Plotting_Model_Predictors.Rmd

Trent Fowler

February 6, 2019

Below are a few examples of using plotting and tables to find the stronger predictors in a data set for use in machine learning prediction. This is meant as a resource guide and not as definitive analysis of the R data set Wage.

```
library(ISLR)
library(ggplot2)
library(caret)
```

```
## Loading required package: lattice
```

```
data(Wage)
summary(Wage)
```

```
##       year           age                      maritl           race
##  Min.   :2003   Min.   :18.00   1. Never Married: 648   1. White:2480
##  1st Qu.:2004   1st Qu.:33.75   2. Married      :2074   2. Black: 293
##  Median :2006   Median :42.00   3. Widowed      :  19   3. Asian: 190
##  Mean   :2006   Mean   :42.41   4. Divorced     : 204   4. Other:  37
##  3rd Qu.:2008   3rd Qu.:51.00   5. Separated    :  55
##  Max.   :2009   Max.   :80.00
##
##                education                     region
##  1. < HS Grad      :268   2. Middle Atlantic   :3000
##  2. HS Grad        :971   1. New England       :   0
##  3. Some College   :650   3. East North Central:   0
##  4. College Grad   :685   4. West North Central:   0
##  5. Advanced Degree:426   5. South Atlantic    :   0
##                           6. East South Central:   0
##                           (Other)              :   0
##            jobclass               health      health_ins      logwage
##  1. Industrial :1544   1. <=Good     : 858   1. Yes:2083   Min.   :3.000
##  2. Information:1456   2. >=Very Good:2142   2. No : 917   1st Qu.:4.447
##                                                            Median :4.653
##                                                            Mean   :4.654
##                                                            3rd Qu.:4.857
##                                                            Max.   :5.763
##
##       wage
##  Min.   : 20.09
##  1st Qu.: 85.38
##  Median :104.92
##  Mean   :111.70
##  3rd Qu.:128.68
##  Max.   :318.34
##
```

```
names(Wage)
```

```
##  [1] "year"       "age"        "maritl"     "race"       "education"
```

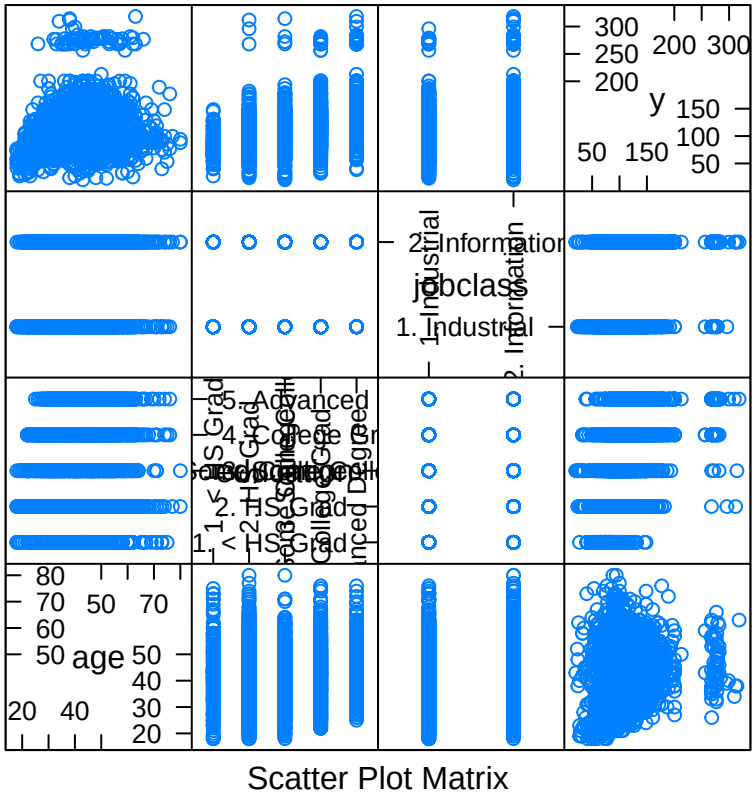
```
##  [6] "region"     "jobclass"   "health"     "health_ins" "logwage"
## [11] "wage"
```

```
# Get training and test sets
inTrain <- createDataPartition(y = Wage$wage, p = 0.7, list = FALSE)
training <- Wage[inTrain,]
testing <- Wage[-inTrain,]
dim(training); dim(testing)
```

```
## [1] 2102   11
```

