Data Mining & Machine Learning

CS37300
Purdue University

September 8, 2017

Data exploration and visualization

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**

Visualization

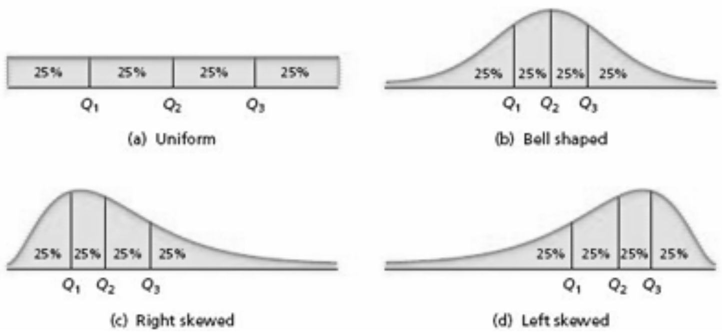
- Human eye/brain have evolved powerful methods to detect structure in nature
- Display data in ways that exploit human pattern recognition abilities
- Limitation: Can be difficult to apply if data size (number of dimensions or instances) is large

Visualizing/summarizing data

- Low-dimensional data
    - Summarizing data with simple statistics
    - Plotting raw data (1D, 2D, 3D)
- Higher-dimensional data
    - Principal component analysis
    - Multidimensional scaling

Data summarization

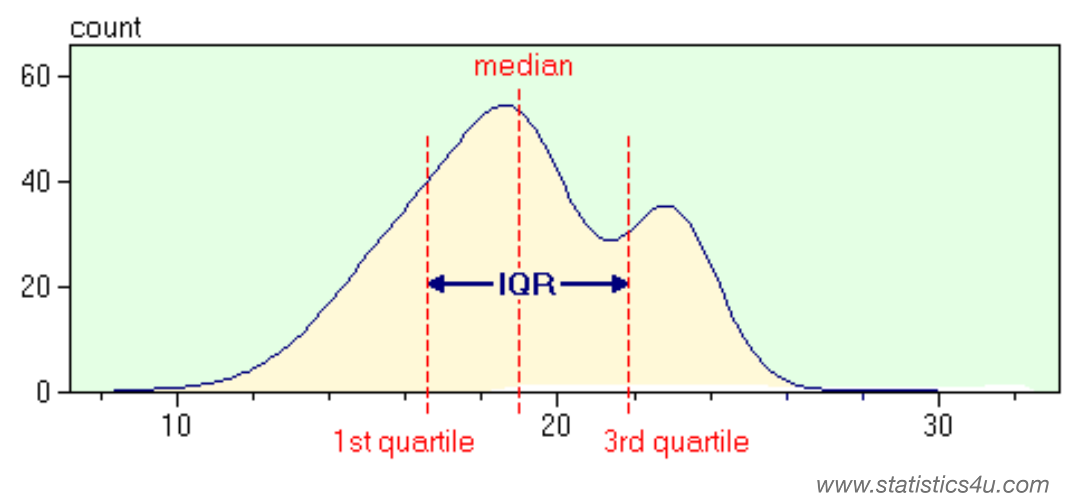
- Measures of location
  - Mean: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x(i)$
  - Median: value with 50% of points above and below
  - Quartile: value with 25% (75%) points above and below
  - Mode: most common value



www.brainfuse.com

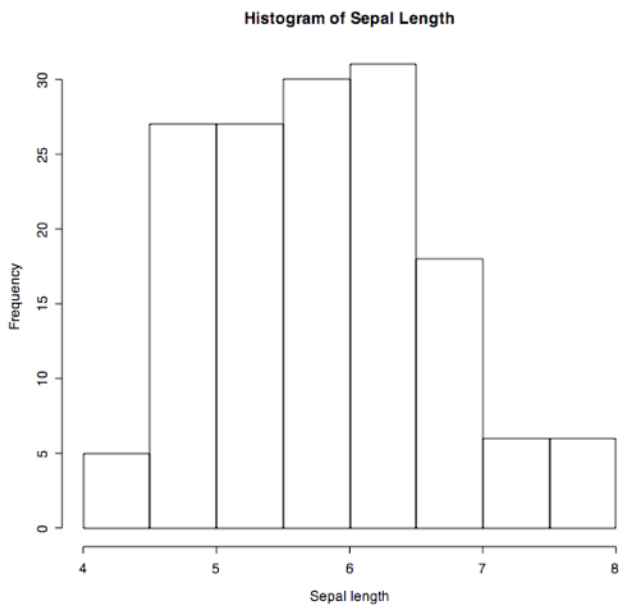
Data summarization

- 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: $\frac{\sum_{i=1}^{n}(x(i)-\hat{\mu})^3}{(\sum_{i=1}^{n}(x(i)-\hat{\mu})^2)^{\frac{3}{2}}}$

