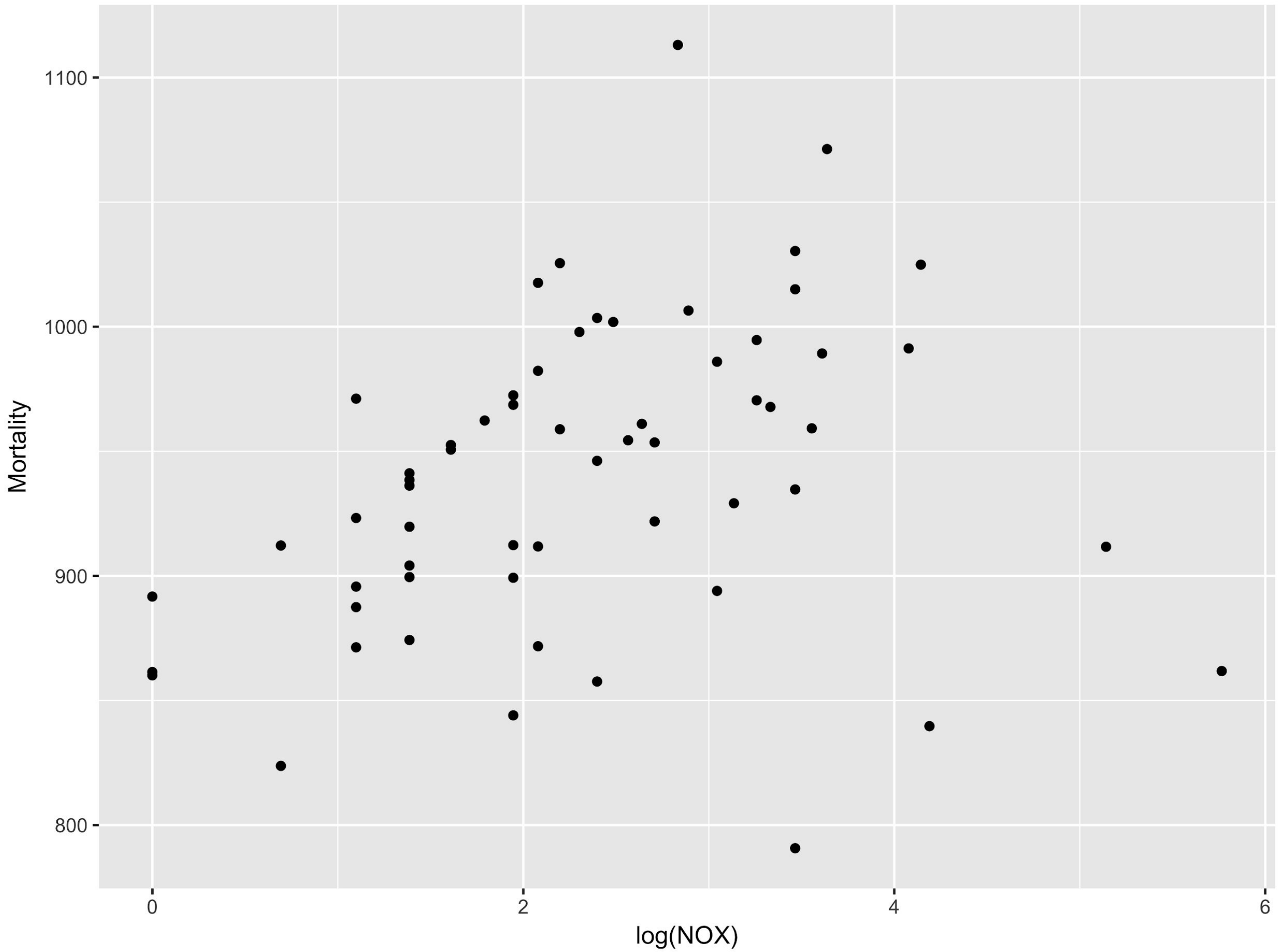
# Training Machines to See What You See

# Let’s start with some data

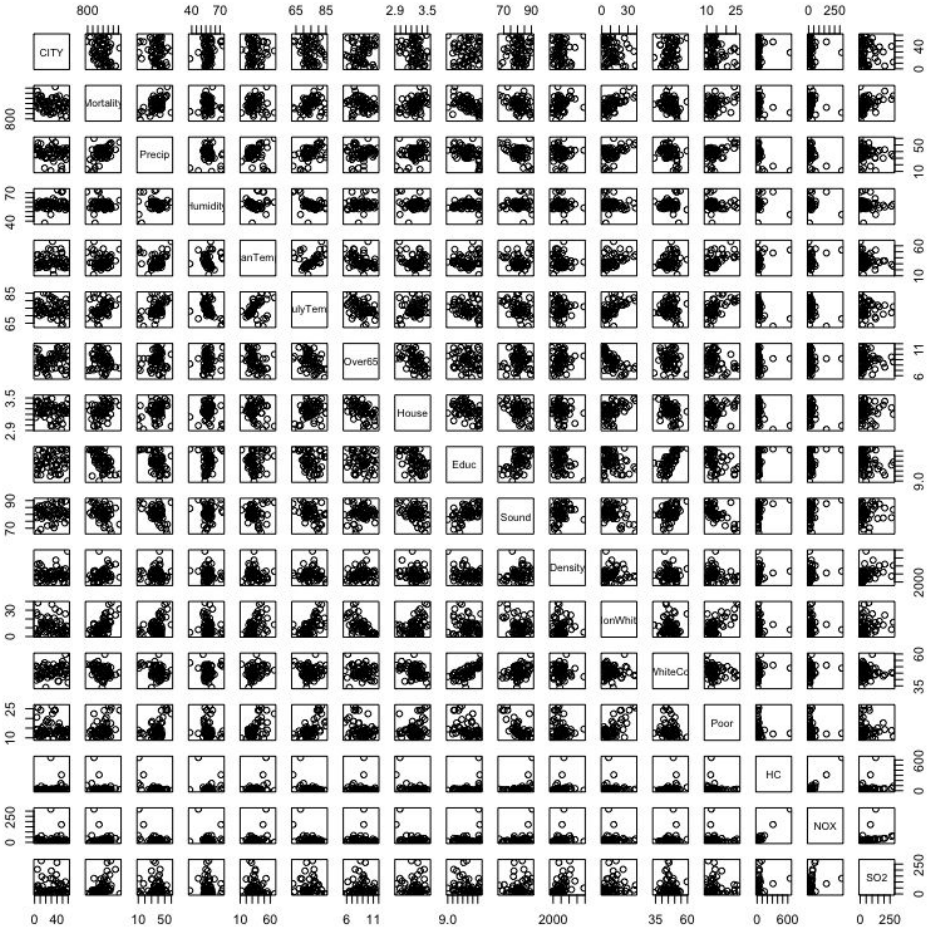
# How does air pollution affect mortality?

Data set with both pollution and socioeconomic data for 5 Standard Metropolitan Statistical Areas in the U.S between 1959–1961.

60 observations of 17 variables.



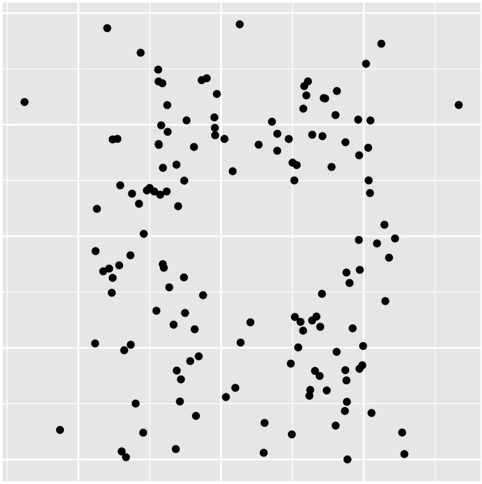
Mortality
1100
1000
900
800
0
2
4
6
log(NOX)



CITY
Mortality
Precip
Humidity
JanTemp
JulyTemp
Over65
House
Educ
Sound
Density
NonWhite
WhiteCol
Poor
HC
NOX
SO2

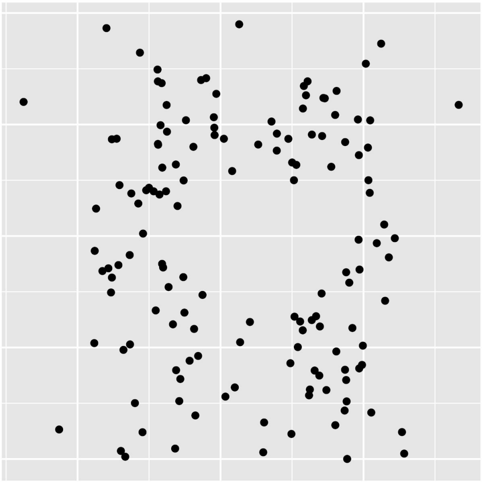
Can we train a computer to detect patterns more effectively and efficiently than humans?

# Summary Statistics



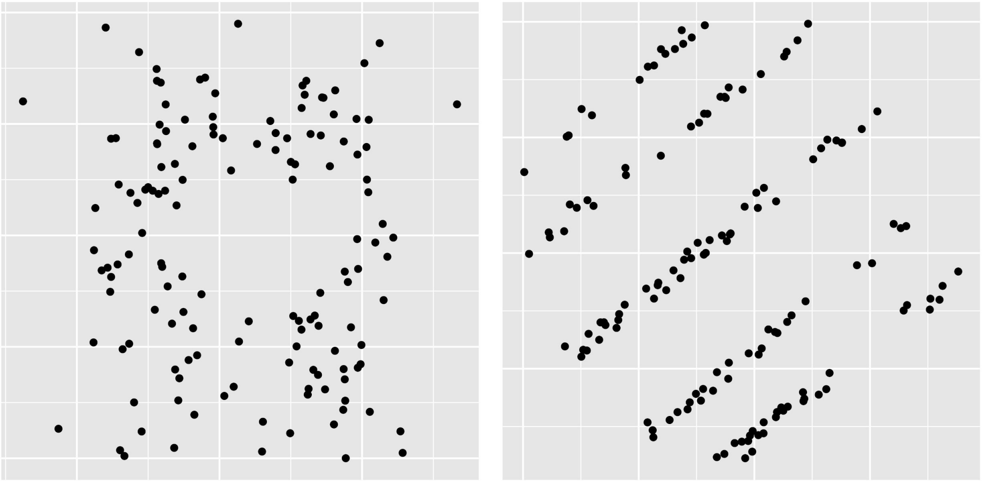
# Summary Statistics

| | |
|---|---|
| Mean(X) | 54.26 |
| Mean(Y) | 47.83 |
| Std.Dev(X) | 16.76 |
| Std.Dev(Y) | 26.93 |
| Correlation | -0.06 |



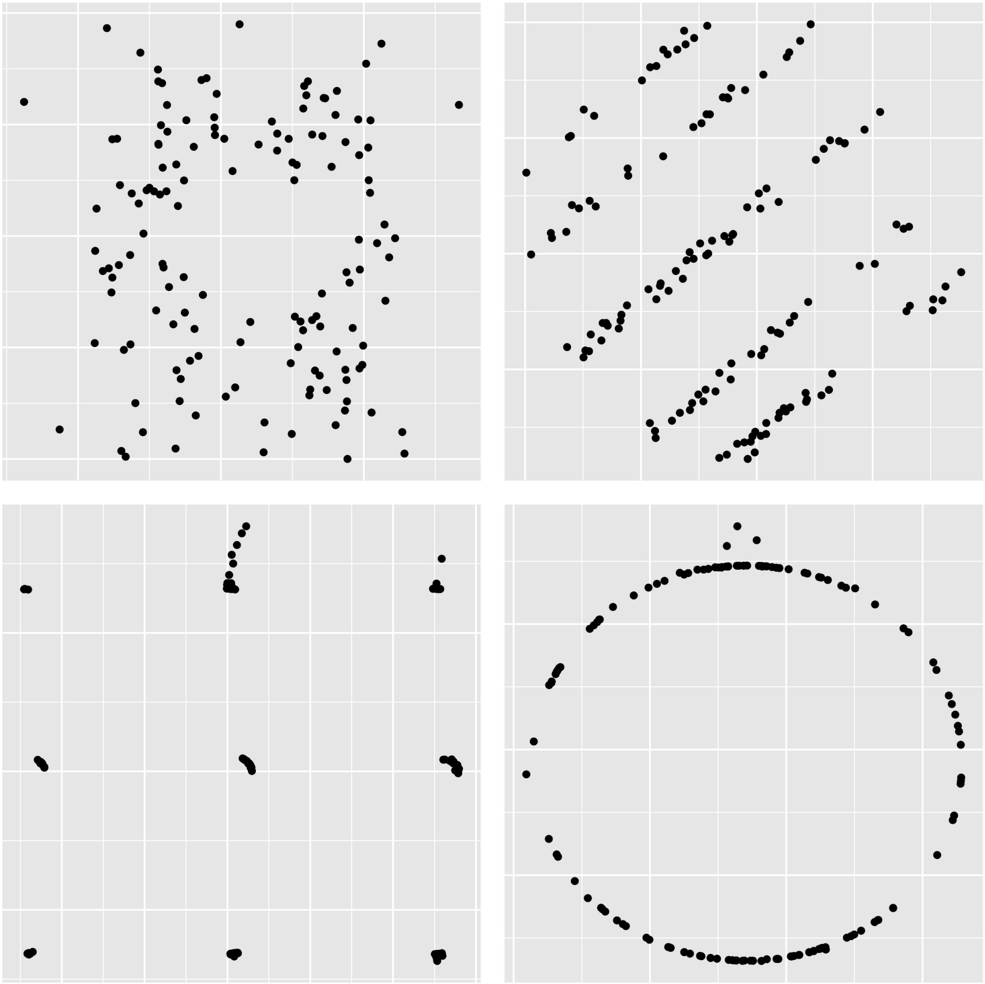
# Summary Statistics

| | |
|---|---|
| Mean(X) | 54.26 |
| Mean(Y) | 47.83 |
| Std.Dev(X) | 16.76 |
| Std.Dev(Y) | 26.93 |
| Correlation | -0.06 |

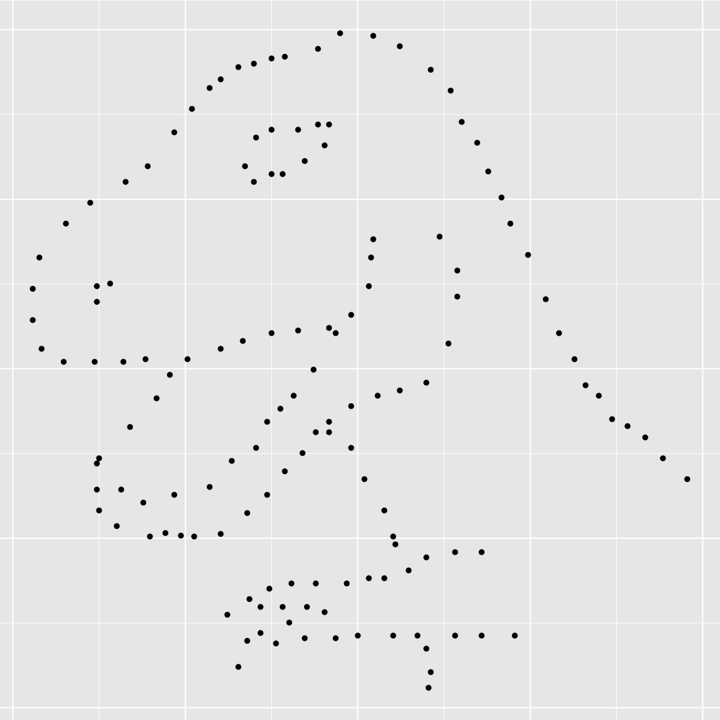


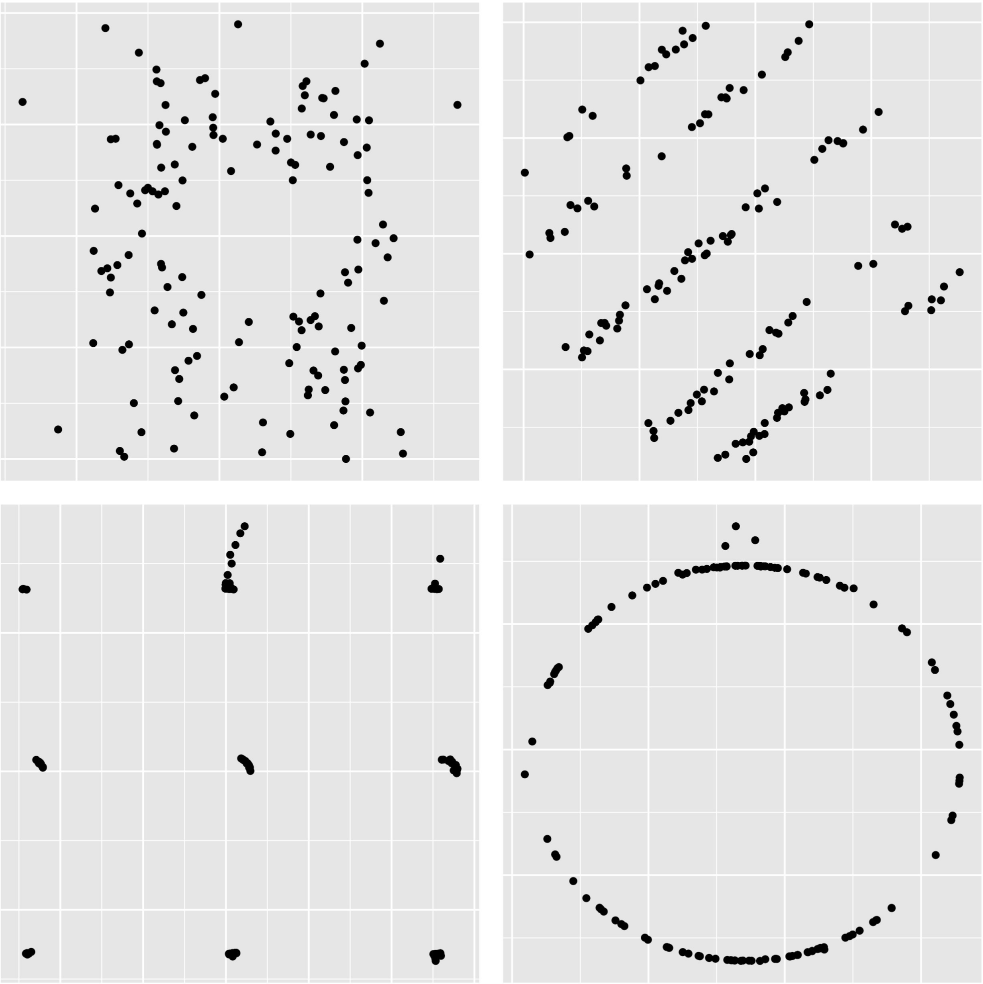
# Summary Statistics

| | |
|---|---|
| Mean(X) | 54.26 |
| Mean(Y) | 47.83 |
| Std.Dev(X) | 16.76 |
| Std.Dev(Y) | 26.93 |
| Correlation | -0.06 |

