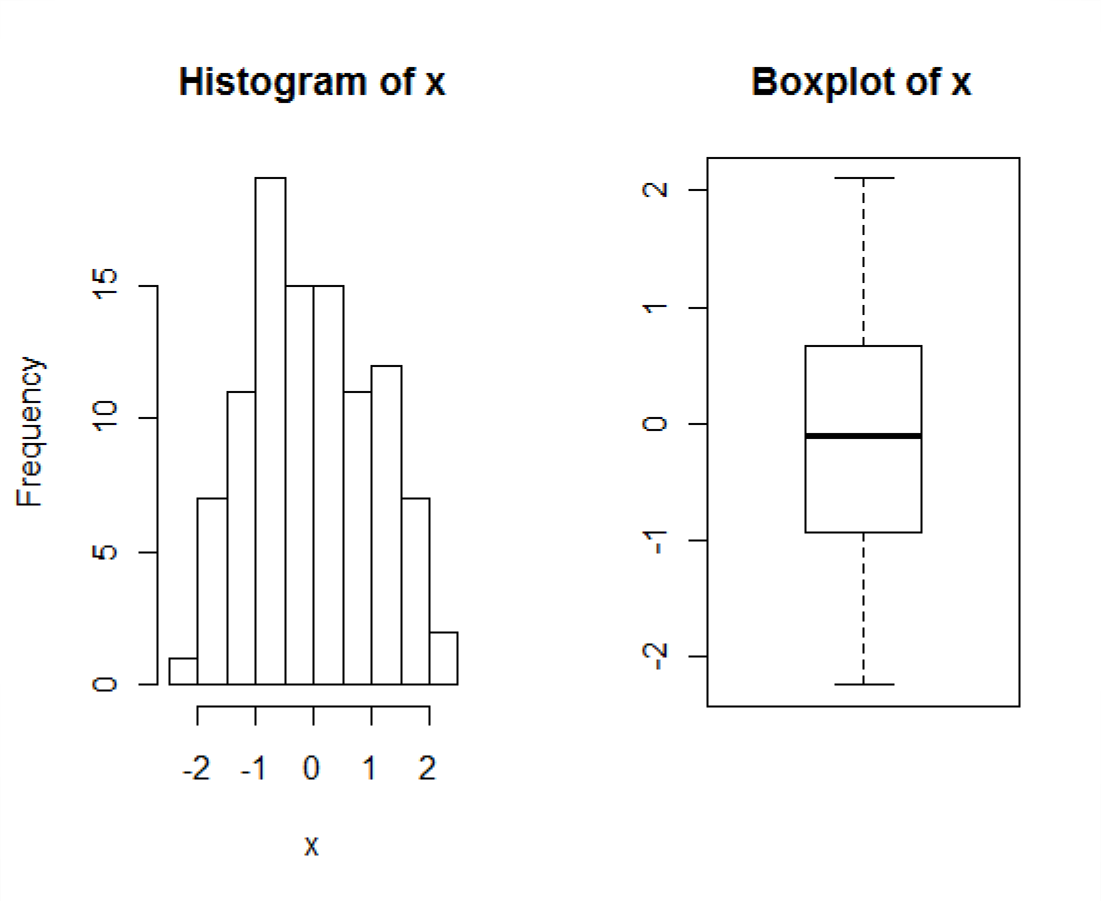
1. Survey
   (https://goo.gl/forms/0qy64czEQ9)
2. Office Hours @ 903 SSW
3. Email Subject: **[4415 MSI]**
4. Waiting List
5. Programming
6. CourseWorks → **Github**

Github.com/MRandomMax/MSI

MULTIVARIATE DISTRIBUTION & MULTIVARIATE GAUSSIAN

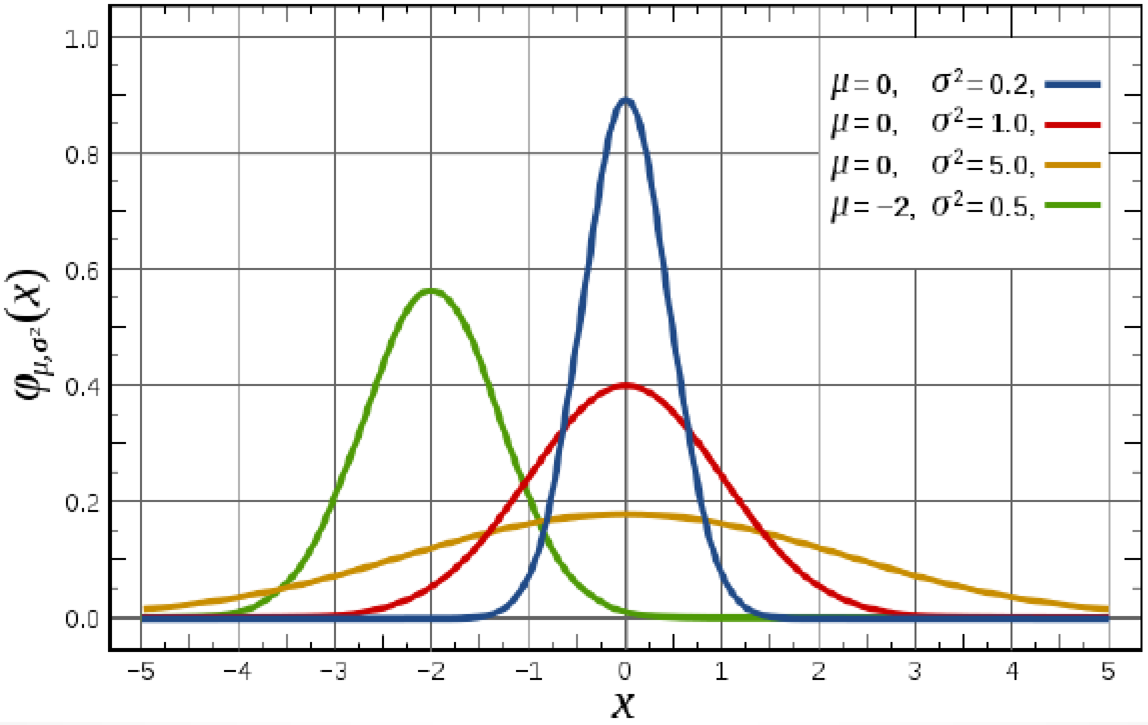
Mengqian LU

Visualization in 1D



Normal Distribution in 1D

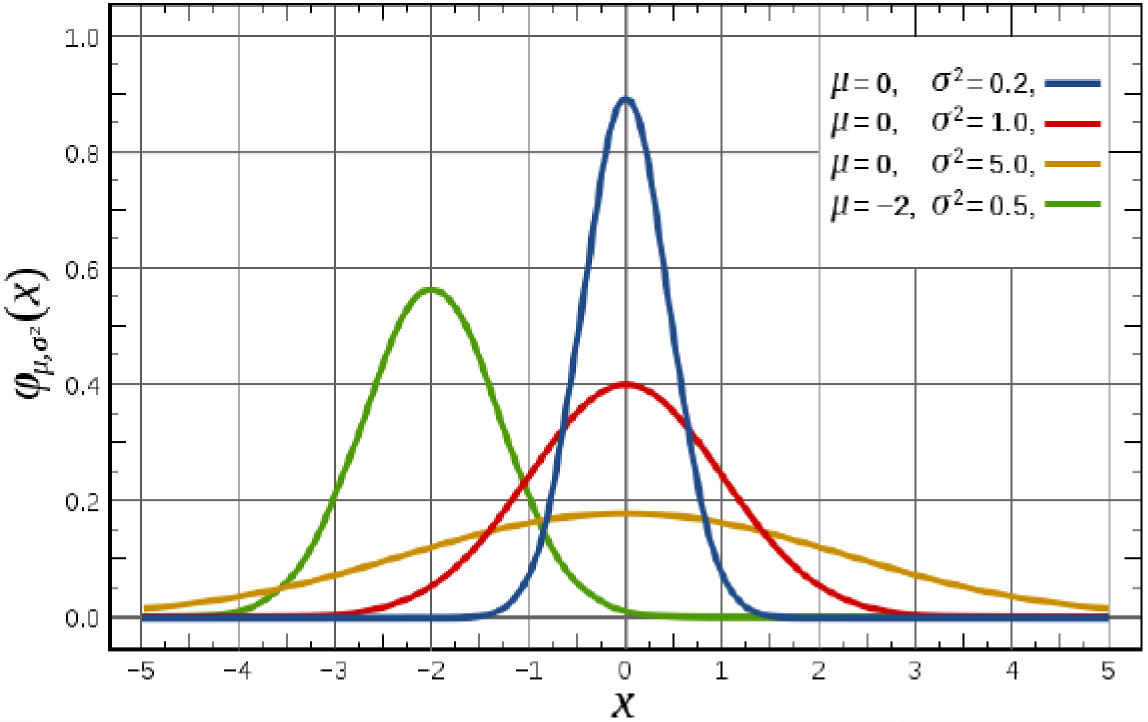
$$\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}\right)$$



Normal Distribution in 1D

(Mahalanobis Distance)^2

$$\varphi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \cdot \boxed{\frac{(x-\mu)^2}{\sigma^2}}\right)$$



