Introduction to Exploratory (Spatial) Data Analysis - Summary

Overview

Exploratory Data Analysis (EDA) and its application in spatial data. The presentation covers the fundamentals of EDA, the importance of EDA before modelling, and the specific techniques used in spatial data analysis.

Learning Objectives

The primary learning objectives of the lesson are:

- Explaining the fundamentals and importance of EDA to peers.
- Applying statistical and visualization methods to different types of data.
- Developing familiarity with Python for data analysis tasks.

Data Analysis Workflow

The presentation outlines a typical data analysis workflow, starting from data preparation, which involves ingesting and cleaning data, followed by EDA to summarize data characteristics using statistical numbers and visualizations.

Data Ingestion and Cleaning

Data ingestion involves reading data from various formats using Python libraries such as:

- pandas.read_csv() for CSV files.
- pandas.read_excel() for Excel files.
- scipy.io.loadmat() for MATLAB files.
- geopandas.read_file() for shapefiles and GeoJSON files.
- rasterio.open() for GeoTIFF files.
- matplotlib.pyplot.imread() for images.

Data cleaning is emphasized as a crucial step to transform messy data into tidy data suitable for modeling.

Exploratory Data Analysis (EDA)

EDA is described as a method to summarize data characteristics with statistical measures and visualizations. Key benefits of EDA include:

- Providing an overview of the data.
- Guiding further analysis and method selection.
- Generating hypotheses.
- Identifying data problems.
- Understanding variable properties and relationships.

Statistical Analysis and Visualization

The presentation highlights the importance of combining statistical analysis with visualization to maximize data insights and uncover underlying structures. Examples include:

- Histograms and Probability Density Functions (PDFs) for univariate analysis.
- Box plots for summarizing data distributions.
- Bar plots for categorical data.

Bi-Variate Analysis

Bi-variate analysis techniques are discussed to understand relationships between two variables. Methods include:

- Correlation analysis to quantify relationships.
- 2-D scatter plots to visualize linear relationships.
- Pair plots to show pairwise relationships and identify patterns and outliers.

Exploratory Spatial Data Analysis (ESDA)

ESDA applies traditional EDA techniques to spatial datasets, connecting variables to specific locations or times and considering spatial autocorrelation. Key concepts include:

- Spatial autocorrelation: Describing how variable values are correlated across space.
    - Positive spatial autocorrelation: Similar values cluster together.
    - Zero spatial autocorrelation: Random distribution of values.
    - Negative spatial autocorrelation: Dissimilar values disperse.

Visualization Techniques in ESDA

Several ESDA mapping techniques are introduced, including:

- Box maps to identify outliers and visualize data distribution.
- Connection maps to show spatial relationships.
- Various advanced mapping methods like conditional choropleth maps and Voronoi diagrams.

Case Study: Ghelgheli’s Teahouse

The presentation includes a team-based learning assignment involving a hypothetical scenario where Ghelgheli, a tea lover, uses data analysis to find a suitable location for his teahouse. The case study emphasizes:

- Data collection on potential locations, foot traffic, competitor locations, rent prices, and demographics.
- Data cleaning and imputation to handle missing and anomalous values.
- Statistical analysis to extract descriptive statistics.

- Visualization techniques to identify patterns and trends.
- Decision-making based on data insights to select the best location.

Key Takeaways

The document concludes with several important lessons:

- The critical role of data in decision-making processes.
- The effectiveness of EDA techniques in uncovering insights.
- The transformation of messy data into valuable insights through proper cleaning and analysis.

This comprehensive presentation provides a solid foundation for understanding and applying EDA and ESDA techniques in various data analysis scenarios.

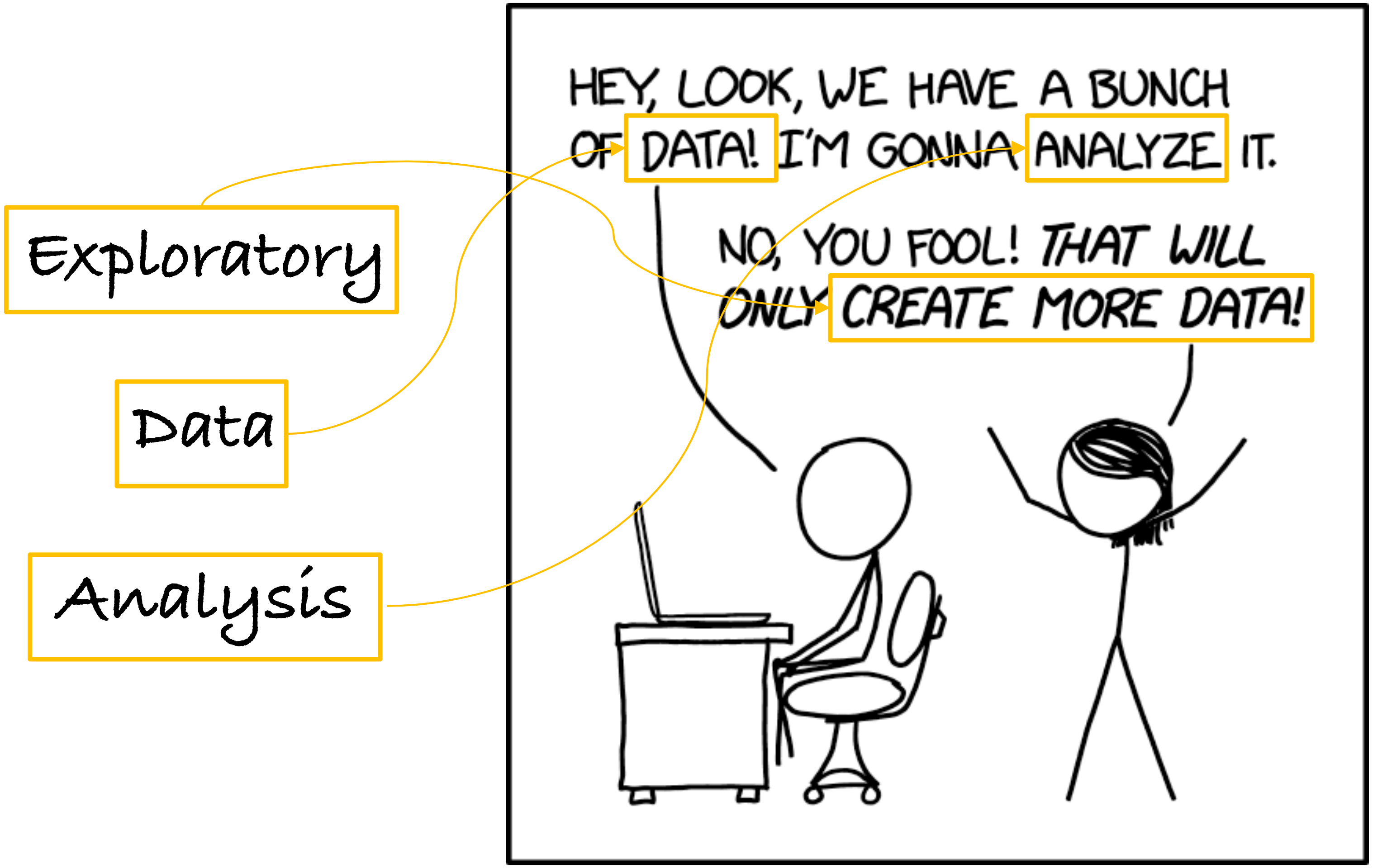
UNIVERSITY OF TWENTE.

Introduction to Exploratory (Spatial) Data Analysis

Mahdi KHODADADZADEH
Assistant Professor
Faculty of Geo-Information Science and Earth Observation (ITC)
Department of Geo-information Processing (GIP)
m.khodadadzadeh@utwente.nl

May 2024

ITC FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION



Exploratory

Data

Analysis

HEY, LOOK, WE HAVE A BUNCH OF DATA! I'M GONNA ANALYZE IT.
NO, YOU FOOL! THAT WILL ONLY CREATE MORE DATA!

