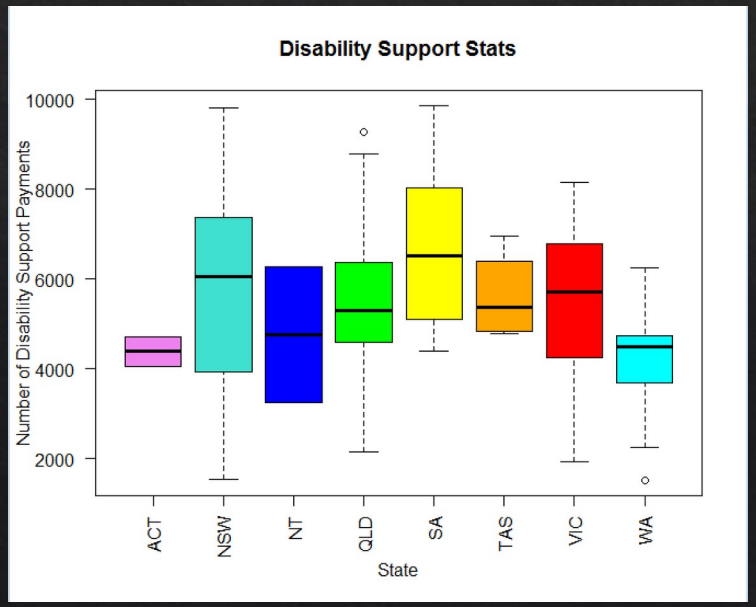
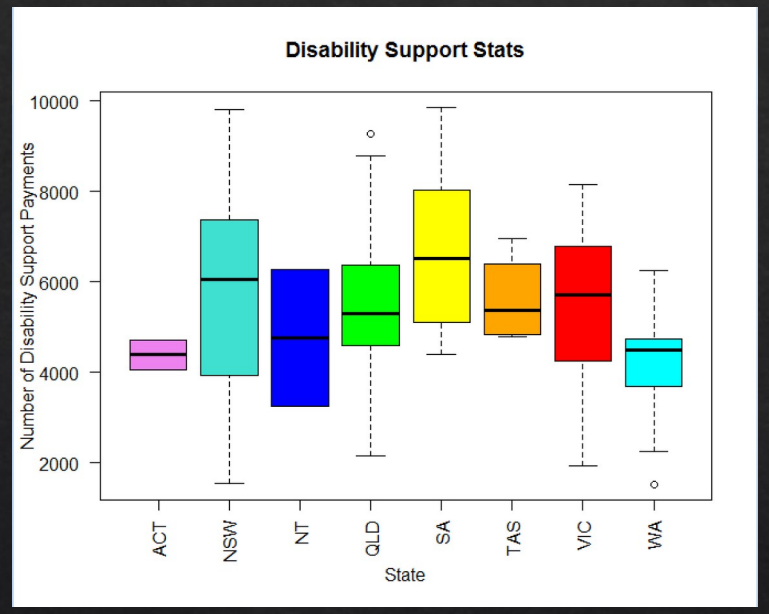
Ready for Box Plots!

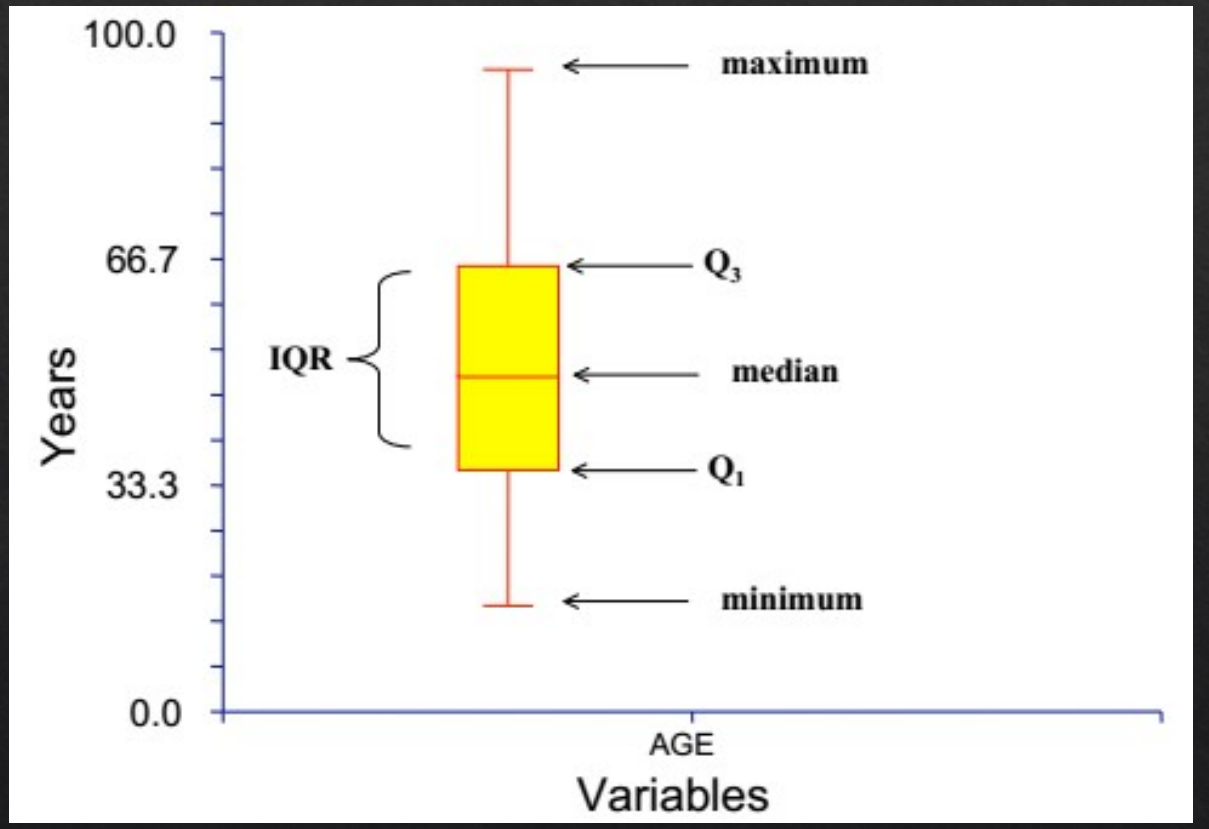
John Tukey - 1977

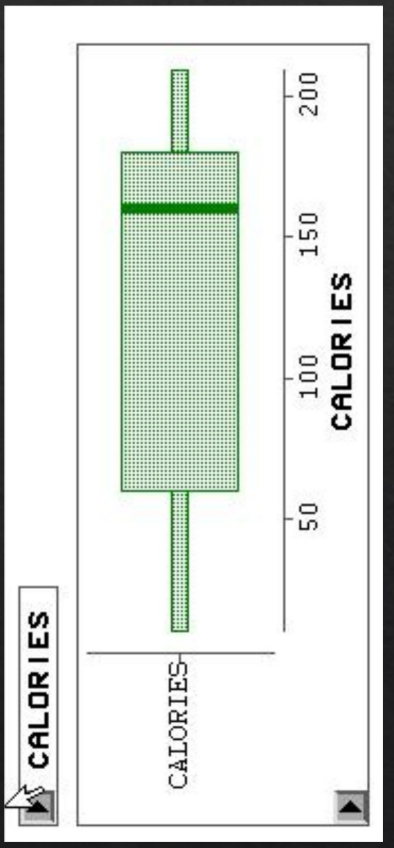


Box Plot for Calories

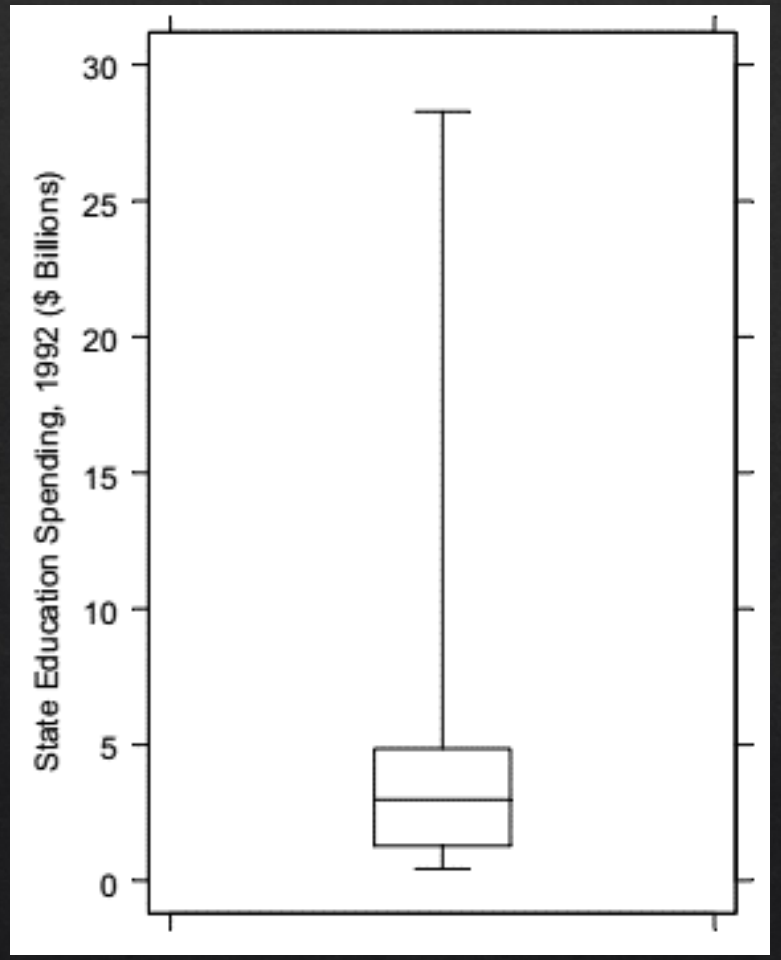
A visualization of most of the basic statistics.

