Plotting predictors

Jeffrey Leek

May 18, 2016

Example: predicting wages

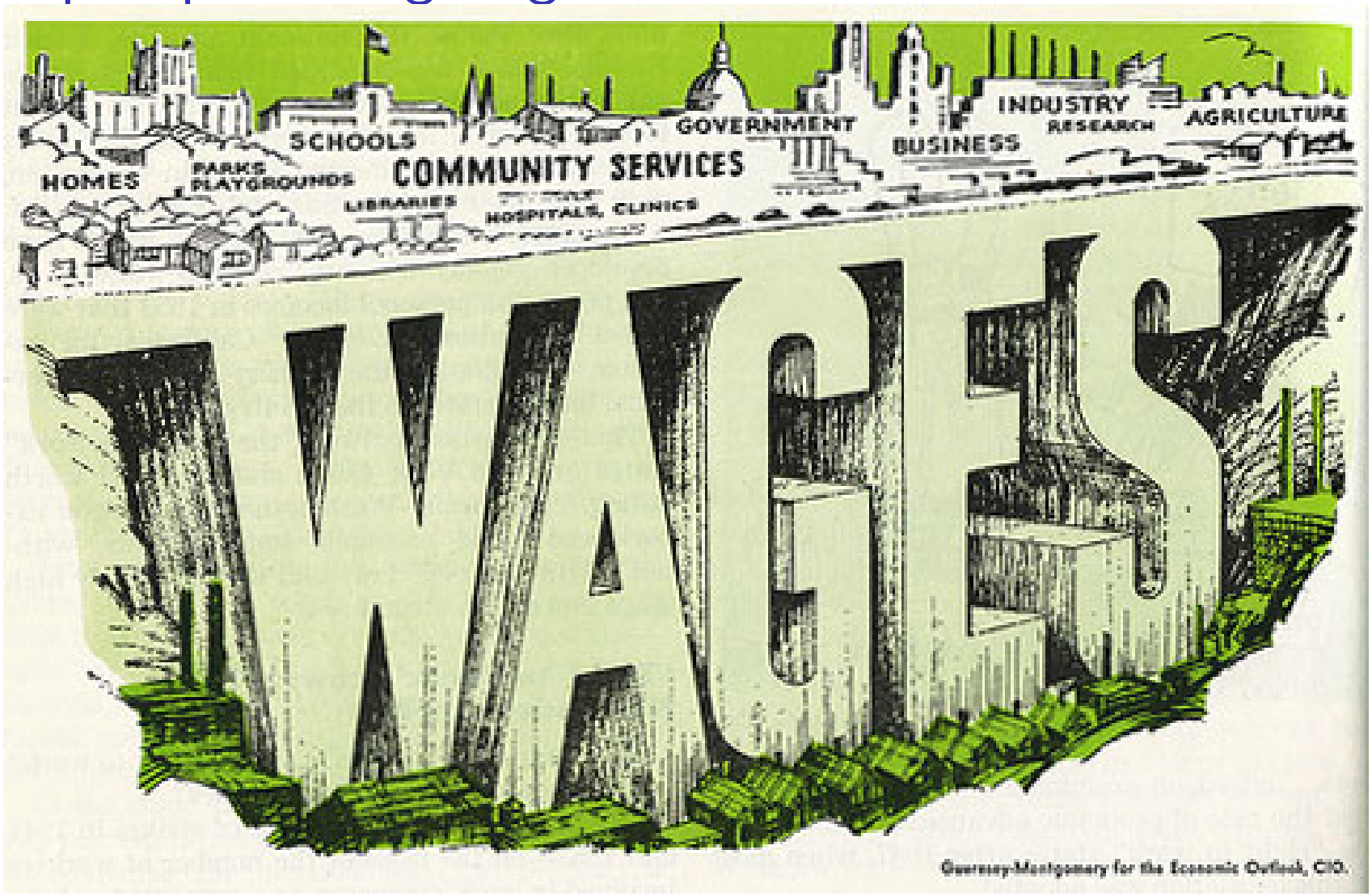


Image Credit

`http://www.cahs-media.org/the-high-cost-of-low-wages`

Data from: ISLR package from the book: Introduction to statistical

Example: Wage data

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

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

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

```
##       year           age                sex
##  Min.   :2003   Min.   :18.00   1. Male  :3000   1. Neve
##  1st Qu.:2004   1st Qu.:33.75   2. Female:   0   2. Marr
##  Median :2006   Median :42.00                    3. Wido
##  Mean   :2006   Mean   :42.41                    4. Divo
##  3rd Qu.:2008   3rd Qu.:51.00                    5. Sepa
##  Max.   :2009   Max.   :80.00
##
##        race                   education
##  1. White:2480   1. < HS Grad      :268   2. Middle Atla
##  2. Black: 293   2. HS Grad        :971   1. New Englan
```

Get training/test sets

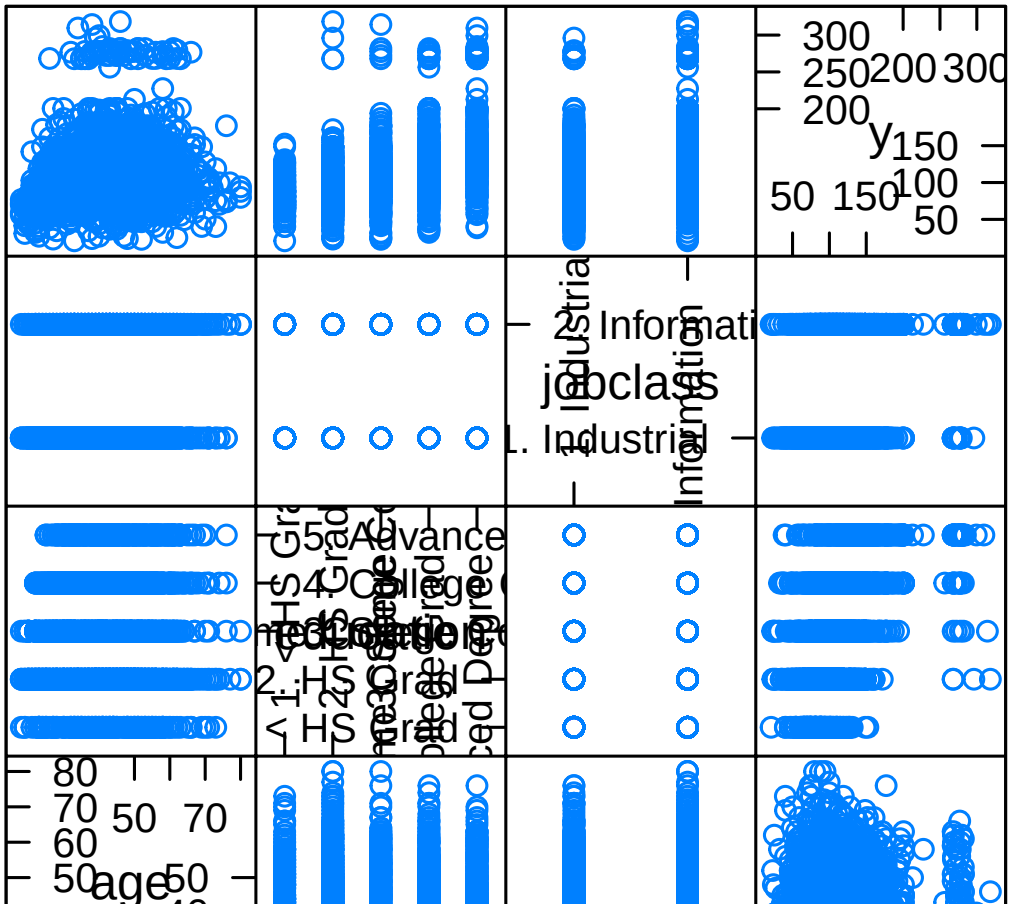
```
inTrain <- createDataPartition(y=Wage$wage,
                               p=0.7, list=FALSE)
training <- Wage[inTrain,]
testing <- Wage[-inTrain,]
dim(training); dim(testing)
```