Covariance & Correlation

Covariance: $Cov(X,Y) = E[(X - E[X])(Y - E[Y])] \quad \in [-\infty;\infty]$

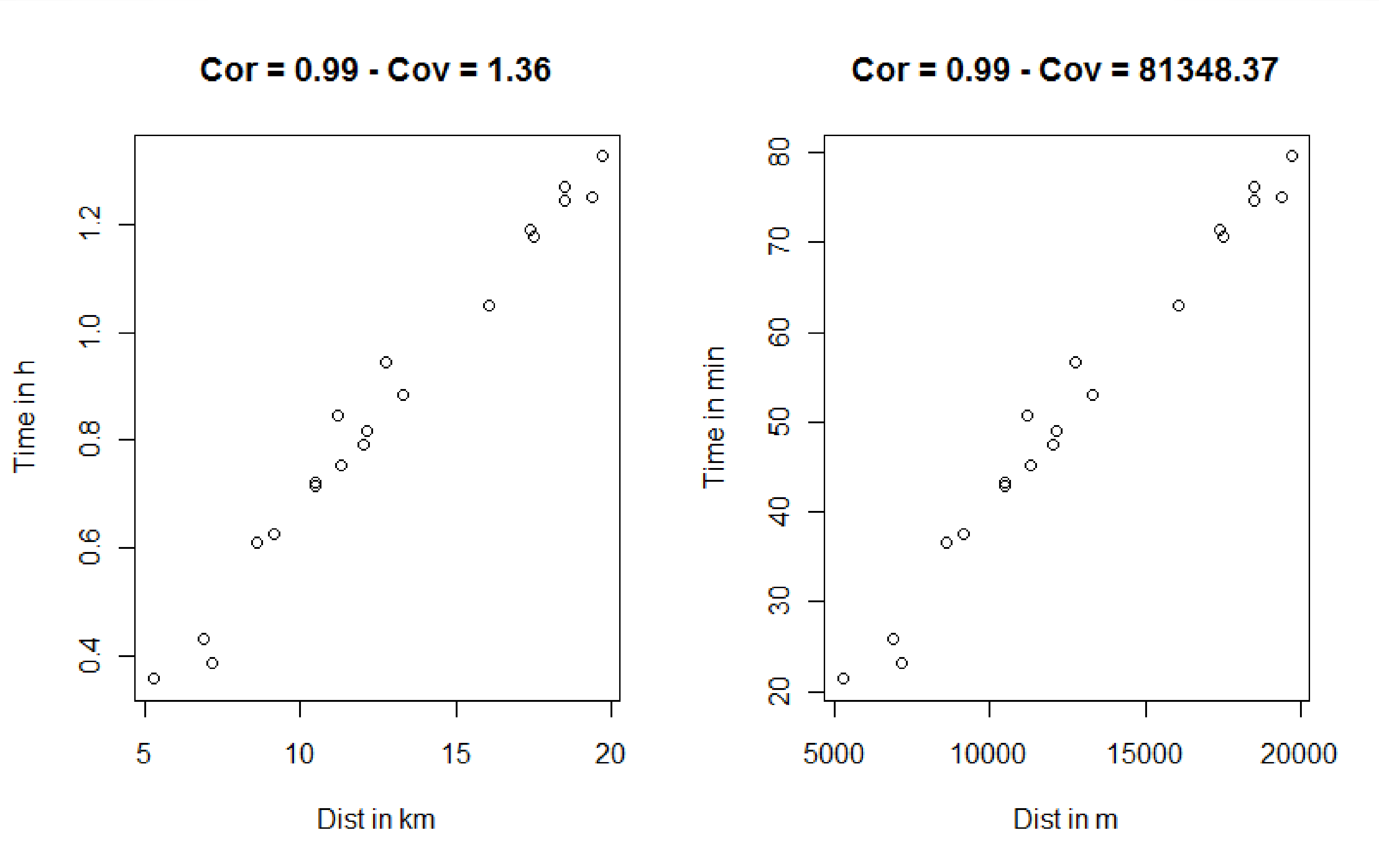
Correlation: $Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \quad \in [-1;1]$

Sample covariance: $\widehat{Cov}(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$

Sample correlation: $r_{xy} = \widehat{Cor}(x,y) = \frac{\widehat{Cov}(x,y)}{\hat{\sigma}_x \hat{\sigma}_y}$

Correlation is invariant to changes in units, covariance is not.

Correlation is scale invariant



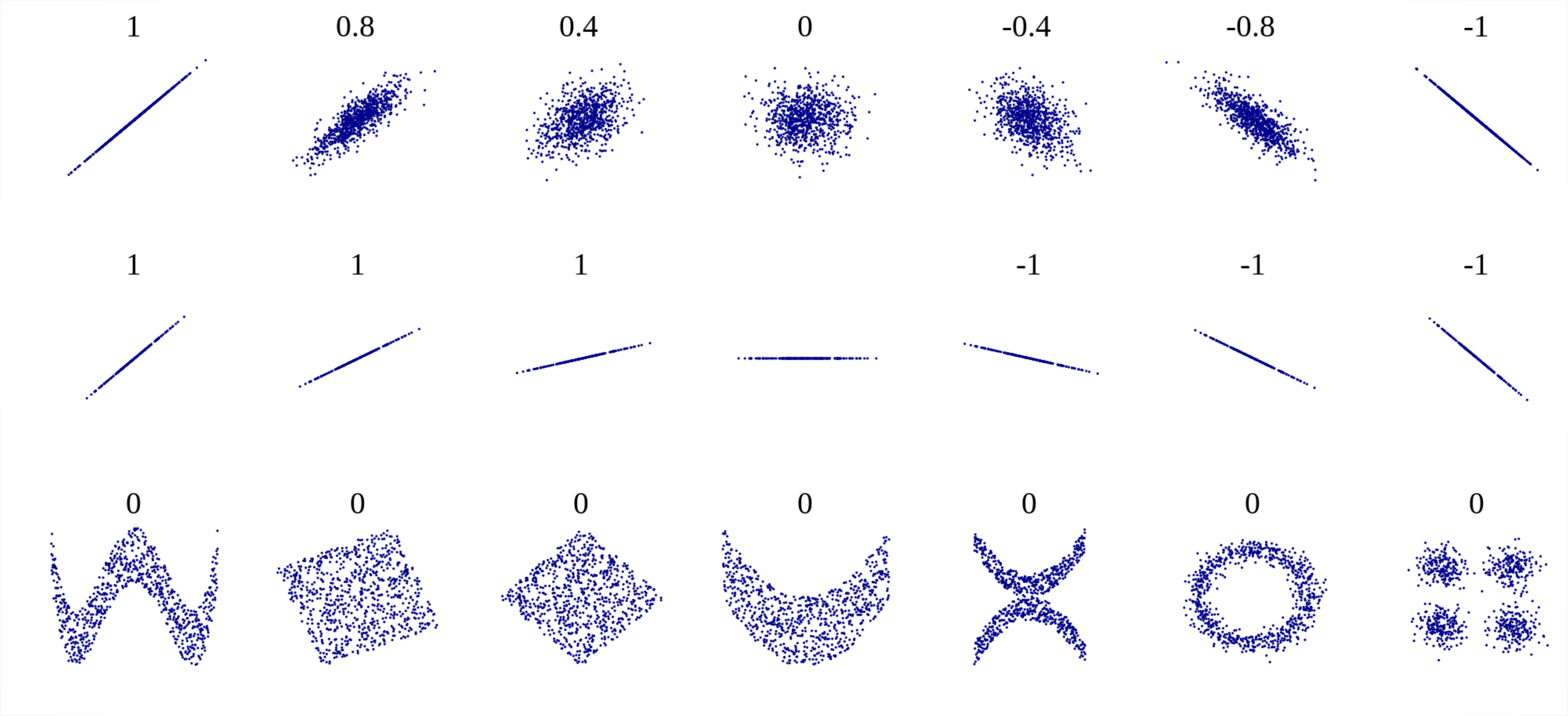
Q1: If correlation (x, y) is close to ONE, the slope is close to ONE?

Y/N, WHY?

Q2: If correlation (x, y) is close to ZERO, the relationship is weak, less hope for a model, at least not easy?

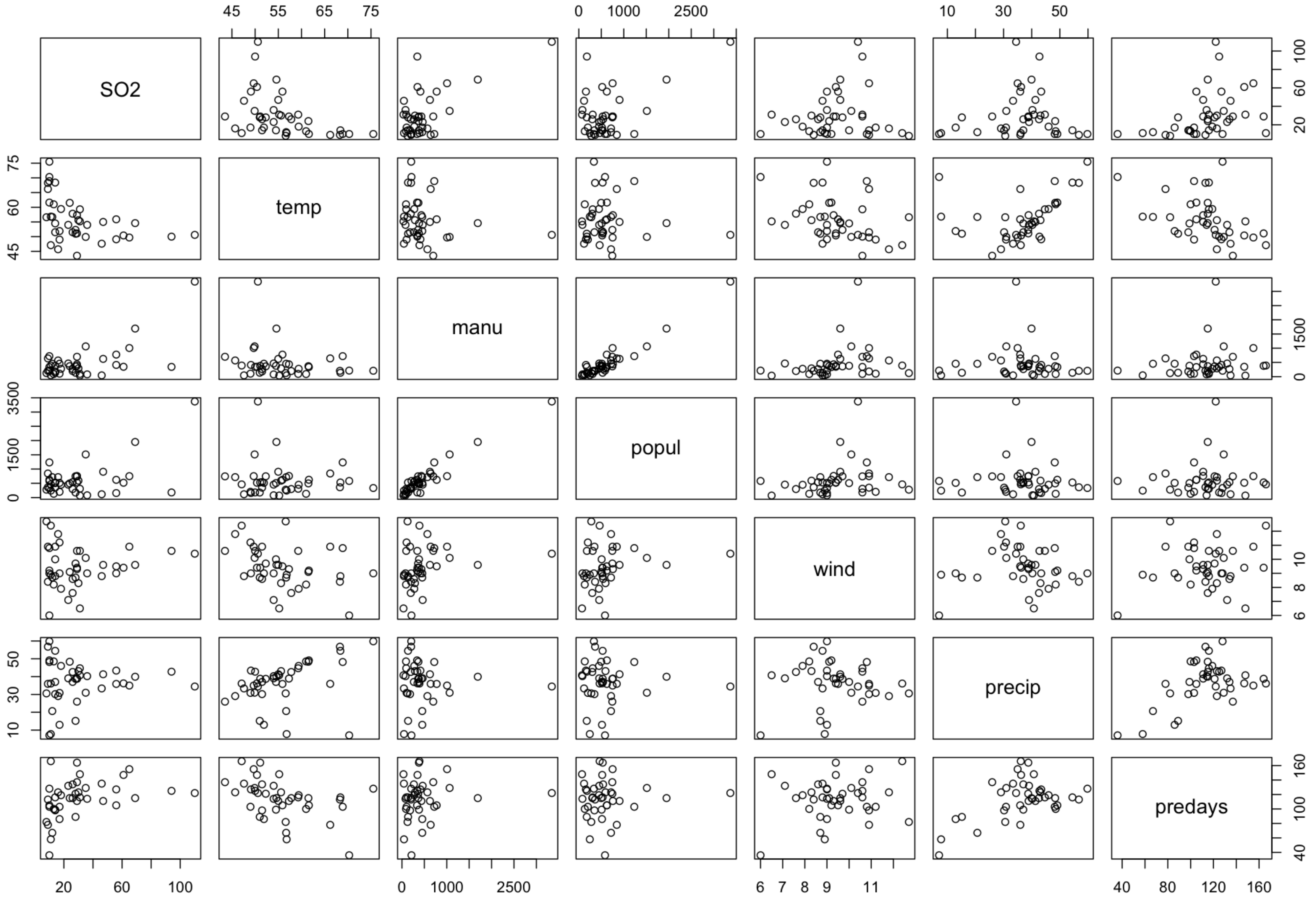
Y/N, WHY?

