analysis_6_code

May 2, 2019

In [164]:
```
library("dslabs")
library("dplyr")
library("Hmisc")
library("ggplot2")
library("DescTools")
library("aod")
library("car")
library("lmtest")
library("pROC")
library("xtable")
library("gam")
library("caret")
library("glmnet")
library("glmnetUtils")
library("pander")
library("lme4")
theme_update(plot.title = element_text(hjust = 0.5))
```

0.1 Read and Clean Data

In [165]:
```
framingham_data = read.csv("framingham_multi.csv")
# summary of data
summary(framingham_data)
```

```
     RANDID             SEX           TOTCHOL           AGE       
 Min.   :   6238   Min.   :1.000   Min.   :135.0   Min.   :33.00  
 1st Qu.: 641740   1st Qu.:1.000   1st Qu.:209.0   1st Qu.:47.00  
 Median :1213108   Median :2.000   Median :238.0   Median :54.00  
 Mean   :1232440   Mean   :1.532   Mean   :239.8   Mean   :54.11  
 3rd Qu.:1833807   3rd Qu.:2.000   3rd Qu.:266.0   3rd Qu.:61.00  
 Max.   :2460331   Max.   :2.000   Max.   :625.0   Max.   :79.00  
     SYSBP            DIABP           CURSMOKE        CIGPDAY      
 Min.   : 92.5   Min.   : 52.00   Min.   :0.000   Min.   : 0.000  
 1st Qu.:120.0   1st Qu.: 75.00   1st Qu.:0.000   1st Qu.: 0.000  
 Median :131.5   Median : 82.00   Median :0.000   Median : 0.000  
 Mean   :134.7   Mean   : 82.73   Mean   :0.416   Mean   : 8.187  
 3rd Qu.:147.1   3rd Qu.: 90.00   3rd Qu.:1.000   3rd Qu.:20.000  
```

```
 Max.   :246.0   Max.   :127.00   Max.   :1.000   Max.   :90.000
      BMI           DIABETES          HEARTRTE         GLUCOSE
 Min.   :15.16   Min.   :0.00000   Min.   : 50.00   Min.   : 45.00
 1st Qu.:23.09   1st Qu.:0.00000   1st Qu.: 68.00   1st Qu.: 73.00
 Median :25.16   Median :0.00000   Median : 75.00   Median : 80.00
 Mean   :25.61   Mean   :0.03667   Mean   : 76.76   Mean   : 84.47
 3rd Qu.:27.78   3rd Qu.:0.00000   3rd Qu.: 85.00   3rd Qu.: 91.00
 Max.   :52.94   Max.   :1.00000   Max.   :220.00   Max.   :420.00
      EDUC          PREVHYP           TIME          PERIOD
 Min.   :1.000   Min.   :0.000   Min.   :   0   Min.   :1
 1st Qu.:1.000   1st Qu.:0.000   1st Qu.:   0   1st Qu.:1
 Median :2.000   Median :0.000   Median :2177   Median :2
 Mean   :2.022   Mean   :0.454   Mean   :2177   Mean   :2
 3rd Qu.:3.000   3rd Qu.:1.000   3rd Qu.:4312   3rd Qu.:3
 Max.   :4.000   Max.   :1.000   Max.   :4607   Max.   :3
```

```
In [166]: describe(framingham_data)
```

```
framingham_data

 16  Variables      1500  Observations
--------------------------------------------------------------------------------
RANDID
       n  missing distinct     Info     Mean      Gmd      .05      .10
    1500        0      500        1  1232440   807581   126460   224615
     .25      .50      .75      .90      .95
  641740  1213108  1833807  2213833  2318599

lowest :    6238   11263   14367   16365   23727
highest: 2422371 2428234 2437351 2441847 2460331
--------------------------------------------------------------------------------
SEX
       n  missing distinct     Info     Mean      Gmd
    1500        0        2    0.747    1.532   0.4983

Value          1     2
Frequency    702   798
Proportion 0.468 0.532
--------------------------------------------------------------------------------
TOTCHOL
       n  missing distinct     Info     Mean      Gmd      .05      .10
    1500        0      210        1    239.8    49.42    171.0    185.0
     .25      .50      .75      .90      .95
   209.0    238.0    266.0    298.1    318.0

lowest : 135 139 141 142 145, highest: 386 390 398 415 625
--------------------------------------------------------------------------------
```

```
AGE
       n  missing distinct     Info     Mean      Gmd      .05      .10
    1500        0       47    0.999    54.11    11.07       39       41
     .25      .50      .75      .90      .95
      47       54       61       67       71

lowest : 33 34 35 36 37, highest: 75 76 77 78 79
------------------------------------------------------------------------------
SYSBP
       n  missing distinct     Info     Mean      Gmd      .05      .10
    1500        0      175        1    134.7     23.4    106.0    110.0
     .25      .50      .75      .90      .95
   120.0    131.5    147.1    163.0    173.0

lowest :  92.5  95.0  96.0  96.5  97.0, highest: 208.0 217.5 220.0 227.0 246.0
------------------------------------------------------------------------------
DIABP
       n  missing distinct     Info     Mean      Gmd      .05      .10
    1500        0      113    0.999    82.73    12.61       66       69
     .25      .50      .75      .90      .95
      75       82       90       98      102

lowest :  52.0  54.0  55.0  57.0  57.5, highest: 120.0 122.0 123.0 124.0 127.0
------------------------------------------------------------------------------
CURSMOKE
       n  missing distinct     Info      Sum     Mean      Gmd
    1500        0        2    0.729      624    0.416   0.4862

------------------------------------------------------------------------------
CIGPDAY
       n  missing distinct     Info     Mean      Gmd      .05      .10
    1500        0       32    0.798    8.187    11.87        0        0
     .25      .50      .75      .90      .95
       0        0       20       30       35

lowest :  0  1  2  3  4, highest: 43 45 50 60 90
------------------------------------------------------------------------------
BMI
       n  missing distinct     Info     Mean      Gmd      .05      .10
    1500        0      813        1    25.61    4.214    20.07    21.16
     .25      .50      .75      .90      .95
   23.09    25.16    27.78    30.29    32.35

lowest : 15.16 15.54 16.30 16.59 17.40, highest: 43.00 43.79 45.43 45.80 52.94
------------------------------------------------------------------------------
DIABETES
       n  missing distinct     Info      Sum     Mean      Gmd
    1500        0        2    0.106       55  0.03667  0.07069
```

```
---------------------------------------------------------------------------
HEARTRTE
       n  missing distinct     Info     Mean      Gmd      .05      .10
    1500        0       62    0.996    76.76    14.15       60       60
     .25      .50      .75      .90      .95
      68       75       85       95      100

lowest :  50  52  53  54  55, highest: 115 120 125 150 220
---------------------------------------------------------------------------
GLUCOSE
       n  missing distinct     Info     Mean      Gmd      .05      .10
    1500        0      115    0.999    84.47    19.66       63       66
     .25      .50      .75      .90      .95
      73       80       91      105      116

lowest :  45  47  49  50  51, highest: 230 265 300 334 420
---------------------------------------------------------------------------
EDUC
       n  missing distinct     Info     Mean      Gmd
    1500        0        4    0.901    2.022      1.1

Value          1     2     3     4
Frequency    582   486   249   183
Proportion 0.388 0.324 0.166 0.122
---------------------------------------------------------------------------
PREVHYP
       n  missing distinct     Info      Sum     Mean      Gmd
    1500        0        2    0.744      681    0.454   0.4961

---------------------------------------------------------------------------
TIME
       n  missing distinct     Info     Mean      Gmd      .05      .10
    1500        0      411    0.963     2177     1954        0        0
     .25      .50      .75      .90      .95
       0     2177     4312     4392     4423

lowest :    0 1799 1847 1891 1914, highest: 4564 4573 4580 4582 4607
---------------------------------------------------------------------------
PERIOD
       n  missing distinct     Info     Mean      Gmd
    1500        0        3    0.889        2   0.8895

Value          1     2     3
Frequency    500   500   500
Proportion 0.333 0.333 0.333
---------------------------------------------------------------------------
```

```
In [264]: framingham_data_clean = framingham_data
          framingham_data_clean$SEX = as.factor(framingham_data_clean$SEX)
          framingham_data_clean$CURSMOKE = as.factor(framingham_data_clean$CURSMOKE)
          framingham_data_clean$DIABETES = as.factor(framingham_data_clean$DIABETES)
          framingham_data_clean$PREVHYP = as.factor(framingham_data_clean$PREVHYP)
          framingham_data_clean$EDUC = as.factor(framingham_data_clean$EDUC)
          framingham_data_clean$PERIOD = as.factor(framingham_data_clean$PERIOD)
```

```
In [168]: train_idx = createDataPartition(framingham_data_clean$PREVHYP, p=0.75, list=F)[,1]
          train_data = framingham_data_clean[train_idx,]
          test_data = framingham_data_clean[-train_idx,]
```

0.2 EDA

```
In [169]: # tables of each input feature with the response
          chisq_cont_table = function(feature, row_names) {
              tbl = table(feature, framingham_data_clean$PREVHYP)
              colnames(tbl) = c("free of disease", "prevalent disease")
              rownames(tbl) = row_names
              print(tbl)
              print(chisq.test(tbl))
              print(pandoc.table(tbl))
          }
```

```
In [170]: chisq_cont_table(framingham_data_clean$SEX, c("Male", "Female"))
```

```
feature  free of disease prevalent disease
  Male               366               336
  Female             453               345

        Pearson's Chi-squared test with Yates' continuity correction

data:  tbl
X-squared = 3.0458, df = 1, p-value = 0.08094



|        | free of disease | prevalent disease |
|:----------:|:---------------:|:-----------------:|
|  **Male**  |       366       |        336        |
| **Female** |       453       |        345        |

NULL
```

```
In [171]: chisq_cont_table(framingham_data_clean$CURSMOKE, c("Yes", "No"))
```

```
feature free of disease prevalent disease
```

```
    Yes             420               456
    No              399               225

        Pearson's Chi-squared test with Yates' continuity correction

data:  tbl
X-squared = 36.979, df = 1, p-value = 1.194e-09
```

```
|    | free of disease | prevalent disease |
|:-------:|:---------------:|:-----------------:|
| **Yes** |       420       |        456        |
| **No**  |       399       |        225        |
```

```
NULL
```

```
In [172]: chisq_cont_table(framingham_data_clean$DIABETES, c("Not Diabetic", "Diabetic"))
```

```
feature        free of disease prevalent disease
  Not Diabetic             811               634
  Diabetic                   8                47

        Pearson's Chi-squared test with Yates' continuity correction

data:  tbl
X-squared = 35.294, df = 1, p-value = 2.835e-09
```

```
|             | free of disease | prevalent disease |
|:----------------:|:---------------:|:-----------------:|
| **Not Diabetic** |       811       |        634        |
|   **Diabetic**   |        8        |        47         |
```

```
NULL
```

```
In [173]: chisq_cont_table(framingham_data_clean$EDUC, c("0-11 years", "High School Diploma or
```

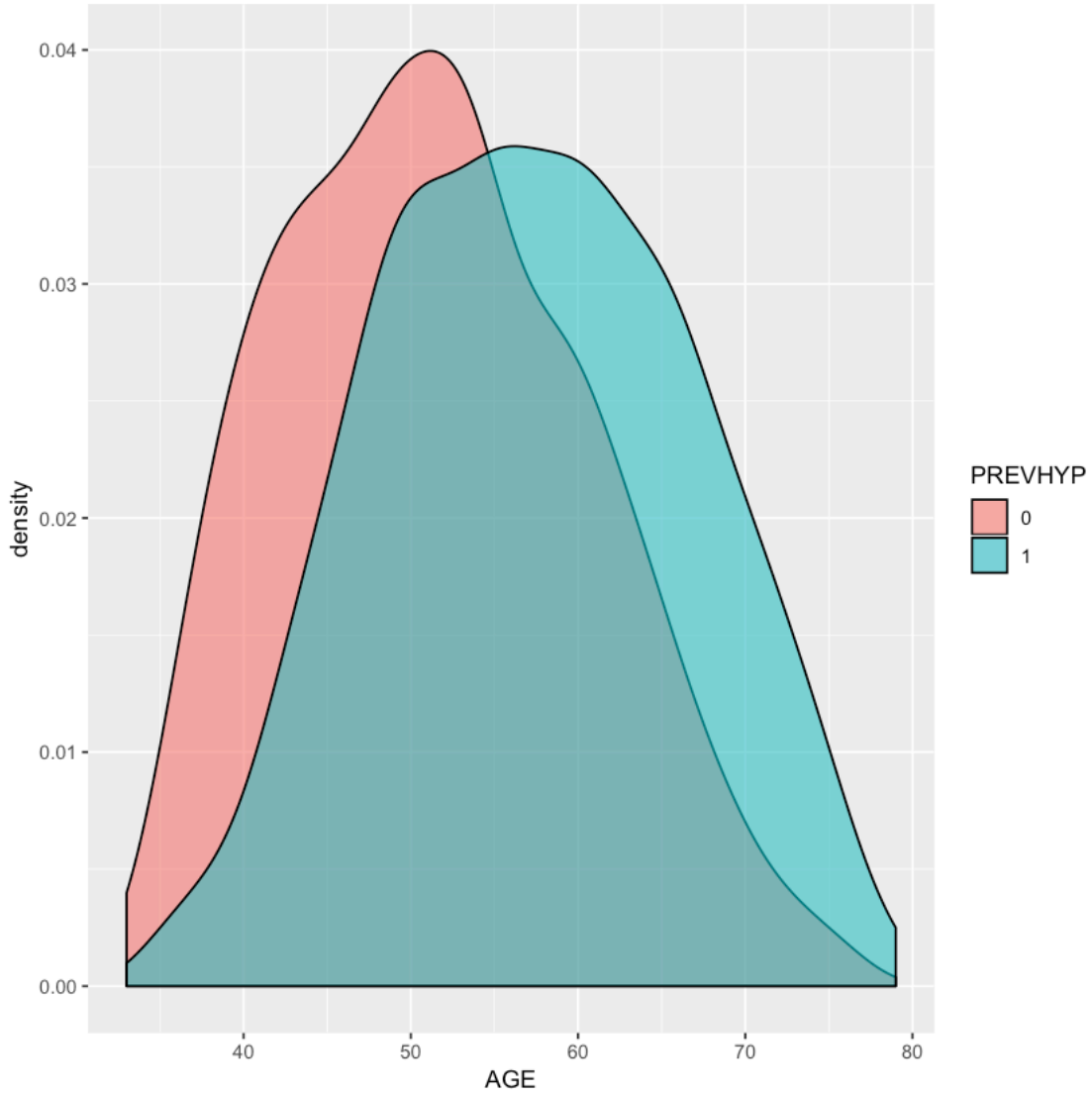
```
feature                             free of disease prevalent disease
  0-11 years                                    287               295
  High School Diploma or GED                    274               212
  Some College or Vocational School             153                96
  College degree or more                        105                78
```

```
        Pearson's Chi-squared test

data:  tbl
X-squared = 12.461, df = 3, p-value = 0.005961
```

