Lab 3: Displaying Multivariate Data

Spring 2018 - Multivariate Data Analysis

Variables

- Variable : characteristic or property that is possible to measure
  - p: the number of variables
  - n: the number of observations
- Type of variables
  - Categorical variable
    nominal/ordinal
  - Continuous variable
  - Discrete variable
- How variables are used in data analysis
  - Outcome variable / dependent variable
  - Explanatory variable / predictor / independent variable
- Outliers
  - Robust methods : insensitive to departures from underlying model assumptions

Example: tipping data

- Data description
  - Food server's tips in restaurants may be influenced by many factors including the nature of the restaurant, size of the party, table locations in the restaurant,... To make appropriate assignments (which tables the food server waits on) for the food servers, restaurant managers need to know what these factors are.
  - In one restaurant, a food server recorded the following data on all customers he had served during a interval of two and a half months in early 1990, resulting in observations on 244 dining parties
- Variables
  - TOTBILL : Total bill, including tax, in dollars
  - TIP : Tip in dollars
  - SEX : Sex of person paying bill (0=male, 1=female)
  - SMOKER : Smoker in party (0=No, 1=Yes)
  - DAY : 3=Thur, 4=Fri, 5=Sat, 6=Sun
  - TIME : 0=day, 1=night
  - SIZE : Size of the party

```
tips<-read.csv("./data/Tipping.csv")
summary(tips)
```

```
##       OBS             TOTBILL           TIP             SEX      SMOK
##  Min.   :  1.00   Min.   : 3.07   Min.   : 1.000   Female: 87   No :
##  1st Qu.: 61.75   1st Qu.:13.35   1st Qu.: 2.000   Male  :157   Yes:
##  Median :122.50   Median :17.80   Median : 2.900
##  Mean   :122.50   Mean   :19.79   Mean   : 2.998
##  3rd Qu.:183.25   3rd Qu.:24.13   3rd Qu.: 3.562
##  Max.   :244.00   Max.   :50.81   Max.   :10.000
##    DAY         TIME          SIZE          TIPRATE
##  Fri :19   Day  : 68   Min.   :1.00   Min.   : 3.56
##  Sat :87   Night:176   1st Qu.:2.00   1st Qu.:12.91
##  Sun :76               Median :2.00   Median :15.47
##  Thur:62               Mean   :2.57   Mean   :16.08
##                        3rd Qu.:3.00   3rd Qu.:19.15
##                        Max.   :6.00   Max.   :71.03
```

Example: titanic data

- Data description
  - Information on the fate of 2201 passengers on the fatal maiden voyage of the ocean liner 'Titanic'
  - in R data(Titanic) has 4-dim array resulting from cross-tabulating 2201 observations on 4 variables
- Variables
  - class : 1st, 2nd, 3rd, crew
  - age : adult, child
  - sex : female, male
  - survived : no, yes

```r
titanic<-read.csv("./data/titanic.csv")
head(titanic)
```

```
##   class   age  sex survived
## 1   1st adult male      yes
## 2   1st adult male      yes
## 3   1st adult male      yes
## 4   1st adult male      yes
## 5   1st adult male      yes
## 6   1st adult male      yes
```

```r
summary(titanic)
```

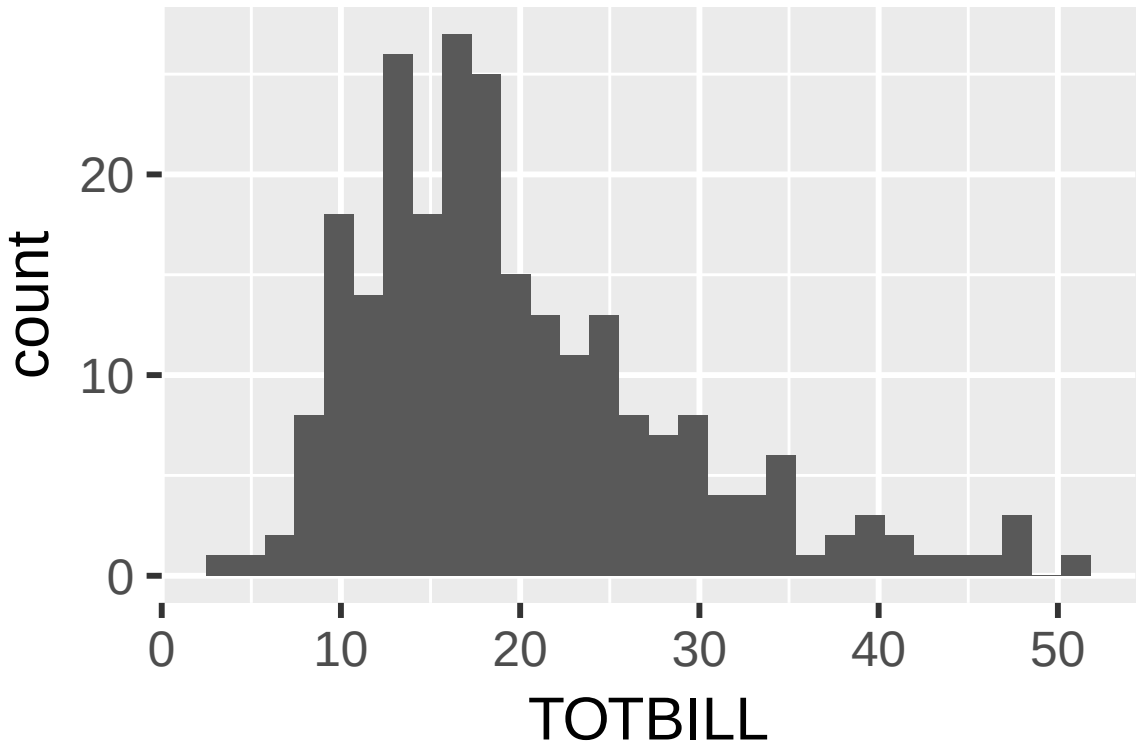
```
##   class        age           sex       survived
##  1st :325   adult:2092   female: 470   no :1490
##  2nd :285   child: 109   male  :1731   yes: 711
##  3rd :706
##  crew:885
```

one continuous variable (1)

- Histogram

```
library(ggplot2)
ggplot(data = tips, aes(x=TOTBILL))+geom_histogram()
```

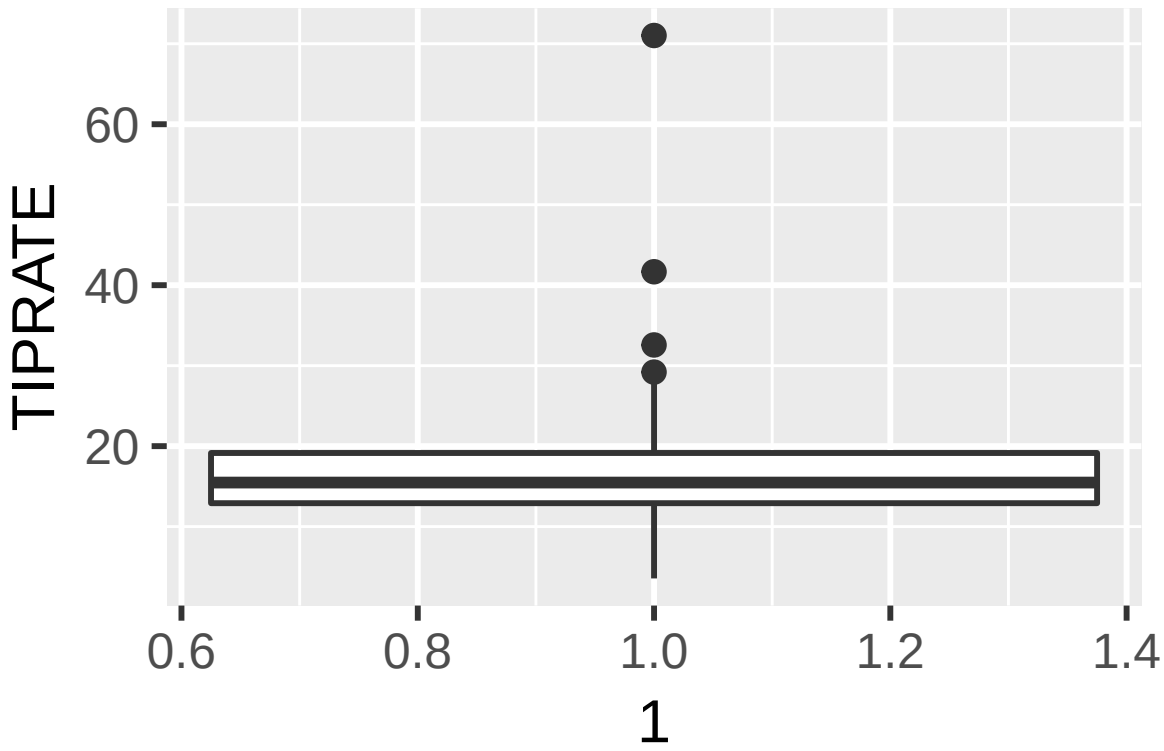
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



one continuous variable (2)

