# homework2

Na Yun Cho

```
library(caret)
```

```
## Loading required package: lattice

## Loading required package: ggplot2
```

```
library(splines)
library(mgcv)
```

```
## Loading required package: nlme

## This is mgcv 1.8-33. For overview type 'help("mgcv-package")'.
```

```
library(pdp)
library(earth)
```

```
## Loading required package: Formula

## Loading required package: plotmo

## Loading required package: plotrix

## Loading required package: TeachingDemos
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::collapse() masks nlme::collapse()
## x dplyr::filter()   masks stats::filter()
## x dplyr::lag()      masks stats::lag()
## x purrr::lift()     masks caret::lift()
## x purrr::partial()  masks pdp::partial()
```

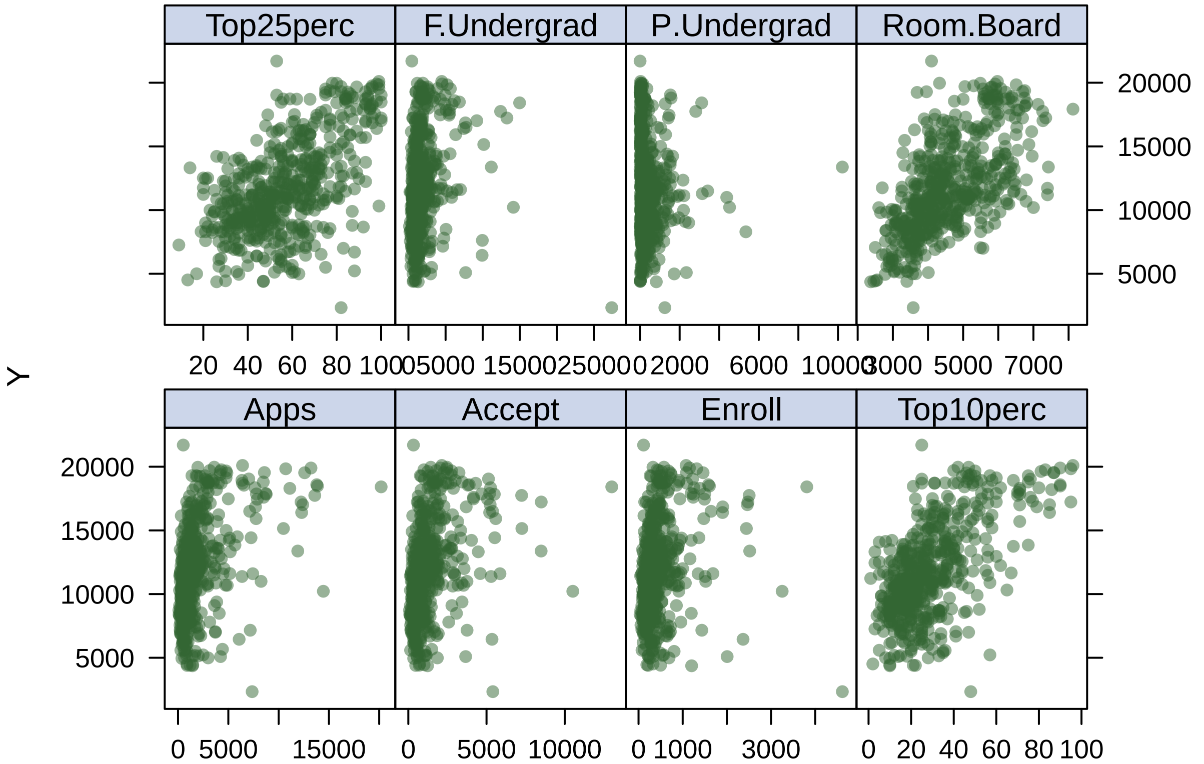
```
library(ggplot2)
```

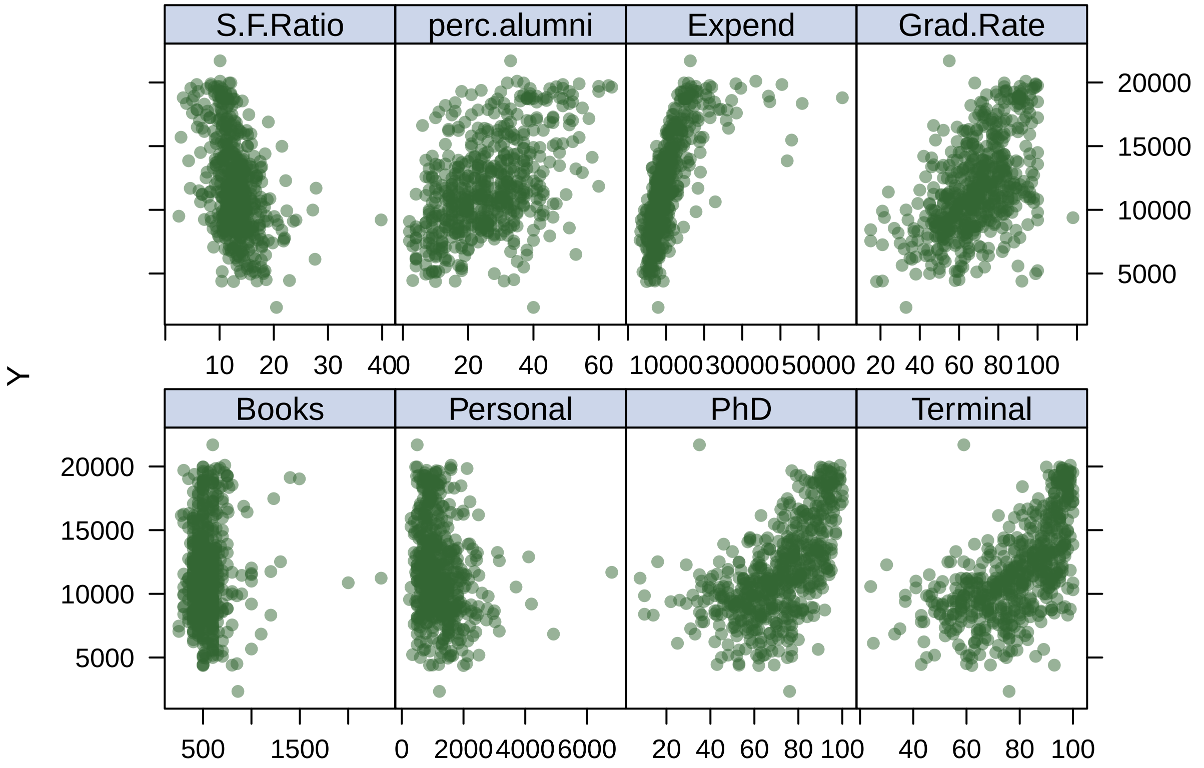
### (a) Exploratory data analysis

```
college = read.csv("./data/data.csv")
college1 <- college[-125,]

college2 = data.matrix(college1, rownames.force = NA)
