# Outpatient 'No Show' Analysis

*James Skane*

*5/18/2017*

## Introduction

This project focused on outpatient appointment no-shows, specifically outpatient endoscopy procedures requiring anesthesia. Simulated data was derived from the following article: Prevalence and predictors of patient no-shows to outpatient endoscopic procedures scheduled with anesthesia. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4589132/

### Relationships being investigated:

Investigating the relationship between between particular behavioral and social determinents of health, and patient no-shows to outpatient endoscopic procedures that require anesthesia. The primary goal is to identify variables that have a statistically significant effect on patient no-shows, and quantify the magnitude and direction of these relationships. Relationships were assessed using chi-square, ANOVA, t-tests, and Logistic Regression. As stated previously, all was derived from the descriptive statistics contained in the following article: Prevalence and predictors of patient no-shows to outpatient endoscopic procedures scheduled with anesthesia.

---

### Dataset Description

The simulated data set includes 511 rows and 18 columns.

Below are the predictor variables I plan to utilize. The distribution information listed below reflects the entire patient population. Class specific distribution information can be found within the artcle's Table 1

## Predictor Variables:

***"age"***

- type=numeric, continous
- distribution=normal, rnorm(n = 511,mean = 55.4,sd = 11.1)
- no missing data

***"male"***

- type=binary factor, Levels: 0 1
- distribution=Bernoulli, rbinom(n = 511,prob= 0.554,size = 1)
- no missing data

***"race"***

- type=factor, Levels: asian, black, hisp, other, white
- distribution=Binomial, sample(x=c("white","black", "hisp", "asian", "other"), size=n, replace=TRUE, prob=c(0.298, 0.341, 0.168, 0.166, 0.027))
- no missing data

***"lang"***

- type=factor, Levels: asian, english, spanish
- distribution=Binomial, lang <- sample(x=c("english","spanish", "asian"), size=n, replace=TRUE,prob=c(.773,.11,.094))
- no missing data

***"immigrant"***

- type=binary factor, Levels: 0 1
- distribution=Bernoulli, rbinom(n = 511,prob = 0.322,size = 1)
- no missing data

***"employed"***

- type=binary factor, Levels: 0 1 -distribution=Bernoulli, rbinom(n = 511,prob = 0.196,size = 1)
- no missing data

***"homelessness"***

- type=binary factor, Levels: 0 1
- distribution=Bernoulli, rbinom(n = 511,prob = 0.121,size = 1)
- no missing data

***"substance" (Active Substance Abuse)***

- type=binary factor, Levels: 0 1
- distribution=Bernoulli,rbinom(n = 511,prob = 0.311,size = 1)
- no missing data

- definition = *medical records revealed self-report of active substance abuse* OR *positive drug toxicology test within 1 year of the pre-endoscopy GI clinic encounter.*

***"opiod_benzo" (Heavy use of prescription opioids or benzodiazepines)***

- type=binary factor, Levels: 0 1
- distribution=Bernoulli, rbinom(n = 511, prob = .327, size = 1)
- no missing data
- Heavy use of prescription opioids or benzodiazepines
- definition = *use of prescription opioids or benzodiazepines for treatment of chronic pain, substance abuse, or psychiatric illness that was determined to be a hindrance to adequate moderate sedation by the evaluating clinician during the pre-endoscopy GI clinic encounter*

***"psych" (History of mental illness)***

- type=binary factor, Levels: 0 1
- distribution=Bernoulli, rbinom(n = 511, prob = .382 ,size = 1)
- no missing data

***"insurance"***

- type=factor, Levels: Medical medicare uninsured
- distribution=Binomial,sample(x=c("uninsured","medicare", "Medical"), size=511, replace=TRUE, prob=c(155/n,132/n,224/n))
- no missing data

***"sympotmatic" (Patient symptoms were indication for procedure)***

- type=binary factor, Levels: 0 1
- distribution=Bernoulli,rbinom(n = 511, prob = .517 ,size = 1)
- no missing data
- context: Non-Symptomatic indications limited to asymptomatic iron deficiency anemia, positive fecal occult blood/fecal immunohistochemical test, history of adenomatous polyp or cancer, and family history of colon cancer.

***"preop_attend" (Patient attended preop Appt with Anesthesiologist)***

- type=binary factor, Levels: 0 1
- distribution=Bernoulli, rbinom(n = 511, prob = .438 ,size = 1)
- no missing data

***"past_endo_hx" (Surgerical History includes endoscopic procedures)***

- type=binary factor, Levels: 0 1
- distribution=Bernoulli, rbinom(n = 511, prob = .587 ,size = 1)
- no missing data

***"hx_noshow" (Patient has previously not shown up for Appt)***

- type=binary factor, Levels: 0 1
- distribution=Bernoulli, rbinom(n = 511, prob = .049 ,size = 1)
- no missing data

***"proc_type"***

- type=factor, Levels: Advanced Routine
- distribution=Bernoulli, sample(x=c("Routine","Advanced"), size=511, replace=TRUE, prob=c(404/511,107/511))
- no missing data

- definitions: *Routine* = esophagogastroduodenoscopy (EGD) and colonoscopy grouped as routine procedures. *Advanced* = endoscopic ultrasound (EUS), endoscopic retrograde cholangiopancreatography (ERCP), and single balloon-assisted enteroscopy grouped as advanced procedures.

***"ref_source" (Source of refferal)***

- type=factor, Levels: gi pcp special
- distribution=Binomial, sample(x=c("pcp","special", "gi"), size=n, replace=TRUE, prob=c(357/n,42/n,112/n))
- no missing data

***"wait_time" (time b/w preop appt & procedure appt measured in weeks)***

- type=continuous numeric
- distribution=normal, rnorm(n = N,mean = 10.9,sd = 6.5)
- no missing data

### Response Variable

***"no_show"***

- type=binary factor, Levels: 0 1,
- distribution=Bernoulli, rbinom(n = 511, prob = .27 ,size = 1)
- no missing data

---

## Data Simulation Process

First I identified the distributions of all variables included in the dataset. Following this I then used the descriptive figures provided in the article to simulate values for each variable. In the article, continuous data was presented as means with standard deviations, whereas categorical data were presented as numbers and proportions. Therefore the datatype of the variables assisted in determining what type of distribution function should be used when simulating values.

Given that the article provides different descriptive statistics for each patient class ('show' & 'no_show'), I generated each class seperately. This ensured that I preserved the characteristics of each type of patient during the simulation process. Below are the functions I used to accomplish simulation, using the means and standard deviations for continuous variables, and the n size and proportions given for categorical variables:

```
generateNO_ShowDataset <- function(N) {
  age <- rnorm(n = N,mean = 54.5,sd = 11.5)
  male <- rbinom(n = N,prob= 0.659,size = 1)
  race <- factor(sample(x=c("white","black", "hisp", "asian", "other"), size=N, replace=TRUE,
                        prob=c(.319,.464,.094,.094, 0.029)))
  lang <- factor(sample(x=c("english","spanish", "asian"), size=N, replace=TRUE,
                        prob=c(.884,.058,.044)))
  immigrant <- rbinom(n = N,prob = 0.177,size = 1)
  employed <- rbinom(n = N,prob = .086,size = 1)
  homelessness <- rbinom(n = N,prob = 0.188,size = 1)
  substance <- rbinom(n = N,prob = 0.493,size = 1)
  psych <- rbinom(n = N, prob = .355 ,size = 1)
  opiod_benzo <- rbinom(n = N, prob = .464, size = 1)

  preop_attend <- rbinom(n = N, prob = .319 ,size = 1)
  past_endo_hx <-rbinom(n = N, prob = .529 ,size = 1)
  hx_noshow <- rbinom(n = N, prob = .123 ,size = 1)
  proc_type <- factor(sample(x=c("Routine","Advanced"), size=N, replace=TRUE,
                             prob=c(.935,.065)))
  sympotmatic <- rbinom(n = N, prob = .471 ,size = 1)
  ref_source <- factor(sample(x=c("pcp","special", "gi"), size=N, replace=TRUE,
                              prob=c(.761,.073,.167)))
  wait_time <- rnorm(n = N,mean = 10.9,sd = 6.5)
  no_show <- rbinom(n = N, prob = 1 ,size = 1)
  data.frame(no_show,age,male,race,lang,immigrant,employed,homelessness,substance,opiod_benzo,psych,symp
}

generateShowDataset <- function(N) {
  age <- rnorm(n = N,mean = 55.7,sd = 10.9)
  male <- rbinom(n = N,prob= 0.544,size = 1)
  race <- sample(x=c("white","black", "hisp", "asian", "other"), size=N, replace=TRUE,
```

```
                prob=c(.29,.295,.196,.193, 0.01565558))
  lang <- sample(x=c("english","spanish", "asian"), size=N, replace=TRUE,
                prob=c(.732,.129,.113))
  immigrant <- rbinom(n = N,prob = 0.376,size = 1)
  employed <- rbinom(n = N,prob = 0.246,size = 1)
  homelessness <- rbinom(n = N,prob = 0.097,size = 1)
  substance <- rbinom(n = N,prob = 0.244,size = 1)
  psych <- rbinom(n = N, prob = .391 ,size = 1)
  opiod_benzo <- rbinom(n = N, prob = .276, size = 1)
  preop_attend <- rbinom(n = N, prob = .483 ,size = 1)
  past_endo_hx <-rbinom(n = N, prob = .609 ,size = 1)
  hx_noshow <- rbinom(n = N, prob = .021 ,size = 1)
  proc_type <- sample(x=c("Routine","Advanced"), size=N, replace=TRUE,
                      prob=c(.737,.263))
  sympotmatic <- rbinom(n = N, prob = .534 ,size = 1)
  ref_source <- sample(x=c("pcp","special", "gi"), size=N, replace=TRUE,
                       prob=c(.676,.086,.239))
  wait_time <- rnorm(n = N,mean = 8.7,sd = 6.2)
  no_show <- rbinom(n = N, prob = 0 ,size = 1)
  data.frame(no_show,age,male,race,lang,immigrant,employed,homelessness,substance,opiod_benzo,psych,sym
}
```

17 of the 18 variables were generated by these functions, however, the 'insurance' variable needed to be generated independently. When it was included in these functions, it was difficult to ensure only individuals with an age >= 65 were the only people labeled as being on Medicare.

To ensure only senior citizens were labeled as having Medicare I used an ifelse statement. This labeled everyone younger than 65 as having 'other' insurance, and those 65 or older as having 'Medicare' insurance. I then used the check categorical function to obtain the n size of the individuals who did not have Medicare. This subgroup n size was then used as the denominator to recalculate the proportions of the remaing insurance levels (Medical, and uninsured). This process is demonstrated by the code shown below:

```
set.seed(616)
N = 138
df_noshow <- generateNO_ShowDataset(N = 138)
df_noshow$insurance <- ifelse(df_noshow$age >= 65, 'MediCare', 'other')


check_categorical <- function(df_noshow, insurance) {
  resultset <- group_by(df_noshow, insurance)
  summarize(resultset,
            min.age = min(age,na.rm = T),
            avg.age = mean(age, na.rm=T),
            med.age = median(age, na.rm=T),
            max.age = max(age,na.rm = T),
            n_size = length(age) )
}


check_categorical(df_noshow, insurance)
```

```
## # A tibble: 2 x 6
##   insurance  min.age  avg.age  med.age  max.age n_size
##       <chr>    <dbl>    <dbl>    <dbl>    <dbl>  <int>
## 1  MediCare 65.45629 72.28182 71.43210 83.56685     29
```

```
## 2     other 21.86373 51.84197 51.02854 64.93001    109
```

```
N = nrow(filter(df_noshow, insurance=='other'))
others<- filter(df_noshow, insurance=='other')
insurance <- sample(x=c("uninsured","Medical"), size=N, replace=TRUE,
                    prob=c(0.3138075,0.6875872))



df_noshow[which(df_noshow$insurance == 'other'),'insurance'] <- insurance
check_categorical(df_noshow, insurance)
```

```
## # A tibble: 3 x 6
##   insurance  min.age  avg.age  med.age  max.age n_size
##       <chr>    <dbl>    <dbl>    <dbl>    <dbl>  <int>
## 1   Medical 21.86373 52.03786 52.30633 64.93001     79
## 2  MediCare 65.45629 72.28182 71.43210 83.56685     29
## 3 uninsured 37.98009 51.32612 48.91758 64.77413     30
```

```
df_show <- generateShowDataset(N=373)
df_show$insurance <- ifelse(df_show$age >= 65, 'MediCare', 'other')
insurance <- sample(x=c("uninsured","Medical"), size=373, replace=TRUE,
                    prob=c(.3677991,0.6174298))



resultset <- group_by(df_noshow, insurance)
summarize(resultset,
          min.age = min(age,na.rm = T),
          avg.age = mean(age, na.rm=T),
          med.age = median(age, na.rm=T),
          max.age = max(age,na.rm = T),
          n_size = length(age) )
```

```
## # A tibble: 3 x 6
##   insurance  min.age  avg.age  med.age  max.age n_size
##       <chr>    <dbl>    <dbl>    <dbl>    <dbl>  <int>
## 1   Medical 21.86373 52.03786 52.30633 64.93001     79
## 2  MediCare 65.45629 72.28182 71.43210 83.56685     29
## 3 uninsured 37.98009 51.32612 48.91758 64.77413     30
```

```
N = nrow(filter(df_show, insurance=='other'))
others<- filter(df_show, insurance=='other')
others_insurance <- sample(x=c("uninsured","Medical"), size=N, replace=TRUE,
                           prob=c(0.4420772,0.5565912))

df_show[which(df_show$insurance == 'other'),'insurance'] <- others_insurance
```

I then created my final data set by merging these data frames.