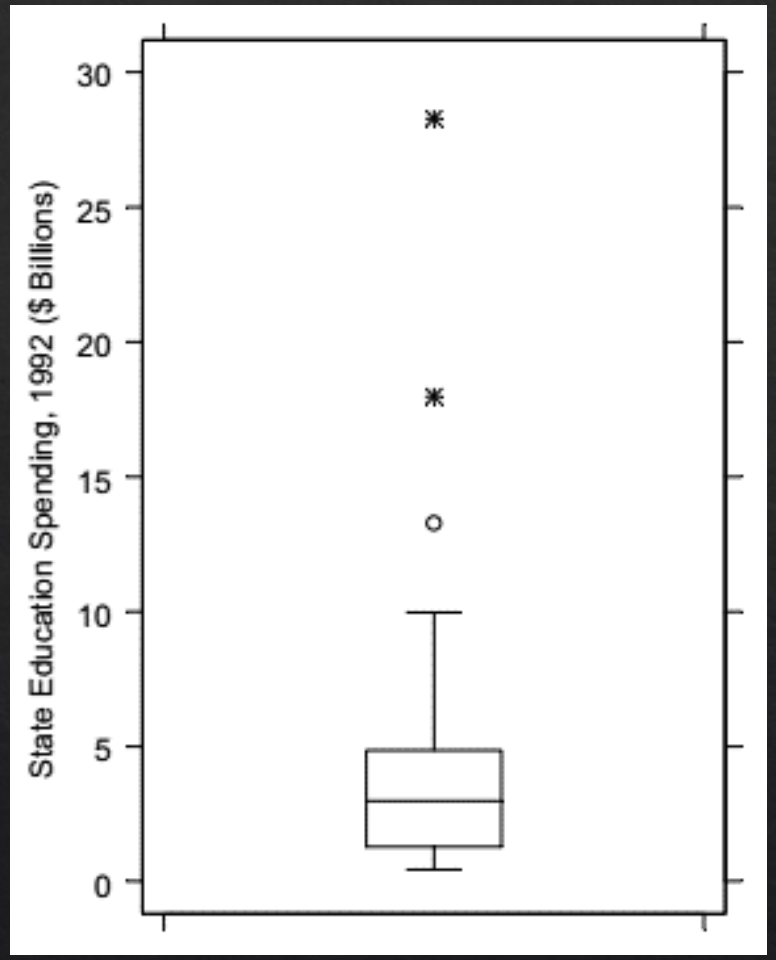


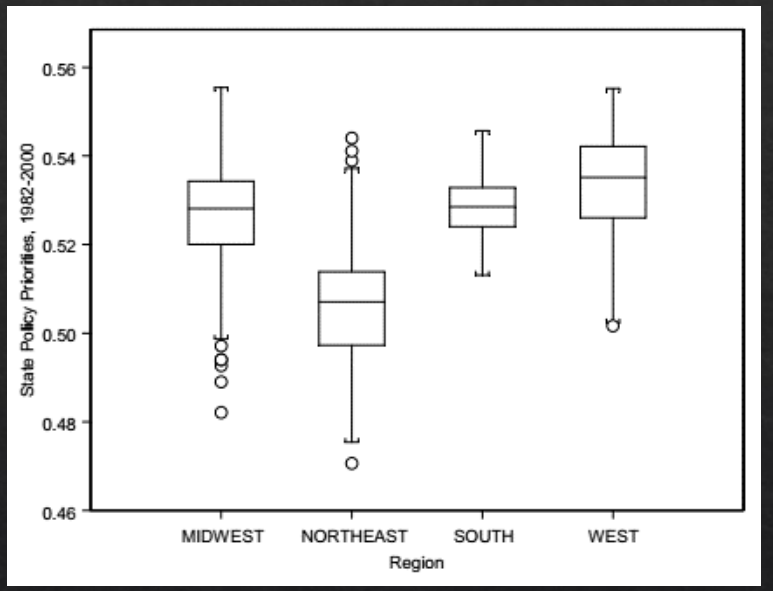




Box Plot and Outliers



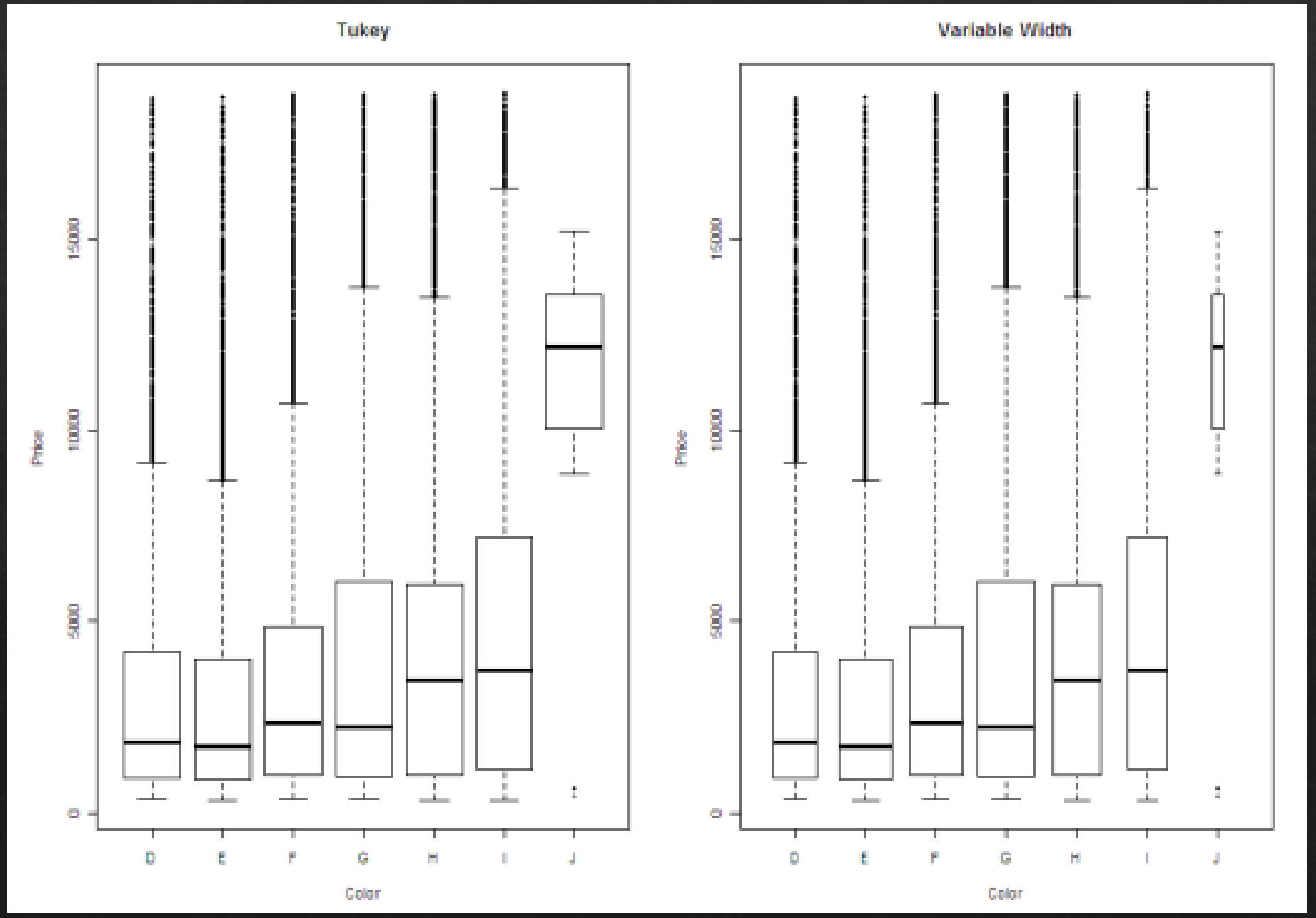




Examples

From the first plot it appears that the overall median price might be between 5,000 and 10,000. However, the actual overall median price is 2,374.