x <- college2 [ ,-c(1,9)]
y <- college2 [ , 9]


theme1 <- trellis.par.get()
theme1$plot.symbol$col <- rgb(.2, .4, .2, .5)
theme1$plot.symbol$pch <- 16
theme1$plot.line$col <- rgb(.8, .1, .1, 1)
theme1$plot.line$lwd <- 2
theme1$strip.background$col <- rgb(.0, .2, .6, .2)
trellis.par.set(theme1)
featurePlot(x, y, plot = "scatter", labels = c("","Y"),
            type = c("p"), layout = c(4, 2))
```





Interpretation: From this exploratory data analysis, I could see that the predictors ‘F.Undergrad’, ‘P.Undergrad’, ‘Apps’, ‘Accept’, ‘Enroll’, ‘Terminal’, and ‘Books’ show a relatively non-linear trend compared to other predictors. The predictors ‘Top25perc’, ‘Room.Board’, ‘Top10perc’, ‘perc.alumni’, ‘Grad.Rate’, ‘Expend’, and ‘PhD’ showed a generally increasing trend that looks quite linear. On the other hand, ‘S.F.Ratio’ and ‘Personal’ seemed to show a slightly decreasing trend that is quite linear. To check the associations of each predictor with the outcome ‘Outstate’ in more detail, further analyses would have to be done.

