DAT200 – Applied Machine Learning I

Raw Data Inspection

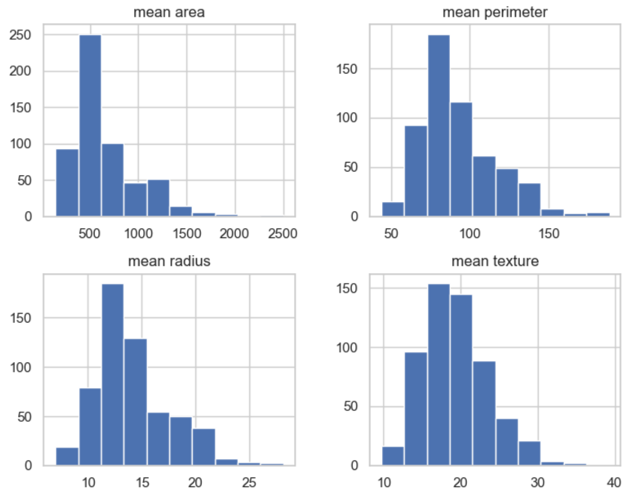
Raw data inspection

- Understand your data if you want to obtain best possible results with your machine learning models
- Visualise your data – the most effective way to learn more about your data
- Absolutely necessary to do this before training your machine learning models
- NOTE: your compulsory assignment submissions will not be accepted without raw data inspection at the beginning of your Juptyer Notebook
- BEWARE: The teachers might include one or more rows with fictious data that produces outliers and will lead you poorer models

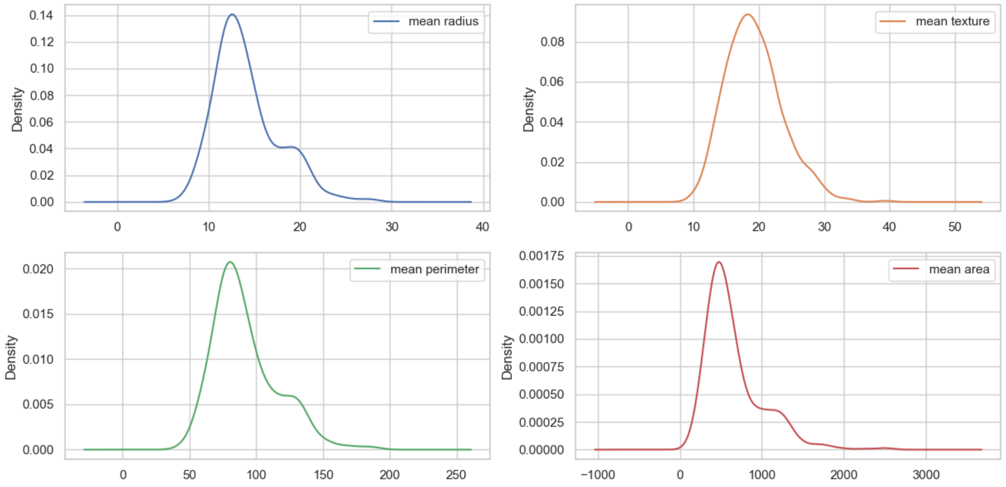
- See python script: «Ch00 – 2 - raw data inspection.py»

Raw data inspection – univariate plots

- Compute descriptive statistics
- Histograms
  - Inspect distribution of each attribute
  - Groups data into bins
  - Count number of observations in each bin

