STATS 415: Exploring Data, Part 2

Prof. Liza Levina

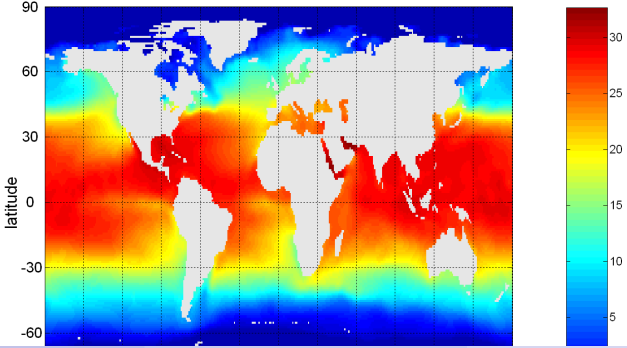
Department of Statistics, University of Michigan

Visualization

- Visualization can mean 2d or 3d plots, movies, or sometimes even well-designed tablesi
- One of the most powerful and appealing techniques for data exploration:
- Humans have a well developed ability to analyze large amounts of information that is presented visually.
  - Can detect general patterns and trends
  - Can detect unusual patterns and outliers

Example: a heatmap of sea surface temperature

- July sea surface temperature across the world: a heatmap
- Tens of thousands of data points in a single figure
- Color represents value; coordinates represent location
- In general, heatmaps are great for plotting 3 variables at a time
- It is important to choose the right range and the right color scheme for your data; always include the color bar!



Representation

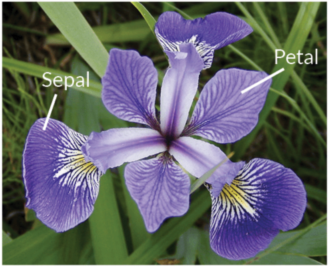
- Data objects, their variables, and the relationships among data objects are translated into graphical elements such as points, lines, shapes, and colors.
- Objects are often represented as points.
- Variable values can be represented as position of the points (x, y, z coordinates) or the characteristics of the points, e.g., color, size, and shape.
- Position is especially useful for seeing clusters and outliers
- Big data warning: can only look at a few variable at a time, and thus patterns may not translate from plot to plot
- Better visualization with dimension reduction techniques - later this term

Selection

- Plotting all the data often does not lead to good visualization
- Choosing a subset of variables
  - In a supervised problem, might want to choose the variables most correlated with response
  - In an unsupervised problem, may select higher variance variables
  - Dimensionality reduction does this in a principled way
  - Can always consider pairs of variables
- Choosing a subset of data points
  - A region of the screen can only show so many points before becoming a mess
  - Can sample, but want to preserve points in sparse areas
  - Sometimes with discrete variables can add jitter to values make a plot look better
- Choosing the range of variables (axes): the axes can change the visual message

Iris Data Example

- Historically important; first classification algorithm by Fisher
- Available in R
- Three iris types: Setosa, Virginica, Versicolour
- Four variables: sepal width and length, petal width and length



Iris Versicolor

Iris Setosa

Iris Virginica

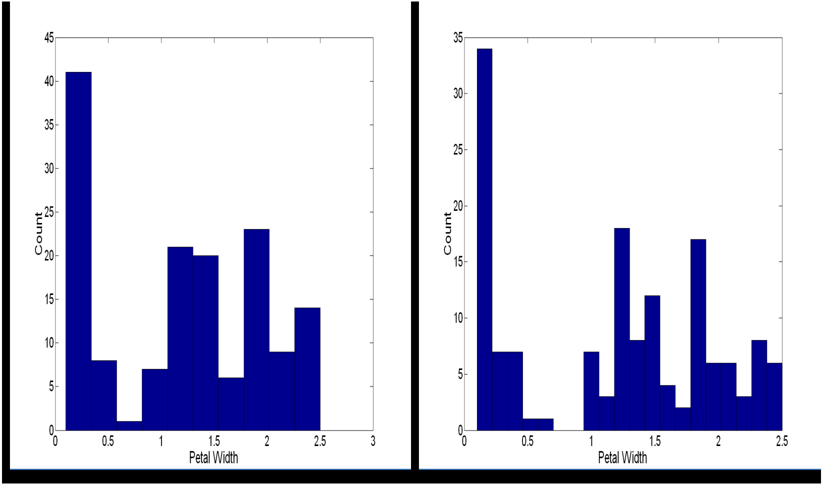
Pima Indians Data Example

- Data collected on 768 adult female Pima Indians
- Variables: number of times pregnant, plasma glucose concentration, diastolic blood pressure, skin fold thickness, 2-hour serum insulin, body mass index, diabetes pedigree function (a continuous score), age, and a test for diabetes

