Hospital-Level Analysis

**Load necessary packages

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.2

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(tidyverse)
```

```
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr

## Conflicts with tidy packages ----------------------------------------------

## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
library(broom)
library(RColorBrewer)
library(car)
```

```
## Warning: package 'car' was built under R version 3.4.3

##
## Attaching package: 'car'

## The following object is masked from 'package:purrr':
##
##     some

## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library(caret)
```

```
## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

Load data

We merged subset of CMS datast that is composed only of practitioners who are affliated with hospitals and a dataset of hospitals in the United States in the American Hospital Directory (AHD). The two datasets were inner-joined based on NPI and the City variables. Each row in the "hosp_merged" dataset is a practitioner, and if there are multiple practitioners in one hospital, he/she appears in multiple rows. While merging, some of the hospitals that has zero gross patient revenue in the AHD data were removed from the data. Number of observations in the final dataset was 266,953.

```
hosp_demo <-readRDS("/Users/jiminyoo/Desktop/BST260-FALL2017/BST-260-Final-Project/hosp_demo_full.rds")
hosp_merged <- readRDS("/Users/jiminyoo/Desktop/BST260-FALL2017/BST-260-Final-Project/Hosp_merged_data.
table(hosp_merged$State.x)
```

```
##
##    AK    AL    AR    AZ    CA    CO    CT    DC    DE    FL    GA    HI
##  1230  5305  2270  4811 23030  2651  5914  2014   104 18365  8005   867
##    IA    ID    IL    IN    KS    KY    LA    MA    MD    ME    MI    MN
##  1353   857 11756  4212  2754  6247  7522 12790  7037  6276  6045  4777
##    MO    MS    NC    ND    NE    NH    NJ    NM    NV    NY    OH    OK
##  5143  2350  8793  1011  1697  2641  5943  2028  2412 13304 10889  1981
##    OR    PA    RI    SC    SD    TN    TX    UT    VA    VT    WA    WI
##  4661 14242   919  4067   839  5967 12864  1499  9036   320  5430  1973
##    WY
##   752
```

Aggregate data into hospital-level

We now aggregate 'hosp_merged' data at hospital level so that each row is a unique hospital.

```
#Group hosp_merged data by "Hospital.affiliation.LBN.1," which is hospital name, and "City_trimmed," wh
full_data <- hosp_merged %>%
  group_by(Hospital.affiliation.LBN.1, City_trimmed)

#Recompute Gender and EHR-use variables into numerics
full_data$Gender_num <- ifelse(full_data$Gender == "F", 1, 0)
full_data$EHR_num <- ifelse(full_data$Used.electronic.health.records == "Y", 1, 0)

#created aggregate-level data
agg_data <- summarise(full_data, num_phys = n_distinct(NPI), female_prop = round(mean(Gender_num),2), a

# Checking the if the EHR_use variable is accurate
# EHR_y_list <- agg_data[agg_data$EHR_use == 1, ]$Hospital.affiliation.LBN.1
# EHR_y_data <- subset(hosp_merged, Hospital.affiliation.LBN.1 %in% EHR_y_list)
# table(EHR_y_data$Used.electronic.health.records)
#
# EHR_n_hosp <- agg_data[agg_data$EHR_use == 0, ]
# test = inner_join(hosp_merged, EHR_n_hosp, by=c("Hospital.affiliation.LBN.1" = "Hospital.affiliation.
# table(test$Used.electronic.health.records)
```

We created new variables EHR_char: character vector with two levels "Y" if the hospital uses EHR and 'yrs_since_grad': average of practitioner's years since medical-school graduation to 2017, for those who have the record.

```r
#RECODE Using EHR_use==1 -> Y, 0 -> ""
agg_data$EHR_char <- ifelse(agg_data$EHR_use == 1, "Y", "Blank")
#RECODE Years since medical school graduation
agg_data$yrs_since_grad = 2017 - agg_data$avg_grad_year

#setwd("~Desktop/BST260-FALL2017/BST-260-Final-Project")
#saveRDS(agg_data, "Hosps_aggregated.rds")
```

**Description of final hospital level dataset: There are 1,746 unique hospitals in the dataset.

The aggregated-level variables are number of physicians in each hospital, number of unique specialties among physicians, proportion of female, average years since graduation, number of staffed beds, total discharge, patient days, gross patient revenue for each hospital. The variable of our interest "EHR_use (the hospital uses the electronic health system)" is calculated as 1 if at least one practitioner in the hopstial uses EHR and 0 if none in the hospital uses EHR. Reminder that for practitioners affiliated with hospitals, we assumed that EHR use is the hospital-level adoption and not individual's. Thus it makes sense that if at least one of the practitioners is recorded in the data as using EHR, we will assume the hospital uses EHR.

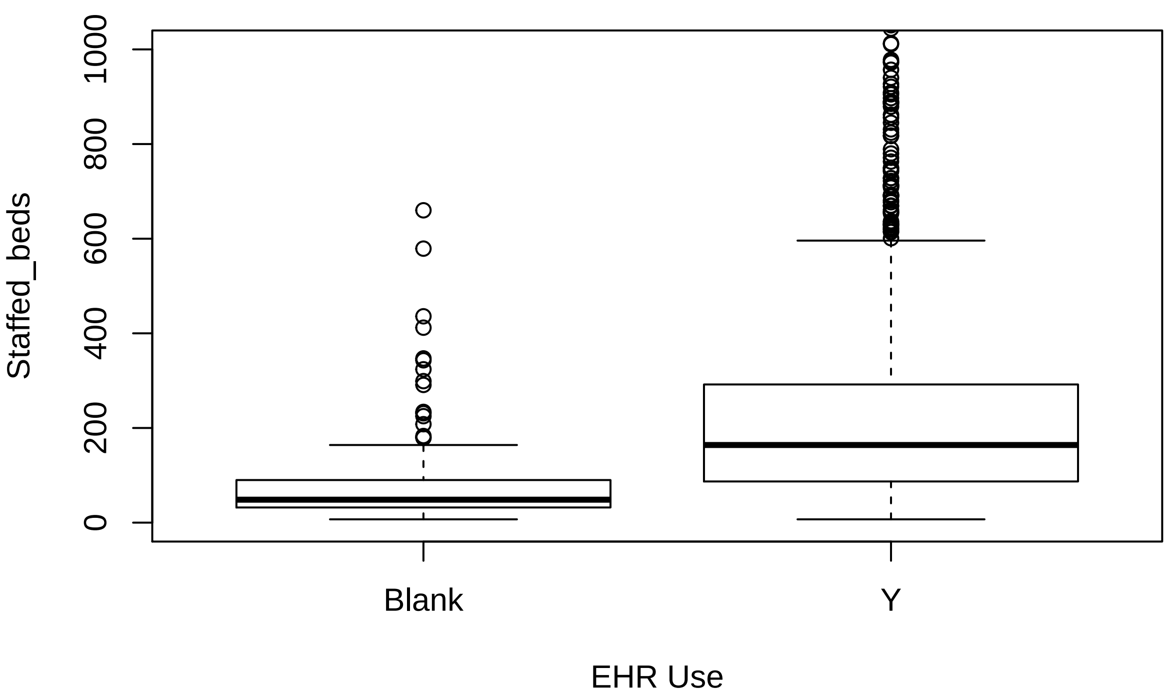
In addition, we created 'EHR_char' and 'yrs_since_grad' variables.

Exploratory Analysis - Hospital Data

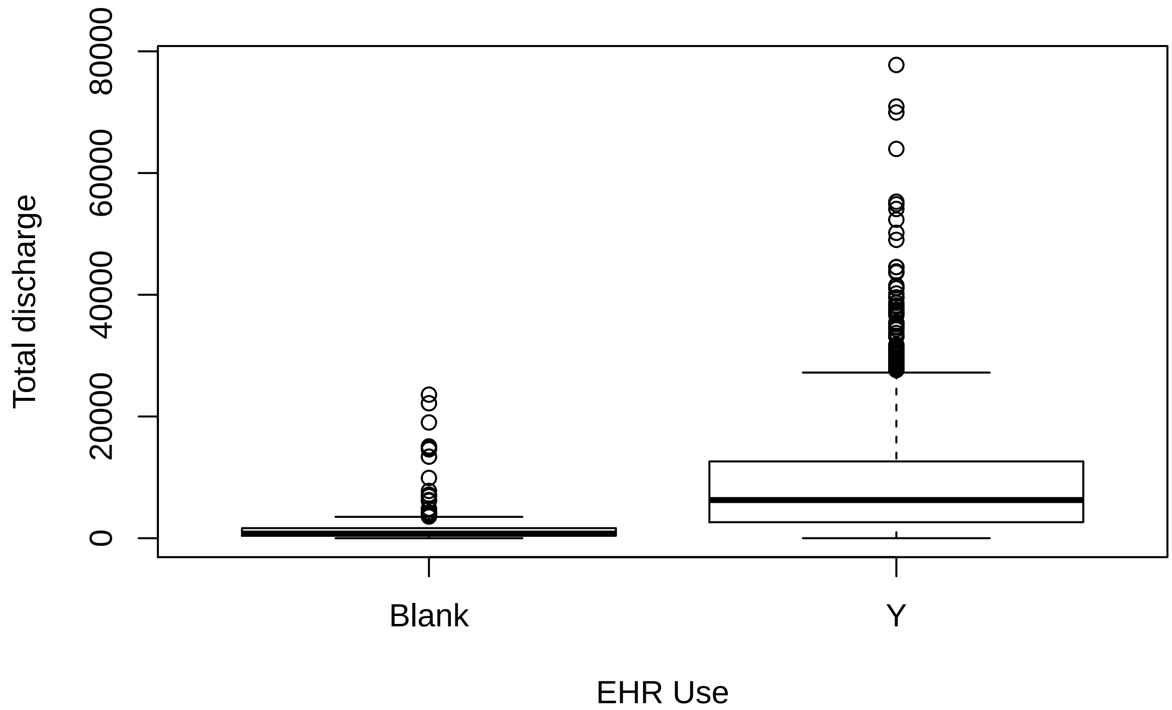
```r
table(agg_data$EHR_use)
```

```
##
##    0    1
##  166 1601
```

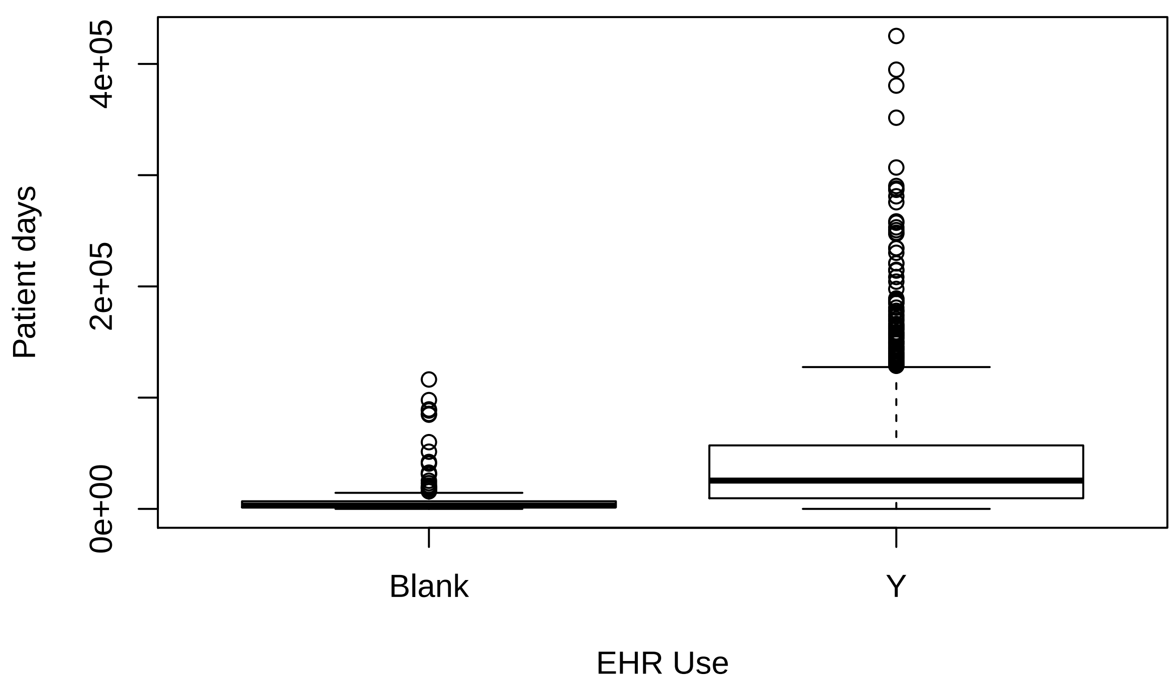
```r
boxplot(staffed_beds~EHR_char,data=agg_data,
        xlab="EHR Use", ylab="Staffed_beds", ylim = c(0,1000))
```



