PIMA Dataset

Vitaly Druker

9/24/2020

Lab with pima dataset (question 3 of exercise)

```
d <- rpart::kyphosis
```

```
d %>% glimpse
```

```
## Rows: 81
## Columns: 4
## $ Kyphosis <fct> absent, absent, present, absent, absent, absent, absent, a...
## $ Age      <int> 71, 158, 128, 2, 1, 1, 61, 37, 113, 59, 82, 148, 18, 1, 16...
## $ Number   <int> 3, 3, 4, 5, 4, 2, 2, 3, 2, 6, 5, 3, 5, 4, 3, 3, 6, 5, 5, 4...
## $ Start    <int> 5, 14, 5, 1, 15, 16, 17, 16, 16, 12, 14, 16, 2, 12, 18, 16...
```

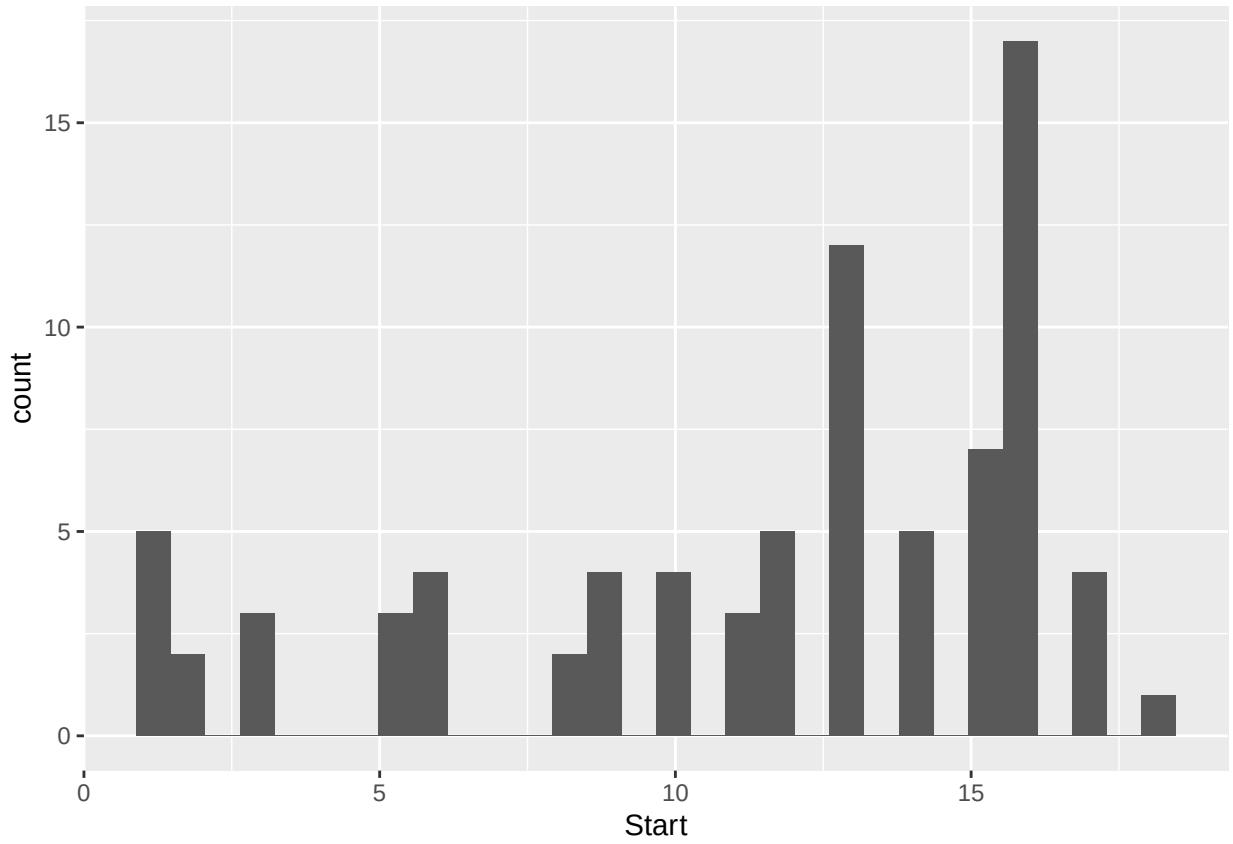
```
?rpart::kyphosis
```

```
summary(d)
```

```
##     Kyphosis       Age              Number           Start
##  absent :64   Min.   :  1.00   Min.   : 2.000   Min.   : 1.00
##  present:17   1st Qu.: 26.00   1st Qu.: 3.000   1st Qu.: 9.00
##               Median : 87.00   Median : 4.000   Median :13.00
##               Mean   : 83.65   Mean   : 4.049   Mean   :11.49
##               3rd Qu.:130.00   3rd Qu.: 5.000   3rd Qu.:16.00
##               Max.   :206.00   Max.   :10.000   Max.   :18.00
```

