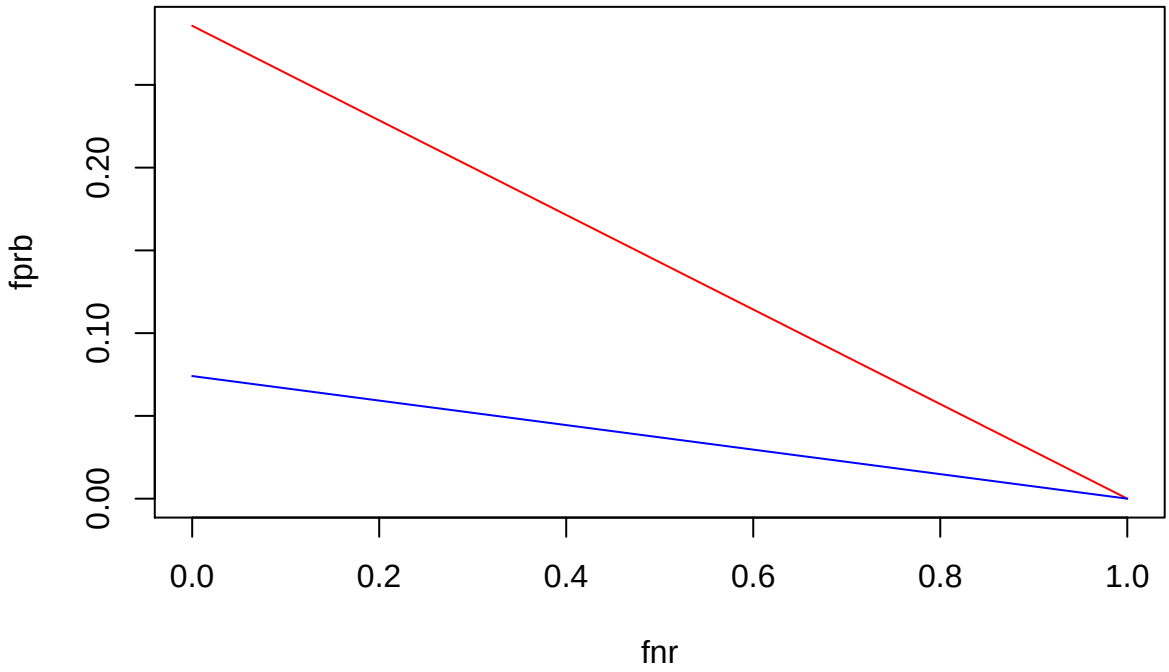
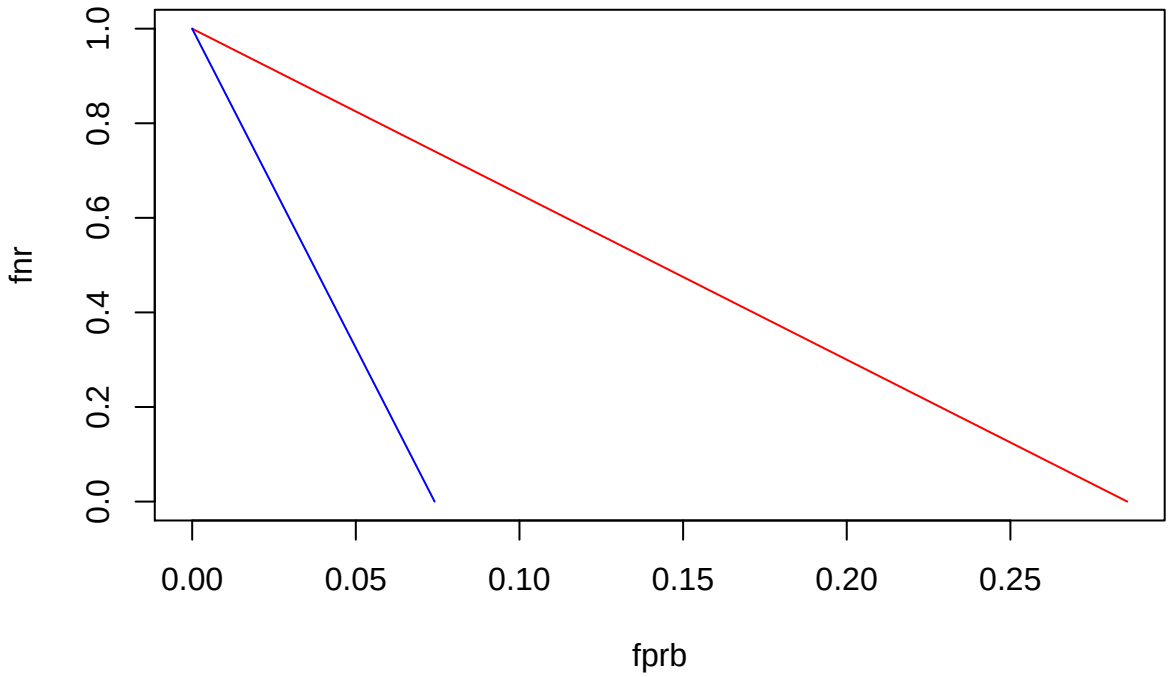
Compa Analysis Exploration

FPR and FNR bounds

```
pw = 0.1
pb = 0.3
PPV = 0.6

fnr = seq(0,1,by=0.01)
fprw = pw/(1-pw) * (1-PPV)/PPV * (1-fnr)
fprb = pb/(1-pb) * (1-PPV)/PPV * (1-fnr)

plot(fnr, fprb, type="l", col="red")
lines(fnr, fprw, col="blue")
```




```
#the larger the prevalence, the bigger the FPR for a given FNR

plot(fprb, fnr, type="l", col="red")
lines(fprw, fnr, col="blue")
```




```
#install.packages("gridExtra")
#install.packages("ggfortify")
#install.packages("dplyr")
#install.packages("ggplot2")
#install.packages("xtable")
#install.packages("texreg")
library(texreg)
```

```
## Warning: package 'texreg' was built under R version 3.3.2
```

```
## Version:  1.36.23
## Date:     2017-03-03
## Author:   Philip Leifeld (University of Glasgow)
##
## Please cite the JSS article in your publications -- see citation("texreg").
```

```
library(ggfortify)
```

```
## Warning: package 'ggfortify' was built under R version 3.3.2
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 3.3.2
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.3.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:gridExtra':
##
```

```
##     combine
```

```
## 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(xtable)
#only keep people who have recidivated in the past two years or have at least two years outside a corre
raw_data <- read.csv(file="~/Desktop/Senior Year/Comp Stats/thesis/compas-scores-two-years.csv", header

nrow(raw_data)
```

```
## [1] 7214
```

Subset data

Remove rows that meet the following:

-If the charge date of a defendants Compas scored crime was not within 30 days from when the person was arrested, we assume that because of data quality reasons, that we do not have the right offense. -We coded the recidivist flag – is_recid – to be -1 if we could not find a compas case at all. -In a similar vein, ordinary traffic offenses – those with a c_charge_degree of 'O' – will not result in Jail time are removed (only two of them). -We filtered the underlying data from Broward county to include only those rows representing people who had either recidivated in two years, or had at least two years outside of a correctional facility. -Since there are not very many observations for other races, keep only cases for Black and White defendants.

```
df <- dplyr::select(raw_data, age, c_charge_degree, race, age_cat, score_text, sex, priors_count,
                    days_b_screening_arrest, decile_score, is_recid, two_year_recid, c_jail_in, c_jail_
        filter(days_b_screening_arrest <= 30) %>%
        filter(days_b_screening_arrest >= -30) %>%
        filter(is_recid != -1) %>%
        filter(c_charge_degree != "O") %>%
        filter(score_text != 'N/A') %>%
        filter(race == "Caucasian" | race == "African-American")
```

```
## Warning: package 'bindrcpp' was built under R version 3.3.2
```

```
nrow(df)
```

```
## [1] 5278
```

Add variable for time spent in jail in units of weeks

```
jail_in <- as.POSIXct(df$c_jail_in,
                                      format='%Y-%m-%d %H:%M:%S')
```

```
## Warning in strptime(x, format, tz = tz): unknown timezone 'zone/tz/2017c.
## 1.0/zoneinfo/America/Los_Angeles'
```

```
jail_out <- as.POSIXct(df$c_jail_out,
                                      format='%Y-%m-%d %H:%M:%S')
df <- mutate(df, jail_sentence = difftime(jail_out,jail_in,units="weeks"))
```

Look for cases where defendants recidivated after the two-year threshold

```
table(df$two_year_recid, df$is_recid)
```

```
##
##        0    1
##   0 2631  164
##   1    0 2483
```

Remove cases where defendants recidivated sometime after two years

```
df <- df %>% filter((two_year_recid != 1 & is_recid != 1) | (two_year_recid != 0 & is_recid != 0) )

nrow(df)
```

```
## [1] 5114
```

Add factor variables that will later be used in logistic model

```
df_bw <- mutate(df, crime_factor = factor(c_charge_degree)) %>%
      mutate(age_factor = as.factor(age_cat)) %>%
      within(age_factor <- relevel(age_factor, ref = 1)) %>%
      mutate(race_factor = factor(race)) %>%
      within(race_factor <- relevel(race_factor, ref = 2)) %>%
      mutate(gender_factor = factor(sex, labels= c("Female","Male"))) %>%
      within(gender_factor <- relevel(gender_factor, ref = 2)) %>%
      mutate(score_factor = factor(score_text != "Low", labels = c("LowScore","HighScore")))
```

Summary Statistics

```
df_bw$length_of_stay <- as.numeric(as.Date(df_bw$c_jail_out) - as.Date(df_bw$c_jail_in))
cor(df_bw$length_of_stay, df_bw$decile_score)
```

```
## [1] 0.2037935
```

```
summary(df_bw$age_cat)
```

```
##         25 - 45 Greater than 45    Less than 25
##            2928            1065            1121
```

```
summary(df_bw$race)
```

```
## African-American            Asian        Caucasian         Hispanic
##             3063                0             2051                0
##  Native American            Other
##                0                0
```

```
print(paste("Black defendants:",round((3063 / 5114 * 100),2), "%"))
```

```
## [1] "Black defendants: 59.89 %"
```

```
print(paste("White defendants:",round((2051 / 5114 * 100),2), "%"))
```

```
## [1] "White defendants: 40.11 %"
```

```
summary(df_bw$score_text)
```

```
##   High    Low Medium
##   1042   2665   1407
```

```
xtabs(~ sex + race, data=df)
```

```
##         race
## sex      African-American Asian Caucasian Hispanic Native American Other
##   Female              536     0       475        0               0     0
##   Male               2527     0      1576        0               0     0
```

```
summary(df_bw$sex)
```

```
## Female   Male
##   1011   4103
```

```
print(paste("Men:",round((4997 / 6172 * 100),2), "%"))
```

```
## [1] "Men: 80.96 %"
```

```
print(paste("Women:",round((1175 / 6172 * 100),2), "%"))
```

```
## [1] "Women: 19.04 %"
```

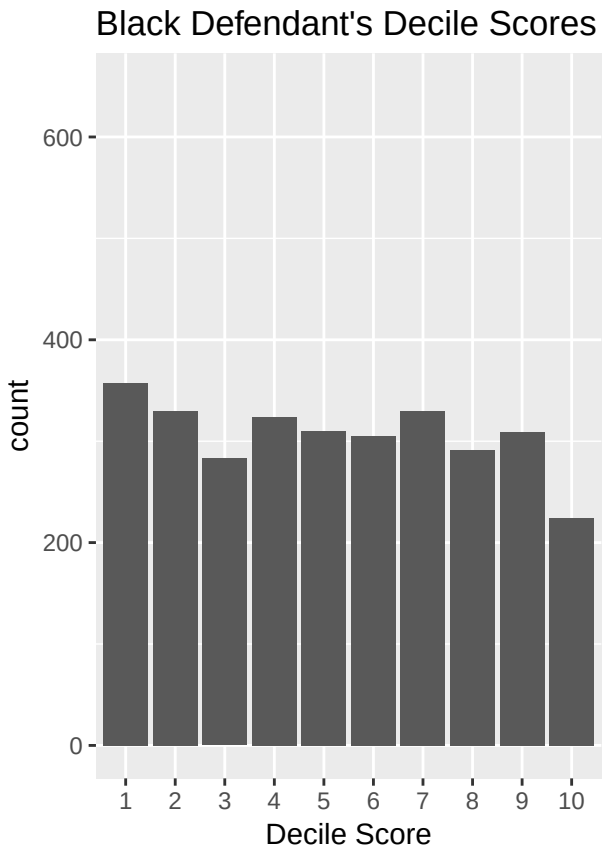
```
nrow(filter(df_bw, two_year_recid == 1))
```

