# Correlation

Flood Frequency Analysis PROSPECT

May 2022

Gregory S. Karlovits, *P.E.*, *PH*, *CFM*

Statistical Hydrologist

US Army Corps of Engineers

Hydrologic Engineering Center

US Army Corps of Engineers®

# Purpose

- Understand how to measure the strength of relation between two variables

US Army Corps of Engineers®

# Outline

1. Definition/understanding correlation
2. Measuring correlation
3. Visualizing correlation, transformations

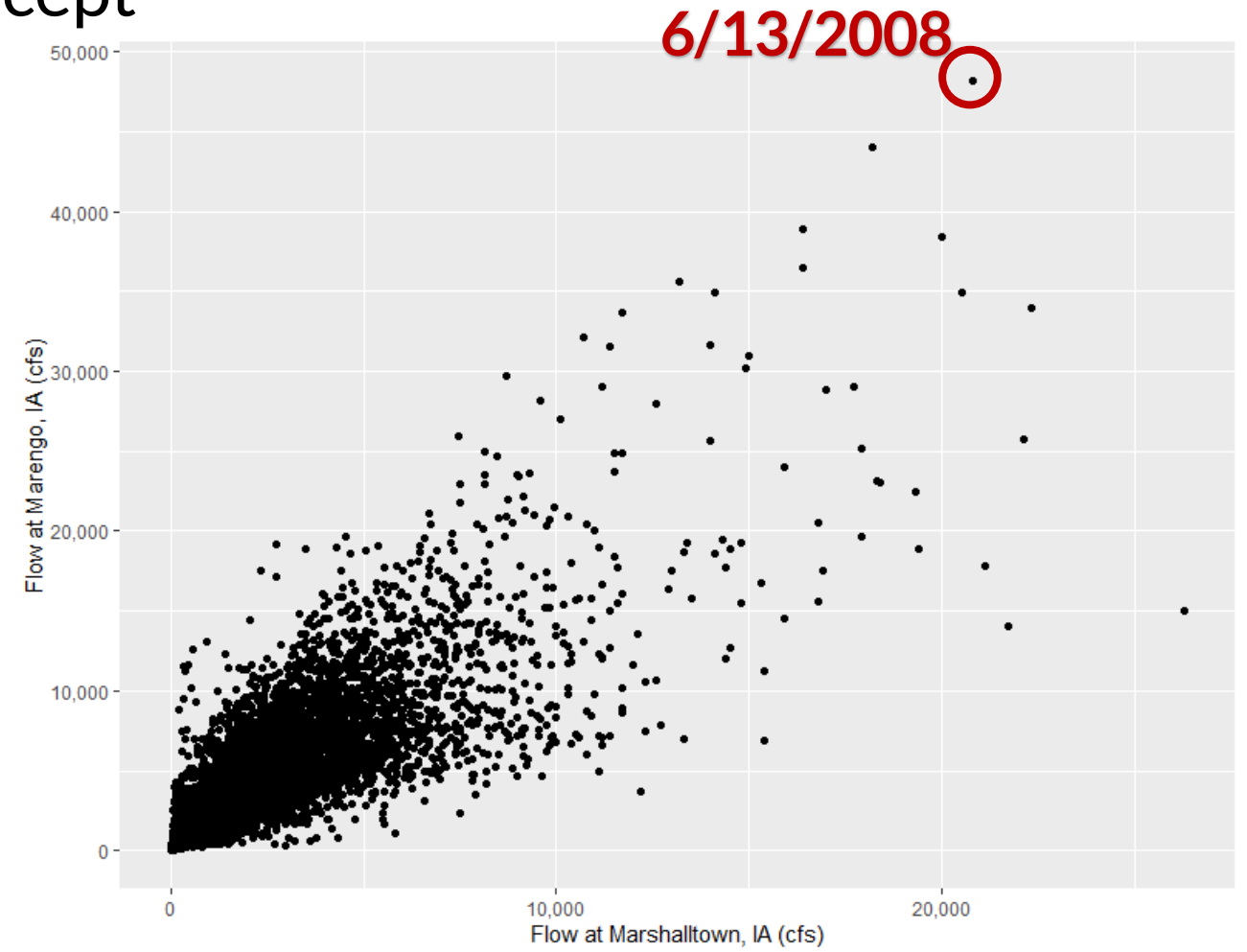
US Army Corps
of Engineers®

# Defining and Understanding Correlation

US Army Corps of Engineers®

# Background

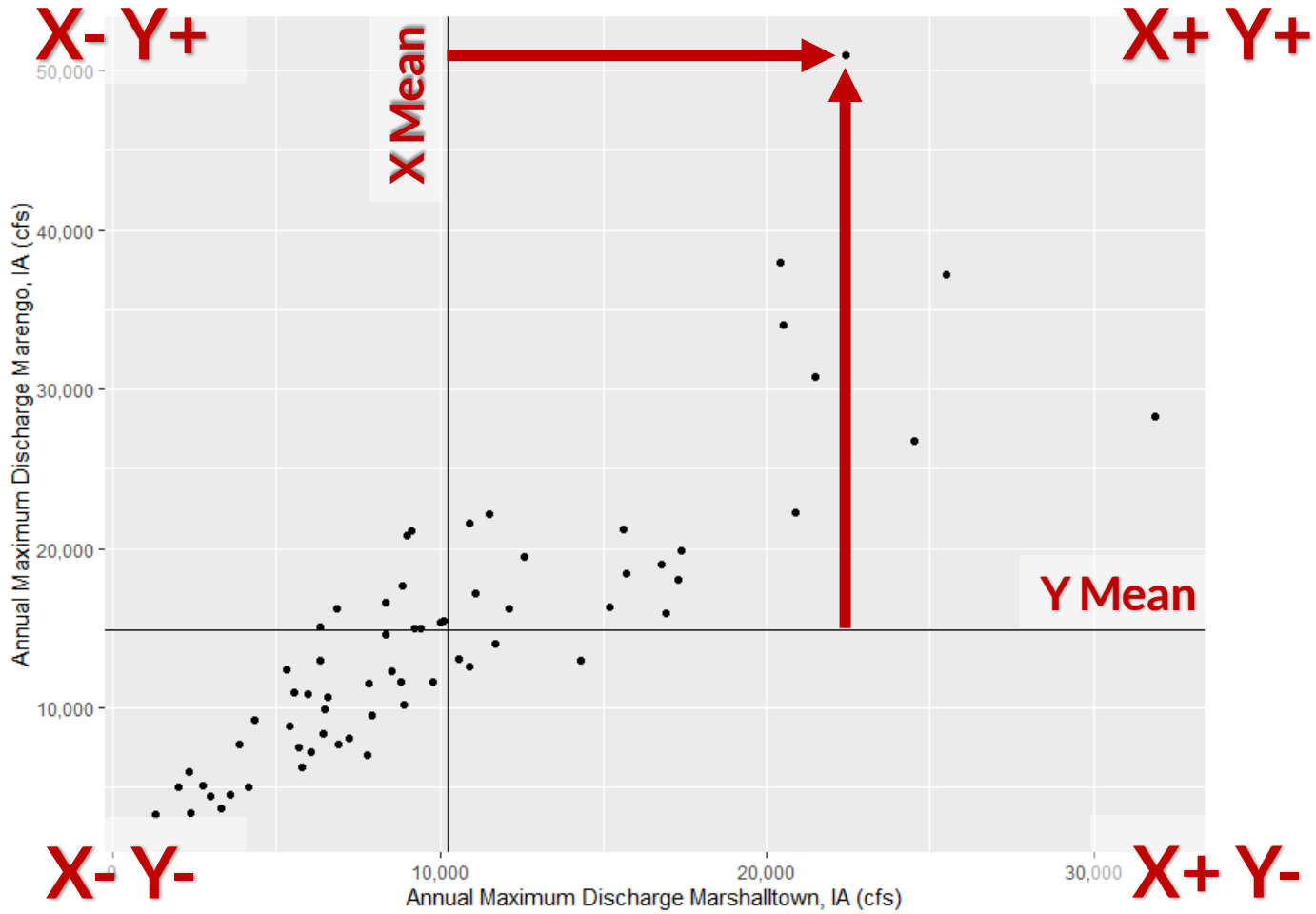
- Correlation is a *multivariate* concept*
- Two or more variables coincident in some way
- Each variable has its own sample of multiple data points




US Army Corps of Engineers®

# “Co-relation”

- When variable X departs from its mean, does variable Y depart from its mean in a similar manner?