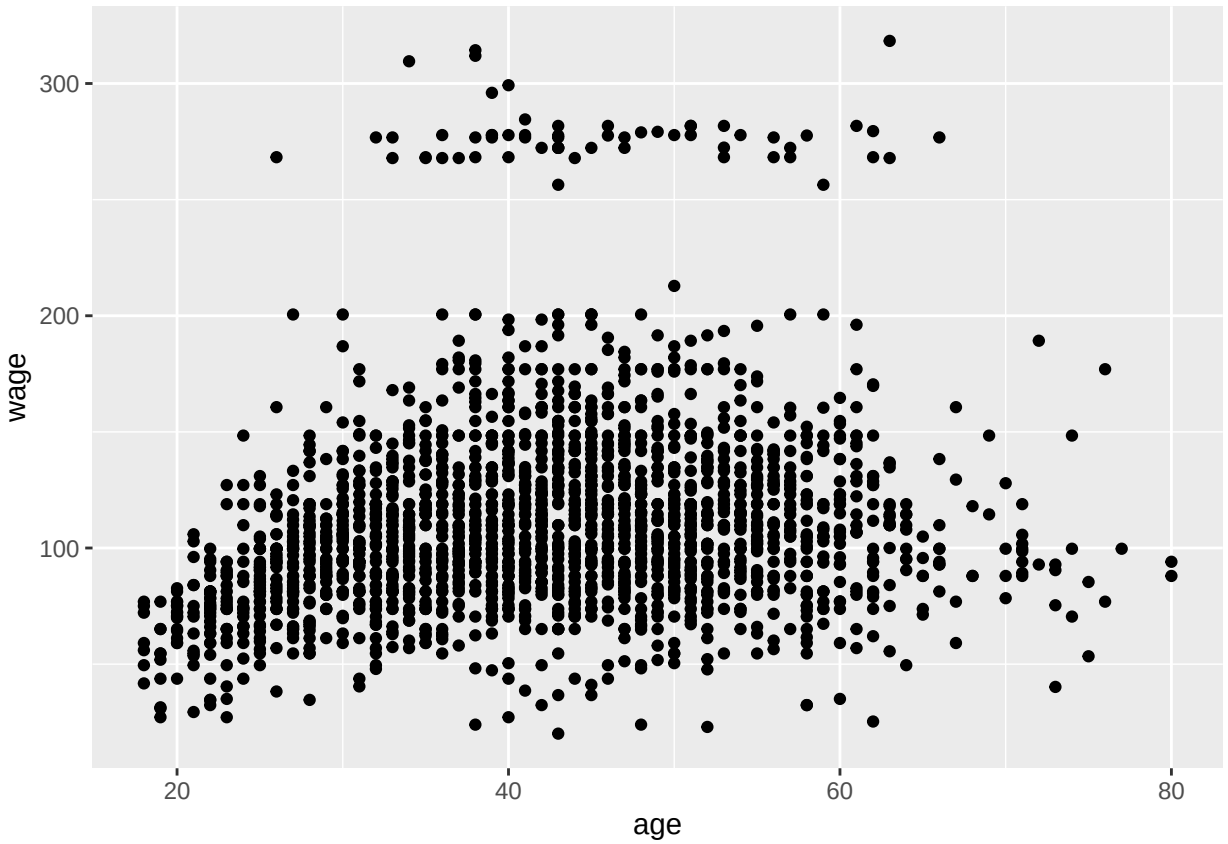
```
## [1] 898  11
```

```
# plot out training set
featurePlot(x = training[, c("age", "education", "jobclass")],
            y = training$wage, plot = "pairs")
```