- Boxplot

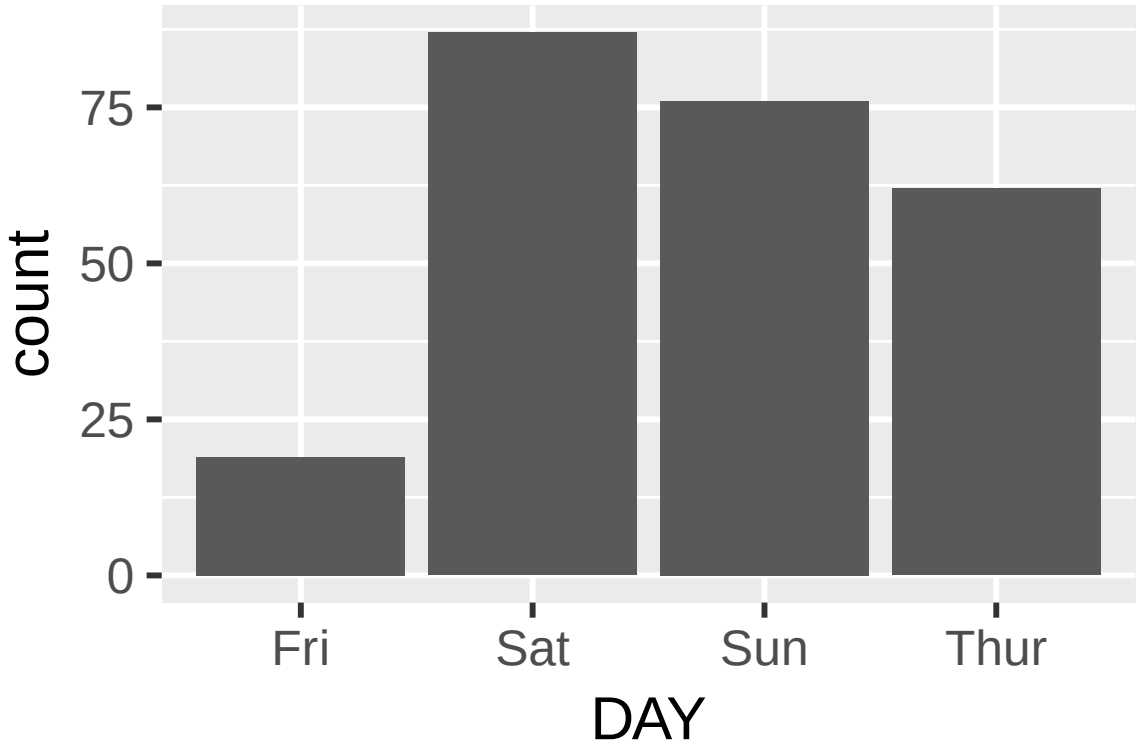
```
ggplot(data = tips, aes(x=1,y=TIPRATE))+geom_boxplot()
```



One categorical variable

- Barchart

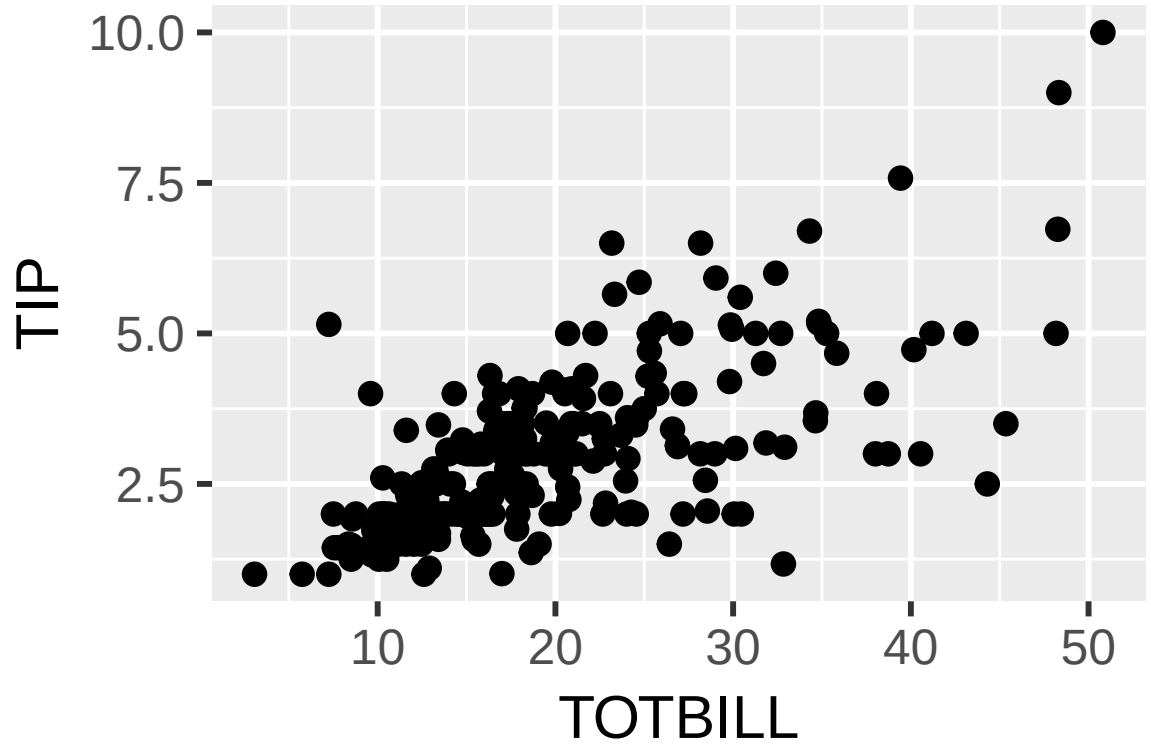
```
ggplot(data = tips, aes(x=DAY))+geom_bar()
```



Two or more continuous variables (1)

- Scatterplot

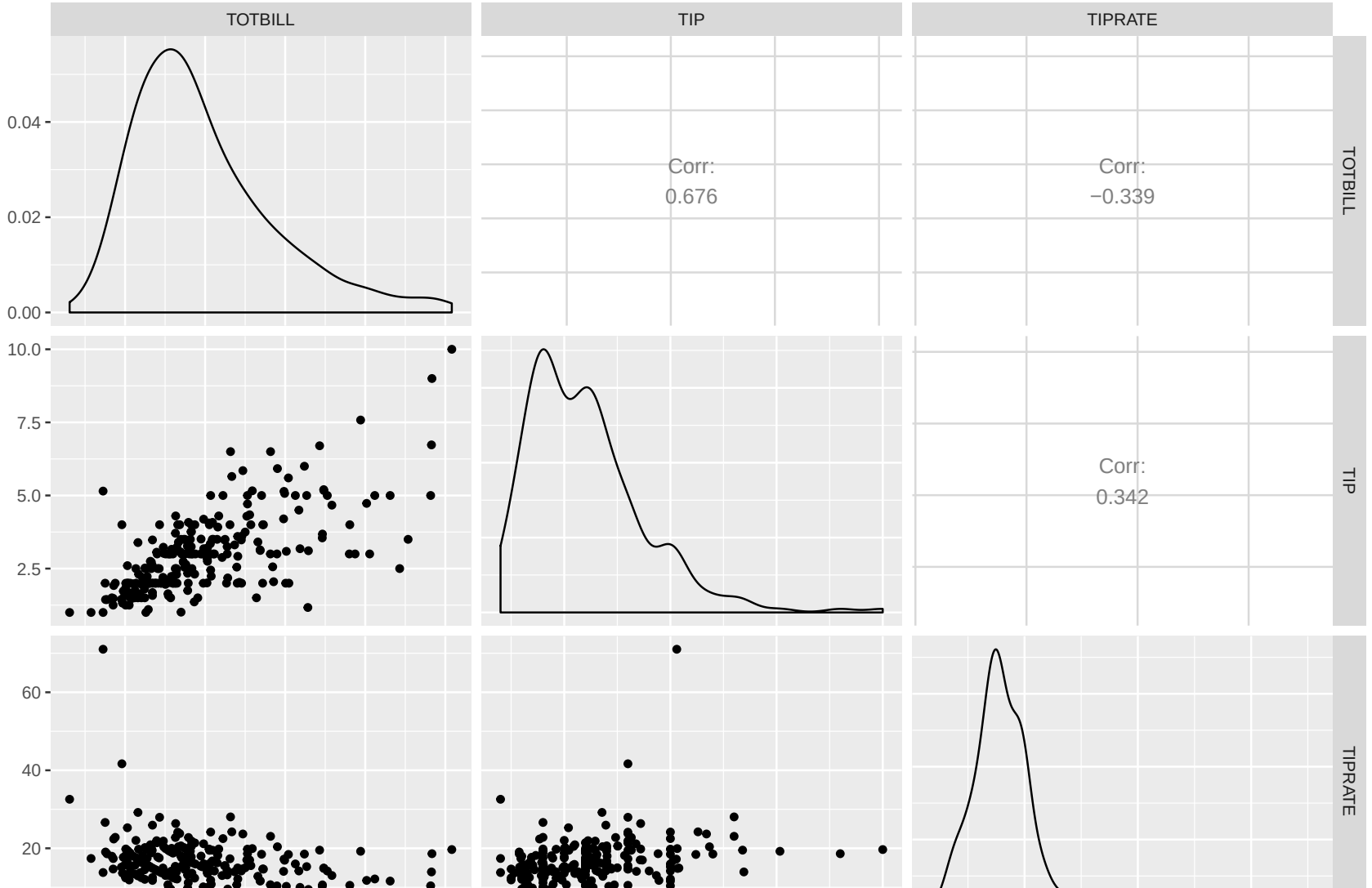
```
ggplot(data = tips, aes(x=TOTBILL,y=TIP))+geom_point()
```