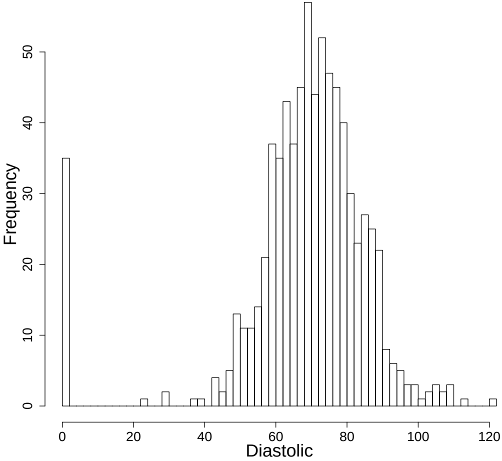
Histogram

- Usually shows the distribution of values of a single variable
- The height of each bar indicates the number of objects (or proportion or percentage or density).
- Shape of histogram depends on the number of bins; need to balance level of detail with noise

Iris data: petal width histogram (10 and 20 bins)

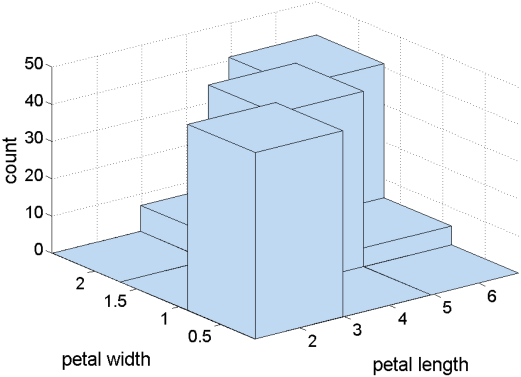


Pima data: diastolic pressure histogram



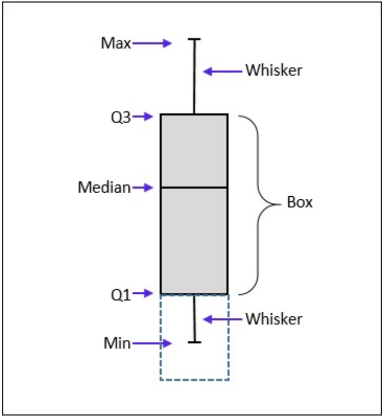
Two-Dimensional Histogram

- Shows the joint distribution of two attributes
- Example: petal width and petal length



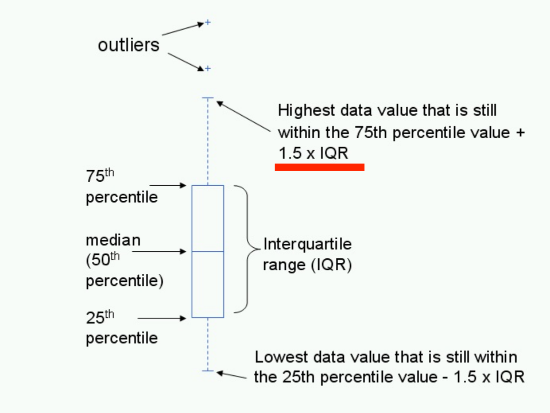
Boxplot

- Invented by Tukey
- Another way of displaying the distribution of data
- A simple boxplot: 5-number summary, Min, 25th percentile (1st quartile), Median, 75th percentile (3rd quartile), Max



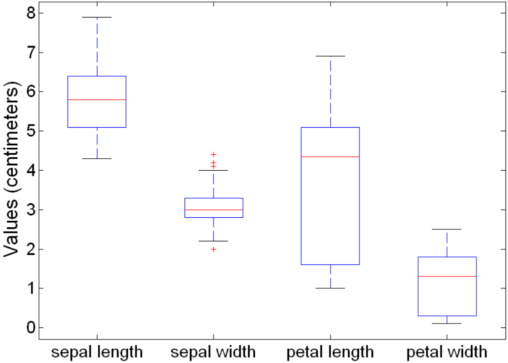
Boxplot

- A boxplot with outliers
- Length of whiskers = multiplier of IQR can be changed in R