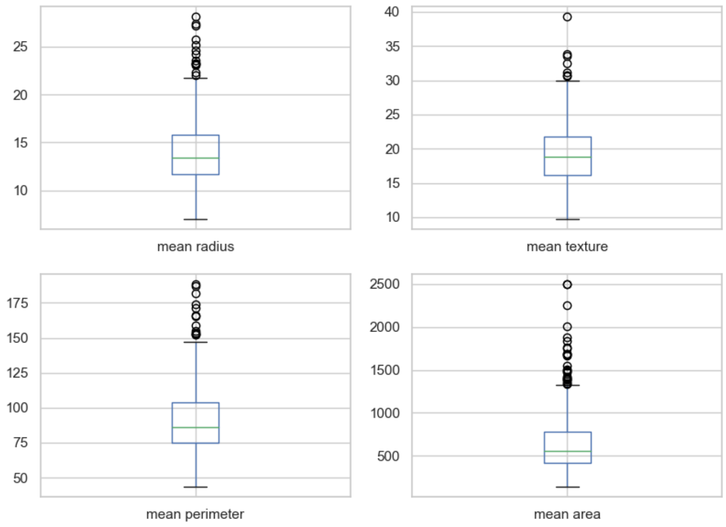


- Density plots
  - Another way of inspecting distributions
  - Smoothed curve



Raw data inspection – univariate plots

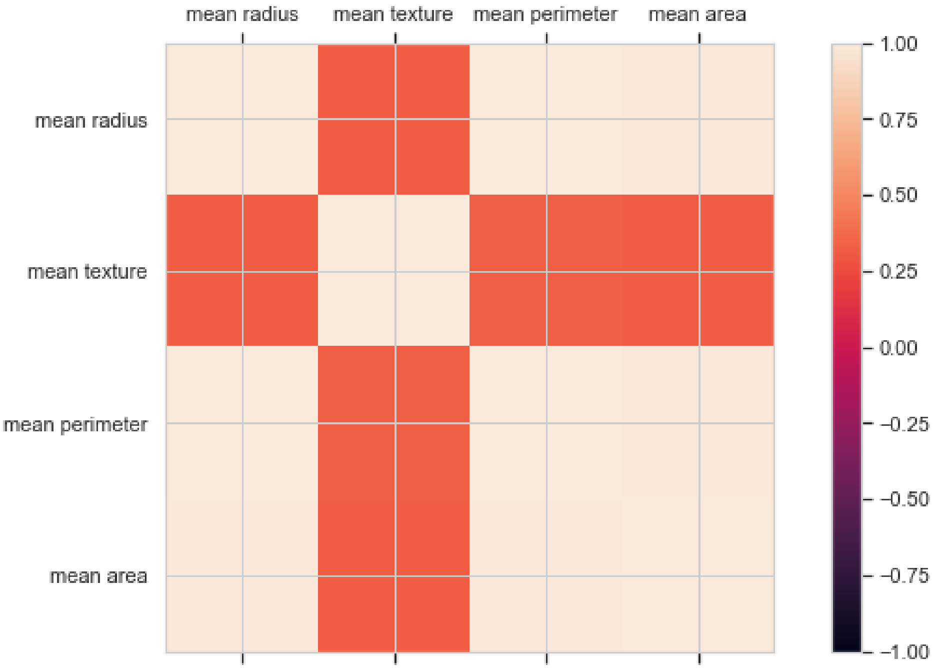
- Box and Whisker plots
    - Inspect distribution of each attribute
    - Boxplots summarise the distribution
        - Line for the median
        - Box around 25th and 75th percentile (middle 50% of the data)
        - Whiskers: 1.5 greater than size of spread ot the middle 50% of the data
        - Dots outside whiskers show candidate outlier values



- Violin plots: a better alternative to Box and Whisker plots
    - Another version/aspect of density plots
    - Give a more complete/precise description of the data
    - Examples: https://www.autodeskresearch.com/publications/samestats