Correlation = LINEAR relation

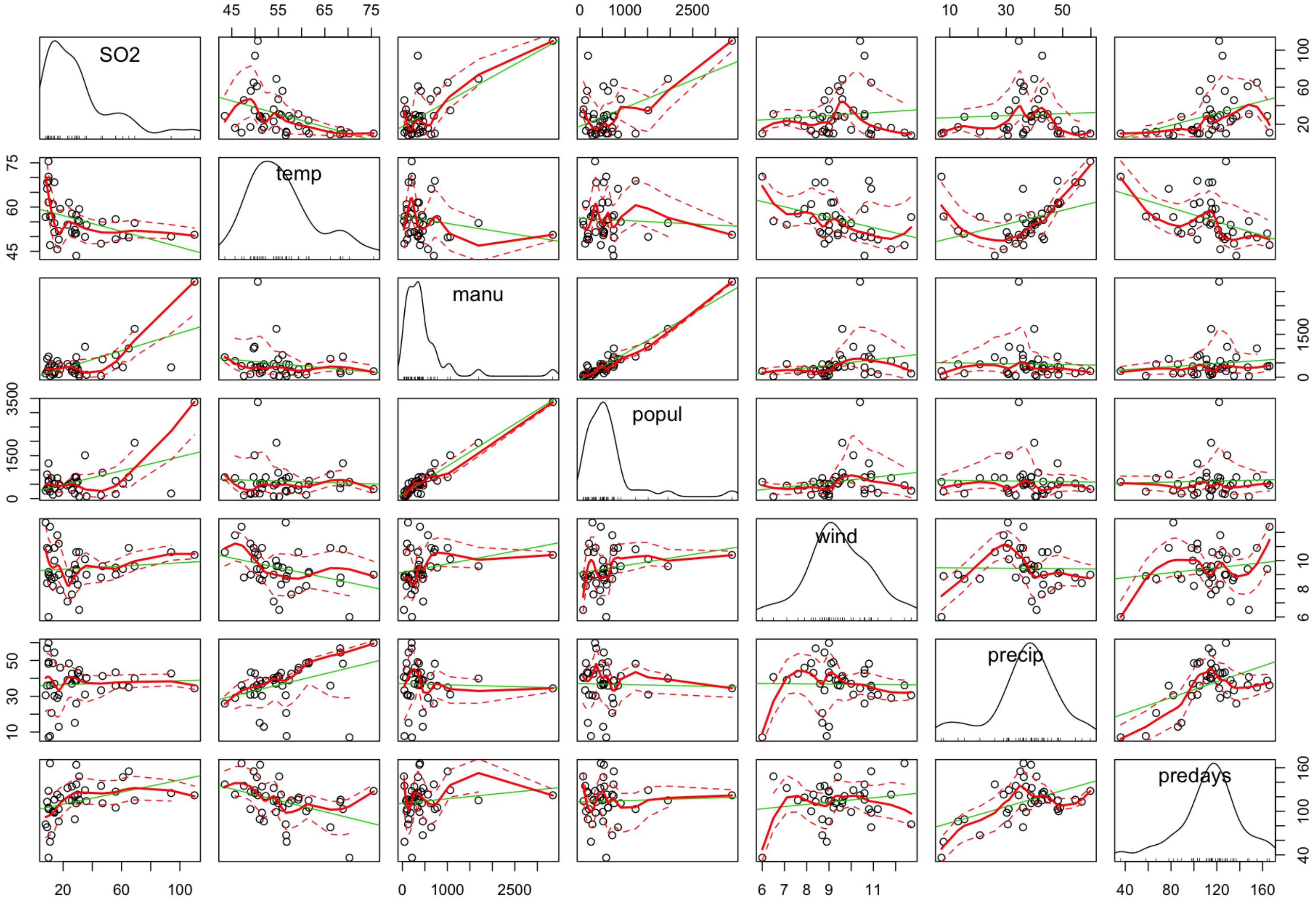


Use cor.test() in R to test for zero correlation (Fisher's z-Test) with confidence interval

data("USairpollution", package = "HSAUR2")



data("USairpollution", package = "HSAUR2")



Covariance/Correlation Matrix

Pairwise values

Covariance matrix: $\Sigma_{ij} = Cov(X_i, X_j)$

Correlation matrix: $C_{ij} = Cor(X_i, X_j)$

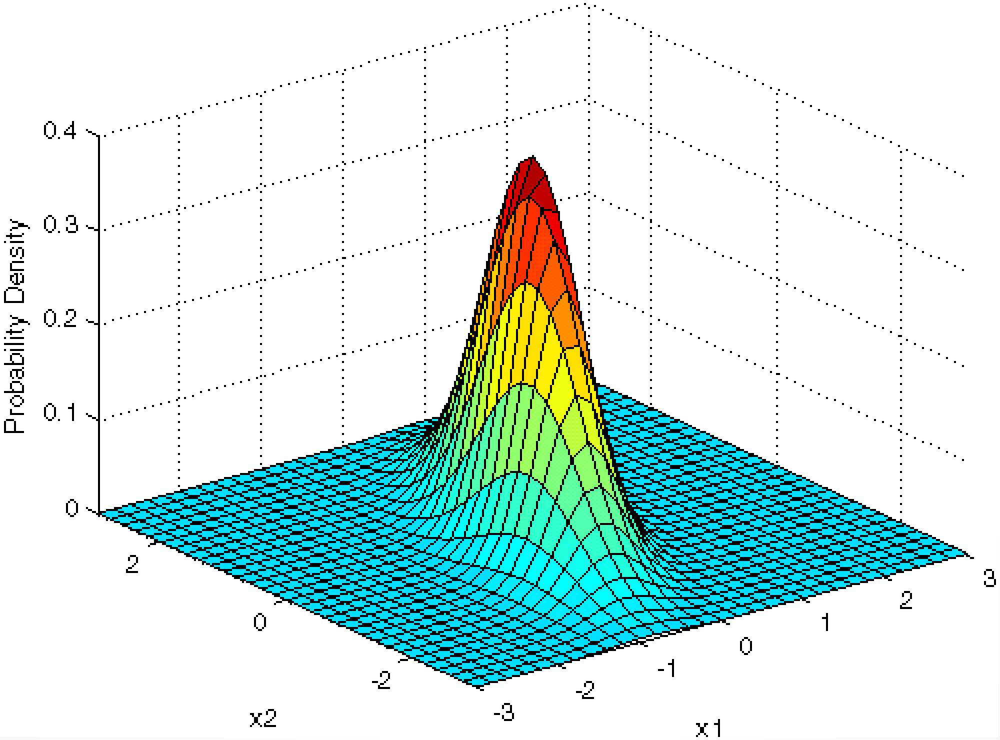
Sample covariance matrix: $S_{ij} = \widehat{Cov}(x_i, x_j)$

Sample correlation matrix: $R_{ij} = \widehat{Cor}(x_i, x_j)$

Correlation is invariant to changes in units, covariance is not.

Multivariate Gaussian

$$f(x;\mu,\Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}}\exp\left(-\frac{1}{2}\cdot(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$



Properties

If $\vec{X} \sim \mathcal{M}\mathcal{V}\mathcal{N}(\vec{\mu}, \Sigma)$

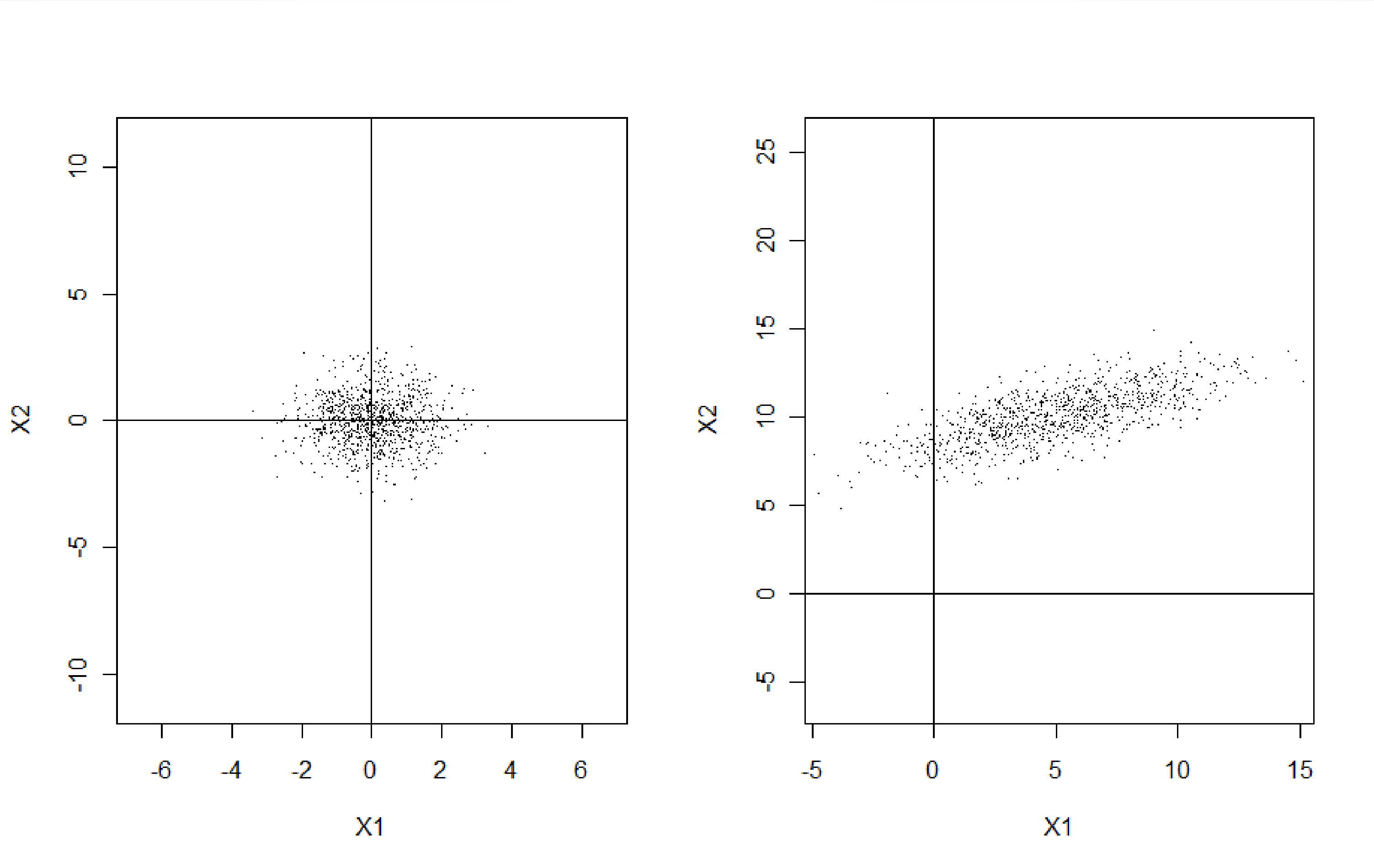
1. every linear combination e.g. $Y = aX+b$ is normally distributed, with

$$X \sim \mathcal{N}(\mu, \sigma^2) \Leftrightarrow aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$$

2. every projection on a subspace is multivariate normally distributed

However. If margins follow normal distribution, it is NOT guaranteed that the underlying distribution of “the Space” is multivariate Gaussian.

