Plotting Predictors Demo - Wage Example

Jiachang (Ernest) Xu

6/23/2017

Example: Wage Data

```
## require ISLR package for machine learning data
require(ISLR)
```

```
## Loading required package: ISLR
```

```
## require caret package for machine learning algorithms
require(caret)
```

```
## Loading required package: caret
```

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

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

```
## require ggplot2 package for plotting
require(ggplot2)
## data loading
data(Wage)
summary(Wage)
```

```
##       year           age               sex                    maritl
##  Min.   :2003   Min.   :18.00   1. Male  :3000   1. Never Married: 648
##  1st Qu.:2004   1st Qu.:33.75   2. Female:   0   2. Married      :2074
##  Median :2006   Median :42.00                    3. Widowed      :  19
##  Mean   :2006   Mean   :42.41                    4. Divorced     : 204
##  3rd Qu.:2008   3rd Qu.:51.00                    5. Separated    :  55
##  Max.   :2009   Max.   :80.00
##
##        race                     education                     region
##  1. White:2480   1. < HS Grad      :268   2. Middle Atlantic   :3000
##  2. Black: 293   2. HS Grad        :971   1. New England       :   0
##  3. Asian: 190   3. Some College   :650   3. East North Central:   0
##  4. Other:  37   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
##
```

Get Training/Testing Sets

```
inTrain <- createDataPartition(y = Wage$wage, p = 0.7, list = FALSE)
training <- Wage[inTrain, ]
testing <- Wage[-inTrain, ]
dim(training)
```

```
## [1] 2102   12
```

```
dim(testing)
```

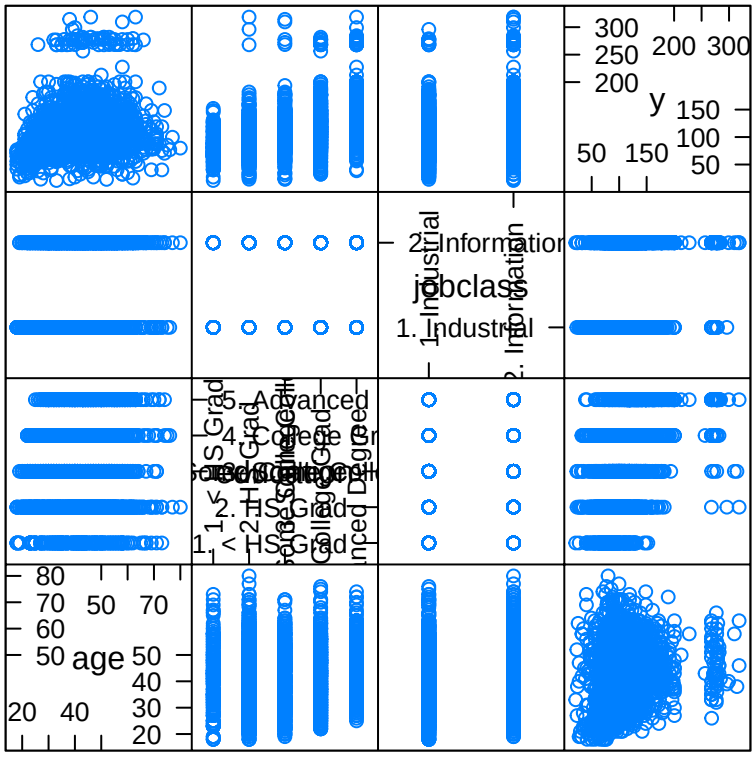
```
## [1] 898  12
```

featurePlot (caret Package)

```
args(featurePlot)
```

```
## function (x, y, plot = if (is.factor(y)) "strip" else "scatter",
##     labels = c("Feature", ""), ...)
## NULL
```