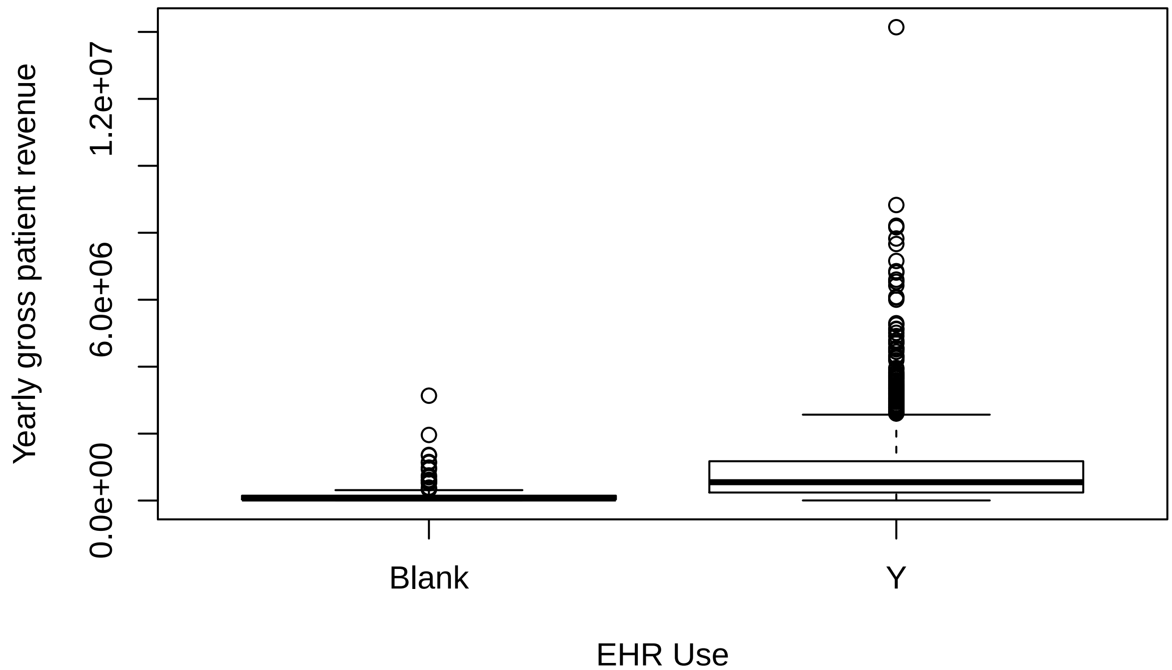
```r
boxplot(total_discharge~EHR_char,data=agg_data,
        xlab="EHR Use", ylab="Total discharge")
```



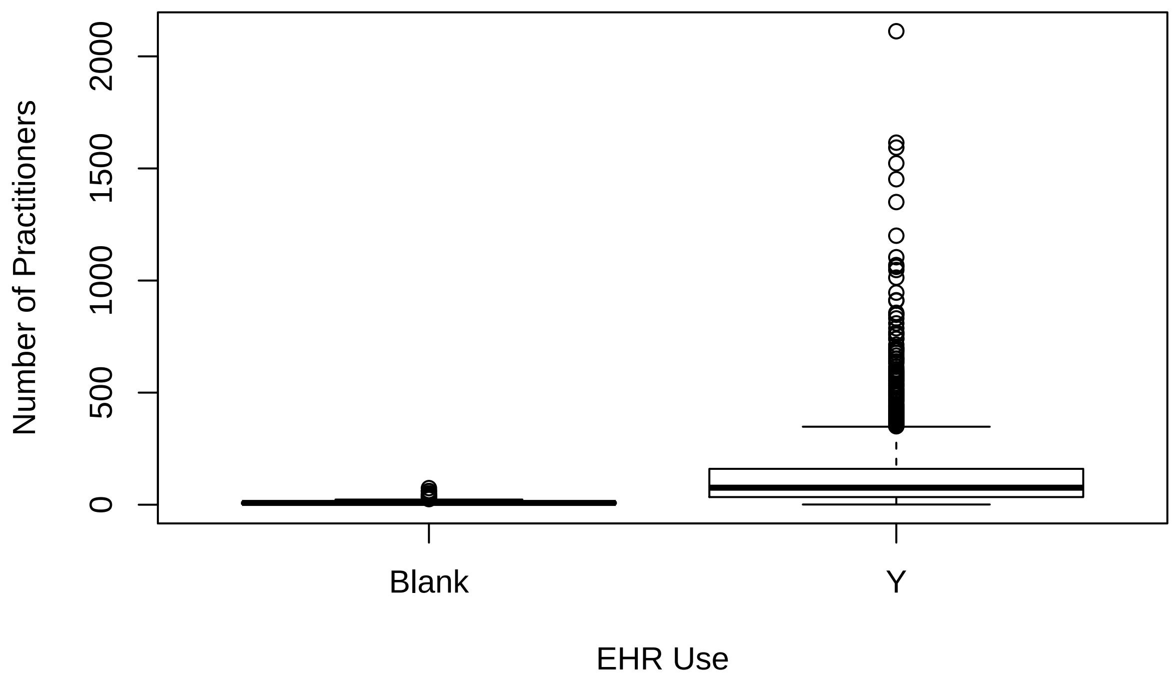

```r
boxplot(patient_days~EHR_char,data=agg_data,
        xlab="EHR Use", ylab="Patient days")
```



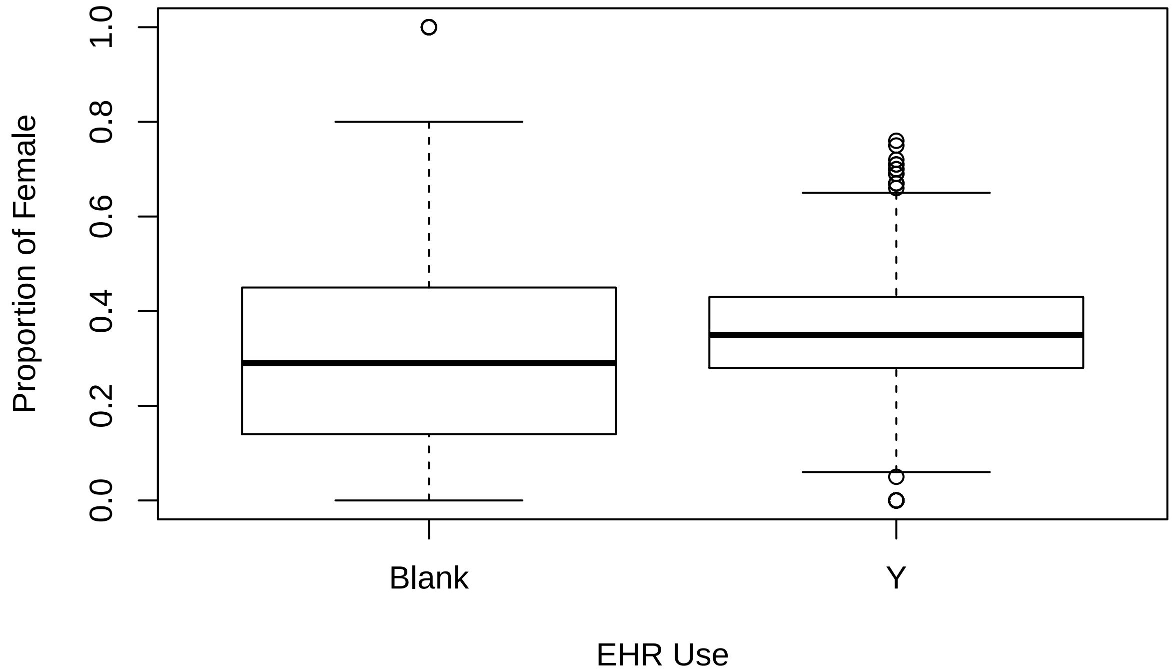

```r
boxplot(gross_patient_rev~EHR_char,data=agg_data,
        xlab="EHR Use", ylab="Yearly gross patient revenue")
```



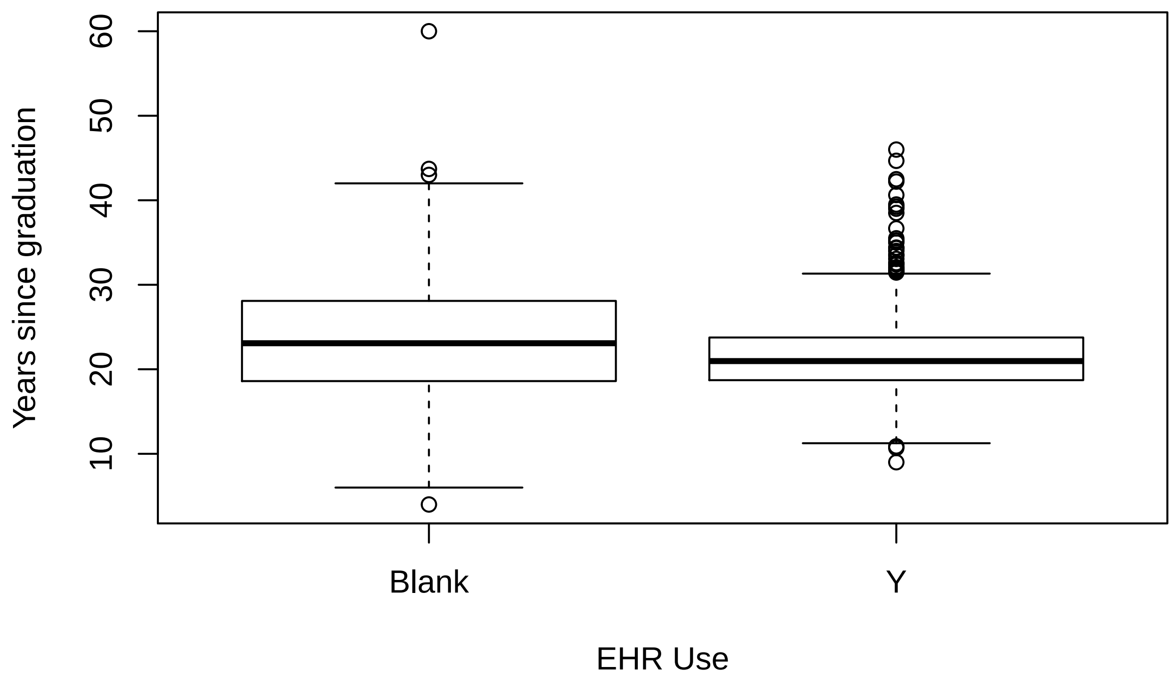

```r
boxplot(num_phys~EHR_char ,data=agg_data,
        xlab="EHR Use", ylab="Number of Practitioners")
```