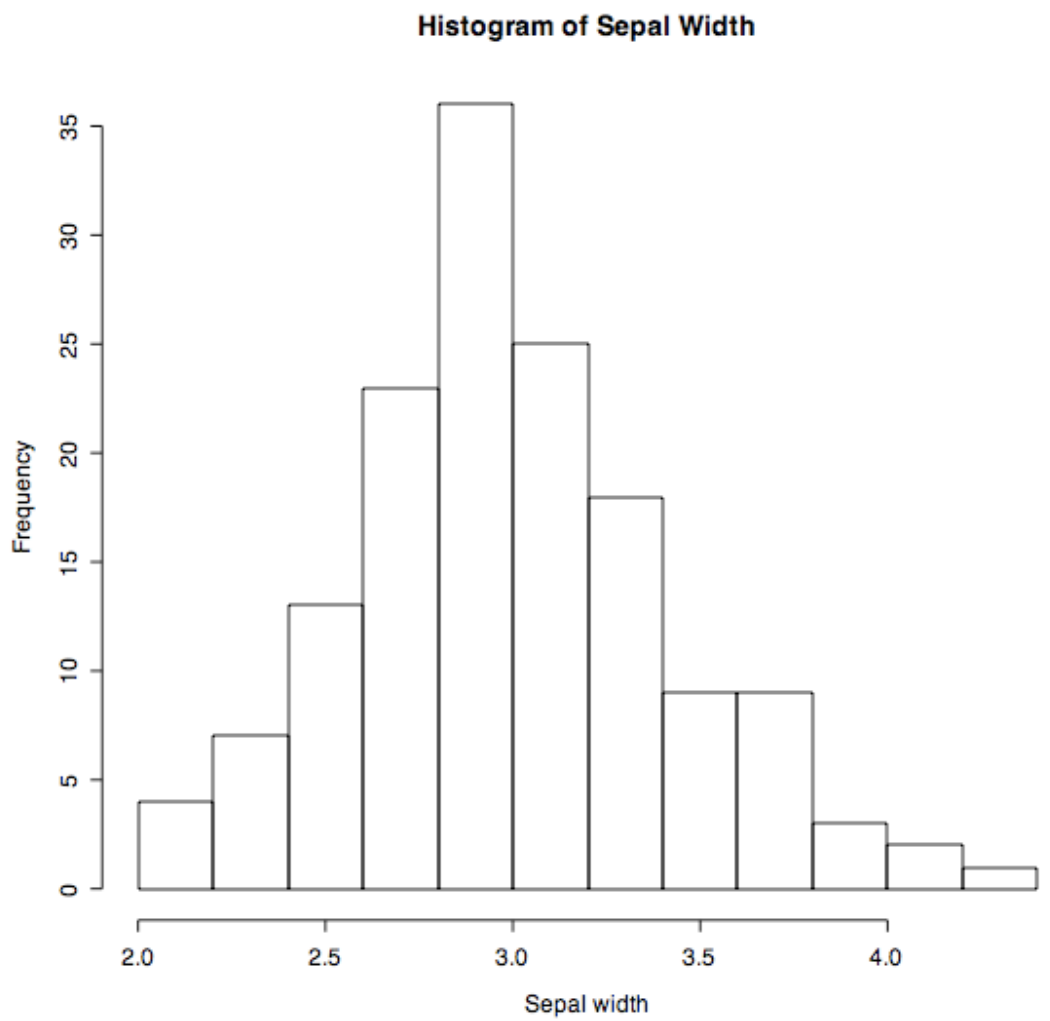


Histograms (1D)