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

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

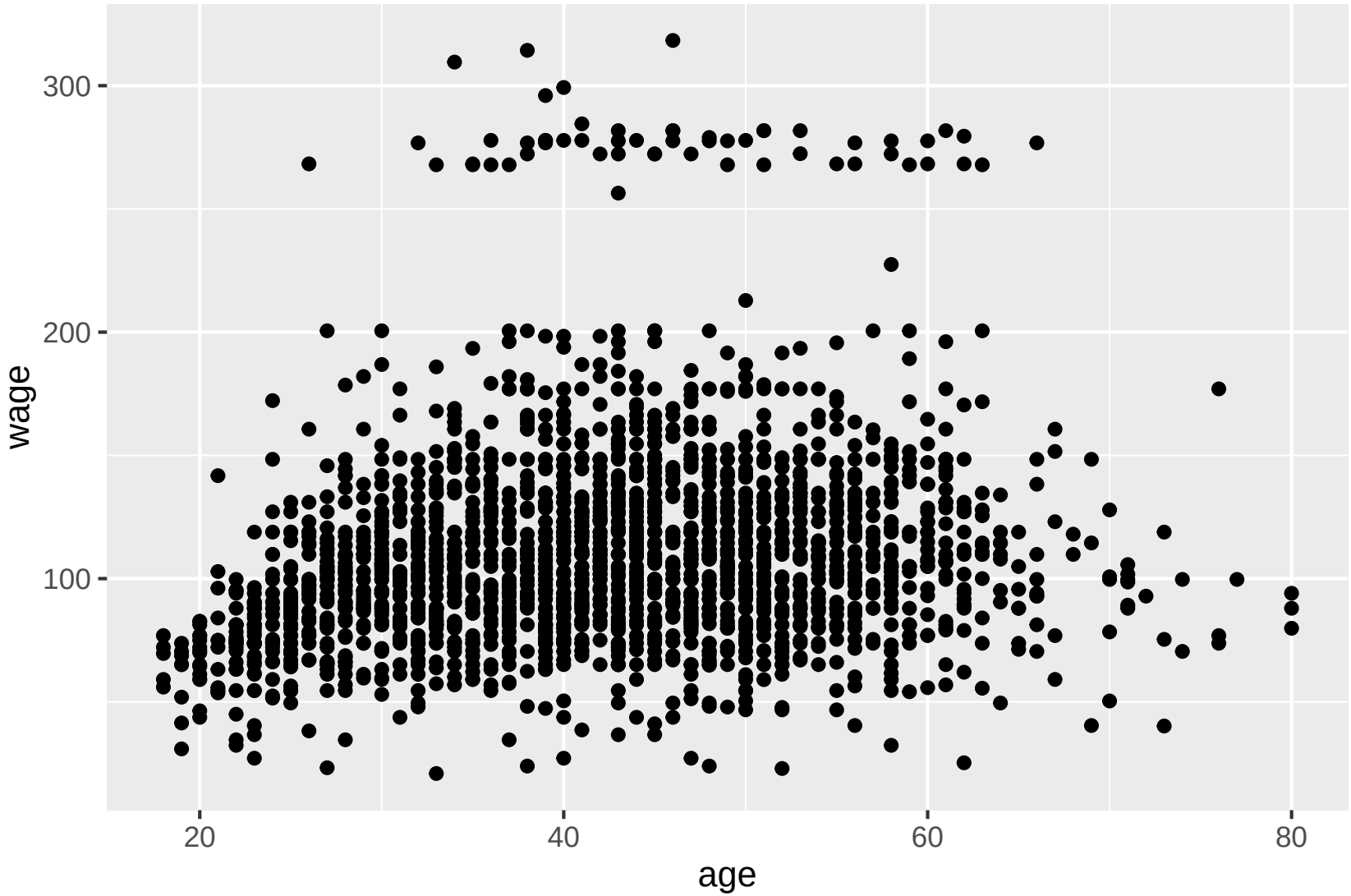
Feature plot (*caret* package)

