Lecture 6

–

Visual EDA

(Exploratory Data Analysis)

Star / Starling

(*Sturnus vulgaris*)

- It is about 20 cm long and has glossy black plumage with a metallic sheen, which is speckled with white at some times of year. The legs are pink and the bill is black in winter and yellow in summer.
- It is a noisy bird, especially in communal roosts and other gregarious situations, with an unmusical but varied song. Its gift for mimicry has been noted in literature including the the works of William Shakespeare.
- The common starling is a highly gregarious species, especially in autumn and winter. Although flock size is highly variable, huge, noisy flocks may form near roosts.
- Due to numerous naturalizations on other continents, the starling is now one of the most common birds in the world.
- The star was "Bird of the Year 2018" in Germany.

Sources:
* Photo by Сергей Шабанов on pixabay
* https://en.wikipedia.org/wiki/Common_starling

Visual EDA

1. Introduction
2. Univariate Visualizations
3. Bivariate Visualizations
4. Multivariate Visualizations

EDA - Intro

- „**Exploratory Data Analysis** refers to the critical process of performing initial investigations on data so as
  - to discover patterns,
  - to spot anomalies,
  - to test hypothesis and
  - to check assumptions with the help of

  summary statistics and graphical representations.“

Source: https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

Matplotlib and Seaborn

- “Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.”

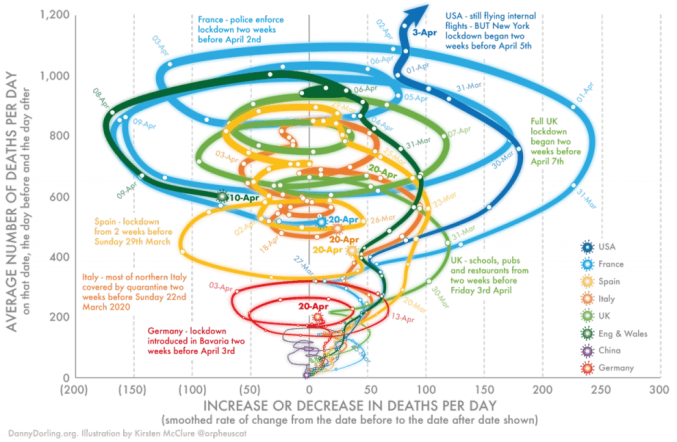
  See https://matplotlib.org/

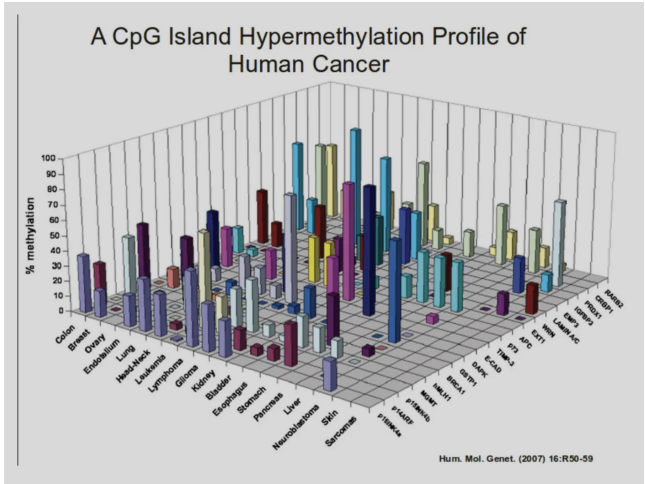
- “Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.” “It builds on top of matplotlib and integrates closely with pandas data structures.”