“Multivariate” is stronger than “Normal Margins”

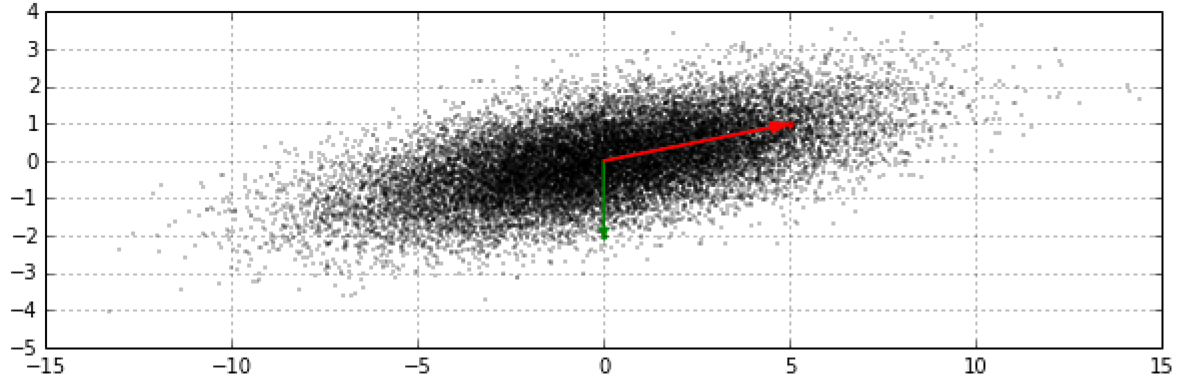


$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \mu = \begin{pmatrix} 5 \\ 10 \end{pmatrix}, \Sigma = \begin{pmatrix} 10 & 3 \\ 3 & 2 \end{pmatrix}$$

Multivariate Gaussian

(Mahalanobis Distance)^2

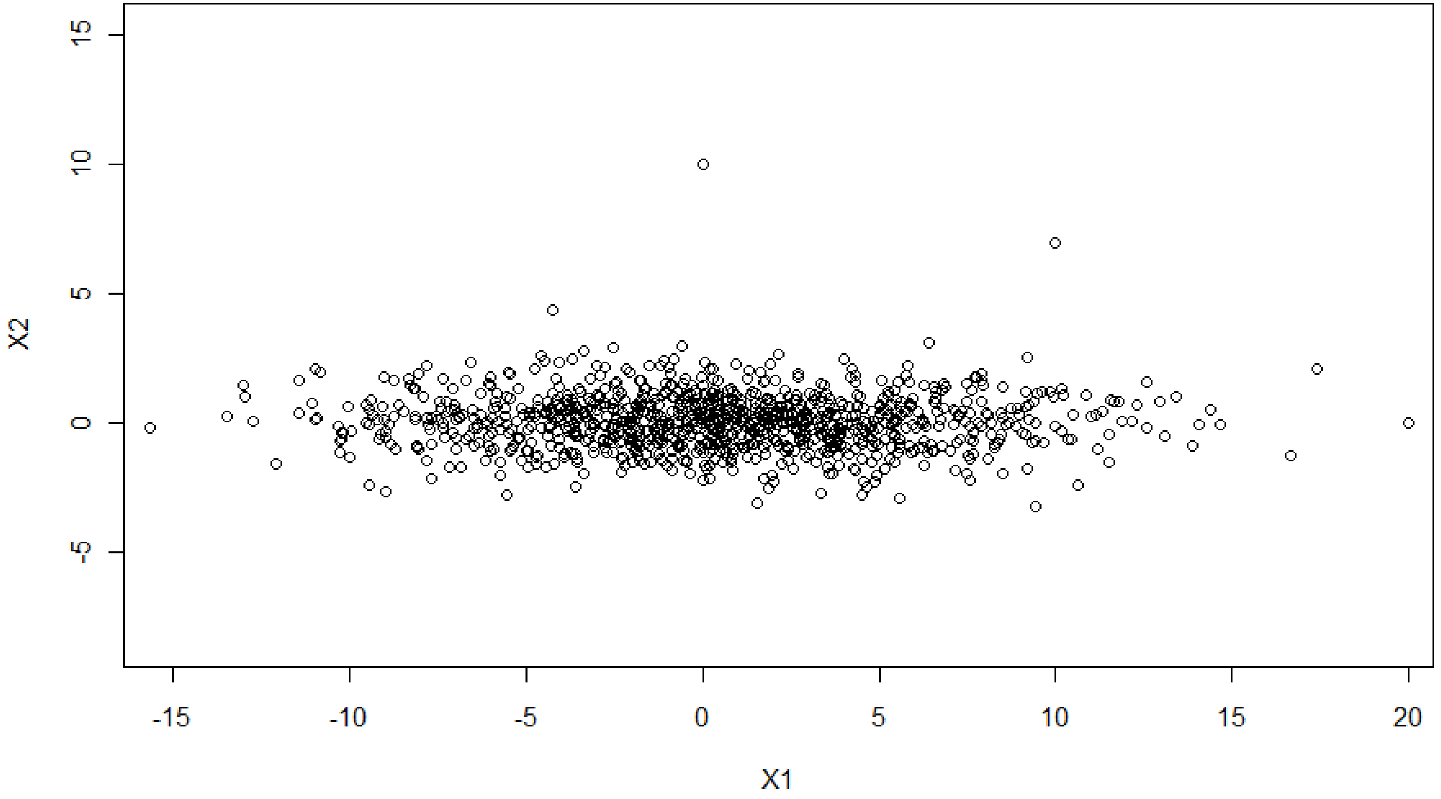
$$f(x;\mu,\Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left(-\frac{1}{2}\cdot \boxed{(x-\mu)^T\Sigma^{-1}(x-\mu)}\right)$$



Euclidean distance

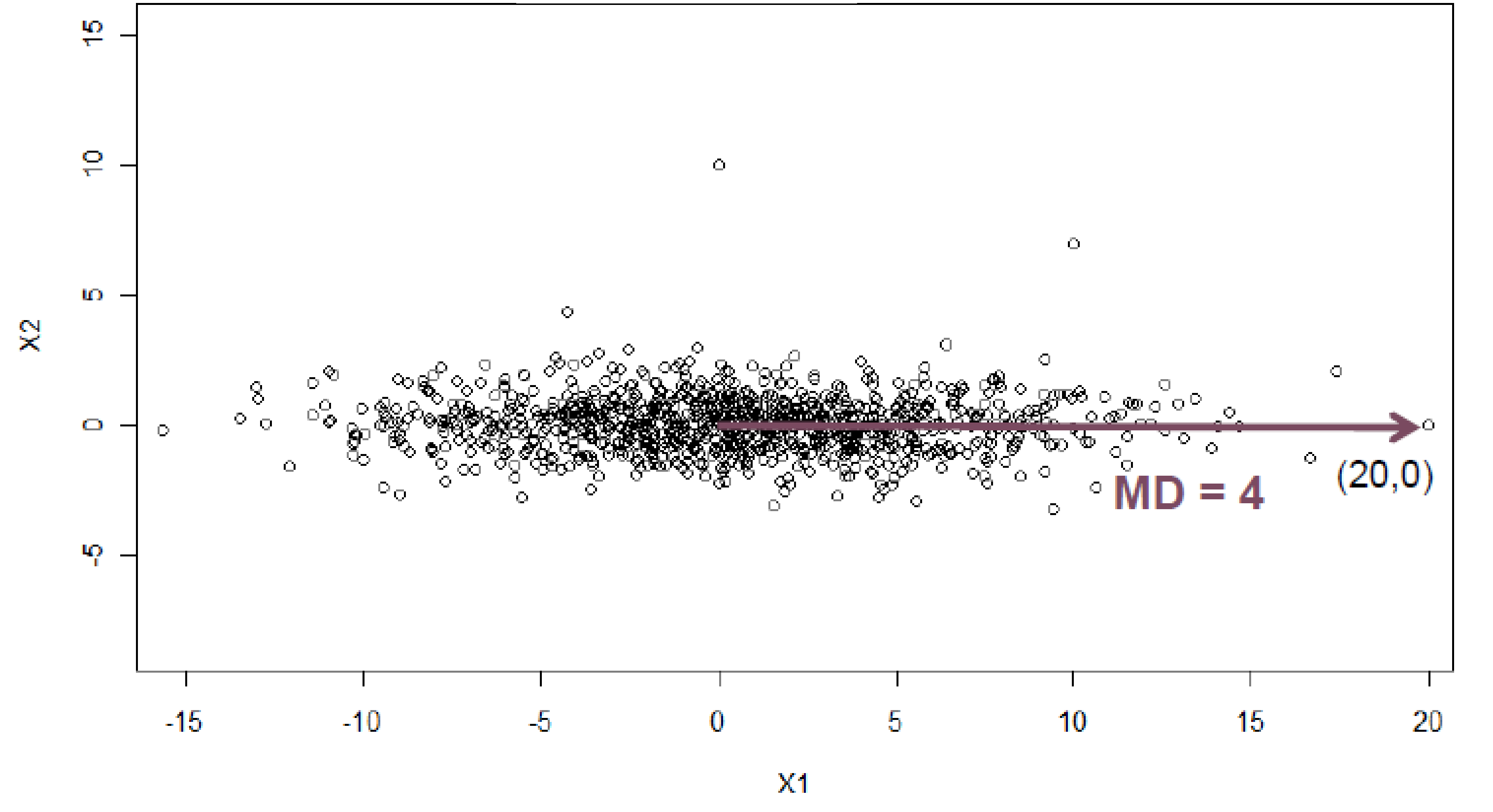
Mahalanobis dist & Correlation

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 25 & 0 \\ 0 & 1 \end{pmatrix}$$