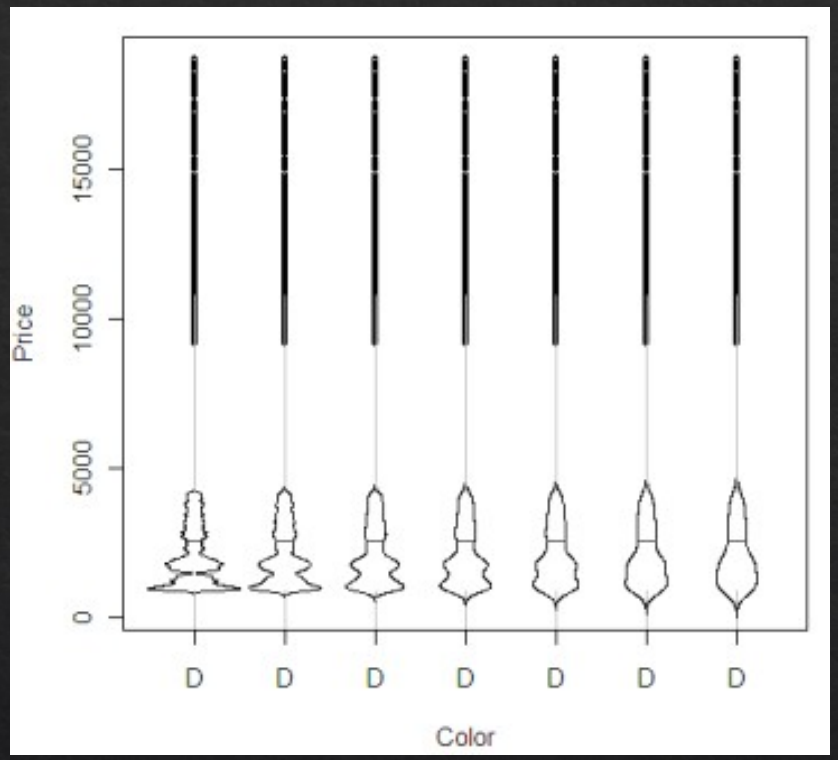
Looking at the variable width boxplot in the middle of figure four it is obvious that J has far fewer observations than the other groups.

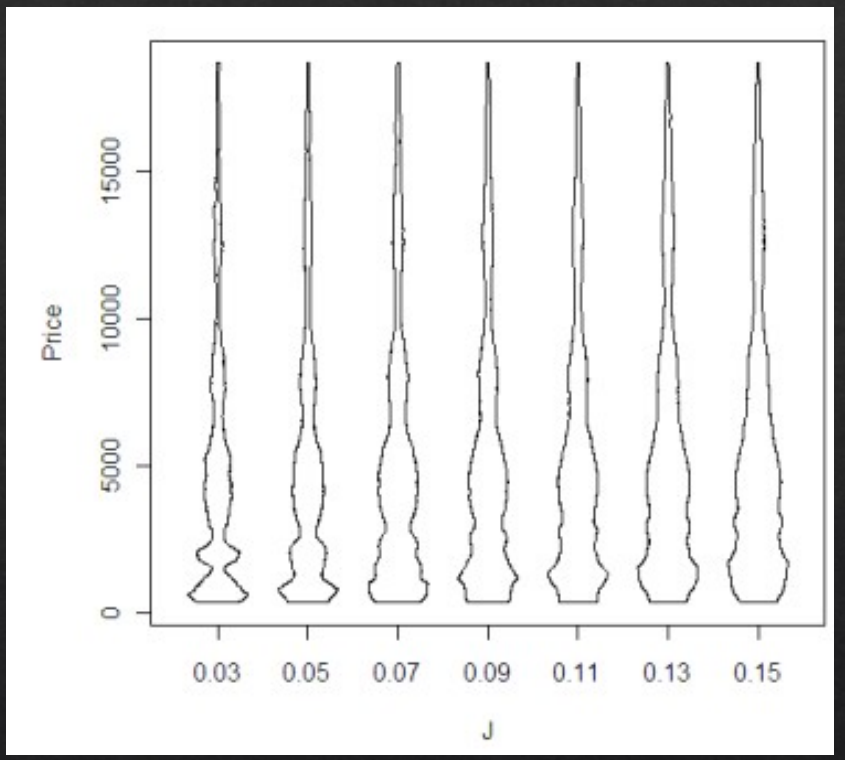


Examples

VasePlot: The width of the box at each point is
proportional to the estimated density

Violin Plot: use *all* the data to plot a density curve. The amount of smoothing to use is at the discretion of the analyst.

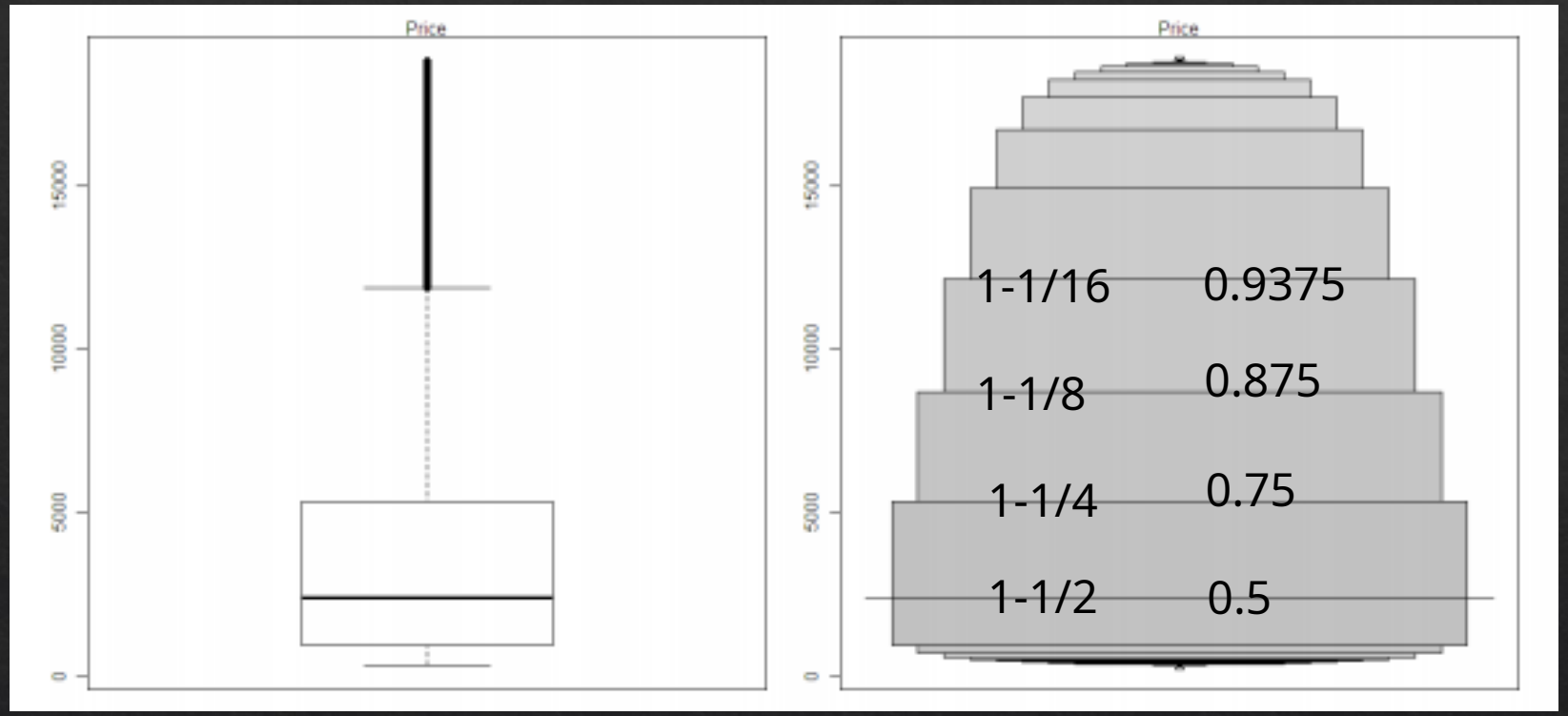




Hofmann, Kafadar, and Wickham, 2006

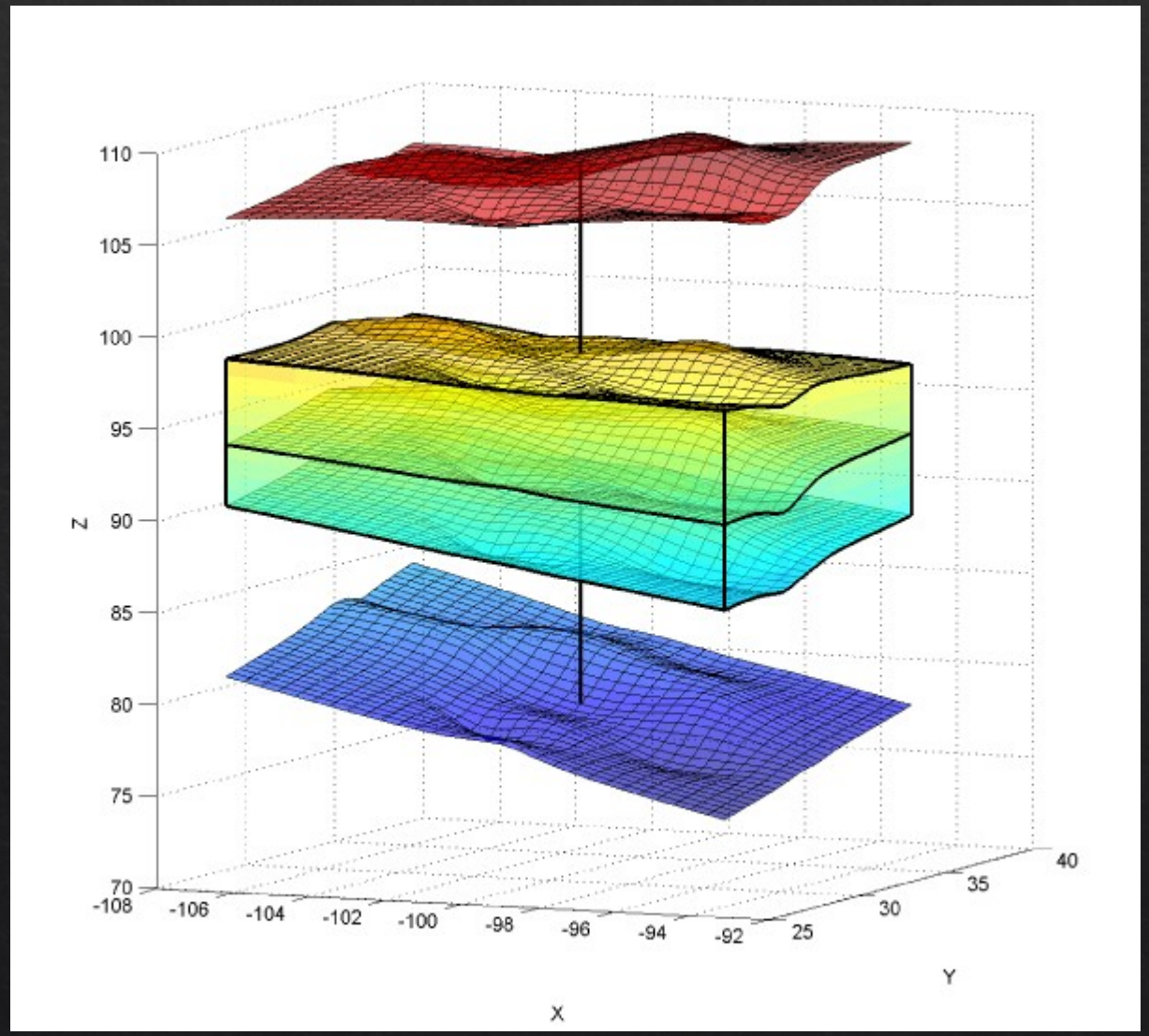
Letter-value Boxplot:

To accommodate large datasets. For moderate sized datasets (n < 1,000), estimates about the behavior in the tails are not reliable. Large datasets (n = 10,000 100,000) yield much more reliable estimates about the behavior beyond the quartiles. The simple boxplot
is not a good method for large datasets, since there is often too much over-plotting in the outlier region.



Letter-value boxplots of color vs. price display the data past the quartiles in a more meaningful manner than
does the original boxplot.

Surface Box Plot



The surface boxplot with the box in the middle representing the 50% central region in R3, the middle surface inside the box denoting the median surface, and the upper and lower surfaces indicating the maximum non-outlying envelope.

Normalization and Standardization

Normalization of data is to rescale the data to have values between 0 and 1. This can be done by subtracting the minimum data value from each data point and then dividing the subtracted value by the range of the data.

$$X_{new} = (X_{old} - X_{min}) / (X_{max} - X_{min})$$

Standardization of data is to rescale the data to have mean 0 and standard deviation 1. This can be done by subtracting the mean M from each data point and then dividing the subtracted value by the standard deviation S.

$$X_{new} = (X_{old} - M) / S$$

Normalization and Standardization

Normalization of data is to rescale the data to have values between 0 and 1. This can be done by subtracting the minimum data value from each data point and then dividing the subtracted value by the range of the data.

$$X_{new} = (X_{old} - X_{min}) / (X_{max} - X_{min})$$

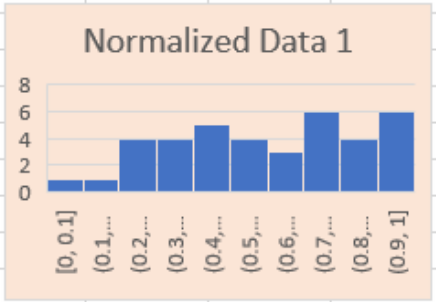
Standardization of data is to rescale the data to have mean 0 and standard deviation 1. This can be done by subtracting the mean M from each data point and then dividing the subtracted value by the standard deviation S.

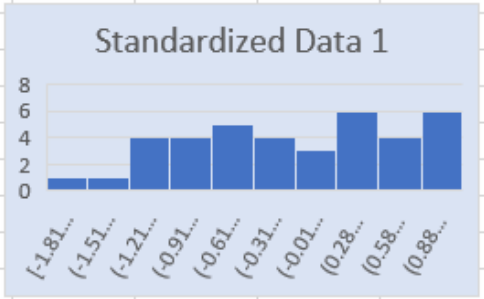
$$X_{new} = (X_{old} - M) / S$$

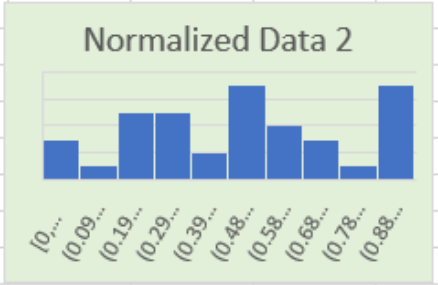
Class Activity

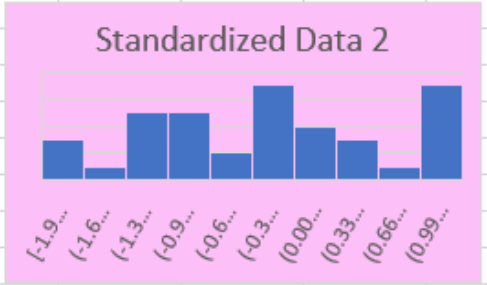
- Let H1, H2, H3 be the histograms of a set X of positive numbers.
  - What are the visual differences between H1 and H2?
  - What are the visual differences between H2 and H3?
- What is the benefit of normalization and Standardization?

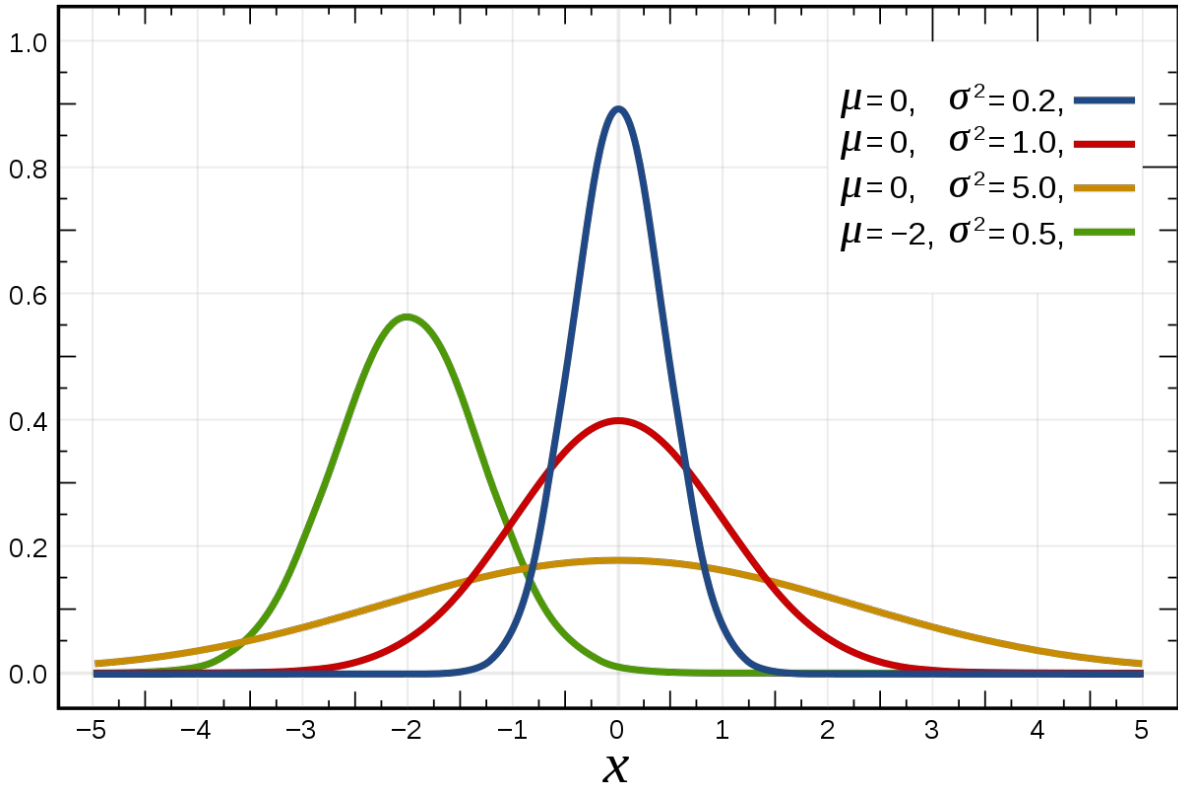
1	0.111111	-1.48495	12	0.102041	-1.625086
2	0.222222	-1.150502	24	0.22449	-1.213845
4	0.444444	-0.481605	68	0.673469	0.294037
3	0.333333	-0.816054	34	0.326531	-0.871145
7	0.777778	0.521739	89	0.887755	1.013708
9	1	1.190635	8	0.061224	-1.762166
6	0.666667	0.187291	56	0.55102	-0.117204
5	0.555556	-0.147157	34	0.326531	-0.871145
4	0.444444	-0.481605	98	0.979592	1.322138
2	0.222222	-1.150502	54	0.530612	-0.185744
3	0.333333	-0.816054	32	0.306122	-0.939685
4	0.444444	-0.481605	57	0.561224	-0.082934
5	0.555556	-0.147157	98	0.979592	1.322138
7	0.777778	0.521739	45	0.438776	-0.494174
8	0.888889	0.856187	23	0.214286	-1.248115
9	1	1.190635	56	0.55102	-0.117204
7	0.777778	0.521739	97	0.969388	1.287868
2	0.222222	-1.150502	76	0.755102	0.568197
8	0.888889	0.856187	53	0.520408	-0.220014
9	1	1.190635	32	0.306122	-0.939685
9	1	1.190635	68	0.673469	0.294037
7	0.777778	0.521739	45	0.438776	-0.494174
5	0.555556	-0.147157	23	0.214286	-1.248115
3	0.333333	-0.816054	78	0.77551	0.636737
7	0.777778	0.521739	98	0.979592	1.322138
4	0.444444	-0.481605	53	0.520408	-0.220014
2	0.222222	-1.150502	23	0.214286	-1.248115
8	0.888889	0.856187	67	0.663265	0.259767
6	0.666667	0.187291	89	0.887755	1.013708
6	0.666667	0.187291	4	0.020408	-1.899246
5	0.555556	-0.147157	32	0.306122	-0.939685
3	0.333333	-0.816054	76	0.755102	0.568197
4	0.444444	-0.481605	98	0.979592	1.322138
7	0.777778	0.521739	64	0.632653	0.156957
8	0.888889	0.856187	2	0	-1.967786
0	0	-1.819398	53	0.520408	-0.220014
9	1	1.190635	87	0.867347	0.945168
9	1	1.190635	24	0.22449	-1.213845