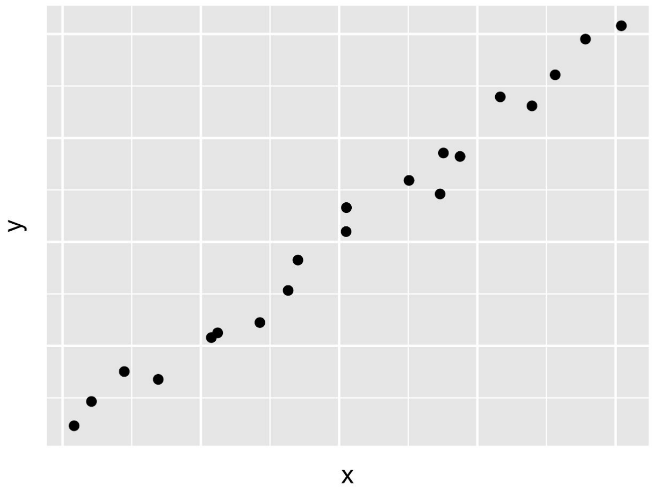


# Summary Statistics

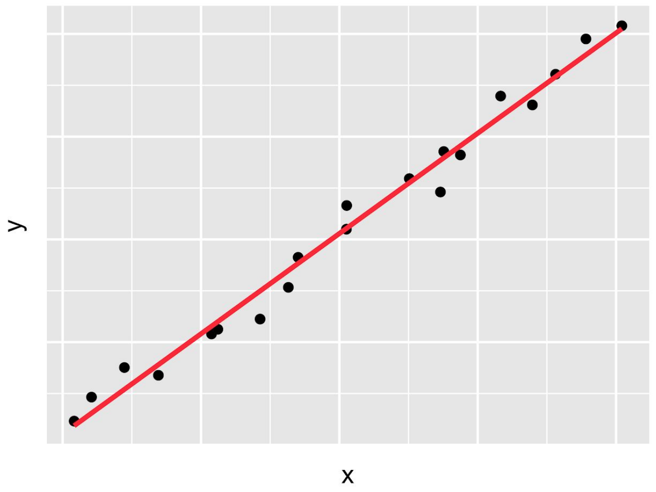


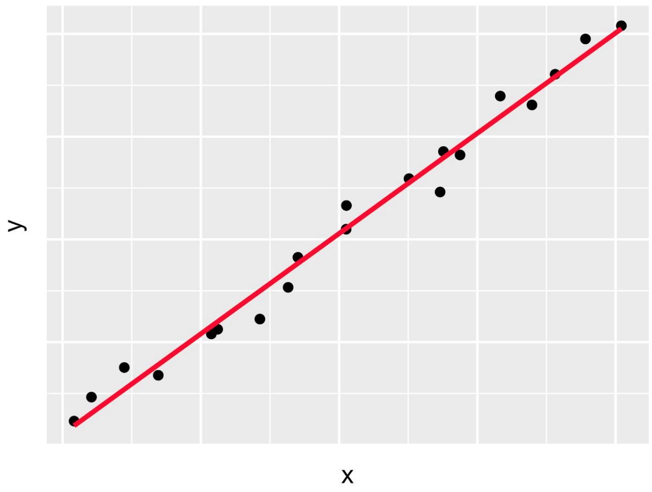


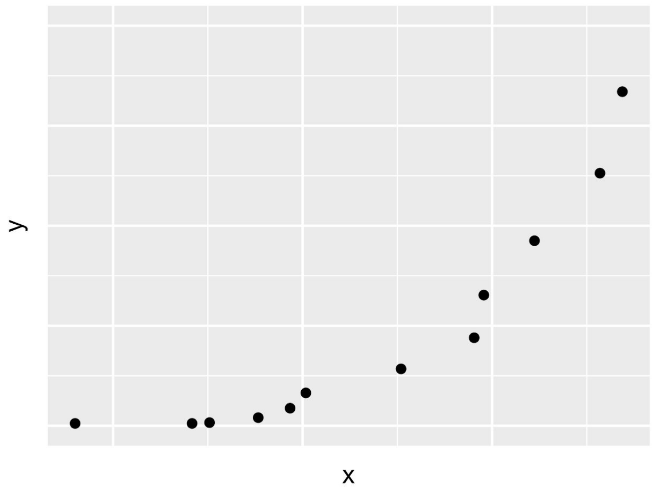
*Datasaurus Dozen*, Alberto Cairo

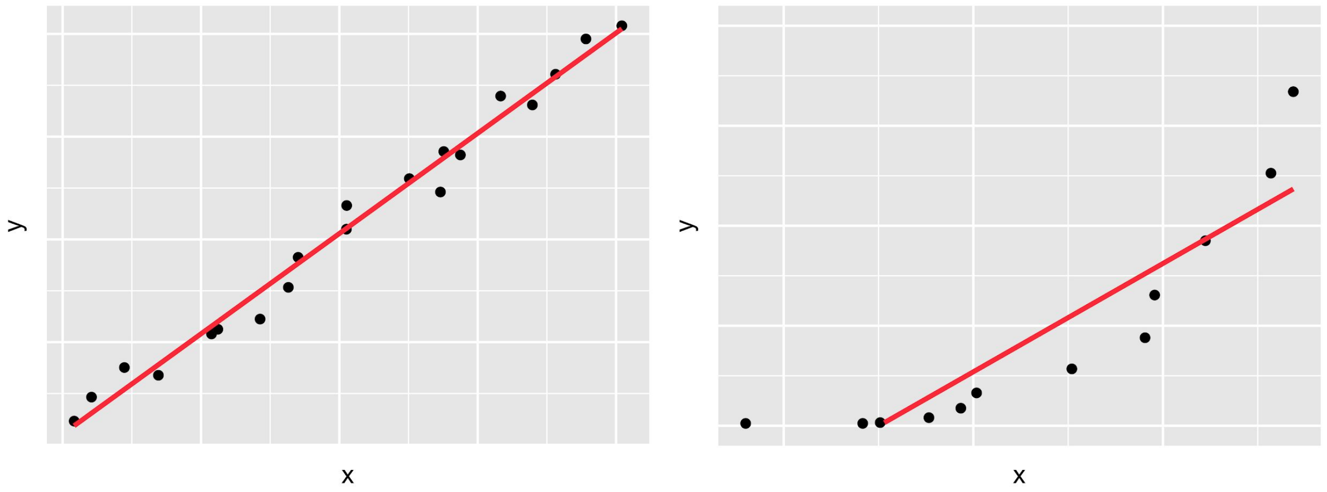


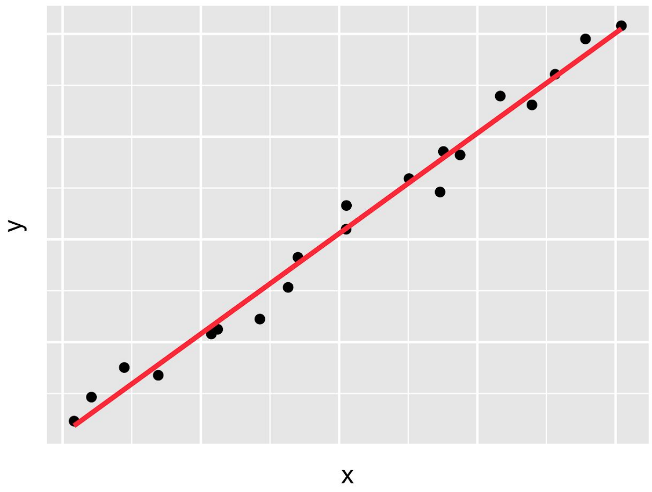
y
x

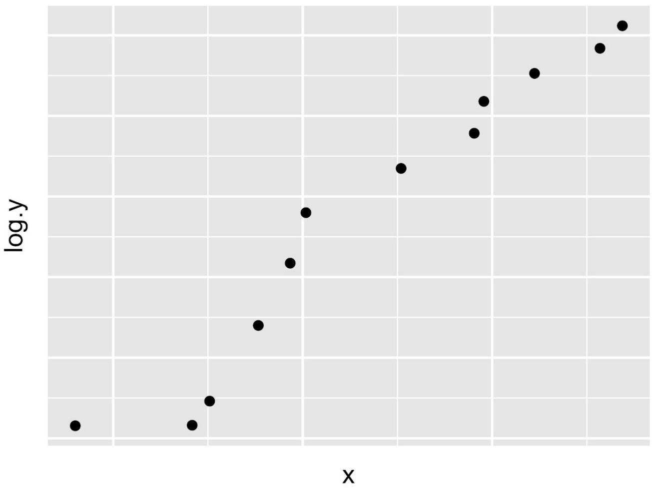


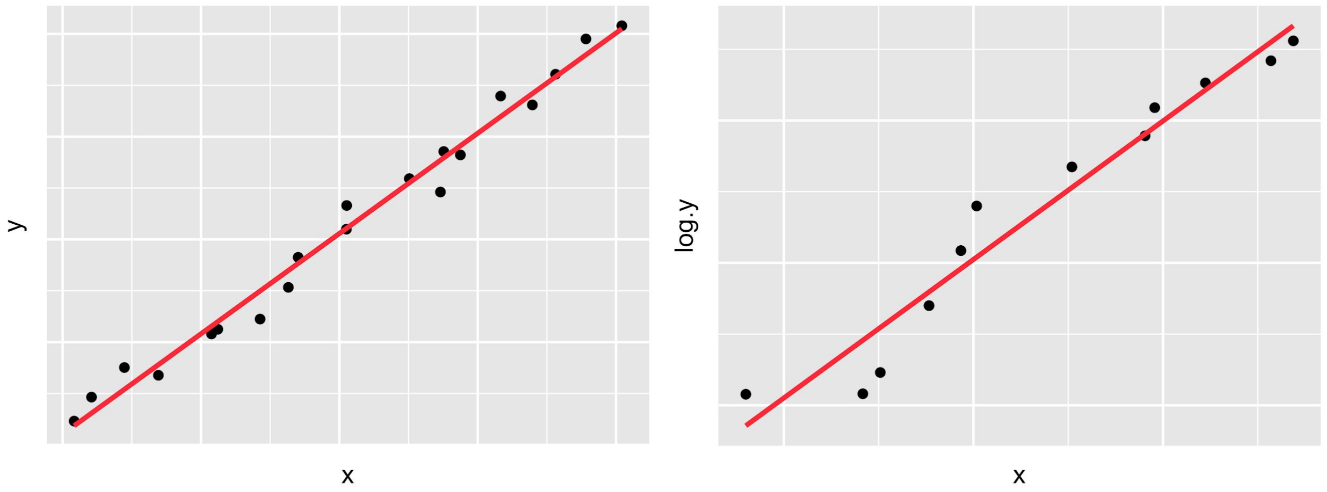

y
x

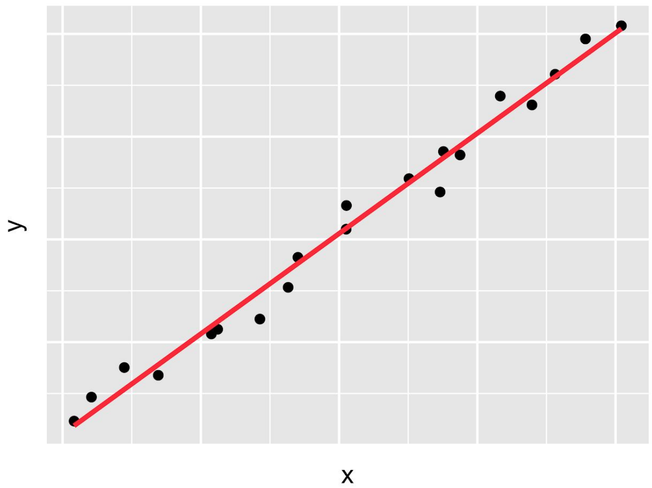

y
x


y
x
y
x

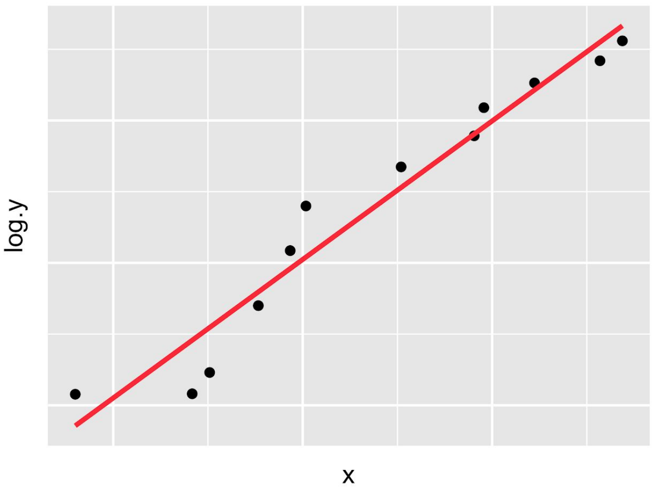

y
x


log.y
x

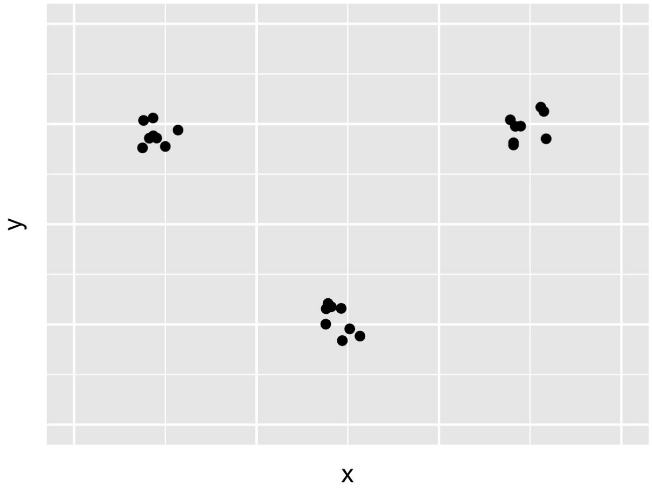

y
x
log.y
x



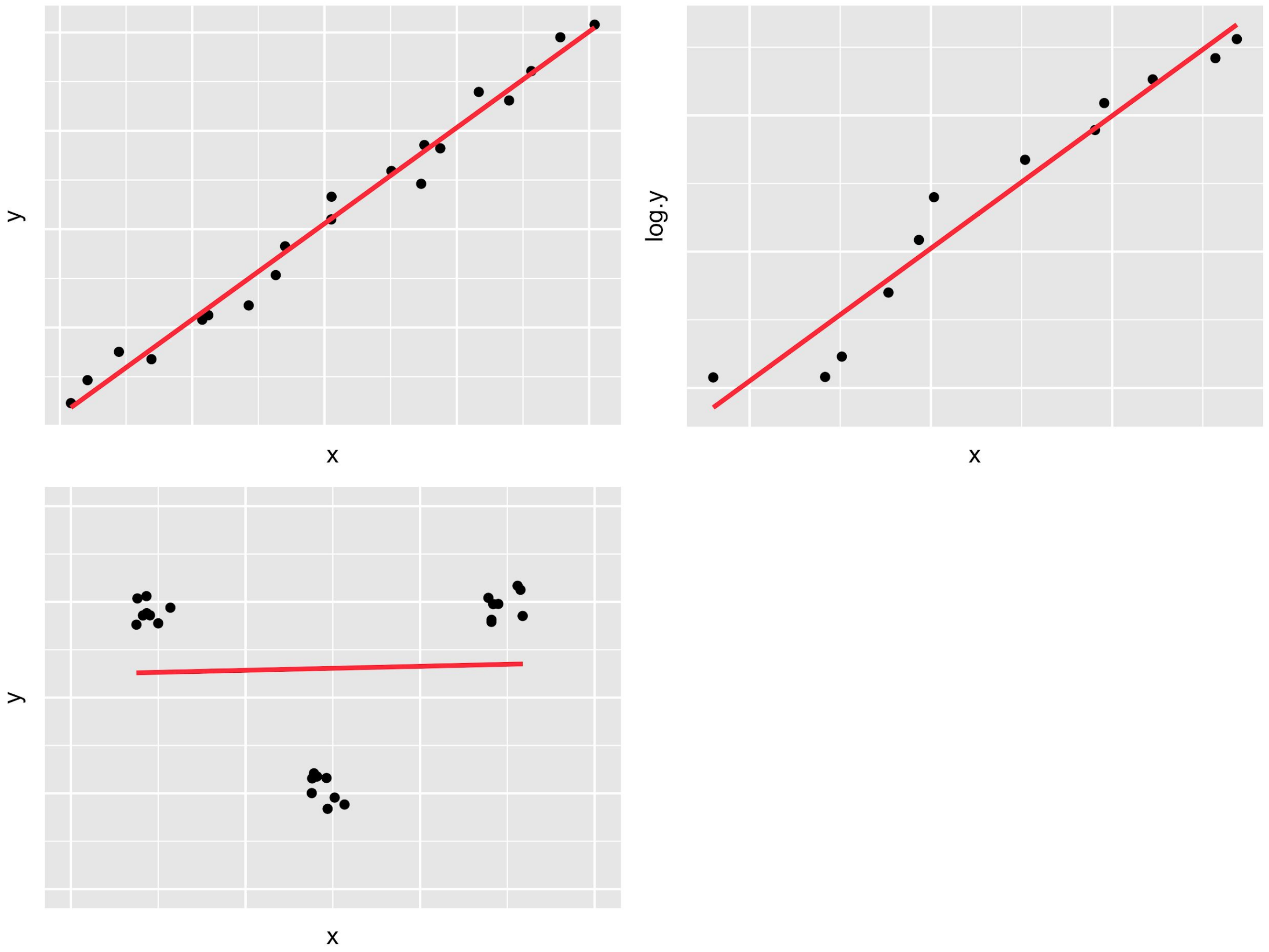
y
x

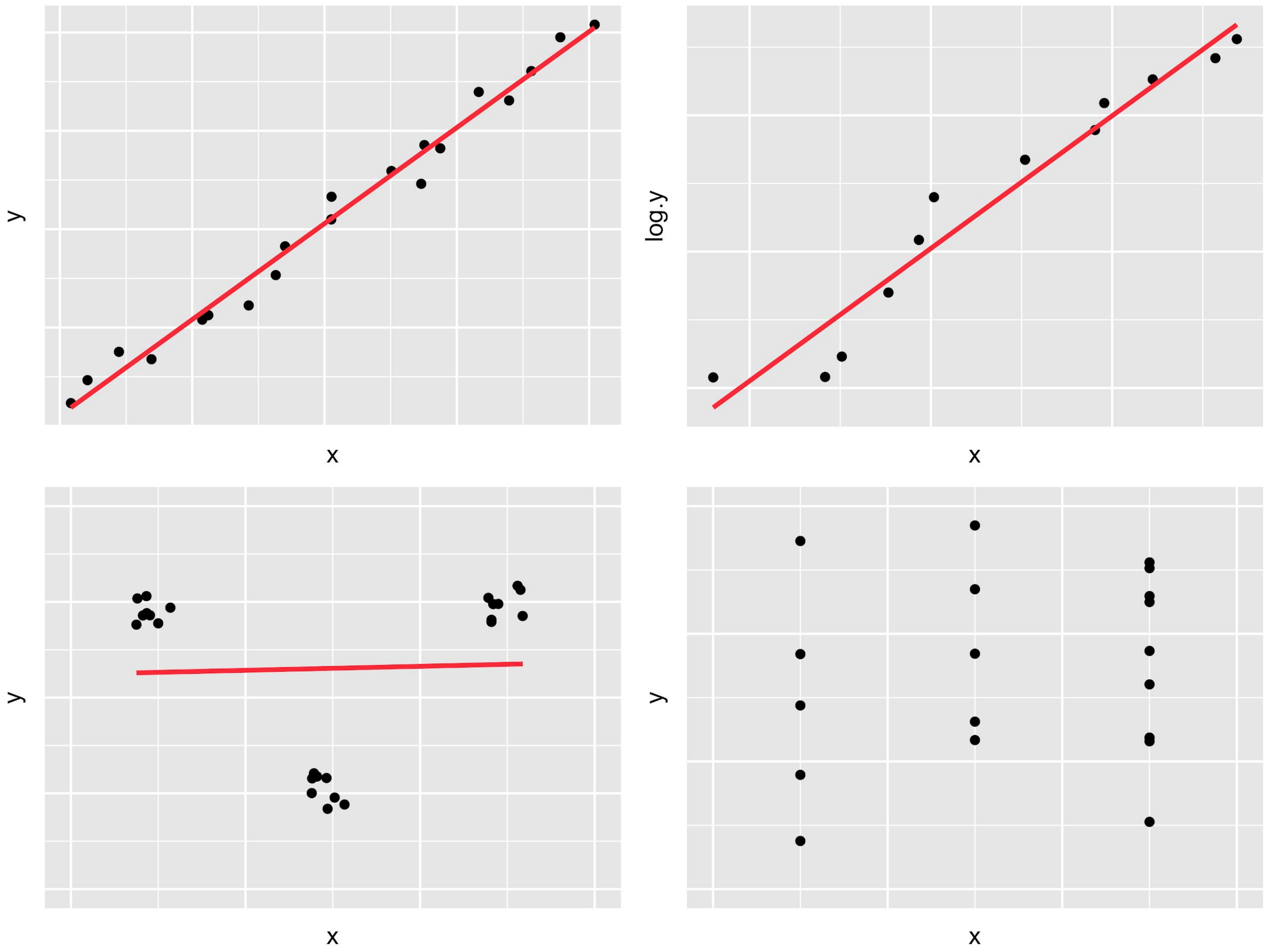



log.y
x

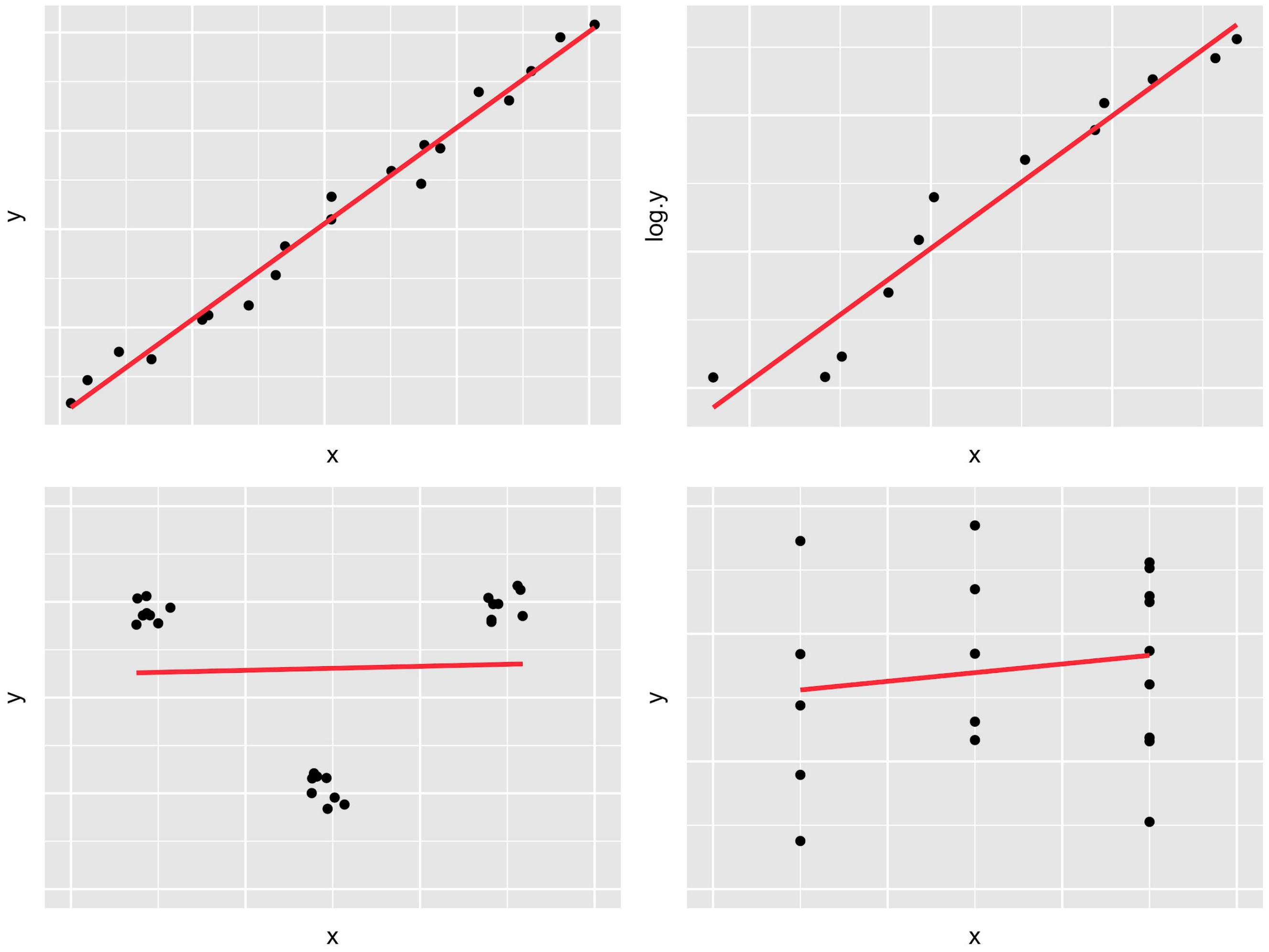



y
x

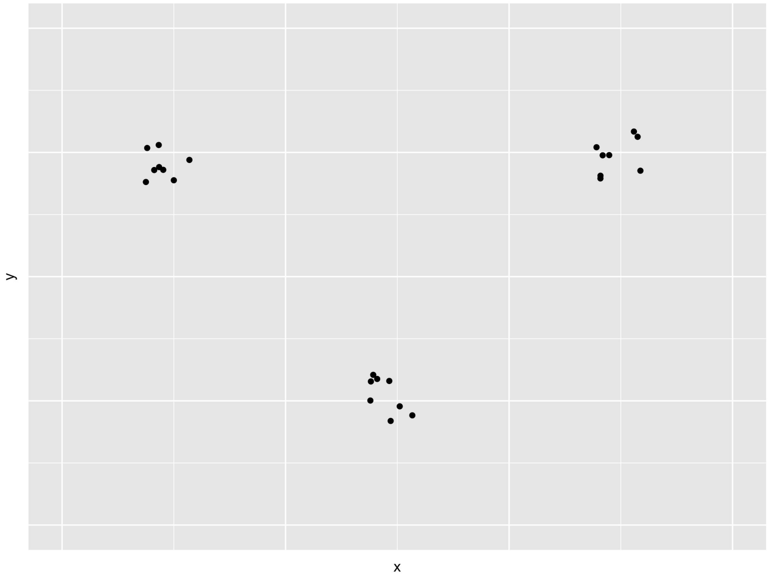

y
x
log.y
x
y
x

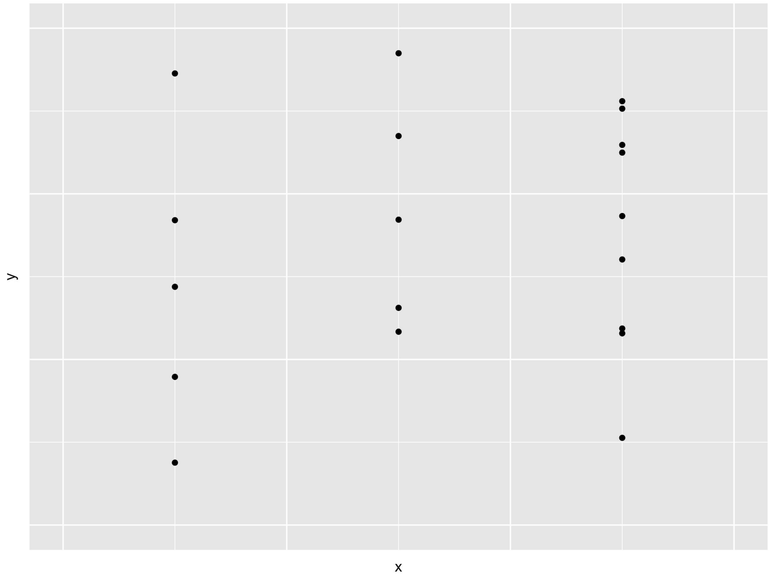


y
x
log.y
x
y
x
y
x


y
x
log.y
x
y
x
y
x

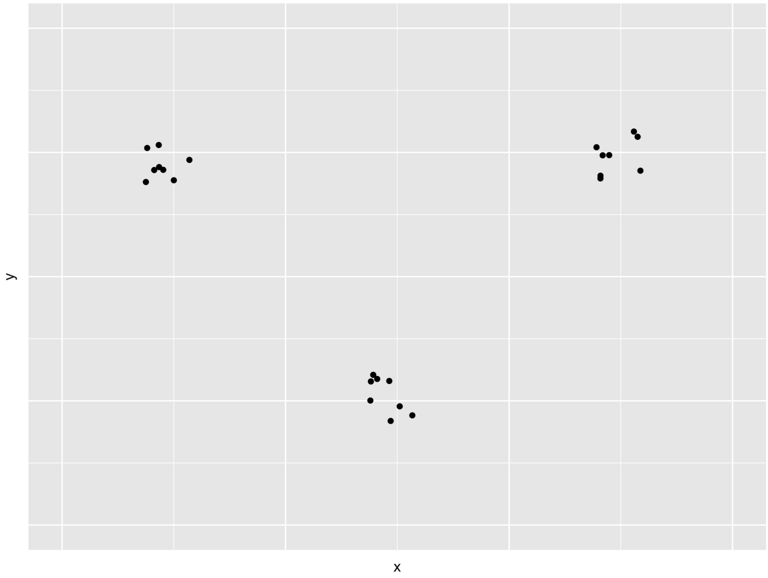
# How would you approach this problem?

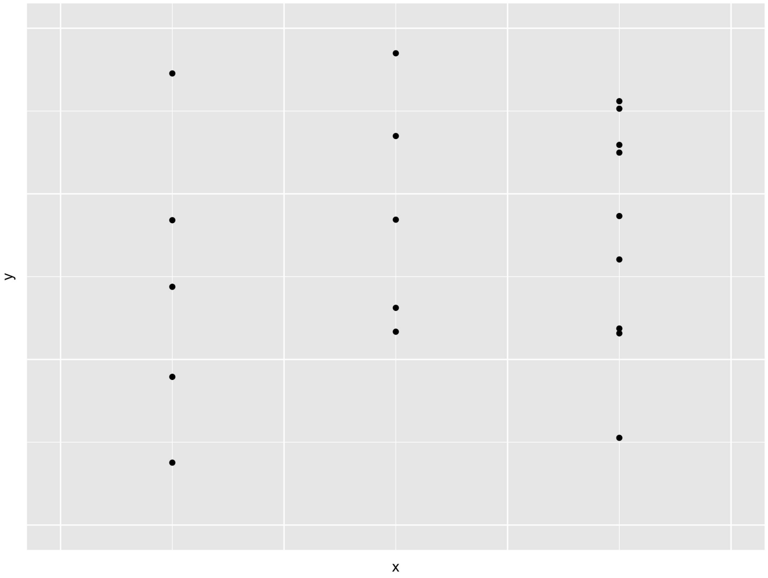


# Scatterplot Diagnostics

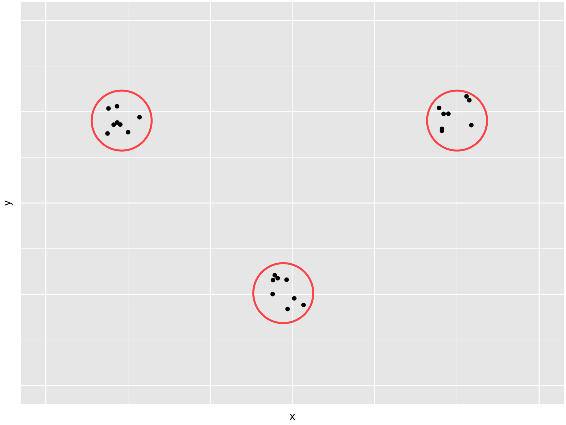
# Patterns

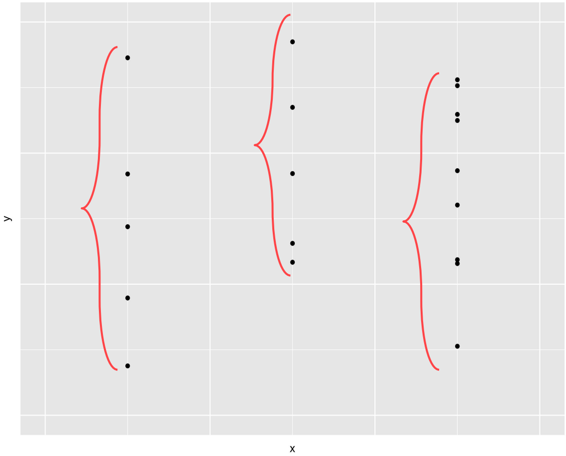


# Scagnostics

- Tukey and Tukey (1985) coined “scagnostics” - scatterplot diagnostics
- Further defined by Wilkinson, Anand, and Grossman (2005, 2008)



# Graphs

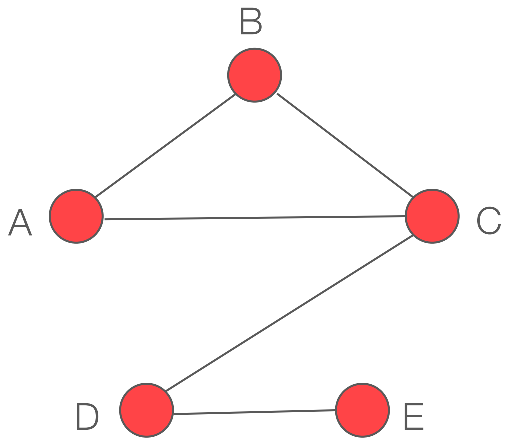
# What is a geometric graph?

- A *graph* is a set of vertices *V* which are related by edges *e(v,w)* in *E* and *v,w* in *V*
- *Geometric graphs* can be represented as points and lines in a metric space *S*

$V$ = {A,B,C,D,E}

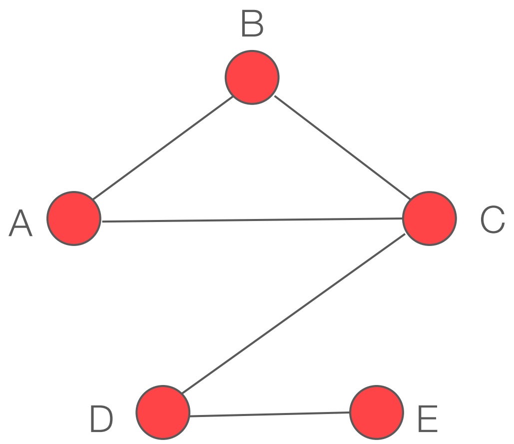
$E$ = {(A,B), (A,C), (B,C), (C,D), (D,E)}

$S$ = 2 dimensional space



# Graphs for Scagnostics

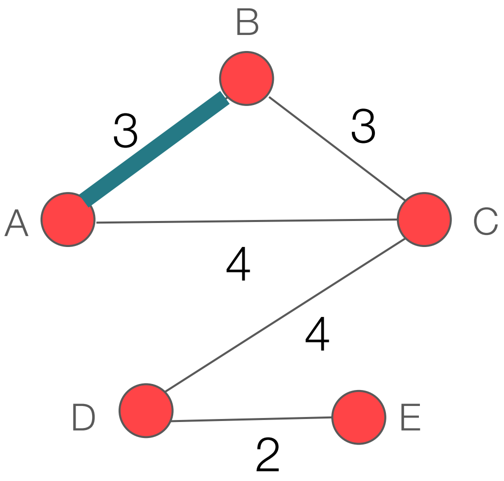
- Undirected
- Simple
- Planar
- Straight
- Finite



# Graph Feature Measures

*Length(e)* is the Euclidean distance between the vertices of an edge *e*

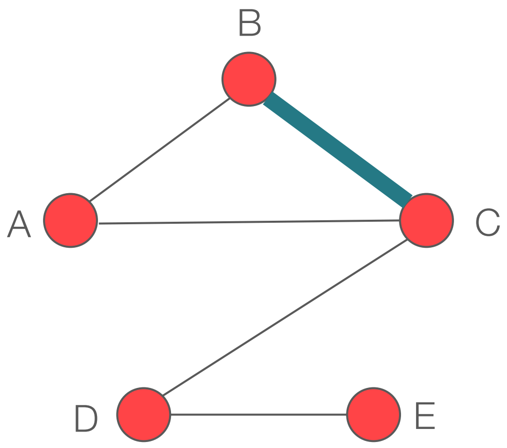
*Length(G)* is the total length of all edges of a graph *G*



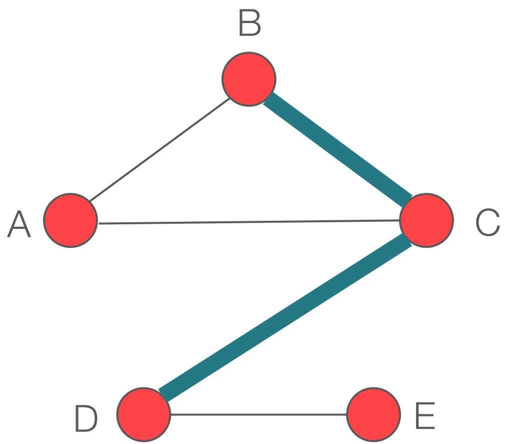

Length(AB) = 3

Length(G) = 3 + 3 + 4 + 4 + 2 = 16

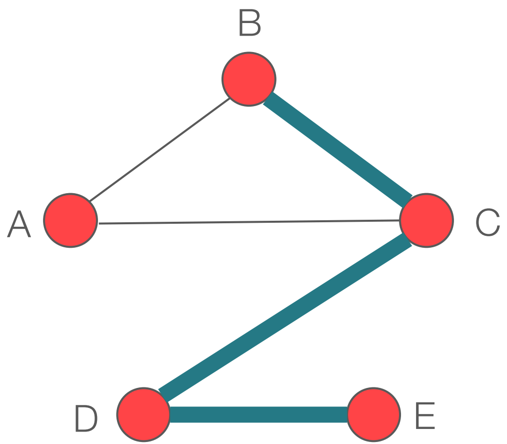
A *path* is a list of vertices such that all successive pairs are an edge



A *path* is a list of vertices such that all successive pairs are an edge

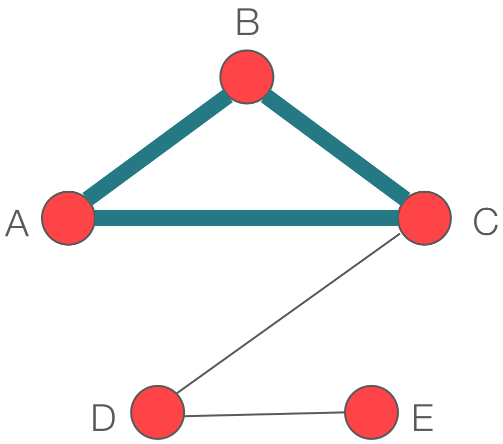


A *path* is a list of vertices such that all successive pairs are an edge



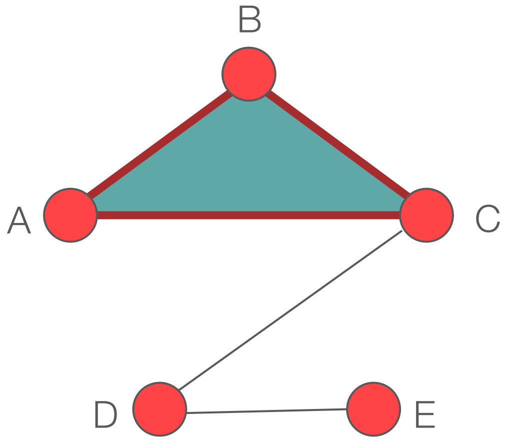
A path is *closed* if its first and last vertices are the same

A *polygon* is the boundary of a closed path



*Area(P)* is the area of polygon *P*

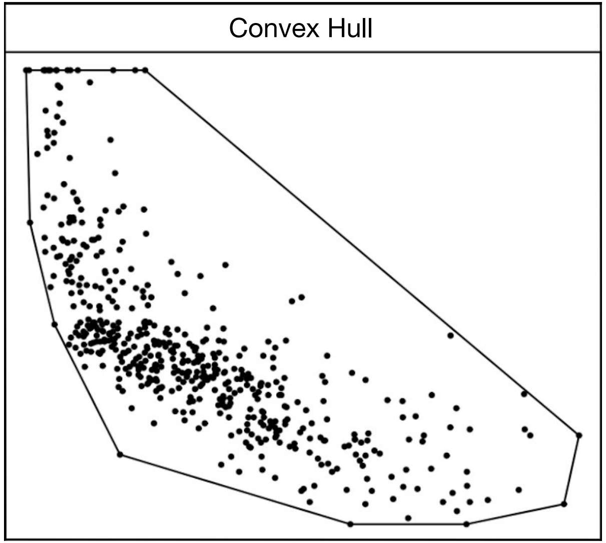
*Perimeter(P)* is the length of the boundary of polygon *P*.



