Review.R

soonmi

2020-11-20

```r
library(mosaicData)
library(dplyr) #functions like arrange
```

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

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

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

```r
library(tidyverse) #for ggplot
```

```
## -- Attaching packages ------------------------------------------------------------------------------

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v stringr 1.4.0
## v tidyr   1.1.2     v forcats 0.5.0
## v readr   1.3.1

## -- Conflicts ---------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(ggplot2) #for ggplot
library(knitr) #for kable

data(package = "mosaicData")
data(SAT)

#random numbers
set.seed(120)
random_sample <- sample(nrow(SAT), size=10, replace=FALSE)

#basics of data
SAT[1:5, ]
```

```
##        state expend ratio salary frac verbal math  sat
## 1    Alabama  4.405  17.2 31.144    8    491  538 1029
## 2     Alaska  8.963  17.6 47.951   47    445  489  934
## 3    Arizona  4.778  19.3 32.175   27    448  496  944
## 4   Arkansas  4.459  17.1 28.934    6    482  523 1005
## 5 California  4.992  24.0 41.078   45    417  485  902
```

```
attach(SAT)
names(SAT)
```

```
## [1] "state"  "expend" "ratio"  "salary" "frac"   "verbal" "math"   "sat"
```

```
dim(SAT)
```

```
## [1] 50  8
```

```
summary(SAT)
```

```
##         state        expend          ratio           salary
##  Alabama   : 1   Min.   :3.656   Min.   :13.80   Min.   :25.99
##  Alaska    : 1   1st Qu.:4.882   1st Qu.:15.22   1st Qu.:30.98
##  Arizona   : 1   Median :5.768   Median :16.60   Median :33.29
##  Arkansas  : 1   Mean   :5.905   Mean   :16.86   Mean   :34.83
##  California: 1   3rd Qu.:6.434   3rd Qu.:17.57   3rd Qu.:38.55
##  Colorado  : 1   Max.   :9.774   Max.   :24.30   Max.   :50.05
##  (Other)   :44
##       frac            verbal           math            sat
##  Min.   : 4.00   Min.   :401.0   Min.   :443.0   Min.   : 844.0
##  1st Qu.: 9.00   1st Qu.:427.2   1st Qu.:474.8   1st Qu.: 897.2
##  Median :28.00   Median :448.0   Median :497.5   Median : 945.5
##  Mean   :35.24   Mean   :457.1   Mean   :508.8   Mean   : 965.9
##  3rd Qu.:63.00   3rd Qu.:490.2   3rd Qu.:539.5   3rd Qu.:1032.0
##  Max.   :81.00   Max.   :516.0   Max.   :592.0   Max.   :1107.0
##
```

```
#correlation
cor(math, verbal)
```

```
## [1] 0.970256
```

```
cor(SAT[,2:8])
```

```
##            expend        ratio       salary       frac      verbal        math
## expend  1.0000000 -0.371025386  0.869801513  0.5926274 -0.41004987 -0.34941409
## ratio  -0.3710254  1.000000000 -0.001146081 -0.2130536  0.06376664  0.09542173
## salary  0.8698015 -0.001146081  1.000000000  0.6167799 -0.47696364 -0.40131282
## frac    0.5926274 -0.213053607  0.616779867  1.0000000 -0.89326296 -0.86938393
## verbal -0.4100499  0.063766636 -0.476963635 -0.8932630  1.00000000  0.97025604
## math   -0.3494141  0.095421730 -0.401312817 -0.8693839  0.97025604  1.00000000
## sat    -0.3805370  0.081253823 -0.439883381 -0.8871187  0.99150325  0.99350238
##                  sat
## expend -0.38053700
## ratio   0.08125382
## salary -0.43988338
## frac   -0.88711868
## verbal  0.99150325
## math    0.99350238
## sat     1.00000000
```

```
#continuous into groups
salary_split <- matrix(0, nrow=nrow(SAT), ncol=1)
for(i in 1: nrow(SAT)) {
  if(salary[i]>=38.55){salary_split[i] <-1}
  else if((salary[i]<38.55) & (salary[i]>=34.83)){salary_split[i] <-2}
  else if((salary[i]<34.83) & (salary[i]>=33.29)){salary_split[i] <-3}
  else if((salary[i]<33.29) & (salary[i]>=30.98)){salary_split[i] <-4}
  else {salary_split[i] <-5}
}

salary_split <-as.factor(salary_split)
table(salary_split)
```

```
## salary_split
##  1  2  3  4  5
## 13  8  4 12 13
```

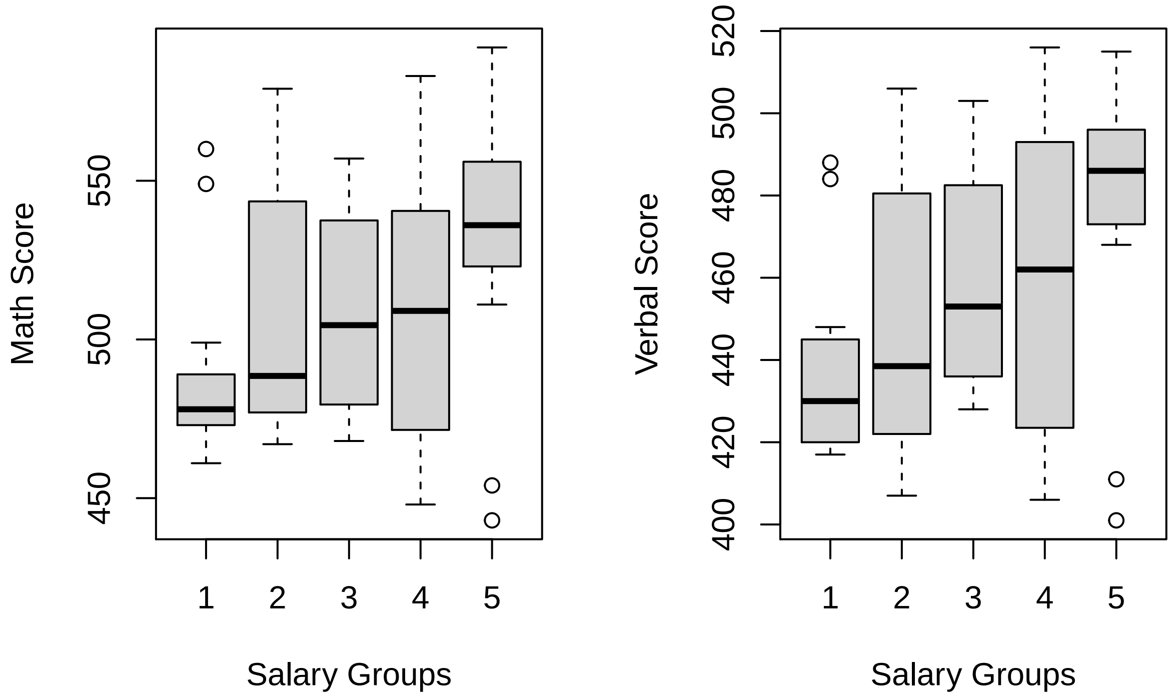
```
#adding new column
SAT <- cbind(SAT, salary_split)
#SAT <- drop(salary_split)

SAT[1:10,]
```

```
##          state expend ratio salary frac verbal math  sat salary_split
## 1      Alabama  4.405  17.2 31.144    8    491  538 1029            4
## 2       Alaska  8.963  17.6 47.951   47    445  489  934            1
## 3      Arizona  4.778  19.3 32.175   27    448  496  944            4
## 4     Arkansas  4.459  17.1 28.934    6    482  523 1005            5
## 5   California  4.992  24.0 41.078   45    417  485  902            1
## 6     Colorado  5.443  18.4 34.571   29    462  518  980            3
## 7  Connecticut  8.817  14.4 50.045   81    431  477  908            1
## 8     Delaware  7.030  16.6 39.076   68    429  468  897            1
## 9      Florida  5.718  19.1 32.588   48    420  469  889            4
## 10     Georgia  5.193  16.3 32.291   65    406  448  854            4
```