From: https://xkcd.com

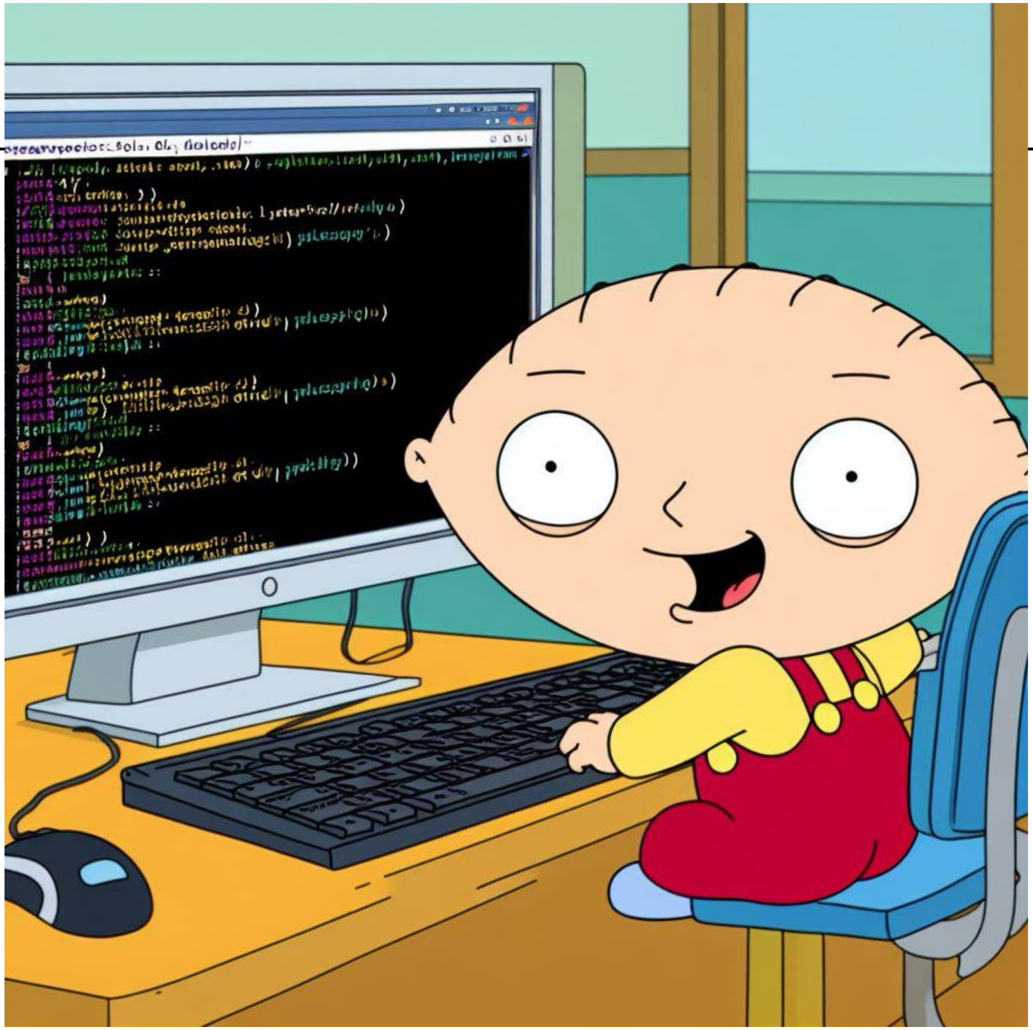
This lesson’s learning objectives

Explain to peers

- the fundamentals of E(S)DA
- the importance of E(S)DA before modelling

Apply statistical and visualization methods on different types of data

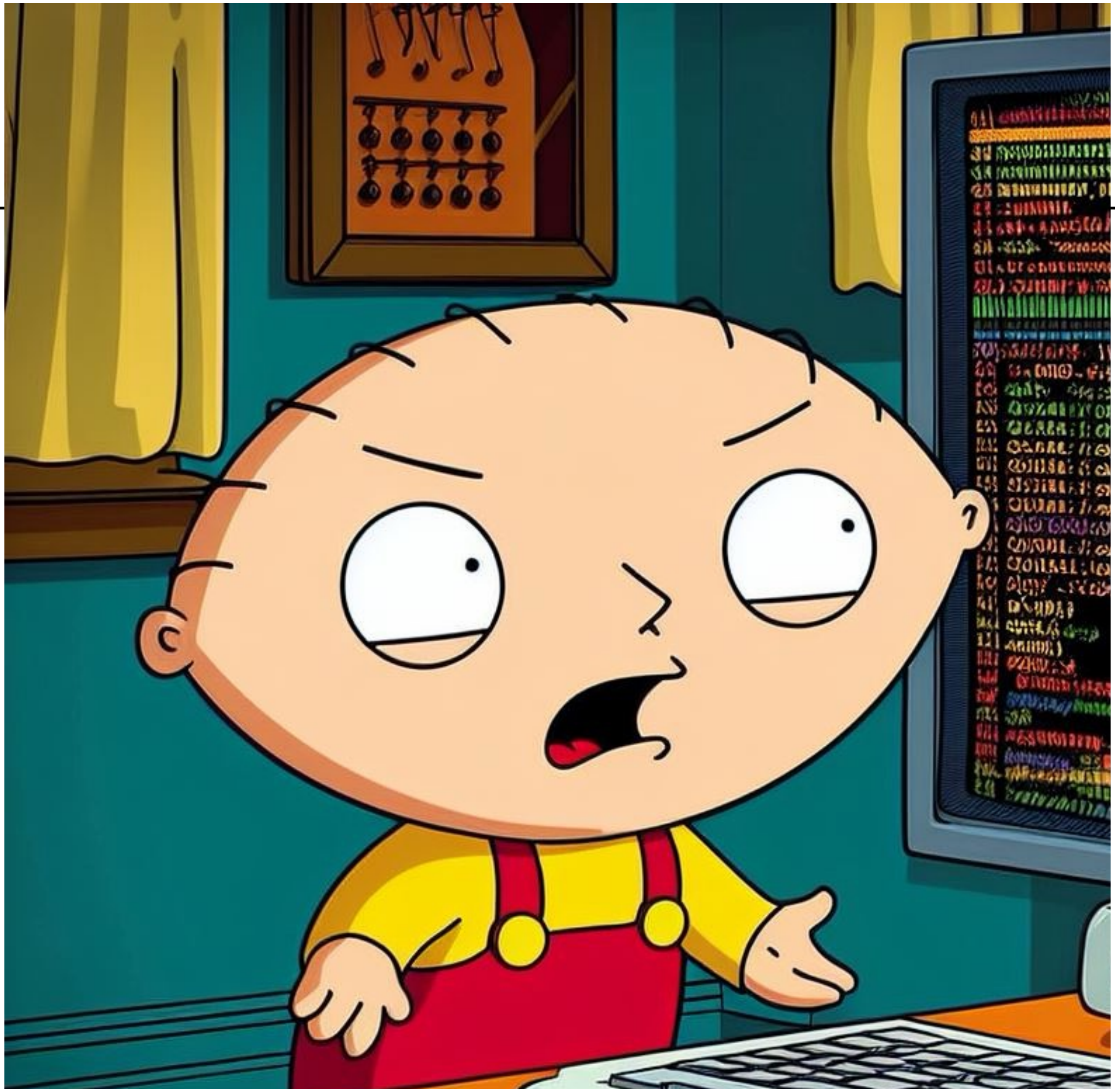
Develop familiarity with Python



You are a Python master. Congrats!

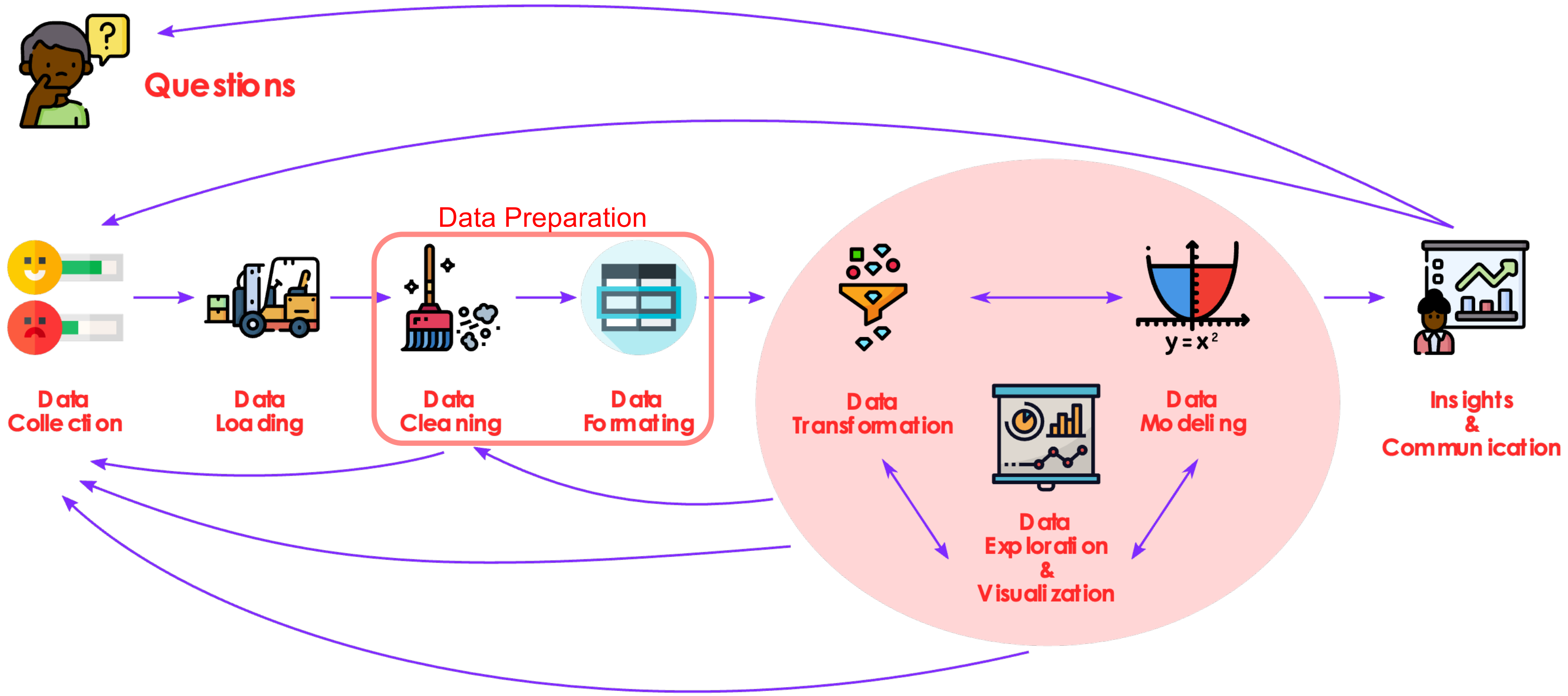
Model = Algorithm_X(Data)

You've learned how to build a model in Python. Congrats!