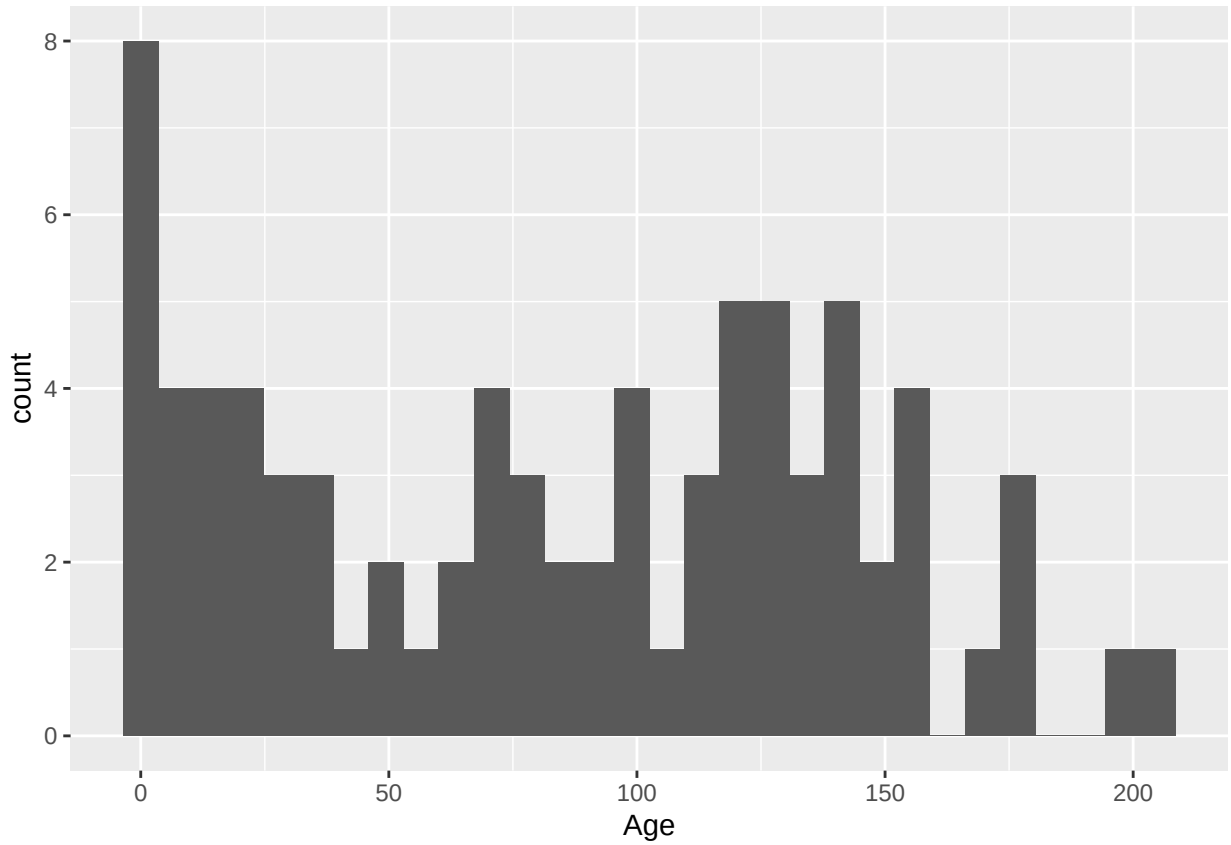
```
# Histograms
d %>%
  ggplot(aes(x = Start)) +
  geom_histogram()
```

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

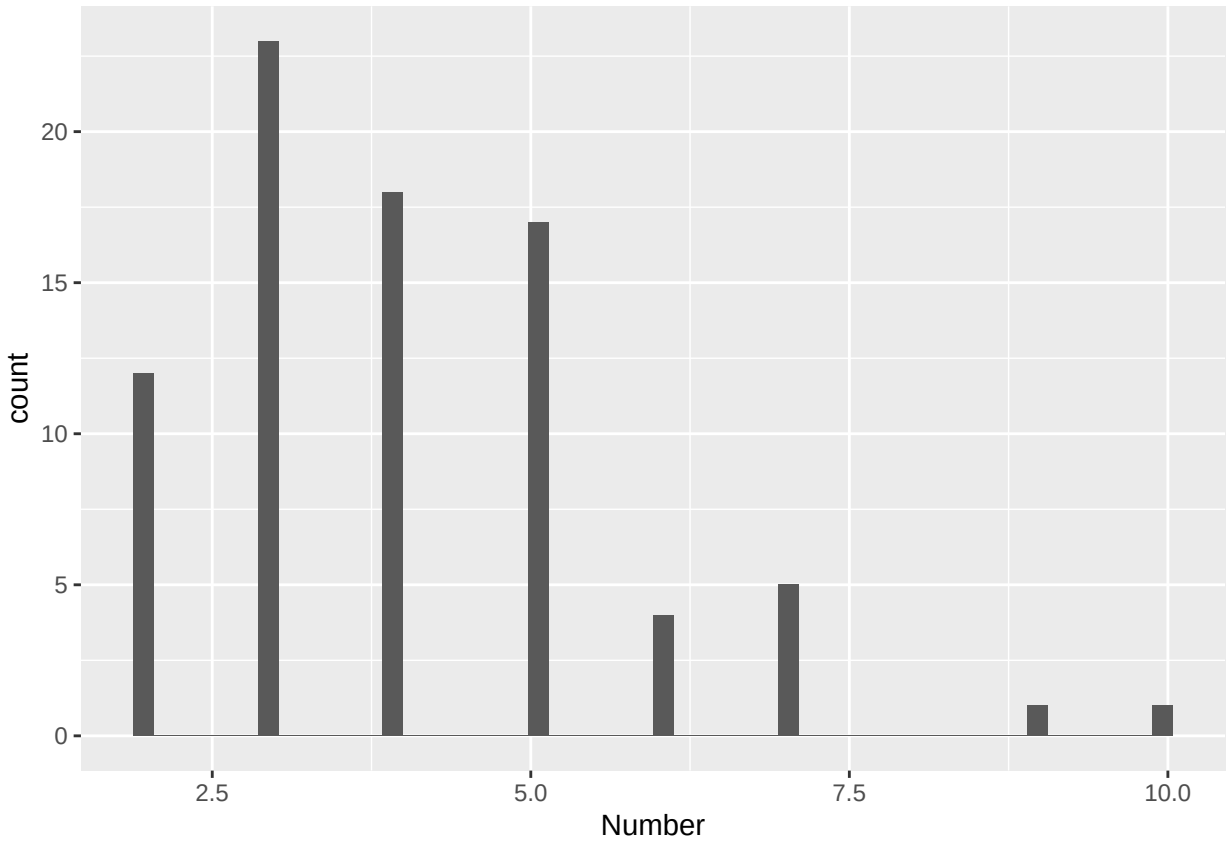



```
d %>%
  ggplot(aes(x = Age)) +
  geom_histogram()
```