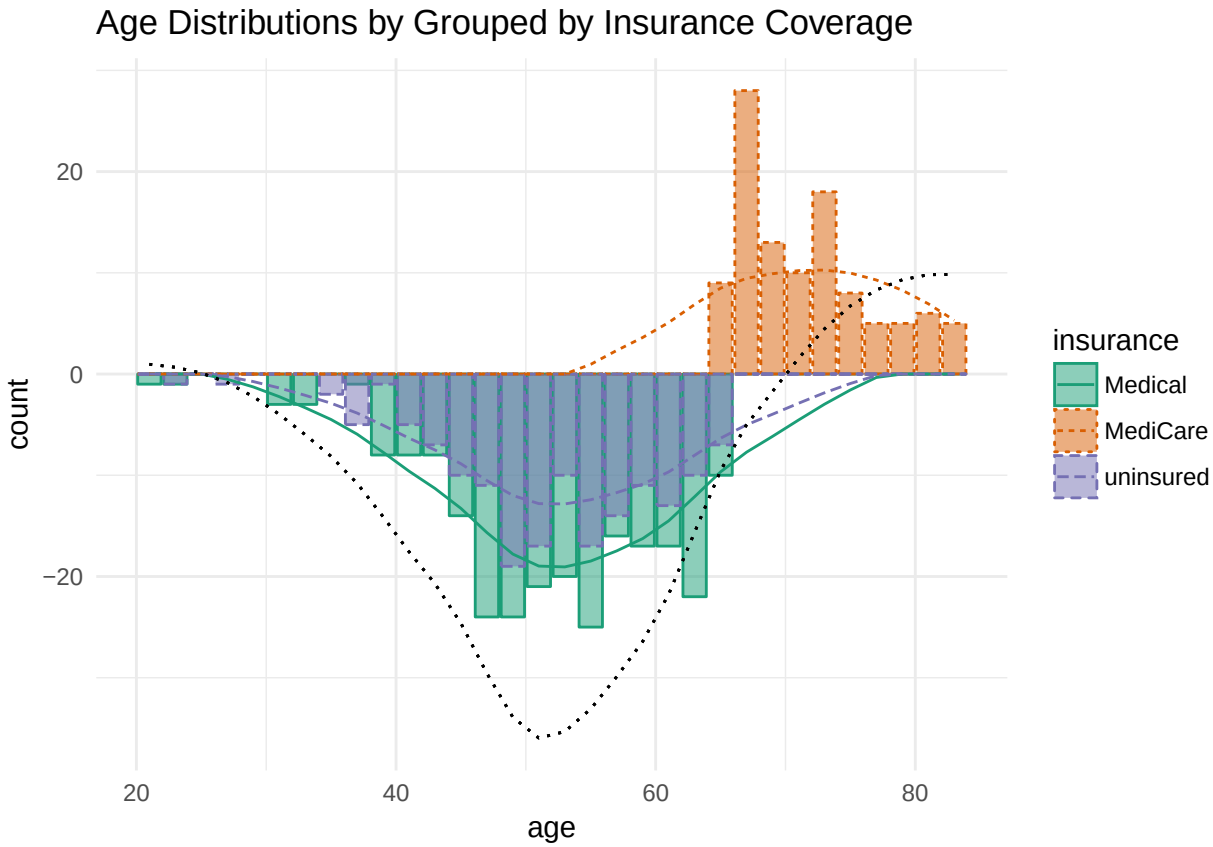
```
# Merged Datasets
df<-merge(df_show, df_noshow,all.x = T,all.y = T)

head(df)
```

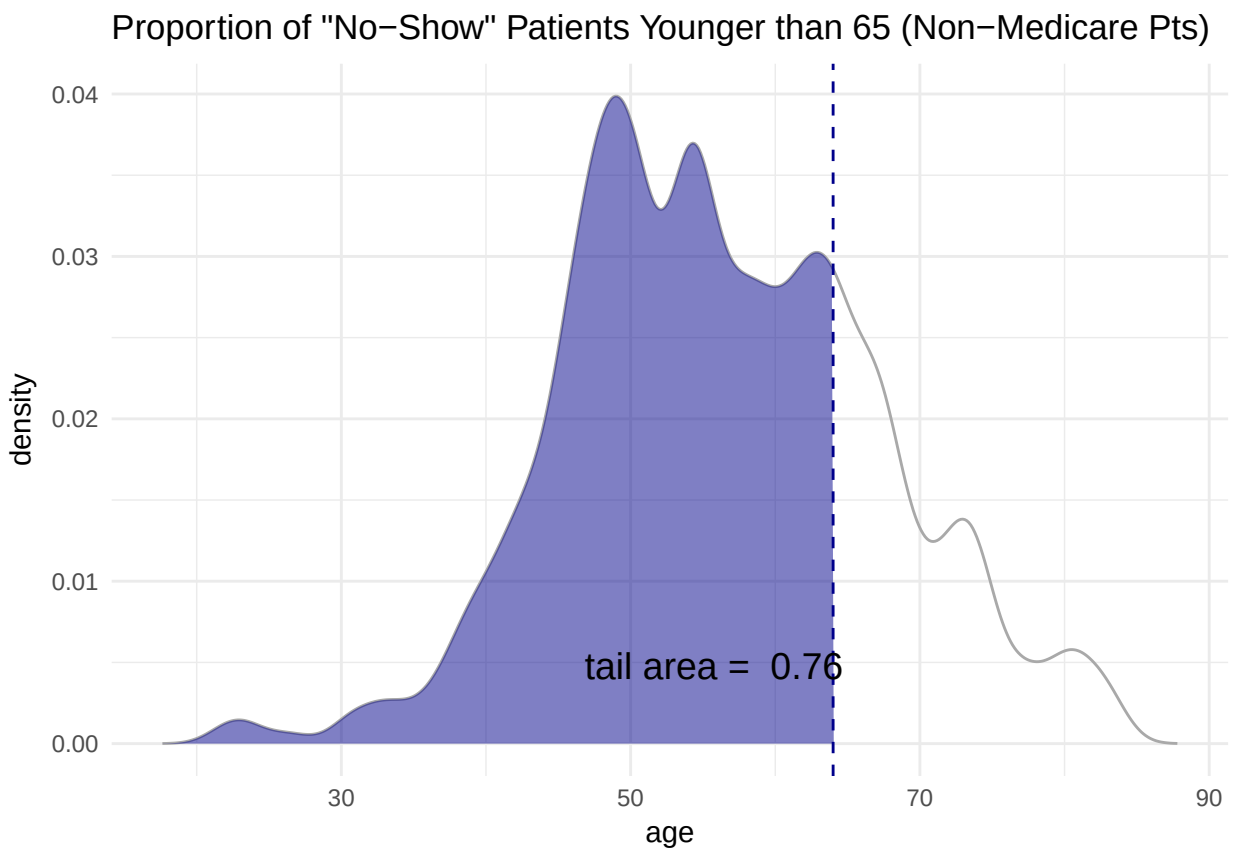
```
##   no_show      age male  race    lang immigrant employed homelessness
## 1       0 23.73438    0 black english         1        1             0
```

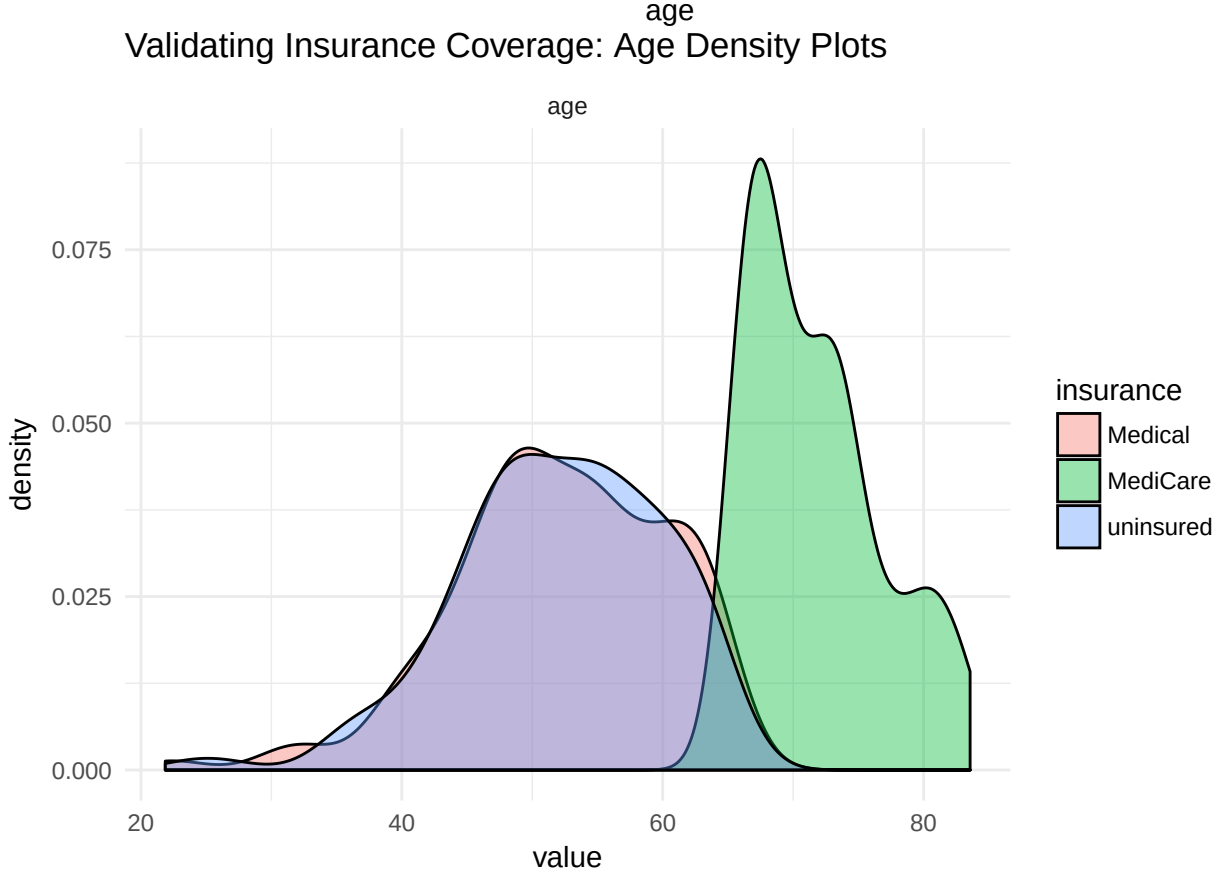
```
## 2       0 26.47902    1 white english         0        0             1
## 3       0 30.48444    1 asian   asian         0        1             0
## 4       0 30.48851    0 white english         1        0             1
## 5       0 32.91458    0  hisp spanish         0        0             0
## 6       0 33.79053    1  hisp english         0        0             0
##   substance opiod_benzo psych sympotmatic preop_attend past_endo_hx
## 1         0           0     0           1            1            0
## 2         0           0     0           1            0            1
## 3         0           0     0           1            0            0
## 4         0           0     0           0            1            1
## 5         1           1     0           0            1            1
## 6         0           0     1           1            0            1
##   hx_noshow proc_type ref_source wait_time insurance
## 1         0   Routine        pcp 10.135932 uninsured
## 2         0   Routine         gi  3.407263 uninsured
## 3         0   Routine         gi  9.014334   Medical
## 4         0   Routine        pcp 16.153565   Medical
## 5         0  Advanced        pcp  8.274184   Medical
## 6         0   Routine        pcp 15.168498   Medical
```