Two or more continuous variables (2)

- Scatterplot matrix

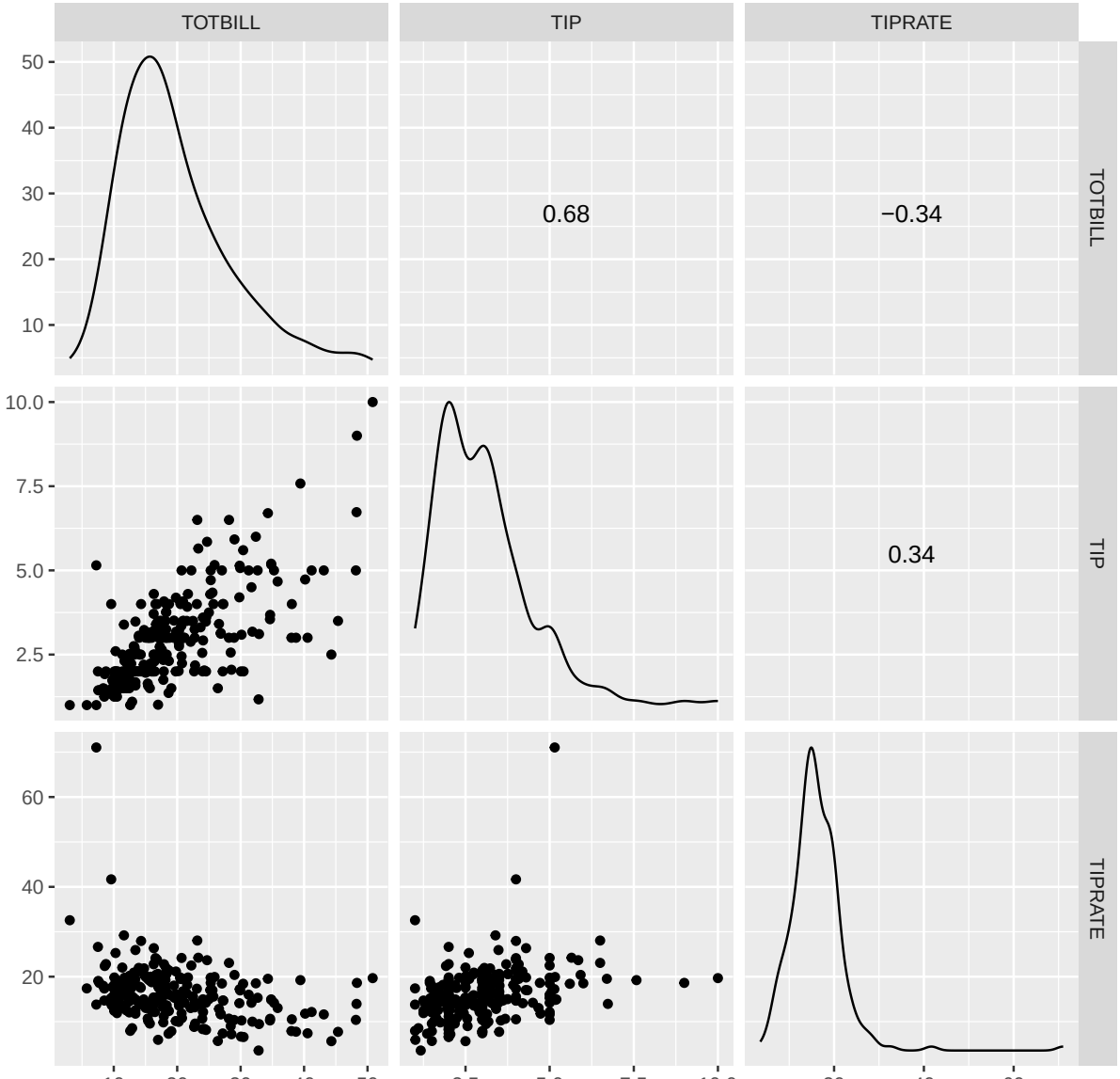
```
library(GGally)
ggpairs(tips[,c(2,3,9)])
```



Two or more continuous variables (3)

- Scatterplot matrix

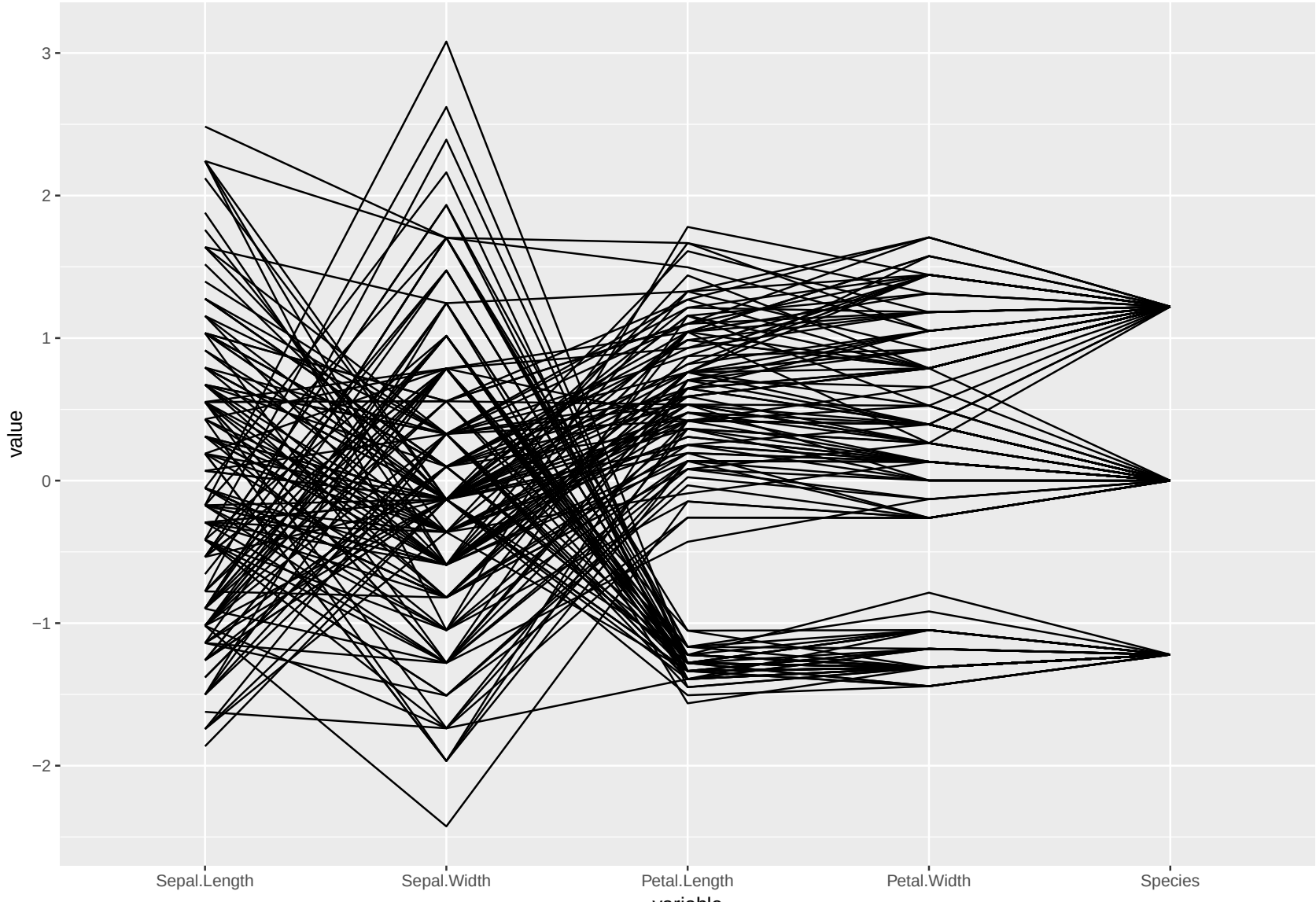
```
ggscatmat(tips[,c(2,3,9)])
```



Two or more continuous variables (4)