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



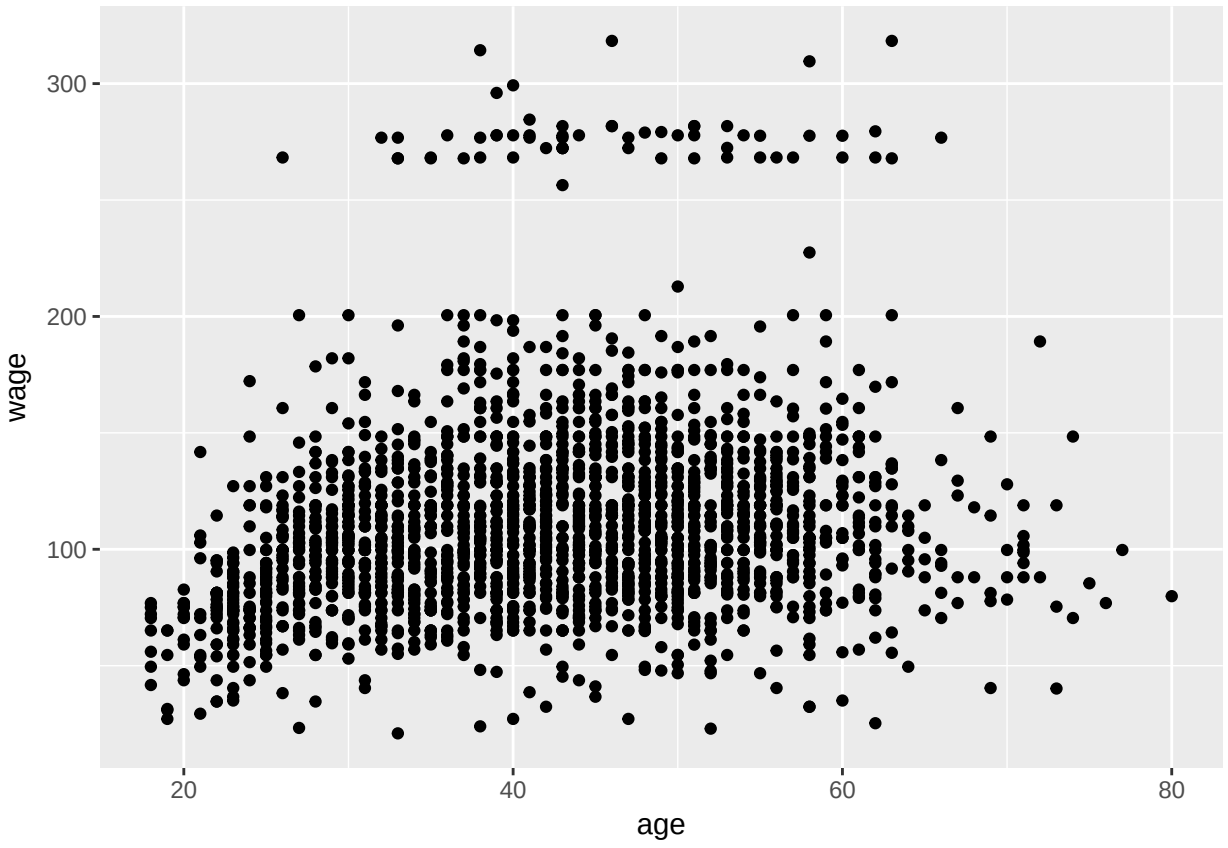
Scatter Plot Matrix

qplot (ggplot2 Package)

```
args(qplot)
```

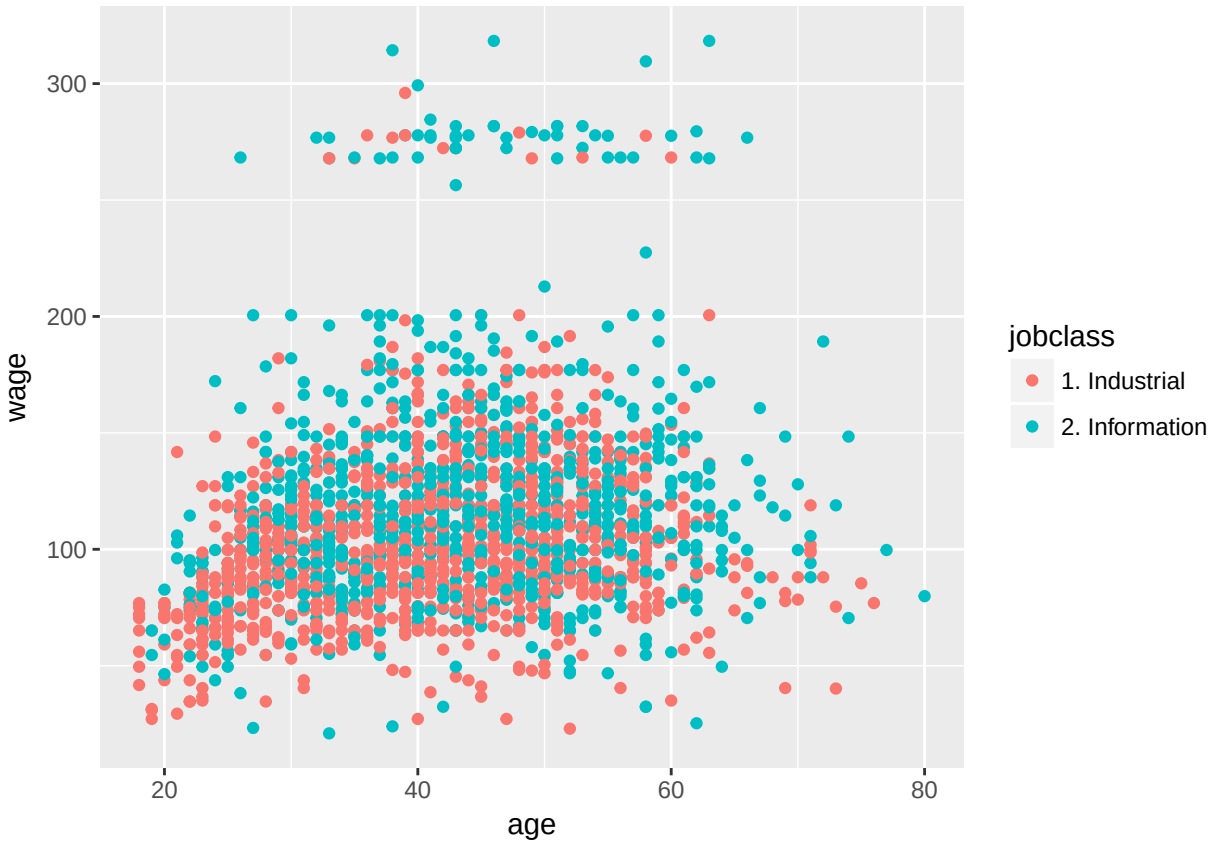
```
## function (x, y = NULL, ..., data, facets = NULL, margins = FALSE,
##     geom = "auto", xlim = c(NA, NA), ylim = c(NA, NA), log = "",