But you run into some issues!

Data Analysis Workflow



From: https://davpy.netlify.app/3-data-workflow.html

Ingesting Data

Getting data in a shape that we can use to start our analysis.

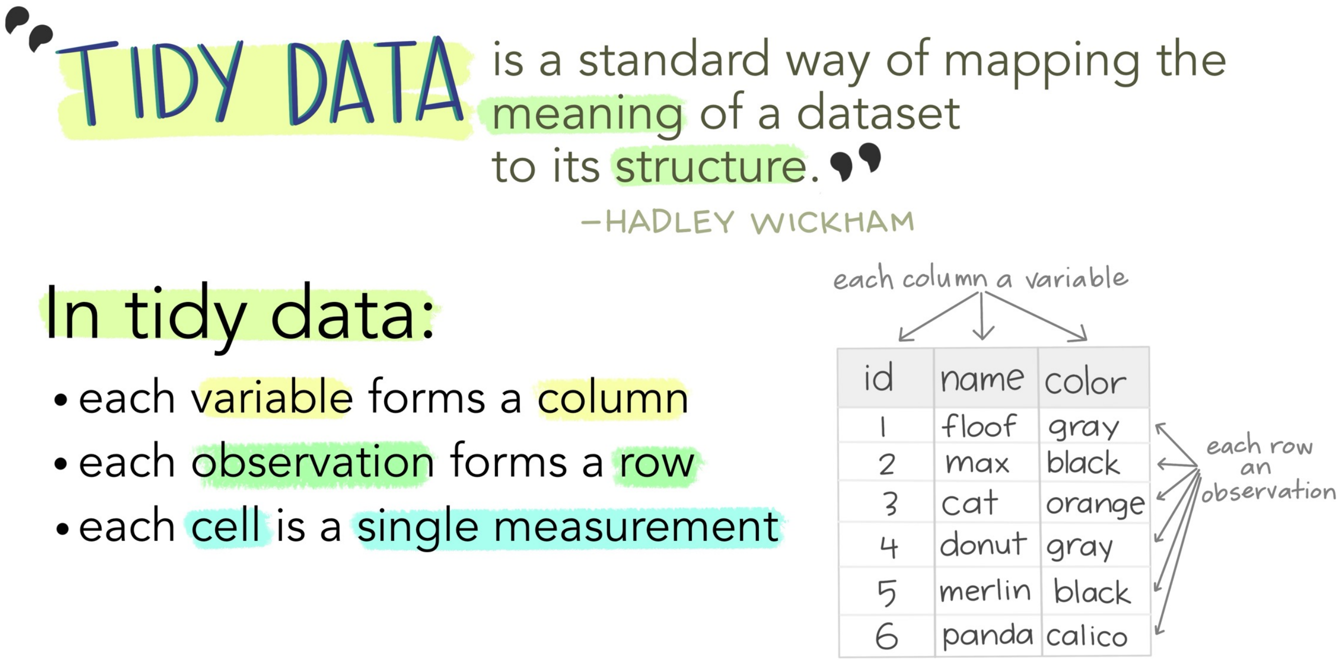
Python:

- 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: rasterio.open()
- Reading an image: matplotlib.pyplot.imread()

Data Cleaning

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

id	name	color
1	floof	gray
2	max	black
3	cat	orange
4	donut	gray
5	merlin	black
6	panda	calico

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

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

Exploratory Data Analysis (EDA)

EDA aims at **summarizing** the characteristics of a dataset with **statistical numbers** and **graphs**

Statistical Analysis + Visualization

Get an overview of the data

Orient further analysis → choose correct methods/approaches

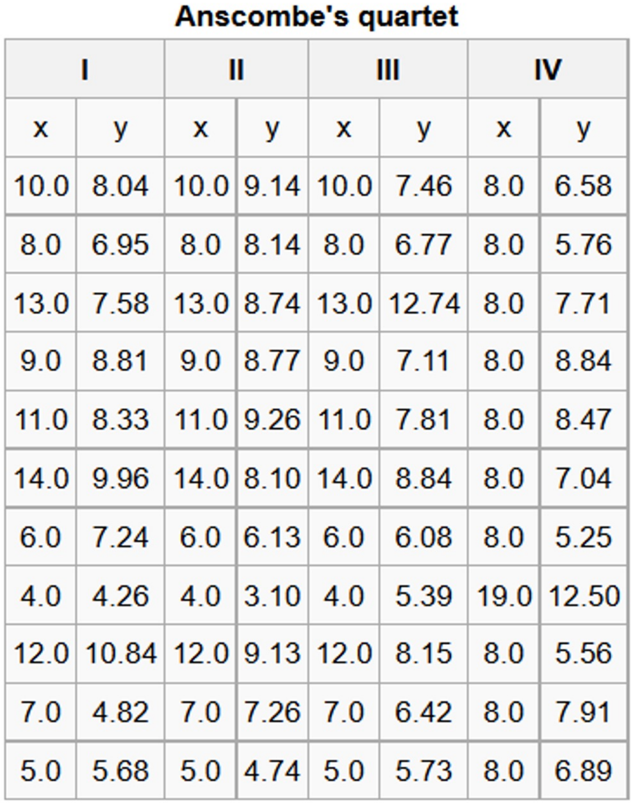
Help you to generate hypothesis

Spot problems in data

Understand properties of the variables (e.g., mean)

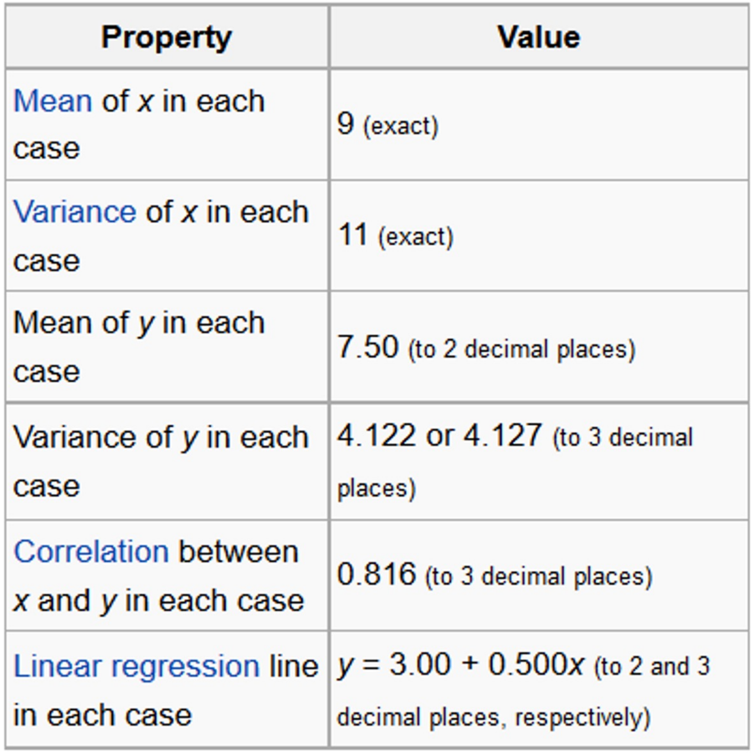
Understand relationships between variables

