UNIVERSITY OF TWENTE.

EXPLORATORY DATA ANALYSIS & EXPLORATORY SPATIAL DATA ANALYSIS WITH PYTHON

Mahdi KHODADADZADEH

October 2021

ITC FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION

INGESTING DATA

- Getting data into Python in a shape that we can use to start our analysis.
  - Reading comma separated value (CSV) data: pandas.read_csv()
  - Reading an Excel file: pandas.read_excel()
  - Reading a MATLAB file: scipy.io.loadmat()
  - Reading shapefile and GeoJSON files: geopandas.read_file()
  - Reading GeoTIFF: gdal.Open()
  - Reading an image: matplotlib.pyplot.imread()

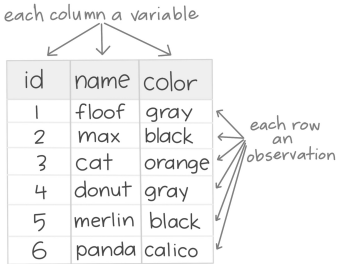
TIDYING DATA

- Data preparation: messy data → tidy data
- Rectangular data structures → Data modelling

> "TIDY DATA is a standard way of mapping the meaning of a dataset to its structure."
>
> –HADLEY WICKHAM

In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement



Wickham, H. (2014). Tidy Data. Journal of Statistical Software 59 (10). DOI: 10.18637/jss.v059.i10

https://www.openscapes.org/blog/2020/10/12/tidy-data/

EXPLORATORY DATA ANALYSIS (EDA)

- Better understanding the (tidy) data at hand
- Statistics + Visualization
  - Get an overview of the data
  - Transform variables
  - Orient further analysis → choose correct methods/approaches
  - Help you to generate hypothesis
  - Spot problems in data
  - Understand variable properties (e.g., mean, variance and outliers)
  - Understand relationships between variables

A PRACTICAL EXAMPLE

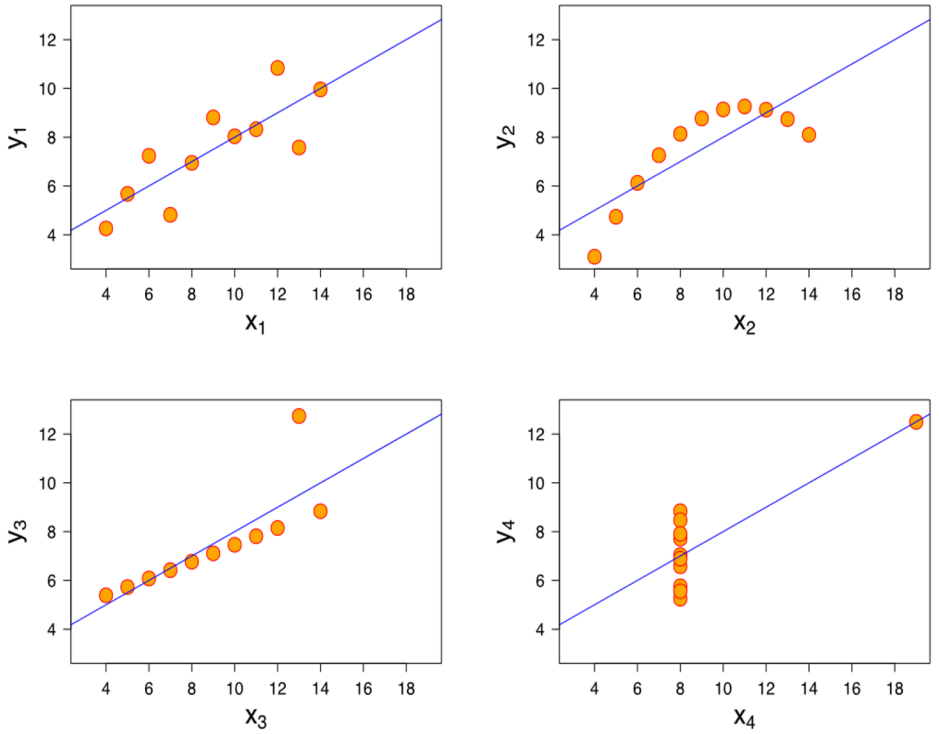
Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Property	Value
Mean of *x* in each case	9 (exact)
Variance of *x* in each case	11 (exact)
Mean of *y* in each case	7.50 (to 2 decimal places)
Variance of *y* in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between *x* and *y* in each case	0.816 (to 3 decimal places)
Linear regression line in each case	$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)

A PRACTICAL EXAMPLE

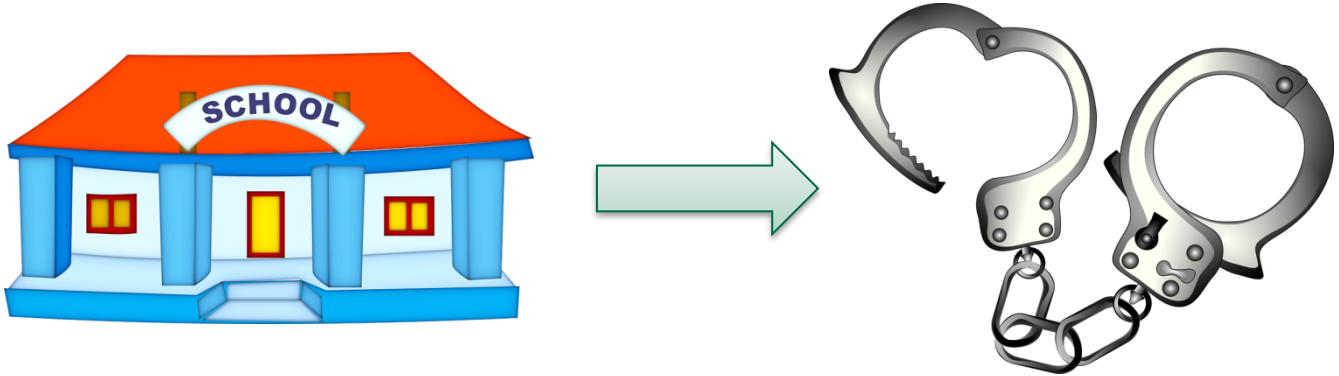
- Visualization
  - Maximize insight into a data set
  - Uncover underlying structure