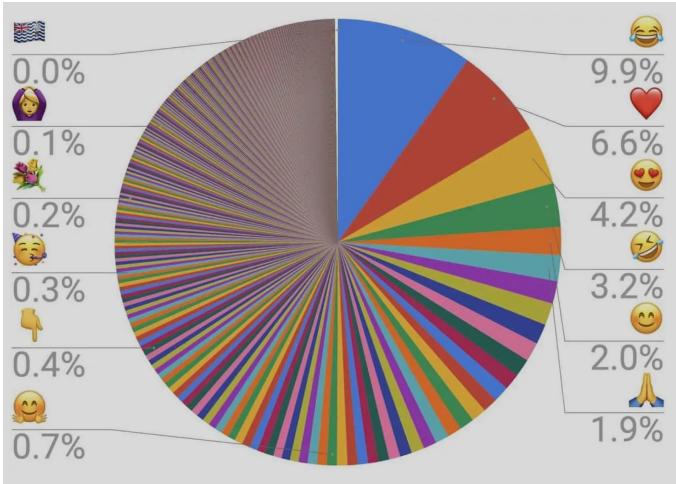
  See https://seaborn.pydata.org/

Caveats

- Many visualizations have been proposed
  - And easy to generate – powerful libraries available
- Many are visually pretty, but very hard to interpret
  - Use visualizations that are easy to interpret
  - You will need them for communication with domain experts from the business side later
  - We only look at the most widely used visualizations
- Exploring data is fun!
  - Do not get carried away
  - Focus on what to explore
  - Spend your time wisely!







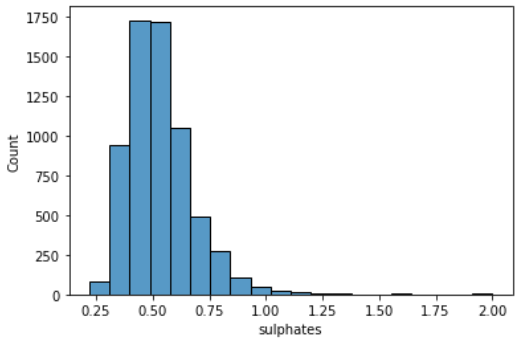
Source: https://www.oldstreetsolutions.com/good-and-bad-data-visualization

Visual EDA

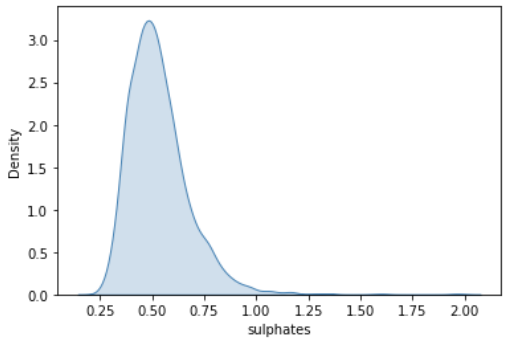
1. Introduction
2. Univariate Visualizations
3. Bivariate Visualizations
4. Multivariate Visualizations

Univariate Visualizations – Numerical Data (1)

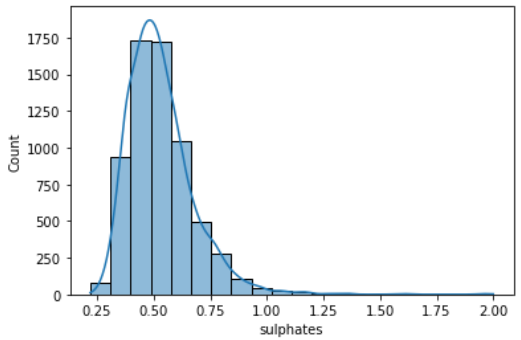
- Numerical Data – Analyse the Distribution
  - Histograms
  - Density Plots
  - Combined Histograms and Density Plots