## (b) Fit a smoothing spline model using ‘Terminal’ as the only predictor

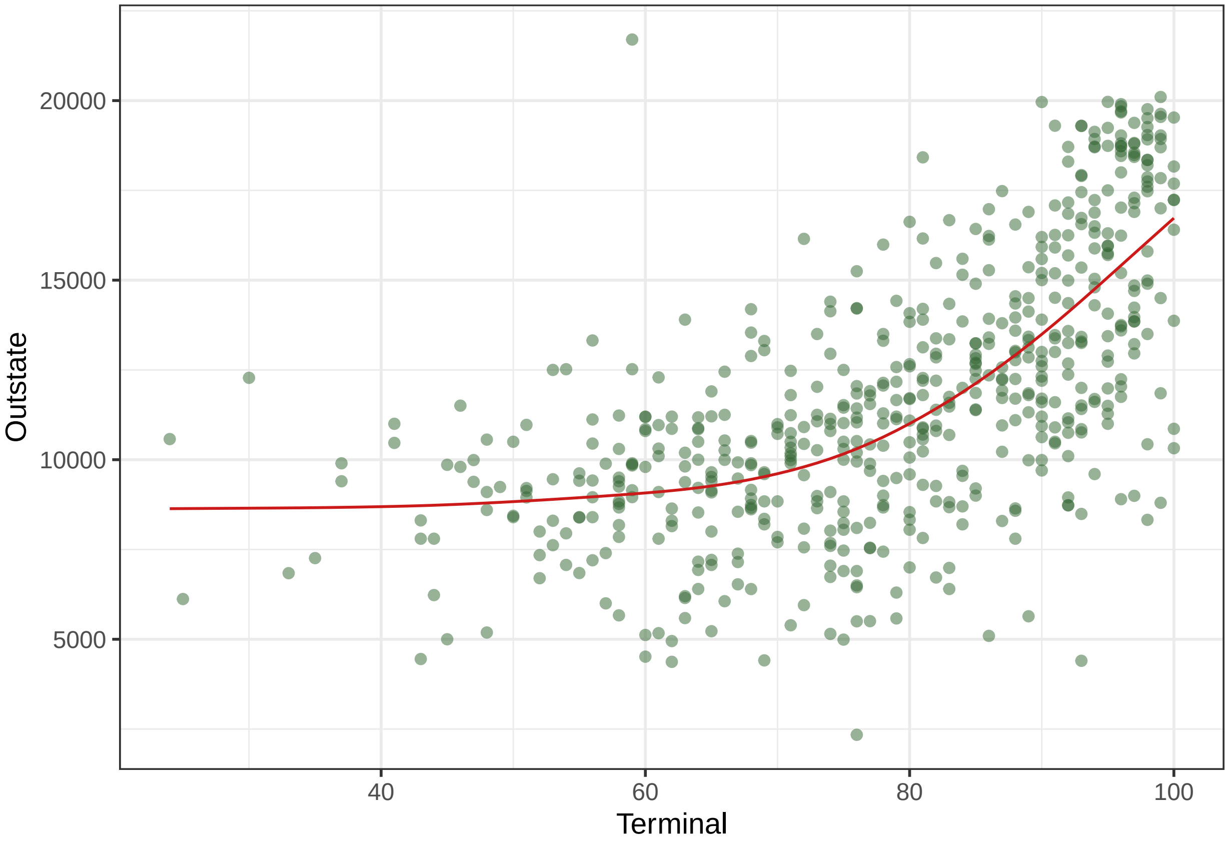
```
# using GCV method
fit.ss <- smooth.spline(college1$Terminal, college1$Outstate)
fit.ss$df
```

```
## [1] 4.468629
```

```
Terminallims <- range(college1$Terminal)
Terminal.grid <- seq(from = Terminallims[1],to = Terminallims[2])

pred.ss <- predict(fit.ss, x = Terminal.grid)
pred.ss.df <- data.frame(pred = pred.ss$y, Terminal = Terminal.grid)

p <- ggplot(data= college1, aes(x = Terminal, y = Outstate)) +geom_point(color = rgb(0.2, 0.4, 0.2, 0.5
p + geom_line(aes(x = Terminal, y =pred), data = pred.ss.df, color = rgb(0.8, 0.1, 0.1, 1)) +theme_bw()
```