---

### EDA & Validating Data Simulation Through Plots

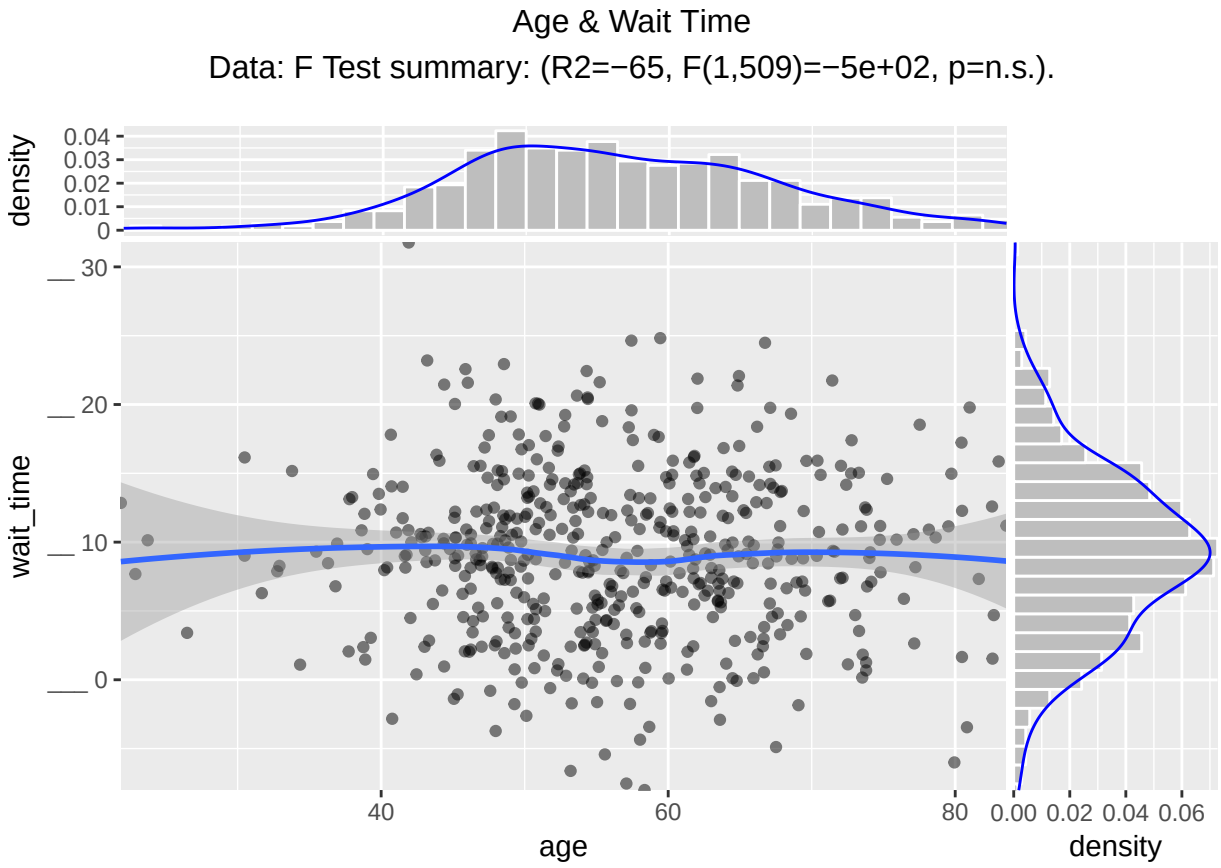



Proportion of "No−Show" Patients Younger than 65 (Non−Medicare Pts)
0.04
0.03
0.02
0.01
0.00
density
tail area = 0.76
30
50
70
90
age

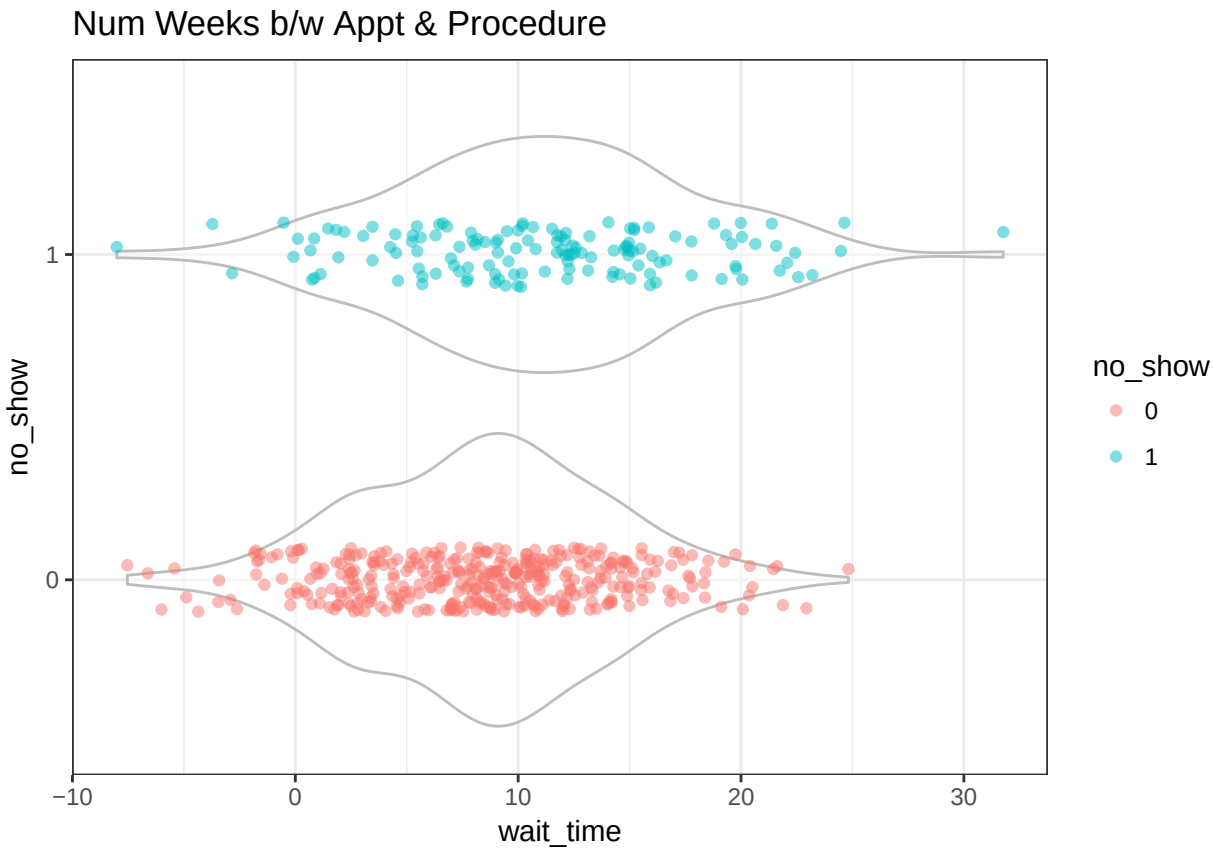

Validating Insurance Coverage: Age Density Plots
age
0.075
0.050
0.025
0.000
density
20
40
60
80
value
insurance
Medical
MediCare
uninsured

```
ScatterHist(df, "age", "wait_time", title="Age & Wait Time")
```

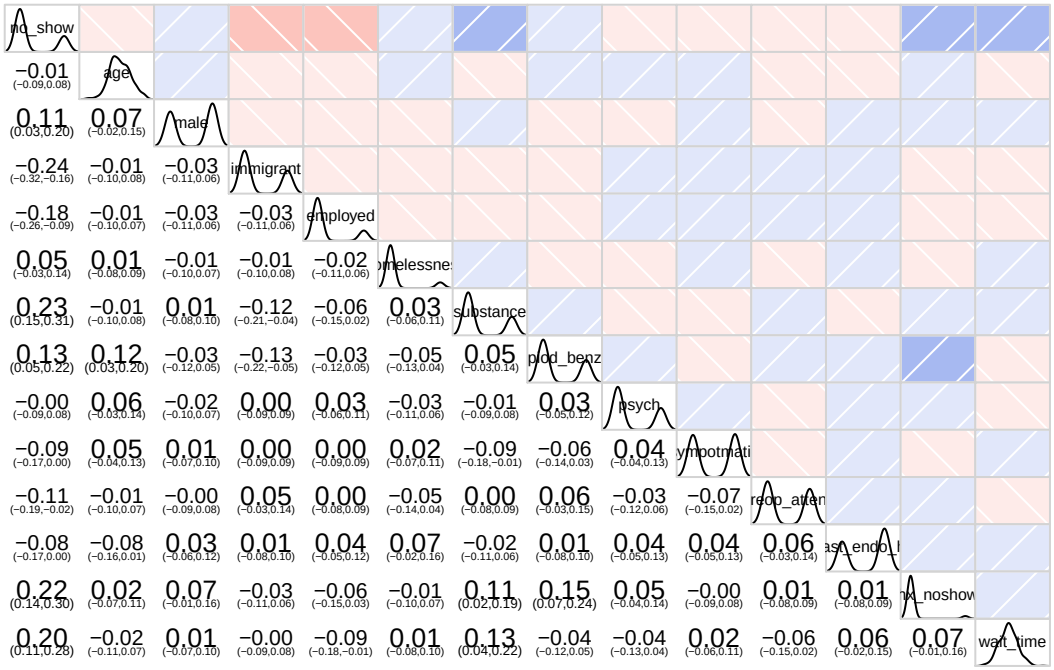
```
## `geom_smooth()` using method = 'loess'
## `geom_smooth()` using method = 'loess'
```




```
# Shows no_shows tend to wait longer between preop appointments and actual endoscopic procedures.
df$no_show <- as.factor(df$no_show)
g<-ggplot(df, aes(x=no_show, y=wait_time))
g+geom_violin(alpha=0.5, color="gray")+geom_jitter(alpha=0.5, aes(color=no_show),position = position_ji
```