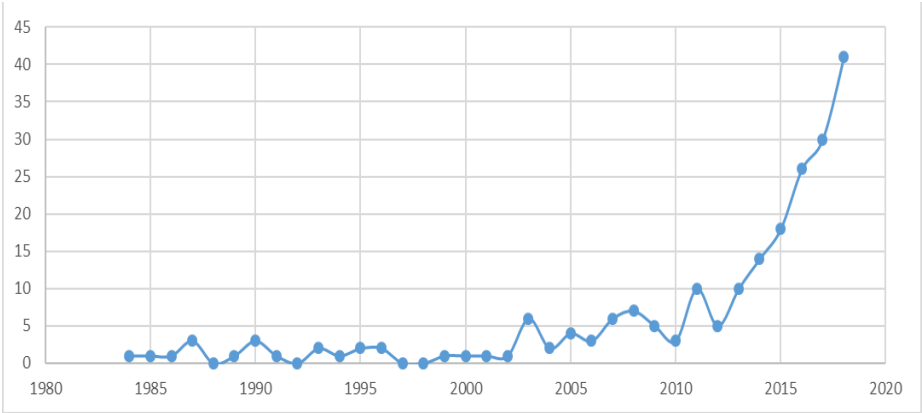
MISUSES OF STATISTICS

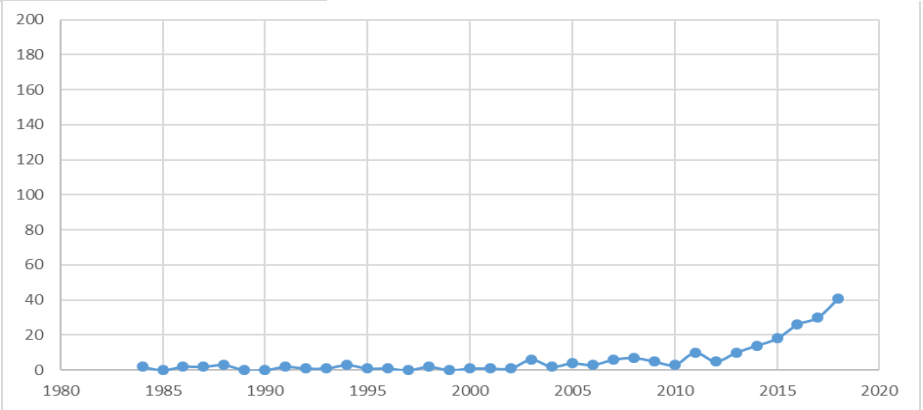
- Interpretation matters…!

  The more schools in a city, the more crime there is. Thus, schools lead to crime



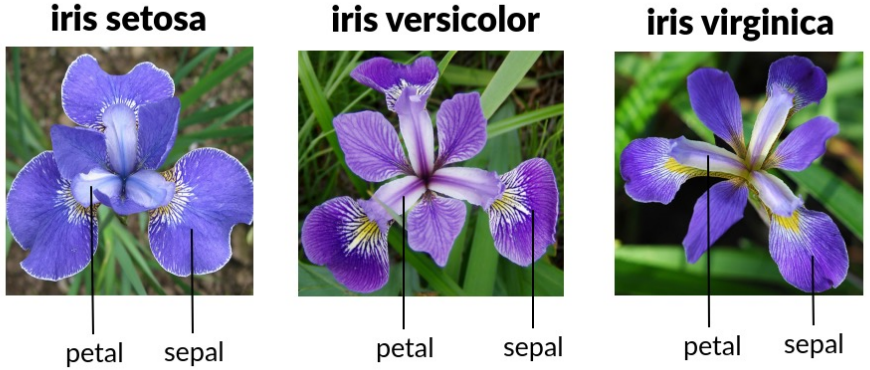
MISUSES OF VISUALIZATION





IRIS DATASET

- 4 features
- 3 classes
- 150 rows



id	species	sepal length	sepal width	petal length	petal width
1	Iris-setosa	5.1	3.5	1.4	0.2
2	Iris-setosa	4.9	3.0	1.4	0.2
3	Iris-setosa	4.7	3.2	1.3	0.2
4	Iris-setosa	4.6	3.1	1.5	0.2
5	Iris-setosa	5.0	3.6	1.4	0.2

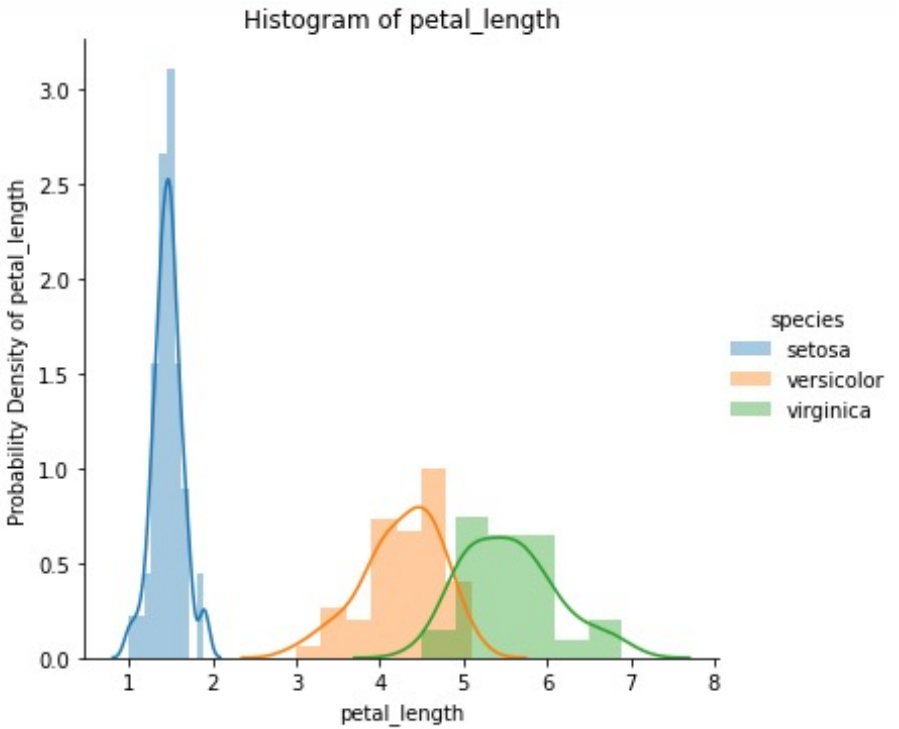
FIRST QUESTIONS

- What is the size of the dataset?
- How many features?
- What are the column (feature) names?
- How many classes?
- How many samples in each class?
- Is there any missing (nan) value?

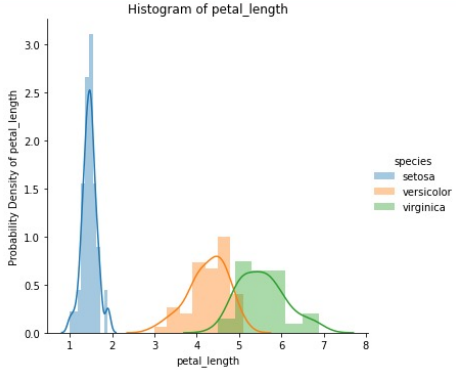
id	species	sepal length	sepal width	petal length	petal width
1	Iris-setosa	5.1	3.5	1.4	0.2
2	Iris-setosa	4.9	3.0	1.4	0.2
3	Iris-setosa	4.7	3.2	1.3	0.2
4	Iris-setosa	4.6	3.1	1.5	0.2
5	Iris-setosa	5.0	3.6	1.4	0.2

