

Plot and Look!

Trust your Data more than your Models

martin@theusRus.de

www.theusRus.de

Telefónica O₂ Germany

Augsburg University

Outline

- Why use Graphics for Data Analysis?
- Foundations of Interactive Statistical Graphics
- Escaping Flatland ... ways to cope with multidimensionality
- Statistification of Graphics
- The Mondrian graphical data analysis tool
- *Demo*

Why use Graphics (anyway) ?

- Classical ➤ **Presentation**

The most common use of graphics is clearly in presenting qualitative or quantitative results to a broad audience

- Statistical ➤ **Diagnostics**

In statistics, graphics are often used to check the quality and properties of statistical procedures or models

- Analytical ➤ **Exploration**

During the exploration process of an analysis, graphics aid to generate insights and deduce properties and relationships

- Essential ➤ **Data Cleaning**

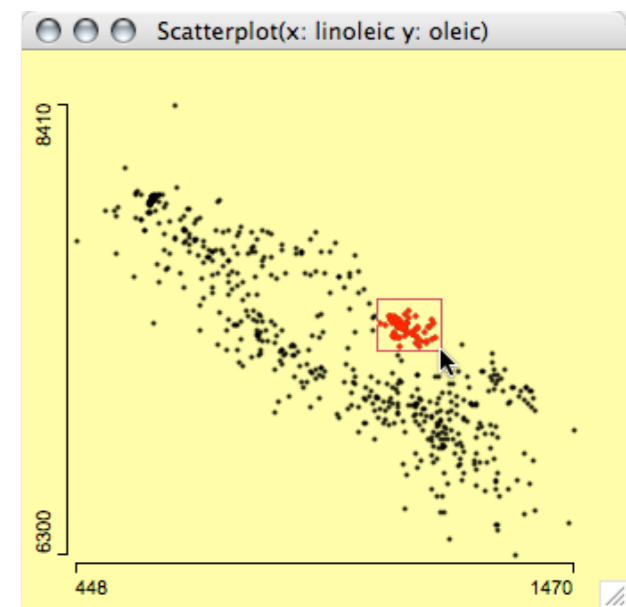
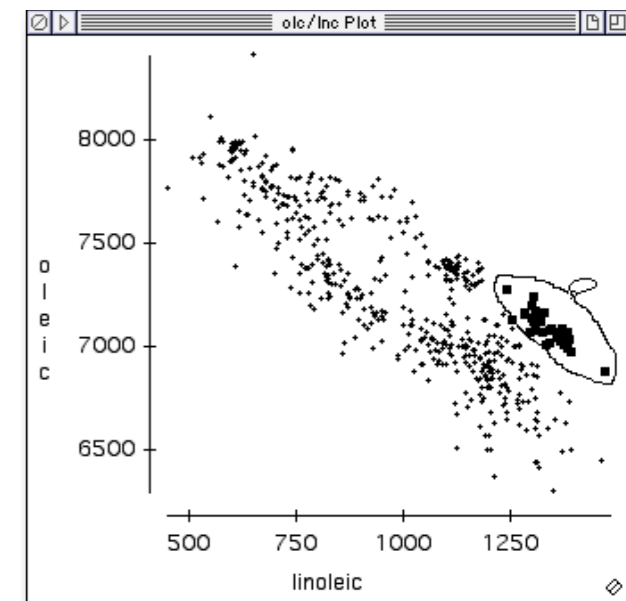
Whenever we get to work on raw (dirty) data, it is essential to find, understand and clean up artifacts and errors

Prerequisites for Graphical Data Analysis

- The general principle which we use in statistics (and thus in data analysis in general) is to compare groups, i.e., look at **conditional distributions**
- In traditional statistics, we assume underlying distributions and given these distributional assumptions, we can then **test hypotheses** and estimate parameters
- Graphical methods often show that these distributional assumptions may be far off from being fulfilled
- Thus for graphical data analysis we (at least) need
 - Tools which allow **selecting** subsets which might be interesting
 - Methods which can **highlight** a subset in specific plots
 - Means to **query** exact information from graphical / statistical elements
 - Mechanism to modify **plot parameters** on the fly

Foundations of Graphical Data Analysis: Selections

- Selections as such are not really interesting – but they are the necessary step to specify subsets of interest
- In an exploratory set-up we often want to look at the properties of specific subgroups, like
“Find all customers, who paid less than 15% tip, at night, except on weekdays!”
- The flexibility with which we can select data directly determines the how successful we may solve the exploratory analysis.
- Obviously we need different selection tools and selection modes

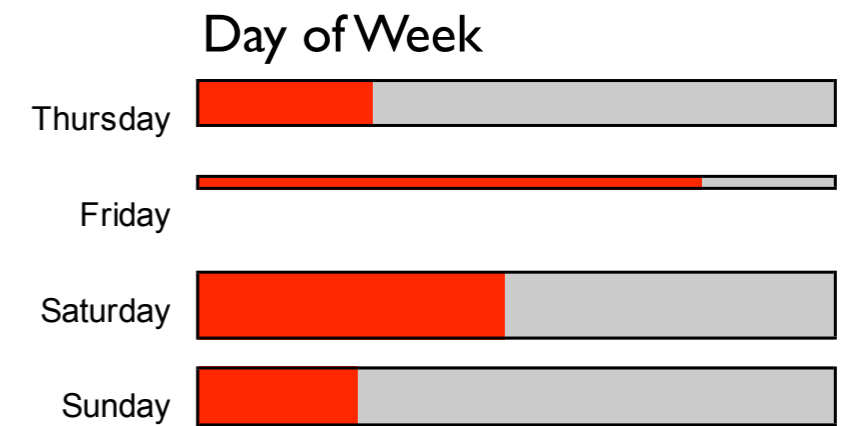
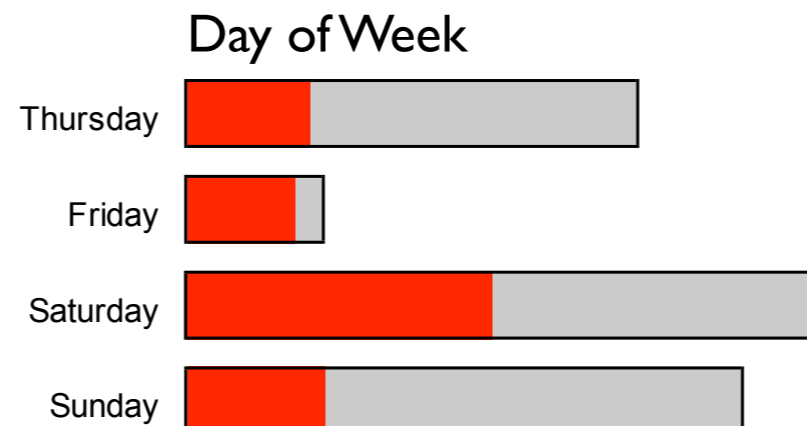
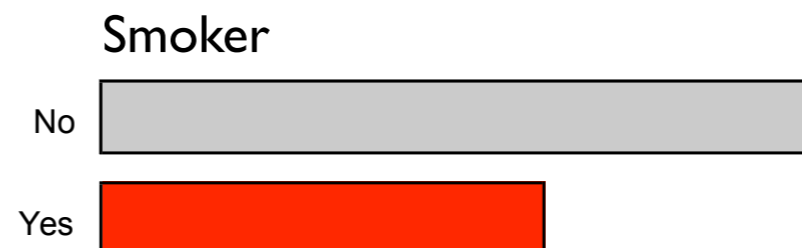