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

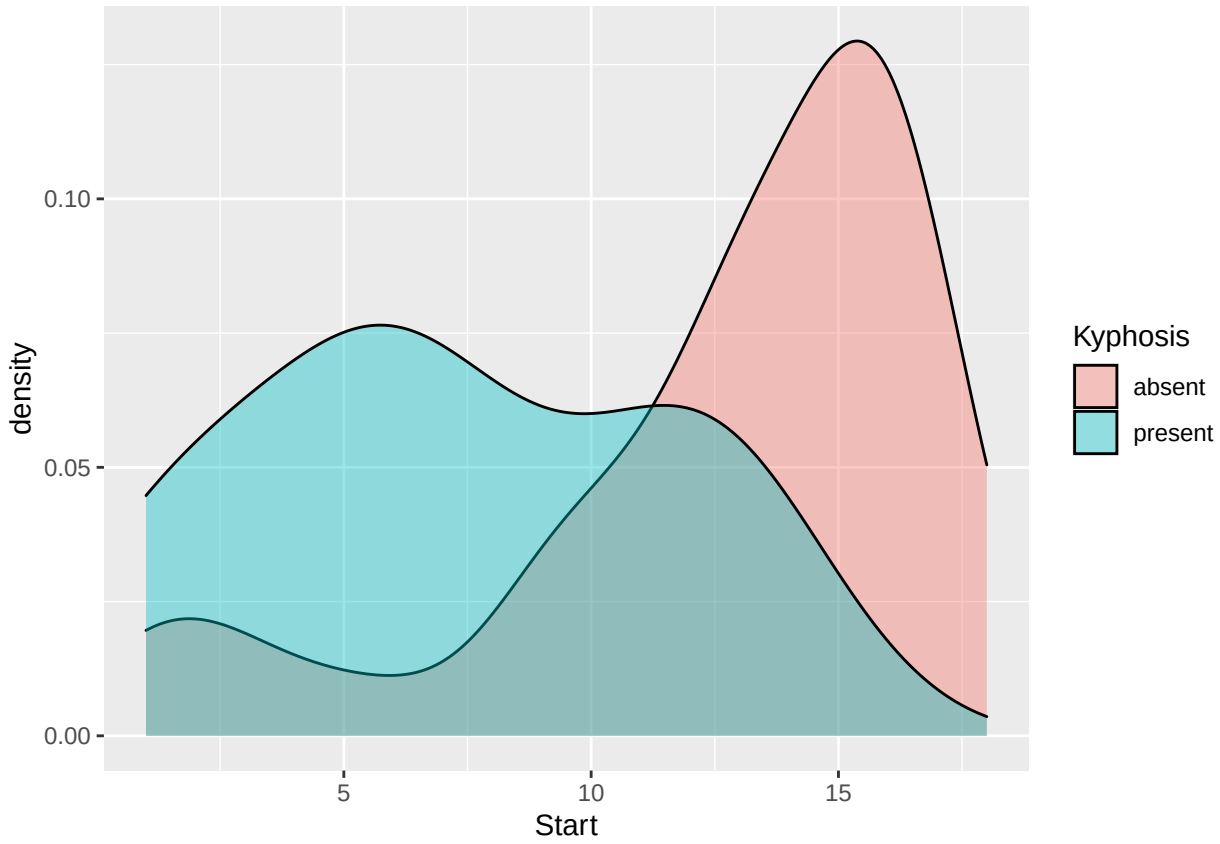



```
d %>%
  ggplot(aes(x = Number)) +
  geom_histogram(bins = 50)
```