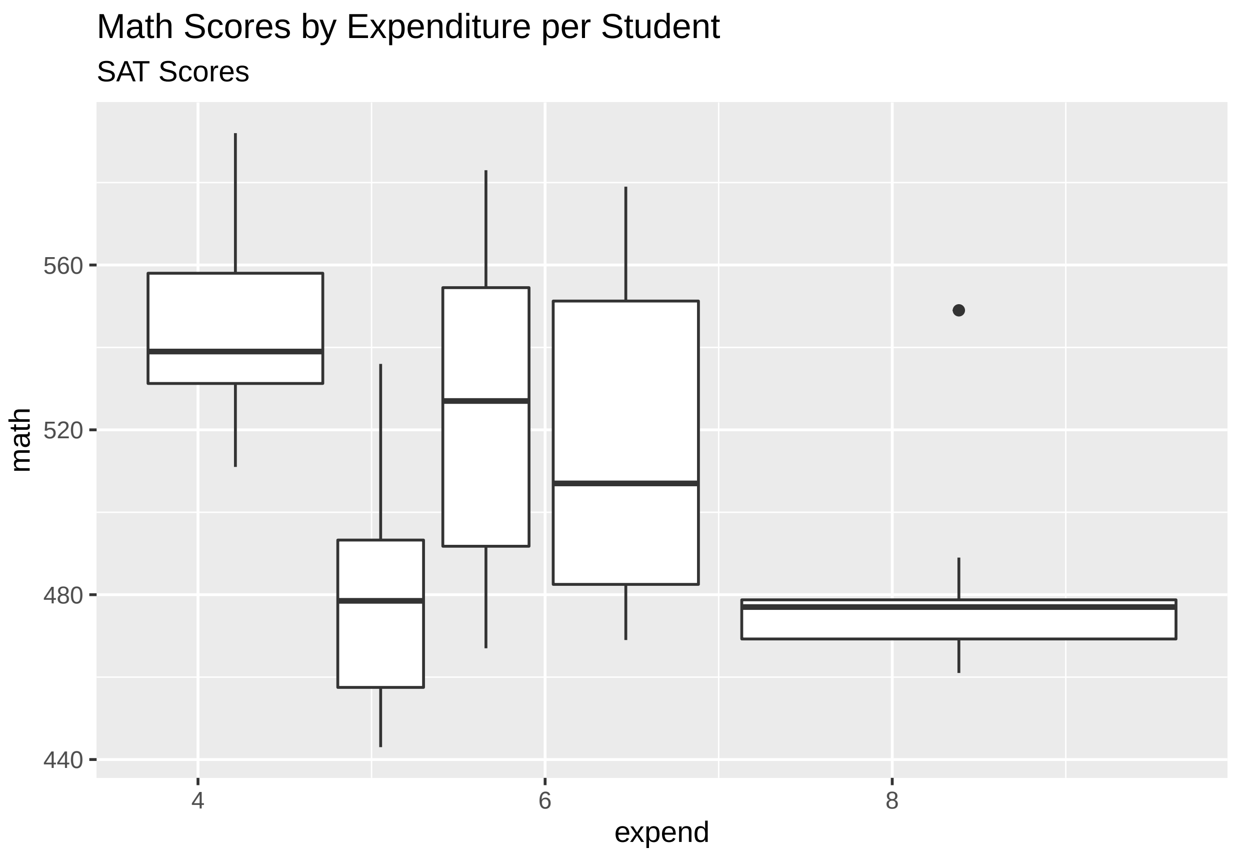
```
SAT_bysal <- arrange(SAT, desc(salary))

SAT_bysal[1:10,]
```

```
##            state expend ratio salary frac verbal math  sat salary_split
## 1    Connecticut  8.817  14.4 50.045   81    431  477  908            1
## 2         Alaska  8.963  17.6 47.951   47    445  489  934            1
## 3       New York  9.623  15.2 47.612   74    419  473  892            1
## 4     New Jersey  9.774  13.8 46.087   70     420  478  898             1
## 5   Pennsylvania  7.109  17.1 44.510   70     419  461  880             1
## 6       Michigan  6.994  20.1 41.895   11     484  549 1033             1
## 7     California  4.992  24.0 41.078   45     417  485  902             1
## 8  Massachusetts  7.287  14.8 40.795   80     430  477  907             1
## 9   Rhode Island  7.469  14.7 40.729   70     425  463  888             1
## 10      Maryland  7.245  17.0 40.661   64     430  479  909             1
```

