# STAT-S470 Final Project

Zining Qi, Yexin Lu, and Mengchu Li

April 30, 2021

**Abstract**

In this experiment we studied the data set of medical appointments. We focus on exploring no-show data and the relationships between whether people show up when they have medical appointments and some special factors (i.e., gender, age, disease, etc).

## 1 Goal

With the Medical Appointment No Shows data set, we are trying to find the answers to these questions:

- Why about 20% of people miss their medical appointment?
- Are there any variable that in this data set can directly explain why the patients do not show up for their medical appointment? If there's not, whether there are more complicated factors to influence the no-show data.
- How does the variable in the data set influence the no-show result? If there shows a relationship between these variables, is there any reasonable explanations?
- Based on the questions before, would there be multiple factors influencing no-show result together?

## 2 Description of the Data Set

This Medical Appointment No Show data set we get from the Kaggle: https://www.kaggle.com/joniarroba/noshowappointments. This data set includes 110,527 medical appointments and the relevant 14 variables (characteristics). Here we have the introduction for every characteristics:

- PatientId – Identification of a patient
- AppointmentID – Identification of each appointment
- AppointmentDay – time data

- ScheduledDay – time data
- Gender – Male or Female
- Age – How old is the patient
- Hipertension – True or False. Whether the patient has this diseases.
- Diabetes – True or False. Whether the patient has this diseases.
- Alcoholism – True or False. Whether the patient has this diseases.
- Handcap – True or False. Whether the patient has this diseases.
- SMS_received – True or False. Whether the patient receive the text message reminder for the appointment.
- Noshow – True(not show up) or False (show up).
- Neighbourhood
- Scholarship

The first 12 characteristics are useful for our analysis and the last two are not useful for our analysis. Then let's take a more detailed look of the data set. Here is the summary of the data set:

```
  PatientId         AppointmentID     Gender     ScheduledDay
Min.   :3.920e+04   Min.   :5030230   M:38687   Min.   :2015-11-10 07:13:56
1st Qu.:4.173e+12   1st Qu.:5640286   F:71840   1st Qu.:2016-04-29 10:27:01
Median :3.173e+13   Median :5680573             Median :2016-05-10 12:13:17
Mean   :1.475e+14   Mean   :5675305             Mean   :2016-05-09 07:49:15
3rd Qu.:9.439e+13   3rd Qu.:5725524             3rd Qu.:2016-05-20 11:18:37
Max.   :1.000e+15   Max.   :5790484             Max.   :2016-06-08 20:07:23
AppointmentDay                     Age          Neighbourhood      Scholarship     Hipertension
Min.   :2016-04-29 00:00:00   Min.   : -1.00   Length:110527      Mode :logical   Mode :logical
1st Qu.:2016-05-09 00:00:00   1st Qu.: 18.00   Class :character   FALSE:99666     FALSE:88726
Median :2016-05-18 00:00:00   Median : 37.00   Mode  :character   TRUE :10861     TRUE :21801
Mean   :2016-05-19 00:57:50   Mean   : 37.09
3rd Qu.:2016-05-31 00:00:00   3rd Qu.: 55.00
Max.   :2016-06-08 00:00:00   Max.   :115.00
 Diabetes       Alcoholism       Handcap         SMS_received    Noshow
Mode :logical   Mode :logical   Mode :logical   Mode :logical   No :88208
FALSE:102584    FALSE:107167    FALSE:108286    FALSE:75045     Yes:22319
TRUE :7943      TRUE :3360      TRUE :2241      TRUE :35482
```

Figure 1: Summary of the Medical Appointment No Show Data Set

The three useful numerical characteristics are PatientID, AppointmentID, and Age. And most of the characteristics are logical, which are binary (0 or 1).

# 3 Analysis

To answer the above goal questions, we did two parts of analysis based on our data set. And first let's see the overall no-show data distribution.