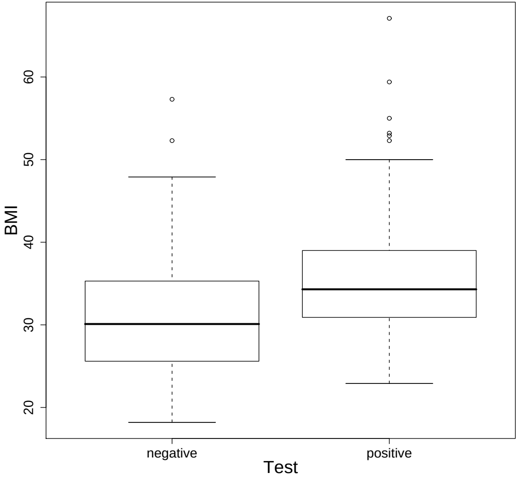

outliers
Highest data value that is still within the 75th percentile value + 1.5 x IQR
75th percentile
median (50th percentile)
25th percentile
Interquartile range (IQR)
Lowest data value that is still within the 25th percentile value - 1.5 x IQR

Examples of boxplots

- Box plots are useful for comparing variables.



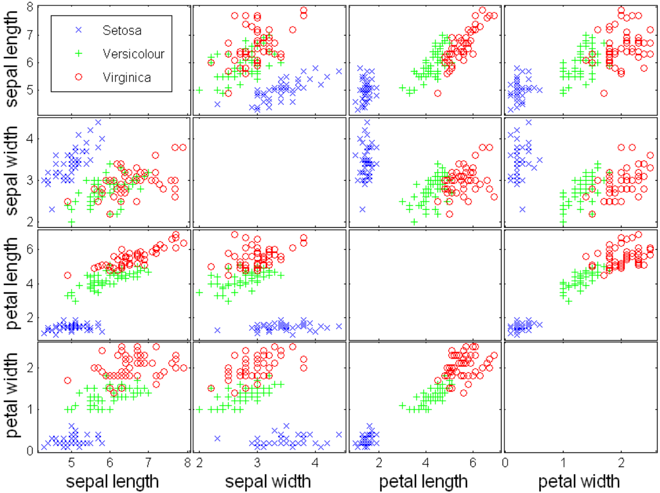
In R, plotting a quantitative variable against a categorical one produces side-by-side boxplots



Scatter Plot

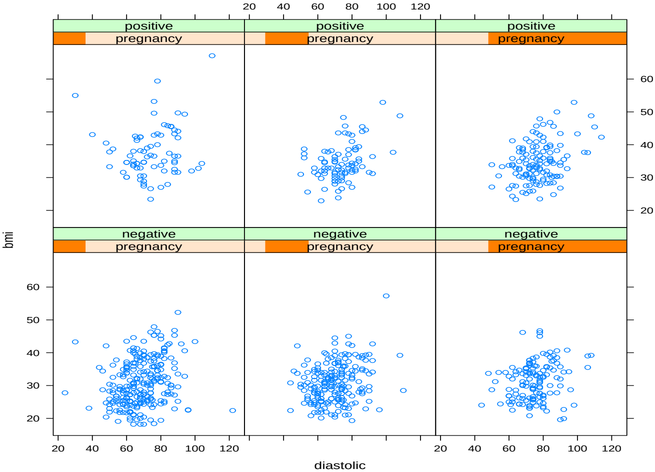
- Variable values determine the position.
- Two-dimensional scatter plots are the most common, but can have three-dimensional scatter plots
- Additional variables can be displayed by using the size, shape, and color of the markers that represent the objects.
- Arrays of scatter plots are useful for compactly summarizing relationships of multiple pairs of variables.

Scatter Plot Array of Iris Variables



Trellis Plot

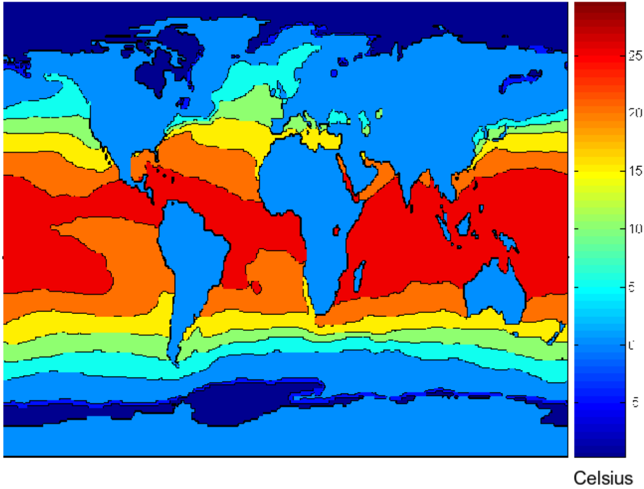
- Fix a particular pair of variables that is to be displayed and produce a series of scatter plots conditioned on levels of one or more other variables
- Can also produce other types of plots, such as histograms, time series plots, contour plots, etc.


positive
pregnancy
negative
bmi
diastolic
20
40
60
80
100
120
20
30
40
50
60

Contour plot

- Useful when a continuous variable is measured on a spatial grid.
- Partition the plane into regions of similar values.
- The contour lines that form the boundaries of these regions connect points with equal values.
- The most common example is contour maps of elevation.
- Can also display temperature, rainfall, air pressure, etc: sea surface temperature.