- 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



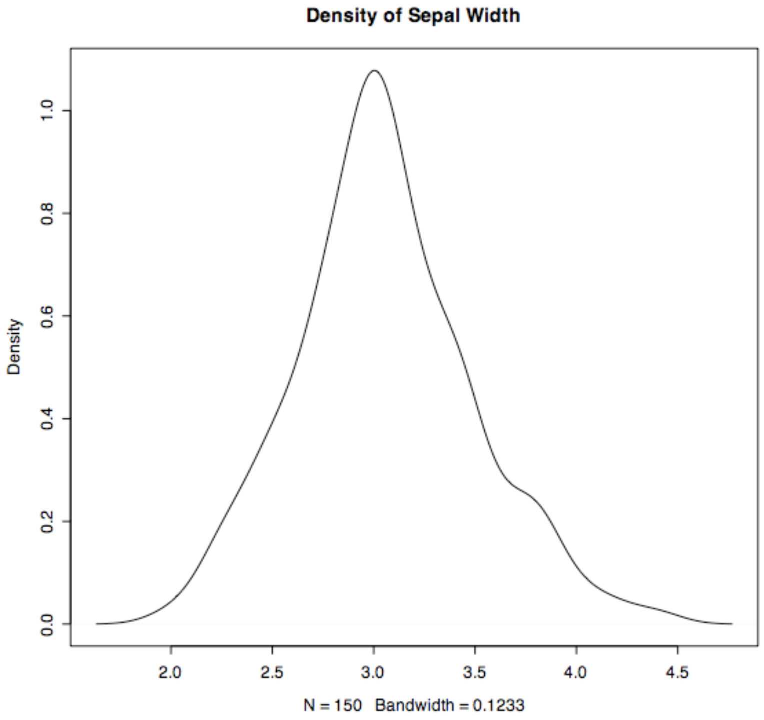
Example histogram



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

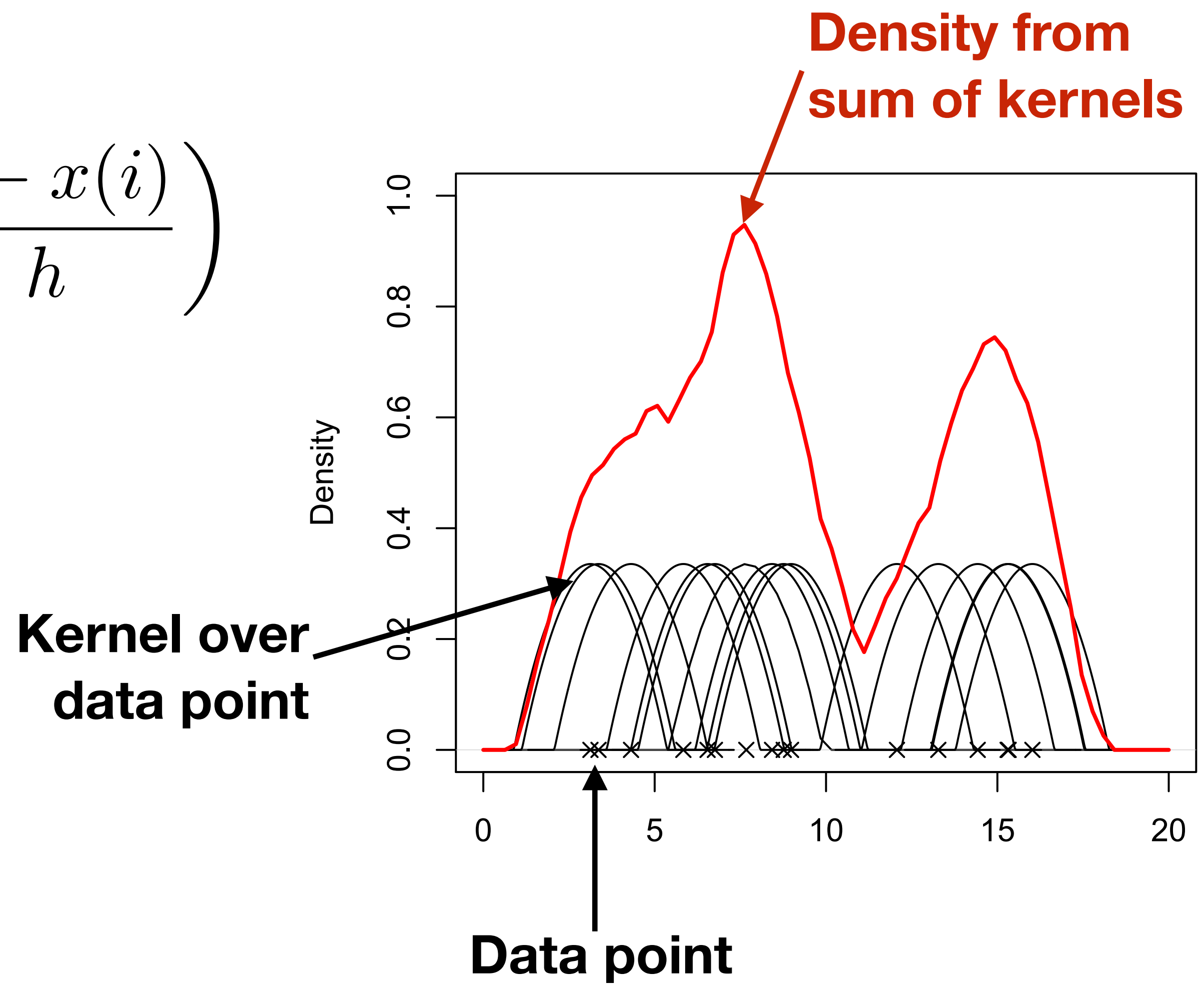


Density plots

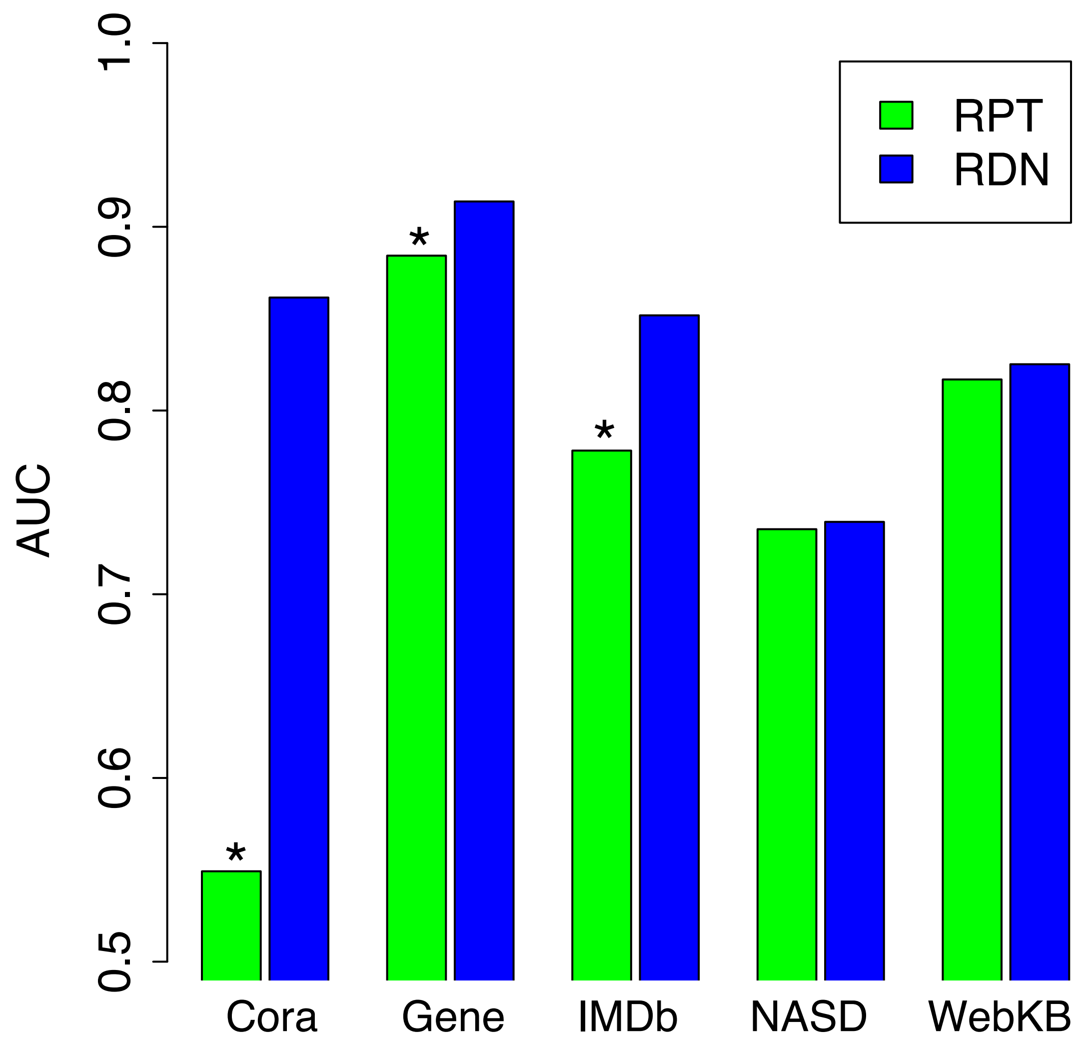
- Estimated density is:

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

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

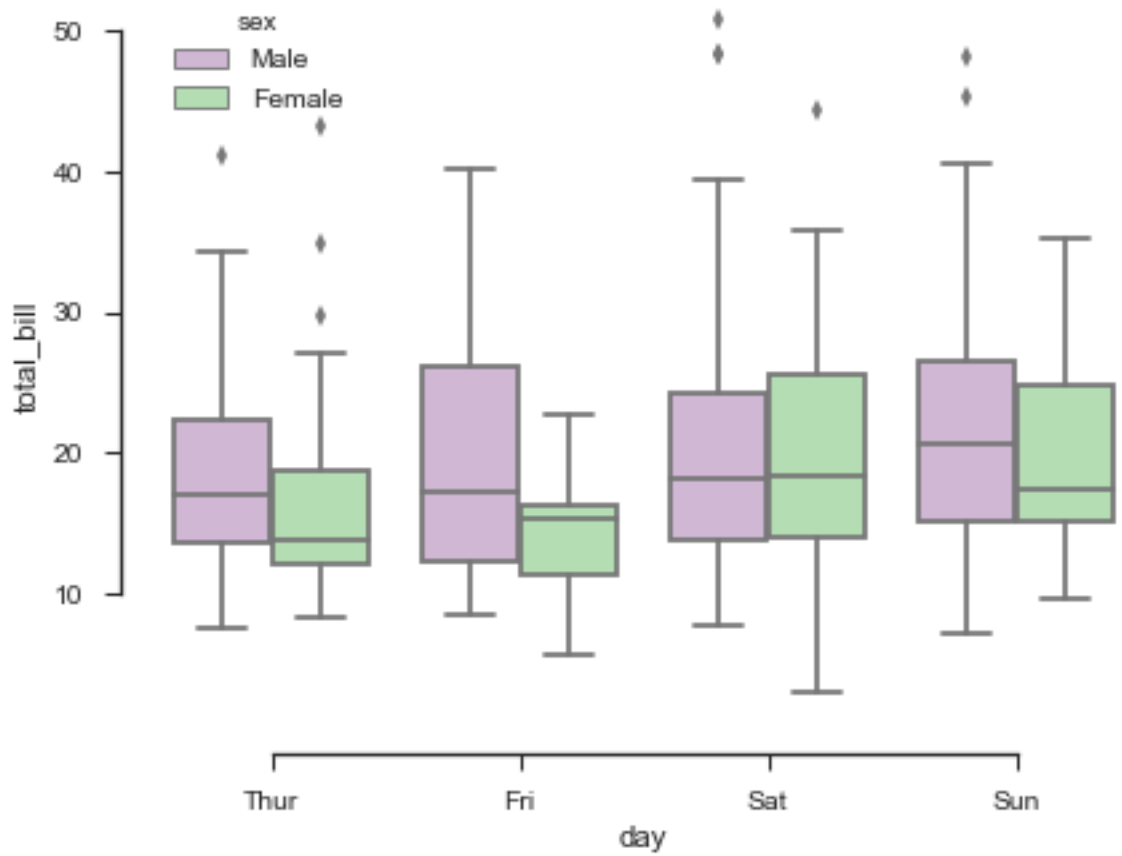


Bar plots



Box plot (2D)

- Display relationship between discrete and continuous variables
- For each discrete value X, calculate quartiles and range of associated Y values

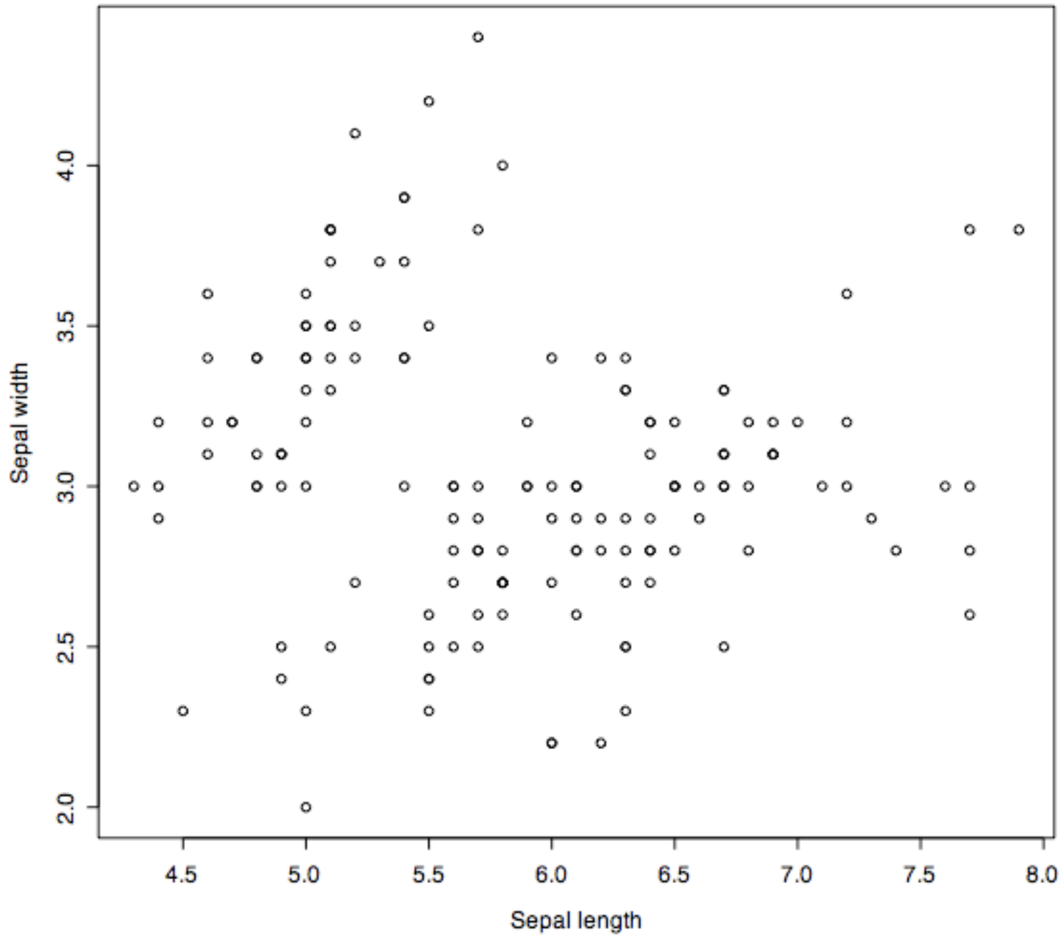



```
import seaborn as sns
import matplotlib.pyplot as plt
sns.set(style="ticks")

# Load the example tips dataset
tips = sns.load_dataset("tips")

# Draw a nested boxplot to show bills by day and sex
sns.boxplot(x="day", y="total_bill", hue="sex", data=tips, palette="PRGn")
sns.despine(offset=10, trim=True)
plt.show()
```