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



Qplot (*ggplot2* package)

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

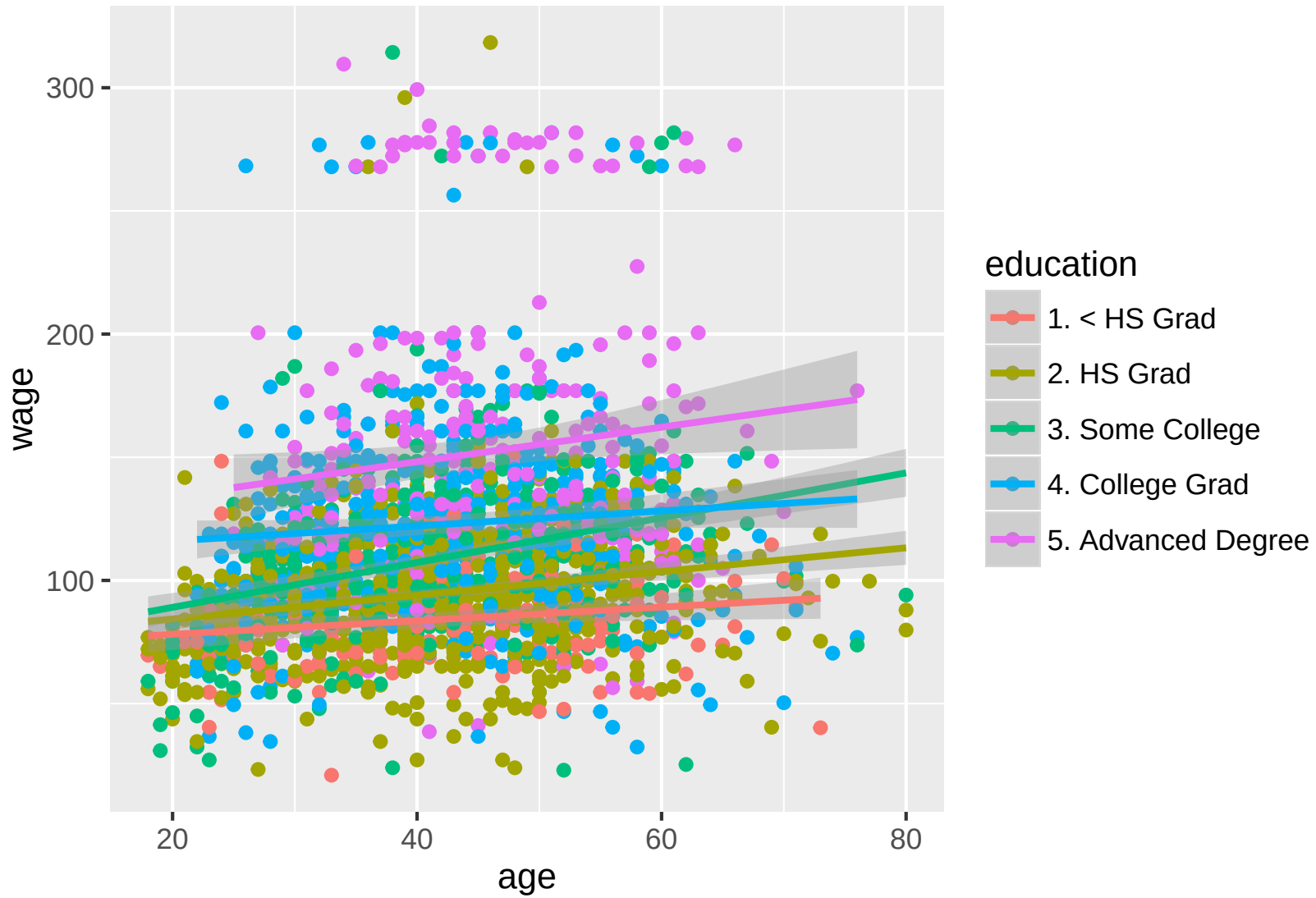


Qplot with color (*ggplot2* package)