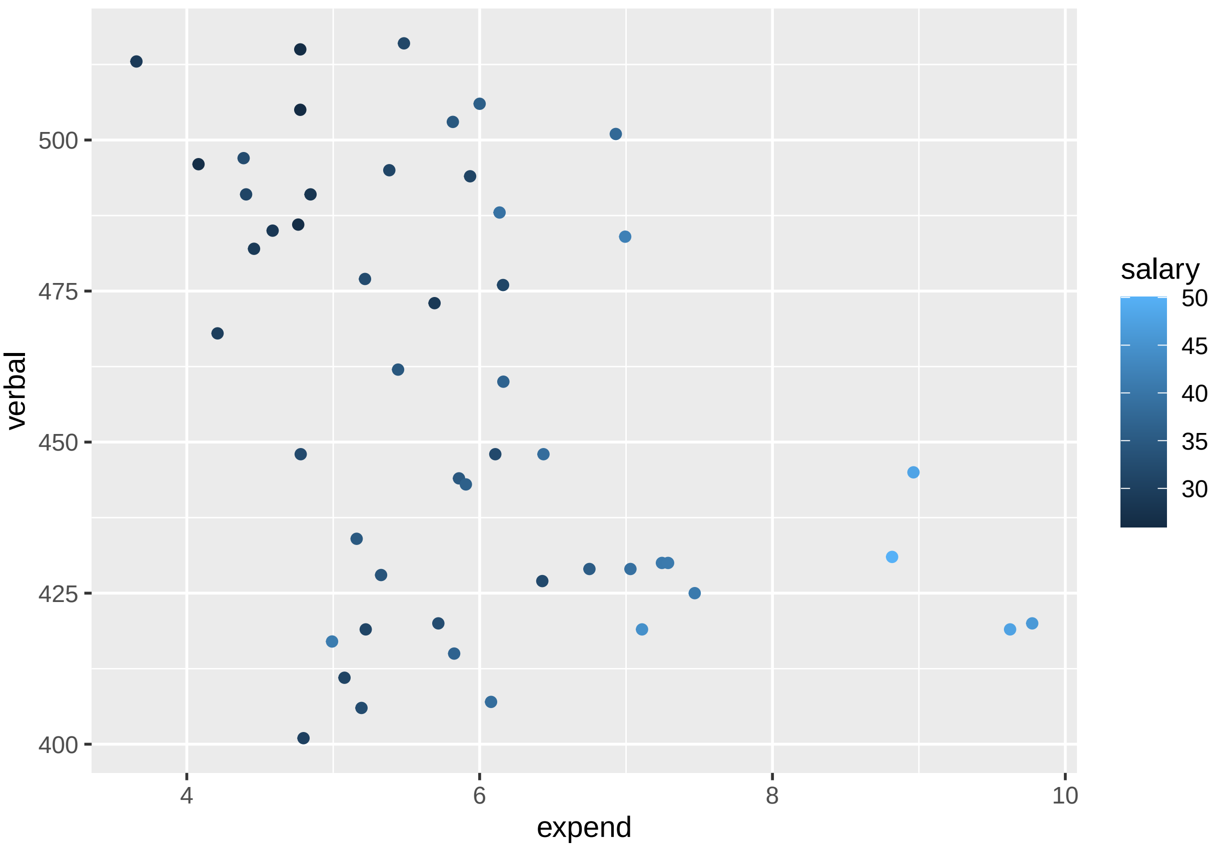
```
#boxplots without ggplot
par(mfrow=c(1,2))
boxplot(math~salary_split, xlab = "Salary Groups", ylab = "Math Score")
boxplot(verbal~salary_split, xlab = "Salary Groups", ylab = "Verbal Score")
```



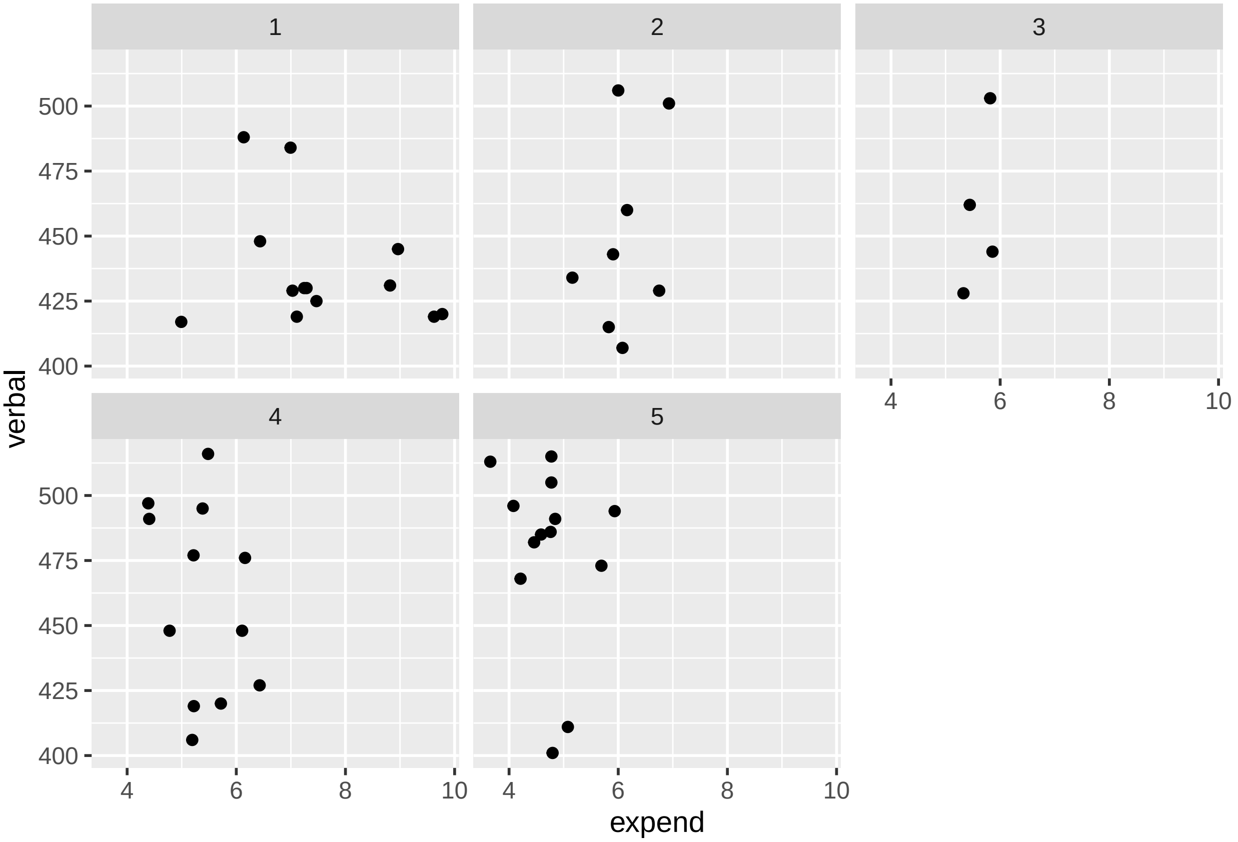
```
#ggplot exploration
ggplot(data=SAT, mapping = aes(x=expend, y=math)) +
  geom_boxplot(mapping = aes(group = cut_number(expend,5))) +
  labs(title = "Math Scores by Expenditure per Student", subtitle = "SAT Scores")
```