Foundations of Graphical Data Analysis: Selections

- Tools to select data:
 - **Pointer**
... is used to select single points.
 - **Drag-Box**
... selects rectangular regions in a graphics window.
 - **Brush**
... allows a dynamic change (movement) of the selected region – usually a rectangle.
 - **Slicer**
... selects intervals along an axis dynamically.
 - **Lasso**
... allows the most flexible definition of the selection area.
Startpoint and endpoint are always connected.
- Modes to select data:
 - **Simple / Standard / Default**
... only points in the selected region are selected.
 - **Intersection / AND / \cap**
... only points that already were selected and are within the new selection stay selected.
 - **Union / OR / \cup**
... the newly selected points are added to the current selection.
 - **Toggle / XOR / \oplus**
... selected points are deselected, unselected are selected.
 - **Negation / NOT / \neg**
... points in the selection region are taken out of the current selection set.

Foundations of Graphical Data Analysis: Highlighting

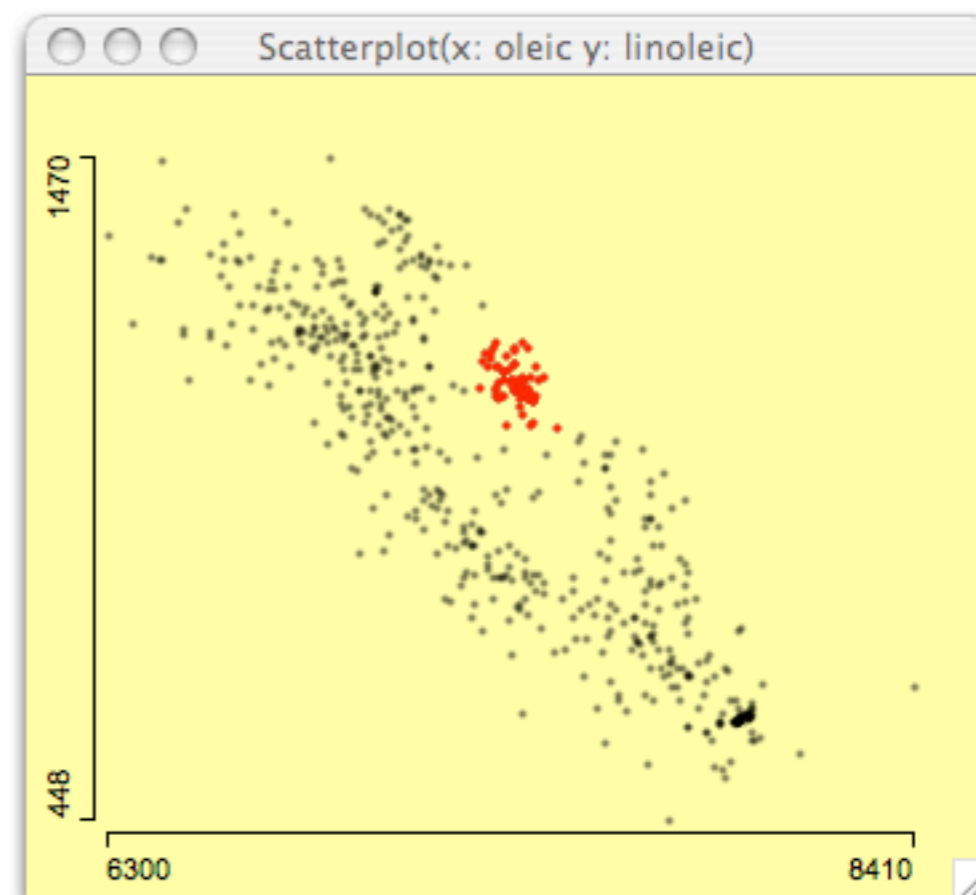
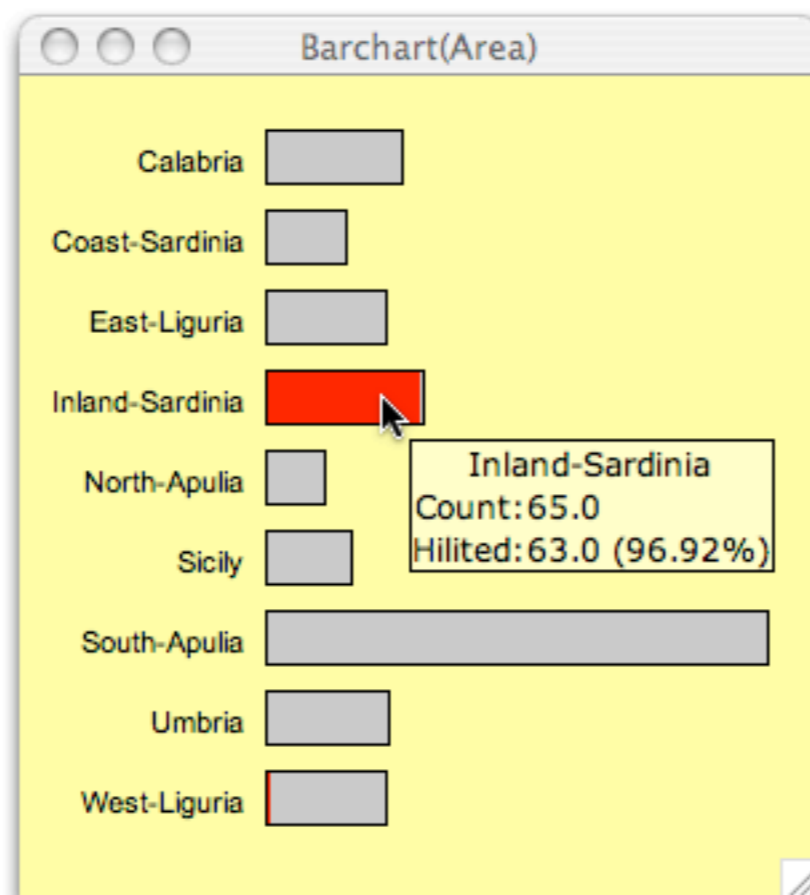
- Once a selection is defined, it needs to be propagated to all other plots
- All plots need to know how to highlight a subgroup
- Highlighting may be
 - **transient** (only changes when a new selection is performed)
 - **persistent** (a new state explicitly must be assigned to the involved cases)
- A clear rule how highlighting is performed is desirable, but exceptions have proven to be very powerful