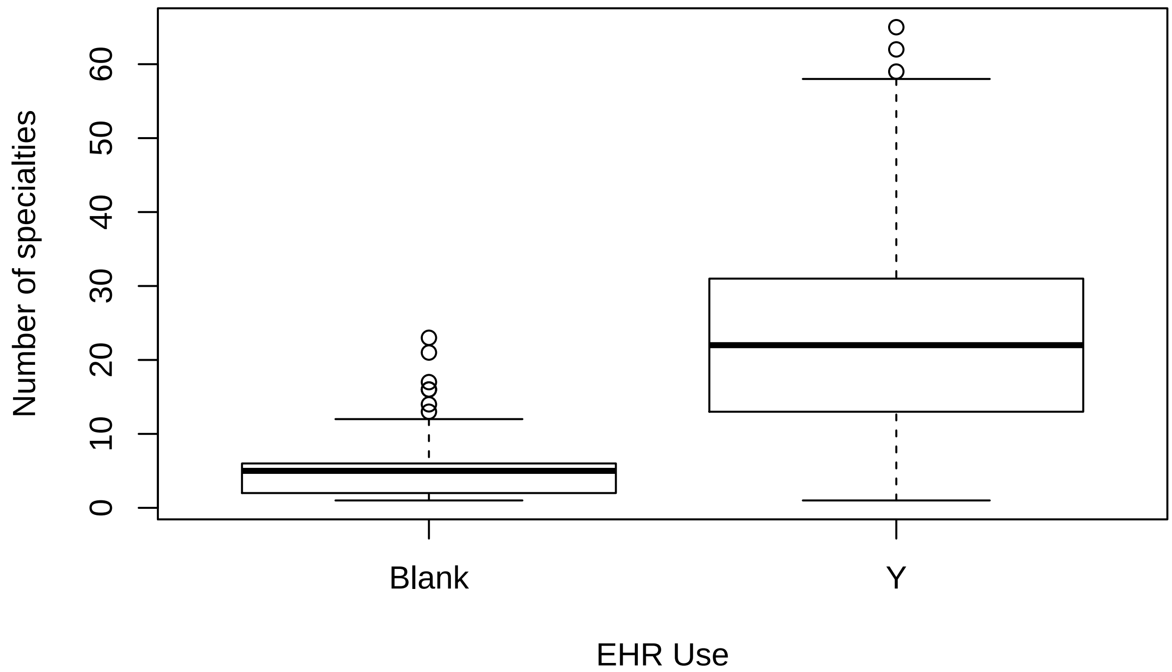

```r
boxplot(female_prop~EHR_char ,data=agg_data,
        xlab="EHR Use", ylab="Proportion of Female")
```




```
boxplot(yrs_since_grad~EHR_char ,data=agg_data,
        xlab="EHR Use", ylab="Years since graduation")
```




```
boxplot(n_specialty~EHR_char ,data=agg_data,
        xlab="EHR Use", ylab="Number of specialties")
```



Correlations

```
#Check for correlations
X <- c("num_phys", "female_prop", "avg_grad_year", "n_specialty", "staffed_beds", "total_discharge", "pa
cor(agg_data$num_phys, agg_data$EHR_use)
```

```
## [1] 0.2057702
```

```
cor(agg_data$female_prop, agg_data$EHR_use)
```

```
## [1] 0.08895285
```

```
cor(agg_data$yrs_since_grad, agg_data$EHR_use, use="complete.obs")
```

```
## [1] -0.1480465
```

```
cor(agg_data$n_specialty, agg_data$EHR_use)
```

```
## [1] 0.3981854
```

```
cor(agg_data$staffed_beds, agg_data$EHR_use)
```

```
## [1] 0.04956017
```

```
cor(agg_data$total_discharge, agg_data$EHR_use)
```

```
## [1] 0.2294191
```

```
cor(agg_data$patient_days, agg_data$EHR_use)
```

```
## [1] 0.2010488
```

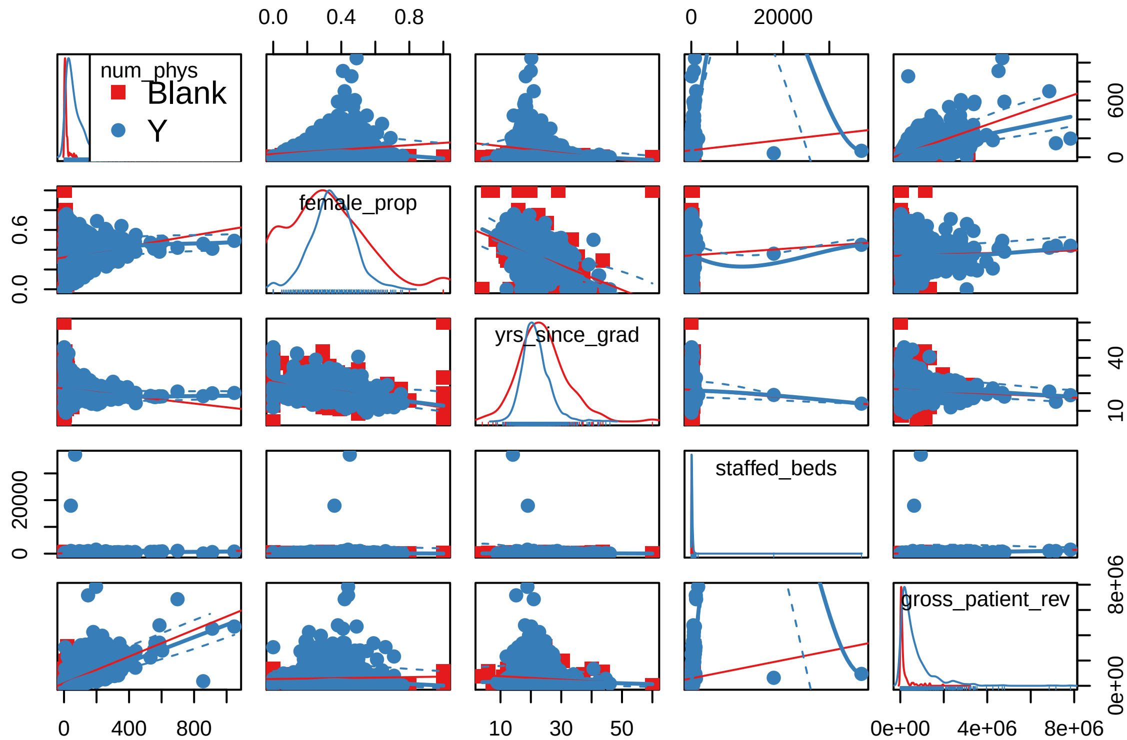
```
cor(agg_data$gross_patient_rev, agg_data$EHR_use)
```

```
## [1] 0.2001997
```

```r
#Correlation Matrix
my_colors <- brewer.pal(nlevels(as.factor(agg_data$EHR_char)), "Set1")
```

```
## Warning in brewer.pal(nlevels(as.factor(agg_data$EHR_char)), "Set1"): minimal value for n is 3, retu
```

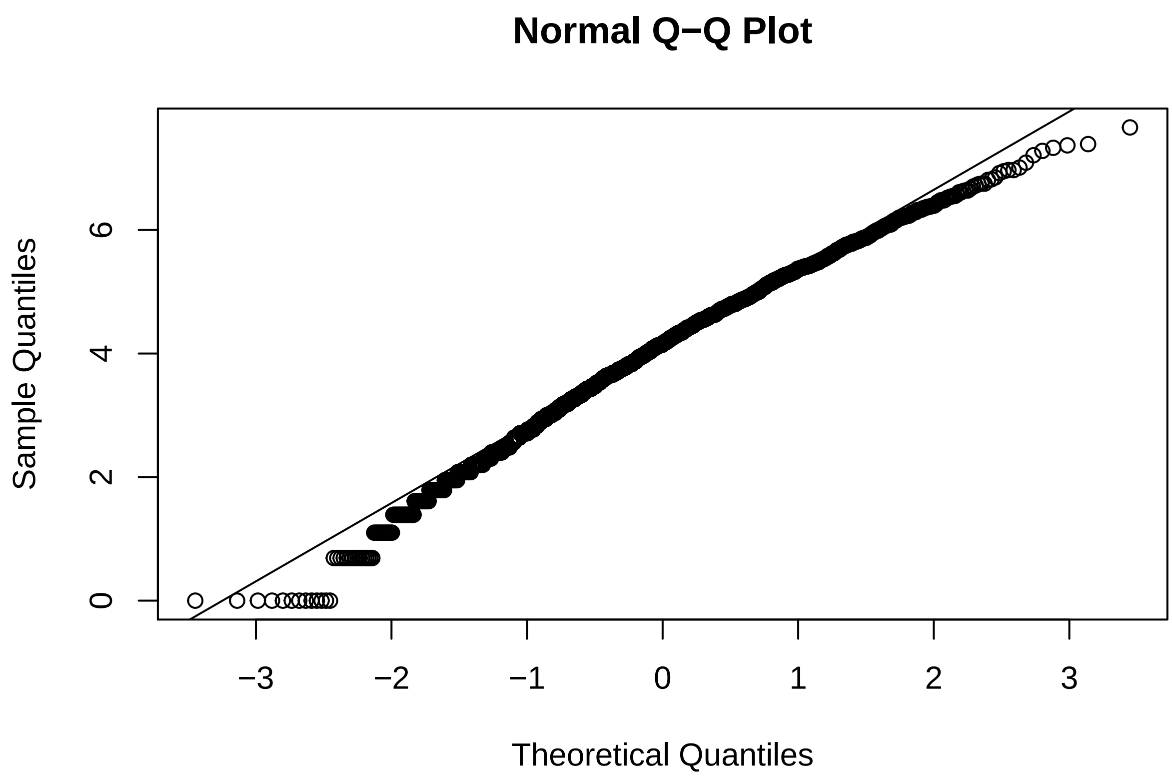
```r
scatterplotMatrix(~num_phys+female_prop+yrs_since_grad+staffed_beds+gross_patient_rev|EHR_char, data=ag
```



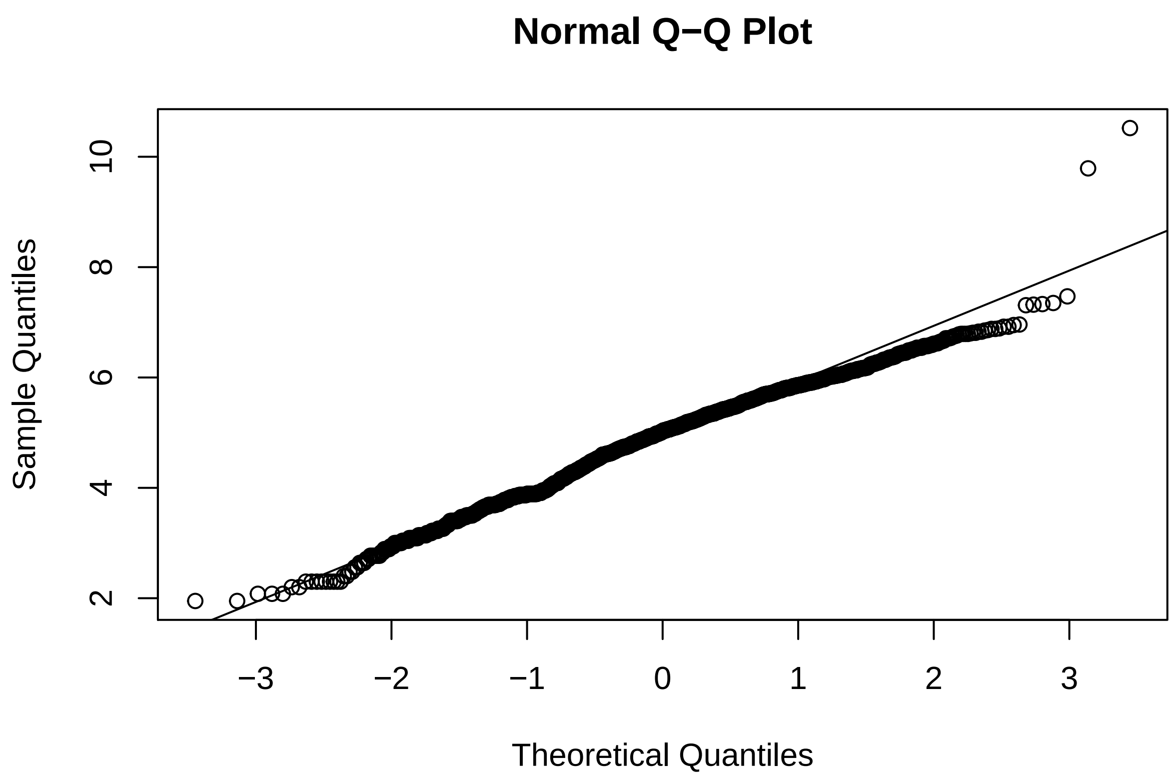
We notice that many of the predictor variables are not normally distributed. We check normalities of the variables.