##     main = NULL, xlab = deparse(substitute(x)), ylab = deparse(substitute(y)),
##     asp = NA, stat = NULL, position = NULL)
## NULL
```

```
qplot(x = age, y = wage, data = training)
```



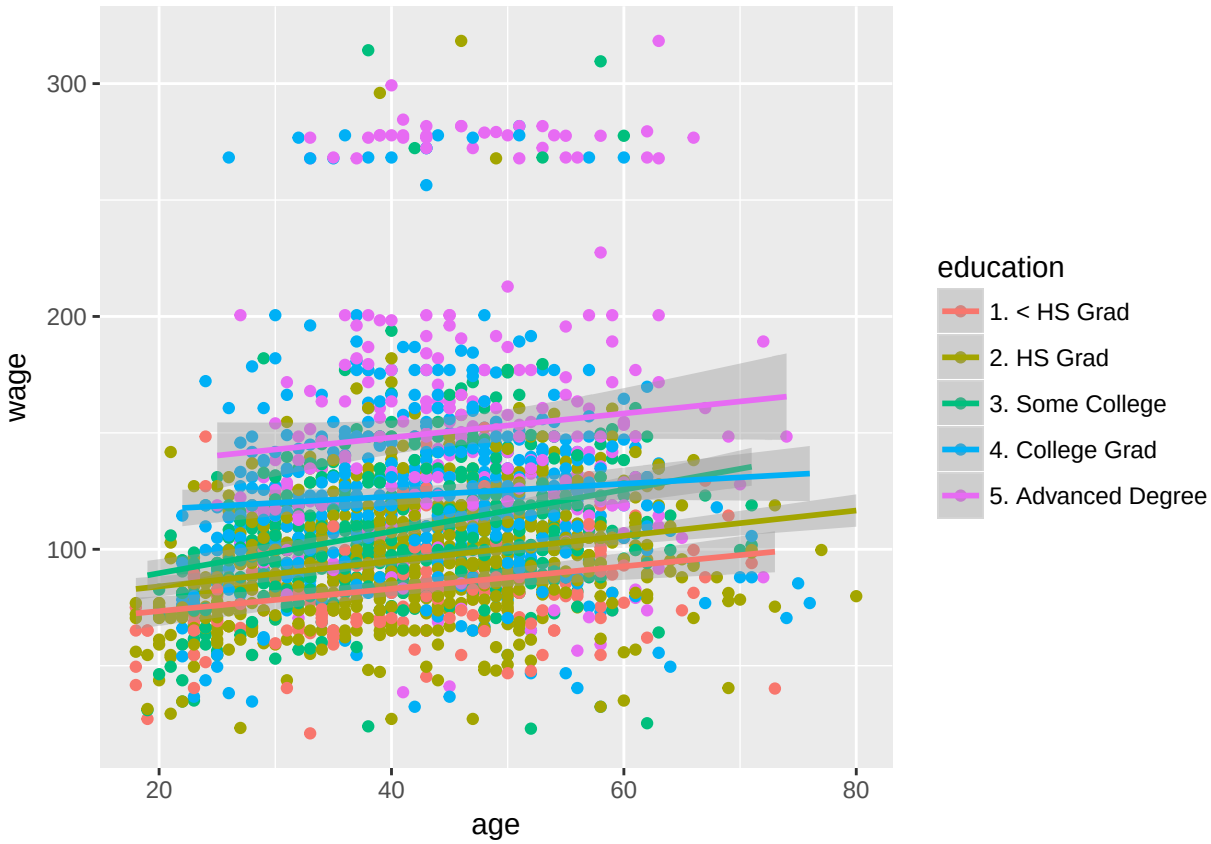
qplot with Colour (ggplot2 Package)

```
qplot(x = age, y = wage, colour = jobclass, data = training)
```