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

Add regression smoothers (*ggplot2* package)

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



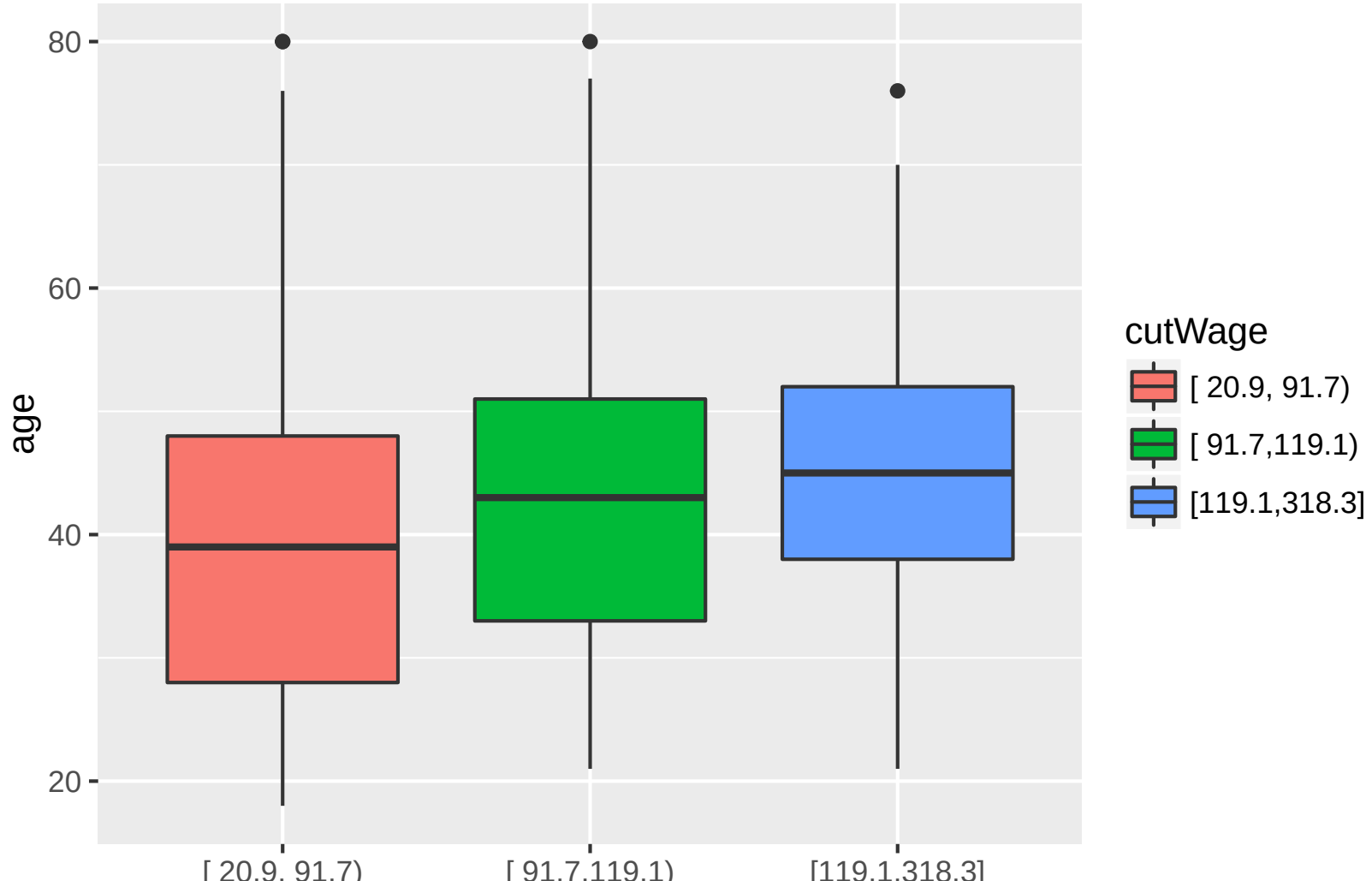
cut2, making factors (*Hmisc* package)

```
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, round.POSIXt, trunc.POSIXt, units
```

