AIML231/DATA302 — Week 04

Exploratory Data Analysis

Dr Bach Hoai Nguyen

School of Engineering and Computer Science

Victoria University of Wellington

Bach.Nguyen@vuw.ac.nz

Office Hour:1-2pm, Friday, Week 4-Week 7

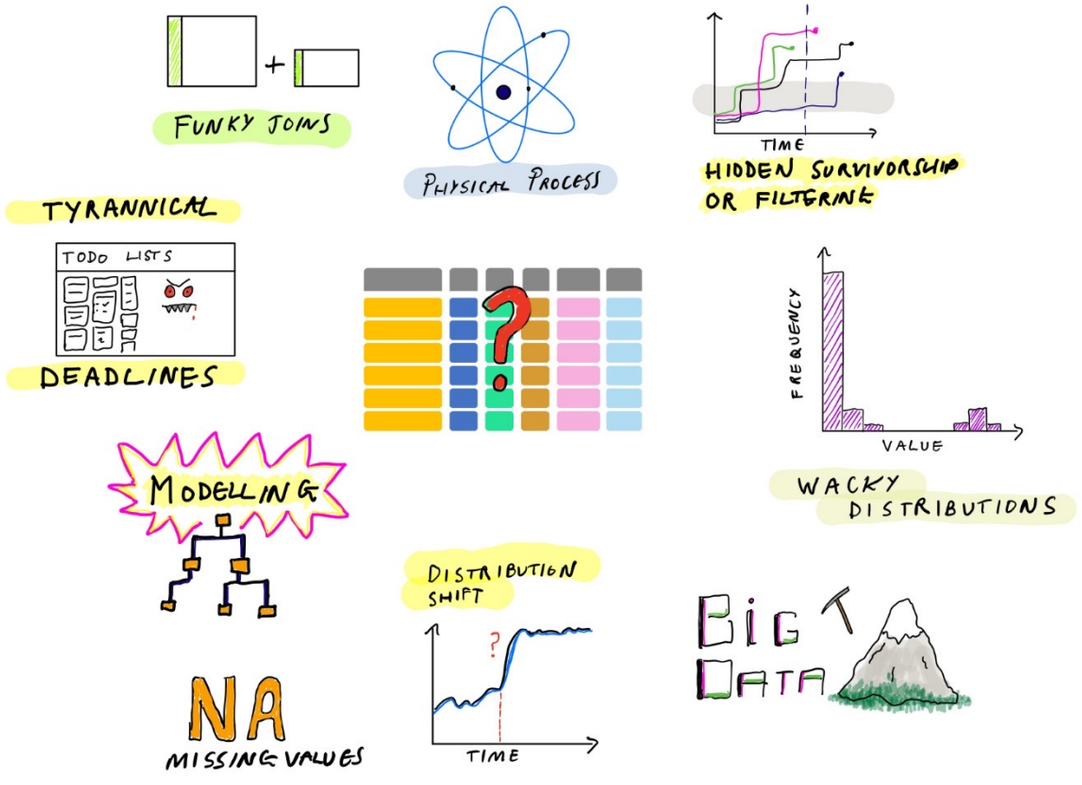
Room: CO364

Week Overview

- Exploratory Data Analysis (EDA) Introduction
  - What is EDA
  - Why need EDA, what can do with EDA
  - How to do EDA
- EDA Techniques
  - Groups of EDA Methods
  - Visualisation Methods
- EDA Tools
  - Python Modules for EDA
- EDA Case Studies

Thinking

- How to choose the most suitable algorithms for your dataset?

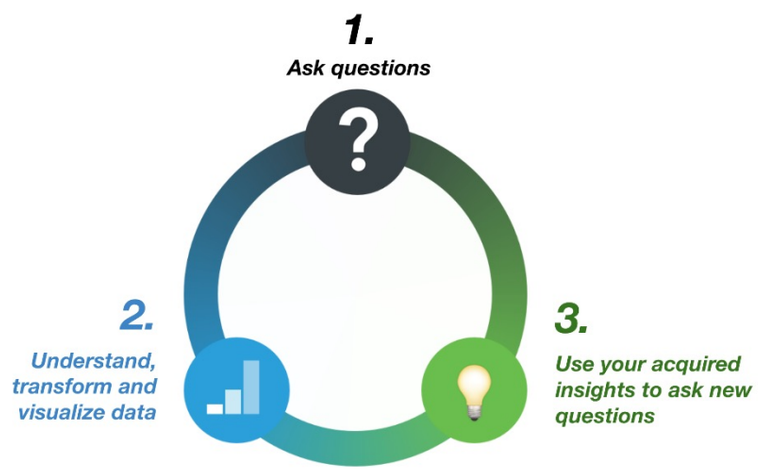


https://alastairrushworth.github.io/exploring_eda/EDA.html#1

- How to ensure you are ready to use machine learning techniques in a new project?
- Answer: Exploratory Data Analysis (EDA) helps to answer

Exploratory Data Analysis

- Exploratory Data Analysis (EDA) is a process for summarising, visualising, and becoming intimately familiar with the important characteristics of the data
- EDA is an iterative cycle:
  - Generate questions about your data:
  - Search for answers by visualising, transforming, and modelling your data
  - Use what you learn to refine your questions and/or generate new questions



- What type of variation occurs within my variables?
- What type of covariation occurs between my variables?

https://duo.com/labs/research/gamifying-data-science-education

- EDA is *statisticians' way of story telling* where you explore data, find patterns and tells insights

Key Concepts of Exploratory Data Analysis

- 4 Objectives of EDA
  - Discover Patterns
  - Spot Anomalies
  - Frame Hypothesis
  - Check Assumptions
- Stuff done during EDA
  - Measures of central tendency: mean, median, mode
  - Spread measurement : standard deviation, variance
  - Shape of distribution: distribution, trends
  - Outlier
  - Correlations
  - Visual Exploration

Making Sense of Data – Distinguish Types of Attributes

- input takes the form of instances and attributes/features
  - information to a machine learning learner takes the form of a set of instances
  - each instance is described by a fixed predefined set of features or attributes:
  - Types: Numerical (Discrete and Continuous) and Nominal (Categorical)

Model	Year	Engine Fuel Type	Engine HP	Engine Cylinders	Transmission Type	Driven_Wheels	Number of Doors	Market Category	Vehicle Size	Vehicle Style	highway MPG	city mpg	Popularity	MSRP
1 Series M	2011	premium unleaded (required)	335.0	6.0	MANUAL	rear wheel drive	2.0	Factory Tuner,Luxury,High-Performance	Compact	Coupe	26	19	3916	46135
1 Series	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,Performance	Compact	Convertible	28	19	3916	40650
1 Series	2011	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,High-Performance	Compact	Coupe	28	20	3916	36350
1 Series	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,Performance	Compact	Coupe	28	18	3916	29450
1 Series	2011	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	Luxury	Compact	Convertible	28	18	3916	34500
1 Series	2012	premium unleaded (required)	230.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,Performance	Compact	Coupe	28	18	3916	31200
1 Series	2012	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,Performance	Compact	Convertible	26	17	3916	44100
1 Series	2012	premium unleaded (required)	300.0	6.0	MANUAL	rear wheel drive	2.0	Luxury,High-Performance	Compact	Coupe	28	20	3916	39300

Types of EDA Methods

- EDA methods: generally classified into two ways
  - graphical or non-graphical/quantitative : summarising data in a visual way or calculation of summary statistics
  - univariate or multivariate: summary statistics for each feature/attribute or find relationship between features

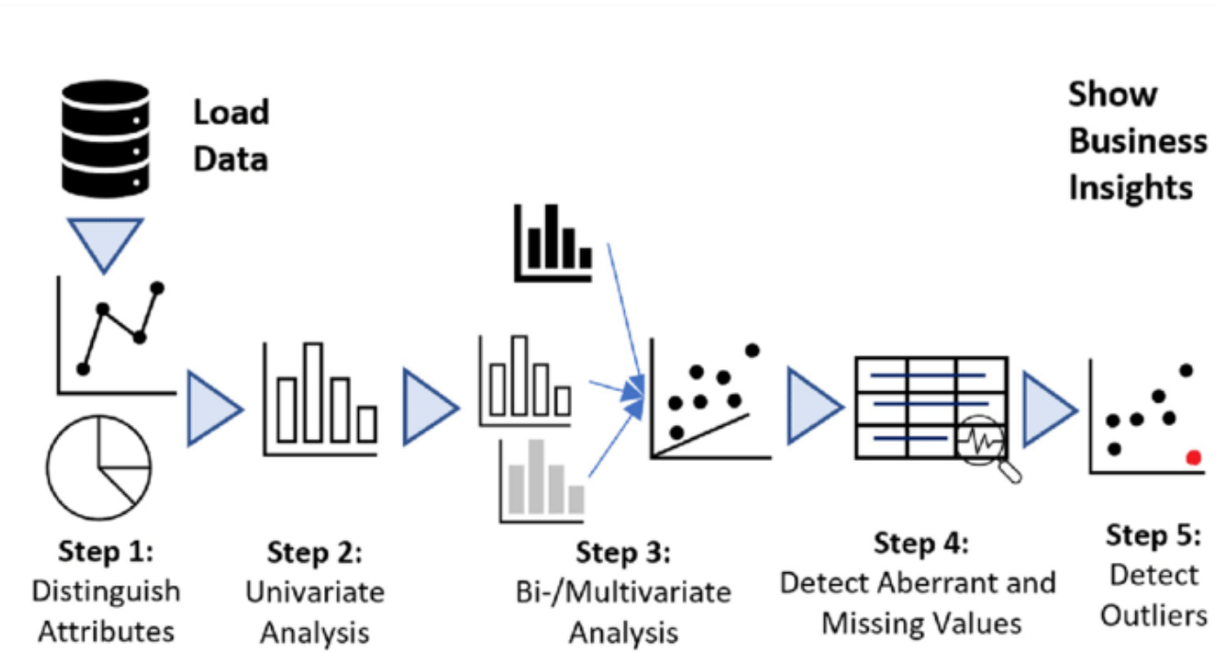
	Univariate	Multivariate
Non-Graphical	**Categorical Variable**: tabular representation of frequency **Quantitative Variable:** • Location (mean, median) • Shape and Spread • Modality • Outliers …	**One Categorical Variable and One Quantitative Variable**: Standard univariable non-graphical statistics for the quantitative variable separately for each level of the categorial variable **Two and more Quantitative Variable:** • Correlation, • Covariance, • …
Graphical	**Categorical Variable**: Bar Chart **Quantitative Variable:** • Histogram • Boxplot • …	**One Categorical Variable and One Quantitative Variable**: • Side-by-side Boxplots **Two and more Categorical Variable**: • Grouped Bar Chart **Two and more Quantitative Variable:** • Scatter plot, Correlation Heatmap, Pairplot **…**