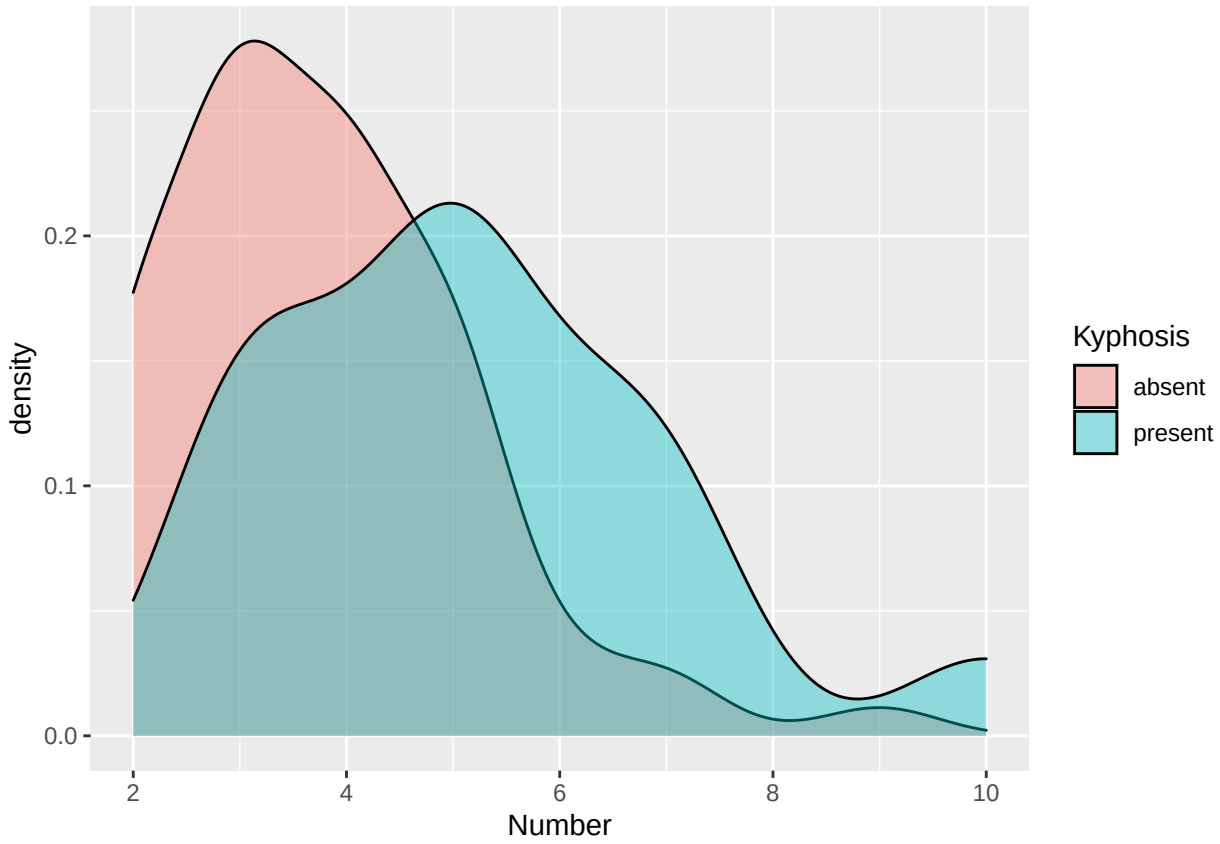


The first one shows a pretty flat graph, with that big spike starting at 12. The second one skews a little bit younger but it's pretty evenly distributed The third one shows a little peak and some outliers.

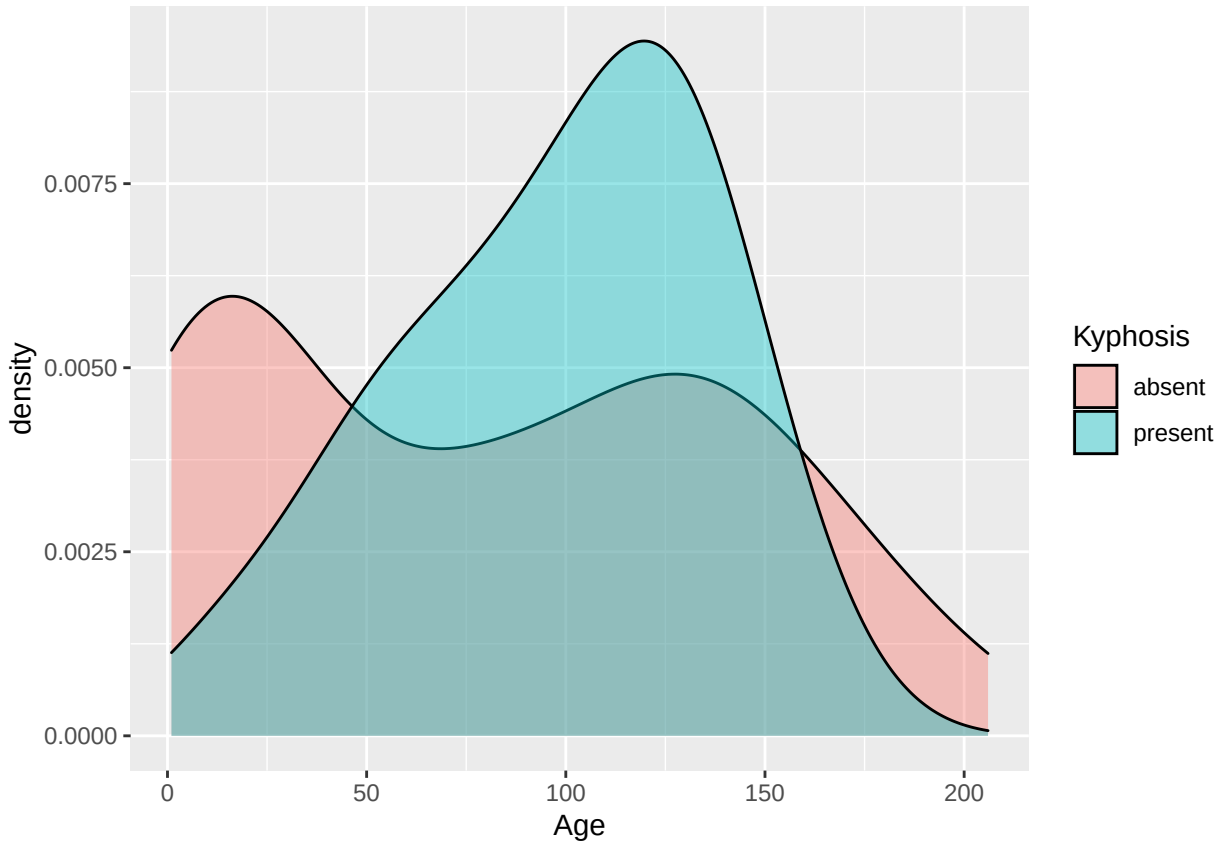
```
# Univariate graphs
d %>%
  ggplot(aes(x = Start, fill = Kyphosis)) +
  geom_density(alpha = .4)
```




```
d %>%
  ggplot(aes(x = Number, fill = Kyphosis)) +
  geom_density(alpha = .4)
```

