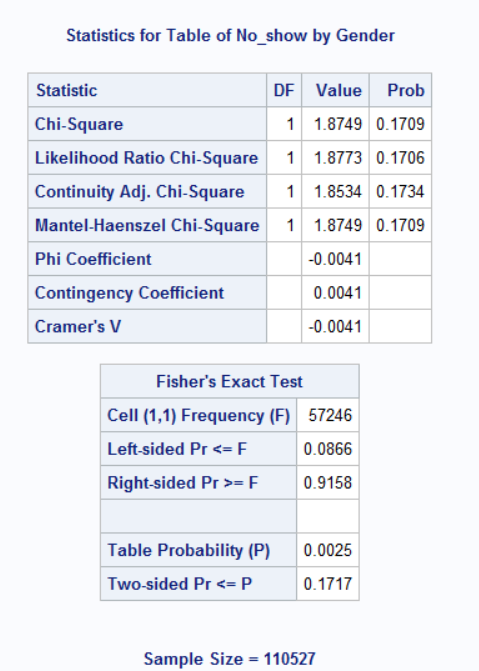




**Chi Square test – Gender Vs No show**

A screenshot of a computer

Description automatically generated



H0 : Gender and whether they miss an appointment are independent

Ha : Gender and whether they miss an appointment are not independent

Gender and No\_show has 2 levels, so degree of freedom for both gender and No\_show are 2-1=1

For Chi square test, degree of freedom =1\*1=1

Since the p-value is more than 5% we cannot reject the null hypothesis as there is not enough evidence to conclude that the gender and no\_show are associated

**Shortcoming & Data Limitations**

* The data doesn’t state the exact hour of the appointment which would have been very useful to try to find out which hours have the most missing appointments and which doesn’t. It could also be very useful to know the difference between scheduling and the appointment since many of the scheduling are on the same day.
* The age column had a negative value but according to the data creator, it means a baby not born yet (a pregnant woman).
* When calculating the day difference between the scheduling and appointment days we had some negative value which makes no sense and might mean that the records of questions have wrong data.

**CONCLUSIONS**

* After analyzing the dataset here are some findings:
* Percentage of patients who didn’t show up for their appointment is 20.19%.
* The percentage of females missing their appointment is nearly two times the number of males. So females are more likely to miss their appointment.
* It appears that the longer the period between the scheduling and appointment the more likely the patient won’t show up.
* It seems that patients with scholarships are actually more likely to miss their appointment.
* A strange finding here suggests that patients who received an SMS are more likely to miss their appointment.
* There is no clear relation between the age and whether the patients show up or not but patients between 6 and 42 years are more likely to miss their appointments.