PURDUE UNIVERSITY

CS37300:
Data Mining and Machine Learning

Exploratory Data Analysis

Sep 13 2023

PURDUE UNIVERSITY

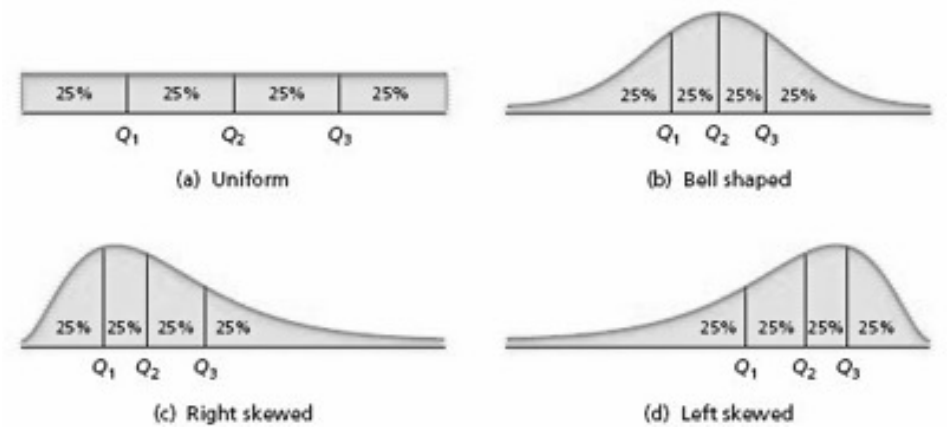
Exploratory data analysis

- Data analysis approach that employs a number of (mostly graphical) techniques to:
  - Maximize insight into data
  - Uncover underlying structure
  - Identify important variables
  - Detect outliers and anomalies
  - Test underlying modeling assumptions
  - Develop parsimonious models
  - **Generate hypotheses from data**

PURDUE UNIVERSITY

Data summarization

- Measures of location
  - Mean: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x(i)$
  - Median: value with 50% of points above and below
  - Quartile: value with 25% (75%) points above and below
  - Mode: most common value

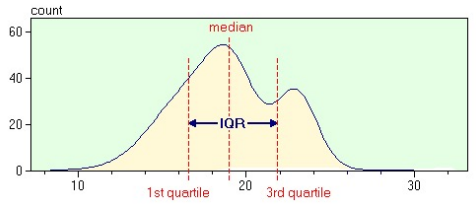


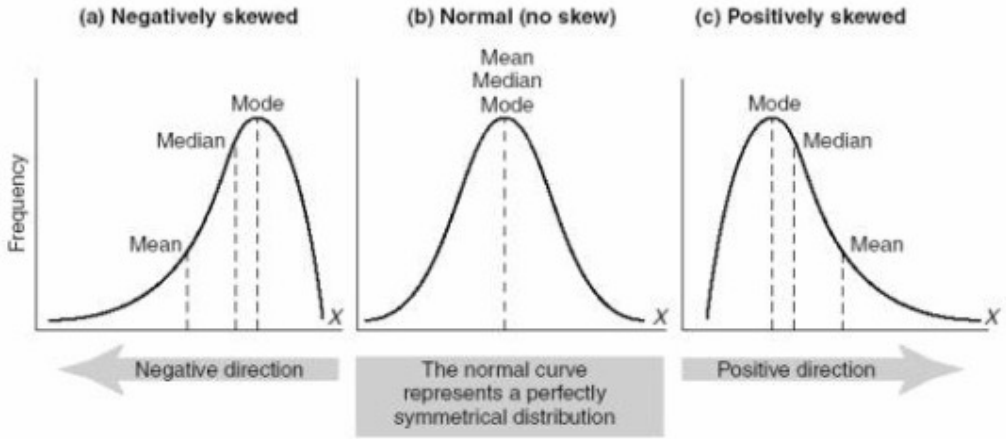
www.brainfuse.com

PURDUE UNIVERSITY

Data summarization

- Measures of dispersion or variability
  - Variance: $\hat{\sigma}_k^2 = \frac{1}{n}\sum_{i=1}^{n}(x(i)-\mu)^2$
  - Standard deviation: $\hat{\sigma}_k = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x(i)-\mu)^2}$
  - Range: difference between max and min point
  - Interquartile range: difference between 1st and 3rd Q
  - Skew: $\mathbf{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] \Leftrightarrow \frac{\sum_{i=1}^{n}(x(i)-\hat{\mu})^3}{(\sum_{i=1}^{n}(x(i)-\hat{\mu})^2)^{\frac{3}{2}}}$