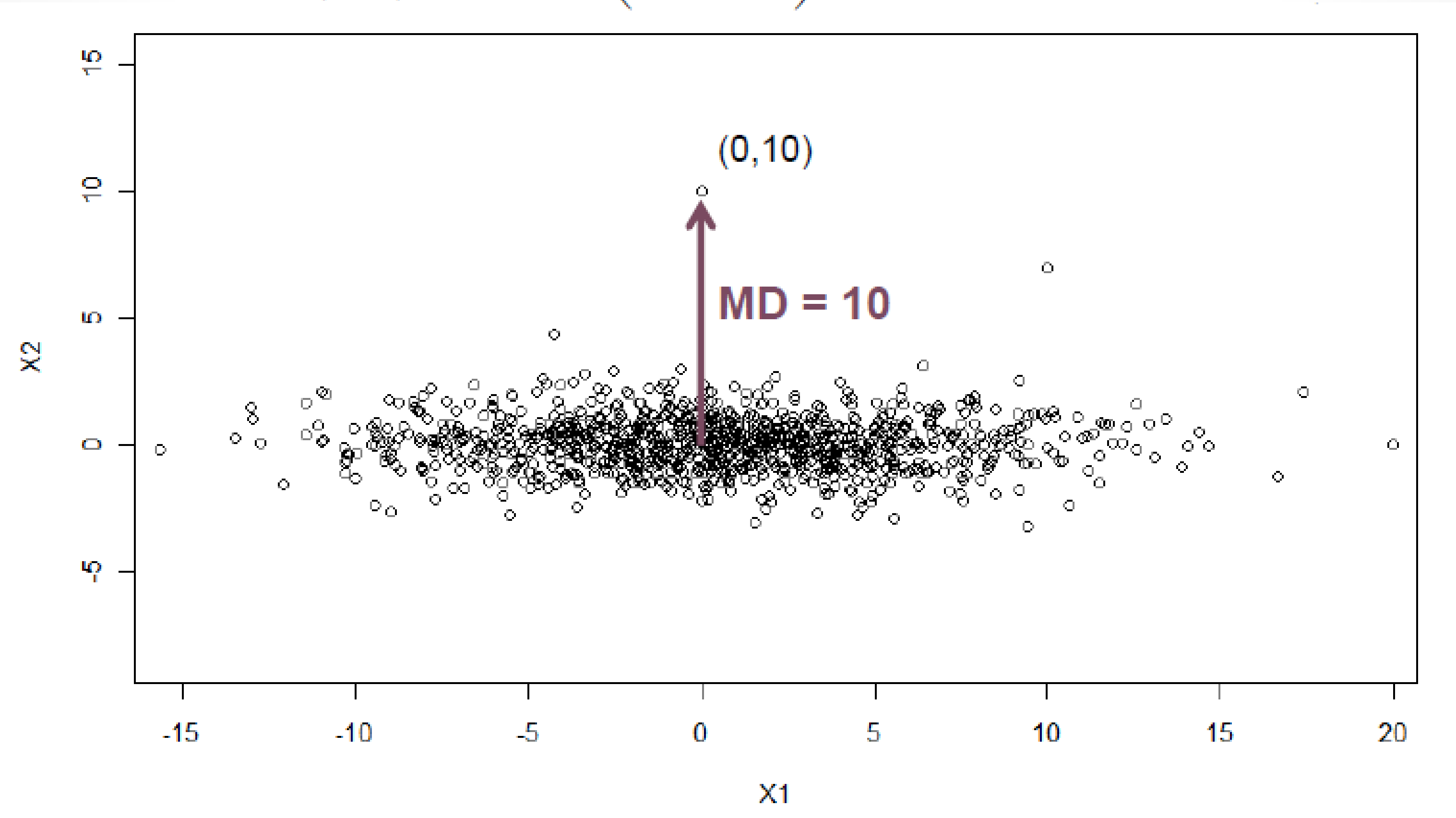
Mahalanobis dist & Correlation

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 25 & 0 \\ 0 & 1 \end{pmatrix}$$



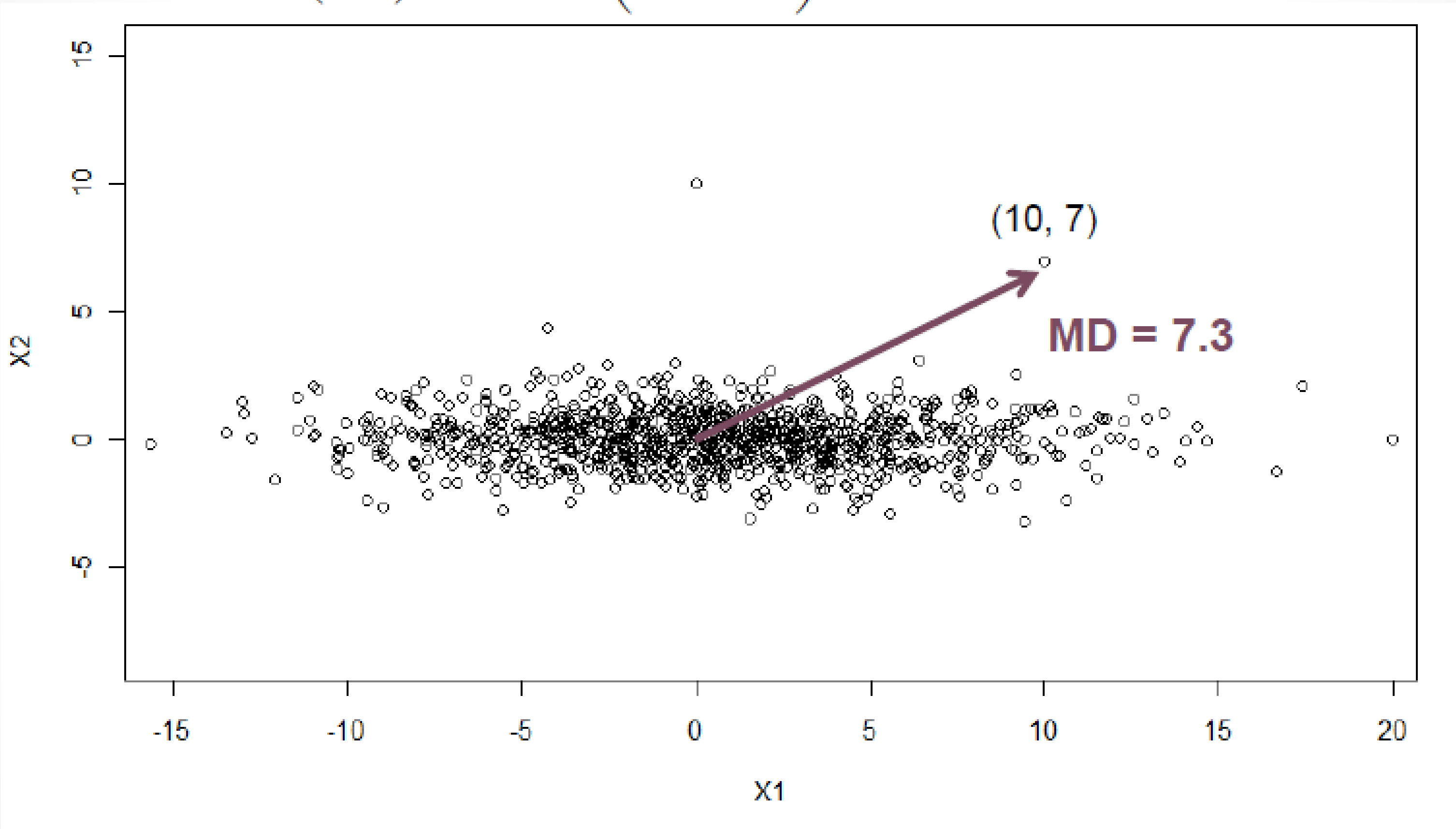
Mahalanobis dist & Correlation

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 25 & 0 \\ 0 & 1 \end{pmatrix}$$

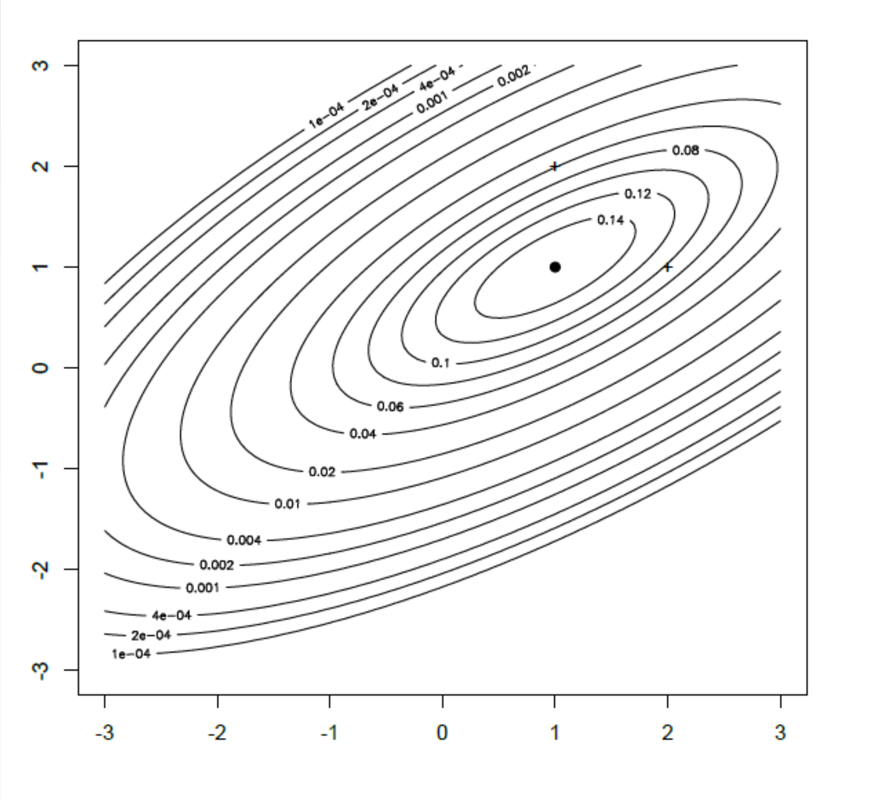


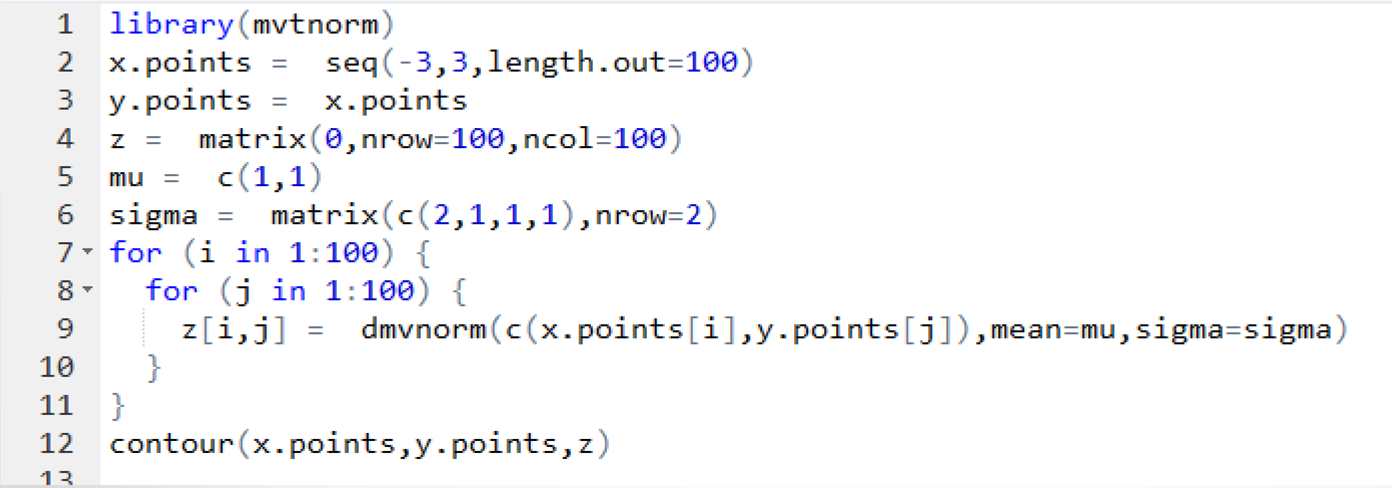
Mahalanobis dist & Correlation

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 25 & 0 \\ 0 & 1 \end{pmatrix}$$