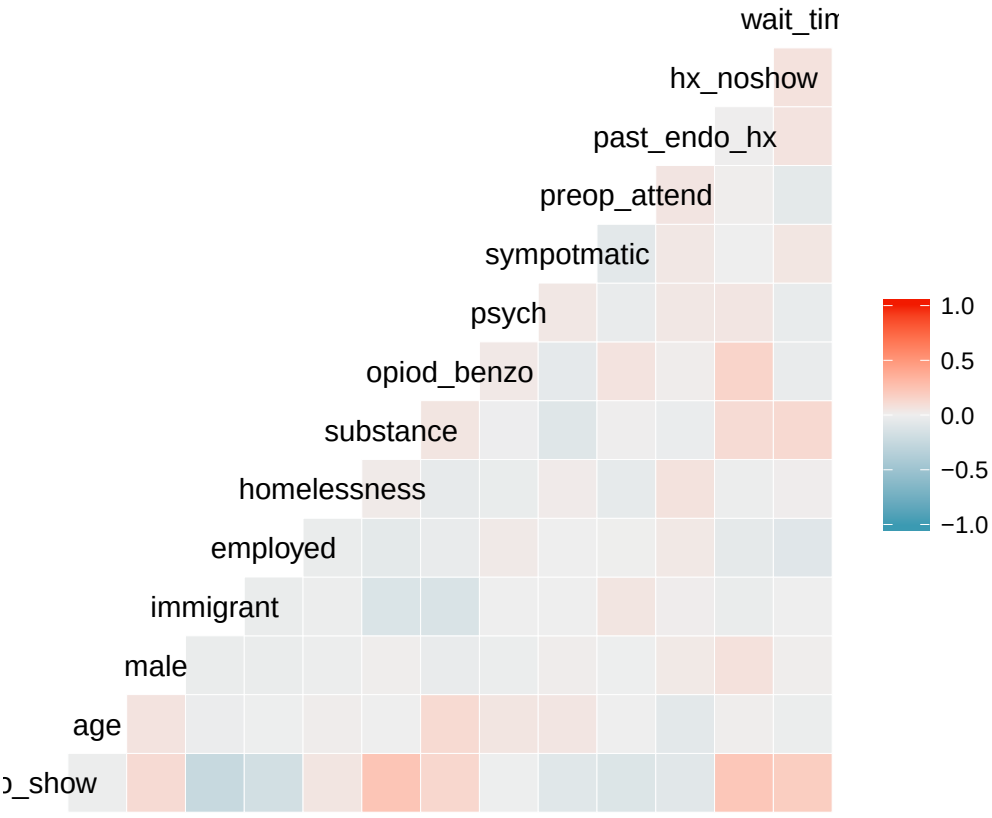



```
# Gives an overall summary of the relationships that exist among the variables in the dataset.
df$no_show <- as.numeric(df$no_show)

corrgram(df,lower.panel=panel.conf,
         upper.panel=panel.shade,
         diag.panel=panel.density)
```




```
ggcorr(df, method = c("all.obs", "spearman"),label_size = 1)
```

```
## Warning in ggcorr(df, method = c("all.obs", "spearman"), label_size = 1):
## data in column(s) 'race', 'lang', 'proc_type', 'ref_source', 'insurance'
## are not numeric and were ignored
```



## Bivariate Analysis

### Categorical

#### Binary - Chi-square test

Chi-square test was used to identify significant binary variables. If the chi-square statistic is less than 0.05, then we reject the hypothesis that the predictor variable is an independent factor in patient no shows. Each binary variable was assessed using the function below:

```
#If the chi-square statistic is less than 0.05, then we reject the hypothesis that the predictor variab

get_chisq <- function(y,x) {
  test<-table(y, x)
  rownames(test) <- c('Show', 'No-Show')
  chisq.test(test)
}

#Significant binary predictors (NOT independent factor in determing whether a patient will not show up
get_chisq(df$no_show, df$substance)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  test
## X-squared = 26.958, df = 1, p-value = 2.08e-07
```

```
get_chisq(df$no_show, df$employed)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  test
## X-squared = 14.753, df = 1, p-value = 0.0001226
```

```
get_chisq(df$no_show, df$opiod_benzo)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  test
## X-squared = 8.58, df = 1, p-value = 0.003399
```

```
get_chisq(df$no_show, df$preop_attend)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  test
## X-squared = 5.6598, df = 1, p-value = 0.01736
```

```
get_chisq(df$no_show, df$hx_noshow)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  test
## X-squared = 23.339, df = 1, p-value = 1.358e-06
```

```r
get_chisq(df$no_show, ifelse(df$proc_type=='Advanced',1,0))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  test
## X-squared = 22.56, df = 1, p-value = 2.036e-06
```

```r
#Other variables assessed and their chisq results (were later identified as insignificant):

#We reject the null hypothesis that being male is an independent factor in not showing up to an outpati
get_chisq(df$no_show, df$male)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  test
## X-squared = 6.1159, df = 1, p-value = 0.0134
```

```r
#We reject the null hypothesis that being an immagrant is an independent factor in not showing up to an
get_chisq(df$no_show, df$immigrant)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  test
## X-squared = 28.595, df = 1, p-value = 8.922e-08
```

```r
#We fail to reject the null hypothesis that being an immagrant is an independent factor in not showing
get_chisq(df$no_show, df$homelessness)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  test
## X-squared = 1.041, df = 1, p-value = 0.3076
```

```r
#We fail to reject the null hypothesis that having a history of mental illness is an independent factor
get_chisq(df$no_show, df$psych)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  test
## X-squared = 8.162e-30, df = 1, p-value = 1
```

```r
#We fail to reject the null hypothesis that being symptomatic is an independent factor in not showing u
get_chisq(df$no_show, df$sympotmatic)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  test
## X-squared = 3.3432, df = 1, p-value = 0.06748
```

```r
#We fail to reject the null hypothesis that having a history of endoscopic procedures is an independent
get_chisq(df$no_show, df$past_endo_hx)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  test
## X-squared = 3.1669, df = 1, p-value = 0.07514
```

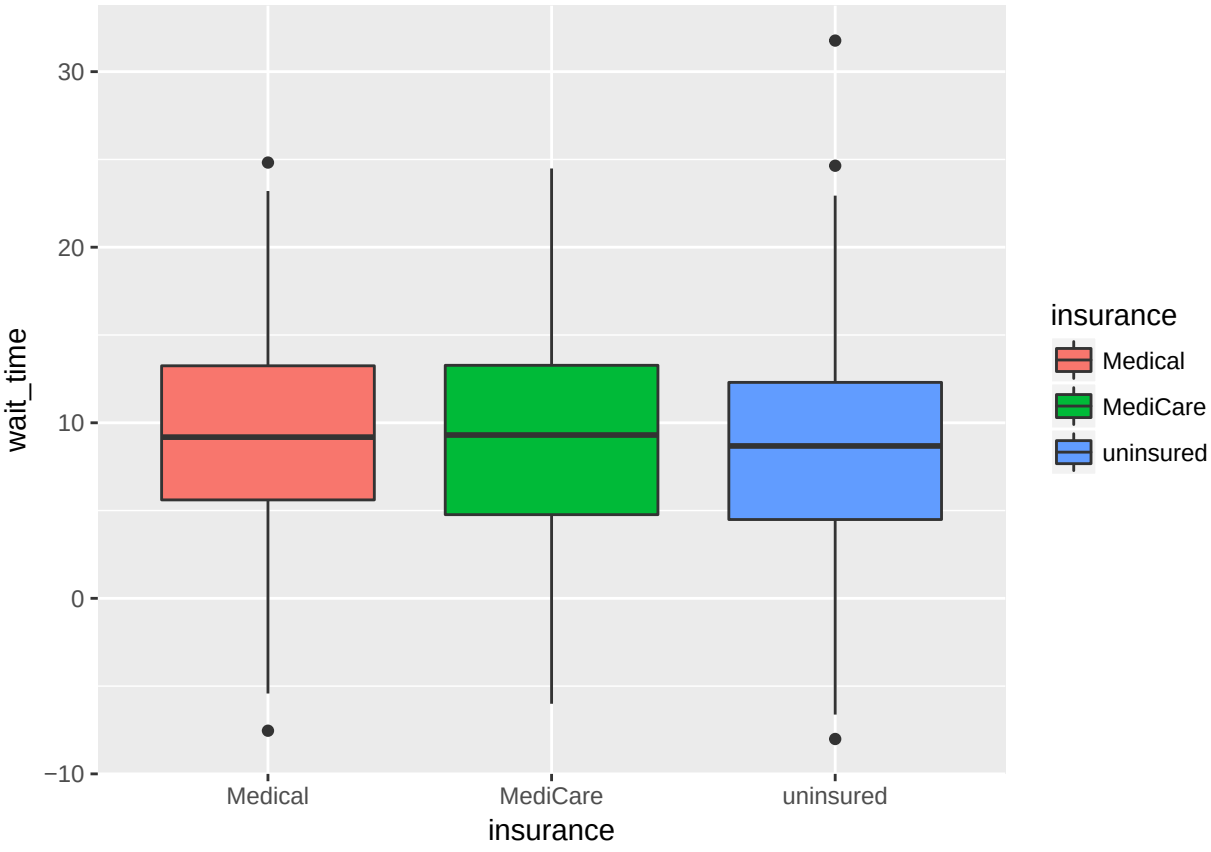
### Non-Binary Categorical - ANOVA

Anova test was used to compare the mean value of continuous variables among different groups. This test was only used to compare groups if there was more than two possible groups. A signicant result from this test does not signify that all of the means are significantly different. Instead, it only signifies that one of the means is significantly different from one of the other means.

```
#The mean wait time between preop appointments and outpatient endo procedures are NOT significantly dif
summary(aov(wait_time ~ insurance, data=df))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## insurance     2     55   27.57   0.765  0.466
## Residuals   508  18322   36.07
```

```
qplot(insurance, wait_time, data=df, geom="boxplot", fill=insurance)
```