US Army Corps of Engineers®

# Correlation

- How can we **quantify** the strength of a relationship between two variables?
- Does not imply causation…

US Army Corps of Engineers®

# Data

- Two or more variables coincident in some way
- Example here: coincident in time
- Correlation measured on coincident pairs of data

| WY | Marshalltown | Marengo |
|---|---|---|
| **1993** | 20400 | 38000 |
| **1994** | 6420 | 8350 |
| **1995** | 4330 | 9210 |
| **1996** | 5970 | 10900 |
| **1997** | 8890 | 10200 |
| **1998** | 12600 | 19500 |
| **1999** | 9790 | 11600 |
| **2000** | 5790 | 6290 |
| **2001** | 7910 | 9540 |
| **2002** | 2750 | 5180 |
| **2003** | 5420 | 8880 |
| **2004** | 8330 | 14600 |
| **2005** | 7240 | 8130 |
| **2006** | 3610 | 4580 |
| **2007** | 16900 | 15900 |
| **2008** | 22400 | 51000 |
| **2009** | 10100 | 15500 |
| **2010** | 12100 | 16200 |
| **2011** | 6480 | 9950 |
| **2012** | 2400 | 3430 |
| **2013** | 25500 | 37200 |
| **2014** | 24500 | 26800 |
| **2015** | 14300 | 13000 |
| **2016** | 16800 | 19000 |
| **2017** | 8510 | 12300 |
| **2018** | 9000 | 20800 |
| **2019** | 15700 | 18400 |
| **2020** | 9250 | 15000 |
| **2021** | 2010 | 5040 |

US Army Corps of Engineers®

# Computing Correlation

US Army Corps
of Engineers®

# Population Covariance

- How much do two variables vary from their mean *at the same time?*
  - Notated $cov[X, Y]$
- $\sigma_{xy} = cov[X, Y] = E[(X - E[X])(Y - E[Y])]$
  - where $E[X]$ is the *expected value* of random variable $X$
  - $cov[X, X] = var[X]$
  - $\sigma_{xx} = \sigma_x^2$

US Army Corps of Engineers®

# Sample Covariance

- Sample $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
- Note:
  - $s_{xx} = s_x^2$
  - Has units of (unit of x) * (unit of y)
  - No bounds on range of values

**Sample Variance**

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

US Army Corps of Engineers®

# Correlation

- Normalization of covariance to be on the range [-1, 1]
- Normalize using the product of the two variables' variance

US Army Corps of Engineers®

# Three Ways to Measure Sample Correlation

- **Pearson's product-moment correlation coefficient (*r*)**
- Kendall's rank correlation coefficient ($\tau$)
- Spearman's rank correlation coefficient ($\rho$)

US Army Corps of Engineers®

# Pearson's Correlation Coefficient

- Measurement of linear correlation between two variables
  - Performs badly in non-linear situations
- Normalization of covariance between variables

- $\rho_{X,Y} = \frac{cov[X,Y]}{\sqrt{var[X]}\sqrt{var[Y]}}$ $\qquad r_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}$

US Army Corps of Engineers®

**R:**
`cor(x)`
**x is a data frame with 2 or more columns**

# Kendall's Tau

- Measurement of ordinal association between variables
- $\tau = \frac{2}{n(n-1)} \sum_{i<j} sgn(x_i - x_j) sgn(y_i - y_j)$

Term is -1 if both x and y do not increase (or decrease), +1 otherwise

- Special modifications for ties ($x_i = x_j$ or $y_i = y_j$)

US Army Corps of Engineers®

R:

```
cor(x, method = "kendall")
```

x is a data frame with 2 or more columns

# Spearman's Rho

- Pearson's correlation on rank-transform of the data
- Assesses monotonic relationships whether or not they are linear
- $\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$ where $d_i = rank(X_i) - rank(Y_i)$
- Special modifications for ties