Histogram



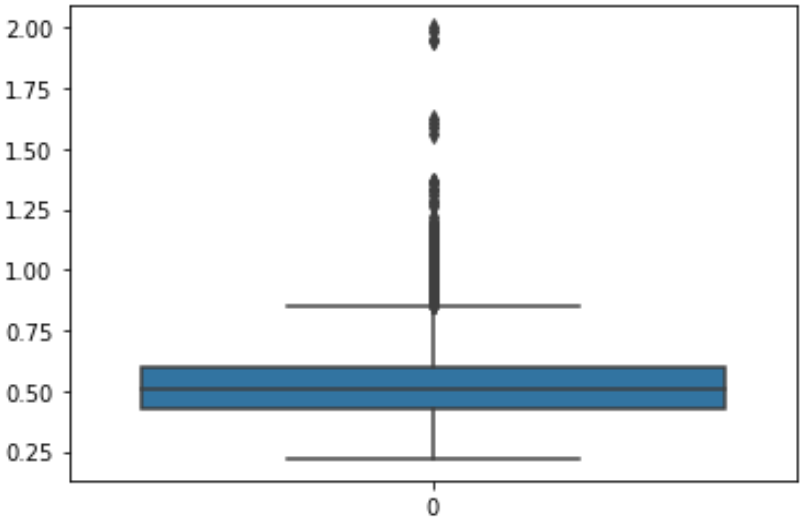
Density Plot



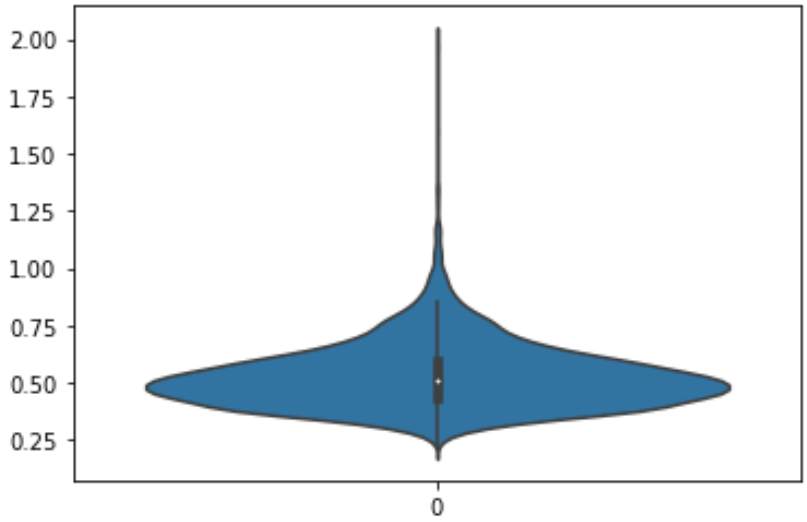
Combined Histogram and Density Plot

Univariate Visualizations – Numerical Data (2)

- Numerical data – Distribution and Outliers
  - Box Plots
  - Violin Plots



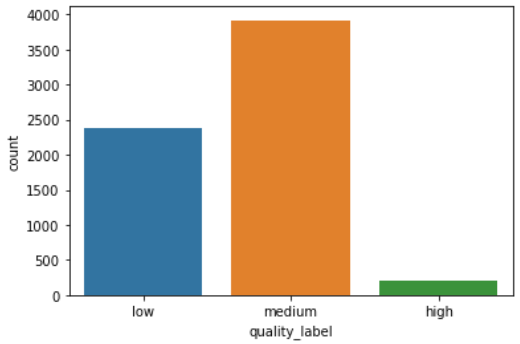
Box Plot



Violin Plot

Univariate Visualizations - Categorical Data

- Categorical data
  - Frequency Plots



Frequency Plot

Demo

Demo

Visual EDA (Univariate)

