Homework for Machine Learning

5: Applied: Exploratory Data Analysis and KDE a) Summary Data:

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## Parsed with column specification:
## cols(
##   EmployeeID = col_integer(),
##   recorddate_key = col_character(),
##   birthdate_key = col_date(format = ""),
##   orighiredate_key = col_date(format = ""),
##   terminationdate_key = col_date(format = ""),
##   age = col_integer(),
##   length_of_service = col_integer(),
##   city_name = col_character(),
##   department_name = col_character(),
##   job_title = col_character(),
##   store_name = col_character(),
##   gender_short = col_character(),
##   gender_full = col_character(),
##   termreason_desc = col_character(),
##   termtype_desc = col_character(),
##   STATUS_YEAR = col_integer(),
##   STATUS = col_character(),
##   BUSINESS_UNIT = col_character()
## )

##    EmployeeID   recorddate_key     birthdate_key
##  Min.   :1318   Length:49653       Min.   :1941-01-15
##  1st Qu.:3360   Class :character   1st Qu.:1958-05-28
##  Median :5031   Mode  :character   Median :1968-12-04
##  Mean   :4859                      Mean   :1969-01-09
##  3rd Qu.:6335                      3rd Qu.:1979-07-18
##  Max.   :8336                      Max.   :1994-12-31
##  orighiredate_key     terminationdate_key       age
##  Min.   :1989-08-28   Min.   :1900-01-01   Min.   :19.00
##  1st Qu.:1995-06-02   1st Qu.:1900-01-01   1st Qu.:31.00
##  Median :2000-03-31   Median :1900-01-01   Median :42.00
##  Mean   :2000-09-04   Mean   :1916-05-10   Mean   :42.08
##  3rd Qu.:2005-10-13   3rd Qu.:1900-01-01   3rd Qu.:53.00
```

```
##  Max.   :2013-12-11   Max.   :2015-12-30   Max.   :65.00
##  length_of_service  city_name         department_name
##  Min.   : 0.00     Length:49653       Length:49653
##  1st Qu.: 5.00     Class :character   Class :character
##  Median :10.00     Mode  :character   Mode  :character
##  Mean   :10.43
##  3rd Qu.:15.00
##  Max.   :26.00
##   job_title          store_name        gender_short
##  Length:49653       Length:49653       Length:49653
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  gender_full        termreason_desc    termtype_desc       STATUS_YEAR
##  Length:49653       Length:49653       Length:49653       Min.   :2006
##  Class :character   Class :character   Class :character   1st Qu.:2008
##  Mode  :character   Mode  :character   Mode  :character   Median :2011
##                                                           Mean   :2011
##                                                           3rd Qu.:2013
##                                                           Max.   :2015
##    STATUS          BUSINESS_UNIT
##  Length:49653       Length:49653
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

b) b1) the percent of terminated employees out of all employees for each year;

```
## Adding missing grouping variables: `STATUS_YEAR`
```

