HR Data

Reading in Data (Use Step Function to Find Significant Variables)

```
hr = read.csv("HR_comma_sep.csv")
library(ggplot2)
```

Binning variables and Creating Tables

For example: Satisfaction vs Number of Projects

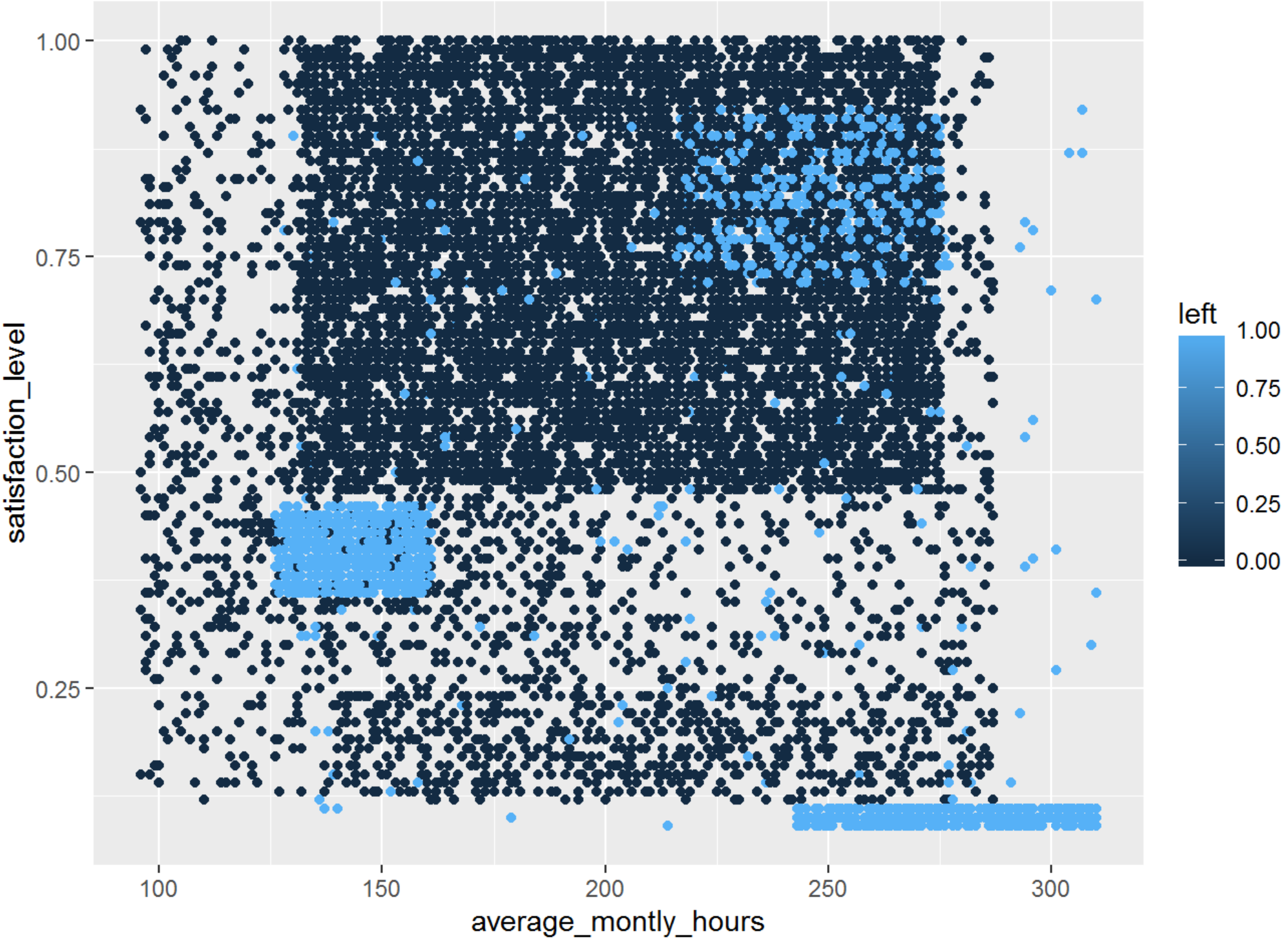
```
satisfaction = cut(hr$satisfaction_level, c(0, .1, .2, .3, .4, .5, .6, .7, .8, .9, 1)
)
table = table(satisfaction, hr$number_project)
table
```

```
## 
## satisfaction   2   3   4   5   6   7
##    (0,0.1]     3   3   0  24 368 155
##    (0.1,0.2]  24  95 158 184 378  86
##    (0.2,0.3]  36  90 114 125  91   7
##    (0.3,0.4] 840 131  93  61  55   3
##    (0.4,0.5] 913 315 270 132  58   0
##    (0.5,0.6] 160 719 680 315  39   4
##    (0.6,0.7] 124 716 720 357  54   1
##    (0.7,0.8] 113 733 837 565  48   0
##    (0.8,0.9]  91 628 849 613  37   0
##    (0.9,1]    84 625 644 385  46   0
```