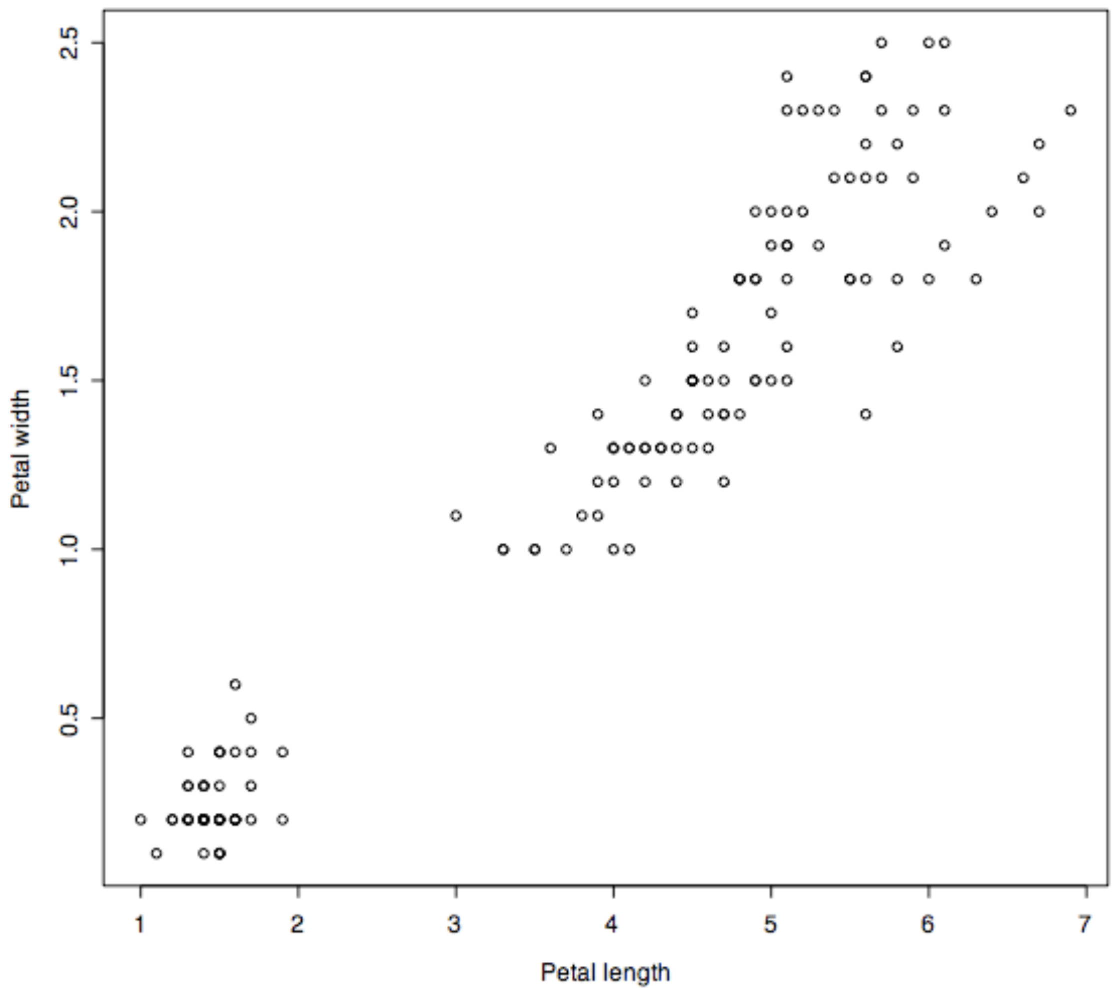
Scatter plot (2D)