```
#Using LOOCV method
fit.ss <- smooth.spline(college1$Terminal, college1$Outstate, cv = TRUE)
```

```
## Warning in smooth.spline(college1$Terminal, college1$Outstate, cv = TRUE):
## cross-validation with non-unique 'x' values seems doubtful
```

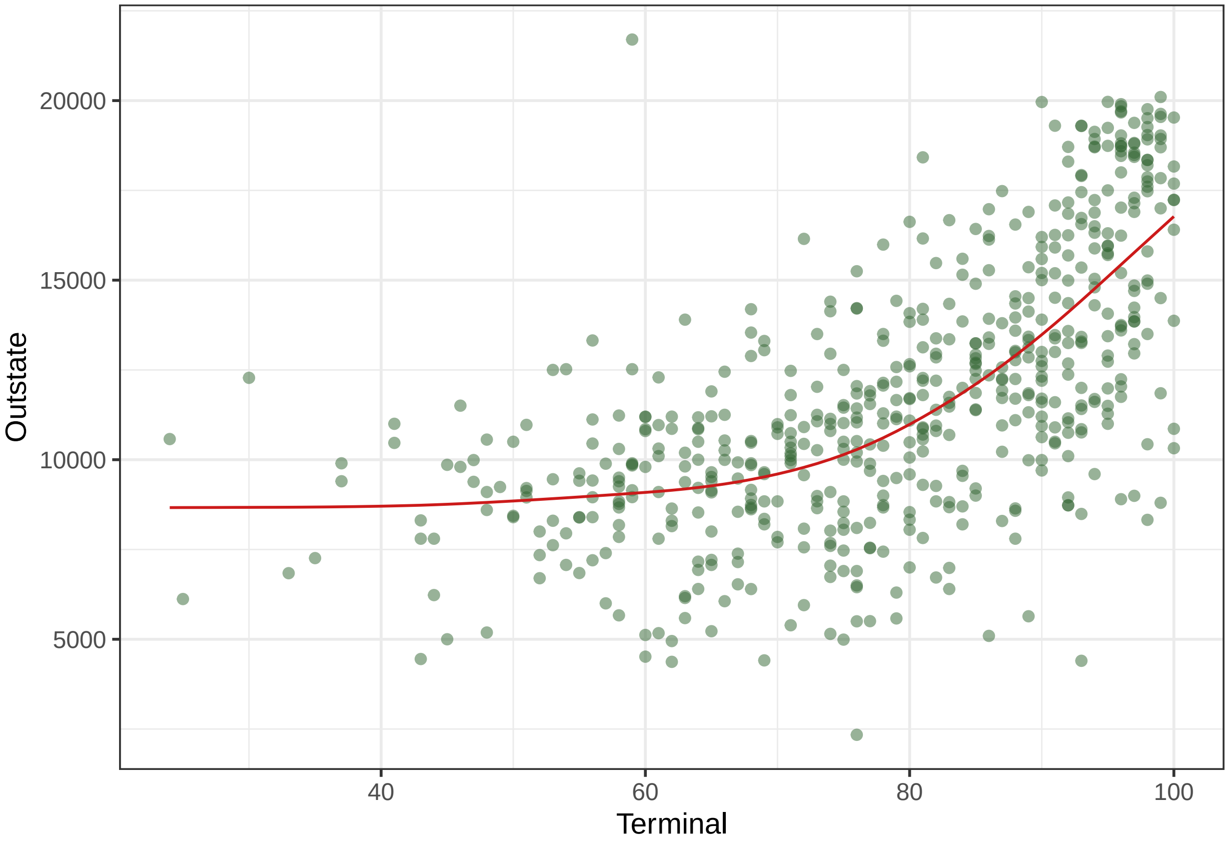
```
fit.ss$df
```

```
## [1] 4.686019
```

```
Terminallims <- range(college1$Terminal)
Terminal.grid <- seq(from = Terminallims[1],to = Terminallims[2])

pred.ss <- predict(fit.ss, x = Terminal.grid)
pred.ss.df <- data.frame(pred = pred.ss$y, Terminal = Terminal.grid)

p <- ggplot(data= college1, aes(x = Terminal, y = Outstate)) +geom_point(color = rgb(0.2, 0.4, 0.2, 0.5