Boxplots with cut2

```
p1 <- qplot(cutWage,age, data=training,fill=cutWage,
      geom=c("boxplot"))
p1
```

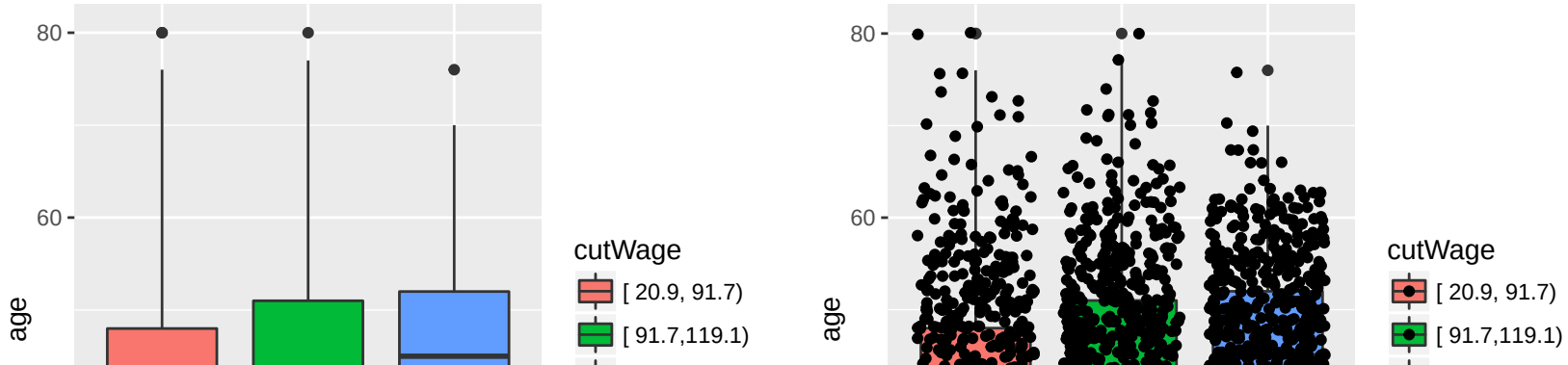


Boxplots with points overlayed

```
library(gridExtra)
```

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

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

```
p2 <- qplot(cutWage,age, data=training,fill=cutWage,
      geom=c("boxplot","jitter"))
grid.arrange(p1,p2,ncol=2)
```