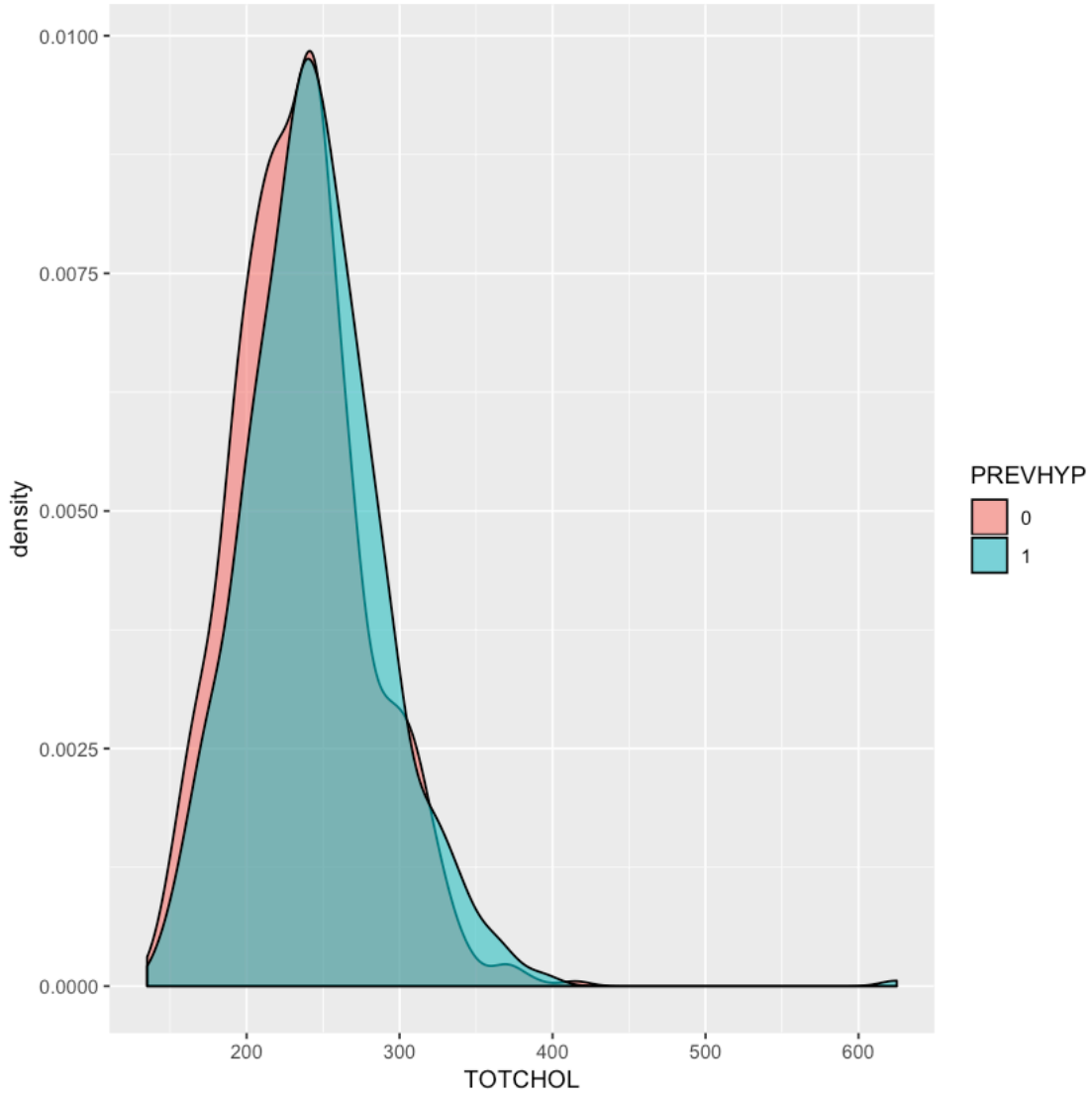
	free of disease	prevalent disease
0-11 years	287	295
High School Diploma or GED	274	212
Some College or Vocational School	153	96
College degree or more	105	78

```
NULL
```

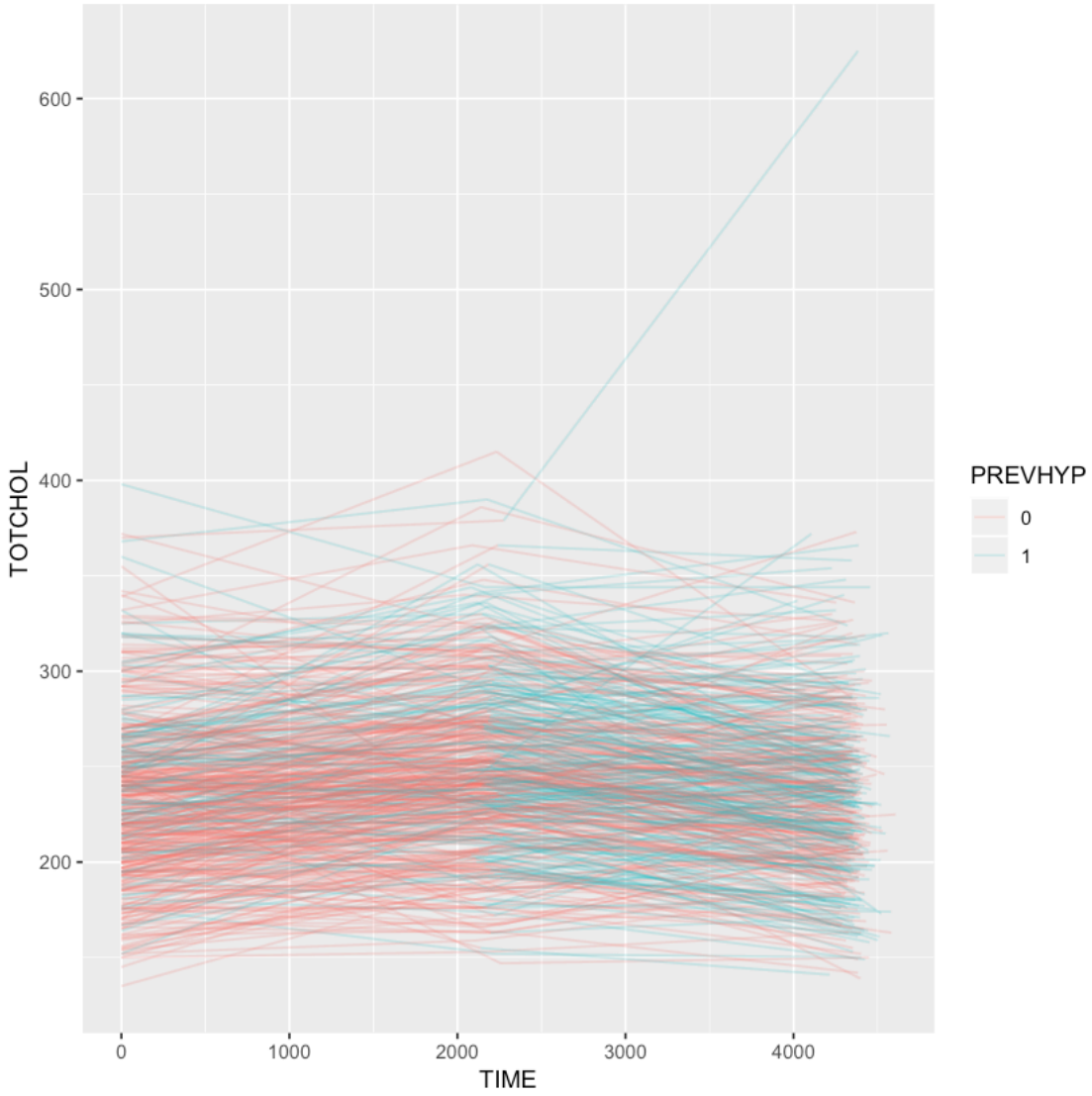
```
In [175]: ggplot(data=framingham_data_clean, aes(x=AGE, fill=PREVHYP)) +
              geom_density(alpha=.6)
```



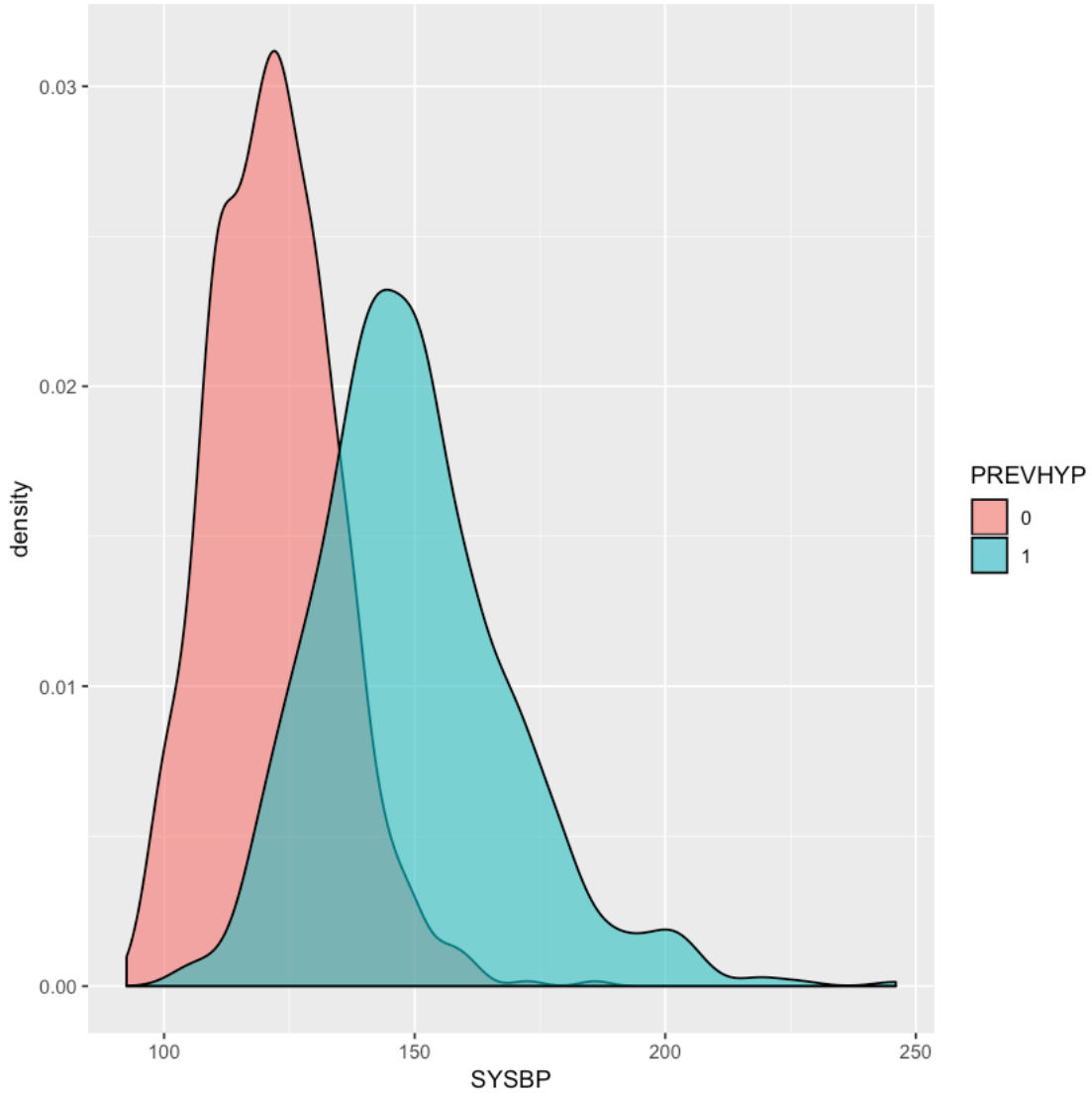

```
In [176]: ggplot(data=framingham_data_clean, aes(x=TOTCHOL, fill=PREVHYP)) +
              geom_density(alpha=.6)
```



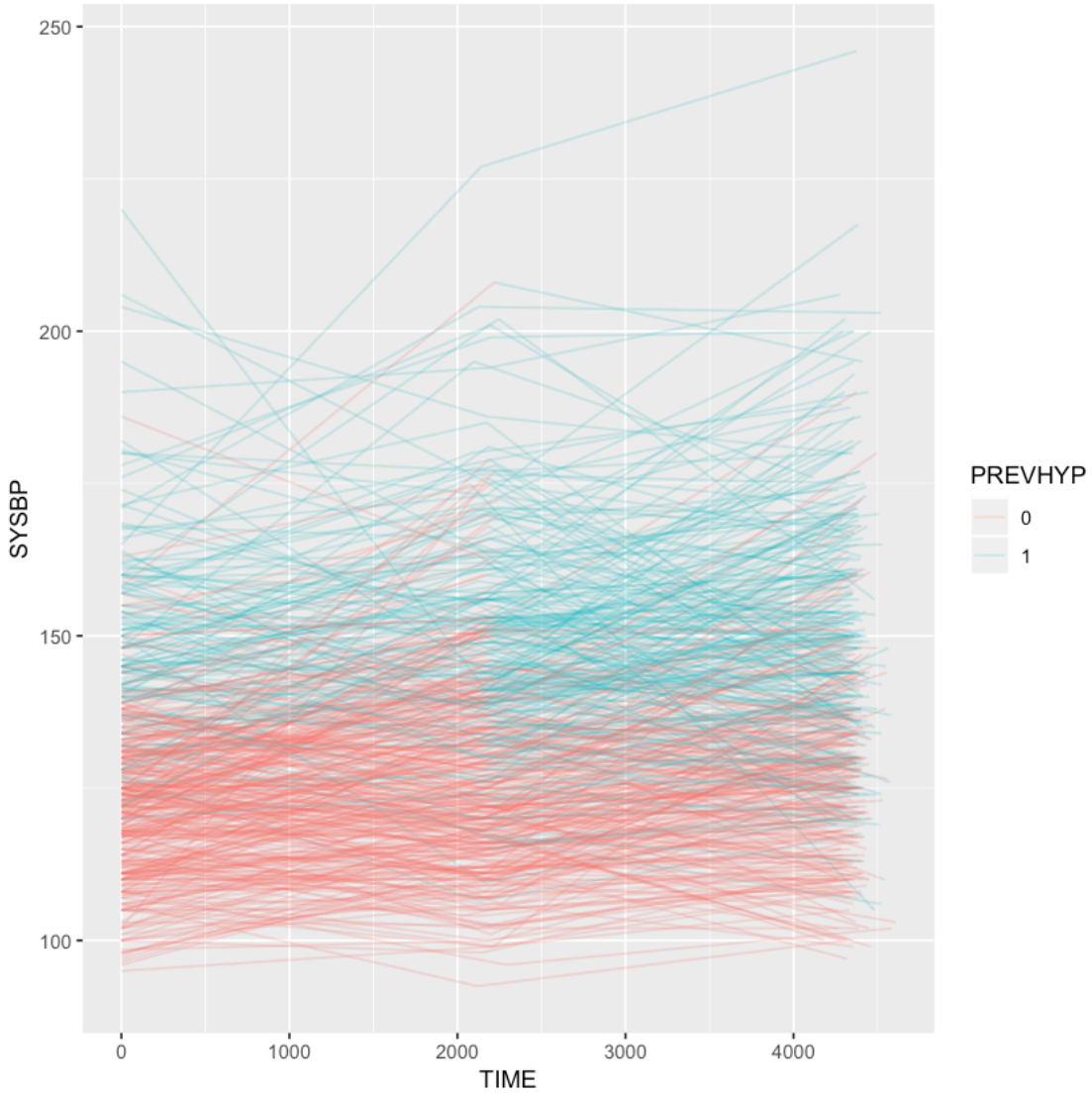

```
In [177]: ggplot(data=framingham_data_clean, aes(x=TIME, y=TOTCHOL, group=RANDID, color=PREVHYP
              geom_line(alpha=.2)
```



