PGE 338 Data Analytics and Geostatistics

Lecture 8: Bivariate Distributions

Lecture outline . . .

- Bivariate Statistics
- Correlation

Introduction

General Concepts

Univariate

Bivariate

Correlation

Regression

Model Checking

Time Series Analysis

Spatial Analysis

Machine Learning

Uncertainty Analysis

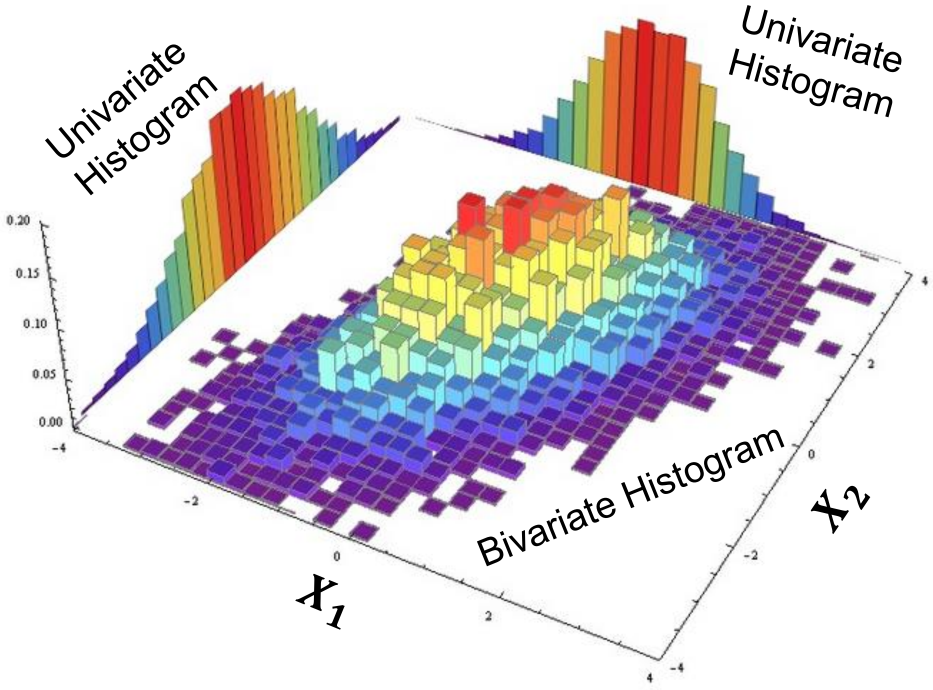
Michael Pyrcz, The University of Texas at Austin

Motivation

We deal with more than one variable or feature?

We must use calculate bivariate statistics and use them in our models.

Note, ‘bivariate’ is pertaining to 2 variables or features at a time



Univariate and bivariate distributions (modified from ElenaPhys on StackOverflow).

PGE 338 Data Analytics and Geostatistics

Lecture 8: Bivariate Distributions

Lecture outline . . .

- Bivariate Statistics

Introduction

General Concepts

Univariate

Bivariate

Correlation

Regression

Model Checking

Time Series Analysis

Spatial Analysis

Machine Learning

Uncertainty Analysis

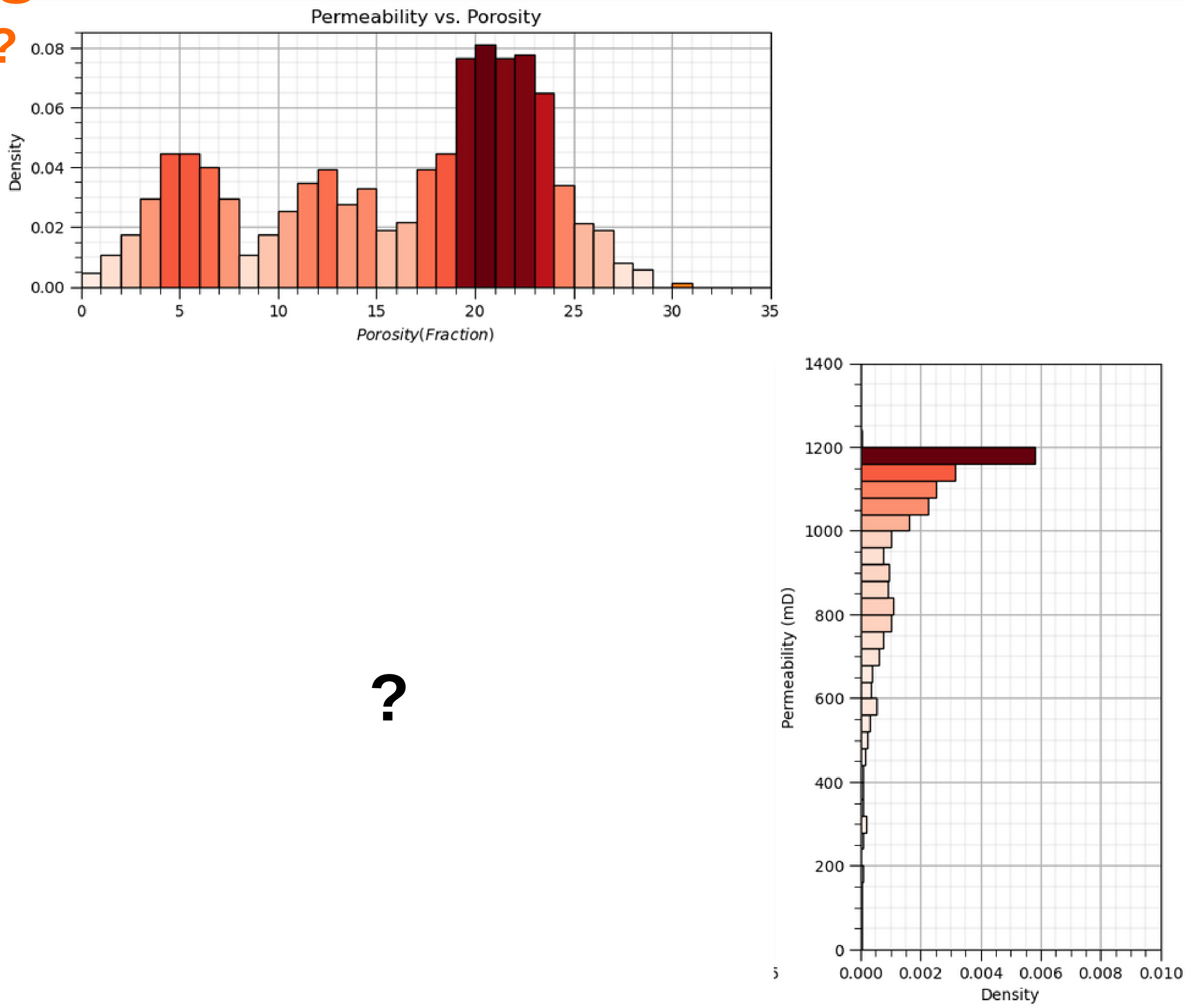
Michael Pyrcz, The University of Texas at Austin

Bivariate Statistics

What is Bivariate Analysis?

Bivariate Analysis: Understand and quantify the relationship between two variables

- Example: Porosity and permeability data
- How can we use this relationship? What would we miss if we only looked at the 2 histograms?
- Is there a realizationship between permeability and porosity?



Porosity and permeability PDFs univariate plot modified from PythonDataBasics_Bivariate _Visualization.ipynb.

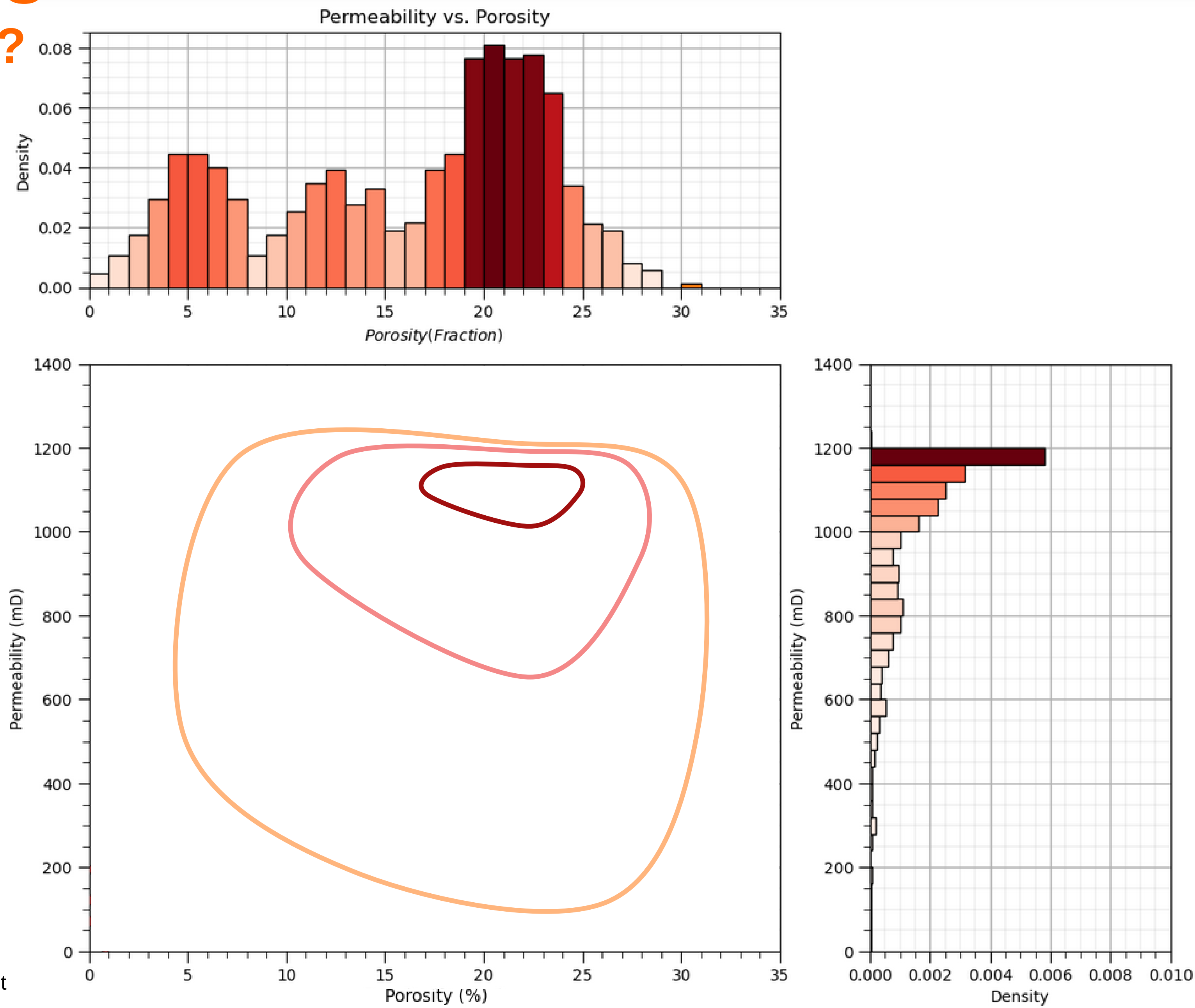
Bivariate Statistics

What is Bivariate Analysis?

Bivariate Analysis: Understand and quantify the relationship between two variables

- Example: Porosity and permeability data
- How can we use this relationship? What would we miss if we only looked at the 2 histograms?
- Is there a realizationship between permeability and porosity?

No Relationship / Independence



Porosity and permeability PDFs univariate and bivariate (schematic) plot modified from PythonDataBasics_Bivariate _Visualization.ipynb.

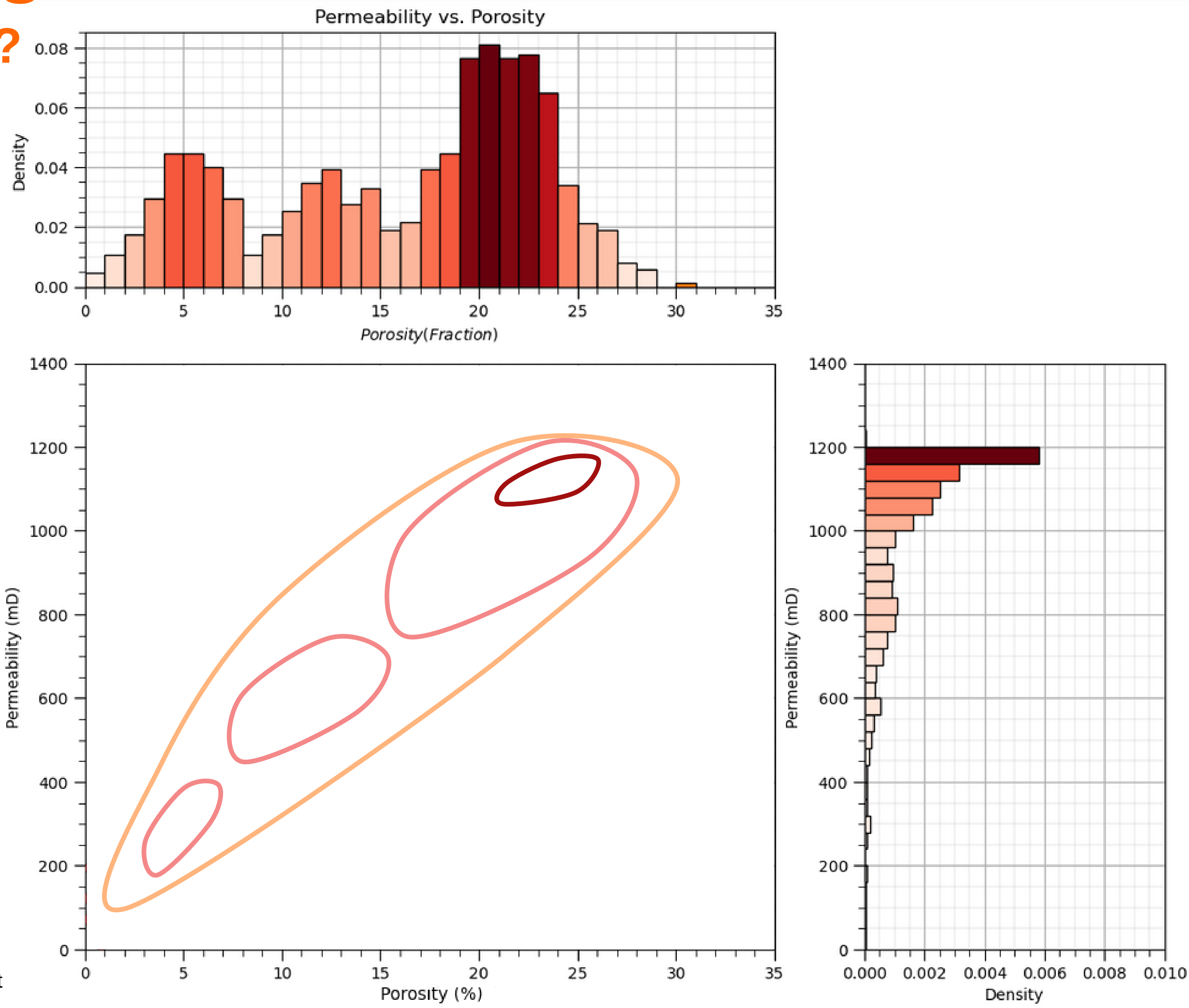
Bivariate Statistics

What is Bivariate Analysis?

Bivariate Analysis: Understand and quantify the relationship between two variables

- Example: Porosity and permeability data
- How can we use this relationship? What would we miss if we only looked at the 2 histograms?
- Is there a realizationship between permeability and porosity?

Positive Relationship
as Porosity Increases,
Permeability Increases



Porosity and permeability PDFs univariate and bivariate (schematic) plot modified from PythonDataBasics_Bivariate _Visualization.ipynb.

Bivariate Statistics

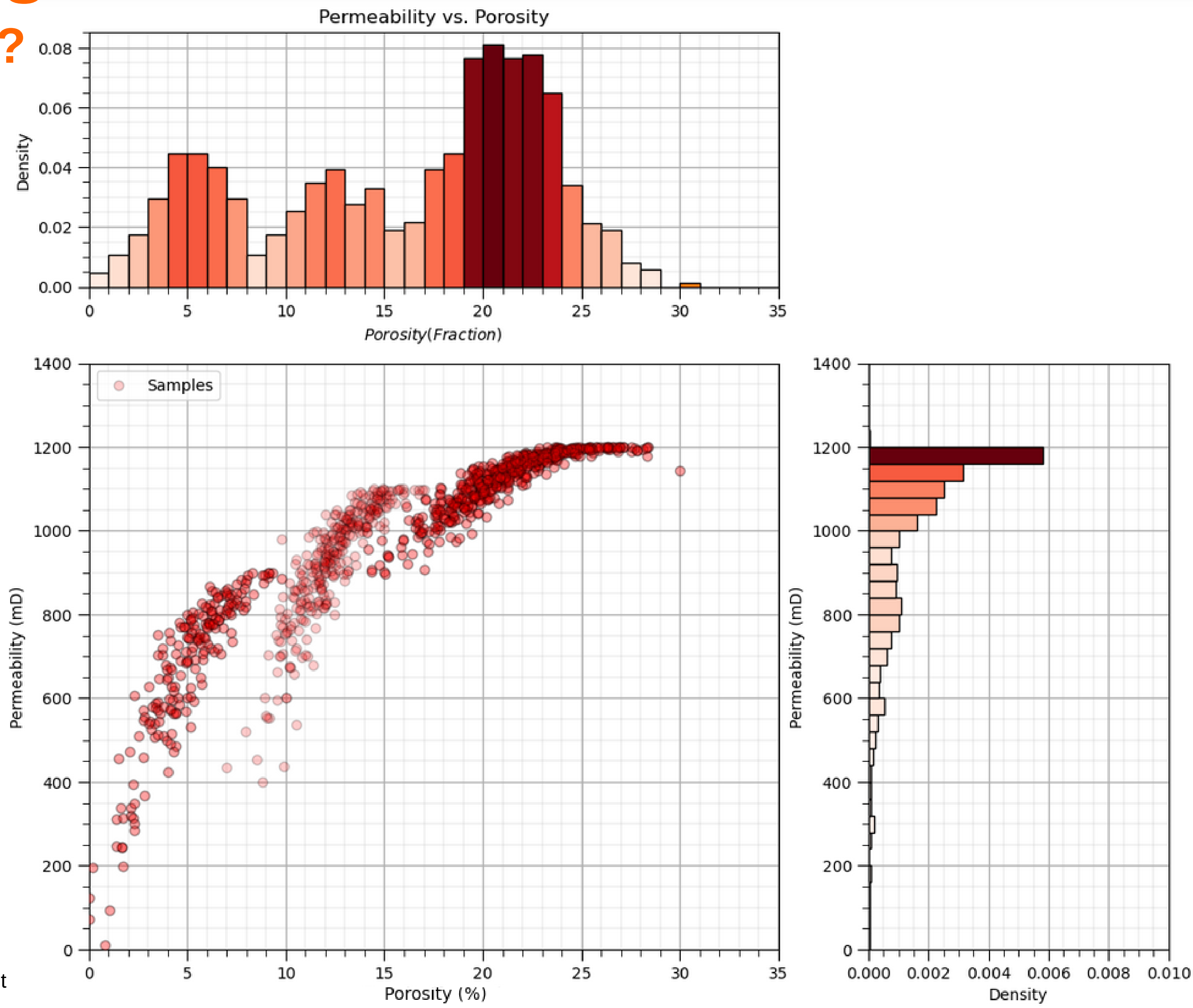
What is Bivariate Analysis?

