Medical Appointment No-show

The Project I picked to do for my capstone involves predicting potential no-shows in the healthcare industry. Doctor offices and other medical institutions generate revenue by scheduling appropriate number of patient appointments in any given day for their medical practice. This ensures that the right amount of staff and resources are available to care for their potential patient but also ensures that there is enough revenue to sustain the medical practice. So when patients do not show up, there are two main problems.

1. Loss of Revenue from the patient that did not show-up. The medical institution essentially loses revenue when one of their potential “customers” does not show up. It also prevents the institution from scheduling another patient in that time slot. In both these cases there is a loss of revenue for the medical practice.
2. Waste of Resources: Since staffing and operational decicions for a medical institution is largely dependent on patient volume, when some of those patients don’t show up it causes these institutions to overstaff their offices and commit resources that could be redirected to more needy areas of the medical practice.

Because of these two main reasons, medical practices are very interested in reducing the number of no shows to their practice. They are interested in not only predicting who has a high probability of not showing up but also in ways that they can help remind their patients of their appointments. One of the potential ways they can remind their patients is to remind them of their upcoming appointment closer to the date of the appointment and this ensuring the patient is available and able to keep the appointment.

Dataset

The data set for this analysis came from the kaggle website and was sufficiently big enough to do my analysis. Below is a list of the variables in the dataset and a sampling of the observations in the different variables. There were a total of 300,000 observations and 15 variables in the data set.

```
> glimpse(noshowraw)
Observations: 300,000
Variables: 15
$ Age                    <int> 19, 24, 4, 5, 38, 5, 46, 4, 20, 51, 33, 58, 62, 62, 38, 73, 4...
$ Gender                 <chr> "M", "F", "F", "M", "M", "F", "F", "F", "F", "F", "M", "F", "...
$ AppointmentRegistration <dttm> 2014-12-16 14:46:25, 2015-08-18 07:01:26, 2014-02-17 12:53:4...
$ ApointmentData         <dttm> 2015-01-14, 2015-08-19, 2014-02-18, 2014-08-07, 2015-10-27, ...
$ DayOfTheWeek           <chr> "Wednesday", "Wednesday", "Tuesday", "Thursday", "Tuesday", "...
$ Status                 <chr> "Show-Up", "Show-Up", "Show-Up", "Show-Up", "Show-Up", "No-Sh...
$ Diabetes               <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0...
$ Alcoolism              <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
$ HiperTension           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0...
$ Handcap                <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
$ Smokes                 <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
$ Scholarship            <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0...
$ Tuberculosis           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
$ Sms_Reminder           <int> 0, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0, 0...
$ AwaitingTime           <int> -29, -1, -1, -15, -6, -35, -18, -14, -14, -4, -39, -22, -17, ...
```

```
> summary(noshowraw)
      Age            Gender          AppointmentRegistration       ApointmentData                DayOfTheWeek
 Min.   : -2.00   Length:300000      Min.   :2013-05-29 15:14:11   Min.   :2014-01-02 00:00:00   Length:300000
 1st Qu.: 19.00   Class :character   1st Qu.:2014-06-24 07:45:05   1st Qu.:2014-07-04 00:00:00   Class :character
 Median : 38.00   Mode  :character   Median :2014-12-03 15:35:36   Median :2014-12-16 00:00:00   Mode  :character
 Mean   : 37.81                      Mean   :2014-12-15 04:43:13   Mean   :2014-12-28 13:09:07
 3rd Qu.: 56.00                      3rd Qu.:2015-06-11 09:06:08   3rd Qu.:2015-06-25 00:00:00
 Max.   :113.00                      Max.   :2015-12-29 12:08:58   Max.   :2015-12-30 00:00:00
    Status             Diabetes          Alcoolism         HiperTension        Handcap            Smokes
 Length:300000      Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
 Class :character   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
 Mode  :character   Median :0.00000   Median :0.00000   Median :0.0000   Median :0.00000   Median :0.00000
                    Mean   :0.07797   Mean   :0.02501   Mean   :0.2159   Mean   :0.02052   Mean   :0.05237
                    3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
                    Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :4.00000   Max.   :1.00000
  Scholarship       Tuberculosis       Sms_Reminder      AwaitingTime
 Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :-398.00
 1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.: -20.00
 Median :0.0000   Median :0.00000   Median :1.0000   Median :  -8.00
 Mean   :0.0969   Mean   :0.00045   Mean   :0.5742   Mean   : -13.84
 3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:  -4.00
 Max.   :1.0000   Max.   :1.00000   Max.   :2.0000   Max.   :  -1.00
```