How to do EDA

Steps/activities involved in EDA:

- identification of variables and data types
- non-graphical and graphical univariate analysis
- bi-/multivariate analysis, correlation analysis
- detect missing values and anomalies
- detect outliers

A typical example:

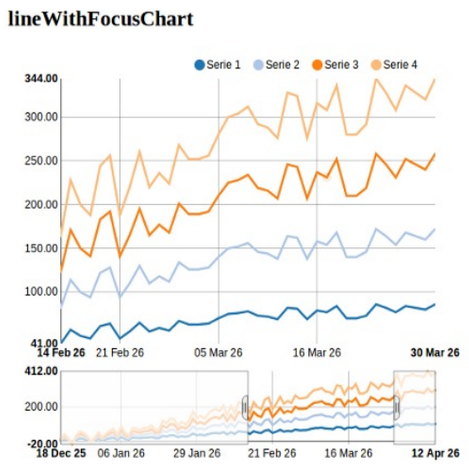


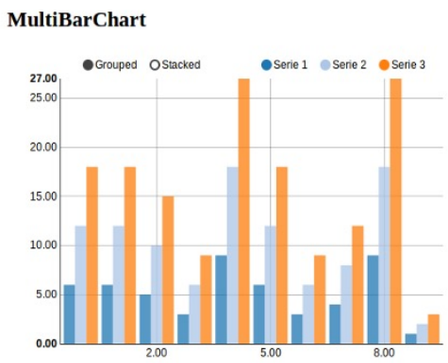
https://www.researchgate.net/publication/342282008_Exploratory_Data_Analysis_and_Data_Envelopment_Analysis_of_Construction_and_Demolition_Waste_Management_in_the_European_Economic_Area/figures?lo=1

Visualisation

There are four basic presentation types

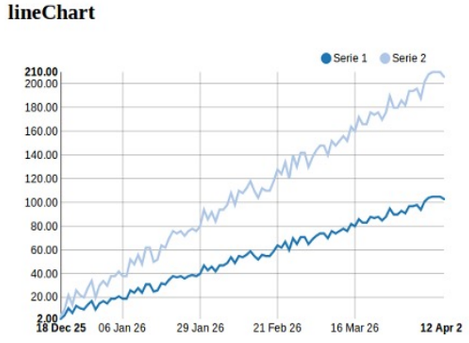
- Composition: pie chart…
- Comparison: bar chart, line chart…
- Distribution: histograms, box plot…
- Relationship: scatter plot…

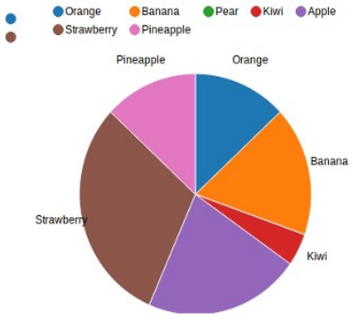




To determine which is best suited

- How many variables in a single chart?
- How many data points display for each variable?
- Will you display values over a period of time, or among items or groups





https://raw.githubusercontent.com/areski/python-nvd3
https://www.tatvic.com/blog/7-visualizations-learn-r/

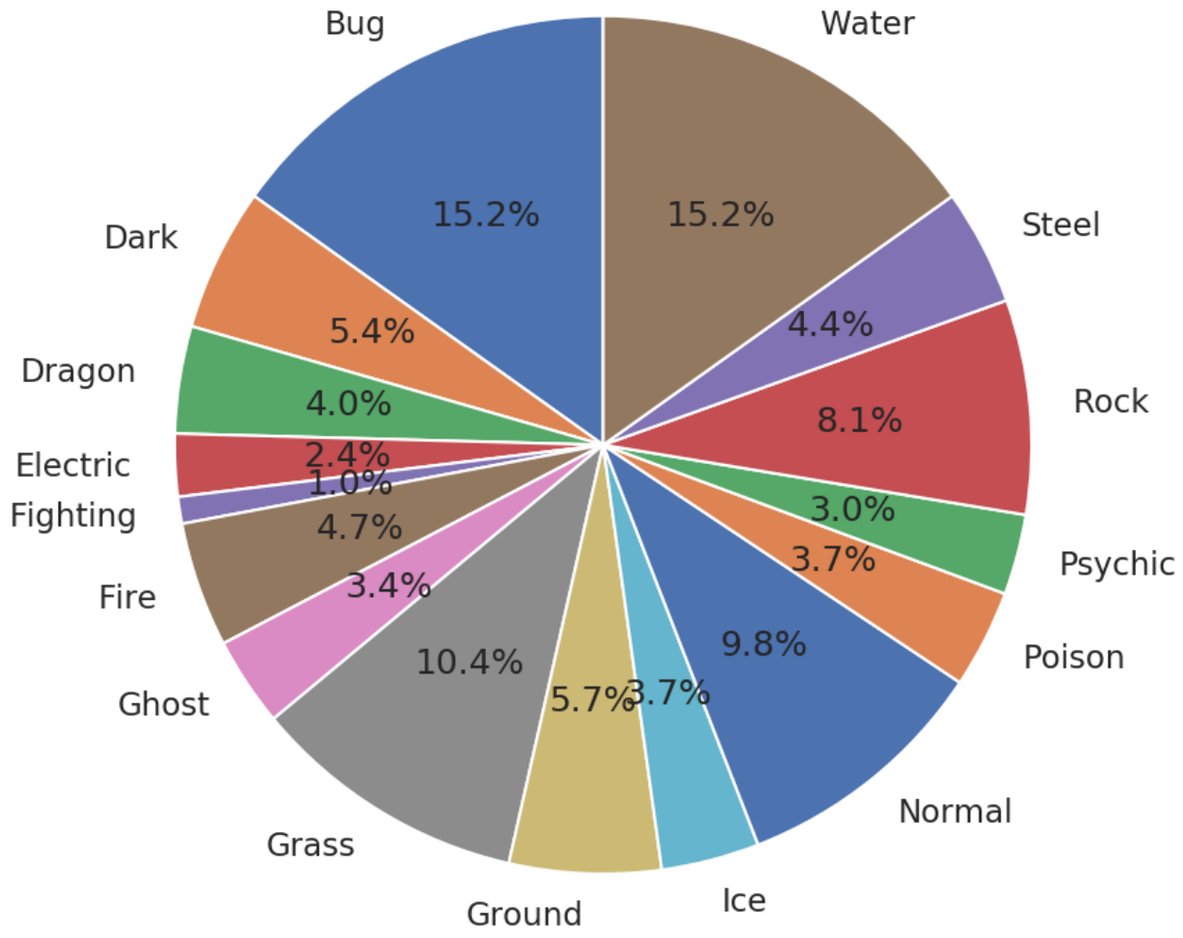
Visual Aids

Common charts in EDA:

- Pie chart
- Histogram
- Bar & Stack Bar Chart
- Box Plot & Violin plot
- Area Chart
- Scatter Plot
- Correlogram
- Heatmap

Pie chart

- Pie chart: circle divided to sectors, to communicate proportions
- a common method for representing categorical variables

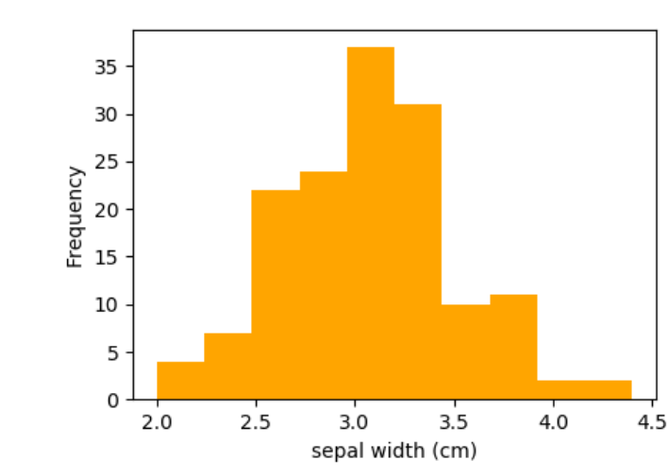


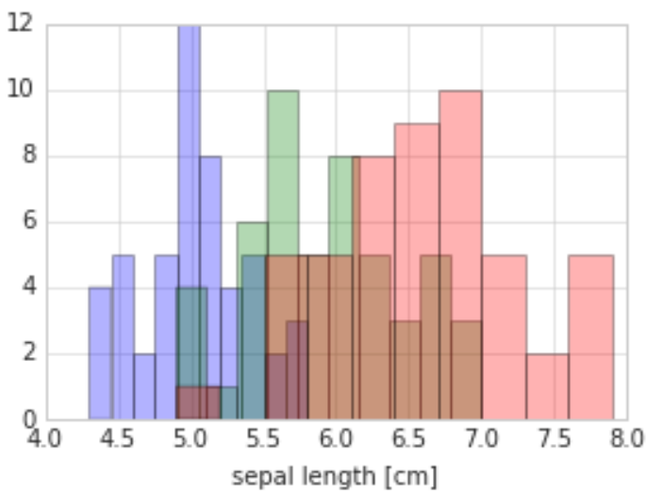
✓ simple and easy-to-understand
✓ understand information quickly