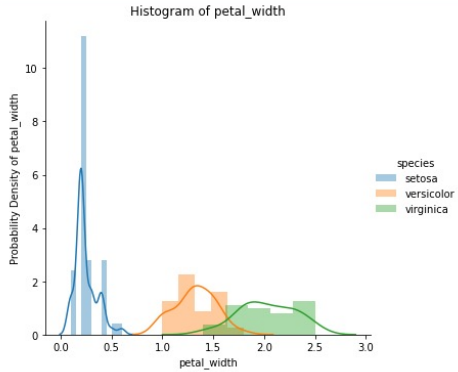
UNIVARIATE ANALYSIS

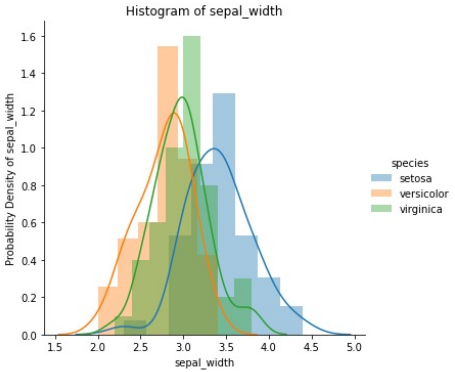
- **Histogram and PDF**
- Python: seaborn and matplotlib

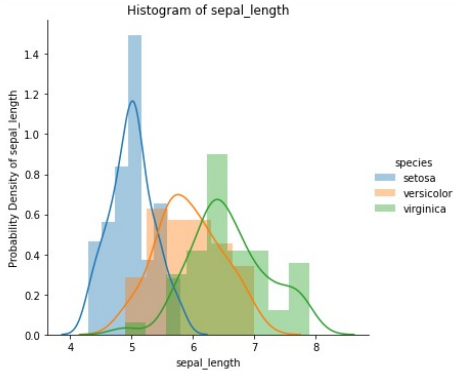


UNIVARIATE ANALYSIS









UNIVARIATE ANALYSIS

- **Mean and Standard Deviation**
- Python: NumPy, pandas

	species	sepal length	sepal width	petal length	petal width
mean	versicolor	5.936	2.770	4.26	1.325
	setosa	5.006	3.418	1.464	0.244
	virginica	6.587	2.974	5.552	2.026
std	versicolor	0.510	0.310	0.465	0.195
	setosa	0.348	0.377	0.171	0.106
	virginica	0.629	0.319	0.546	0.271

Sample Standard Deviation Formulas

Formula to calculate mean

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Formula to Estimate Sample Standard Deviation

$$\sigma_{n-1} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Formula to Estimate Sample SD Variance

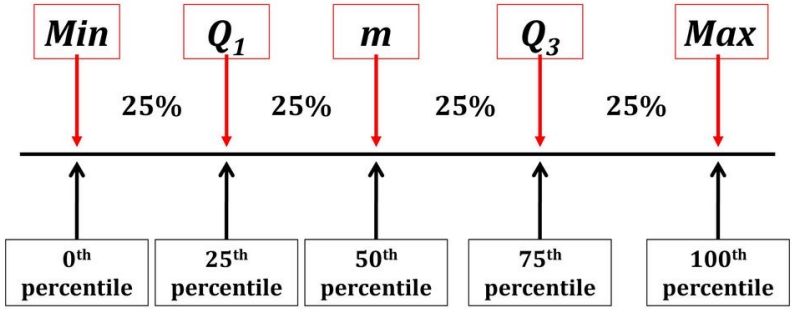
$$(\sigma_{n-1})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

https://standard-deviation-calculator.com/

UNIVARIATE ANALYSIS

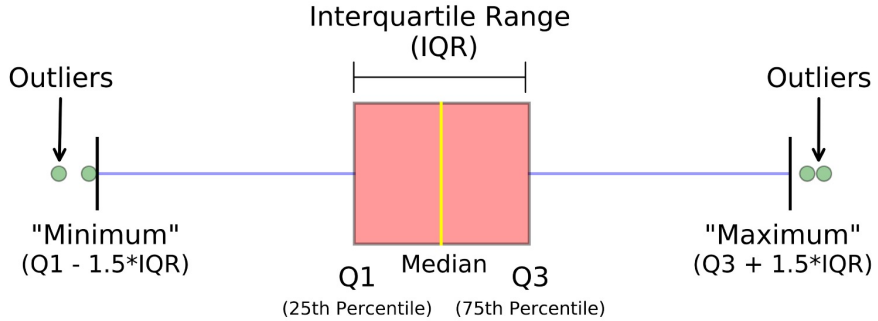
- **Min, Max, Median, Percentile, Quartile**
- Python: NumPy, pandas
- Percentile: Given a vector V of length N, the q-th percentile of V is the value q/100 of the way from the minimum to the maximum in a sorted copy of V.
- Quartile: The q-th quantile of V is the value q of the way from the minimum to the maximum in a sorted copy of V.

five-number summary →



UNIVARIATE ANALYSIS

- **Box plot**: displays the five-number summary (the minimum, first quartile, median, third quartile, and maximum) of a set of data.
- Within the BOX 25th percentile to 75th percentile values
- It can tell you about your outliers and what their values are
- Python: seaborn, matplotlib, pandas