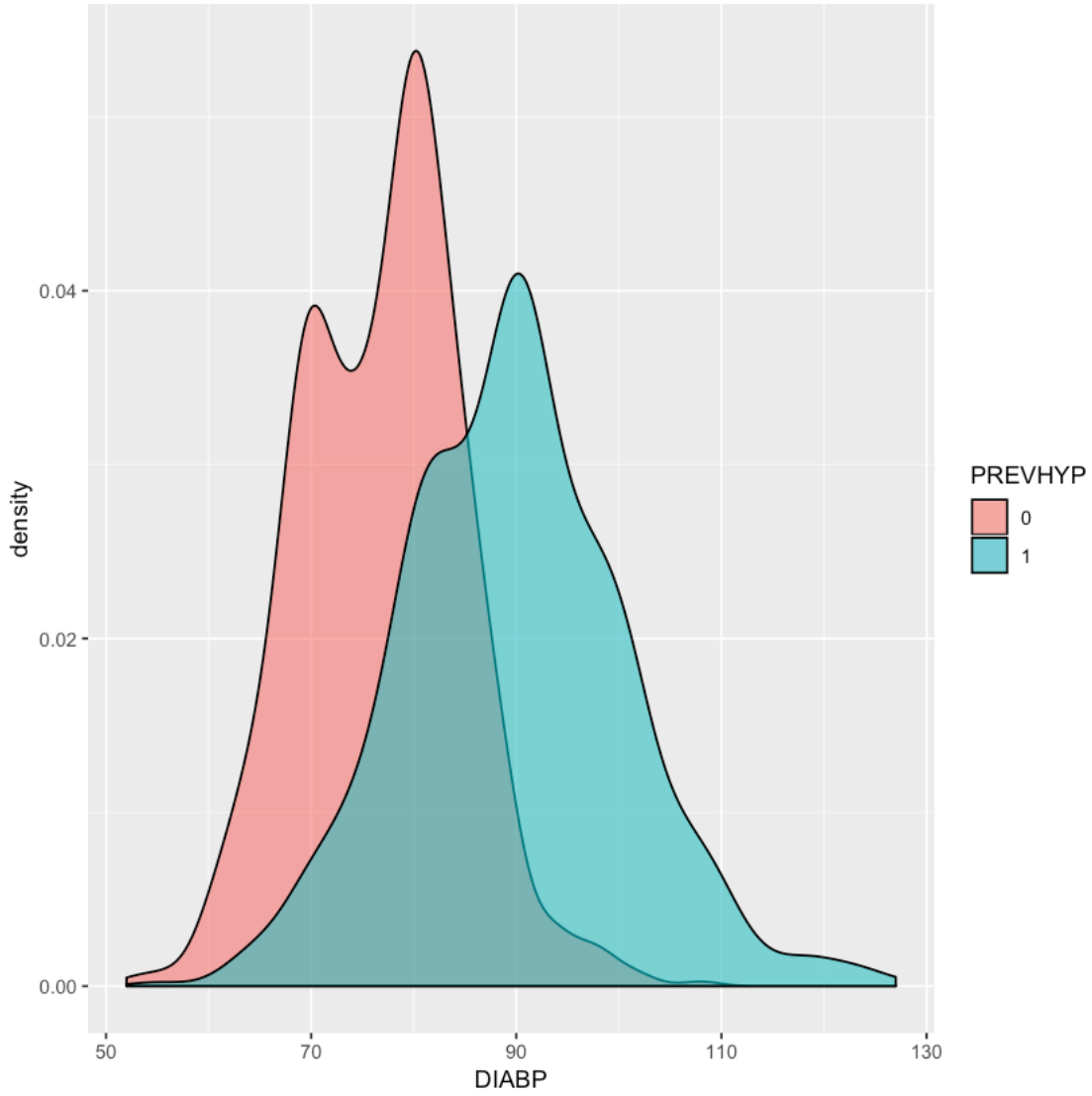

```
In [178]: ggplot(data=framingham_data_clean, aes(x=SYSBP, fill=PREVHYP)) +
              geom_density(alpha=.6)
```



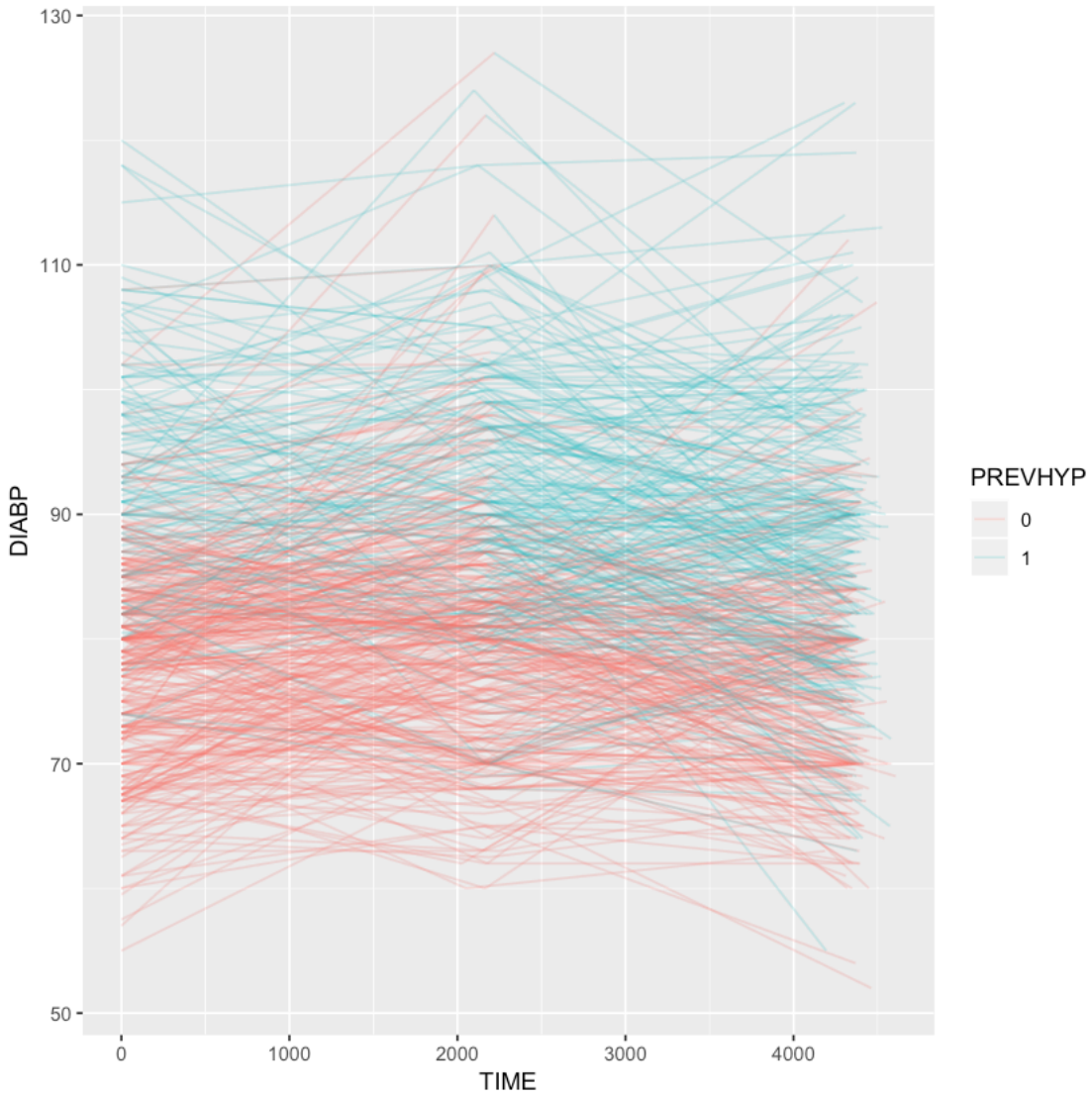

```
In [181]: ggplot(data=framingham_data_clean, aes(x=TIME, y=SYSBP, group=RANDID, color=PREVHYP))
              geom_line(alpha=.2)
```



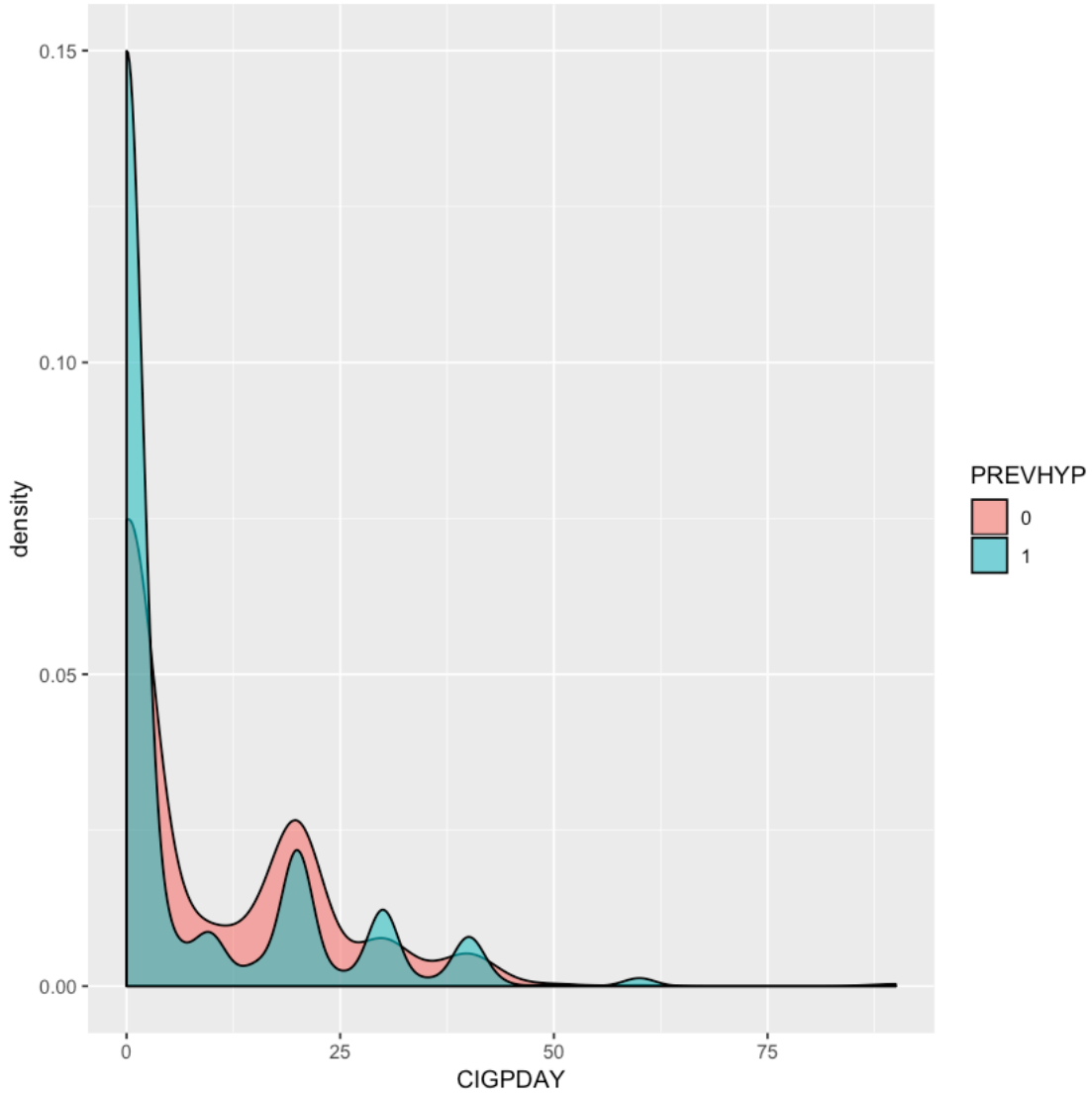

```
In [64]: ggplot(data=framingham_data_clean, aes(x=DIABP, fill=PREVHYP)) +
             geom_density(alpha=.6)
```



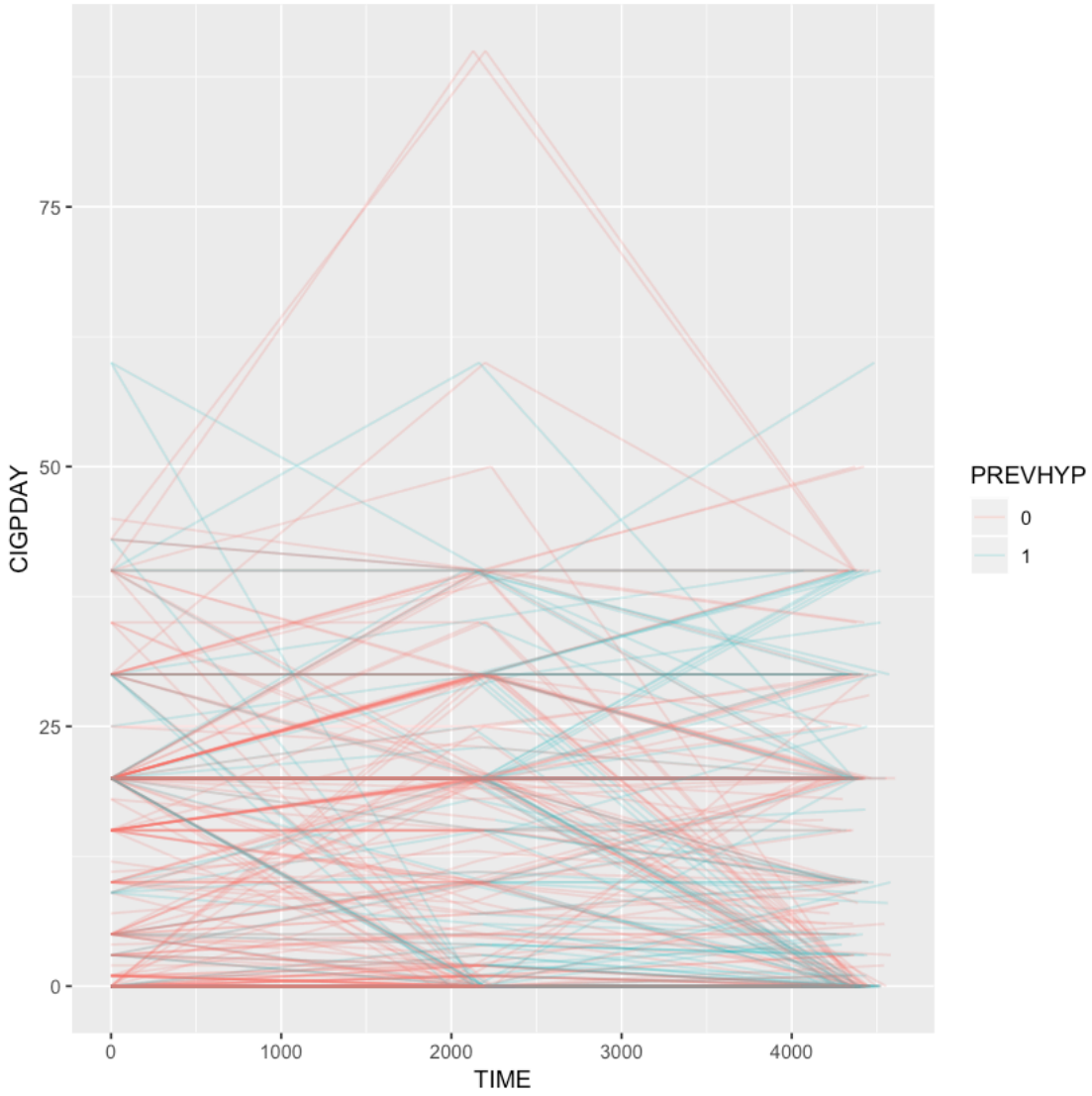

```
In [182]: ggplot(data=framingham_data_clean, aes(x=TIME, y=DIABP, group=RANDID, color=PREVHYP))
              geom_line(alpha=.2)
```