Statistics + Visualization



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



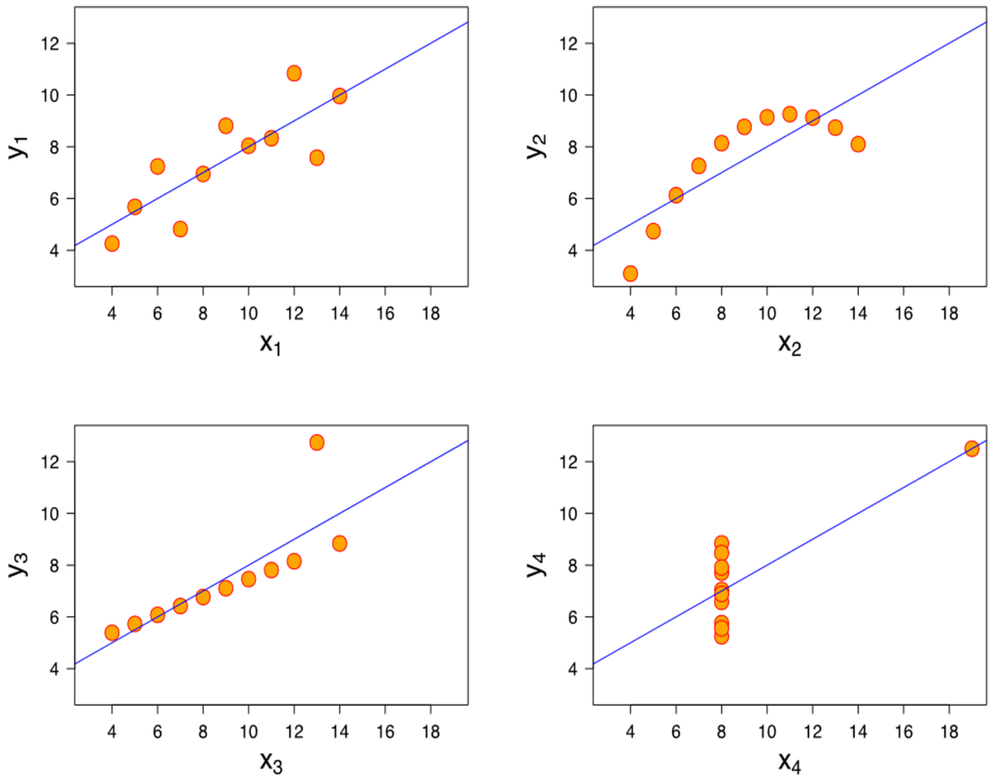
Property	Value
Mean of *x* in each case	9 (exact)
Variance of *x* in each case	11 (exact)
Mean of *y* in each case	7.50 (to 2 decimal places)
Variance of *y* in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between *x* and *y* in each case	0.816 (to 3 decimal places)
Linear regression line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

From: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Statistics + Visualization

Visualization

- Maximize insight into a data set
- Uncover underlying structure



From: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Univariate Analysis

Mean and Standard Deviation

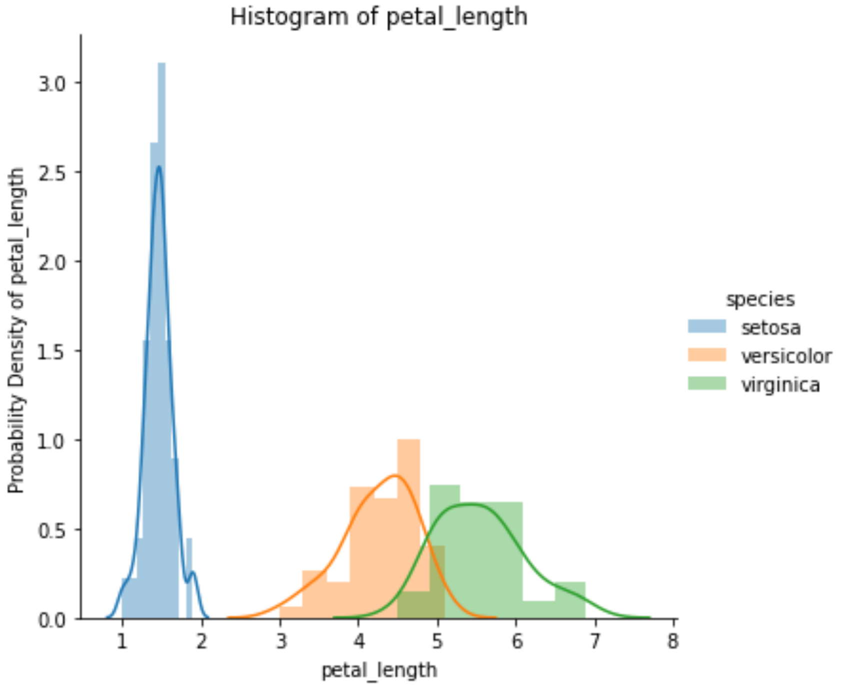
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

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

Histogram and PDF

distribution of the data, showing the number of observations that fall within each bin.

PDF is the continuous version of the histogram



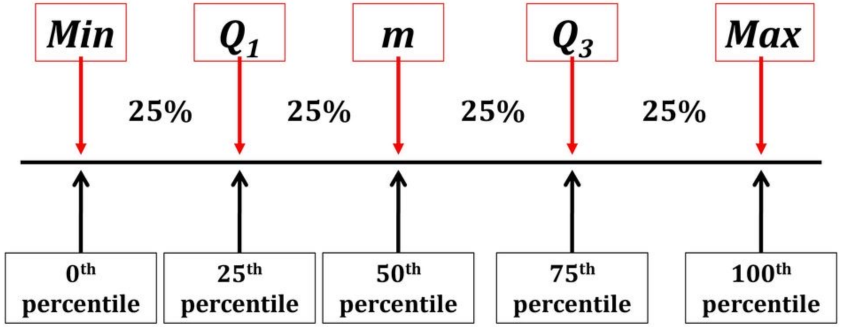
Univariate Analysis

Min, Max, Median, Percentile, Quartile

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.

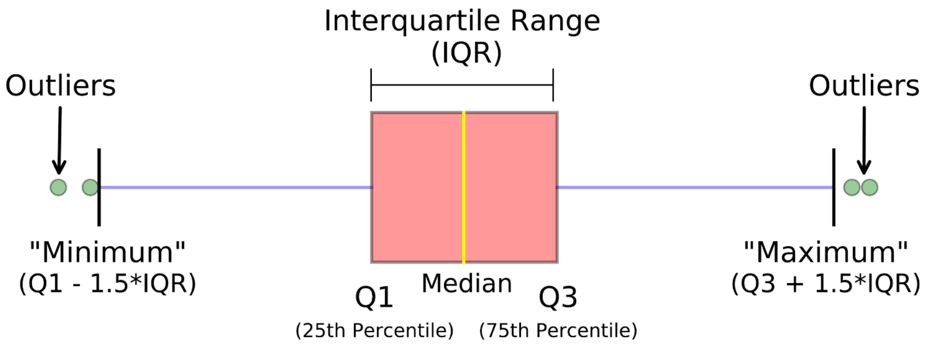
Quartile: The q-th quantile of V is the value q of the way from the minimum to the maximum in a sorted copy of V.

five-number summary →



Univariate Analysis

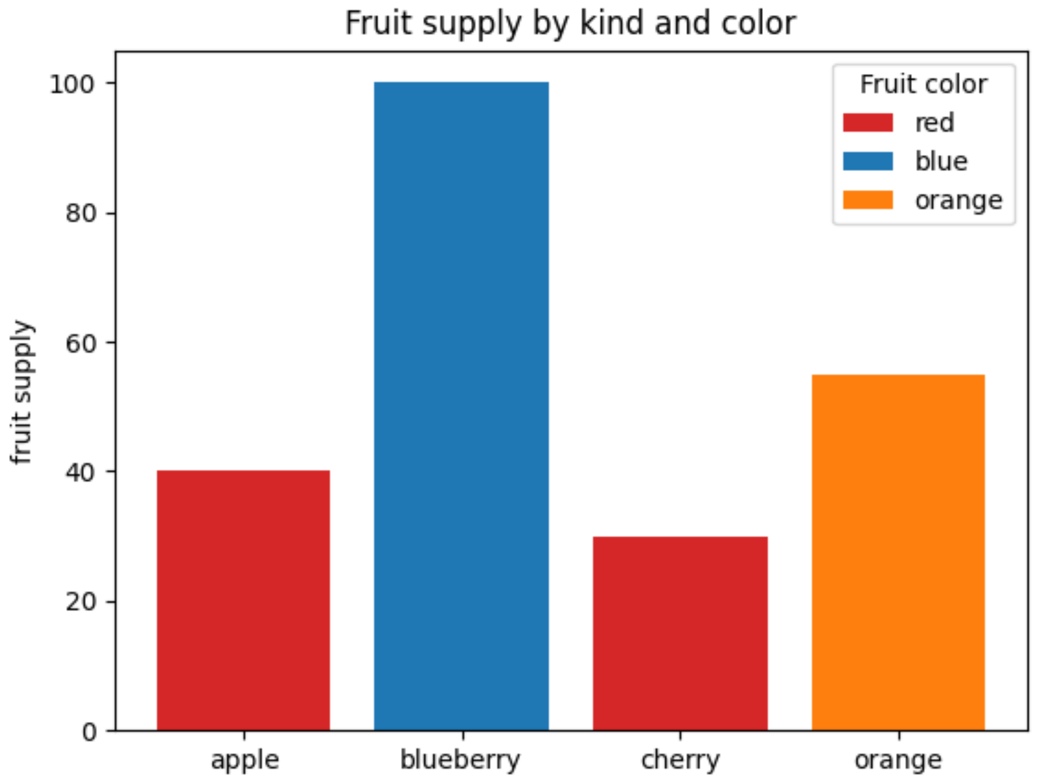
Box plot: displays the five-number summary (the minimum, first quartile, median, third quartile, and maximum) of a set of data. It can tell you about your outliers and what their values are