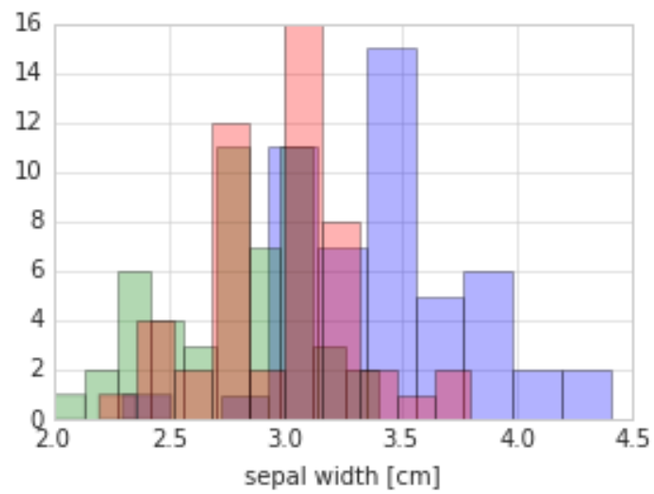
○ difficult to compare a few pieces
○ unhelpful when observing trends over time

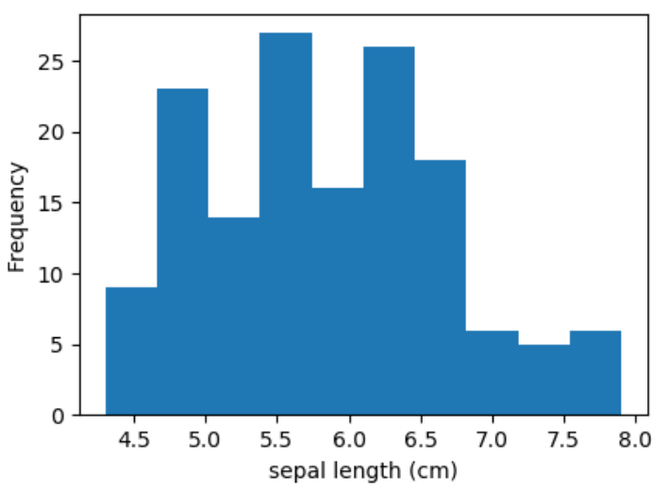
Histogram

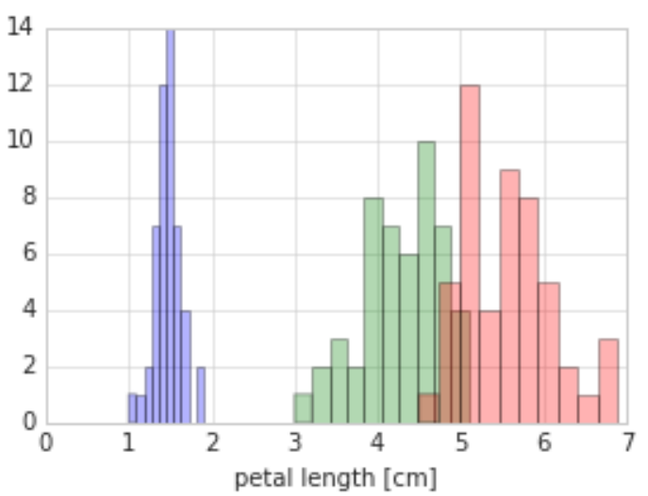
- Histogram: a plot of the frequency distribution of numeric variable by splitting values to small equal-sized bins, provide a visual summary of central tendency, spread, and shape.

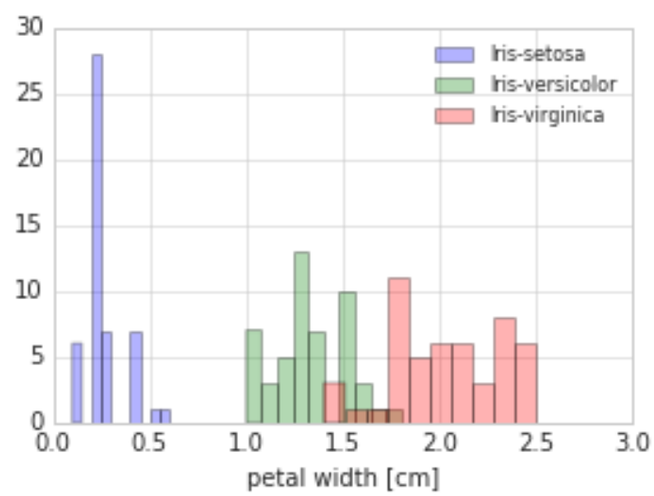








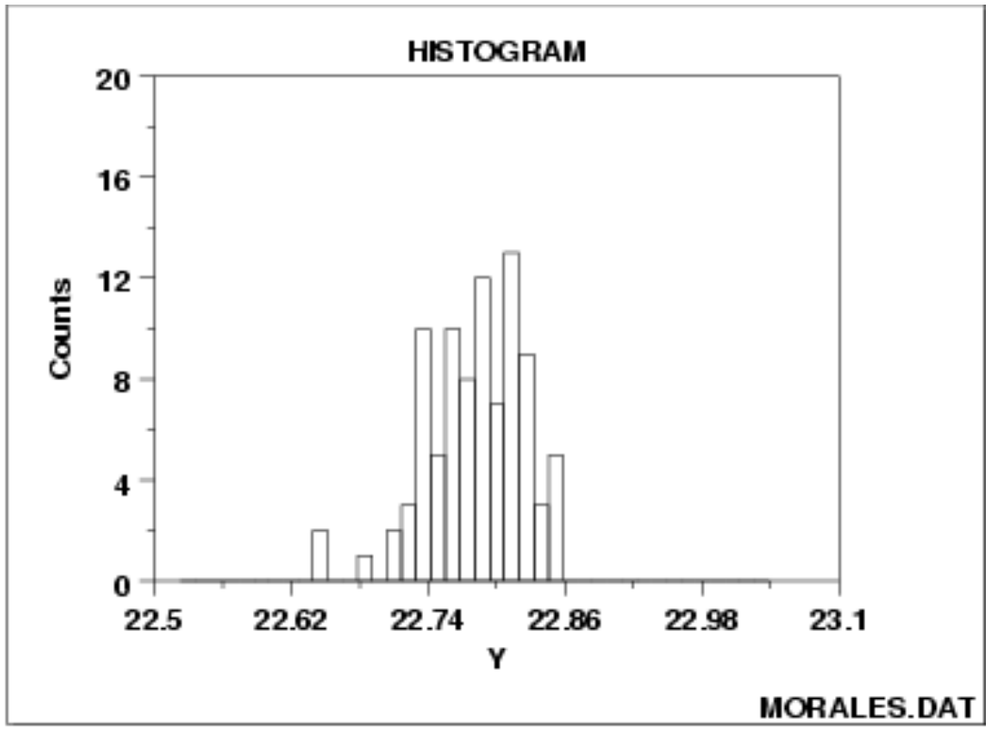


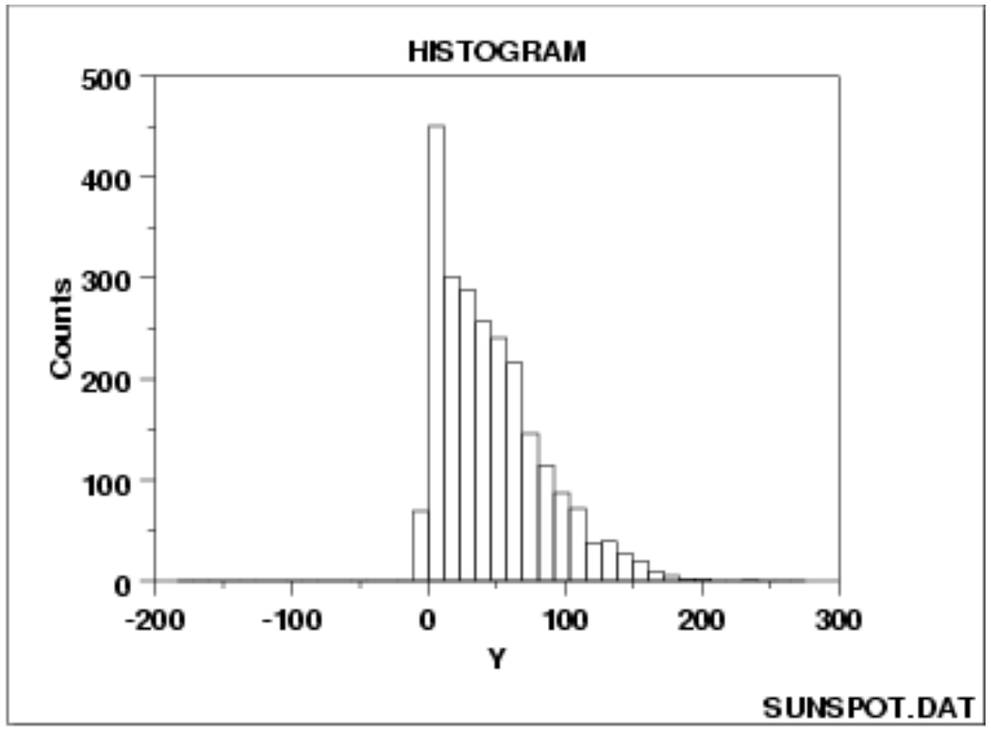


https://stackoverflow.com/questions/37911731/seaborn-histogram-with-4-panels-2-x-2-in-python