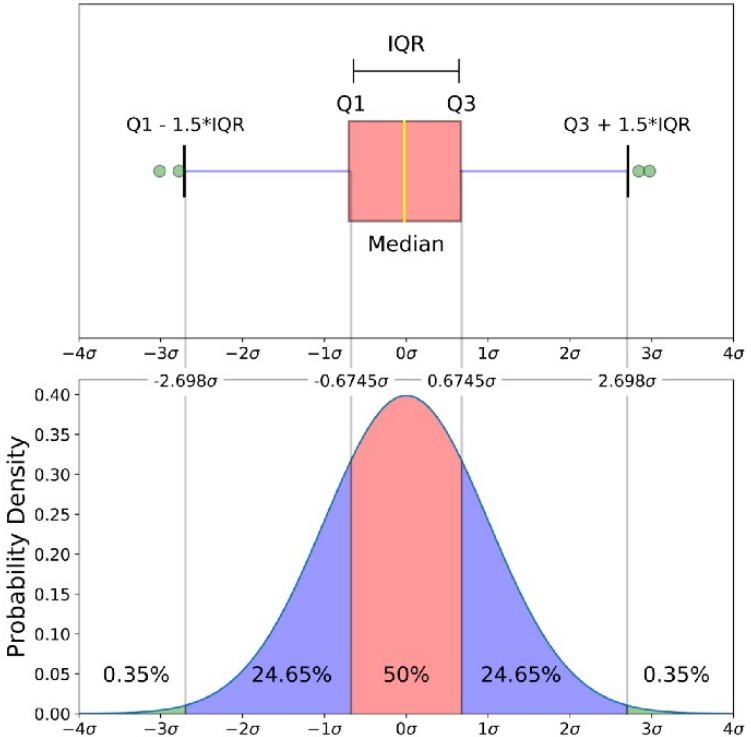


https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

UNIVARIATE ANALYSIS

Q2-median = Q3-median

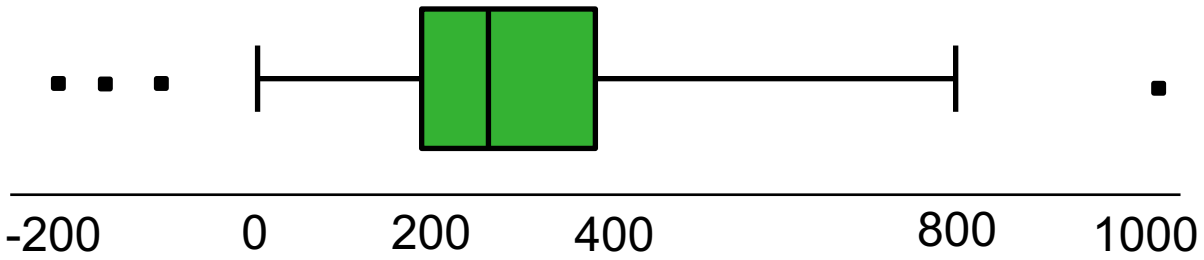
mean = median = mode



https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

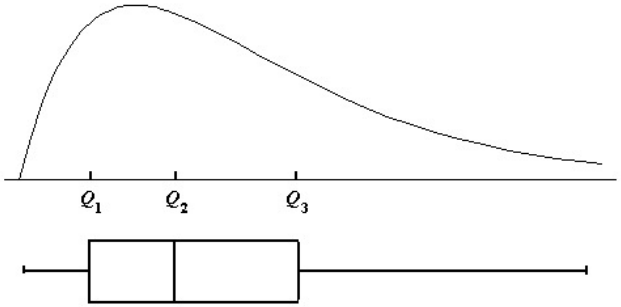
UNIVARIATE ANALYSIS

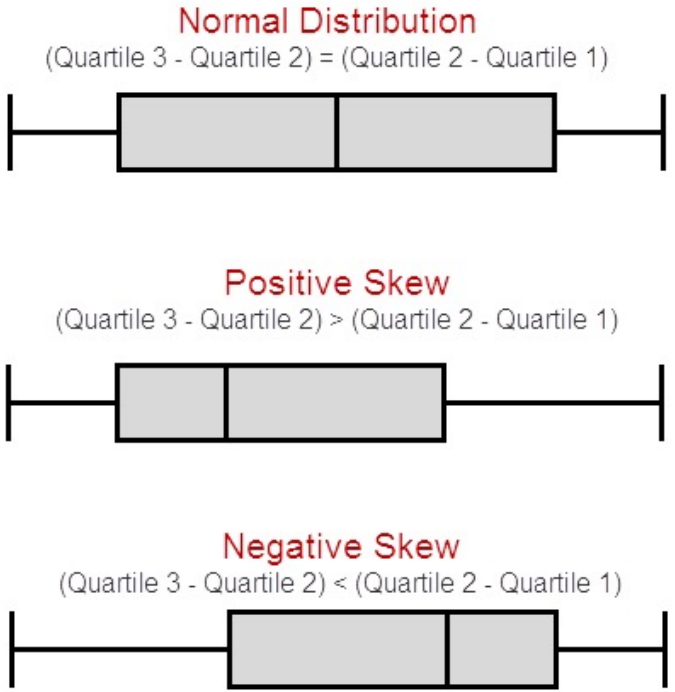
- What is the range?
- What is the interquartile range?



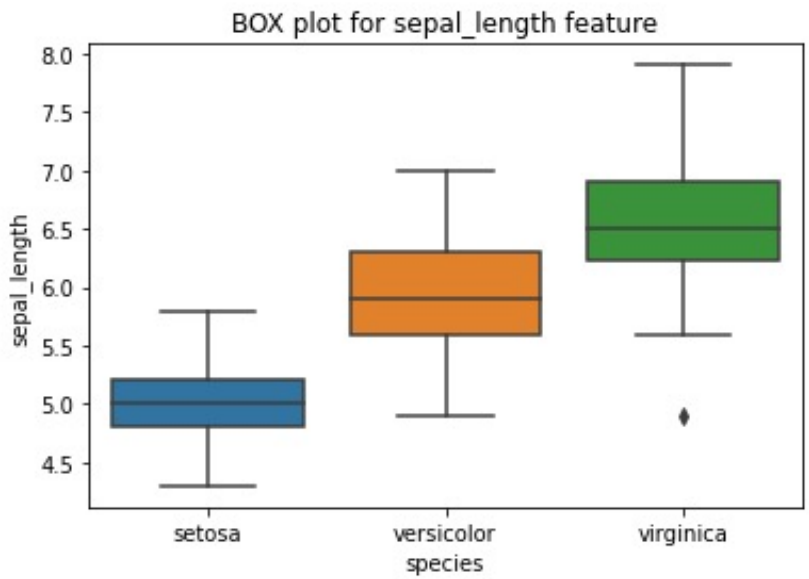
UNIVARIATE ANALYSIS

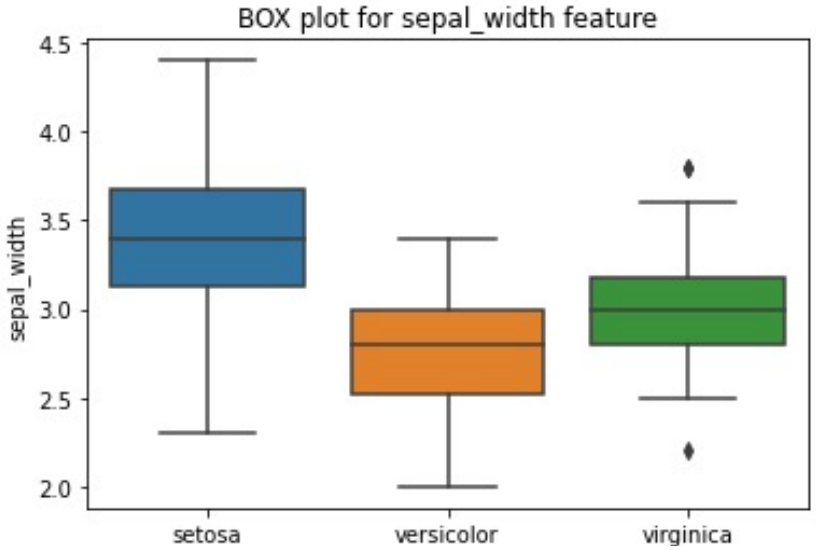
- **A positively skewed** (or right-skewed) distribution is a type of distribution in which most values are clustered around the left tail