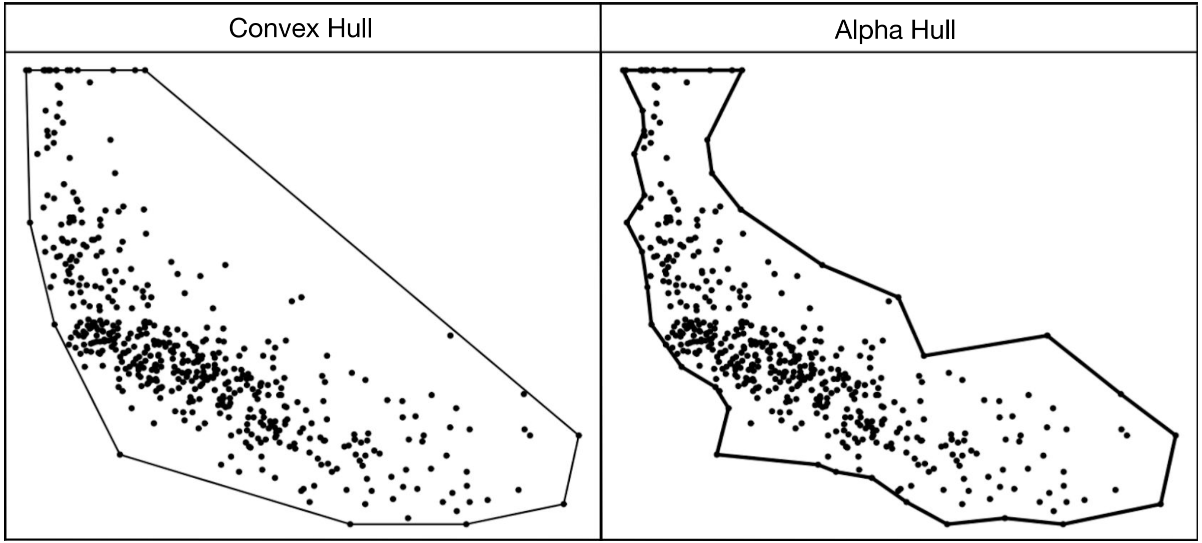
# Geometric Graphs of Interest

- Convex Hull
- Alpha Hull
- Minimum Spanning Tree

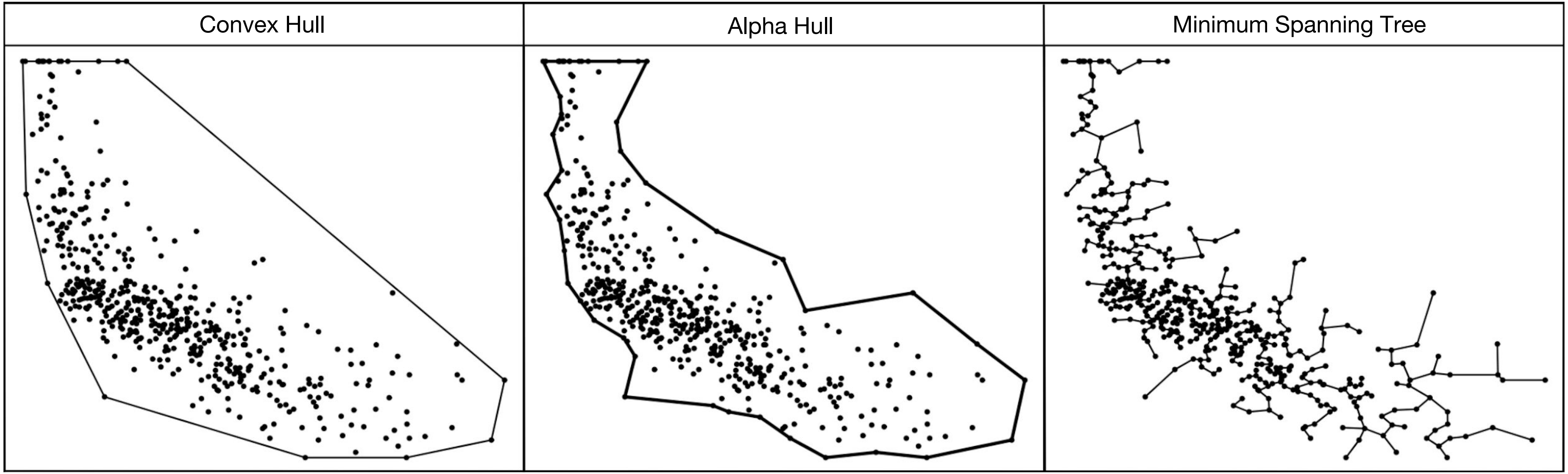
# Convex Hull

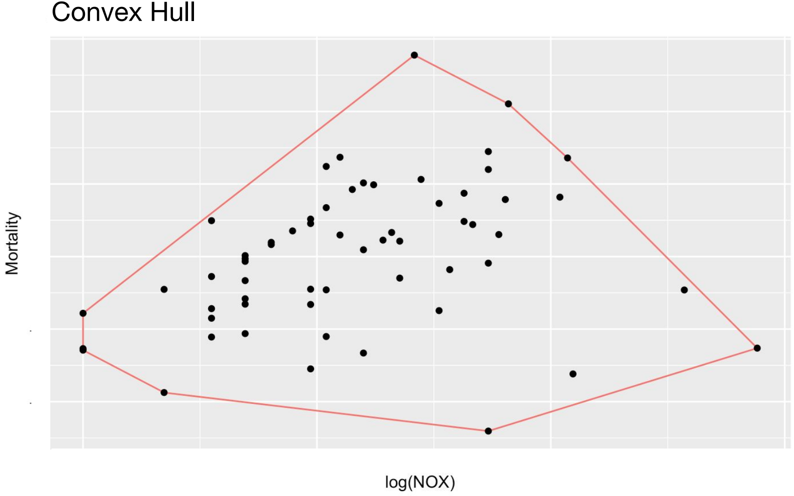


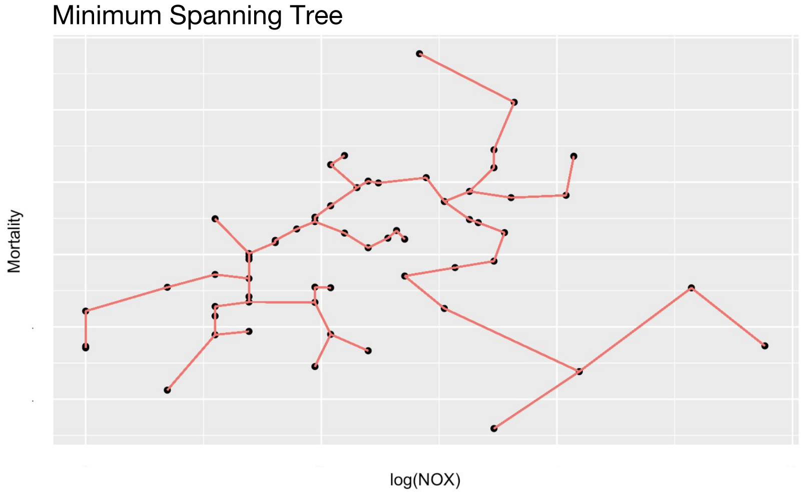
# Alpha Hull

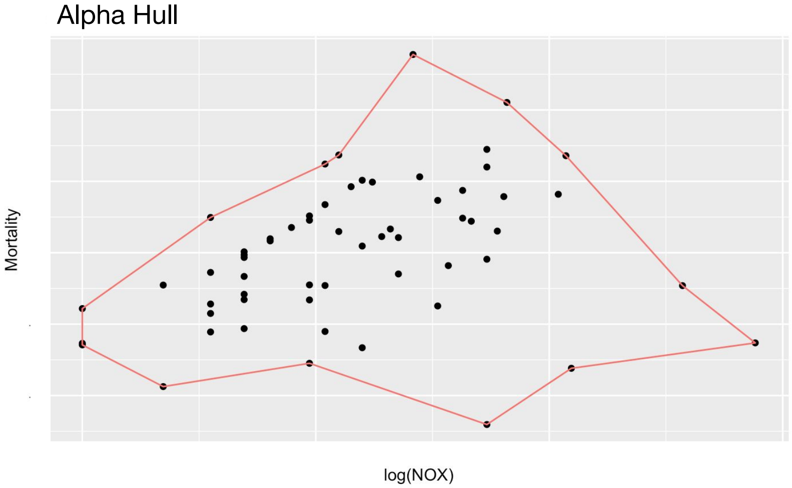


# Minimum Spanning Tree



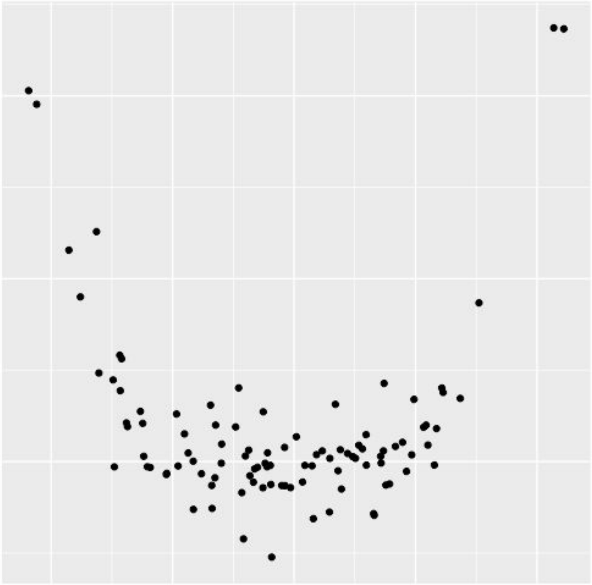

Convex Hull
Mortality
log(NOX)


Minimum Spanning Tree
Mortality
log(NOX)

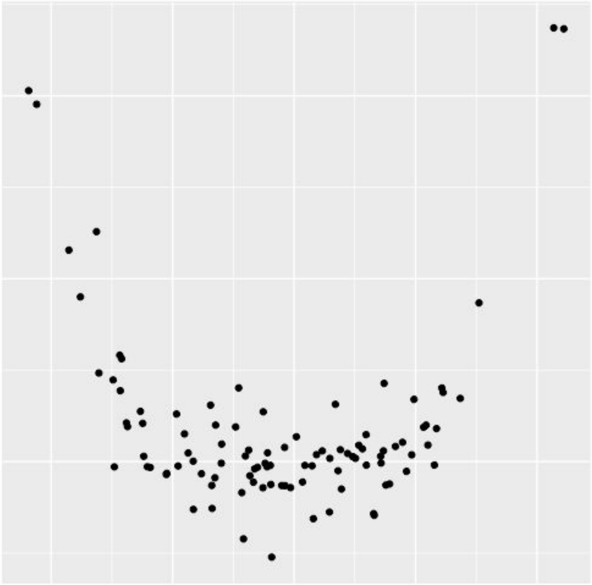

Alpha Hull
Mortality
log(NOX)

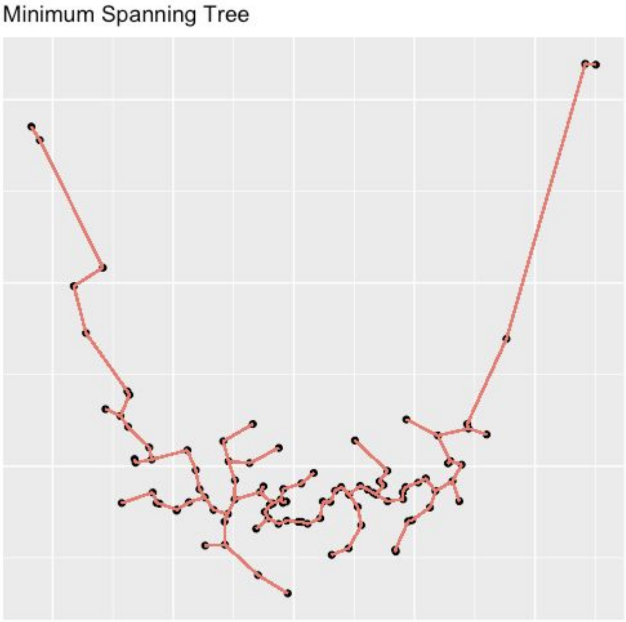
# Calculating Scagnostics

# How do we quantify patterns?

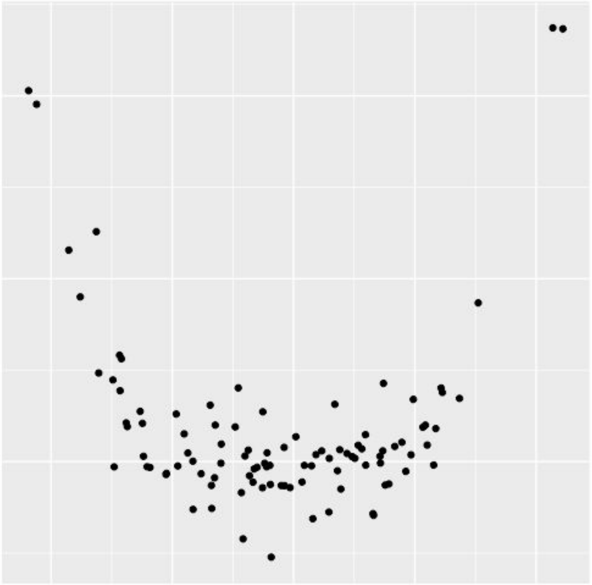


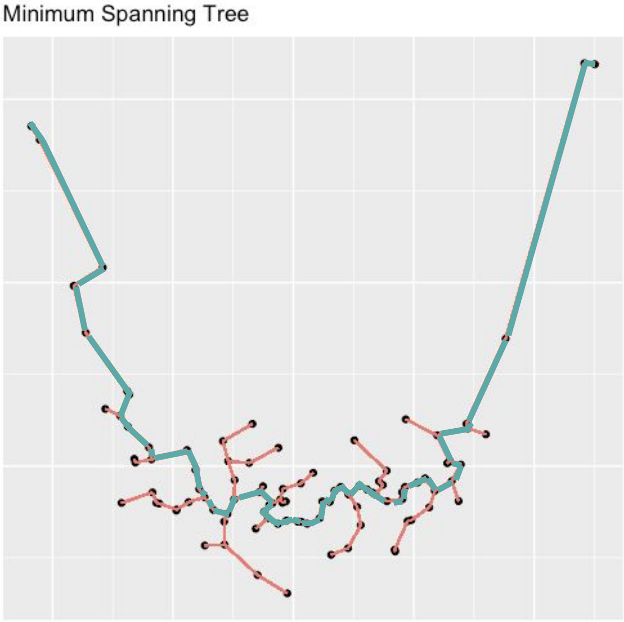
# Suppose we want to measure how “stringy” a plot is



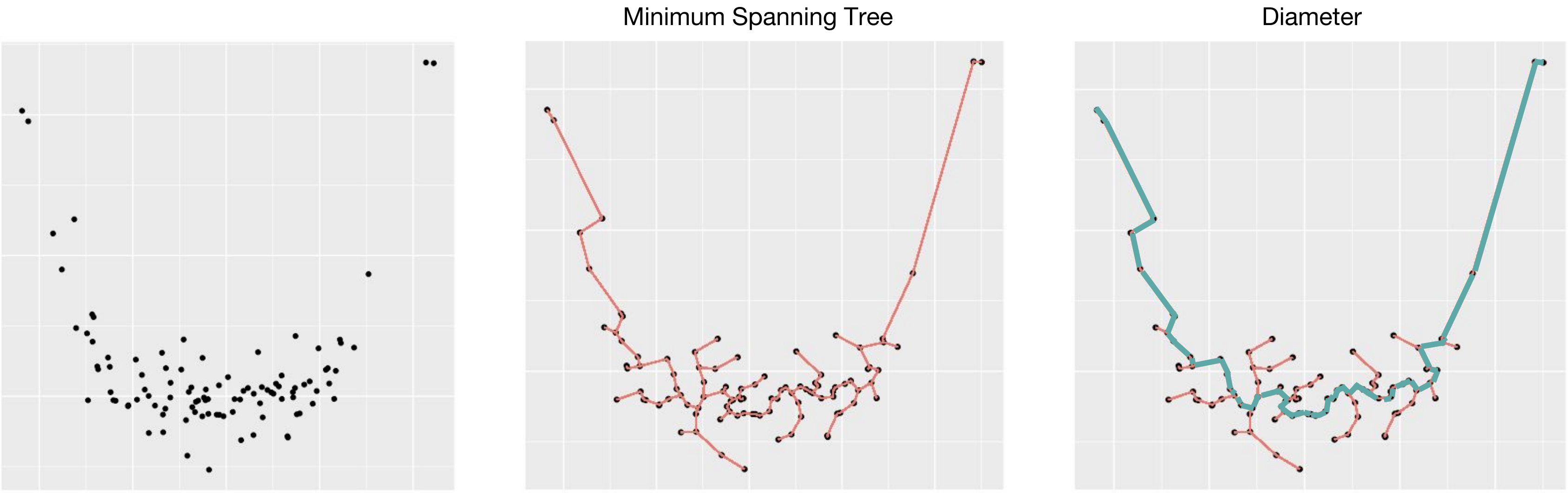


# Suppose we want to measure how "stringy" a plot is





$$c_{stringy} = \frac{diameter(T)}{length(T)}$$



# Scagnostic Measures

Shape

- Stringy
- Convex
- Skinny
- Clumpy
- Striated

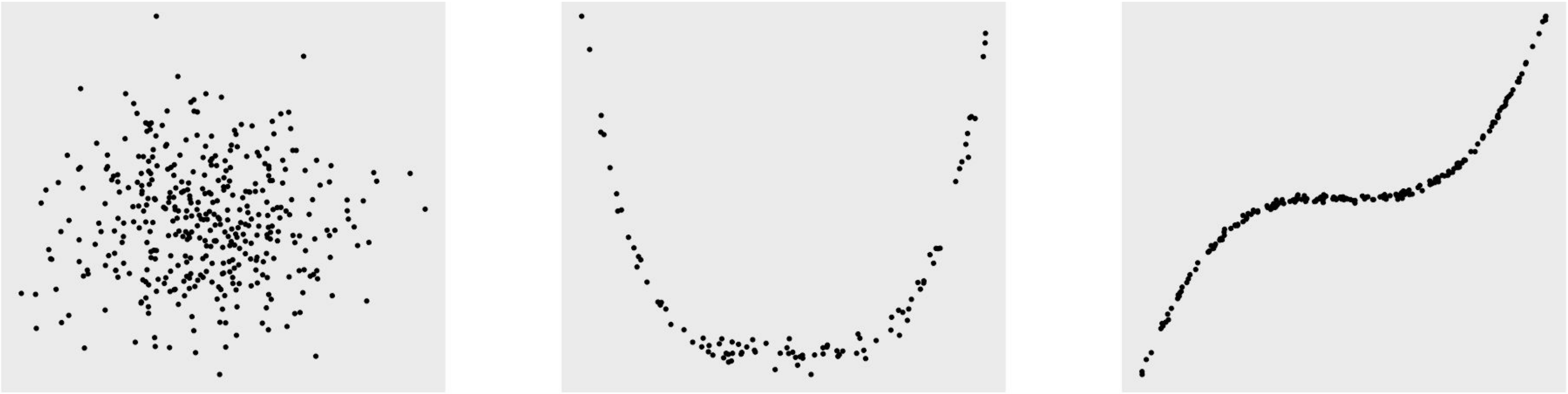
Density and Association

- Monotonic
- Outlying
- Sparse
- Skewed

# Stringy

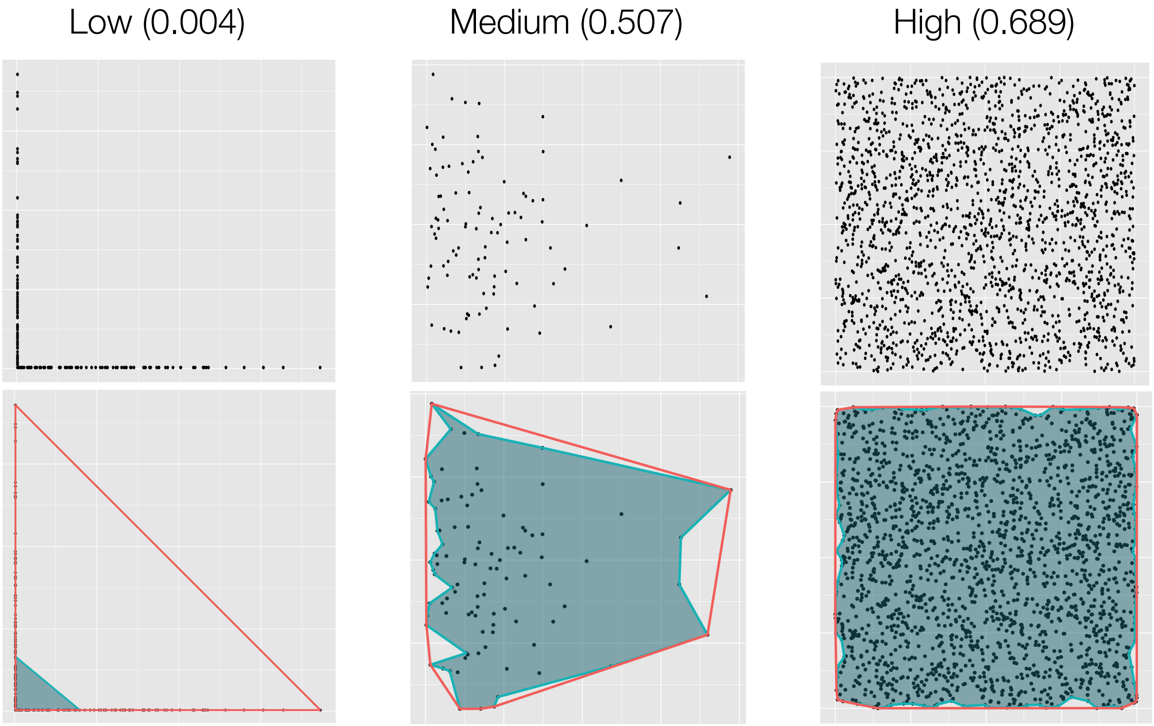
$$c_{stringy} = \frac{diameter(T)}{length(T)}$$

Low (0.361) Medium (0.611) High (0.894)



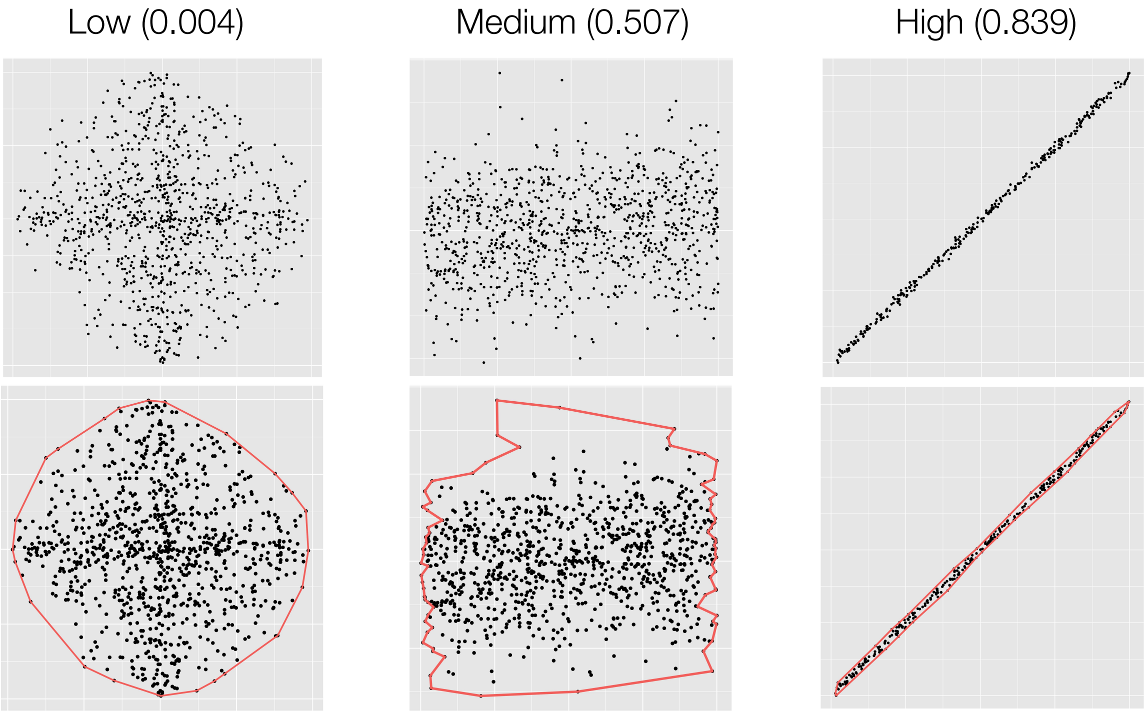
# Scagnostics: Shape

$$c_{convex} = \frac{area(A)}{area(H)}$$



# Scagnostics: Shape

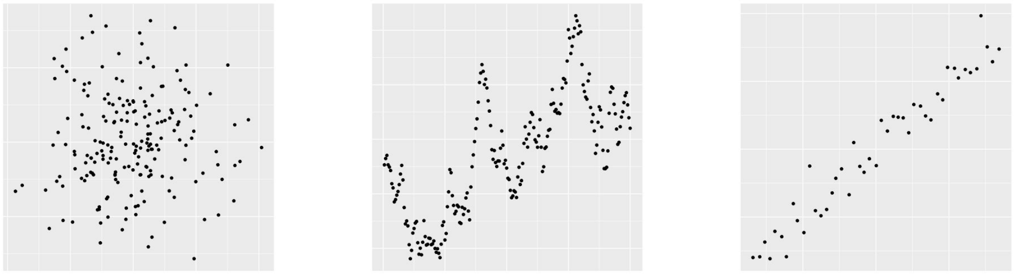
$$c_{skinny} = 1 - \frac{\sqrt{4\pi area(A)}}{perimeter(A)}$$



# Scagnostics: Association

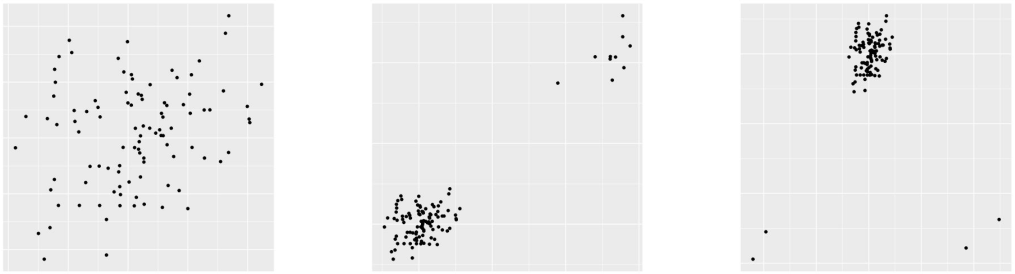
$$c_{monotonic} = {r_{spearman}}^2$$

Low (0.001) | Medium (0.506) | High (0.948)



$$c_{outlying} = \frac{length(T_{outliers})}{length(T)}$$

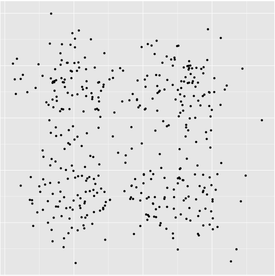
Low (0.052) | Medium (0.543) | High (0.976)



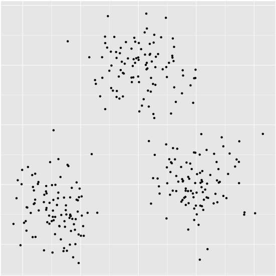
# Scagnostics: Shape

$$c_{clumpy} = \max_j \left[ 1 - \frac{\max_k [length(e_k)]}{length(e_j)} \right]$$

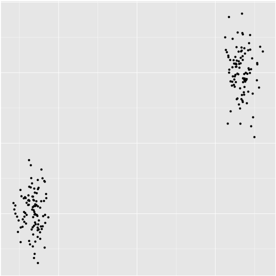
Low (0.007)



Medium (0.446)

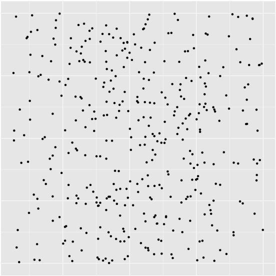


High (0.900)



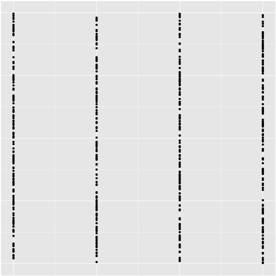
$$c_{striated} = \frac{1}{|V|} \sum_{v \in V^{(2)}} I(cos\theta_{e(v,a)e(v,b)} < -.75)$$

Low (0.035)



Medium (0.514)

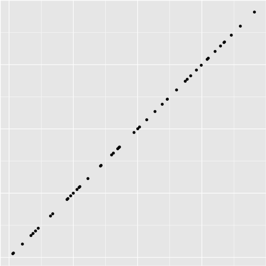
High (0.928)



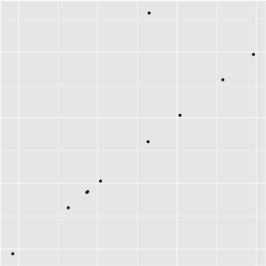
# Scagnostics: Density

$$c_{sparse} = q_{90}(T)$$

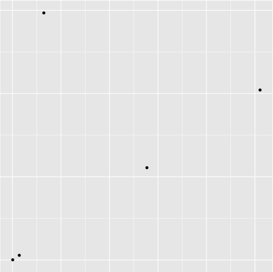
Low (0.080)



Medium (0.415)

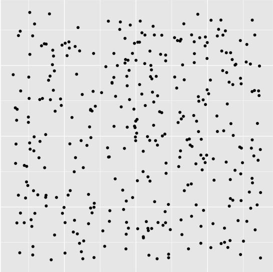


High (0.754)

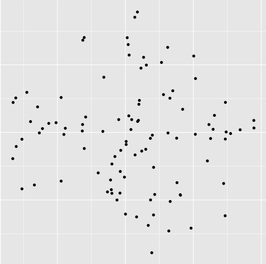


$$c_{skew} = \frac{q_{90}(T) - q_{50}(T)}{q_{90}(T) - q_{10}(T)}$$

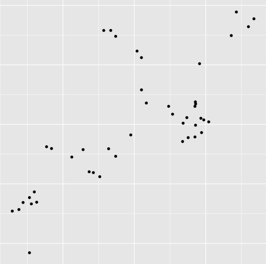
Low (0.382)



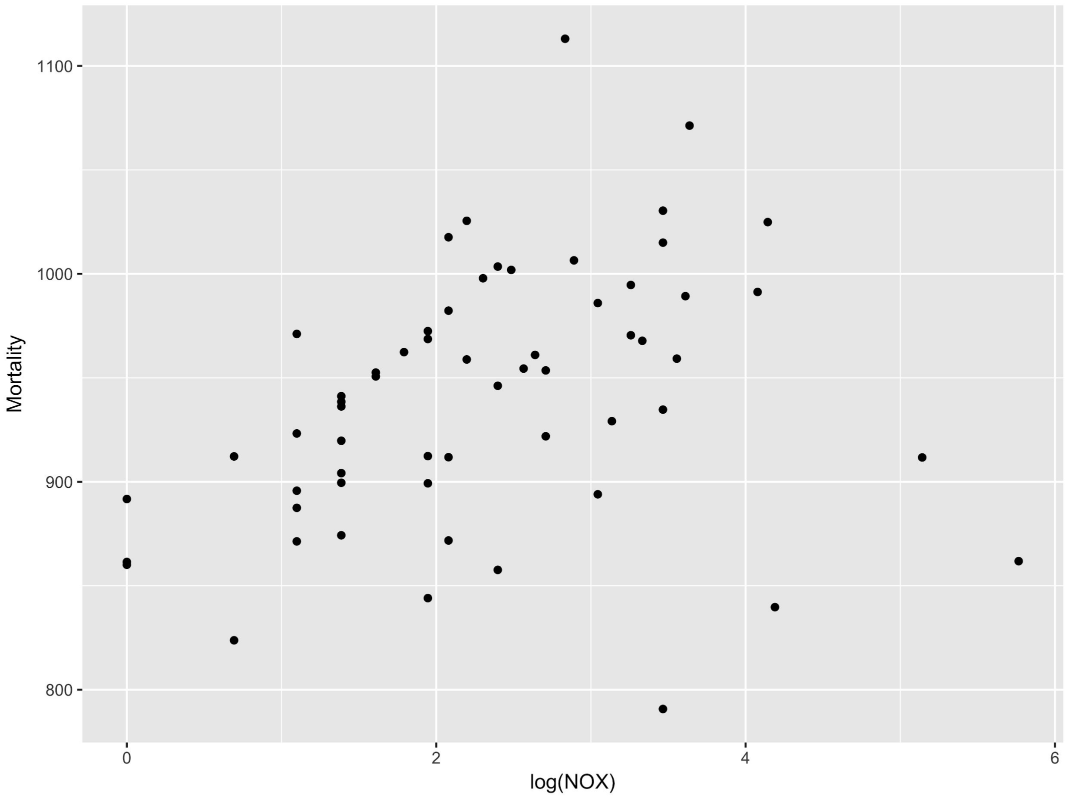
Medium (0.526)