- Parallel coordinate plot

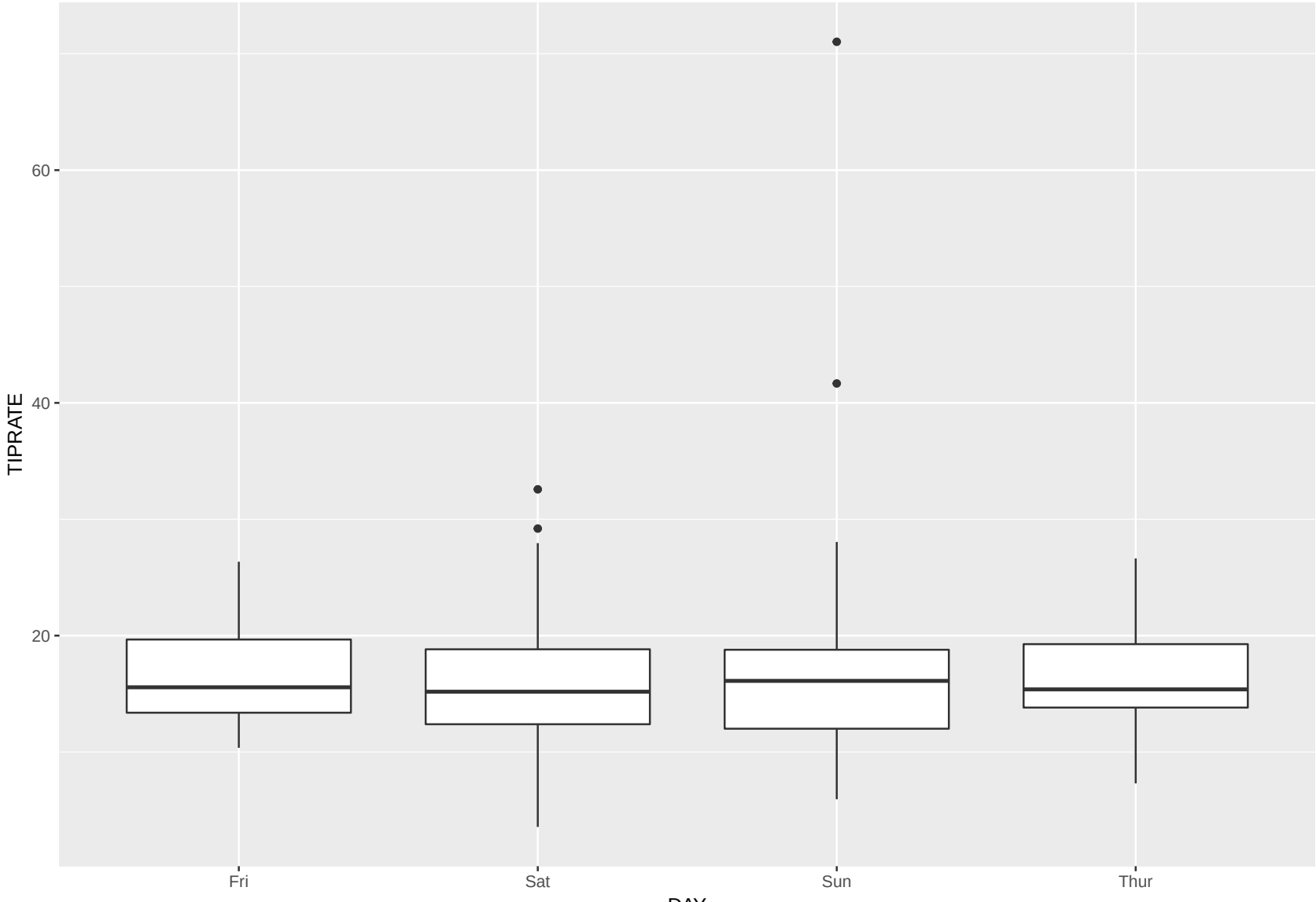
```
ggparcoord(iris)
```



One continuous variable + one categorical variable

- parallel boxplot

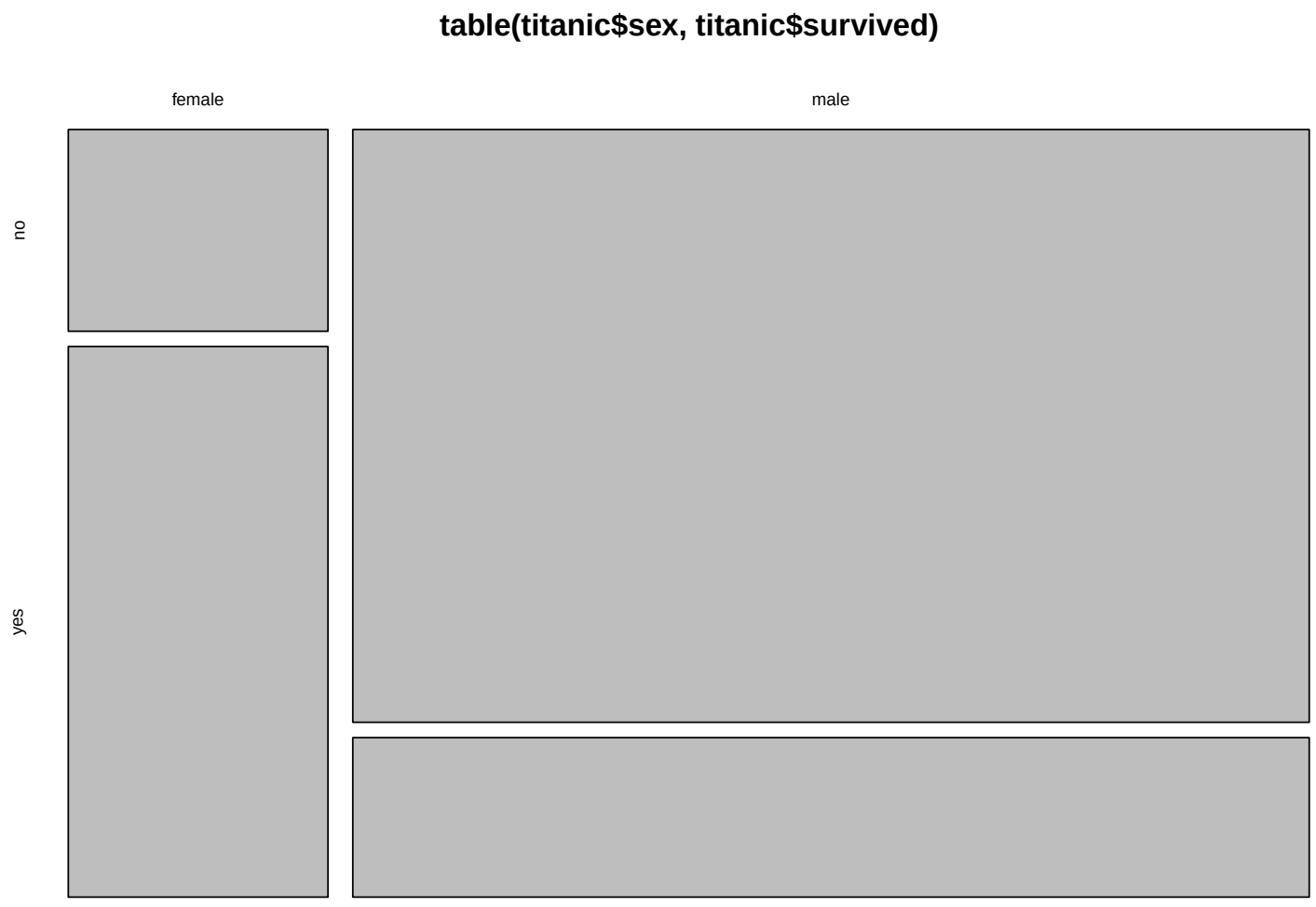
```
ggplot(data = tips, aes(x=DAY,y=TIPRATE))+geom_boxplot()
```



Two or more categorical variables (1)