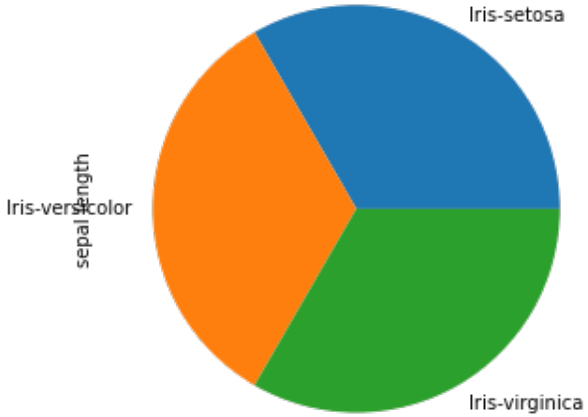
UNIVARIATE ANALYSIS

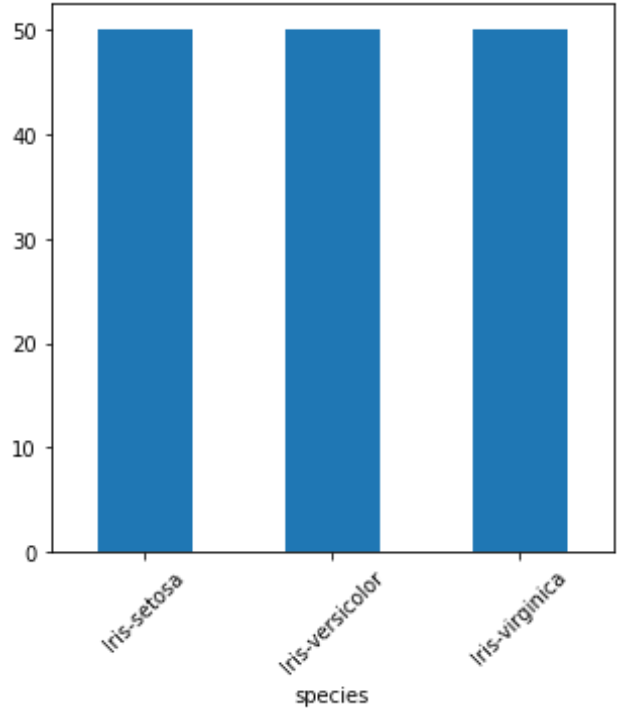




UNIVARIATE ANALYSIS

- **Bar plots, Pie chart**
- Python: seaborn, matplotlib



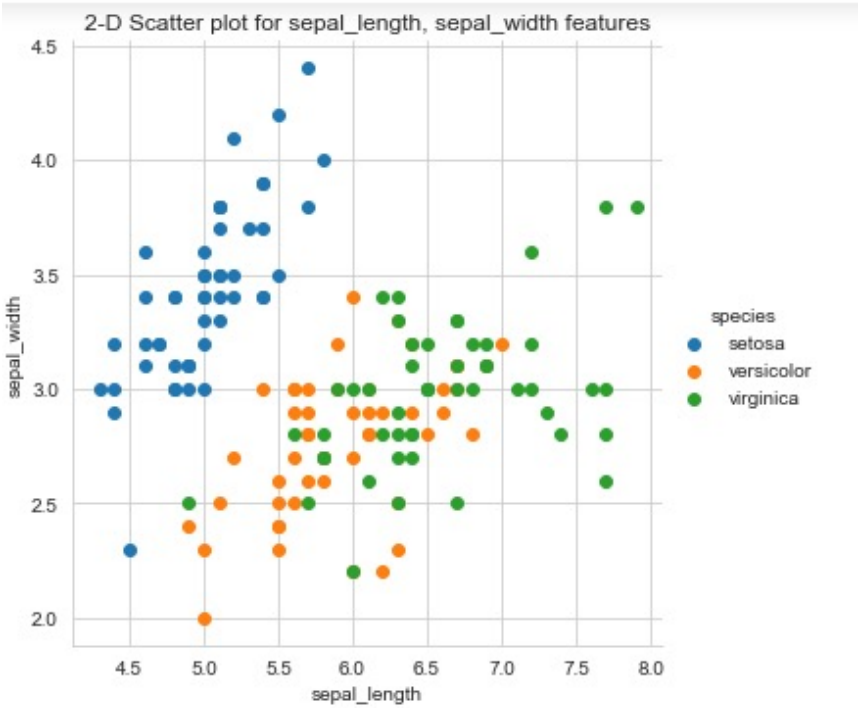


BI-VARIATE ANALYSIS

- **2-D Scatter Plots**
- Python: seaborn, matplotlib, pandas
- They can show correlation between the variables:
  - no correlation exists → randomly scattered
  - a large correlation exists → concentrate near a straight line

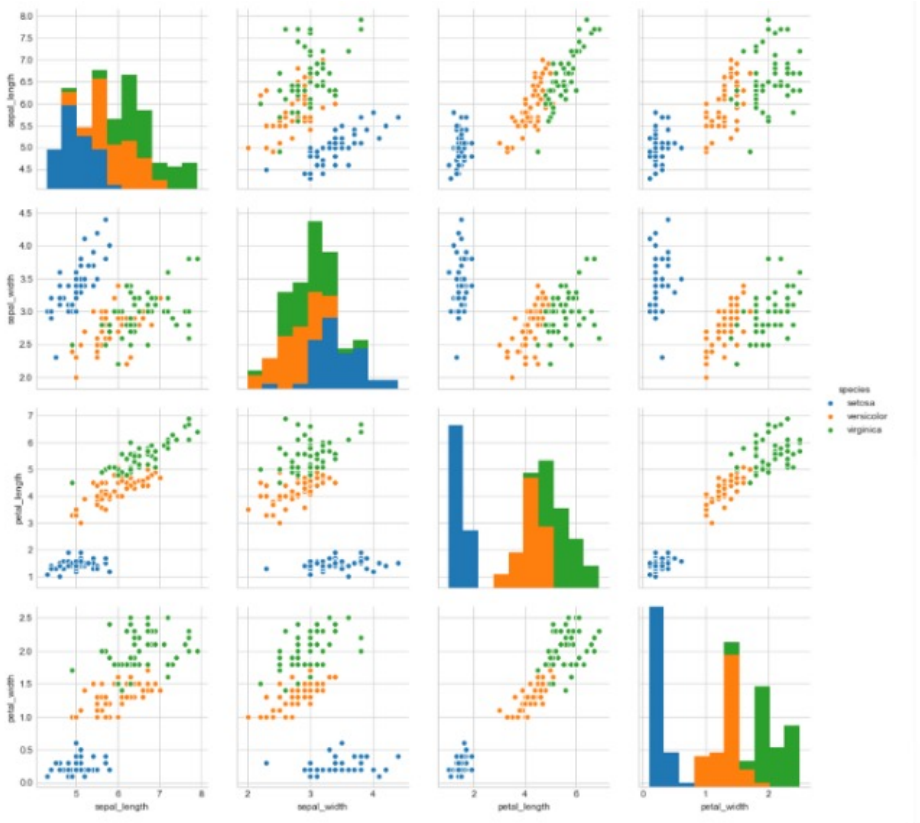
BI-VARIATE ANALYSIS

- **2-D Scatter Plots**



BI-VARIATE ANALYSIS

- **Pair-plot**



BI-VARIATE ANALYSIS

- **Covariance and correlation**
- Relationship between two variables quantitatively
- NumPy, pandas

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

$$cor(x,y) = \frac{cov(x,y)}{sd(x)sd(y)}$$

DATA TRANSFORMATIONS

- **Standardization, mean normalization and min-max scaling**
- Python: scikit learn

Standardization:

$$x' = \frac{x - \bar{x}}{\sigma}$$

Mean Normalization:

$$x' = \frac{x - \bar{x}}{max(x) - min(x)}$$

Min-Max Scaling:

$$x' = \frac{x - min(x)}{max(x) - min(x)}$$

DATA TRANSFORMATIONS

- Fit the scaler using the training set, then apply the same scaler to transform the test set.
- Do not scale the training and test sets using different scalers: this could lead to random skew in data
- Do not fit the scaler using any part of the test data: referencing the test data can lead to a form of data leakage.