High (0.877)



# How Scagnostics Work




Outlying: 0.496

Skewed: 0.556

Clumpy: 0.038

Sparse: 0.098

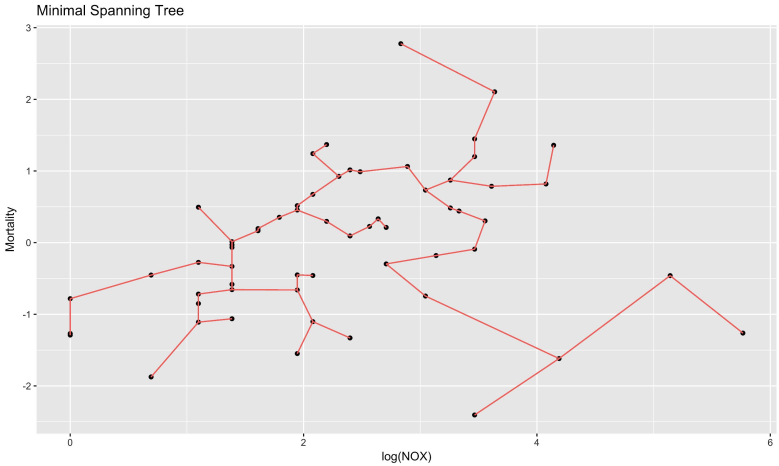
Striated: 0.100

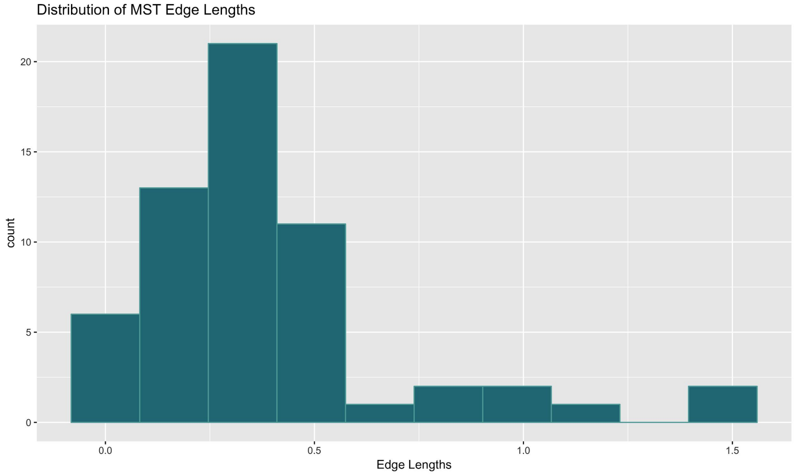
Convex: 0.718

Skinny: 0.236

Stringy: 0.521

Monotonic: 0.340




Outlying: 0.496

**Skewed: 0.556**

Clumpy: 0.038

Sparse: 0.098

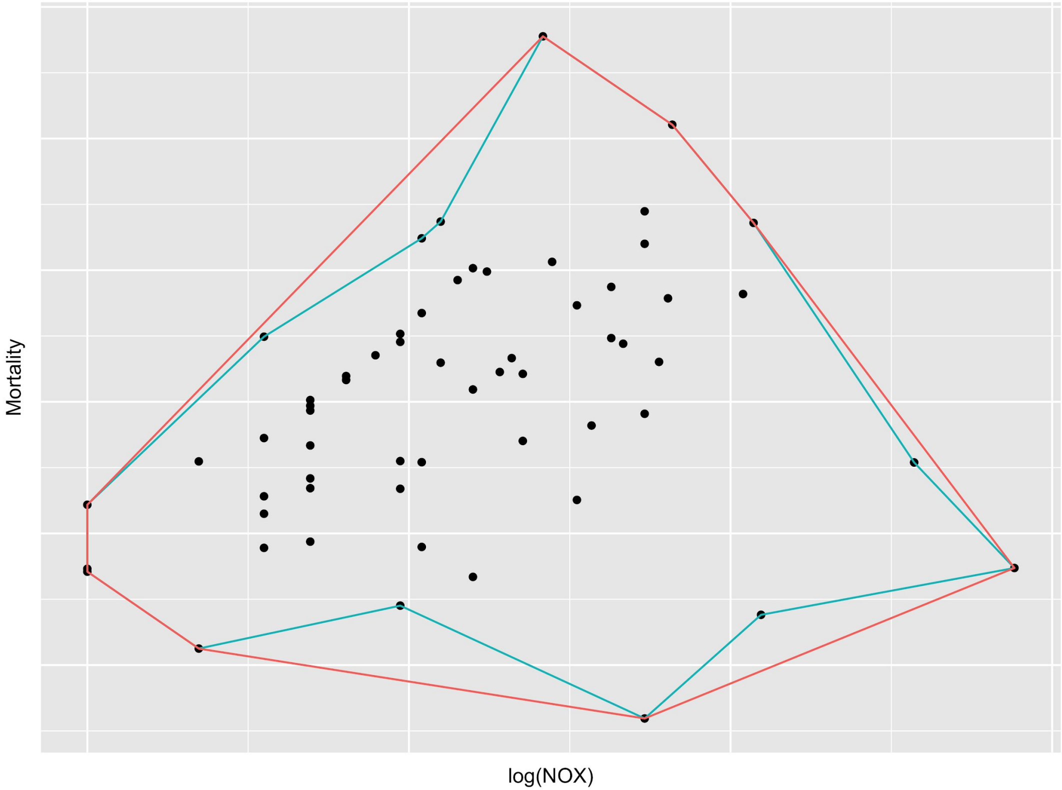
Striated: 0.100

**Convex: 0.718**

Skinny: 0.236

**Stringy: 0.521**

Monotonic: 0.340




Outlying: 0.496

**Skewed: 0.556**

Clumpy: 0.038

Sparse: 0.098

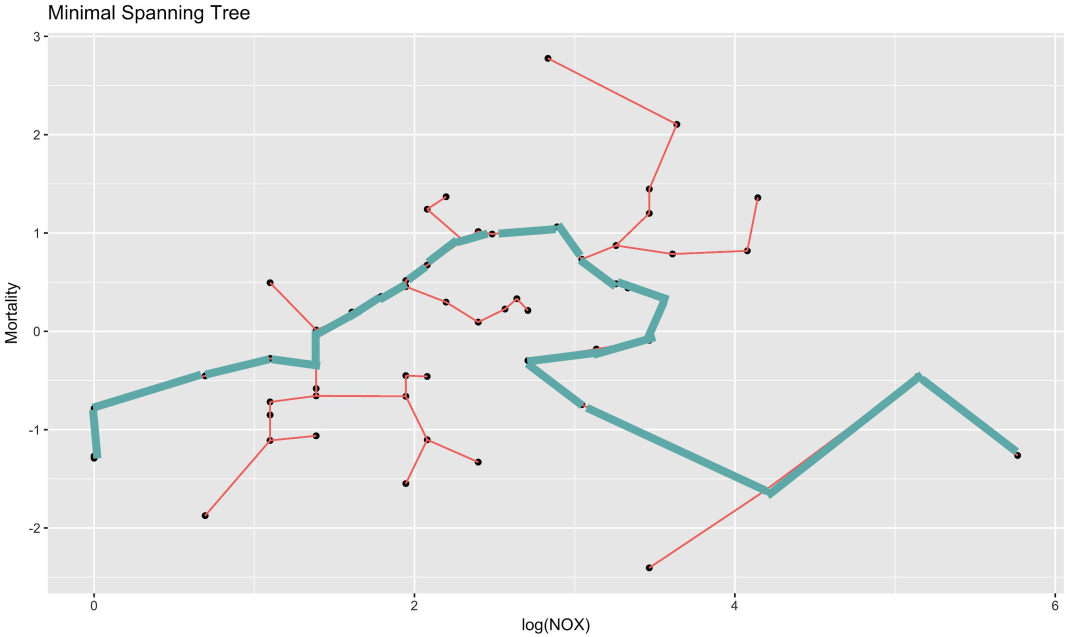
Striated: 0.100

**Convex: 0.718**

Skinny: 0.236

**Stringy: 0.521**

Monotonic: 0.340




Outlying: 0.496

**Skewed: 0.556**

Clumpy: 0.038

Sparse: 0.098

Striated: 0.100

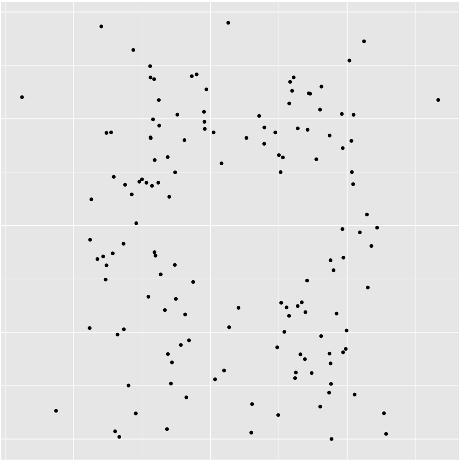
**Convex: 0.718**

Skinny: 0.236

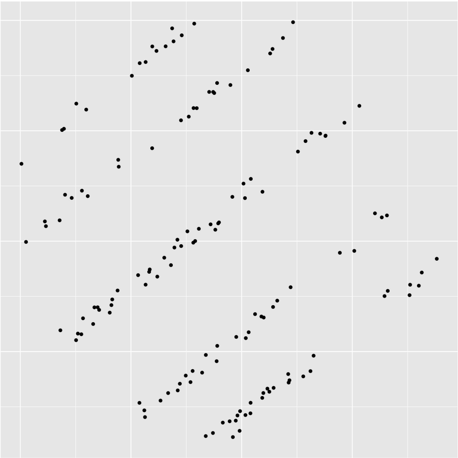
**Stringy: 0.521**

Monotonic: 0.340

# How Scagnostics Differentiate Plots



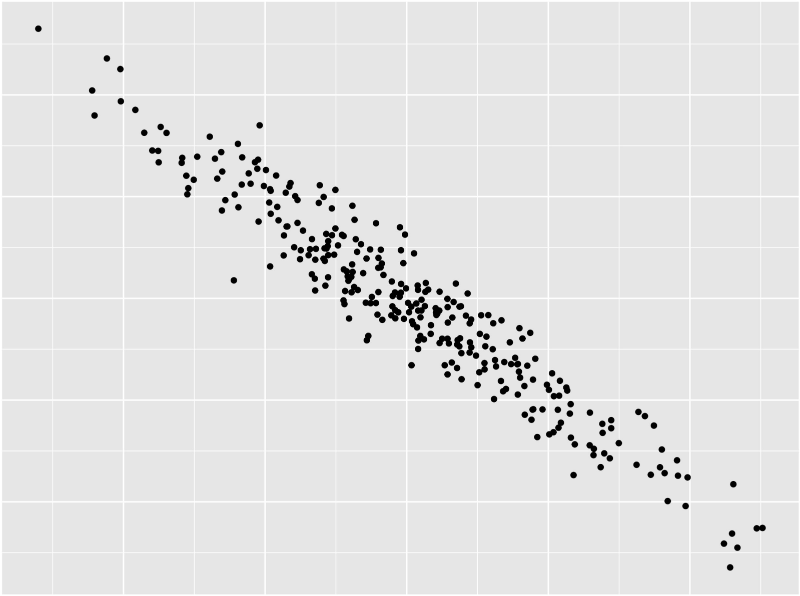
Outlying: 0.108
Skewed: 0.617
**Clumpy: 0.002**
Sparse: 0.078
Striated: 0.076
**Convex: 0.522**
**Skinny: 0.571**
**Stringy: 0.369**
Monotonic: 0.008



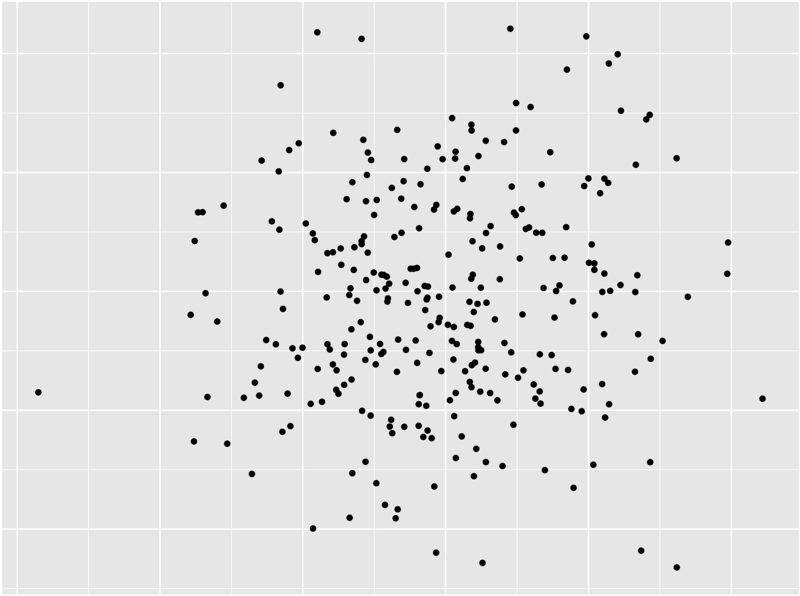
Outlying: 0.088
Skewed: 0.749
**Clumpy: 0.142**
Sparse: 0.067
Striated: 0.172
**Convex: 0.094**
**Skinny: 0.838**
**Stringy: 0.559**
Monotonic: 0.003