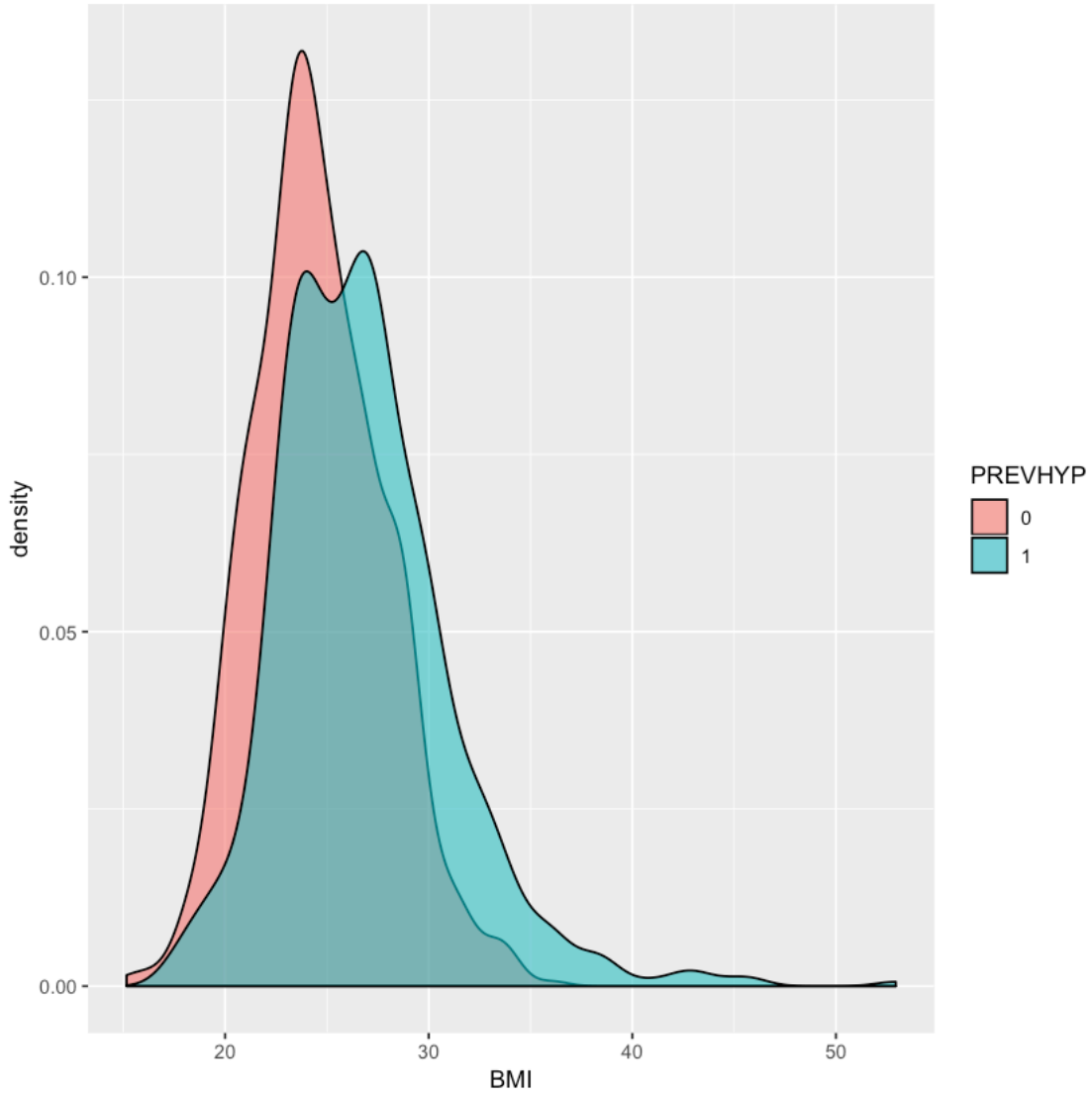

```
In [23]: ggplot(data=framingham_data_clean, aes(x=CIGPDAY, fill=PREVHYP)) +
             geom_density(alpha=.6)
```



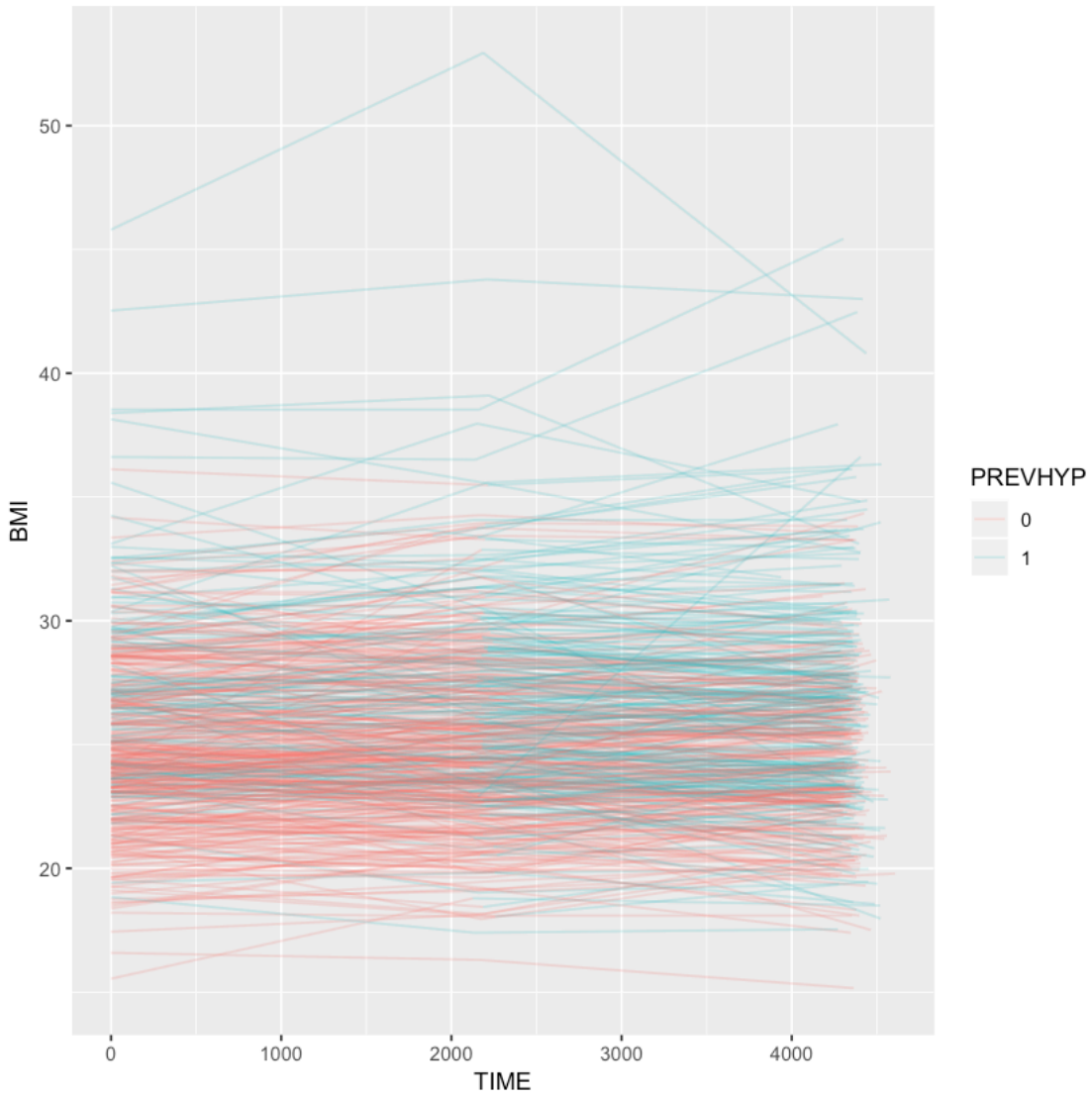

```
In [183]: ggplot(data=framingham_data_clean, aes(x=TIME, y=CIGPDAY, group=RANDID, color=PREVHYP
              geom_line(alpha=.2)
```



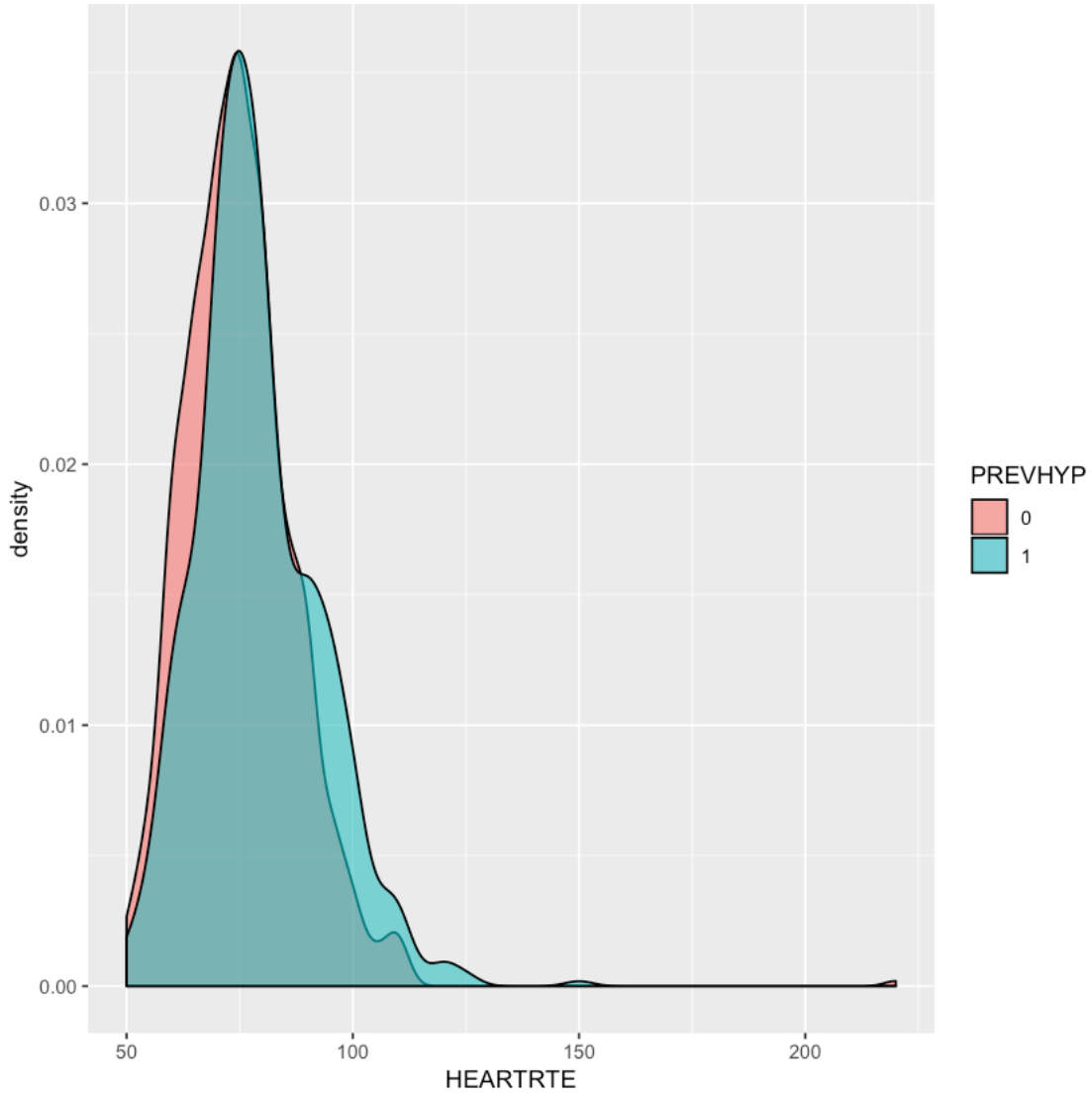

```
In [24]: ggplot(data=framingham_data_clean, aes(x=BMI, fill=PREVHYP)) +
             geom_density(alpha=.6)
```