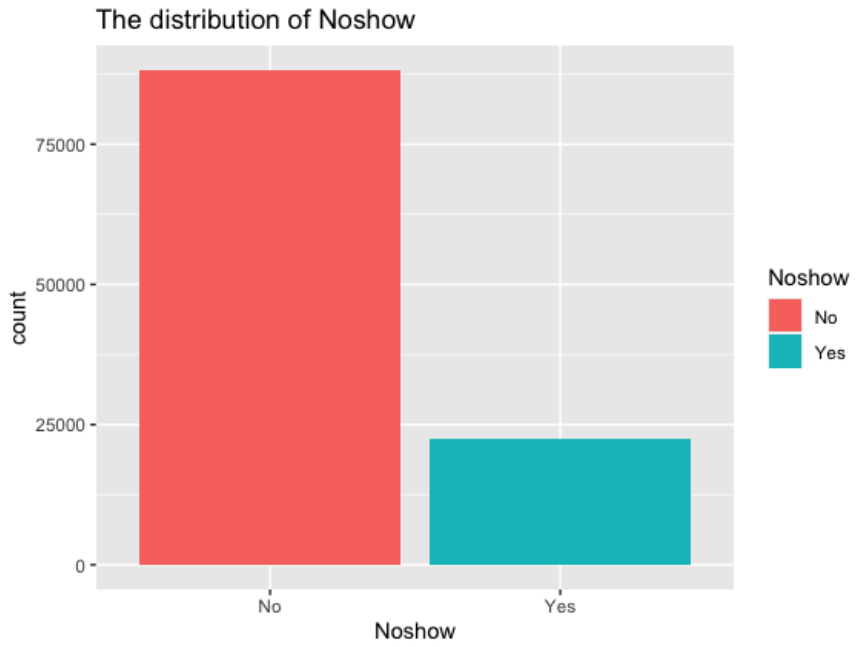


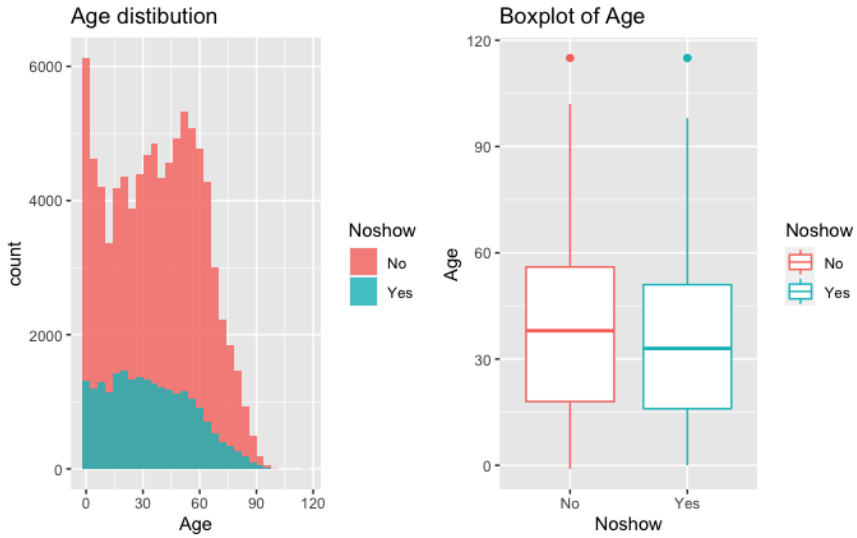

Figure 2: Overall No-show Distribution

## 3.1 Overall Distribution

Using the data set to calculate, we can get that 79.73586% would not show up for their medical appointment and 20.26414% would show up for their medical appointment and also see from the Figure 2.