Add Regression Smoothers (ggplot2 Package)

```
qq <- qplot(x = age, y = wage, colour = education, data = training)
qq + geom_smooth(method = "lm", formula = y ~ x)
```



cut2 Making Factors (Hmisc Package)

```
require(Hmisc)
```

```
## Loading required package: 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, round.POSIXt, trunc.POSIXt, units
```

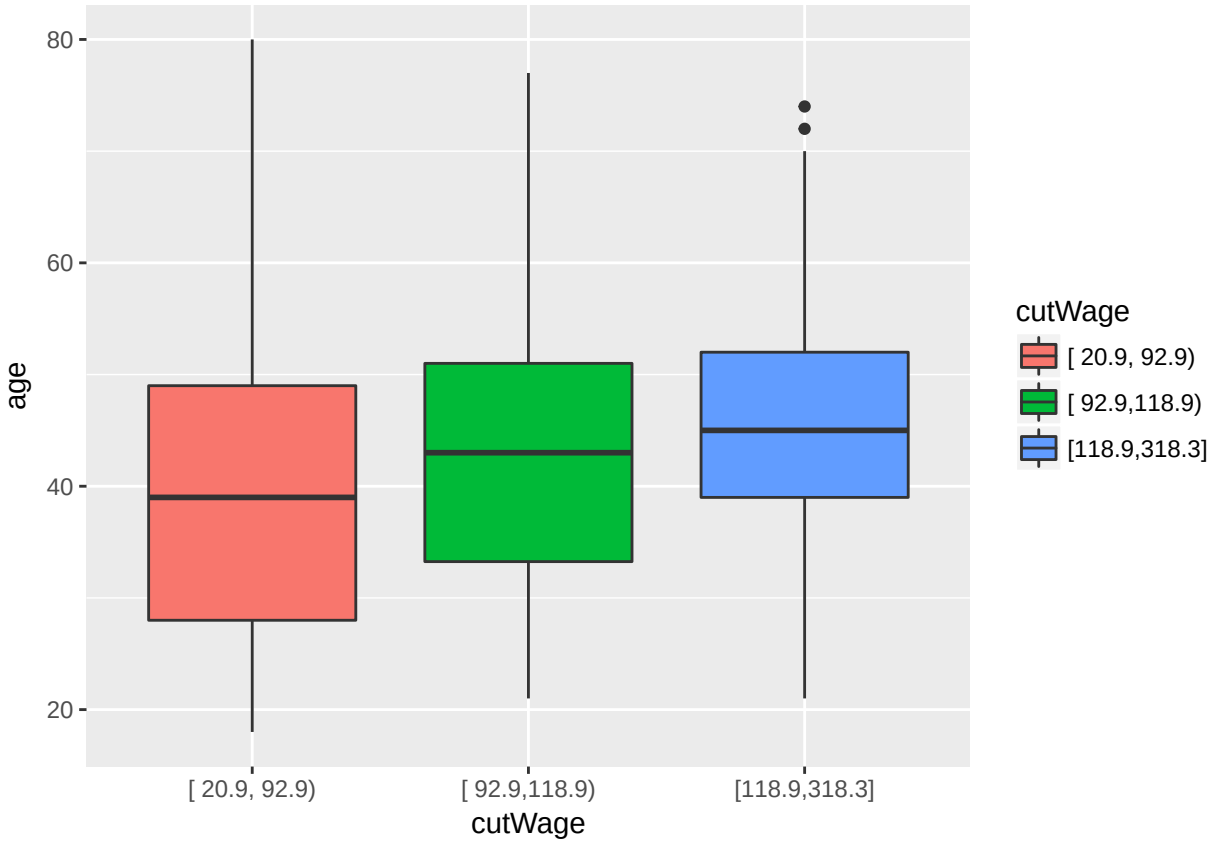
```
cutWage <- cut2(training$wage, g = 3)
table(cutWage)
```

```
## cutWage
## [ 20.9, 92.9) [ 92.9,118.9) [118.9,318.3]
```

```
##          702          734          666
```