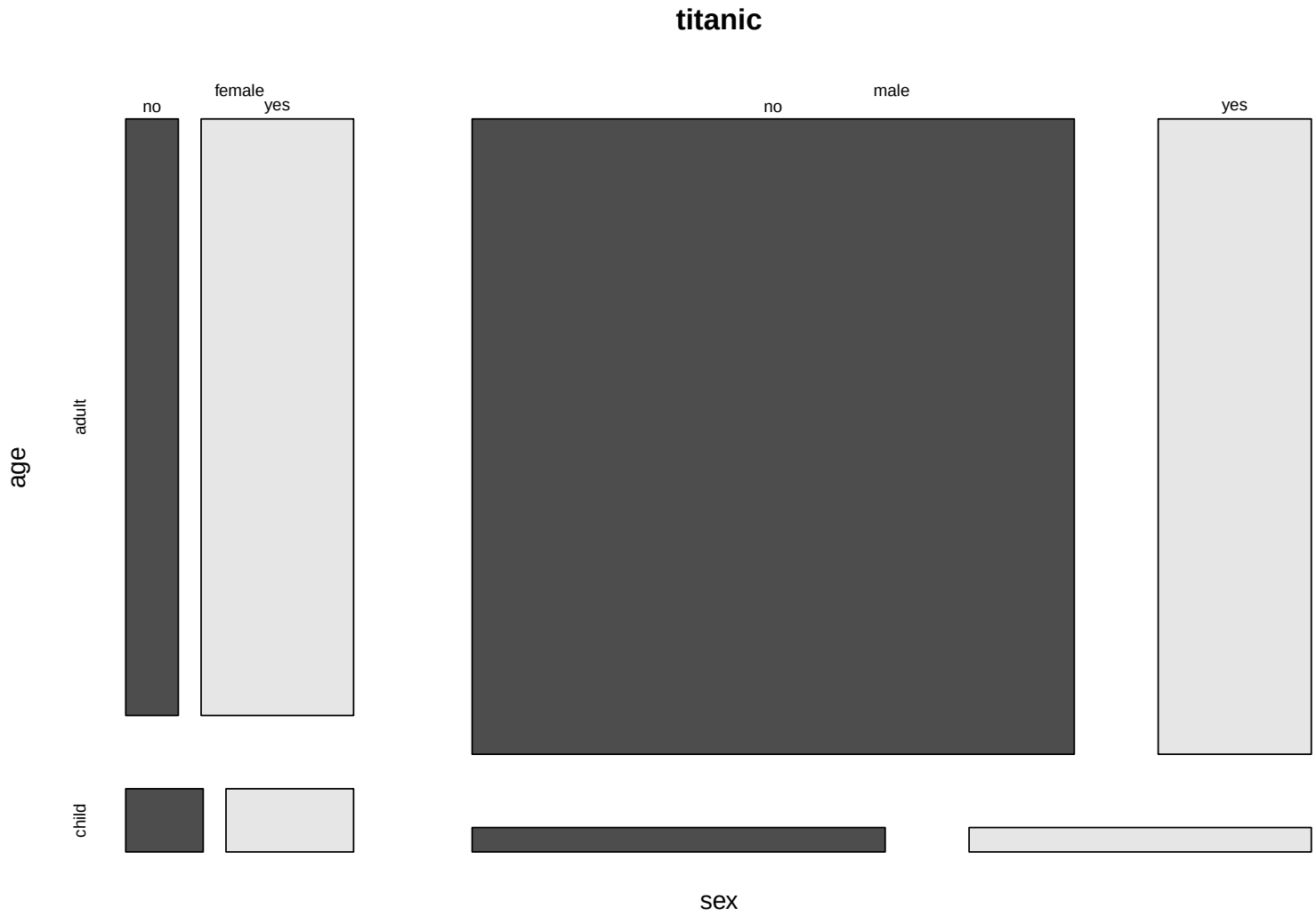
- Mosaic plot

```
mosaicplot(table(titanic$sex,titanic$survived))
```



Two or more categorical variables (2)

```
mosaicplot(~sex+age+survived,data = titanic,color=TRUE)
```



Dynamic graphics

- GGobi : www.ggobi.org
  - Open source visualization program for exploring high-dimensional data
  - Dynamic and interactive graphics
- Example: lizard data
  - A zoologist obtained measurements on 25 lizards know scientifically as Cophosaurus texanus
  - X1 : Mass(grams)
  - X2 : SVL(Snout-vent Length, mm)
  - X3 : HLS(hind limb span, mm)

```r
lizard<-read.csv("./data/lizard.csv")
head(lizard)
```

```
##   Lizard   Mass  SVL   HLS sex
## 1      1  5.526 59.0 113.5   f
## 2      2 10.401 75.0 142.0   m
## 3      3  9.213 69.0 124.0   f
## 4      4  8.953 67.5 125.0   f
## 5      5  7.063 62.0 129.5   m
## 6      6  6.610 62.0 123.0   f
```

```r
summary(lizard)
```

```
##      Lizard          Mass             SVL             HLS          sex
##  Min.   : 1   Min.   : 2.447   Min.   :47.0   Min.   : 97.0   f:12
##  1st Qu.: 7   1st Qu.: 6.978   1st Qu.:63.0   1st Qu.:118.0   m:13
##  Median :13   Median : 8.953   Median :68.0   Median :129.5
##  Mean   :13   Mean   : 8.687   Mean   :68.4   Mean   :129.3
##  3rd Qu.:19   3rd Qu.:10.091   3rd Qu.:74.0   3rd Qu.:137.0
##  Max.   :25   Max.   :15.493   Max.   :86.5   Max.   :162.0
```

```r
#library(rggobi)
#ggobi(lizard)
```

univariate vs. multivariate (1)

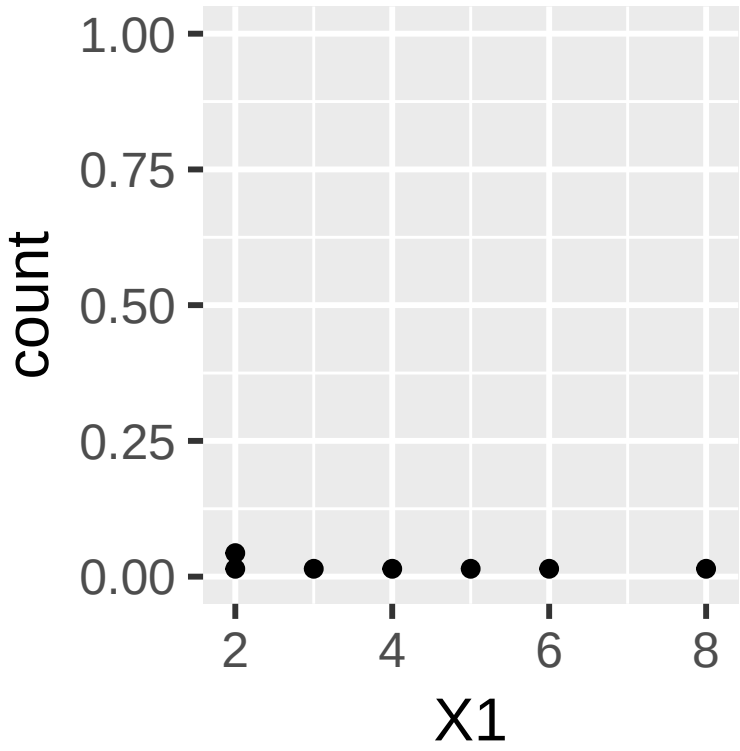
- Example 1

X1	3	4	2	6	8	2	5
X2	5	5.5	4	7	10	5	7.5

```
X1<-c(3,4,2,6,8,2,5)
X2<-c(5,5.5,4,7,10,5,7.5)
ex1.data<-data.frame(X1=X1,X2=X2)
```

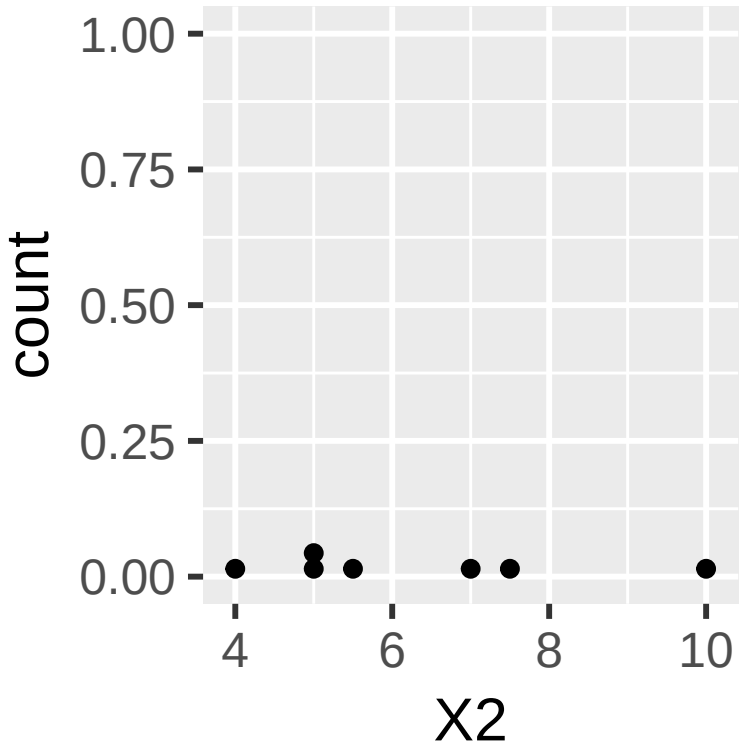
```
ggplot(ex1.data,aes(x=X1))+geom_dotplot()
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`
```

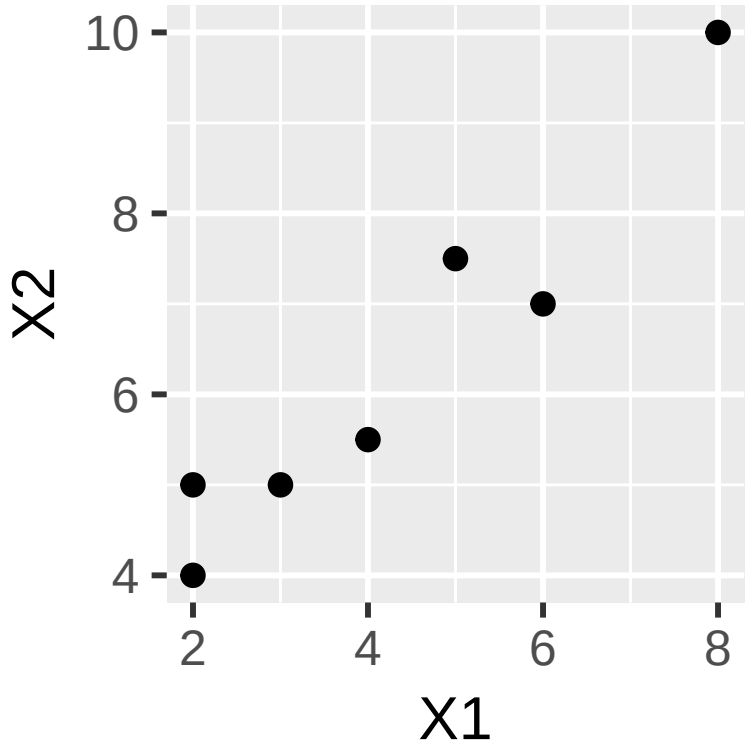


```
ggplot(ex1.data,aes(x=X2))+geom_dotplot()
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`
```



```
ggplot(ex1.data,aes(x=X1,y=X2))+geom_point()
```



univariate vs. multivariate (2)