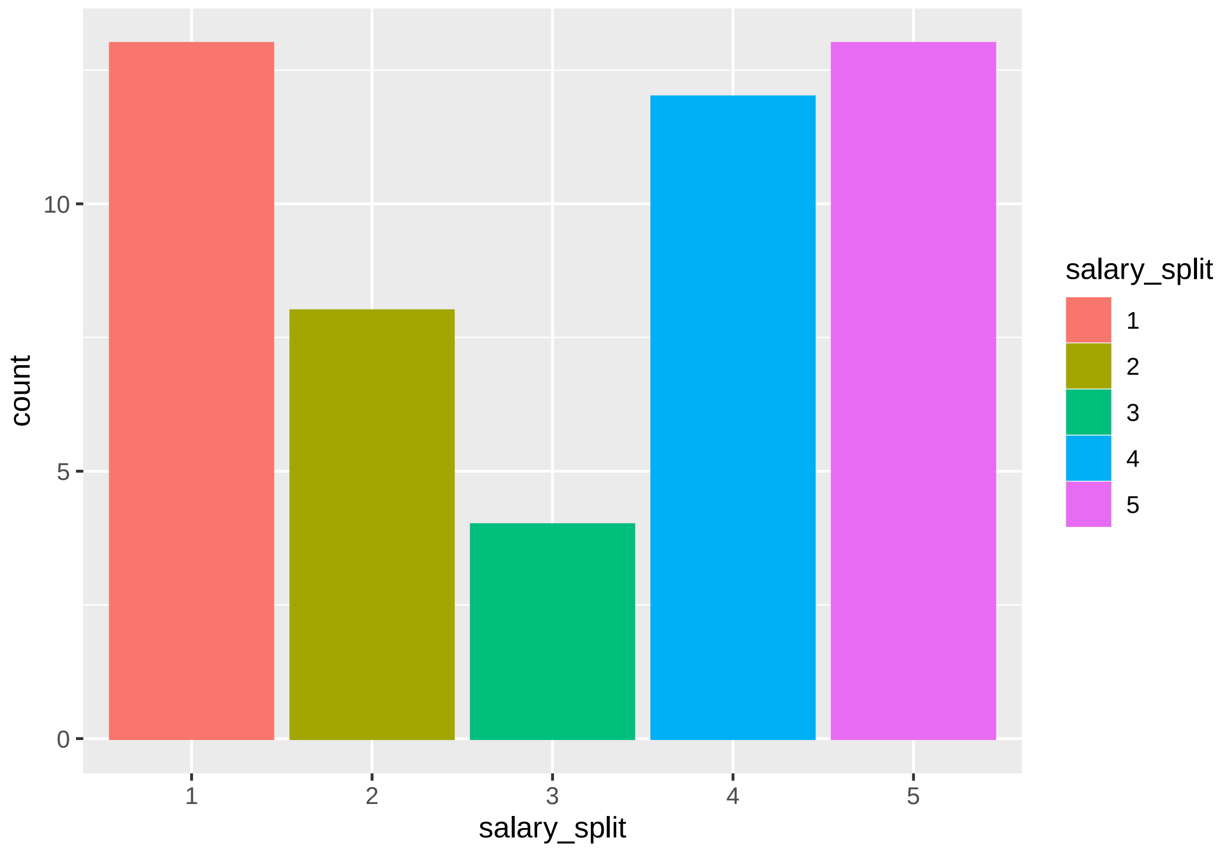
```r
ggplot(data=SAT) + geom_point(mapping = aes(x=expend, y=verbal, color = salary))
```



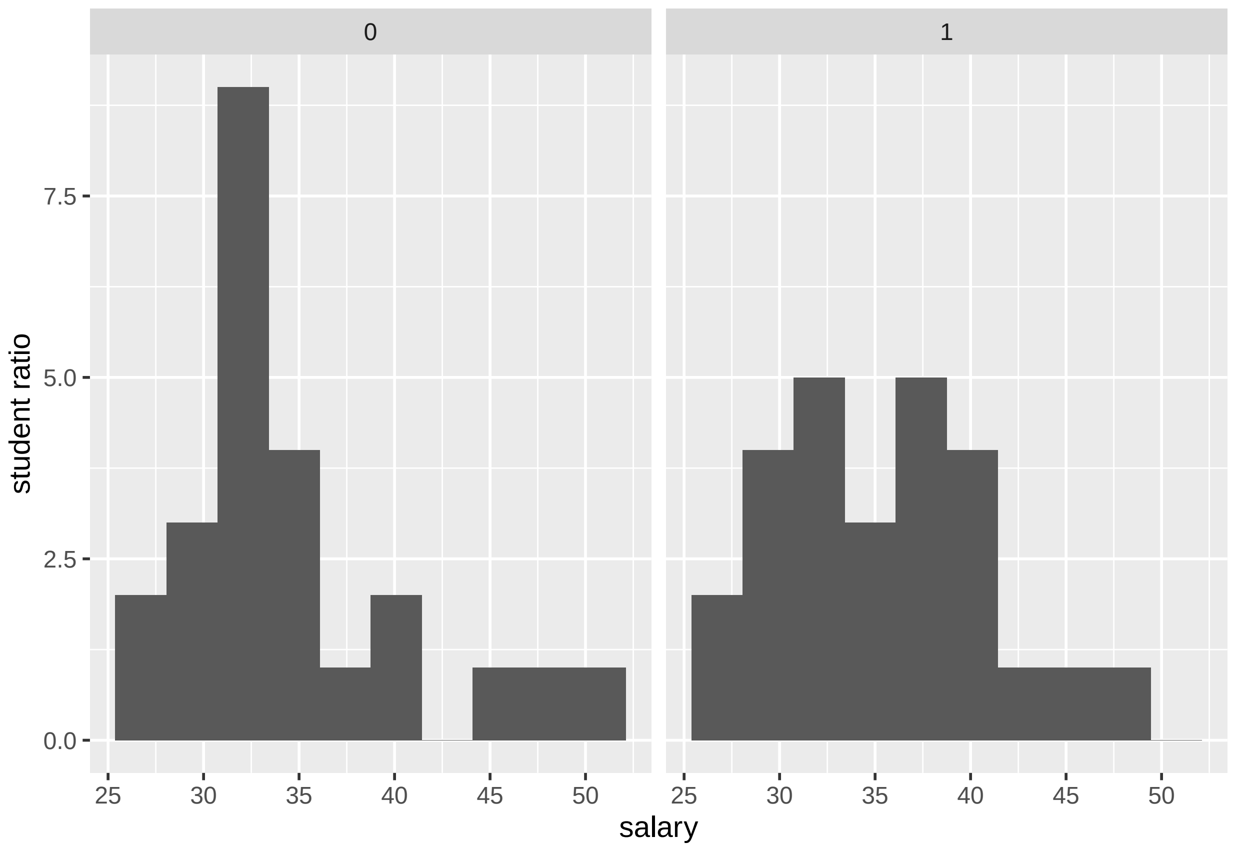
```r
ggplot(data=SAT) + geom_point(mapping = aes(x=expend, y=verbal)) + facet_wrap(~salary_split, nrow=2)
```