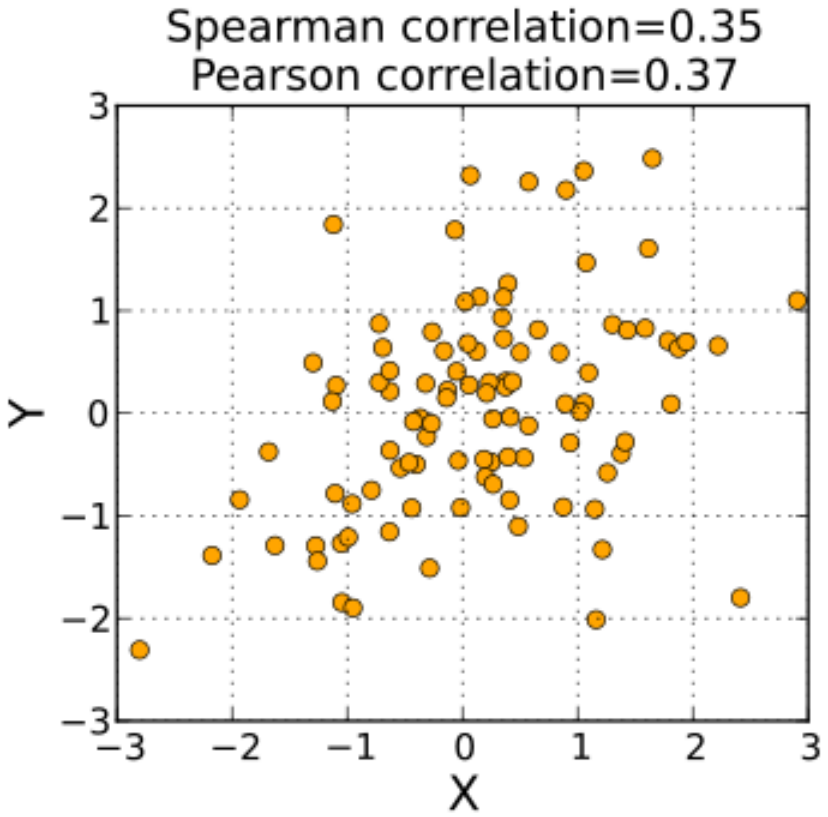
US Army Corps of Engineers®

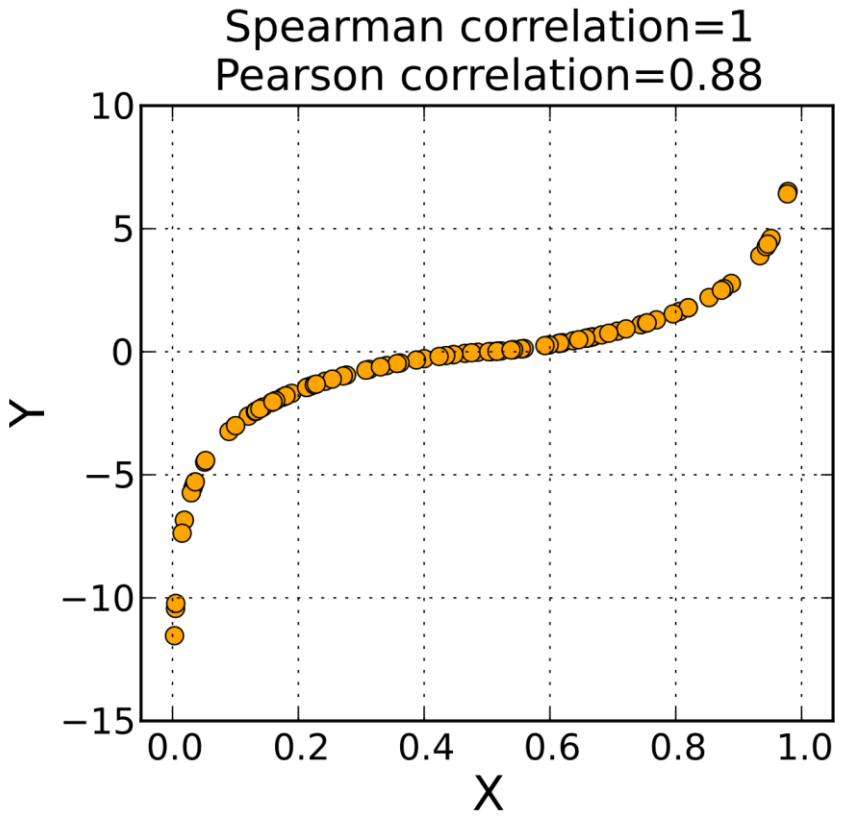
**R:**

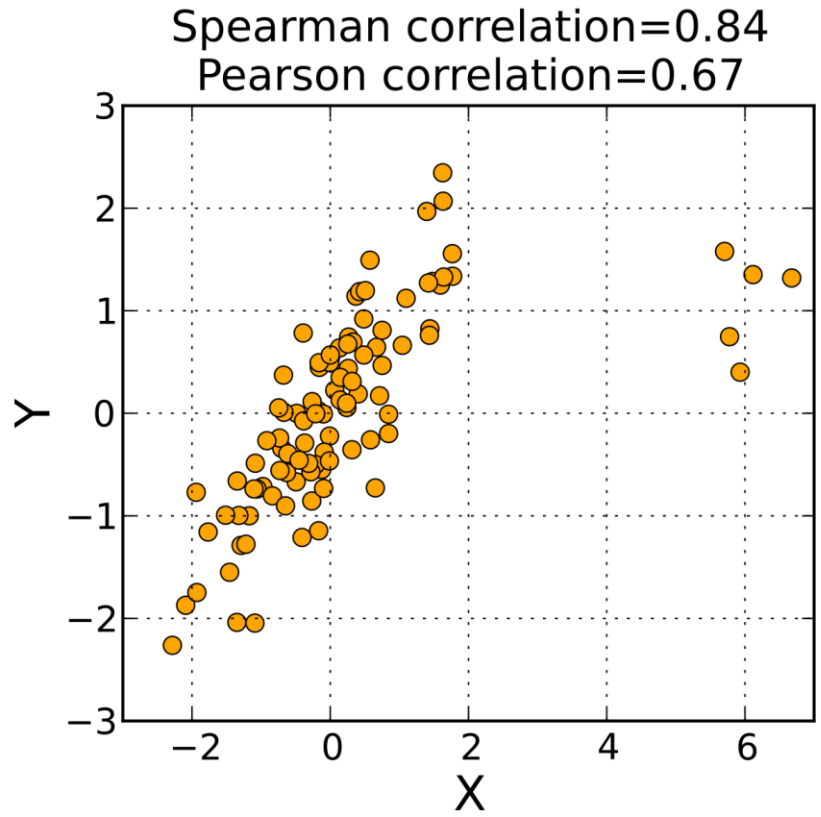
```
cor(x, method = "spearman")
```

**x is a data frame with 2 or more columns**

# Product-Moment vs Rank Correlation




US Army Corps of Engineers®

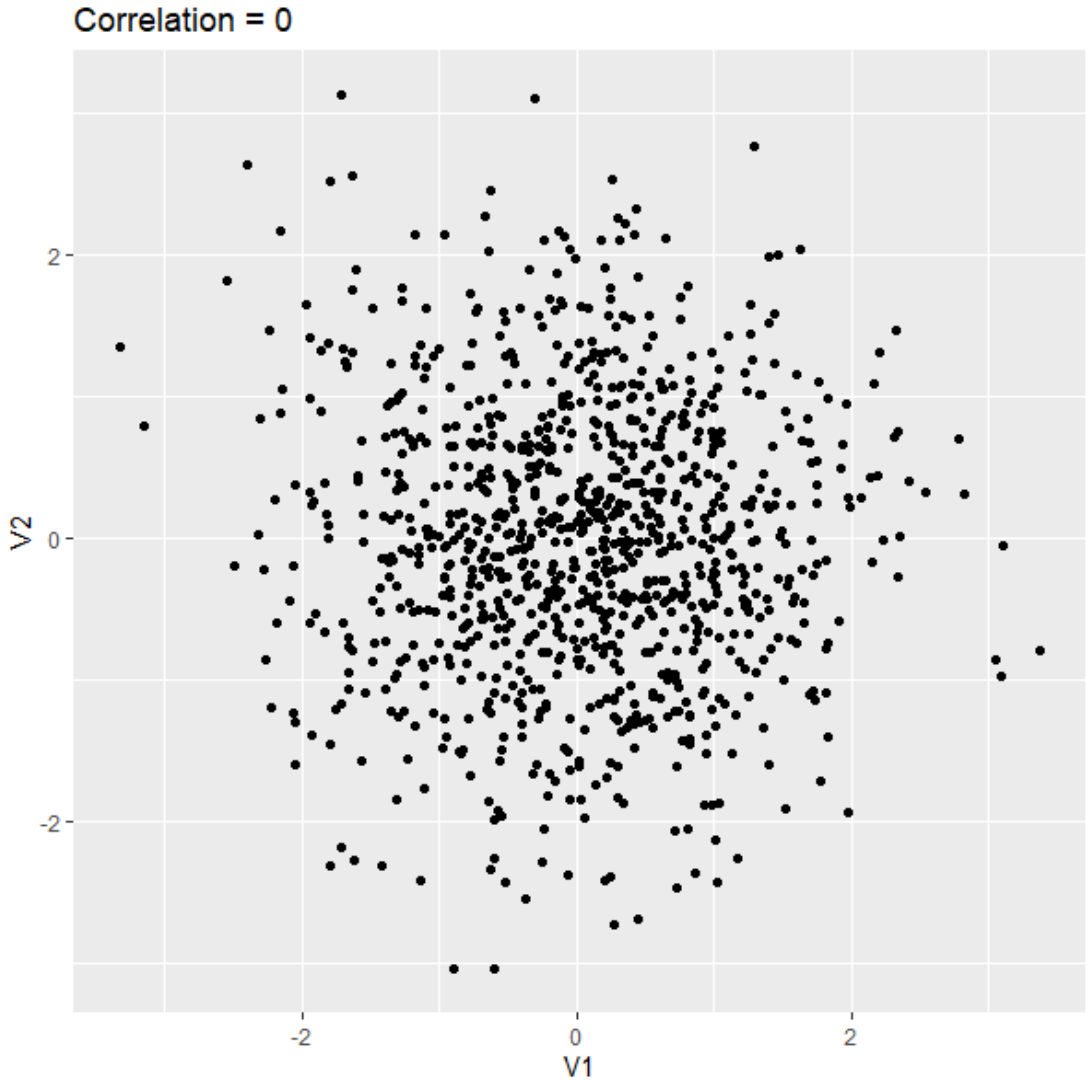