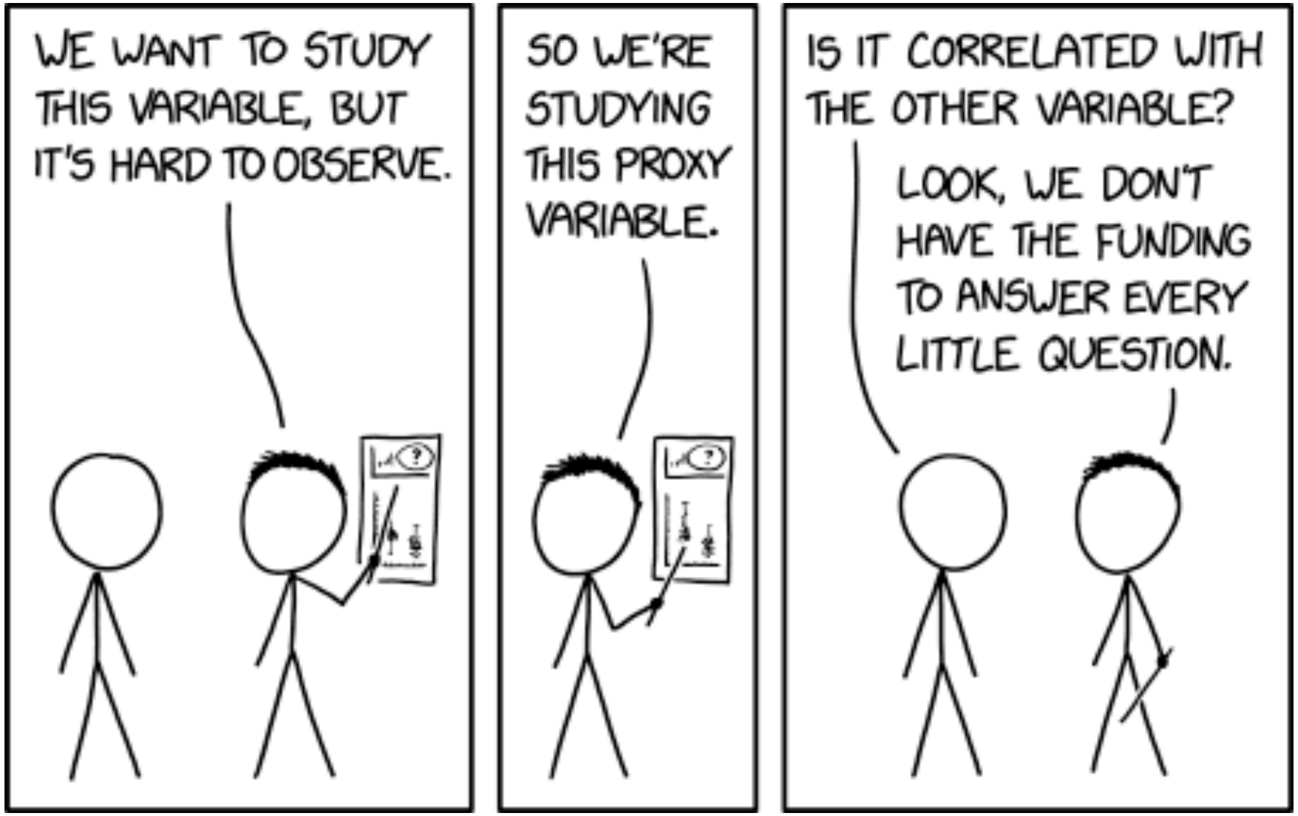
https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Univariate Analysis

Bar plots



From: https://matplotlib.org/


WE WANT TO STUDY THIS VARIABLE, BUT IT'S HARD TO OBSERVE.
SO WE'RE STUDYING THIS PROXY VARIABLE.
IS IT CORRELATED WITH THE OTHER VARIABLE?
LOOK, WE DON'T HAVE THE FUNDING TO ANSWER EVERY LITTLE QUESTION.

From: https://xkcd.com

Bi-Variate Analysis

Correlation

Relationship between two variables quantitatively

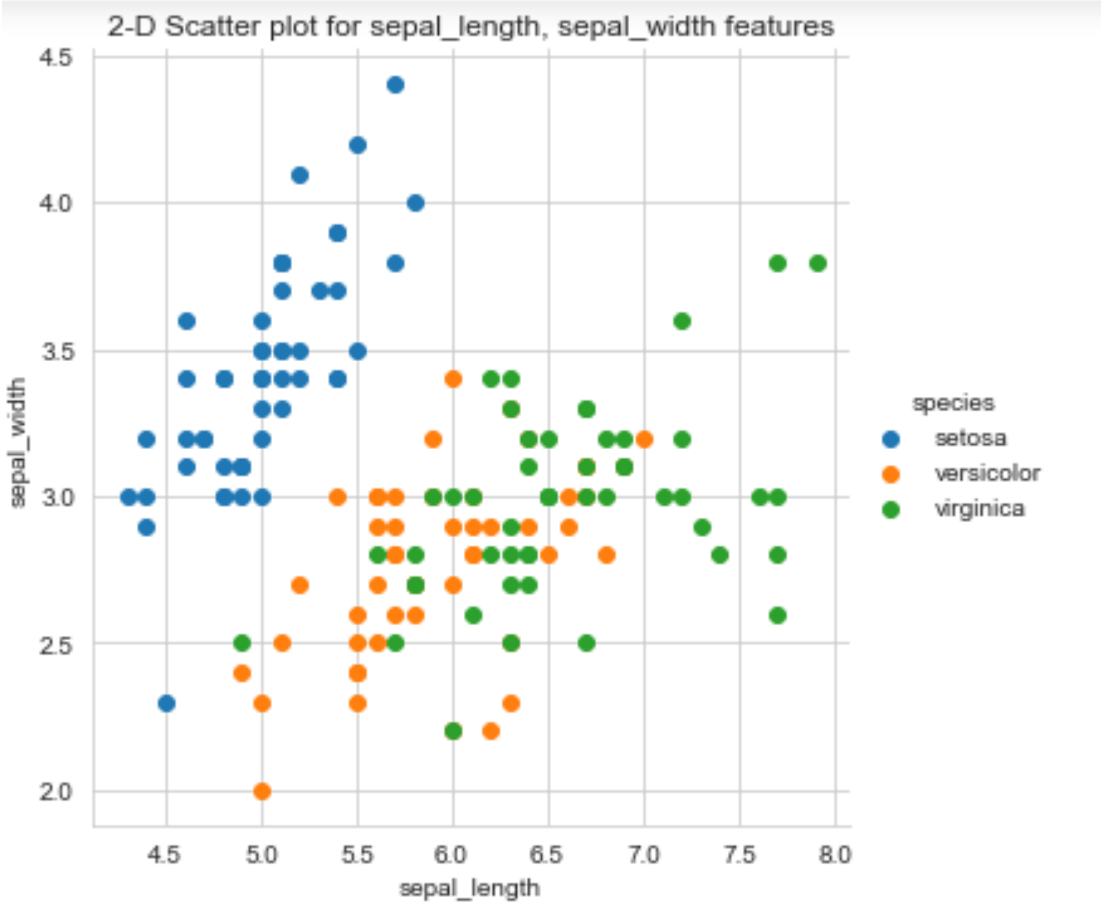
$$cor(x, y) = \frac{cov(x, y)}{sd(x)sd(y)}$$