Shape of Data - Skewness

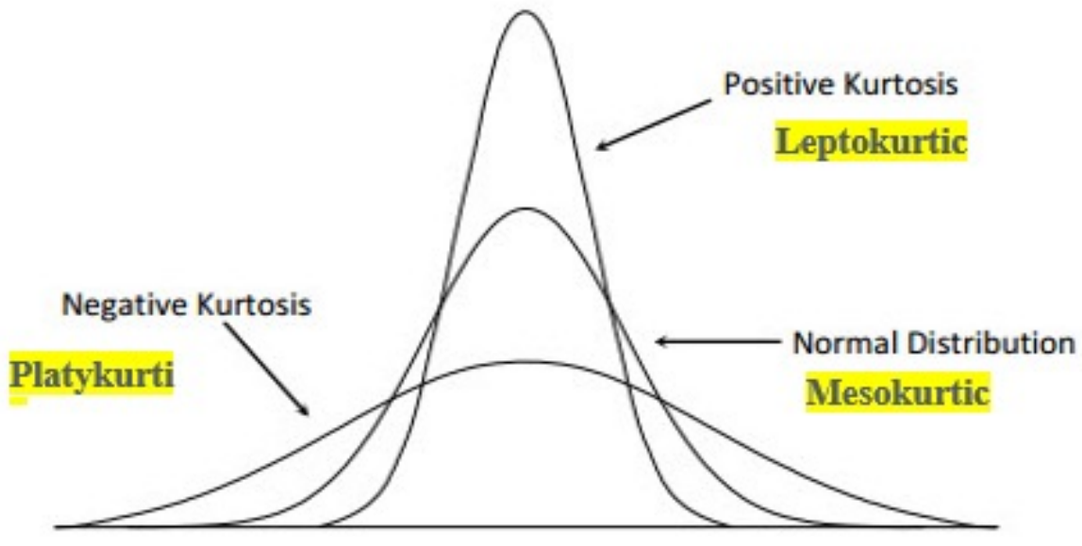
- skewness is a measure of the lack of symmetry
- a distribution, or data set, is symmetric if it looks the same to the left and right of the center point (mean≈median≈mode)
- Positive vs Negative vs Zero skewness





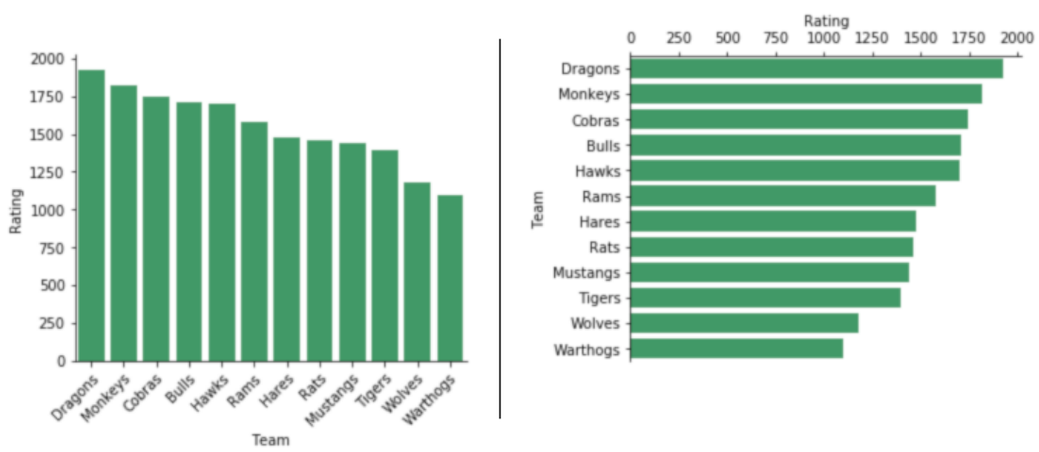
Shape of Data - Kurtosis

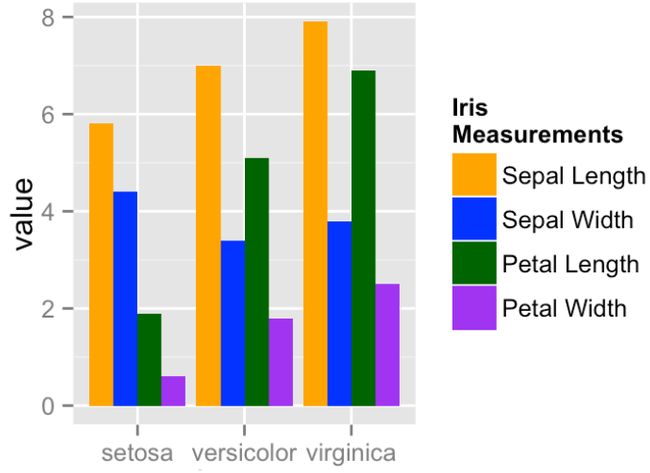
- Kurtosis: measures the degree to which the tails of a distribution differ from those of a normal distribution.
- Leptokurtic (positive) kurtosis: heavier tale
- Platykurtic (negative) kurtosis: lighter tale
- Mesokurtic (zero) kurtosis: normal tale

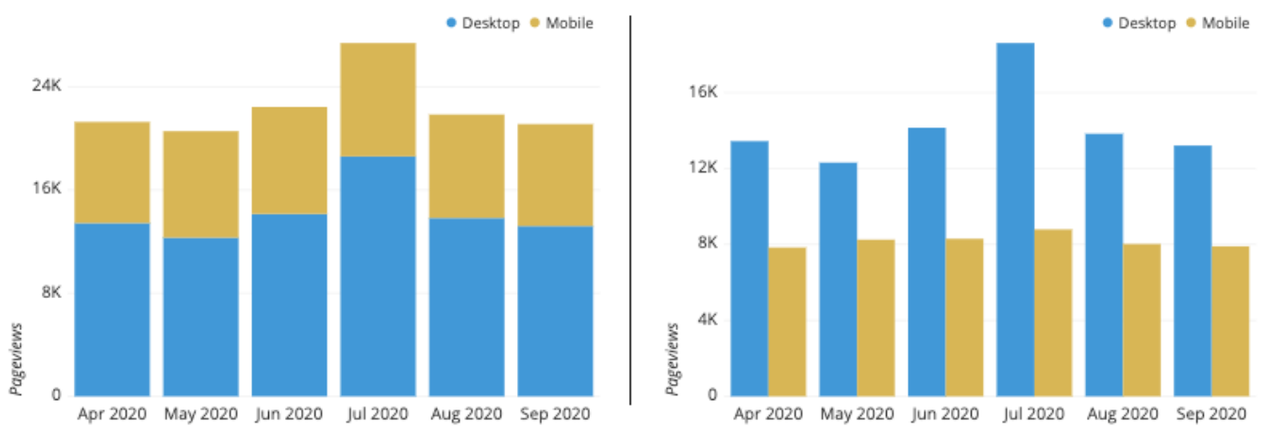


Bar & Stacked Bar Chart

- Bar charts: a way of summarizing a set of categorical data, displays data using bars, each representing a particular category, the height is proportional to a specific aggregation
- Bars can be horizontal or vertical