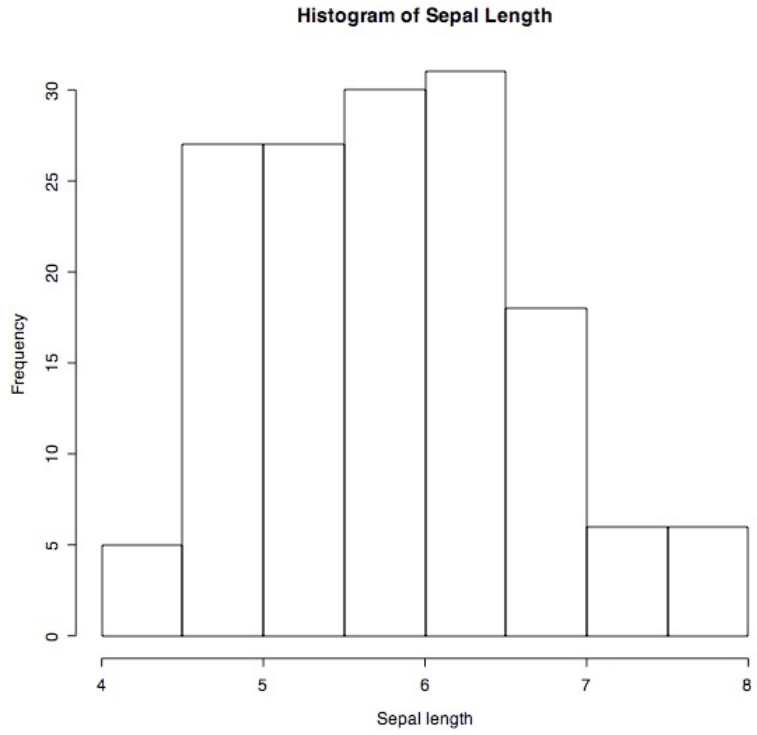




PURDUE UNIVERSITY

Histograms (1D)

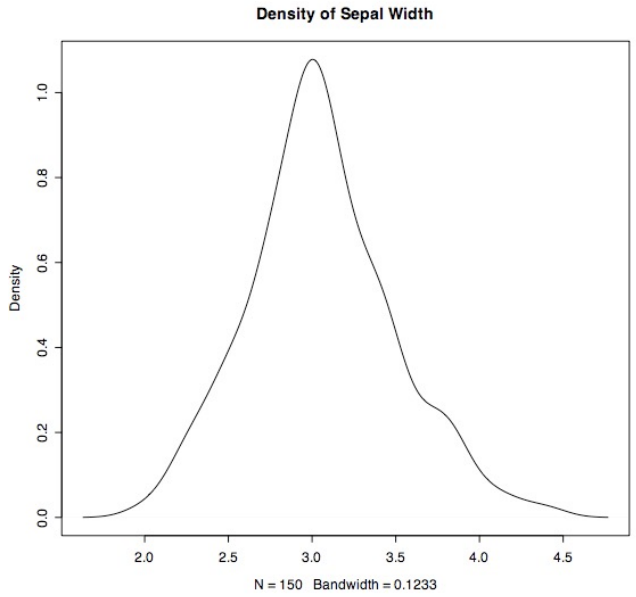
- Most common plot for univariate data
- Split data range into equal-sized bins, count number of data points that fall into each bin
- Graphically shows:
  - Center (location)
  - Spread (scale)
  - Skew
  - Outliers
  - Multiple modes



PURDUE UNIVERSITY

Histogram limitations

- Histograms can be misleading for small datasets
  - Slight changes in the data or binning approach can result in different histograms
- Solution: smoothed density plots
  - Use kernel function to estimate density at each point x, pools information from neighboring points



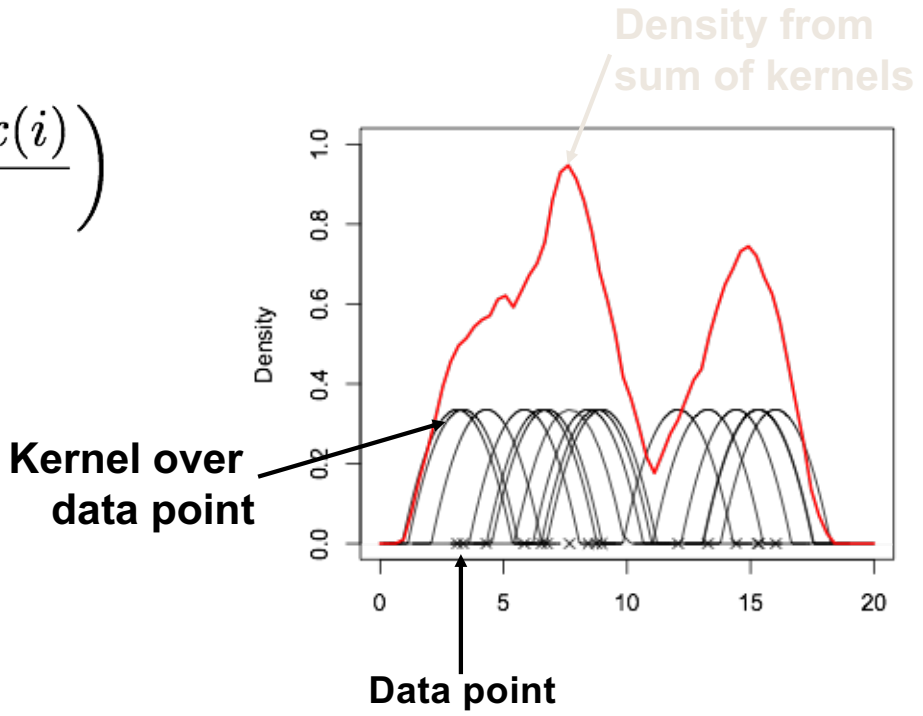
PURDUE UNIVERSITY

Density plots

- Estimated density is:

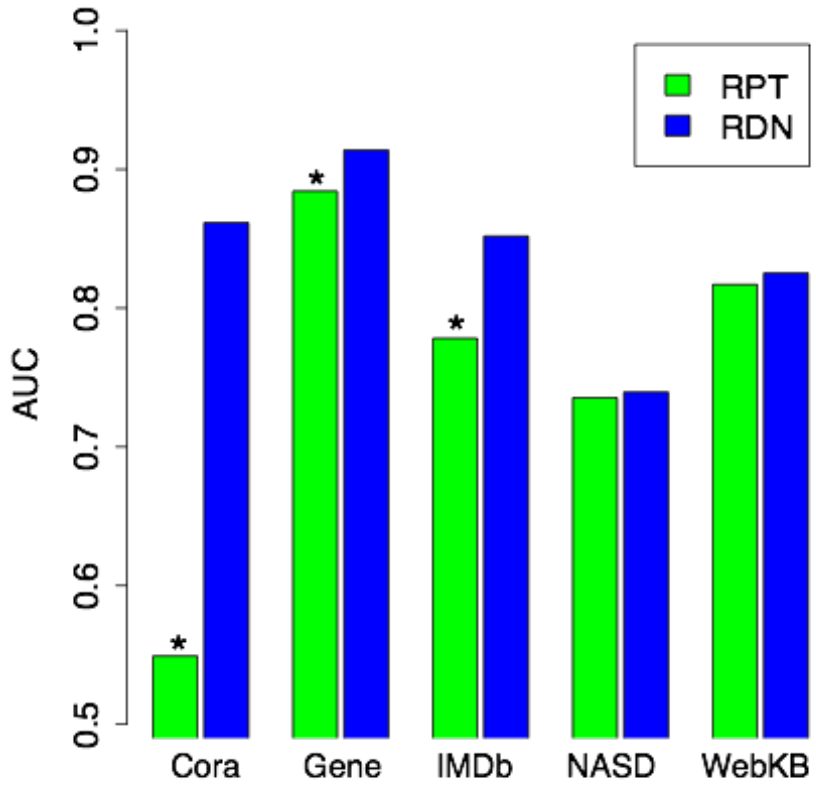
$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - x(i)}{h}\right)$$

- Two parameters:
  - **Kernel function** K (e.g., Gaussian, Epanechnikov)
  - Bandwidth h



PURDUE UNIVERSITY

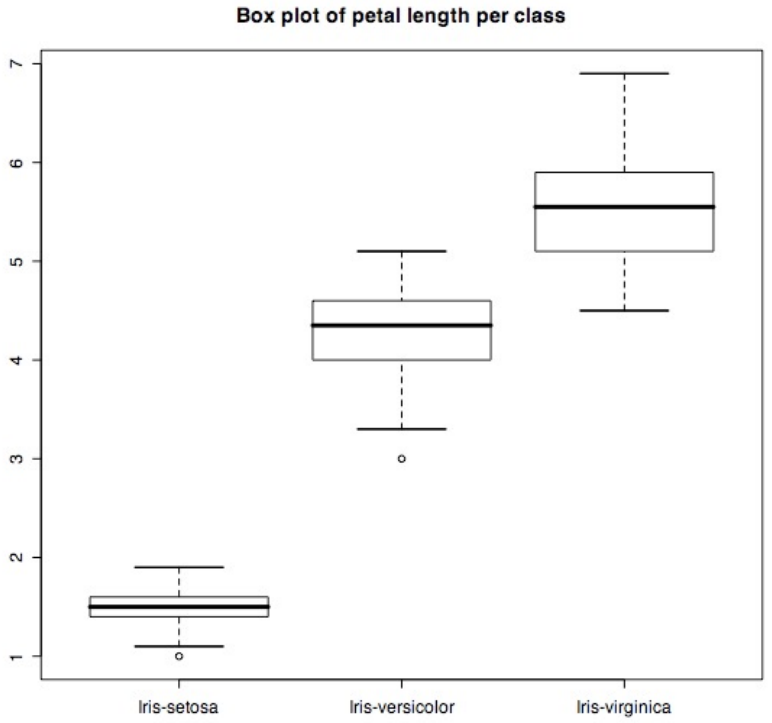
Bar plots



PURDUE UNIVERSITY

Box plot (2D)

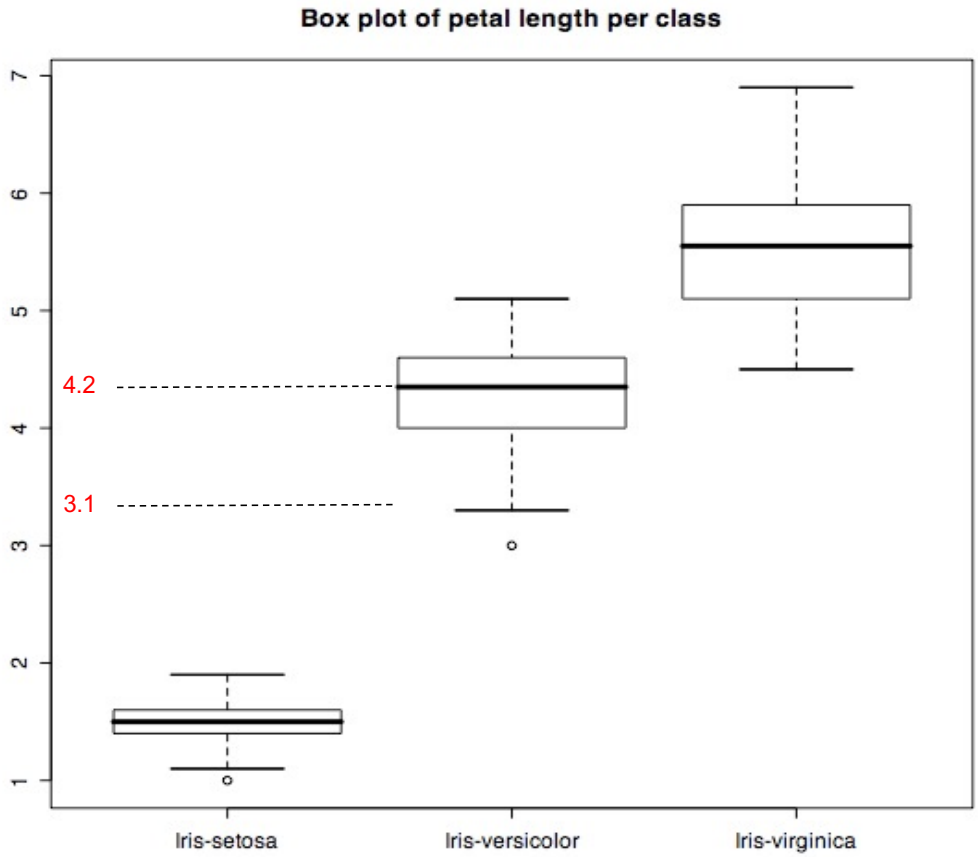
- For each discrete value X, calculate quartiles and range of associated Y values
- Data summary for: *minimum, first quartile, median, third quartile, and maximum*
- Can also plot outliers separately