# Building the Model

# Why?



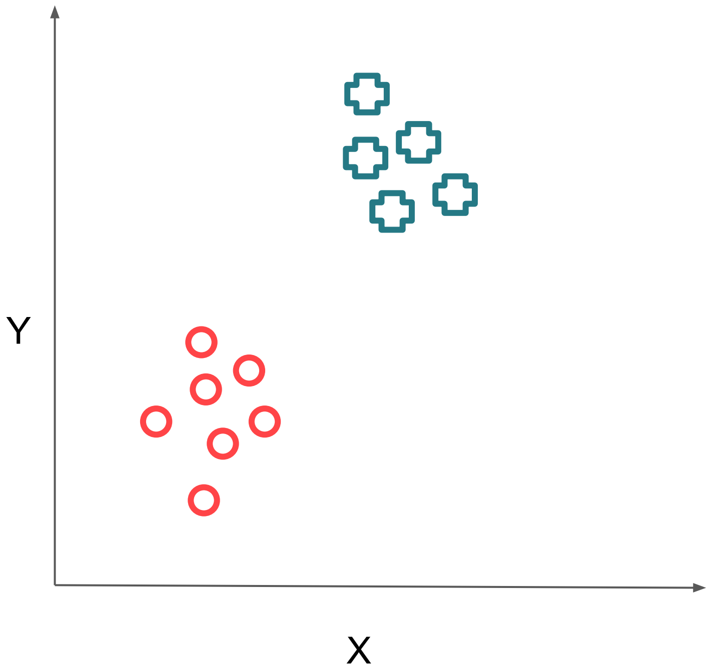
Signal



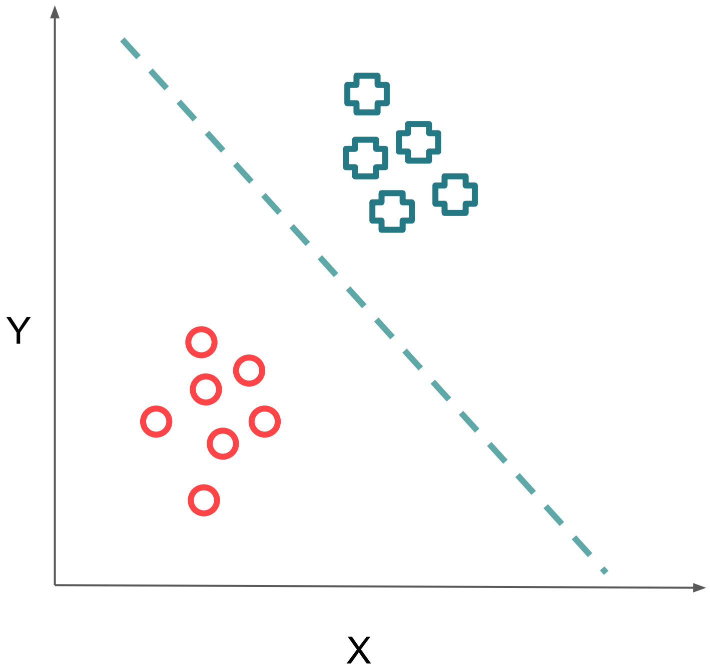
Null

# Statistical Learning

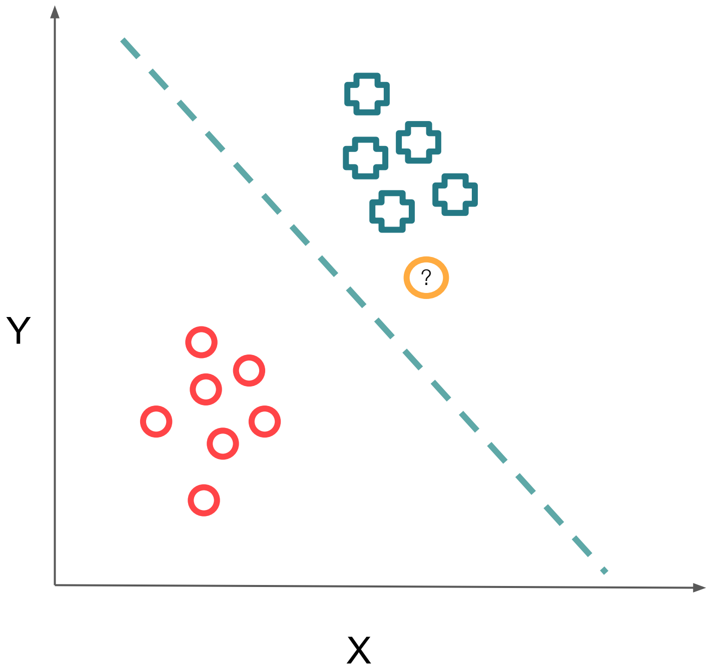
# Supervised Learning



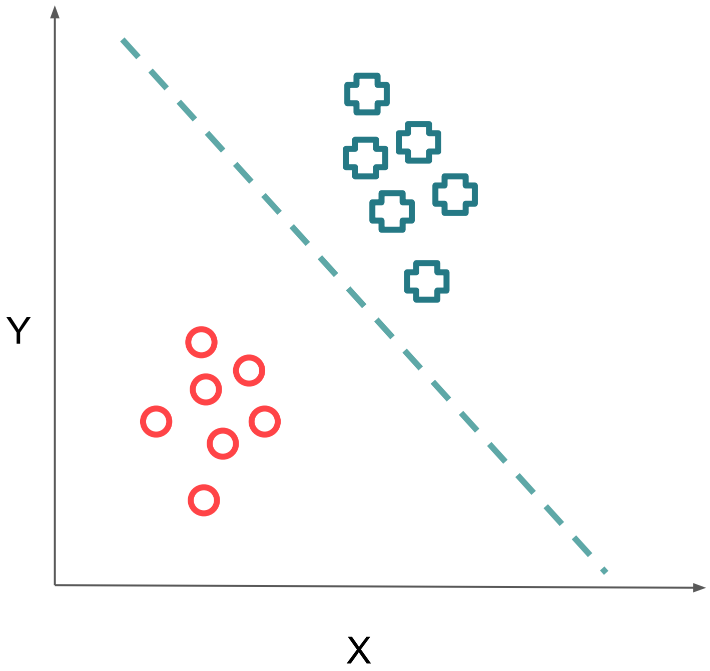
# Supervised Learning



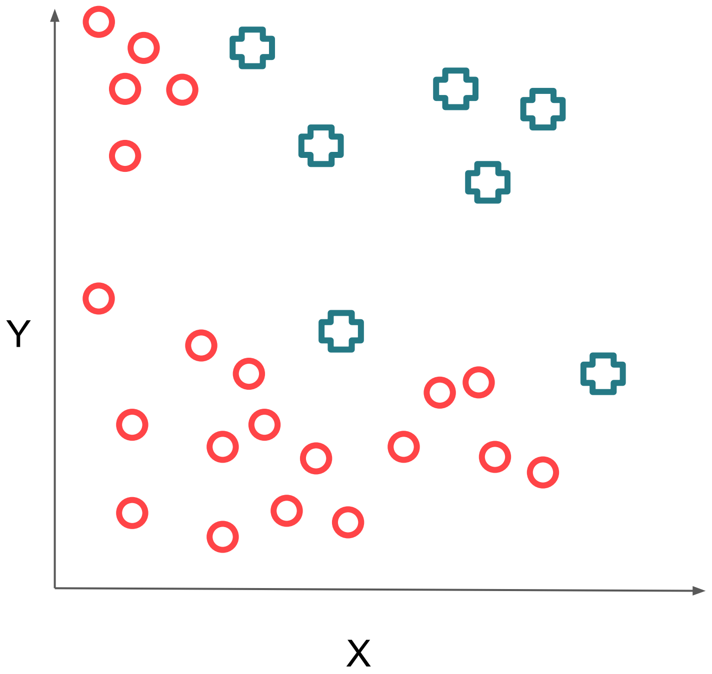
# Supervised Learning



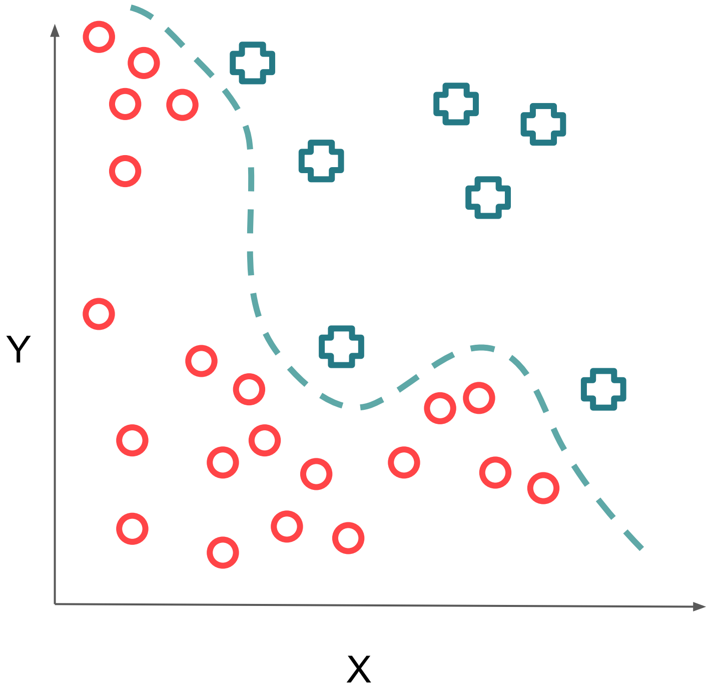
# Supervised Learning



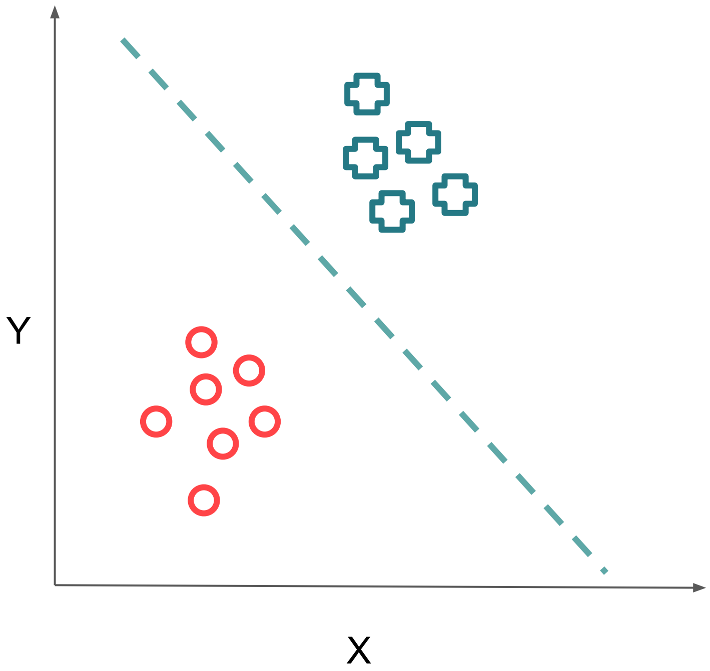
# Supervised Learning



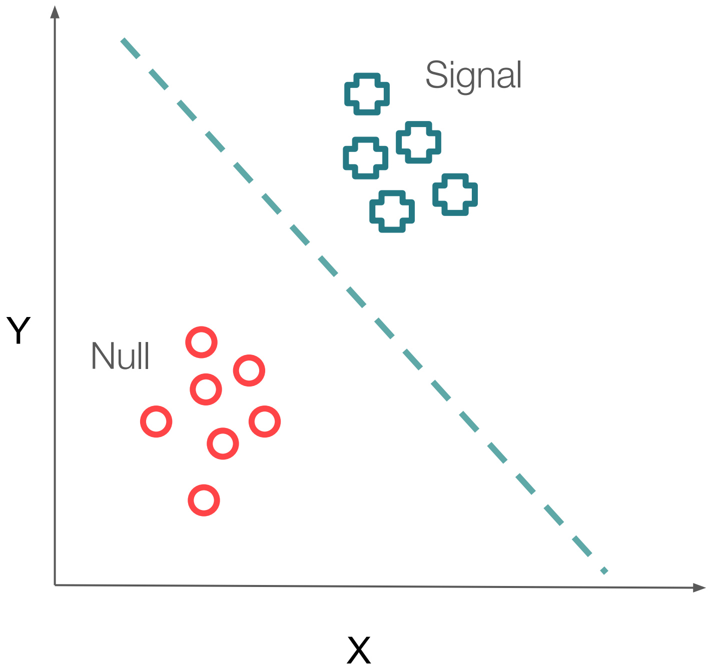
# Supervised Learning



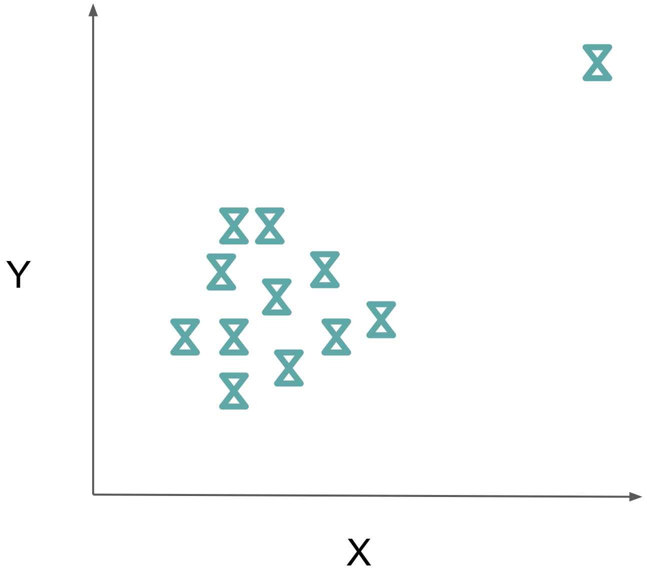
# Supervised Learning



# Supervised Learning

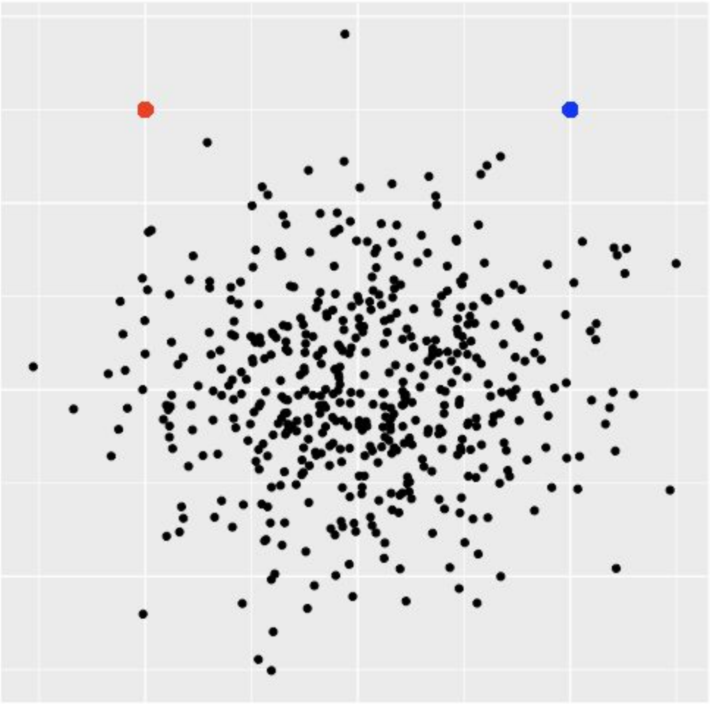


# Unsupervised Learning

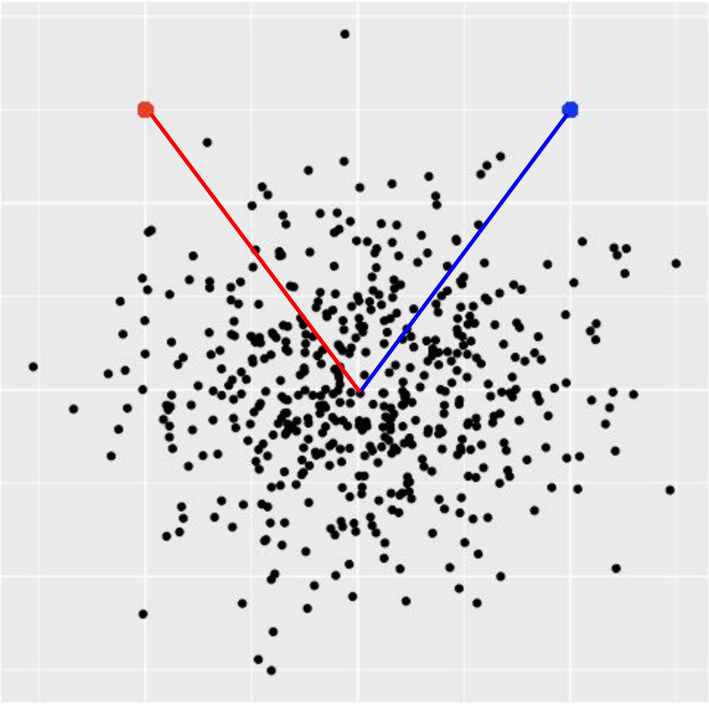


# Unsupervised Method: Distance

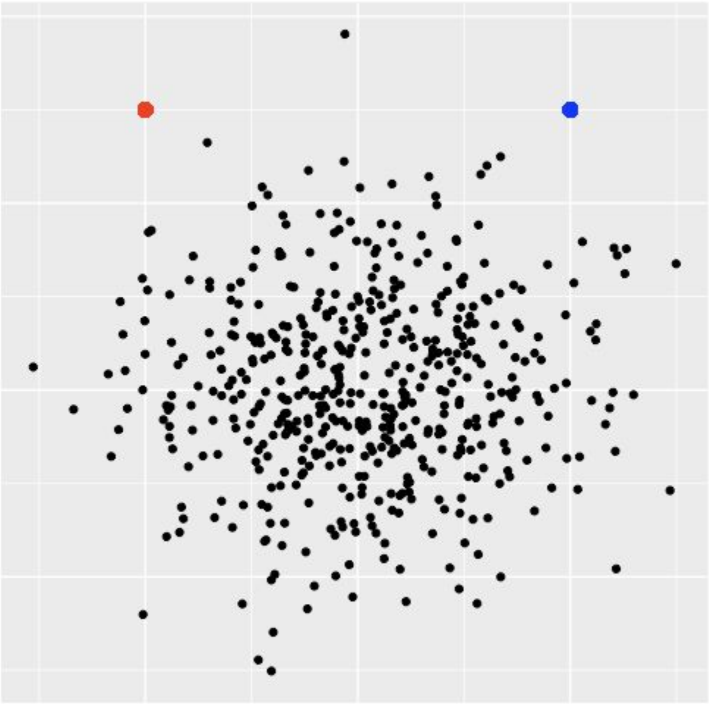
# Euclidean Distance

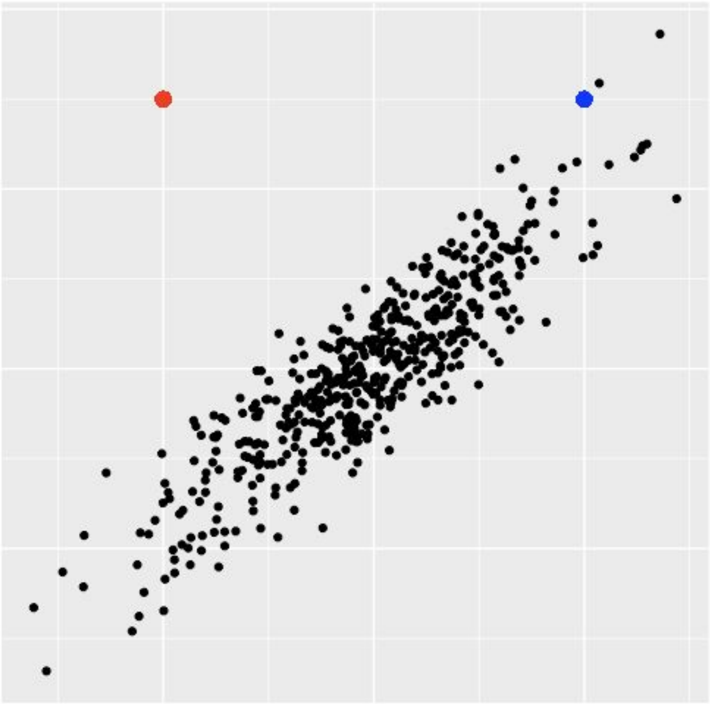


# Euclidean Distance

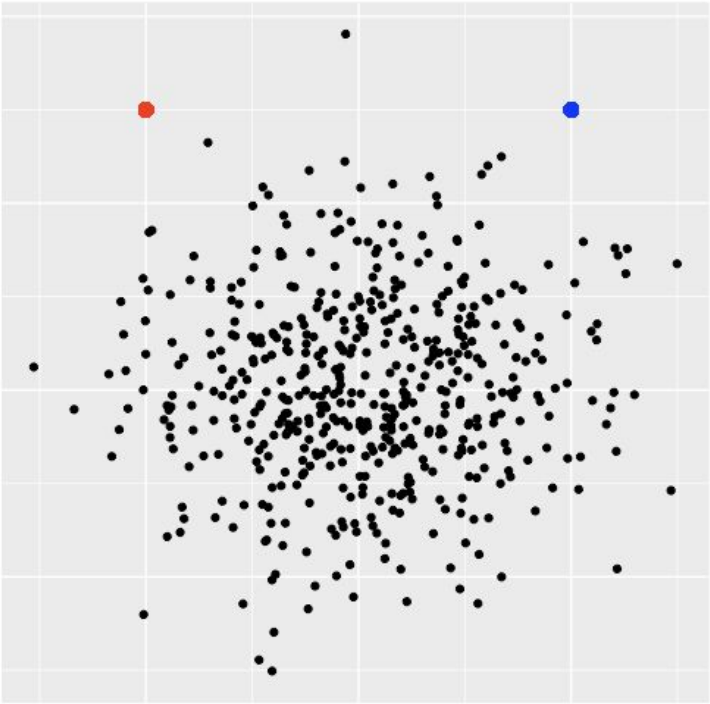


# Euclidean Distance

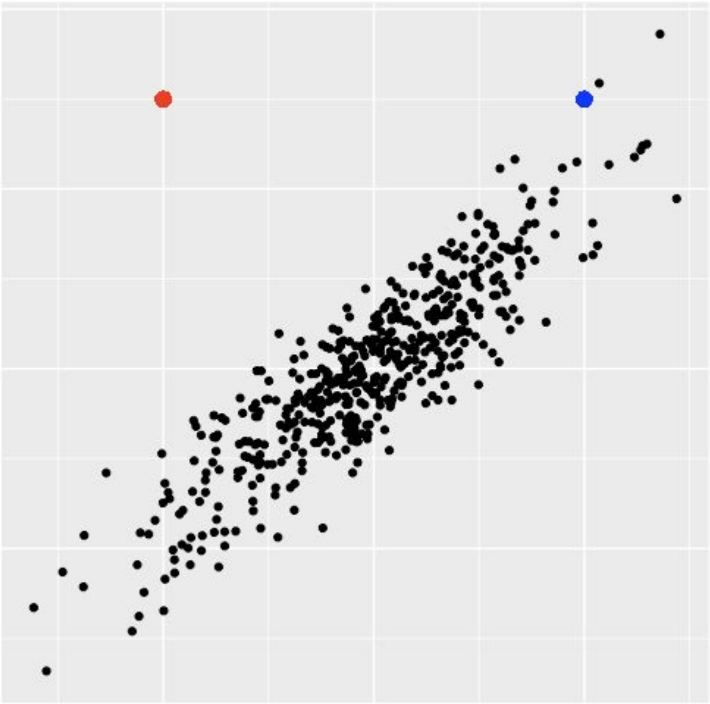




## Euclidean Distance



## Mahalanobis Distance

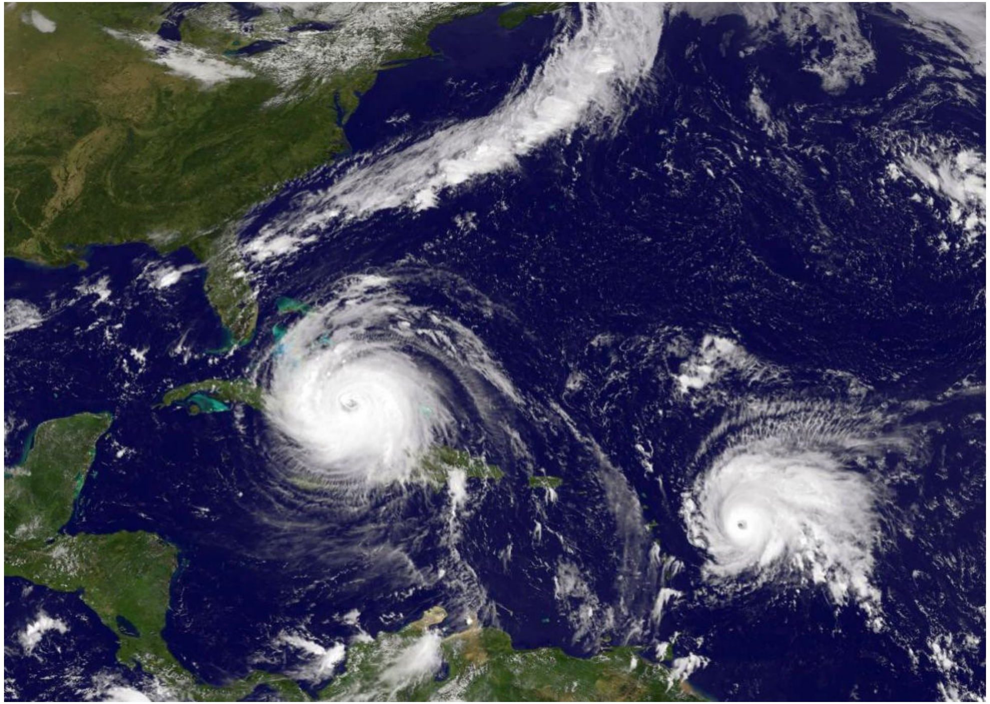


# Supervised Method: Random Forest

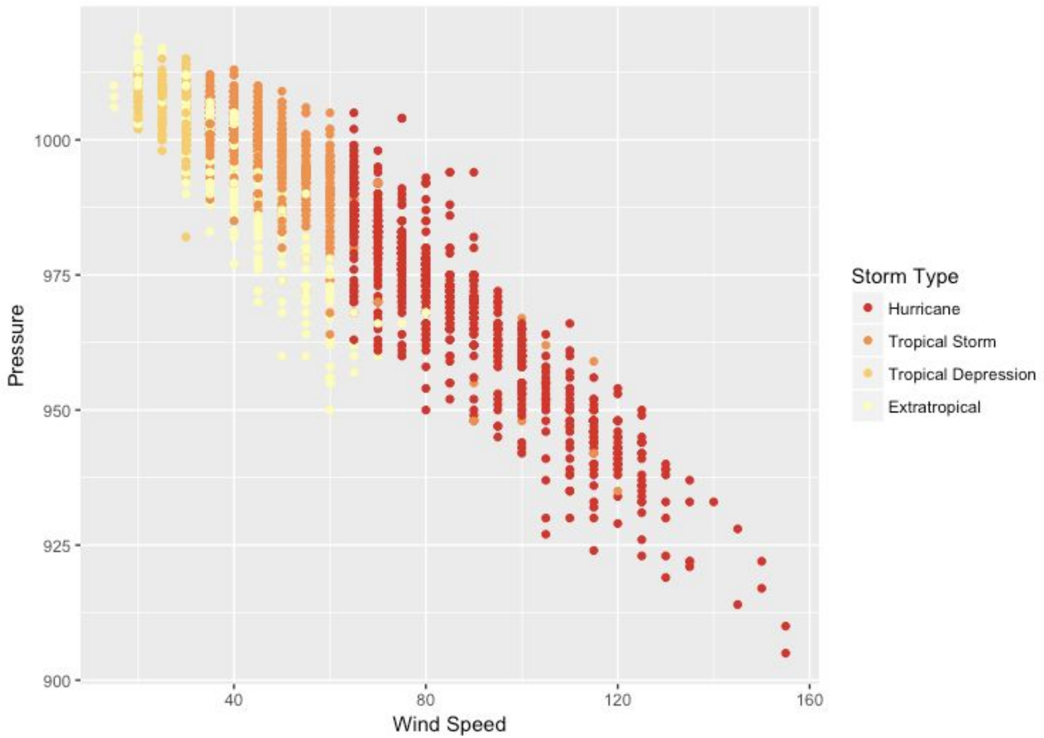
# Decision Tree

# NASA Hurricane Data

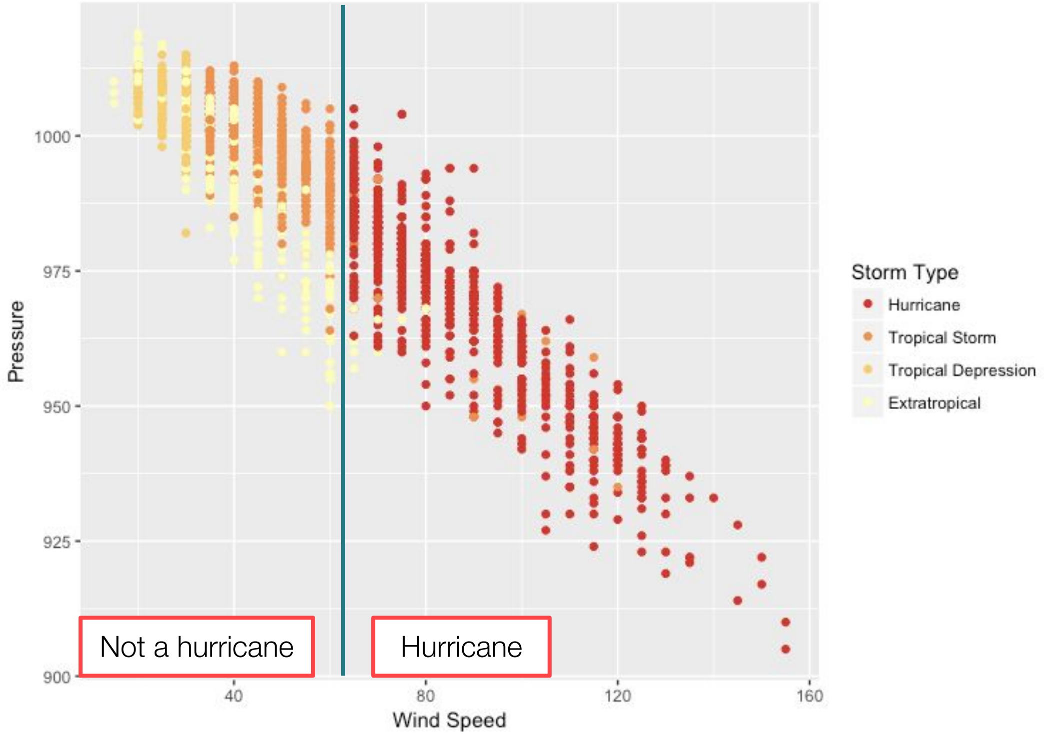
- Storm data from the  National Hurricane Center's archive of Tropical Cyclone Reports (1995-2005).
- Hurricanes, tropical storms, tropical depressions, and extratropical storms were tracked through the Atlantic Ocean, Caribbean Sea and Gulf of Mexico.



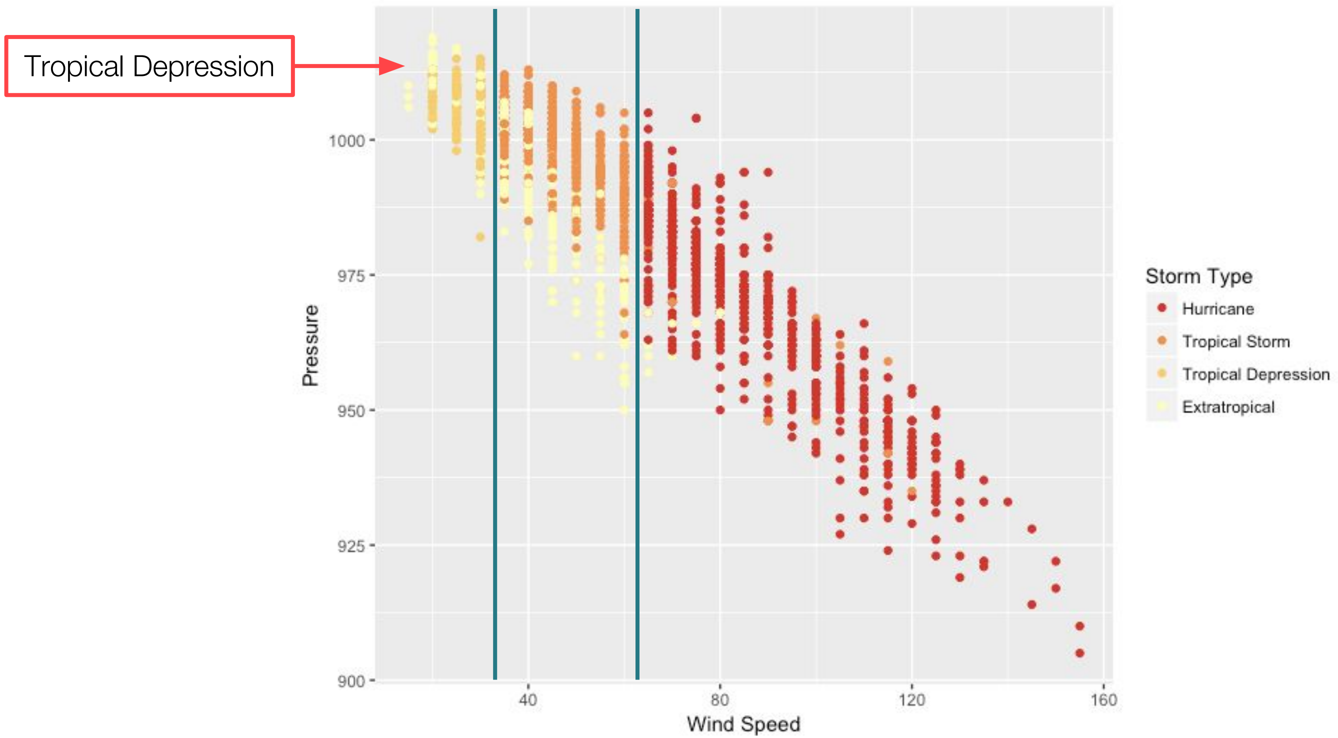
# How might we split this data?



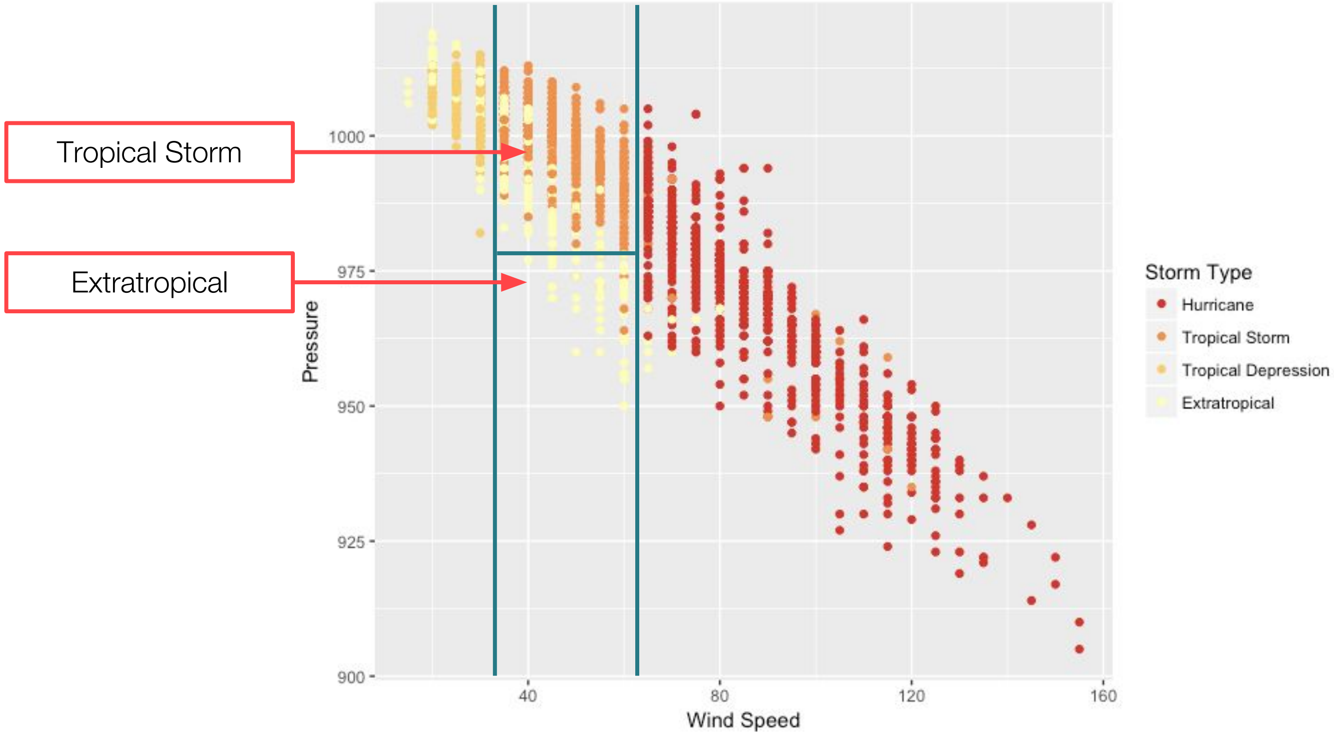
# How might we split this data?



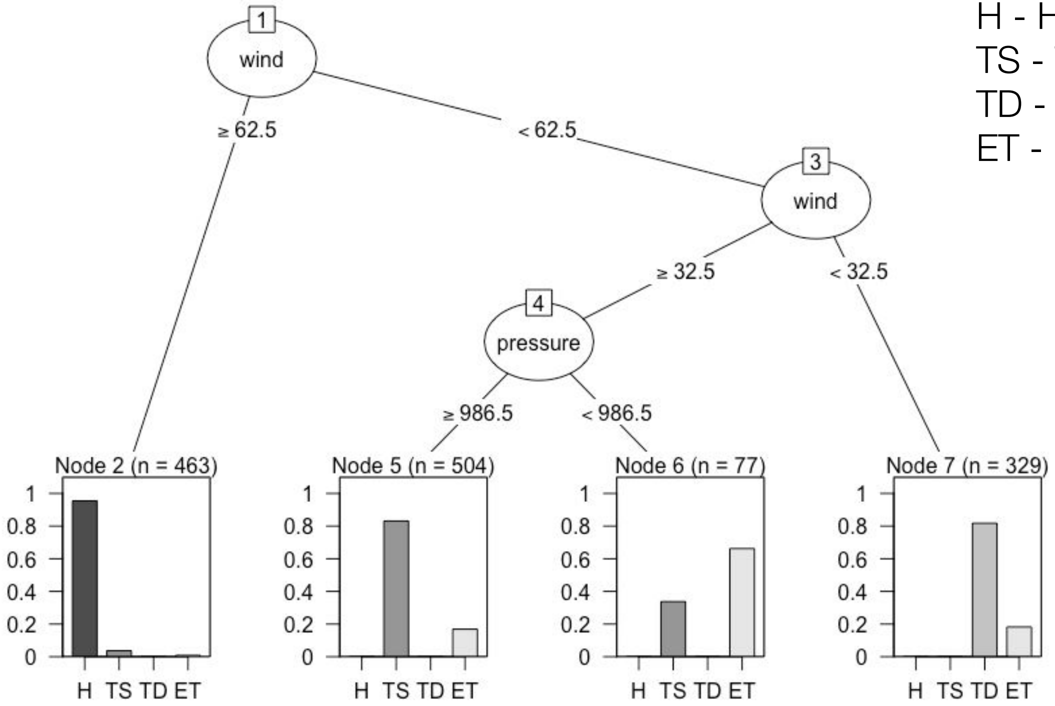
# How might we split this data?



# How might we split this data?



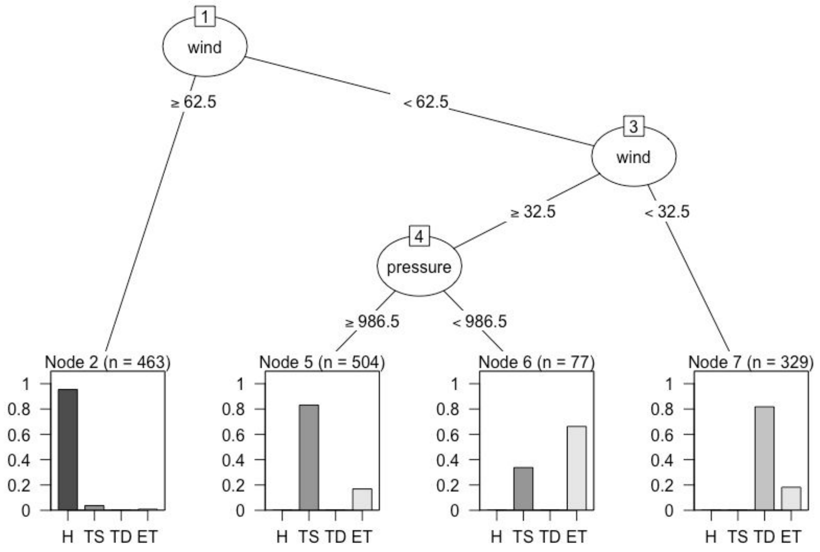
# Decision Tree



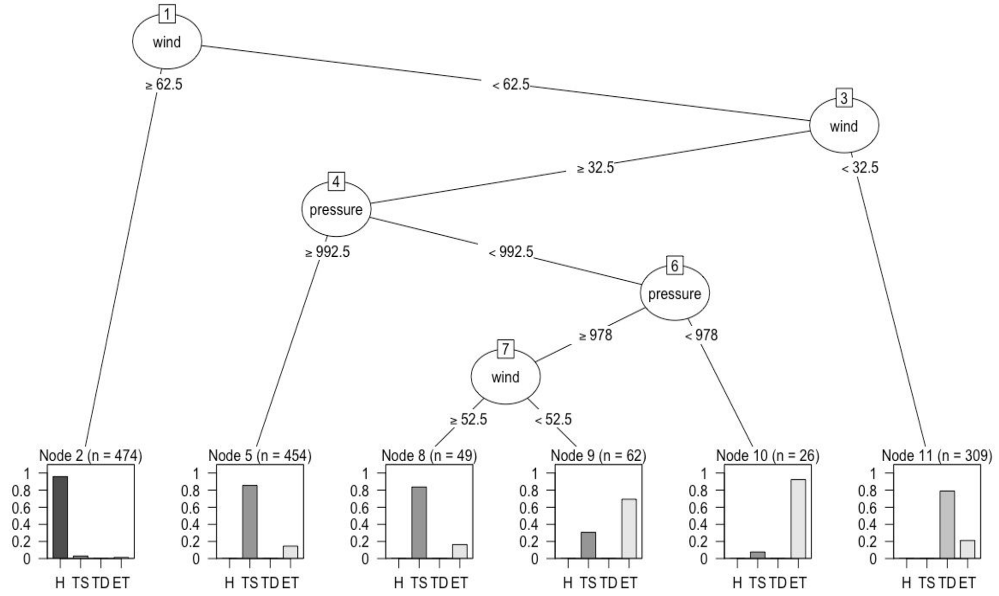

H - Hurricane
TS - Tropical Storm
TD - Tropical Depression
ET - Extra Tropical

# Problem: Decision trees are not very robust

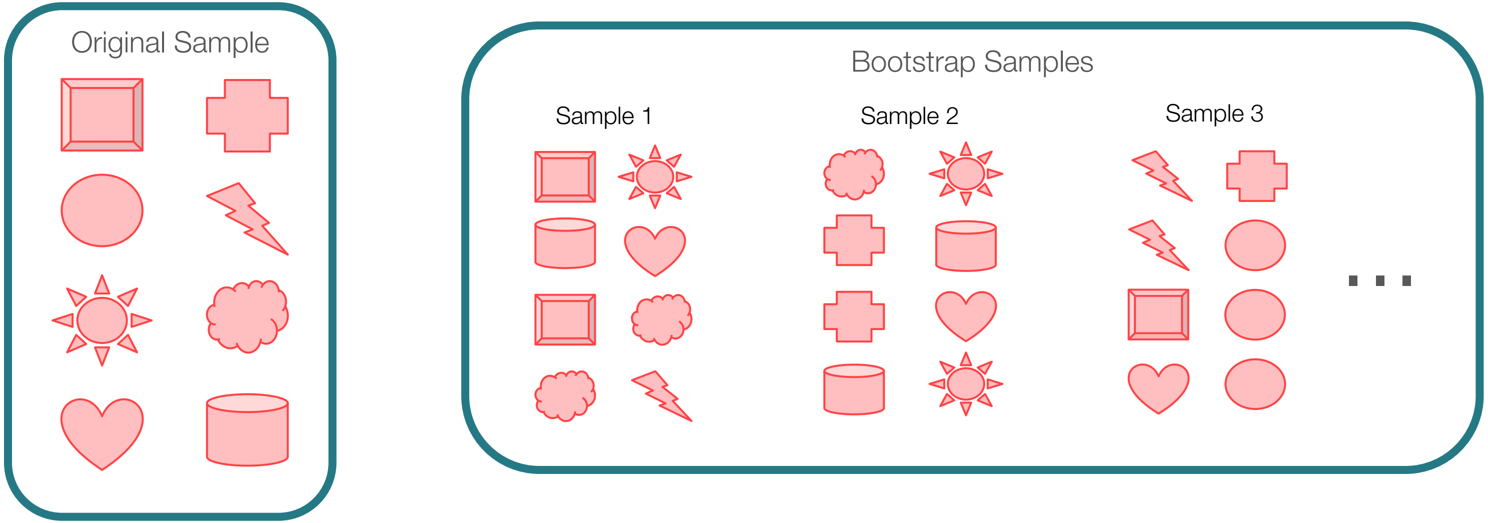
Sample 1



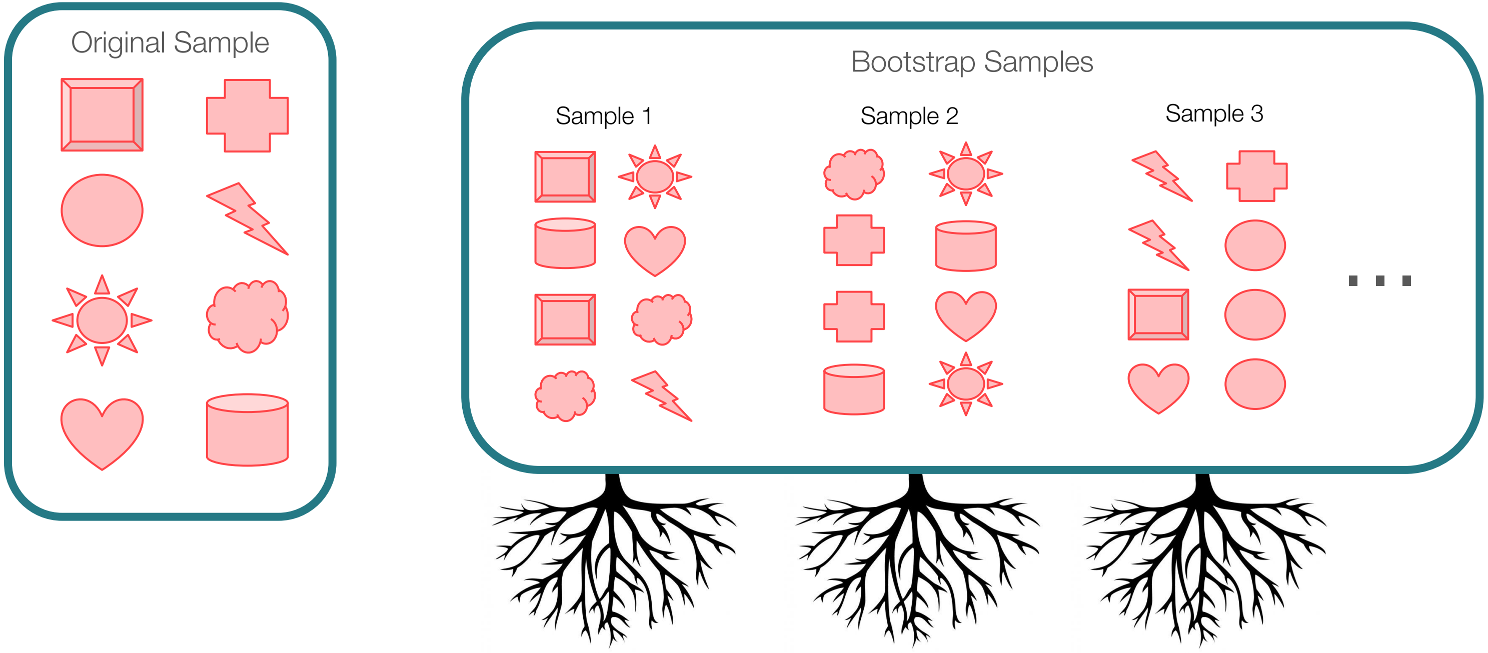

Sample 2



# Bagging


Original Sample
Bootstrap Samples
Sample 1
Sample 2
Sample 3
...