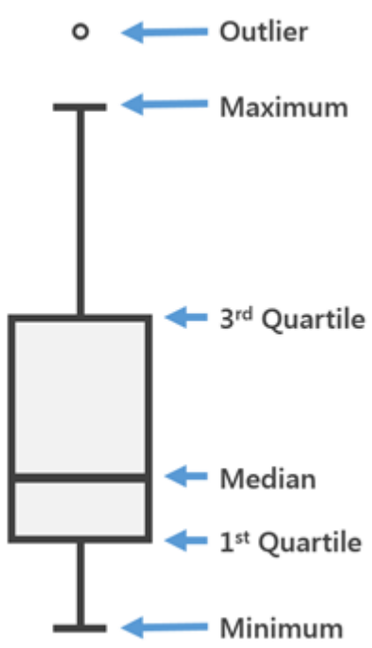


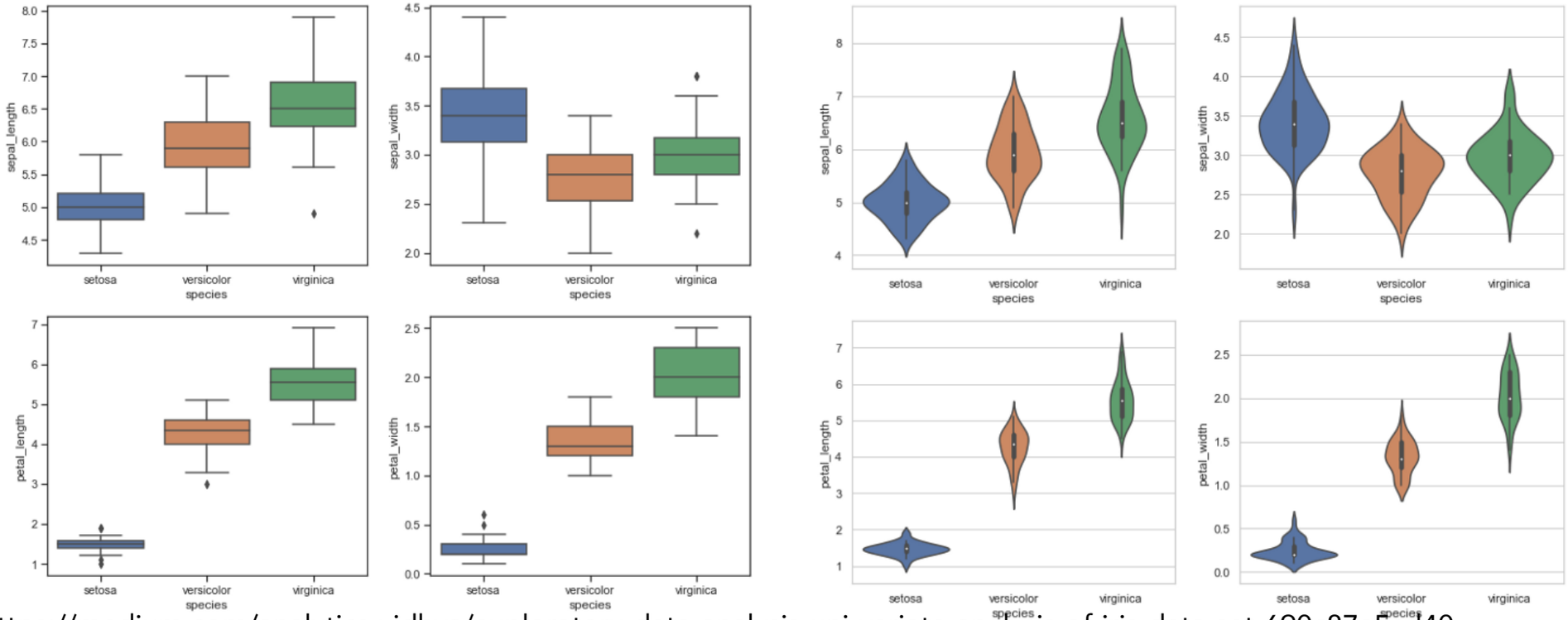


https://chartio.com/learn/charts/bar-chart-complete-guide/

Box Plot & Violin Plot

- Box plot: box and whisker plot, displays a summary of a large amount of data in five numbers , a good indication of how the values in the data are spread out with in groups
- plot a combination of categorical and continuous variables
- Violin plot: similar as box plot, additionally shows the kernel density estimation of the underlying distribution

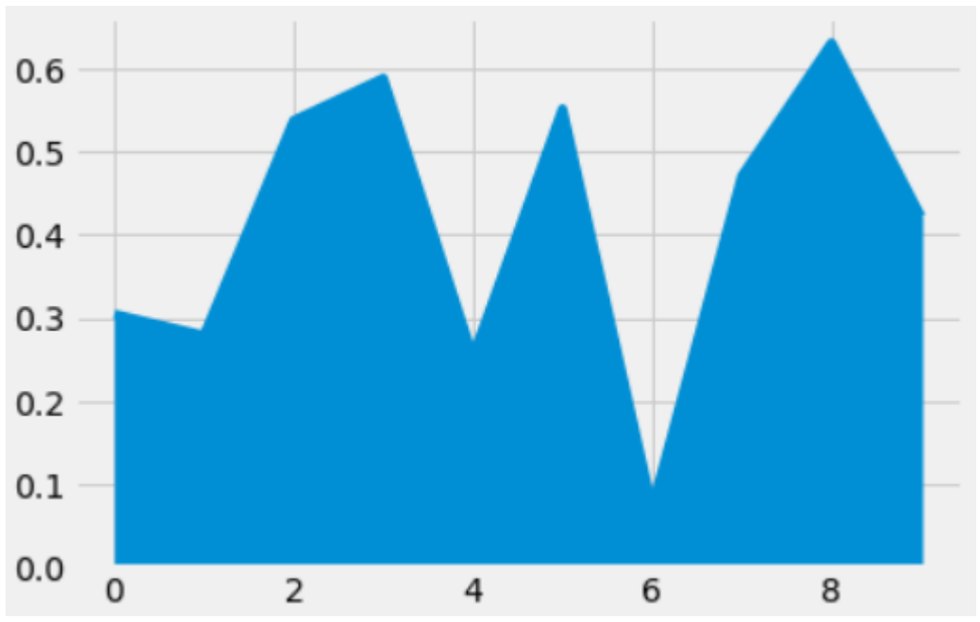


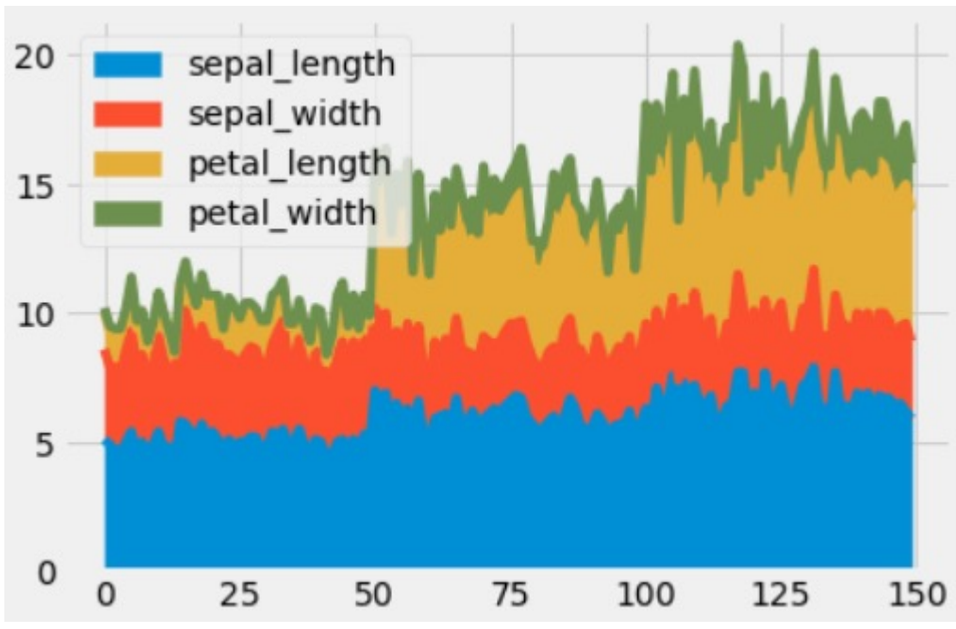


https://medium.com/analytics-vidhya/exploratory-data-analysis-uni-variate-analysis-of-iris-data-set-690c87a5cd40

Area Chart/Stacked Chart

- base on the line chart, areas between axis and line are commonly emphasized with colours
- share features with bar charts and line charts, compare two or more quantities, work better for large difference and multiple values over time