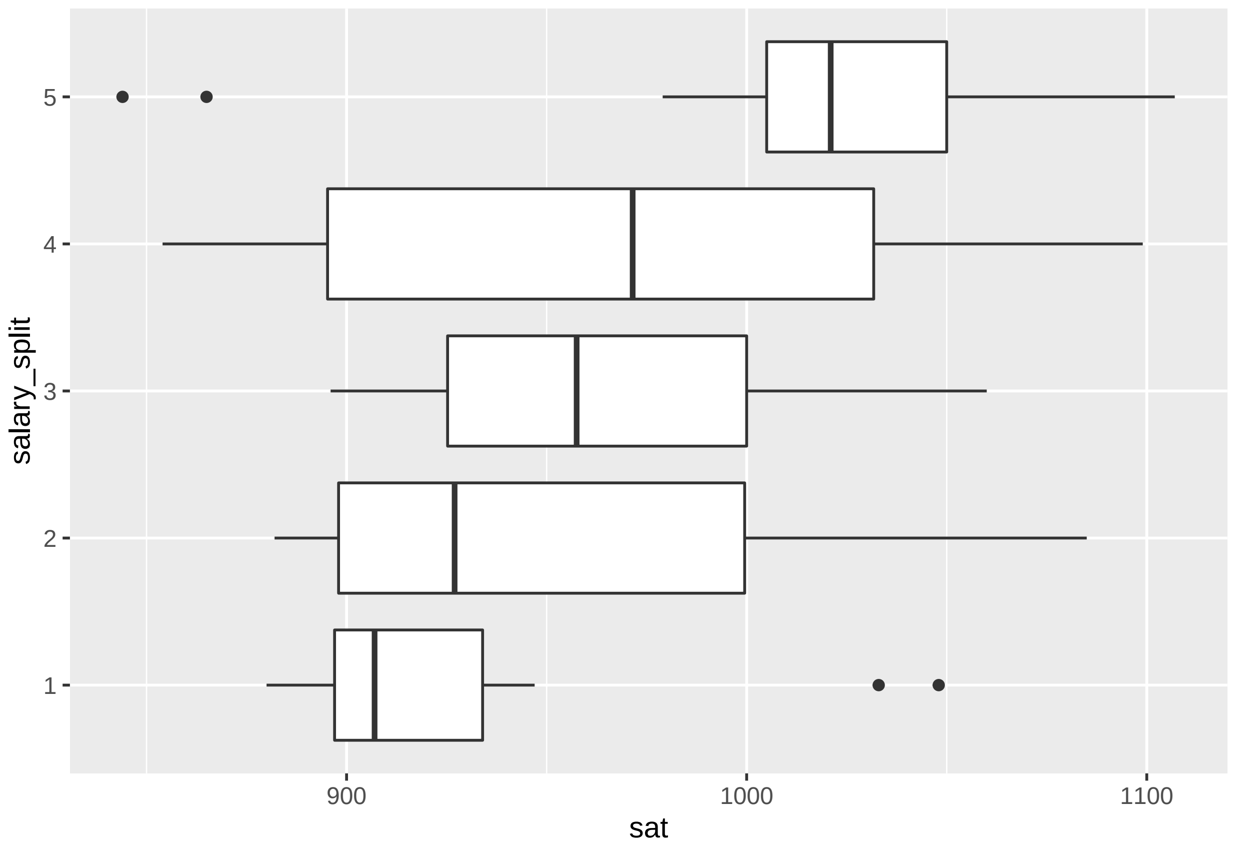
```
ggplot(data=SAT) + geom_bar(mapping = aes(x=salary_split, color=salary_split, fill=salary_split))
```



```
#making another column
ratio_split <- matrix(0, nrow=nrow(SAT), ncol=1)
med_ratio <- median(ratio)
for(i in 1: nrow(SAT)){
  if (ratio[i] >= med_ratio){ratio_split[i] <- 1}
  else {ratio_split[i] <- 0}
}
SAT <- cbind(SAT, ratio_split)

#more ggplot
ggplot(data=SAT) + geom_histogram(mapping = aes(x=salary), bins=10) + facet_wrap(~ratio_split, ncol=2) +
  ylab("student ratio")
```



```
ggplot(data = SAT, mapping = aes(x=salary_split, y=sat)) + geom_boxplot() + coord_flip()
```



```r
##switching over to SaratogaHouses data in mosaicData

summarize(SaratogaHouses, mean_bedrooms = mean(bedrooms))
```

```
##   mean_bedrooms
## 1      3.154514
```

```r
#using pipe operator
by_NC <- SaratogaHouses %>% group_by(newConstruction) %>%
  summarize(mean=mean(bedrooms))
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

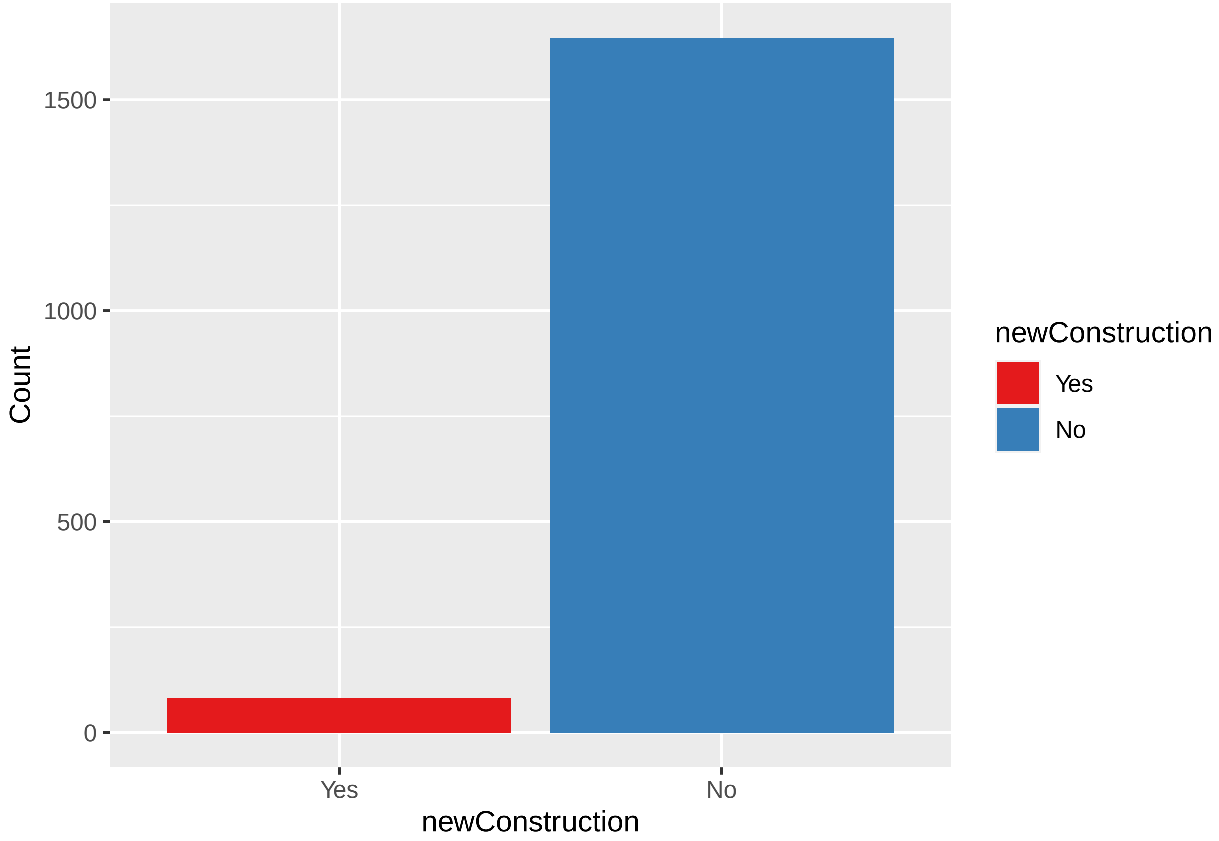
```r
by_NC
```