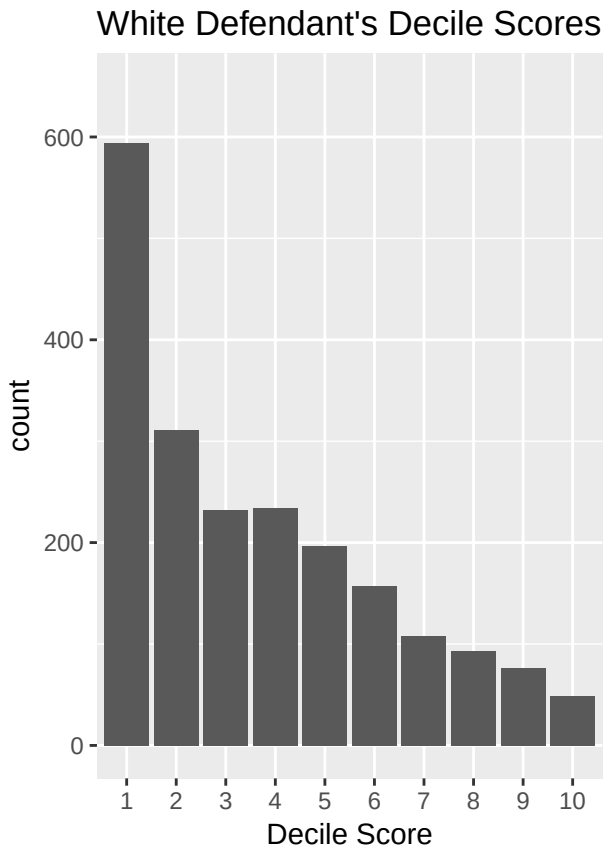
```
## [1] 2483
```

```
nrow(filter(df_bw, two_year_recid == 1)) / nrow(df) * 100
```

```
## [1] 48.55299
```

```
library(grid)
library(gridExtra)
pblack <- ggplot(data=filter(df, race =="African-American"), aes(ordered(decile_score))) +
          geom_bar() + xlab("Decile Score") +
          ylim(0, 650) + ggtitle("Black Defendant's Decile Scores")
pwhite <- ggplot(data=filter(df, race =="Caucasian"), aes(ordered(decile_score))) +
          geom_bar() + xlab("Decile Score") +
          ylim(0, 650) + ggtitle("White Defendant's Decile Scores")
grid.arrange(pblack, pwhite,  ncol = 2)
```






```
xtabs(~ decile_score + race, data=df)
```

```
##             race
## decile_score African-American Asian Caucasian Hispanic Native American
##           1               357     0       594        0               0
##           2               330     0       311        0               0
##           3               283     0       232        0               0
##           4               324     0       234        0               0
##           5               310     0       197        0               0
##           6               305     0       157        0               0
##           7               330     0       108        0               0
##           8               291     0        93        0               0
##           9               309     0        76        0               0
##           10              224     0        49        0               0
##             race
## decile_score Other
##           1      0
##           2      0
##           3      0
##           4      0
##           5      0
##           6      0
##           7      0
##           8      0
##           9      0
##           10     0
```

```
summary(df)
```

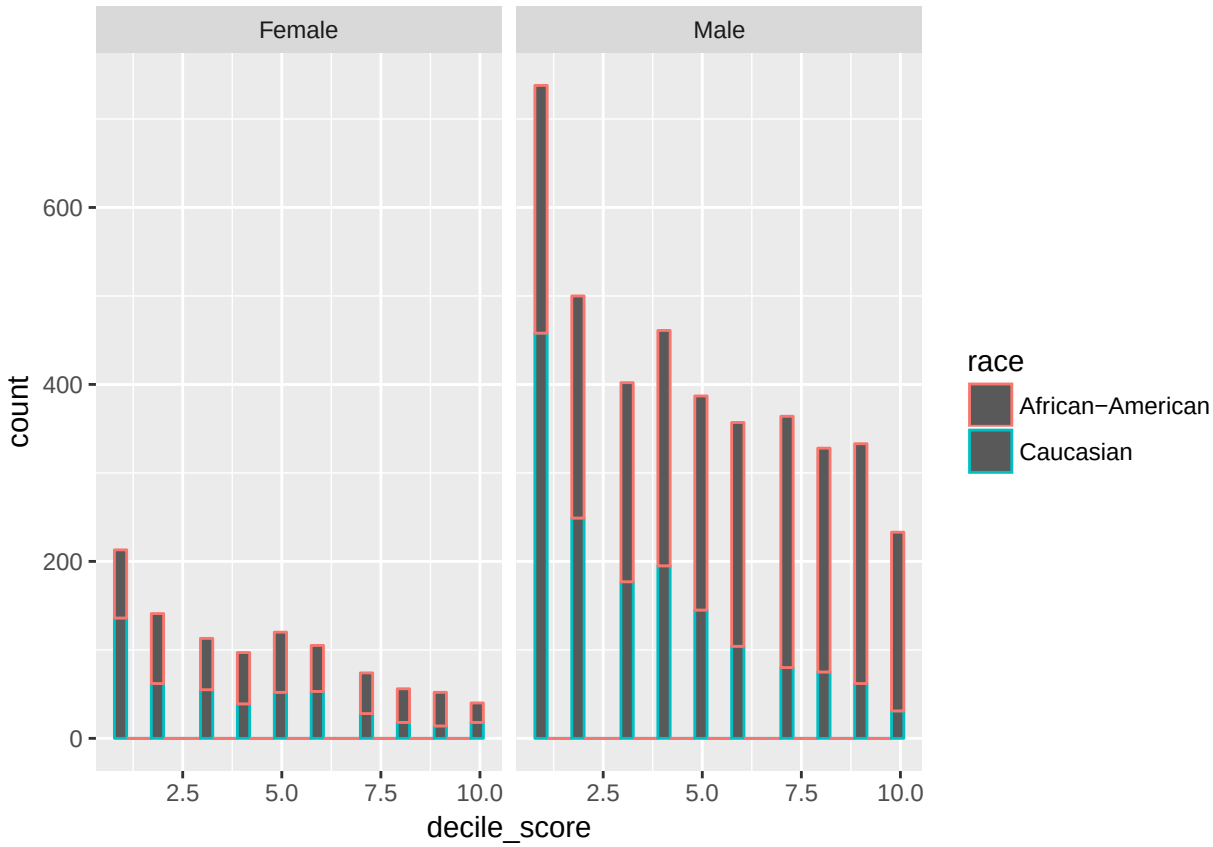
```
##       age        c_charge_degree               race
##  Min.   :18.00   F:3340          African-American:3063
##  1st Qu.:25.00   M:1774          Asian           :   0
##  Median :31.00                   Caucasian       :2051
##  Mean   :34.48                   Hispanic        :   0
##  3rd Qu.:42.00                   Native American :   0
##  Max.   :80.00                   Other           :   0
##
##             age_cat      score_text       sex        priors_count
##  25 - 45        :2928   High  :1042   Female:1011   Min.   : 0.000
##  Greater than 45:1065   Low   :2665   Male  :4103   1st Qu.: 0.000
##  Less than 25   :1121   Medium:1407                 Median : 1.000
##                                                     Mean   : 3.452
##                                                     3rd Qu.: 5.000
##                                                     Max.   :38.000
##
##  days_b_screening_arrest  decile_score       is_recid
##  Min.   :-30.000         Min.   : 1.000   Min.   :0.0000
##  1st Qu.: -1.000         1st Qu.: 2.000   1st Qu.:0.0000
##  Median : -1.000         Median : 4.000   Median :0.0000
##  Mean   : -1.725         Mean   : 4.625   Mean   :0.4855
##  3rd Qu.: -1.000         3rd Qu.: 7.000   3rd Qu.:1.0000
##  Max.   : 30.000         Max.   :10.000   Max.   :1.0000
##
##  two_year_recid                 c_jail_in                  c_jail_out
##  Min.   :0.0000   2013-01-01 01:31:55:   1   2013-09-14 05:58:00:   3
##  1st Qu.:0.0000   2013-01-01 03:16:15:   1   2013-02-06 10:01:51:   2
##  Median :0.0000   2013-01-01 03:28:03:   1   2013-08-13 10:05:00:   2
##  Mean   :0.4855   2013-01-01 04:17:22:   1   2013-09-14 05:54:00:   2
##  3rd Qu.:1.0000   2013-01-01 04:29:04:   1   2013-11-09 02:08:17:   2
##  Max.   :1.0000   2013-01-01 05:21:55:   1   2014-02-06 09:10:58:   2
##                   (Other)            :5108   (Other)            :5101
##  jail_sentence
##  Length:5114
##  Class :difftime
##  Mode  :numeric
##
##
##
##
```

Risk Score Distributions

Risk Score Distribution by Sex (colored by race)

```
ggplot(df, aes(decile_score)) + geom_histogram(aes(color=race)) + facet_wrap('sex')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

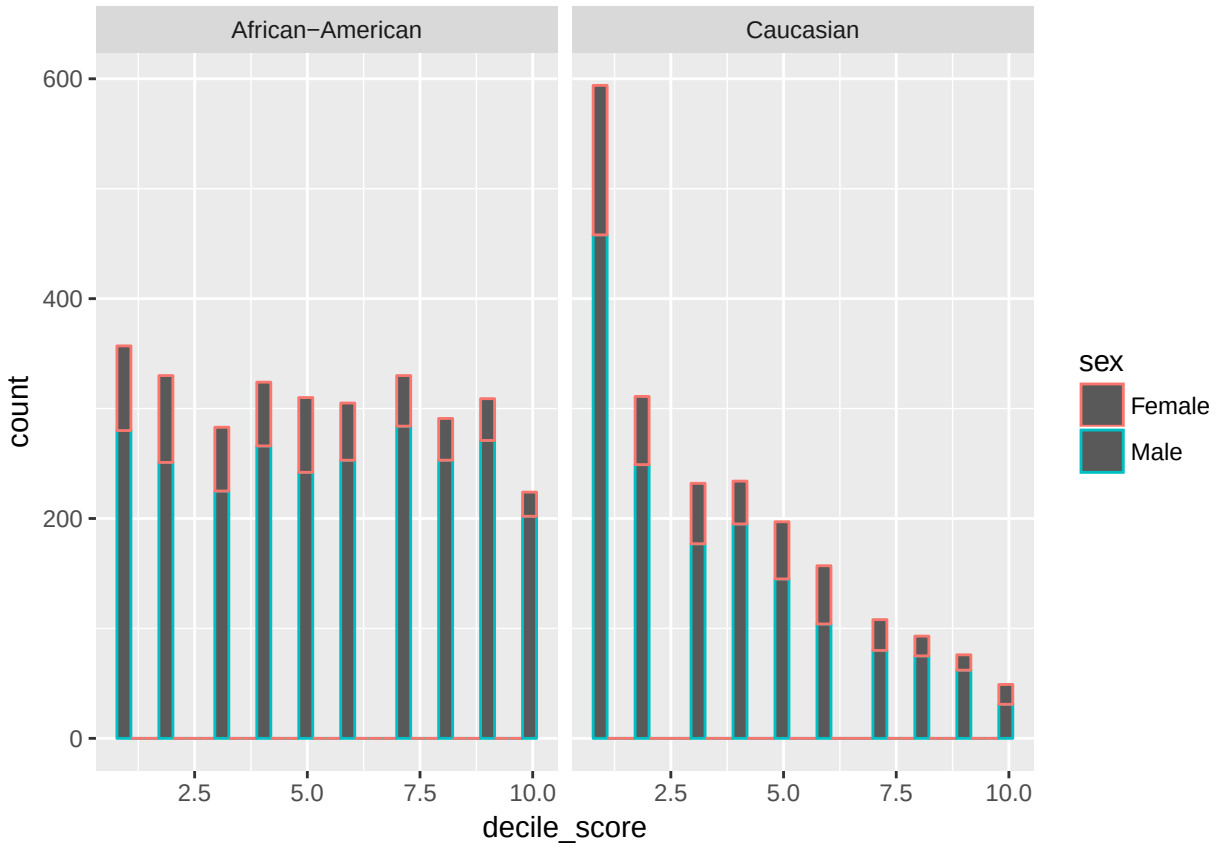


-Very few white people were given a score of 1 for both sexes -Defendant count tapers off as decile score increases for Black females, but not for Black males

Risk Score Distribution by Race (colored by sex)

```
ggplot(df, aes(decile_score)) + geom_histogram(aes(color=sex)) + facet_wrap('race')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