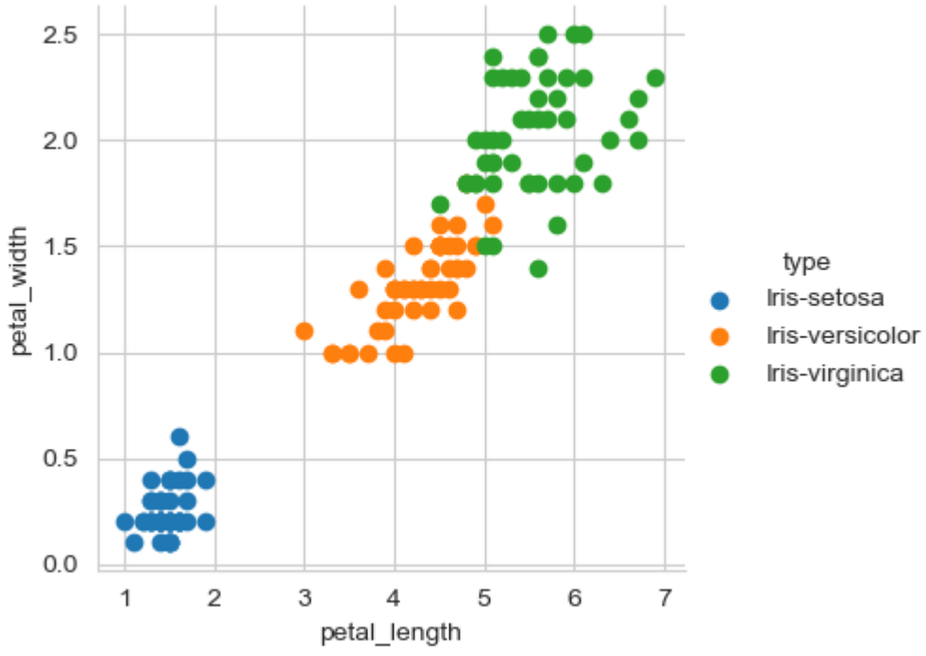


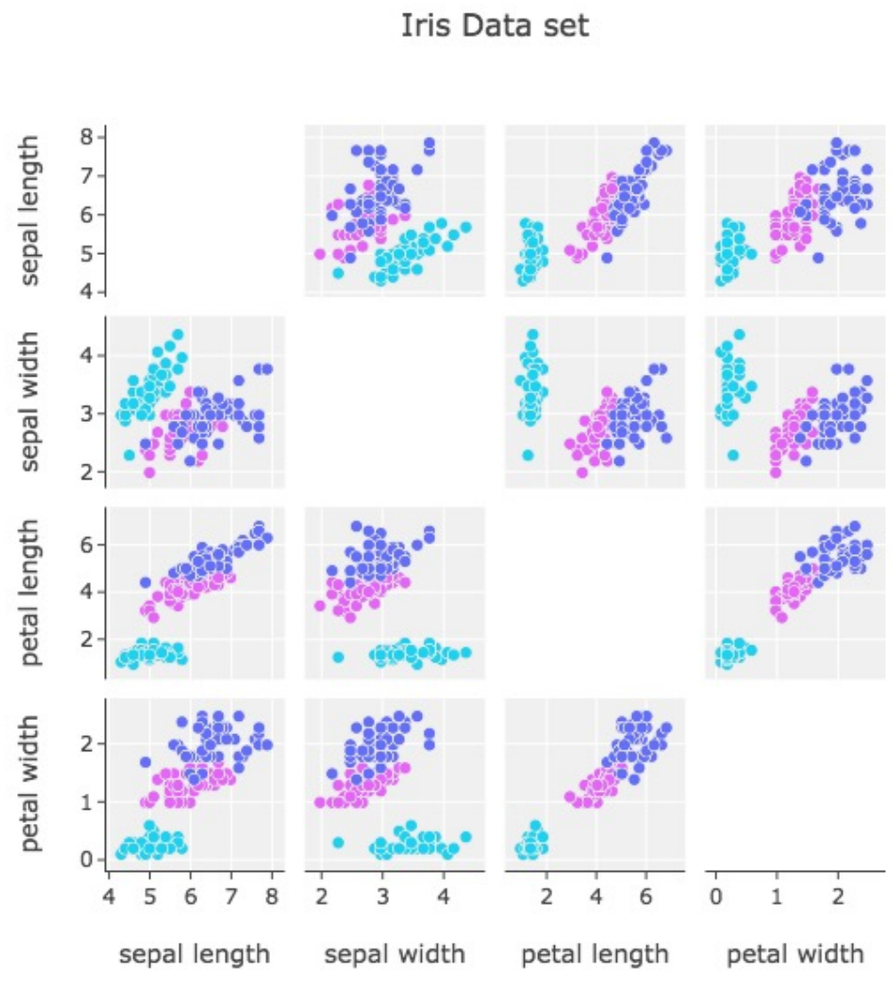


https://levelup.gitconnected.com/data-visualization-with-pandas-in-action-part-2-2cc8674da1d0

Scatter Plot

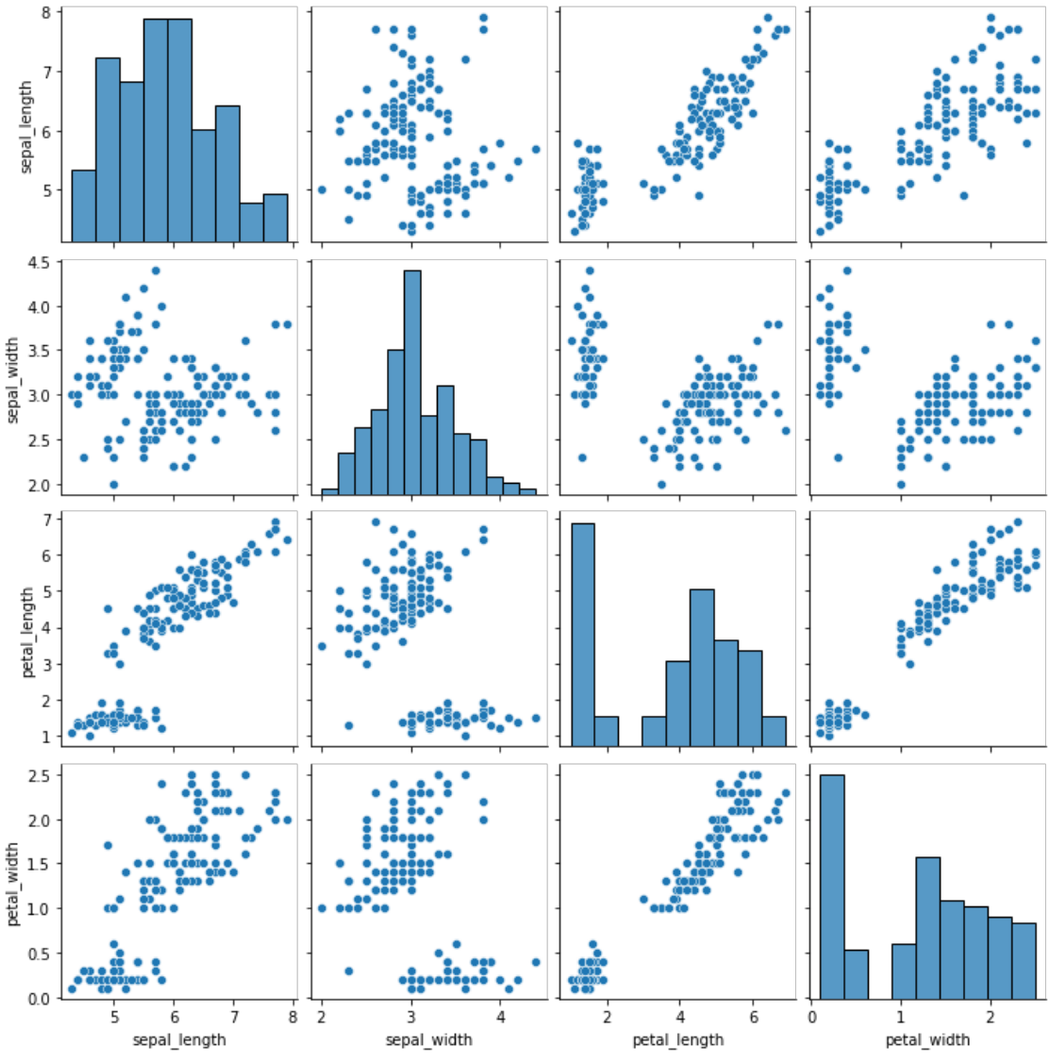
- use a Cartesian coordinates system to display values of two variables for a set of data
- show the relationship between two variables, referred to as correlation plots