# Visualizing Correlation

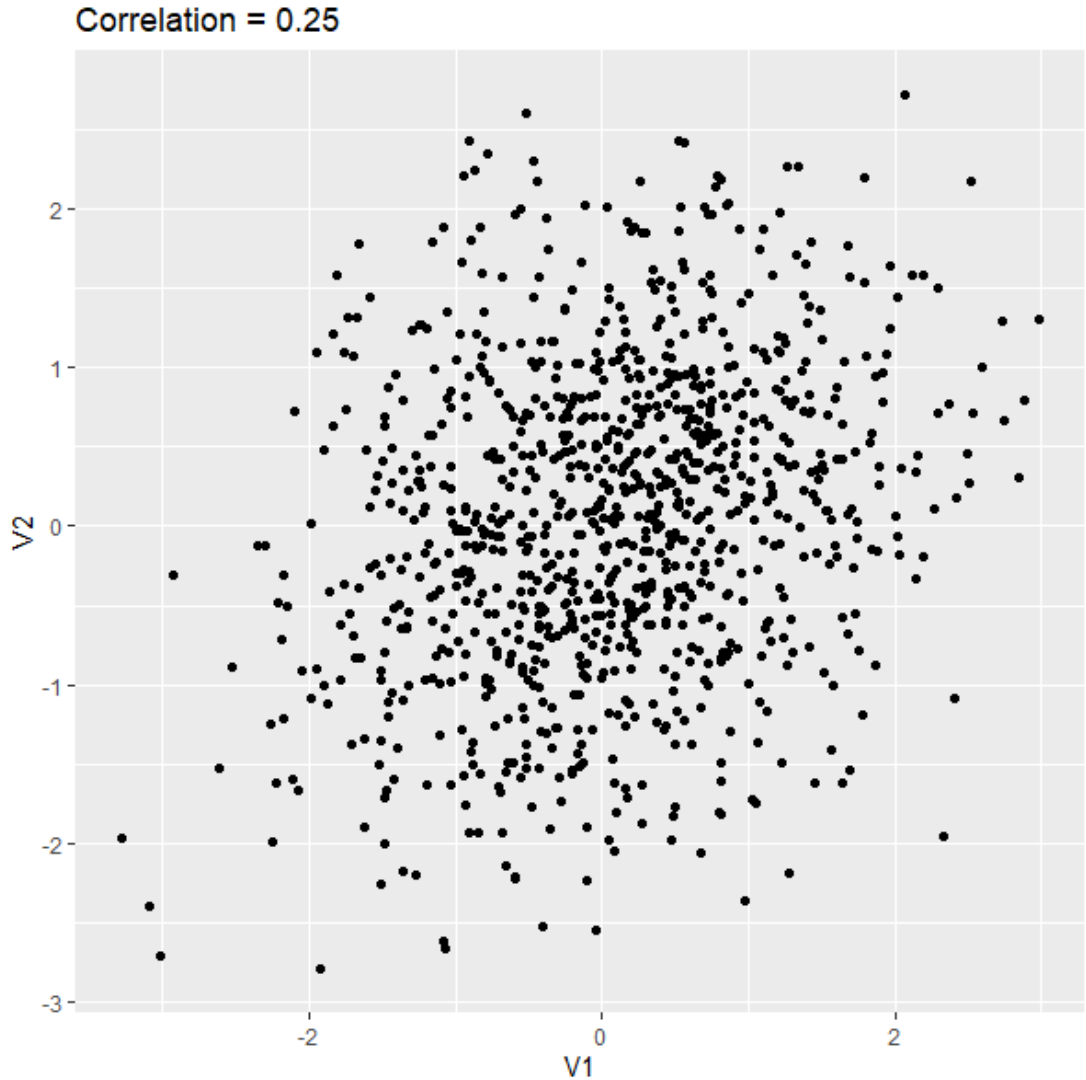
US Army Corps
of Engineers®

# Visualizing Correlation

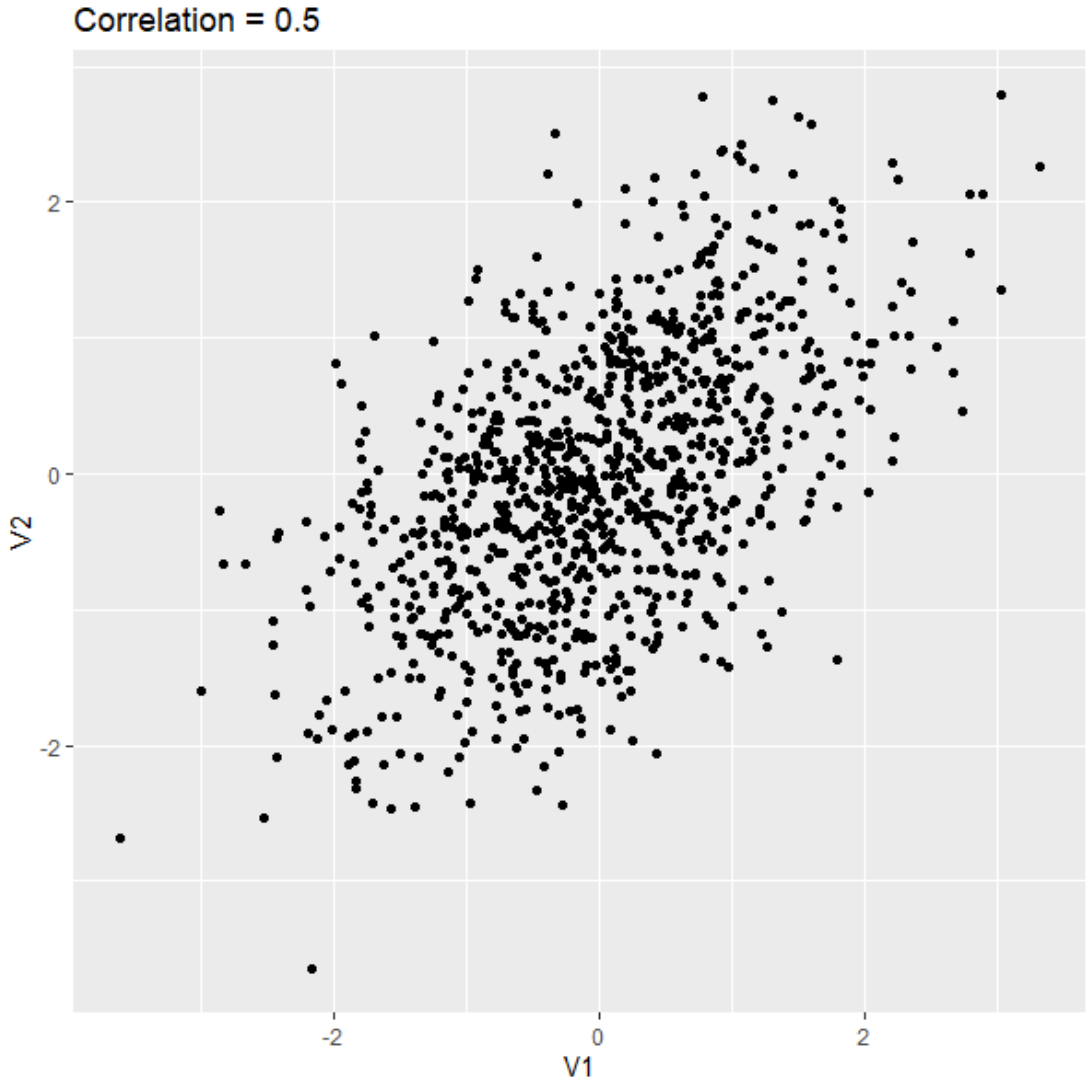
- Is there a pattern between X and Y?
- Is it positive (large X tends with large Y)?
- Or is it negative (large X tends with small Y)?
- Is it linear?
  - Can it be transformed?