PURDUE UNIVERSITY

Interpreting Box Plots

- Petals of Iris-Versicolor are:
  - Always longer than 3.1?
  - At least 50% of the petals are longer than 4



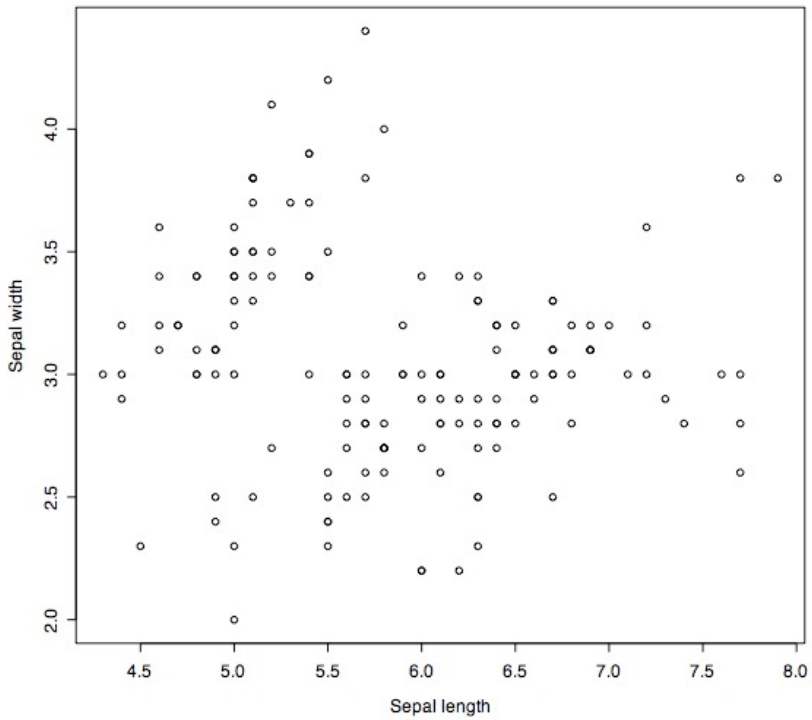
PURDUE UNIVERSITY®

Scatter plot (2D)

- Most common plot for bivariate data
  - Horizontal X axis: the suspected **independent** variable
  - Vertical Y axis: the suspected **dependent** variable
- Graphically shows:
  - If X and Y are related
  - Linear or non-linear relationship
  - If the variation in Y depends on X
  - Outliers

PURDUE UNIVERSITY

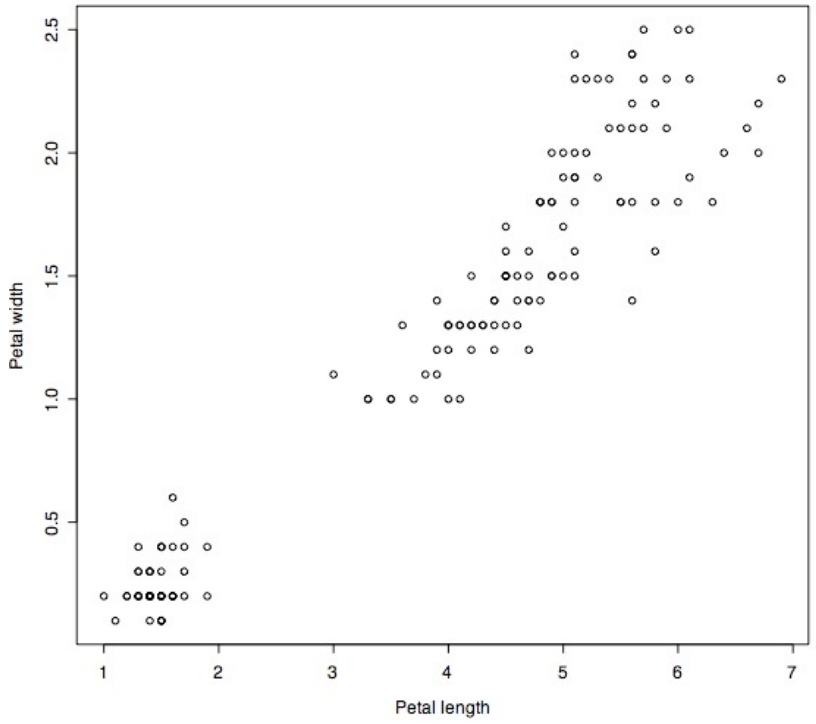
No relationship



Should this variable be excluded or “pre-pruned” when building a decision tree? Why or why not?