Data Wrangling

After doing a summary of the data set, I noticed a few variables that need to be cleaned up or modified to work for my analysis. Below are the steps I took to clean the dataset

1. Age: In the age variable, I noticed that there were a few ages listed as negative. Since ages can't be negative and because it was only a handful of records, I decided to keep them after transforming them by taking their absolute value.

    a. noshowraw1 %<>%
       mutate(Age=abs(Age))

    b. Below is a summary of the transformation

```
> summary(noshowraw1$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   19.00   38.00   37.81   56.00  113.00
```

2. Converting columns with binary values from integer to factor: Some of the variables in the dataset that had binary values were read into R as an integer instead of factor. These variables would be most useful to the analysis if they were seen as factors (1 or 0)
    a.

```
noshowraw1<-noshowraw
```

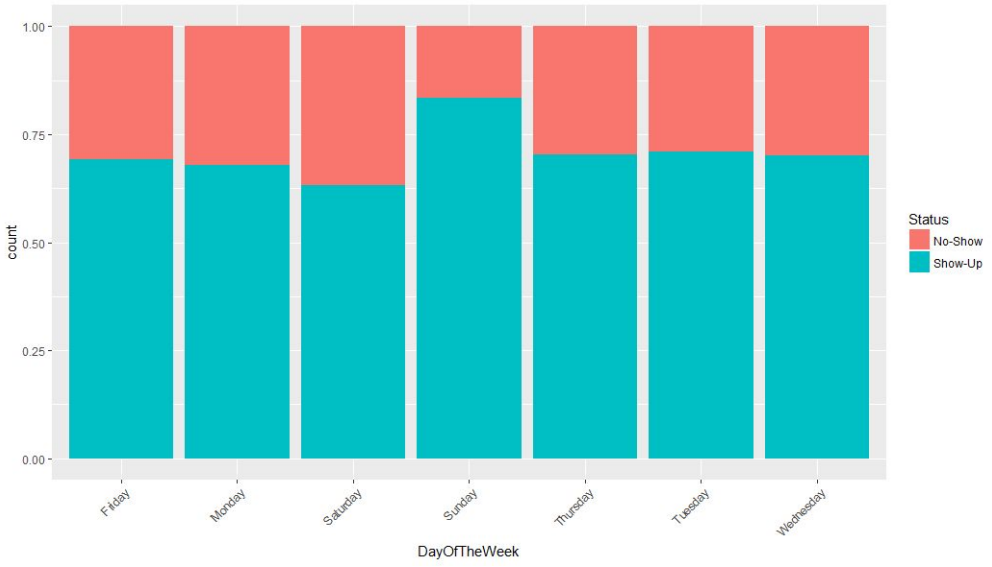
```
> str(noshowraw1)
Classes 'tbl_df', 'tbl' and 'data.frame':       300000 obs. of  15 variables:
 $ Age                   : int  19 24 4 5 38 5 46 4 20 51 ...
 $ Gender                : Factor w/ 2 levels "F","M": 2 1 1 2 2 1 1 1 1 1 ...
 $ AppointmentRegistration: POSIXct, format: "2014-12-16 14:46:25" "2015-08-18 07:01:26" "2014-02-17 12:53:46" "2014-07-23 17:02:11" ...
 $ ApointmentData        : POSIXct, format: "2015-01-14" "2015-08-19" "2014-02-18" "2014-08-07" ...
 $ DayOfTheWeek          : chr  "Wednesday" "Wednesday" "Tuesday" "Thursday" ...
 $ Status                : Factor w/ 2 levels "No-Show","Show-Up": 2 2 2 2 2 1 2 2 2 2 ...
 $ Diabetes              : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 2 ...
 $ Alcoolism             : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ HiperTension          : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 2 ...
 $ Handcap               : Factor w/ 5 levels "0","1","2","3",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Smokes                : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ Scholarship           : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...
 $ Tuberculosis          : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ Sms_Reminder          : int  0 0 0 1 1 1 1 1 0 1 ...
 $ AwaitingTime          : int  -29 -1 -1 -15 -6 -35 -18 -14 -14 -4 ...
```