Normalities of Variables

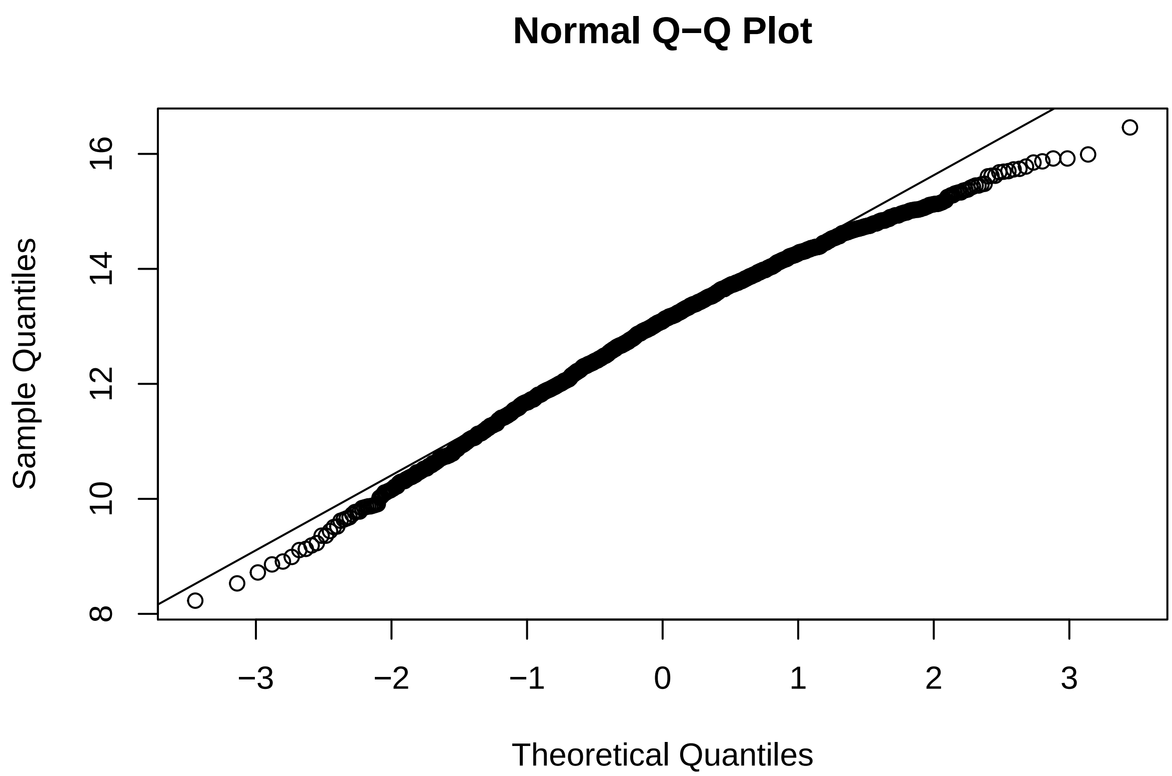
```r
#Variables that are not normally distributed are logged: num_phys, staffed_bed, gross_patient_rev
agg_data$num_phys_log <- round(log(agg_data$num_phys),2)
agg_data$staffed_beds_log <- round(log(agg_data$staffed_beds),2)
agg_data$gross_patient_rev_log <- round(log(agg_data$gross_patient_rev),2)
agg_data$total_discharge_log <- round(log(agg_data$total_discharge),2)
agg_data$patient_days_log<- round(log(agg_data$patient_days),2)

#Check for normality after logging
qqnorm(agg_data$num_phys_log)
qqline(agg_data$num_phys_log)
```



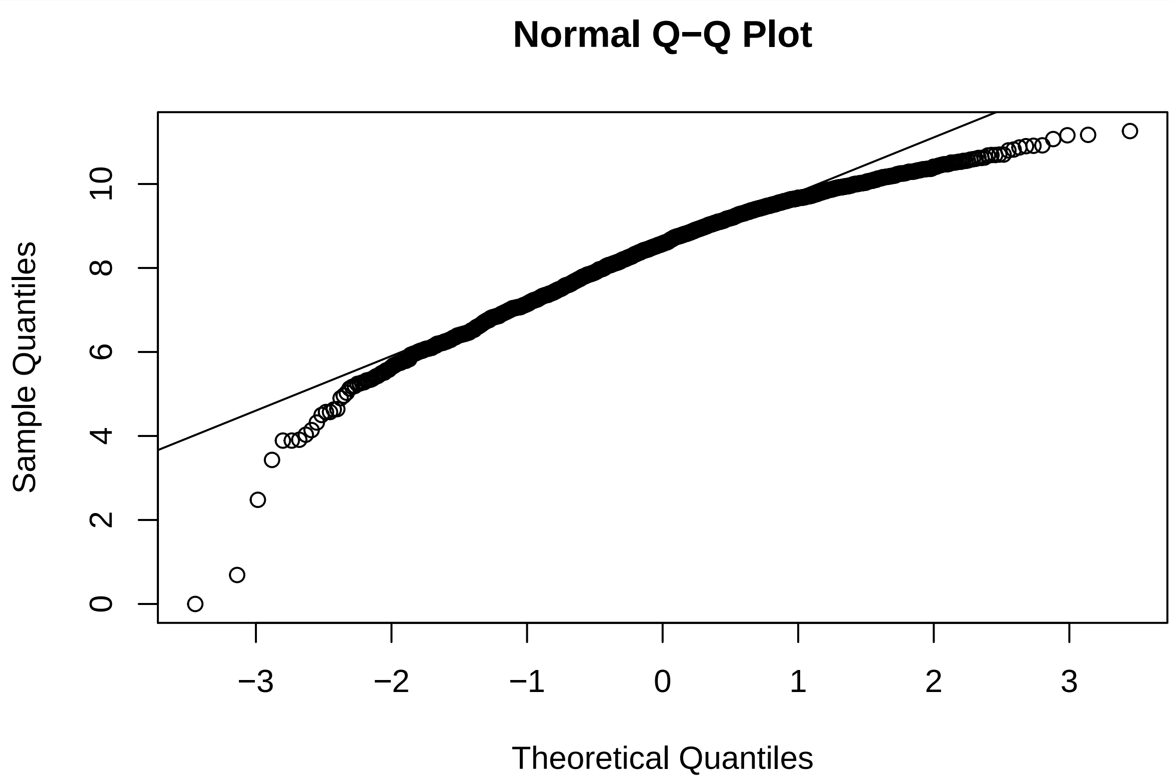

```
qqnorm(agg_data$staffed_beds_log)
qqline(agg_data$staffed_beds_log)
```



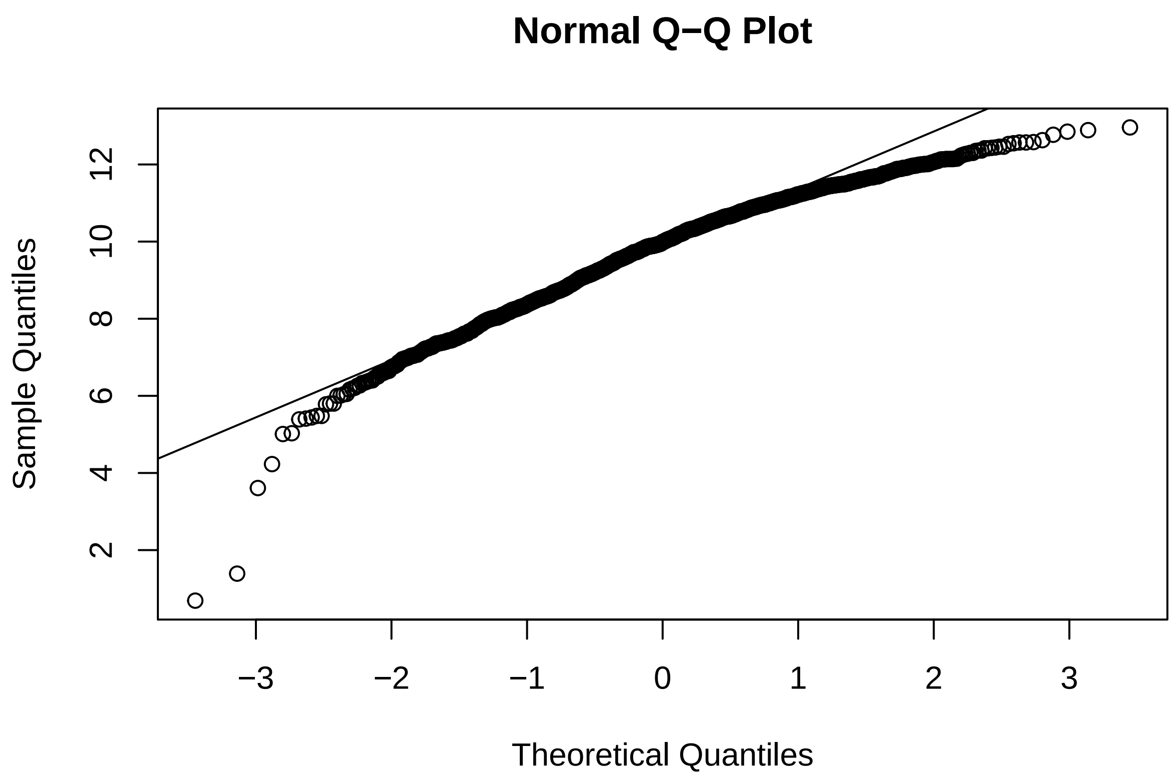

```
qqnorm(agg_data$gross_patient_rev_log)
qqline(agg_data$gross_patient_rev_log)
```




```r
qqnorm(agg_data$total_discharge_log)
qqline(agg_data$total_discharge_log)
```




```r
qqnorm(agg_data$patient_days_log)
qqline(agg_data$patient_days_log)
```