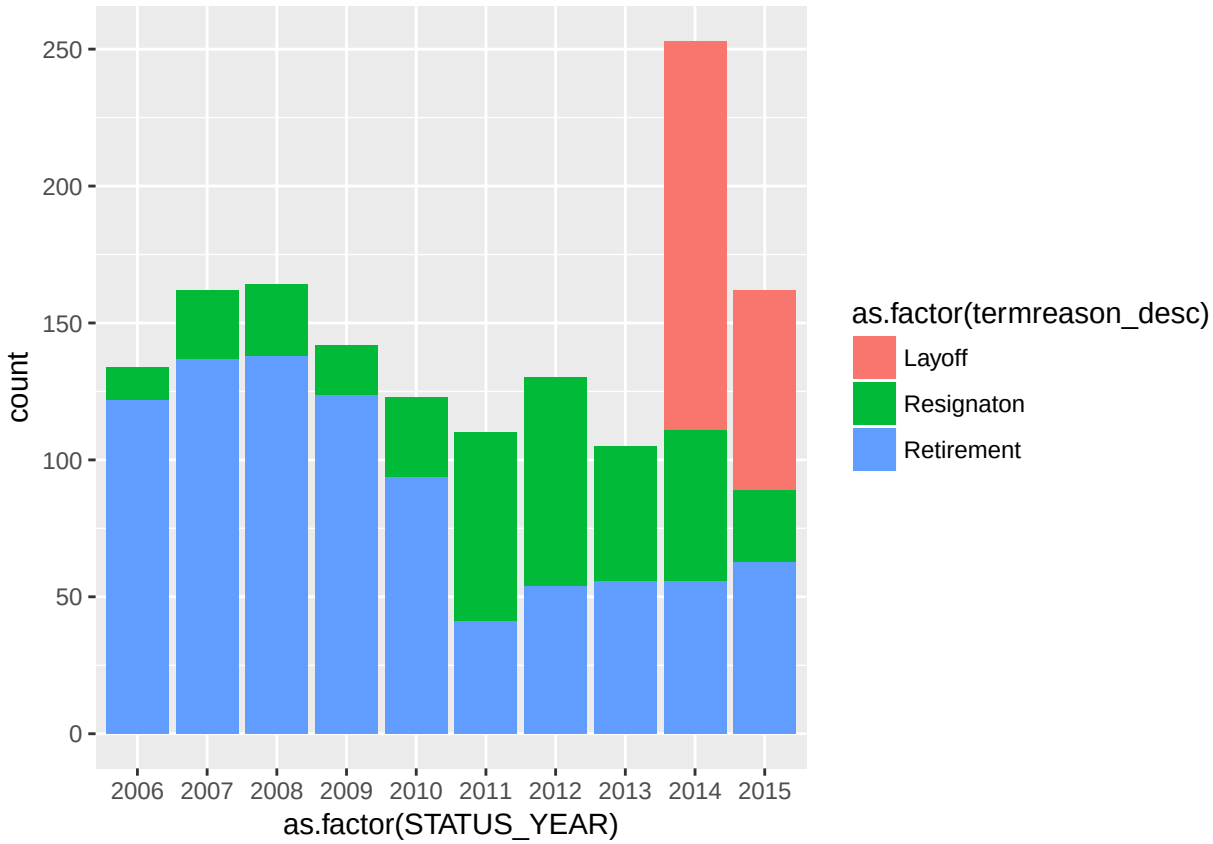
```
##      ACTIVE TERMINATED TOTAL PercentTerminated
## 2006   4445        134  4579          2.926403
## 2007   4521        162  4683          3.459321
## 2008   4603        164  4767          3.440319
## 2009   4710        142  4852          2.926628
## 2010   4840        123  4963          2.478340
## 2011   4972        110  5082          2.164502
## 2012   5101        130  5231          2.485184
## 2013   5215        105  5320          1.973684
## 2014   4962        253  5215          4.851390
## 2015   4799        162  4961          3.265471
```

The percent terminated ranges from 1.97 to 4.85. This is not a huge percent a year, and there doesnt seem to be a trend in terminations.

b2) average termination rate over the 10 years?