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

Bi-Variate Analysis

2-D Scatter Plots

They can show the linear relationship between two variables



Bi-Variate Analysis

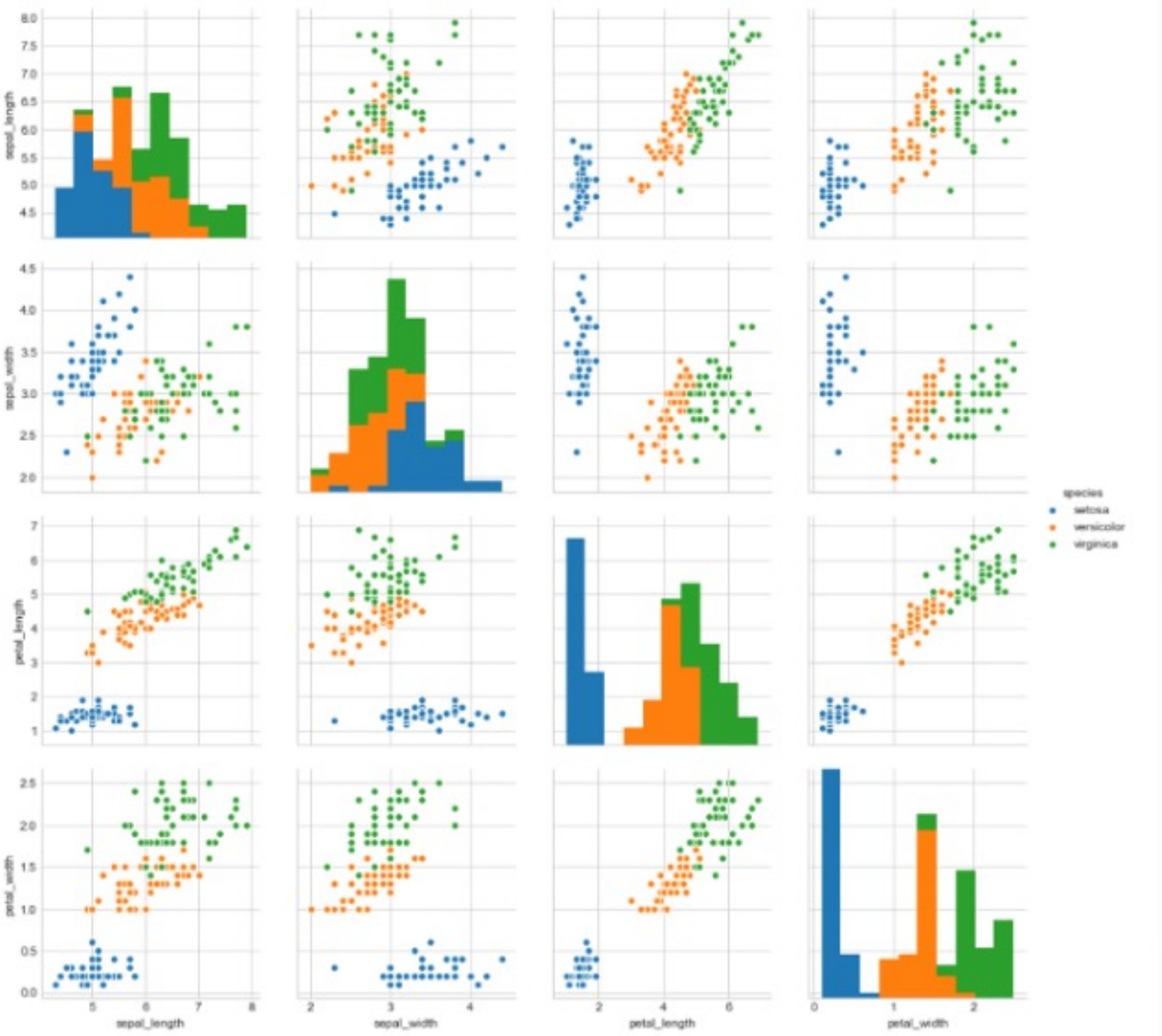
Pair-plot

Note: -

A pair plot is a visualization that shows pairwise relationships between variables in a dataset. It’s a great way to explore how different variables correlate with each other.

What is a Pair Plot?

A pair plot displays scatterplots, histograms, or kernel density estimates for each variable pair in your dataset.
It’s useful for identifying patterns, correlations, and potential outliers.



Exploratory Spatial Data Analysis

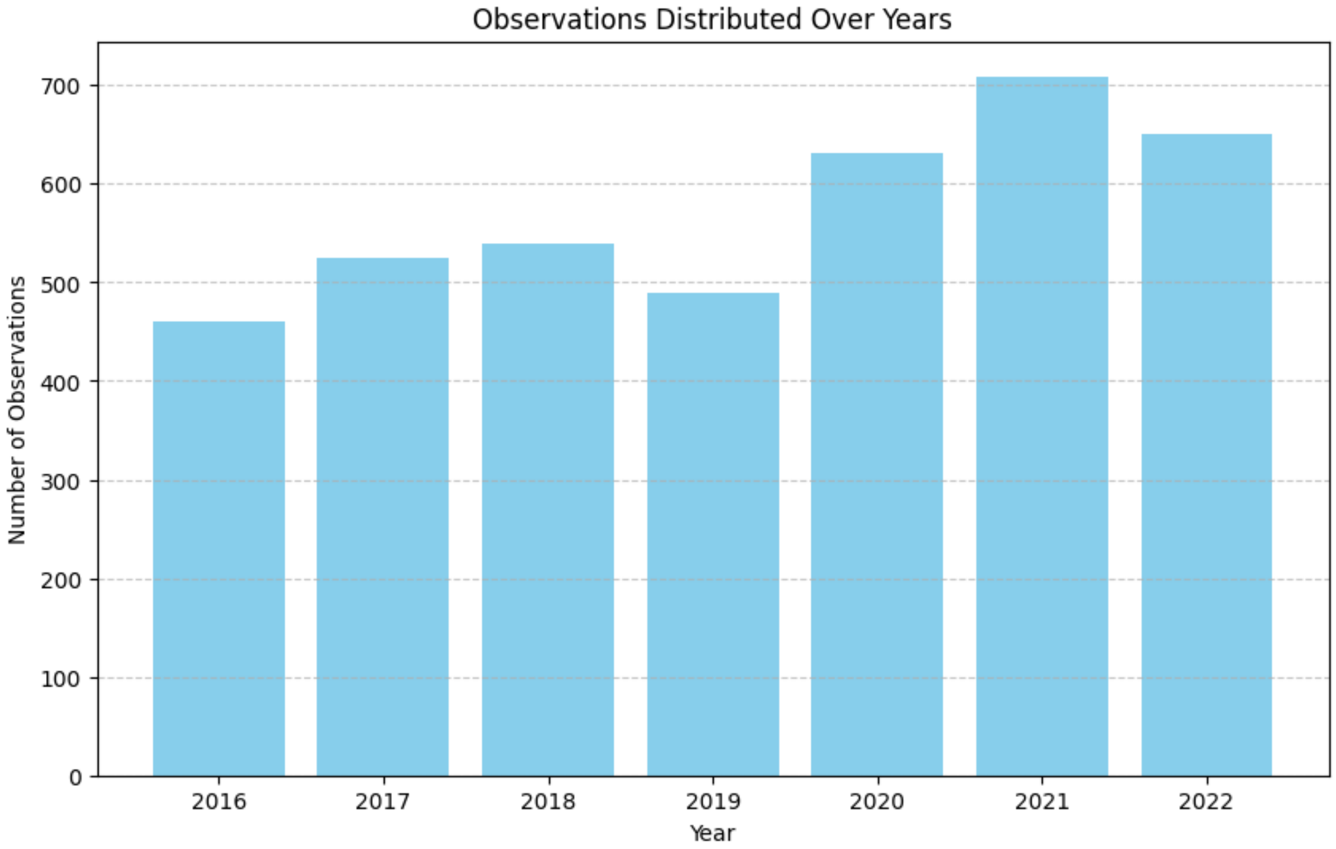
Geospatial data → ESDA

“Traditional” EDA can be applied to spatial datasets for obtaining statistics and basic plots (barplot, histograms, boxplots,..).

ESDA tools connects a specific variable to a location/time

It takes into account the values of the same variable in different locations/time.

Applying EDA to geospatial data



Spatial autocorrelation

Correlation of a variable with itself across space (in different places in space) → relationships to neighbors

Positive spatial autocorrelation

- values are similar to their neighbors or other close objects
- **clusters** of similar values on the map

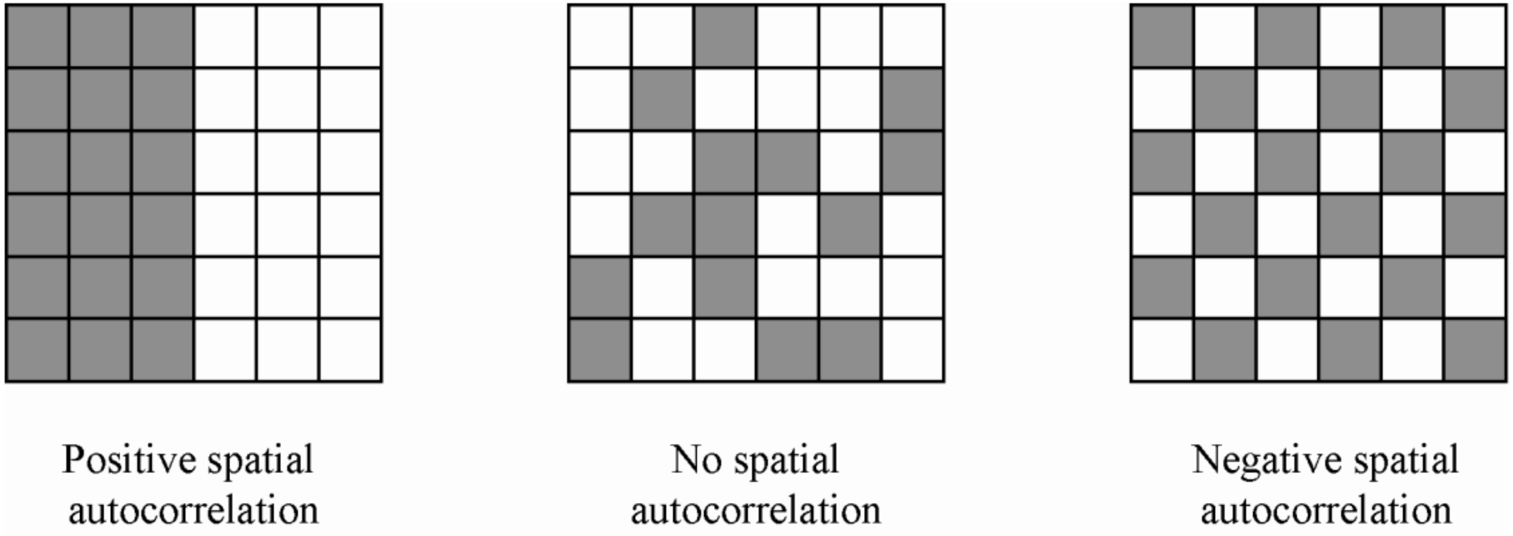
Zero or no spatial autocorrelation

- **random** values of close objects or neighbors
- no clear pattern visually

Negative spatial autocorrelation

- values are dissimilar to their neighbors or close objects
- **dispersed** patterns of values on the map