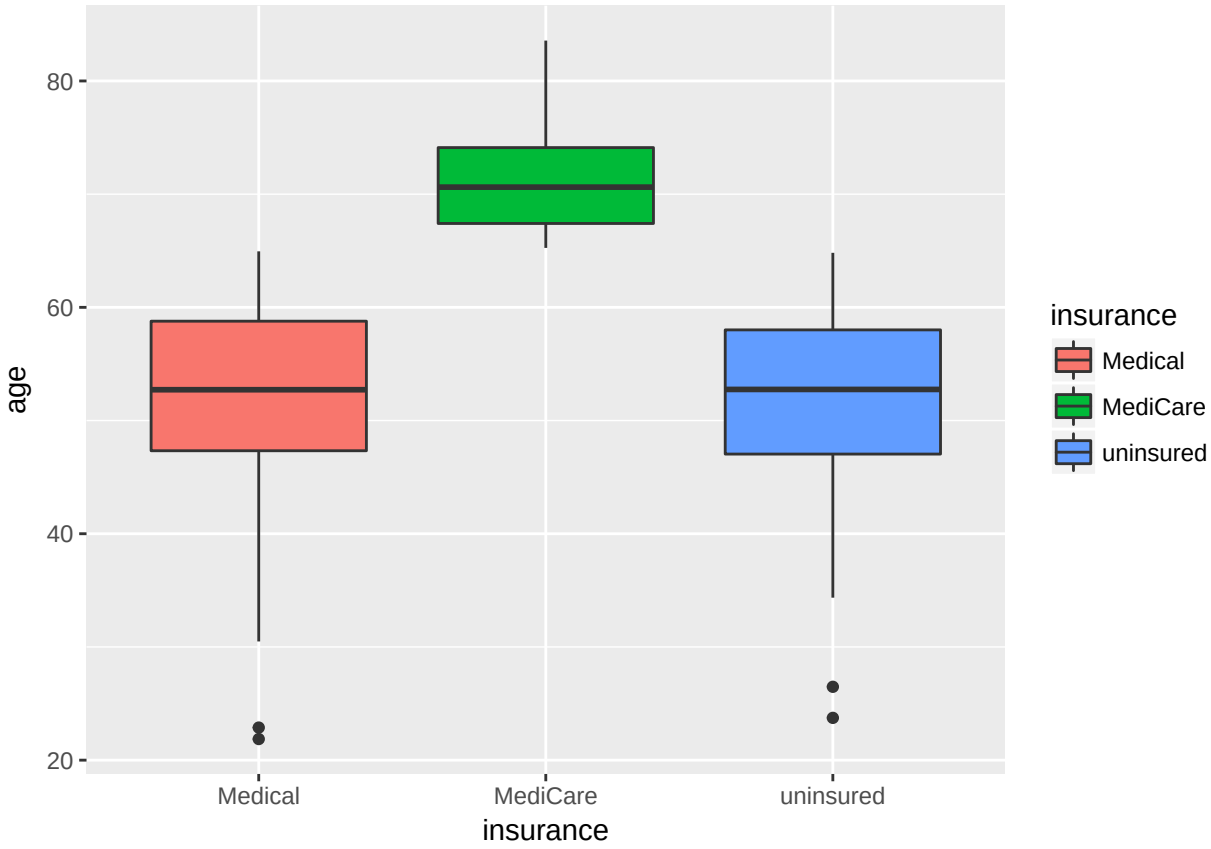



```
#At least one of the insurance groups have a significantly different mean age. This is expected given t
summary(aov(age ~ insurance, data=df))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## insurance     2  32204   16102   286.1 <2e-16 ***
## Residuals   508  28594      56
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

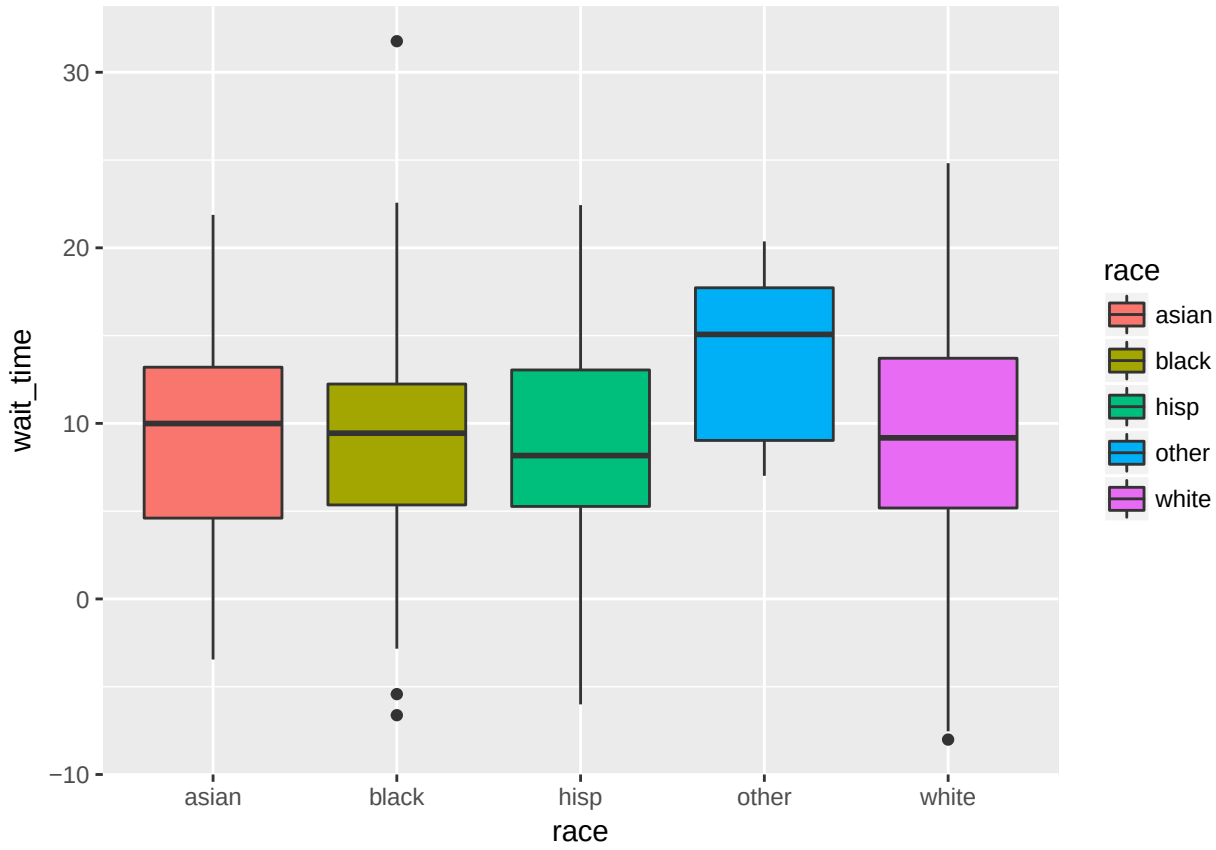
```
qplot(insurance, age, data=df, geom="boxplot", fill=insurance)
```




```
#The mean wait time between preop appointments and outpatient endo procedures are NOT significantly dif
summary(aov(wait_time ~ race, data=df))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## race          4    245   61.34   1.712  0.146
## Residuals   506  18132   35.83
```

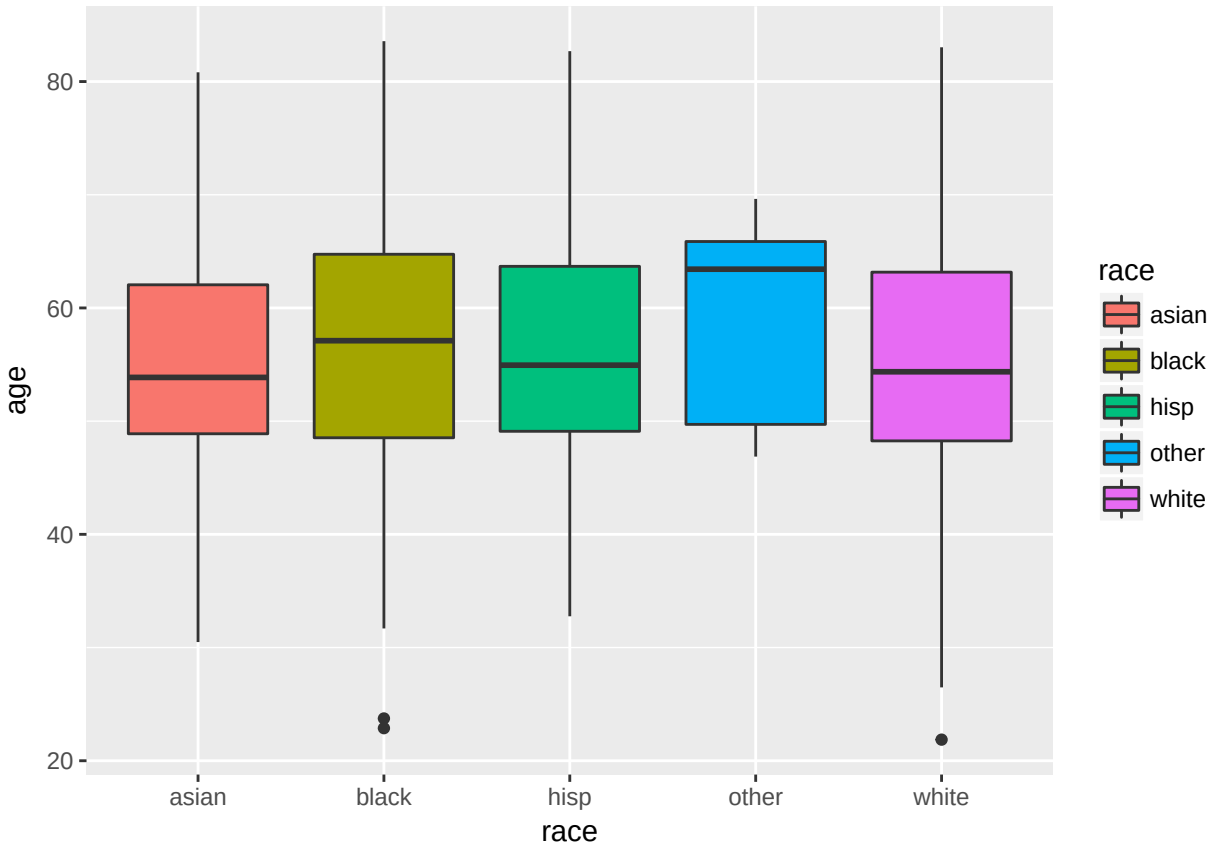
```
qplot(race, wait_time, data=df, geom="boxplot", fill=race)
```




```
#The mean age is NOT significantly different among diferent racial groups.
summary(aov(age ~ race, data=df))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## race          4    339   84.64   0.708  0.586
## Residuals   506  60460  119.49
```

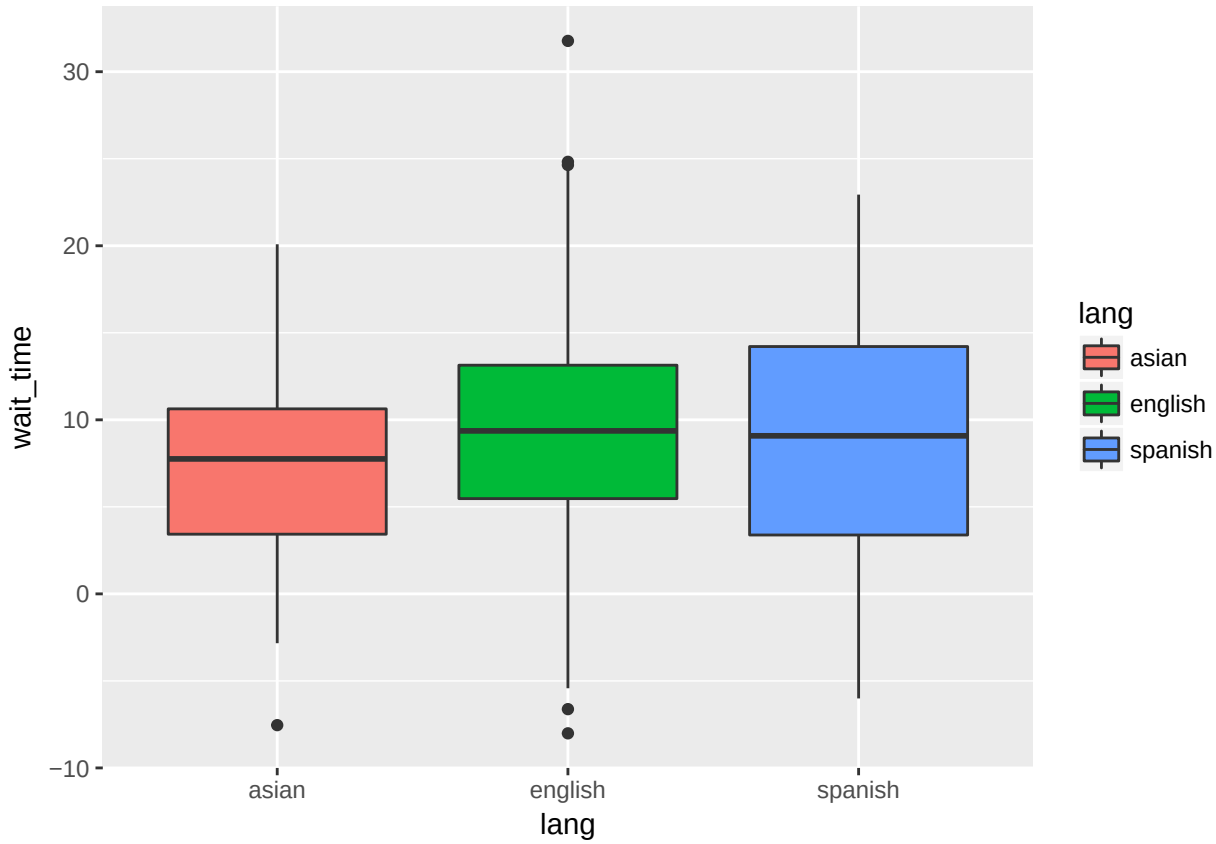
```
qplot(race, age, data=df, geom="boxplot", fill=race)
```




```
#The mean wait time between preop appointments and outpatient endo procedures are NOT significantly dif
summary(aov(wait_time ~ lang, data=df))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## lang          2    183   91.28   2.549 0.0792 .
## Residuals   508  18194   35.82
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