Example:
Barchart/Spineplot

Foundations of Graphical Data Analysis: Queries

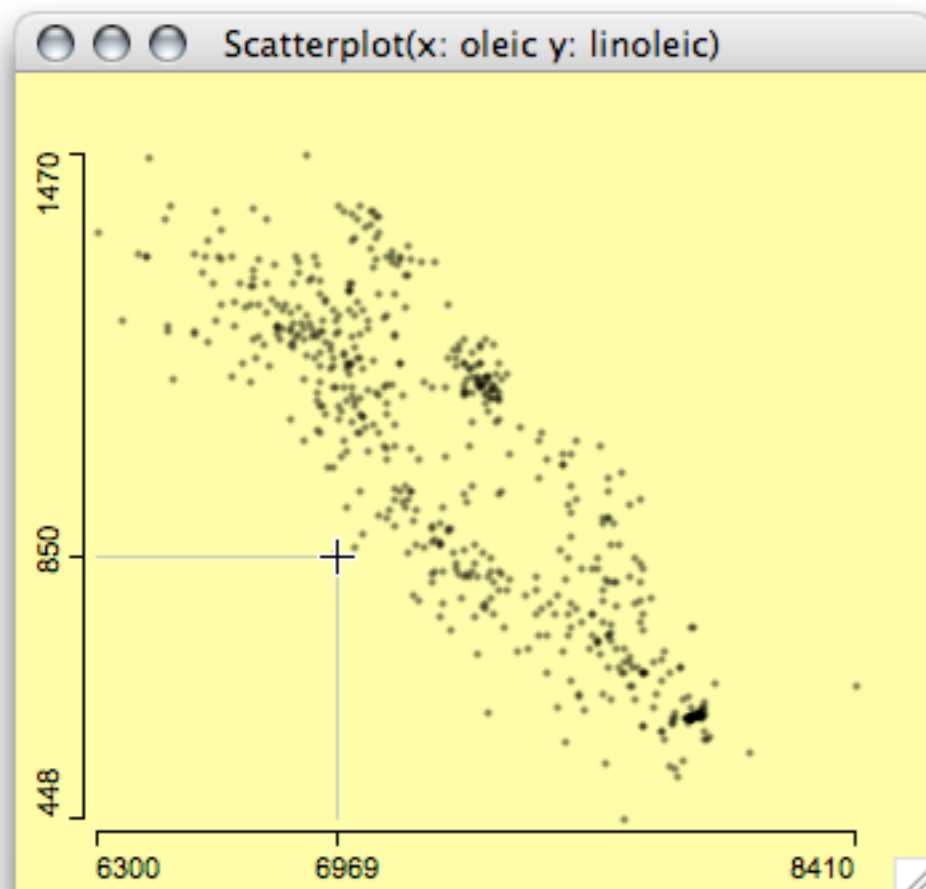
- Graphics are good at communicating qualitative information but fail to give exact quantities \Rightarrow need queries to get exact values
- Gridlines can help (only) for the variables within the plot
- Interactive graphics often display very little scale information (cf. Tufte's "data-ink-ratio")
- Examples:



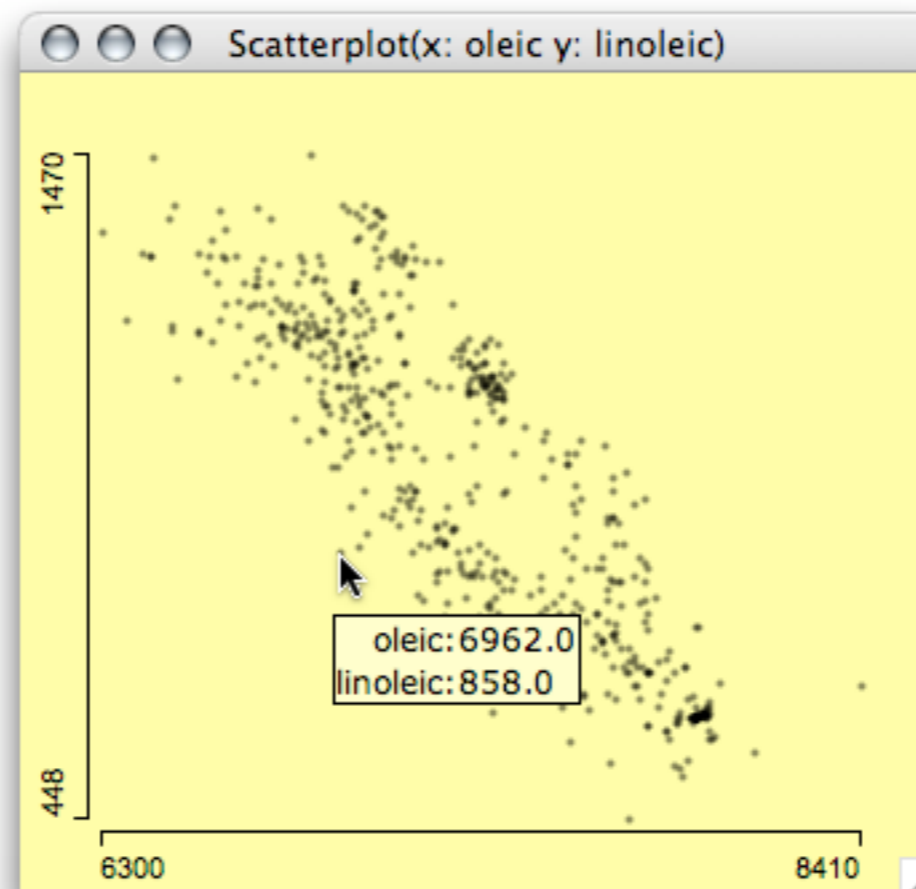
Foundations of Graphical Data Analysis: Queries

- The level of detail of a query should have optional granularities:
 - **orientation**, “what are the coordinates at the mouse pointer” (interactive grid)
 - **standard**, “what are the coordinates of a particular value”
 - **extended**, “what are the values for an object beyond the variables in the plot”
- Example: scatterplot