```
## # A tibble: 2 x 2
##   newConstruction  mean
##   <fct>           <dbl>
## 1 Yes              3.68
## 2 No               3.13
```

```r
as.data.frame(by_NC)
```

```
##   newConstruction     mean
## 1             Yes 3.679012
## 2              No 3.128719
```

```
SaratogaHouses %>% ggplot(aes(x=newConstruction, fill=newConstruction)) +geom_bar()+
  ylab("Count") + scale_fill_brewer(palette="Set1")
```



```
#filter search
with_fpwf <- SaratogaHouses %>%
  filter(fireplaces==1, waterfront =="Yes")
dim(with_fpwf)
```

```
## [1]  8 16
```

```
with_fpwf
```

```
##    price lotSize age landValue livingArea pctCollege bedrooms fireplaces
## 1 457000    0.43  53      2700       2461         39        4          1
## 2 490000    0.34  18     79700       1346         52        3          1
## 3 319000    0.50   5     40200       1681         57        3          1
## 4 290000    1.00  33     21700        944         27        1          1
## 5 775000    0.00   5    412600       2472         57        3          1
## 6 320900    0.47   5     20400       1885         21        2          1
## 7 430000    1.34  15     75700       2649         21        3          1
## 8 325000    0.27 105     56500       1391         40         2           1
##   bathrooms rooms         heating     fuel              sewer waterfront
## 1       2.0    10         hot air      oil public/commercial        Yes
## 2       2.0     6         hot air      oil public/commercial        Yes
## 3       2.5     4         hot air      gas public/commercial        Yes
## 4       1.0     4         hot air      oil            septic        Yes
## 5       2.5     9         hot air      gas            septic        Yes
## 6       2.0     7 hot water/steam      oil            septic        Yes
## 7       3.0     7        electric electric            septic        Yes
## 8       1.0     4        electric electric public/commercial        Yes
##   newConstruction centralAir
## 1              No         No
## 2              No         No
## 3              No        Yes
## 4              No         No
## 5              No        Yes
## 6              No         No
## 7              No         No
## 8              No         No
```

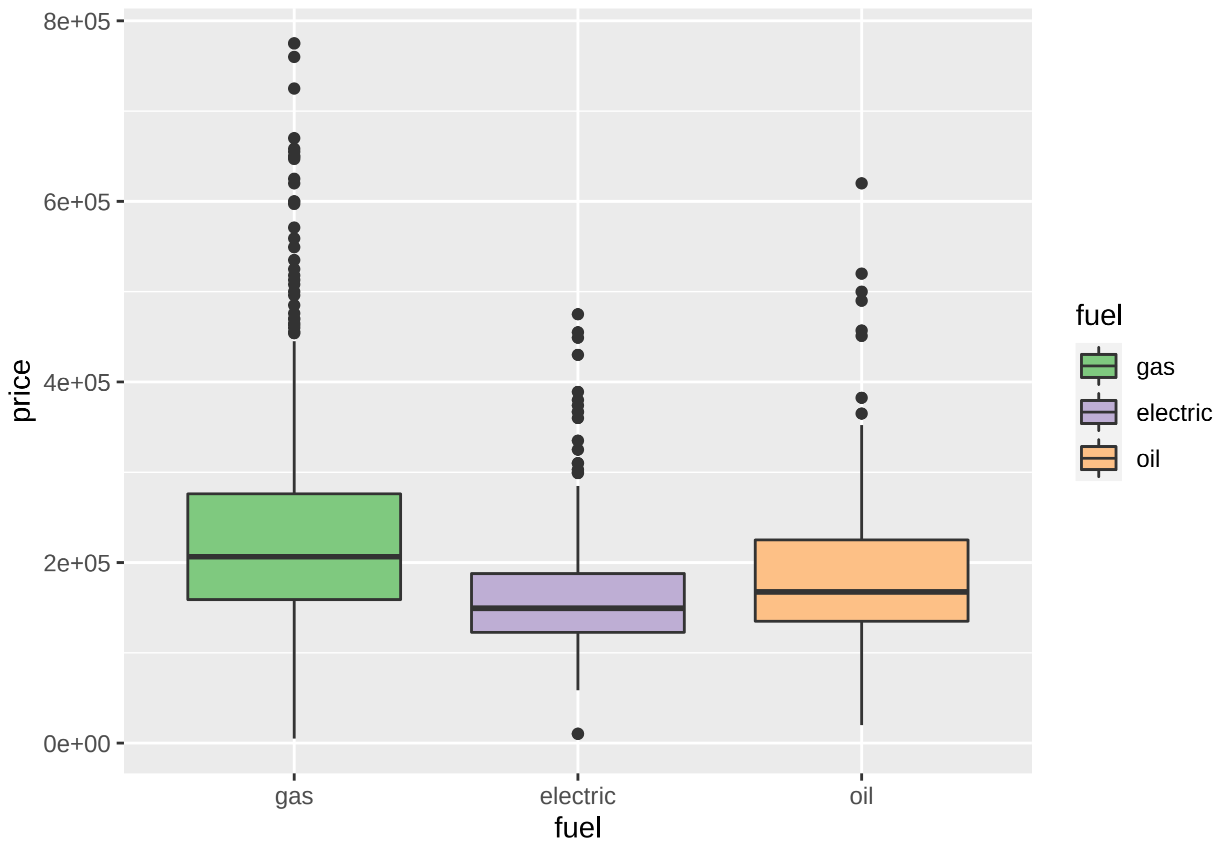
```
#categorical variables
price_fuel_heat <- SaratogaHouses %>% group_by(fuel, heating) %>%
  summarize(mean_p=mean(price), freq=n(), mean_a=mean(age))
```

```
## ‘summarise()‘ regrouping output by ’fuel’ (override with ‘.groups‘ argument)
```

```
kable(price_fuel_heat)
```

fuel	heating	mean_p	freq	mean_a
gas	hot air	231363.7	961	21.55151
gas	hot water/steam	218346.4	230	44.20435
gas	electric	166050.0	6	16.50000
electric	hot air	221131.4	16	15.25000
electric	hot water/steam	237500.0	1	19.00000
electric	electric	161676.9	298	21.10403
oil	hot air	193512.5	144	46.50694
oil	hot water/steam	178885.0	71	55.33803
oil	electric	200000.0	1	84.00000

```
SaratogaHouses %>% ggplot(aes(x=fuel, y=price, fill=fuel)) + geom_boxplot() +
  scale_fill_brewer(palette="Accent")
```



```
#data details
SaratogaHouses %>% head(5)
```

```
##    price lotSize age landValue livingArea pctCollege bedrooms fireplaces
## 1 132500    0.09  42     50000        906         35        2          1
## 2 181115    0.92   0     22300       1953         51        3          0
## 3 109000    0.19 133      7300       1944         51        4          1
## 4 155000    0.41  13     18700       1944         51        3          1
## 5  86060    0.11   0     15000        840         51        2          0
##   bathrooms rooms          heating     fuel              sewer waterfront
## 1       1.0     5         electric electric             septic         No
## 2       2.5     6 hot water/steam      gas             septic         No
## 3       1.0     8 hot water/steam      gas public/commercial         No
## 4       1.5     5          hot air      gas             septic         No
## 5       1.0     3          hot air      gas public/commercial         No
##   newConstruction centralAir
## 1              No         No
## 2              No         No
## 3              No         No
## 4              No         No
## 5             Yes        Yes
```