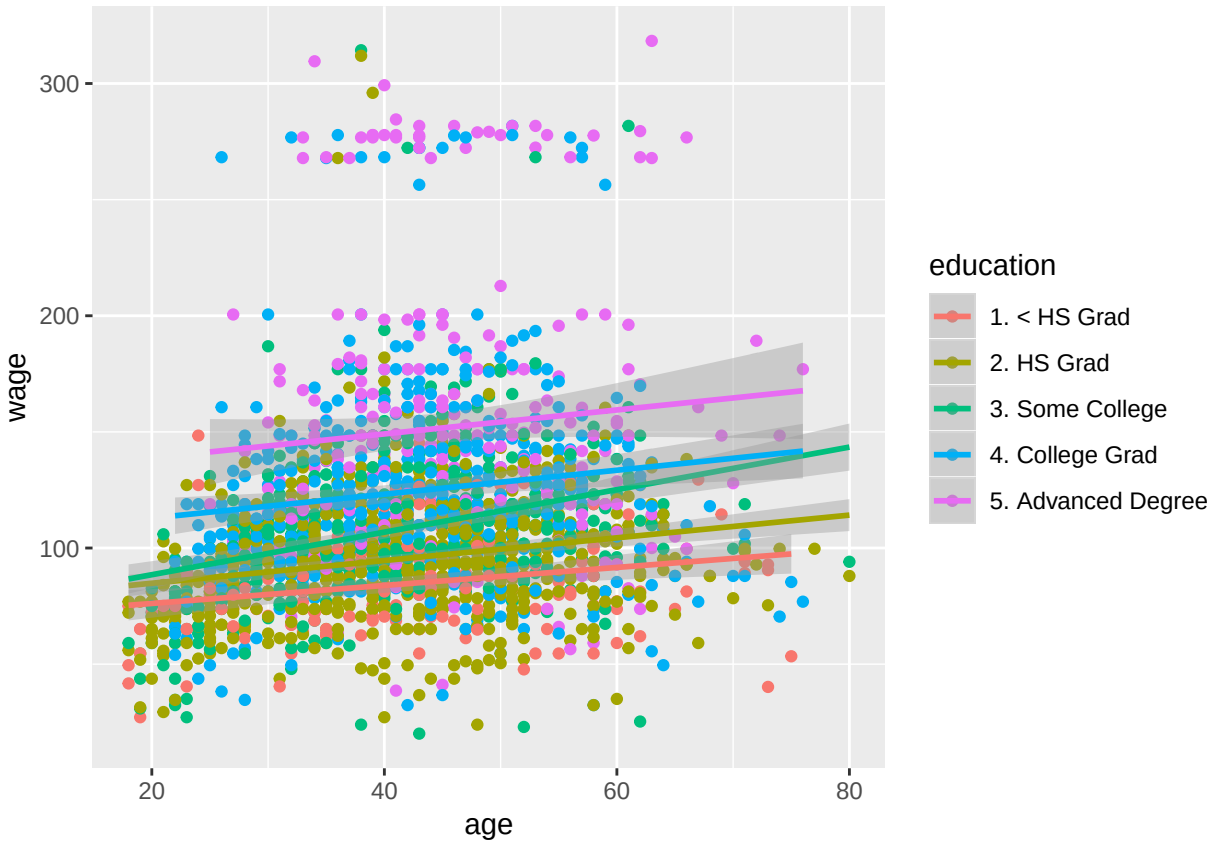
Scatter Plot Matrix

```
#these pair plots are a bit confusing though
# using ggplot2
qplot(age, wage, data = training)
```




```
# plotting with color might help resolution of data
qplot(age, wage, colour = jobclass, data = training)
```


```
# adding regression smoothers might help also
qq <- qplot(age, wage, colour = education, data = training)
qq + geom_smooth(method = "lm", formula = y ~ x)
```




```
# cutting data in to different categories when it is clear there is a relationship
# making factors
library(Hmisc)
```

```
## Loading required package: survival
```

```
##
## Attaching package: 'survival'
```

```
## The following object is masked from 'package:caret':
##
##     cluster
```

```
## Loading required package: Formula
```

```
##
## Attaching package: 'Hmisc'
```