Spatial autocorrelation



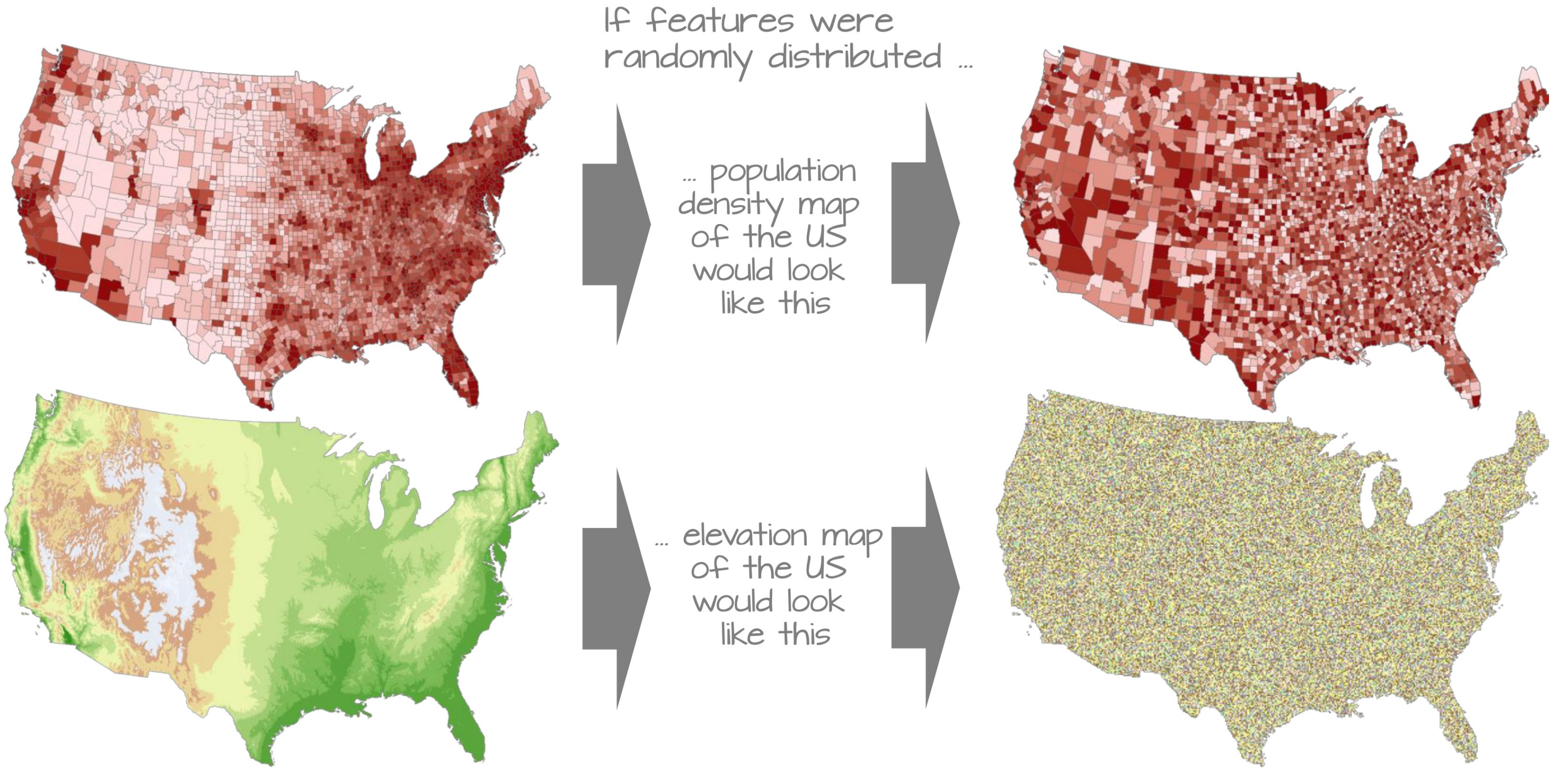
Positive spatial autocorrelation

No spatial autocorrelation

Negative spatial autocorrelation

From: (Radil, 2011)

Spatial autocorrelation



If features were randomly distributed …

… population density map of the US would look like this

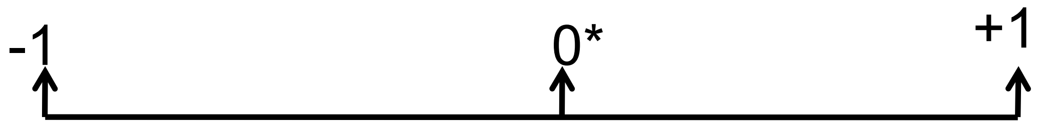
… elevation map of the US would look like this