- Check out this page:

  http://www.faqs.org/faqs/ai-faq/neural-nets/part2/section-16.html

DATA TRANSFORMATIONS

- **Convert categorical variable** into dummy/indicator variables
- Python: pandas

Land Cover
Forest
Forest
Agriculture
Water
Water

Forest	Agriculture	Water
1	0	0
1	0	0
0	1	0
0	0	1
0	0	1

DATA TRANSFORMATIONS

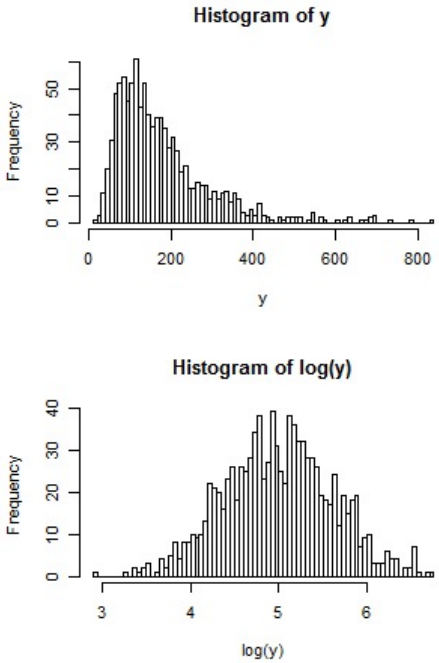
- **Discretizing** continuous values
- Python: pandas

id	age
0	2
1	67
2	40
3	32
4	4
5	15
6	82
7	99
8	26
9	30

id	age	age_group
0	2	(1.903, 34.333]
1	67	(66.667, 99.0]
2	40	(34.333, 66.667]
3	32	(1.903, 34.333]
4	4	(1.903, 34.333]
5	15	(1.903, 34.333]
6	82	(66.667, 99.0]
7	99	(66.667, 99.0]
8	26	(1.903, 34.333]
9	30	(1.903, 34.333]

DATA TRANSFORMATIONS

- **Log, square root, and Box-Cox transformations**
- Applying a transformation to reduce skewness
- Python: NumPy, Scipy, scikit learn



- Check out this page:

  https://www.statisticshowto.com/box-cox-transformation/

EXPLORATORY SPATIAL DATA ANALYSIS

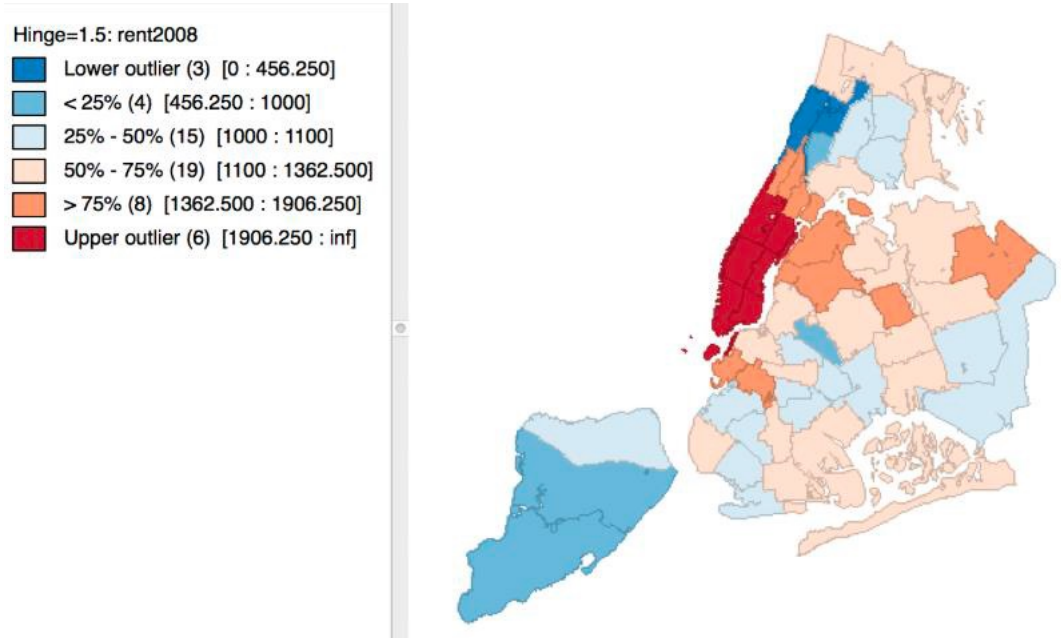
- EDA applied to spatial data → ESDA

- "Traditional" EDA can be applied to spatial datasets calculate statistics, basic plots (histograms, boxplots).
- Do not always have a statistical nature

Convex hull, variety, majority, minority, min, max, frequency, mean, median, standard deviation, standard deviational ellipse

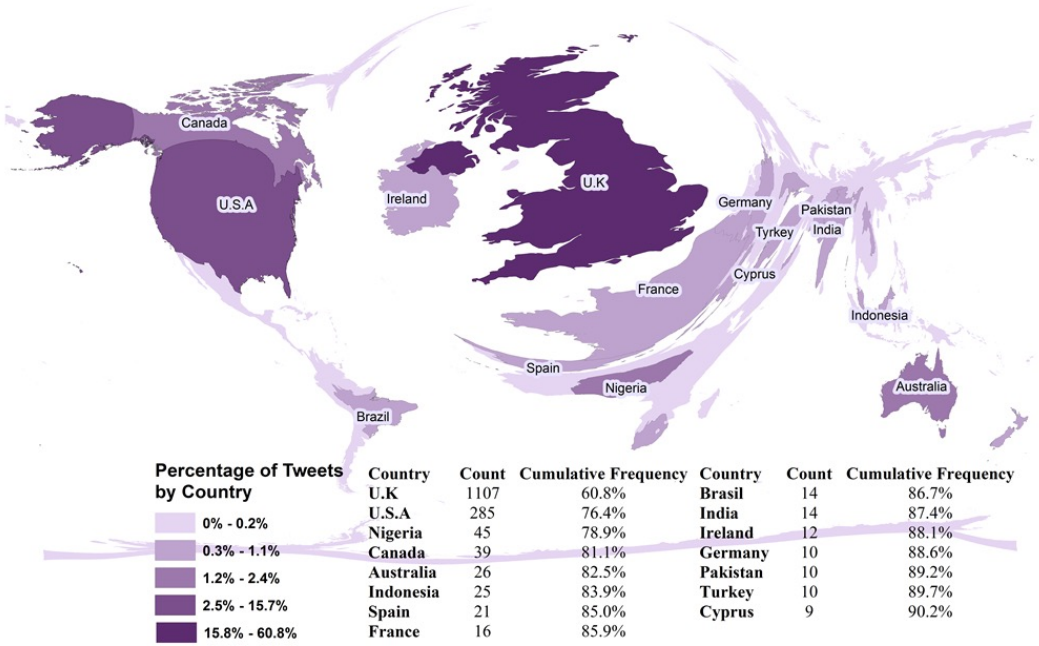
EXAMPLES

- Box map
  - Which value gives you the Q1?



EXAMPLES

- Cartogram



Country	Count	Cumulative Frequency	Country	Count	Cumulative Frequency
U.K	1107	60.8%	Brasil	14	86.7%
U.S.A	285	76.4%	India	14	87.4%
Nigeria	45	78.9%	Ireland	12	88.1%
Canada	39	81.1%	Germany	10	88.6%
Australia	26	82.5%	Pakistan	10	89.2%
Indonesia	25	83.9%	Turkey	10	89.7%
Spain	21	85.0%	Cyprus	9	90.2%
France	16	85.9%			

The word map above shows the percentage of tweets, related to crime incidents in London, by country. The size of each country is increasing with higher intensities.