After taking log on the variables, we get much closer to normality for each variables. Now note some of correlations.

```
#Noticeable correlations(more than 0.3):
#EHR_use: EHR_use - num_phys_log, EHR_use-n_specialty, EHR_use-staffed_beds_log, EHR_use-gross_patient_
#Confounding (correlation over 0.7):
#num_phys_log-n_specialty/staffed_Bed _log/total_discharge/gross_patient_rev
#n_specialty-staffed_beds_log/ total_discharge/patient_days/gross_patient_rev_log
#staffed_bed_log-total_discharge, patient_days,gross_patient_rev_log
#total_discharge-patient_days, gross_patient_rev
#patient_days-gross_patient_rev_log
#Conclusion: Confounding factor is the size of the hospital that influences all number of physicians, n
#Potential highest confounding factors are number of physicians-number of speciaties, staffed_beds - gr

agg_data_cor <- agg_data[, c("EHR_use", "num_phys_log","female_prop","n_specialty","staffed_beds_log","
round(cor(agg_data_cor), 2)
```

```
##                     EHR_use num_phys_log female_prop n_specialty
## EHR_use                1.00         0.52        0.09        0.40
## num_phys_log           0.52         1.00        0.24        0.94
## female_prop            0.09         0.24        1.00        0.19
## n_specialty            0.40         0.94        0.19        1.00
## staffed_beds_log       0.32         0.71        0.10        0.72
## total_discharge        0.23         0.70        0.11        0.74
## patient_days           0.20         0.67        0.11        0.72
## gross_patient_rev_log  0.44         0.79        0.08        0.77
## yrs_since_grad           NA           NA          NA          NA
##                     staffed_beds_log total_discharge patient_days
## EHR_use                         0.32            0.23         0.20
## num_phys_log                    0.71            0.70         0.67
```

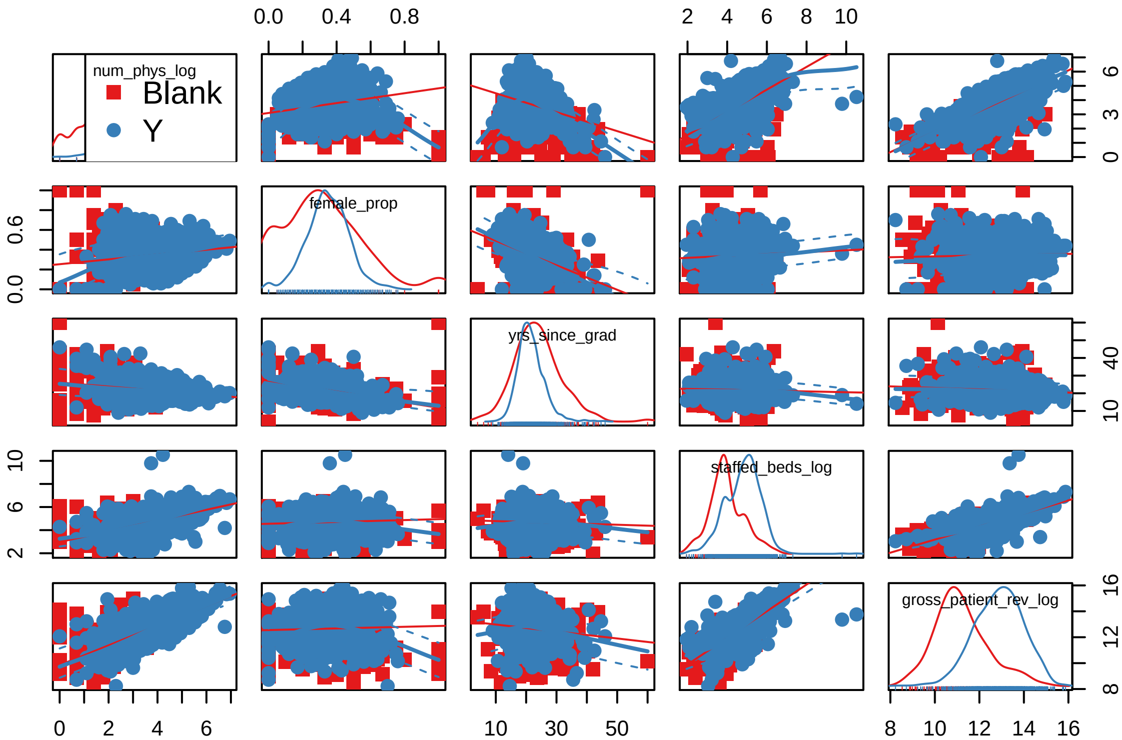
```
## female_prop                    0.10            0.11         0.11
## n_specialty                    0.72            0.74         0.72
## staffed_beds_log               1.00            0.77         0.75
## total_discharge                0.77            1.00         0.98
## patient_days                   0.75            0.98         1.00
## gross_patient_rev_log          0.82            0.77         0.74
## yrs_since_grad                   NA              NA           NA
##                      gross_patient_rev_log yrs_since_grad
## EHR_use                               0.44             NA
## num_phys_log                          0.79             NA
## female_prop                           0.08             NA
## n_specialty                           0.77             NA
## staffed_beds_log                      0.82             NA
## total_discharge                       0.77             NA
## patient_days                          0.74             NA
## gross_patient_rev_log                 1.00             NA
## yrs_since_grad                          NA              1
```

```r
#*Years since graduation comes out as NAs, so pull out yrs_since_grad correlation with only complete ob
round(cor(agg_data_cor, use="complete.obs"), 2)
```

```
##                      EHR_use num_phys_log female_prop n_specialty
## EHR_use                 1.00         0.55        0.07        0.43
## num_phys_log            0.55         1.00        0.21        0.92
## female_prop             0.07         0.21        1.00        0.15
## n_specialty             0.43         0.92        0.15        1.00
## staffed_beds_log        0.29         0.61        0.06        0.63
## total_discharge         0.23         0.59        0.07        0.65
## patient_days            0.20         0.57        0.06        0.63
## gross_patient_rev_log   0.45         0.74        0.04        0.72
## yrs_since_grad         -0.15        -0.28       -0.44       -0.19
##                      staffed_beds_log total_discharge patient_days
## EHR_use                          0.29            0.23         0.20
## num_phys_log                     0.61            0.59         0.57
## female_prop                      0.06            0.07         0.06
## n_specialty                      0.63            0.65         0.63
## staffed_beds_log                 1.00            0.73         0.72
## total_discharge                  0.73            1.00         0.99
## patient_days                     0.72            0.99         1.00
## gross_patient_rev_log            0.76            0.76         0.72
## yrs_since_grad                  -0.04           -0.10        -0.07
##                      gross_patient_rev_log yrs_since_grad
## EHR_use                               0.45          -0.15
## num_phys_log                          0.74          -0.28
## female_prop                           0.04          -0.44
## n_specialty                           0.72          -0.19
## staffed_beds_log                      0.76          -0.04
## total_discharge                       0.76          -0.10
## patient_days                          0.72          -0.07
## gross_patient_rev_log                 1.00          -0.11
## yrs_since_grad                       -0.11           1.00
```

```r
#conclusion: weak negative correlation with all variables with all but female between -0.3 and 0.
#strongest correlation is thefemale proportion, with is -0.44
```

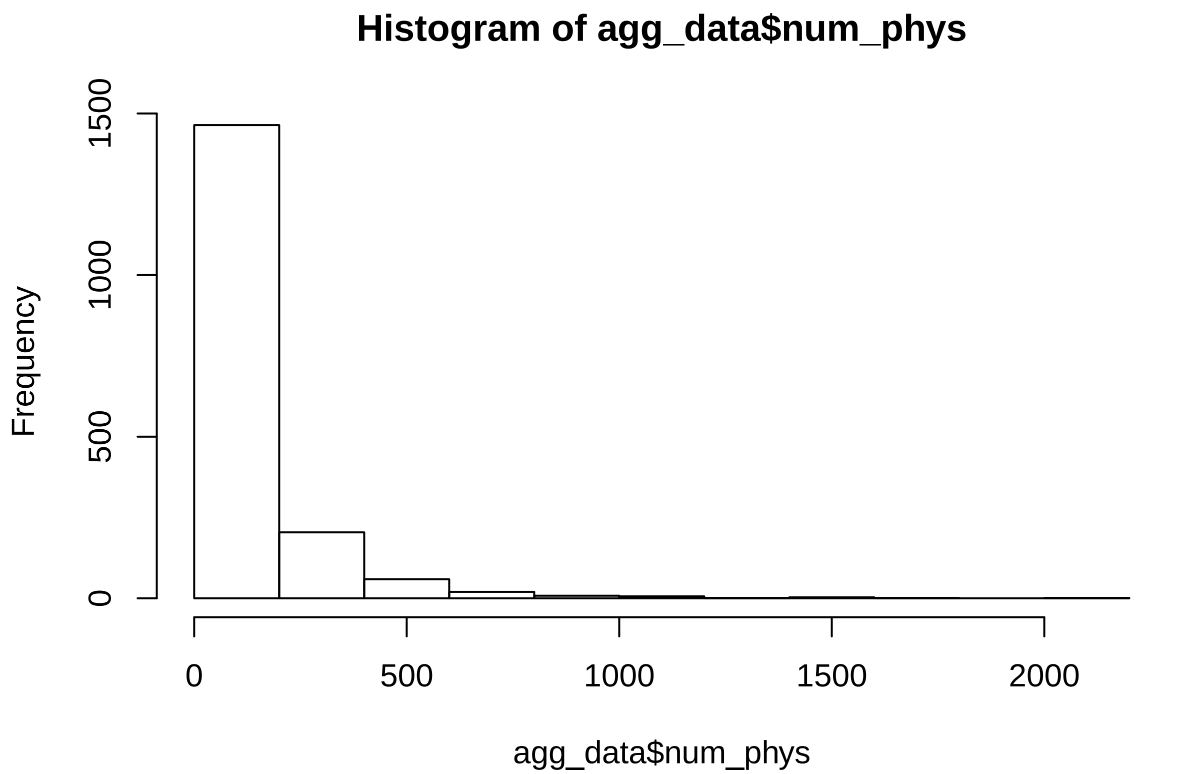
```
#Correlation Matrix with logged variables
scatterplotMatrix(~num_phys_log+female_prop+yrs_since_grad+staffed_beds_log
                  +gross_patient_rev_log|EHR_char, data=agg_data, col=my_colors , smoother.args=list(co
```



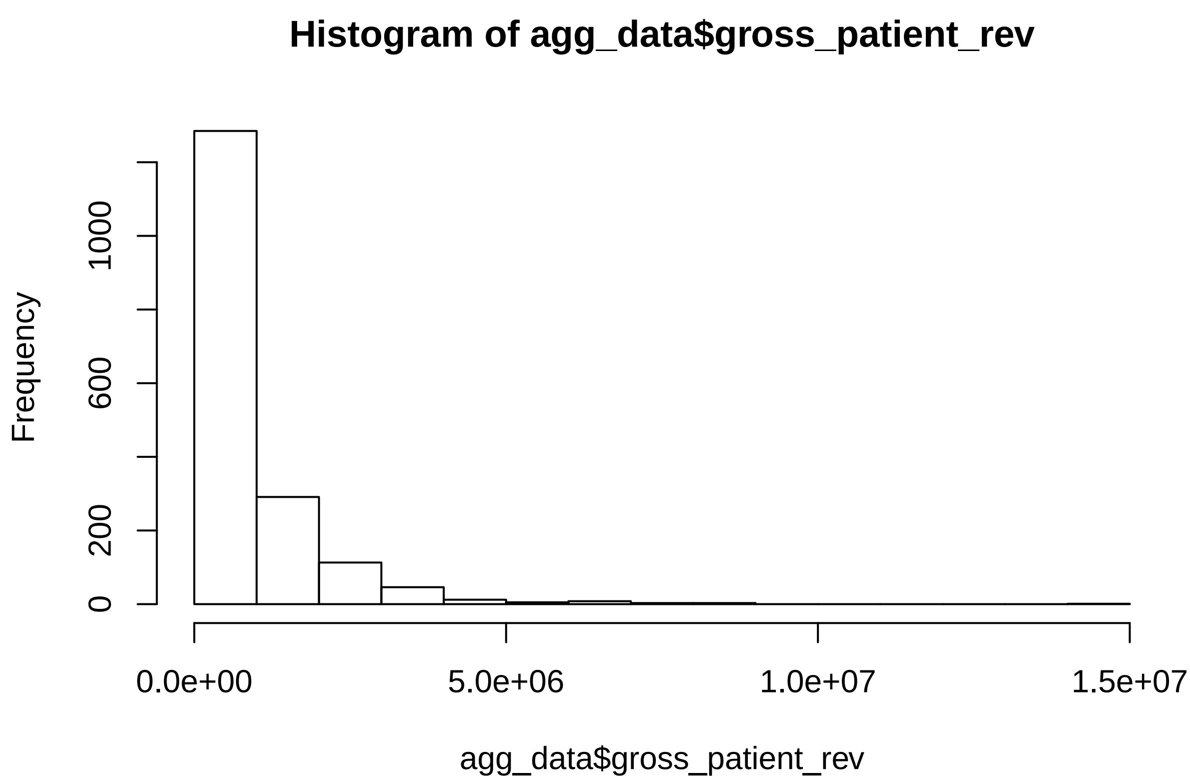
Stratification

Two variables most strongly correlated with EHR use–number of physicians(corr=0.55), and gross_patient_rev_log(corr=0.45)–are also correlated to each other. We believed that the hospital size is a confounding factor that affects both the number of physicians and gross patient revenue. Thus, we will test this theory by stratifying on gross patient revenue.

```
hist(agg_data$num_phys)
```