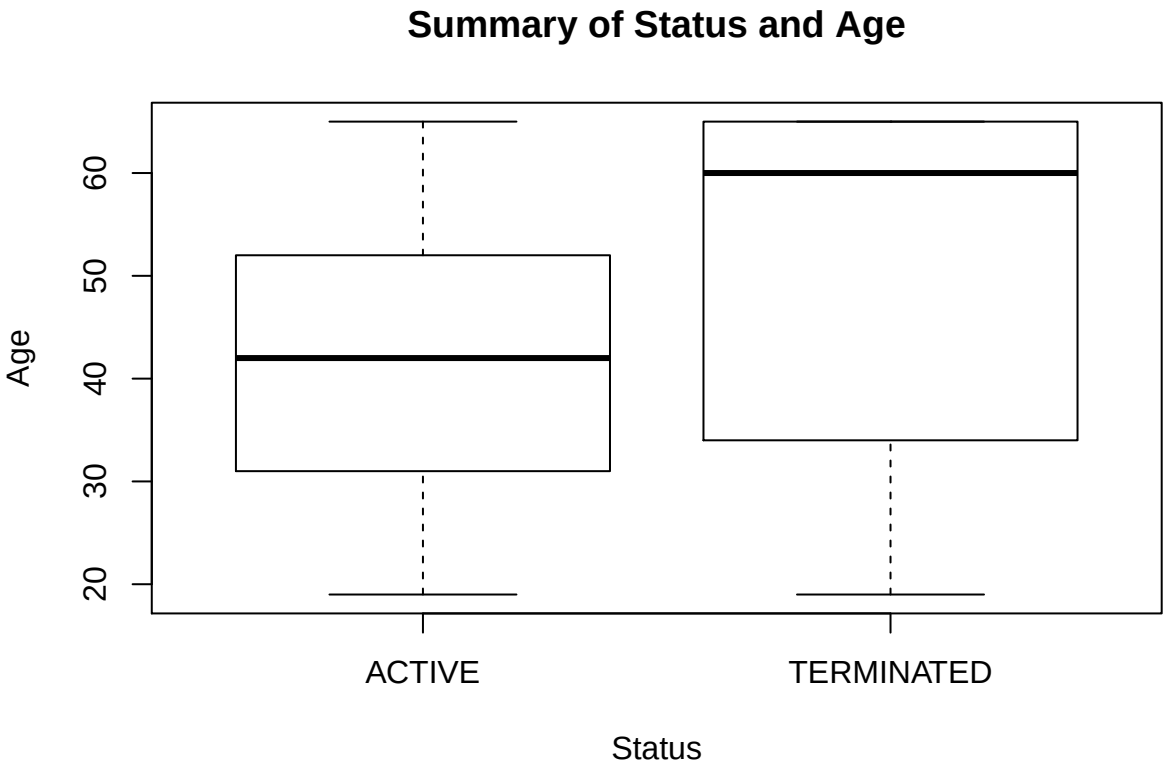
```
## [1] 2.997124
```

b3)



There were high layoffs in 2014 & 2015 that led to the terminations that we see in those years.

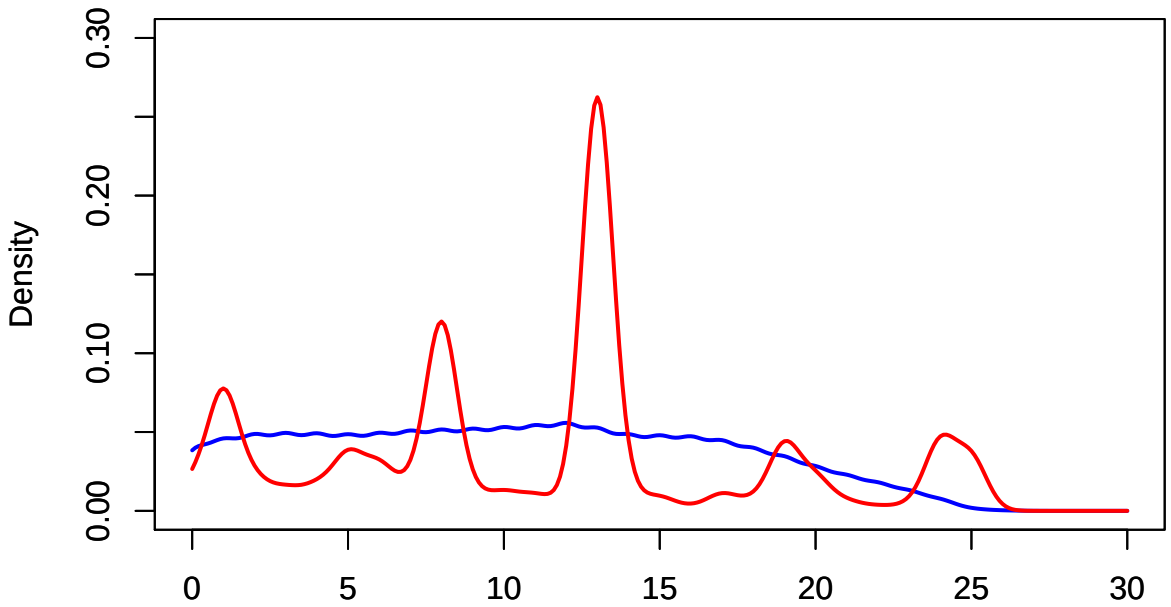
C) Does Age affect termination? : Box-plots of Age for active and terminated employees