EXAMPLES

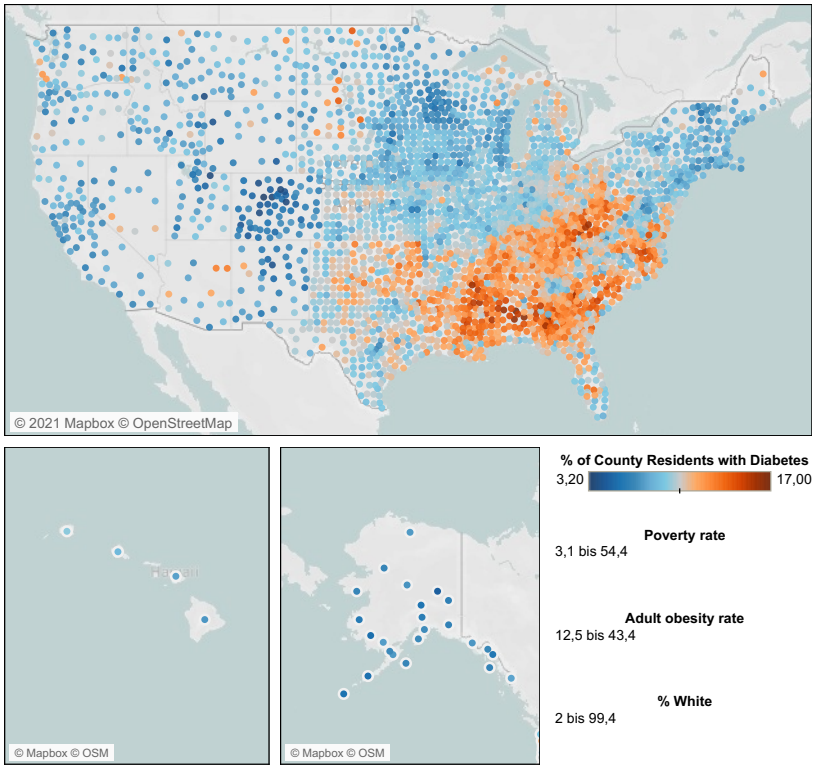
- Connection map

From: https://www.data-to-viz.com/story/MapConnection.html

EXAMPLES

- Conditioned map

The Geography of Diabetes



From: http://publichealthintelligence.org/content/geography-diabetes-us-conditioned-map

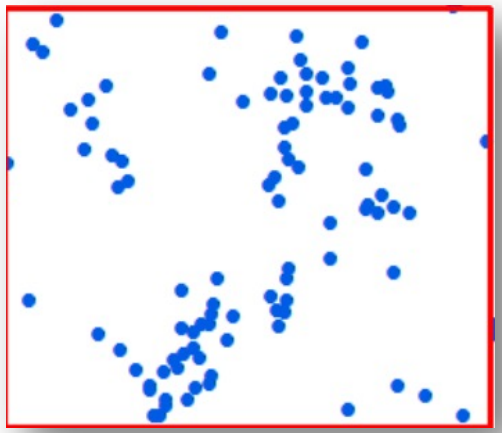
EXAMPLES

- A list of additional examples of ESDA maps:
  - Box Map: https://geodacenter.github.io/workbook/3a_mapping/lab3a.html#extreme-value-maps
  - Brushing & linking: https://www.spatialanalysisonline.com/HTML/eda__esda_and_estda.htm
  - Conditional choropleth mapping: http://publichealthintelligence.org/content/geography-diabetes-us-conditioned-map
  - Voronoi analysis: https://www.gislounge.com/voronoi-diagrams-and-gis/
  - Cartograms: https://gisgeography.com/cartogram-maps/
  - Connection map: https://www.data-to-viz.com/story/MapConnection.html

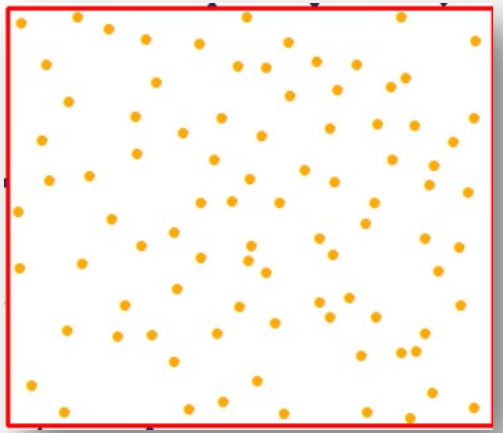
SPATIAL PATTERNS

- Three main ways to describe the spatial pattern of objects:
  - Clustered: occurs when objects exist in close proximity to one another.
  - Dispersed: occurs when objects are spread out from one another.
  - Random: occurs when objects exist in neither a clustered or dispersed pattern.

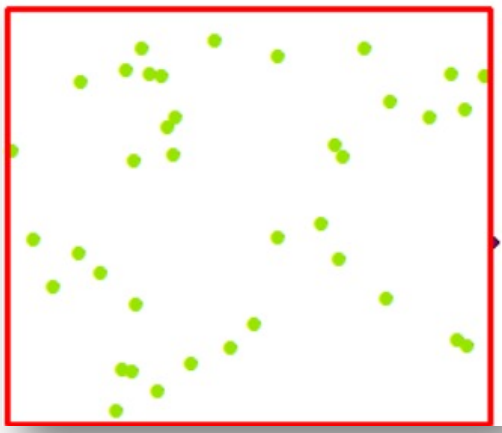
SPATIAL PATTERNS



Clustered



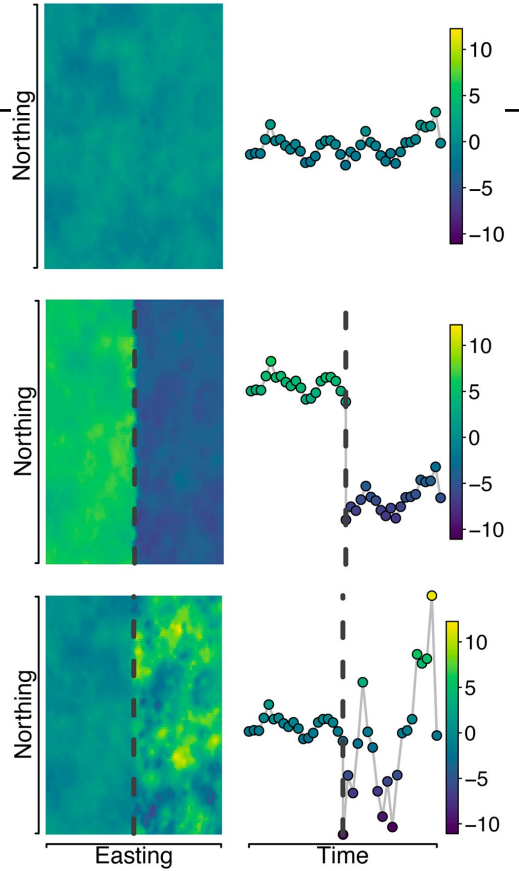
Dispersed



Random

DATASET'S VARIABILITY

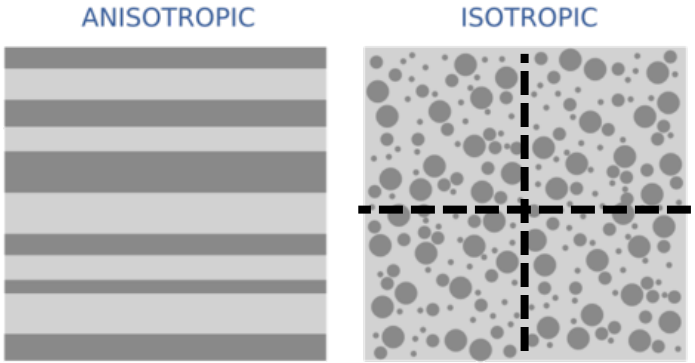
- Stationary process
  - Its statistical properties (mean, variance) are independent of absolute location.
  - properties do not vary in space (non-stationarity is the opposite)