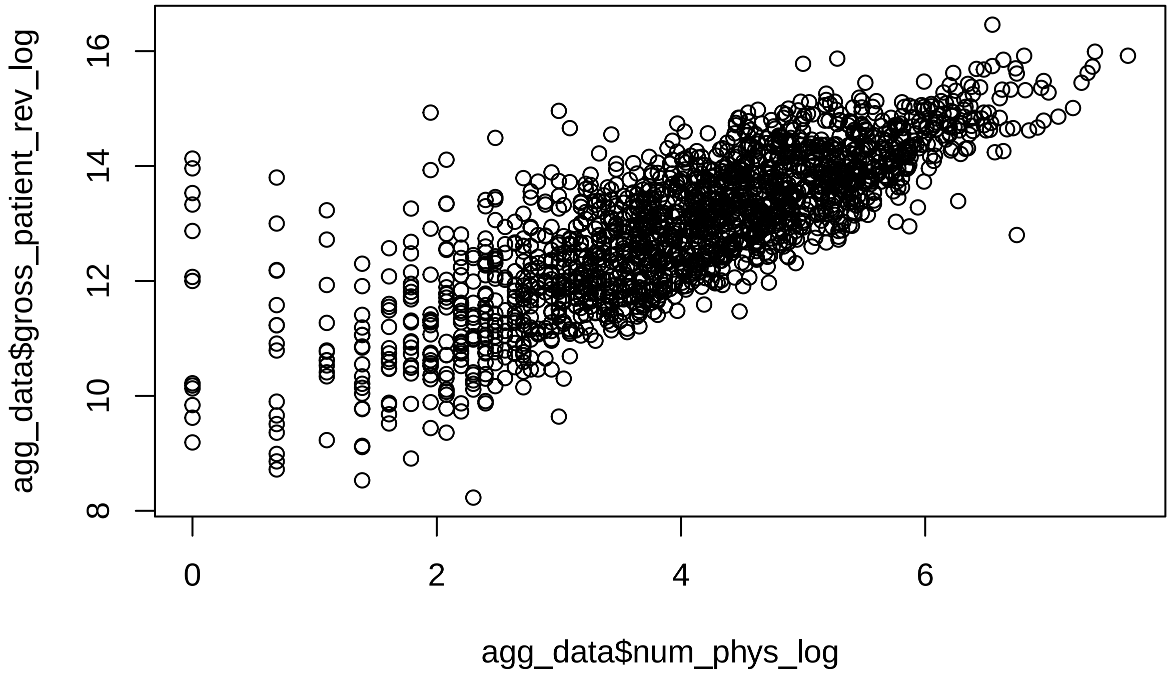



```
hist(agg_data$gross_patient_rev)
```




```
agg_data$num_phys_grp = cut(agg_data$num_phys, quantile(agg_data$num_phys, prob = seq(0, 1, .2)), incluc
agg_data$gpr_grp = cut(agg_data$gross_patient_rev, quantile(agg_data$gross_patient_rev, prob = seq(0, 1
#correlation between number of physicians and gross patientrevenue
```

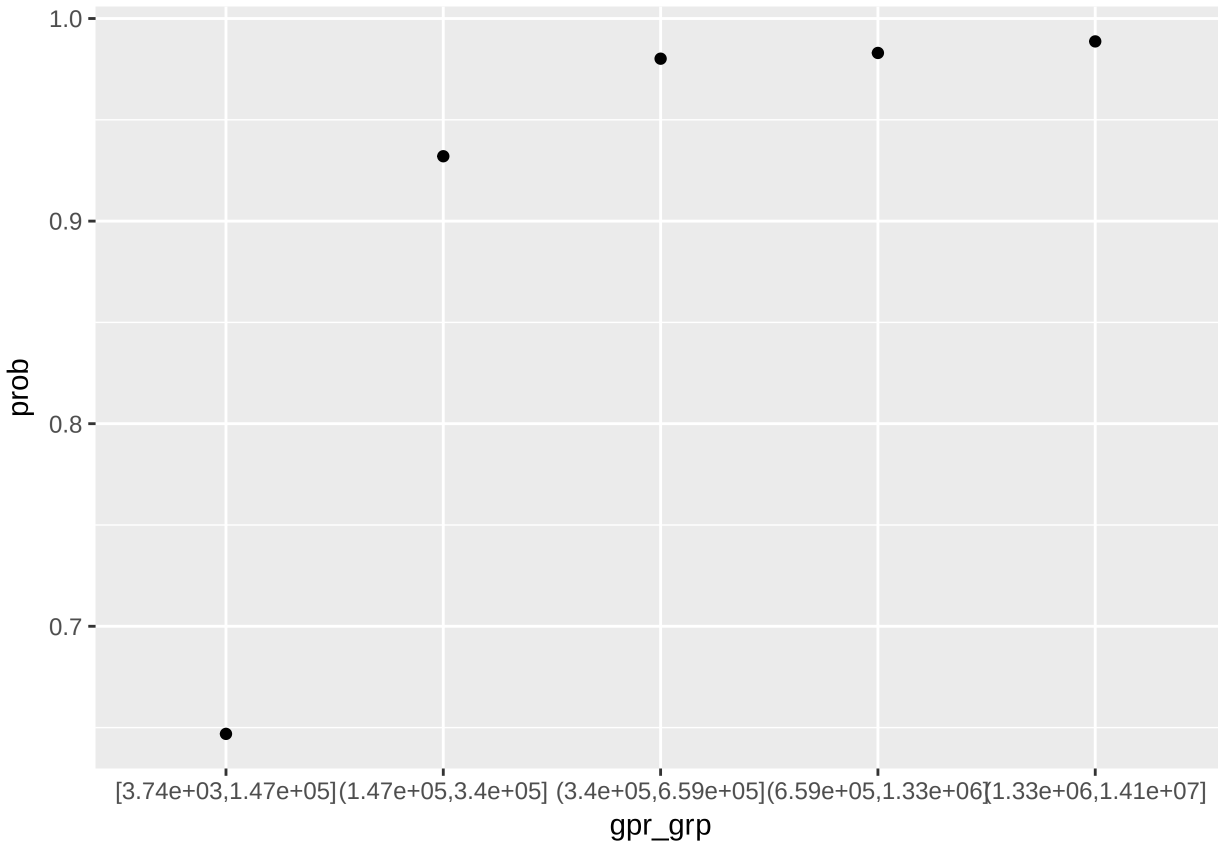
```r
plot(agg_data$num_phys_log, agg_data$gross_patient_rev_log)
```



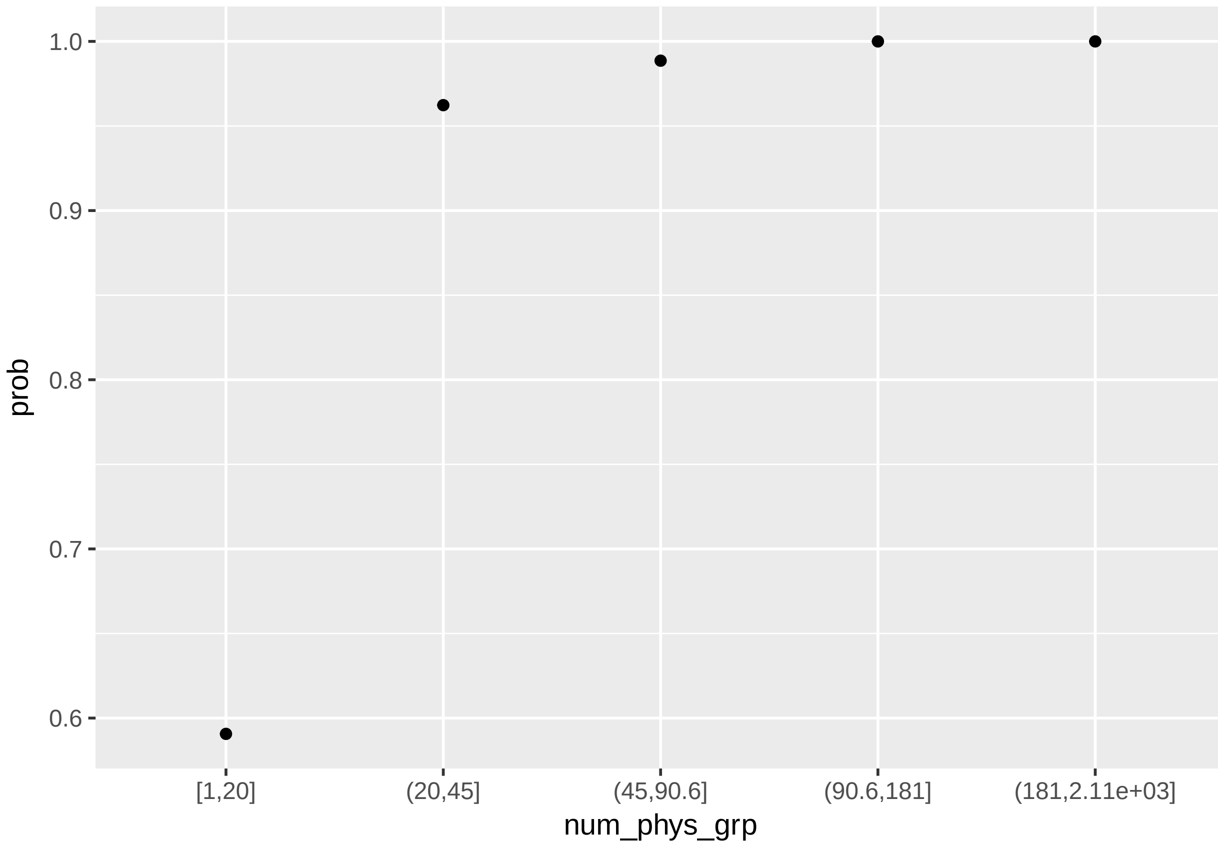

```r
#Table: counts in num_phys and gpr groups
table(agg_data$num_phys_grp, agg_data$gpr_grp)
```

```
##
##                  [3.74e+03,1.47e+05] (1.47e+05,3.4e+05] (3.4e+05,6.59e+05]
##   [1,20]                         239                 80                 26
##   (20,45]                        102                123                 82
##   (45,90.6]                       13                117                106
##   (90.6,181]                       0                 32                108
##   (181,2.11e+03]                   0                  1                 31
##
##                  (6.59e+05,1.33e+06] (1.33e+06,1.41e+07]
##   [1,20]                          14                   5
##   (20,45]                         34                   4
##   (45,90.6]                       87                  28
##   (90.6,181]                     113                 100
##   (181,2.11e+03]                 105                 217
```