## 3.2 Single Feature Influence on No-show data

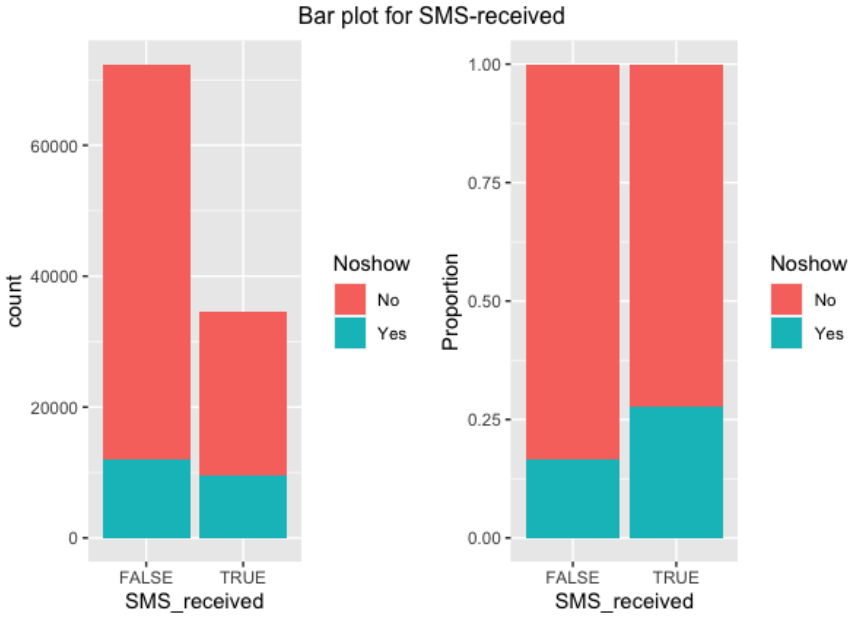
- **AGE**




We use red to indicate patients show up, blue indicates patients who didn't show up. From the distribution of age by noshow and showup groups, we can see that the

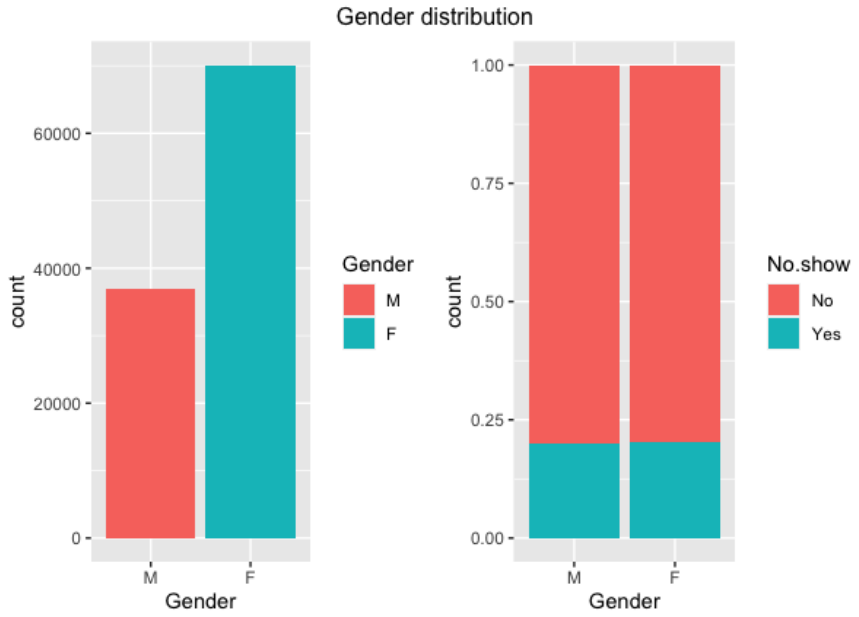
patients who's age are between 0-10 and 45-60 are more likely to showup for their medical appointment. And the age of the patients are 10-45 are more likely to not show up for their medical appointment. And the distribution of age is right skewed, so we might to perform a transformation in the further analysis. From the box plot of age, we can see that although there are some outliers in both noshow and showup group, the average age of noshow is younger than the average age of showup, which means that age might be a determinant of "showup" chance.

- **SMS_Received**
  Then let's take a look of the influence of text message reminder.