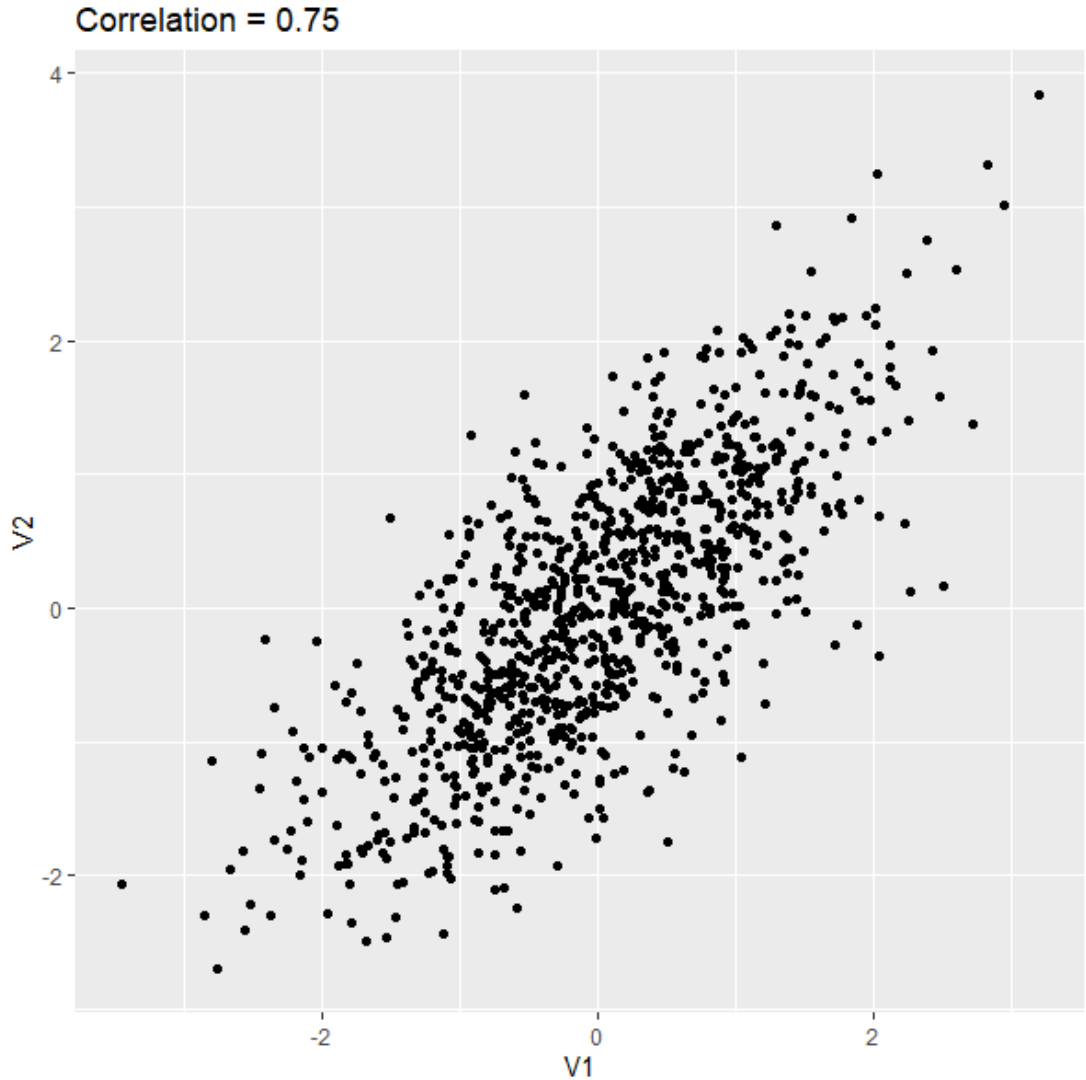
US Army Corps of Engineers®



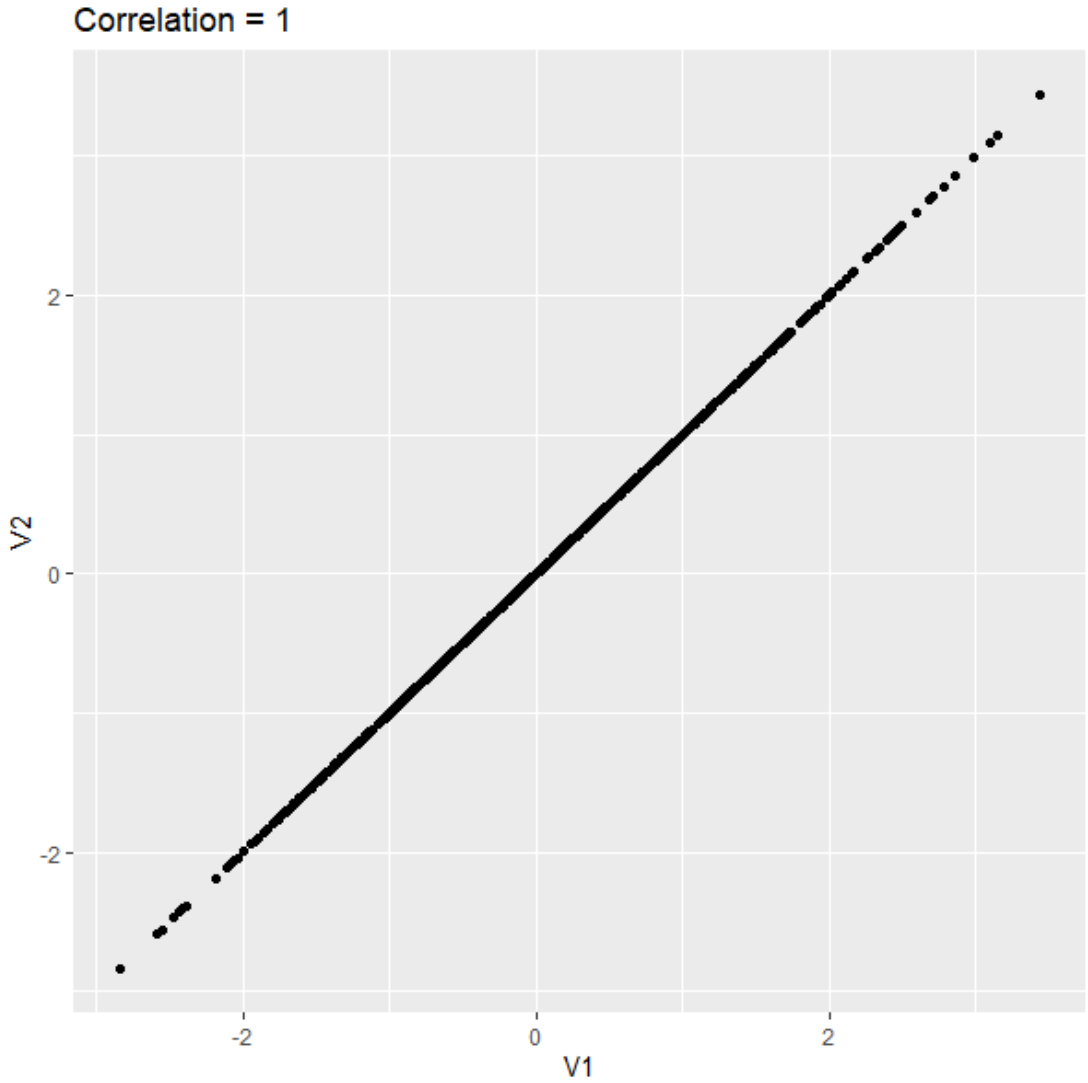



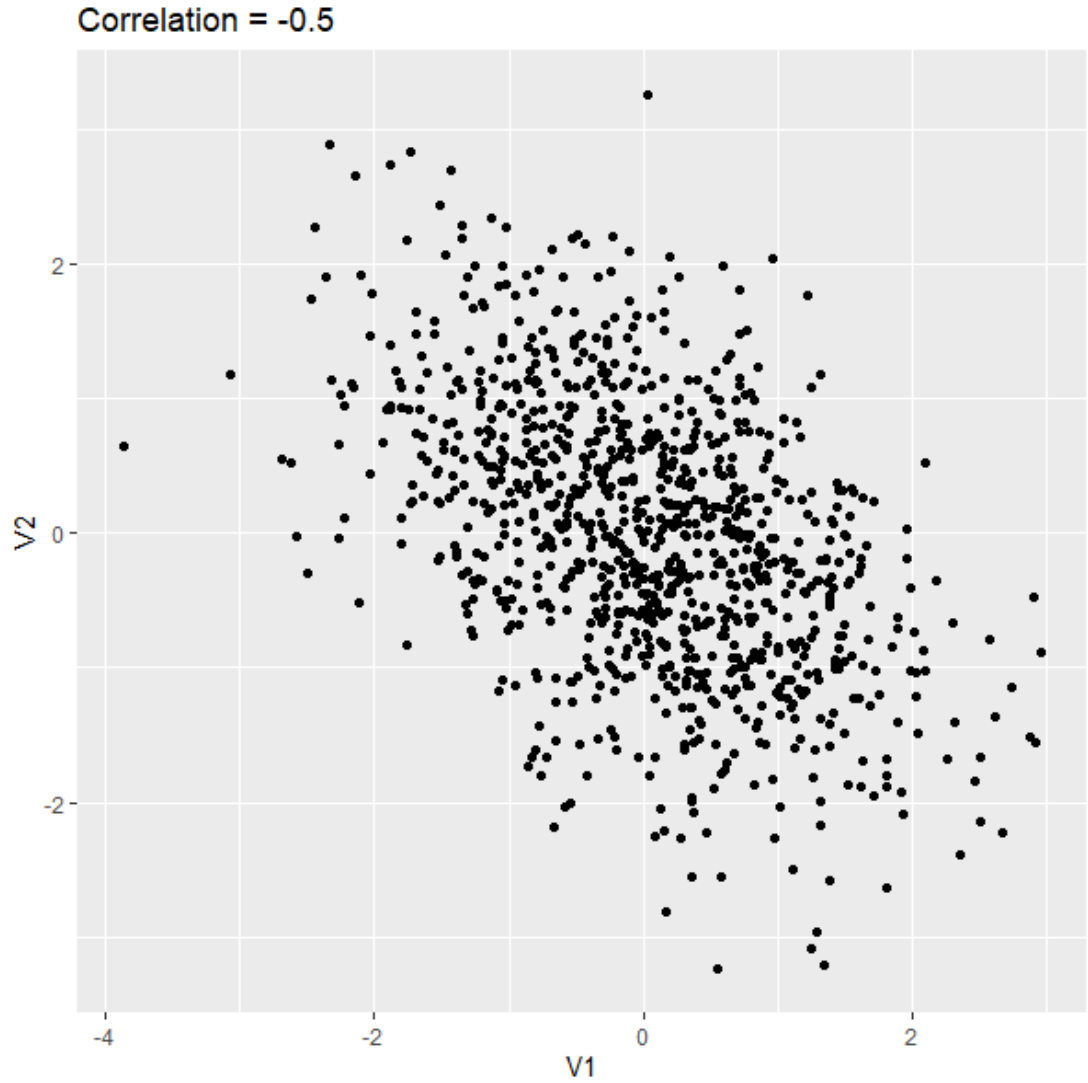

US Army Corps of Engineers®



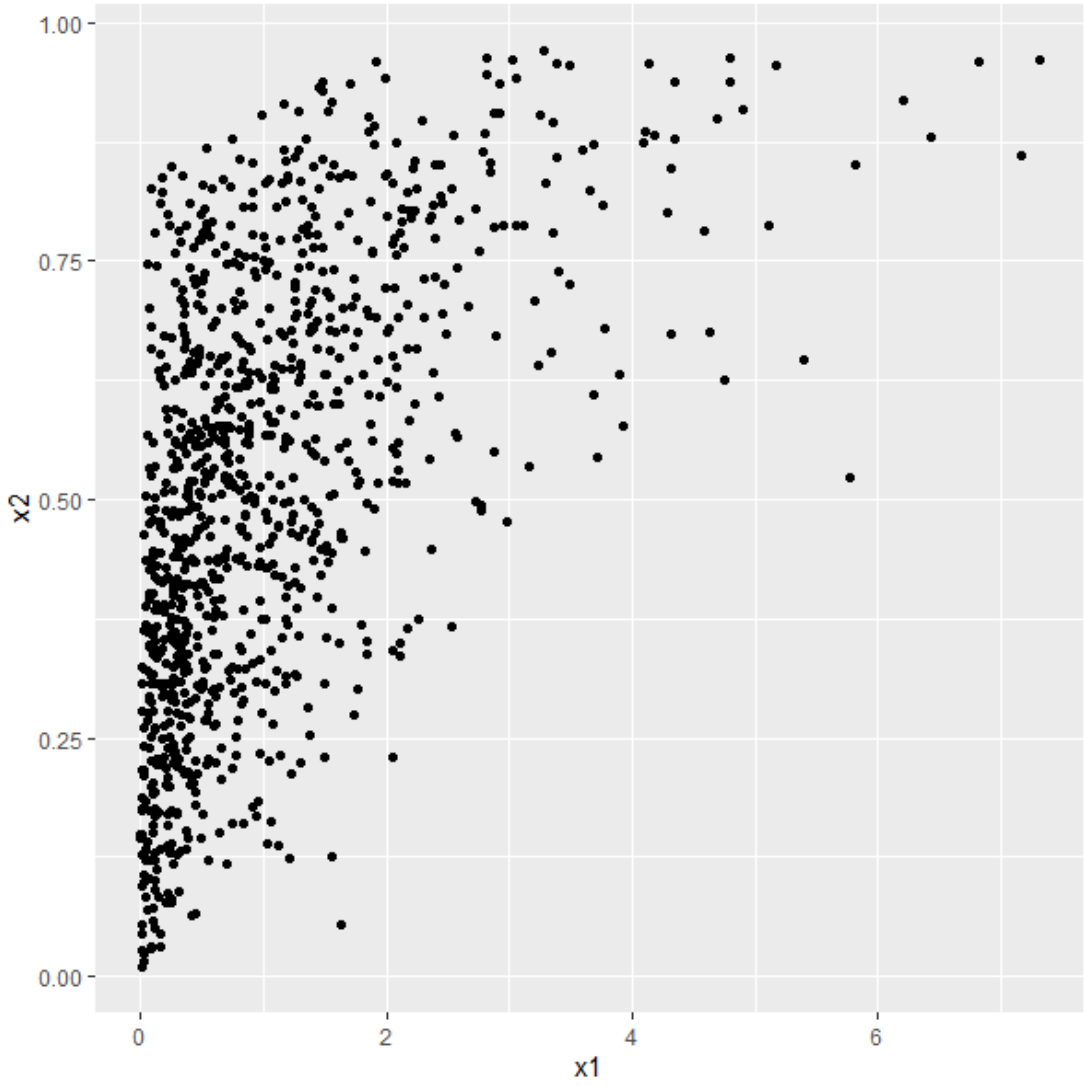

US Army Corps of Engineers®




US Army Corps of Engineers®

# Is this linear?

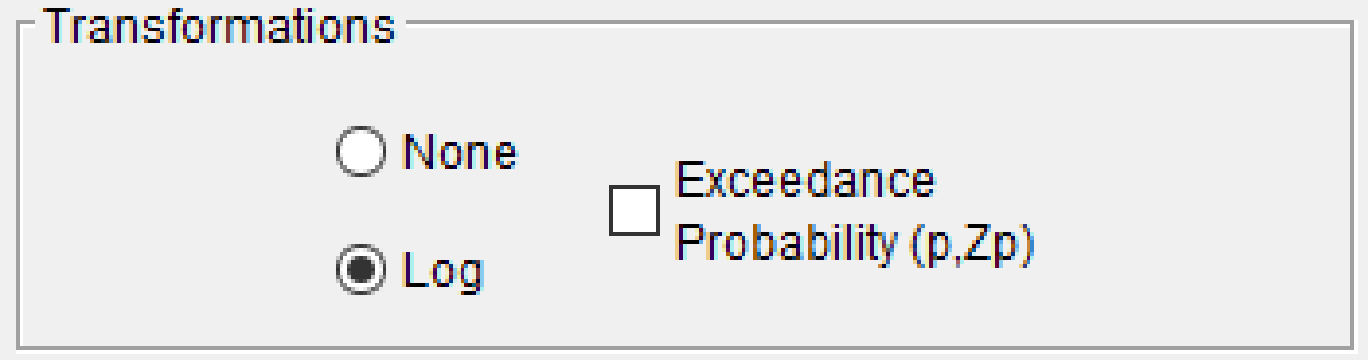