Normal Distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

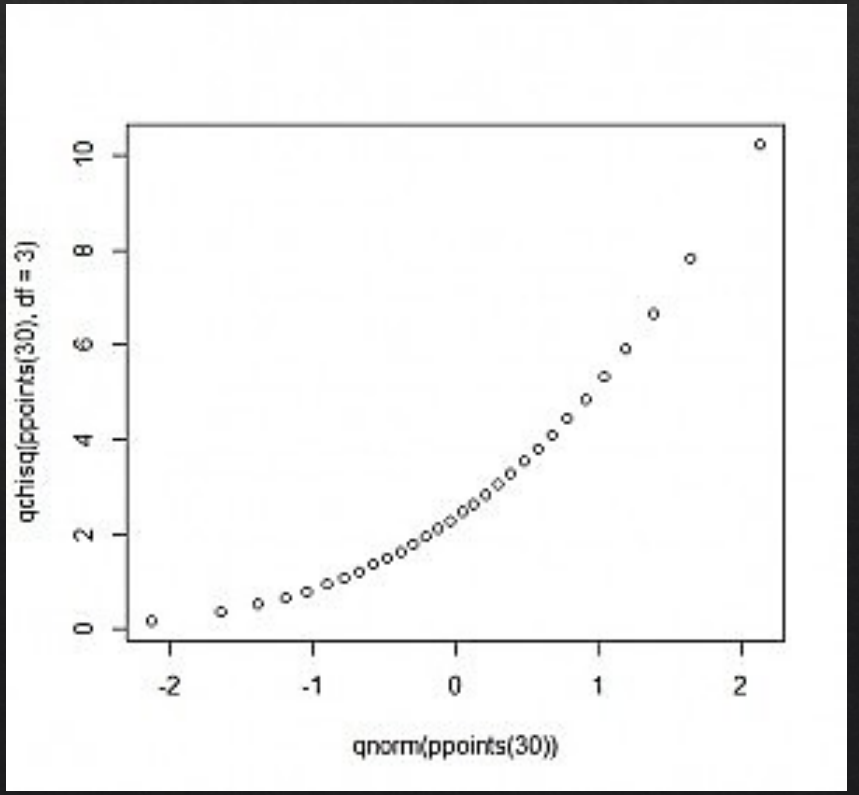
Q - Q Plots!

Quantile-Quantile Plot

A scatterplot created by plotting two sets of quantiles against one another.

Let's try to check whether the data comes from a normal distribution.

- Q-Q plots that look like this usually mean your sample data are skewed.

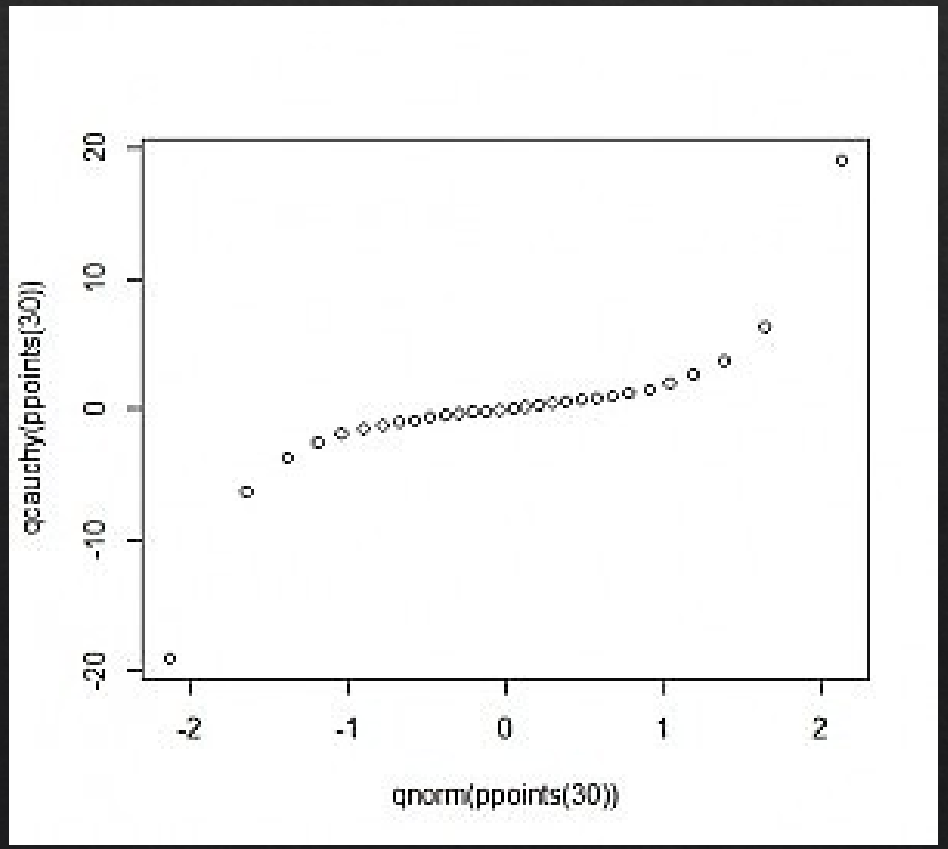


Quantile-Quantile Plot

A scatterplot created by plotting two sets of quantiles against one another.

Let's try to check whether the data comes from a normal distribution.

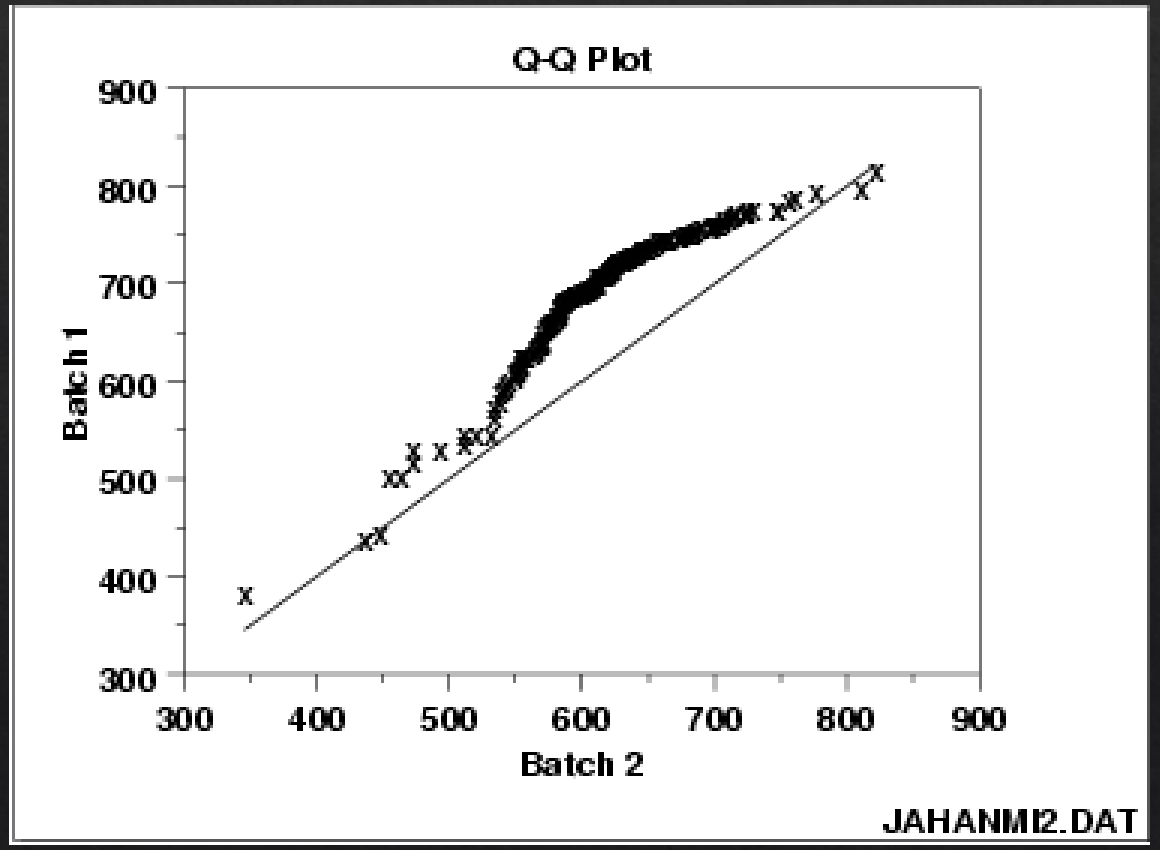
- Q-Q plots that exhibit this behavior usually mean your data have more extreme values than the data coming from a Normal distribution.