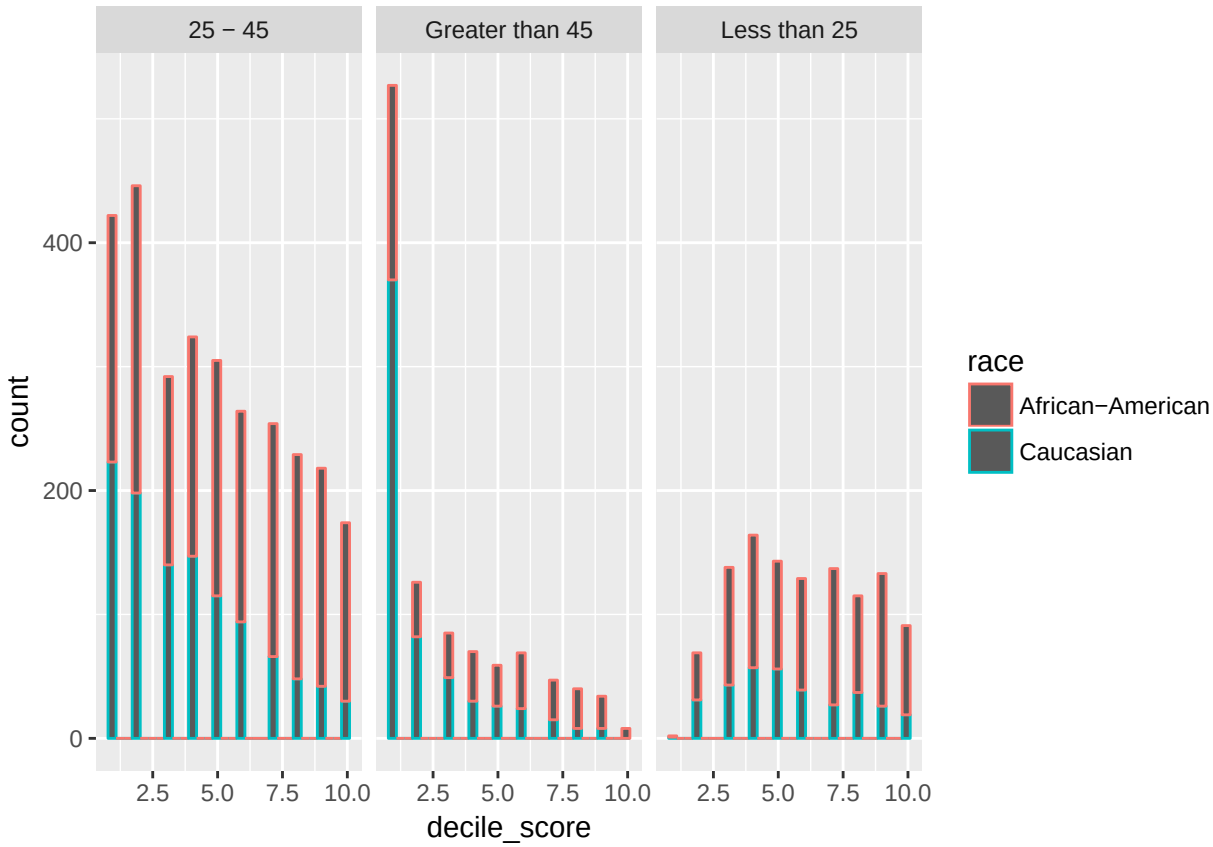


-Most observations are from male defendants

Risk Score Distribution by Age Category (colored by race)

```
ggplot(df, aes(decile_score)) + geom_histogram(aes(color=race)) + facet_wrap('age_cat')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

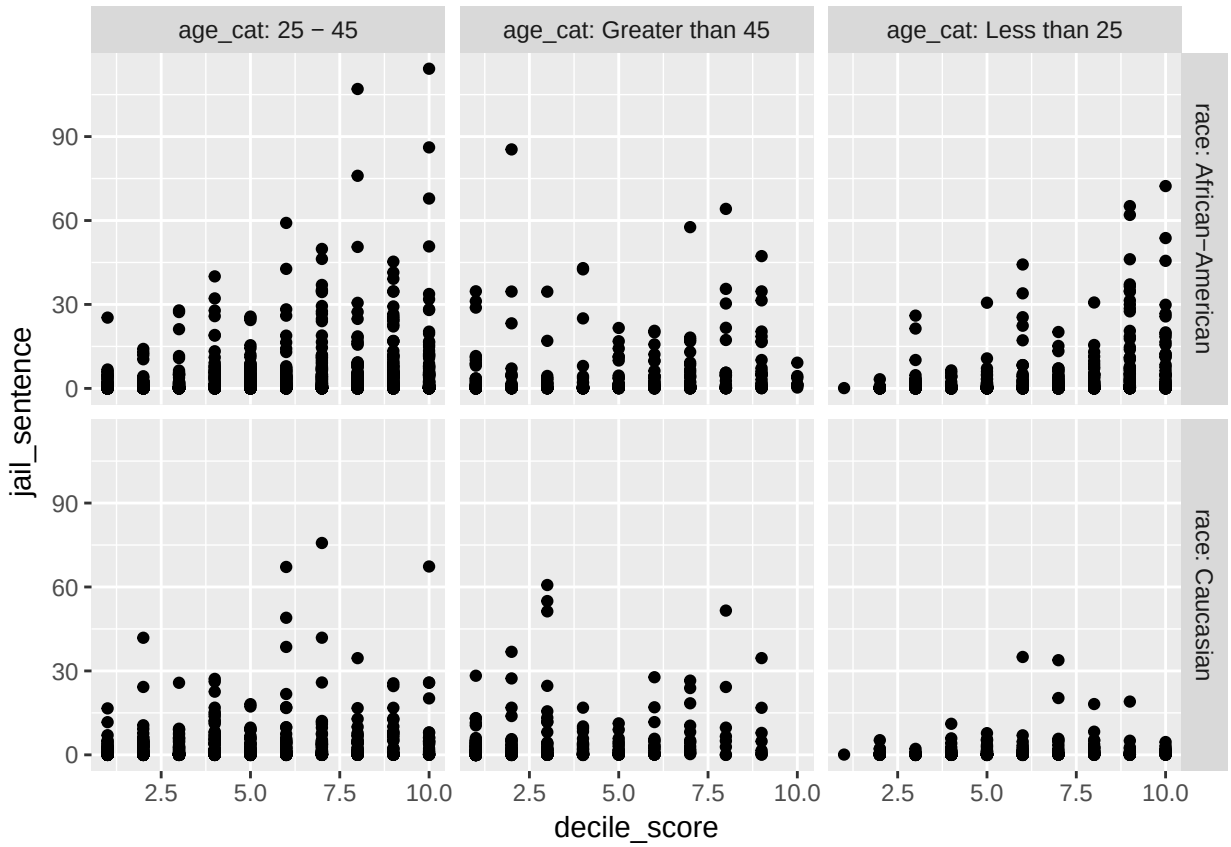


-There is a drastic spike in the number of older age defendants who were given a score of 1 -For older and middle-age classified defendants, the higher the decile score, the less people were given that score For younger classified defendants, little to no people were given a score of 1 and there does no downward trend in count number as decile score increases. Decile scores seem to be pretty evenly distributed

Jail Sentence by Age, Decile Score, and Race

```
p <- ggplot(df, aes(decile_score, jail_sentence)) + geom_point()
p + facet_grid(race ~ age_cat, labeller = label_both)
```

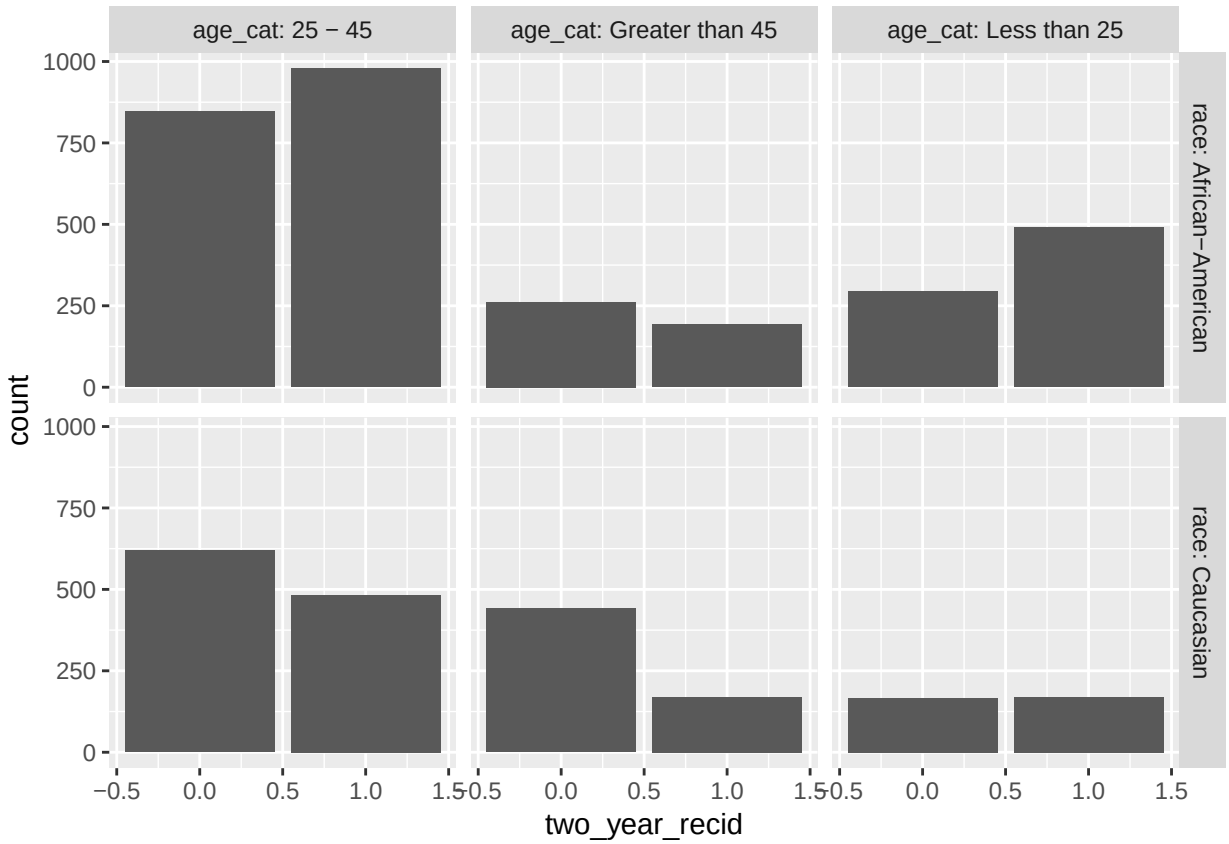
```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.
```



-More extreme sentences were given disproportionately to younger and middle-aged Black defendants with higher risk scores -One extreme case of a Black defendants with a very low risk score given a very high jail sentence

Actual Recidivism Distributions

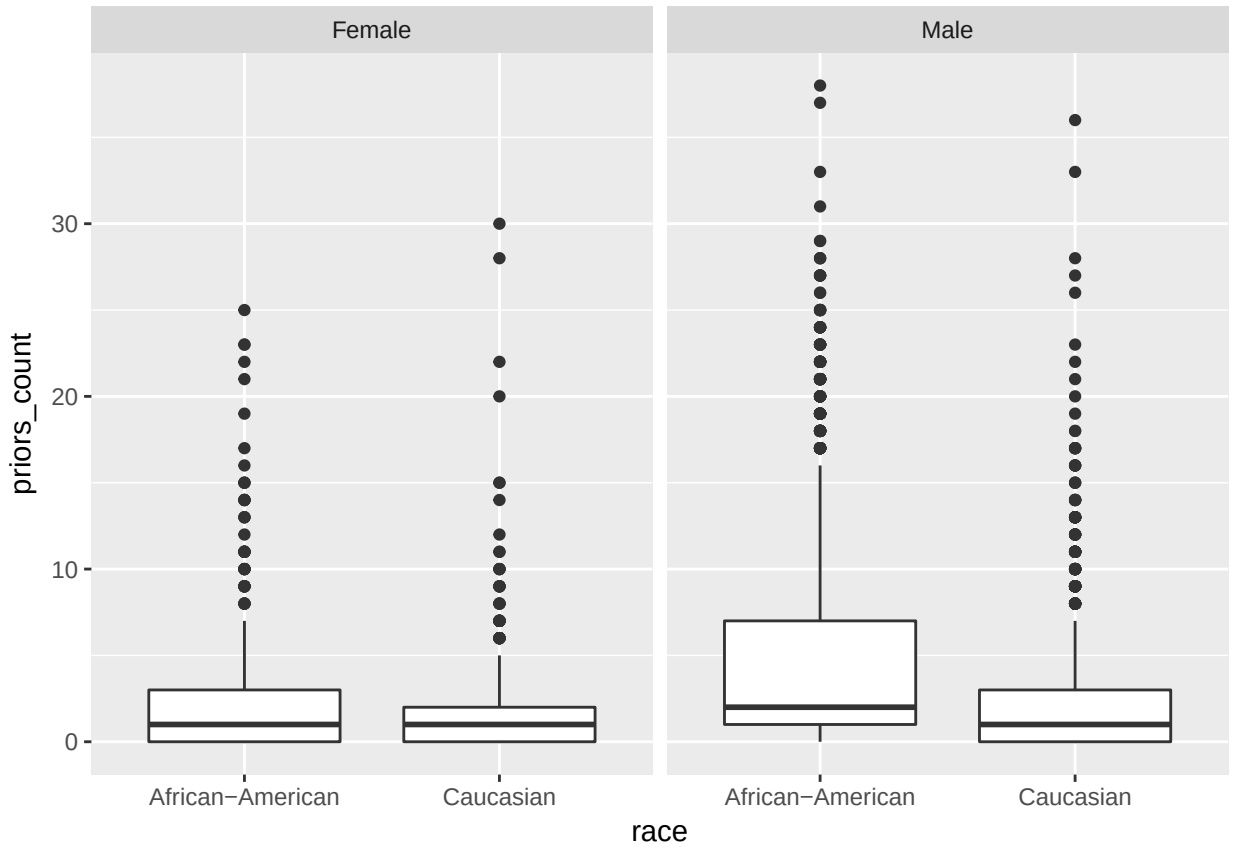
```
a <- ggplot(df, aes(two_year_recid)) + geom_bar()
a + facet_grid(race ~ age_cat, labeller = label_both)
```



Priors Count Distributions

Priors Count by Sex and Race

```
ggplot(df, aes(race, priors_count)) + geom_boxplot() + facet_wrap('sex')
```



-The average number of priors is about the same for female defendants across race and slightly higher for Black male defendants than White male defendants -The priors count distribution is skewed right by about the same amount for Black females and White males -The priors count distribution for Black males is is skewed right about 2.5 more than the priors count distribution for White males -There are more extreme outliers for White females than Black females and more extreme outliers for Black males than White Males

Logistic regression models for actual recidivism

```
log_actual <- glm(two_year_recid ~ crime_factor + age_factor + race_factor + gender_factor + score_facto
summary(log_actual)
```

```
##
## Call:
## glm(formula = two_year_recid ~ crime_factor + age_factor + race_factor +
##     gender_factor + score_factor + priors_count, family = "binomial",
##     data = df_bw)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6917  -0.9432  -0.5661   0.9996   2.0141
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -0.766658   0.070835 -10.823  < 2e-16 ***
## crime_factorM                -0.133228   0.065797  -2.025   0.0429 *
## age_factorGreater than 45    -0.554743   0.085368  -6.498 8.12e-11 ***
```

```
## age_factorLess than 25       0.553411   0.079212   6.986 2.82e-12 ***
## race_factorAfrican-American  0.049777   0.065220   0.763   0.4453
## gender_factorFemale         -0.432637   0.078646  -5.501 3.78e-08 ***
## score_factorHighScore        0.738701   0.068739  10.746  < 2e-16 ***
## priors_count                 0.137294   0.009317  14.736  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7085.2  on 5113  degrees of freedom
## Residual deviance: 6146.3  on 5106  degrees of freedom
## AIC: 6162.3
##
## Number of Fisher Scoring iterations: 4
```

While type of crime, age category, gender, and COMPAS classification were significant factors in determining whether or not a defendant actually recidivated, being Black or White was not a significant factor.

FPR/FNR rates with changes in classification threshold with each race

Original Thresholds

Confusion Matrix for Black defendants, original risk scores