Boxplot with cut2

```
boxplot.1 <- qplot(cutWage, age, data = training, fill = cutWage,
                   geom = c("boxplot"))
boxplot.1
```



Boxplot with Points Overlayed

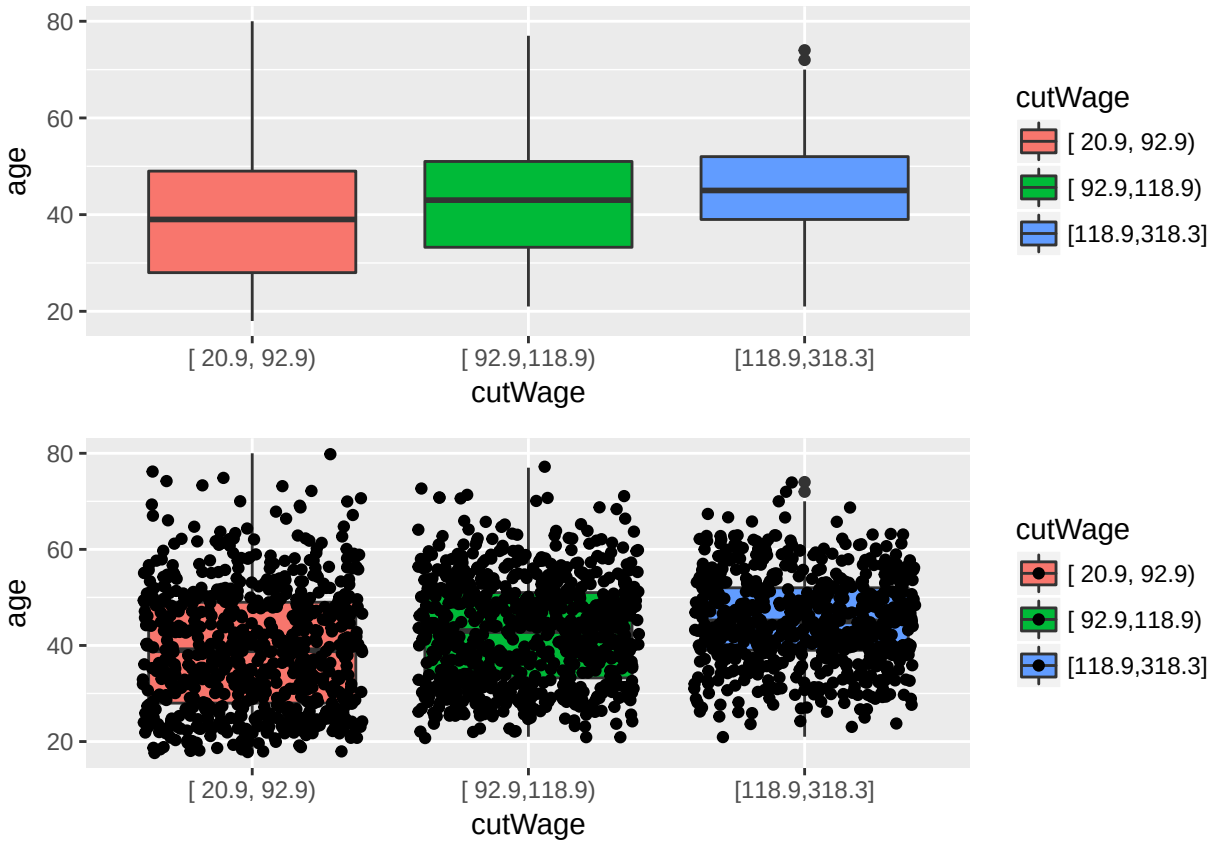
```
require(gridExtra)
```

```
## Loading required package: gridExtra
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:Hmisc':
##
##     combine
```

```
boxplot.2 <- qplot(cutWage, age, data = training, fill = cutWage,
                   geom = c("boxplot", "jitter"))
grid.arrange(boxplot.1, boxplot.2, nrow = 2)
```