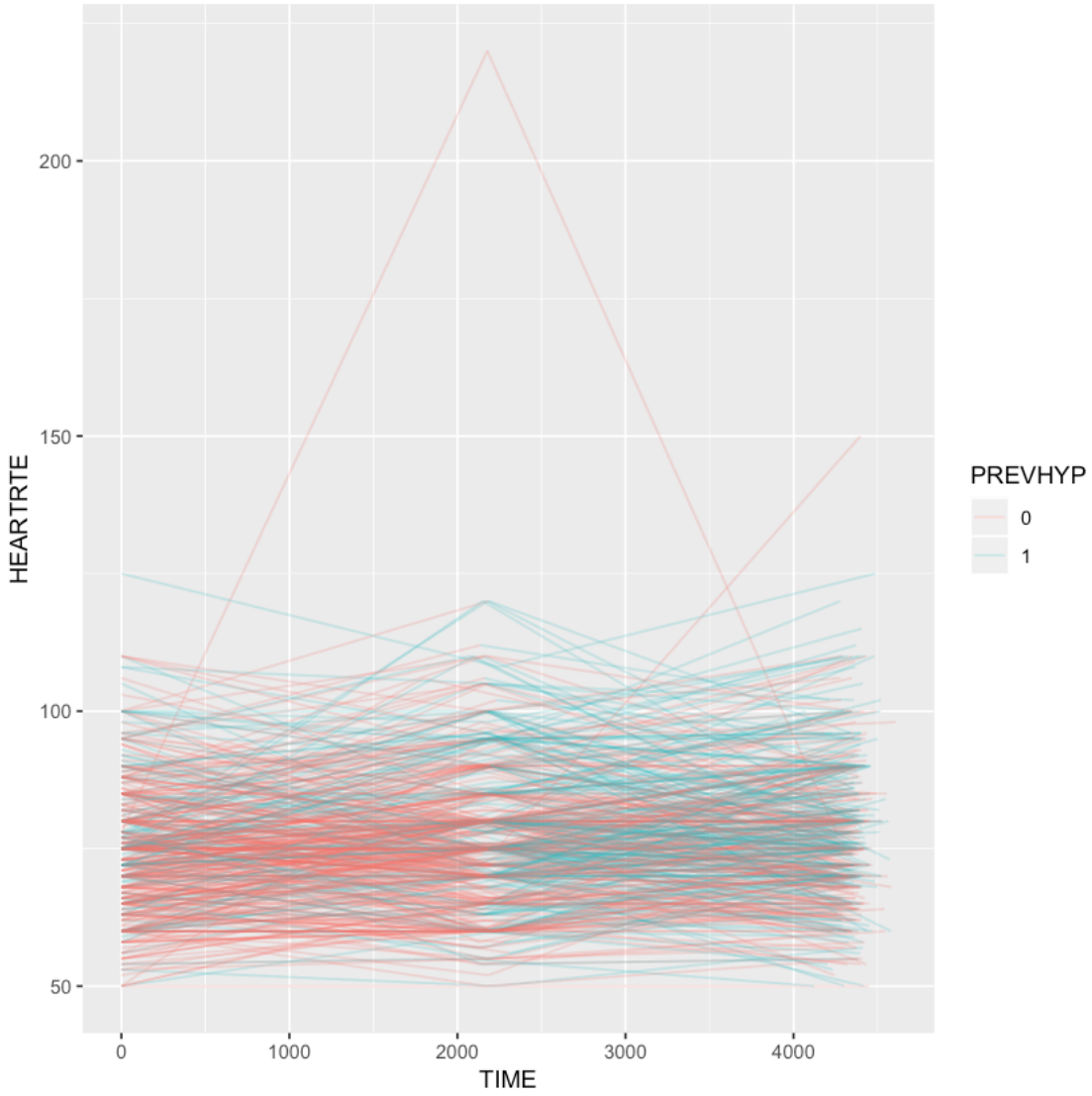

```
In [184]: ggplot(data=framingham_data_clean, aes(x=TIME, y=BMI, group=RANDID, color=PREVHYP)) +
              geom_line(alpha=.2)
```



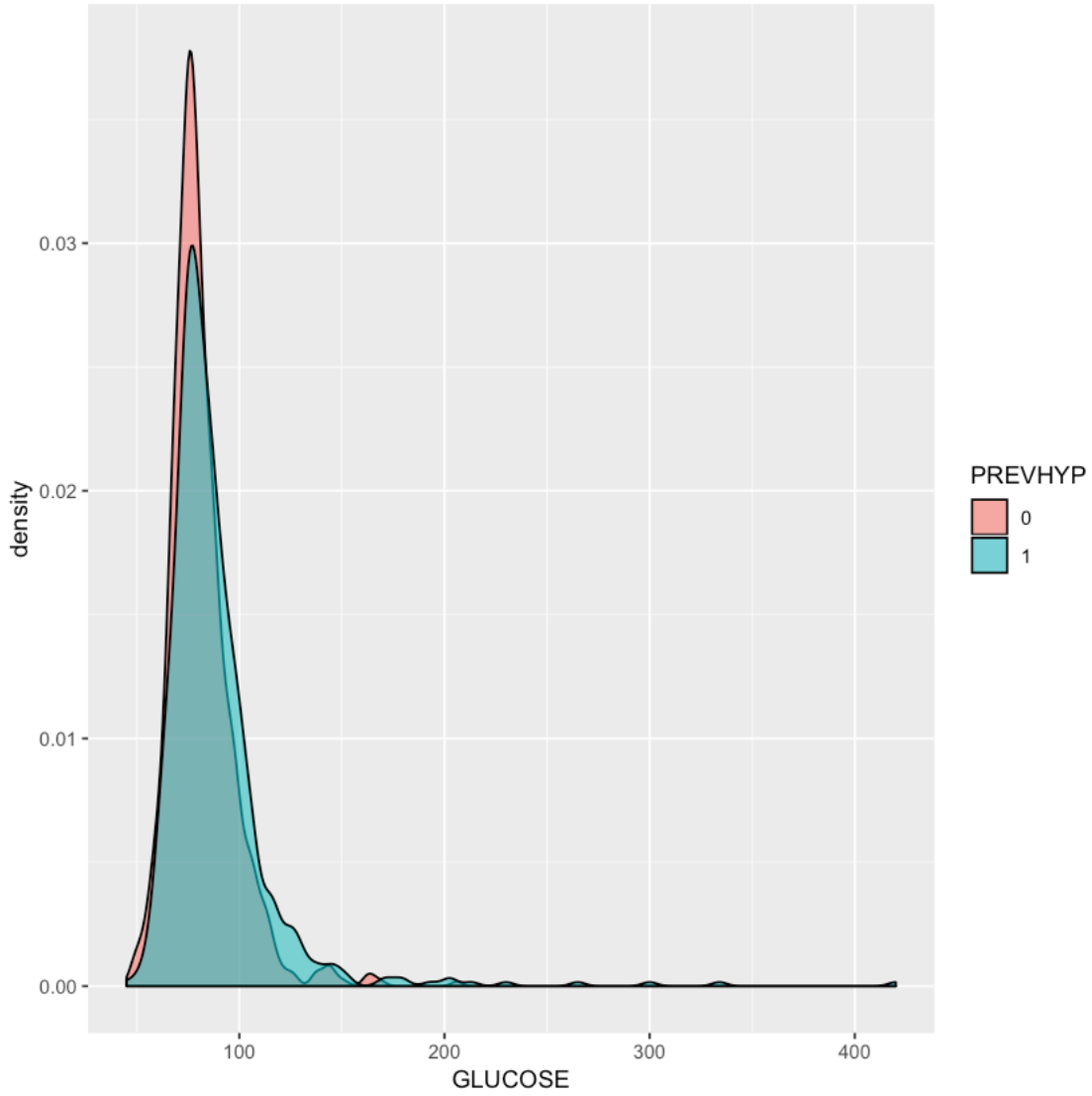

```
In [25]: ggplot(data=framingham_data_clean, aes(x=HEARTRTE, fill=PREVHYP)) +
             geom_density(alpha=.6)
```



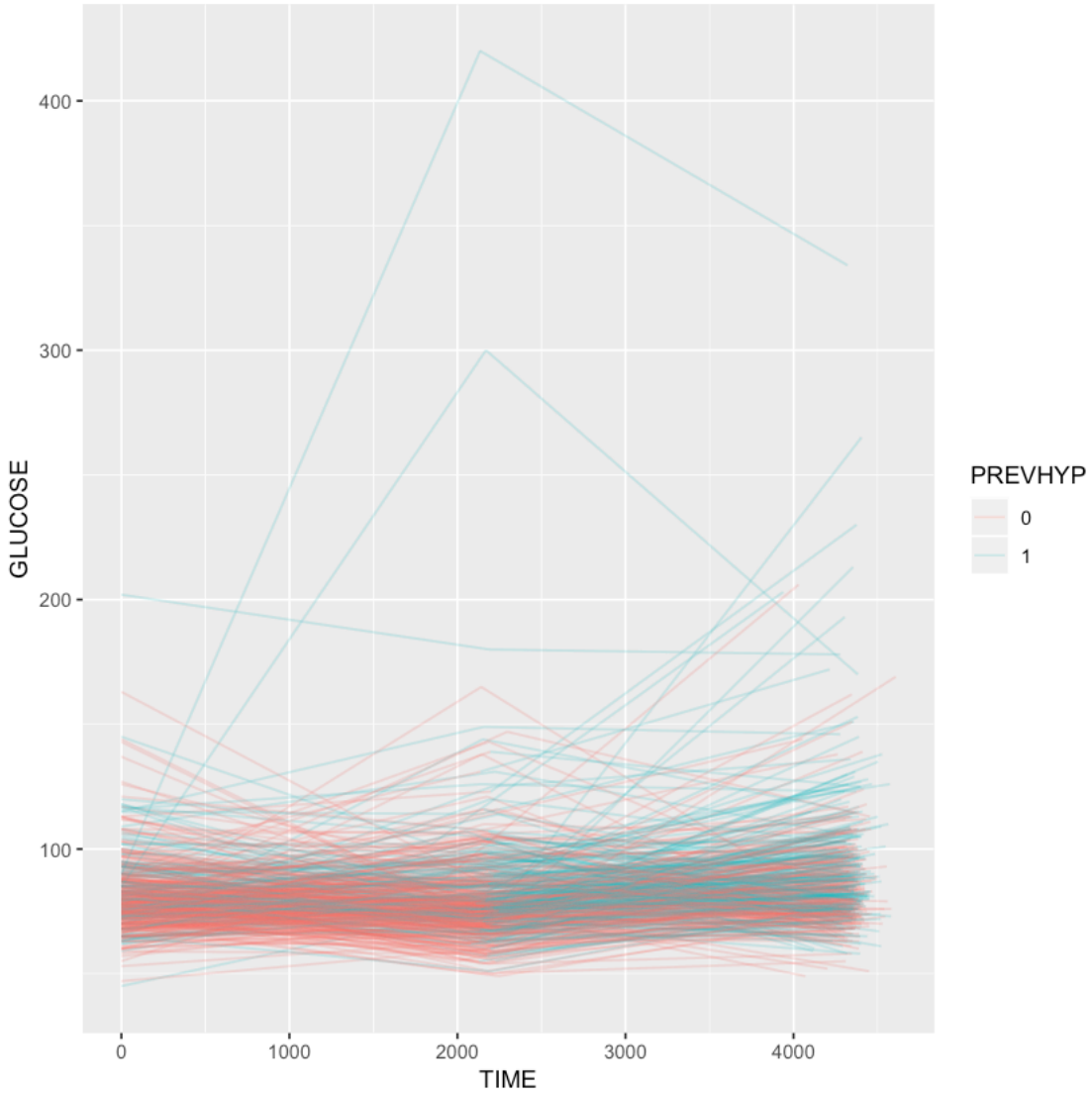

```
In [185]: ggplot(data=framingham_data_clean, aes(x=TIME, y=HEARTRTE, group=RANDID, color=PREVHY
              geom_line(alpha=.2)
```



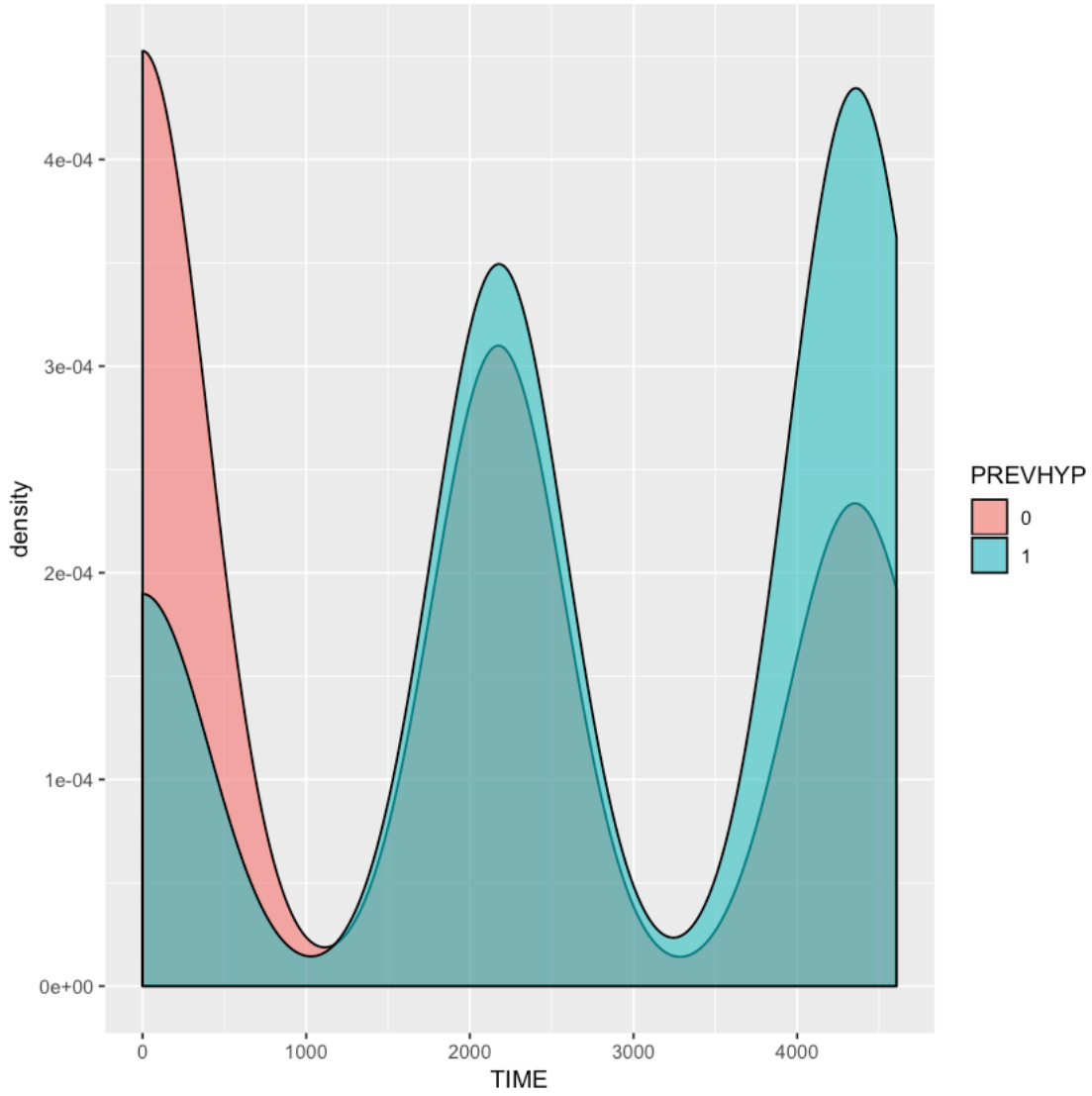

```
In [26]: ggplot(data=framingham_data_clean, aes(x=GLUCOSE, fill=PREVHYP)) +
             geom_density(alpha=.6)
```