Source: Foto von Markus Spiske auf Unsplash

Exercise

Exercise 1

Iris Data – Univariate Visualizations for EDA

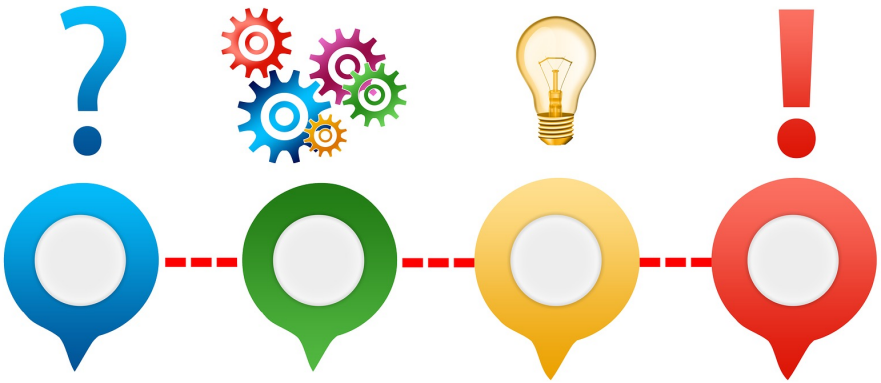


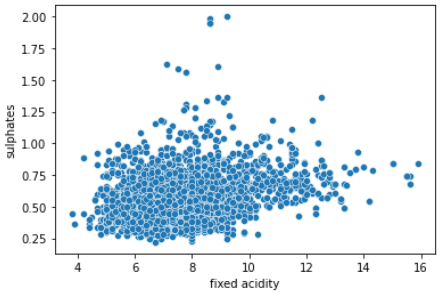
Photo by Gerd Altmann on Pixabay

Visual EDA

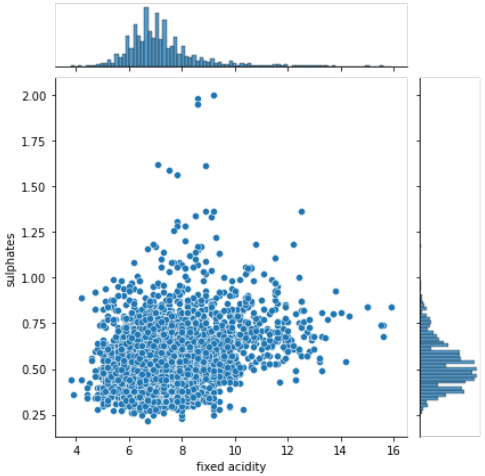
1. Introduction
2. Univariate Visualizations
3. Bivariate Visualizations
4. Multivariate Visualizations

Bivariate Visualizations (1)

- Two Attributes - Numerical
  - Scatter Plot
  - Joint Plot



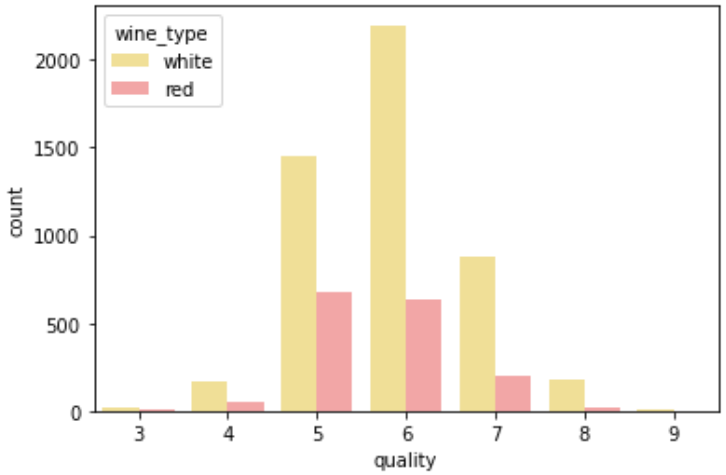
Scatter Plot



Joint Plot

Bivariate Visualizations (2)

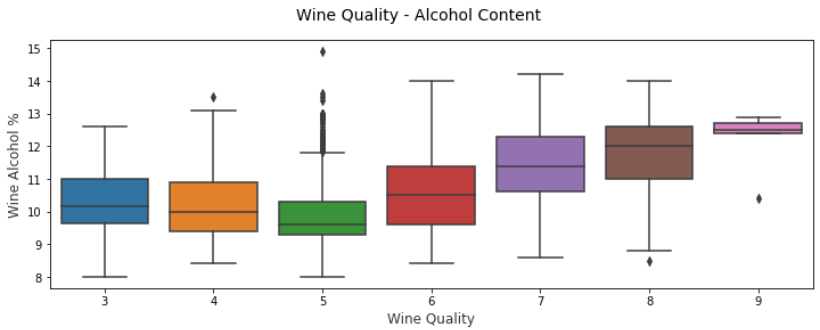
- Two Attributes – Categorical
  - Multi Bar Plot