```
SaratogaHouses %>% glimpse
```

```
## Rows: 1,728
## Columns: 16
## $ price           <int> 132500, 181115, 109000, 155000, 86060, 120000, 1530...
## $ lotSize         <dbl> 0.09, 0.92, 0.19, 0.41, 0.11, 0.68, 0.40, 1.21, 0.8...
## $ age             <int> 42, 0, 133, 13, 0, 31, 33, 23, 36, 4, 123, 1, 13, 1...
## $ landValue       <int> 50000, 22300, 7300, 18700, 15000, 14000, 23300, 146...
## $ livingArea      <int> 906, 1953, 1944, 1944, 840, 1152, 2752, 1662, 1632,...
## $ pctCollege      <int> 35, 51, 51, 51, 51, 22, 51, 35, 51, 44, 51, 51, 41,...
## $ bedrooms        <int> 2, 3, 4, 3, 2, 4, 4, 4, 3, 3, 7, 3, 2, 3, 3, 3, 3, ...
## $ fireplaces      <int> 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ bathrooms       <dbl> 1.0, 2.5, 1.0, 1.5, 1.0, 1.0, 1.5, 1.5, 1.5, 1.5, 1...
## $ rooms           <int> 5, 6, 8, 5, 3, 8, 8, 9, 8, 6, 12, 6, 4, 5, 8, 4, 7,...
## $ heating         <fct> electric, hot water/steam, hot water/steam, hot air...
## $ fuel            <fct> electric, gas, gas, gas, gas, gas, oil, oil, electr...
## $ sewer           <fct> septic, septic, public/commercial, septic, public/c...
## $ waterfront      <fct> No, No, No, No, No, No, No, No, No, No, No, No, No,...
## $ newConstruction <fct> No, No, No, No, Yes, No, No, No, No, No, No, No, No...
## $ centralAir      <fct> No, No, No, No, Yes, No, No, No, No, No, No, No, No...
```

```
SaratogaHouses %>% str
```

```
## 'data.frame':    1728 obs. of  16 variables:
##  $ price          : int  132500 181115 109000 155000 86060 120000 153000 170000 90000 122900 ...
##  $ lotSize        : num  0.09 0.92 0.19 0.41 0.11 0.68 0.4 1.21 0.83 1.94 ...
##  $ age            : int  42 0 133 13 0 31 33 23 36 4 ...
##  $ landValue      : int  50000 22300 7300 18700 15000 14000 23300 14600 22200 21200 ...
##  $ livingArea     : int  906 1953 1944 1944 840 1152 2752 1662 1632 1416 ...
##  $ pctCollege     : int  35 51 51 51 51 22 51 35 51 44 ...
##  $ bedrooms       : int  2 3 4 3 2 4 4 4 3 3 ...
##  $ fireplaces     : int  1 0 1 1 0 1 1 1 0 0 ...
##  $ bathrooms      : num  1 2.5 1 1.5 1 1 1.5 1.5 1.5 1.5 ...
##  $ rooms          : int  5 6 8 5 3 8 8 9 8 6 ...
##  $ heating        : Factor w/ 3 levels "hot air","hot water/steam",..: 3 2 2 1 1 1 2 1 3 1 ...
##  $ fuel           : Factor w/ 3 levels "gas","electric",..: 2 1 1 1 1 1 3 3 2 1 ...
##  $ sewer          : Factor w/ 3 levels "septic","public/commercial",..: 1 1 2 1 2 1 1 1 1 3 ...
##  $ waterfront     : Factor w/ 2 levels "Yes","No": 2 2 2 2 2 2 2 2 2 2 ...
##  $ newConstruction: Factor w/ 2 levels "Yes","No": 2 2 2 2 1 2 2 2 2 2 ...
##  $ centralAir     : Factor w/ 2 levels "Yes","No": 2 2 2 2 1 2 2 2 2 2 ...
```

```
SaratogaHouses %>% nrow
```

```
## [1] 1728
```

```
SaratogaHouses %>% names
```

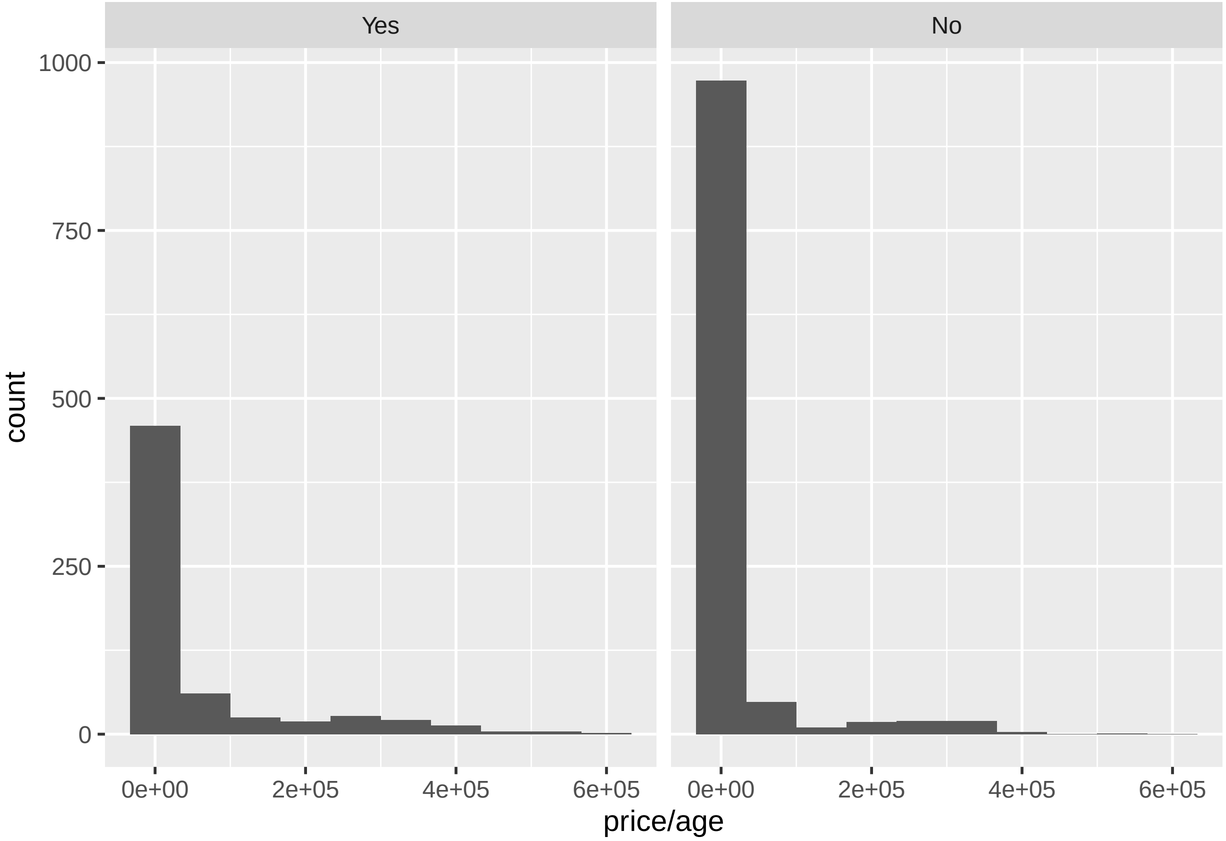
```
##  [1] "price"           "lotSize"         "age"             "landValue"
##  [5] "livingArea"      "pctCollege"      "bedrooms"        "fireplaces"
##  [9] "bathrooms"       "rooms"           "heating"         "fuel"
## [13] "sewer"           "waterfront"      "newConstruction" "centralAir"
```