- 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

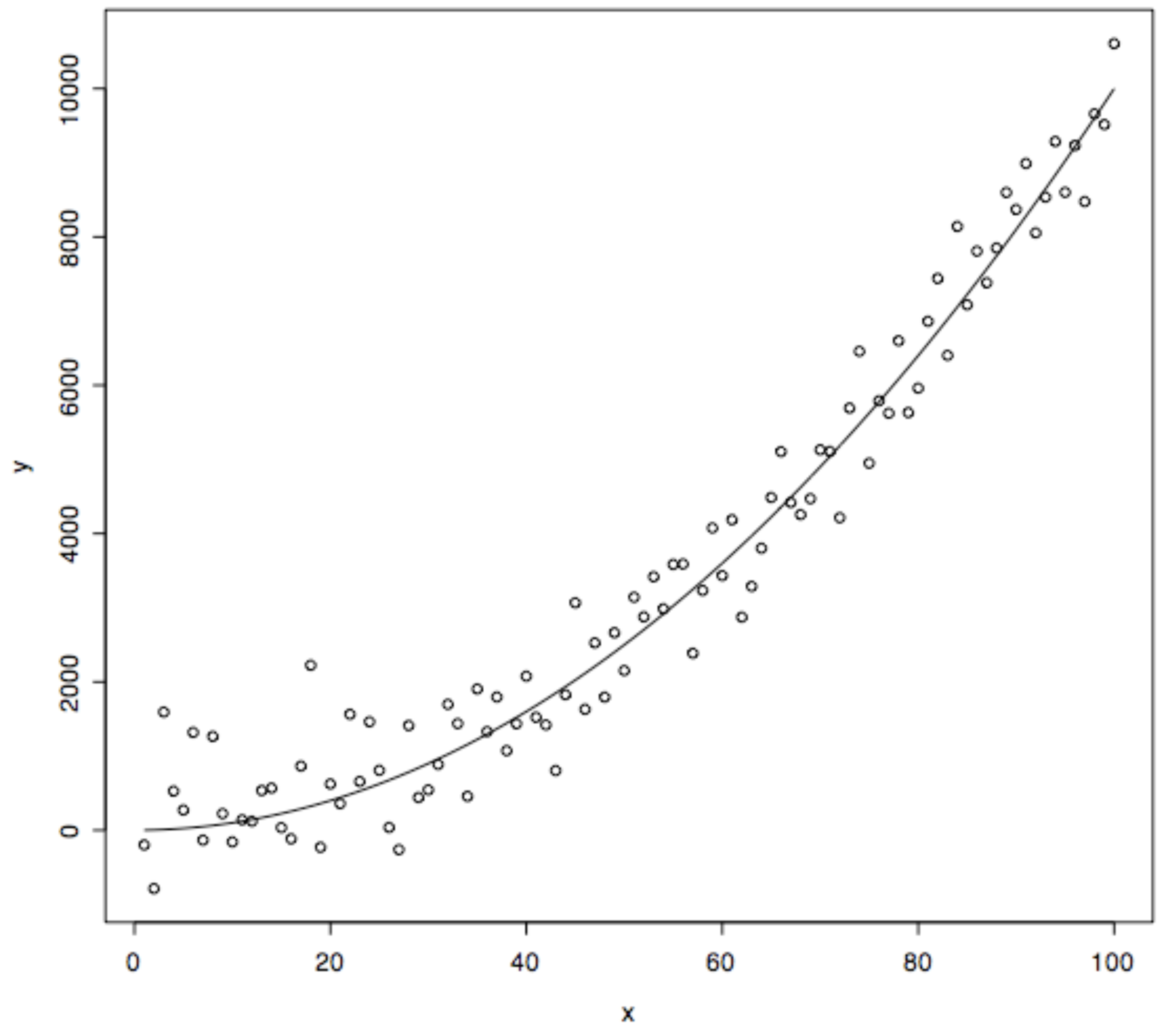
No relationship



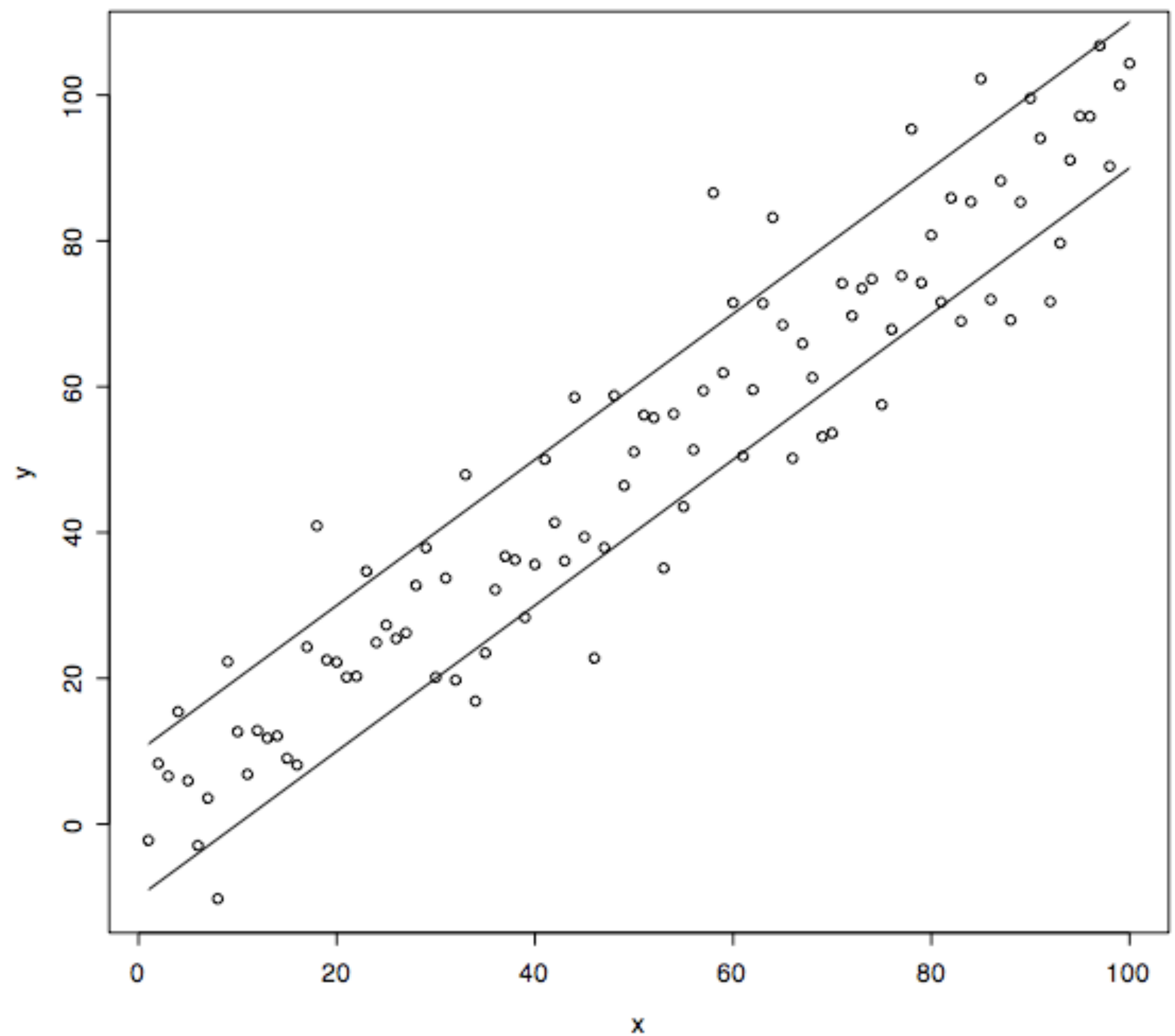
Linear relationship



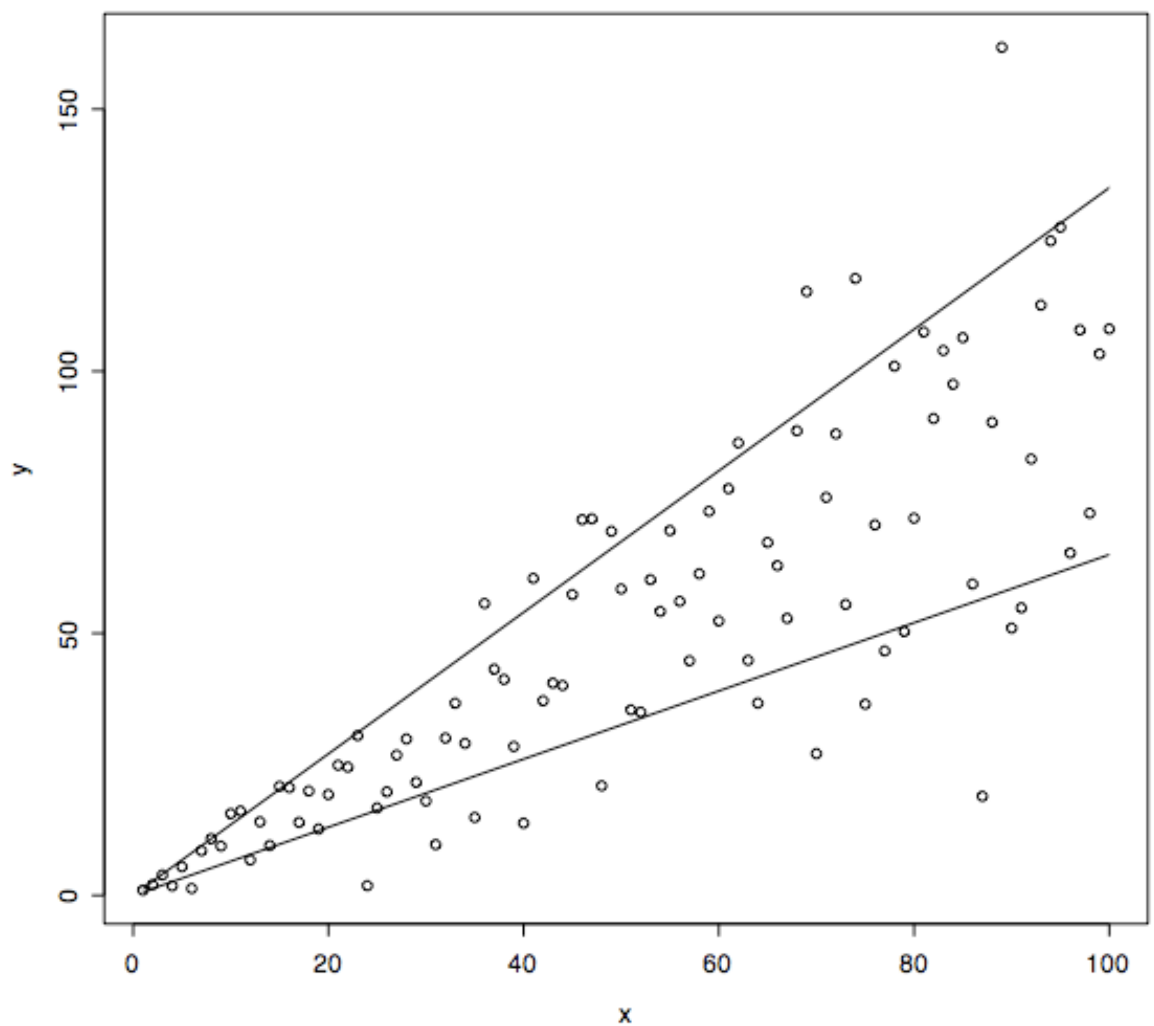
Non-linear relationship



Homoskedastic