    b. Below is the result of the transformation
3. Outliers: I checked the variables for outliers and noticed that for the most part the data did not have a lot of outliers to treat.

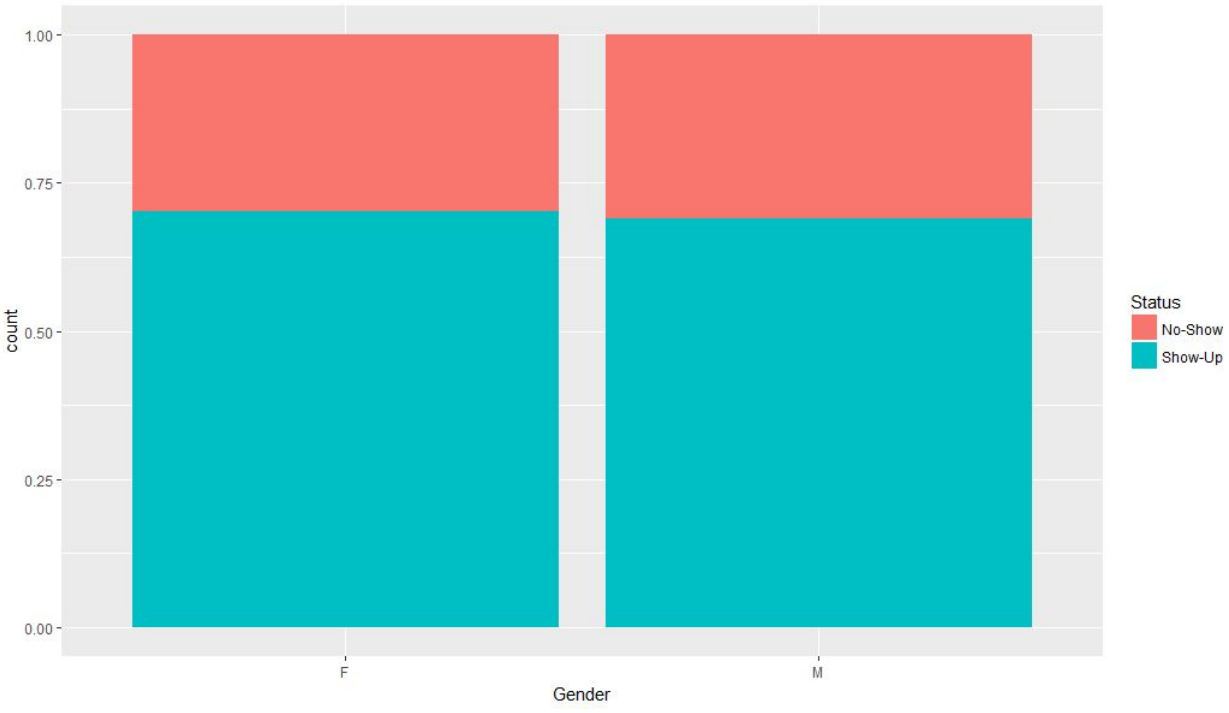
Exploratory Data Analysis

After cleaning up the variables that I needed to, I proceeded to explore the variables in the dataset to better understand them.