```
cm_b <- df_bw %>% filter(race == "African-American") %>%
  select(two_year_recid, score_factor) %>%
  table()
xtable(cm_b)
```

```
## % latex table generated in R 3.3.1 by xtable 1.8-2 package
## % Sat Dec  9 00:50:44 2017
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrr}
##   \hline
##  & LowScore & HighScore \\
##   \hline
## 0 & 821 & 581 \\
##   1 & 473 & 1188 \\
##    \hline
## \end{tabular}
## \end{table}
```

FPR: 581/1402 = 41.44% FNR: 473/1661 = 28.48% PPV: 1188/1769 = 67.16% p: 1661/3063 = 54.23%

Confusion Matrix for Caucasian, original risk scores

```
cm_w <- df_bw %>% filter(race == "Caucasian") %>%
  select(two_year_recid, score_factor) %>%
  table()
xtable(cm_w)
```

```
## % latex table generated in R 3.3.1 by xtable 1.8-2 package
## % Sat Dec  9 00:50:45 2017
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrr}
##   \hline
##  & LowScore & HighScore \\
##   \hline
## 0 & 963 & 266 \\
##   1 & 408 & 414 \\
##    \hline
## \end{tabular}
## \end{table}
```

FPR: 266/1229 = 21.64% FNR: 408/822 = 49.64% PPV: 414/680 = 60.88% p: 822/2051 = 40.08%

New thresholds for Black defendants

Subtract 1 from AA risk score

```
df_b1 <- mutate(df_bw, decile1 = ifelse(race=="African-American", decile_score - 1, decile_score)) %>%
          mutate(riskclass1 = ifelse(decile1 > 4, "HighScore", "LowScore"))
```

Compute new confusion matrices

```
cm_b1 <- df_b1 %>% filter(race == "African-American") %>%
  select(two_year_recid, riskclass1) %>%
  table()
xtable(cm_b1)
```

```
## % latex table generated in R 3.3.1 by xtable 1.8-2 package
## % Sat Dec  9 00:50:45 2017
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrr}
##   \hline
##  & HighScore & LowScore \\
##   \hline
## 0 & 429 & 973 \\
##   1 & 1030 & 631 \\
##    \hline
## \end{tabular}
## \end{table}
```

FPR: 429/1402 = 30.60% FNR: 631/1661 = 37.99% PPV: 1030/1459 = 70.60% p: 1661/3063 = 54.23%

Subtract 2 from AA risk score

```
df_b2 <- mutate(df_bw, decile2 = ifelse(race=="African-American", decile_score - 2, decile_score)) %>%
          mutate(riskclass2 = ifelse(decile2 > 4, "HighScore", "LowScore"))
```

Compute new confusion matrices

```
cm_b2 <- df_b2 %>% filter(race == "African-American") %>%
  select(two_year_recid, riskclass2) %>%
  table()
xtable(cm_b2)
```

```
## % latex table generated in R 3.3.1 by xtable 1.8-2 package
## % Sat Dec  9 00:50:45 2017
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrr}
##   \hline
##  & HighScore & LowScore \\
##   \hline
## 0 & 311 & 1091 \\
##   1 & 843 & 818 \\
##    \hline
## \end{tabular}
## \end{table}
```

FPR: 311/1402 = 22.18% FNR: 818/1661 = 49.25% PPV: 843/1154 = 73.05% p: 1661/3063 = 54.23%

Subtract 3 from AA risk score

```
df_b3 <- mutate(df_bw, decile3 = ifelse(race=="African-American", decile_score - 3, decile_score)) %>%
          mutate(riskclass3 = ifelse(decile3 > 4, "HighScore", "LowScore"))
```

Compute new confusion matrices

```
cm_b3 <- df_b3 %>% filter(race == "African-American") %>%
  select(two_year_recid, riskclass3) %>%
  table()
xtable(cm_b3)
```

```
## % latex table generated in R 3.3.1 by xtable 1.8-2 package
## % Sat Dec  9 00:50:45 2017
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrr}
##   \hline
##  & HighScore & LowScore \\
##   \hline
## 0 & 190 & 1212 \\
##   1 & 634 & 1027 \\
##    \hline
## \end{tabular}
## \end{table}
```

FPR: 190/1402 = 13.55% FNR: 1027/1661 = 61.83% PPV: 634/8241 = 76.85% p: 1661/3063 = 54.23%

Subtract 4 from AA risk score

```
df_b4 <- mutate(df_bw, decile4 = ifelse(race=="African-American", decile_score - 4, decile_score)) %>%
          mutate(riskclass4 = ifelse(decile4 > 4, "HighScore", "LowScore"))
```

Compute new confusion matrices

```
cm_b4 <- df_b4 %>% filter(race == "African-American") %>%
  select(two_year_recid, riskclass4) %>%
  table()
xtable(cm_b4)
```

```
## % latex table generated in R 3.3.1 by xtable 1.8-2 package
```

```
## % Sat Dec  9 00:50:45 2017
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrr}
##   \hline
##  & HighScore & LowScore \\
##   \hline
## 0 & 114 & 1288 \\
##   1 & 419 & 1242 \\
##    \hline
## \end{tabular}
## \end{table}
```

FPR: 114/1402 = 8.13% FNR: 1242/1661 = 74.77% PPV: 419/533 = 78.61% p: 1661/3063 = 54.23%

New thresholds for white defendants

White risk scores + 1

```
df_w1 <- mutate(df_bw, decile1 = ifelse(race=="Caucasian", decile_score + 1, decile_score)) %>%
          mutate(riskclass1 = ifelse(decile1 > 4, "HighScore", "LowScore"))
```

Compute new confusion matrices

```
cm_w1 <- df_w1 %>% filter(race == "Caucasian") %>%
  select(two_year_recid, riskclass1) %>%
  table()
xtable(cm_w1)
```

```
## % latex table generated in R 3.3.1 by xtable 1.8-2 package
## % Sat Dec  9 00:50:45 2017
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrr}
##   \hline
##  & HighScore & LowScore \\
##   \hline
## 0 & 402 & 827 \\
##   1 & 512 & 310 \\
##    \hline
## \end{tabular}
## \end{table}
```

FPR: 402/1229 = 32.71% FNR: 310/822 = 37.71% PPV: 512/914 = 56.02% p: 822/2051 = 40.08%

White risk scores + 2

```
df_w2 <- mutate(df_bw, decile2 = ifelse(race=="Caucasian", decile_score + 2, decile_score)) %>%
          mutate(riskclass2 = ifelse(decile2 > 4, "HighScore", "LowScore"))
```

Compute new confusion matrices

```
cm_w2 <- df_w2 %>% filter(race == "Caucasian") %>%
  select(two_year_recid, riskclass2) %>%
  table()
xtable(cm_w2)
```

```
## % latex table generated in R 3.3.1 by xtable 1.8-2 package
## % Sat Dec  9 00:50:45 2017
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrr}
##   \hline
##  & HighScore & LowScore \\
##   \hline
## 0 & 552 & 677 \\
##   1 & 594 & 228 \\
##    \hline
## \end{tabular}
## \end{table}
```

FPR: 552/1229 = 44.91% FNR: 228/822 = 27.74% PPV: 594/1146 = 51.83% p: 822/2051 = 40.08%

White risk scores + 3

```
df_w3 <- mutate(df_bw, decile3 = ifelse(race=="Caucasian", decile_score + 3, decile_score)) %>%
          mutate(riskclass3 = ifelse(decile3 > 4, "HighScore", "LowScore"))
```

Compute new confusion matrices

```
cm_w3 <- df_w3 %>% filter(race == "Caucasian") %>%
  select(two_year_recid, riskclass3) %>%
  table()
xtable(cm_w3)
```

```
## % latex table generated in R 3.3.1 by xtable 1.8-2 package
## % Sat Dec  9 00:50:45 2017
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrr}
##   \hline
##  & HighScore & LowScore \\
##   \hline
## 0 & 763 & 466 \\
##   1 & 694 & 128 \\
##    \hline
## \end{tabular}
## \end{table}
```

FPR: 763/1229 = 62.08% FNR: 128/822 = 15.57% PPV: 694/1457 = 47.63% p: 822/2051 = 40.08%

White risk scores + 4

```
df_w4 <- mutate(df_bw, decile4 = ifelse(race=="Caucasian", decile_score + 4, decile_score)) %>%
          mutate(riskclass4 = ifelse(decile4 > 4, "HighScore", "LowScore"))
```

Compute new confusion matrices

```
cm_w4 <- df_w4 %>% filter(race == "Caucasian") %>%
  select(two_year_recid, riskclass4) %>%
  table()
xtable(cm_w4)
```

```
## % latex table generated in R 3.3.1 by xtable 1.8-2 package
## % Sat Dec  9 00:50:45 2017
## \begin{table}[ht]
```

```
## \centering
## \begin{tabular}{rr}
##   \hline
##  & HighScore \\
##   \hline
## 0 & 1229 \\
##   1 & 822 \\
##    \hline
## \end{tabular}
## \end{table}
```

FPR: 763/1229 = 100% FNR: 128/822 = 0% PPV: 822/2051 = 40.08% p: 822/2051 = 40.08%

DIRECTION OF RACIAL BIAS (ROC CURVES)

Subset for Black defendants

```
df_b <- df_bw %>% filter(race_factor=="African-American")
```

Subset for White defendants

```
df_w <- df_bw %>% filter(race_factor=="Caucasian")
```

```
library(ROCR)
```

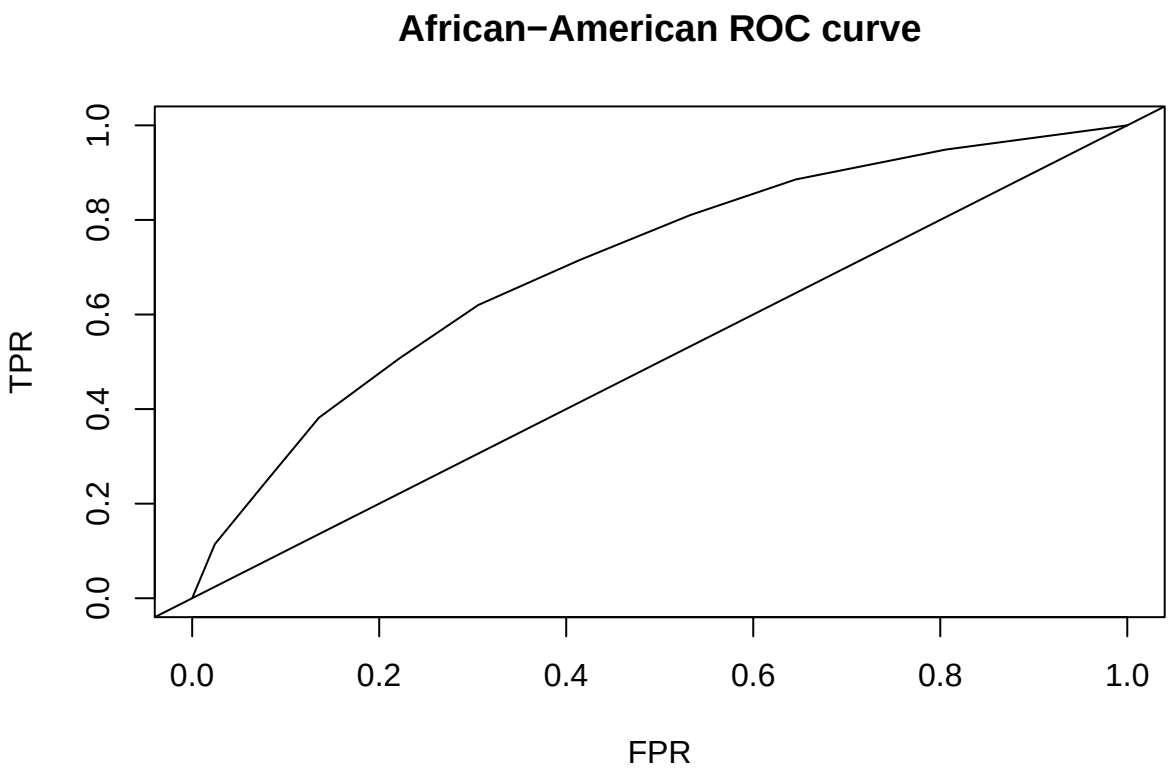
```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

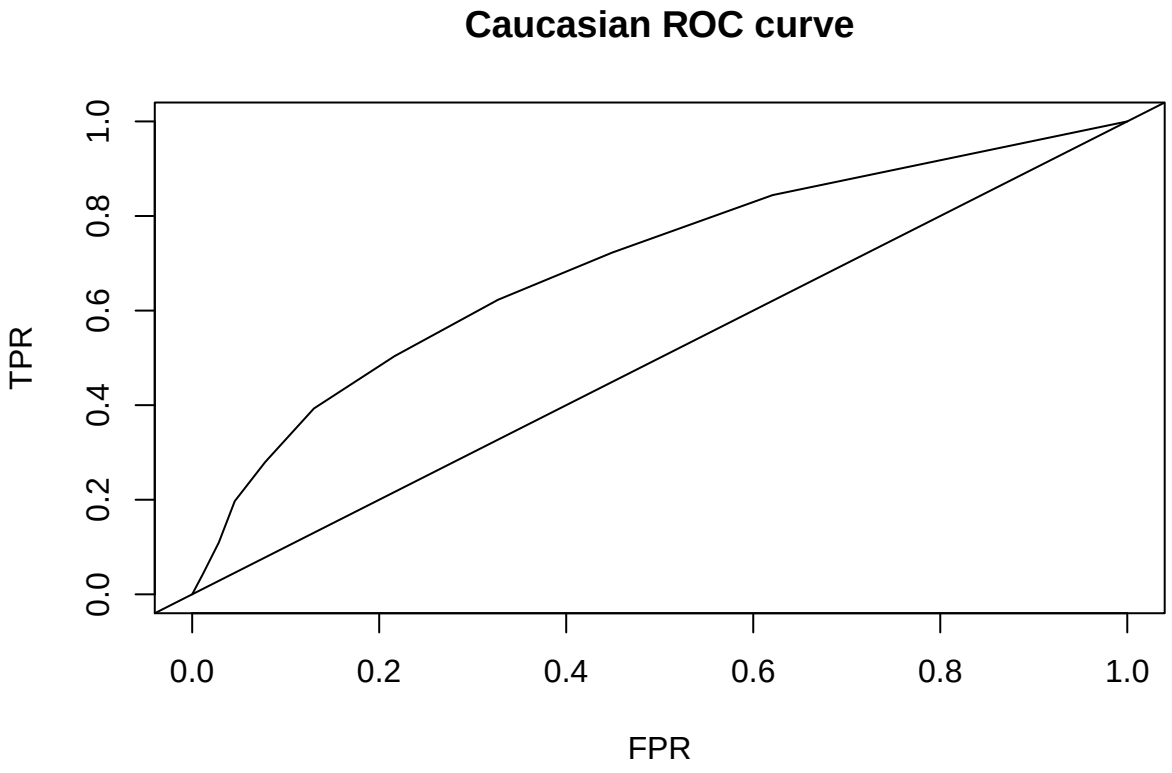
Black ROC Curve

```
recid.pred.b <- prediction(df_b$decile_score,df_b$two_year_recid)
recid.perf.b <- performance(recid.pred.b,measure="tpr",x.measure="fpr")
plot(recid.perf.b,xlab="FPR",ylab="TPR",main="African-American ROC curve")
abline(a=0,b=1)
```