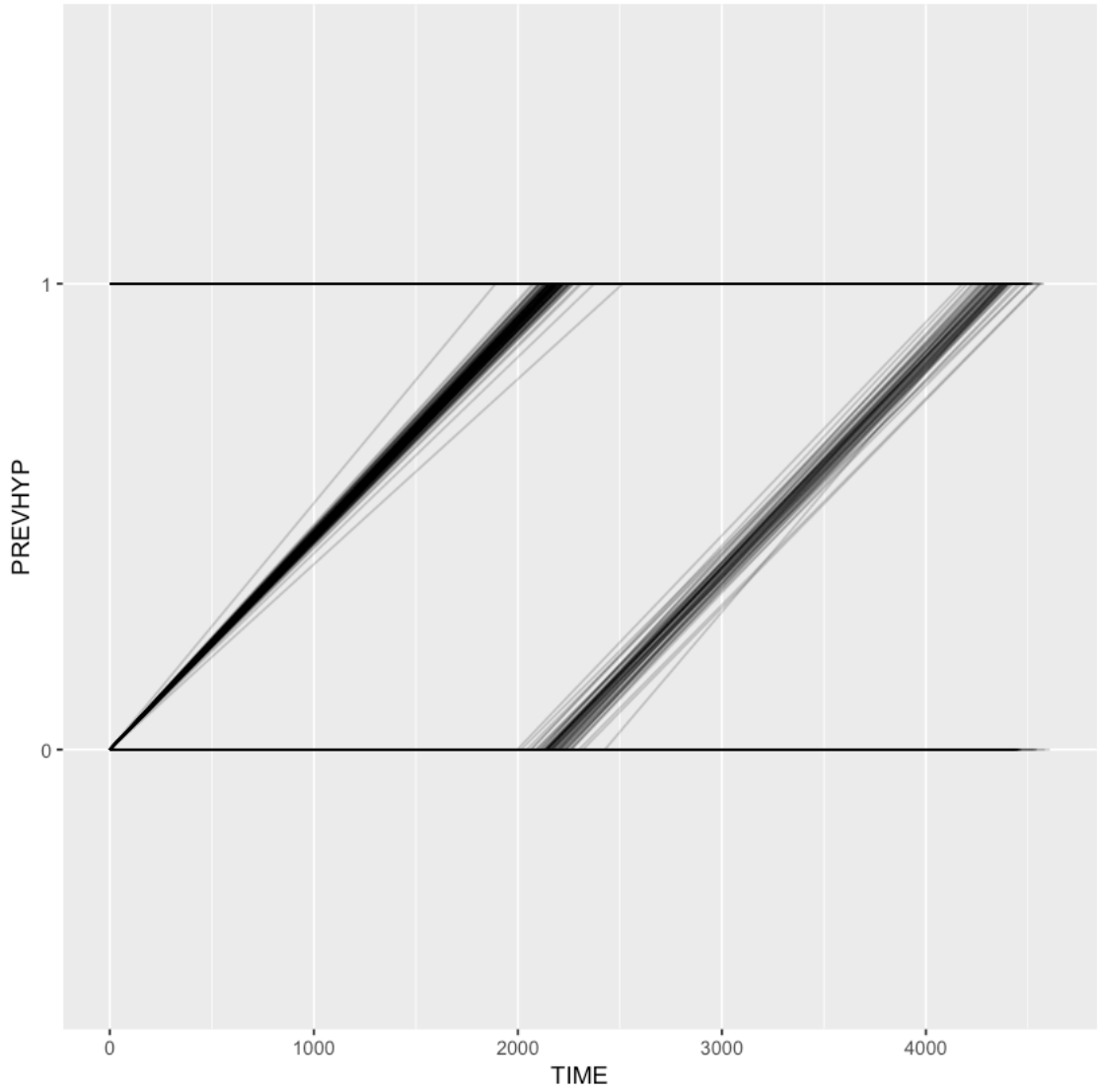

```
In [186]: ggplot(data=framingham_data_clean, aes(x=TIME, y=GLUCOSE, group=RANDID, color=PREVHYP
              geom_line(alpha=.2)
```




```
In [27]: ggplot(data=framingham_data_clean, aes(x=TIME, fill=PREVHYP)) +
             geom_density(alpha=.6)
```




```
In [188]: ggplot(data=framingham_data_clean, aes(x=TIME, y=PREVHYP, group=RANDID)) +
              geom_line(alpha=.2)
```



0.3 Model Fitting

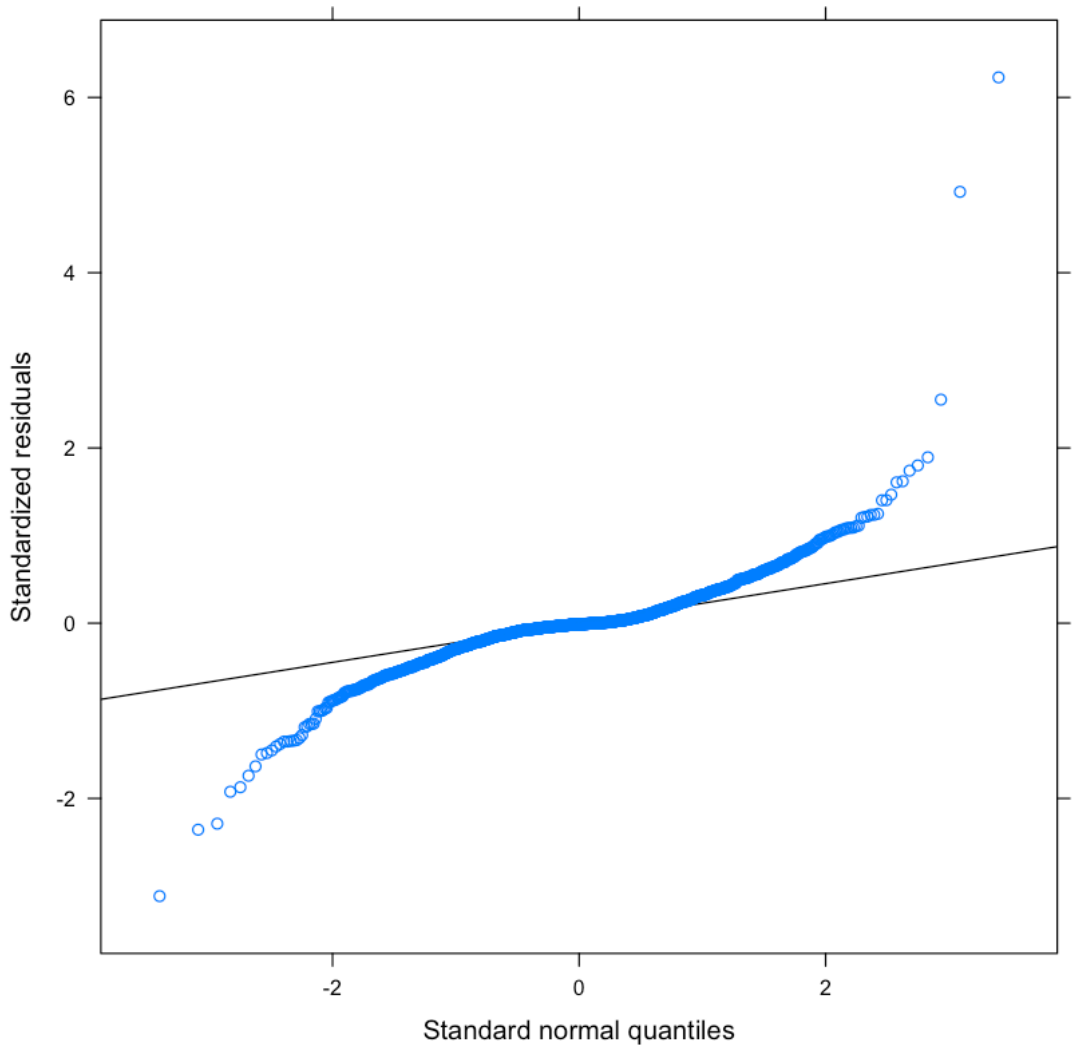
```
In [267]: # all variables, random intercept
          model_1 = glmer(PREVHYP ~ SEX + TOTCHOL + AGE + SYSBP + DIABP + CURSMOKE + CIGPDAY +
                          data=framingham_data_clean,
                          family=binomial)
          summary(model_1)
```

```
Warning message:
Some predictor variables are on very different scales: consider rescalingWarning message in che
unable to evaluate scaled gradientWarning message in checkConv(attr(opt, "derivs"), opt$par, ct
Model failed to converge: degenerate  Hessian with 1 negative eigenvaluesWarning message in vco
variance-covariance matrix computed from finite-difference Hessian is
```

```
not positive definite or contains NA values: falling back to var-cov estimated from RXWarning m
variance-covariance matrix computed from finite-difference Hessian is
not positive definite or contains NA values: falling back to var-cov estimated from RX
Correlation matrix not shown by default, as p = 18 > 12.
Use print(obj, correlation=TRUE)  or
    vcov(obj)        if you need it



Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: PREVHYP ~ SEX + TOTCHOL + AGE + SYSBP + DIABP + CURSMOKE + CIGPDAY +
    BMI + DIABETES + HEARTRTE + GLUCOSE + EDUC + TIME + PERIOD +
    (1 | RANDID)
   Data: framingham_data_clean

     AIC      BIC   logLik deviance df.resid
   952.4   1053.4   -457.2    914.4     1481

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.1150 -0.1491 -0.0124  0.1540  6.2290

Random effects:
 Groups Name        Variance Std.Dev.
 RANDID (Intercept) 5.652    2.377
Number of obs: 1500, groups:  RANDID, 500

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -37.576625   2.937844 -12.791  < 2e-16 ***
SEX2         -0.492451   0.341600  -1.442   0.1494
TOTCHOL       0.001272   0.003520   0.361   0.7178
AGE           0.031290   0.022034   1.420   0.1556
SYSBP         0.136116   0.013869   9.814  < 2e-16 ***
DIABP         0.103752   0.021319   4.867 1.14e-06 ***
CURSMOKE1    -0.659590   0.452927  -1.456   0.1453
CIGPDAY      -0.002847   0.017957  -0.159   0.8740
BMI           0.183682   0.045491   4.038 5.40e-05 ***
DIABETES1     1.768149   1.024053   1.727   0.0842 .
HEARTRTE      0.019370   0.011318   1.712   0.0870 .
GLUCOSE       0.007604   0.007423   1.024   0.3056
EDUC2        -0.082261   0.400183  -0.206   0.8371
EDUC3        -0.300380   0.472999  -0.635   0.5254
EDUC4        -0.115724   0.543927  -0.213   0.8315
TIME          0.001433   0.002242   0.639   0.5228
PERIOD2      -0.870727   4.895082  -0.178   0.8588
```

```
PERIOD3      -2.860799   9.782842  -0.292   0.7700
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
fit warnings:
Some predictor variables are on very different scales: consider rescaling
convergence code: 0
unable to evaluate scaled gradient
Model failed to converge: degenerate  Hessian with 1 negative eigenvalues
```

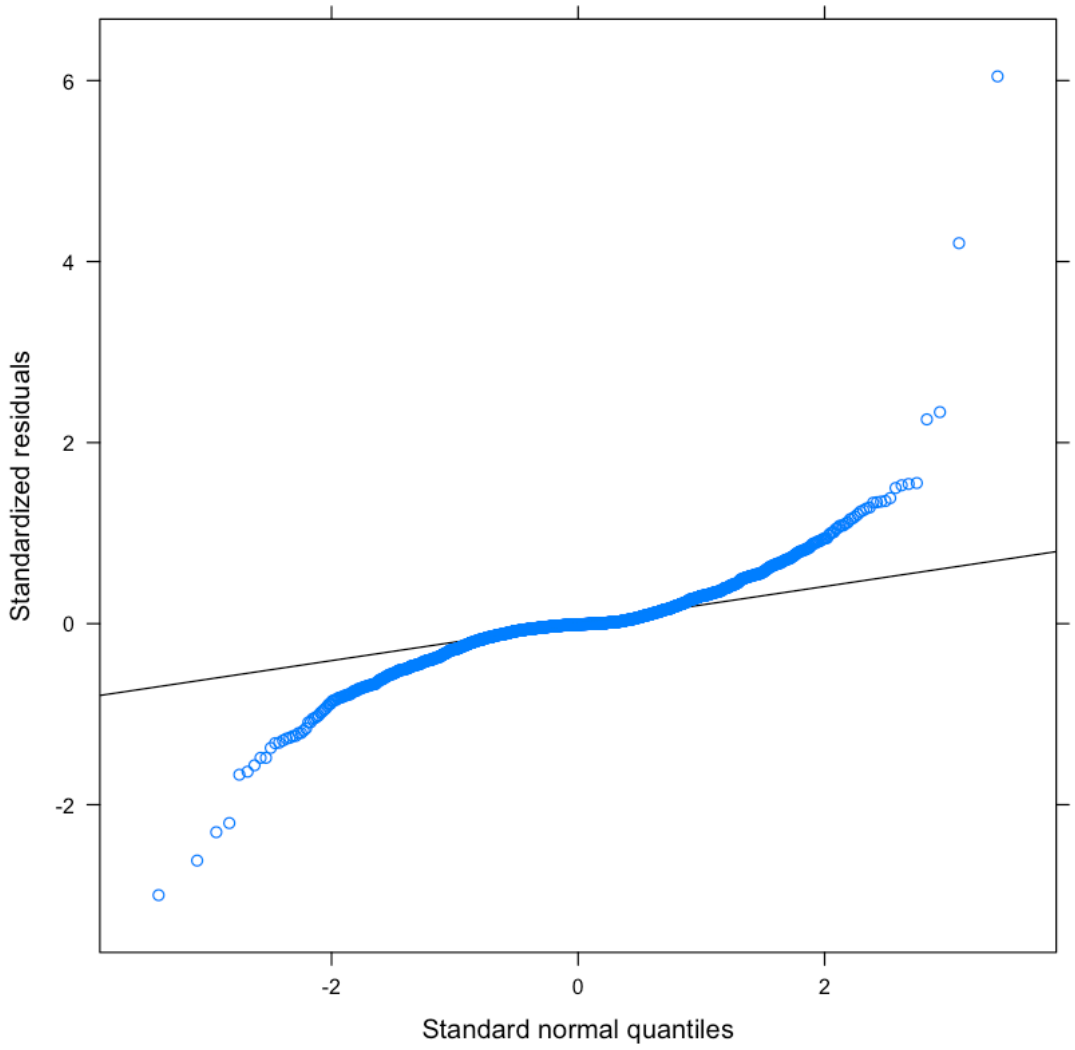
```
In [268]: qqmath(model_1)
```



```
In [269]: # all variables, random intercept, no period as it is just a categorical of time
          model_2 = glmer(PREVHYP ~ SEX + TOTCHOL + AGE + SYSBP + DIABP + CURSMOKE + CIGPDAY +
                          data=framingham_data_clean,
                          family=binomial)
          summary(model_2)

Warning message:
Some predictor variables are on very different scales: consider rescalingWarning message in ch
Model failed to converge with max|grad| = 39.8421 (tol = 0.001, component 1)Warning message in
Model is nearly unidentifiable: very large eigenvalue
 - Rescale variables?;Model is nearly unidentifiable: large eigenvalue ratio
 - Rescale variables?
Correlation matrix not shown by default, as p = 16 > 12.
Use print(obj, correlation=TRUE)  or
    vcov(obj)        if you need it



Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: PREVHYP ~ SEX + TOTCHOL + AGE + SYSBP + DIABP + CURSMOKE + CIGPDAY +
    BMI + DIABETES + HEARTRTE + GLUCOSE + EDUC + TIME + (1 |      RANDID)
   Data: framingham_data_clean

     AIC      BIC   logLik deviance df.resid
   950.3   1040.6   -458.1    916.3     1483

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.9992 -0.1366 -0.0087  0.1399  6.0459

Random effects:
 Groups Name        Variance Std.Dev.
 RANDID (Intercept) 6.491    2.548
Number of obs: 1500, groups:  RANDID, 500

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.236e+01  5.261e+00  -8.051 8.20e-16 ***
SEX2        -8.678e-01  3.858e-01  -2.249 0.024514 *
TOTCHOL      3.201e-03  3.520e-03   0.909 0.363100
AGE          7.351e-02  2.785e-02   2.639 0.008312 **
SYSBP        1.268e-01  1.617e-02   7.843 4.40e-15 ***
DIABP        1.470e-01  2.551e-02   5.762 8.34e-09 ***
CURSMOKE1   -4.321e-01  4.748e-01  -0.910 0.362865
CIGPDAY     -1.023e-02  1.884e-02  -0.543 0.587180
BMI          1.859e-01  5.276e-02   3.524 0.000425 ***
```

```
DIABETES1    1.027e+00  9.989e-01   1.028 0.303721
HEARTRTE     1.980e-02  1.216e-02   1.628 0.103494
GLUCOSE      5.534e-03  6.650e-03   0.832 0.405270
EDUC2        8.611e-02  4.145e-01   0.208 0.835402
EDUC3        2.130e-01  4.951e-01   0.430 0.667087
EDUC4        7.400e-02  5.682e-01   0.130 0.896387
TIME         7.399e-04  1.251e-04   5.914 3.34e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
fit warnings:
Some predictor variables are on very different scales: consider rescaling
convergence code: 0
Model failed to converge with max|grad| = 39.8421 (tol = 0.001, component 1)
Model is nearly unidentifiable: very large eigenvalue
 - Rescale variables?
Model is nearly unidentifiable: large eigenvalue ratio
 - Rescale variables?
```

```
In [270]: qqmath(model_2)
```