Tables

```
table.1 <- table(cutWage, training$jobclass)
table.1
```

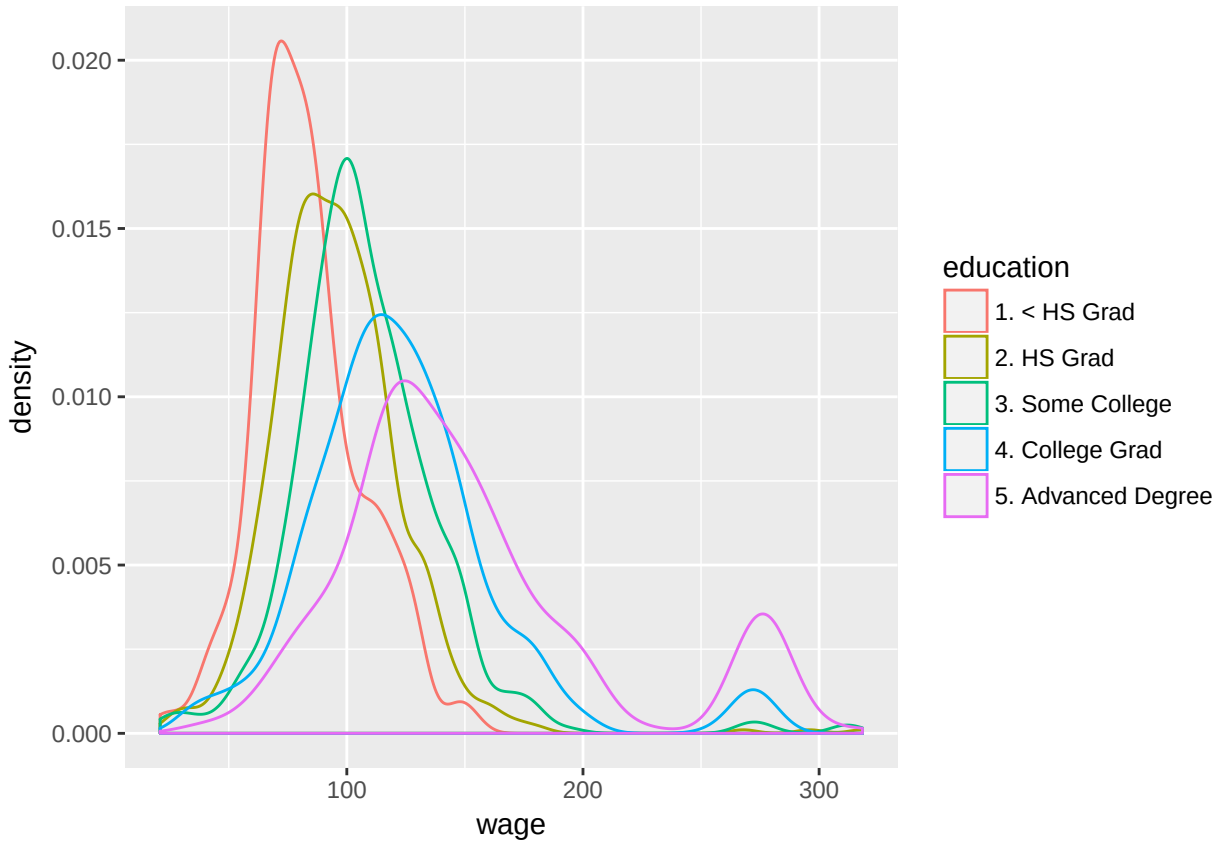
```
##
## cutWage         1. Industrial 2. Information
##   [ 20.9, 92.9)           448            254
##   [ 92.9,118.9)           378            356
##   [118.9,318.3]           270            396
```

```
prop.table(table.1, 1)
```

```
##
## cutWage         1. Industrial 2. Information
##   [ 20.9, 92.9)     0.6381766      0.3618234
##   [ 92.9,118.9)     0.5149864      0.4850136
##   [118.9,318.3]     0.4054054      0.5945946
```

Density Plot

```
qplot(wage, colour = education, data = training, geom = "density")
```


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