Bivariate Analysis: Understand and quantify the relationship between two variables

- Example: Porosity and permeability data
- How can we use this relationship? What would we miss if we only looked at the 2 histograms?

- Is there a realizationship between permeability and porosity?

Now Look at The Data!



Porosity and permeability PDFs univariate and bivariate (schematic) plot modified from PythonDataBasics_Bivariate _Visualization.ipynb.

Bivariate Statistics

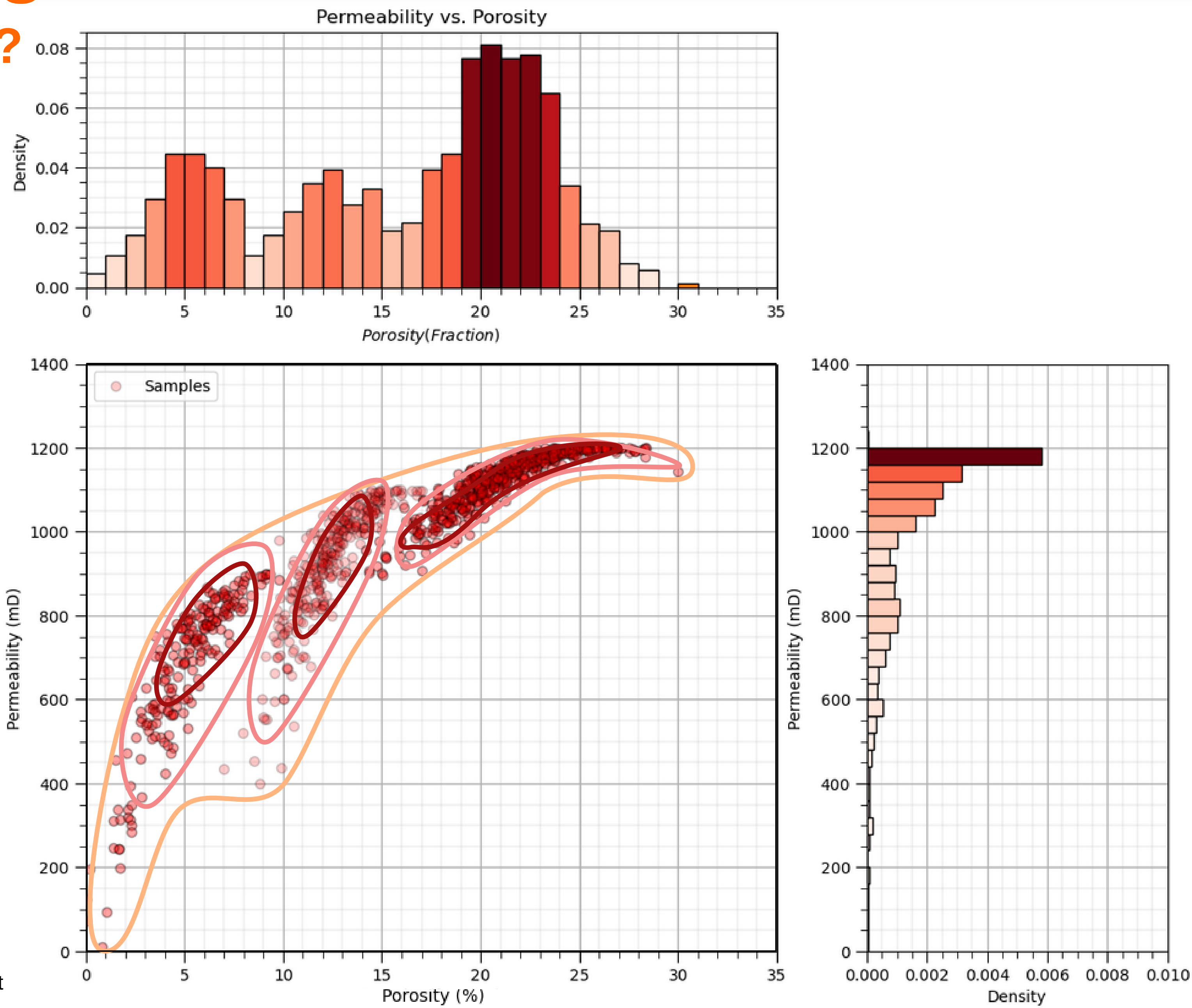
What is Bivariate Analysis?

Bivariate Analysis: Understand and quantify the relationship between two variables

- Example: Porosity and permeability data
- How can we use this relationship? What would we miss if we only looked at the 2 histograms?
- Is there a realizationship between permeability and porosity?

Now Look at The Data!

Positive Relationship and…



Porosity and permeability PDFs univariate and bivariate (schematic) plot modified from PythonDataBasics_Bivariate _Visualization.ipynb.

Bivariate Statistics

What is Bivariate Analysis?

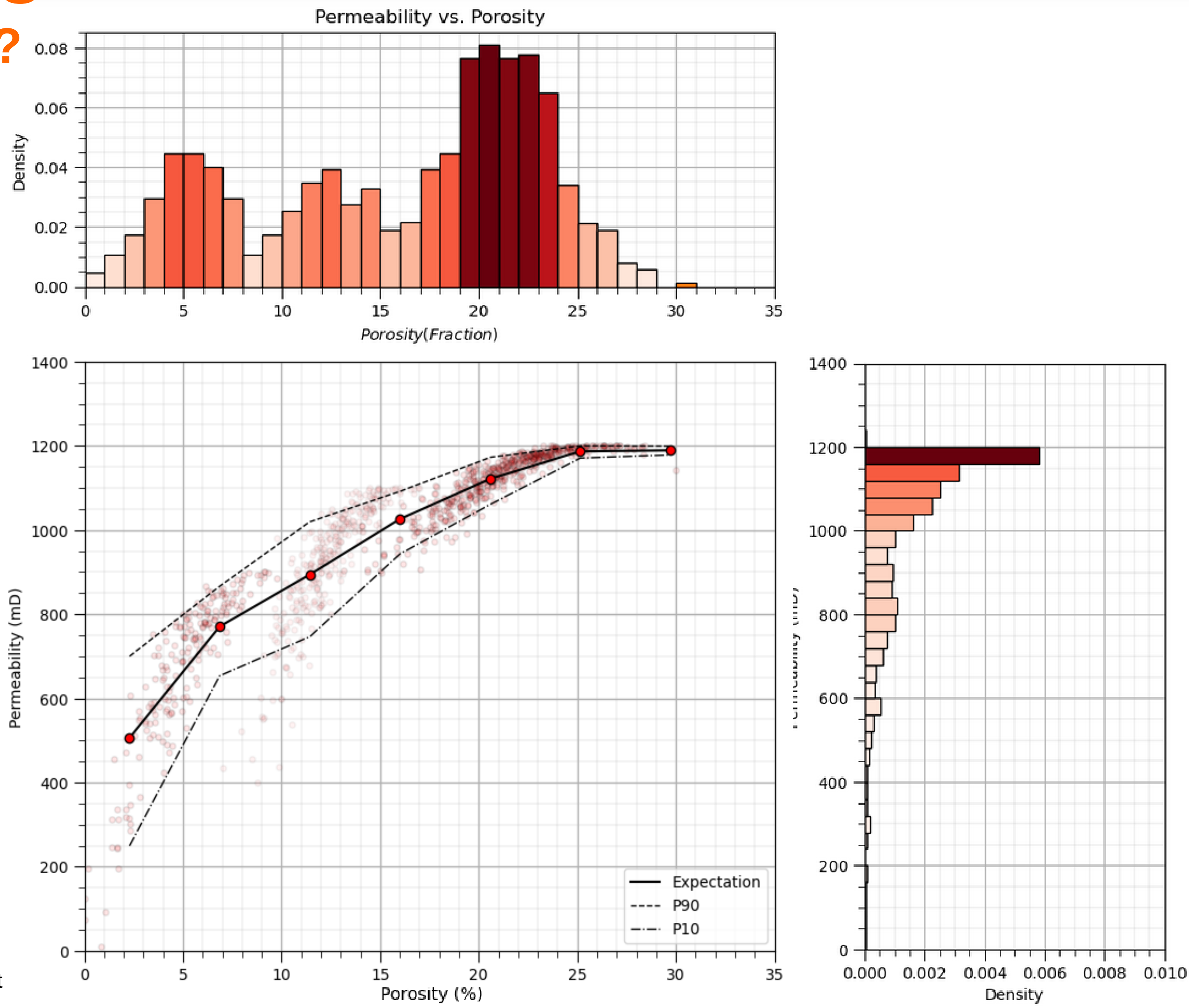
Bivariate Analysis: Understand and quantify the relationship between two variables

- Example: Porosity and permeability data
- How can we use this relationship? What would we miss if we only looked at the 2 histograms?
- Is there a realizationship between permeability and porosity?

Now Look at The Data!

Positive Relationship and

We Could Model It with Conditional Distributions!



Porosity and permeability PDFs univariate and bivariate (schematic) plot modified from PythonDataBasics_Bivariate _Visualization.ipynb.

Bivariate Statistics

What is Bivariate Analysis?

Bivariate Analysis: Understand and quantify the relationship between two variables

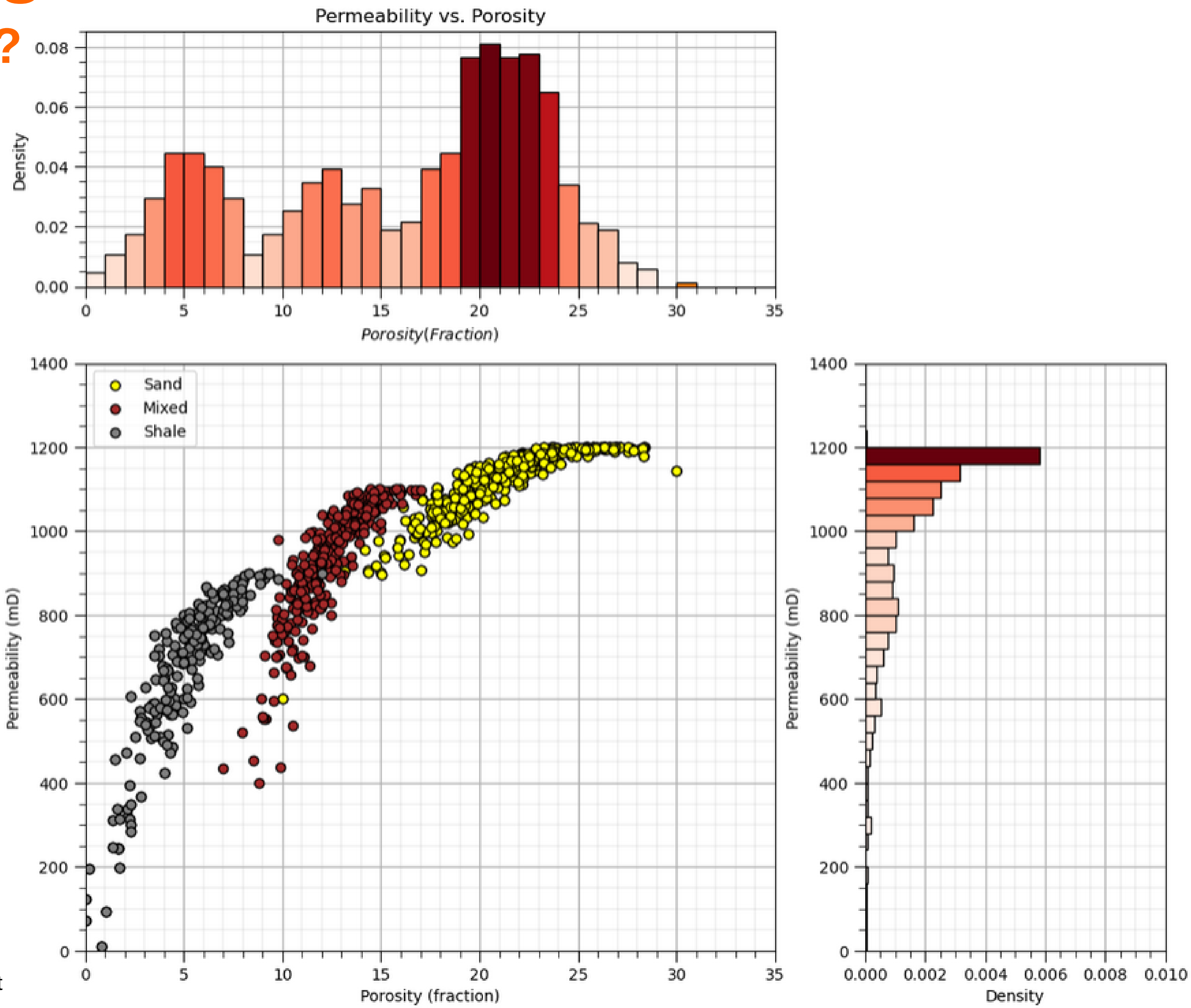
- Example: Porosity and permeability data
- How can we use this relationship? What would we miss if we only looked at the 2 histograms?
- Is there a realizationship between permeability and porosity?

Now Look at The Data!

Positive Relationship and

Multiple Facies / Populations

We use all of this to build better subsurface models.



Porosity and permeability PDFs univariate and bivariate (schematic) plot modified from PythonDataBasics_Bivariate _Visualization.ipynb.

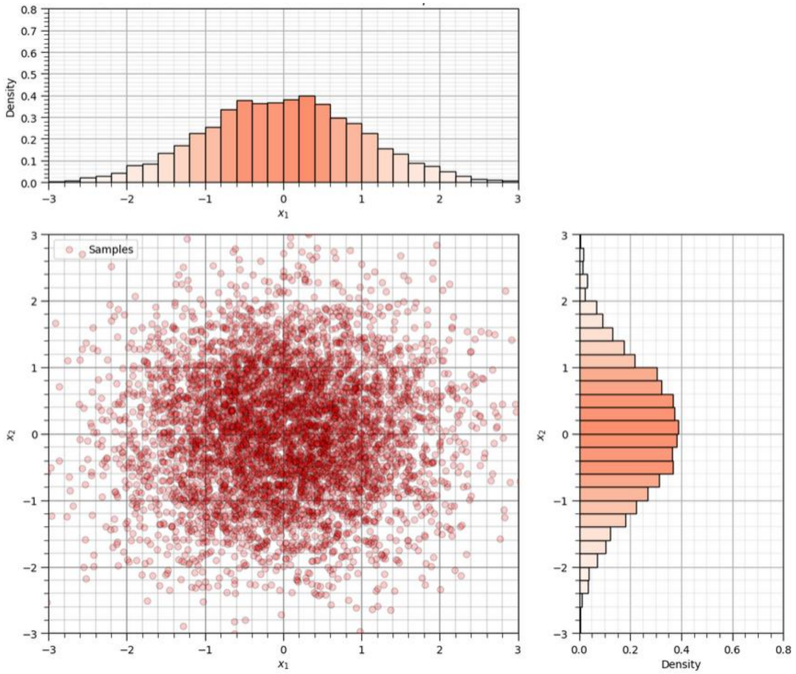
Bivariate Statistics

How to Quantify Relationship between two Variables?

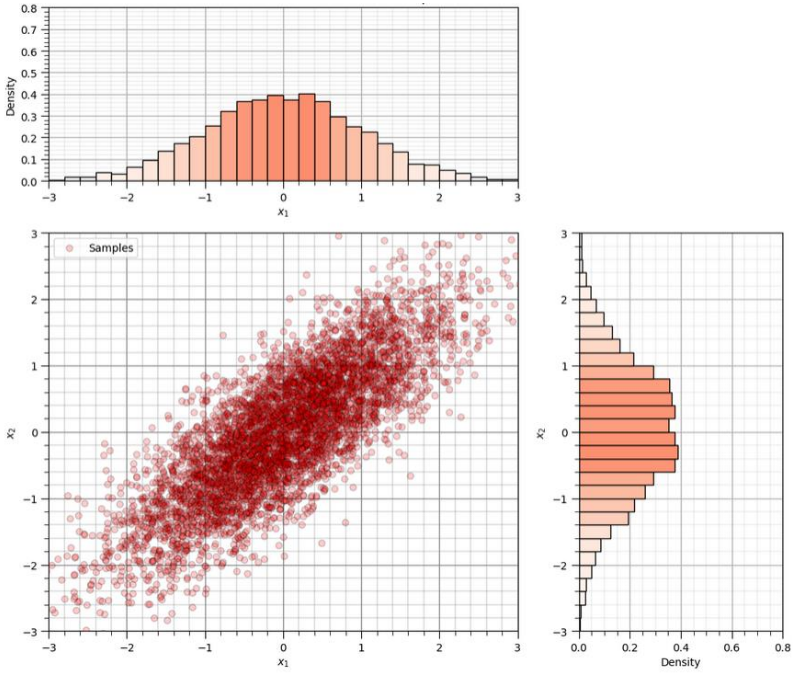
Bivariate Analysis: Quantifying the strength of the relationship between 2 features

- Example: porosity and permeability data from a carbonate and a sandstone formation.
- In which case are the features X1 and X2 best correlated? I.e., a stronger relationship?