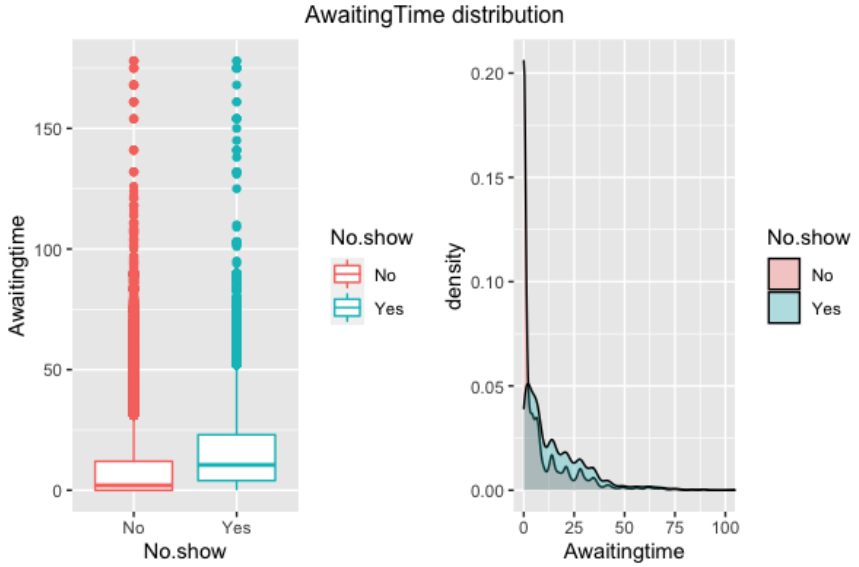



From the above plot, we can see that the number of patients who received SMS message but didn't show up is about 10000, but the number of patients who didn't receive SMS message but didn't show up is more than 10000. Since the difference between these two numbers are very small, we cannot get any useful conclusion by that plot.
Thus, we plot the proportion of show up and not show up by SMS_received. From that plot, we can see that the proportion of not show up in didn't receive SMS message group is smaller than the proportion of not show up in received SMS message group. Thus, SMS_received might be a factor to impact the "showup" chance.

- **Gender**
  From the plot of distribution of gender, we can see that the number of male patients is much less than the female patients. Thus, we plot the proportion of show up and noshow by gender. We can see that the proportion of men and women that didn't show up are very similar. Hence, we cannot get any conclusion about the relationship between gender and showup.

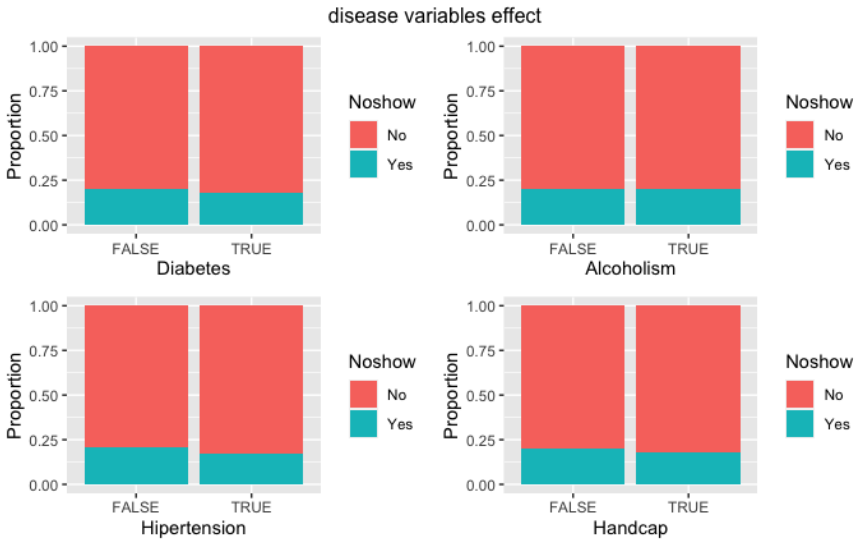



- **Awaiting Time**
  Awaiting time is the time difference between ScheduledDay and AppointmentDay.




From the above plots we can see that the Awaiting time for noshow is larger than Awaiting time for show up. So longer awaiting time might lead to the higher chance to no show.