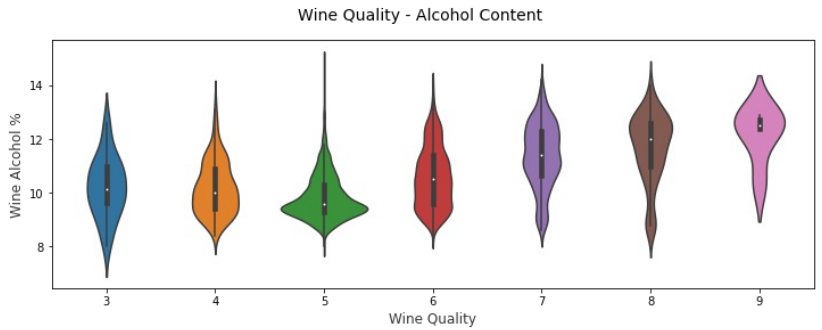
Multi Bar Plot

Bivariate Visualizations (3)

- Two Attributes – Mixed
  - Multi Box Plot
  - Multi Violin Plot



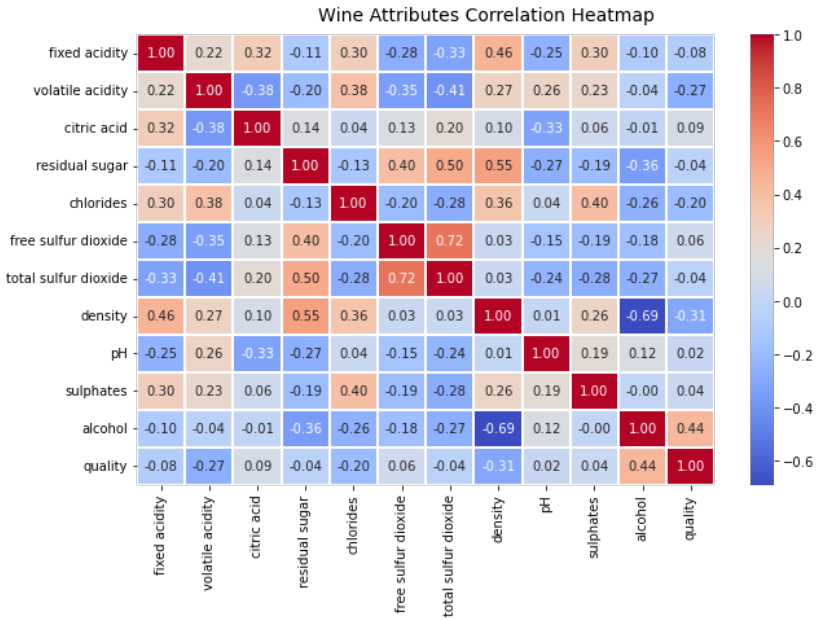
Multi Box Plot

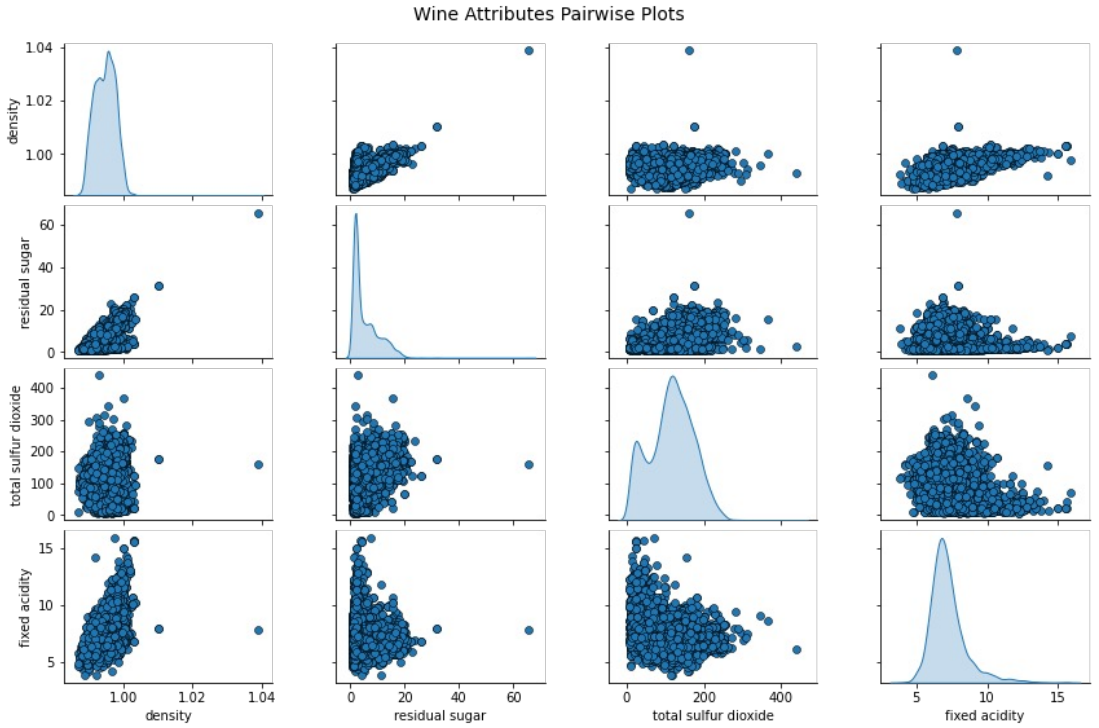


Multi Violin Plot