The ones that are older are more likely terminated.

D) Kernal Density Function: Does Length of Service affect termination?

```
#samples


Activex <- subset(termination$length_of_service, termination$STATUS == "ACTIVE")
Terminatedx <- subset(termination$length_of_service, termination$STATUS == "TERMINATED")

xseq <-seq(0, 30, .1)

dataset<-Activex

 KDEG<- function(x){
  diff = x - dataset
  densitylist <- sapply(diff, dnorm, mean=0 , sd=0.5)
  density <- sum(densitylist)/(46000)
 }

 yseq <- sapply(xseq, KDEG)

plot(xseq, yseq, col="blue", xlab= "", ylab = "Density", type = "l", lwd=2, cex=2, ylim = c(0, 0.3))

dataset<-Terminatedx

 KDEG<- function(x){
  diff = x - dataset
  densitylist <- sapply(diff, dnorm, mean=0 , sd=0.5)
  density <- sum(densitylist)/(1485)
 }

 yseq <- sapply(xseq, KDEG)
par(new=TRUE)
plot(xseq, yseq, col="red", xlab= "", ylab = "Density", type = "l", lwd=2, cex=2, ylim = c(0, 0.3))
```

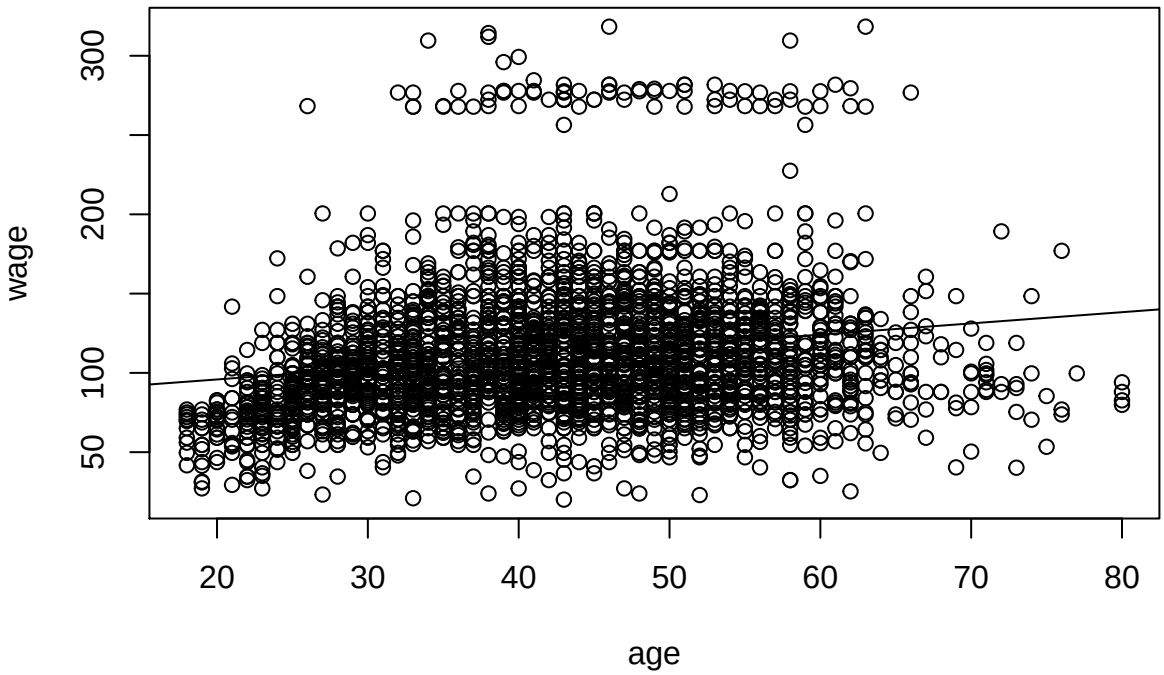


The lengths of service are more spread out for Active Status members. People who are terminated have peaks right before 10, 15, 20, and 25 years of service, with a really high peak around the 15 year mark.

6 Applied: Linear Regression a)

```
## Parsed with column specification:
## cols(
##   ID = col_integer(),
##   year = col_integer(),
##   age = col_integer(),
##   sex = col_character(),
##   maritl = col_character(),
##   race = col_character(),
##   education = col_character(),
##   region = col_character(),
##   jobclass = col_character(),
##   health = col_character(),
##   health_ins = col_character(),
##   logwage = col_double(),
##   wage = col_double()
## )
```

```
##        ID              year           age            sex
##  Min.   :  7373   Min.   :2003   Min.   :18.00   Length:3000
##  1st Qu.: 85622   1st Qu.:2004   1st Qu.:33.75   Class :character
##  Median :228800   Median :2006   Median :42.00   Mode  :character
##  Mean   :218883   Mean   :2006   Mean   :42.41
##  3rd Qu.:374760   3rd Qu.:2008   3rd Qu.:51.00
##  Max.   :453870   Max.   :2009   Max.   :80.00
##     maritl              race           education
##  Length:3000        Length:3000        Length:3000
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##     region            jobclass            health
##  Length:3000        Length:3000        Length:3000
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##   health_ins           logwage           wage
##  Length:3000        Min.   :3.000   Min.   : 20.09
##  Class :character   1st Qu.:4.447   1st Qu.: 85.38
##  Mode  :character   Median :4.653   Median :104.92
##                     Mean   :4.654   Mean   :111.70
##                     3rd Qu.:4.857   3rd Qu.:128.68
##                     Max.   :5.763   Max.   :318.34
```




```
##
## Call:
## lm(formula = wage ~ age, data = Wage)
##
## Coefficients:
## (Intercept)          age
##     81.7047       0.7073
```