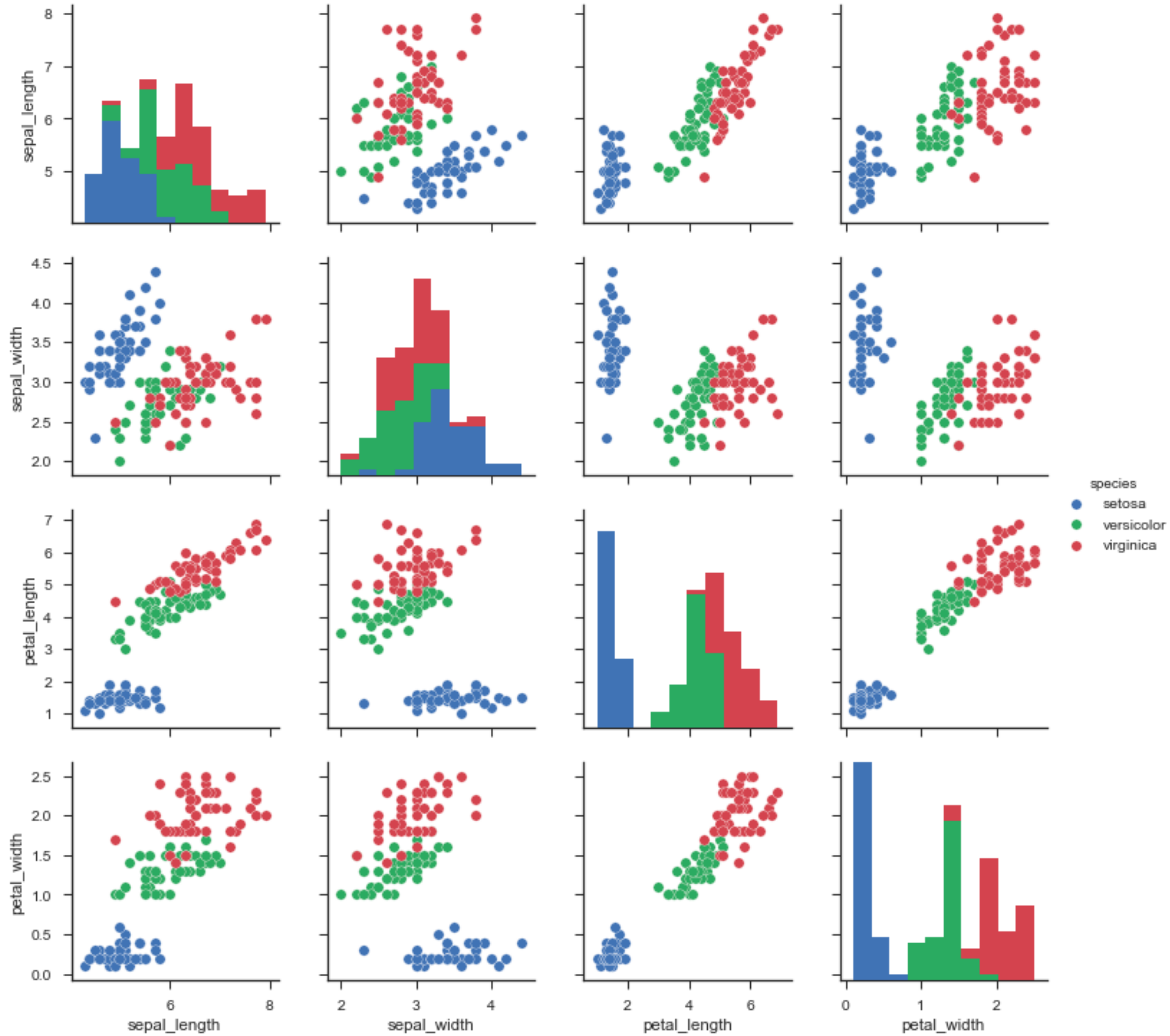
Heteroskedastic



Scatterplot matrix

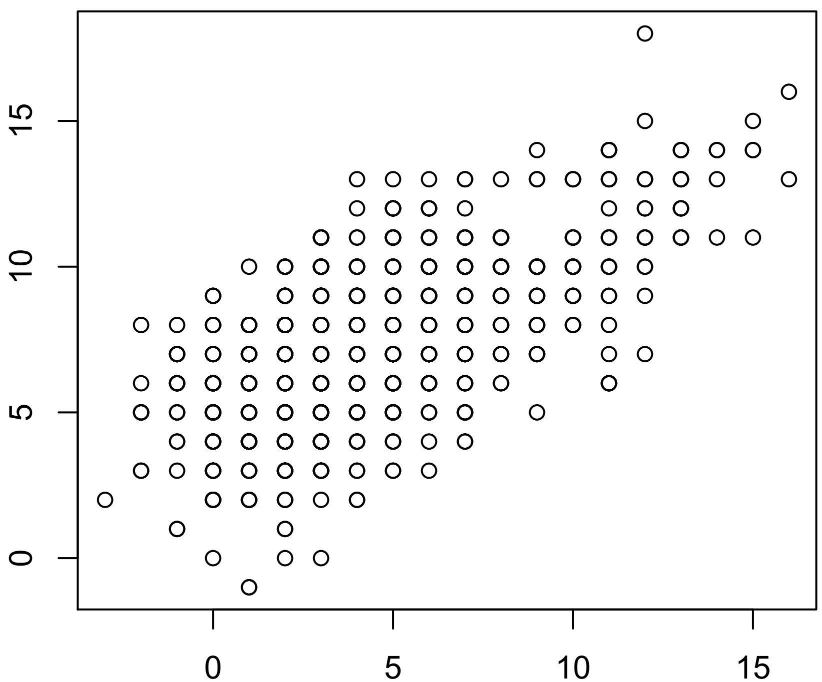
http://seaborn.pydata.org/generated/seaborn.pairplot.html

Good to check for
linear relationships
in multivariate datasets

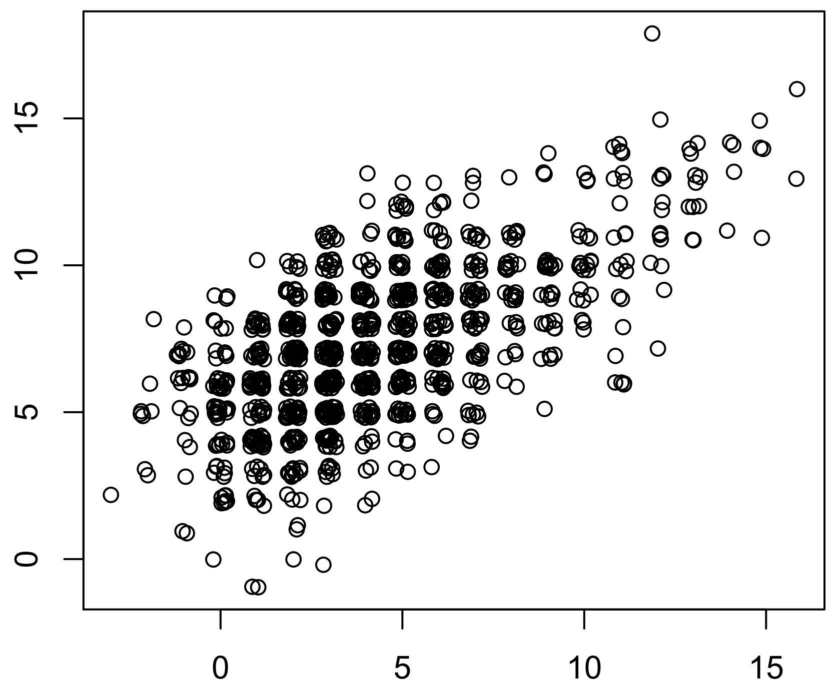



```
import seaborn as sns
import matplotlib.pyplot as plt
sns.set(style="ticks")

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

Scatterplot limitations



Overprinting



Solution: Jitter points

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
- Contour with KDE