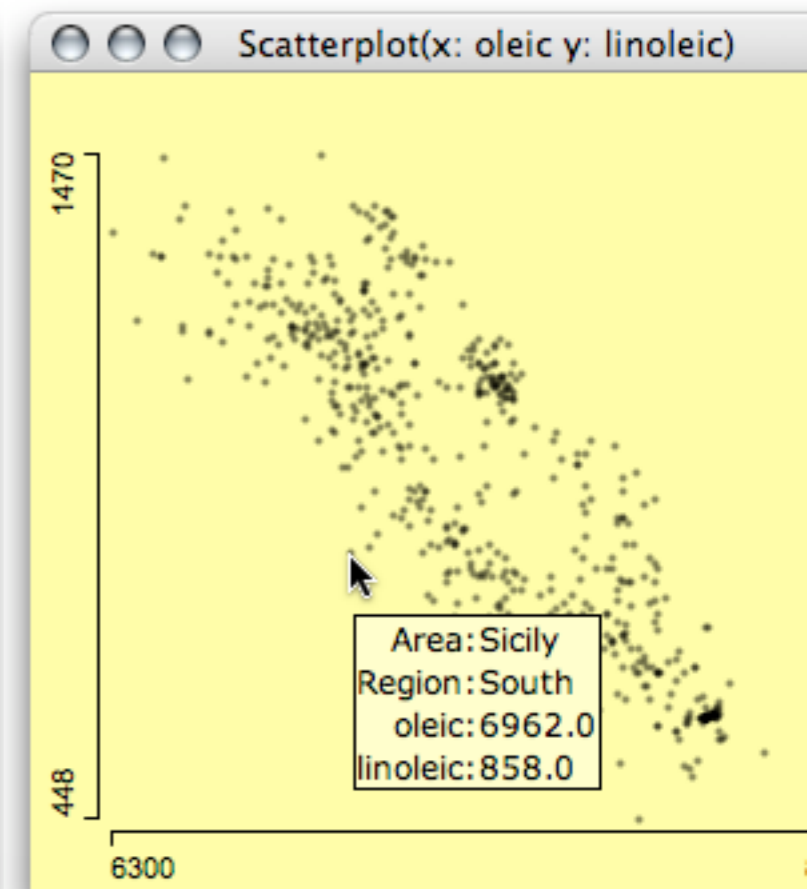
orientation



standard



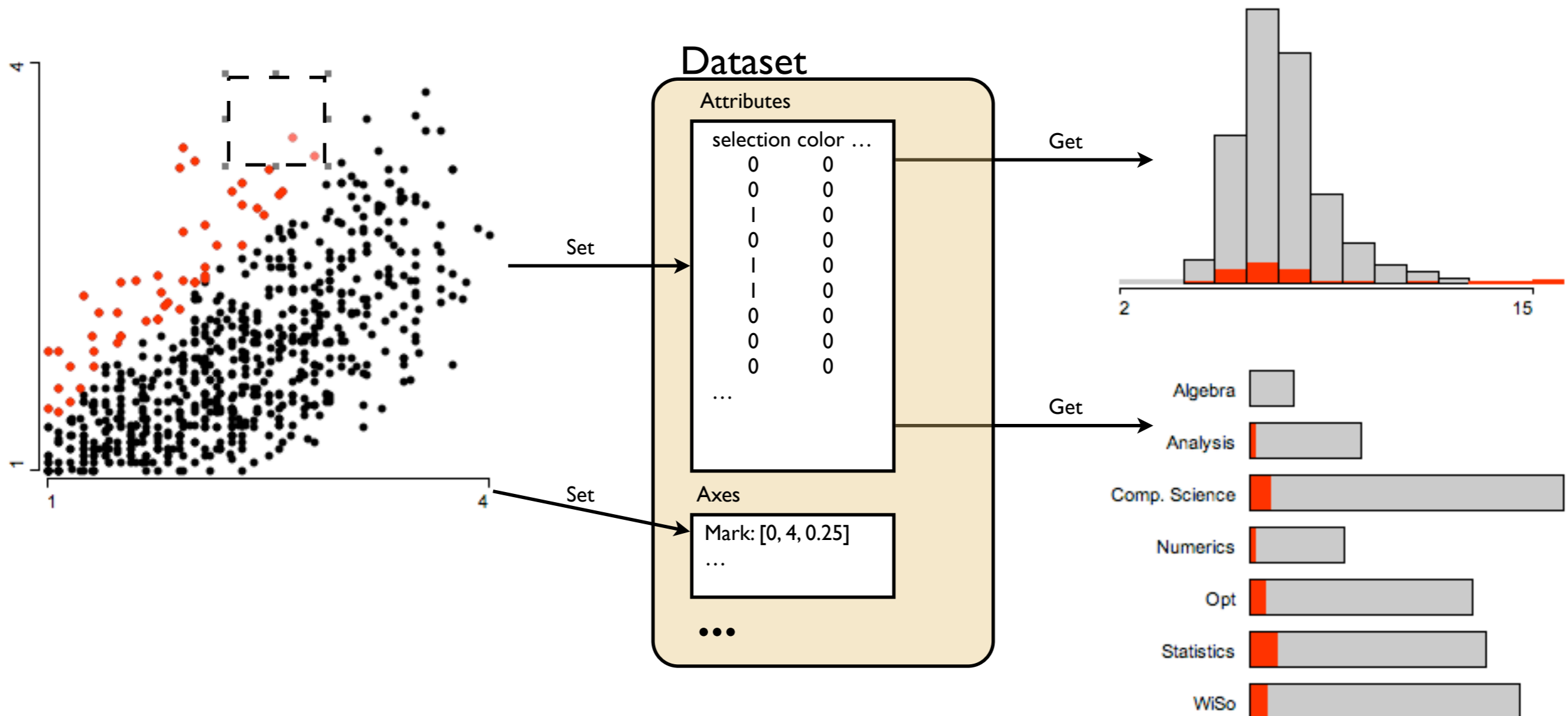
extended



Escaping Flatland I

- Selection ➤ Linking ➤ Highlighting

Example:



Escaping Flatland II

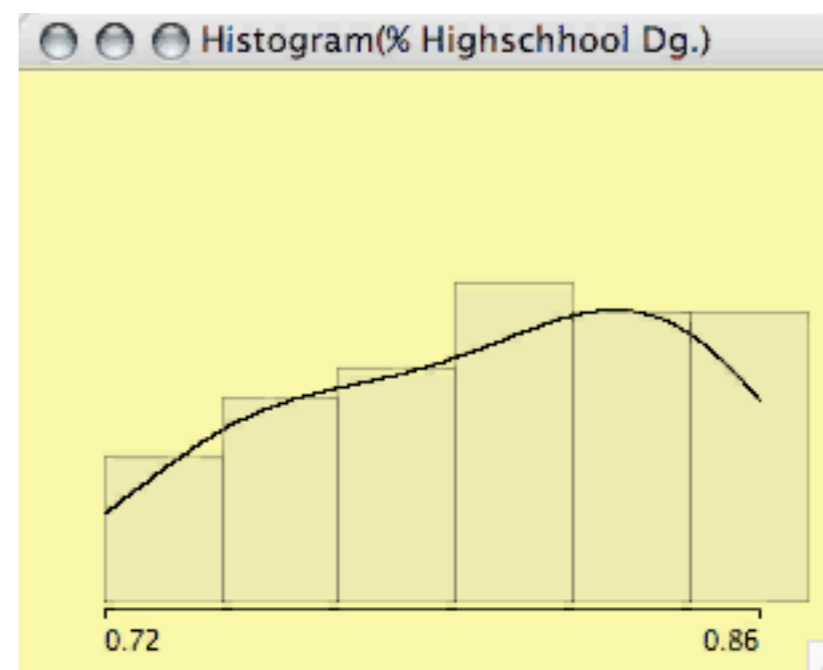
- The two essential multivariate plots are
 - parallel coordinate plots (for continuous data)
 - mosaic plots (for categorical data)
- Both of these plots are not very powerful for exploratory work as long as they are not implemented interactively.
- Essential interactive features are
 - Parallel coordinates
 - rearrangement of axes (manual, automatic permutations)
 - scaling of axes (common, individual, inversion)
 - alignment of axes (mean, median, constants)
 - sorting (min, max, mean, median, range, std.dev.)
 - Mosaic plots
 - include and exclude variables
 - permute variable order
 - (censored) zooming
- Linking with these plots increases dimensionality even more