# Bagging


Original Sample
Bootstrap Samples
Sample 1
Sample 2
Sample 3
...

# Growing a forest

Predictors

$$X_1, X_2, ..., X_k$$

# Growing a forest

Predictors

$$X_1, X_2, \ldots, X_k$$

Sample of Predictors

$$X_{j_1}, X_{j_2}, \ldots, X_{j_{\lceil \sqrt{k} \rceil}}$$

# Growing a forest

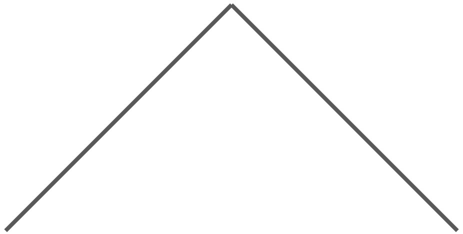
Predictors

$$X_1, X_2, ..., X_k$$

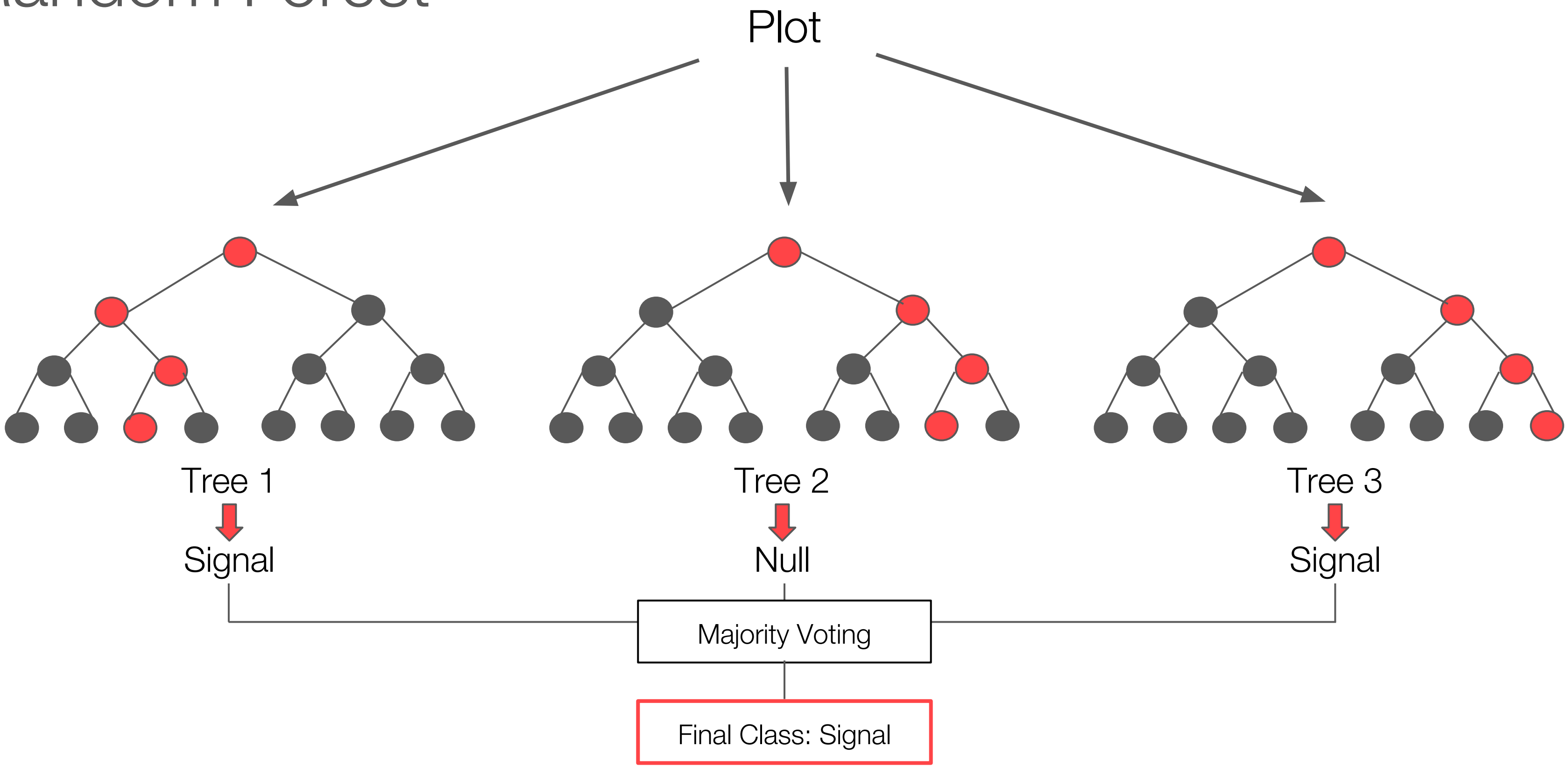
Sample of Predictors

$$X_{j_1}, X_{j_2}, ..., X_{j_{\lceil \sqrt{k} \rceil}}$$

Split! $\rightarrow$ $X_{j_i}$



# Random Forest



# Back to hurricane data...

Accuracy of decision tree:

89%

Accuracy of random forest:

96%

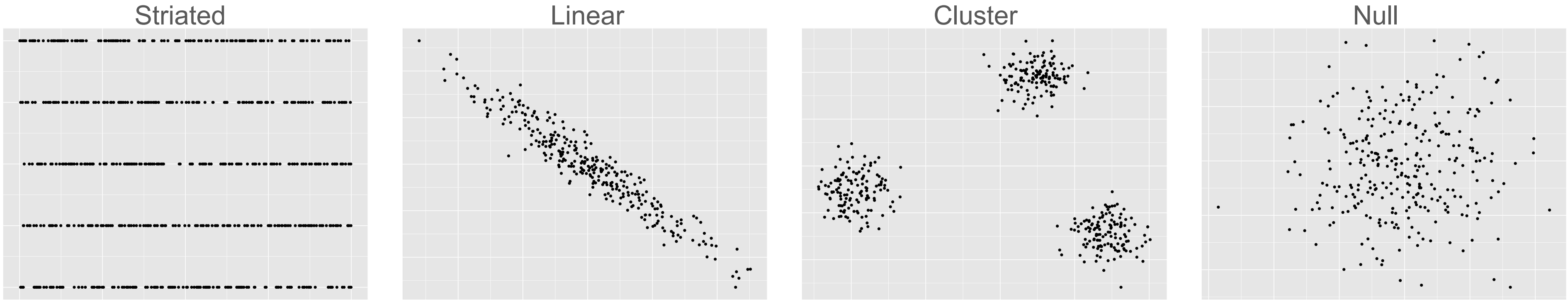
# Training Data

# Primary Family Data

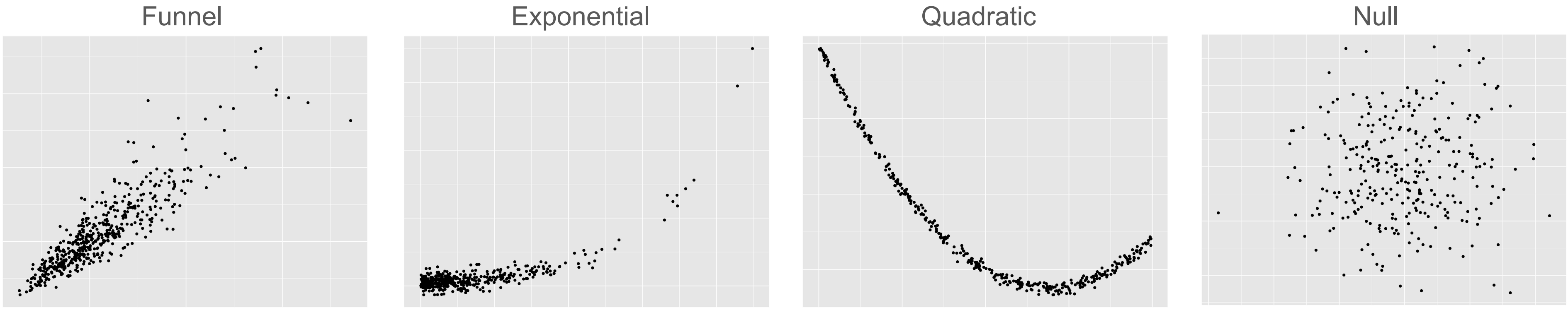
- Striated
- Linear
- Cluster
- Funnel
- Exponential
- Quadratic

- 14865 signal plots
- 14865 null plots

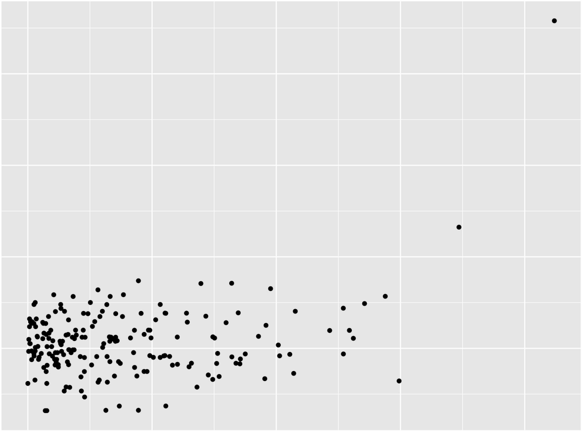
# Primary Family Data



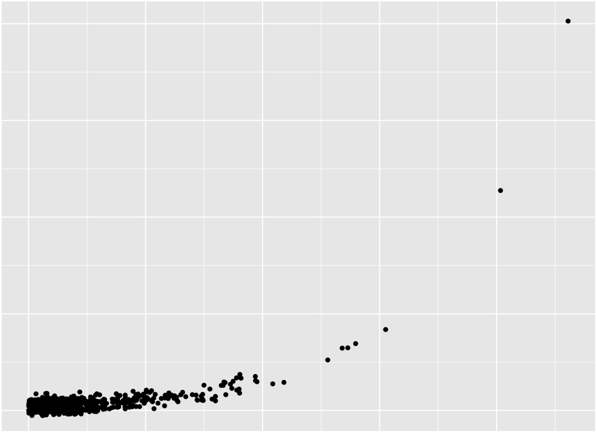
# Primary Family Data



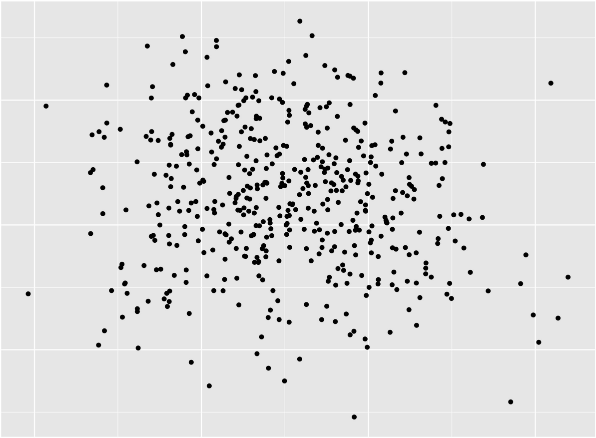
# Training Data: Exponential



Signal

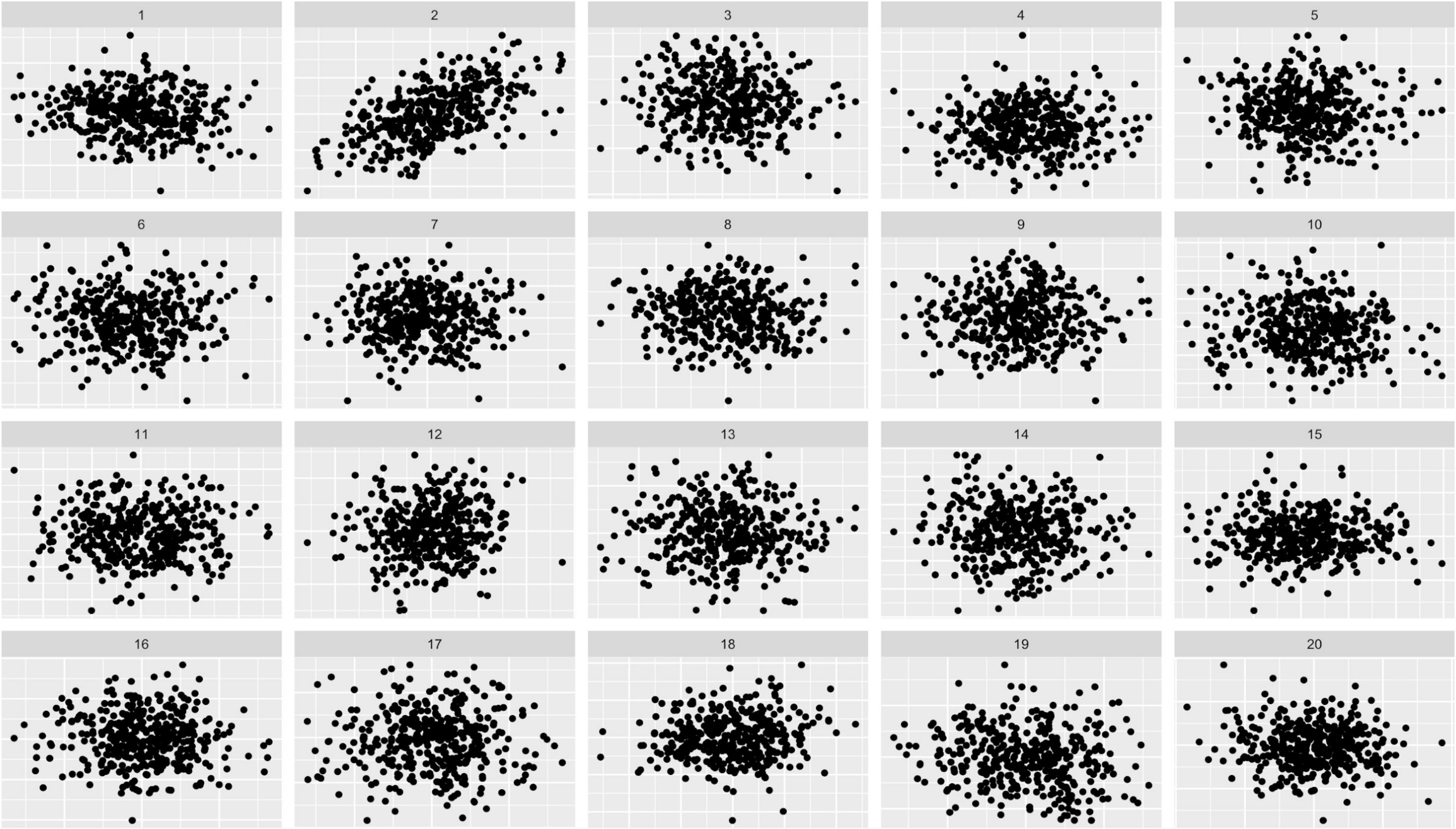


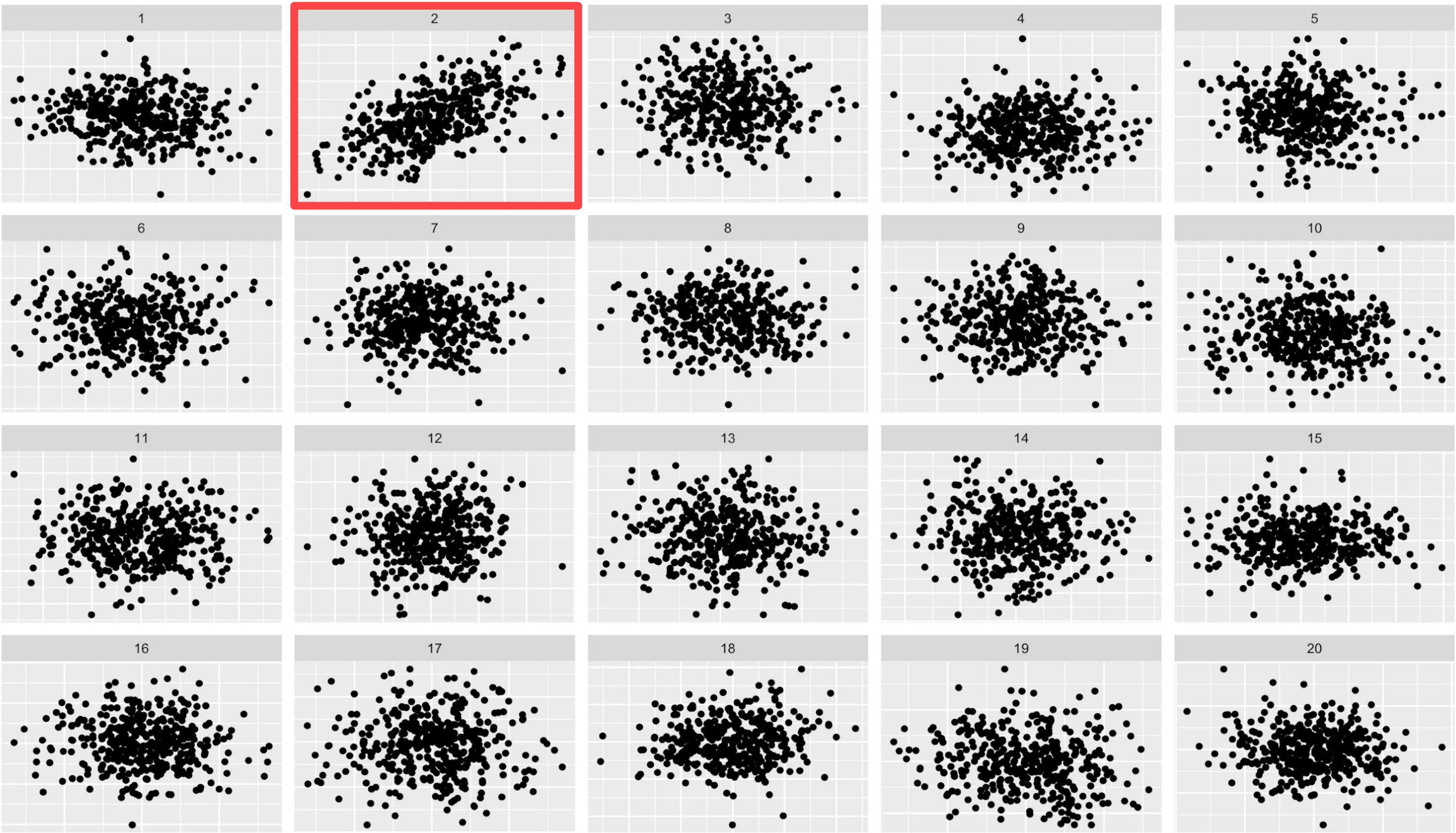
Signal

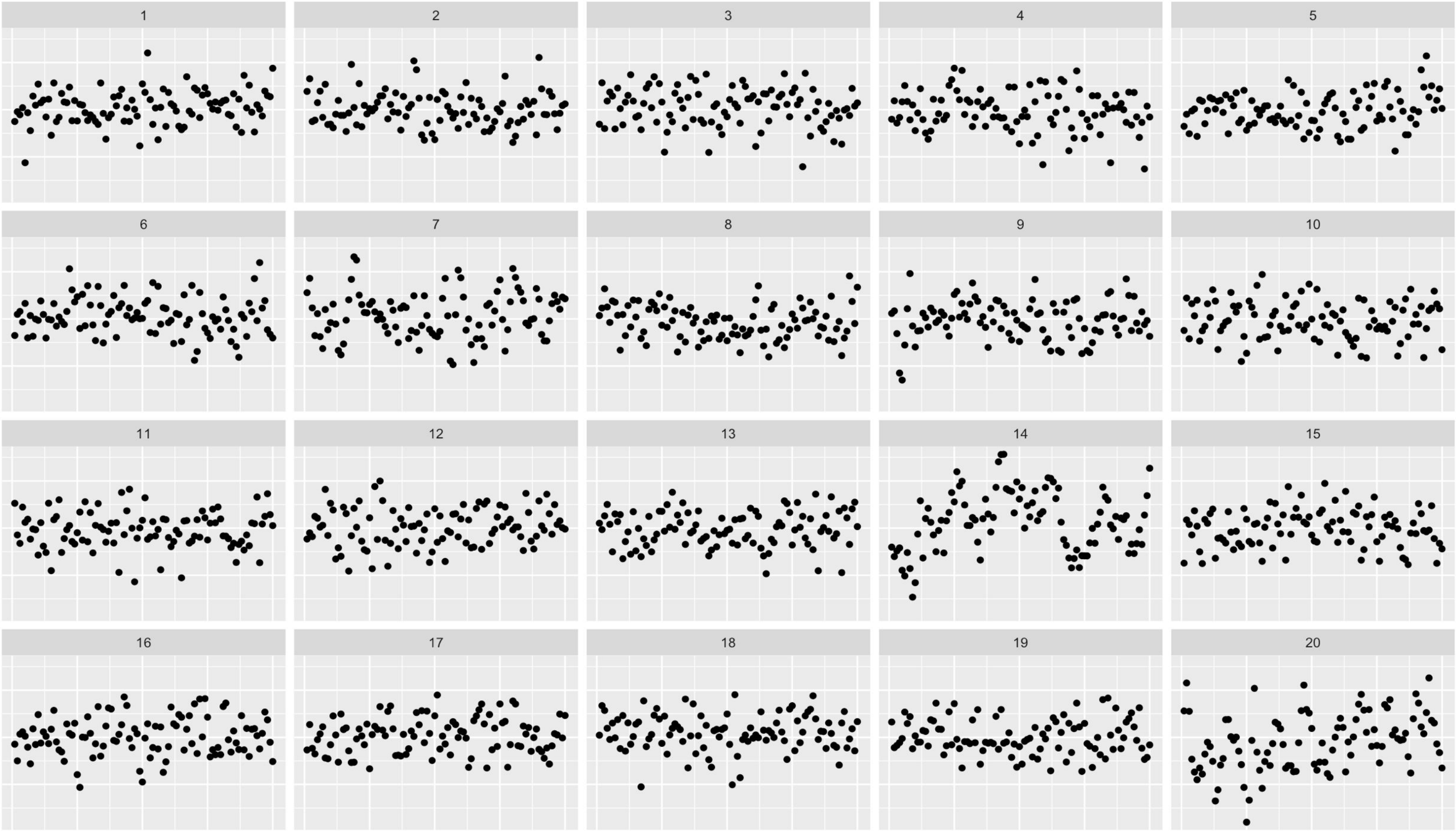


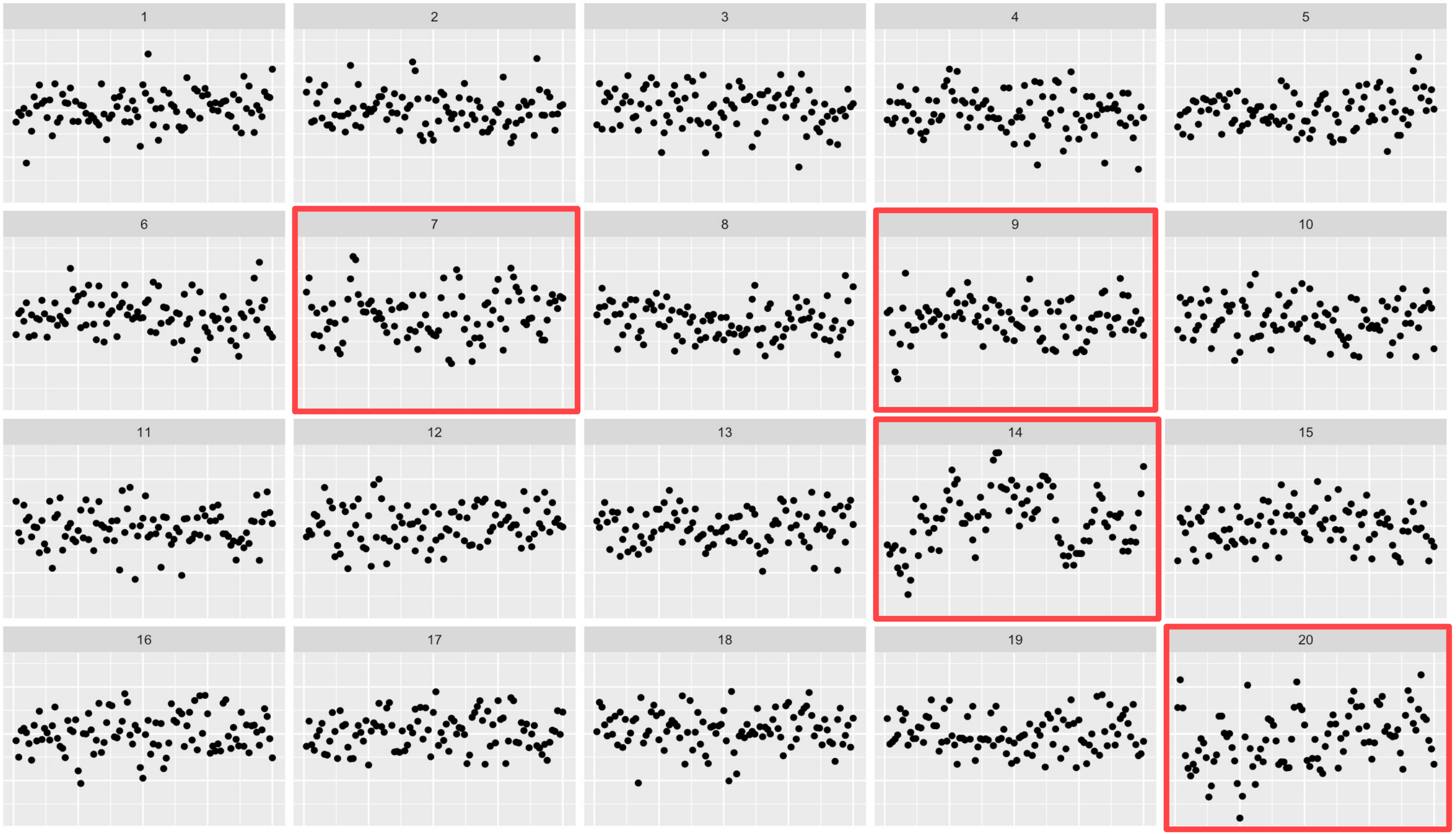
Null

# Scatterplot Lineups: Testing the Model


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

# Lineups: Model Accuracy

| *One Signal Plot* | Mahalanobis Accuracy | Random Forest Accuracy |
|---|---|---|
| Linear Trend | .857 | .992 |
| Primary Family | .952 | .972 |
| *Unknown Signal Plots* | | |
| Linear Trend | .628 (.372) | .932 (.053) |
| Primary Family | .722 (.278) | .979 (.030) |
| Lineups Per Dataset | 1000 | |