Multivariate Gaussian density with p = 2

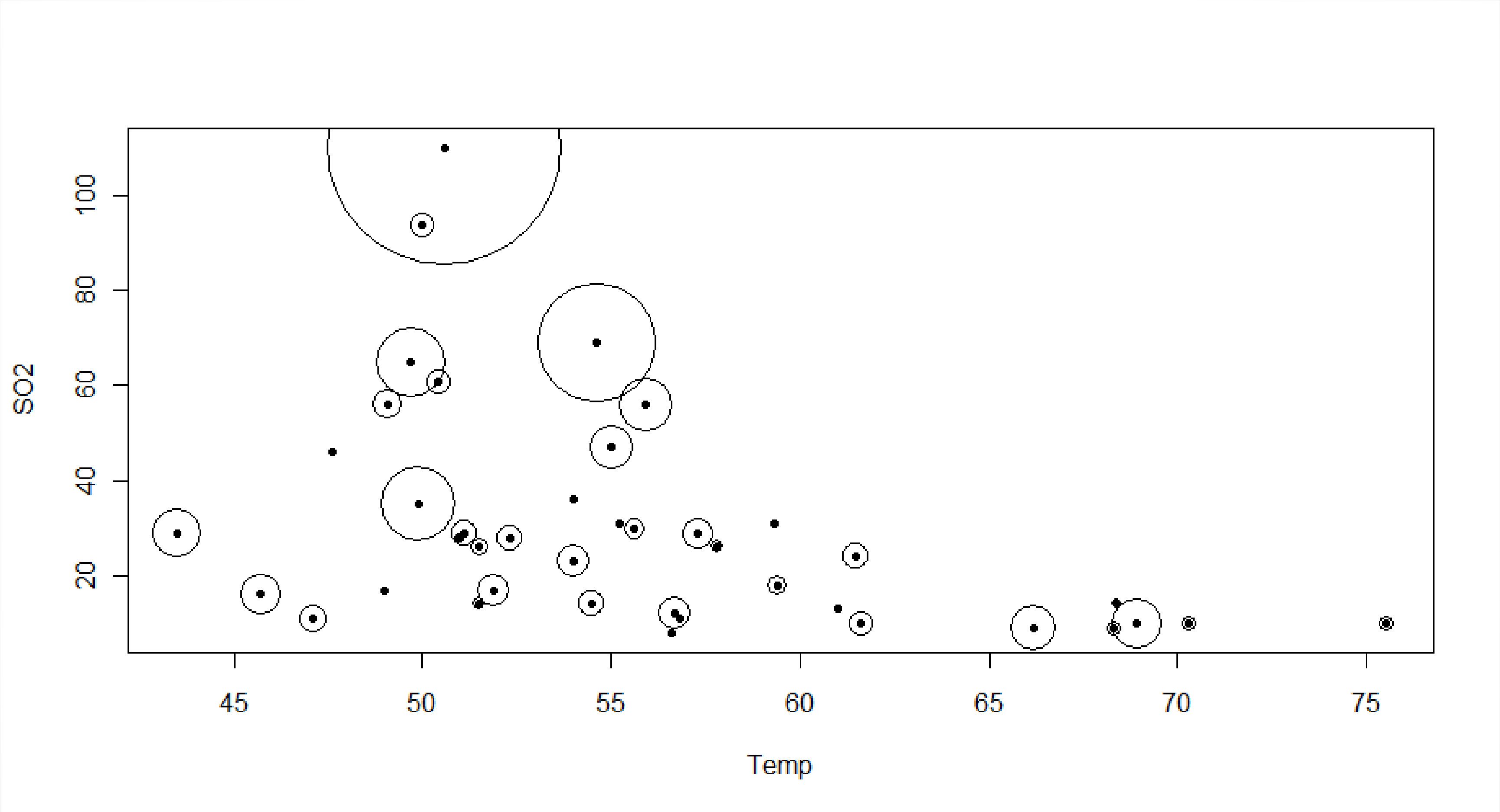




```r
library(mvtnorm)
x.points =  seq(-3,3,length.out=100)
y.points =  x.points
z =  matrix(0,nrow=100,ncol=100)
mu =  c(1,1)
sigma =  matrix(c(2,1,1,1),nrow=2)
for (i in 1:100) {
  for (j in 1:100) {
    z[i,j] =  dmvnorm(c(x.points[i],y.points[j]),mean=mu,sigma=sigma)
  }
}
contour(x.points,y.points,z)
```

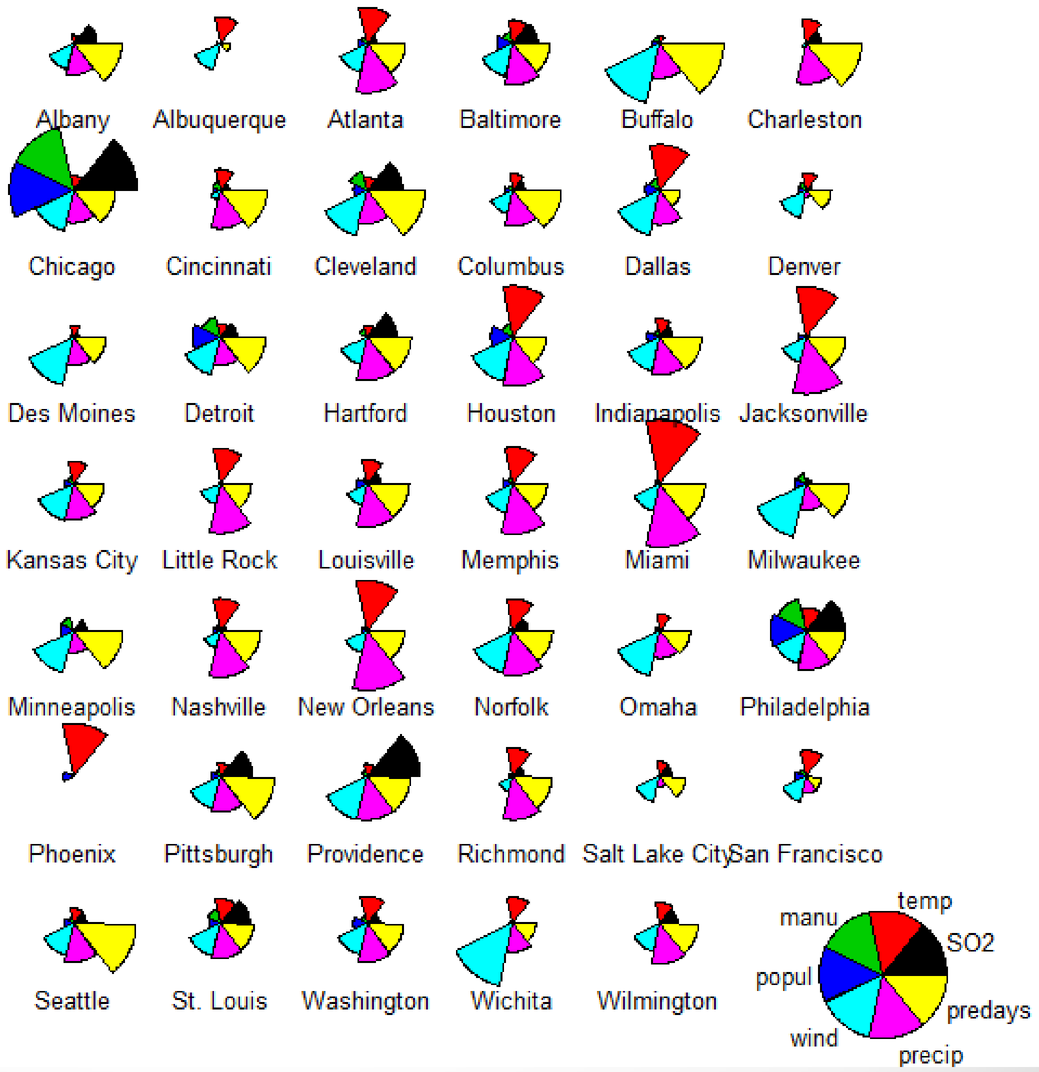
Exploratory Data Analysis & Visualization (EDAV) for Multivariate

Bubbleplot



Glyphplots

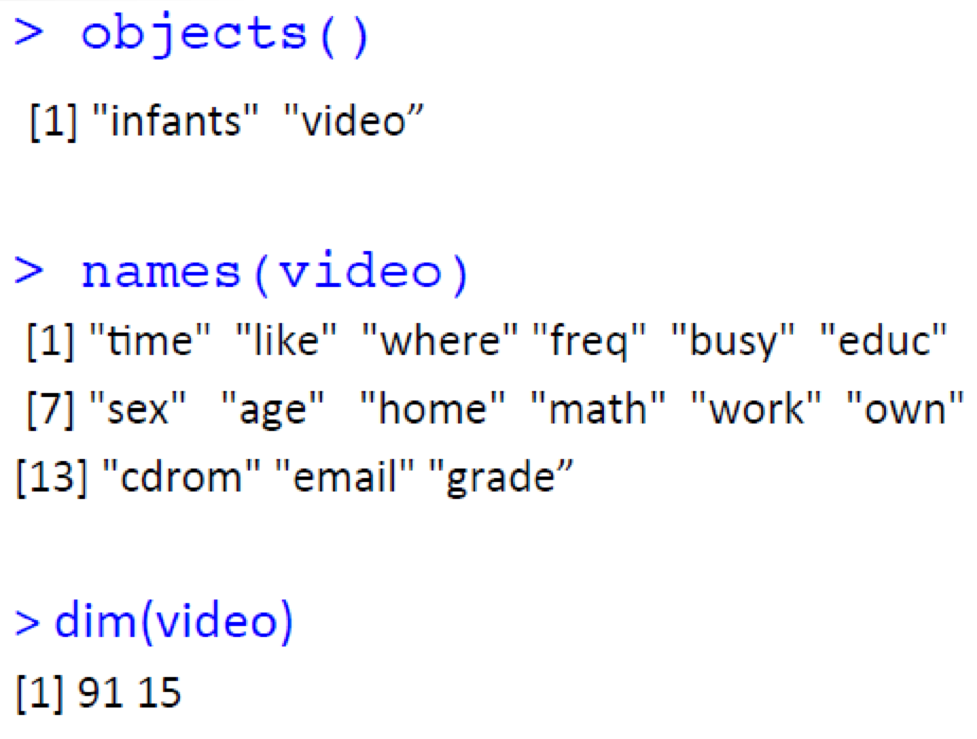
Good for continuous data, what if data is not continuous?