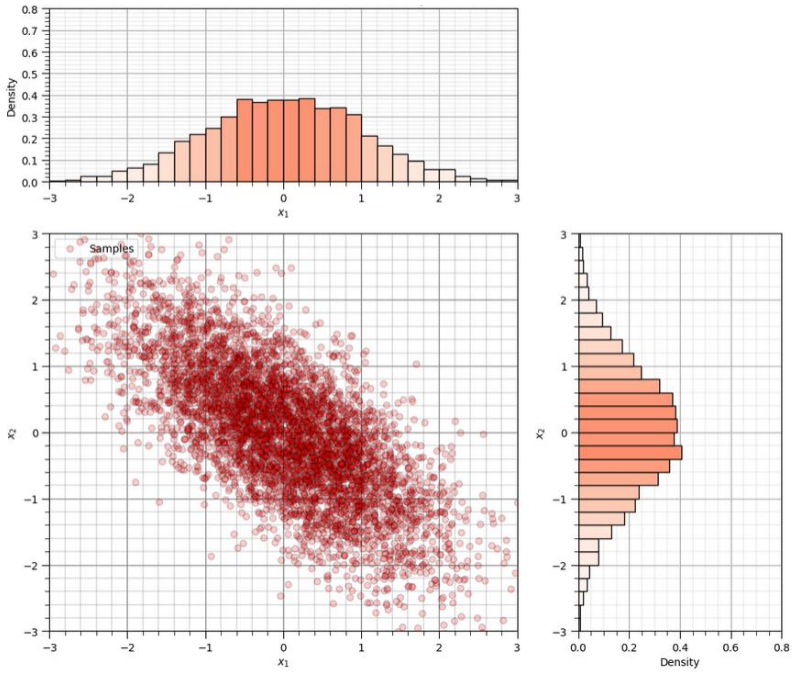
Data Set 1



Data Set 2



Data Set 3



Univariate and bivariate distributions for 3 data sets, generated with Interative_Correlation_Coefficient.ipynb.

PGE 338 Data Analytics and Geostatistics

Lecture 8: Bivariate Distributions

Lecture outline . . .

- Correlation

Introduction

General Concepts

Univariate

Bivariate

Correlation

Regression

Model Checking

Time Series Analysis

Spatial Analysis

Machine Learning

Uncertainty Analysis

Michael Pyrcz, The University of Texas at Austin

Bivariate Statistics

Pearson’s Correlation Coefficient

How Do We Quantify the Relationships Between 2 Features, Bivariate Relationships?

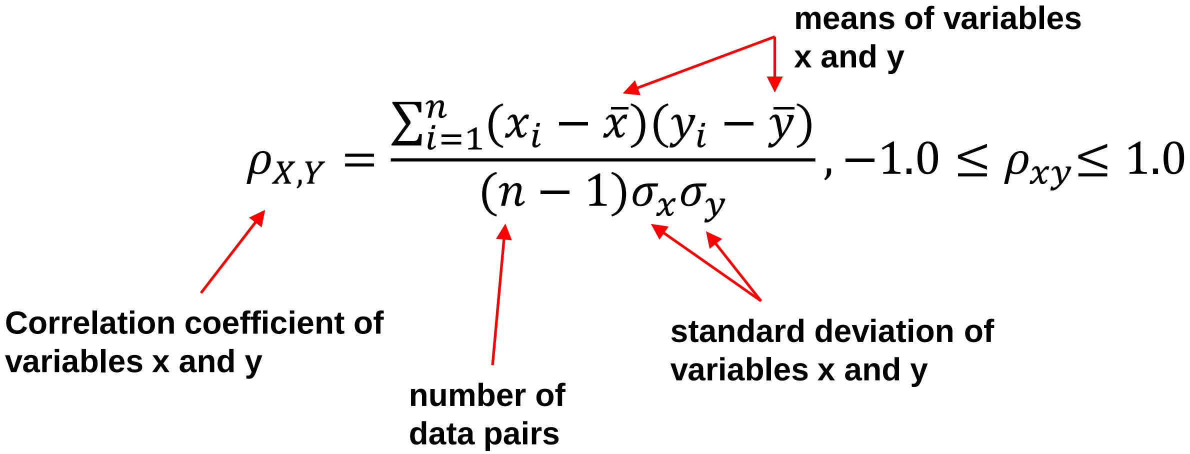
- We need to go beyond qualitative descriptions, good, bad, strong, weak, none…

Bivariate Statistics

Pearson’s Correlation Coefficient

Definition: Pearson’s Product-moment Correlation Coefficient ($\rho_{X,Y}$)

- Provides a measure of the degree of linear relationship.



$$\rho_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\sigma_x\sigma_y}, -1.0 \le \rho_{xy} \le 1.0$$

means of variables x and y

Correlation coefficient of variables x and y

number of data pairs

standard deviation of variables x and y

- Correlation coefficient is a “standardized” covariance. The covariance ($\boldsymbol{C_{X,Y}}$):

Covariance: $$C_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$$

Correlation Coefficient: $$\rho_{X,Y} = \frac{C_{X,Y}}{\sigma_X \sigma_Y}$$

Correlation coefficient is covariance standardized by dividing by $\sigma_X \sigma_Y$

Bivariate Statistics

Variance and Covariance

We can see that covariance and variance are related.

1. Sample Variance:

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

A measure of how 1 variable varies with itself.

2. Sample Covariance:

- Replace the second term in the square with another feature.

$$\boldsymbol{C_{xy}} = \frac{1}{n-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

A measure of how 2 features vary together.

Bivariate Statistics

Spearman's Rank Correlation Coefficient

Definition: Spearman's Rank Correlation Coefficient

- Provides a measure of the degree of monotonic relationship.

means of rank transform of variables x and y

$$\rho_{R_x,R_y} = \frac{\sum_{i=1}^{n}(R_{x_i} - \overline{R_x})(R_{y_i} - \overline{R_y})}{(n-1)\sigma_{R_x}\sigma_{R_y}}, -1.0 \leq \rho_{xy} \leq 1.0$$

Rank correlation coefficient of variables x and y

number of data pairs

standard deviation of Rank transform of variables x and y

- Rank transform, e.g. R_{x_i}, sort the data in ascending order and replace the data with the index, $i = 1, \ldots, n$.
- Spearman's rank correlation coefficient is more robust in the presence of outliers and some nonlinear features than the Pearson's correlation coefficient

Bivariate Statistics

Covariance

Let's think about covariance.

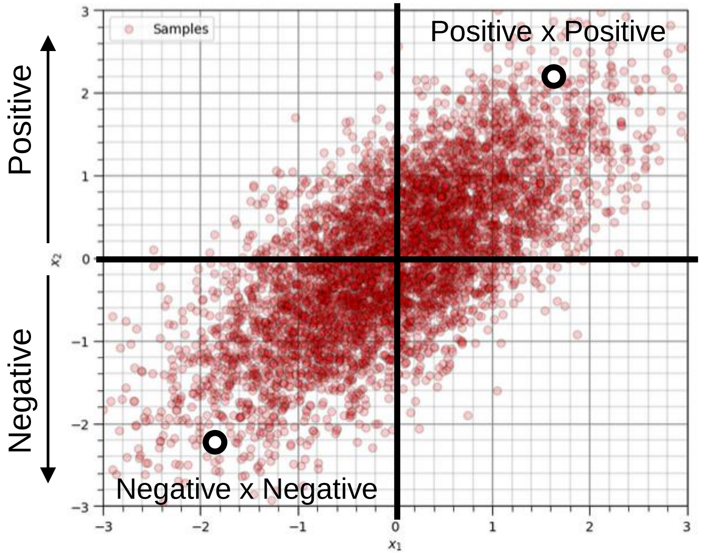
- For a thought experiment, consider 2 standard normal variables, N[0,1].

$$C_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$$

the means, $\bar{x} = \bar{y} = 0\ \therefore$

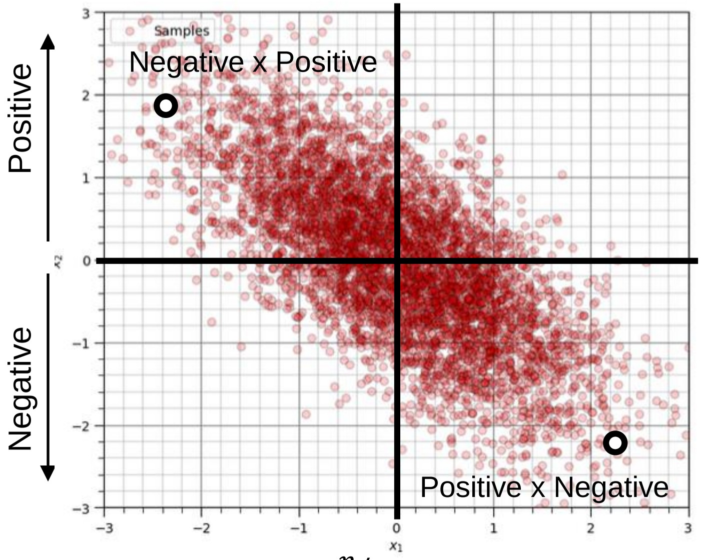
$$C_{xy} \sim E\{X\ Y\} \text{ and } \rho_{xy} \sim \frac{E\{X\ Y\}}{\sigma_x \sigma_y}$$

- Positive covariance / correlation if we pair high-high and low-low feature values over the samples.
- Negative covariance / correlation if we pair high-low and low-high feature values over the samples.



$$\rho_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y} > 0.0$$

$$if\ \rho > 0,\ \uparrow C_{xy},\ \sum_{i=1}^{n/2}[\,x^- \times\ y^-] + \sum_{i=n/2}^{n}[\,x^+ \times\ y^+]$$



$$\rho_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y} < 0.0$$

$$if\ \rho < 0,\ \uparrow C_{xy},\ \sum_{i=1}^{n/2}[(\,x^- \times\ y^+] + \sum_{i=n/2}^{n}[(\,x^+ \times\ y^-]$$

Interactive Correlation Demonstration

What does correlation coefficient actually see?

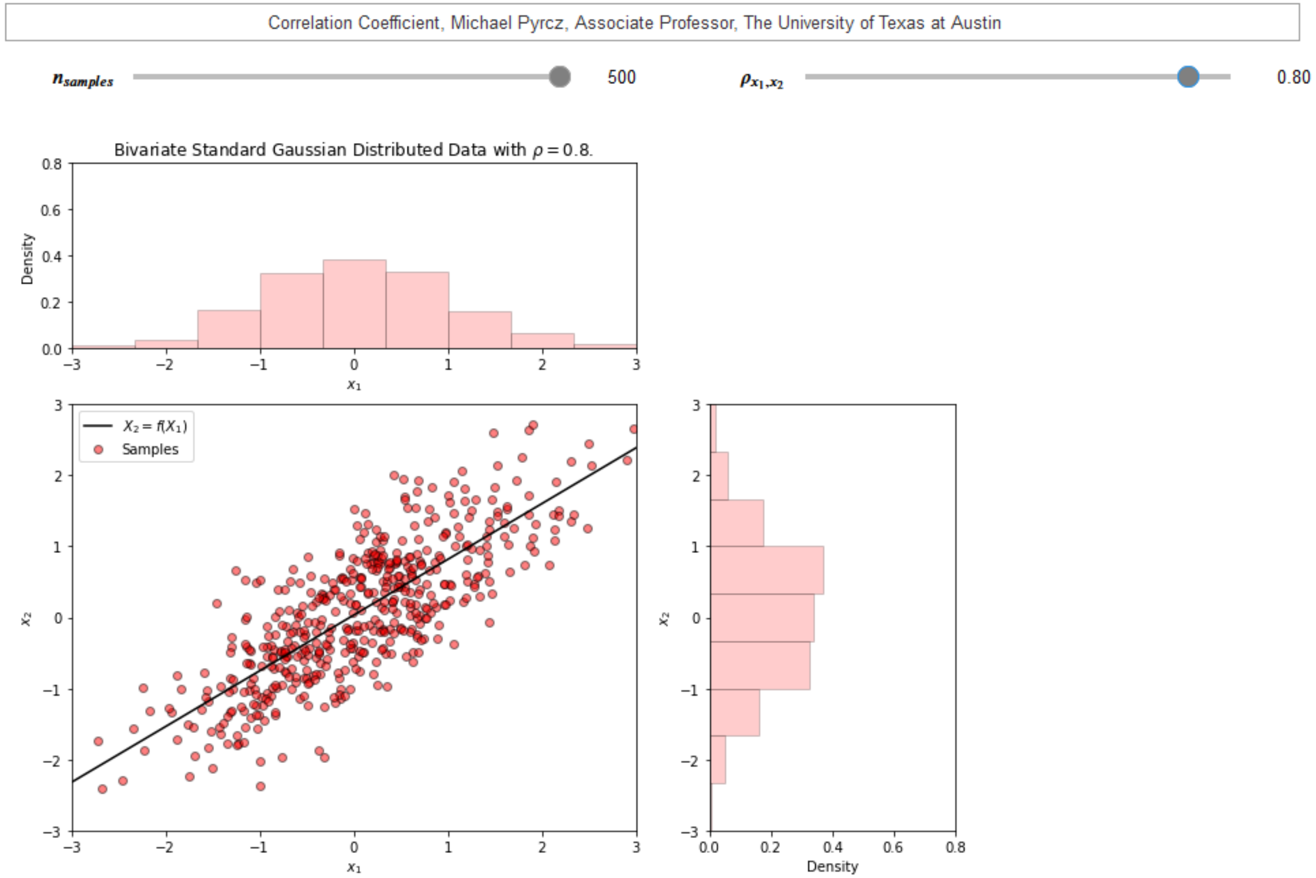
- The correlation coefficient measure linear, homoscedastic relationships between 2 features.

- For the bivariate Gaussian distribution the correlation coefficient completely describes the relationship between features.

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{x,y}^2}}\exp\left[-\frac{z}{2(1-\rho_{x,y}^2)}\right]$$

where:

$$z = \frac{(x-\mu_x)^2}{\sigma_x^2} - \frac{2\rho_{x,y}(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}$$



Correlation Coefficient, Michael Pyrcz, Associate Professor, The University of Texas at Austin

$n_{samples}$ 500

ρ_{x_1,x_2} 0.80

Interactive Correlation demonstration, the file is Interactive_Correlation_Coefficient.ipynb.

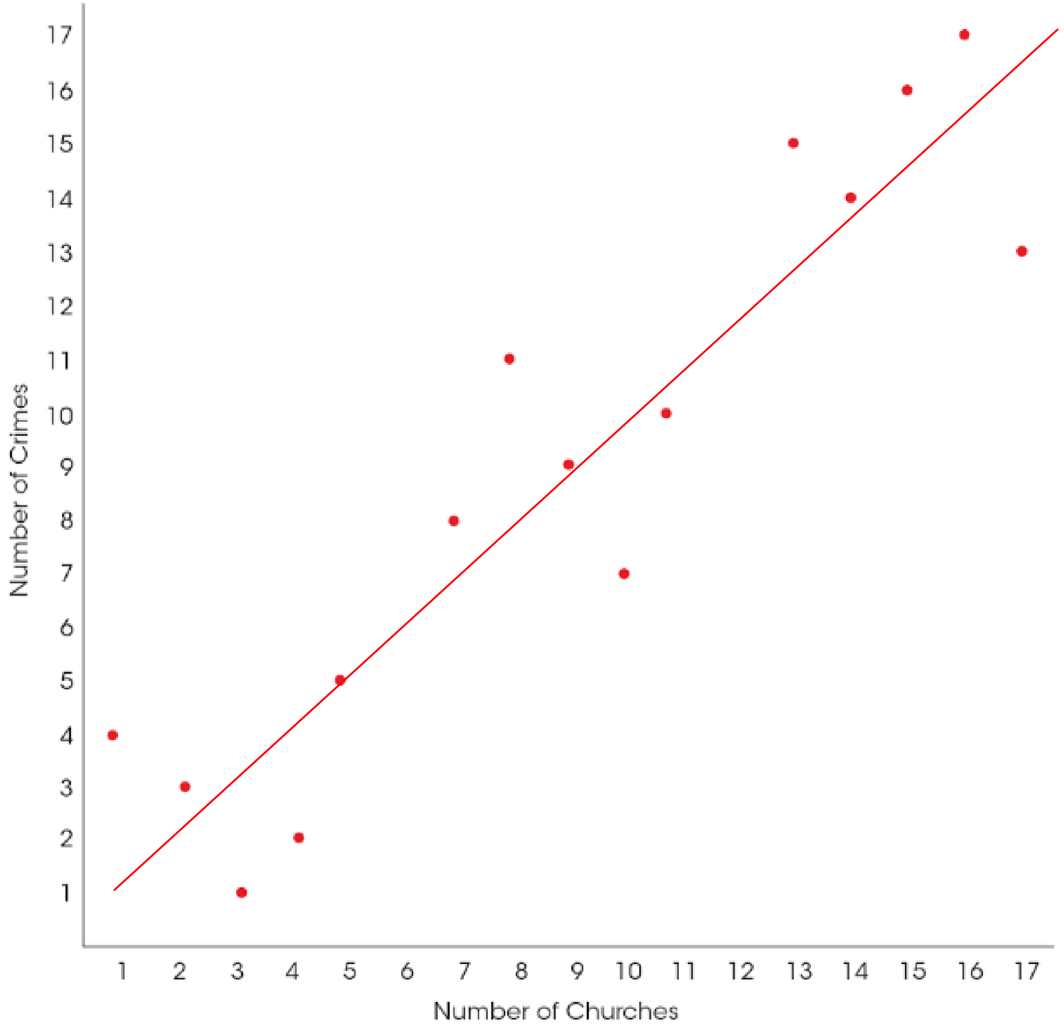
Bivariate Statistics
Correlation and Causation

Does correlation imply causation?