Bivariate Visualizations (4)

- Pair-wise Correlation Matrix Heatmap
- Pair-Wise Scatter Plots





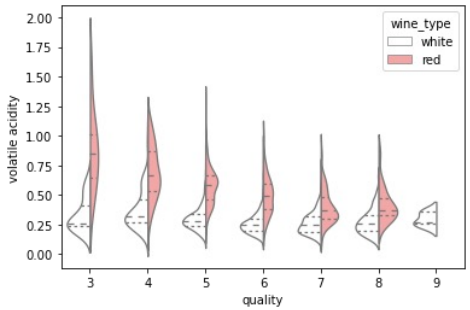
Visual EDA

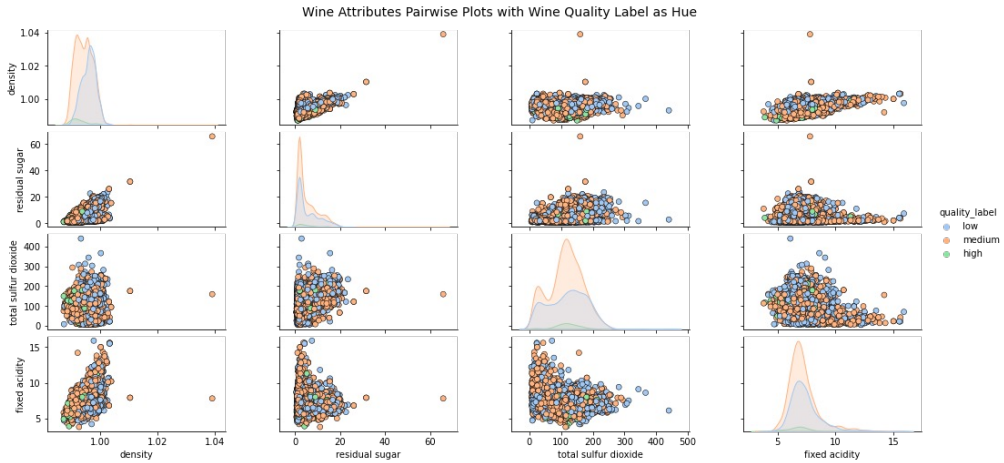
1. Introduction
2. Univariate Visualizations
3. Bivariate Visualizations
4. Multivariate Visualizations