From: https://mgimond.github.io/Spatial/spatial-autocorrelation.html

SPATIAL 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*

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



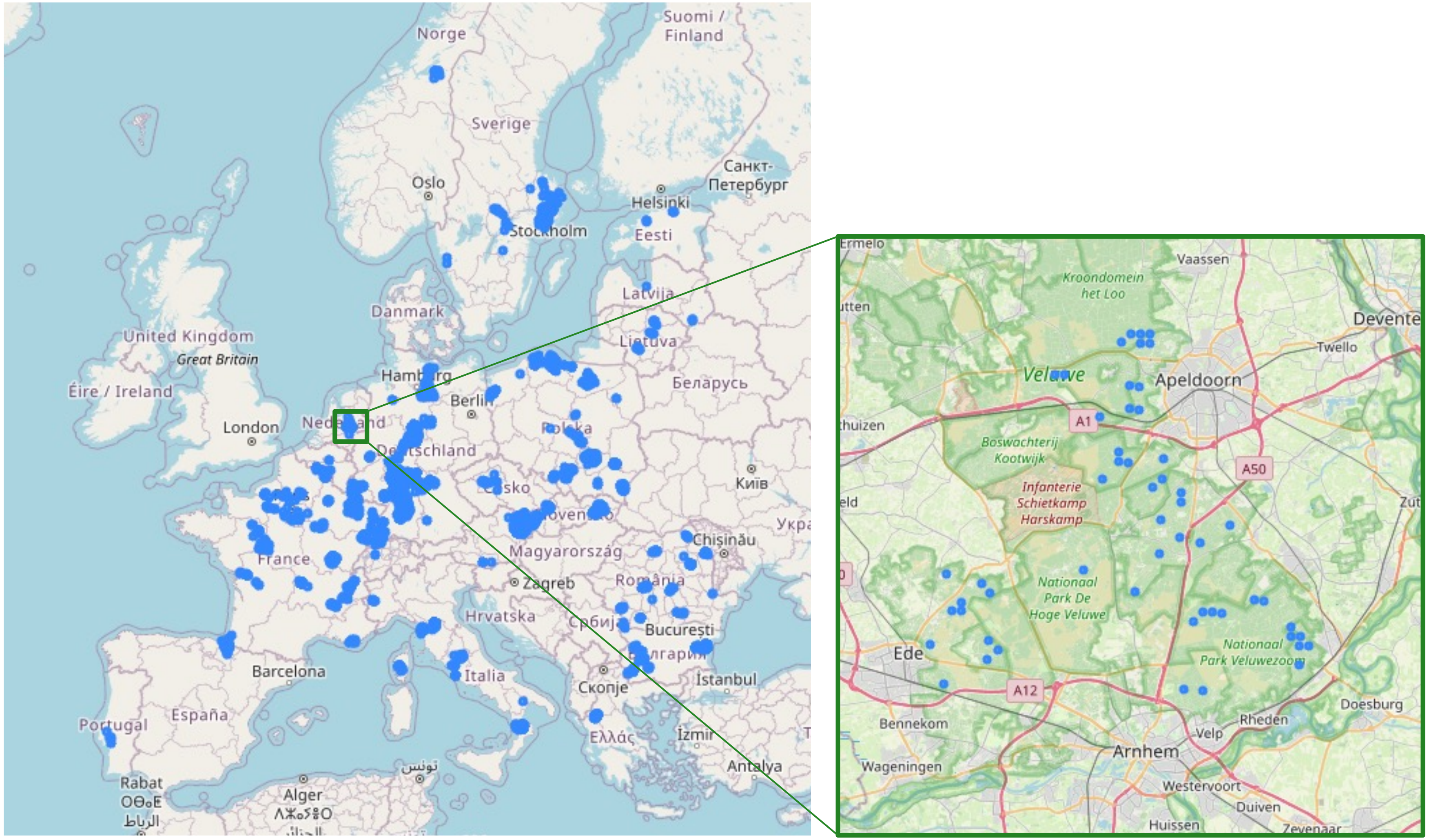
high negative spatial autocorrelation

*no spatial autocorrelation**

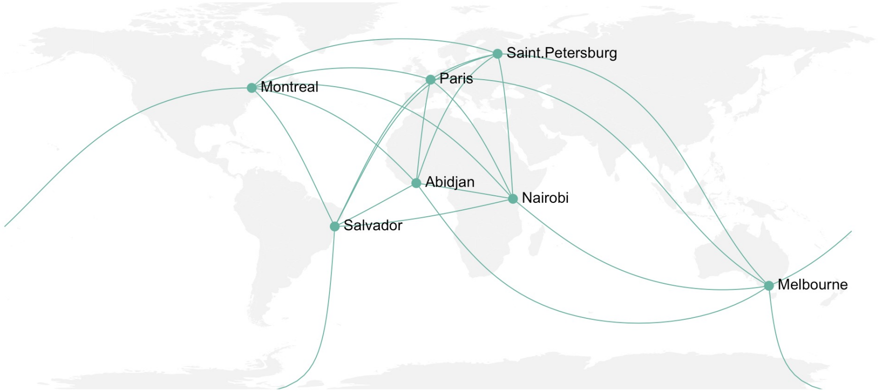
high positive spatial autocorrelation

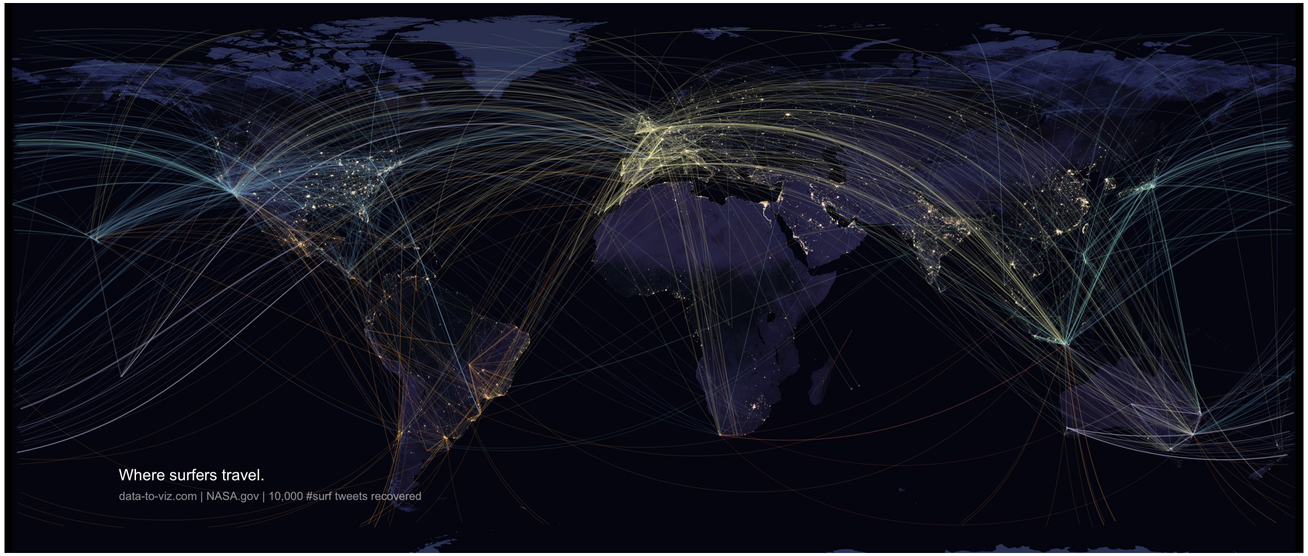
Check out the link below for more in-depth explanation:
https://rpubs.com/corey_sparks/105700

Visualization on map



Connection map





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

Box map

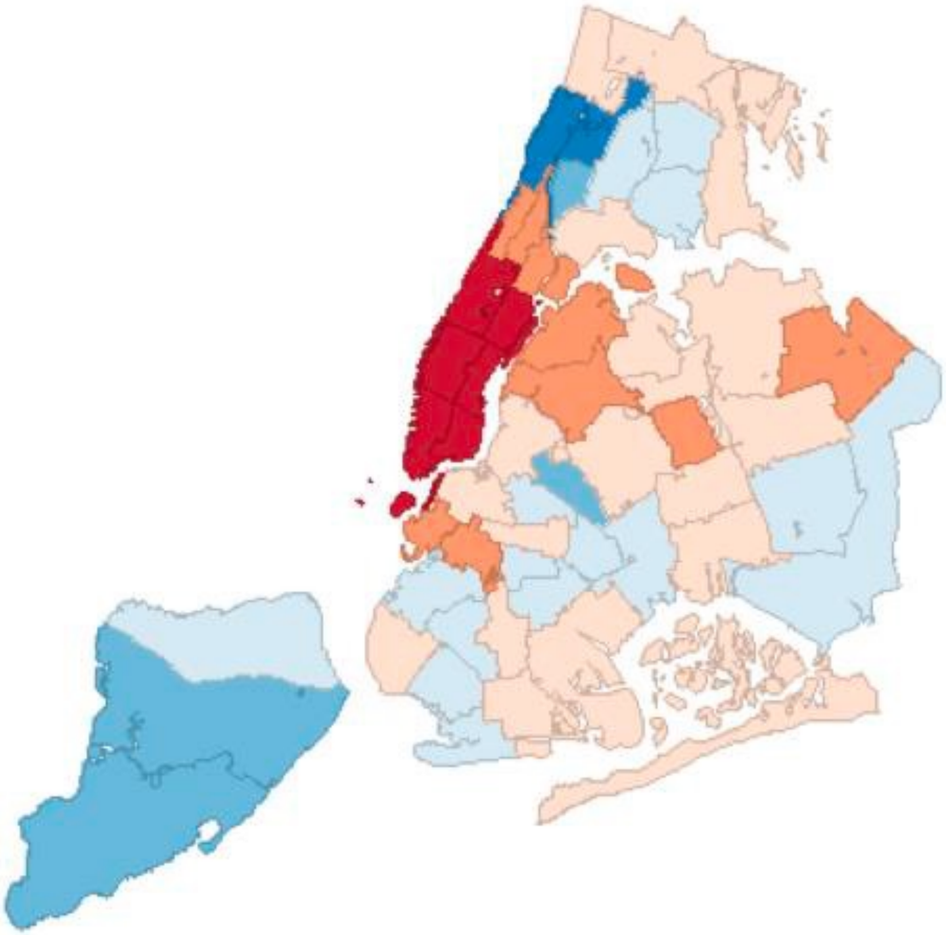
Hinge=1.5: rent2008

- 🟦 Lower outlier (3) [0 : 456.250]
- 🟦 < 25% (4) [456.250 : 1000]
- ⬜ 25% - 50% (15) [1000 : 1100]
- ⬜ 50% - 75% (19) [1100 : 1362.500]
- 🟧 > 75% (8) [1362.500 : 1906.250]
- 🟥 Upper outlier (6) [1906.250 : inf]

Note: -

A box map (Anselin 1994) is the mapping counterpart of the idea behind a box plot. The point of departure is again a quantile map, more specifically, a quartile map. But the four categories are extended to six bins, to separately identify the lower and upper outliers. The definition of outliers is a function of a multiple of the inter-quartile range (IQR), the difference between the values for the 75 and 25 percentile. As we will see in a later chapter in our discussion of the box plot, we use two options for these cut-off values, or hinges, 1.5 and 3.0. The box map uses the same convention.

The box map in Figure separates the three lower outliers (the observations with zero values) from the other four observations in the first quartile. They are depicted in dark blue. Similarly, it separates the six outliers in Manhattan from the eight other observations in the upper quartile. The upper outliers are colored dark red.



ESDA maps

Some 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

UNIVERSITY OF TWENTE.

Team Based Learning

FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION

Team based learning assignment

Ghelgheli decided to change his job, and as a tea lover, he opted to open a teahouse. He aimed to find the right location for his business, where many people were passing by and not many competitors around.

Ghelgheli started by collecting data, organizing it into rows and columns within a table on his computer. However, the data was somewhat messy, containing several missing values and even some anomalies. Nevertheless, Ghelgheli was enthusiastic about working with such a dataset. He used some cool techniques to clean the data, extract statistical measures, and generate plots and maps.

Through his analysis, Ghelgheli pinpointed a suitable location for his teahouse, and soon after opening, it became a local favorite.

Which data and methods do you think Ghelgheli utilized for his analysis?

What interesting learnings did you derive from Ghelgheli's story?

Can you provide some real-life examples similar to Ghelgheli's experience?

Data Collection: Ghelgheli started by collecting data on potential locations for his teahouse. This could include foot traffic data, competitor locations, rent prices, demographic information of the area, etc.

Data Cleaning: The data Ghelgheli collected was described as messy, with missing and strange values. Ghelgheli likely employed techniques like data imputation, outlier detection, and data validation to clean the dataset.

Statistical Analysis: Ghelgheli extracted statistical measures from the cleaned dataset. This could involve calculating means, medians, standard deviations, and other descriptive statistics to understand the characteristics of the data.

Visualization: Ghelgheli created plots and maps to visualize the data. This could include scatter plots, histograms, heatmaps, and geographical maps to identify patterns and trends in the data.

Decision Making: Through the analysis, Ghelgheli identified a suitable location for his teahouse based on the insights gained from the data analysis.

- The importance of data in decision-making processes
- The power of EDA techniques in uncovering insights and making informed decisions.
- How messy data can be transformed into valuable insights through proper cleaning and analysis.