p + geom_line(aes(x = Terminal, y =pred), data = pred.ss.df, color = rgb(0.8, 0.1, 0.1, 1)) +theme_bw()
```



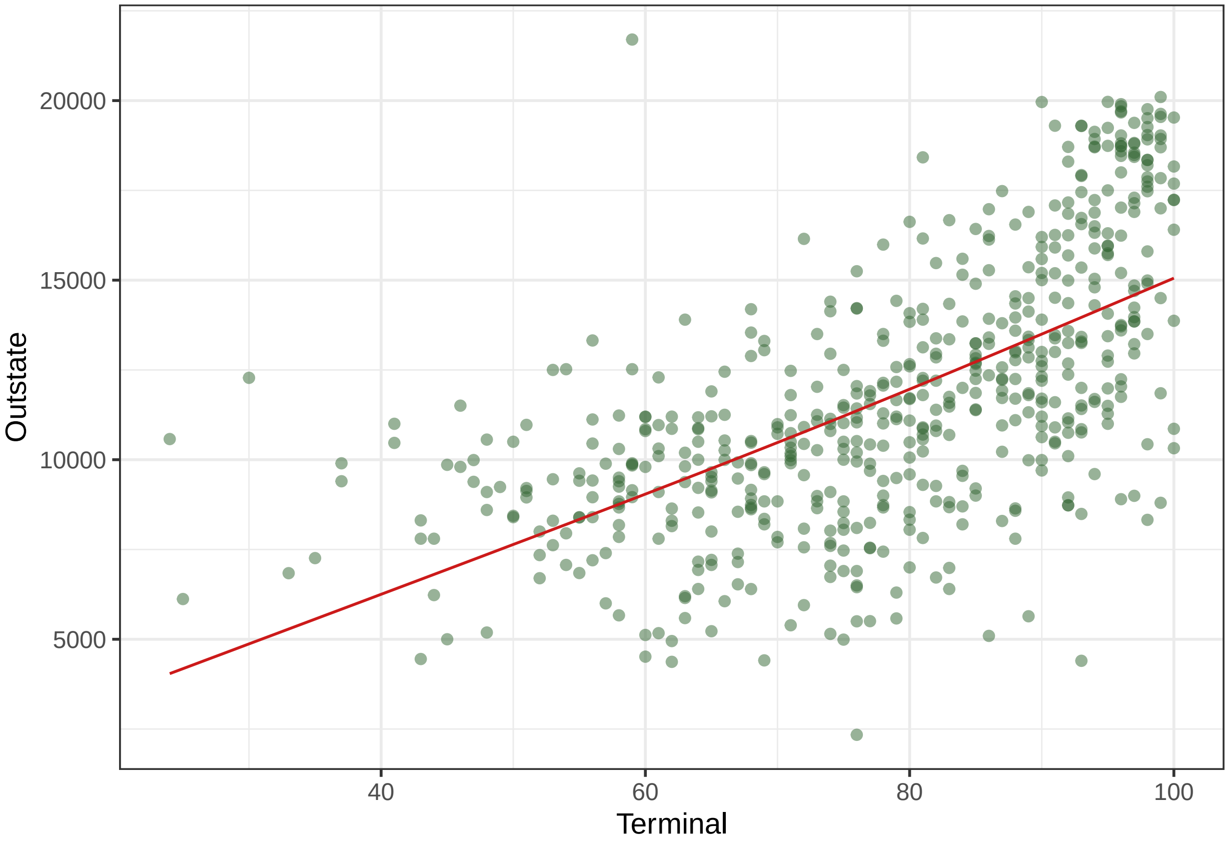
```
#Using arbitrary lambda values
#Using lambda = 10
fit.ss <- smooth.spline(college1$Terminal, college1$Outstate, lambda=10)
fit.ss$df
```

```
## [1] 2.06511
```

```
Terminallims <- range(college1$Terminal)
Terminal.grid <- seq(from = Terminallims[1],to = Terminallims[2])

pred.ss <- predict(fit.ss, x = Terminal.grid)
pred.ss.df <- data.frame(pred = pred.ss$y, Terminal = Terminal.grid)

p <- ggplot(data= college1, aes(x = Terminal, y = Outstate)) +geom_point(color = rgb(0.2, 0.4, 0.2, 0.5