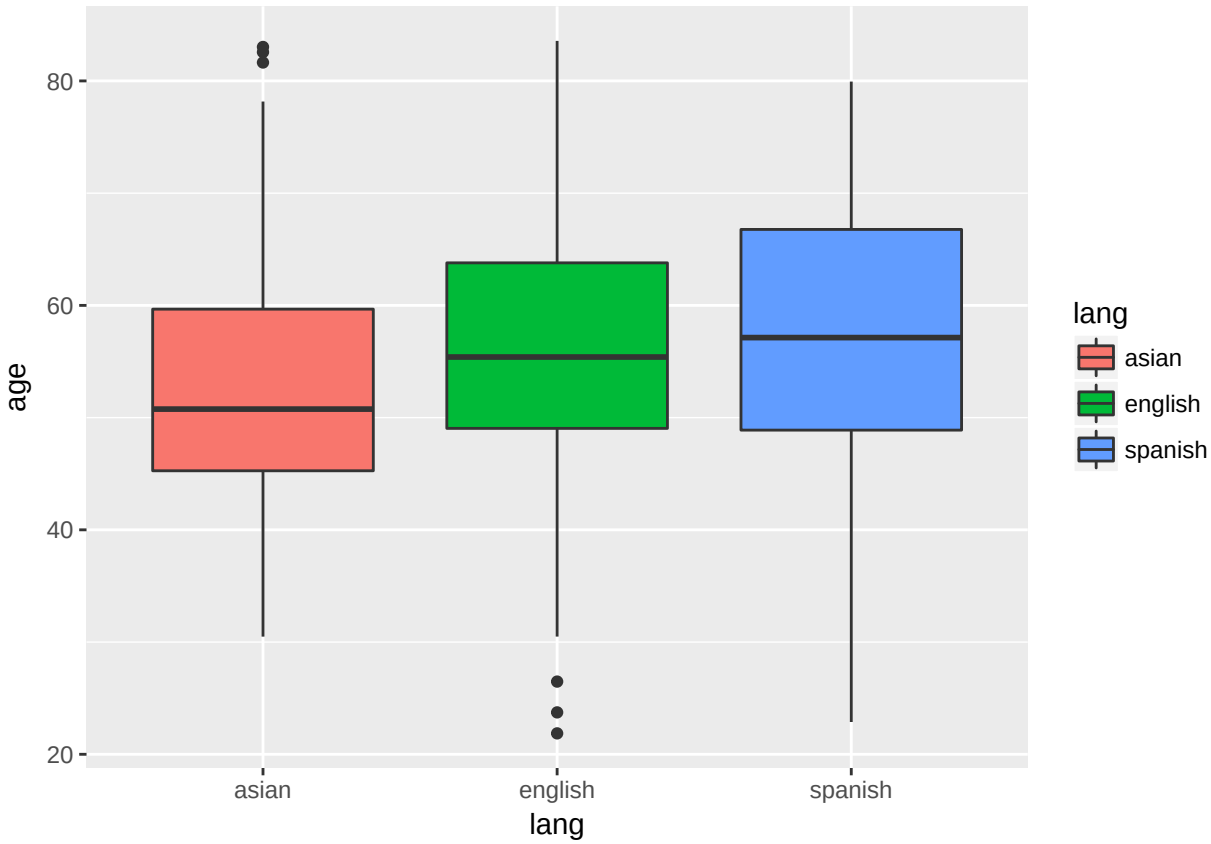
```
qplot(lang, wait_time, data=df, geom="boxplot", fill=lang)
```




```
#The mean age is NOT significantly different among diferent language groups.
summary(aov(age ~ lang, data=df))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## lang          2    393   196.5   1.653  0.193
## Residuals   508  60405   118.9
```

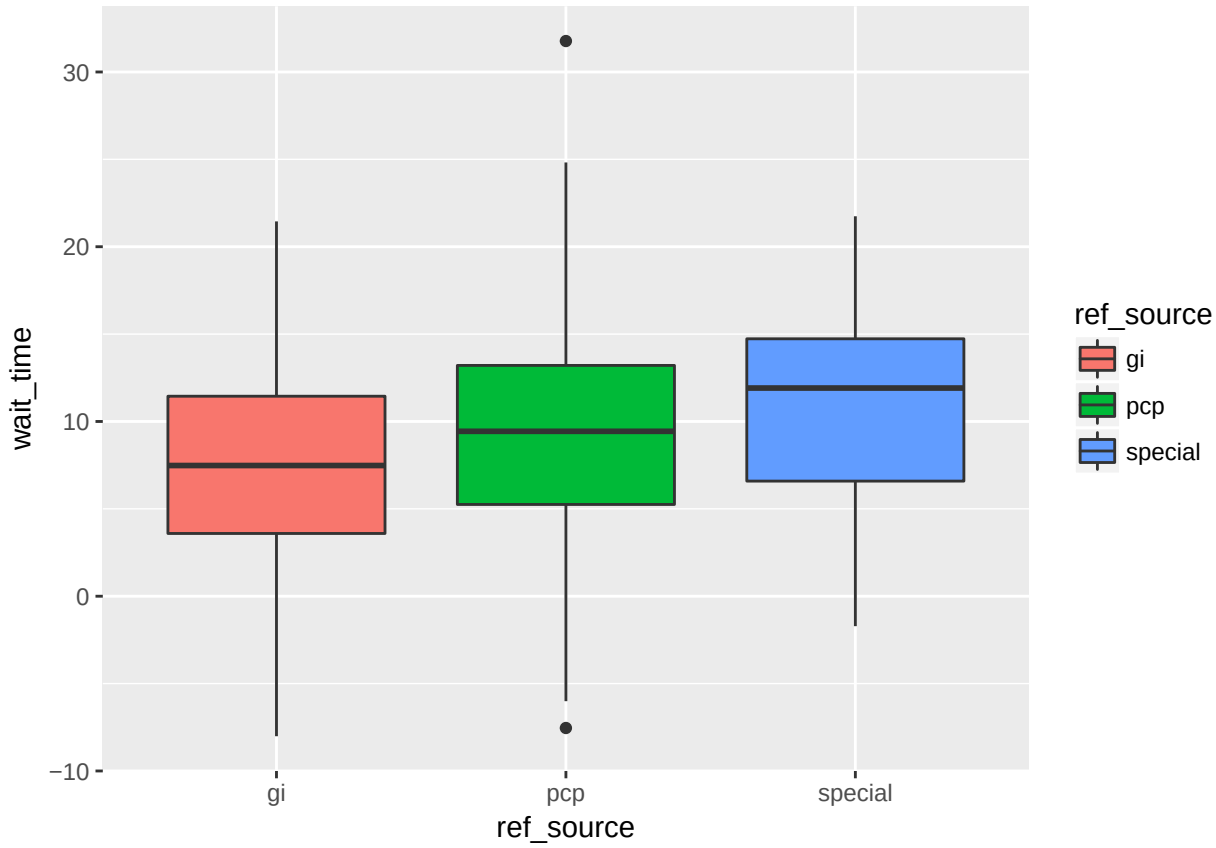
```
qplot(lang, age, data=df, geom="boxplot", fill=lang)
```




```
#The mean wait time between preop appointments and outpatient endo procedures IS significantly differen
summary(aov(wait_time ~ ref_source, data=df))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## ref_source    2    315  157.28   4.424 0.0125 *
## Residuals   508  18062   35.56
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

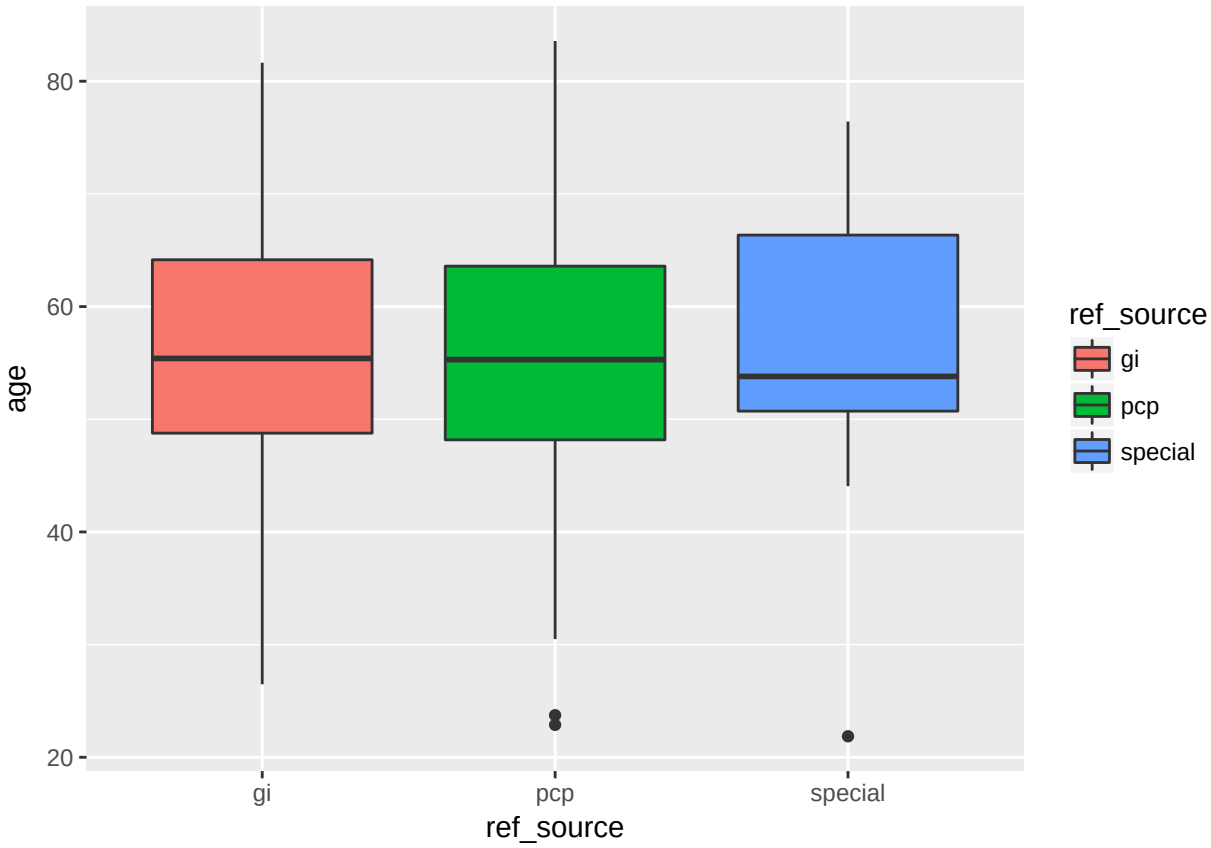
```
qplot(ref_source, wait_time, data=df, geom="boxplot", fill=ref_source)
```




```
#The mean age is NOT significantly different among diferent referral groups.
summary(aov(age ~ ref_source, data=df))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## ref_source    2     47   23.28   0.195  0.823
## Residuals   508  60752  119.59
```

```
qplot(ref_source, age, data=df, geom="boxplot", fill=ref_source)
```




**Continuous - t-tests**

Next I determine if the mean age and wait time is significantly different among the two patient groups of interest (shows, and no_shows). These groups are compared using an independent samples t-test. A significant result in this test signifies that the mean value for the continuous variable is significantly different among the two groups.