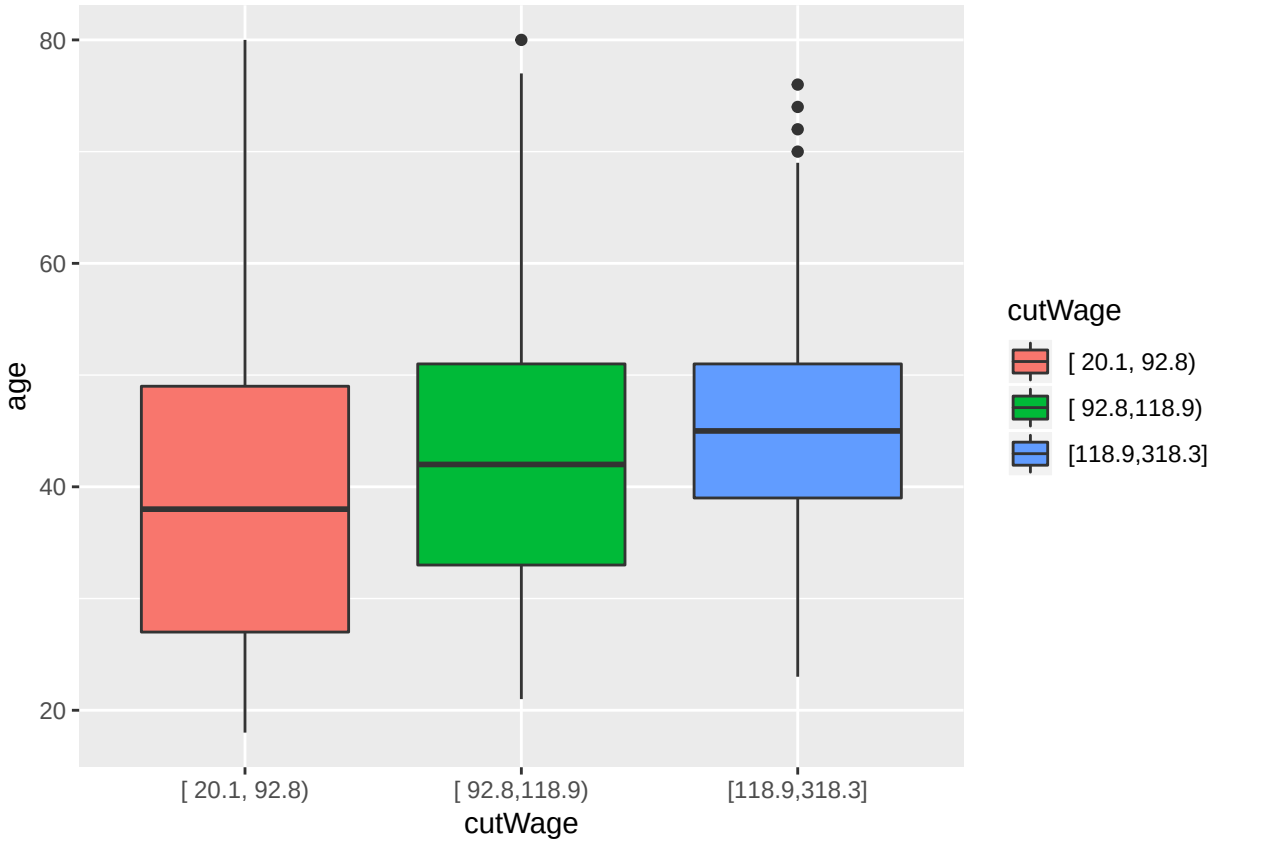
```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
cutWage <- cut2(training$wage, g = 3)
table(cutWage)
```

```
## cutWage
## [ 20.1, 92.8) [ 92.8,118.9) [118.9,318.3]
##           701           731           670
```

```
# use these factor groups to get another view of the data
p1 <- qplot(cutWage, age, data = training, fill = cutWage, geom = c("boxplot"))
```

```
p1
```



There is some trend in wages with age, as expected, but in this particular data set age is not a clear cut predictor.