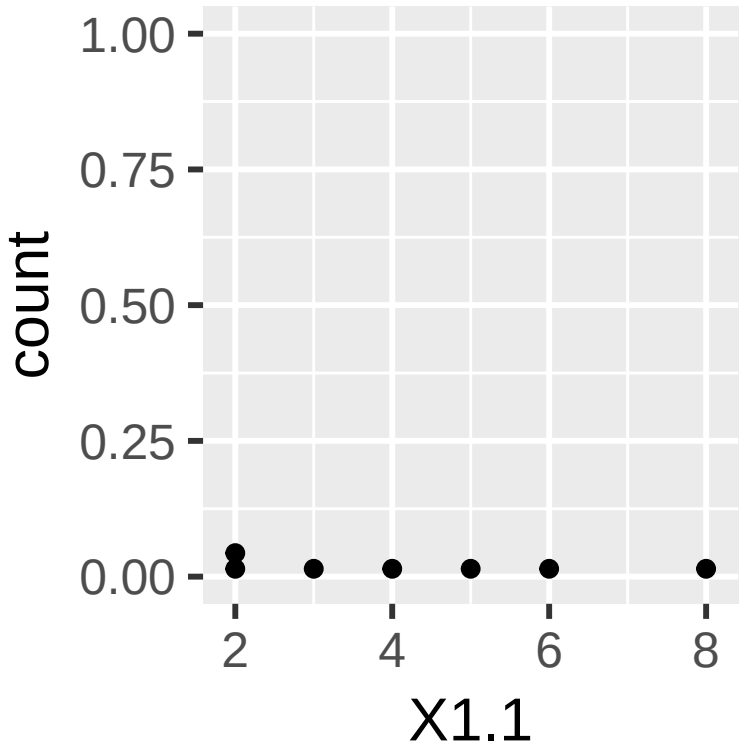
- Example 2

X1	5	4	6	2	2	8	3
X2	5	5.5	4	7	10	5	7.5

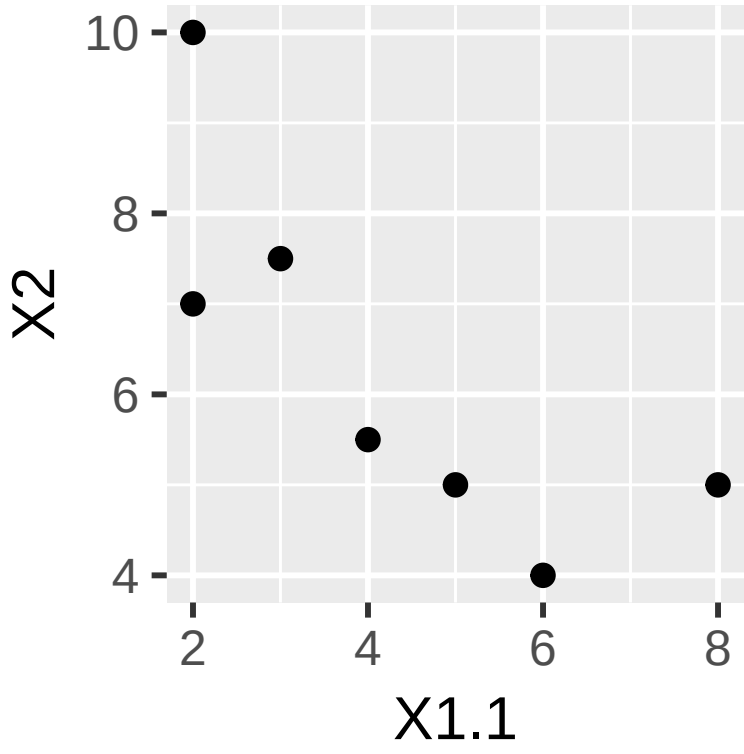
```
X1.1<-c(5,4,6,2,2,8,3)
X2<-c(5,5.5,4,7,10,5,7.5)
ex2.data<-data.frame(X1.1=X1.1,X2=X2)
```

```
ggplot(ex2.data,aes(x=X1.1))+geom_dotplot()
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`
```



```
ggplot(ex2.data,aes(x=X1.1,y=X2))+geom_point()
```



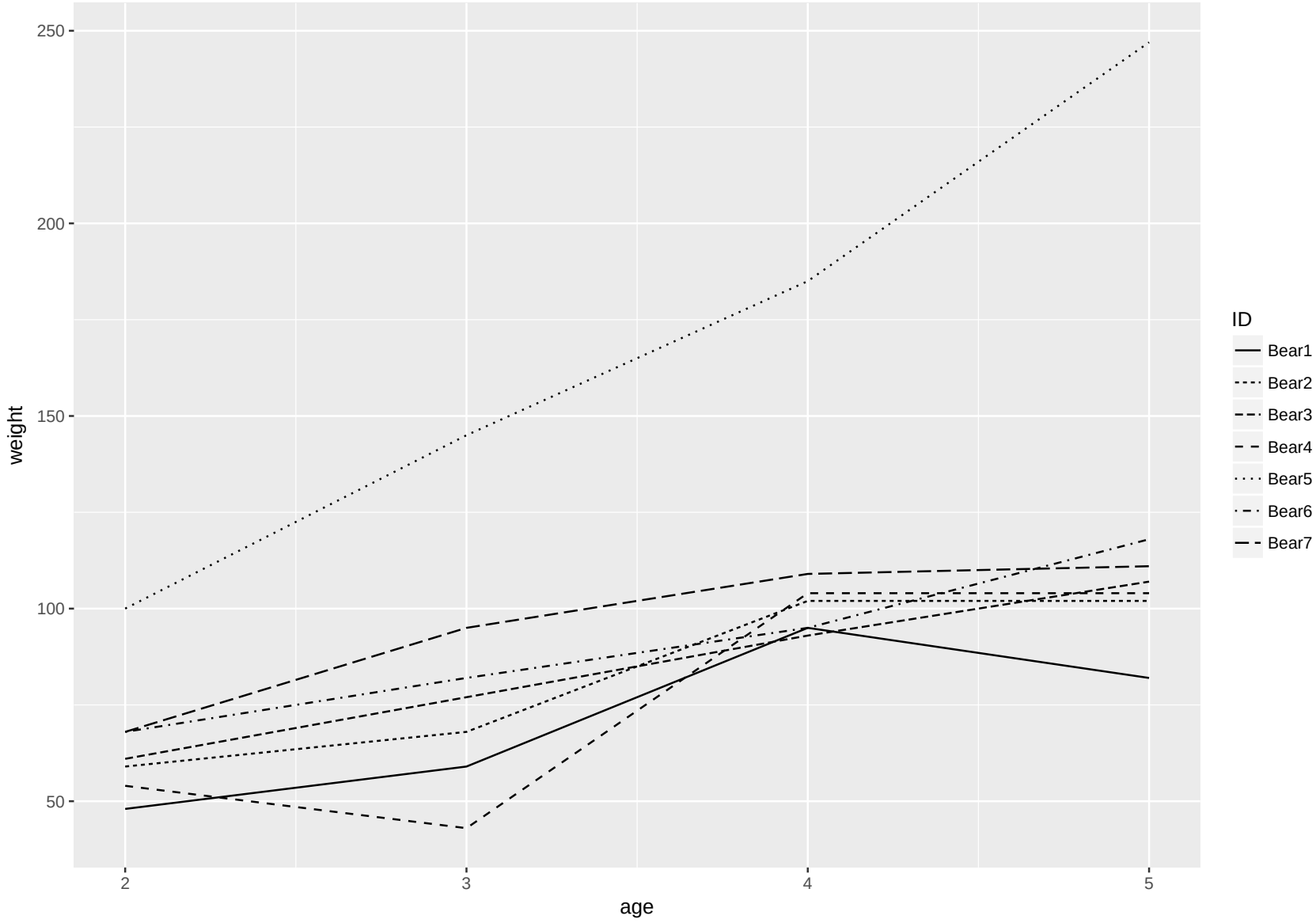
growth curve

- Example: Grizzly Bear
  - The Alaska Fish and Game department monitors grizzly bears with the goal of maintaining a healthy population
  - Bears are shot with a dart to include sleep and weighted on a scale hanging from a tripod
  - ID/ age / weight / height