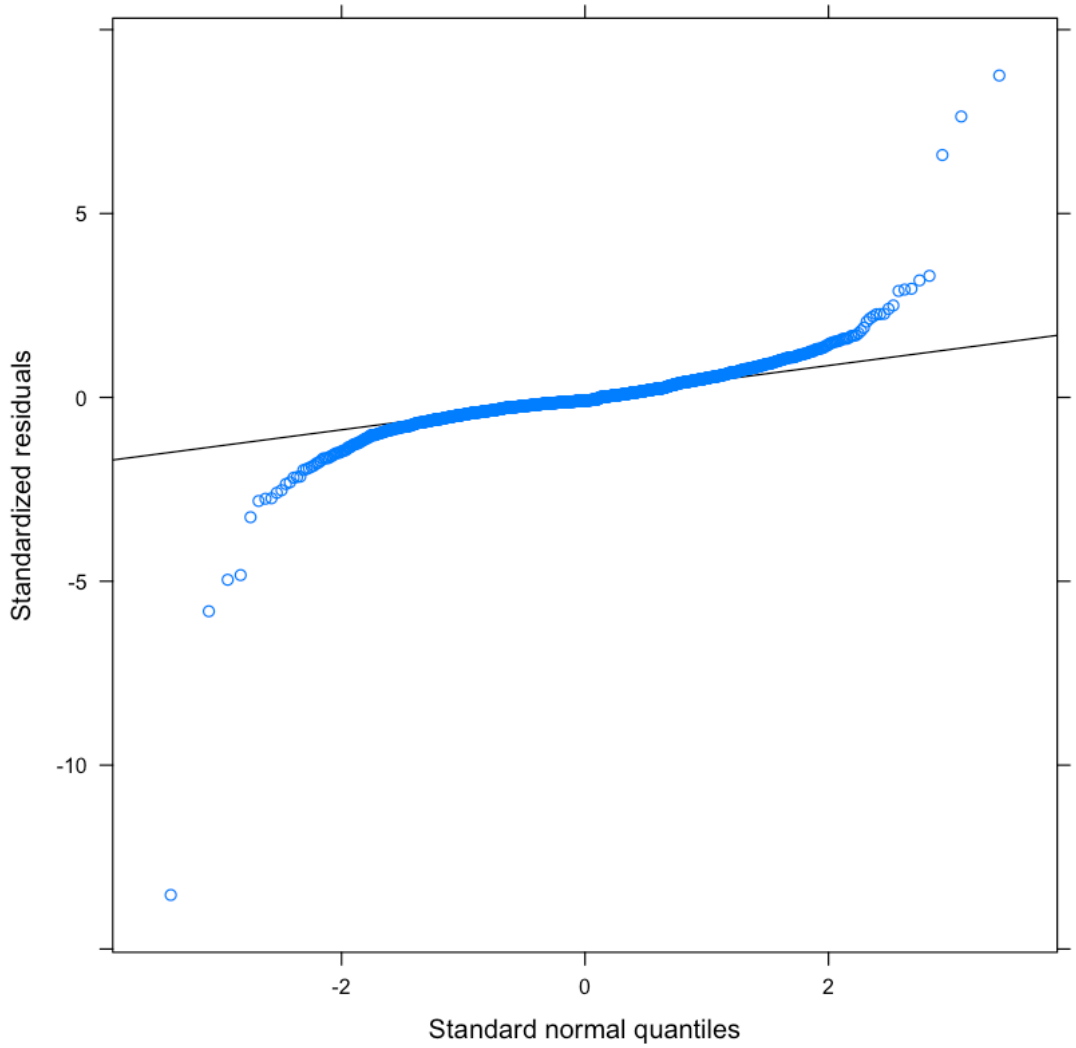



```
In [271]: # all with significant p values below .05 and trajectories that look different in th
          model_3 = glmer(PREVHYP ~ CIGPDAY + SYSBP + CURSMOKE + DIABETES + EDUC + (1 | RANDID)
                          data=framingham_data_clean,
                          family=binomial)
          summary(model_3)
```

```
Warning message in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
Model failed to converge with max|grad| = 3.80408 (tol = 0.001, component 1)Warning message in
Model is nearly unidentifiable: very large eigenvalue
 - Rescale variables?;Model is nearly unidentifiable: large eigenvalue ratio
 - Rescale variables?

Generalized linear mixed model fit by maximum likelihood (Laplace
```

```
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: PREVHYP ~ CIGPDAY + SYSBP + CURSMOKE + DIABETES + EDUC + (1 |
    RANDID)
   Data: framingham_data_clean

     AIC      BIC   logLik deviance df.resid
  1149.5   1197.3   -565.7   1131.5     1491

Scaled residuals:
     Min       1Q   Median       3Q      Max
-13.5308  -0.3015  -0.0856   0.2880   8.7529

Random effects:
 Groups Name        Variance Std.Dev.
 RANDID (Intercept) 1.899    1.378
Number of obs: 1500, groups:  RANDID, 500

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -19.922627   1.343785 -14.826  < 2e-16 ***
CIGPDAY       0.007149   0.012501   0.572  0.56739
SYSBP         0.147722   0.009812  15.055  < 2e-16 ***
CURSMOKE1    -0.807659   0.322537  -2.504  0.01228 *
DIABETES1     2.064571   0.680872   3.032  0.00243 **
EDUC2         0.113750   0.259084   0.439  0.66063
EDUC3        -0.059485   0.317380  -0.187  0.85133
EDUC4         0.043190   0.364247   0.119  0.90561
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Correlation of Fixed Effects:
          (Intr) CIGPDA SYSBP  CURSMO DIABET EDUC2  EDUC3
CIGPDAY   -0.021
SYSBP     -0.990  0.023
CURSMOKE1  0.052 -0.753 -0.091
DIABETES1 -0.101  0.039  0.091 -0.059
EDUC2     -0.140 -0.057  0.059 -0.019  0.013
EDUC3     -0.131  0.020  0.060 -0.002  0.008  0.370
EDUC4     -0.094 -0.095  0.035  0.052 -0.010  0.333  0.266
convergence code: 0
Model failed to converge with max|grad| = 3.80408 (tol = 0.001, component 1)
Model is nearly unidentifiable: very large eigenvalue
 - Rescale variables?
Model is nearly unidentifiable: large eigenvalue ratio
 - Rescale variables?
```

In [272]:
```
qqmath(model_3)
```

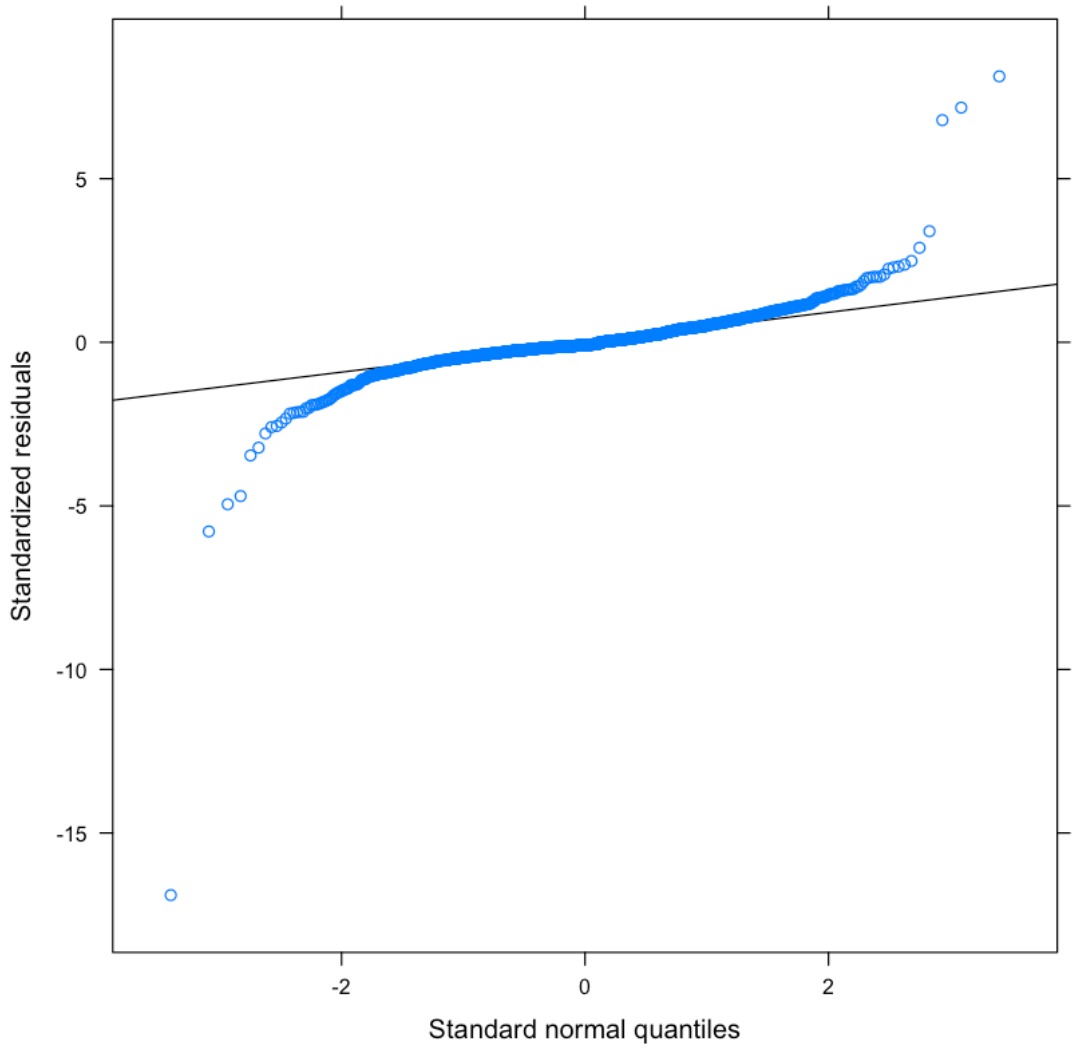


In [273]:
```
# just differing trajectories variables, rand intercept
model_4 = glmer(PREVHYP ~ CIGPDAY + SYSBP + (1 | RANDID),
                data=framingham_data_clean,
                family=binomial)
summary(model_4)
```

```
Warning message in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
Model is nearly unidentifiable: very large eigenvalue
 - Rescale variables?;Model is nearly unidentifiable: large eigenvalue ratio
 - Rescale variables?

Generalized linear mixed model fit by maximum likelihood (Laplace
```

```
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: PREVHYP ~ CIGPDAY + SYSBP + (1 | RANDID)
   Data: framingham_data_clean

     AIC      BIC   logLik deviance df.resid
  1157.5   1178.8   -574.8   1149.5     1496

Scaled residuals:
     Min       1Q   Median       3Q      Max
-16.8991  -0.3056  -0.0868   0.3112   8.1287

Random effects:
 Groups Name        Variance Std.Dev.
 RANDID (Intercept) 2.103    1.45
Number of obs: 1500, groups:  RANDID, 500

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -20.311035   1.347110 -15.077   <2e-16 ***
CIGPDAY      -0.016804   0.008393  -2.002   0.0453 *
SYSBP         0.150212   0.009932  15.125   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Correlation of Fixed Effects:
        (Intr) CIGPDA
CIGPDAY  0.016
SYSBP   -0.995 -0.066
convergence code: 0
Model is nearly unidentifiable: very large eigenvalue
 - Rescale variables?
Model is nearly unidentifiable: large eigenvalue ratio
 - Rescale variables?
```

In [274]: `qqmath(model_4)`




```
In [275]: anova(model_1, model_2, model_3, model_4)
```

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)
model_4	4	1157.5267	1178.780	-574.7633	1149.5267	NA	NA	NA
model_3	9	1149.4816	1197.301	-565.7408	1131.4816	18.045024	5	2.890503e-03
model_2	17	950.2637	1040.588	-458.1318	916.2637	215.217954	8	3.940688e-42
model_1	19	952.4222	1053.373	-457.2111	914.4222	1.841491	2	3.982220e-01