*Note:* Rate of false positives is given in parentheses

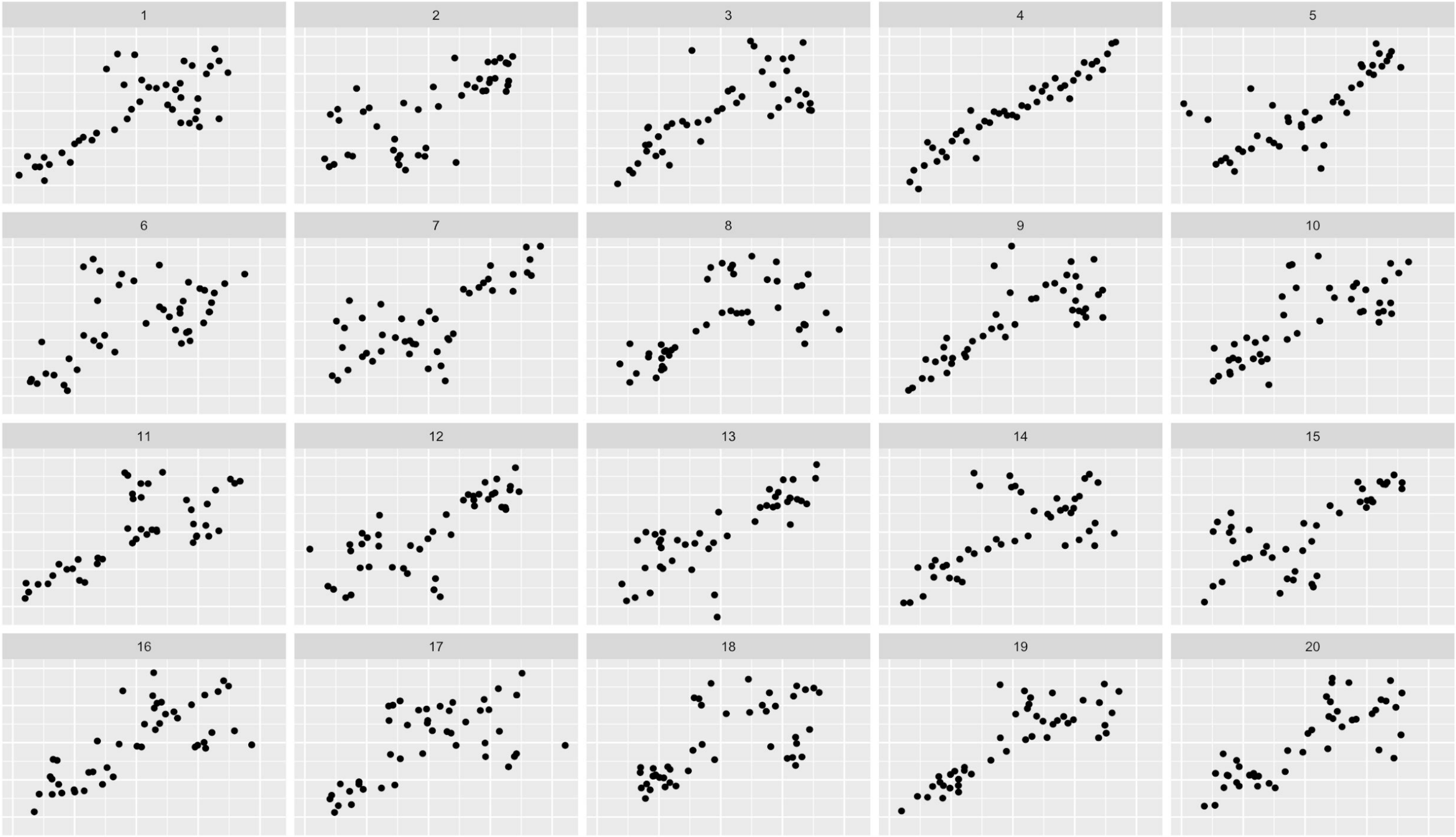
# ISU Data

Lineup perception study from Iowa State

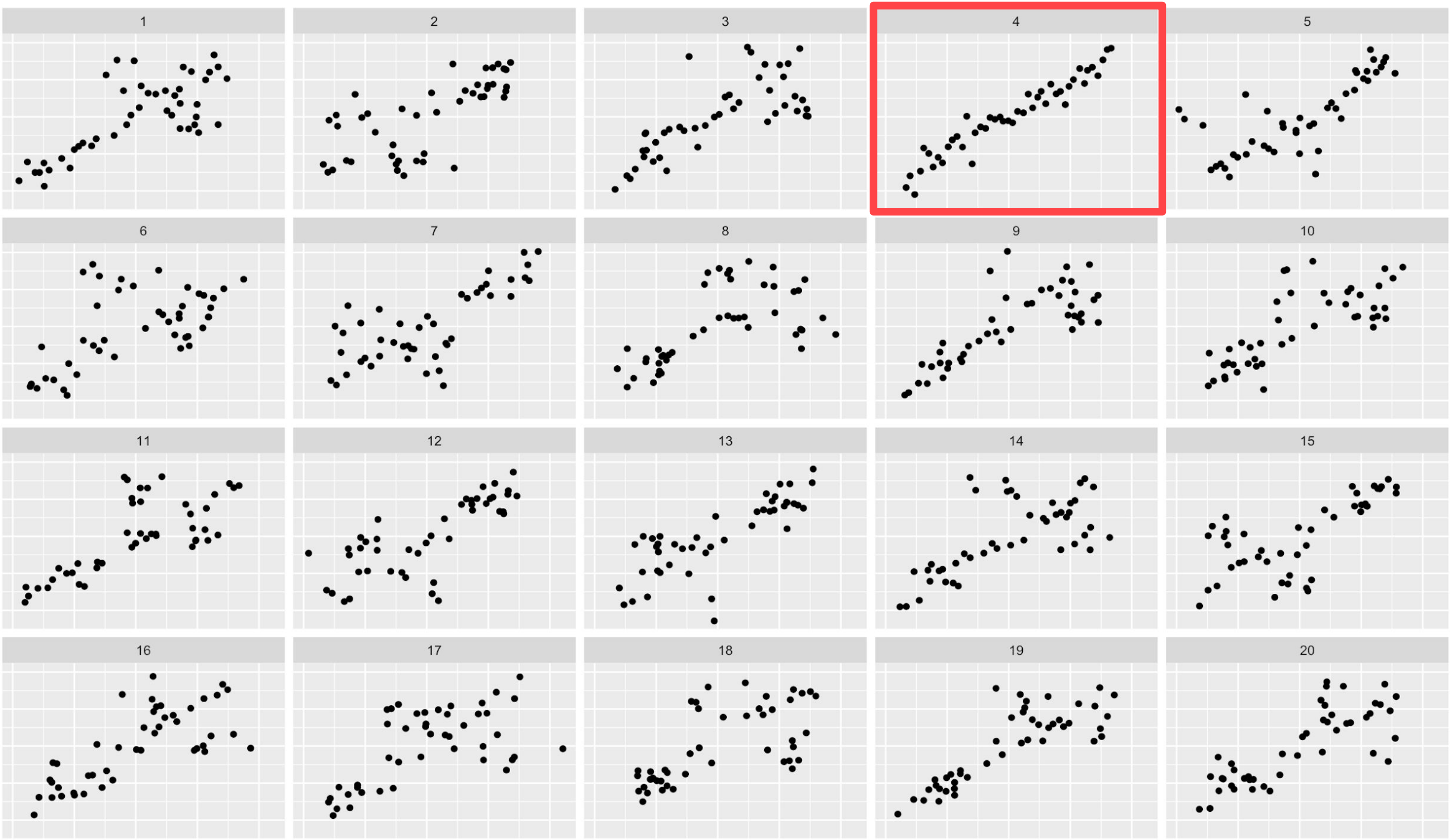
Hybrid linear + cluster plots

- 20 One-Signal Lineups
- 27 Multiple-Signal Lineups

# ISU Data



# ISU Data



# ISU Data: Model Accuracy

| | Single Signal Plot | Unknown Signal Plots |
|---|---|---|
| Mahalanobis Accuracy | .600 | .537 (.433) |
| Random Forest Accuracy | 1.000 | .926 (.077) |
| Total Number of Lineups | 20 | 27 |

*Note:* Rate of false positives is given in parentheses

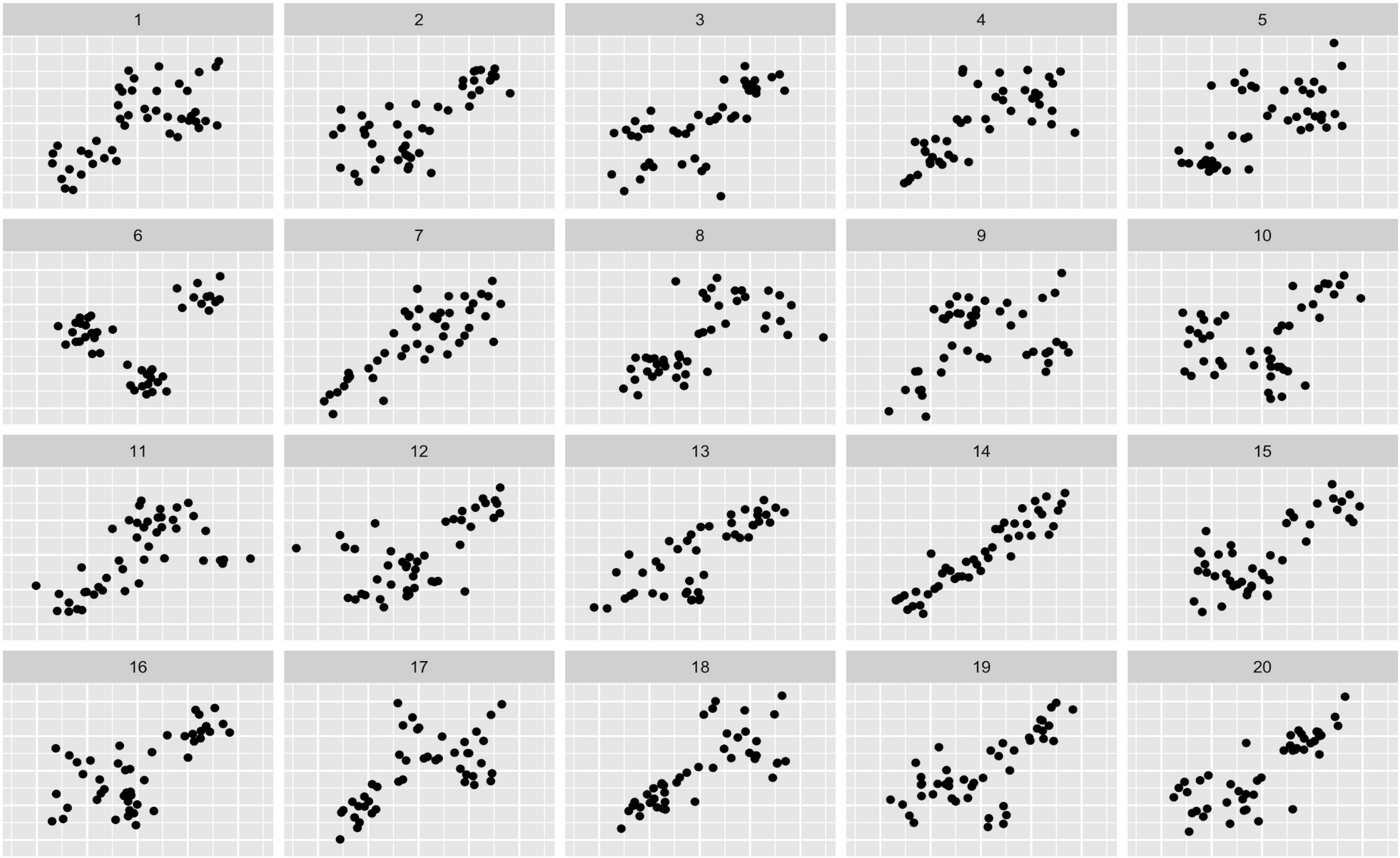
# Carleton Study

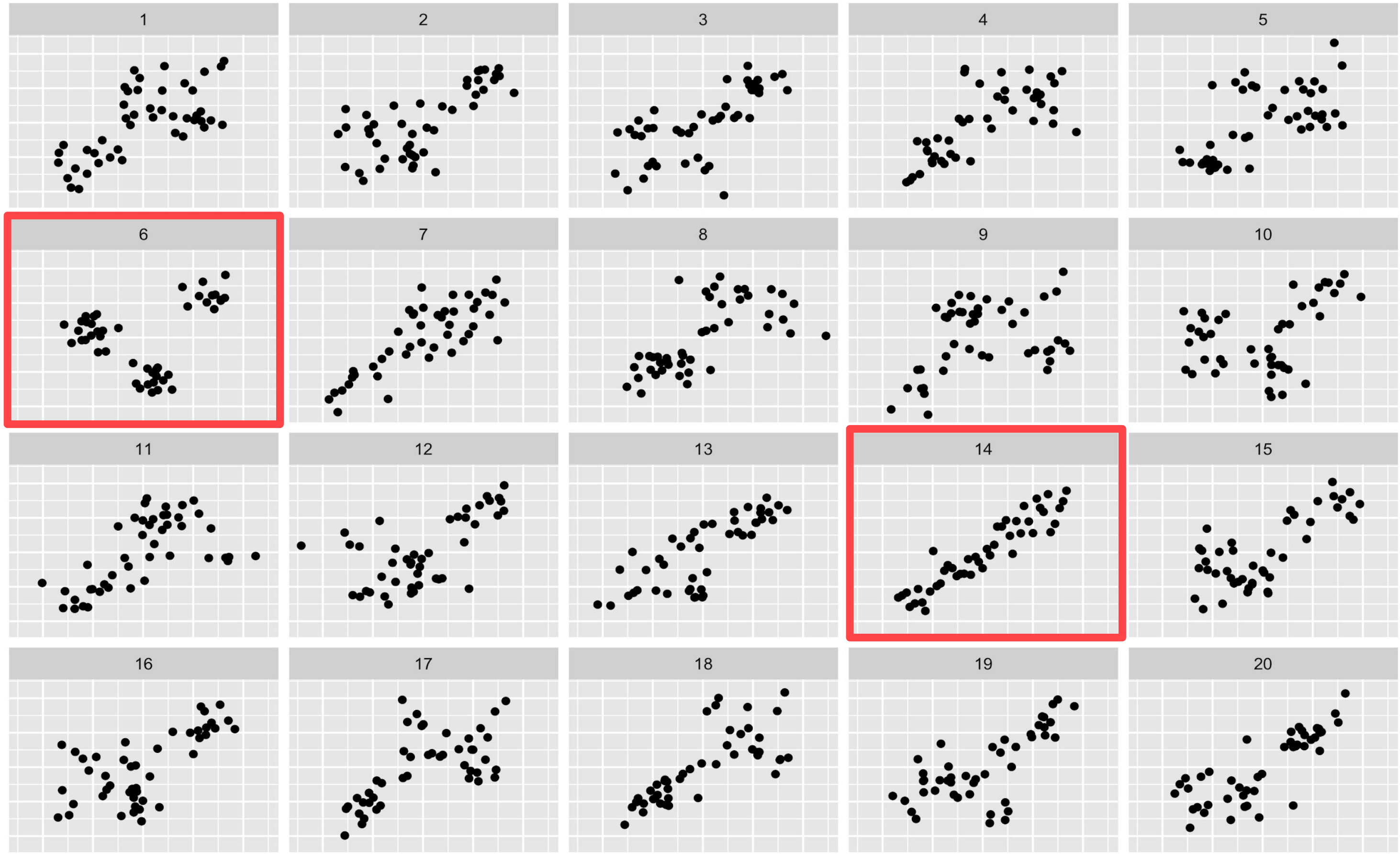
# Comparison to Human Perception

Participants: 50 Carleton students

Procedure: 9 lineups shown to each person

- 6 lineups had 1 target plot
- 3 lineups had unknown number of target plots


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

# Results

| | Single Signal Plot | Unknown Signal Plots |
|---|---|---|
| Participant Accuracy | .805 | .720 (.105) |
| Mahalanobis Accuracy | .750 | .628 (.291) |
| Random Forest Accuracy | .917 | .917 (.083) |

*Note:* Rate of false positives is given in parentheses

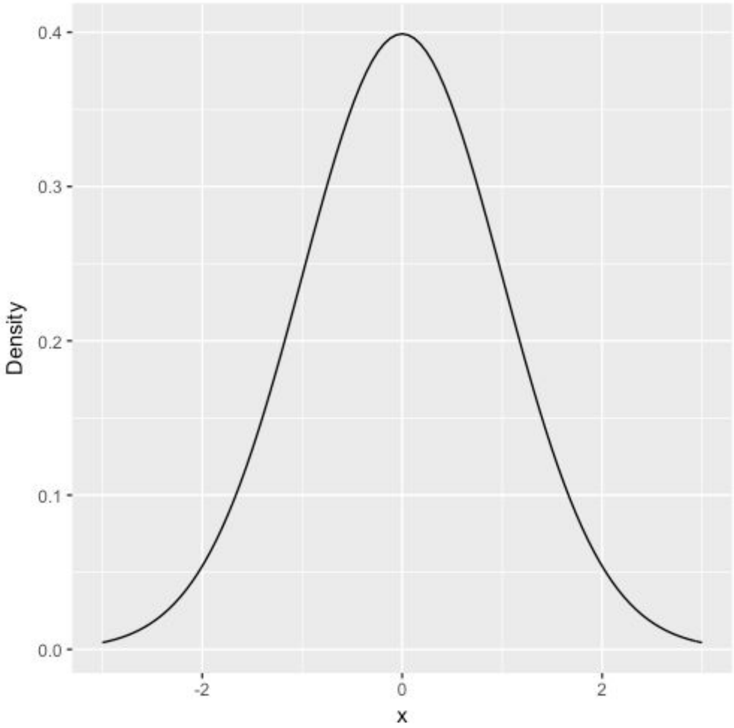
# Lineups: Model Accuracy

| *One Signal Plot* | Mahalanobis Accuracy | Random Forest Accuracy |
|---|---|---|
| Time Series | .248 | .425 |
| QQ Plots | .210 | .500 |
| *Unknown Signal Plots* | | |
| Time Series | .579 (.421) | .566 (.464) |
| QQ Plots | .589 (.411) | .761 (.248) |
| Lineups Per Dataset | 1000 | |

*Note:* Rate of false positives is given in parentheses

Can scagnostics help with other types of plots?