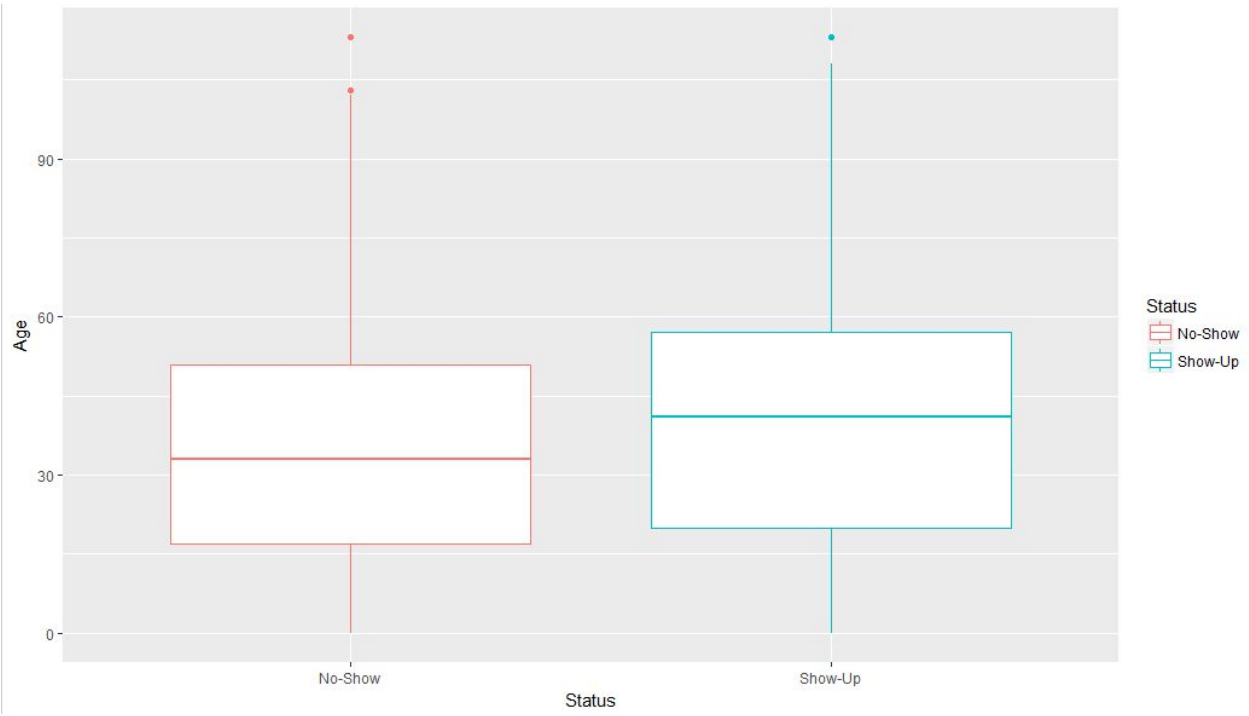
1. No shows by Day of the week



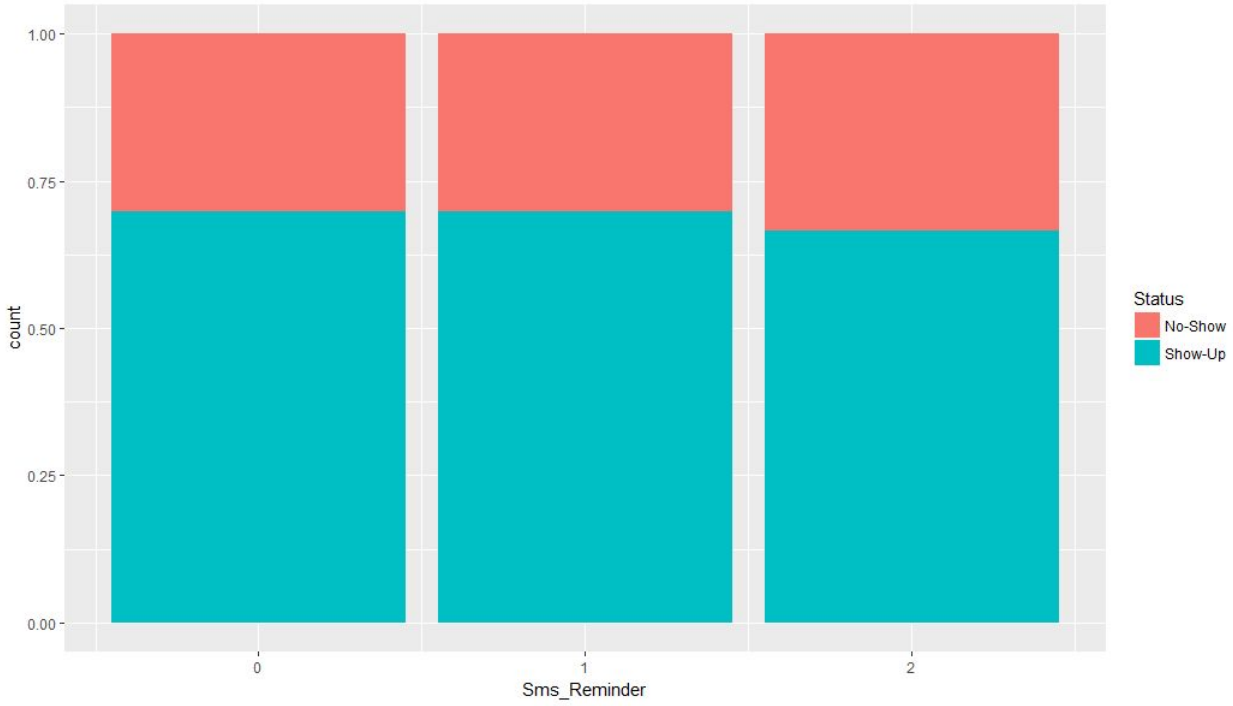
2. No shows by Gender



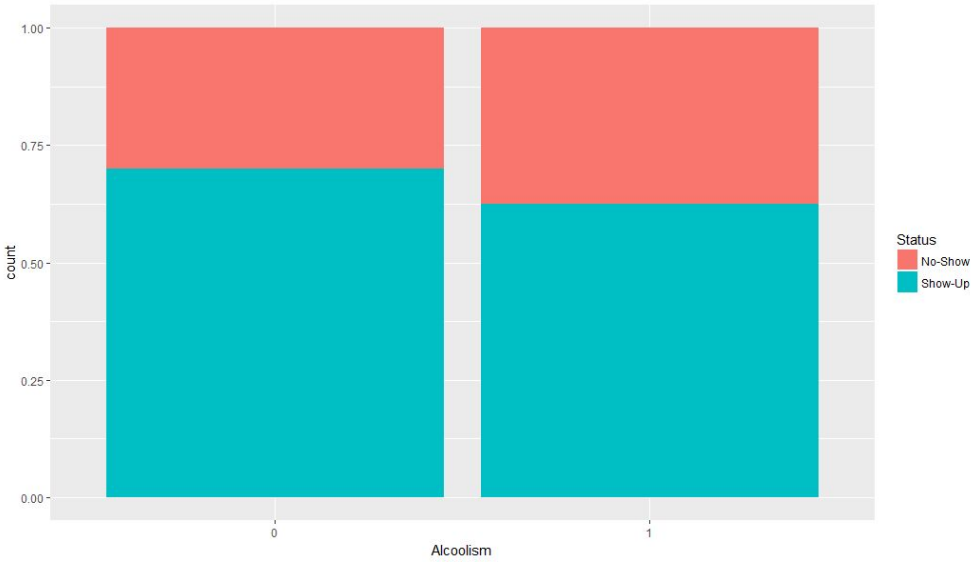