**r (Pearson) = 0.55**
**ρ (Spearman) = 0.59**

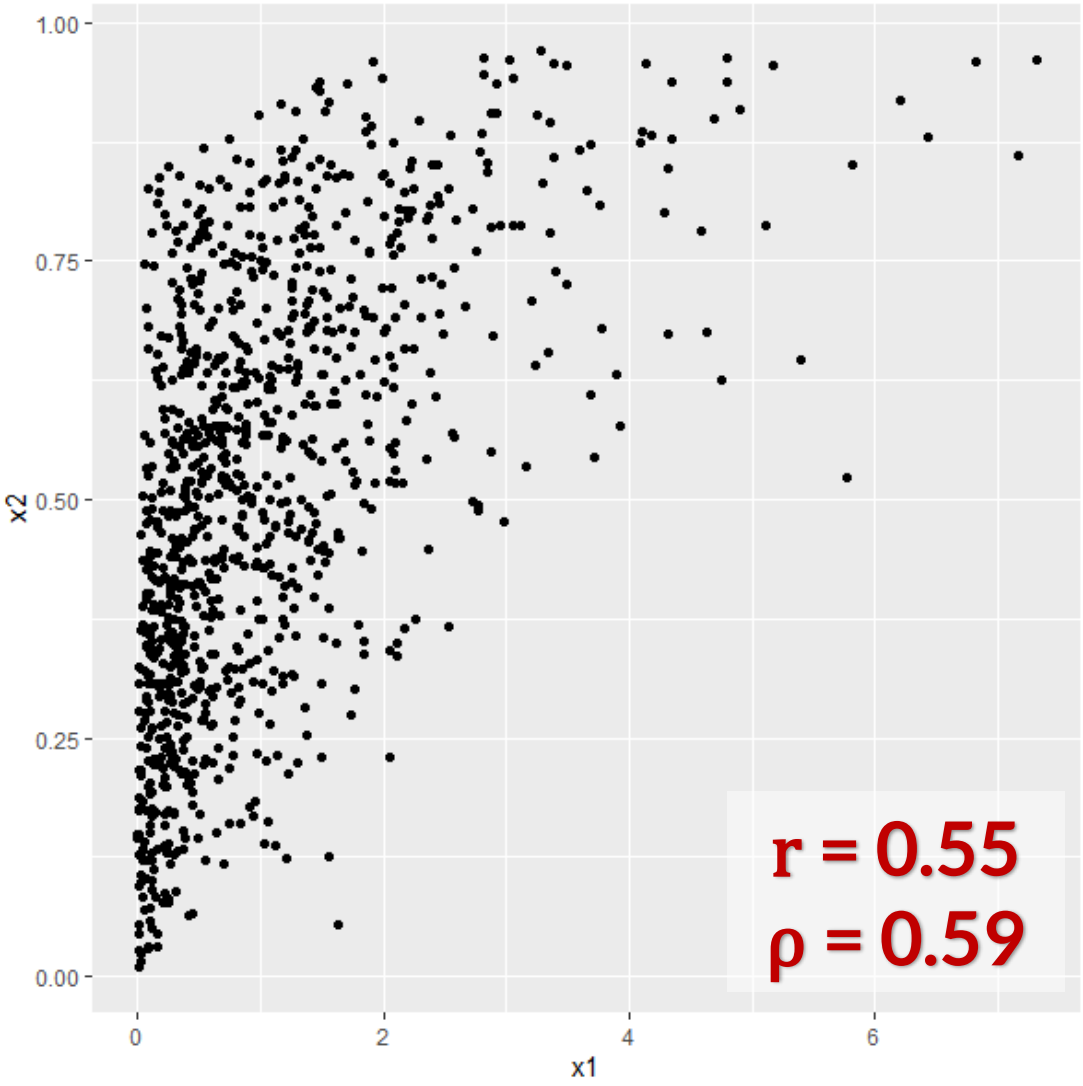
US Army Corps of Engineers®

# Transformation

- Use a monotonic function on the variables to improve linearity
  - Only affects Pearson (linear) correlation!
- Two common functions:
  - Logarithm
  - Exceedance probability/standard normal
    - (requires a probability distribution)