White ROC Curve

```
recid.pred.w <- prediction(df_w$decile_score,df_w$two_year_recid)
recid.perf.w <- performance(recid.pred.w,measure="tpr",x.measure="fpr")
plot(recid.perf.w,xlab="FPR",ylab="TPR",main="Caucasian ROC curve")
abline(a=0,b=1)
```

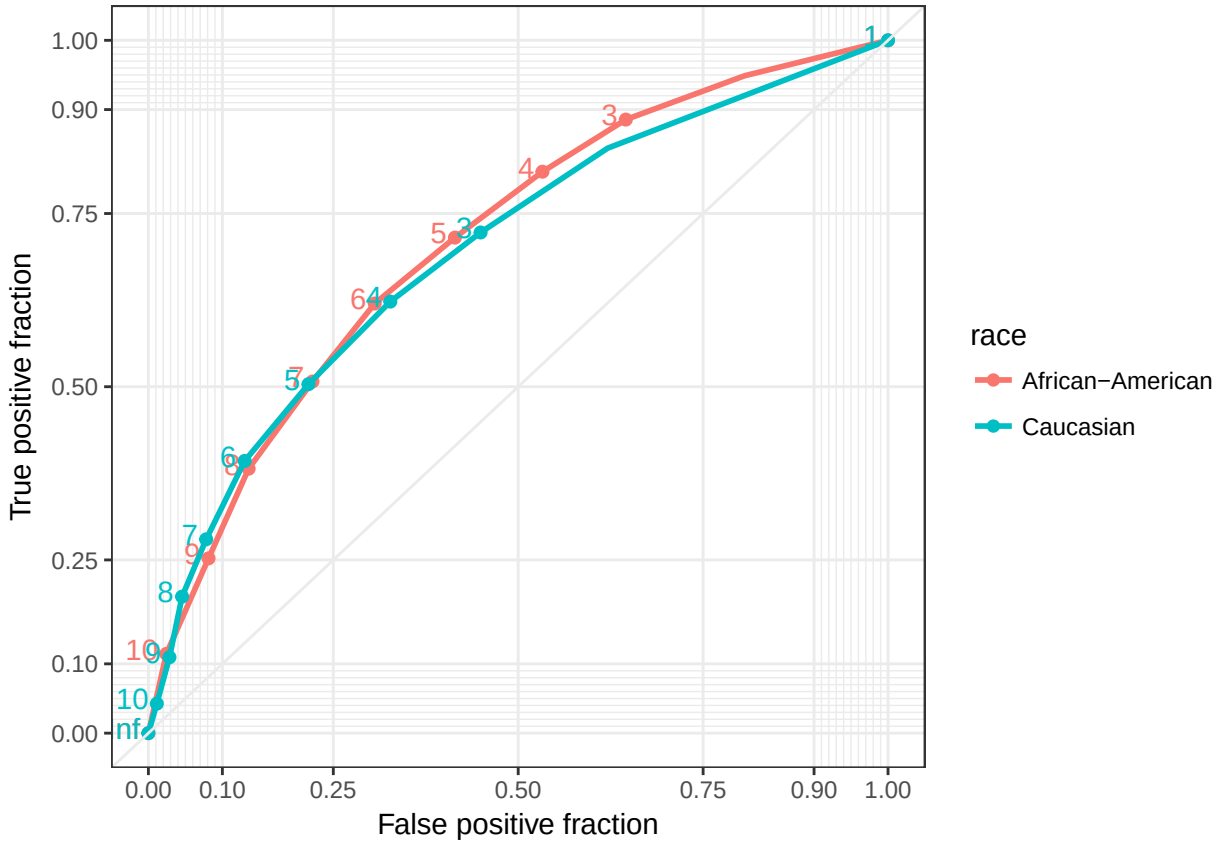


```
require(plotROC)
```

```
## Loading required package: plotROC
```

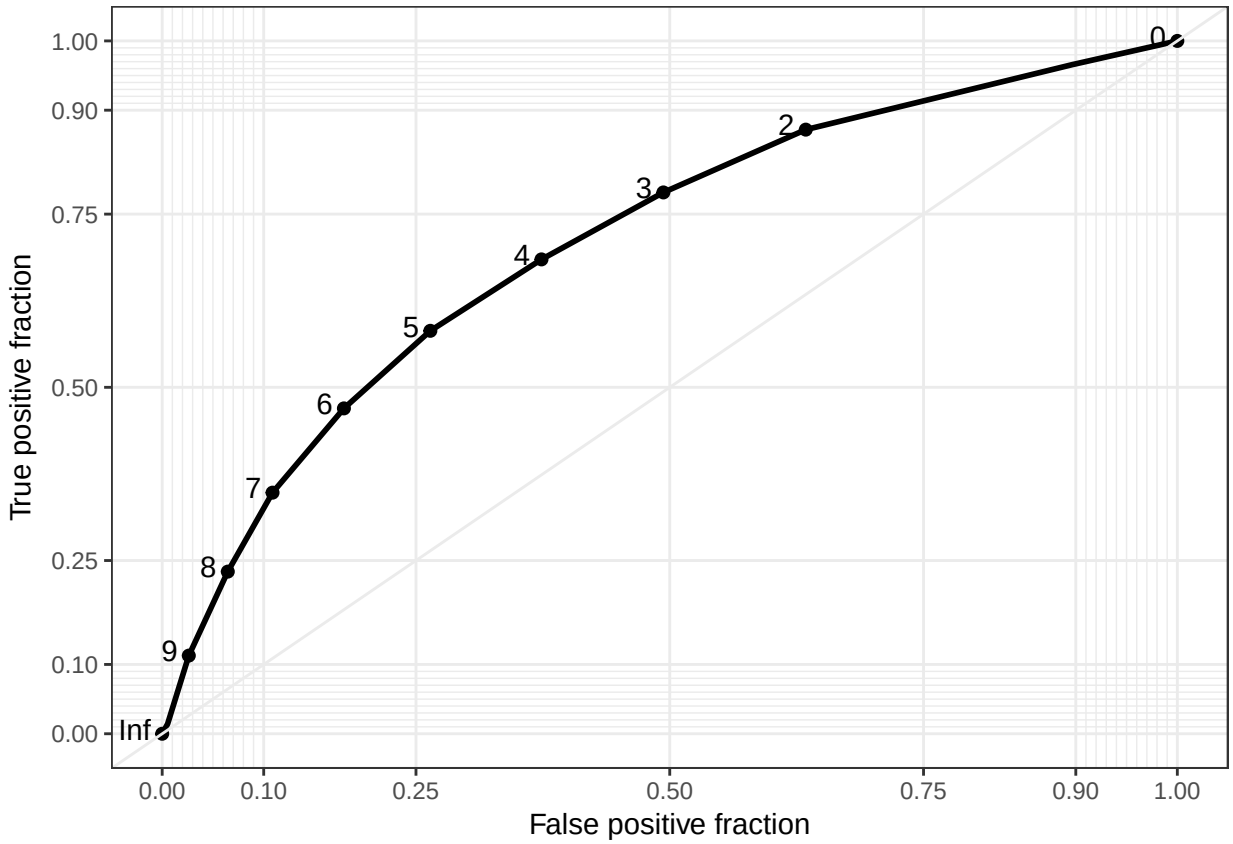
```
## Warning: package 'plotROC' was built under R version 3.3.2
```

```
ggplot(df_bw, aes(d = two_year_recid, m = decile_score, color = race)) + geom_roc() + style_roc()
```



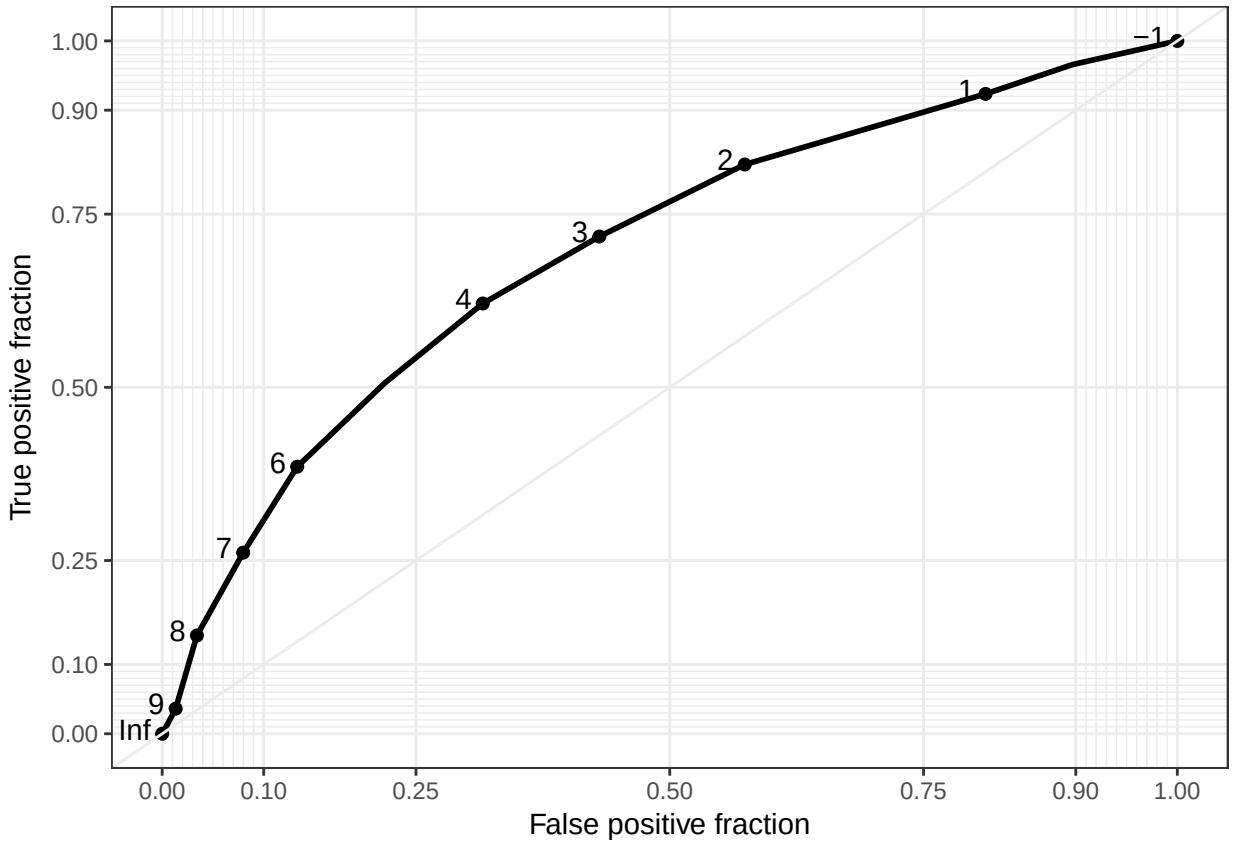
Adjustted Black ROC Curve (-1 from score)

```
require(plotROC)
ggplot(df_b1, aes(d = two_year_recid, m = decile1)) + geom_roc() + style_roc()
```