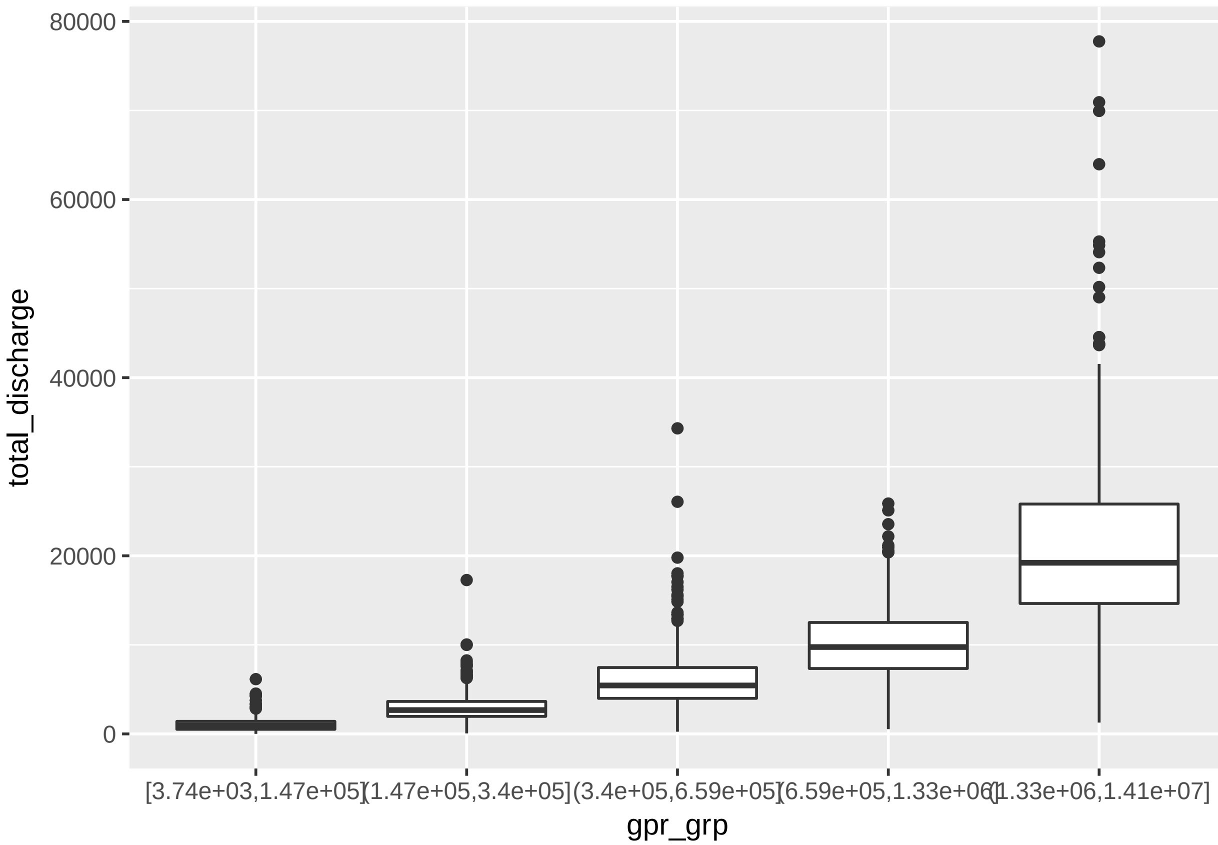
```r
#Heatmap: num_phys, gross_patient_rev, EHR_use proportion

##Proportion of EHR use depending on the gross patient revenue group
agg_data %>%
  group_by(gpr_grp) %>%
  #filter(n() >= 10) %>%
  summarize(prob = mean(EHR_use)) %>%
  ggplot(aes(gpr_grp, prob)) +
  geom_point()
```



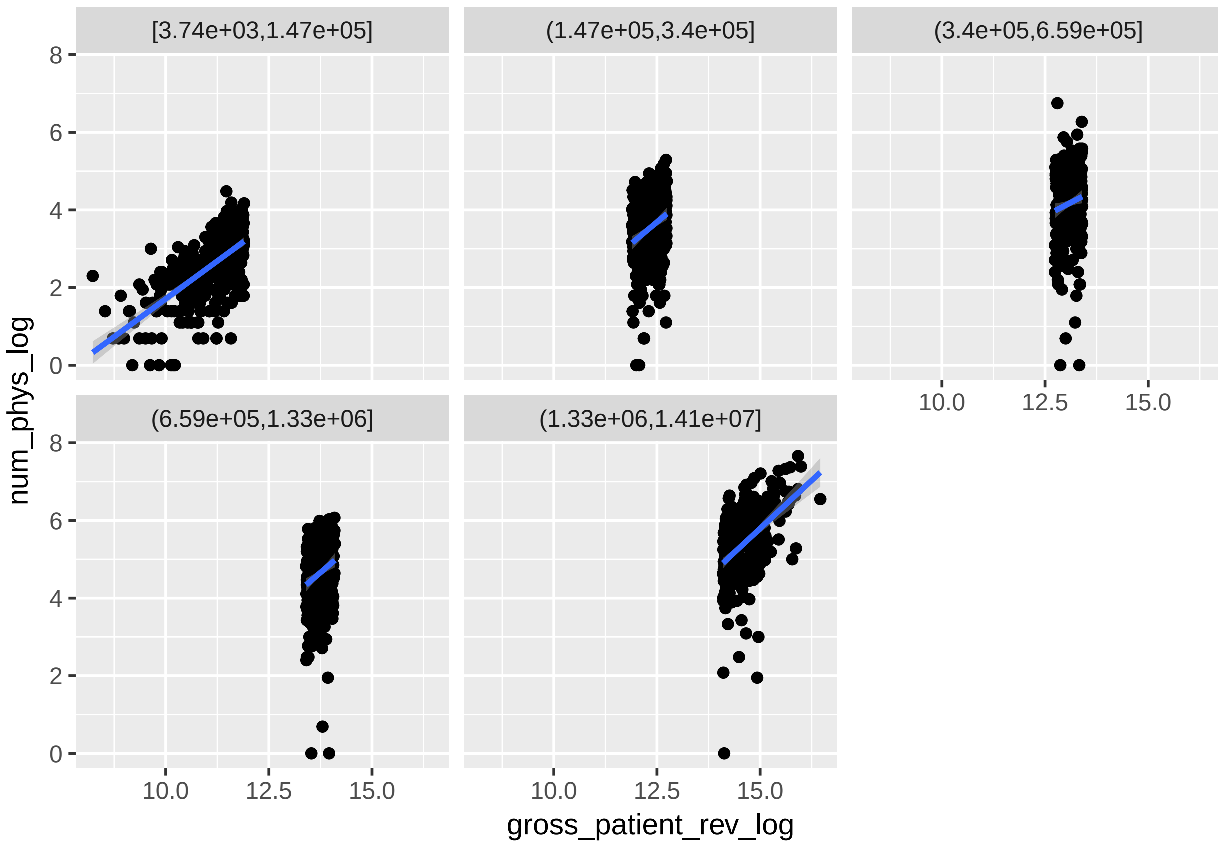

```
##Proportion of EHR use depending on the number of physicians group
agg_data %>%
  group_by(num_phys_grp) %>%
  #filter(n() >= 10) %>%
  summarize(prob = mean(EHR_use)) %>%
  ggplot(aes(num_phys_grp, prob)) +
  geom_point()
```




```
###Boxplot: Gross Revenue vs. Num_phys, staffed_beds_log + total_discharge + patient_days
agg_data %>%
  ggplot(aes(gpr_grp, total_discharge)) +
  geom_boxplot()
```




```
###Correlation Plot
agg_data %>%
  ggplot(aes(gross_patient_rev_log, num_phys_log)) +
  geom_point() +
  geom_smooth(method = "lm") +
  facet_wrap(~gpr_grp)
```



Making Models

Simply all prediction variables

```
filter_var = "patient_days"
agg_data %>%
    group_by(gpr_grp) %>%
    do(tidy(glm(EHR_use ~ num_phys + staffed_beds_log + total_discharge + patient_days, data = .), conf
    filter(term==filter_var)
```

```
## # A tibble: 5 x 8
## # Groups:   gpr_grp [5]
##               gpr_grp         term      estimate    std.error  statistic
##                <fctr>        <chr>         <dbl>        <dbl>      <dbl>
## 1 [3.74e+03,1.47e+05] patient_days -8.954703e-06 8.992329e-06 -0.9958157
## 2  (1.47e+05,3.4e+05] patient_days -1.553826e-05 4.949480e-06 -3.1393719
## 3  (3.4e+05,6.59e+05] patient_days -6.378796e-06 1.595285e-06 -3.9985297
## 4 (6.59e+05,1.33e+06] patient_days -1.307471e-06 1.100206e-06 -1.1883875
## 5 (1.33e+06,1.41e+07] patient_days -5.169995e-07 3.629587e-07 -1.4244032
## # ... with 3 more variables: p.value <dbl>, conf.low <dbl>,
## #   conf.high <dbl>
```

When stratified by gross patient revenue, NONE appears significant but NUMBER OF PHYSICIANS