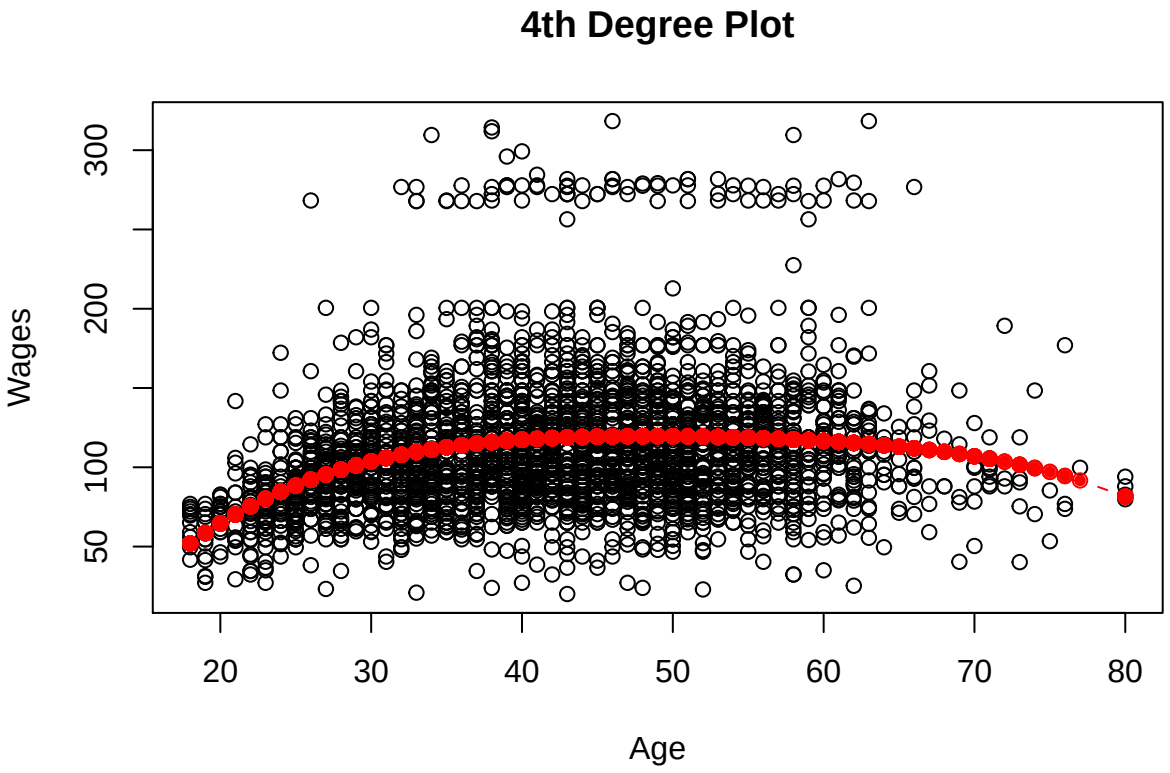
b)

```
##
## Call:
## lm(formula = wage ~ age + jobclass + age * jobclass, data = Wage)
##
## Coefficients:
##                (Intercept)                         age
##                    73.5283                      0.7197
##     jobclass2. Information  age:jobclass2. Information
##                    22.7309                     -0.1602
```

You will earn 22.73 on average more when in joblcass2 as apposed to joblcass 1, and earn .7197 more for every year old year are. Additionally, the interaction means that the effect of age on Height is different for different values of jobclass. c)

```
##
## Call:
## lm(formula = Wage$wage ~ Wage$age + I(Wage$age^2) + I(Wage$age^3) +
##     I(Wage$age^4))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -98.707 -24.626  -4.993  15.217 203.693
##
## Coefficients:
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.842e+02  6.004e+01  -3.067 0.002180 **
## Wage$age       2.125e+01  5.887e+00   3.609 0.000312 ***
## I(Wage$age^2) -5.639e-01  2.061e-01  -2.736 0.006261 **
## I(Wage$age^3)  6.811e-03  3.066e-03   2.221 0.026398 *
## I(Wage$age^4) -3.204e-05  1.641e-05  -1.952 0.051039 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.91 on 2995 degrees of freedom
## Multiple R-squared:  0.08626,    Adjusted R-squared:  0.08504
## F-statistic: 70.69 on 4 and 2995 DF,  p-value: < 2.2e-16
```