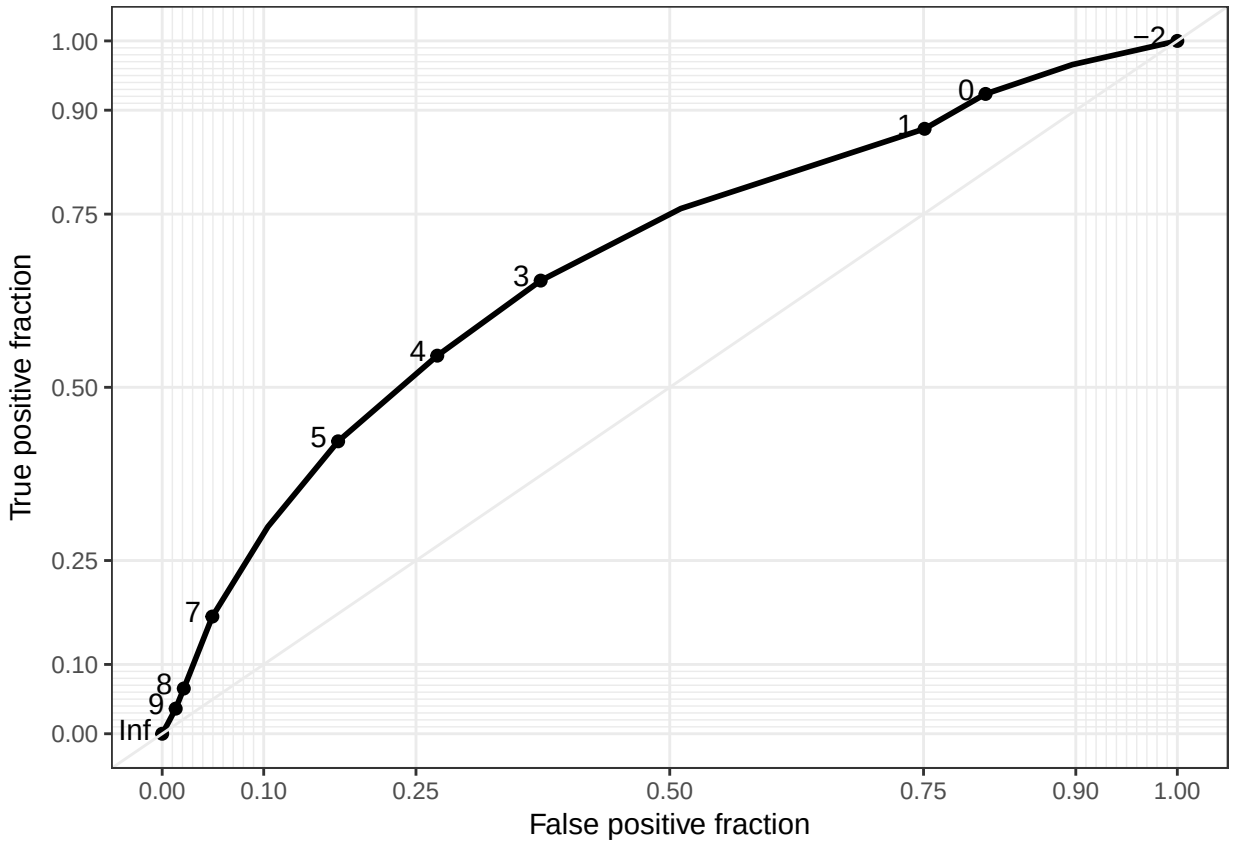
Adjustted Black ROC Curve (-2 from score)

```
ggplot(df_b2, aes(d = two_year_recid, m = decile2)) + geom_roc() + style_roc()
```



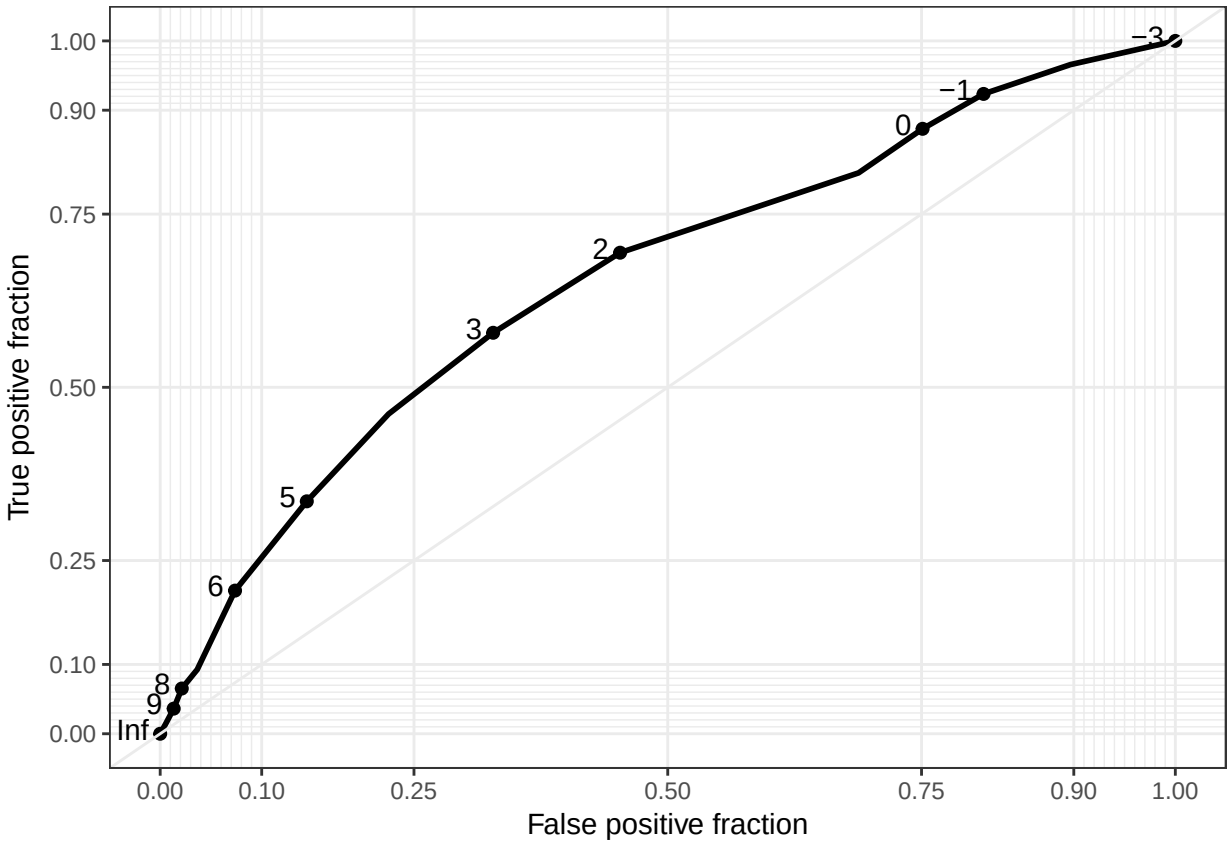
Adjustted Black ROC Curve (-3 from score)

```
ggplot(df_b3, aes(d = two_year_recid, m = decile3)) + geom_roc() + style_roc()
```



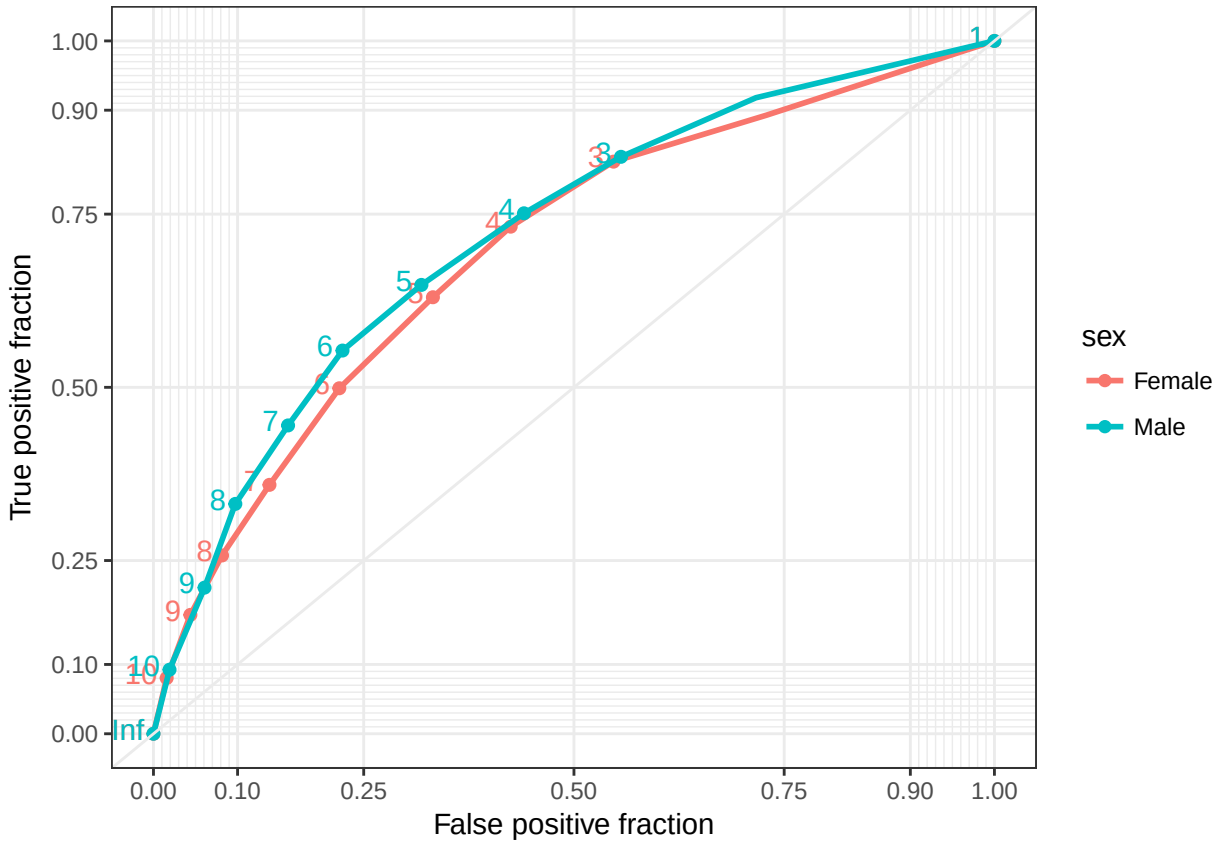
Adjustted Black ROC Curve (-4 from score)

```
ggplot(df_b4, aes(d = two_year_recid, m = decile4)) + geom_roc() + style_roc()
```



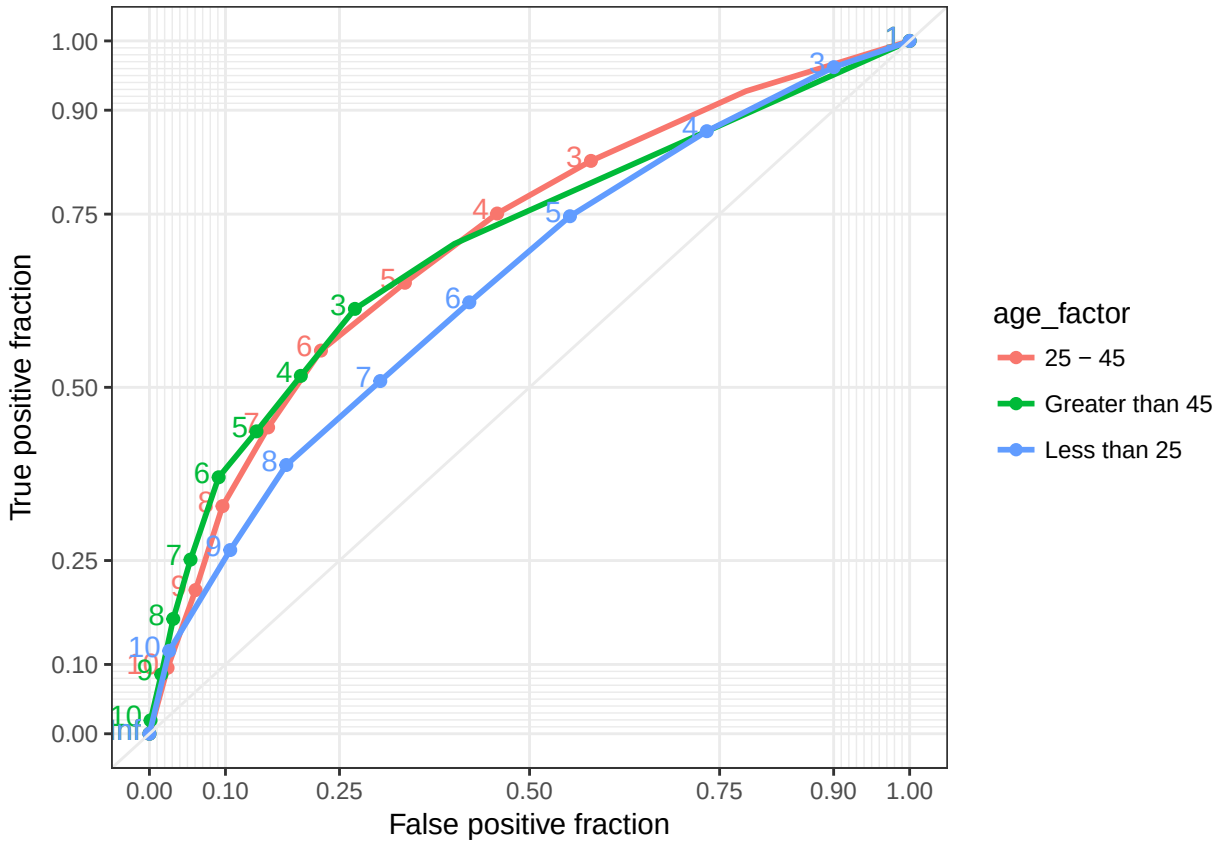
ROC Curves by Gender

```
ggplot(df_bw, aes(d = two_year_recid, m = decile_score, color = sex)) + geom_roc() + style_roc()
```



ROC Curves by Age

```
ggplot(df_bw, aes(d = two_year_recid, m = decile_score, color = age_factor)) + geom_roc() + style_roc()
```



RACIAL BIAS IN COMPAS RISK SCORES

Logistic regression model for COMPAS score

```
log_risk_score <- glm(score_factor ~ gender_factor + age_factor + race_factor +
                        priors_count + crime_factor + two_year_recid, family="binomial", data=df_bw)
summary(log_risk_score)
```

```
##
## Call:
## glm(formula = score_factor ~ gender_factor + age_factor + race_factor +
##     priors_count + crime_factor + two_year_recid, family = "binomial",
##     data = df_bw)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9946  -0.8449  -0.3167   0.8307   2.5588
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -1.53356    0.08243 -18.605  < 2e-16 ***
## gender_factorFemale              0.32074    0.08483   3.781 0.000156 ***
## age_factorGreater than 45       -1.37365    0.10585 -12.977  < 2e-16 ***
## age_factorLess than 25           1.23785    0.08295  14.922  < 2e-16 ***
## race_factorAfrican-American      0.47992    0.07043   6.814 9.46e-12 ***
## priors_count                     0.26758    0.01202  22.263  < 2e-16 ***
```

```
## crime_factorM               -0.32791    0.07218  -4.543 5.56e-06 ***
## two_year_recid               0.70957    0.06951  10.208  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7080.4  on 5113  degrees of freedom
## Residual deviance: 5225.4  on 5106  degrees of freedom
## AIC: 5241.4
##
## Number of Fisher Scoring iterations: 5
```