p + geom_line(aes(x = Terminal, y =pred), data = pred.ss.df, color = rgb(0.8, 0.1, 0.1, 1)) +theme_bw()
```



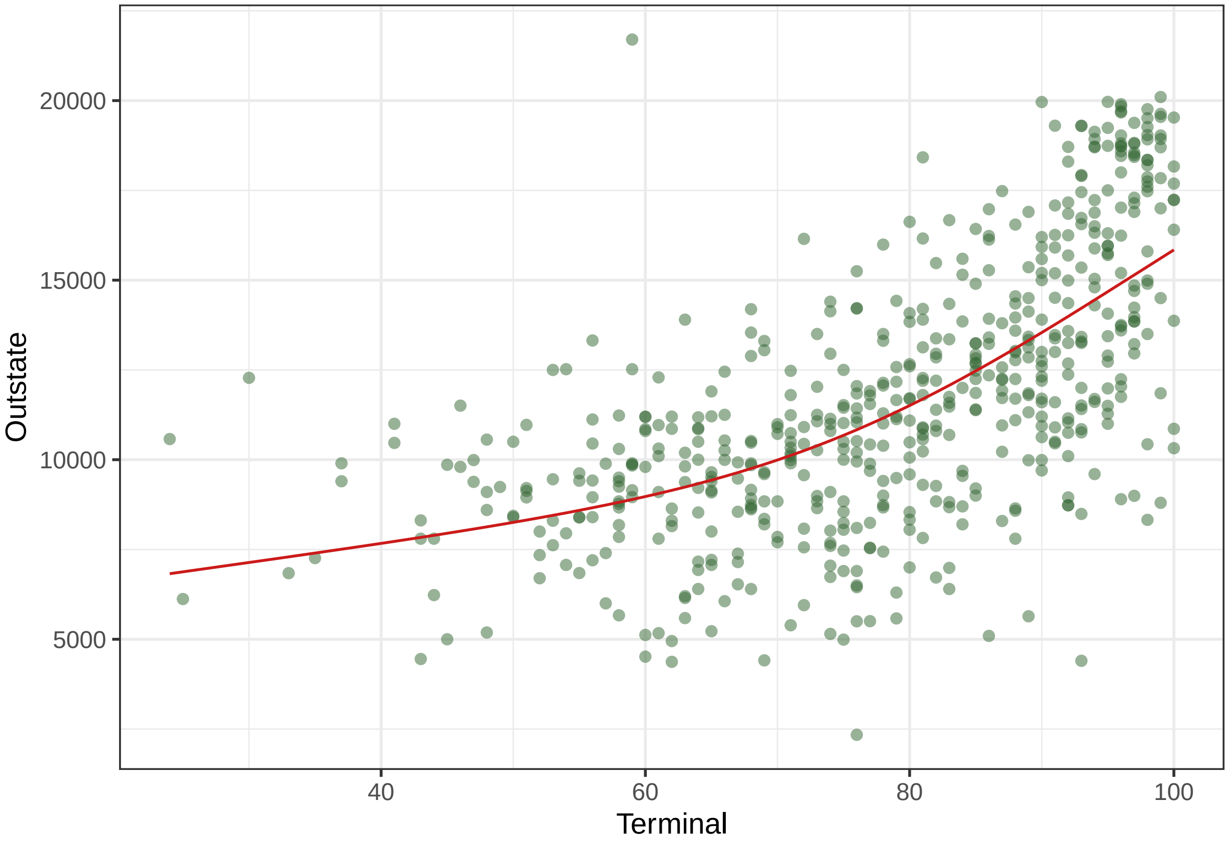
```
#Using lambda = 0.5
fit.ss <- smooth.spline(college1$Terminal, college1$Outstate, lambda=0.5)
fit.ss$df
```

```
## [1] 2.761186
```

```
Terminallims <- range(college1$Terminal)
Terminal.grid <- seq(from = Terminallims[1],to = Terminallims[2])

pred.ss <- predict(fit.ss, x = Terminal.grid)
pred.ss.df <- data.frame(pred = pred.ss$y, Terminal = Terminal.grid)

p <- ggplot(data= college1, aes(x = Terminal, y = Outstate)) +geom_point(color = rgb(0.2, 0.4, 0.2, 0.5
p + geom_line(aes(x = Terminal, y =pred), data = pred.ss.df, color = rgb(0.8, 0.1, 0.1, 1)) +theme_bw()