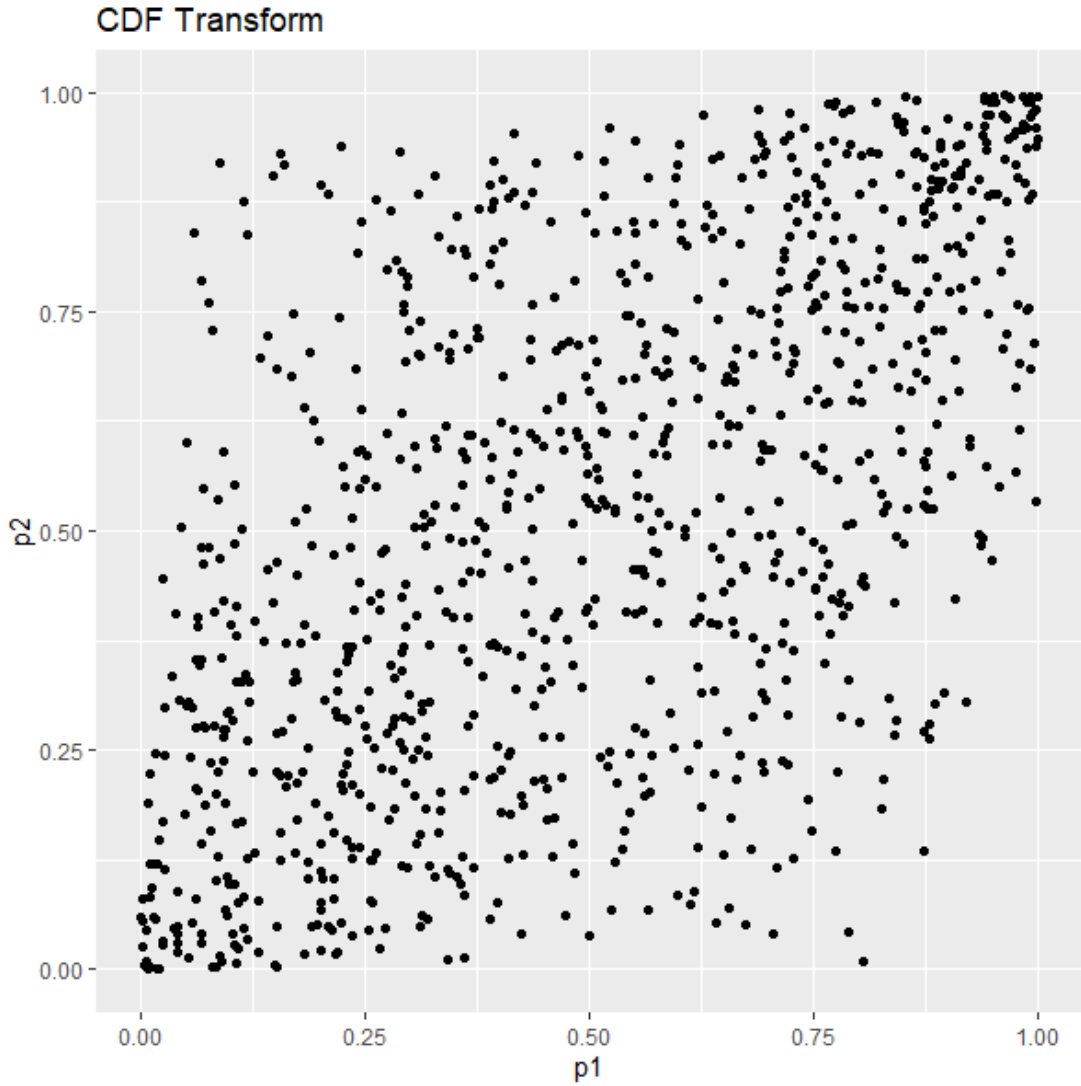

US Army Corps of Engineers®




$F(x2) \sim$ Beta(2,2)

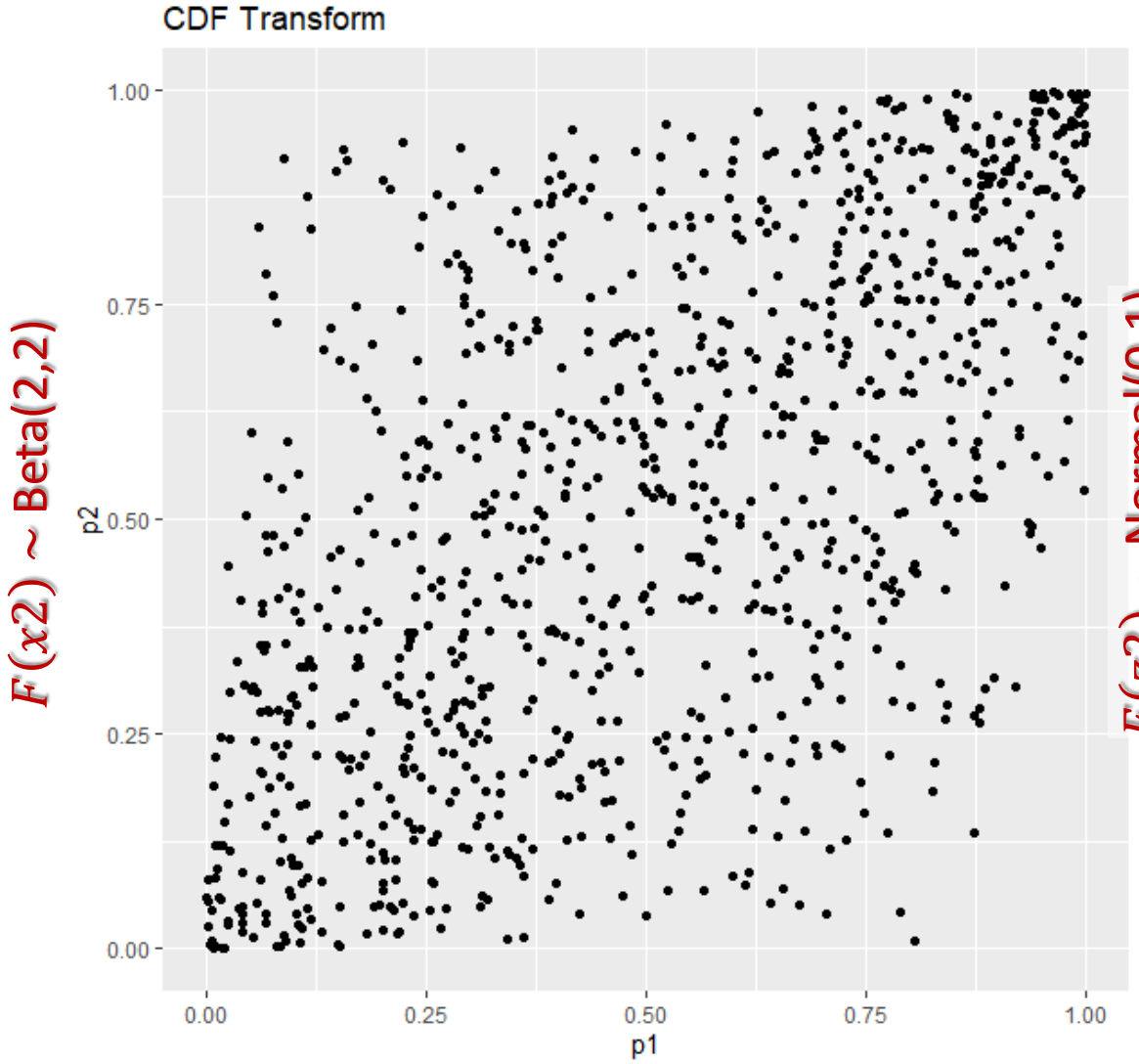
r = 0.55
ρ = 0.59

$F(x1) \sim$ Gamma(1,1)