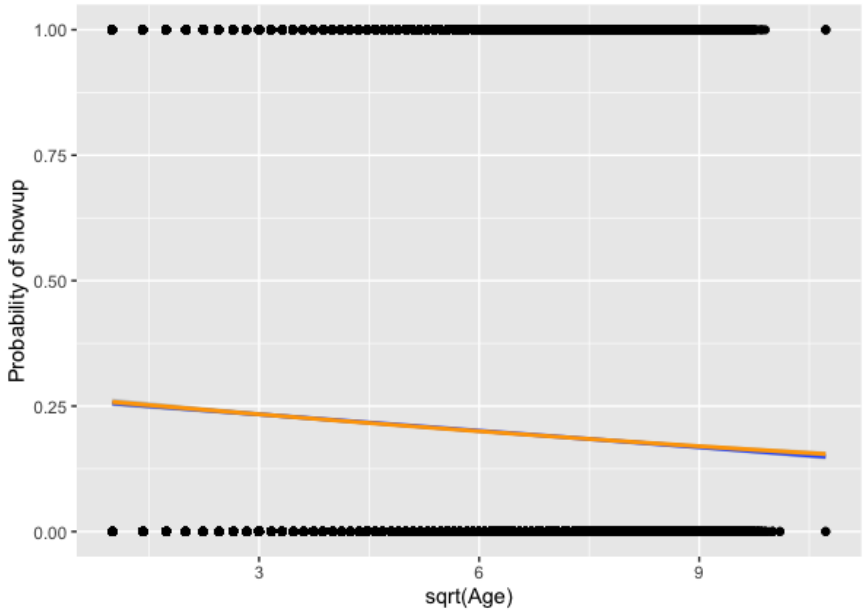
- **Disease: Hipertension, Diabetes, Alcoholism, Handcap**
  From the following set of plots for binary variables, we can see some differences in the proportion for each variables except Alcoholism, but the differences are very small in each of the variables that indicate the type of disease. Thus, we cannot get enough evidence to conclude the relationship between disease and chance of show up.




## 3.3 More relationships: Logistic Regression

- **One Predictor: Age**

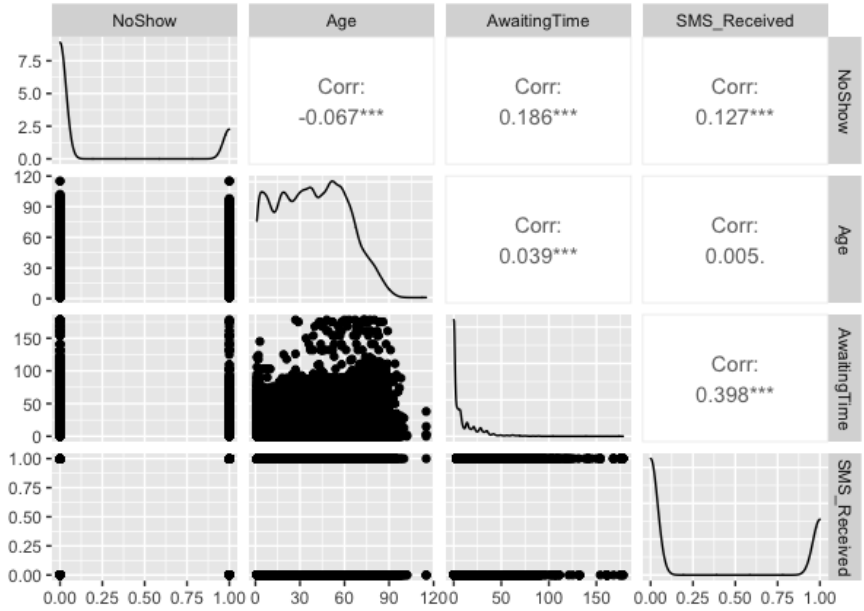
  We choose to use square root transformation on Age in this section, because we noticed that the distribution of age is right skewed in the previous part.