- NO!

- We require a “true experiment” where one variable is manipulated, and others are rigorously controlled!

Here’s an example to demonstrate a potential issue of assigning causation based on the observation of correlation.



Plot of frequency of crimes vs. frequency of churches.

Bivariate Statistics

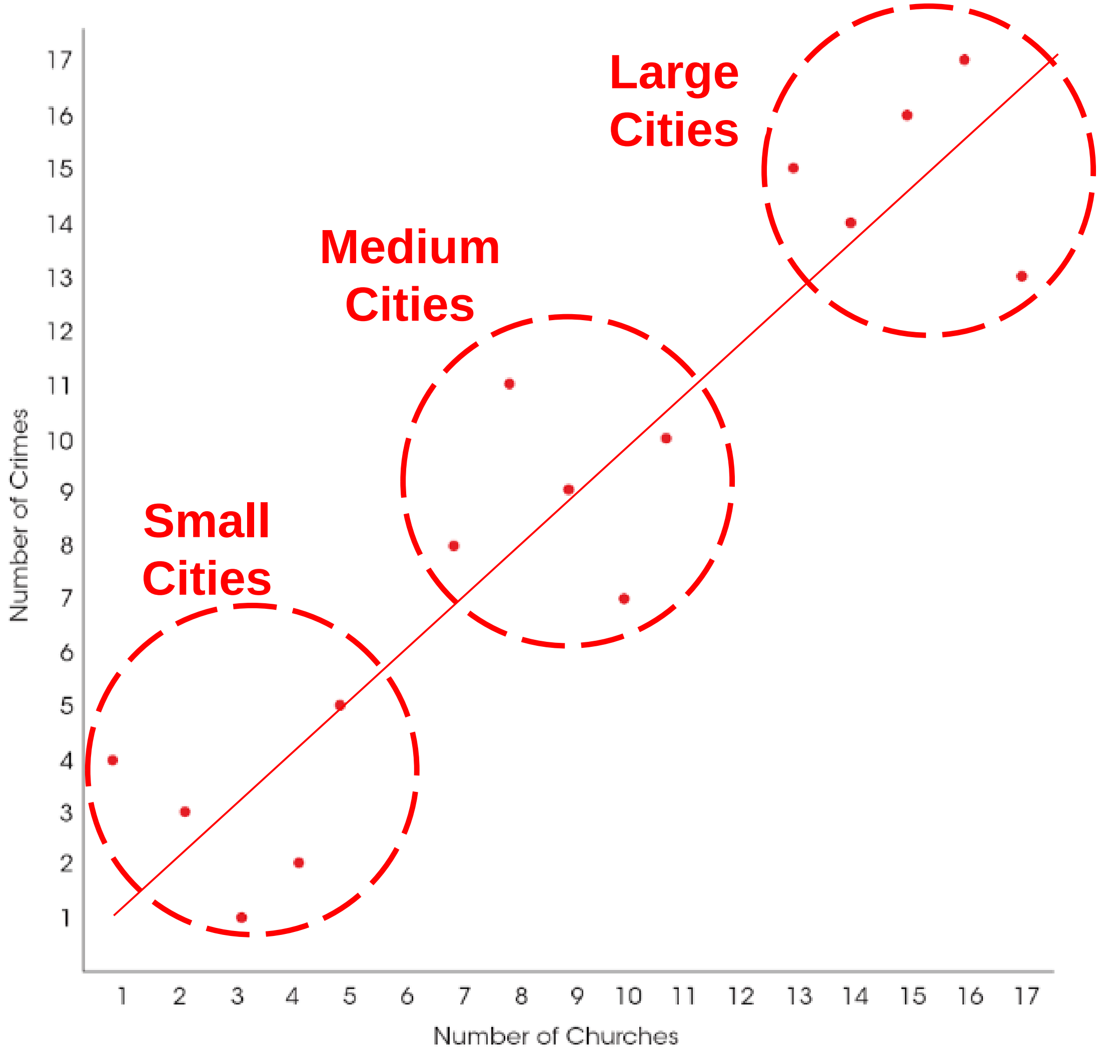
Correlation and Causation

Does correlation imply causation?

- NO!

- We require a “true experiment” where one variable is manipulated, and others are rigorously controlled!

Here’s an example to demonstrate a potential issue of assigning causation based on the observation of correlation.



Plot of frequency of crimes vs. frequency of churches.

Bivariate Statistics
Spurous Correlations

Can correlation be misleading and misused?

Spurious Correlation:

A correlation that seems to be causal but is not.

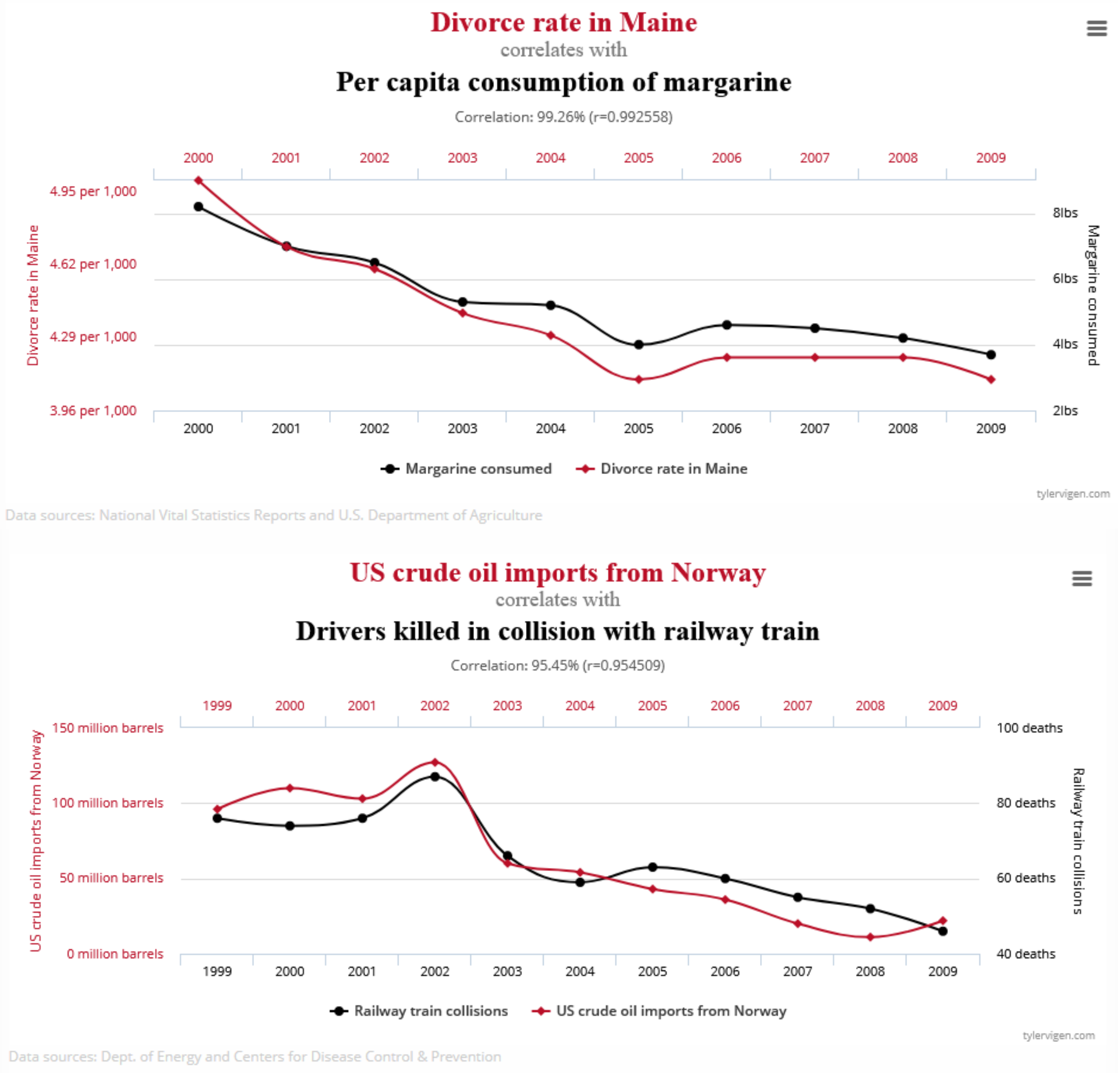
Due to:

- random chance, too few data, data sampling bias
- outliers or artificial truncation of the data
- confounding features

Confounding Features:

feature not part of the study that is related to both the predictor and response feature

- causes spurious correlations



Examples of spurious correlations.

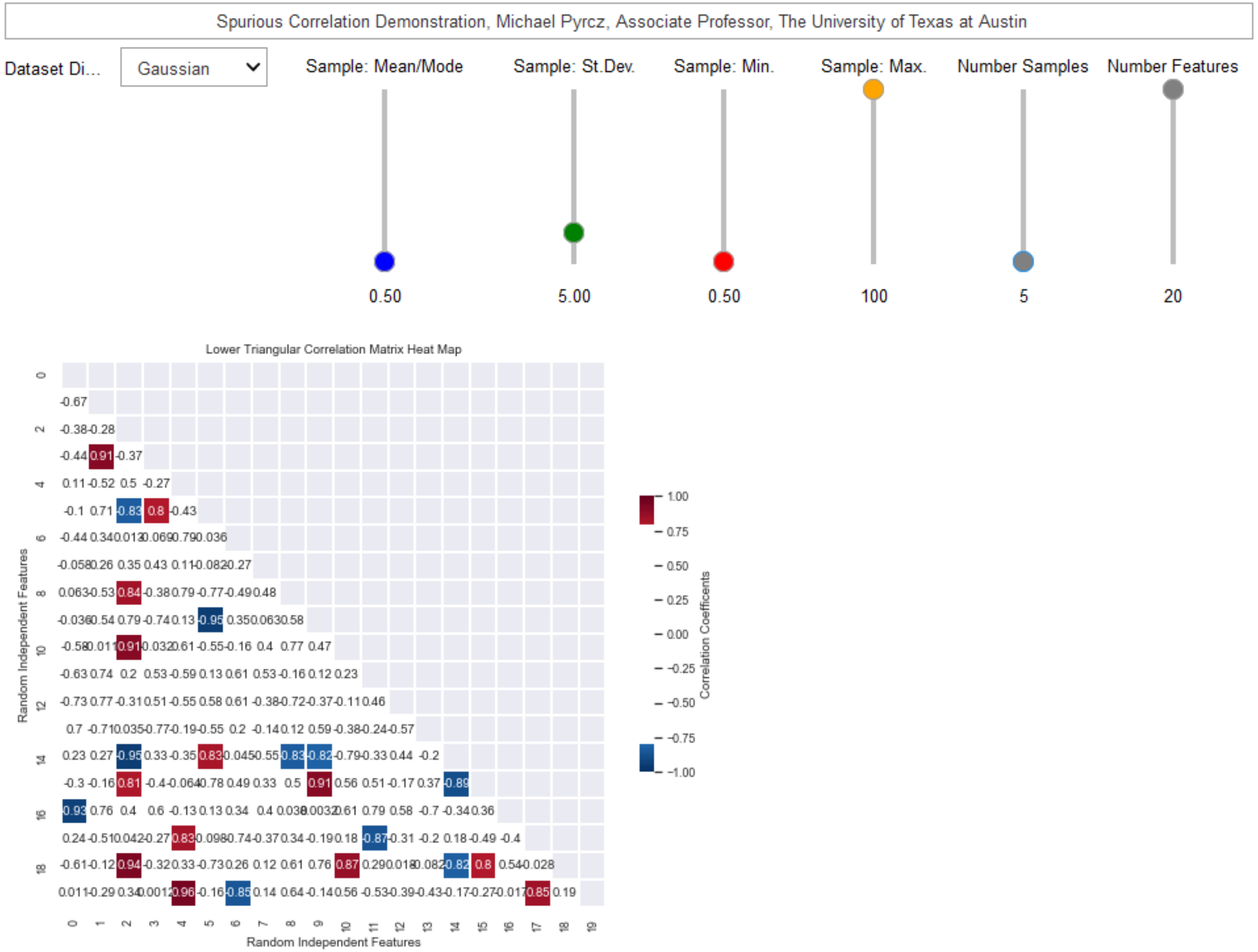
Images from https://www.tylervigen.com/spurious-correlations

Interactive Spurious Correlations

Let’s make uncorrelated random features and check their correlations.

- We can plot all feature pairwise correlations in a matrix color coded by very high or low correlations (>0.8 or <-0.8 respectively)

- What happens has the number of data is small and the number of features is large?



Spurious Correlation Demonstration, Michael Pyrcz, Associate Professor, The University of Texas at Austin

Dataset Di... Gaussian

Sample: Mean/Mode 0.50

Sample: St.Dev. 5.00

Sample: Min. 0.50

Sample: Max. 100

Number Samples 5

Number Features 20

Lower Triangular Correlation Matrix Heat Map
Random Independent Features
Random Independent Features
Correlation Coefficents
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.00

Interactive spurious correlations demonstration, the file is Interactive_Spurious_Correlations.ipynb.

Bivariate Statistics

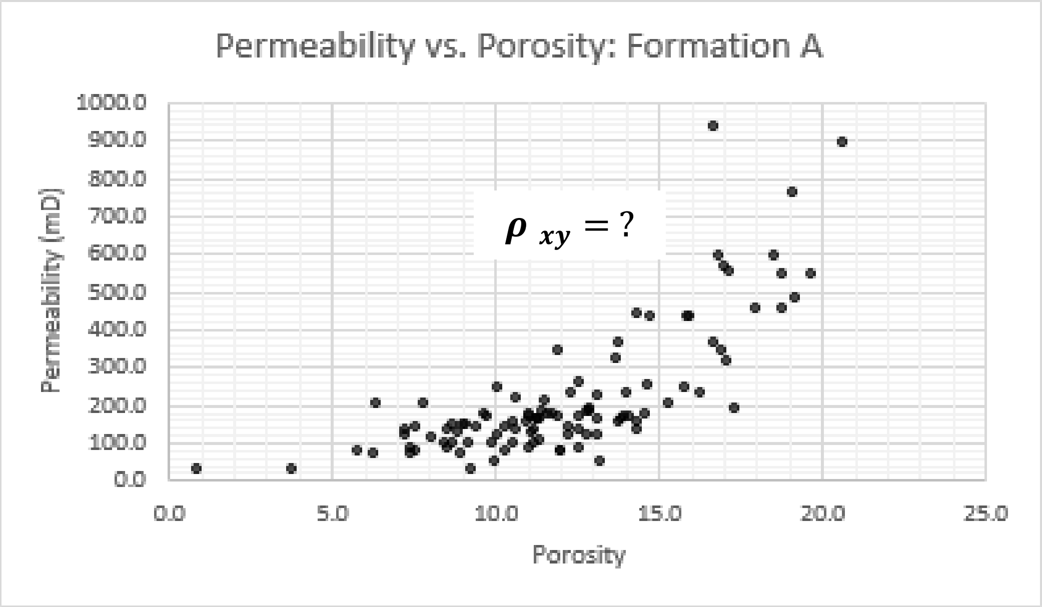
Exercise with Pearson's Correlation Coefficient

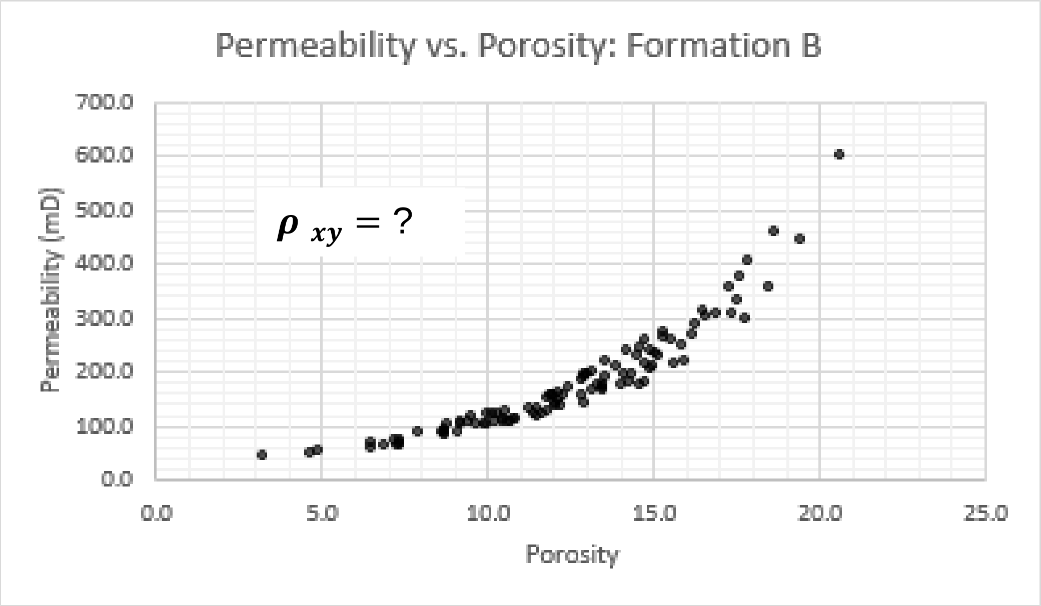
Prepare the scatterplot of the core porosity and permeability data for formations A and B provided to you. Estimate correlation coefficient in both cases. How well is porosity and permeability are correlated in these two formations?