Case study: College Students' Video Game

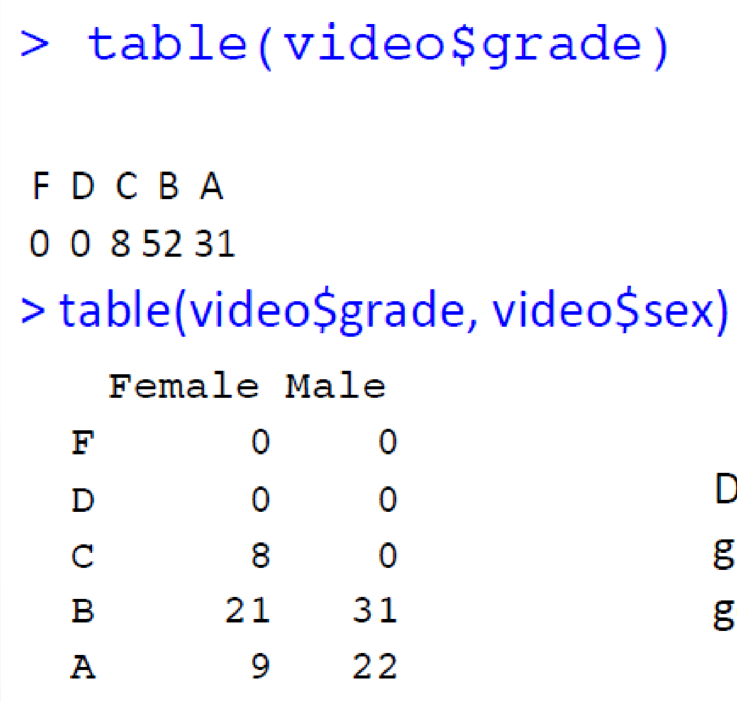
- College students' video game data:
  1. Random sample of 91 out of 314 students
  2. Variables:
     - gender (male/female) – Qualitative (nominal)
     - expected grade (A,B,C,D,F) – Qualitative (ordered)

- Let’s see what are there in our current environment



```
> objects()
 [1] "infants"  "video”

> names(video)
 [1] "time"  "like"  "where" "freq"  "busy"  "educ"
 [7] "sex"   "age"   "home"  "math"  "work"  "own"
[13] "cdrom" "email" "grade”

> dim(video)
[1] 91 15
```

table(…) is helpful for qualitative data

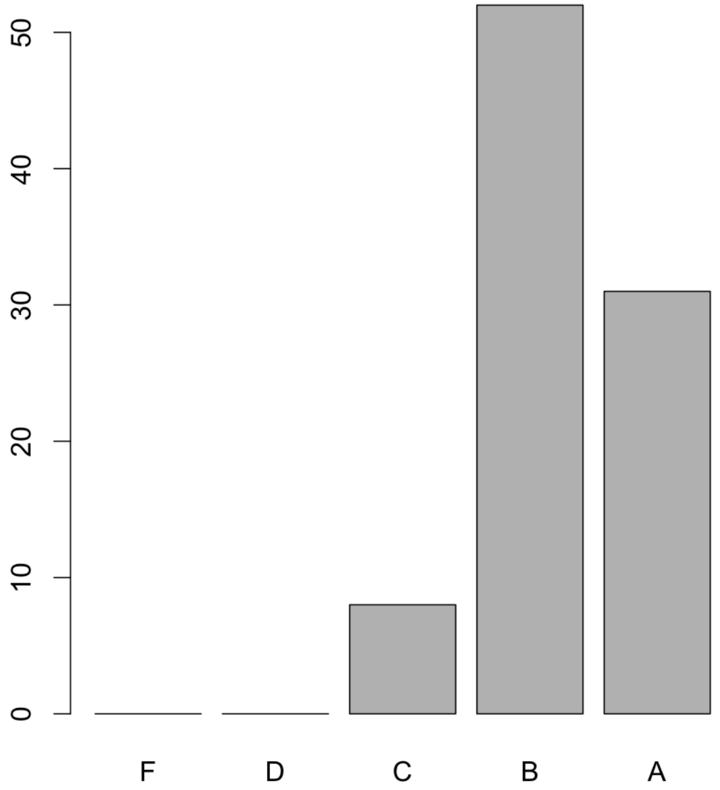


```
> table(video$grade)

 F  D  C  B  A
 0  0  8 52 31
> table(video$grade, video$sex)

       Female Male
  F         0    0
  D         0    0
  C         8    0
  B        21   31
  A         9   22
```

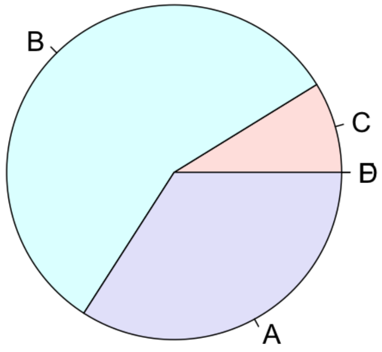
Anything unusual about the expected grade?

Does expected grade depend on gender?

Pie chart pie(…) & Bar chart barplot(…) is helpful for qualitative data, BUT



barplot(table(video$grade))

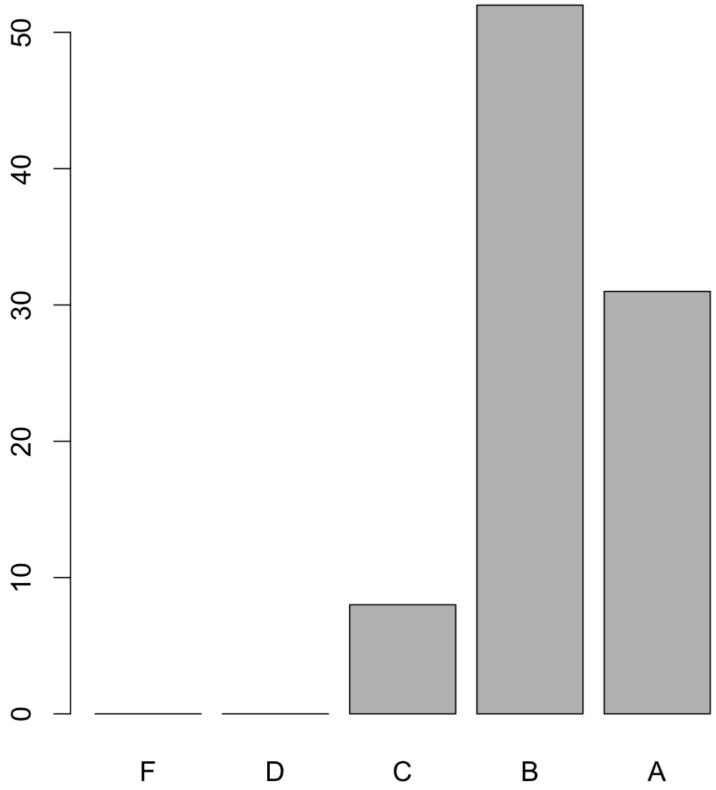


pie(table(video$grade))

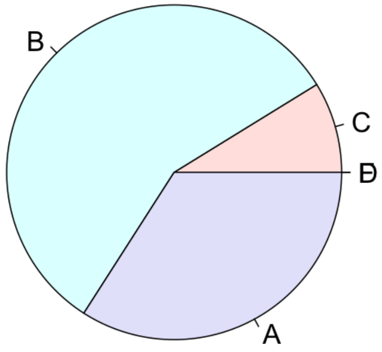
Pie chart pie(…)

& Bar chart barplot(…) is helpful for

qualitative data, BUT



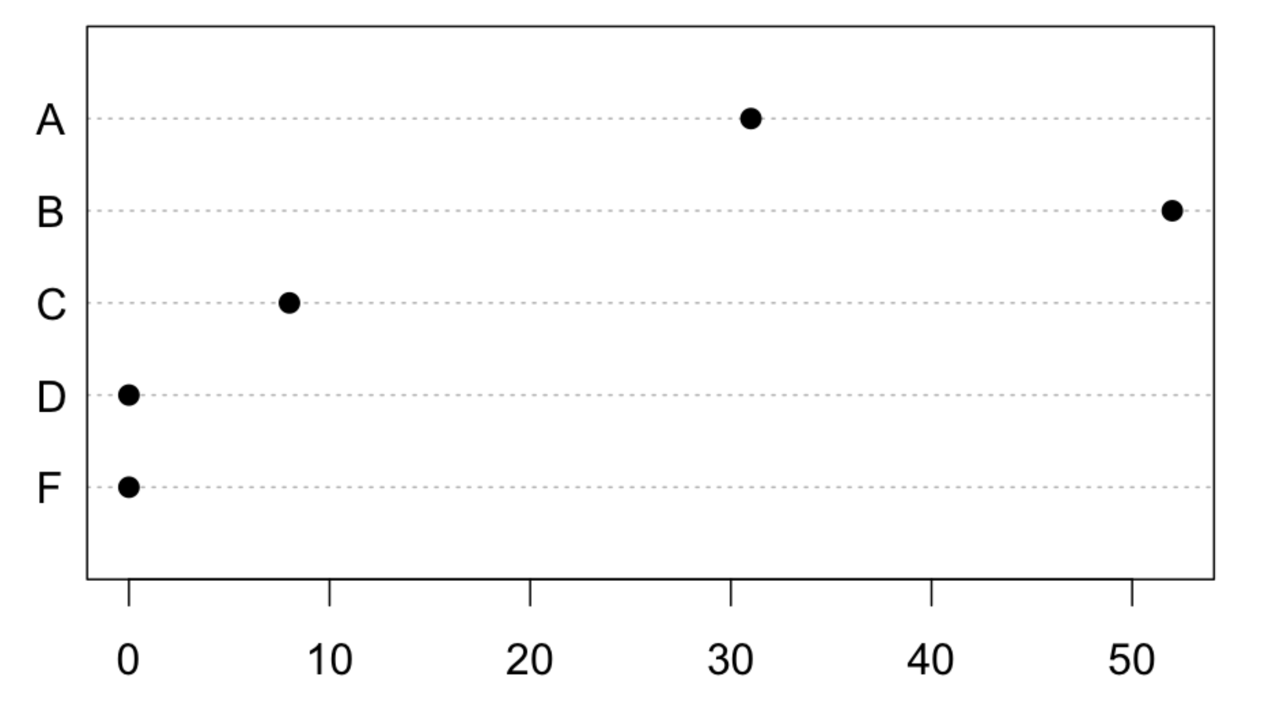
Width of bars have no meaning



Areas can be hard to compare

Dot Chart: focus on comparison of values

```
dotchart(table(video$grade), pch=19)
```