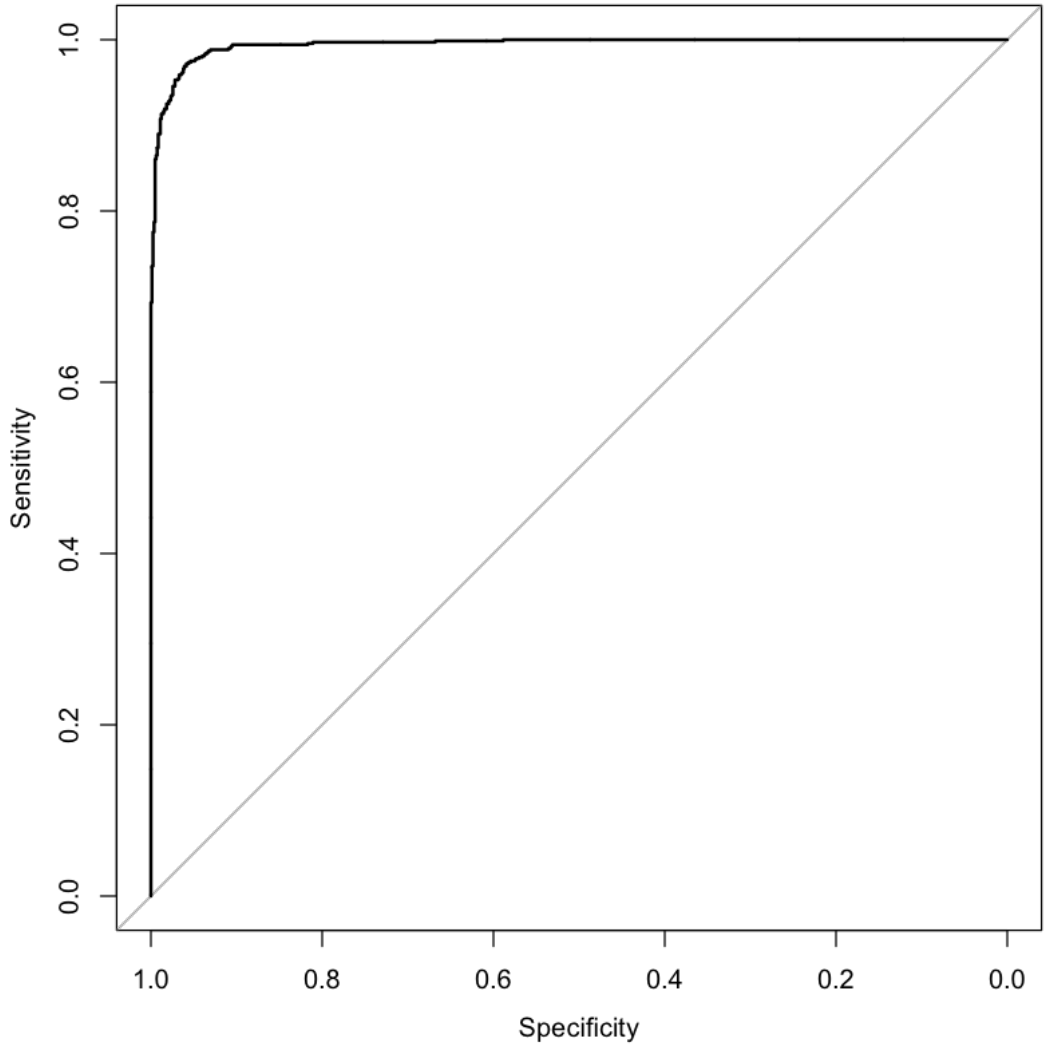
```
In [251]: # AIC and BIC are similar for model_1 and model_2, however, 2 is simpler, has slight
          final_model = model_2
```

0.4 Model Evaluation

```
In [252]: pred = predict(final_model, cbind(framingham_data_clean[1:13], framingham_data_clean
          roc.curve = roc(framingham_data_clean$PREVHYP, pred)
```

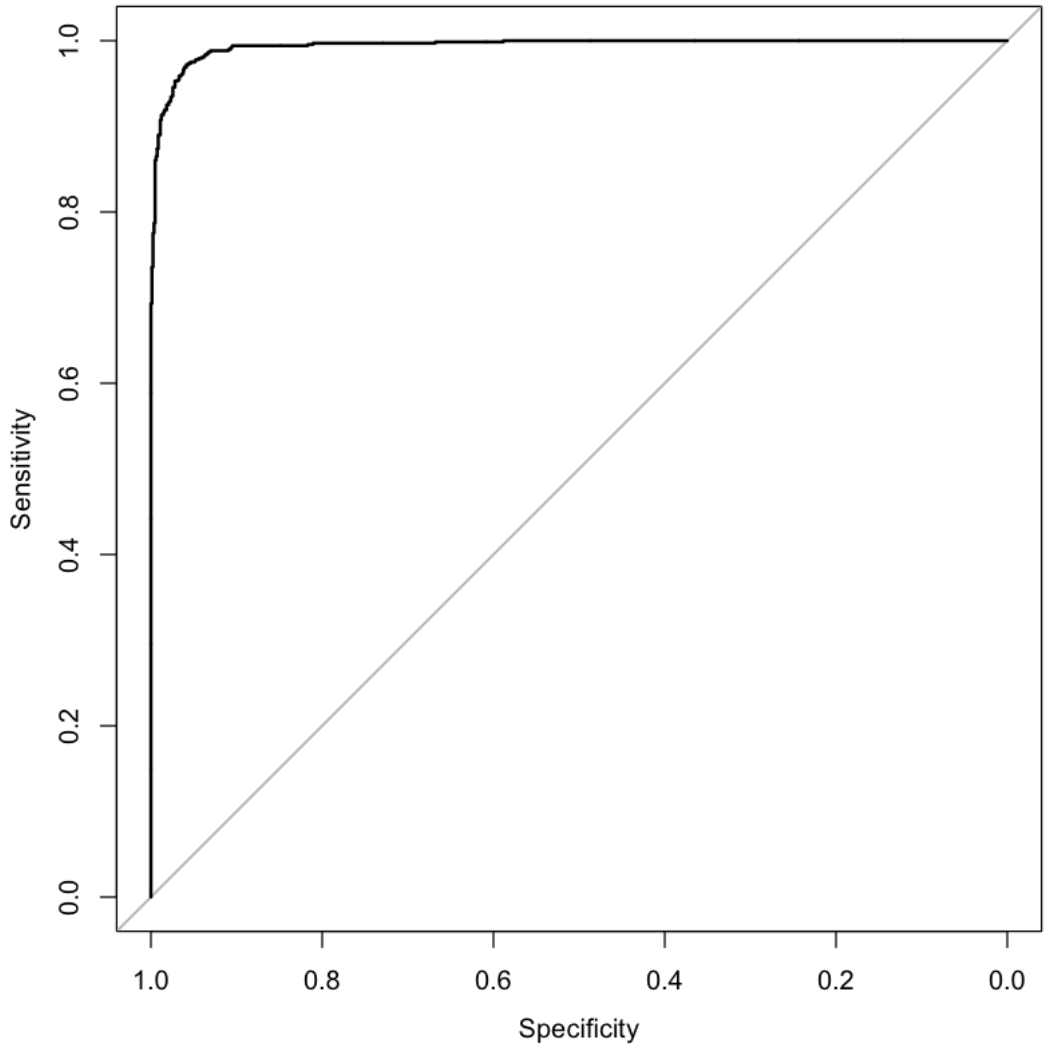
```
        print(pROC::auc(roc.curve))
        plot(roc.curve)
```

```
Area under the curve: 0.9942
```



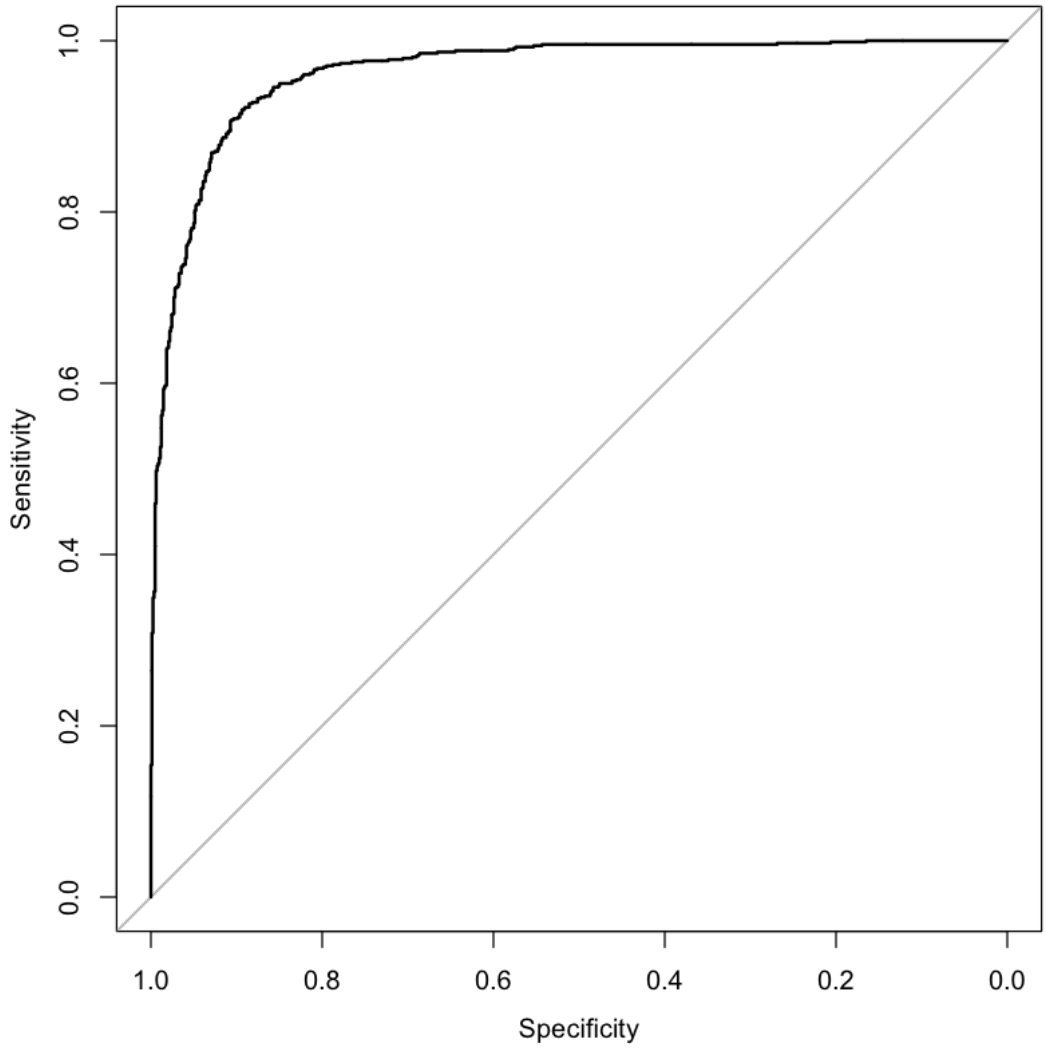

```
In [253]: pred = predict(model_2, cbind(framingham_data_clean[1:13], framingham_data_clean[15:
          roc.curve = roc(framingham_data_clean$PREVHYP, pred)
          print(pROC::auc(roc.curve))
          plot(roc.curve)
```

```
Area under the curve: 0.9942
```



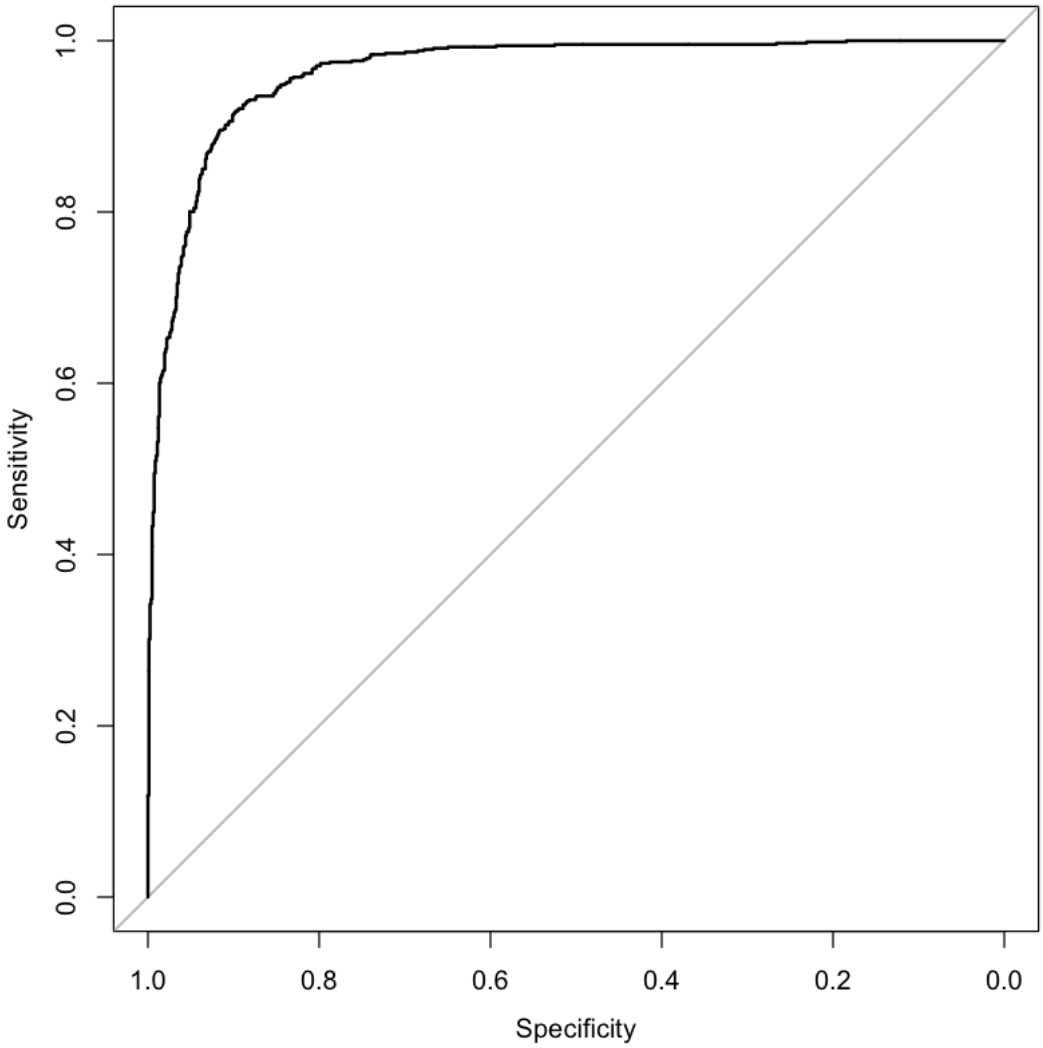

```
In [254]: pred = predict(model_3, cbind(framingham_data_clean[1:13], framingham_data_clean[15:
          roc.curve = roc(framingham_data_clean$PREVHYP, pred)
          print(pROC::auc(roc.curve))
          plot(roc.curve)
```

```
Area under the curve: 0.9631
```




```
In [255]: pred = predict(model_4, cbind(framingham_data_clean[1:13], framingham_data_clean[15:
          roc.curve = roc(framingham_data_clean$PREVHYP, pred)
          print(pROC::auc(roc.curve))
          plot(roc.curve)
```

```
Area under the curve: 0.964
```



In [ ]: *# compare to analysis 5 code in write up/include roc curves*