Foundations of Graphical Data Analysis: Parameters

- Looking at the graphics functions in systems like R, SPSS or SAS, we find a large number of options to set
- Many of these options only apply to the artistic quality of the plots, i.e., fonts, colors, patterns, etc.
- For an exploratory analysis, we need to modify those plot parameters, which relate to statistical aspects of the graph
- Example: **Barchart**

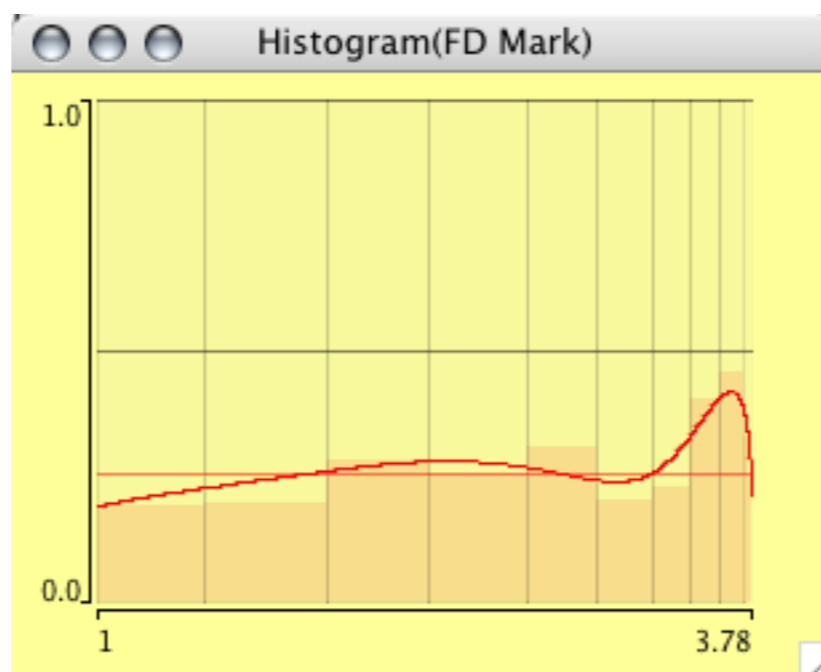
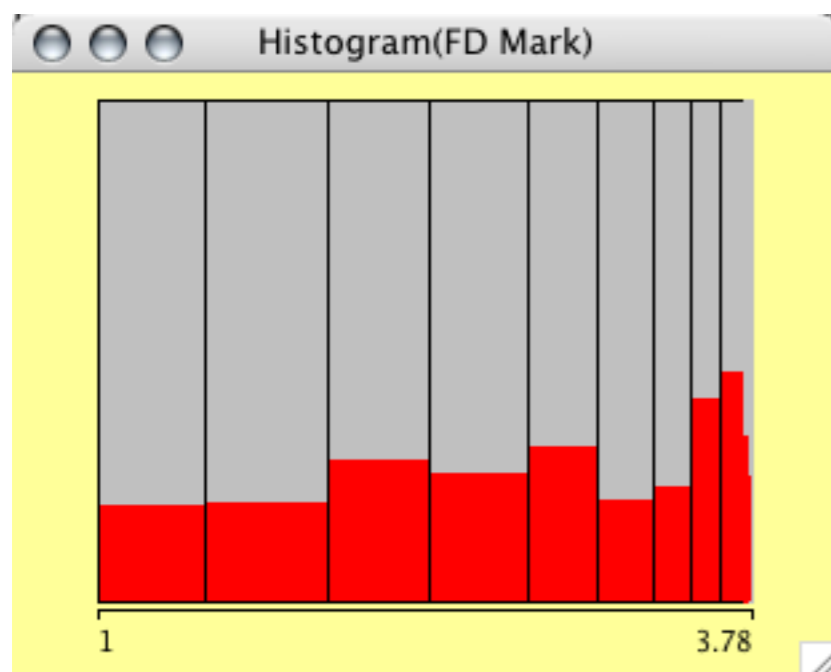
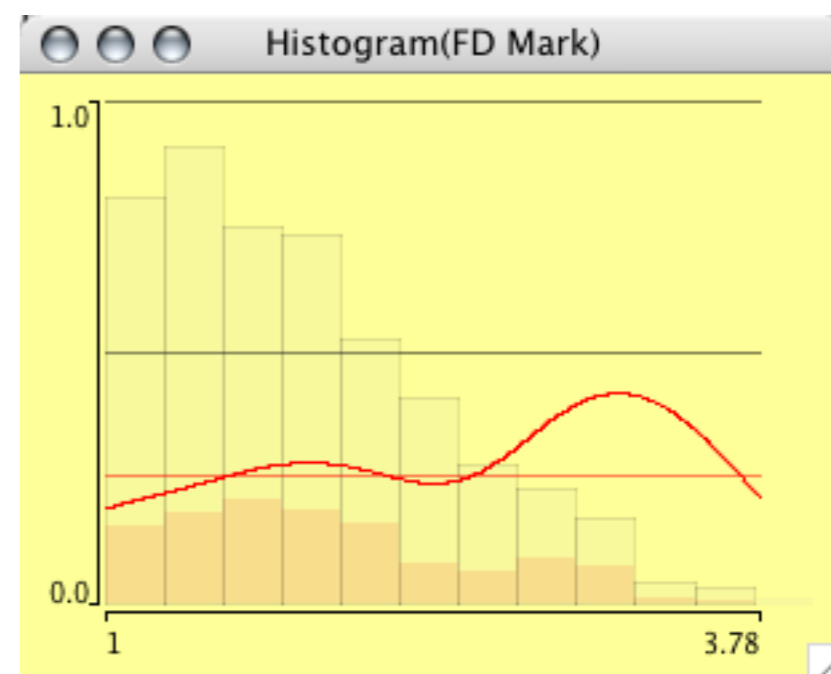
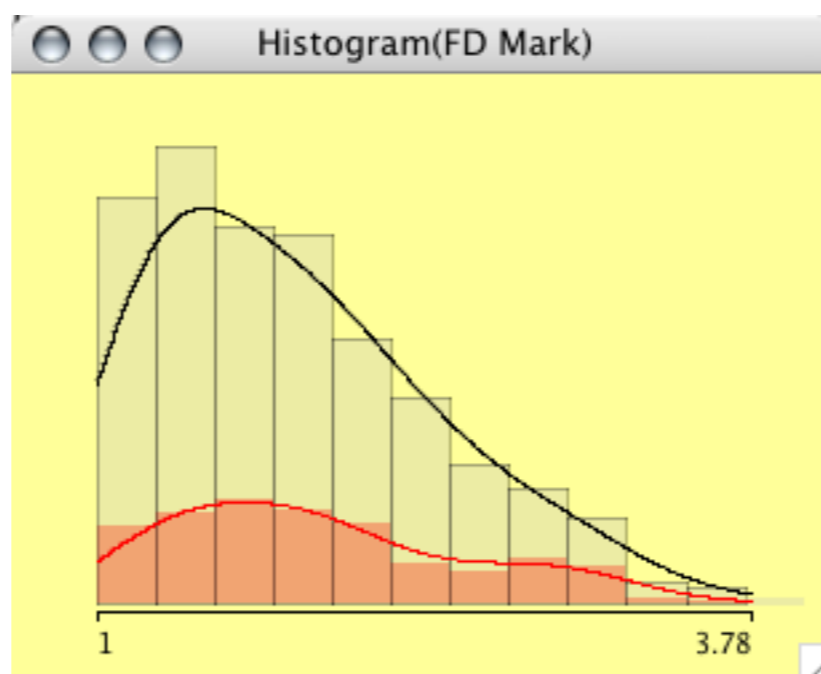
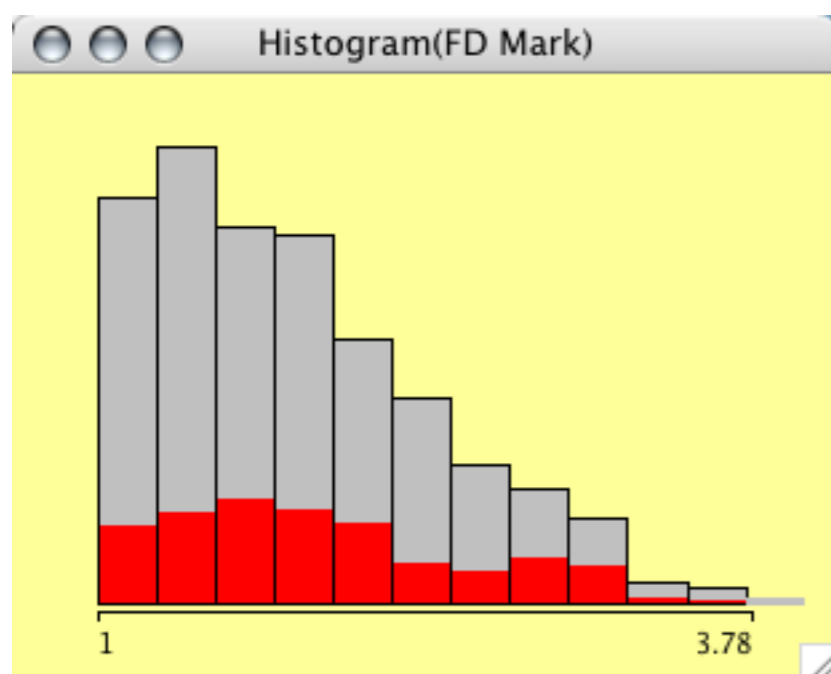
Two parameters:

- anchor point
- bin width / no. of bins



Statistification of Graphics

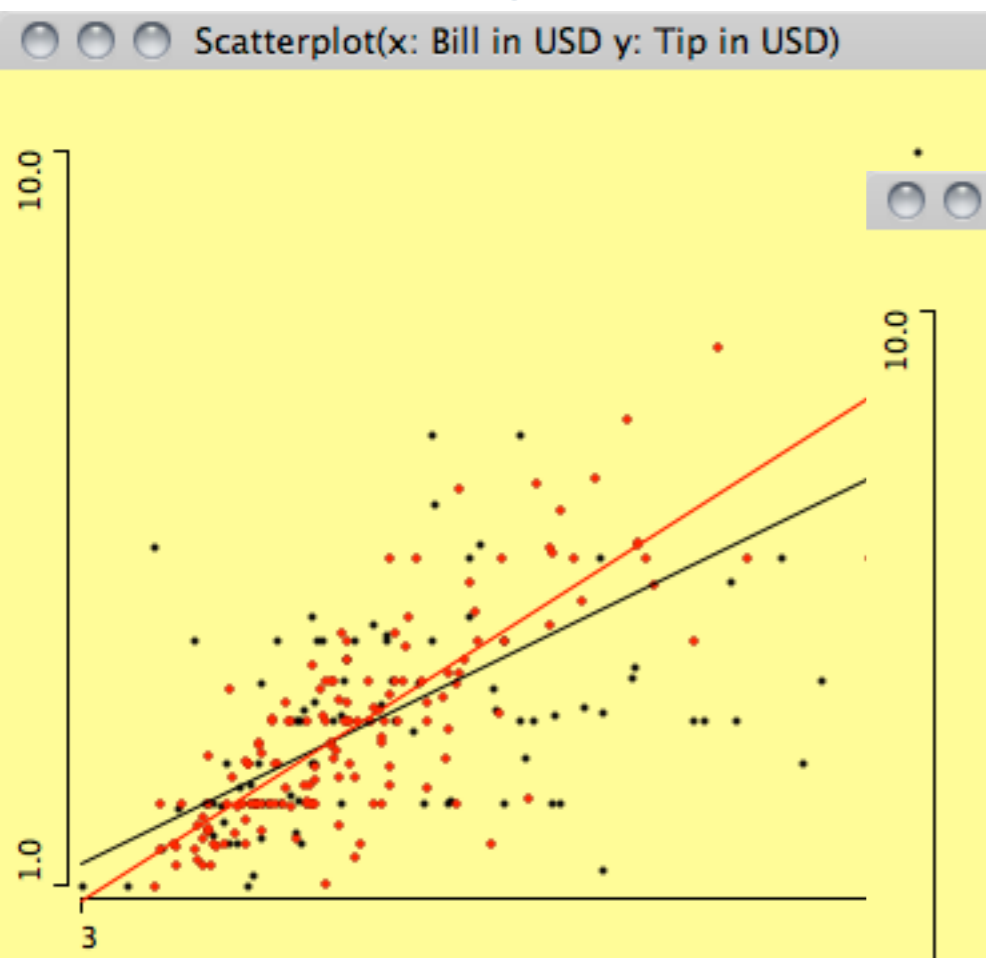
- Example: Density Estimation



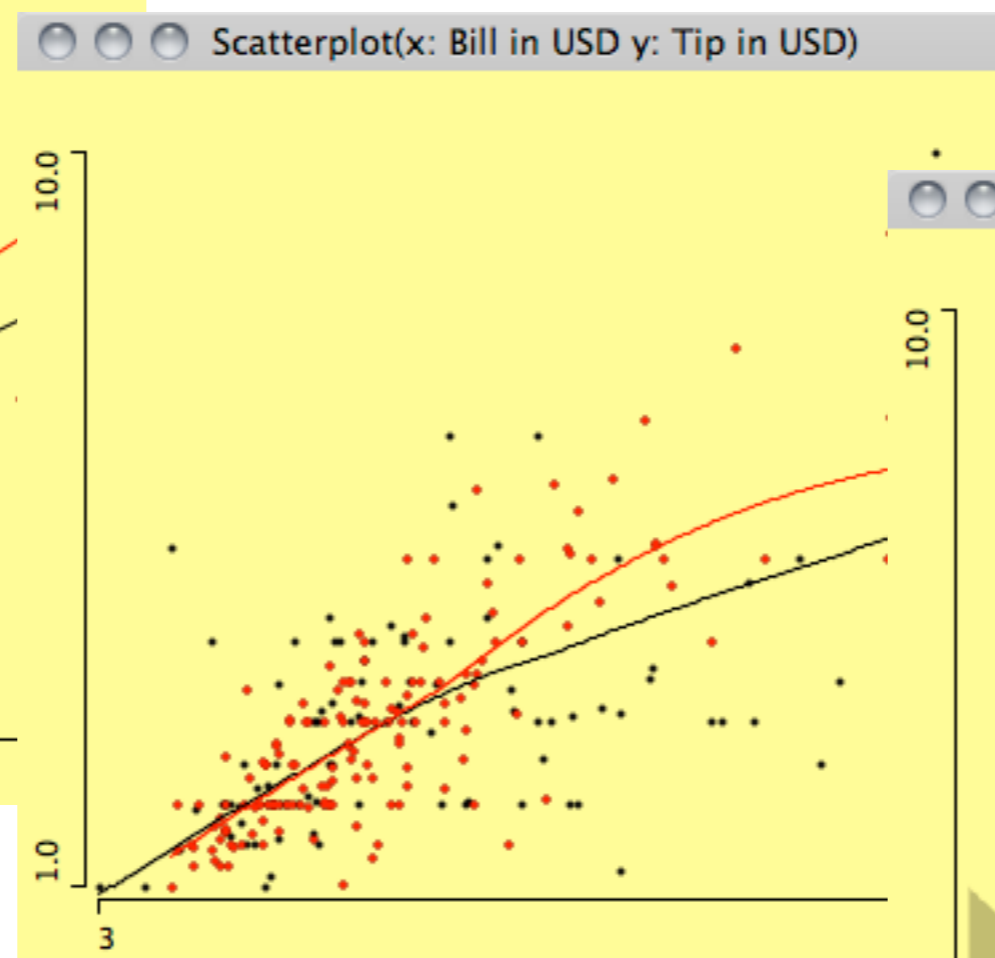
Statistification of Graphics

- Example: Scatterplot Smoothers

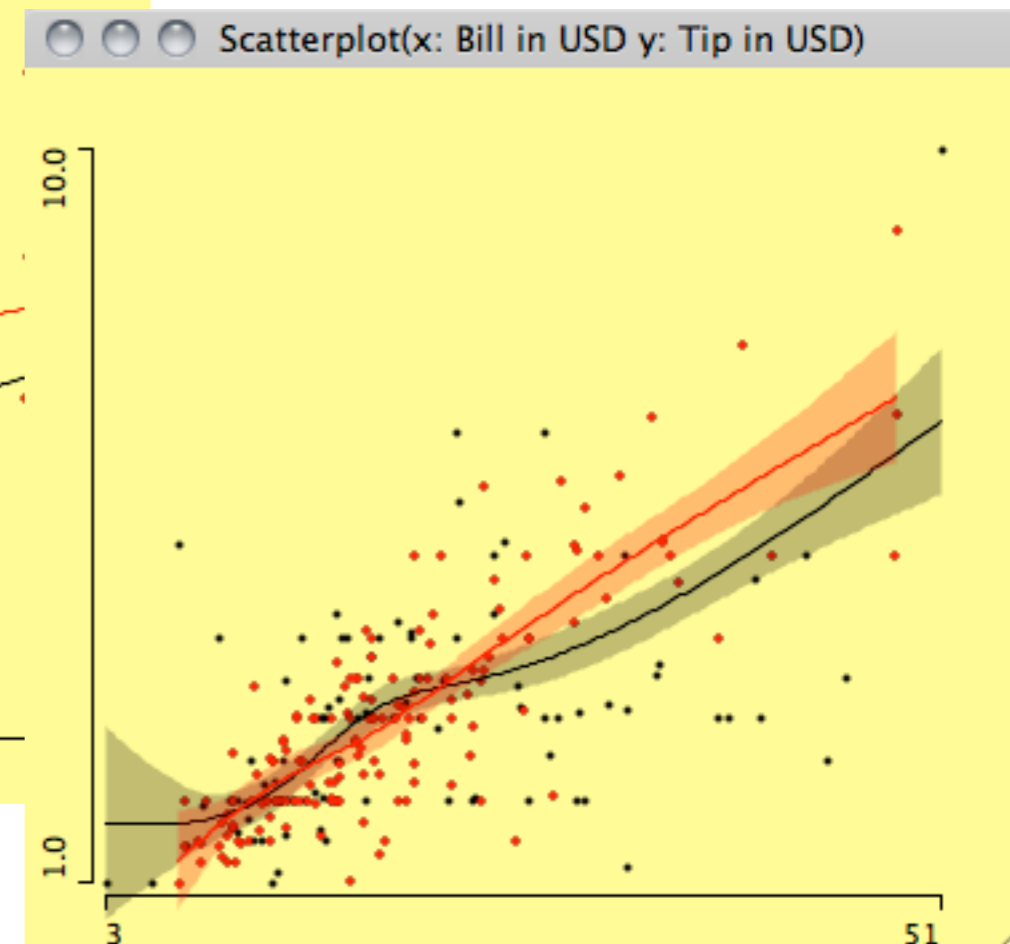
linear regression



loess

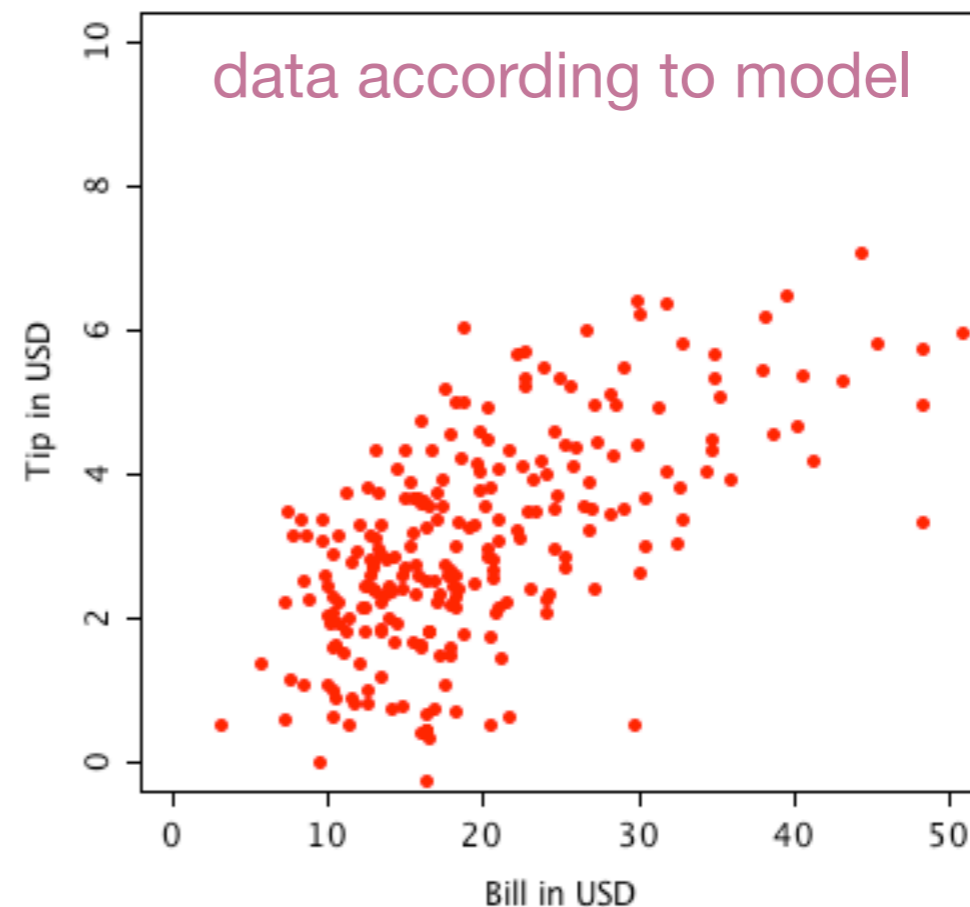
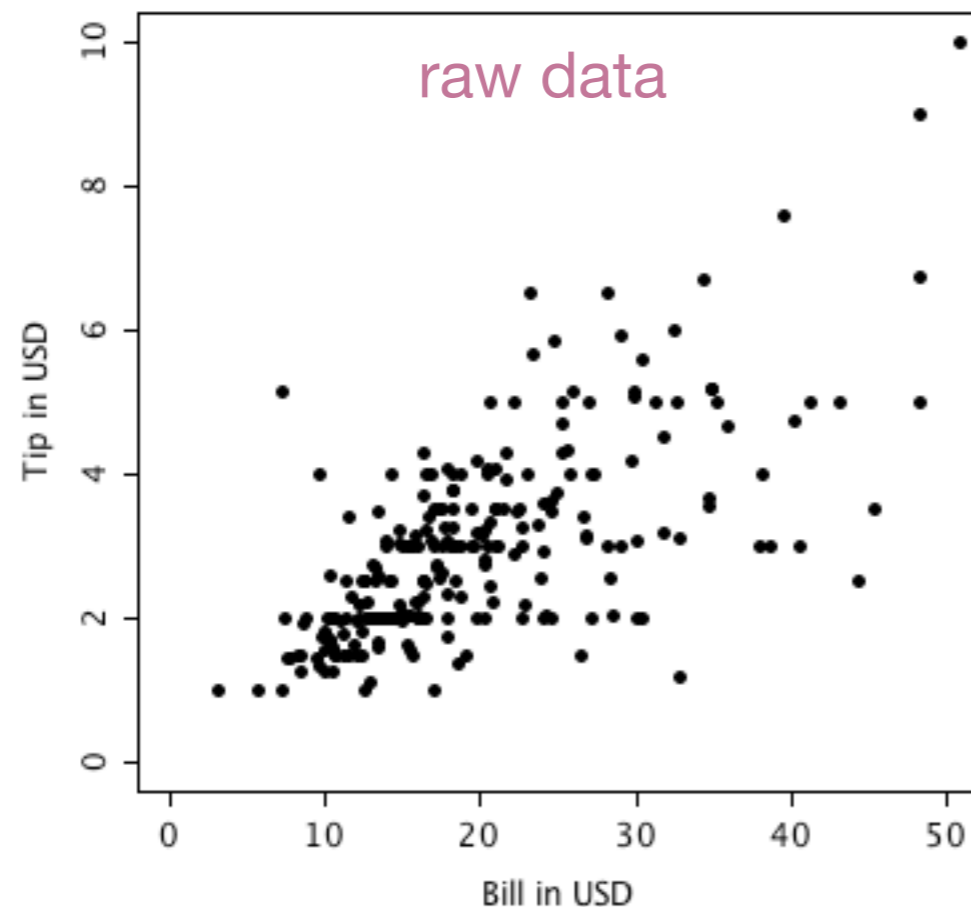


splines



Outlook: Graphical Inference

- Basic Idea:
“Look the sampled data of the model like my raw data?”
- Once we “know” how our raw data “looks like”, we can compare it to the data we sample from a chosen model
- Example: simple linear model for **Tip** ~ **Billsize**



About Mondrian

- Mondrian is a general purpose **graphical data analysis tool**
- It is based on the experiences and tries to expand the concepts and ideas of