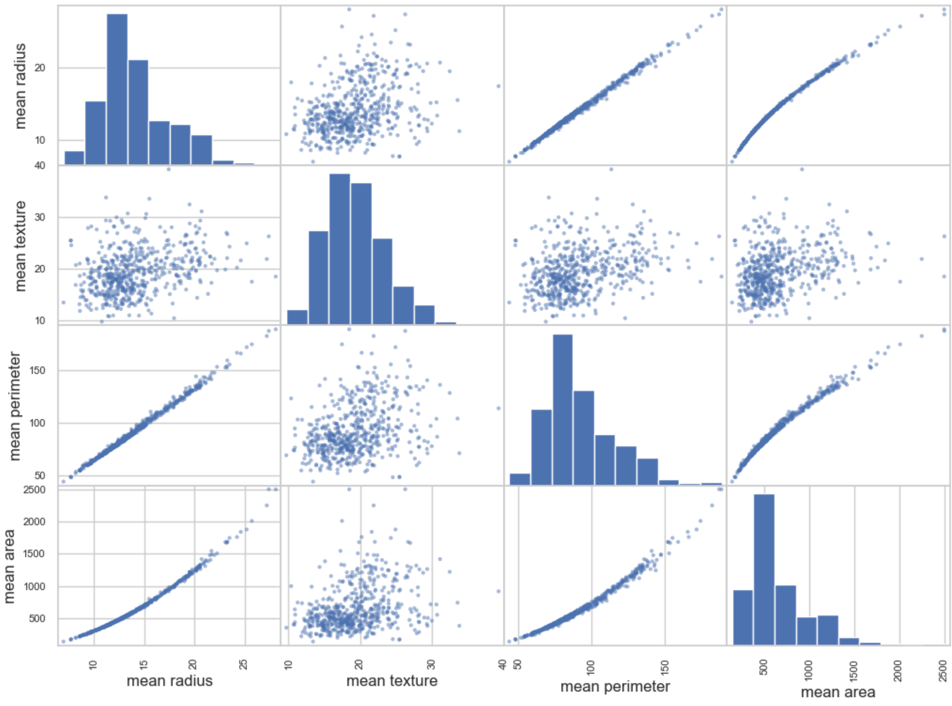
Raw data inspection – multivariate plots

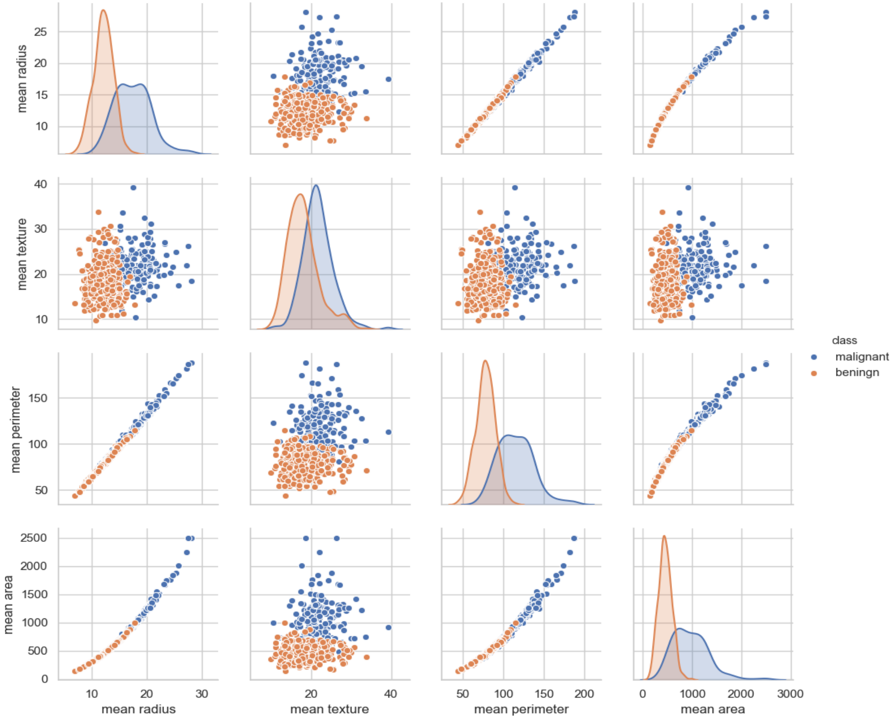
- Correlation Matrix Plot
    - Gives indication of how related changes of two variables are (pairs of variables)
    - Two variables change in same direction: positive correlation
    - Two variables change in opposite direction: negative correlation



Raw data inspection – multivariate plots

- Scatter Plot Matrix
  - Shows relationship between two variables as dots in two dimensions
  - Useful for spotting structured relationships between variables
  - Structural relationships may also be correlalated and good candidats for removal from the dataset


mean radius
mean texture
mean perimeter
mean area


mean radius
mean texture
mean perimeter
mean area
class
malignant
beningn