Black defendants are 46% more likely than white defendants to receive a higher score correcting for the seriousness of their crime, previous arrests, and future criminal behavior.

```
control <- exp(-1.53356) / (1 + exp(-1.53356))
exp(0.47992) / (1 - control + (control * exp(0.47992)))
```

```
## [1] 1.456707
```

```
texreg(log_risk_score)
```

```
##
## \begin{table}
## \begin{center}
## \begin{tabular}{l c }
## \hline
##  & Model 1 \\
## \hline
## (Intercept)                  & $-1.53^{***}$ \\
##                              & $(0.08)$      \\
## gender\_factorFemale         & $0.32^{***}$  \\
##                              & $(0.08)$      \\
## age\_factorGreater than 45   & $-1.37^{***}$ \\
##                              & $(0.11)$      \\
## age\_factorLess than 25      & $1.24^{***}$  \\
##                              & $(0.08)$      \\
## race\_factorAfrican-American & $0.48^{***}$  \\
##                              & $(0.07)$      \\
## priors\_count                & $0.27^{***}$  \\
##                              & $(0.01)$      \\
## crime\_factorM               & $-0.33^{***}$ \\
##                              & $(0.07)$      \\
## two\_year\_recid             & $0.71^{***}$  \\
##                              & $(0.07)$      \\
## \hline
## AIC                          & 5241.43       \\
## BIC                          & 5293.75       \\
## Log Likelihood               & -2612.71      \\
## Deviance                     & 5225.43       \\
## Num. obs.                    & 5114          \\
## \hline
## \multicolumn{2}{l}{\scriptsize{$^{***}p<0.001$, $^{**}p<0.01$, $^*p<0.05$}}
## \end{tabular}
## \caption{Statistical models}
```

```
## \label{table:coefficients}
## \end{center}
## \end{table}
```

risk score logistic model for Black - 1

```
#create a dummy variable for adjusted predicitions
df_b1 <- mutate(df_b1, riskclass1_binary = ifelse(riskclass1=="LowScore", 0, 1))
log_risk_score1 <- glm(riskclass1_binary ~ gender_factor + age_factor + race_factor +
summary(log_risk_score1)
```

```
##
## Call:
## glm(formula = riskclass1_binary ~ gender_factor + age_factor +
##     race_factor + priors_count + crime_factor + two_year_recid,
##     family = "binomial", data = df_b1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8643  -0.8200  -0.3909   0.8995   2.5540
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -1.50837    0.08217 -18.357  < 2e-16 ***
## gender_factorFemale           0.24742    0.08559   2.891  0.00384 **
## age_factorGreater than 45    -1.35725    0.10959 -12.384  < 2e-16 ***
## age_factorLess than 25        1.25605    0.08201  15.317  < 2e-16 ***
## race_factorAfrican-American -0.09340    0.07181  -1.301  0.19339
## priors_count                  0.26074    0.01118  23.318  < 2e-16 ***
## crime_factorM                -0.35669    0.07312  -4.878 1.07e-06 ***
## two_year_recid                0.73322    0.06960  10.534  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6952.2  on 5113  degrees of freedom
## Residual deviance: 5240.5  on 5106  degrees of freedom
## AIC: 5256.5
##
## Number of Fisher Scoring iterations: 5
```

If every decile score for Black defendants is subtracted by 1, White defendants are 7% more likely than Black defendants to receive a higher score correcting for the seriousness of their crime, previous arrests, and future criminal behavior.

```
control <- exp(-1.50837) / (1 + exp(-1.50837))
exp(-0.09340) / (1 - control + (control * exp(-0.09340)))
```

```
## [1] 0.9257861
```

subtract 2 from every AA decile score

```
df_b2 <- mutate(df_b2, riskclass2_binary = ifelse(riskclass2=="LowScore", 0, 1))
log_risk_score2 <- glm(riskclass2_binary ~ gender_factor + age_factor + race_factor +
                           priors_count + crime_factor + two_year_recid, family="binomial", data=df_b2)
summary(log_risk_score2)
```

```
##
## Call:
## glm(formula = riskclass2_binary ~ gender_factor + age_factor +
##     race_factor + priors_count + crime_factor + two_year_recid,
##     family = "binomial", data = df_b2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0582  -0.8138  -0.4654   0.8790   2.6361
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -1.378254   0.081309 -16.951  < 2e-16 ***
## gender_factorFemale             0.219790   0.086908   2.529   0.0114 *
## age_factorGreater than 45      -1.446271   0.113805 -12.708  < 2e-16 ***
## age_factorLess than 25          1.187653   0.082430  14.408  < 2e-16 ***
## race_factorAfrican-American    -0.618446   0.073835  -8.376  < 2e-16 ***
## priors_count                    0.223249   0.009859  22.644  < 2e-16 ***
## crime_factorM                  -0.391345   0.074608  -5.245 1.56e-07 ***
## two_year_recid                  0.742741   0.070891  10.477  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6675.0  on 5113  degrees of freedom
## Residual deviance: 5204.5  on 5106  degrees of freedom
## AIC: 5220.5
##
## Number of Fisher Scoring iterations: 5
```

If every decile score for Black defendants is subtracted by 2, White defendants are 40.1% more likely than white defendants to receive a higher score correcting for the seriousness of their crime, previous arrests, and future criminal behavior.

```
control <- exp(-1.378254) / (1 + exp(-1.378254))
exp(-0.618446 ) / (1 - control + (control * exp(-0.618446)))
```

```
## [1] 0.5939197
```

subtract 3 from every AA decile score

```
df_b3 <- mutate(df_b3, riskclass3_binary = ifelse(riskclass3=="LowScore", 0, 1))
log_risk_score3 <- glm(riskclass3_binary ~ gender_factor + age_factor + race_factor +
                           priors_count + crime_factor + two_year_recid, family="binomial", data=df_b3)
```

```
summary(log_risk_score3)
```

```
##
## Call:
## glm(formula = riskclass3_binary ~ gender_factor + age_factor +
##     race_factor + priors_count + crime_factor + two_year_recid,
##     family = "binomial", data = df_b3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2246  -0.7378  -0.4316   0.7635   2.8445
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -1.356742   0.083288 -16.290  < 2e-16 ***
## gender_factorFemale           0.274071   0.090671   3.023  0.00251 **
## age_factorGreater than 45    -1.426928   0.118290 -12.063  < 2e-16 ***
## age_factorLess than 25        1.128147   0.086584  13.030  < 2e-16 ***
## race_factorAfrican-American -1.239543   0.078983 -15.694  < 2e-16 ***
## priors_count                  0.199327   0.009125  21.844  < 2e-16 ***
## crime_factorM                -0.403278   0.078764  -5.120 3.05e-07 ***
## two_year_recid                0.817972   0.075406  10.848  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6195.9  on 5113  degrees of freedom
## Residual deviance: 4881.0  on 5106  degrees of freedom
## AIC: 4897
##
## Number of Fisher Scoring iterations: 5
```

If every decile score for Black defendants is subtracted by 3, White defendants are 66% more likely than Black defendants to receive a higher score correcting for the seriousness of their crime, previous arrests, and future criminal behavior.

```
control <- exp( -1.356742) / (1 + exp( -1.356742))
exp(-1.239543) / (1 - control + (control * exp(-1.239543)))
```

```
## [1] 0.3388083
```

COMPAS Predictive Accuracy

```
library(survival)
```

```
## Warning: package 'survival' was built under R version 3.3.2
```

```
library(ggfortify)


data <- filter(filter(read.csv(file="~/Desktop/Senior Year/Comp Stats/thesis/cox-parsed.csv", header = T
        mutate(race_factor = factor(race,
```

```
                                  labels = c("African-American",
                                             "Asian",
                                             "Caucasian",
                                             "Hispanic",
                                             "Native American",
                                             "Other"))) %>%
        within(race_factor <- relevel(race_factor, ref = 2)) %>%
        mutate(score_factor = factor(score_text)) %>%
        within(score_factor <- relevel(score_factor, ref=2))

grp <- data[!duplicated(data$id),]
nrow(grp)
```

```
## [1] 10314
```

```
summary(grp$score_factor)
```

```
##    Low   High Medium
##   5751   1952   2611
```

```
summary(grp$race_factor)
```

```
##            Asian African-American        Caucasian         Hispanic
##               51             5147             3569              944
##  Native American            Other
##               32              571
```

```
f <- Surv(start, end, event, type="counting") ~ score_factor
model <- coxph(f, data=data)
summary(model)
```

```
## Call:
## coxph(formula = f, data = data)
##
##   n= 13344, number of events= 3469
##
##                       coef exp(coef) se(coef)     z Pr(>|z|)
## score_factorHigh   1.24969   3.48927  0.04146 30.14   <2e-16 ***
## score_factorMedium 0.79627   2.21725  0.04077 19.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                    exp(coef) exp(-coef) lower .95 upper .95
## score_factorHigh       3.489     0.2866     3.217     3.785
## score_factorMedium     2.217     0.4510     2.047     2.402
##
## Concordance= 0.636  (se = 0.004 )
## Rsquare= 0.068   (max possible= 0.99 )
## Likelihood ratio test= 942.8  on 2 df,   p=0
## Wald test            = 954.8  on 2 df,   p=0
## Score (logrank) test = 1055  on 2 df,   p=0
```

```
decile_f <- Surv(start, end, event, type="counting") ~ decile_score
dmodel <- coxph(decile_f, data=data)
summary(dmodel)
```

```
## Call:
```

```
## coxph(formula = decile_f, data = data)
##
##   n= 13344, number of events= 3469
##
##                  coef exp(coef) se(coef)     z Pr(>|z|)
## decile_score 0.194931  1.215228 0.005801 33.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##              exp(coef) exp(-coef) lower .95 upper .95
## decile_score     1.215     0.8229     1.201     1.229
##
## Concordance= 0.664  (se = 0.005 )
## Rsquare= 0.08   (max possible= 0.99 )
## Likelihood ratio test= 1112  on 1 df,   p=0
## Wald test            = 1129  on 1 df,   p=0
## Score (logrank) test = 1208  on 1 df,   p=0
```

```
f2 <- Surv(start, end, event, type="counting") ~ race_factor + score_factor + race_factor * score_factor
model <- coxph(f2, data=data)
print(summary(model))
```