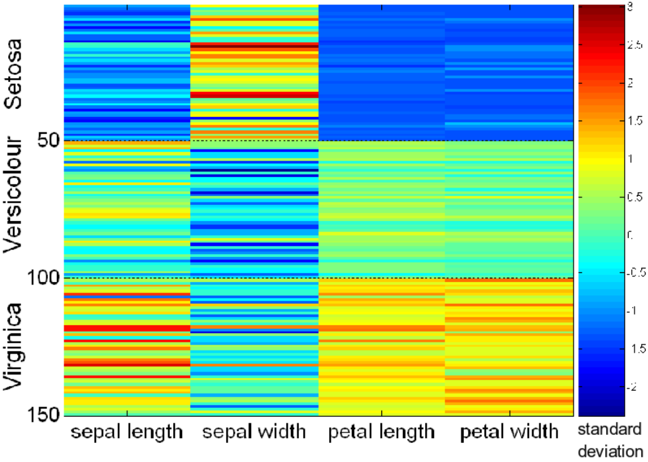
Contour plot of sea surface temperature



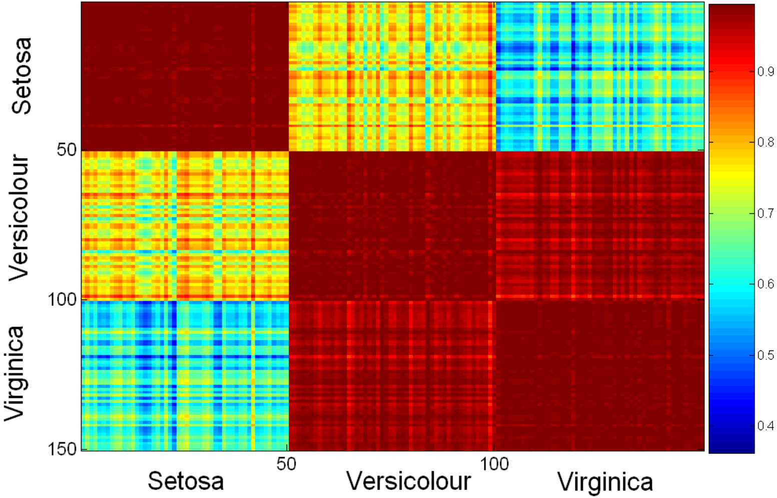
Matrix Plot

- Can plot the data matrix
- This can be useful when objects are sorted according to class.
- Typically, the variables are normalized to prevent one variable from dominating the plot.
- Plots of similarity or distance matrices can also be useful for visualizing the relationships between objects.

Visualization of the Iris Data Matrix



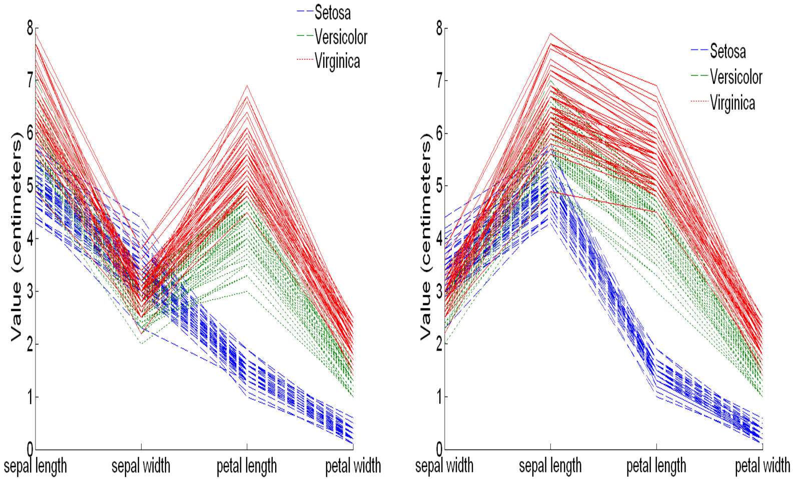
Visualization of the Iris Similarity Matrix



Parallel Coordinates Plot

- Use a set of parallel axes, one for each varialbe
- The variable values corresponding to the same data point are connected by a line.
- Can see whether the lines separate into groups and along which variables
- Ordering of variables can be important.