Train and Test Datasets

```
library(caret)
Train <- createDataPartition(agg_data$EHR_use, p=0.6, list=FALSE)
training <- agg_data[Train, ]
testing <- agg_data[-Train, ]
```

```
#TEST GLM1
glm1 <- glm(EHR_use ~ gross_patient_rev_log + staffed_beds_log + total_discharge_log + patient_days_log
summary(glm1)
```

```
##
## Call:
## glm(formula = EHR_use ~ gross_patient_rev_log + staffed_beds_log +
##     total_discharge_log + patient_days_log + staffed_beds_log:gross_patient_rev_log +
##     total_discharge_log:patient_days_log, family = "binomial",
##     data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1296   0.1378   0.2163   0.3626   2.3953
##
## Coefficients:
##                                        Estimate Std. Error z value
## (Intercept)                           -16.68326    5.62606  -2.965
## gross_patient_rev_log                   1.61629    0.51028   3.167
## staffed_beds_log                        1.84296    1.46103   1.261
## total_discharge_log                     0.83841    0.55278   1.517
## patient_days_log                       -1.12242    0.54695  -2.052
## gross_patient_rev_log:staffed_beds_log -0.17497    0.11933  -1.466
## total_discharge_log:patient_days_log    0.06017    0.05035   1.195
##                                       Pr(>|z|)
## (Intercept)                            0.00302 **
## gross_patient_rev_log                  0.00154 **
## staffed_beds_log                       0.20716
## total_discharge_log                    0.12934
## patient_days_log                       0.04015 *
## gross_patient_rev_log:staffed_beds_log 0.14256
## total_discharge_log:patient_days_log   0.23208
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 676.10  on 1060  degrees of freedom
## Residual deviance: 462.03  on 1054  degrees of freedom
## AIC: 476.03
##
## Number of Fisher Scoring iterations: 7
```

```
p_hat_logit <- predict(glm1, newdata = testing, type="response")
y_hat_logit <- ifelse(p_hat_logit > 0.5, 1,  0)
confusionMatrix(data = y_hat_logit, reference = testing$EHR_use)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  17  11
##          1  46 632
##
##                Accuracy : 0.9193
##                  95% CI : (0.8967, 0.9383)
##     No Information Rate : 0.9108
##     P-Value [Acc > NIR] : 0.2365
##
##                   Kappa : 0.3372
##  Mcnemar's Test P-Value : 6.687e-06
##
##             Sensitivity : 0.26984
##             Specificity : 0.98289
##          Pos Pred Value : 0.60714
##          Neg Pred Value : 0.93215
##              Prevalence : 0.08924
##          Detection Rate : 0.02408
##    Detection Prevalence : 0.03966
##       Balanced Accuracy : 0.62637
##
##        'Positive' Class : 0
##
```

```r
#TEST GLM2
#WE WON'T use this model because num_phys is not very reliable
#BUT the model has the best fit.
glm2 <- glm(EHR_use ~ gross_patient_rev_log + num_phys_log + gross_patient_rev_log*num_phys_log, data=t
summary(glm2)
```

```
##
## Call:
## glm(formula = EHR_use ~ gross_patient_rev_log + num_phys_log +
##     gross_patient_rev_log * num_phys_log, family = "binomial",
##     data = training)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.02498   0.06059   0.12004   0.26656   2.51199
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                        -10.8045     4.1692  -2.592  0.00956 **
## gross_patient_rev_log                0.6374     0.3459   1.842  0.06541 .
## num_phys_log                         3.3523     1.6649   2.013  0.04406 *
## gross_patient_rev_log:num_phys_log  -0.1315     0.1326  -0.992  0.32122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 676.10  on 1060  degrees of freedom
```

```
## Residual deviance: 357.57  on 1057  degrees of freedom
## AIC: 365.57
##
## Number of Fisher Scoring iterations: 8
```

```
p_hat_logit <- predict(glm2, newdata = testing, type="response")
y_hat_logit <- ifelse(p_hat_logit > 0.5, 1,  0)
confusionMatrix(data = y_hat_logit, reference = testing$EHR_use)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  30  14
##          1  33 629
##
##                Accuracy : 0.9334
##                  95% CI : (0.9125, 0.9507)
##     No Information Rate : 0.9108
##     P-Value [Acc > NIR] : 0.01734
##
##                   Kappa : 0.526
##  Mcnemar's Test P-Value : 0.00865
##
##             Sensitivity : 0.47619
##             Specificity : 0.97823
##          Pos Pred Value : 0.68182
##          Neg Pred Value : 0.95015
##              Prevalence : 0.08924
##          Detection Rate : 0.04249
##    Detection Prevalence : 0.06232
##       Balanced Accuracy : 0.72721
##
##        'Positive' Class : 0
##
```

```
anova(glm1, glm2, test ="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: EHR_use ~ gross_patient_rev_log + staffed_beds_log + total_discharge_log +
##     patient_days_log + staffed_beds_log:gross_patient_rev_log +
##     total_discharge_log:patient_days_log
## Model 2: EHR_use ~ gross_patient_rev_log + num_phys_log + gross_patient_rev_log *
##     num_phys_log
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1054     462.03
## 2      1057     357.57 -3   104.47
```

```
#TEST GLM3
glm3 <- glm(EHR_use ~ gross_patient_rev_log + staffed_beds_log  + gross_patient_rev_log*staffed_beds_log
summary(glm3)
```

```
##
## Call:
## glm(formula = EHR_use ~ gross_patient_rev_log + staffed_beds_log +
```

```
##     gross_patient_rev_log * staffed_beds_log, family = "binomial",
##     data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1746   0.1453   0.2310   0.3763   2.6312
##
## Coefficients:
##                                    Estimate Std. Error z value
## (Intercept)                        -19.7509     5.8917  -3.352
## gross_patient_rev_log                1.8379     0.5047   3.642
## staffed_beds_log                     1.3850     1.3273   1.043
## gross_patient_rev_log:staffed_beds_log  -0.1205     0.1092  -1.104
##                                    Pr(>|z|)
## (Intercept)                        0.000801 ***
## gross_patient_rev_log              0.000271 ***
## staffed_beds_log                   0.296723
## gross_patient_rev_log:staffed_beds_log 0.269699
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 676.10  on 1060  degrees of freedom
## Residual deviance: 472.76  on 1057  degrees of freedom
## AIC: 480.76
##
## Number of Fisher Scoring iterations: 7
```

```
p_hat_logit <- predict(glm3, newdata = testing, type="response")
y_hat_logit <- ifelse(p_hat_logit > 0.5, 1,  0)
confusionMatrix(data = y_hat_logit, reference = testing$EHR_use)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  16   9
##          1  47 634
##
##                Accuracy : 0.9207
##                  95% CI : (0.8982, 0.9395)
##     No Information Rate : 0.9108
##     P-Value [Acc > NIR] : 0.1967
##
##                   Kappa : 0.3296
##  Mcnemar's Test P-Value : 7.641e-07
##
##             Sensitivity : 0.25397
##             Specificity : 0.98600
##          Pos Pred Value : 0.64000
##          Neg Pred Value : 0.93098
##              Prevalence : 0.08924
##          Detection Rate : 0.02266
##    Detection Prevalence : 0.03541
```

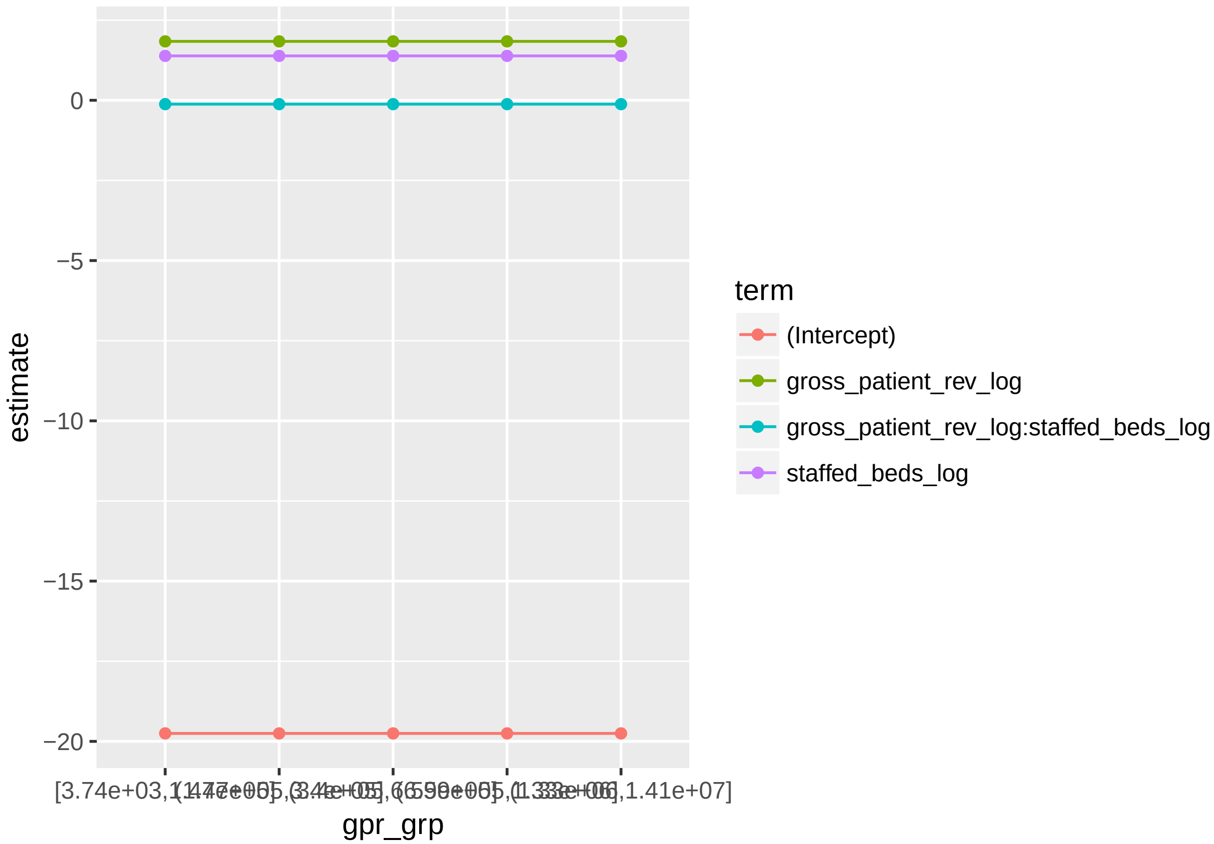
```
##       Balanced Accuracy : 0.61999
##
##        'Positive' Class : 0
##
```

```r
anova(glm1, glm3, test ="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: EHR_use ~ gross_patient_rev_log + staffed_beds_log + total_discharge_log +
##     patient_days_log + staffed_beds_log:gross_patient_rev_log +
##     total_discharge_log:patient_days_log
## Model 2: EHR_use ~ gross_patient_rev_log + staffed_beds_log + gross_patient_rev_log *
##     staffed_beds_log
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1054     462.03
## 2      1057     472.76 -3  -10.725  0.01331 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Estimate Plots

```r
##MODEL ESTIMATE PLOTS
agg_data %>%
  group_by(gpr_grp) %>%
  do(tidy(glm3)) %>%
  #filter(!grepl("Intercept", term))%>%
  ggplot(aes(gpr_grp, estimate, group = term, col = term)) +
  geom_line() +
  geom_point()
```




```
#KATHERINE NEVERMIND THE BELOW THINGS
#Ribbon plots
# newdata1$rankP <- predict(glm3, newdata = agg_data, type = "response")
# newdata1
#
# newdata2 <- with(mydata, data.frame(gre = rep(seq(from = 200, to = 800, length.out = 100),
#     4), gpa = mean(gpa), rank = factor(rep(1:4, each = 100))))
#
# fit <- predict(glm3, newdata = agg_data, type = "link", se = TRUE)$fit
# se.fit <- predict(glm3, newdata = agg_data, type = "link", se = TRUE)$se.fit
# #$fit
# #$se.fit
# newdata3 <- cbind(agg_data, fit)[1:10,]
# newdata3 <- within(new_data3, {
#     PredictedProb <- plogis(glm3)
#     LL <- plogis(fit - (1.96 * se.fit))
#     UL <- plogis(fit + (1.96 * se.fit))
# })
#
# ## view first few rows of final dataset
# head(newdata3)
#
# ggplot(newdata3, aes(x = gre, y = PredictedProb)) + geom_ribbon(aes(ymin = LL,
#     ymax = UL, fill = rank), alpha = 0.2) + geom_line(aes(colour = rank),
#     size = 1)
```

Calculate odds ratio and CI

Our final model is glm3 with two variables: staffed beds and gross revenue.

```r
#confidence intervals with log-likelihood
confint(glm3)
```

```
## Waiting for profiling to be done...
```

```
##                                             2.5 %     97.5 %
## (Intercept)                           -31.0521047 -8.0265438
## gross_patient_rev_log                   0.8279081  2.7982723
## staffed_beds_log                       -1.3086999  3.8911261
## gross_patient_rev_log:staffed_beds_log  -0.3238057  0.1036141
```

```r
#table of 95% confidence intervals

#CIs using standard errors
confint.default(glm3)
```

```
##                                             2.5 %      97.5 %
## (Intercept)                           -31.2984448 -8.20338608
## gross_patient_rev_log                   0.8487190  2.82707205
## staffed_beds_log                       -1.2164140  3.98638904
## gross_patient_rev_log:staffed_beds_log  -0.3344917  0.09347924
```

```r
#table of odds ratios
exp(coef(glm3))
```

```
##                            (Intercept)
##                           2.644152e-09
##                  gross_patient_rev_log
##                           6.283301e+00
##                       staffed_beds_log
##                           3.994776e+00
## gross_patient_rev_log:staffed_beds_log
##                           8.864716e-01
```

```r
#table of odds ratios with 95% CI
exp(cbind(OR = coef(glm3), confint(glm3)))
```

```
## Waiting for profiling to be done...
```

```
##                                                  OR        2.5 %
## (Intercept)                            2.644152e-09 3.267701e-14
## gross_patient_rev_log                  6.283301e+00 2.288526e+00
## staffed_beds_log                       3.994776e+00 2.701711e-01
## gross_patient_rev_log:staffed_beds_log 8.864716e-01 7.233908e-01
##                                              97.5 %
## (Intercept)                            3.266753e-04
## gross_patient_rev_log                  1.641626e+01
## staffed_beds_log                       4.896600e+01
## gross_patient_rev_log:staffed_beds_log 1.109172e+00
```