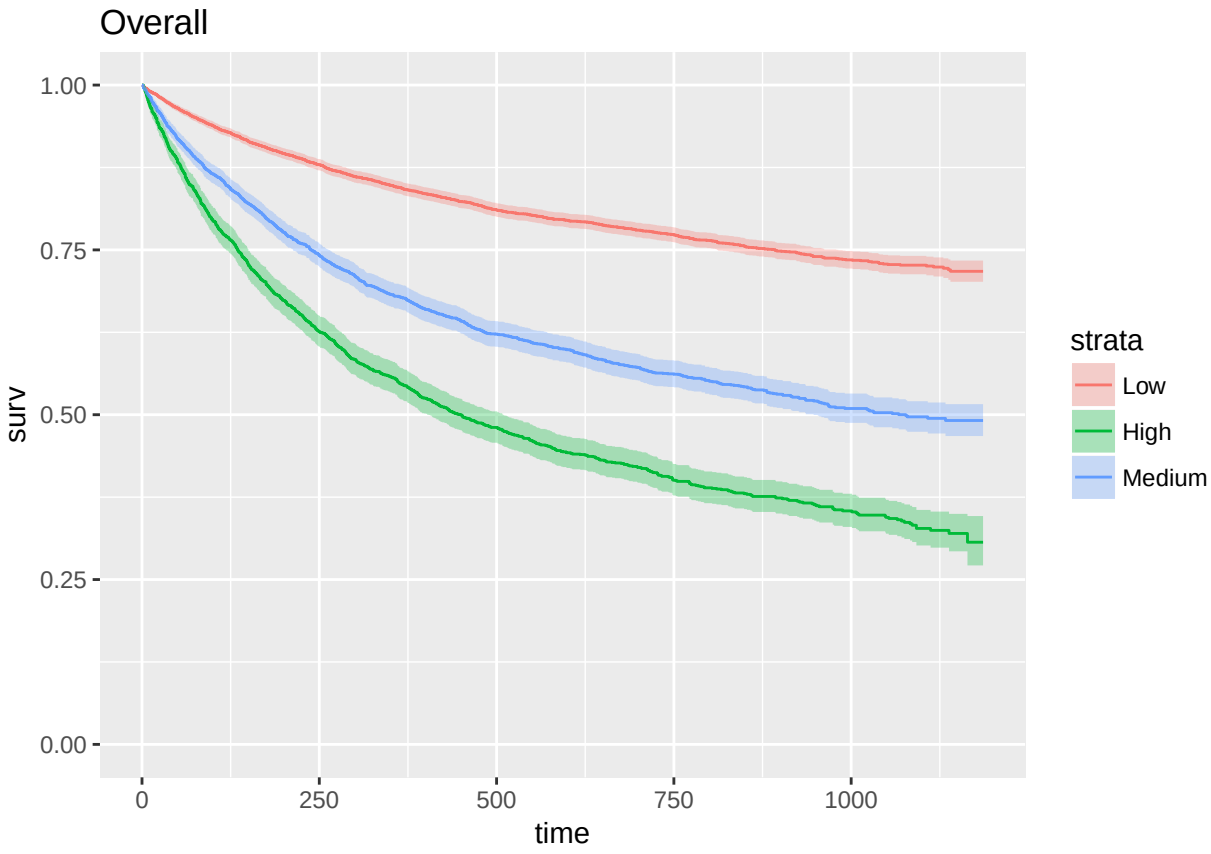
```
## Call:
## coxph(formula = f2, data = data)
##
##   n= 13344, number of events= 3469
##
##                                                    coef exp(coef) se(coef)
## race_factorAfrican-American                      1.0557    2.8741   0.5017
## race_factorCaucasian                             0.7769    2.1748   0.5020
## race_factorHispanic                              0.7134    2.0410   0.5073
## race_factorNative American                      -0.4777    0.6202   1.1180
## race_factorOther                                 0.7911    2.2059   0.5101
## score_factorHigh                                 2.5991   13.4511   0.7638
## score_factorMedium                               1.8291    6.2280   0.7071
## race_factorAfrican-American:score_factorHigh    -1.5053    0.2219   0.7658
## race_factorCaucasian:score_factorHigh           -1.3156    0.2683   0.7684
## race_factorHispanic:score_factorHigh            -1.4347    0.2382   0.7847
## race_factorNative American:score_factorHigh      0.6402    1.8968   1.3229
## race_factorOther:score_factorHigh               -0.9010    0.4062   0.8022
## race_factorAfrican-American:score_factorMedium -1.1588    0.3139   0.7094
## race_factorCaucasian:score_factorMedium         -0.9862    0.3730   0.7107
## race_factorHispanic:score_factorMedium          -0.9209    0.3982   0.7224
## race_factorNative American:score_factorMedium    0.4035    1.4970   1.3229
## race_factorOther:score_factorMedium             -1.3201    0.2671   0.7407
##                                                      z Pr(>|z|)
## race_factorAfrican-American                      2.104 0.035349 *
## race_factorCaucasian                             1.548 0.121701
## race_factorHispanic                              1.406 0.159635
## race_factorNative American                      -0.427 0.669155
## race_factorOther                                 1.551 0.120922
## score_factorHigh                                 3.403 0.000667 ***
## score_factorMedium                               2.587 0.009692 **
## race_factorAfrican-American:score_factorHigh    -1.966 0.049321 *
## race_factorCaucasian:score_factorHigh           -1.712 0.086862 .
```

```
## race_factorHispanic:score_factorHigh           -1.828 0.067495 .
## race_factorNative American:score_factorHigh     0.484 0.628459
## race_factorOther:score_factorHigh              -1.123 0.261391
## race_factorAfrican-American:score_factorMedium -1.634 0.102347
## race_factorCaucasian:score_factorMedium        -1.388 0.165257
## race_factorHispanic:score_factorMedium         -1.275 0.202387
## race_factorNative American:score_factorMedium   0.305 0.760379
## race_factorOther:score_factorMedium            -1.782 0.074719 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                                exp(coef) exp(-coef)
## race_factorAfrican-American                       2.8741    0.34794
## race_factorCaucasian                              2.1748    0.45981
## race_factorHispanic                               2.0410    0.48997
## race_factorNative American                        0.6202    1.61244
## race_factorOther                                  2.2059    0.45333
## score_factorHigh                                 13.4511    0.07434
## score_factorMedium                                6.2280    0.16056
## race_factorAfrican-American:score_factorHigh      0.2219    4.50558
## race_factorCaucasian:score_factorHigh             0.2683    3.72684
## race_factorHispanic:score_factorHigh              0.2382    4.19835
## race_factorNative American:score_factorHigh       1.8968    0.52721
## race_factorOther:score_factorHigh                 0.4062    2.46209
## race_factorAfrican-American:score_factorMedium    0.3139    3.18611
## race_factorCaucasian:score_factorMedium           0.3730    2.68101
## race_factorHispanic:score_factorMedium            0.3982    2.51156
## race_factorNative American:score_factorMedium     1.4970    0.66801
## race_factorOther:score_factorMedium               0.2671    3.74389
##                                                lower .95 upper .95
## race_factorAfrican-American                      1.07512    7.6831
## race_factorCaucasian                             0.81304    5.8173
## race_factorHispanic                              0.75512    5.5163
## race_factorNative American                       0.06932    5.5487
## race_factorOther                                 0.81167    5.9949
## score_factorHigh                                 3.01035   60.1037
## score_factorMedium                               1.55757   24.9030
## race_factorAfrican-American:score_factorHigh     0.04948    0.9955
## race_factorCaucasian:score_factorHigh            0.05952    1.2097
## race_factorHispanic:score_factorHigh             0.05117    1.1088
## race_factorNative American:score_factorHigh      0.14189   25.3561
## race_factorOther:score_factorHigh                0.08430    1.9569
## race_factorAfrican-American:score_factorMedium   0.07815    1.2605
## race_factorCaucasian:score_factorMedium          0.09263    1.5020
## race_factorHispanic:score_factorMedium           0.09664    1.6405
## race_factorNative American:score_factorMedium    0.11199   20.0108
## race_factorOther:score_factorMedium              0.06254    1.1407
##
## Concordance= 0.646  (se = 0.005 )
## Rsquare= 0.072   (max possible= 0.99 )
## Likelihood ratio test= 993.7  on 17 df,   p=0
## Wald test            = 988.8  on 17 df,   p=0
## Score (logrank) test = 1105  on 17 df,   p=0
```

```{r eval=FALSE} import math print("Black High Hazard: %.2f" % (math.exp(-0.18976 + 1.28350))) print("White High Hazard: %.2f" % (math.exp(1.28350))) print("Black Medium Hazard: %.2f" % (math.exp(0.84286-0.17261))) print("White Medium Hazard: %.2f" % (math.exp(0.84286))) #
```

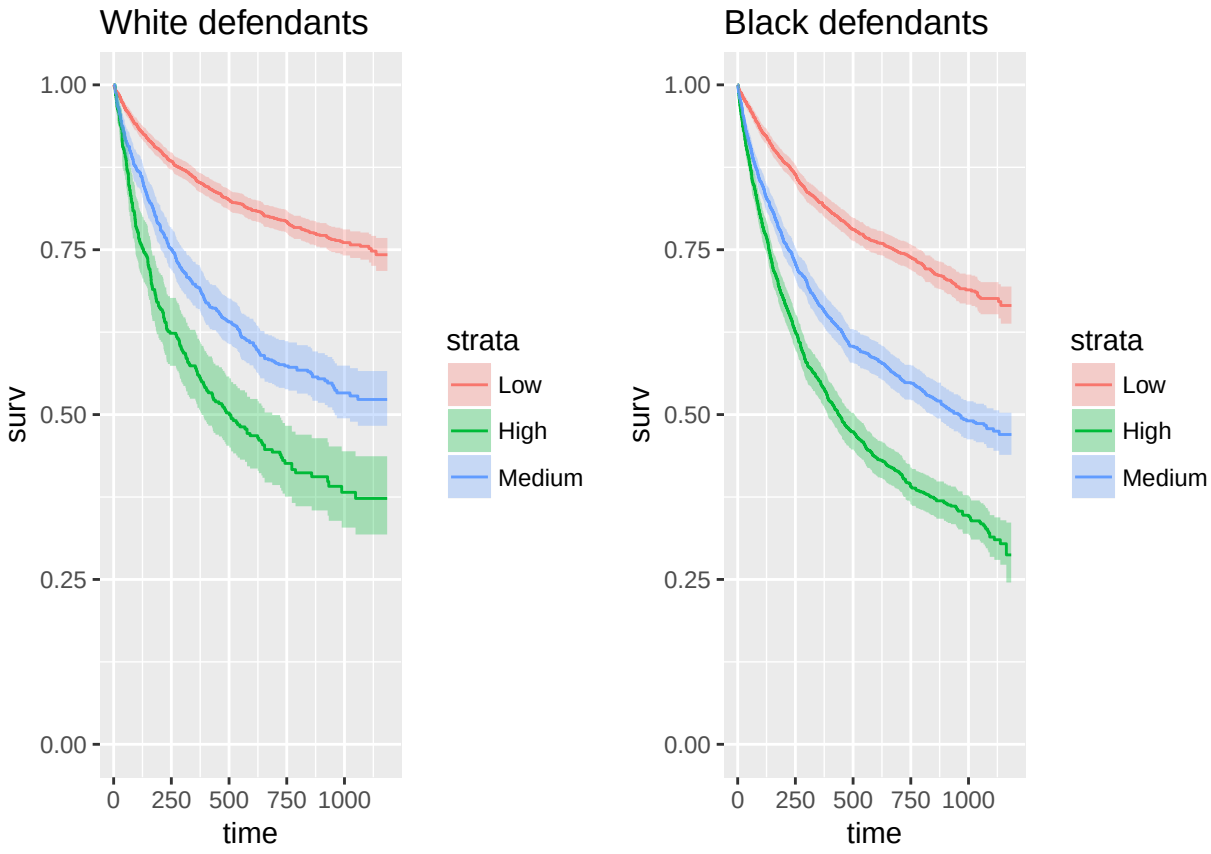
```
fit <- survfit(f, data=data)

plotty <- function(fit, title) {
  return(autoplot(fit, conf.int=T, censor=F) + ggtitle(title) + ylim(0,1))
}
plotty(fit, "Overall")
```

```
## Scale for 'y' is already present. Adding another scale for 'y', which
## will replace the existing scale.
```




```
white <- filter(data, race == "Caucasian")
white_fit <- survfit(f, data=white)

black <- filter(data, race == "African-American")
black_fit <- survfit(f, data=black)

grid.arrange(plotty(white_fit, "White defendants"),
             plotty(black_fit, "Black defendants"), ncol=2)
```

```
## Scale for 'y' is already present. Adding another scale for 'y', which
## will replace the existing scale.
## Scale for 'y' is already present. Adding another scale for 'y', which
## will replace the existing scale.
```




```
summary(fit, times=c(730))
```

```
## Call: survfit(formula = f, data = data)
##
##                 score_factor=Low
##         time       n.risk      n.event     censored     survival
##     7.30e+02     2.75e+03     1.21e+03     2.79e+03     7.76e-01
##      std.err lower 95% CI upper 95% CI
##     5.74e-03     7.64e-01     7.87e-01
##
##                 score_factor=High
##         time       n.risk      n.event     censored     survival
##     7.30e+02     5.05e+02     9.72e+02     1.26e+03     4.08e-01
##      std.err lower 95% CI upper 95% CI
##     1.22e-02     3.85e-01     4.33e-01
##
##                 score_factor=Medium
##         time       n.risk      n.event     censored     survival
##     7.30e+02     9.79e+02     1.02e+03     1.39e+03     5.63e-01
##      std.err lower 95% CI upper 95% CI
##     1.03e-02     5.43e-01     5.84e-01
```

```r
summary(black_fit, times=c(730))
```

```
## Call: survfit(formula = f, data = black)
##
##                 score_factor=Low
##         time       n.risk      n.event     censored     survival
##     7.30e+02     1.02e+03     5.29e+02     1.04e+03     7.43e-01
##      std.err lower 95% CI upper 95% CI
##     9.70e-03     7.24e-01     7.62e-01
##
##                 score_factor=High
##         time       n.risk      n.event     censored     survival
##     730.0000     362.0000     719.0000     914.0000       0.3976
##      std.err lower 95% CI upper 95% CI
##       0.0142       0.3707       0.4265
##
##                 score_factor=Medium
##         time       n.risk      n.event     censored     survival
##     730.0000     578.0000     623.0000     785.0000       0.5485
##      std.err lower 95% CI upper 95% CI
##       0.0134       0.5227       0.5755
```

```r
summary(white_fit, times=c(730))
```

```
## Call: survfit(formula = f, data = white)
##
##                 score_factor=Low
##         time       n.risk      n.event     censored     survival
##     7.30e+02     1.16e+03     4.57e+02     1.18e+03     7.95e-01
##      std.err lower 95% CI upper 95% CI
##     8.63e-03     7.78e-01     8.12e-01
##
##                 score_factor=High
##         time       n.risk      n.event     censored     survival
##     730.0000     102.0000     191.0000     278.0000       0.4347
##      std.err lower 95% CI upper 95% CI
##       0.0272       0.3846       0.4914
##
##                 score_factor=Medium
##         time       n.risk      n.event     censored     survival
##     730.0000     299.0000     306.0000     460.0000       0.5757
##      std.err lower 95% CI upper 95% CI
##       0.0185       0.5405       0.6132
```

```r
summary(coxph(f, data=white))
```

```
## Call:
## coxph(formula = f, data = white)
##
##   n= 4564, number of events= 1023
##
##                       coef exp(coef) se(coef)     z Pr(>|z|)
## score_factorHigh   1.27628   3.58327  0.08365 15.26   <2e-16 ***
## score_factorMedium 0.83965   2.31556  0.07144 11.75   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                    exp(coef) exp(-coef) lower .95 upper .95
## score_factorHigh       3.583     0.2791     3.041     4.222
## score_factorMedium     2.316     0.4319     2.013     2.664
##
## Concordance= 0.625  (se = 0.007 )
## Rsquare= 0.056   (max possible= 0.971 )
## Likelihood ratio test= 262.8  on 2 df,   p=0
## Wald test            = 282.2  on 2 df,   p=0
## Score (logrank) test = 311.7  on 2 df,   p=0
```

```
summary(coxph(f, data=black))
```

```
## Call:
## coxph(formula = f, data = black)
##
##   n= 6862, number of events= 2035
##
##                       coef exp(coef) se(coef)     z Pr(>|z|)
## score_factorHigh   1.09514   2.98959  0.05475 20.00   <2e-16 ***
## score_factorMedium 0.67025   1.95473  0.05636 11.89   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                    exp(coef) exp(-coef) lower .95 upper .95
## score_factorHigh       2.990     0.3345     2.685     3.328
## score_factorMedium     1.955     0.5116     1.750     2.183
##
## Concordance= 0.623  (se = 0.006 )
## Rsquare= 0.059   (max possible= 0.992 )
## Likelihood ratio test= 416.9  on 2 df,   p=0
## Wald test            = 401.3  on 2 df,   p=0
## Score (logrank) test = 432.9  on 2 df,   p=0
```