Quantile-Quantile Plot

A scatterplot created by plotting two sets of quantiles against one another.

- These 2 batches do not appear to have come from populations with a common distribution.
- The batch 1 values are higher than the corresponding batch 2 values.
- The differences are increasing from values 525 to 625. Then the values for the 2 batches get closer again.



Benefits of Quantile-Quantile Plot

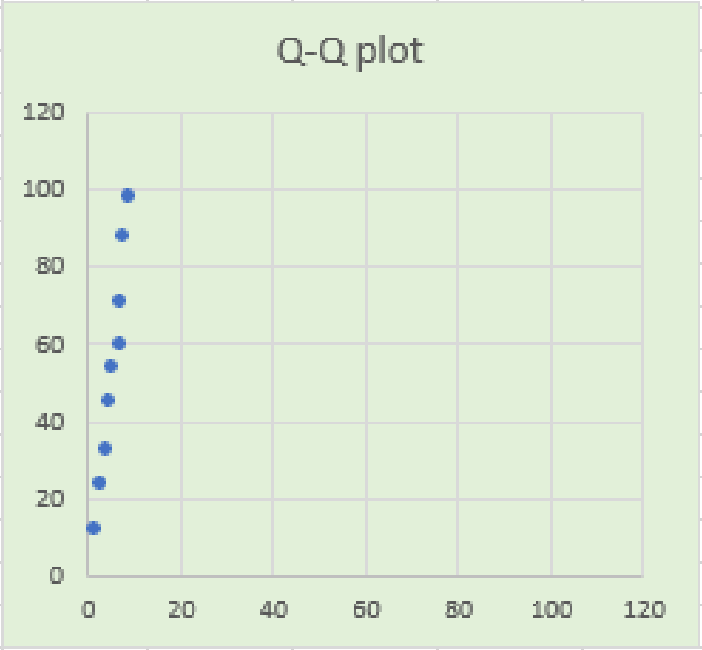
- The sample sizes do not need to be equal.
- Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers
- The q-q plot is used to answer the following questions: Do two data sets come from populations with a common distribution? Do two data sets have common location and scale? Do two data sets have similar distributional shapes? Do two data sets have similar tail behavior?

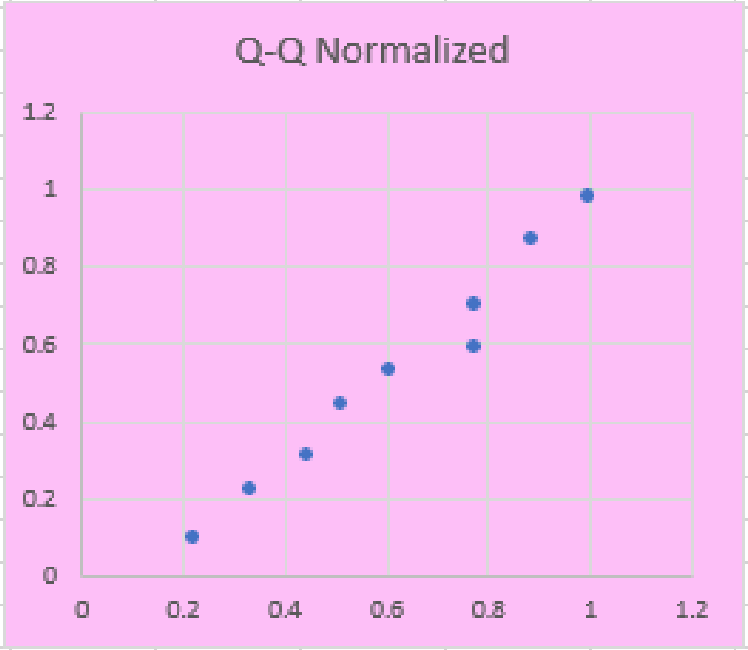
Details Skipped

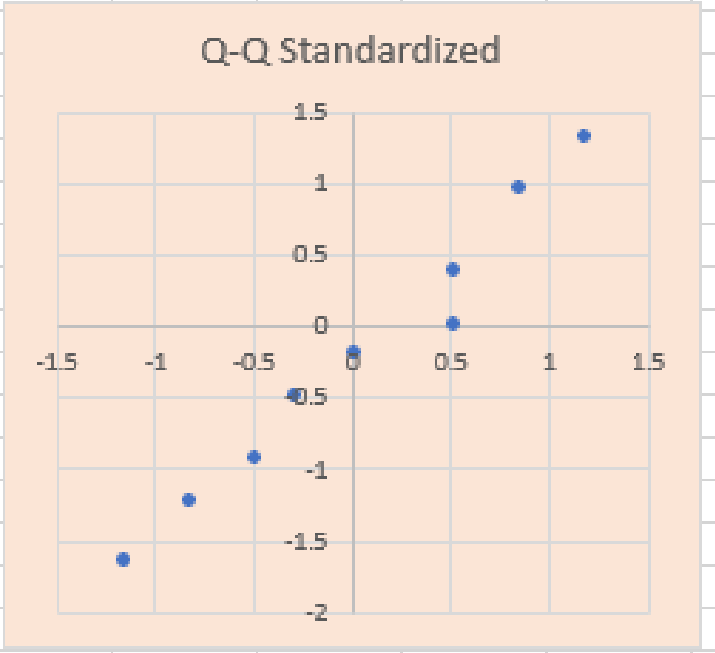
http://www.itl.nist.gov/div898/handbook/eda/section3/qqplot.htm

Normalization and Standardization

2	0.22222	-1.1505	11.6	0.09796	-1.63879
3	0.33333	-0.81605	23.8	0.22245	-1.2207
4	0.44444	-0.48161	32	0.30612	-0.93968
4.6	0.51111	-0.28094	45	0.43878	-0.49417
5.5	0.61111	0.02007	53.5	0.52551	-0.20288
7	0.77778	0.52174	59.8	0.5898	0.01302
7	0.77778	0.52174	70.4	0.69796	0.37629
8	0.88889	0.85619	87.4	0.87143	0.95888
9	1	1.19064	98	0.97959	1.32214







Convention for Bivariate Data

Variable 1	Variable 2	Display
Categorical	Categorical	Crosstabs Stacked Plot
Categorical	Continuous	Boxplot
Continuous	Continuous	Scatterplot Box Plot

Convention for Bivariate Data

Variable 1	Variable 2	Display
Categorical	Categorical	Crosstabs Stacked Plot
Categorical	Continuous	Boxplot
Continuous	Continuous	Scatterplot
		Box Plot

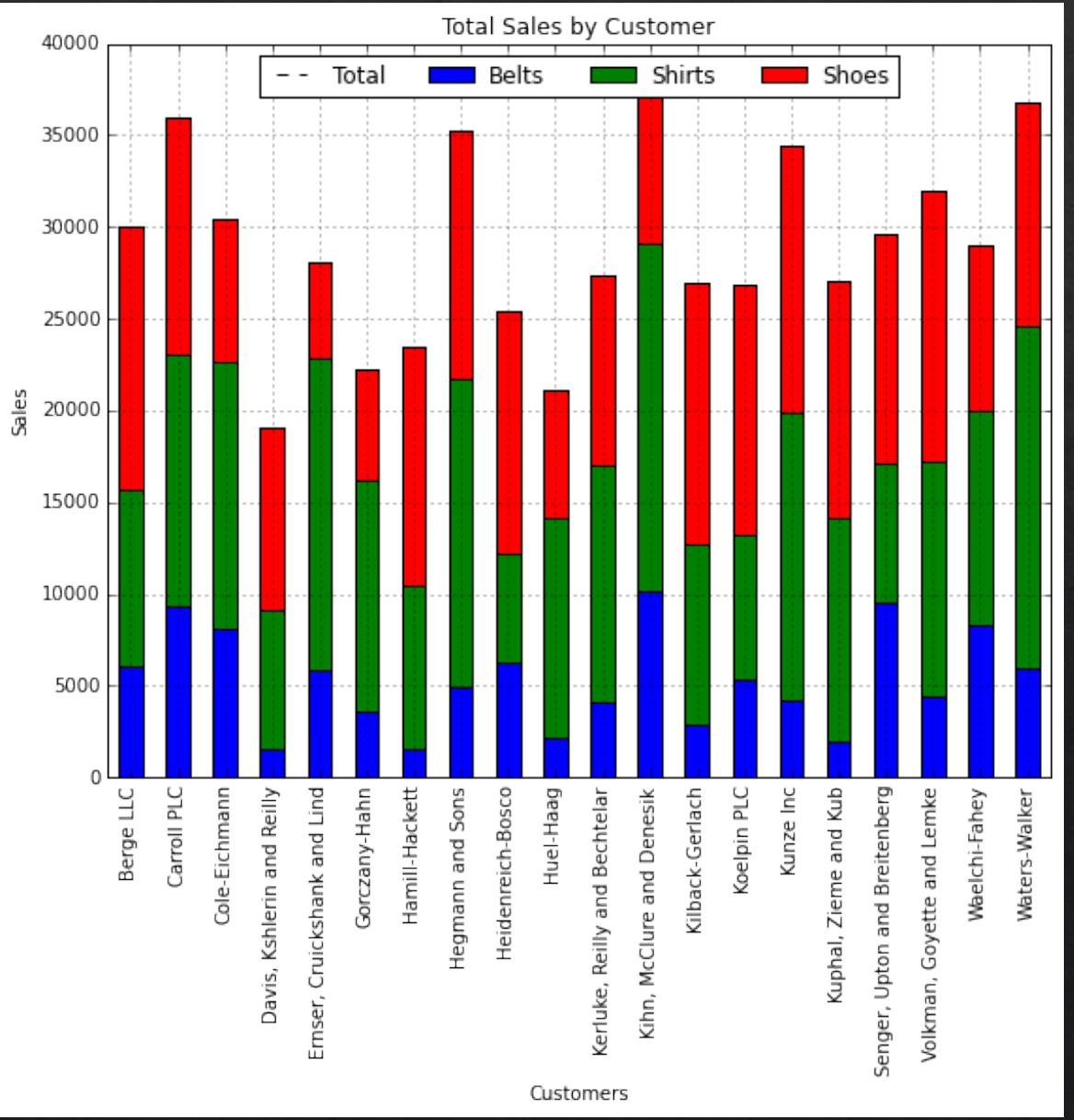
Class Rank * Do you live on campus? * State of residence Crosstabulation