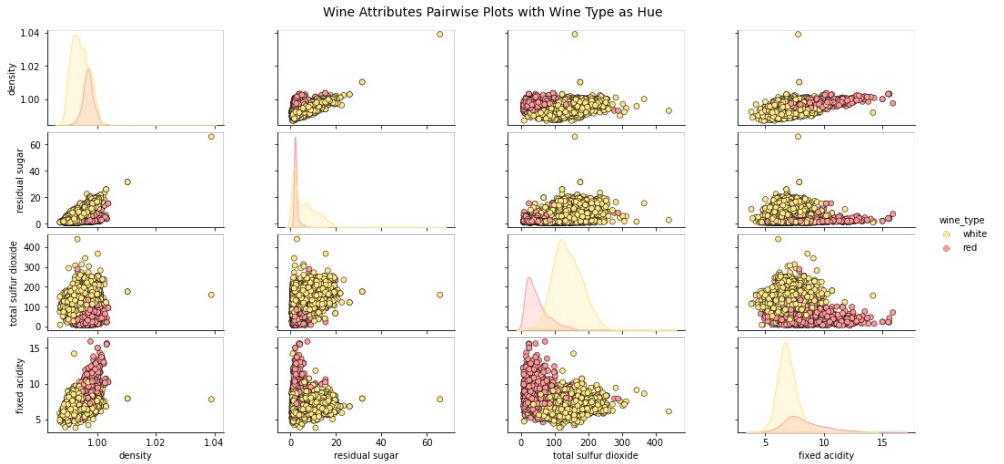
Multivariate Visualizations

- Often hard to implement and interpret
  - Interactive visualizations can be helpful
    - Rotate a 3-d scatterplot
    - Time dimension as animation

- Introduce third dimension as colour / hue into bivariate visualizations
  - Frequently used for the target variable in classification problems
    - Example: Classification: Pair-wise scatter plot of numerical attributes with the target (class) as colour







Demo

Demo

Visual EDA (Bi- and Multivariate)

Source: Foto von Markus Spiske auf Unsplash

Exercise

Exercise 2

Iris Data – Bivariate and Multivariate Visualizations for EDA

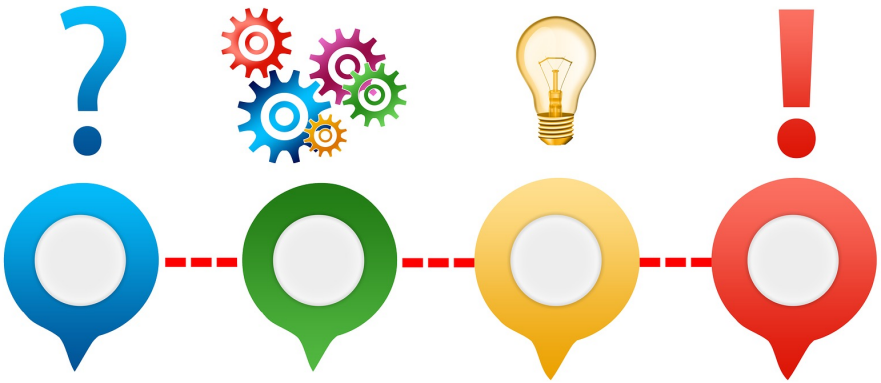


Photo by Gerd Altmann on Pixabay

Key Takeaways

- Visual Analysis of the data is key part of EDA
- Start with univariate visualizations of all features (and the target, if applicable)
  - Start with a Histogram/Density Plot/Frequency Plot
  - Look for Outliers in the Box Plots
- Continue with bivariate/multivariate visualizations
  - Pair-wise scatter plots
  - Colour code/hue code the target, if applicable
- Use easy-to-interpret and easy-to-communicate visualizations
- Focus – explore deep but do not get carried away

Image by Gordon Johnson on pixabay