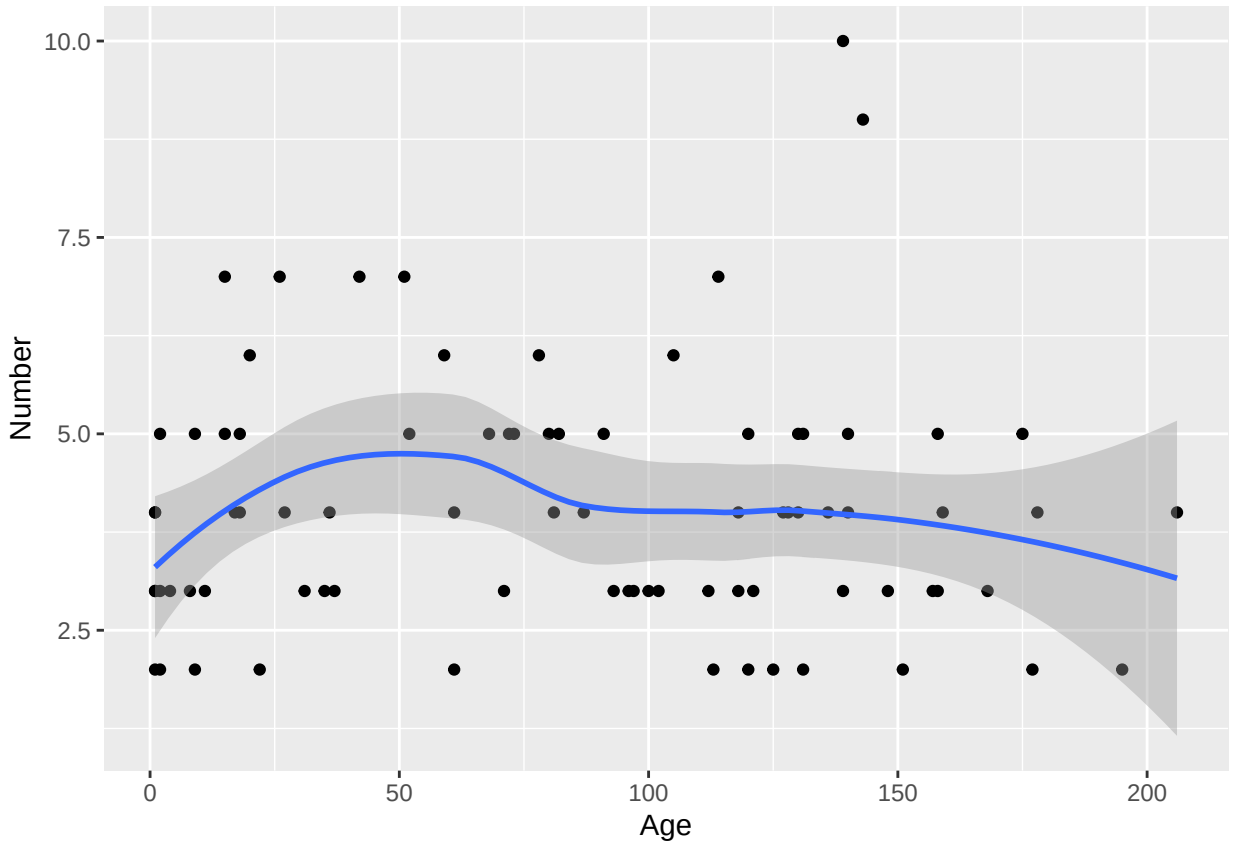



```
d %>%
  ggplot(aes(x = Age, fill = Kyphosis)) +
  geom_density(alpha = .4)
```