d)

```
##
## Call:
## lm(formula = Wage$wage ~ Wage$race)
##
## Coefficients:
##       (Intercept)  Wage$race2. Black  Wage$race3. Asian
##           112.564            -10.962              7.725
## Wage$race4. Other
##           -22.590
```

```
## NULL
```

```
##
## Call:
## lm(formula = Wage$wage ~ Wage$age + Wage$year + Wage$maritl +
##     Wage$race + Wage$jobclass + Wage$health + Wage$health_ins +
##     Wage$age * Wage$year + Wage$age * Wage$maritl + Wage$age *
##     Wage$race + Wage$age * Wage$jobclass + Wage$age * Wage$health +
```

```
##     Wage$age * Wage$health_ins + Wage$year * Wage$maritl + Wage$year *
##     Wage$race + Wage$year * Wage$jobclass + Wage$year * Wage$health +
##     Wage$year * Wage$health_ins + Wage$race * Wage$health + Wage$race *
##     Wage$health_ins + Wage$jobclass * Wage$health + Wage$jobclass *
##     Wage$health_ins + Wage$health * Wage$health_ins)
##
## Coefficients:
##                                          (Intercept)
##                                            -6.919e+03
##                                              Wage$age
##                                             1.157e+02
##                                             Wage$year
##                                             3.491e+00
##                                 Wage$maritl2. Married
##                                            -2.203e+03
##                                 Wage$maritl3. Widowed
##                                             3.341e+03
##                                Wage$maritl4. Divorced
##                                            -3.746e+03
##                               Wage$maritl5. Separated
##                                             4.473e+02
##                                     Wage$race2. Black
##                                            -1.765e+03
##                                     Wage$race3. Asian
##                                             4.878e+03
##                                     Wage$race4. Other
##                                             4.245e+03
##                           Wage$jobclass2. Information
##                                            -2.299e+03
##                             Wage$health2. >=Very Good
##                                             1.888e+03
##                                  Wage$health_ins2. No
##                                             1.614e+03
##                                Wage$age:Wage$year
##                                            -5.755e-02
##                          Wage$age:Wage$maritl2. Married
##                                            -2.596e-01
##                          Wage$age:Wage$maritl3. Widowed
##                                             1.232e-01
##                         Wage$age:Wage$maritl4. Divorced
##                                            -2.484e-01
##                        Wage$age:Wage$maritl5. Separated
##                                             5.536e-01
##                              Wage$age:Wage$race2. Black
##                                            -2.544e-01
##                              Wage$age:Wage$race3. Asian
##                                            -1.711e-01
##                              Wage$age:Wage$race4. Other
##                                            -9.295e-02
##                     Wage$age:Wage$jobclass2. Information
##                                             4.597e-02
##                       Wage$age:Wage$health2. >=Very Good
##                                             3.985e-01
##                            Wage$age:Wage$health_ins2. No
```

```
##                                             1.471e-01
##                         Wage$year:Wage$maritl2. Married
##                                             1.112e+00
##                         Wage$year:Wage$maritl3. Widowed
##                                            -1.670e+00
##                        Wage$year:Wage$maritl4. Divorced
##                                             1.873e+00
##                      Wage$year:Wage$maritl5. Separated
##                                            -2.337e-01
##                             Wage$year:Wage$race2. Black
##                                             8.829e-01
##                             Wage$year:Wage$race3. Asian
##                                            -2.424e+00
##                             Wage$year:Wage$race4. Other
##                                            -2.116e+00
##                   Wage$year:Wage$jobclass2. Information
##                                             1.151e+00
##                     Wage$year:Wage$health2. >=Very Good
##                                            -9.441e-01
##                          Wage$year:Wage$health_ins2. No
##                                            -8.168e-01
##             Wage$race2. Black:Wage$health2. >=Very Good
##                                            -8.566e+00
##             Wage$race3. Asian:Wage$health2. >=Very Good
##                                             3.405e+00
##             Wage$race4. Other:Wage$health2. >=Very Good
##                                            -1.620e+01
##                  Wage$race2. Black:Wage$health_ins2. No
##                                             5.599e+00
##                  Wage$race3. Asian:Wage$health_ins2. No
##                                            -1.201e+01
##                  Wage$race4. Other:Wage$health_ins2. No
##                                            -3.366e+00
## Wage$jobclass2. Information:Wage$health2. >=Very Good
##                                             3.972e+00
##      Wage$jobclass2. Information:Wage$health_ins2. No
##                                            -5.000e+00
##        Wage$health2. >=Very Good:Wage$health_ins2. No
##                                            -2.402e+00
```