```
#change zeros to 1 to make price/age a valid variable
SaratogaHouses$age[SaratogaHouses$age == 0] <- 1
```

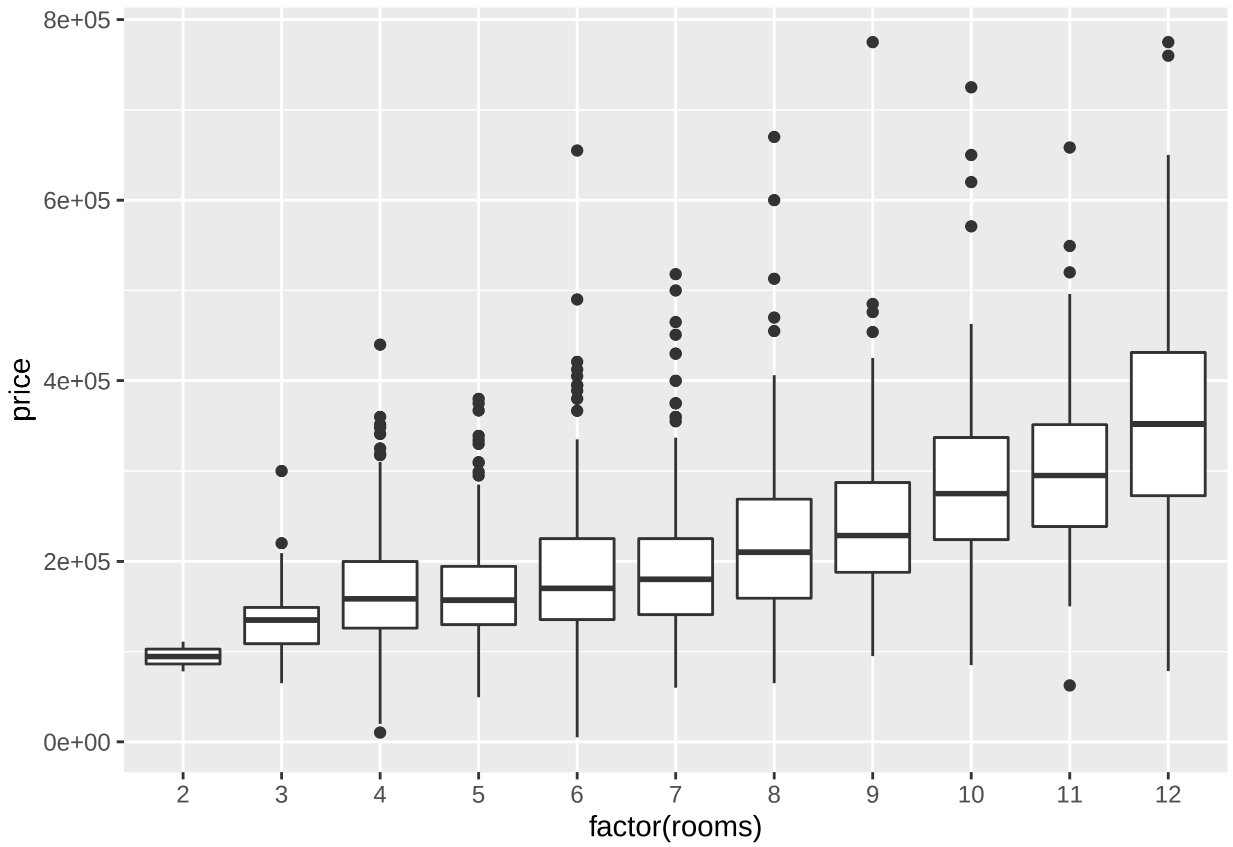
```r
#create new variables in a new data set
SaratogaHouses2 <- SaratogaHouses %>% mutate(price/age)
SaratogaHouses2 %>% head(5)
```

```
##    price lotSize age landValue livingArea pctCollege bedrooms fireplaces
## 1 132500    0.09  42     50000        906         35        2          1
## 2 181115    0.92   1     22300       1953         51        3          0
## 3 109000    0.19 133      7300       1944         51        4          1
## 4 155000    0.41  13     18700       1944         51        3          1
## 5  86060    0.11   1     15000        840         51        2          0
##   bathrooms rooms         heating     fuel             sewer waterfront
## 1       1.0     5        electric electric            septic         No
## 2       2.5     6 hot water/steam      gas            septic         No
## 3       1.0     8 hot water/steam      gas public/commercial         No
## 4       1.5     5         hot air      gas            septic         No
## 5       1.0     3         hot air      gas public/commercial         No
##   newConstruction centralAir   price/age
## 1              No         No   3154.7619
## 2              No         No 181115.0000
## 3              No         No    819.5489
## 4              No         No  11923.0769
## 5             Yes        Yes  86060.0000
```

```r
SaratogaHouses2 %>% ggplot(aes(x=price/age)) + geom_histogram(bins=10) + facet_wrap(~centralAir)
```



```
SaratogaHouses2 %>% ggplot(mapping=aes(x=factor(rooms), y=price)) + geom_boxplot()
```



```
SaratogaHouses2 %>% arrange(price) %>% head(5)
```

```
##   price lotSize age landValue livingArea pctCollege bedrooms fireplaces
## 1  5000    0.29   4     35800       1700         63        3          1
## 2 10300    0.16  20     15700        912         54        2          1
## 3 10300    0.16  20     15700        912         54        2          1
## 4 20000    0.52  59      8000        936         20        2          0
## 5 25000    0.21  75       900        920         44        2          0
##   bathrooms rooms  heating     fuel             sewer waterfront
## 1       2.5     6  hot air      gas public/commercial         No
## 2       1.5     4 electric electric public/commercial         No
## 3       1.5     6 electric electric public/commercial         No
## 4       1.0     4  hot air      oil            septic         No
## 5       1.0     6  hot air      oil            septic         No
##   newConstruction centralAir price/age
## 1              No        Yes 1250.0000
## 2              No         No  515.0000
## 3              No         No  515.0000
## 4              No         No  338.9831
## 5              No         No  333.3333
```