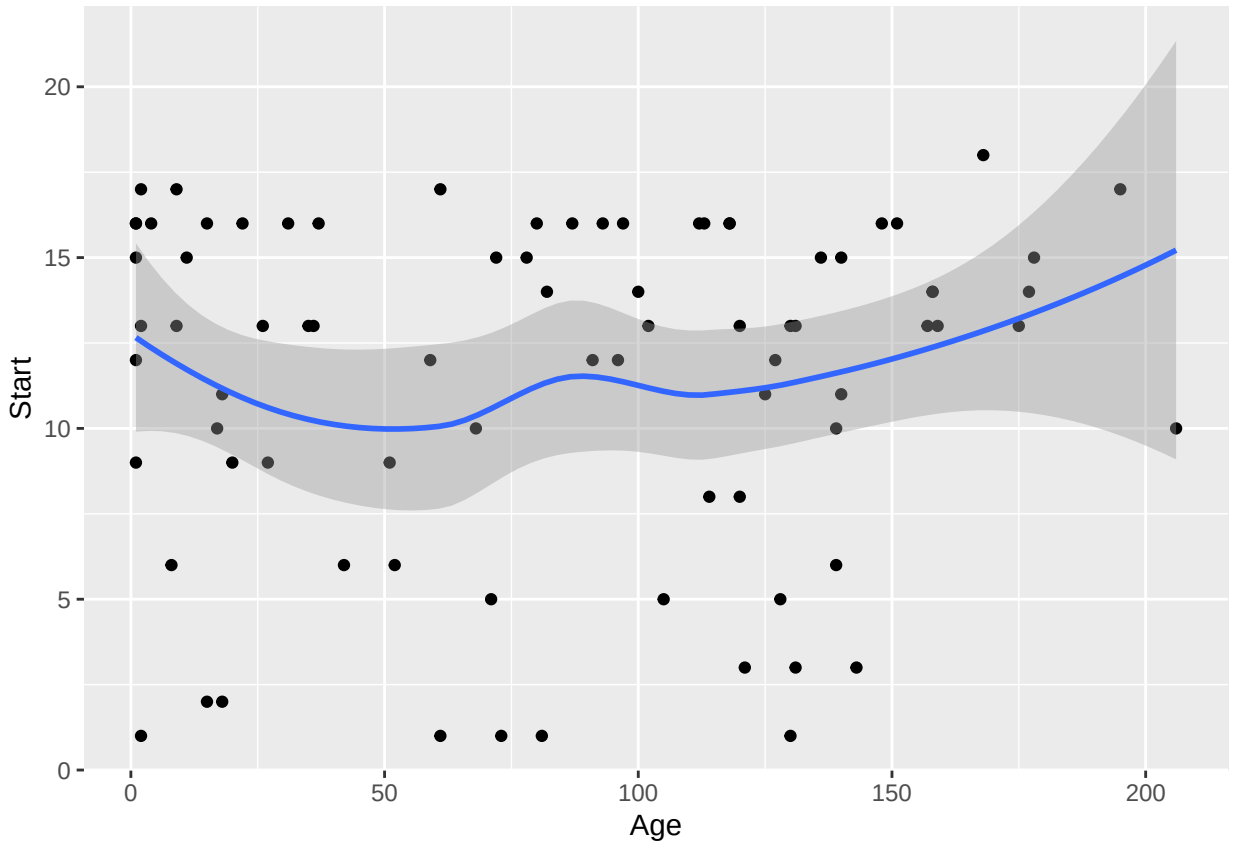


The first one looks like the higher you start, the more likely you're to have kyphosis The second one also looks like there's some separation here. Looks like the more vertebrae you involve, the more likely you'll have issues afterwards The third one looks pretty flat for the people who don't have kyphosis. But seems to peak over at age 100. Looks like the younger people do better than the people who have aged 100 months.