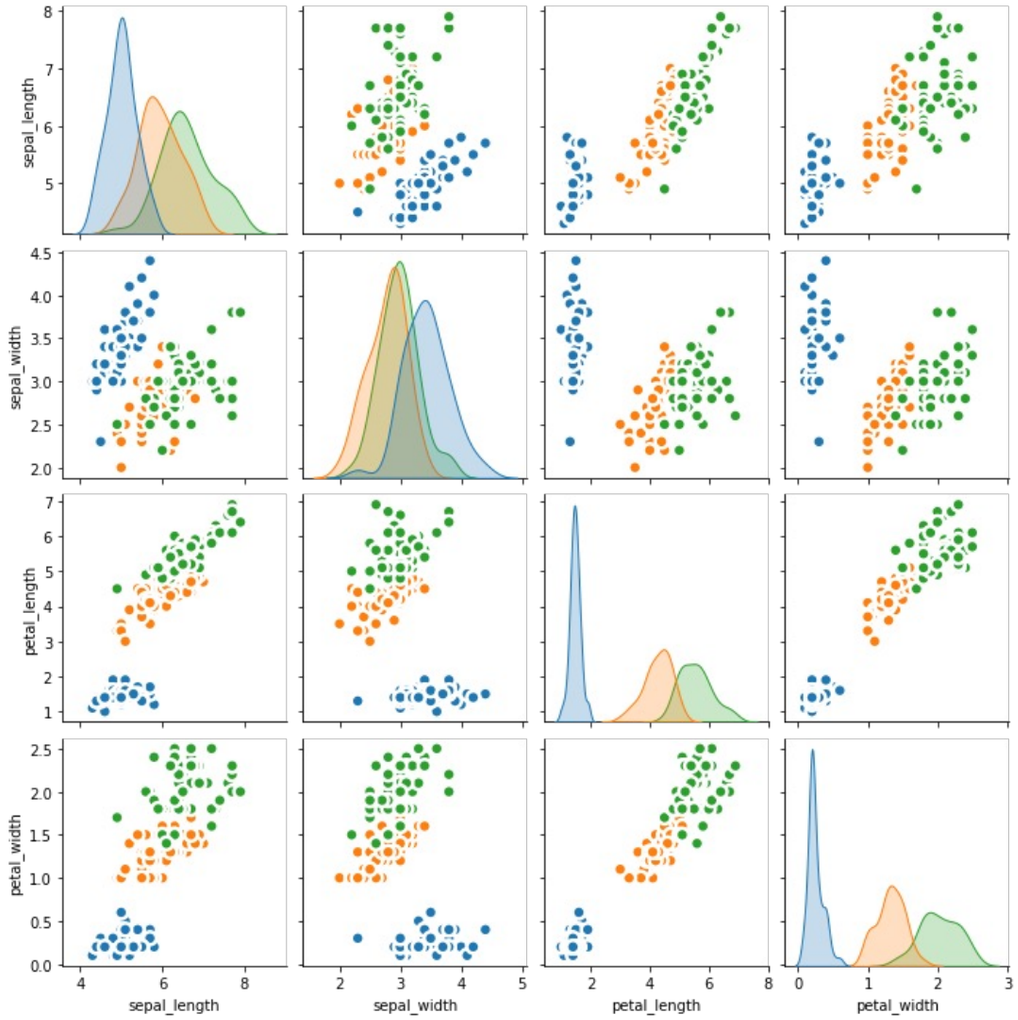




Correlogram

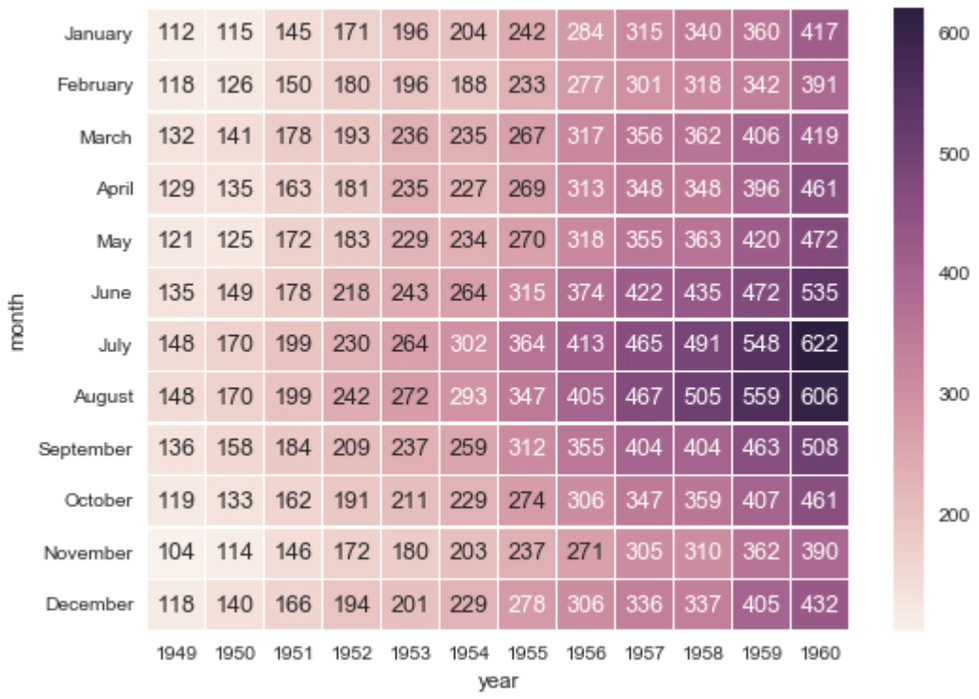
- Correlogram: AKS correlation matrix, to analyse the relationship between each pair of numeric variables

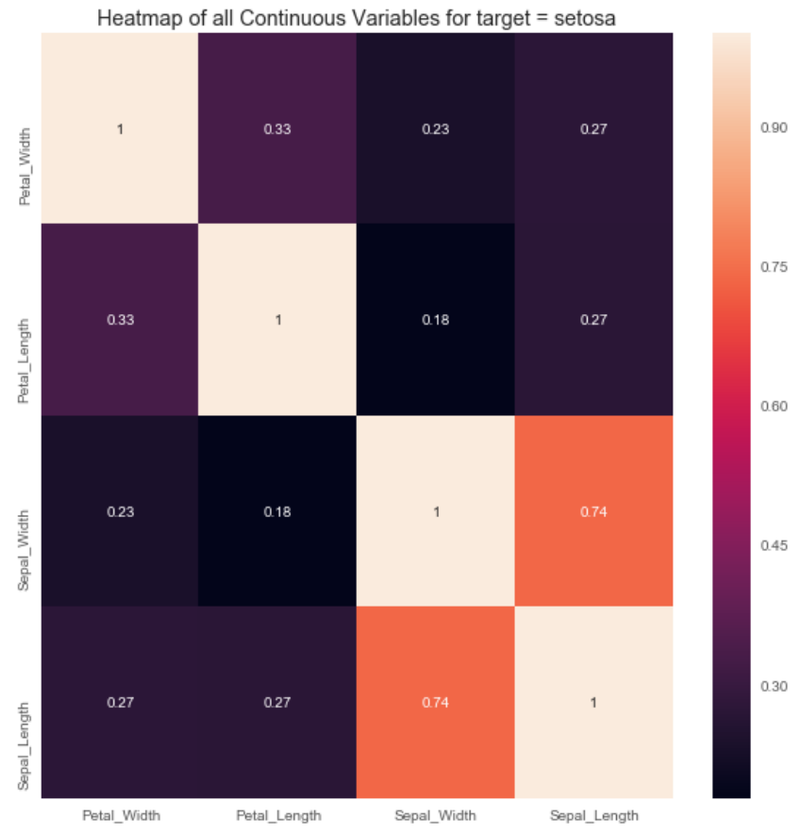




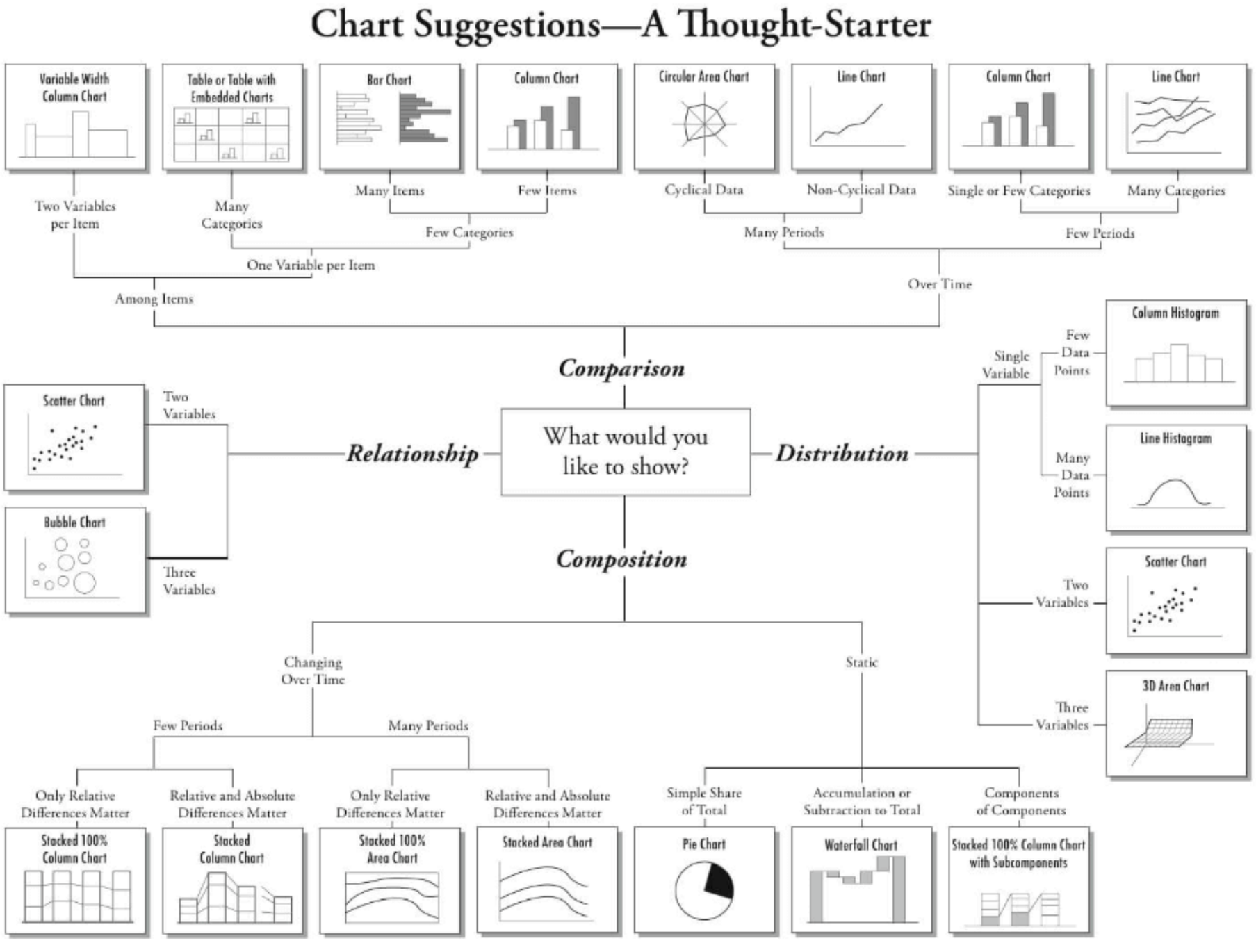
HeatMap

- Heatmap: a two-dimensional graphical representation of data where the individual values that are contained in a matrix are represented as colors
- useful to see which intersections of the categorical values, have higher concentration of the data compared to the others