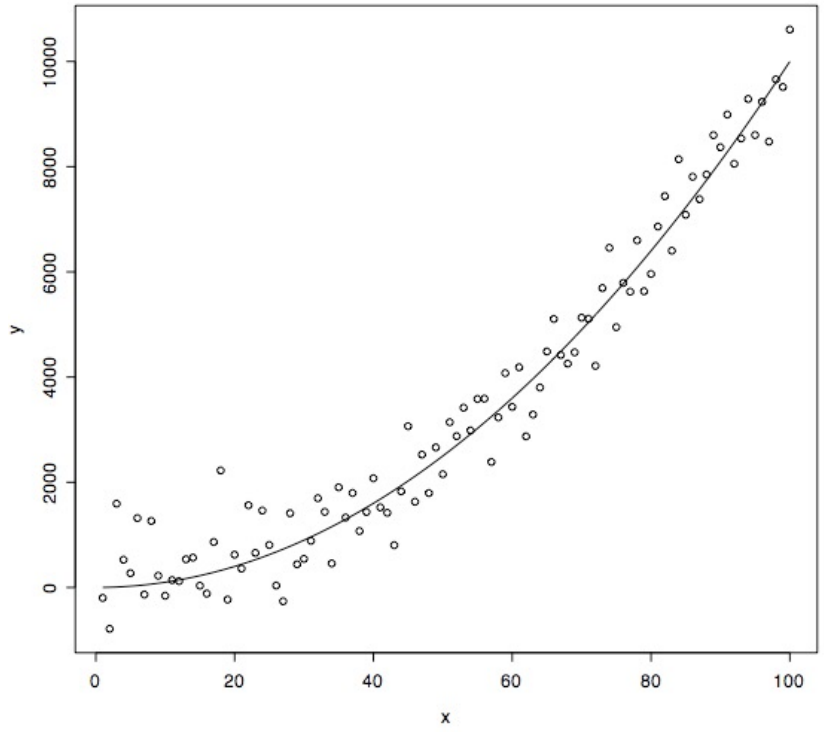
PURDUE UNIVERSITY

Linear relationship



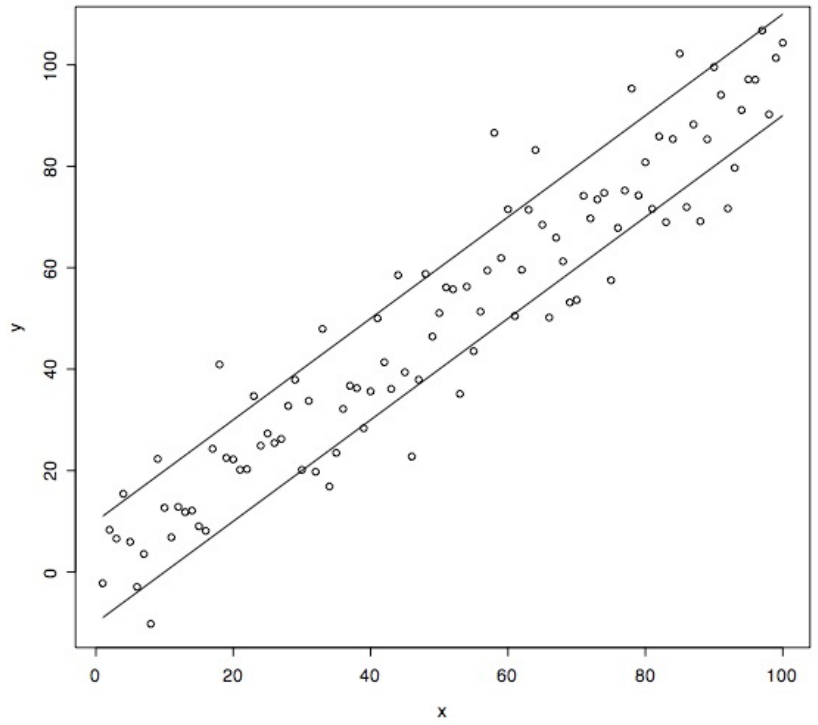
PURDUE UNIVERSITY®

Non-linear relationship



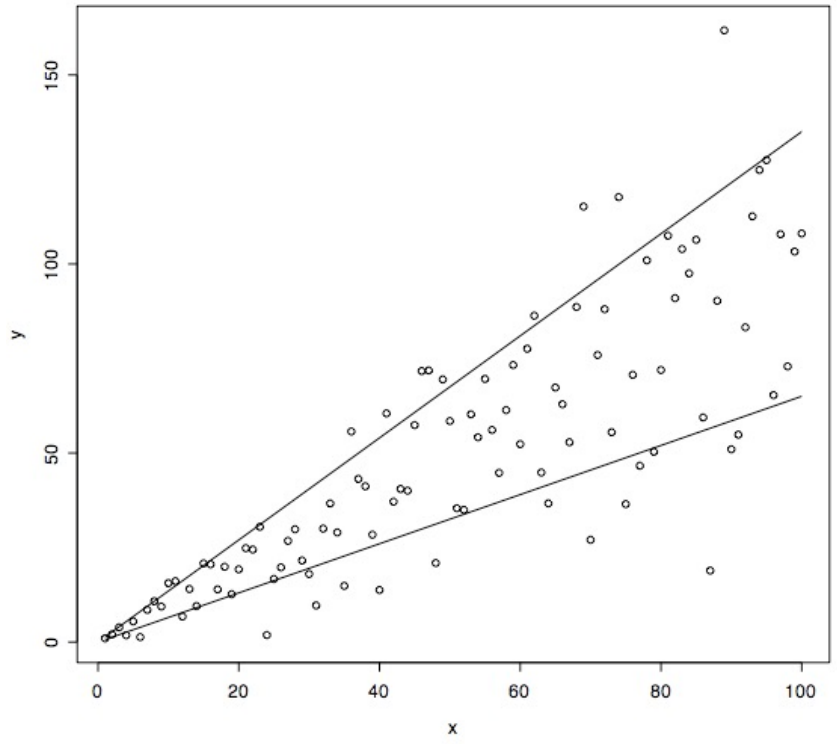
PURDUE UNIVERSITY®

Homoskedastic