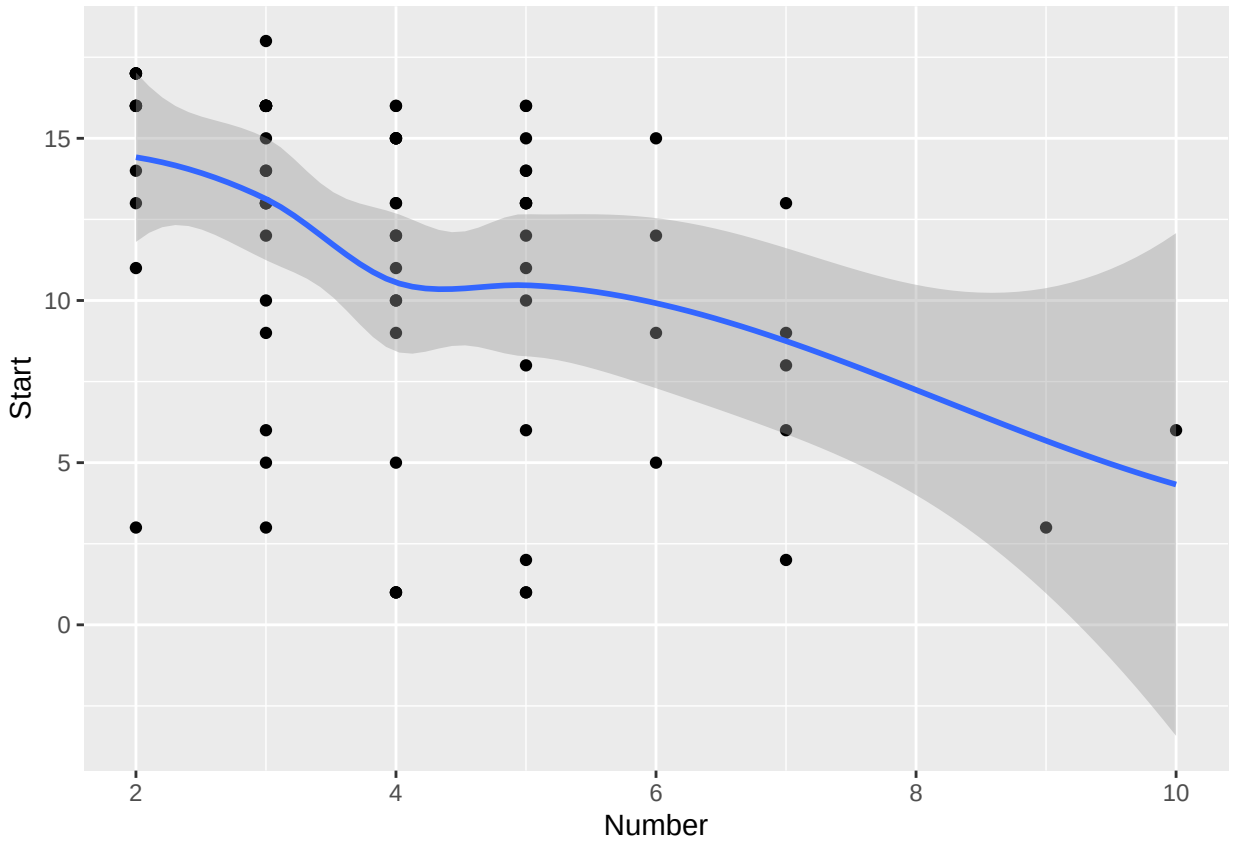
```
# Correlation of the variables together
d %>%
  ggplot(aes(x = Age, y = Number)) +
  geom_point() +
  geom_smooth(method = 'loess', formula = y~ x)
```




```
d %>%
  ggplot(aes(x = Age, y = Start)) +
  geom_point() +
  geom_smooth(method = 'loess', formula = y~ x)
```




```
d %>%
  ggplot(aes(x = Number, y = Start)) +
  geom_point() +
  geom_smooth(method = 'loess', formula = y~ x)
```



The first one, no real pattern between age in number, pretty flat The second one, also pretty flat, maybe a little curve going up but nothing pretty obvious. The third one makes sense because the further you start down the spine the fewer you can do. it has a nice little negative slope

```
library(broom)
# binomial model
# instead of typing out all the predictors, you can just use period which will take everything that is
mod <- glm(Kyphosis ~ ., data = d, family = binomial)

summary(mod)
```

```
##
## Call:
## glm(formula = Kyphosis ~ ., family = binomial, data = d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3124  -0.5484  -0.3632  -0.1659   2.1613
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.036934   1.449575  -1.405  0.15996
## Age          0.010930   0.006446   1.696  0.08996 .
## Number       0.410601   0.224861   1.826  0.06785 .
## Start       -0.206510   0.067699  -3.050  0.00229 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.234  on 80  degrees of freedom
## Residual deviance: 61.380  on 77  degrees of freedom
## AIC: 69.38
##
## Number of Fisher Scoring iterations: 5
```

```
# This is all on a linear scale
mod %>%
  tidy()
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)  -2.04     1.45        -1.41 0.160
## 2 Age           0.0109   0.00645      1.70 0.0900
## 3 Number        0.411    0.225        1.83 0.0678
## 4 Start        -0.207    0.0677      -3.05 0.00229
```

```
# This is all on the odds scale
mod %>%
  tidy(conf.int = T, exponentiate = T)
```

```
## # A tibble: 4 x 7
##   term        estimate std.error statistic p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 (Intercept)    0.130   1.45        -1.41 0.160    0.00604     1.95
## 2 Age            1.01    0.00645      1.70 0.0900   0.999       1.02
## 3 Number         1.51    0.225        1.83 0.0678   1.00        2.45
## 4 Start          0.813   0.0677      -3.05 0.00229  0.705       0.924
```

In the summary function, the start has an estimate from 0.206, plus or minus .06 for the std error with a significant p-value On the odds scale, the start estimate means the increase of one of start decreases the odds. It's 81% of the odds of what we were seeing before.