Tables

```
t1 <- table(cutWage,training$jobclass)
t1
```

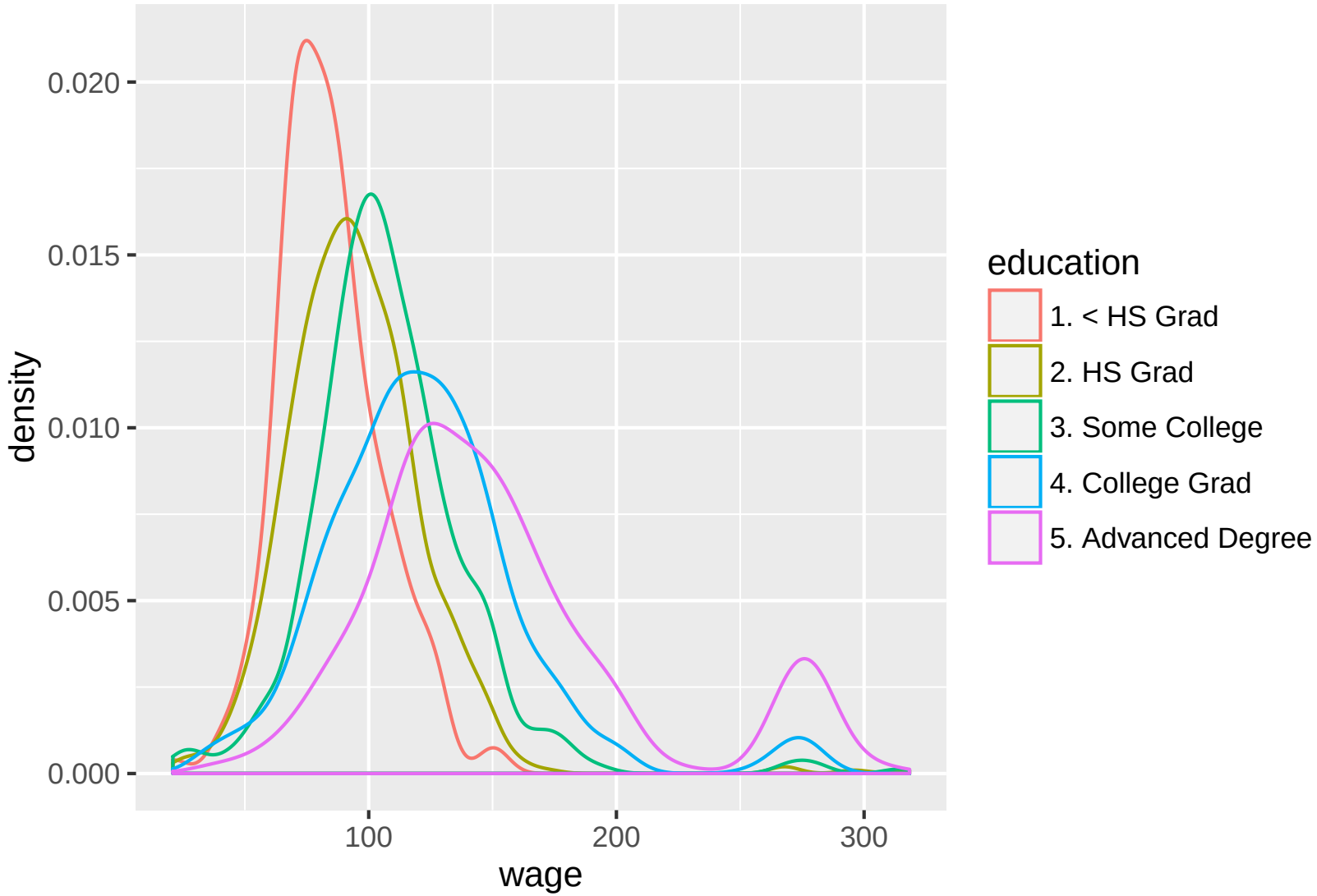
```
##
## cutWage          1. Industrial 2. Information
##   [ 20.9, 91.7)            444             260
##   [ 91.7,119.1)            379             349
##   [119.1,318.3]            270             400
```

```
prop.table(t1,1)
```

```
##
## cutWage          1. Industrial 2. Information
##   [ 20.9, 91.7)       0.6306818       0.3693182
##   [ 91.7,119.1)       0.5206044       0.4793956
##   [119.1,318.3]       0.4029851       0.5970149
```

Density plots

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



Notes and further reading

- Make your plots only in the training set
- Don't use the test set for exploration!
- Things you should be looking for
- Imbalance in outcomes/predictors
- Outliers
- Groups of points not explained by a predictor
- Skewed variables
- ggplot2 tutorial
- caret visualizations