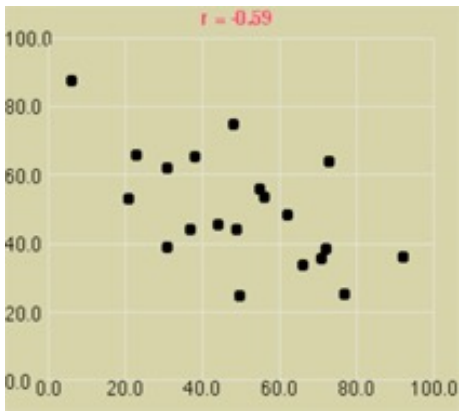
PURDUE UNIVERSITY®

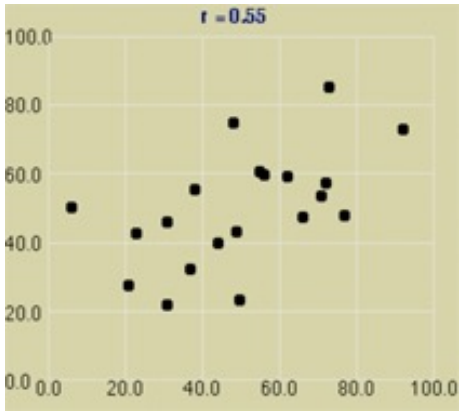
Heteroskedastic

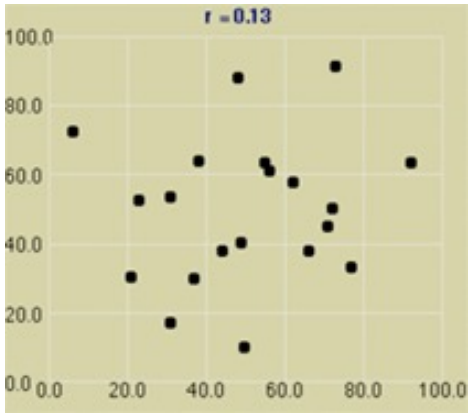


PURDUE UNIVERSITY®

Which one of the plots describes a positive correlation?