```
#The average wait time among patients that show up to outpatient endoscopic procedures requiring anesth
t.test(wait_time ~ no_show, data=df, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  wait_time by no_show
## t = -4.5706, df = 509, p-value = 6.111e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.834619 -1.529085
## sample estimates:
## mean in group 1 mean in group 2
##        8.425402       11.107255
```

```
#The average age among patients that show up to outpatient endoscopic procedures requiring anesthesia a
t.test(age ~ no_show, data=df, var.equal=TRUE)
```

```
##
##  Two Sample t-test
```

```
##
## data:  age by no_show
## t = 0.18676, df = 509, p-value = 0.8519
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.935924  2.342659
## sample estimates:
## mean in group 1 mean in group 2
##        56.34067        56.13730
```

## Correlated Variables

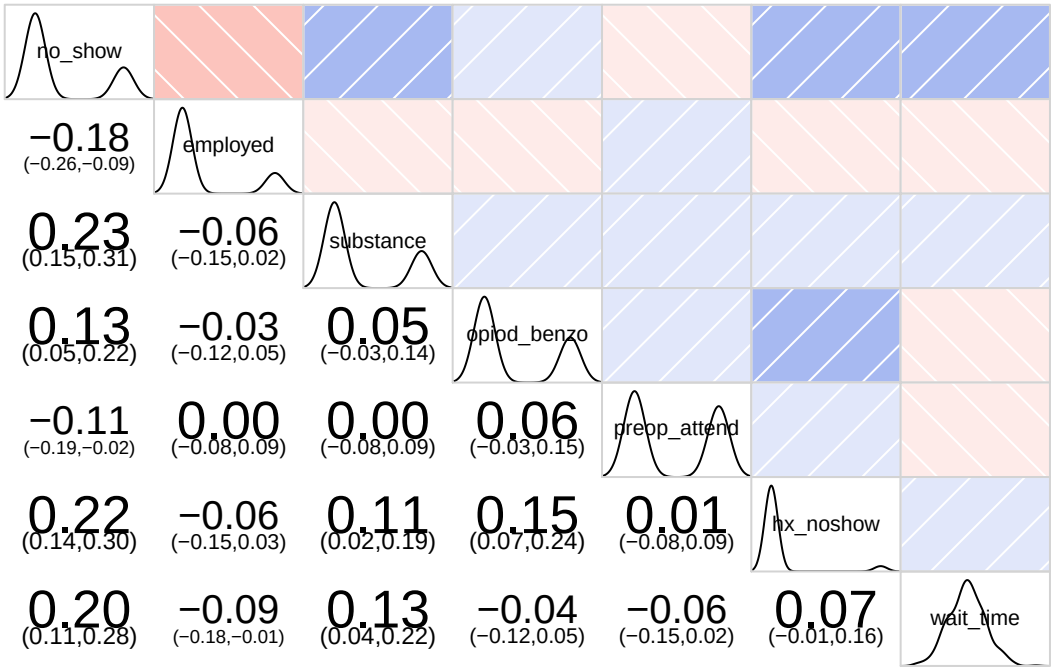
The response variable no-show is correlated with the following variables: - employed - substance - opiod_benzo - preop_attend - hx_noshow - proc_type - wait_time (weeks)

These correlated variables can be found within the artcles Table 2: Multivariable logistic regression of predictors The analysis and plots above also reiterate these findings, and show no multicolinearity or confounding variables in the final data set.

```
df$no_show <- as.numeric(df$no_show)

slim_df <- df[,c('no_show','employed','substance',
                 'opiod_benzo','preop_attend','hx_noshow',
                 'proc_type','wait_time')]

resultset <- group_by(slim_df, no_show)
summarize(resultset,
          emp = mean(employed,na.rm = T),
          substance = mean(substance, na.rm=T),
          opiod_benzo = mean(opiod_benzo, na.rm=T),
          preop_att = mean(preop_attend,na.rm = T),
          hx_noshow = mean(hx_noshow,na.rm = T),
          wait_time = mean(wait_time,na.rm = T))
```

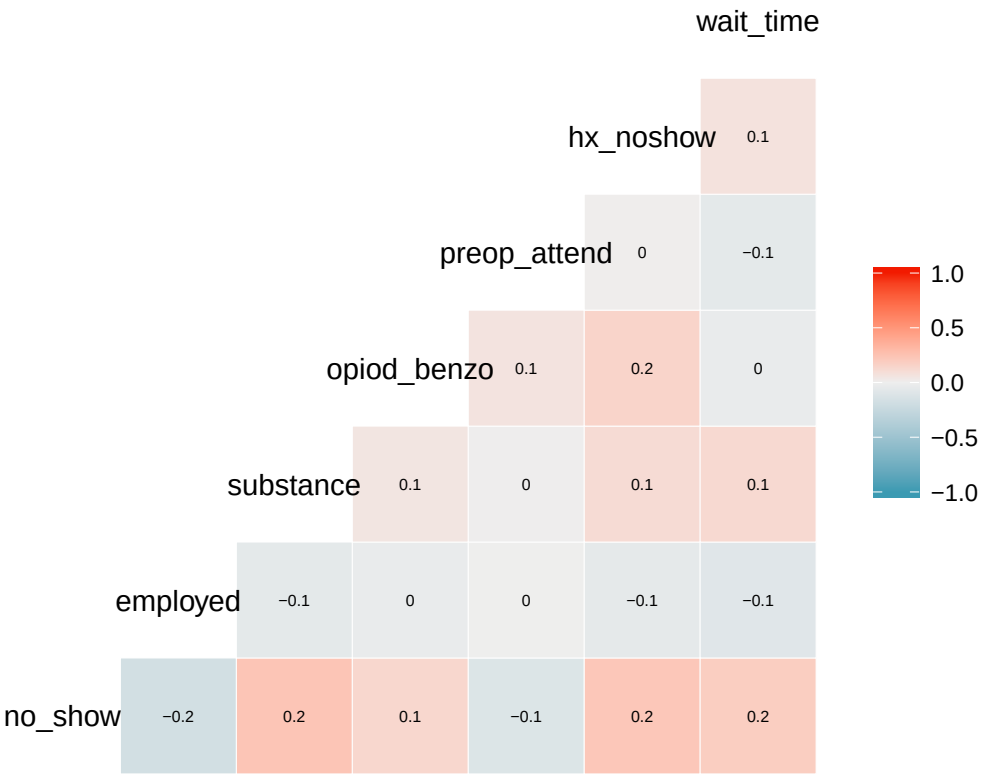
```
## # A tibble: 2 x 7
##   no_show        emp substance opiod_benzo preop_att  hx_noshow wait_time
##     <dbl>      <dbl>     <dbl>       <dbl>     <dbl>      <dbl>     <dbl>
## 1       1 0.23592493 0.2359249   0.3056300 0.4852547 0.02680965  8.425402
## 2       2 0.07971014 0.4782609   0.4492754 0.3623188 0.14492754 11.107255
```

```
corrgram(slim_df,lower.panel=panel.conf,
         upper.panel=panel.shade,
         diag.panel=panel.density)
```




```
ggcorr(slim_df, method = c("all.obs", "spearman"),label_size = 2,label = T)
```

```
## Warning in ggcorr(slim_df, method = c("all.obs", "spearman"), label_size =
## 2, : data in column(s) 'proc_type' are not numeric and were ignored
```



## Comparison with Original Article

Below uses the epi package to calculate the Odds Ratio for every predictor variable use. All results fall within the 95% CI published in the article verifying the simulated data correctly emulates the data used in the research article.

```
get_or_and_corr <- function(x,label_x1,label_x0) {
  test<-table(df$no_show, x)
  rownames(test) <- c('Show', 'No-Show')
  names(test) <- c(label_x1,label_x0)
  print(epi.2by2(test))
  cor.test(df$no_show, x)
}
df$no_show <- as.numeric(df$no_show)
get_or_and_corr(df$employed,'unemployed','employed' )
```

```
##              Outcome +    Outcome -      Total        Inc risk *
## Exposed +          285           88        373              76.4
## Exposed -          127           11        138              92.0
## Total              412           99        511              80.6
##                  Odds
## Exposed +        3.24
## Exposed -       11.55
## Total            4.16
##
## Point estimates and 95 % CIs:
## -------------------------------------------------------------------
## Inc risk ratio                                 0.83 (0.77, 0.89)
## Odds ratio                                     0.28 (0.14, 0.54)
## Attrib risk *                                  -15.62 (-21.87, -9.38)
## Attrib risk in population *                    -11.40 (-17.07, -5.73)
## Attrib fraction in exposed (%)                 -20.44 (-29.80, -11.77)
## Attrib fraction in population (%)              -14.14 (-20.27, -8.33)
## -------------------------------------------------------------------
##  X2 test statistic: 15.737 p-value: < 0.001
##  Wald confidence limits
##  * Outcomes per 100 population units
##
##  Pearson's product-moment correlation
##
## data:  df$no_show and x
## t = -4.0216, df = 509, p-value = 6.656e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.258298 -0.090120
## sample estimates:
##       cor
## -0.175489
```

```
get_or_and_corr(df$substance,'SA','no_SA' )
```

```
##              Outcome +    Outcome -      Total        Inc risk *
## Exposed +          285           88        373              76.4
## Exposed -           72           66        138              52.2
## Total              357          154        511              69.9
```

```
##                  Odds
## Exposed +        3.24
## Exposed -        1.09
## Total            2.32
##
## Point estimates and 95 % CIs:
## -------------------------------------------------------------------
## Inc risk ratio                                 1.46 (1.24, 1.73)
## Odds ratio                                     2.97 (1.97, 4.48)
## Attrib risk *                                  24.23 (14.85, 33.62)
## Attrib risk in population *                    17.69 (8.45, 26.92)
## Attrib fraction in exposed (%)                 31.72 (19.11, 42.36)
## Attrib fraction in population (%)              25.32 (14.45, 34.81)
## -------------------------------------------------------------------
##  X2 test statistic: 28.097 p-value: < 0.001
##  Wald confidence limits
##  * Outcomes per 100 population units
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$no_show and x
## t = 5.442, df = 509, p-value = 8.203e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1508129 0.3148237
## sample estimates:
##       cor
## 0.2344862
```

```
get_or_and_corr(df$opiod_benzo,'opiod_benzo','no_opiod_benzo' )
```

```
##              Outcome +    Outcome -      Total        Inc risk *
## Exposed +          259          114        373              69.4
## Exposed -           76           62        138              55.1
## Total              335          176        511              65.6
##                  Odds
## Exposed +        2.27
## Exposed -        1.23
## Total            1.90
##
## Point estimates and 95 % CIs:
## -------------------------------------------------------------------
## Inc risk ratio                                 1.26 (1.07, 1.49)
## Odds ratio                                     1.85 (1.24, 2.77)
## Attrib risk *                                  14.36 (4.84, 23.89)
## Attrib risk in population *                    10.49 (1.22, 19.75)
## Attrib fraction in exposed (%)                 20.69 (6.45, 32.75)
## Attrib fraction in population (%)              15.99 (4.53, 26.08)
## -------------------------------------------------------------------
##  X2 test statistic: 9.205 p-value: 0.002
##  Wald confidence limits
##  * Outcomes per 100 population units
```

```
##
##  Pearson's product-moment correlation
```

```
##
## data:  df$no_show and x
## t = 3.0557, df = 509, p-value = 0.002363
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.04803518 0.21841477
## sample estimates:
##       cor
## 0.1342168
```

```
get_or_and_corr(df$preop_attend,'preop_attend','preop_noshow' )
```

```
##              Outcome +    Outcome -      Total        Inc risk *
## Exposed +          192          181        373              51.5
## Exposed -           88           50        138              63.8
## Total              280          231        511              54.8
##                  Odds
## Exposed +        1.06
## Exposed -        1.76
## Total            1.21
##
## Point estimates and 95 % CIs:
## -------------------------------------------------------------------
## Inc risk ratio                                0.81 (0.69, 0.95)
## Odds ratio                                    0.60 (0.40, 0.90)
## Attrib risk *                                 -12.29 (-21.78, -2.80)
## Attrib risk in population *                   -8.97 (-18.08, 0.13)
## Attrib fraction in exposed (%)                -23.88 (-45.34, -5.59)
## Attrib fraction in population (%)             -16.38 (-29.90, -4.27)
## -------------------------------------------------------------------
##  X2 test statistic: 6.146 p-value: 0.013
##  Wald confidence limits
##  * Outcomes per 100 population units
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$no_show and x
## t = -2.4893, df = 509, p-value = 0.01312
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.19455990 -0.02314932
## sample estimates:
##       cor
## -0.1096699
```

```
get_or_and_corr(df$hx_noshow, 'phx_noshows', 'no_phx_noshows')
```

```
##              Outcome +    Outcome -      Total        Inc risk *
## Exposed +          363           10        373              97.3
## Exposed -          118           20        138              85.5
## Total              481           30        511              94.1
##                  Odds
## Exposed +        36.3
## Exposed -         5.9
## Total            16.0
```