3. No Show by Age



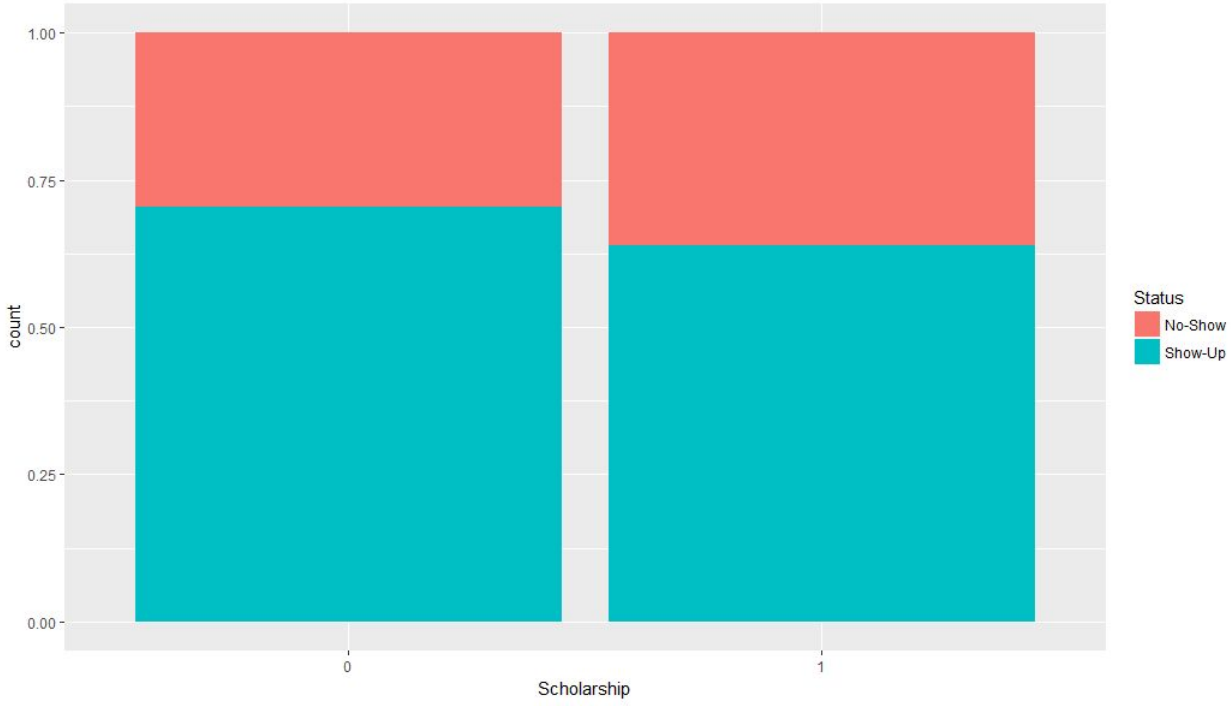
4. No shows by SMS Reminder



5. Effect of Alcoholism



6. Effect of Government assistance



Data Mining - Machine Learning

Since I was trying to predict what the probability was for a patient to not show-up, I decided to use Logistic regression as my algorithm.