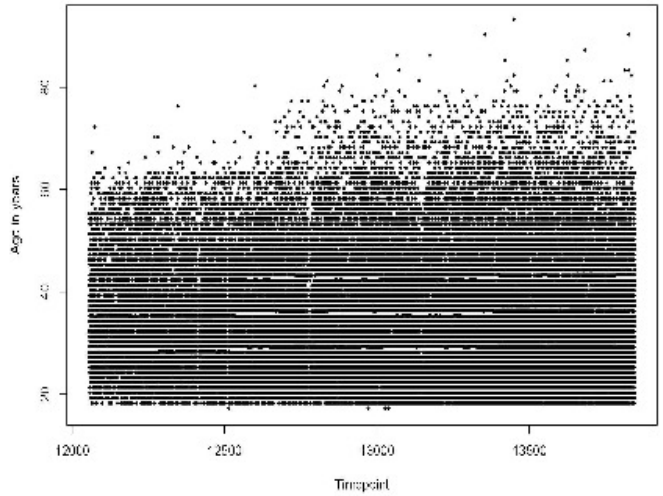




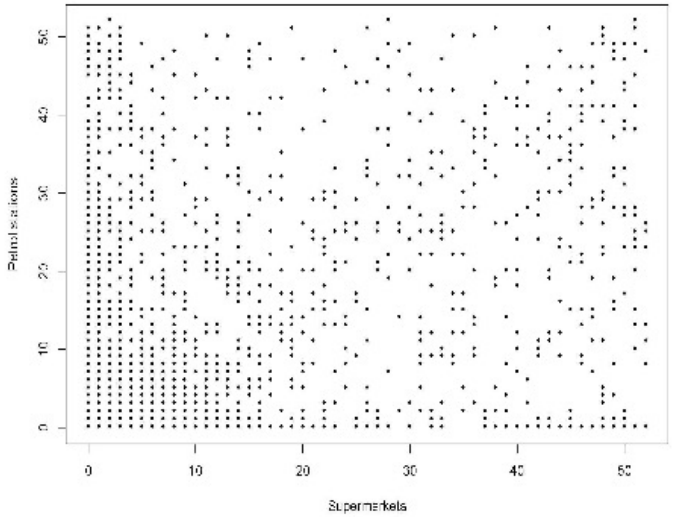


PURDUE UNIVERSITY®

Scatterplot limitations



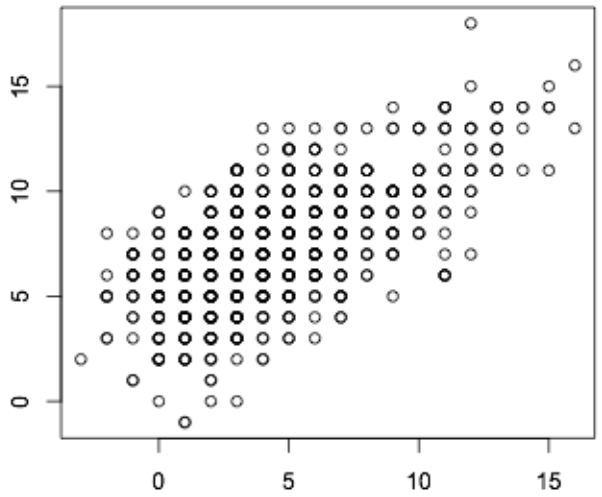
Too much data



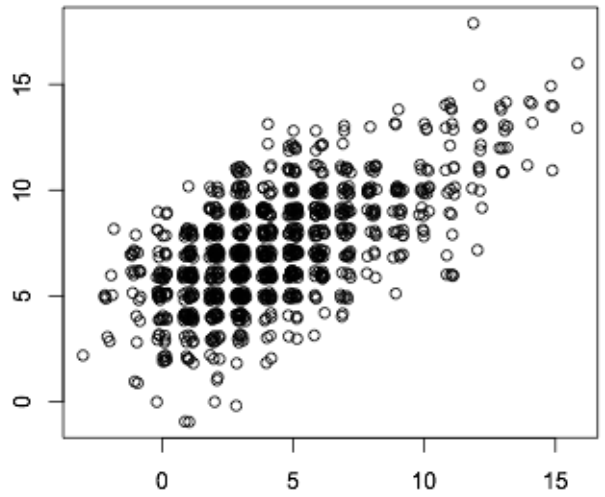
Overprinting

PURDUE UNIVERSITY

Scatterplot limitations



Overprinting

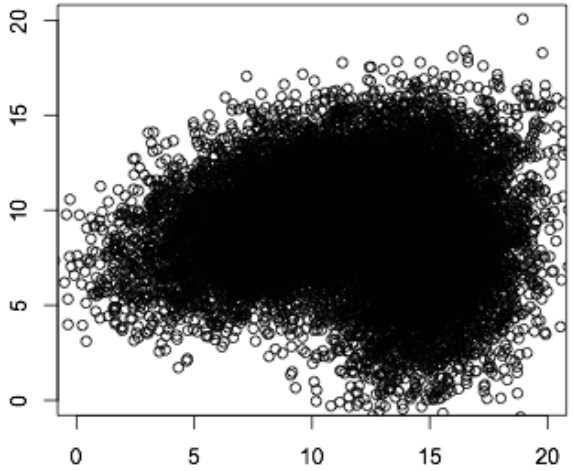


Solution: Jitter points

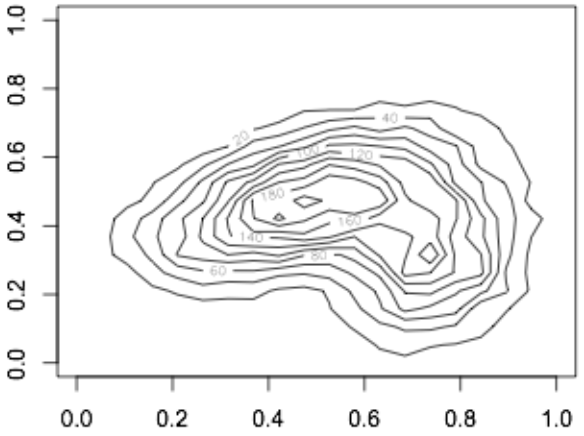
PURDUE UNIVERSITY

Contour plot (3D)

- Limitations of 2D scatterplot (e.g., when there is too much data to discern relationship)



- Solution: represent a 3D surface by plotting constant z slices (contours) in a 2D format