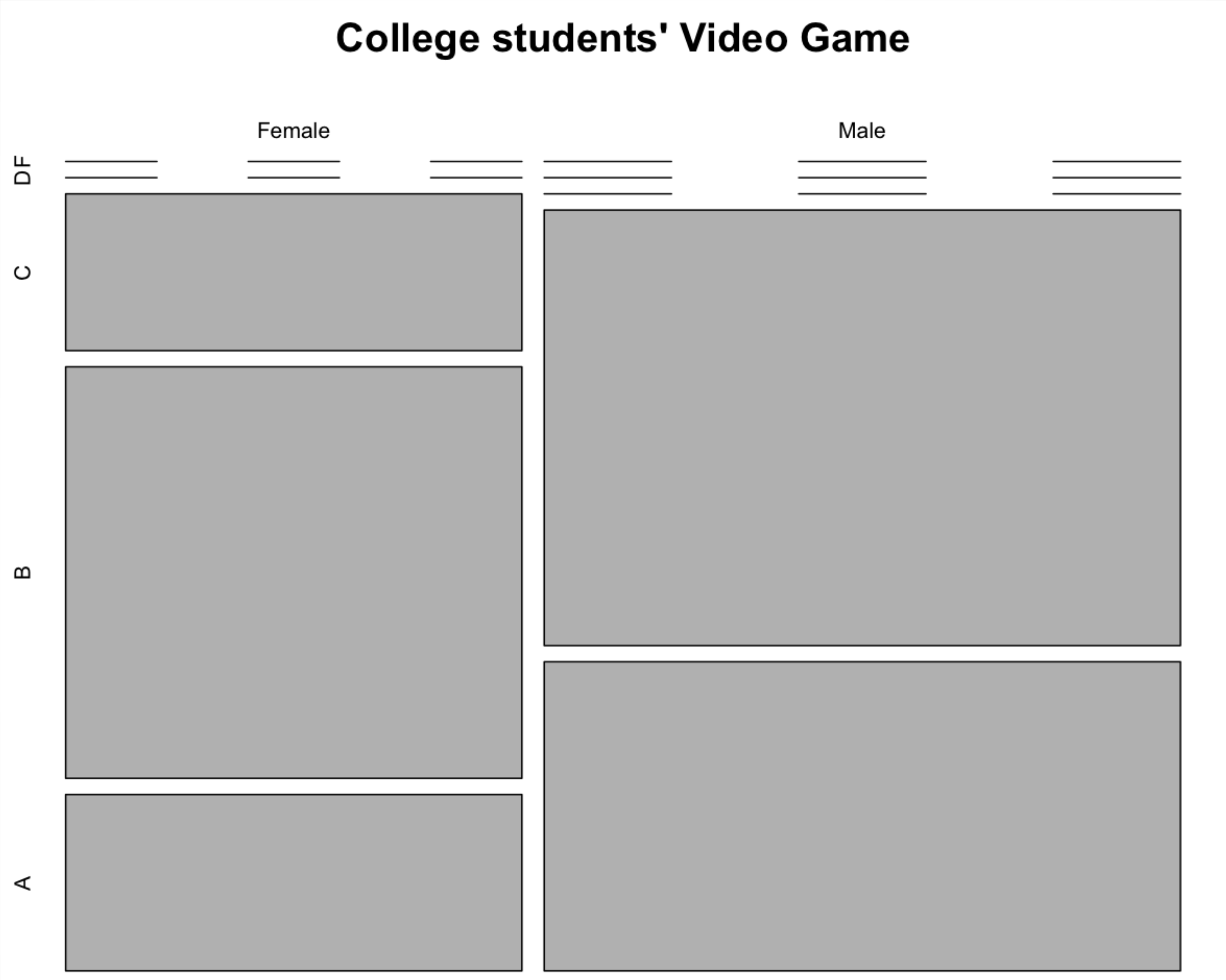
Graphs are comparison…

❑ **Goal of comparison**

1. better understand a distribution
2. Subgroups vs. Population
3. Subgroups vs. Standard

❖ How do you find out the expected grade distribution might vary with gender?

- Two qualitative variables – any clue?

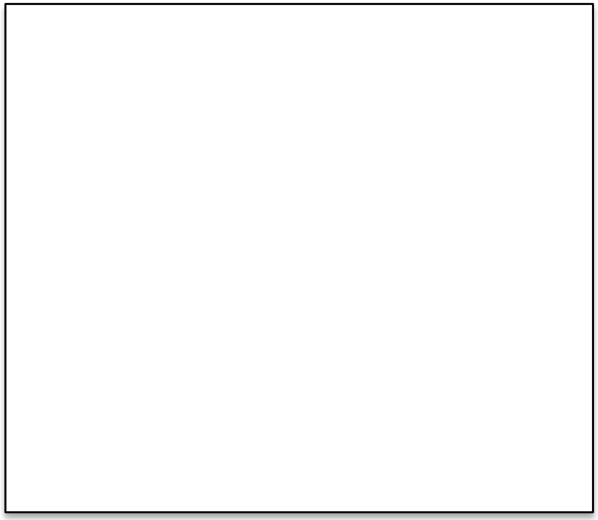


```
mosaicplot(table(video$sex,video$grade),
                    main='College students\' Video Game')
```

How is a Mosaic plotted?

- 91 students

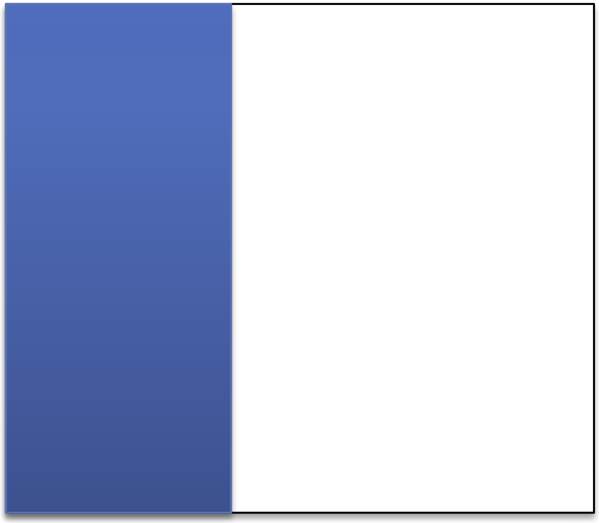
Think of them as spread out evenly over the box



Start to plot a new Mosaic:

- ❑ 38 females

Put all the females on one side of the box.

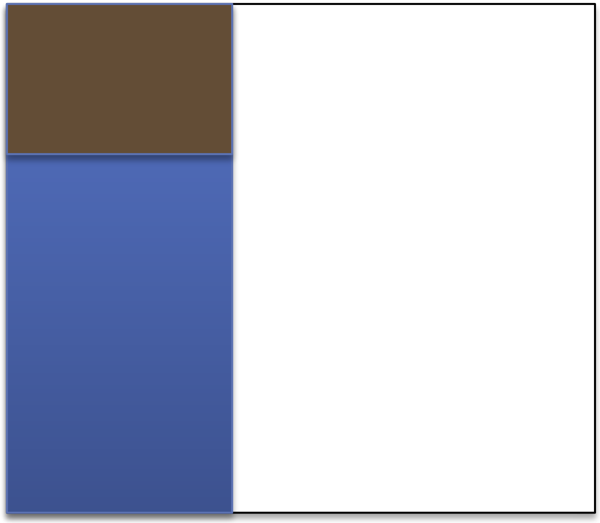


Females 38/91

Continue with the Mosaic:

- ❑ Grades (C):

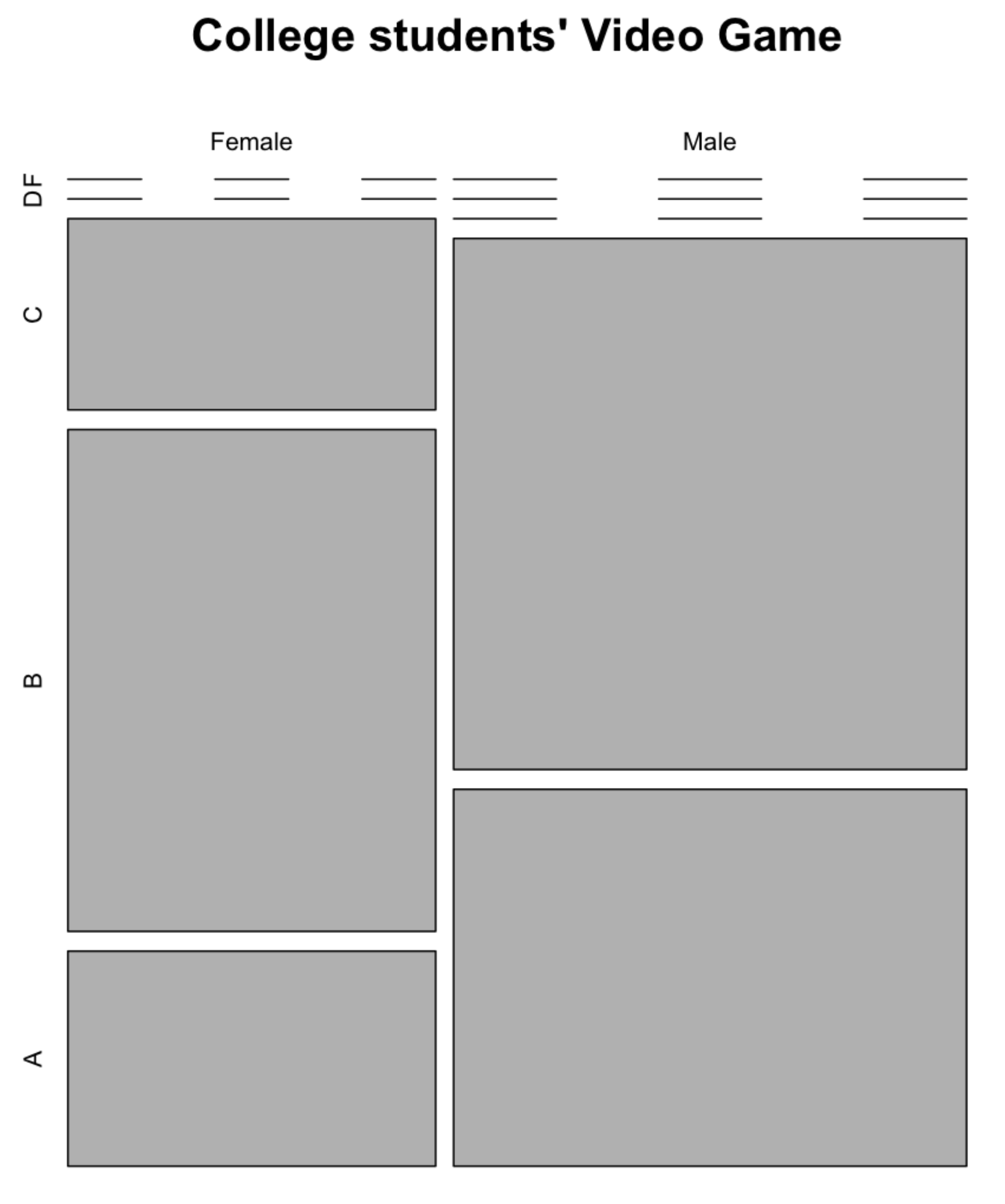
Females Expect C are 8/38



Females 38/91

1. Smaller fraction of females expect an A in comparison to Males

2. None of the males expect a C



AFTER CLASS

1. Complete the survey

2. Get a Github account – learn to fork the repo:

   Github.com/MRandomMax/MSI

3. Read IAMA Ch2

4. Homework starts next week.