Plot of Hours vs Satisfaction Accounting for Left

We see four distinct clusters

```
p = ggplot(data = hr, aes(x = average_montly_hours, y = satisfaction_level, col = lef
t))
p + geom_point()
```



Seperating Clusters By Eye

So we can analyze them individually

```
bigsquare = subset(hr, satisfaction_level >.5)
bigsquare = subset(hr, average_montly_hours > 130)
bigsquare = subset(hr, average_montly_hours < 280)
smallsquare = subset(hr, satisfaction_level < .5)
smallsquare = subset(hr, average_montly_hours > 125)
smallsquare = subset(hr, average_montly_hours <170)
bottomline = subset(hr, satisfaction_level <.1)
bottomline = subset(hr, average_montly_hours > 240)
bigleft = subset(bigsquare, left == 1)
```

Cluster Summaries (Boring Part)

We see that the group in the bottom right are being worked very hard and all eventually quit. The left of center cluster is newer employees who do not get many projects and are yet unhappy which is opposite of the usual correlation. The group of folks who left in the upper right have been at the company a while and are satisfied but leave anyways. These people have not gotten promotons in the last five years and could be leaving for more money. They are happy despite leaving which is the inverse of the overall trend. This is why it is so important to separate these groups. Any stories we make up to why people leave is just speculation, but the data does not lie. To predict who leaves or employ techniques to retain employees, one must analyze each subset.

```
summary(hr)
```

```
##  satisfaction_level last_evaluation  number_project  average_montly_hours
##  Min.   :0.0900     Min.   :0.3600   Min.   :2.000   Min.   : 96.0
##  1st Qu.:0.4400     1st Qu.:0.5600   1st Qu.:3.000   1st Qu.:156.0
##  Median :0.6400     Median :0.7200   Median :4.000   Median :200.0
##  Mean   :0.6128     Mean   :0.7161   Mean   :3.803   Mean   :201.1
##  3rd Qu.:0.8200     3rd Qu.:0.8700   3rd Qu.:5.000   3rd Qu.:245.0
##  Max.   :1.0000     Max.   :1.0000   Max.   :7.000   Max.   :310.0
##
##  time_spend_company Work_accident         left
##  Min.   : 2.000     Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 3.000     1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 3.000     Median :0.0000   Median :0.0000
##  Mean   : 3.498     Mean   :0.1446   Mean   :0.2381
##  3rd Qu.: 4.000     3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :10.000     Max.   :1.0000   Max.   :1.0000
##
##  promotion_last_5years         sales         salary
##  Min.   :0.00000       sales      :4140   high  :1237
##  1st Qu.:0.00000       technical  :2720   low   :7316
##  Median :0.00000       support    :2229   medium:6446
##  Mean   :0.02127       IT         :1227
##  3rd Qu.:0.00000       product_mng: 902
##  Max.   :1.00000       marketing  : 858
##                        (Other)    :2923
```

```
summary(smallsquare)
```

```
##  satisfaction_level last_evaluation number_project  average_montly_hours
##  Min.   :0.1100     Min.   :0.360   Min.   :2.000   Min.   : 96.0
##  1st Qu.:0.4000     1st Qu.:0.510   1st Qu.:2.000   1st Qu.:135.0
##  Median :0.5300     Median :0.580   Median :3.000   Median :146.0
##  Mean   :0.5767     Mean   :0.642   Mean   :3.207   Mean   :144.6
##  3rd Qu.:0.7500     3rd Qu.:0.760   3rd Qu.:4.000   3rd Qu.:156.0
##  Max.   :1.0000     Max.   :1.000   Max.   :7.000   Max.   :169.0
##
##  time_spend_company Work_accident         left
##  Min.   : 2.000     Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 3.000     1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 3.000     Median :0.0000   Median :0.0000
##  Mean   : 3.318     Mean   :0.1366   Mean   :0.3157
##  3rd Qu.: 3.000     3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :10.000     Max.   :1.0000   Max.   :1.0000
##
##  promotion_last_5years         sales         salary
##  Min.   :0.0000        sales      :1429   high  : 392
##  1st Qu.:0.0000        technical  : 894   low   :2564
##  Median :0.0000        support    : 761   medium:2141
##  Mean   :0.0208        IT         : 405
##  3rd Qu.:0.0000        product_mng: 313
##  Max.   :1.0000        marketing  : 298
##                        (Other)    : 997
```

```
summary(bigsquare)
```

```
##  satisfaction_level last_evaluation  number_project  average_montly_hours
##  Min.   :0.0900     Min.   :0.3600   Min.   :2.000   Min.   : 96.0
##  1st Qu.:0.4600     1st Qu.:0.5600   1st Qu.:3.000   1st Qu.:155.0
##  Median :0.6600     Median :0.7100   Median :4.000   Median :196.0
##  Mean   :0.6293     Mean   :0.7119   Mean   :3.727   Mean   :197.3
##  3rd Qu.:0.8200     3rd Qu.:0.8600   3rd Qu.:5.000   3rd Qu.:241.0
##  Max.   :1.0000     Max.   :1.0000   Max.   :7.000   Max.   :279.0
##
##  time_spend_company Work_accident         left
##  Min.   : 2.000     Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 3.000     1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 3.000     Median :0.0000   Median :0.0000
##  Mean   : 3.471     Mean   :0.1478   Mean   :0.2158
##  3rd Qu.: 4.000     3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :10.000     Max.   :1.0000   Max.   :1.0000
##
##  promotion_last_5years         sales         salary
##  Min.   :0.00000       sales      :4000   high  :1208
##  1st Qu.:0.00000       technical  :2588   low   :6971
##  Median :0.00000       support    :2152   medium:6229
##  Mean   :0.02172       IT         :1165
##  3rd Qu.:0.00000       product_mng: 863
##  Max.   :1.00000       marketing  : 827
##                        (Other)    :2813
```

```
summary(bottomline)
```

```
##  satisfaction_level last_evaluation  number_project  average_montly_hours
##  Min.   :0.0900     Min.   :0.3600   Min.   :2.000   Min.   :241.0
##  1st Qu.:0.2100     1st Qu.:0.6600   1st Qu.:3.000   1st Qu.:251.0
##  Median :0.6500     Median :0.8200   Median :4.000   Median :261.0
##  Mean   :0.5684     Mean   :0.7819   Mean   :4.452   Mean   :263.2
##  3rd Qu.:0.8300     3rd Qu.:0.9100   3rd Qu.:5.000   3rd Qu.:272.0
##  Max.   :1.0000     Max.   :1.0000   Max.   :7.000   Max.   :310.0
##
##  time_spend_company Work_accident         left
##  Min.   : 2.000     Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 3.000     1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 4.000     Median :0.0000   Median :0.0000
##  Mean   : 3.792     Mean   :0.1274   Mean   :0.3615
##  3rd Qu.: 5.000     3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :10.000     Max.   :1.0000   Max.   :1.0000
##
##  promotion_last_5years         sales         salary
##  Min.   :0.00000       sales      :1156   high  : 313
##  1st Qu.:0.00000       technical  : 779   low   :2093
##  Median :0.00000       support    : 644   medium:1802
##  Mean   :0.01949       IT         : 354
##  3rd Qu.:0.00000       product_mng: 243
##  Max.   :1.00000       RandD      : 231
##                        (Other)    : 801
```