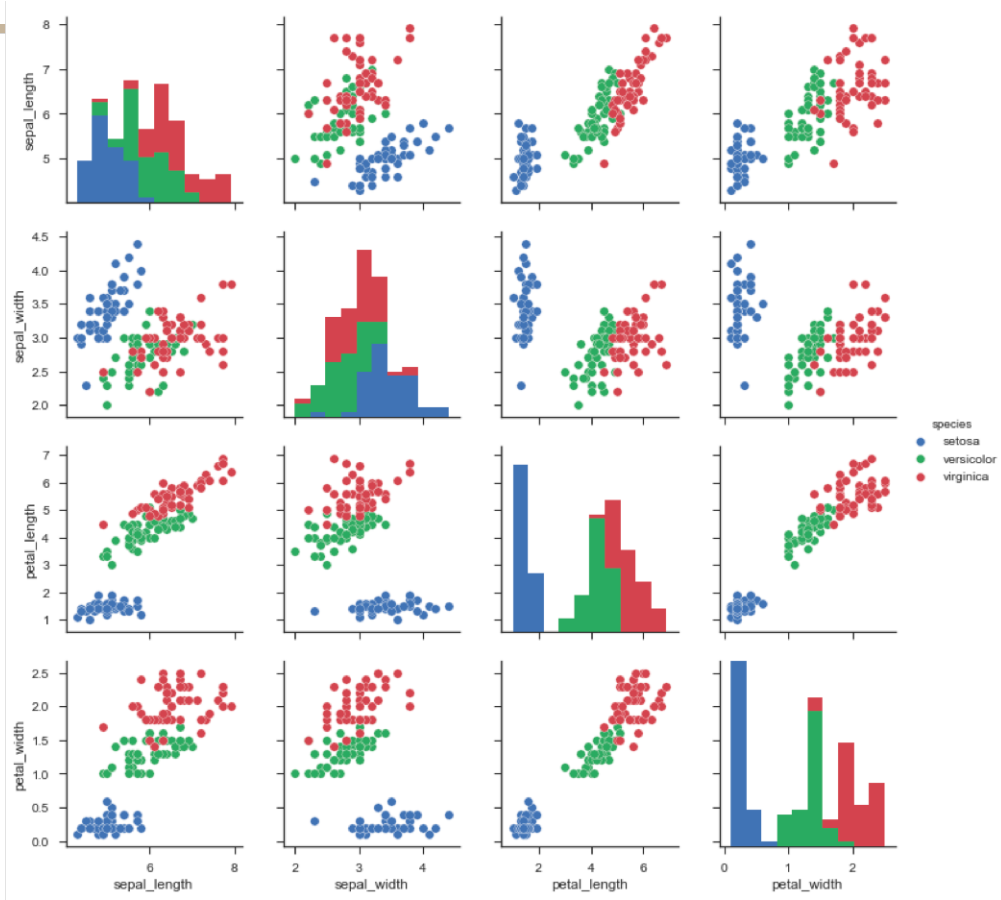
PURDUE UNIVERSITY

Scatterplot matrix

Good to check for
linear relationships
in multivariate datasets

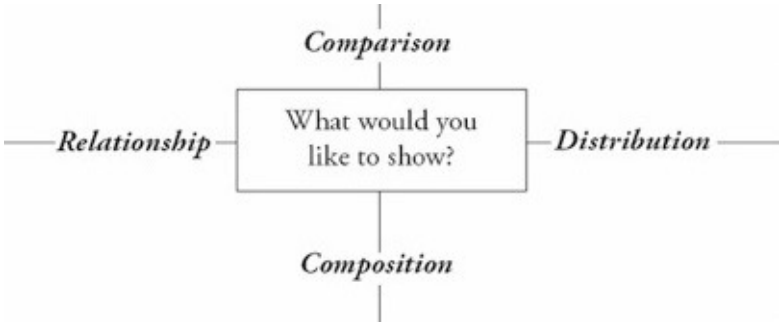
```
import seaborn as sns
import matplotlib.pyplot as plt
sns.set(style="ticks")

df = sns.load_dataset("iris")
sns.pairplot(df, hue="species")
plt.show()
```



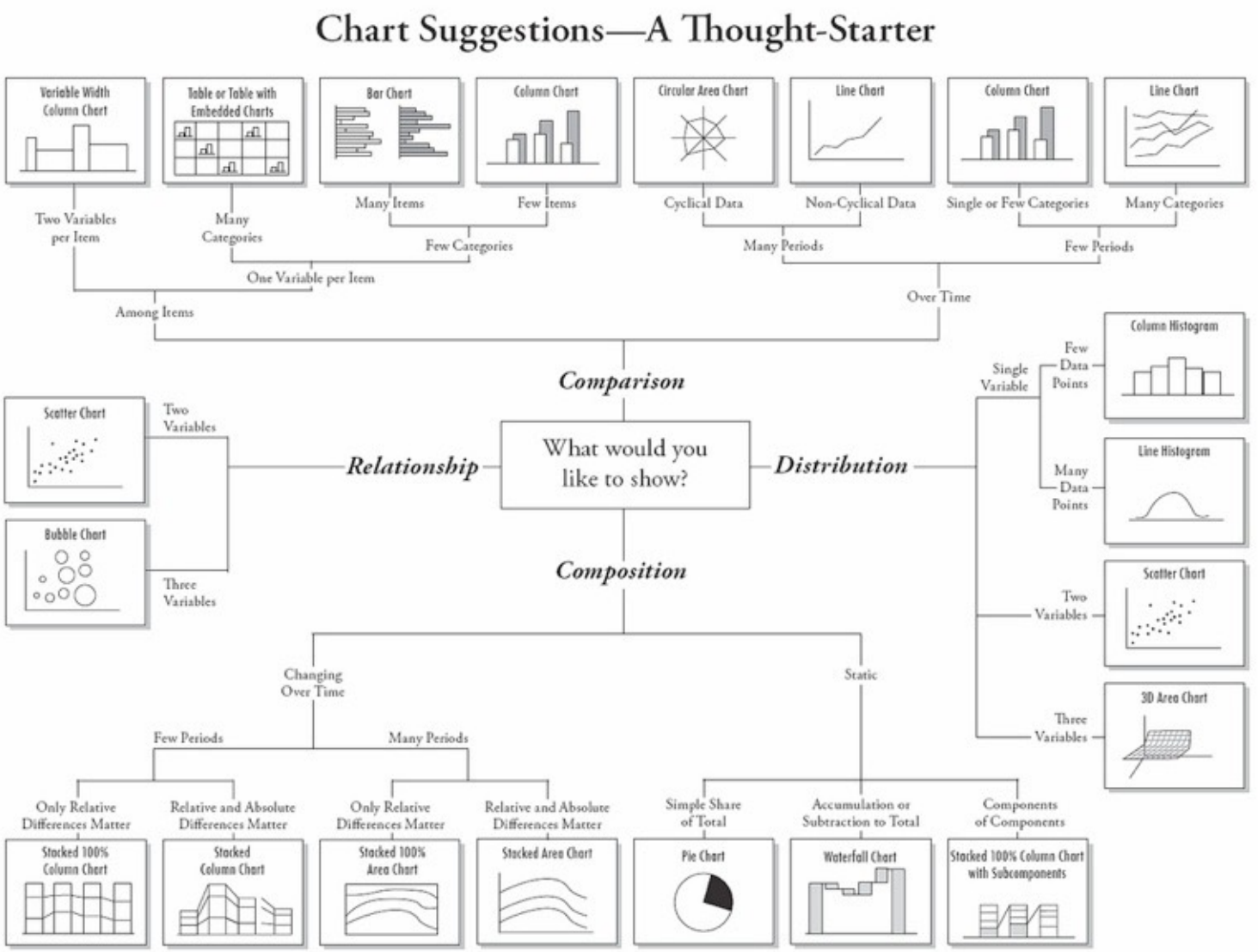
PURDUE UNIVERSITY

Summary – Data Exploration



http://extremepresentation.typepad.com/blog/2006/09/choosing_a_good.html

PURD
UNIVER



http://extremepresentation.typepad.com/blog/2006/09/choosing_a_good.html