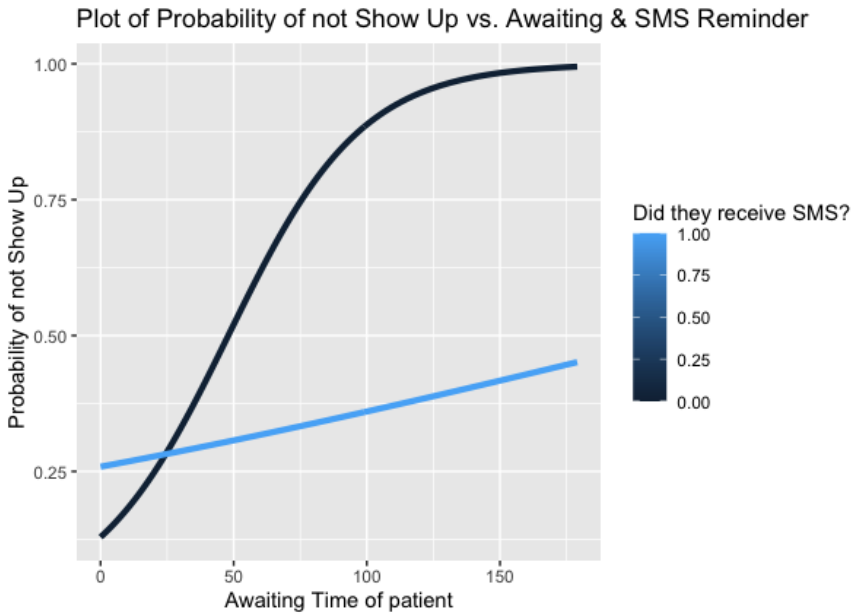

In the above plot, we display the prediction of both liner regression and logistic regression, and we can see that there is not very obvious difference between these two method. And there is a negative relationship between age and chance to showup. This observation is same as the result that we get from the summary of logistic regression. In this plot, 0 means patient did show up, and 1 means did not show up. There is a negative relationship between show up and age. As the age increase, the probability of not showing up decrease, which means patient tent to be more likely show up as age increase. Then, we want to know if we can use two explanatory variables to explain

the chance of show up.

- **Two Predictors**
  Before we do the two-predictor logistic regression, let us check the correlations of some variables first.




From the plot, we can tell the correlation between age and no show is negative(-0.067) and the density plot of age is right-skewed. This could verify the one predictor model we did in last part. Except age, there are a lot of other variables that may affect the probability of show up, such as Awaiting Time and SMS_Received. They both have correlations with show up. And there is a correlation between AwaitingTime and SMS_Received. So, we did a two predictor model with the interaction.




We plot the above graph by modeling the logistic model of show up with AwaitingTime,

SMS_Received and their interaction. Again, in this plot, 1 means did not show up and 0 means show up. The blue line represents patient receiving text message(sms) reminder, and the black line represents patient did not receive sms. From the plot, we can see as awaiting time increase, the probability of not showing up increase. Without sms reminder, probability of not showing up increases dramatically with AwaitingTime, and almost be 1 when AwaitingTime is larger than 150. While, with sms reminder, the probability of not showing up is much more lower as AwaitingTime increasing. This result did make sense, people tend to forget the appointment when AwaitingTime is long, especially when there is no text message reminder.

### 3.4 Conclusion(Brief Answers to Goal Questions)

- The age, gender, text message reminder, awaiting time and kinds of diseases attributes to the no show of medical appointments.
- From the one-prediction logistic regression, age indicates a negative influence on no-show data. However, from the ggpairs plot, we can get the conclusion that the age is not the only one factor that can affect no-show and there may be multiple factor affecting that.
- We use the two-predictor model for awaitingtime and SMS_Received, and we found out that the reasonable result that patient tend to miss the appointment with the increase of Awaitingtime, especially without text message reminder.

## 4 Limitations & Further Possible Analysis

In our analysis, there several limitation that we need to consider:

- Some data in this data set might not very credible. For example, the range of the patients' age is from -1 to 115. And we all know that age cannot be less than 0.
- There are some outliers in data set, but we didn't exclude them from our analysis. That might impact out result.
- We don't pay too much attention to the residual of the model, so we don't know if the model that we build can explain the chance of show up reasonably.

These three limitations could be good topics to do further analysis.