Start of zero is near the head. so the lower you start, the less likely you are to have kyphosis. But the CI high is at 92%, so if you're only saving .08% what are you losing in the surgery?

```
# Step function
mod2 <- step(mod)
```

```
## Start:  AIC=69.38
## Kyphosis ~ Age + Number + Start
##
##          Df Deviance    AIC
## <none>        61.380 69.380
## - Age     1   64.536 70.536
## - Number  1   65.299 71.299
## - Start   1   71.627 77.627
```

```
#
tidy(mod2)
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)  -2.04     1.45        -1.41 0.160
```

```
## 2 Age           0.0109   0.00645      1.70 0.0900
## 3 Number        0.411    0.225        1.83 0.0678
## 4 Start        -0.207    0.0677      -3.05 0.00229
```

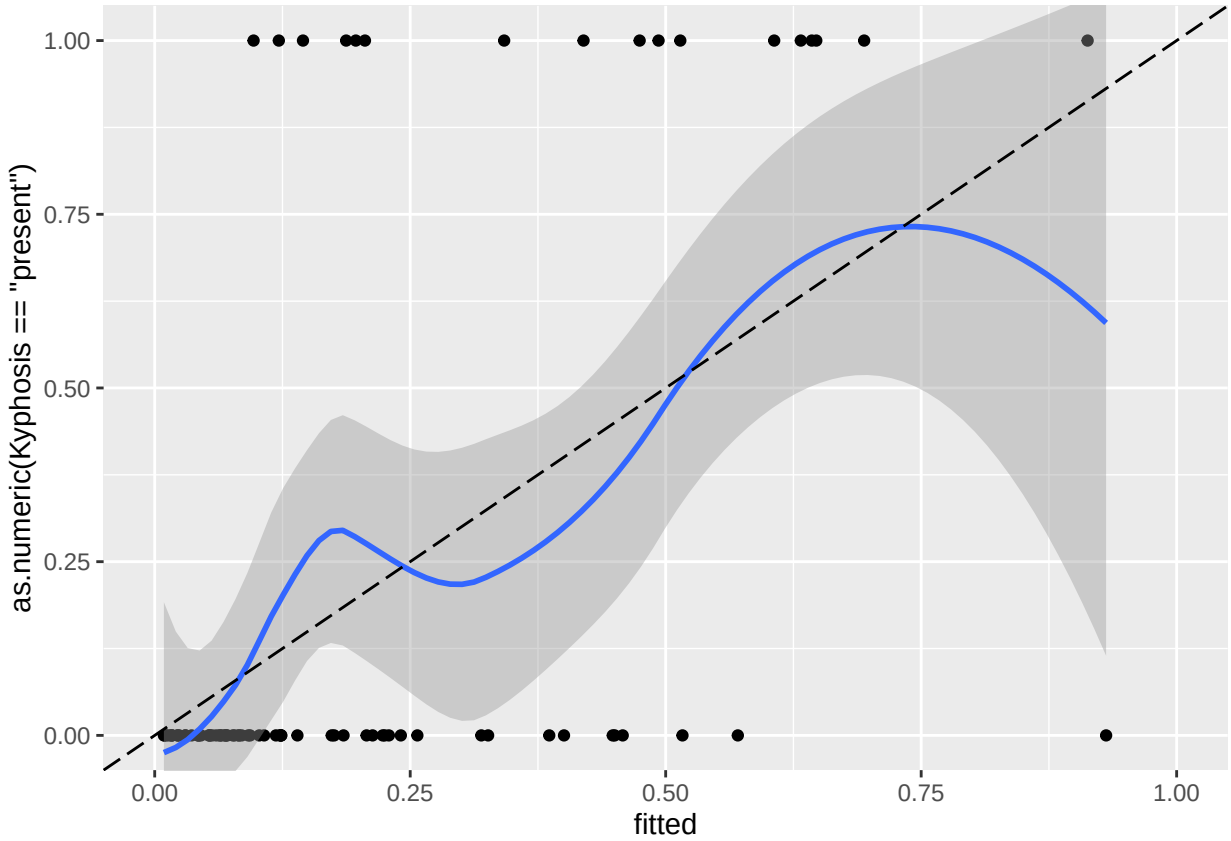
Nothing happens in the step trace. The AIC does not go down at all, aka it's best to leave everything in the model. Basically the AIC is the lowest when we don't subtract anything.

For mod2, our p-values are not significant, but that doesn't mean it's a bad model.

```
# Add some fitted values to this
# Fitted values give you probability right off the bat, you'd have to specify you don't want probably
# Not how the predict function works, which gives you the linear parameter that you didn't have to tran
d$fitted <- fitted.values(mod)

# Calibration curve
d %>%
  ggplot(aes(x = fitted, y = as.numeric(Kyphosis == 'present'))) +
  geom_point() +
  geom_smooth() +
  geom_abline(slope = 1, intercept = 0, linetype = "longdash") +
  coord_cartesian(xlim  = c(0,1), y = c(0, 1))
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



This curve looks a little crazy, but we don't have a lot of data. Ideally, it'd be a little straighter. You can see that the bounds of this curve are pretty wide. And we're actually doing a pretty good job going across the full spectrum here.

Some other things we would do here is look at the sensitivity and the specificity and things like that to see what we could learn. One other thing that'd be interesting to look at is this relationship that we saw between

the number and start. Removing anyone of them wouldn't work because they are both valuable to the model as we saw in the step fn.

One other thing to keep in mind, is that it looks like removing the number has a very small effect on AIC, but removing the start has a much larger effect on AIC. That one way to start thinking about which features are important. For example, what if you were missing a bunch of data for the number, maybe you're okay with that if you have the start variable.

```
?faraway::pima

# To make 0 and 1 into negative and positive
factor(c(0,1,1,0,1), levels = c(1, 0 ), labels = c('pos', 'neg'))
```

```
## [1] neg pos pos neg pos
## Levels: pos neg
```