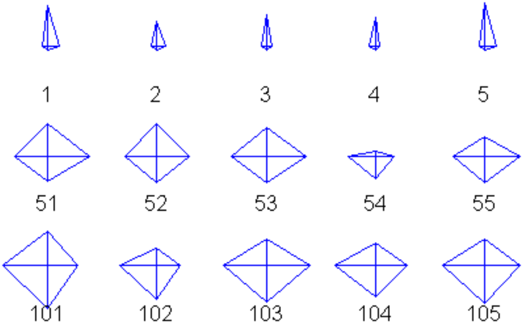
Parallel Coordinates Plot for Iris Data



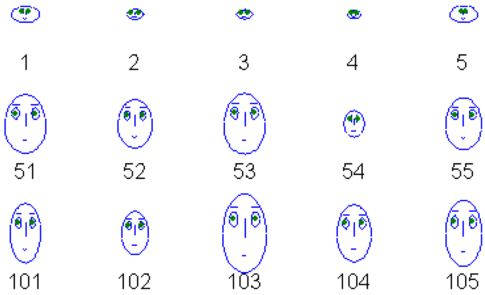
Old-school visualization techniques

- Star plots
  - Axes radiate from a central point.
  - Each object becomes a polygon.
- Chernoff faces
  - This approach associates each variable with a characteristic of a face.
  - The values of each variable determine the appearance of the corresponding facial characteristic.
  - Each object becomes a separate face.

Star Plots for Iris Data



Chernoff Faces for Iris Data



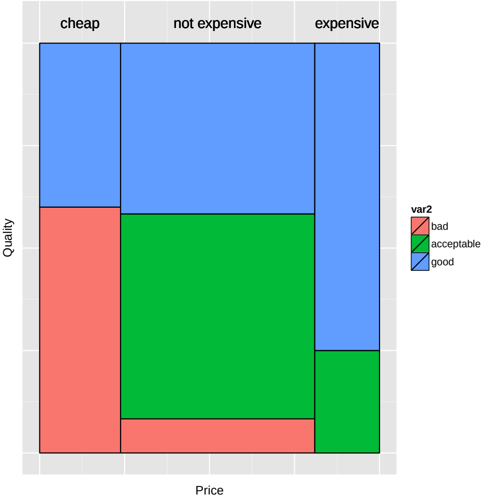
Categorical data example: sleeping bags

- The variables are price, fiber and quality for 21 sleeping bags
- All variables are categorical; cannot do a scatter plot or side-by-side boxplots.
- With a few categorical variables, data are often best summarized in a table, but may still want a picture

	cheap	not expensive	expensive	down fibers	synthetic fibers	good	acceptable	bad
Brand	**Price**			**Fiber**		**Quality**		
One Kilo Bag	1	0	0	0	1	1	0	0
Sund	1	0	0	0	1	0	0	1
Kompakt Basic	1	0	0	0	1	1	0	0
Finmark Tour	1	0	0	0	1	0	0	1
Interlight Lyx	1	0	0	0	1	0	0	1
Kompakt	0	1	0	0	1	0	1	0
Touch the Cloud	0	1	0	0	1	0	1	0
Cat's Meow	0	1	0	0	1	1	0	0
Igloo Super	0	1	0	0	1	0	0	1
Donna	0	1	0	0	1	0	1	0
Tyin	0	1	0	0	1	0	1	0
Travellers Dream	0	1	0	1	0	1	0	0
Yeti Light	0	1	0	1	0	1	0	0
Climber	0	1	0	1	0	0	1	0
Viking	0	1	0	1	0	1	0	0
Eiger	0	0	1	1	0	0	1	0
Climber light	0	1	0	1	0	1	0	0
Cobra	0	0	1	1	0	1	0	0
Cobra Comfort	0	1	0	1	0	0	1	0
Foxfire	0	0	1	1	0	1	0	0
Mont Blanc	0	0	1	1	0	1	0	0

A panel plot sleeping bag data

One panel can only display two variables at a time, but can make multiple panels



Summary

- Visualization is extremeley important
- Different problems require different tools
- Selection of all kinds matters: it can make better plots, but can also be used to “lie with statistics”
- Pretty plots in R are made with the `ggplot` package (taught in Stats 306); you are welcome to use it if you know it but this class will not cover it.