Excel Function Correl(Array1,Array2)

Can I derive a correlation between porosity and permeability?

How reliable that correlation would be?





Correlation for 2 datasets, file is Univariate_NonLinearPor_Perm.xlsx.

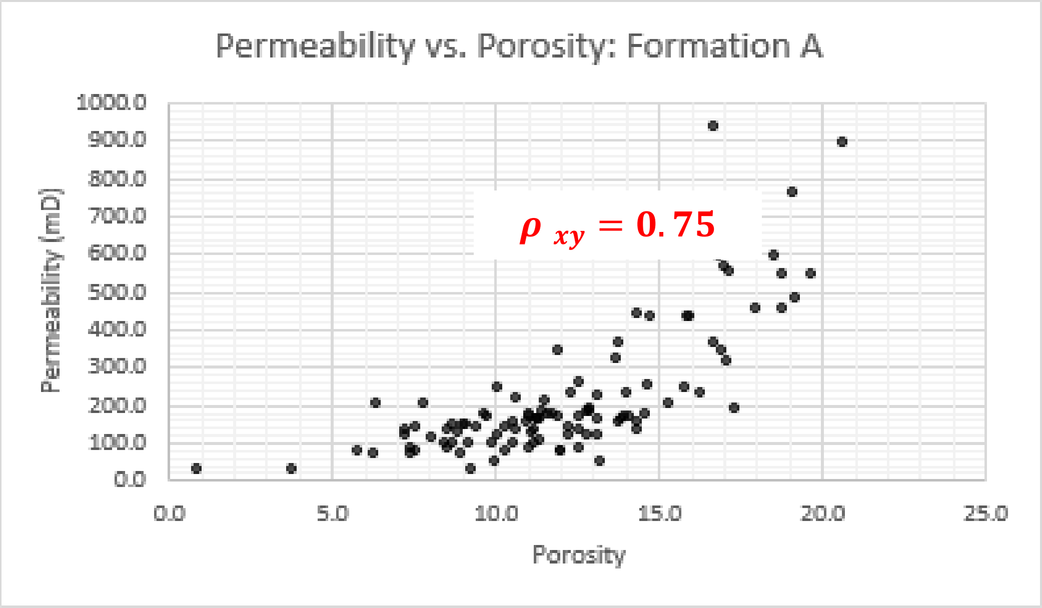
Bivariate Statistics

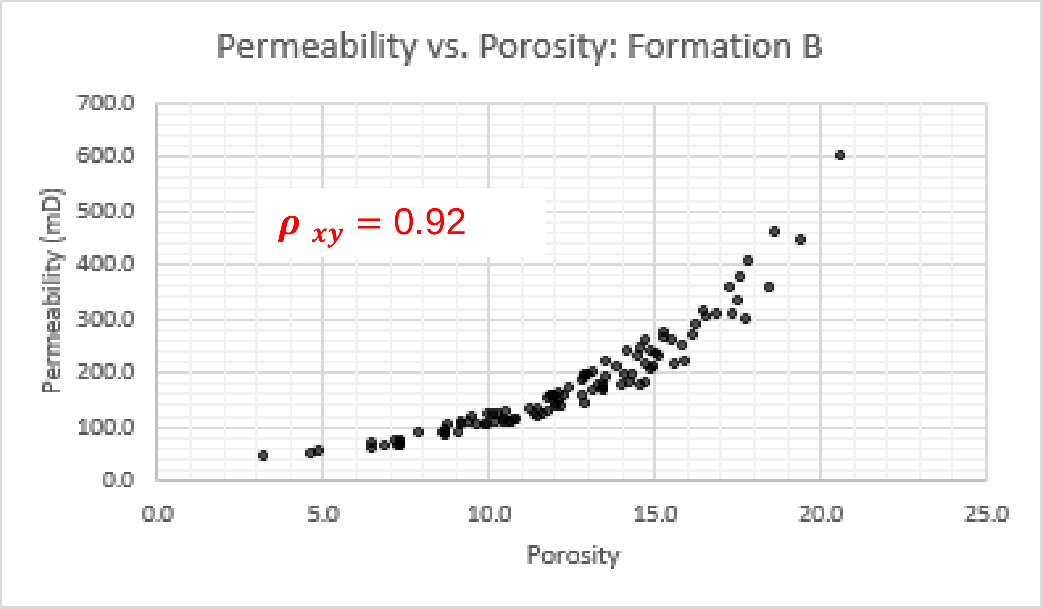
Exercise with Pearson's Correlation Coefficient

Prepare the scatterplot of the core porosity and permeability data for formations A and B provided to you. Estimate correlation coefficient in both cases. How well is porosity and permeability are correlated in these two formations?

Excel Function Correl(Array1,Array2)

Can I derive a correlation between porosity and permeability?
How reliable that correlation would be?



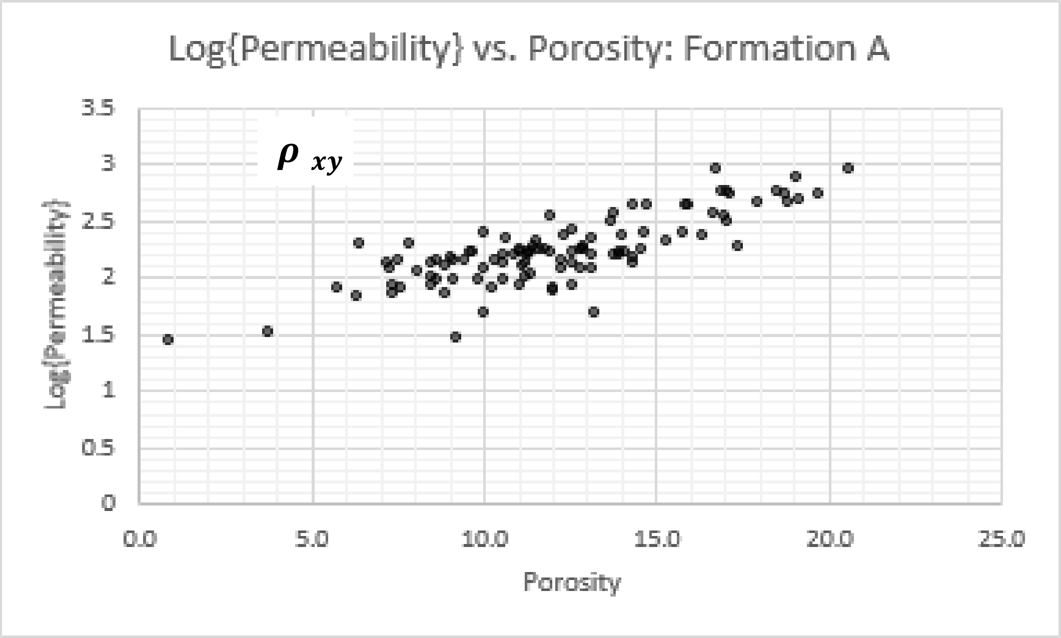


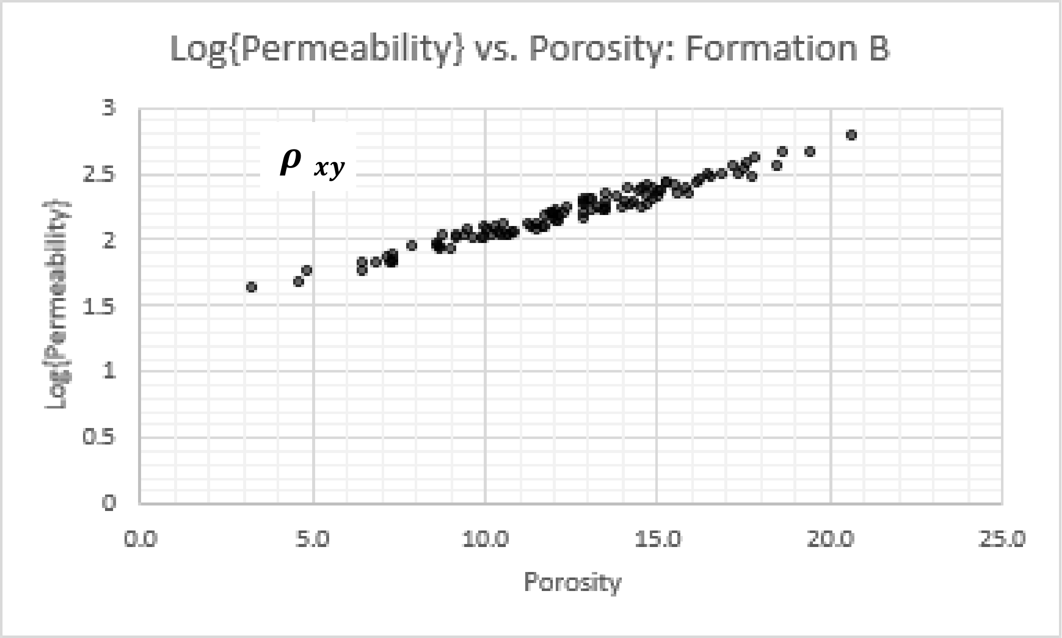
Correlation for 2 datasets, file is Univariate_NonLinearPor_Perm.xlsx.

Bivariate Statistics

Exercise with Pearson's Correlation Coefficient

Does this nonlinear log transform improve our use / characterization with the correlation coefficient?



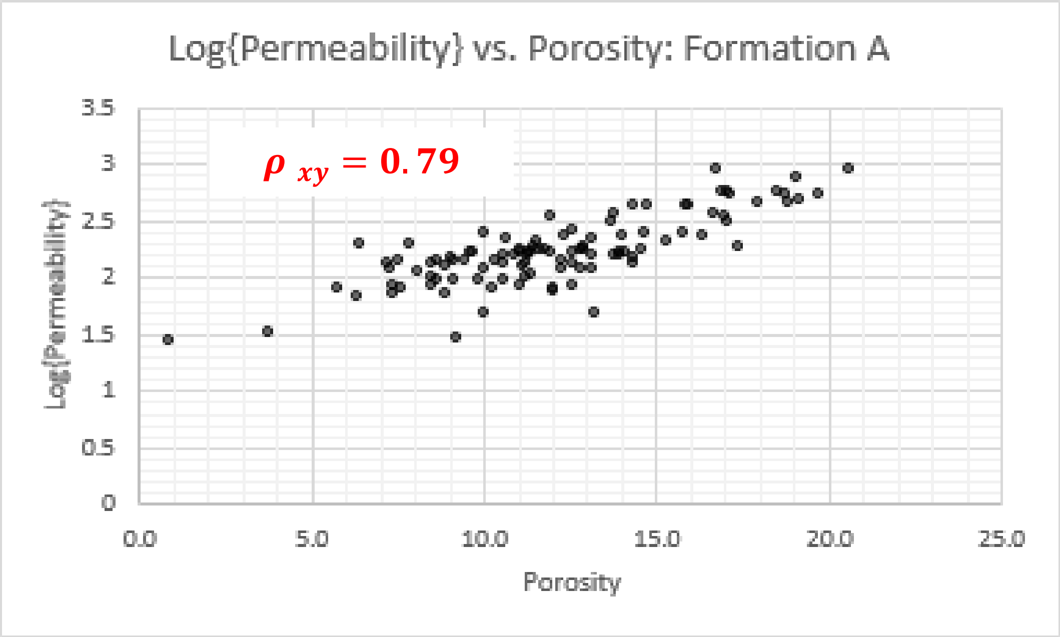


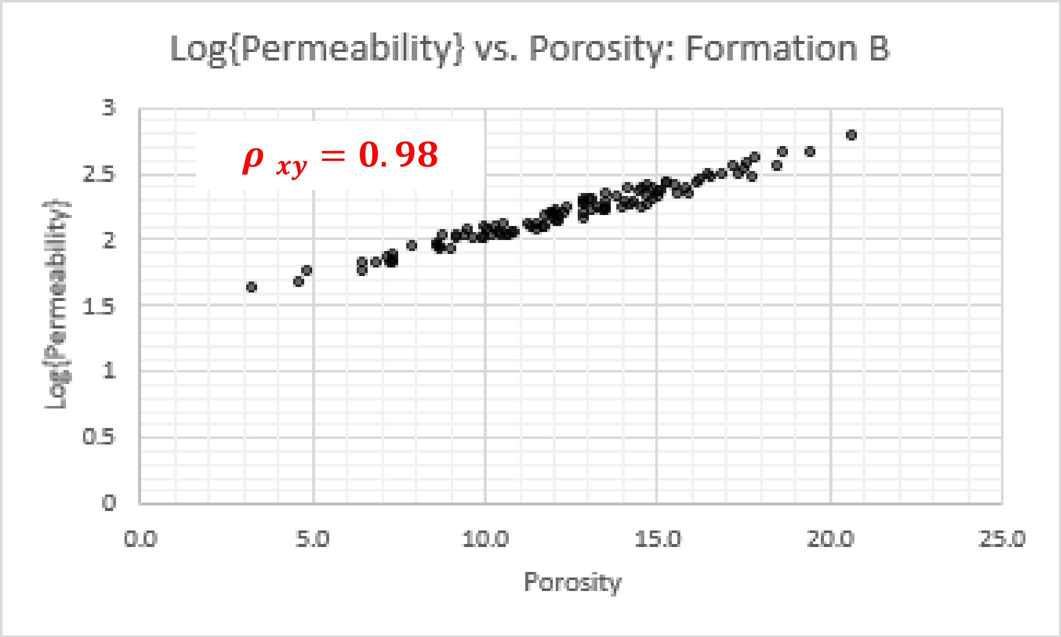
Correlation for 2 datasets, file is Univariate_NonLinearPor_Perm.xlsx.

Bivariate Statistics

Exercise with Pearson's Correlation Coefficient

Does this nonlinear log transform improve our use / characterization with the correlation coefficient?





Correlation for 2 datasets, file is Univariate_NonLinearPor_Perm.xlsx.

Bivariate Statistics

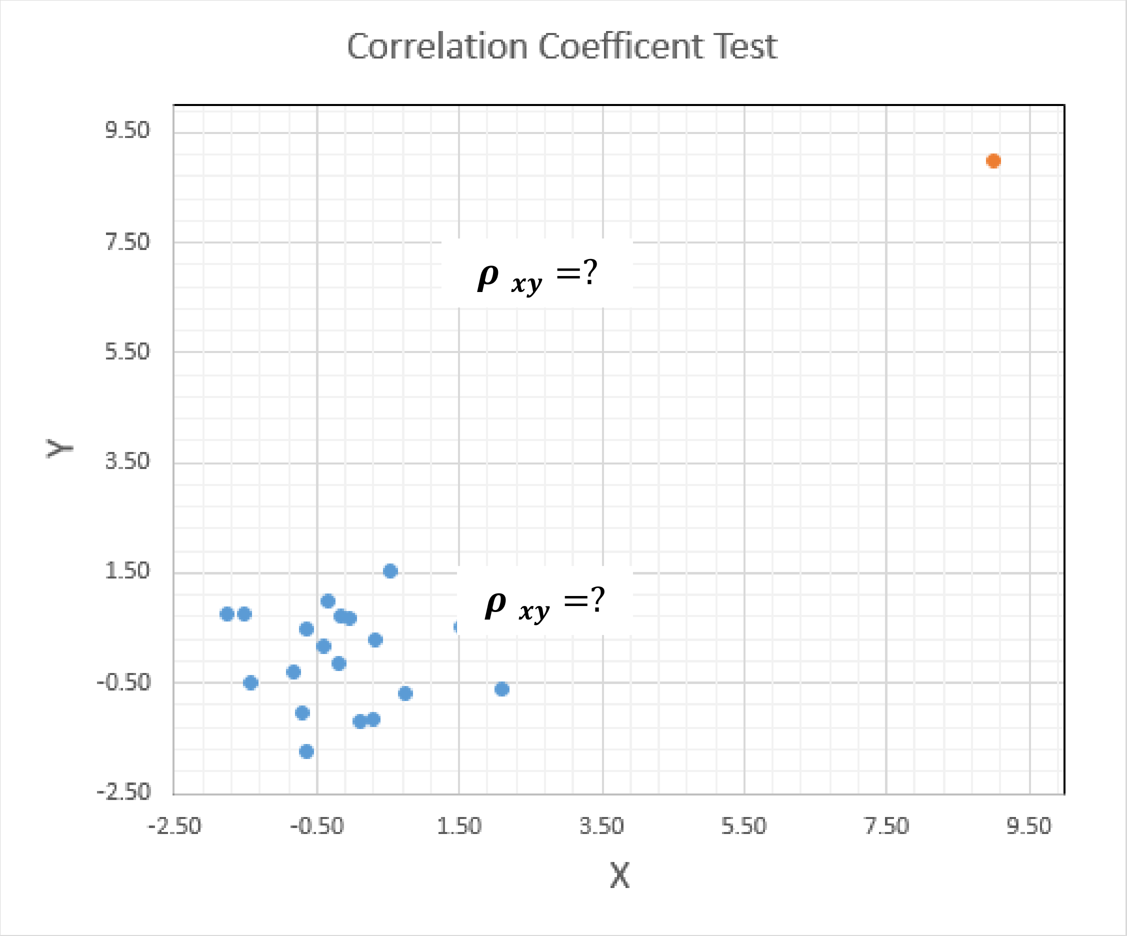
Exercise with Pearson’s Correlation Coefficient

Step 1: Generate a random data set of 19 x and y variables and estimate their correlation coefficient (Hint: Rand() in Excel with N[0,1]).

Excel Function NORM.INV(RAND(),0,1)

Step 2: Now add any desired outlier to the data and estimate the correlation coefficient (see example).