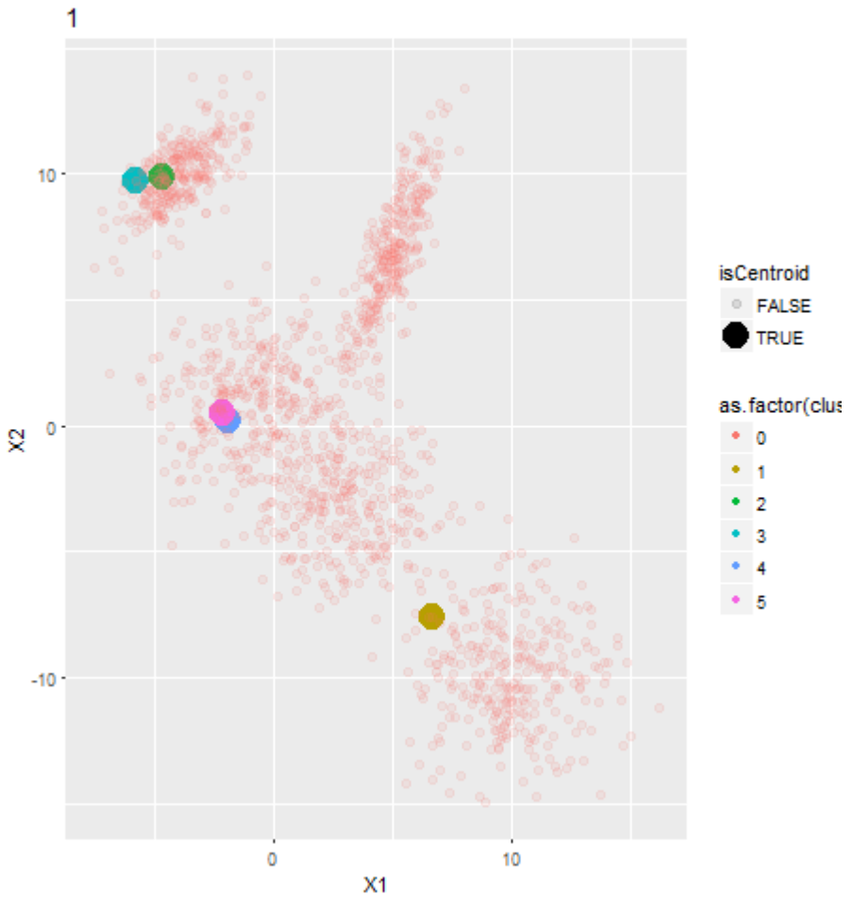
Choose the Most Suitable Plots



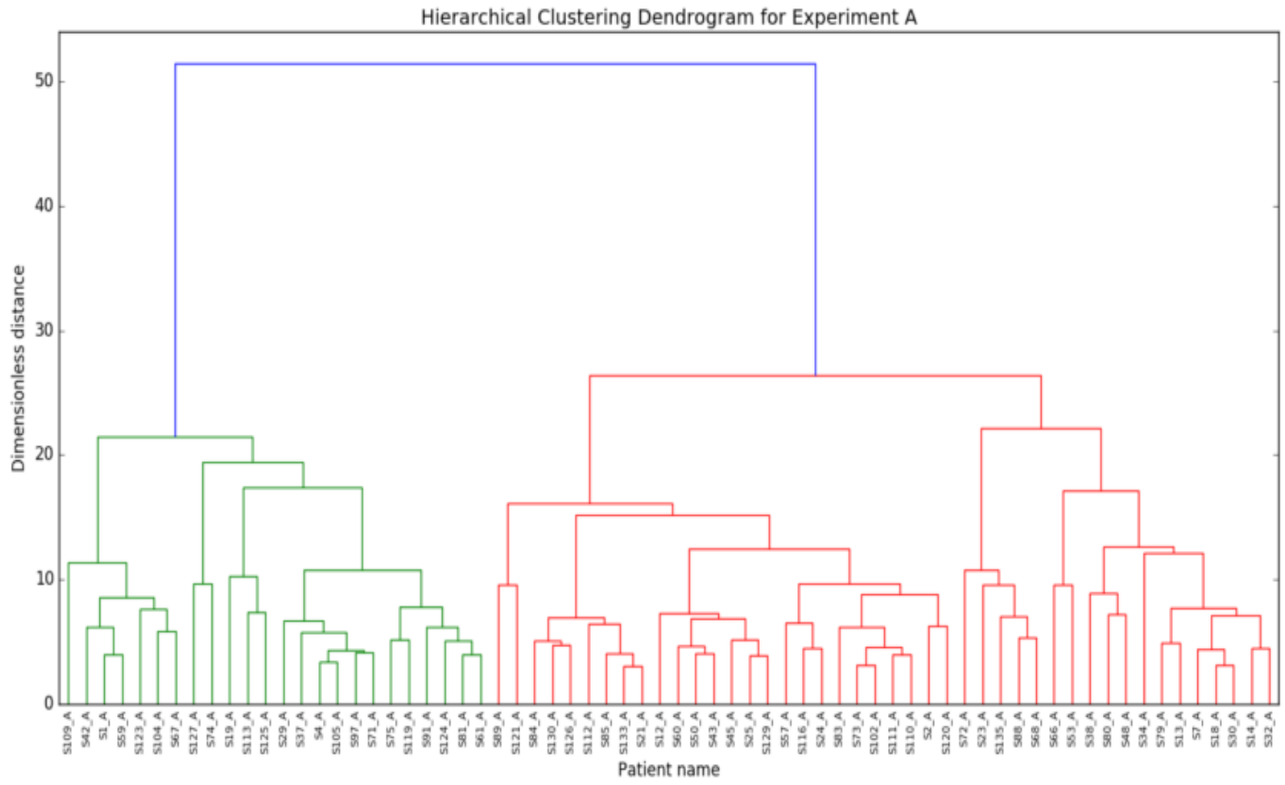
https://www.tatvic.com/blog/7-visualizations-learn-r/

Clustering Analysis for EDA

- Clustering in EDA to find new insights



K-Means clustering
can be used to
detect possible outliers

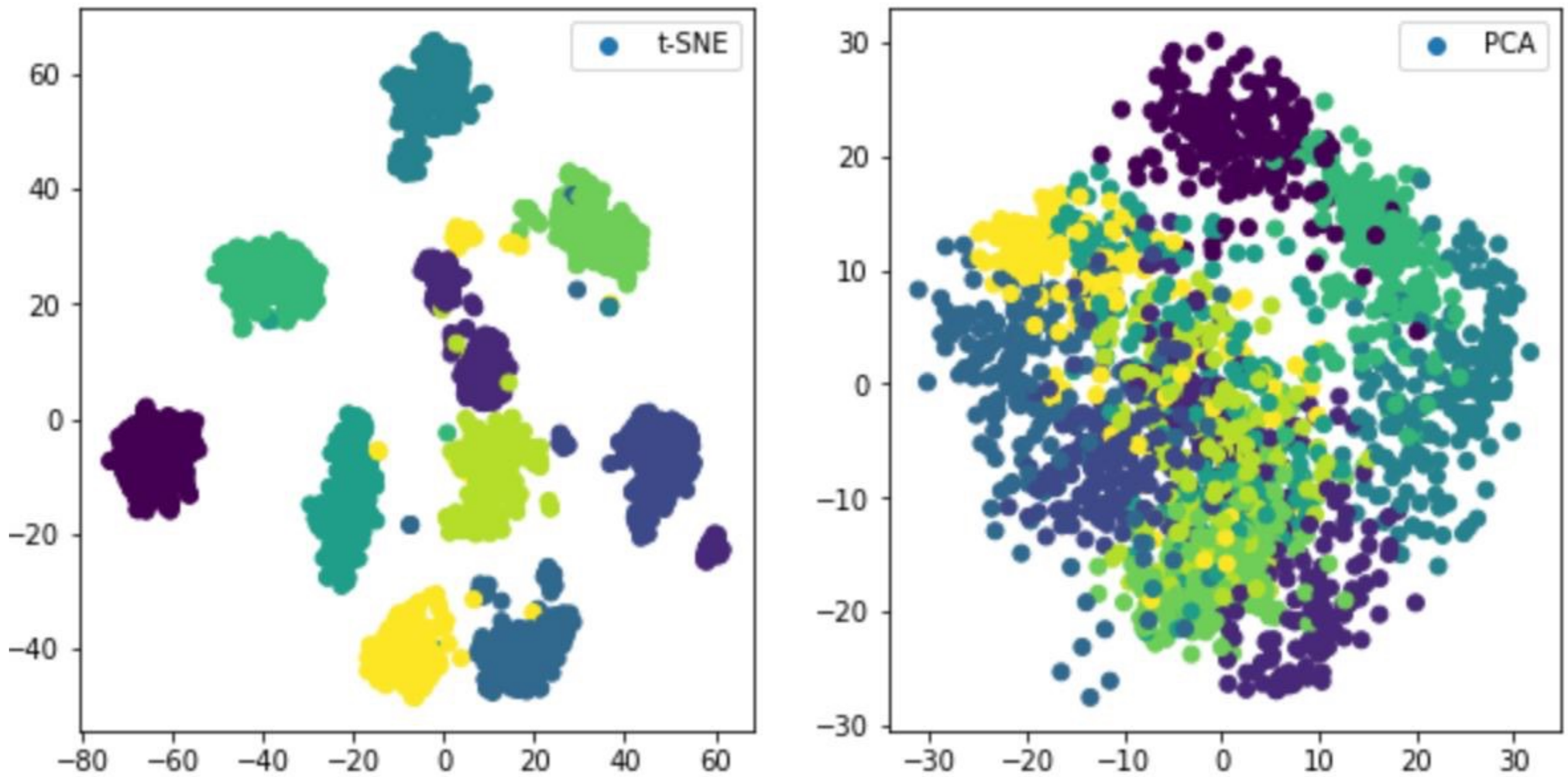


Hierarchical clustering
can be used to
find underlying connectivity properties

https://www.neuroelectrics.com/blog/2018/07/06/clustering-methods-in-exploratory-analysis/

Dimensionality Reduction for EDA

- Reduce the dimensions of the data into fewer dimensions would help describing the relationship between variables



T-distributed stochastic neighbor embedding (T-SNE) and Principal Component Analysis (PCA)

https://www.programmersought.com/article/92363395092/

What to look for in your plots?

- Turn the information into useful questions
  - Which values are the most common? Why?
  - Which values are rare? Why?
  - Can you see any unusual patterns? What might explain them?

- Clusters suggest that subgroups exist in your data.
  - How can you explain or describe the clusters?
  - How are the observations within each cluster similar to each other?
  - How are the observations in separate clusters different from each other?