1. Split data into train and test datasets

```
set.seed(100)
testnrow <- sample(nrow(noshowraw1),0.3*nrow(noshowraw1))
noshowraw1.train <- noshowraw1[-testnrow,]
noshowraw1.test <- noshowraw1[testnrow,]

nrow(noshowraw1.test)
nrow(noshowraw1.train)

summary(noshowraw1.test)
summary(noshowraw1.train)
```

```
> summary(noshowraw1.test)
      Age          Gender     AppointmentRegistration       ApointmentData
 Min.   :  0.00   F:60077   Min.   :2013-08-14 10:49:20   Min.   :2014-01-02 00:00:00
 1st Qu.: 19.00   M:29923   1st Qu.:2014-06-24 15:16:49   1st Qu.:2014-07-08 00:00:00
 Median : 38.00             Median :2014-12-05 07:21:02   Median :2014-12-17 00:00:00
 Mean   : 37.79             Mean   :2014-12-16 06:00:57   Mean   :2014-12-29 14:43:12
 3rd Qu.: 56.00             3rd Qu.:2015-06-12 07:17:18   3rd Qu.:2015-06-26 00:00:00
 Max.   :113.00             Max.   :2015-12-29 10:49:12   Max.   :2015-12-30 00:00:00
 DayOfTheWeek           Status       Diabetes  Alcoolism HiperTension Handcap   Smokes
 Length:90000       No-Show:27191   0:82923   0:87737   0:70437      0:88377   0:85314
 Class :character   Show-Up:62809   1: 7077   1: 2263   1:19563      1: 1484   1: 4686
 Mode  :character                                                    2:  124
                                                                     3:   12
                                                                     4:    3

 Scholarship Tuberculosis  Sms_Reminder     AwaitingTime
 0:81293     0:89950      Min.   :0.0000   Min.   :-349.00
 1: 8707     1:   50      1st Qu.:0.0000   1st Qu.: -20.00
                          Median :1.0000   Median :  -8.00
                          Mean   :0.5743   Mean   : -13.85
                          3rd Qu.:1.0000   3rd Qu.:  -4.00
                          Max.   :2.0000   Max.   :  -1.00
> summary(noshowraw1.train)
      Age          Gender     AppointmentRegistration       ApointmentData
 Min.   :  0.00   F:140428   Min.   :2013-05-29 15:14:11   Min.   :2014-01-02 00:00:00
 1st Qu.: 19.00   M: 69572   1st Qu.:2014-06-23 10:05:36   1st Qu.:2014-07-04 00:00:00
 Median : 38.00              Median :2014-12-03 09:38:17   Median :2014-12-16 00:00:00
 Mean   : 37.82              Mean   :2014-12-14 17:52:45   Mean   :2014-12-28 02:11:40
 3rd Qu.: 56.00              3rd Qu.:2015-06-10 19:00:24   3rd Qu.:2015-06-25 00:00:00
 Max.   :113.00              Max.   :2015-12-29 12:08:58   Max.   :2015-12-30 00:00:00
 DayOfTheWeek           Status       Diabetes   Alcoolism  HiperTension Handcap
 Length:210000      No-Show: 63540   0:193687   0:204760   0:164796     0:206026
 Class :character   Show-Up:146460   1: 16313   1:  5240   1: 45204     1:  3614
 Mode  :character                                                       2:   325
                                                                        3:    27
                                                                        4:     8

 Smokes      Scholarship Tuberculosis  Sms_Reminder     AwaitingTime
 0:198975    0:189638    0:209915     Min.   :0.0000   Min.   :-398.00
 1: 11025    1: 20362    1:    85     1st Qu.:0.0000   1st Qu.: -20.00
                                      Median :1.0000   Median :  -8.00
                                      Mean   :0.5741   Mean   : -13.84
                                      3rd Qu.:1.0000   3rd Qu.:  -4.00
                                      Max.   :2.0000   Max.   :  -1.00
```

2. Building the model

```
#building the model
glm.1 <- glm(Status~.,data=noshowraw1.train,family=binomial)
summary(glm.1)

glm1.probs <- predict(glm.1,noshowraw1.test,type="response")

summary(glm1.probs)
```

3. Analyzing the results/ Confusion Matrix

```
table(noshowraw1.test$Status,glm1.probs>0.6)

#          FALSE  TRUE
#No-Show    147 27044
#Show-Up    192 62617
```

Recommendations & Further Analysis

Based on my analysis of the data, it looks like the three factors that seem to influence No-shows are Age, Alcoholism and Government assistance. The younger the patient, the more likely they are to skip their appointment. And patient struggling with Alcoholism and that are on government subsidies tend to be more likely to be no-shows for their appointments.

However, these are all preliminary findings and there is more room to refine this analysis. I plan on doing more analysis on how SMS reminders affect no-shows given their age. I also want to dig deeper into any effects that might exist with the number of days between the creation of the appointment and the actual appointment.