How does this outlier affect the correlation coefficient?



Random data with one bivariate outlier.

Bivariate Statistics

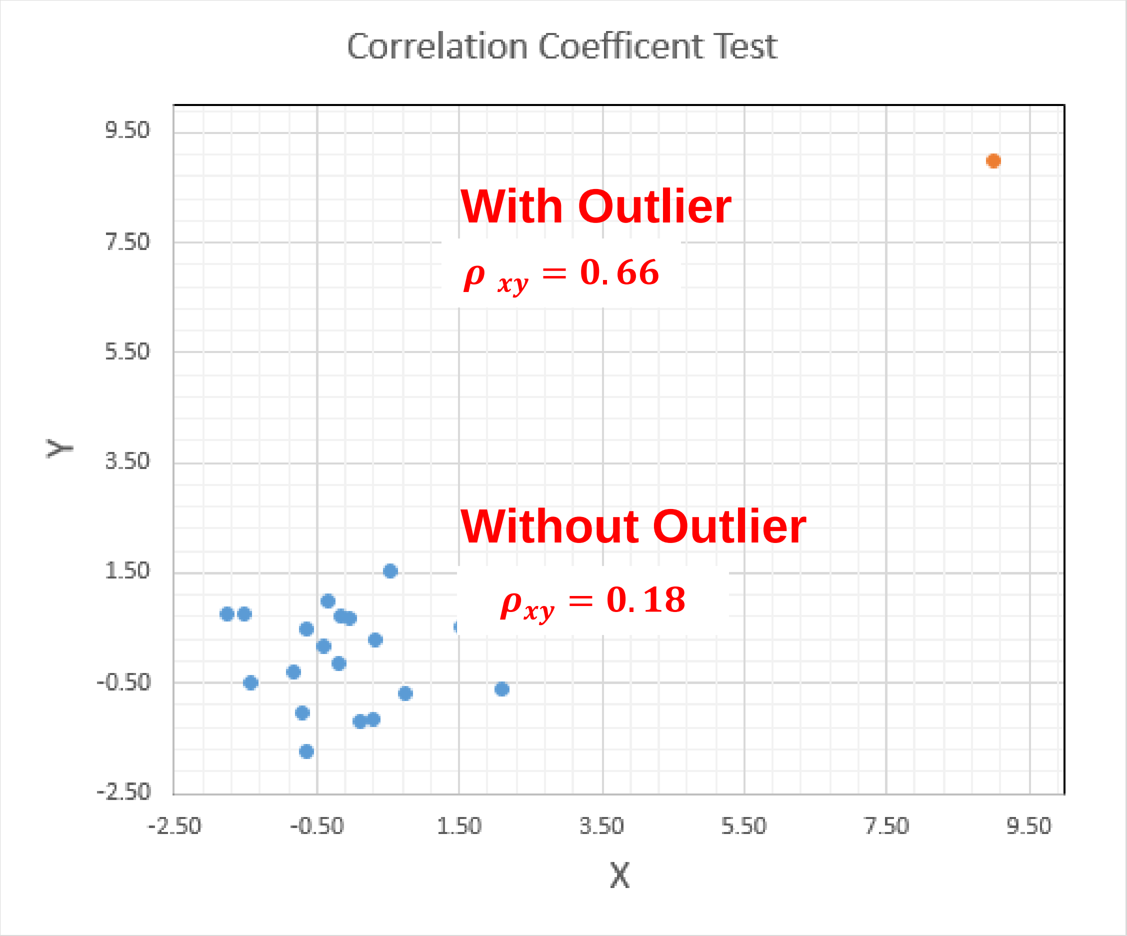
Exercise with Pearson's Correlation Coefficient

Step 1: Generate a random data set of 19 x and y variables and estimate their correlation coefficient (Hint: Rand() in Excel with N[0,1]).

Excel Function NORM.INV(RAND(),0,1)

Step 2: Now add any desired outlier to the data and estimate the correlation coefficient (see example).

How does this outlier affect the correlation coefficient?



Random data with one bivariate outlier.

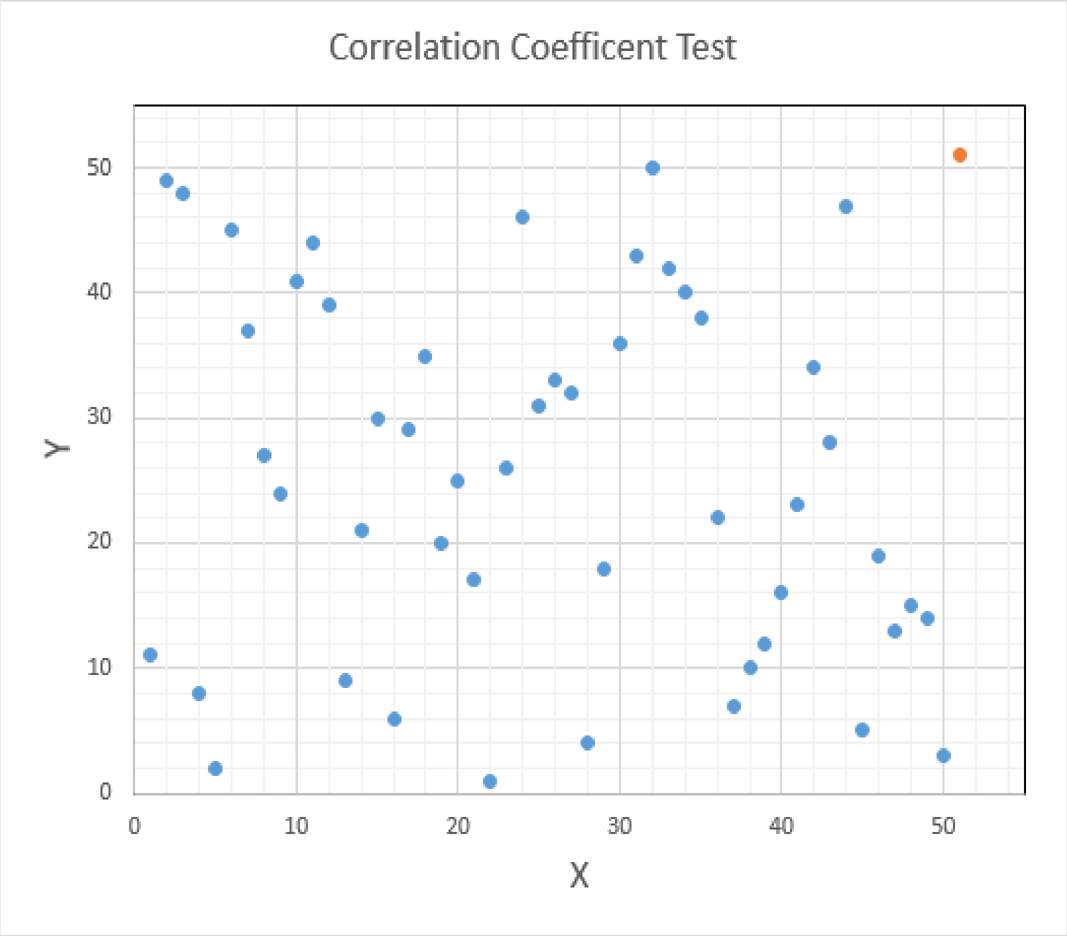
Bivariate Statistics

Exercise with Pearson’s Correlation Coefficient

Step 3: Apply the rank transform to the dataset (Hint: 21-Rank.Avg() in Excel).

How does this outlier now affect the correlation coefficient?

This is a more robust form of the correlation coefficient called the rank correlation coefficient.



Random data rank transformed with one bivariate outlier.

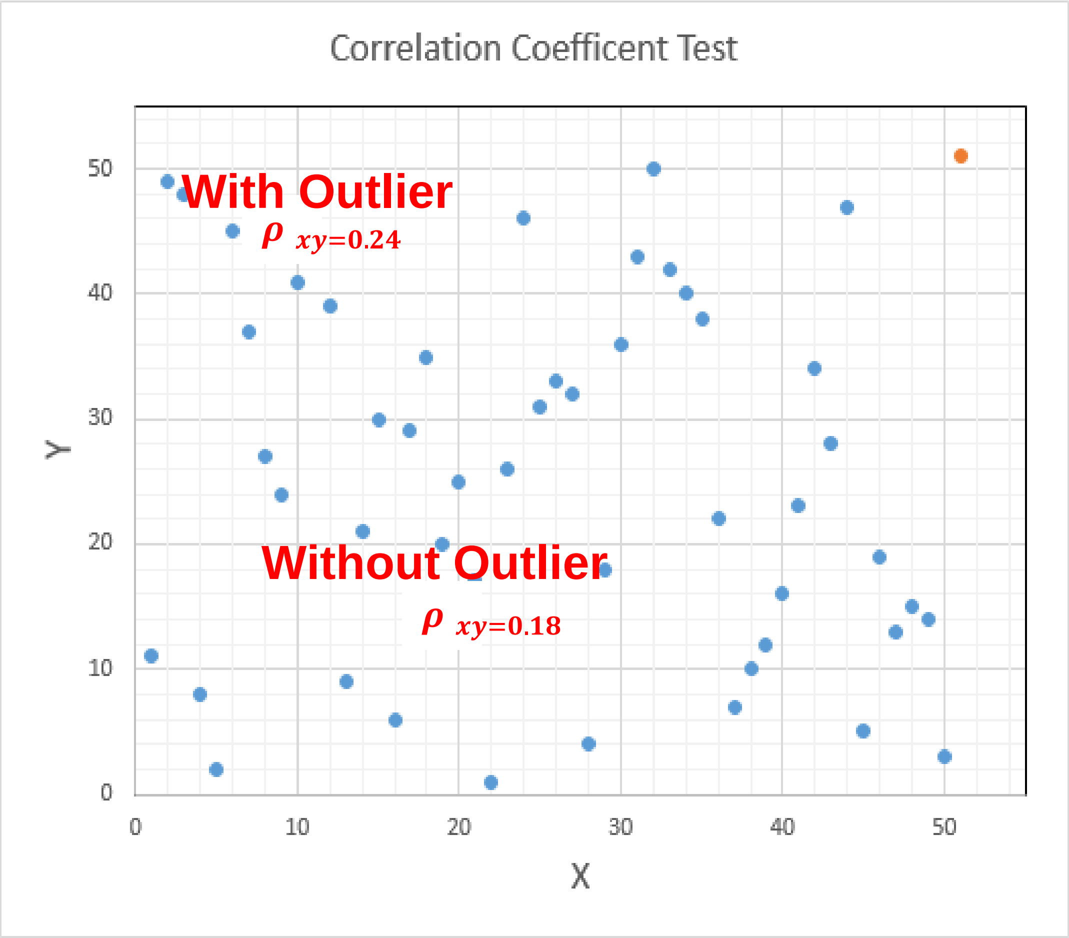
Bivariate Statistics

Exercise with Pearson’s Correlation Coefficient

Step 3: Apply the rank transform to the dataset (Hint: 21-Rank.Avg() in Excel).

How does this outlier now affect the correlation coefficient?

This is a more robust form of the correlation coefficient called the rank correlation coefficient.



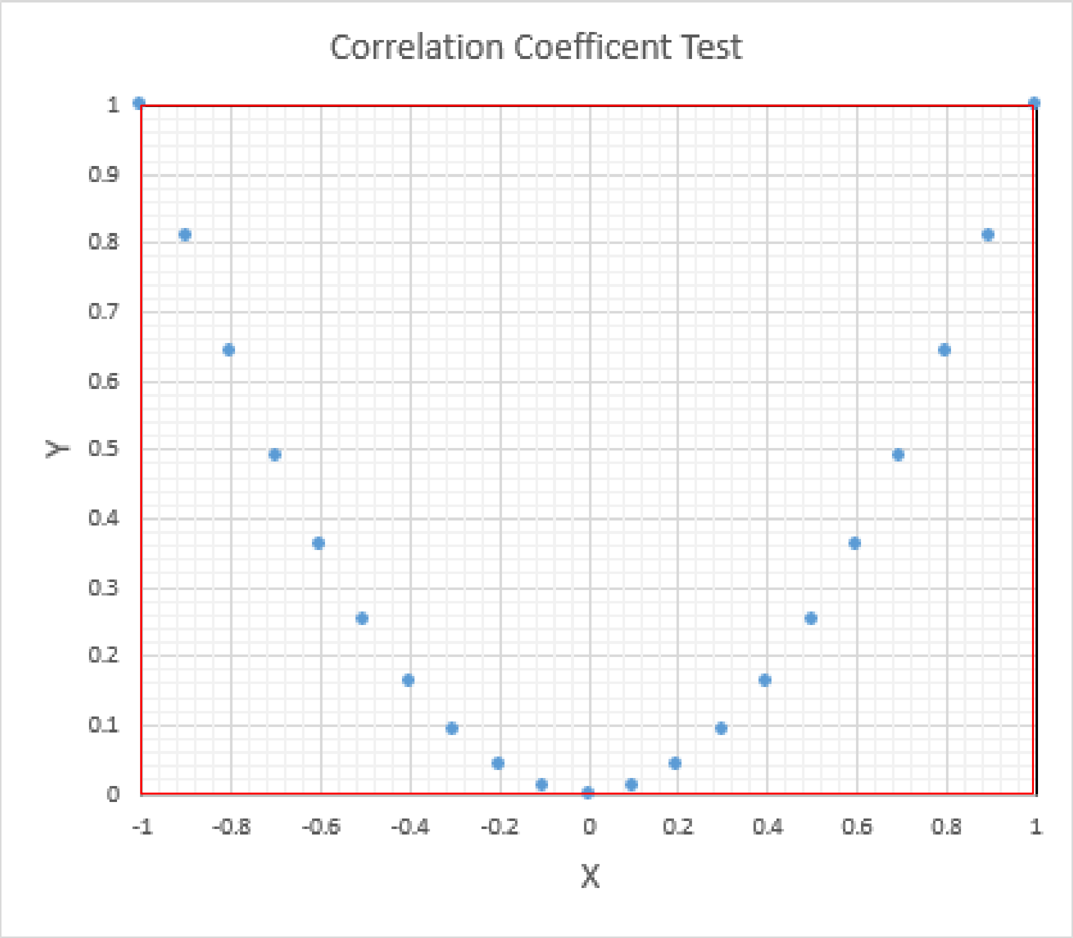
Random data rank transformed with one bivariate outlier.

Bivariate Statistics

Measuring Linear Relationships with the Correlation Coefficient

Correlation / Covariance is a measure of linear relationship

- What is the Correlation / Covariance of $y = x^2$ over range of $[-1, 1]$?



Bivariate data that have follow a parabola.

Excel Function Correl(array1,array2)

Bivariate Statistics

Measuring Linear Relationships with the Correlation Coefficient

Correlation / Covariance is a measure of linear relationship

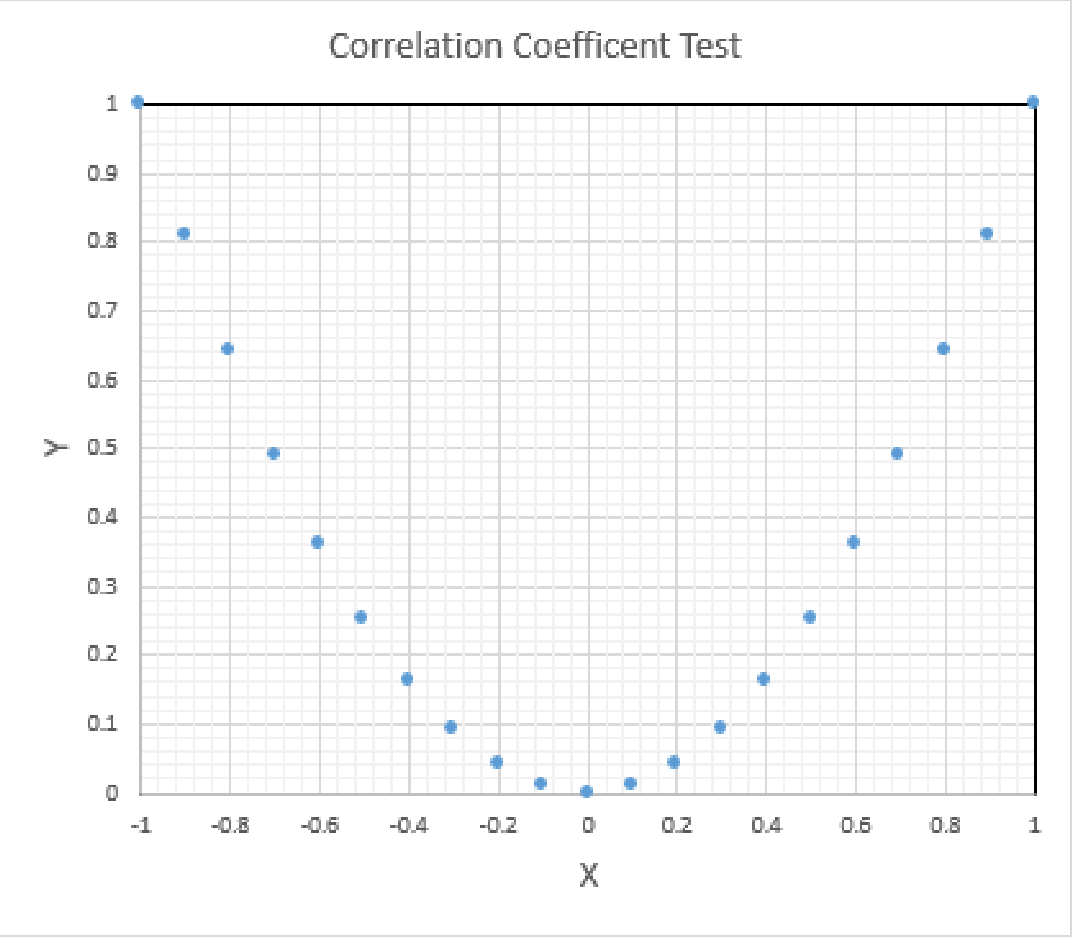
- What is the Correlation / Covariance of

  $y = x^2$ over range of $[-1, 1]$?