Count

State of residence			Do you live on campus?		Total
			Off-campus	On-campus	
In state	Class Rank	Underclassman	58	110	168
		Upperclassman	108	7	115
	Total		166	117	283
Out of state	Class Rank	Underclassman	13	30	43
		Upperclassman	39	2	41
	Total		52	32	84
Total	Class Rank	Underclassman	71	140	211
		Upperclassman	147	9	156
	Total		218	149	367

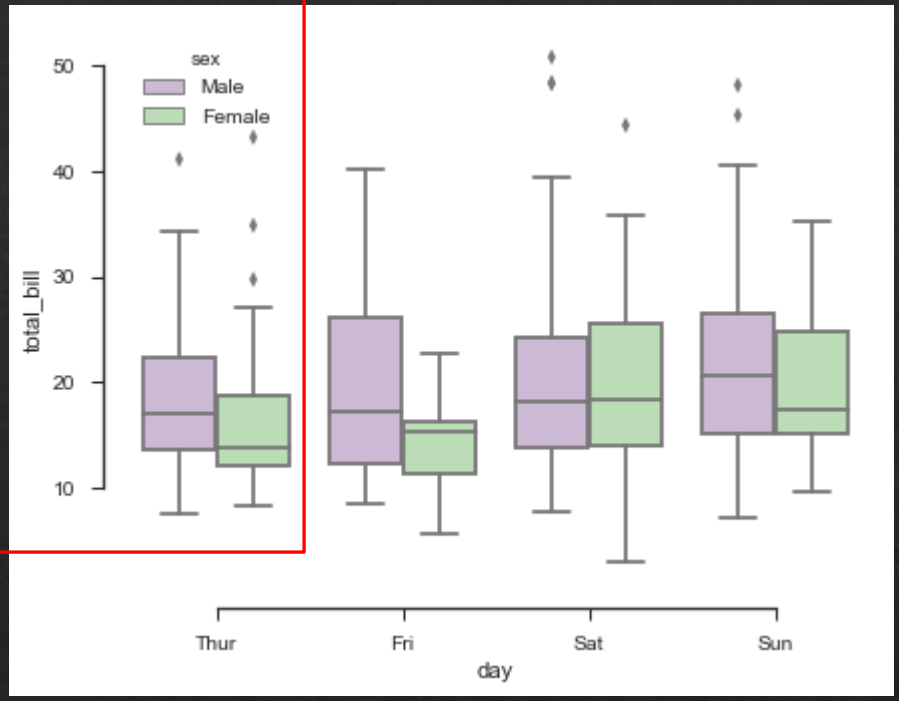
Convention for Bivariate Data

Variable 1	Variable 2	Display
Categorical	Categorical	Crosstabs Stacked Plot
Categorical	Continuous	Boxplot
Continuous	Continuous	Scatterplot Box Plot



Convention for Bivariate Data

Variable 1	Variable 2	Display
Categorical	Categorical	Crosstabs Stacked Plot
Categorical	Continuous	Boxplot
Continuous	Continuous	Scatterplot
		Box Plot

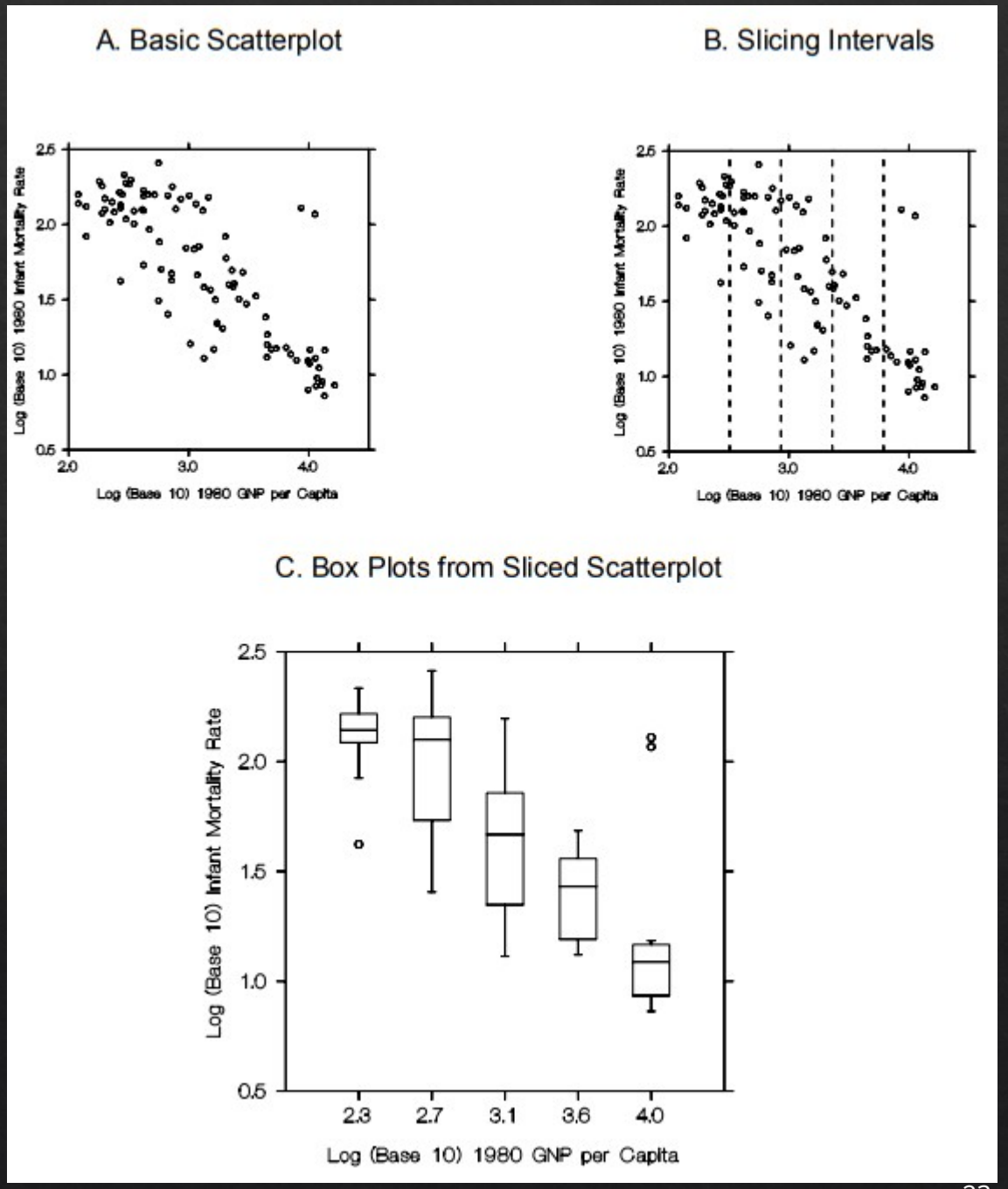


Convention for Bivariate Data

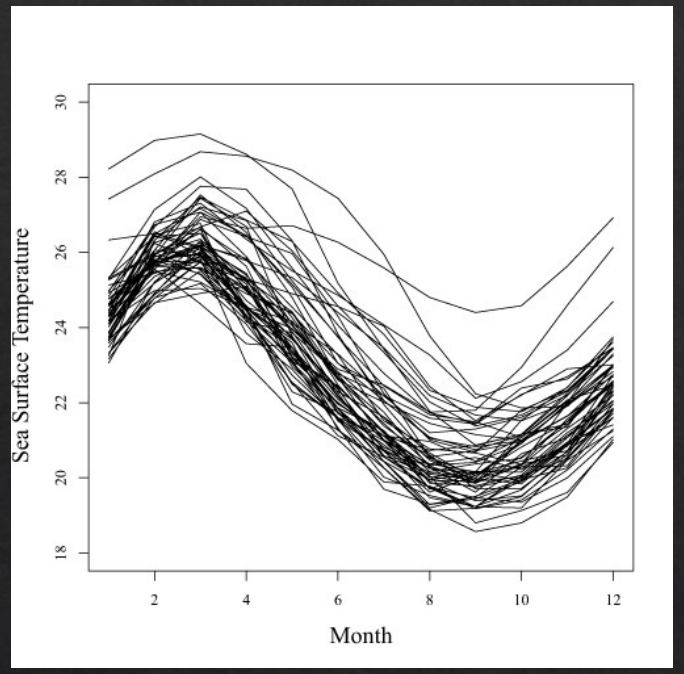
Variable 1	Variable 2		Display
Categorical	Categorical		Crosstabs Stacked Plot
Categorical	Continuous		Boxplot
Continuous	Continuous	?	Scatterplot Box Plot