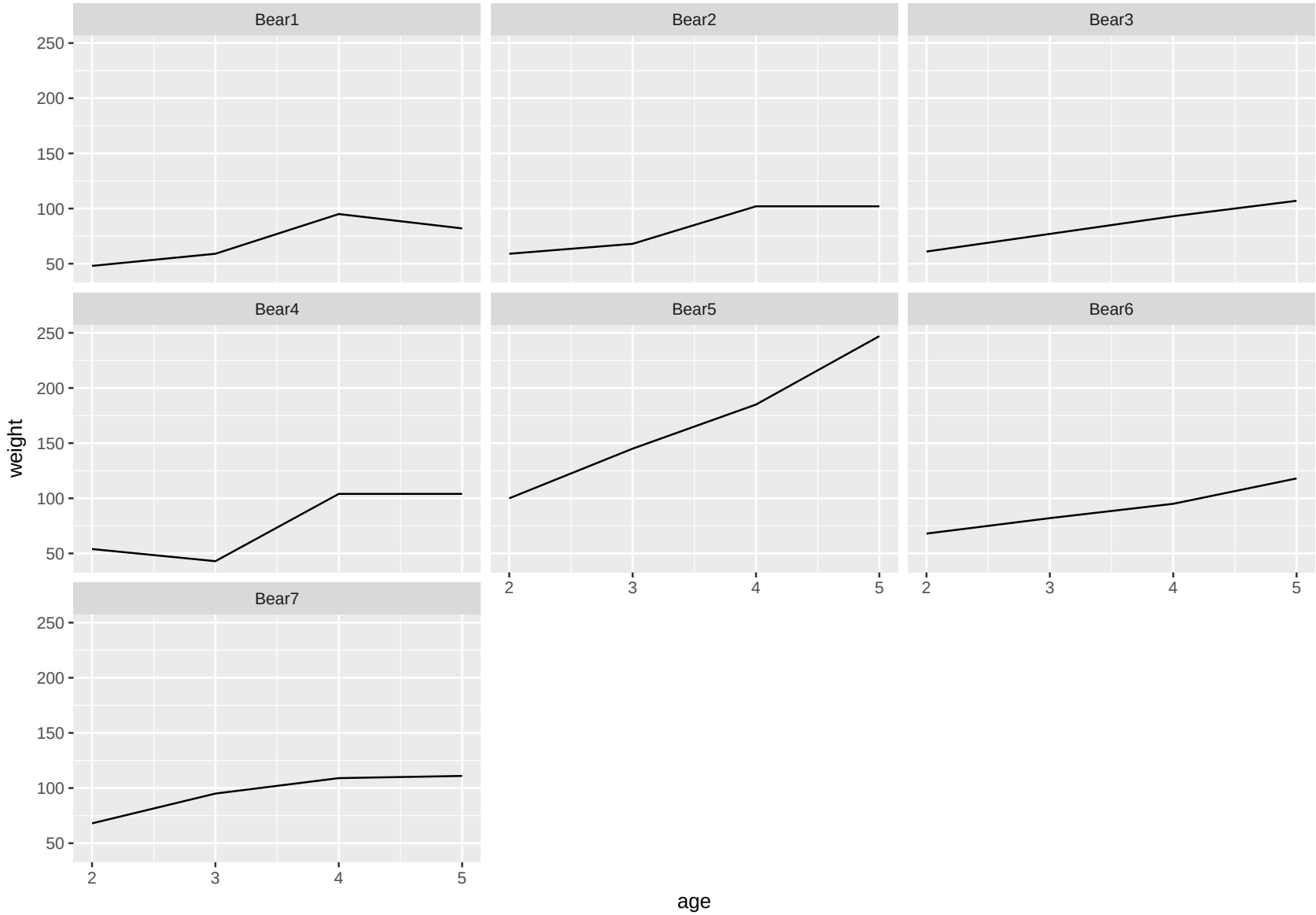
```
bear<-read.csv("./data/bear.csv")
summary(bear)
```

```
##       ID          age              weight           height
##  Bear1:4   Min.   :2.00   Min.   : 43.00   Min.   :139.0
##  Bear2:4   1st Qu.:2.75   1st Qu.: 68.00   1st Qu.:149.0
##  Bear3:4   Median :3.50   Median : 95.00   Median :168.0
##  Bear4:4   Mean   :3.50   Mean   : 95.75   Mean   :163.2
##  Bear5:4   3rd Qu.:4.25   3rd Qu.:104.75   3rd Qu.:175.0
##  Bear6:4   Max.   :5.00   Max.   :247.00   Max.   :189.0
##  Bear7:4
```

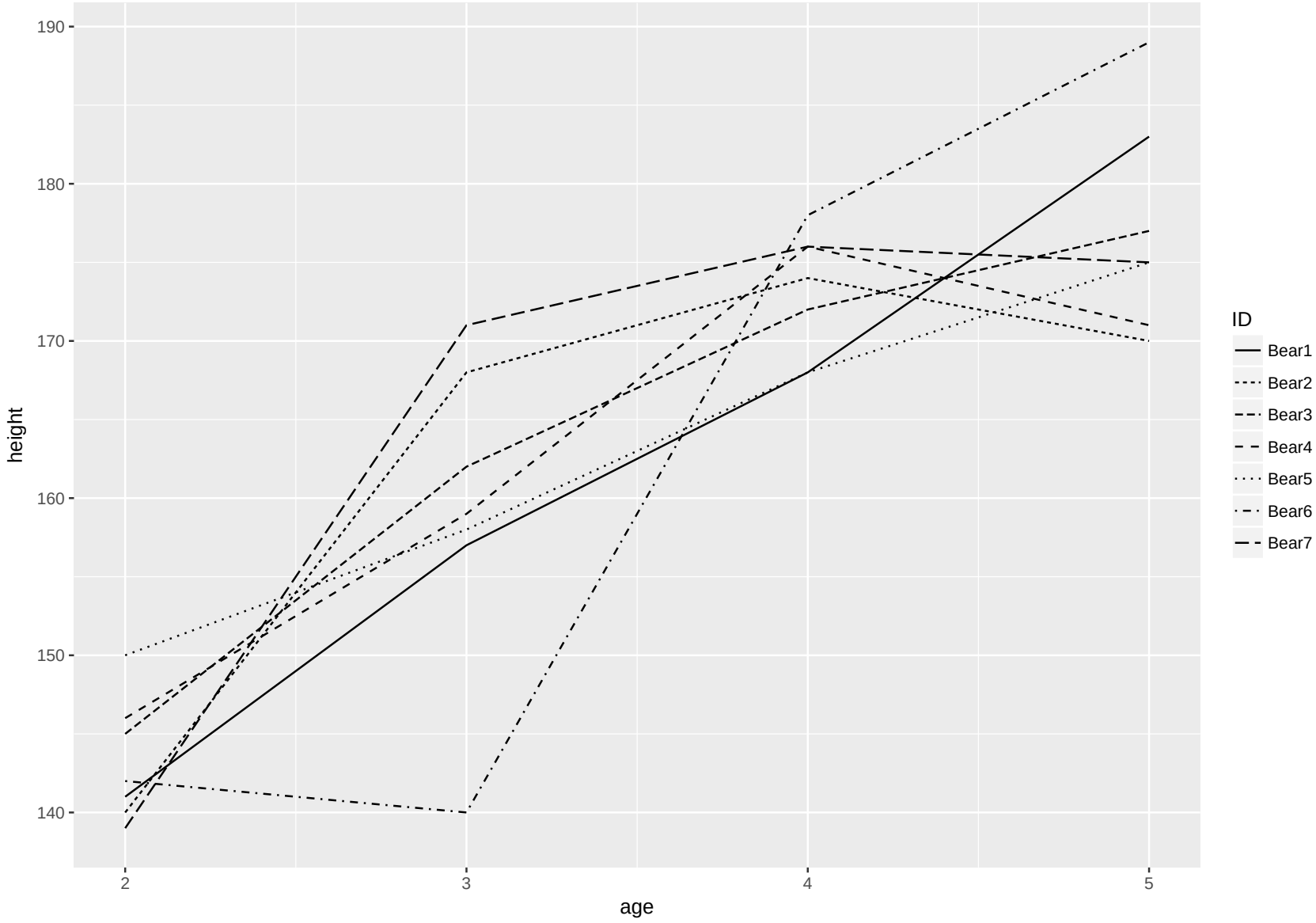
```
ggplot(bear,aes(x=age,y=weight,group=ID))+geom_line(aes(linetype=ID))
```