- Correlation Coefficient $\boldsymbol{\rho}_{x,y}$ = 0.0!

  $y = x^2$ over range of $[0, 1]$?

- Correlation Coefficient $\boldsymbol{\rho}_{x,y}$ = 0.98!



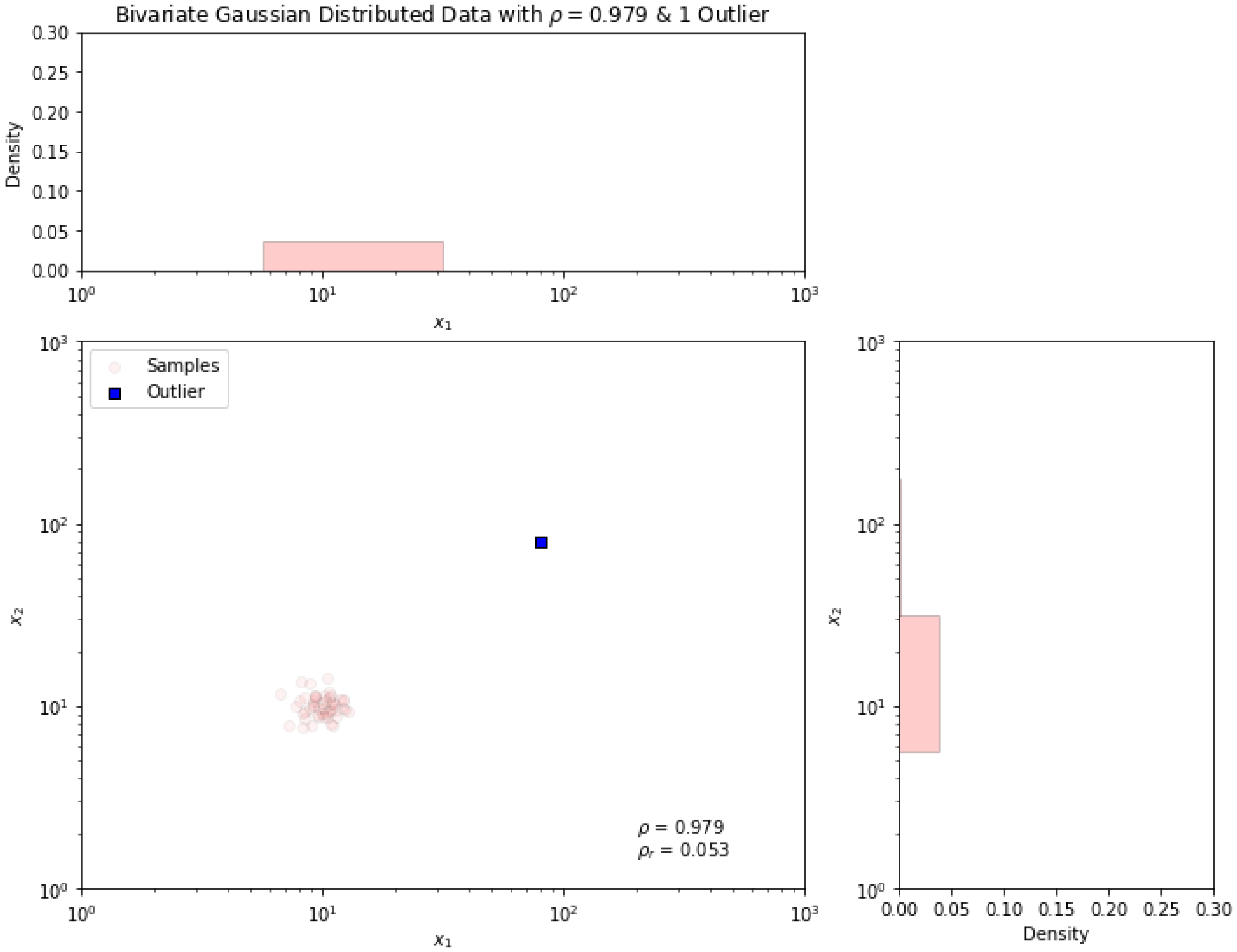
Bivariate data that have follow a parabola.

Excel Function Correl(array1,array2)

Interactive Bivariate Outlier Demonstration

Let’s make a random data set, and add and adjust a single outlier

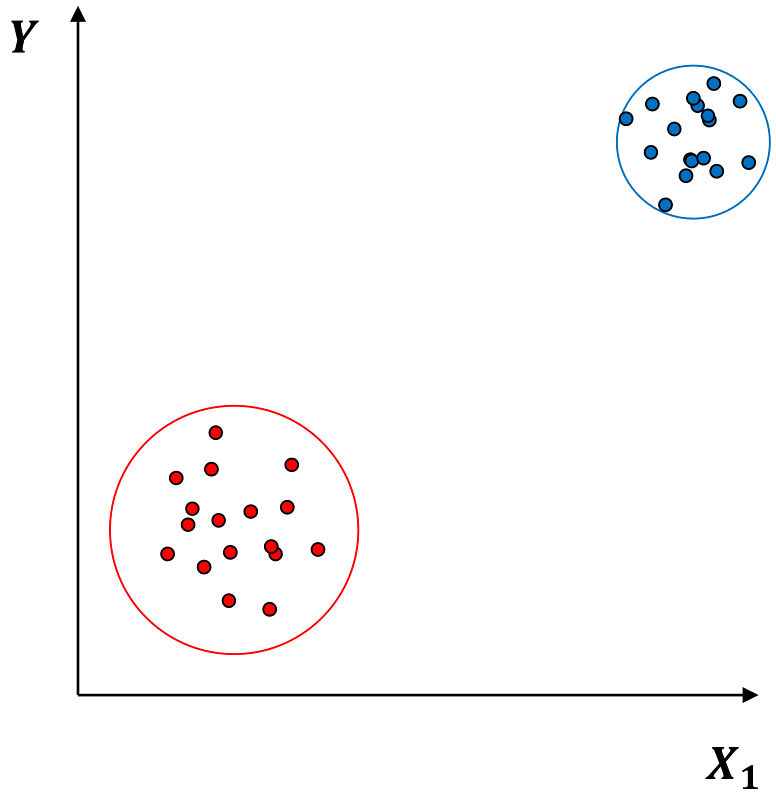
- Watch the Pearson and Spearman correlation coefficients as you increase X and then increase Y for the single outlier.



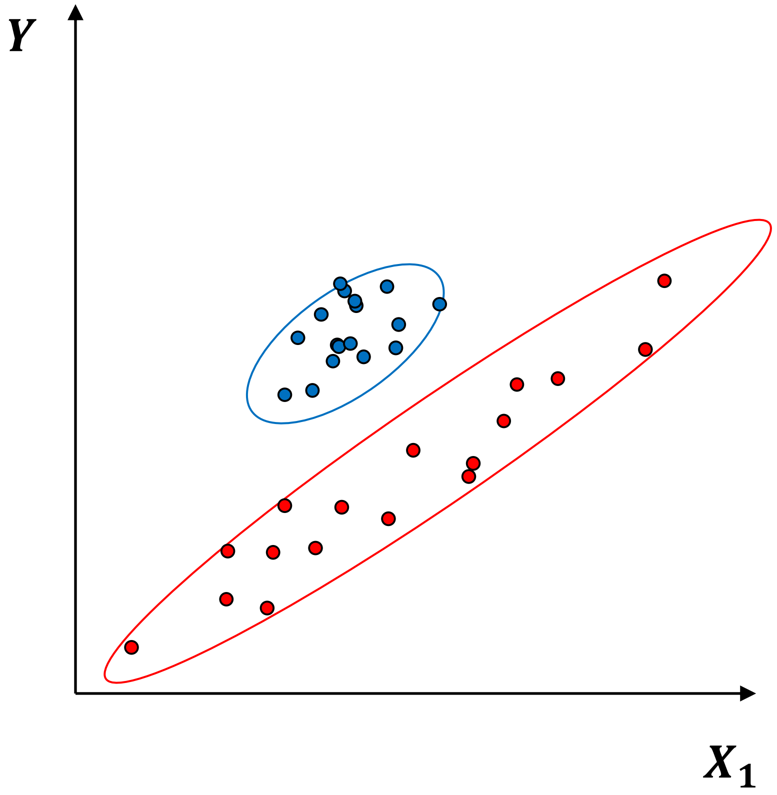
Interactive Outlier demonstration, the file is Interactive_Correlation_Coefficient_Issues.ipynb.

Bivariate Statistics
Mixing Populations

Mixing multiple populations may impact correlation coefficients.



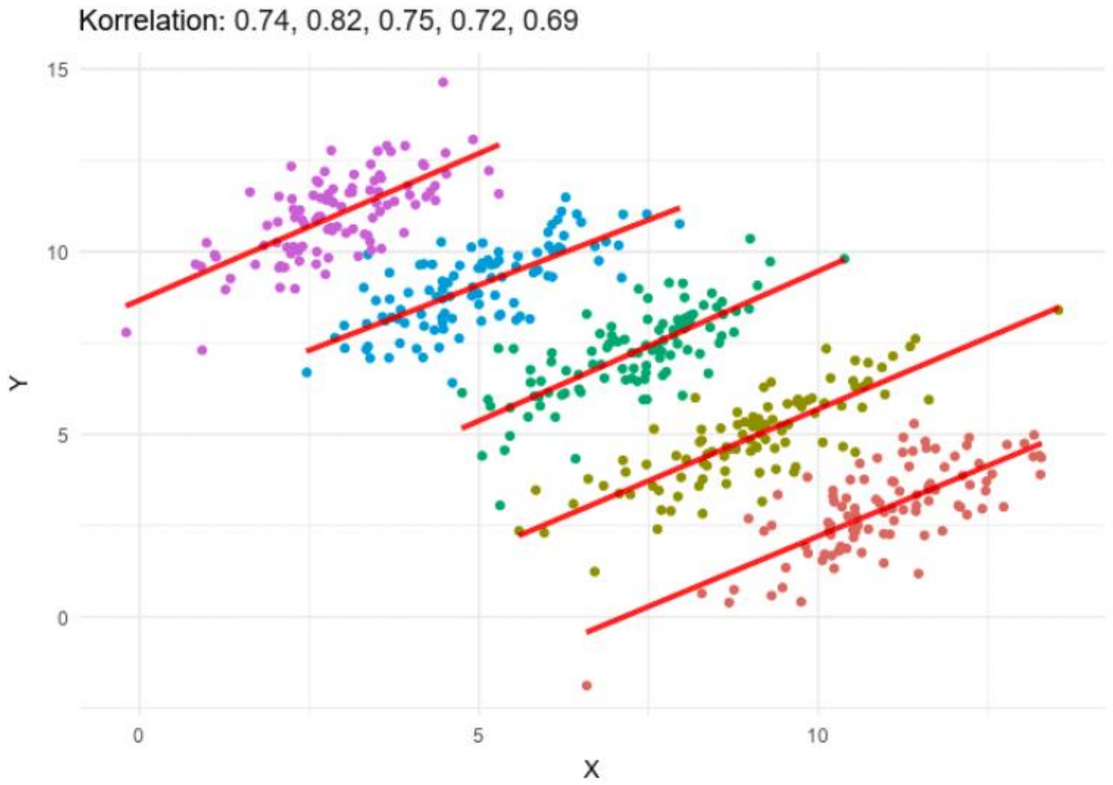
2 populations artificially inflating correlation



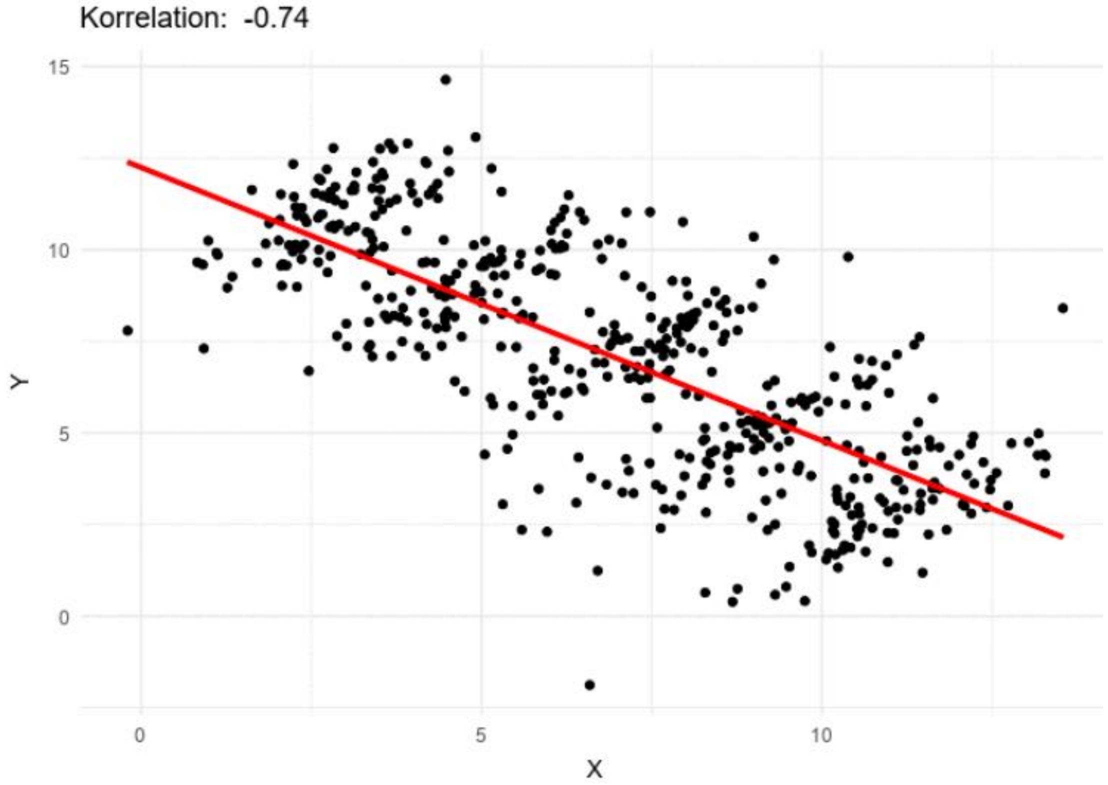
2 populations artificially deflating correlation

Simpson's Paradox

Correlation within multiple groups disappears or reverses when the groups are combined together.



5 groups each with positive correlations.



Combined group with a negative correlation.

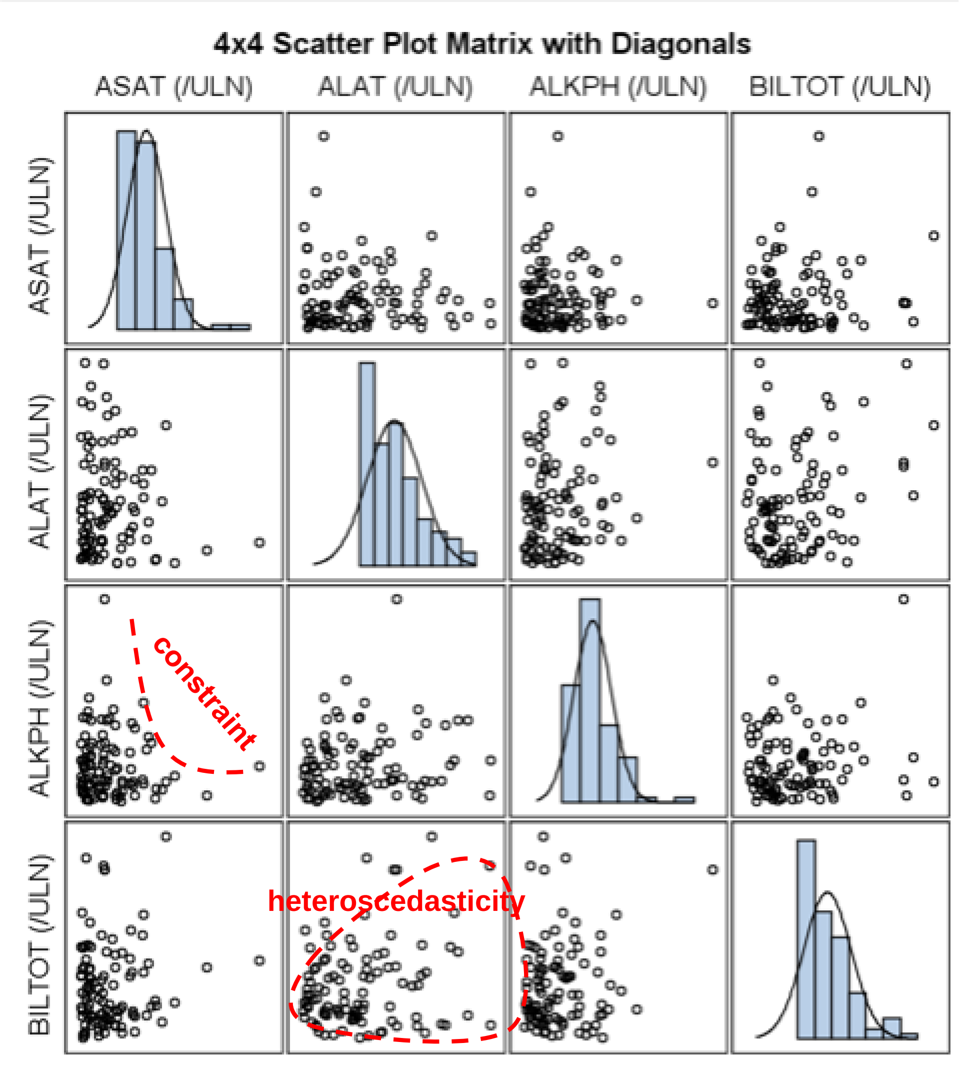
Images from Wikipedia gif available at https://commons.wikimedia.org/wiki/File:Simpsons_paradox_-_animation.gif by Pace~svwiki.