# Assessing normality

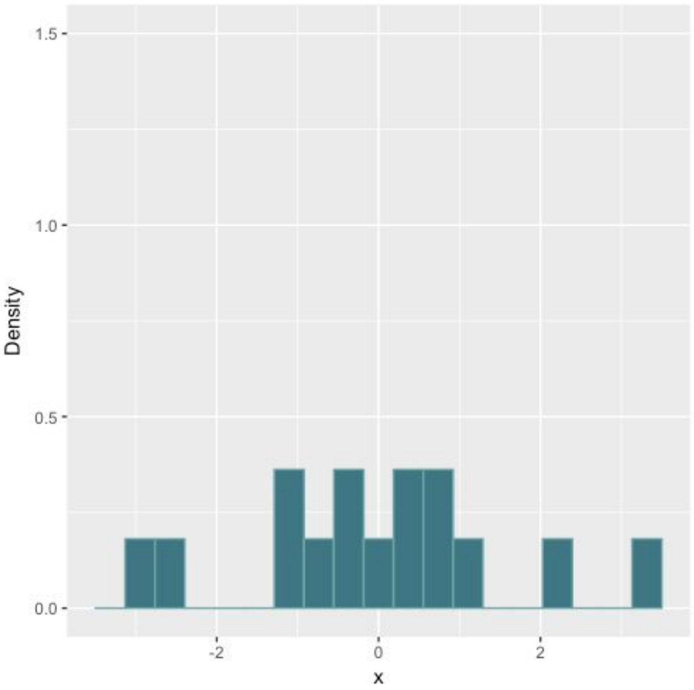


# Assessing normality

Some data

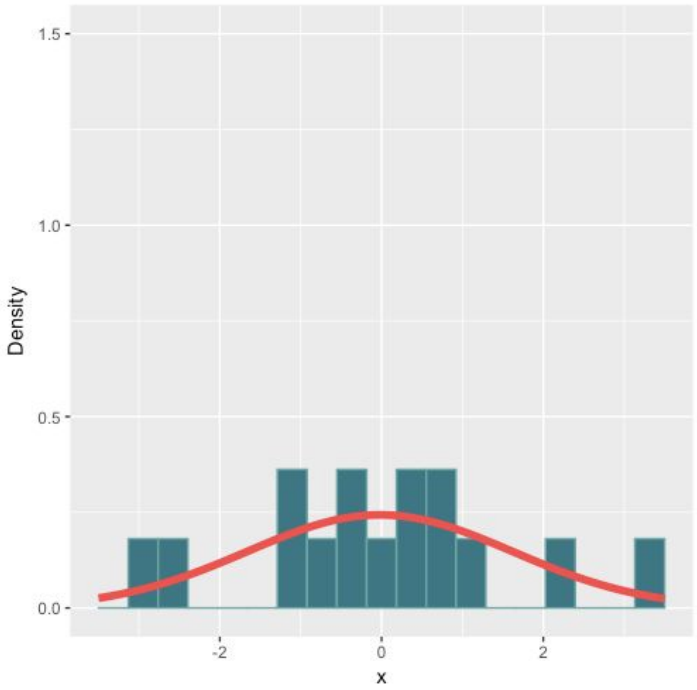
# Assessing normality

Histogram



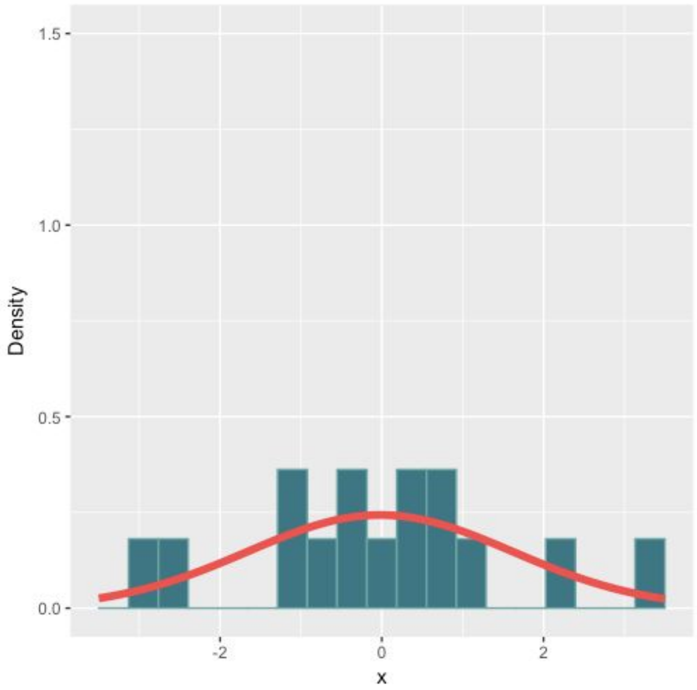
# Assessing normality

Histogram

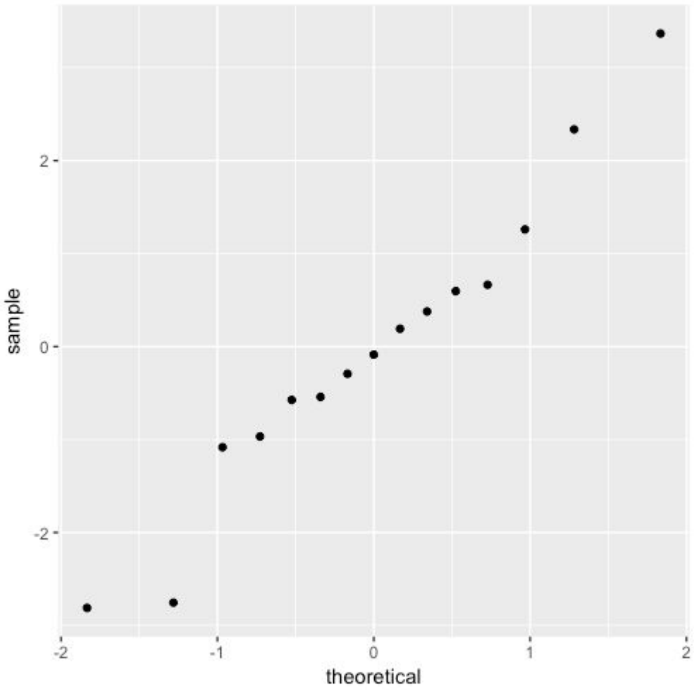


# Assessing normality

Histogram

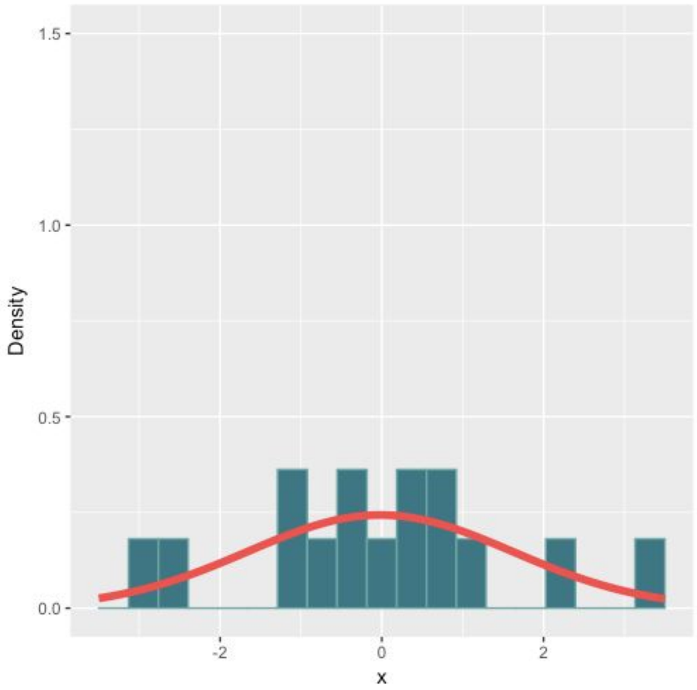



QQ Plot

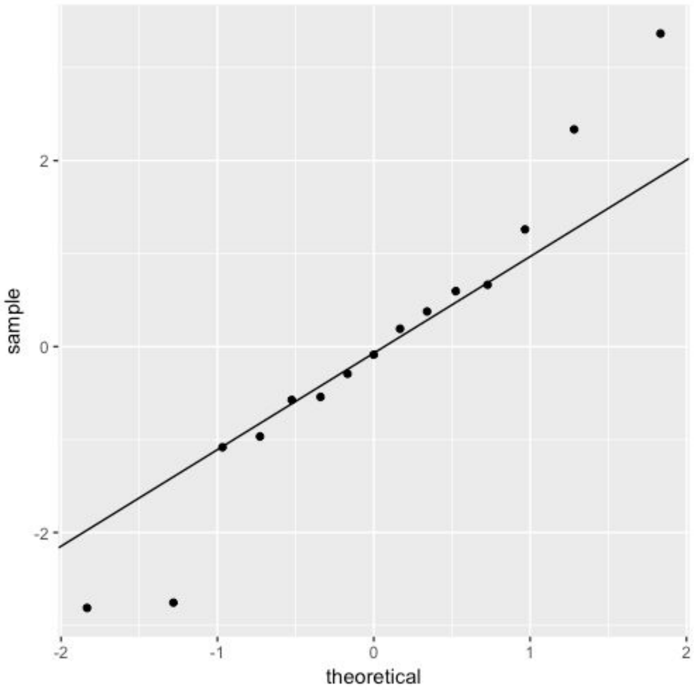


# Assessing normality

Histogram

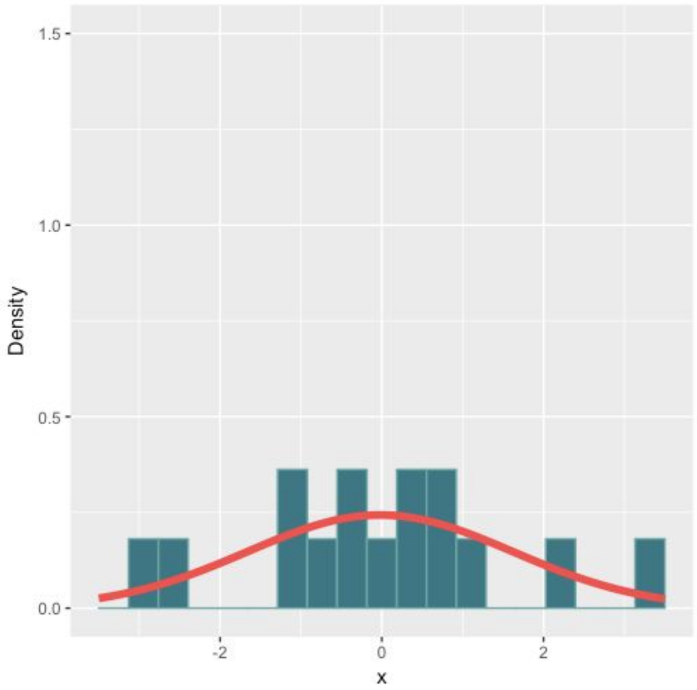



QQ Plot

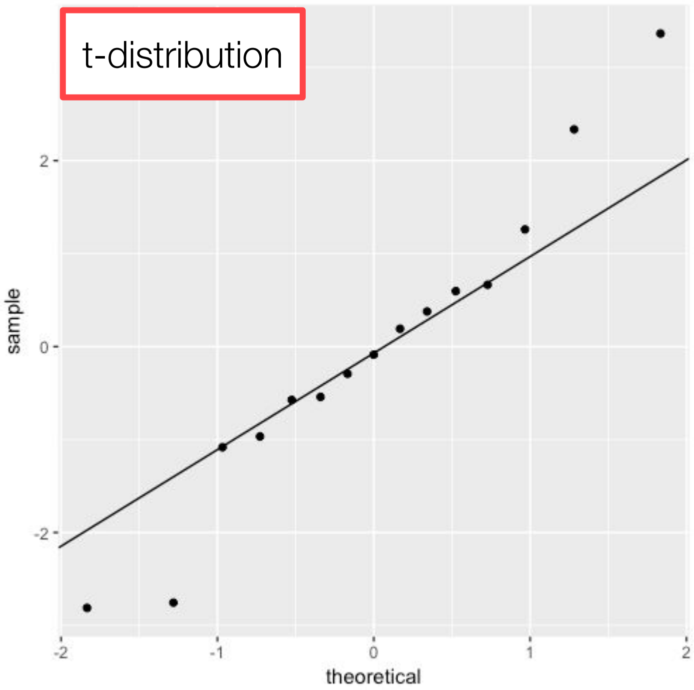


# Assessing normality

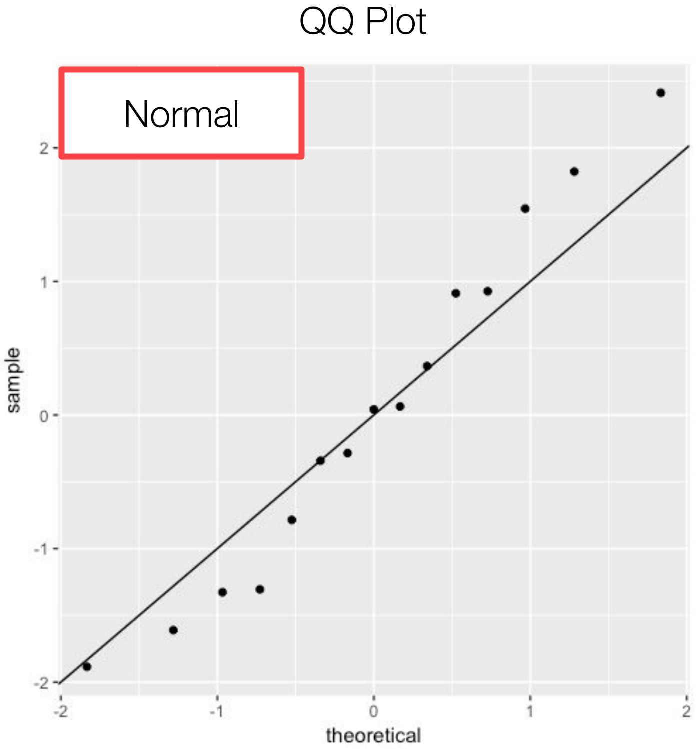
Histogram

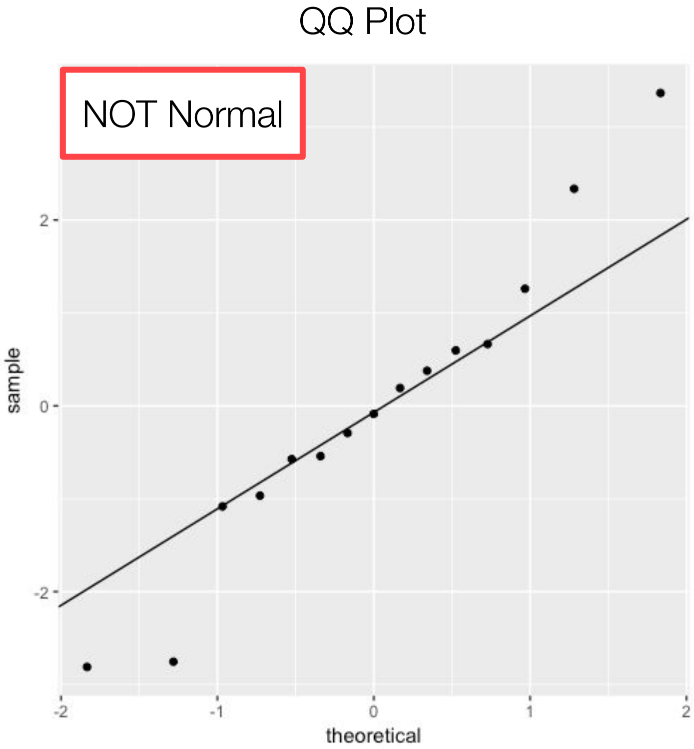


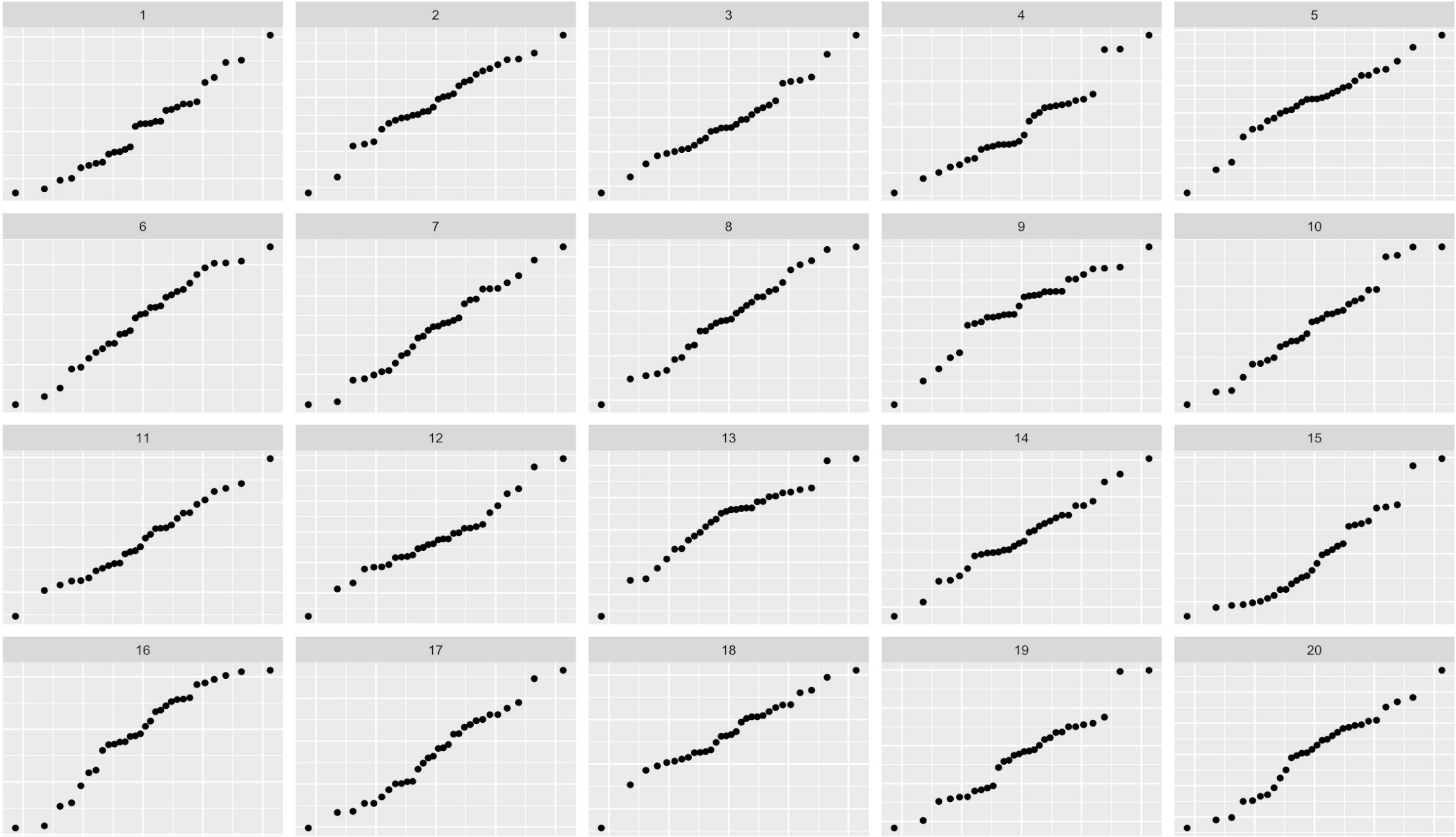

QQ Plot

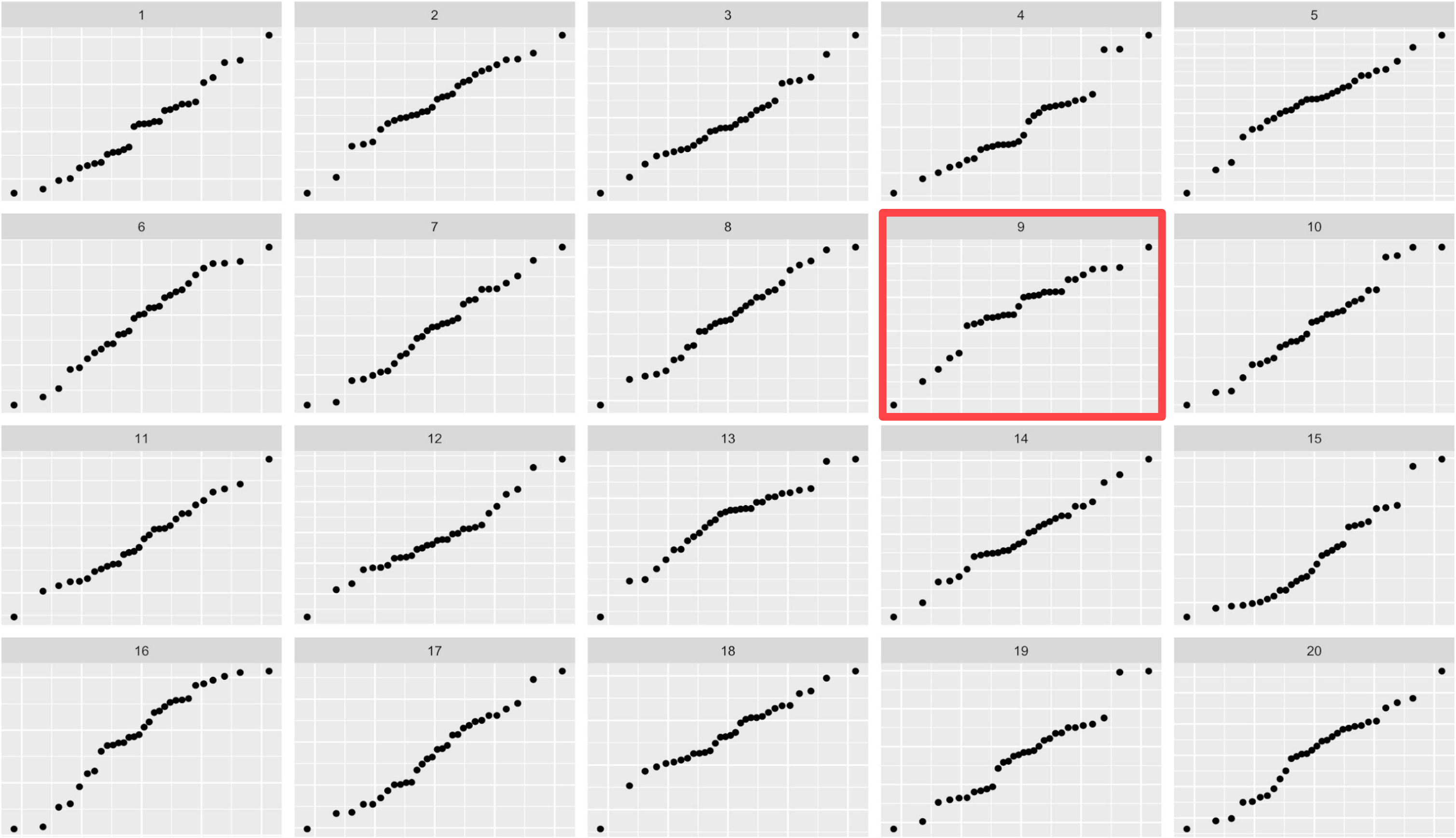


# Assessing normality




1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

# Can scagnostics predict non-normality?

# Can scagnostics predict non-normality?

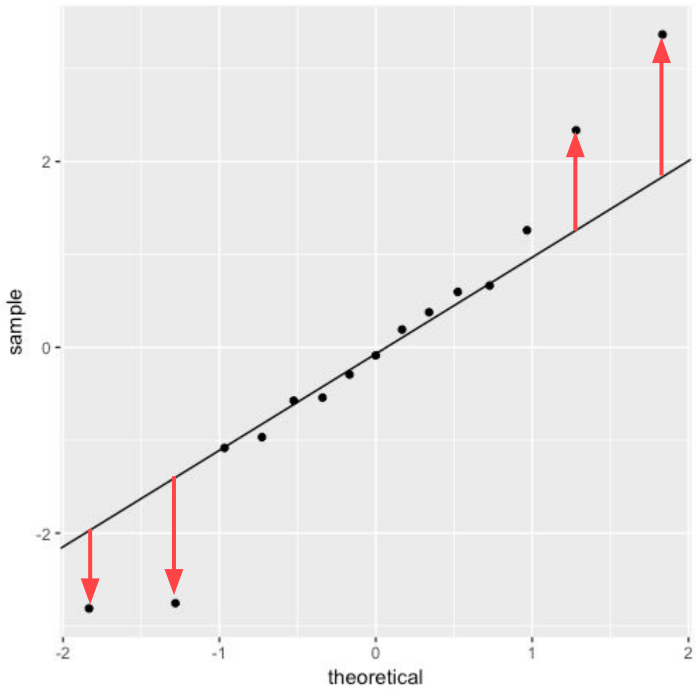
Shape

- Stringy
- Convex
- Skinny
- Clumpy
- Striated

Density and Association

- Monotonic
- Outlying
- Sparse
- Skewed

# Can scagnostics predict non-normality?



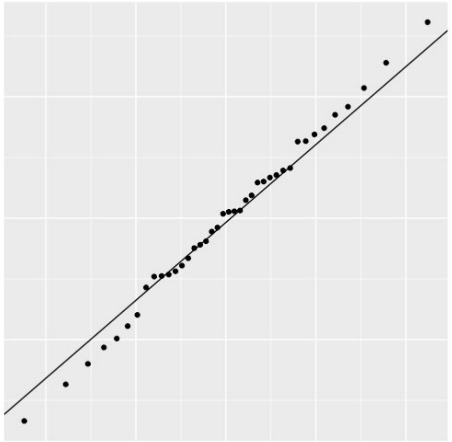
# A new scagnostic

$$c_{deviation} = \frac{1}{n}\sum_{i=1}^{k}((x_i^2 + 1)(y_i - x_i)^2)$$

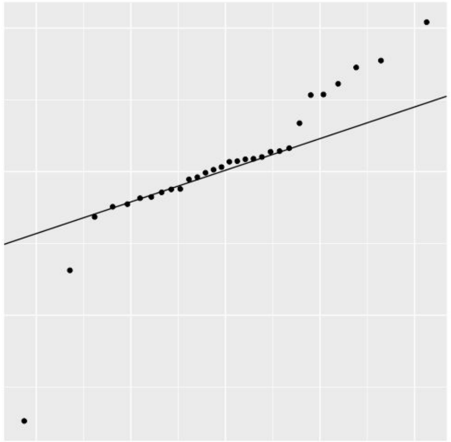
# A new scagnostic

$$c_{deviation} = \frac{1}{n}\sum_{i=1}^{k}((x_i^2+1)(y_i-x_i)^2)$$

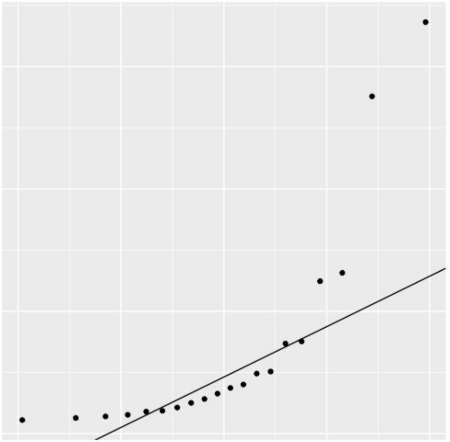
Low (0.007)



Medium (0.549)



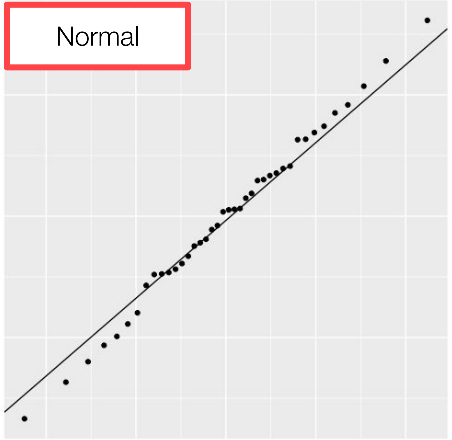
High (1.11)



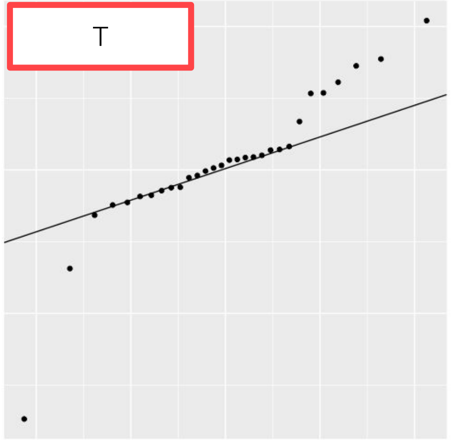
# A new scagnostic

$$c_{deviation} = \frac{1}{n}\sum_{i=1}^{k}((x_i^2+1)(y_i-x_i)^2)$$

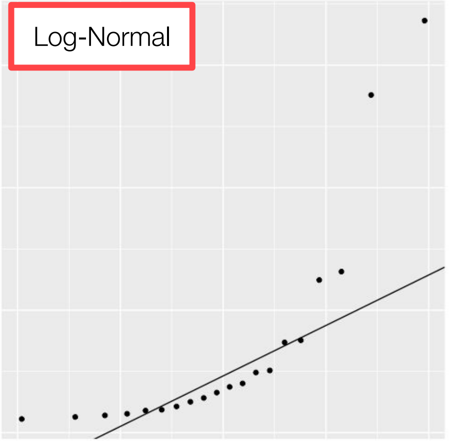
Low (0.007)




Medium (0.549)




High (1.11)



# Performance

16,000 QQPlots were generated from a variety of distributions (normal, t, log-normal, exponential, and Chi-Squared).

# Performance

16,000 QQPlots were generated from a variety of distributions (normal, t, log-normal, exponential, and Chi-Squared).

Accuracy:
Anderson-Darling
Normality Test

81.8%

# Performance

16,000 QQPlots were generated from a variety of distributions (normal, t, log-normal, exponential, and Chi-Squared).

Accuracy:
Anderson-Darling
Normality Test

81.8%

Accuracy:
Model without Deviation
Scagnostic

78.6%

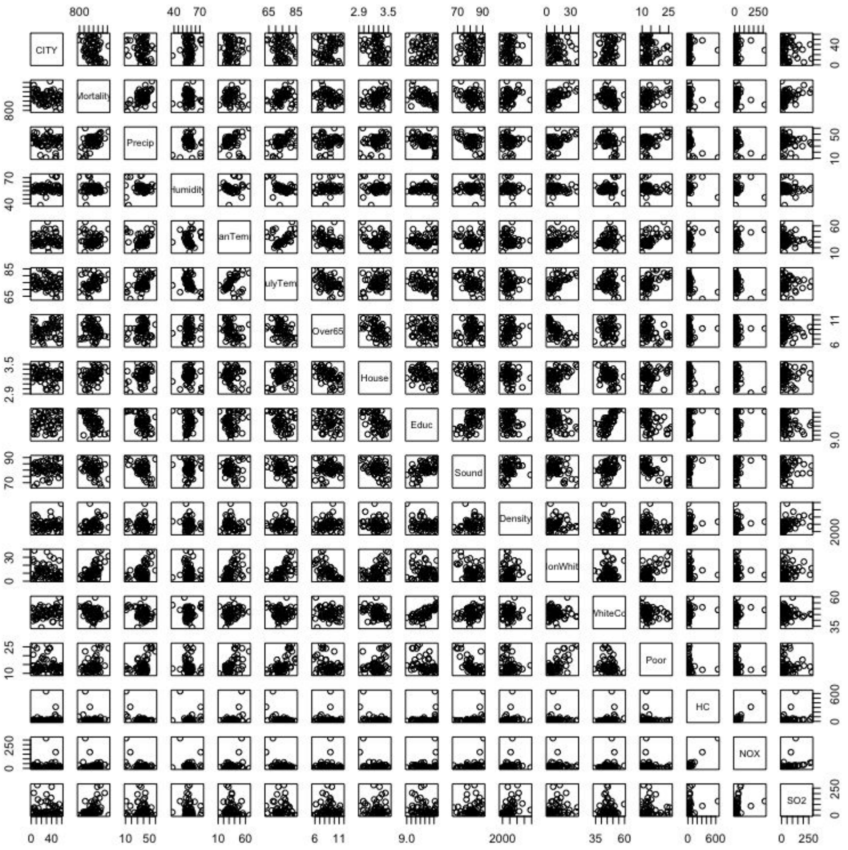
# Performance

16,000 QQPlots were generated from a variety of distributions (normal, t, log-normal, exponential, and Chi-Squared).

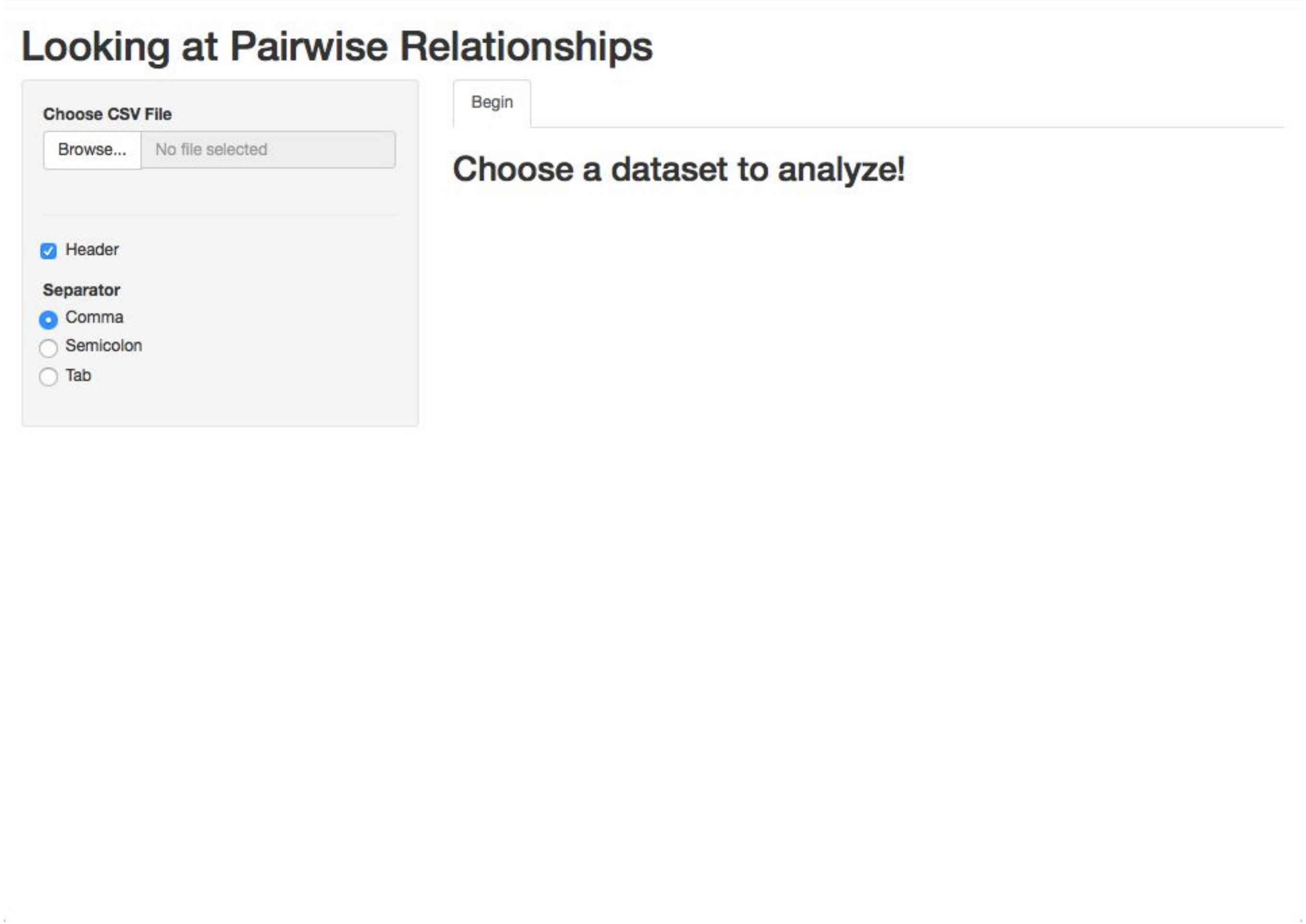
| Accuracy: Anderson-Darling Normality Test | Accuracy: Model without Deviation Scagnostic | Accuracy: Model with Deviation Scagnostic |
|---|---|---|
| 81.8% | 78.6% | 84.0% |

# Conclusions

# Applications



# Our App


Looking at Pairwise Relationships
Choose CSV File
Browse...
No file selected
Header
Separator
Comma
Semicolon
Tab
Begin
Choose a dataset to analyze!

# Our App

## Looking at Pairwise Relationships

**Choose CSV File**

Browse... pollution2.csv

Upload complete

☑ Header

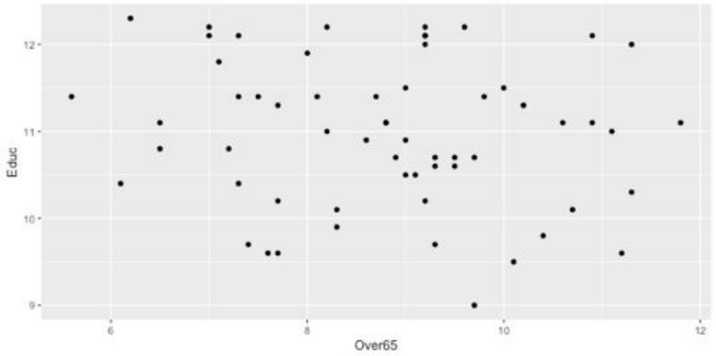
**Separator**

◉ Comma
○ Semicolon
○ Tab

Plot Type: exponential | funnel | linear trend | null | quadratic

## null

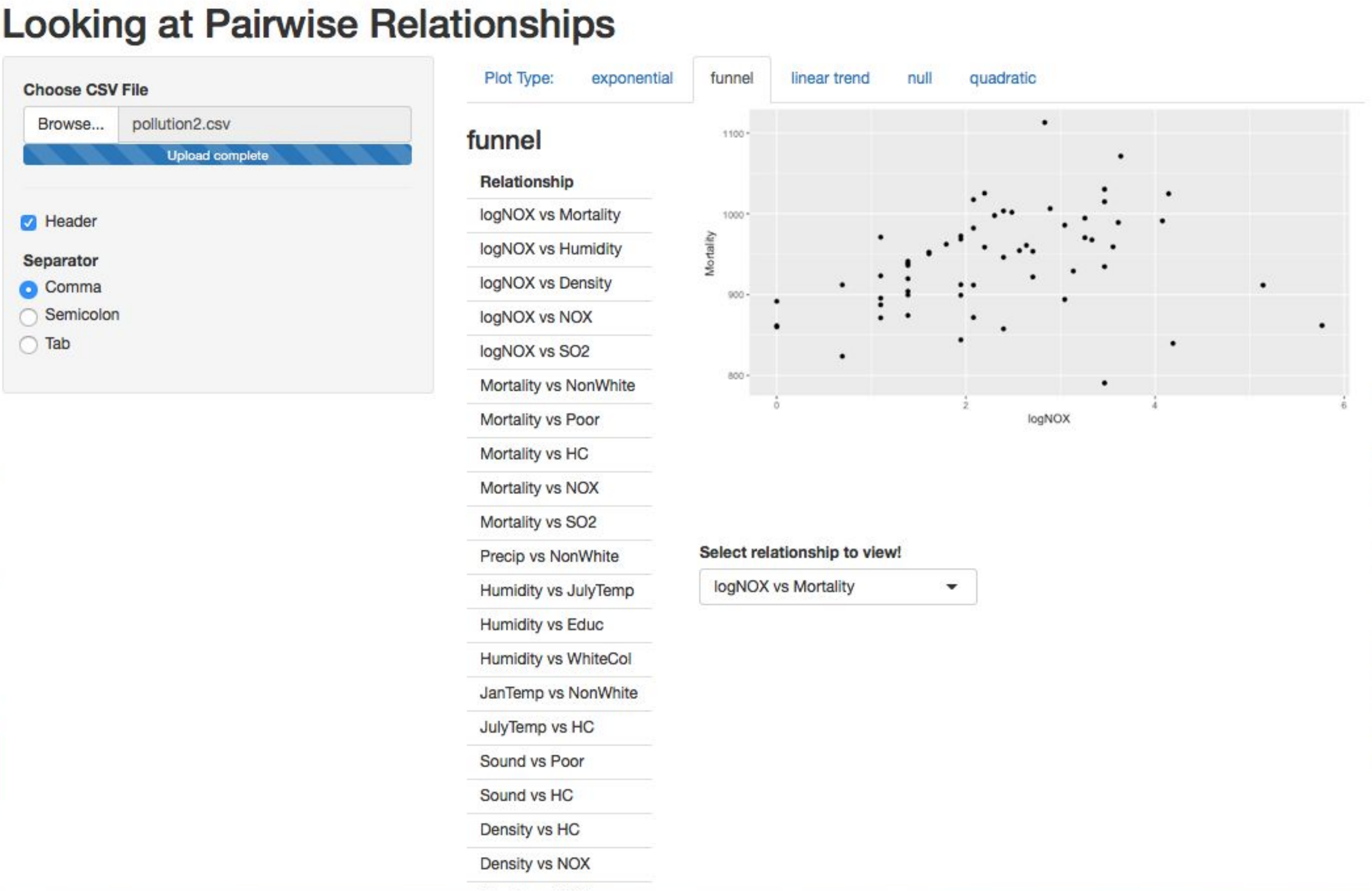
| Relationship |
|---|
| logNOX vs Precip |
| logNOX vs JanTemp |
| logNOX vs JulyTemp |
| logNOX vs Over65 |
| logNOX vs House |
| logNOX vs Educ |
| logNOX vs NonWhite |
| logNOX vs WhiteCol |
| logNOX vs Poor |
| Mortality vs Humidity |
| Mortality vs JanTemp |
| Mortality vs JulyTemp |
| Mortality vs Over65 |
| Mortality vs Sound |
| Mortality vs Density |
| Precip vs Humidity |
| Precip vs JanTemp |
| Precip vs JulyTemp |
| Precip vs Over65 |
| Precip vs House |
| Precip vs Density |




**Select relationship to view!**

Over65 vs Educ ▾

# Our App


Looking at Pairwise Relationships
Choose CSV File
Browse...
pollution2.csv
Upload complete
Header
Separator
Comma
Semicolon
Tab
Plot Type:
exponential
funnel
linear trend
null
quadratic
funnel
Relationship
logNOX vs Mortality
logNOX vs Humidity
logNOX vs Density
logNOX vs NOX
logNOX vs SO2
Mortality vs NonWhite
Mortality vs Poor
Mortality vs HC
Mortality vs NOX
Mortality vs SO2
Precip vs NonWhite
Humidity vs JulyTemp
Humidity vs Educ
Humidity vs WhiteCol
JanTemp vs NonWhite
JulyTemp vs HC
Sound vs Poor
Sound vs HC
Density vs HC
Density vs NOX
Mortality
logNOX
Select relationship to view!
logNOX vs Mortality

# Acknowledgements

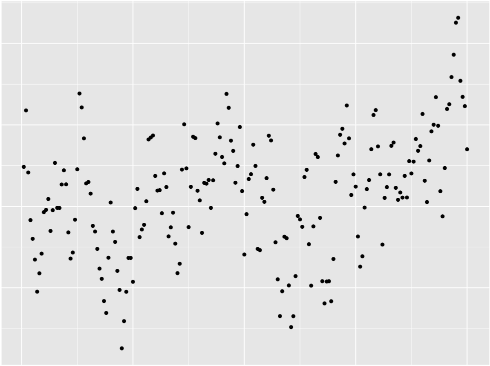
Thank you to Adam Loy, the Carleton Math & Statistics Department Faculty, our classmates, and our families.

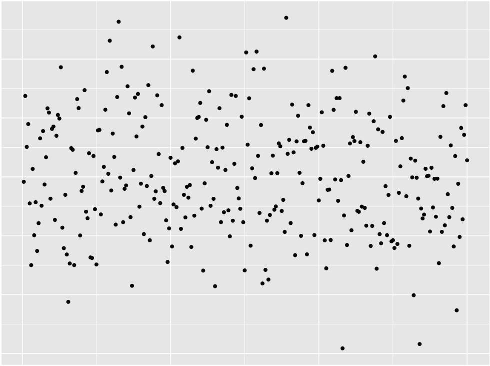
# Primary Family Models

| Model | Accuracy |
|---|---|
| K-Nearest Neighbors | 69.6% |
| Linear Discriminant Analysis | 93.9% |
| Support Vector Machine | 97.3% |
| Logistic Regression | 97.4% |
| Quadratic Discriminant Analysis | 98.1% |
| Random Forest | 98.6% |

# QQ Plots and Time Series

Time Series

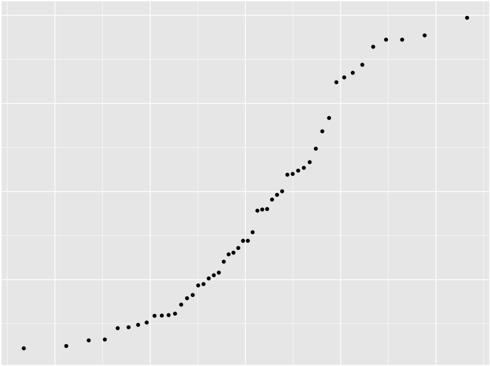


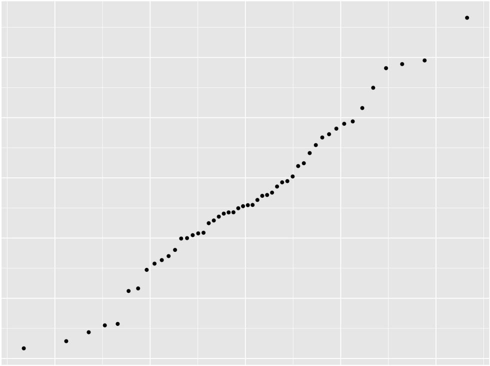


Signal

Null

QQ Plots





#5ea8a7

#ff4447

#257985