From: Rollinson, Christine R., et al. "Working across space and time: nonstationarity in ecological research and application." Frontiers in Ecology and the Environment 19.1 (2021): 66-72.

DATASET'S VARIABILITY

- Isotropic process
  - A stationary process where the covariance depends only on distance and not direction
  - Properties do not vary in direction (anisotropy is the opposite)



From: https://agilescientific.com/blog/2015/2/9/what-is-anisotropy

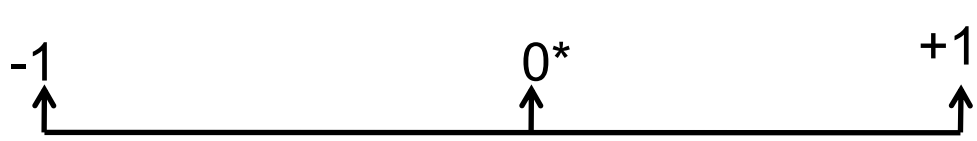
SPATIAL AUTOCORRELATION

- Autocorrelation means correlation of a variable with itself through time and/or space
  - If there is any systematic pattern on the variable's distribution, it is said to be autocorrelated
  - If nearby features are more alike, this is positive autocorrelation
  - Negative autocorrelation describes patterns in which neighboring features are unlike
  - Random patterns exhibit no autocorrelation

MORAN'S I

- ***n*** is the number of cases
- x_i is the variable value at a particular location
- x_j is the variable value at another location
- $\bar{X}$ is the mean of the variable
- w_{ij} is a weight applied to the comparison between location *i* and location *j*

$$I = \frac{n\sum_i\sum_j w_{i,j}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i\sum_j w_{i,j}\sum_i (x_i - \bar{x})^2}$$



**technically it is: –1/(n-1)*

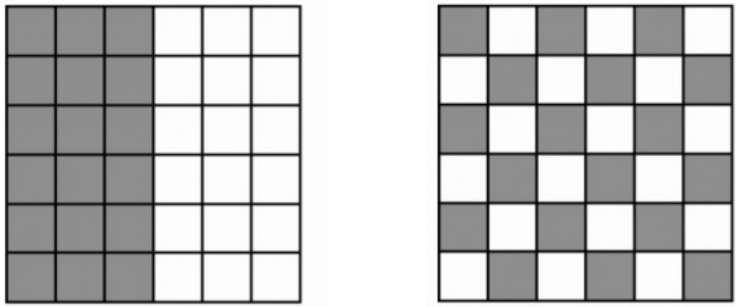
high negative spatial autocorrelation

*no spatial autocorrelation**

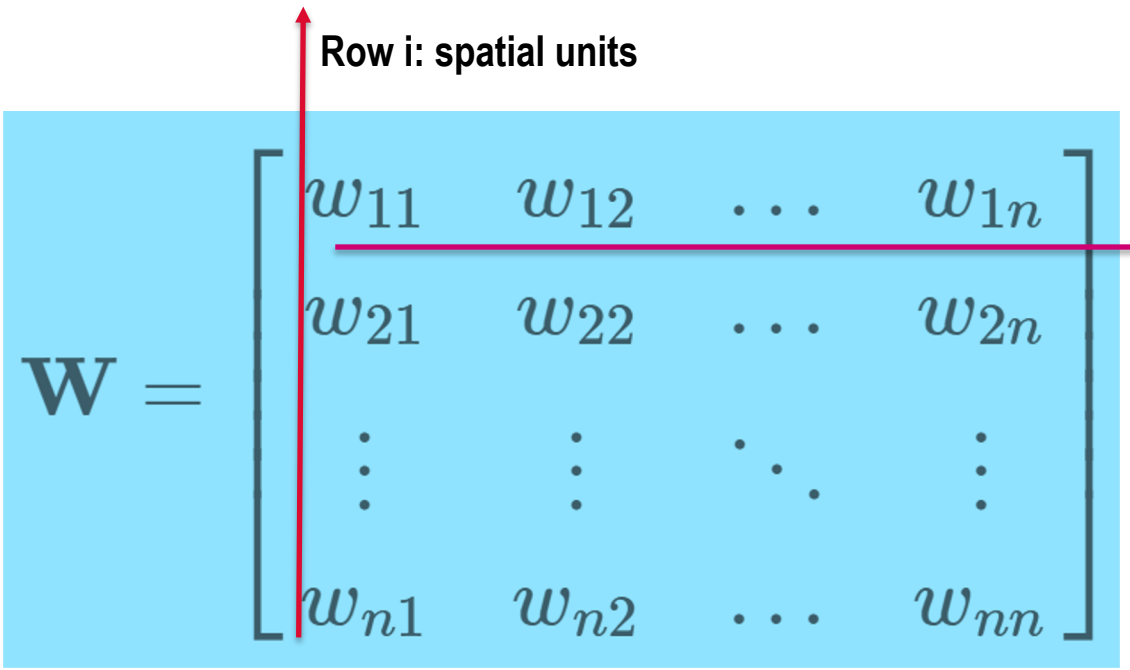
high positive spatial autocorrelation

QUESTION

- Assume that a grey cell = 5 and a white cell = 0.
- Which one has positive autocorrelation?



SPATIAL WEIGHTS



-Simplest form: Is there a neighbor relation? YES = 1 / NO = 0

-Spatial weights are sparse matrices…

SPATIAL WEIGHTS

- Contiguity-based weights
  - Units share a common border

    Rook criterion→common edge

    Queen criterion→common edge OR vertex
- Distance-based Spatial Weights
  - The further away the smaller the influence

    Inverse distance weight OR kernel weight

    Euclidean distance (Cartesian projection)

    Arc distance (geographic)

    Distance-band weights thresholds

SPATIAL LAG

- Spatially lagged variable is a weighted sum or a weighted average of the neighbouring values for that variable

$$[Wy]_i = w_{i1}y_1 + w_{i2}y_2 + \cdots + w_{in}y_n,$$

$$[Wy]_i = \sum_{j=1}^{n} w_{ij}y_j,$$

EXPLORATORY SPATIO-TEMPORAL DATA ANALYSIS

- ESTDA
  - Much less developed discipline
  - Complexity of simultaneously exploring the spatial and temporal dimensions of the data

→ An opportunity for your future career!!