Bivariate Statistics

Other Issues Correlation Coefficient

For more than two variables make matrix scatterplots:

- By hand in Excel or packages in R and Python.
- Look for linear / nonlinear features
- Look for homoscedasticity (constant conditional variance) and heteroscedasticity (conditional variance changes with value)
- Look for constraints



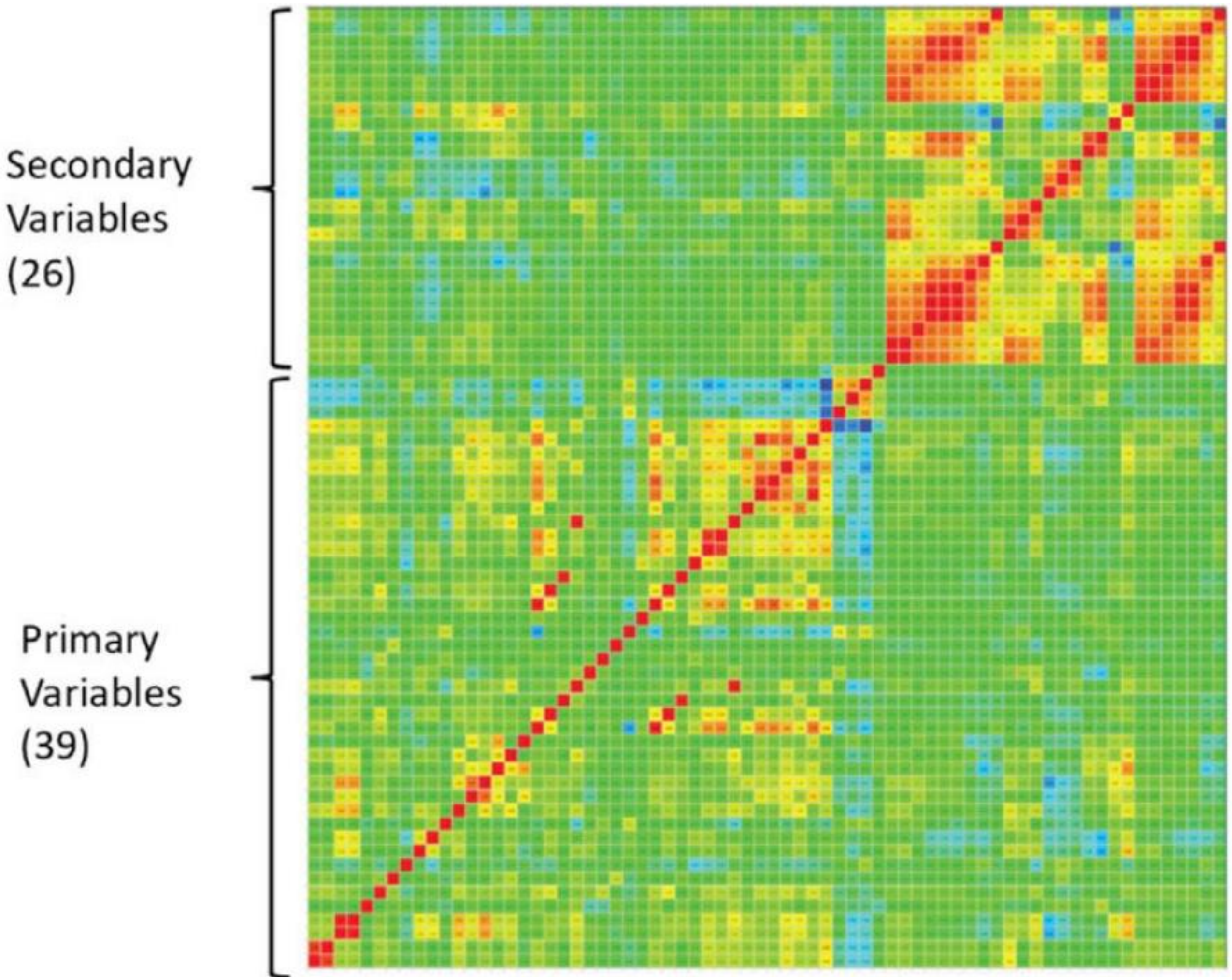
Matrix scatter plot with a couple interpretations.

Bivariate Statistics

Other Issues Correlation Coefficient

Correlation Matrix Plot

- look for high and low correlations
- collinearity, variables with the same correlations with the other variables.

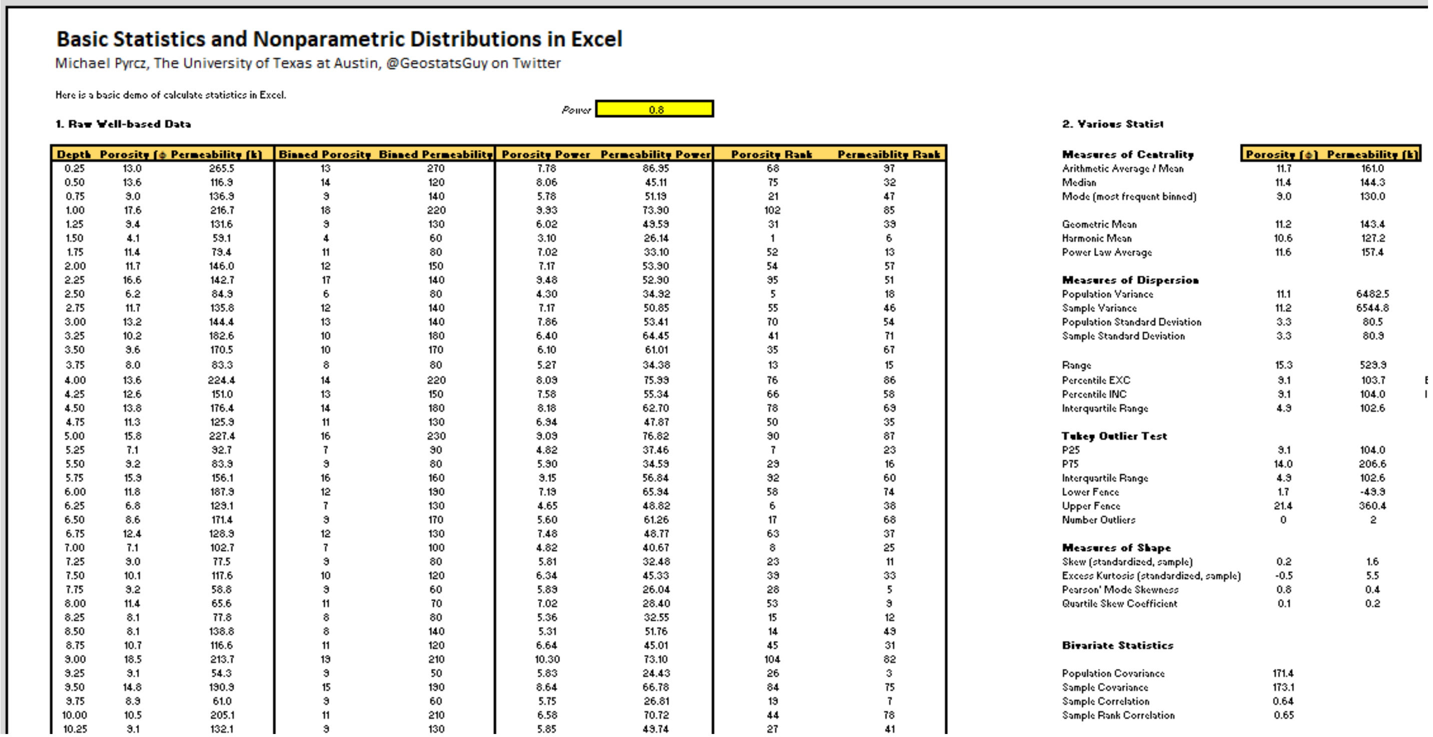


Correlation matrix plot from McLennon et al., (2006).

Bivariate Statistics In Excel

Calculating bivariate statistics in Excel demo

- covariance
- Pearson product-moment correlation coefficient
- Spearman rank correlation coefficient



Basic Statistics and Nonparametric Distributions in Excel

Michael Pyrcz, The University of Texas at Austin, @GeostatsGuy on Twitter

Here is a basic demo of calculate statistics in Excel.

Power 0.8

1. Raw Well-based Data

Depth	Porosity (ϕ	Permeability (k)	Binned Porosity	Binned Permeability	Porosity Power	Permeability Power	Porosity Rank	Permeaiblity Rank
0.25	13.0	265.5	13	270	7.78	86.95	68	97
0.50	13.6	116.9	14	120	8.06	45.11	75	32
0.75	9.0	136.9	9	140	5.78	51.19	21	47
1.00	17.6	216.7	18	220	9.93	73.90	102	85
1.25	9.4	131.6	9	130	6.02	49.59	31	39
1.50	4.1	59.1	4	60	3.10	26.14	1	6
1.75	11.4	79.4	11	80	7.02	33.10	52	13
2.00	11.7	146.0	12	150	7.17	53.90	54	57
2.25	16.6	142.7	17	140	9.48	52.90	95	51
2.50	6.2	84.9	6	80	4.30	34.92	5	18
2.75	11.7	135.8	12	140	7.17	50.85	55	46
3.00	13.2	144.4	13	140	7.86	53.41	70	54
3.25	10.2	182.6	10	180	6.40	64.45	41	71
3.50	9.6	170.5	10	170	6.10	61.01	35	67
3.75	8.0	83.3	8	80	5.27	34.38	13	15
4.00	13.6	224.4	14	220	8.09	75.99	76	86
4.25	12.6	151.0	13	150	7.58	55.34	66	58
4.50	13.8	176.4	14	180	8.18	62.70	78	69
4.75	11.3	125.9	11	130	6.94	47.87	50	35
5.00	15.8	227.4	16	230	9.09	76.82	90	87
5.25	7.1	92.7	7	90	4.82	37.46	7	23
5.50	9.2	83.9	9	80	5.90	34.59	29	16
5.75	15.9	156.1	16	160	9.15	56.84	92	60
6.00	11.8	187.9	12	190	7.19	65.94	58	74
6.25	6.8	129.1	7	130	4.65	48.82	6	38
6.50	8.6	171.4	9	170	5.60	61.26	17	68
6.75	12.4	128.9	12	130	7.48	48.77	63	37
7.00	7.1	102.7	7	100	4.82	40.67	8	25
7.25	9.0	77.5	9	80	5.81	32.48	23	11
7.50	10.1	117.6	10	120	6.34	45.33	39	33
7.75	9.2	58.8	9	60	5.89	26.04	28	5
8.00	11.4	65.6	11	70	7.02	28.40	53	9
8.25	8.1	77.8	8	80	5.36	32.55	15	12
8.50	8.1	138.8	8	140	5.31	51.76	14	49
8.75	10.7	116.6	11	120	6.64	45.01	45	31
9.00	18.5	213.7	19	210	10.30	73.10	104	82
9.25	9.1	54.3	9	50	5.83	24.43	26	3
9.50	14.8	190.9	15	190	8.64	66.78	84	75
9.75	8.9	61.0	9	60	5.75	26.81	19	7
10.00	10.5	205.1	11	210	6.58	70.72	44	78
10.25	9.1	132.1	9	130	5.85	49.74	27	41

2. Various Statist

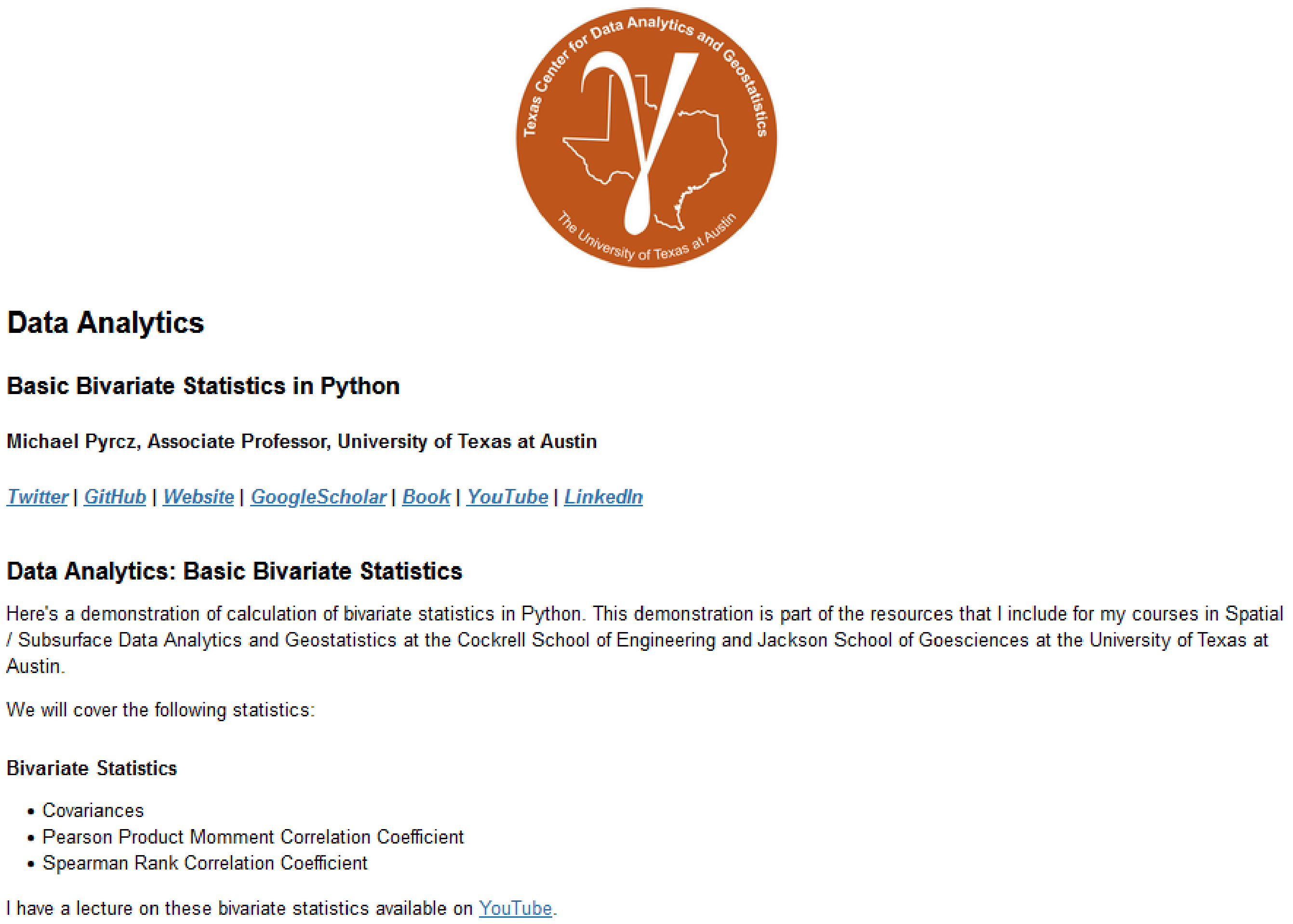
Measures of Centrality	Porosity (ϕ)	Permeability (k)	
Arithmetic Average / Mean	11.7	161.0	
Median	11.4	144.3	
Mode (most frequent binned)	9.0	130.0	
Geometric Mean	11.2	143.4	
Harmonic Mean	10.6	127.2	
Power Law Average	11.6	157.4	
Measures of Dispersion			
Population Variance	11.1	6482.5	
Sample Variance	11.2	6544.8	
Population Standard Deviation	3.3	80.5	
Sample Standard Deviation	3.3	80.9	
Range	15.3	529.9	
Percentile EXC	9.1	103.7	E
Percentile INC	9.1	104.0	I
Interquartile Range	4.9	102.6	
Tukey Outlier Test			
P25	9.1	104.0	
P75	14.0	206.6	
Interquartile Range	4.9	102.6	
Lower Fence	1.7	-49.9	
Upper Fence	21.4	360.4	
Number Outliers	0	2	
Measures of Shape			
Skew (standardized, sample)	0.2	1.6	
Excess Kurtosis (standardized, sample)	-0.5	5.5	
Pearson' Mode Skewness	0.8	0.4	
Quartile Skew Coefficient	0.1	0.2	
Bivariate Statistics			
Population Covariance	171.4		
Sample Covariance	173.1		
Sample Correlation	0.64		
Sample Rank Correlation	0.65		

File is Basic_Statistics_Demo.xlsx.

Bivariate Statistics
In Python

Calculating bivariate statistics in Python demo

- covariance
- Pearson product-moment correlation coefficient
- Spearman rank correlation coefficient



Data Analytics

Basic Bivariate Statistics in Python

Michael Pyrcz, Associate Professor, University of Texas at Austin

Twitter | *GitHub* | *Website* | *GoogleScholar* | *Book* | *YouTube* | *LinkedIn*

Data Analytics: Basic Bivariate Statistics

Here's a demonstration of calculation of bivariate statistics in Python. This demonstration is part of the resources that I include for my courses in Spatial / Subsurface Data Analytics and Geostatistics at the Cockrell School of Engineering and Jackson School of Goesciences at the University of Texas at Austin.

We will cover the following statistics:

Bivariate Statistics

- Covariances
- Pearson Product Momment Correlation Coefficient
- Spearman Rank Correlation Coefficient

I have a lecture on these bivariate statistics available on YouTube.

File is PythonDataBasics_Bivariate_Statistics.ipynb.

PGE 338 Data Analytics and Geostatistics

Lecture 8: Bivariate Distributions

Lecture outline . . .

- Bivariate Statistics
- Correlation

Introduction

General Concepts

Univariate

Bivariate

Correlation

Regression

Model Checking

Time Series Analysis

Spatial Analysis

Machine Learning

Uncertainty Analysis

Michael Pyrcz, The University of Texas at Austin