 - **DataDesk** (Paul F. Velleman, 1985)
 - **MANET** (Unwin et al., 1994)
- The basic building blocks of Mondrian are
 - **uni- and multivariate plots** for variables measured on various scales (including geographical maps)
 - **selection**, and
 - linked **highlighting**
 - fast **parameter** changes
 - **link to R** to add statistical procedures of various kinds
- Mondrian can be used **free of charge**, is **open source** and runs equally well on **Windows**, **MacOS** and **Linux** computers

Demo

Sample of 572 Italian Olive Oils

- The definition of olive oil quality (extra virgin, virgin) is expressed as threshold on acidity measured as the proportion of “oleic” fatty acid
- Composition of fatty acids in olive oils can be measured with the aim of identifying the origin of the olive oil
- Here measurements were taken in 1983 from Italian olive oils originating from 9 different regions aggregated to 3 areas



Summary

- Graphical methods are quite effective when it comes to explore the structure of a dataset
- Interactions with the graphs are crucial to aid a fast and successful exploration of the data
- High-dimensional plots are already quite powerful and linking increases the dimensionality even further
- Without interaction, those high-dimensional plots are far less useful
- After an initial graphical cleaning and exploration of a dataset, we usually end up with much simpler models and avoid modeling artifacts or errors

You may certainly trust your models, but only once you understood your data

Why Mondrian?

- Mondrian (Theus, 2010)

Piet Mondriaan