```
##
## Point estimates and 95 % CIs:
## -------------------------------------------------------------------
## Inc risk ratio                                 1.14 (1.06, 1.22)
## Odds ratio                                     6.15 (2.80, 13.52)
## Attrib risk *                                  11.81 (5.71, 17.91)
## Attrib risk in population *                    8.62 (2.40, 14.84)
## Attrib fraction in exposed (%)                 12.14 (5.70, 18.14)
## Attrib fraction in population (%)              9.16 (4.16, 13.90)
## -------------------------------------------------------------------
##  X2 test statistic: 25.432 p-value: < 0.001
##  Wald confidence limits
##  * Outcomes per 100 population units
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$no_show and x
## t = 5.1632, df = 509, p-value = 3.488e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1390378 0.3039473
## sample estimates:
##       cor
## 0.2230881
```

```
get_or_and_corr(ifelse(df$proc_type=='Advanced',1,0),'Advanced','Routine' )
```

```
##              Outcome +    Outcome -      Total        Inc risk *
## Exposed +          275           98        373              73.7
## Exposed -          129            9        138              93.5
## Total              404          107        511              79.1
##                  Odds
## Exposed +        2.81
## Exposed -       14.33
## Total            3.78
##
## Point estimates and 95 % CIs:
## -------------------------------------------------------------------
## Inc risk ratio                                 0.79 (0.73, 0.85)
## Odds ratio                                     0.20 (0.10, 0.40)
## Attrib risk *                                  -19.75 (-25.83, -13.68)
## Attrib risk in population *                    -14.42 (-19.84, -8.99)
## Attrib fraction in exposed (%)                 -26.79 (-36.65, -17.64)
## Attrib fraction in population (%)              -18.24 (-24.53, -12.26)
## -------------------------------------------------------------------
##  X2 test statistic: 23.739 p-value: < 0.001
##  Wald confidence limits
##  * Outcomes per 100 population units
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$no_show and x
## t = -4.9797, df = 509, p-value = 8.734e-07
## alternative hypothesis: true correlation is not equal to 0
```

```
## 95 percent confidence interval:
##  -0.2967276 -0.1312472
## sample estimates:
##        cor
## -0.2155343
```

```
cor.test(df$no_show, df$wait_time) # quantitative - 2 by 2 table not needed.
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$no_show and df$wait_time
## t = 4.5706, df = 509, p-value = 6.111e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1137743 0.2804660
## sample estimates:
##       cor
## 0.1985556
```

Like the article my final model identified that patients with a history of no-show had the greatest odds of not attending their endoscopy appointment (article 6.4, this analysis 6.15).

## Varibles associated with NOT showing up to procedure

Below compares my findings with the articles among the predictors that were shown to be *positively* associated with a higher no-show rate:

### History of no-show

- Article - (odds ratio [OR] 6.4; 95 % confidence interval [CI], 2.4-17.5)
- Simulation - 6.15 (2.80, 13.52)

### active substance abuse within the past year

- Article (OR 2.2; 95 % CI 1.4-3.6)
- Simulation 2.97 (1.97, 4.48)

### Longer Wait-time in weeks

- Article (OR 1.05; 95 % CI 1.00-1.09)
- Simulation

### Heavy prescription opioids or benzodiazepines use

- Article (OR 1.6; 95 % CI 1.0-2.6)
- Simulation 1.85 (1.24, 2.77)

## Varibles associated with showing up to procedure

Below compares my findings with the articles among the predictors that were shown to be *inversely* associated with no-shows:

**Active Employment**

- Article (OR 0.38; 95 % CI 0.18-0.81) Simulation 0.28 (0.14, 0.54)

**Attended a pre-operative appointment with an anesthesiologist**

- Article (OR 0.52; CI 0.32-0.85),
- Simulation 0.60 (0.40, 0.90)

**Advanced Procedure**

- Article *ADVANCED* Procedures (OR 0.43; 95 % CI 0.19-0.94)
- Simulation *ADVANCED* Procedures 0.20 (0.10, 0.40)

---

## Logistic Regression

The Logistic Regression formula is shown below:

$$\log(\theta/(1-\theta)) = \gamma_0 + \gamma_1 X_1 + \cdots + \gamma_p X_p$$

Here each $X$ signifies a predictor variable and we are calculating its effect on patient no_show while adjusting for other predictors $X_2, \ldots, X_p$.

```
#Logistic Regression
logisticPseudoR2s <- function(LogModel) {
  dev <- LogModel$deviance
  nullDev <- LogModel$null.deviance
  modelN <-  length(LogModel$fitted.values)
  R.l <-  1 -  dev / nullDev
  R.cs <- 1- exp ( -(nullDev - dev) / modelN)
  R.n <- R.cs / ( 1 - ( exp (-(nullDev / modelN))))
  cat("Pseudo R^2 for logistic regression\n")
  cat("Hosmer and Lemeshow R^2  ", round(R.l, 3), "\n")
  cat("Cox and Snell R^2        ", round(R.cs, 3), "\n")
  cat("Nagelkerke R^2           ", round(R.n, 3),    "\n")
}


df<-merge(df_show, df_noshow,all.x = T,all.y = T)

df$no_show <- as.numeric(df$no_show)

slim_df <- df[,c('no_show','employed','substance',
                 'opiod_benzo','preop_attend','hx_noshow',
                 'proc_type','wait_time')]

fit_null <- glm(formula = no_show~1., data = slim_df, family = 'binomial')
fit_full <- glm(formula = no_show~., data = slim_df, family = 'binomial')
fit_step1 = step(fit_null, scope=list(lower=fit_null, upper=fit_full),direction="forward")
```

```
## Start:  AIC=598.15
## no_show ~ 1
```

```
##
##                Df Deviance    AIC
## + proc_type     1   567.87 571.87
## + substance     1   569.28 573.28
## + hx_noshow     1   574.16 578.16
## + wait_time     1   575.69 579.69
## + employed      1   578.05 582.05
## + opiod_benzo   1   587.15 591.15
## + preop_attend  1   589.93 593.93
## <none>              596.15 598.15
##
## Step:  AIC=571.87
## no_show ~ proc_type
##
##                Df Deviance    AIC
## + substance     1   544.19 550.19
## + hx_noshow     1   544.57 550.57
## + wait_time     1   550.99 556.99
## + employed      1   553.16 559.16
## + opiod_benzo   1   558.68 564.68
## + preop_attend  1   563.28 569.28
## <none>              567.87 571.87
##
## Step:  AIC=550.19
## no_show ~ proc_type + substance
##
##                Df Deviance    AIC
## + hx_noshow     1   524.83 532.83
## + employed      1   530.87 538.87
## + wait_time     1   531.45 539.45
## + opiod_benzo   1   535.96 543.96
## + preop_attend  1   539.36 547.36
## <none>              544.19 550.19
##
## Step:  AIC=532.83
## no_show ~ proc_type + substance + hx_noshow
##
##                Df Deviance    AIC
## + employed      1   512.37 522.37
## + wait_time     1   513.22 523.22
## + opiod_benzo   1   519.65 529.65
## + preop_attend  1   519.76 529.76
## <none>              524.83 532.83
##
## Step:  AIC=522.37
## no_show ~ proc_type + substance + hx_noshow + employed
##
##                Df Deviance    AIC
## + wait_time     1   501.70 513.70
## + opiod_benzo   1   507.21 519.21
## + preop_attend  1   507.27 519.27
## <none>              512.37 522.37
##
## Step:  AIC=513.7
```

```
## no_show ~ proc_type + substance + hx_noshow + employed + wait_time
##
##                Df Deviance    AIC
## + opiod_benzo   1   495.51 509.51
## + preop_attend  1   497.42 511.42
## <none>              501.70 513.70
##
## Step:  AIC=509.51
## no_show ~ proc_type + substance + hx_noshow + employed + wait_time +
##     opiod_benzo
##
##                Df Deviance    AIC
## + preop_attend  1   490.71 506.71
## <none>              495.51 509.51
##
## Step:  AIC=506.71
## no_show ~ proc_type + substance + hx_noshow + employed + wait_time +
##     opiod_benzo + preop_attend
```

## Final Model

```
logisticPseudoR2s(fit_step1)
```

```
## Pseudo R^2 for logistic regression
## Hosmer and Lemeshow R^2   0.177
## Cox and Snell R^2         0.186
## Nagelkerke R^2            0.271
```

```
summary(fit_step1)
```

```
##
## Call:
## glm(formula = no_show ~ proc_type + substance + hx_noshow + employed +
##     wait_time + opiod_benzo + preop_attend, family = "binomial",
##     data = slim_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9842  -0.7598  -0.4800   0.6741   2.6177
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -3.21962    0.46132  -6.979 2.97e-12 ***
## proc_typeRoutine  1.56335    0.39393   3.969 7.23e-05 ***
## substance         0.90480    0.22976   3.938 8.22e-05 ***
## hx_noshow         1.64862    0.45451   3.627 0.000286 ***
## employed         -1.12643    0.36248  -3.108 0.001886 **
## wait_time         0.06125    0.01903   3.219 0.001286 **
## opiod_benzo       0.60233    0.23196   2.597 0.009412 **
## preop_attend     -0.49648    0.22869  -2.171 0.029932 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 596.15  on 510  degrees of freedom
## Residual deviance: 490.71  on 503  degrees of freedom
## AIC: 506.71
##
## Number of Fisher Scoring iterations: 5
```

## Final Model Interpretation

The final model's predictor variables are all identified as significant. This mirrors the findings in the research article. Below are the interpretations of each predictor variables effect on the response 'no_show'

- proc_typeRoutine -
  The beta coef of proc_typeRoutine, 1.56335, means that a one unit increase in proc_typeRoutine (aka the procedure to be performed is a routine procedure NOT an Advanced procedure) is associated with an 1.56335 increase of the logarithm of the odds of the patient not showing up to the scheduled outpatient endoscopic procedure. In short, this means those scheduled for Routine procedures are more likely to not show up.

- substance -
  The beta coef of substance, 0.90480, means that a one unit increase in substance (aka a patient having a history of substance abuse) is associated with an 0.90480 increase of the logarithm of the odds of the patient not showing up to the scheduled outpatient endoscopic procedure. In short, this means those with a history of substance abuse are more likely to not show up.

- hx_noshow -
  The beta coef of hx_noshow, 1.64862, means that a one unit increase in hx_noshow (aka a patient having a history of not showing up to appointments) is associated with an 1.64862 increase of the logarithm of the odds of the patient not showing up to the scheduled outpatient endoscopic procedure. In short this means those with a history of missed appointements are more likely to not show up.

- employed -
  The beta coef of employed, -1.12643, means that a one unit increase in employed (aka a patient being employed) is associated with an 1.12643 DECREASE of the logarithm of the odds of the patient not showing up to the scheduled outpatient endoscopic procedure. In short, this means those who are currently employed are LESS likely to miss an appointment.

- wait_time -
  The beta coef of wait_time, 0.06125, means that a one unit increase in wait_time (aka the number of weeks between preop appt and the scheduled procedure) is associated with an 0.06125 increase of the logarithm of the odds of the patient not showing up to the scheduled outpatient endoscopic procedure. In short, this means those with longer periods of time between preop appts and procedures are more likely to not show up.

- opiod_benzo -
  The beta coef of opiod_benzo, 0.60233, means that a one unit increase in opiod_benzo (aka the patient having being a heavy user of opiods or benzodiazipines) is associated with an 0.60233 increase of the logarithm of the odds of the patient not showing up to the scheduled outpatient endoscopic procedure. In short, this means those who use opiods or benzos heavily are more likely to not show up.

- preop_attend -
  The beta coef of preop_attend, -0.49648, means that a one unit increase in preop_attend (aka the patient attended their preop appointment) is associated with an 0.49648 DECREASE of the logarithm of the odds of the patient not showing up to the scheduled outpatient endoscopic procedure. In short, this means those who attend preop appointments are LESS likely to not show up.

---

## References:

Chang JT, Sewell JL, Day LW. Prevalence and predictors of patient no-shows to outpatient endoscopic procedures scheduled with anesthesia. BMC Gastroenterology. 2015;15:123. doi:10.1186/s12876-015-0358-3 (doi:10.1186/s12876-015-0358-3).