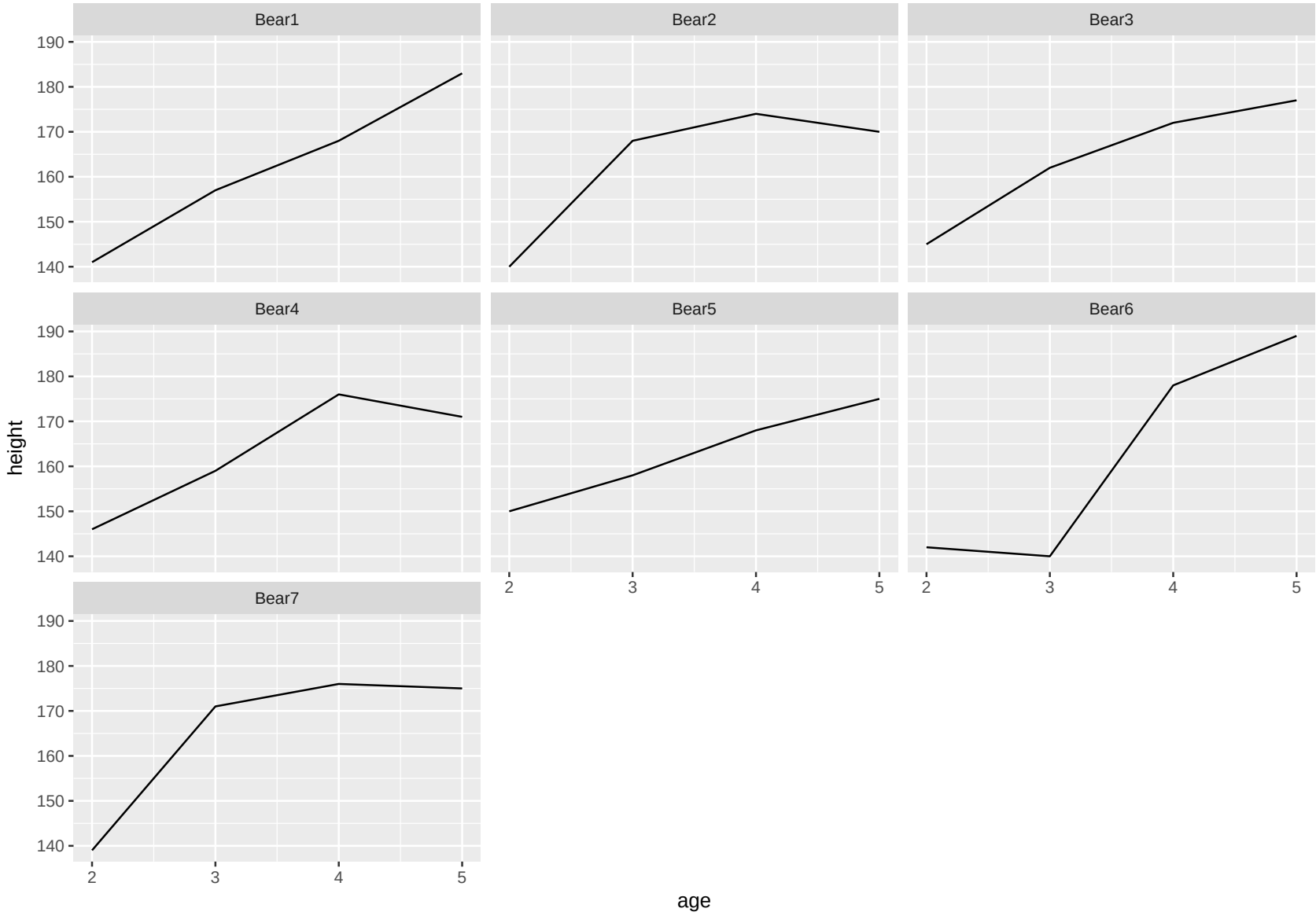
```
ggplot(bear,aes(x=age,y=weight,group=ID))+geom_line()+facet_wrap(~ID)
```



```
ggplot(bear,aes(x=age,y=height,group=ID))+geom_line(aes(linetype=ID))
```



```
ggplot(bear,aes(x=age,y=height,group=ID))+geom_line()+facet_wrap(~ID)
```



Chernoff faces

- Example: Longley's Economic Regression Data (data(longley) in R)
  - A macroeconomic data set which provides a well-known example for a highly collinear regression
  - 7 economical variables, observed yearly from 1947 to 1962 (n=16)
  - GNP.deflator: GNP implicit price deflator (1954=100)
  - GNP: Gross National Product.
  - Unemployed: number of unemployed.
  - Armed.Forces: number of people in the armed forces.
  - Population: noninstitutionalized population $\geq$ 14 years of age.
  - Year: the year (time).
  - Employed: number of people employed.

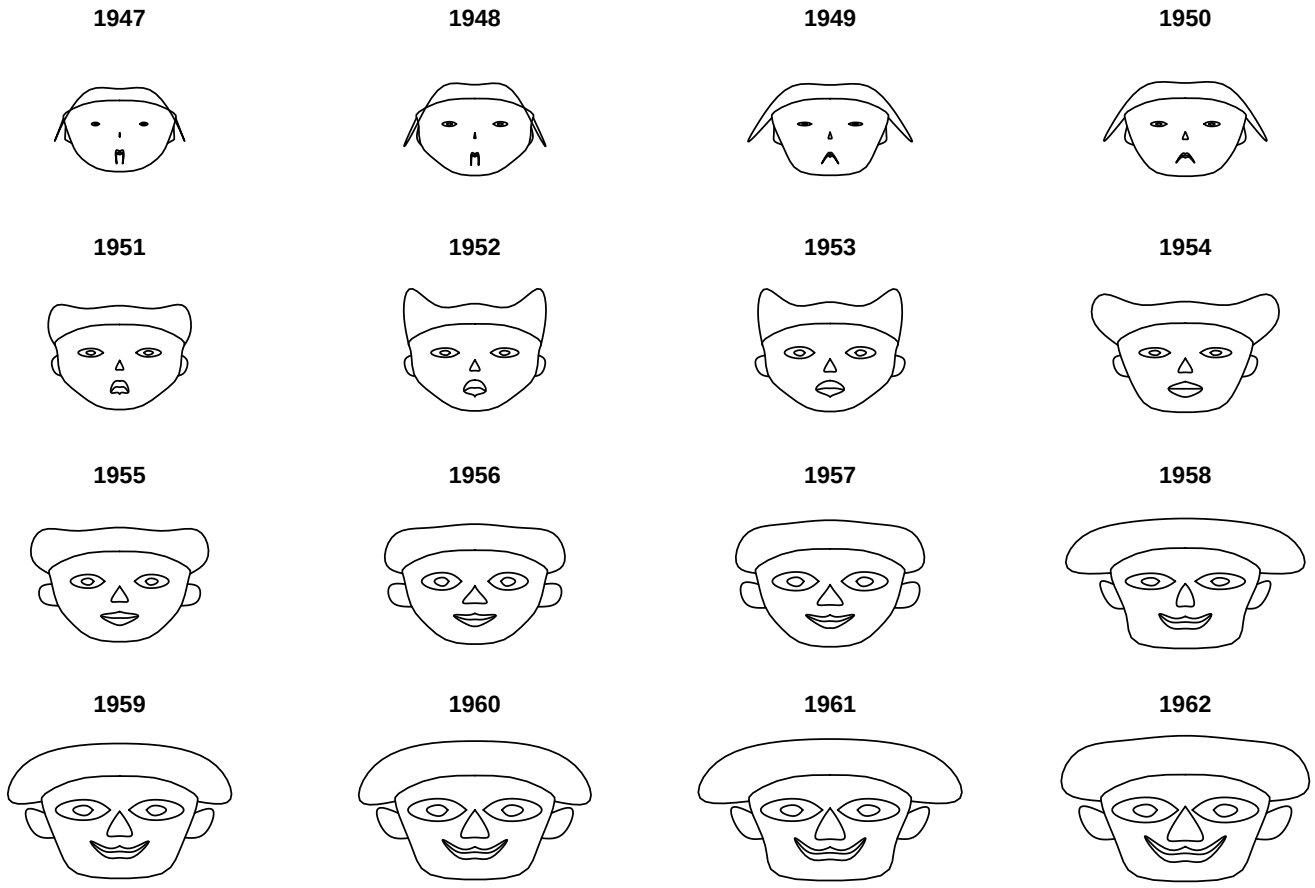
```r
library(aplpack)
```

```
## Loading required package: tcltk
```

```r
data(longley)
head(longley)
```

```
##      GNP.deflator     GNP Unemployed Armed.Forces Population Year Em
## 1947         83.0 234.289      235.6        159.0    107.608 1947
## 1948         88.5 259.426      232.5        145.6    108.632 1948
## 1949         88.2 258.054      368.2        161.6    109.773 1949
## 1950         89.5 284.599      335.1        165.0    110.929 1950
## 1951         96.2 328.975      209.9        309.9    112.075 1951
## 1952         98.1 346.999      193.2        359.4    113.270 1952
```

```
faces(longley,face.type=0) # face.type=1: color
```




```
## effect of variables:
##   modified item       Var
```

```
## effect of variables:
##  modified item       Var
##  "height of face   " "GNP.deflator"
##  "width of face    " "GNP"
##  "structure of face" "Unemployed"
##  "height of mouth  " "Armed.Forces"
##  "width of mouth   " "Population"
##  "smiling          " "Year"
##  "height of eyes   " "Employed"
##  "width of eyes    " "GNP.deflator"
##  "height of hair   " "GNP"
##  "width of hair   "  "Unemployed"
##  "style of hair   "  "Armed.Forces"
##  "height of nose  "  "Population"
##  "width of nose   "  "Year"
##  "width of ear    "  "Employed"
##  "height of ear   "  "GNP.deflator"
```

Star plot

```
stars(longley)
```