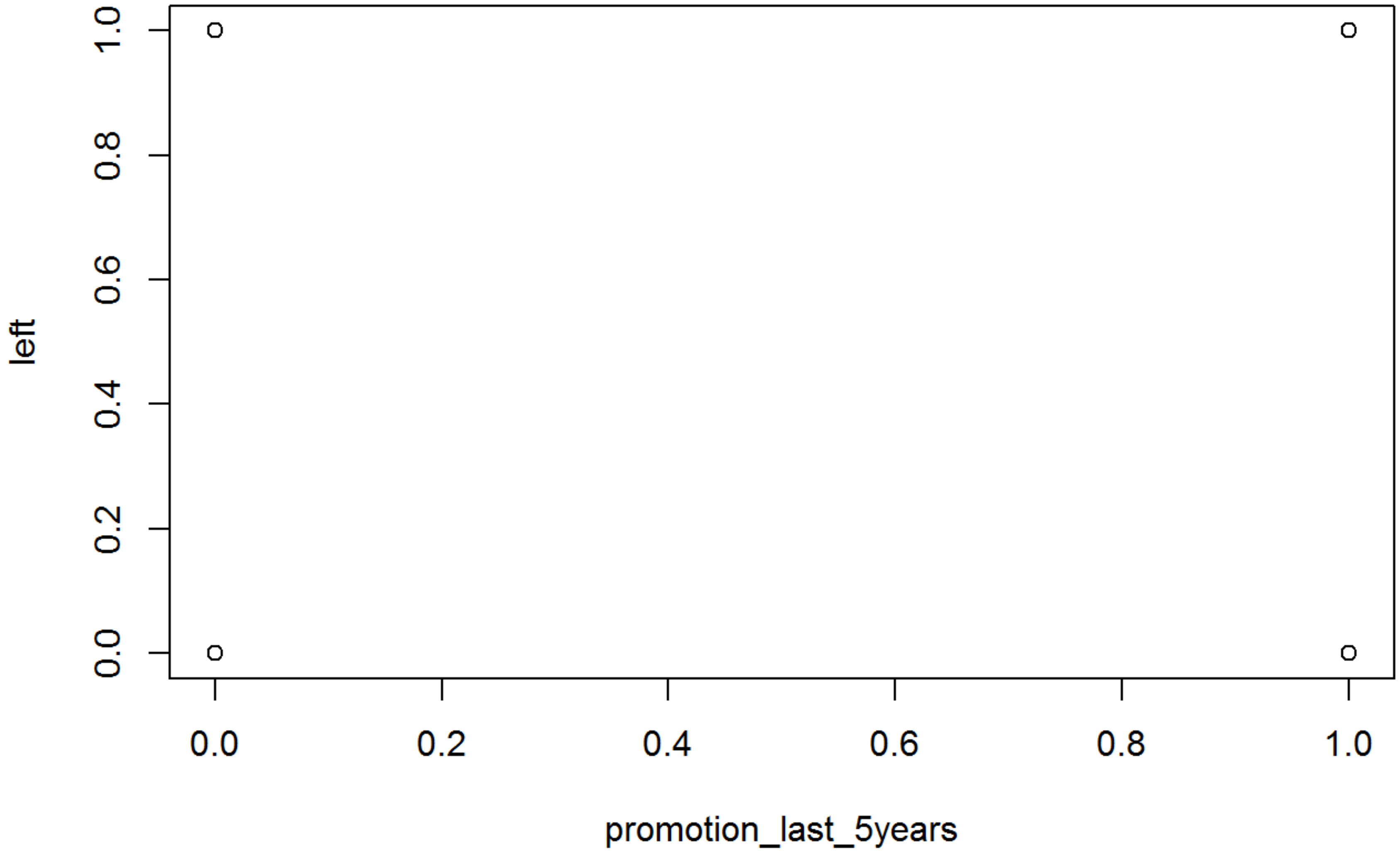
```
summary(bigleft)
```

```
##  satisfaction_level last_evaluation  number_project  average_montly_hours
##  Min.   :0.0900     Min.   :0.4500   Min.   :2.000   Min.   :126.0
##  1st Qu.:0.3700     1st Qu.:0.5100   1st Qu.:2.000   1st Qu.:144.0
##  Median :0.4300     Median :0.5700   Median :2.000   Median :160.0
##  Mean   :0.4857     Mean   :0.6977   Mean   :3.523   Mean   :194.4
##  3rd Qu.:0.7500     3rd Qu.:0.9000   3rd Qu.:5.000   3rd Qu.:251.0
##  Max.   :0.9200     Max.   :1.0000   Max.   :7.000   Max.   :279.0
##
##  time_spend_company Work_accident          left   promotion_last_5years
##  Min.   :2.000      Min.   :0.00000   Min.   :1   Min.   :0.000000
##  1st Qu.:3.000      1st Qu.:0.00000   1st Qu.:1   1st Qu.:0.000000
##  Median :3.000      Median :0.00000   Median :1   Median :0.000000
##  Mean   :3.841      Mean   :0.04825   Mean   :1   Mean   :0.005146
##  3rd Qu.:5.000      3rd Qu.:0.00000   3rd Qu.:1   3rd Qu.:0.000000
##  Max.   :6.000      Max.   :1.00000   Max.   :1   Max.   :1.000000
##
##        sales        salary
##  sales    :909   high  :  72
##  technical:583   low   :1889
##  support  :490   medium:1148
##  IT       :224
##  hr       :183
##  marketing:179
##  (Other)  :541
```

Not All of Your Plots Will Be Good

I made this horrible plot at one point and felt really stupid. That is a part ofS exploratory data analysis

```
plot(left ~ promotion_last_5years, data = hr)
```


left
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
promotion_last_5years