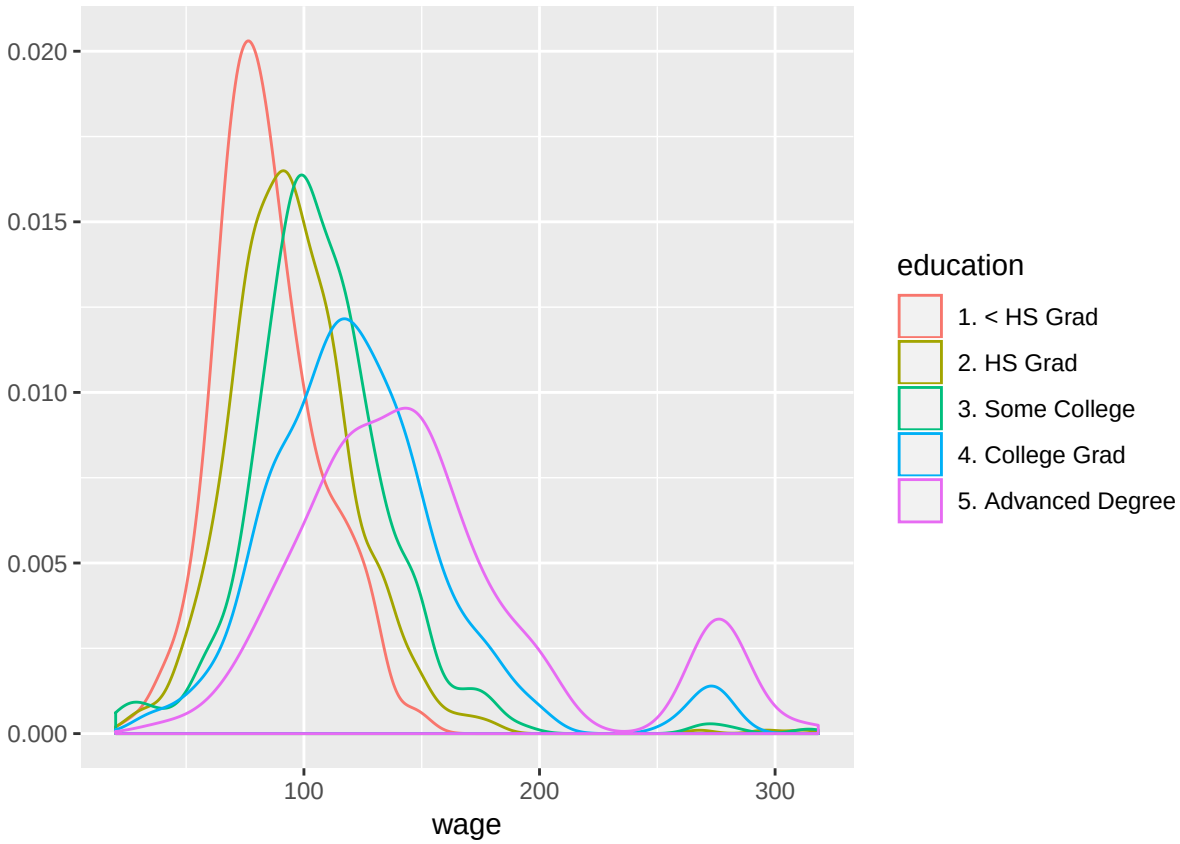
```
#viewing these cut data in tables may also be useful
t1 <- table(cutWage, training$jobclass)
t1
```

```
##
## cutWage         1. Industrial 2. Information
##   [ 20.1, 92.8)           445            256
##   [ 92.8,118.9)           371            360
##   [118.9,318.3]           272            398
```

```
# prop.table() will give the proportions
prop.table(t1, 1)
```

```
##
## cutWage         1. Industrial 2. Information
##   [ 20.1, 92.8)     0.6348074      0.3651926
##   [ 92.8,118.9)     0.5075239      0.4924761
##   [118.9,318.3]     0.4059701      0.5940299
```

```
# lastly density plots with a few factors can illuminate
qplot(wage, colour = education, data = training, geom = "density")
```


0.020
0.015
0.010
0.005
0.000
100
200
300
wage
education
1. < HS Grad
2. HS Grad
3. Some College
4. College Grad
5. Advanced Degree