```



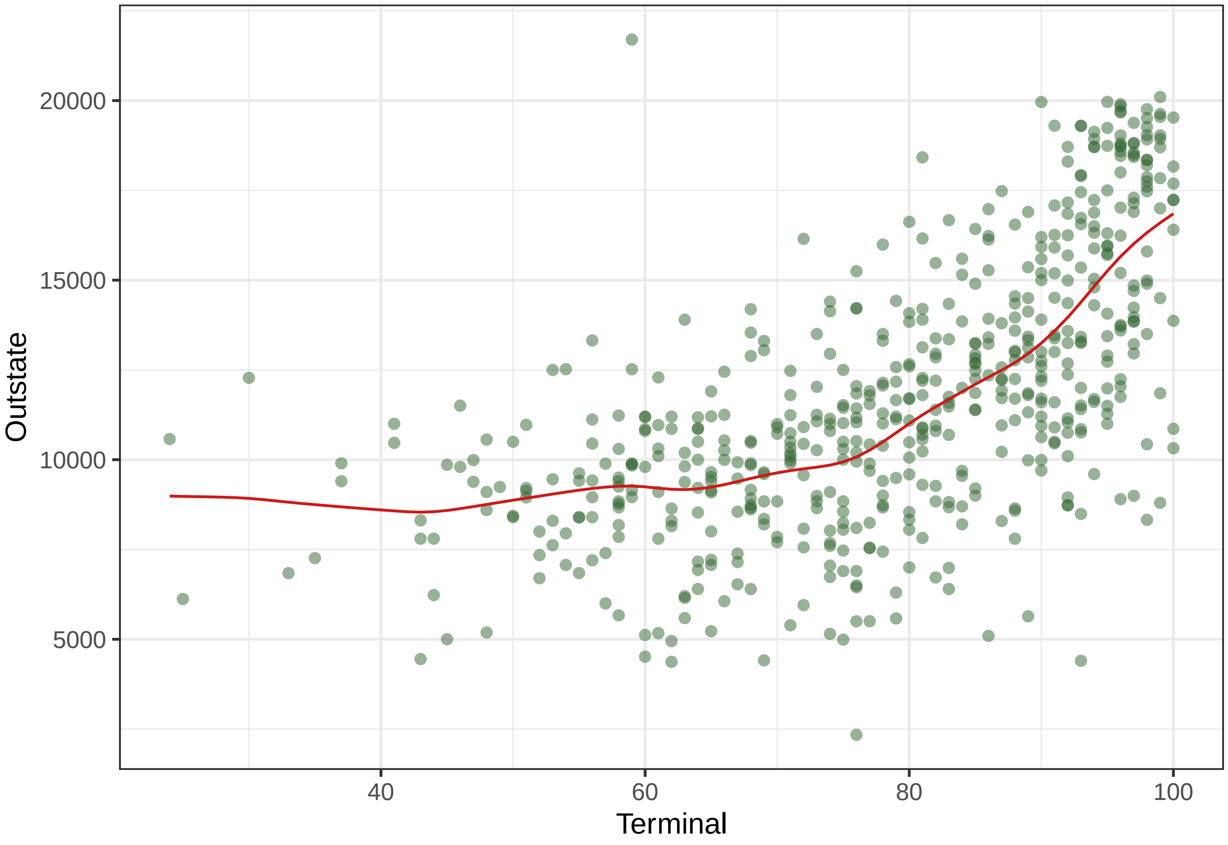
```
#Using lambda = 0.001
fit.ss <- smooth.spline(college1$Terminal, college1$Outstate, lambda=0.001)
fit.ss$df
```

```
## [1] 9.838879
```

```
Terminallims <- range(college1$Terminal)
Terminal.grid <- seq(from = Terminallims[1],to = Terminallims[2])

pred.ss <- predict(fit.ss, x = Terminal.grid)
pred.ss.df <- data.frame(pred = pred.ss$y, Terminal = Terminal.grid)

p <- ggplot(data= college1, aes(x = Terminal, y = Outstate)) +geom_point(color = rgb(0.2, 0.4, 0.2, 0.5
p + geom_line(aes(x = Terminal, y =pred), data = pred.ss.df, color = rgb(0.8, 0.1, 0.1, 1)) +theme_bw()
```


20000
15000
10000
5000
Outstate
40
60
80
100
Terminal