r = 0.59 ρ = 0.59

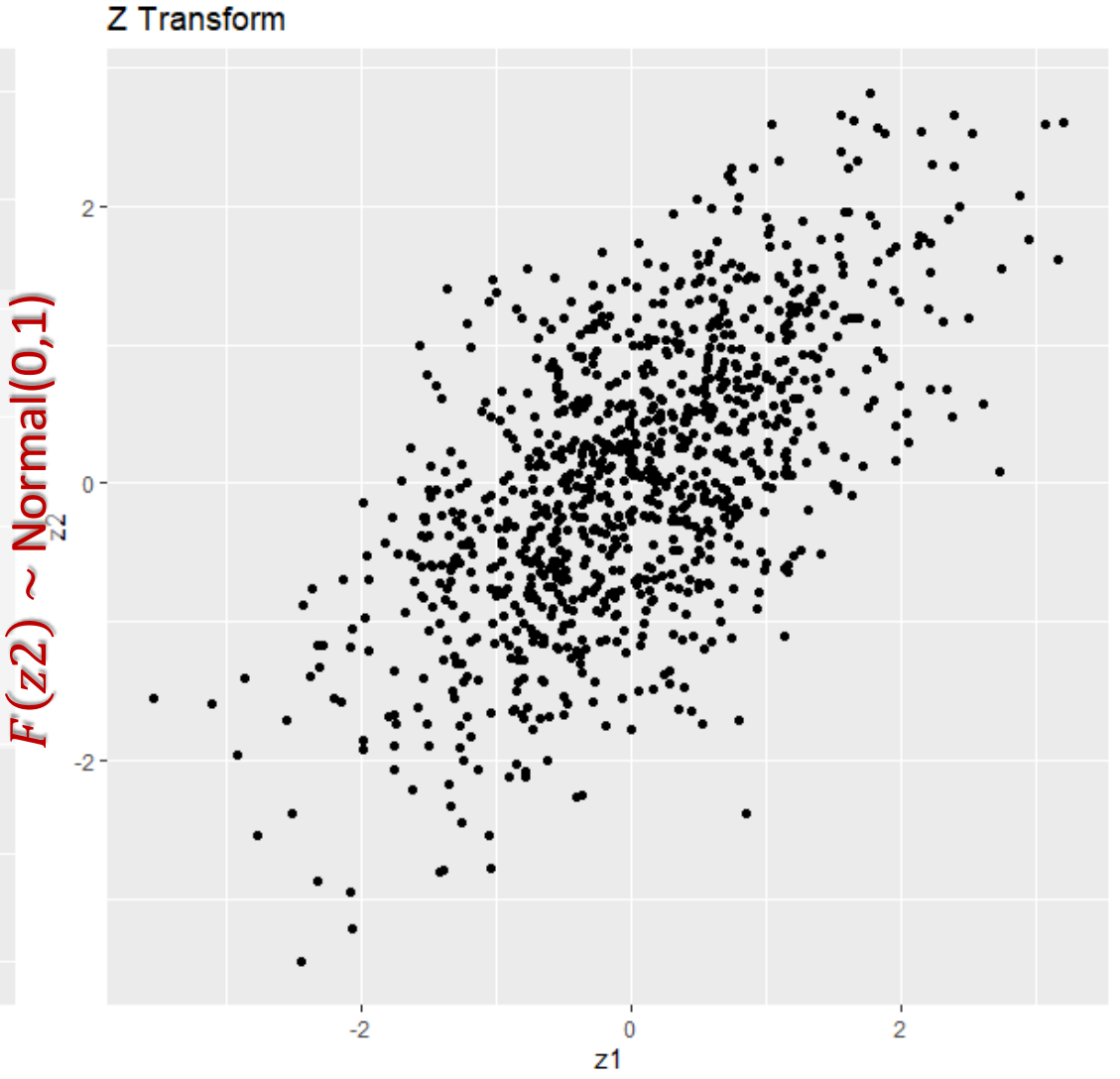
US Army Corps of Engineers®




$F(x2) \sim$ Beta(2,2)

$F(x1) \sim$ Gamma(1,1)

**r = 0.59 ρ = 0.59**




$F(z2) \sim$ Normal(0,1)

$F(z1) \sim$ Normal(0,1)

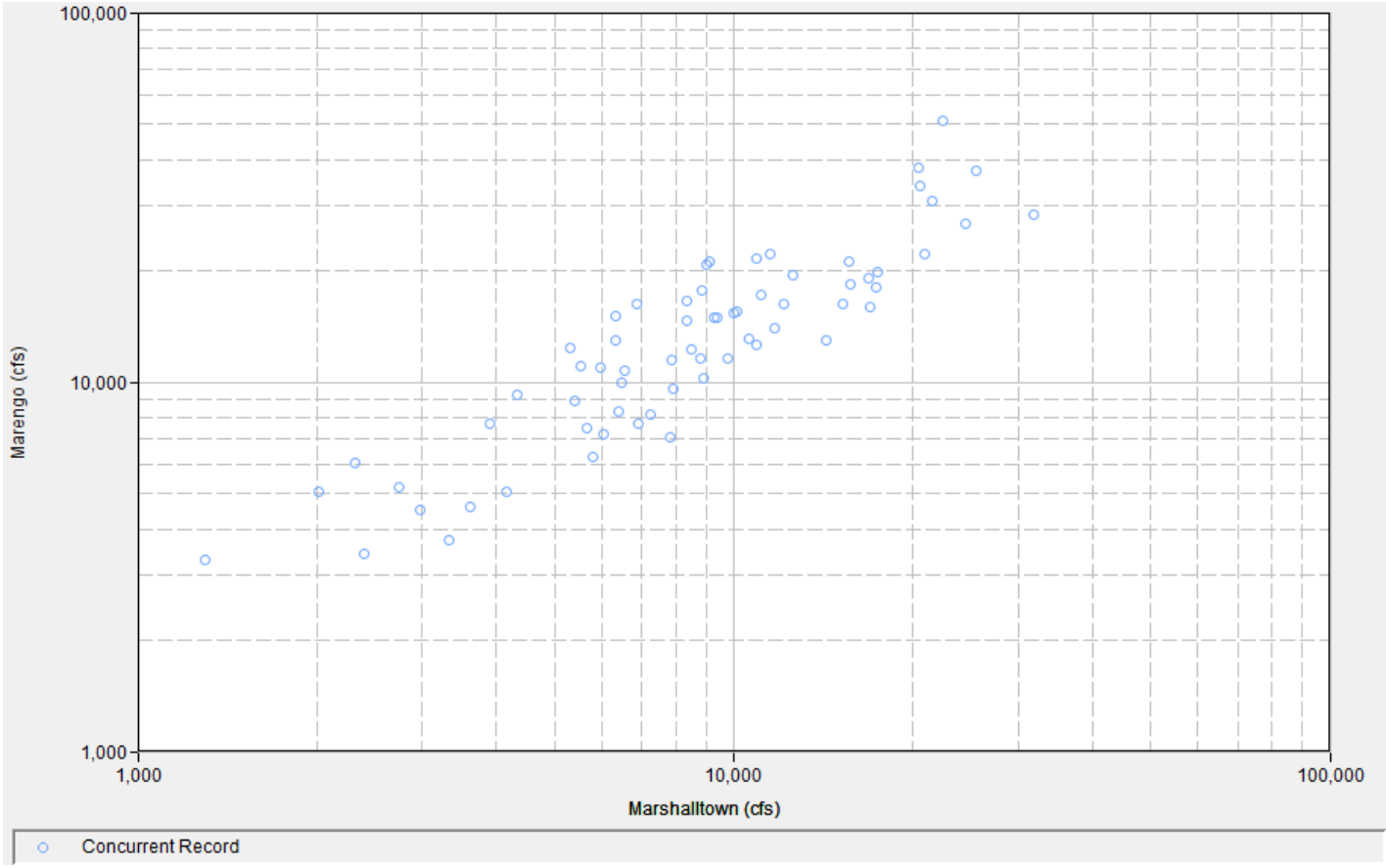
**r = 0.62 ρ = 0.59**

US Army Corps of Engineers®

# Exceedance Probability Transform

Correlation in two-site peak streamflow?

- Start with log transform
- Consider exceedance probability transform
  - Fitted log-Pearson type III distribution




US Army Corps of Engineers®

# Questions?

US Army Corps
of Engineers®