Slicing a Scatterplot

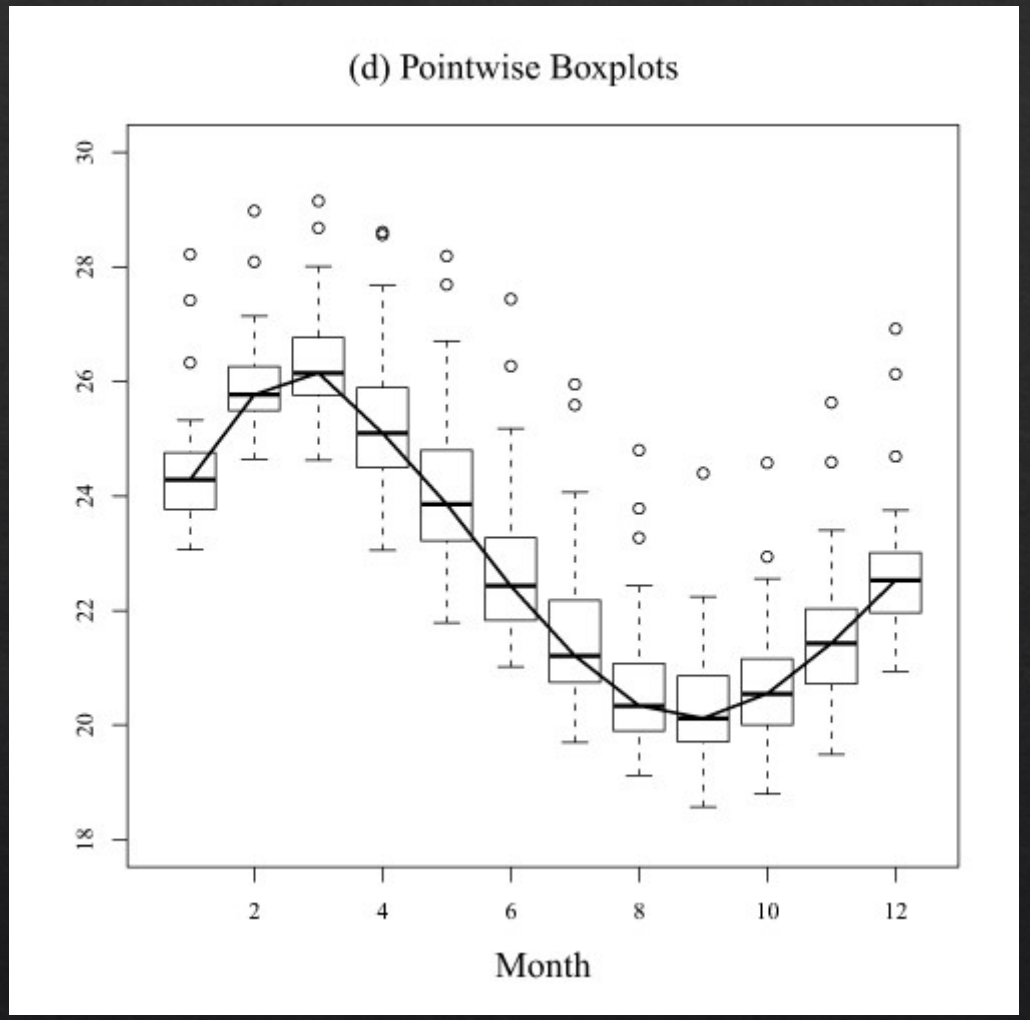
- Scatterplots are usually used to examine functional dependence between variables
- Visual assessment of functional dependence in "raw" data is problematic
- Dividing the plotting region of the scatterplot into a series of vertical "slices" enables visual display of conditional Y distributions



Box Plot



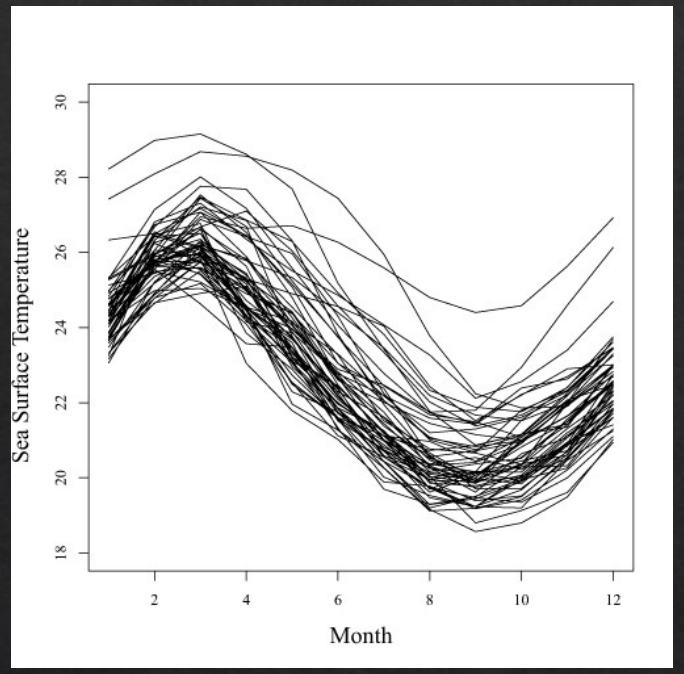
Data of monthly sea surface temperatures Pacific Ocean from 1951 to 2007.



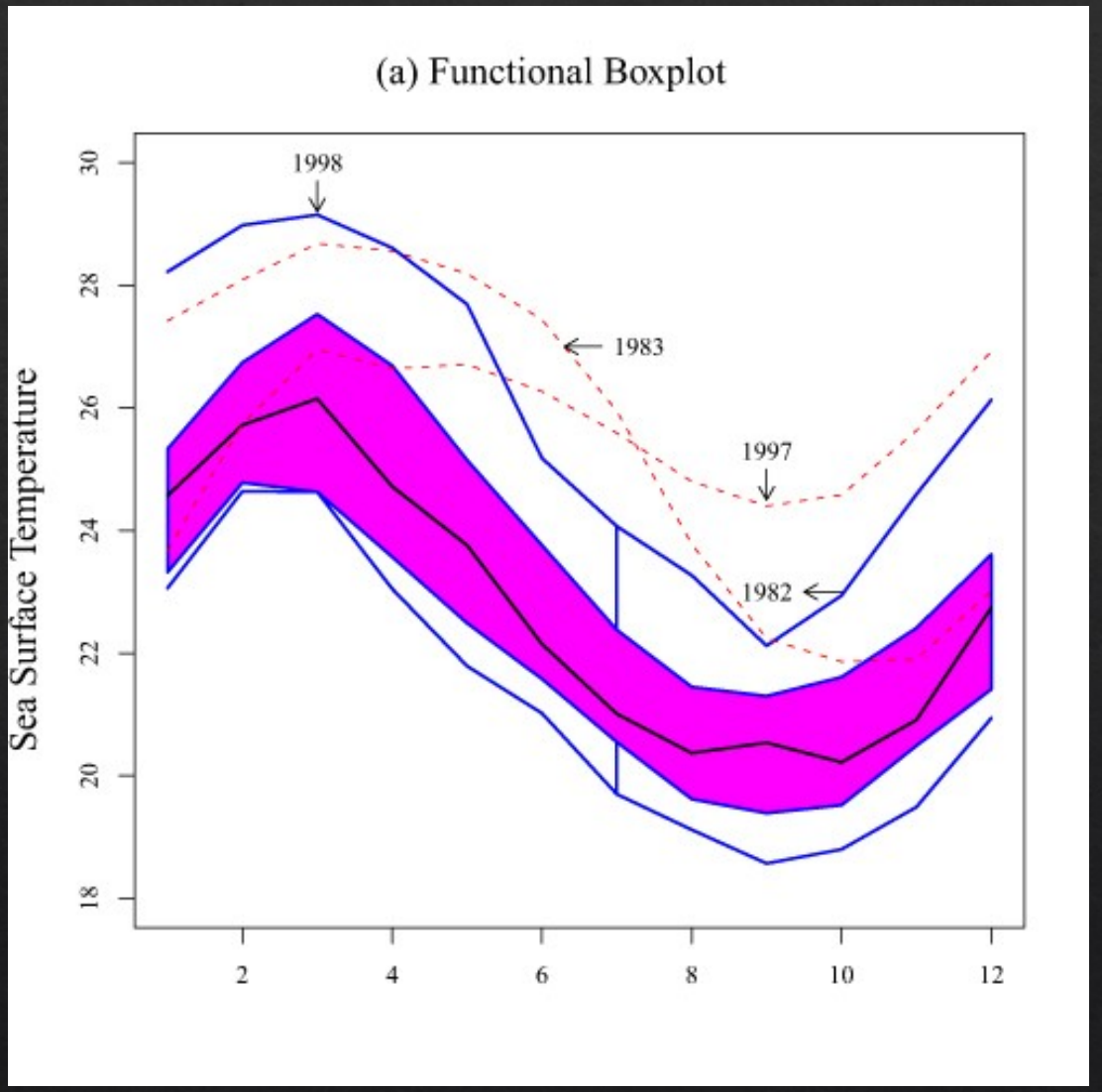
The pointwise boxplots of sea surface temperatures (SST) with medians connected by a black line.

https://en.wikipedia.org/wiki/Functional_boxplot

Functional Box Plot

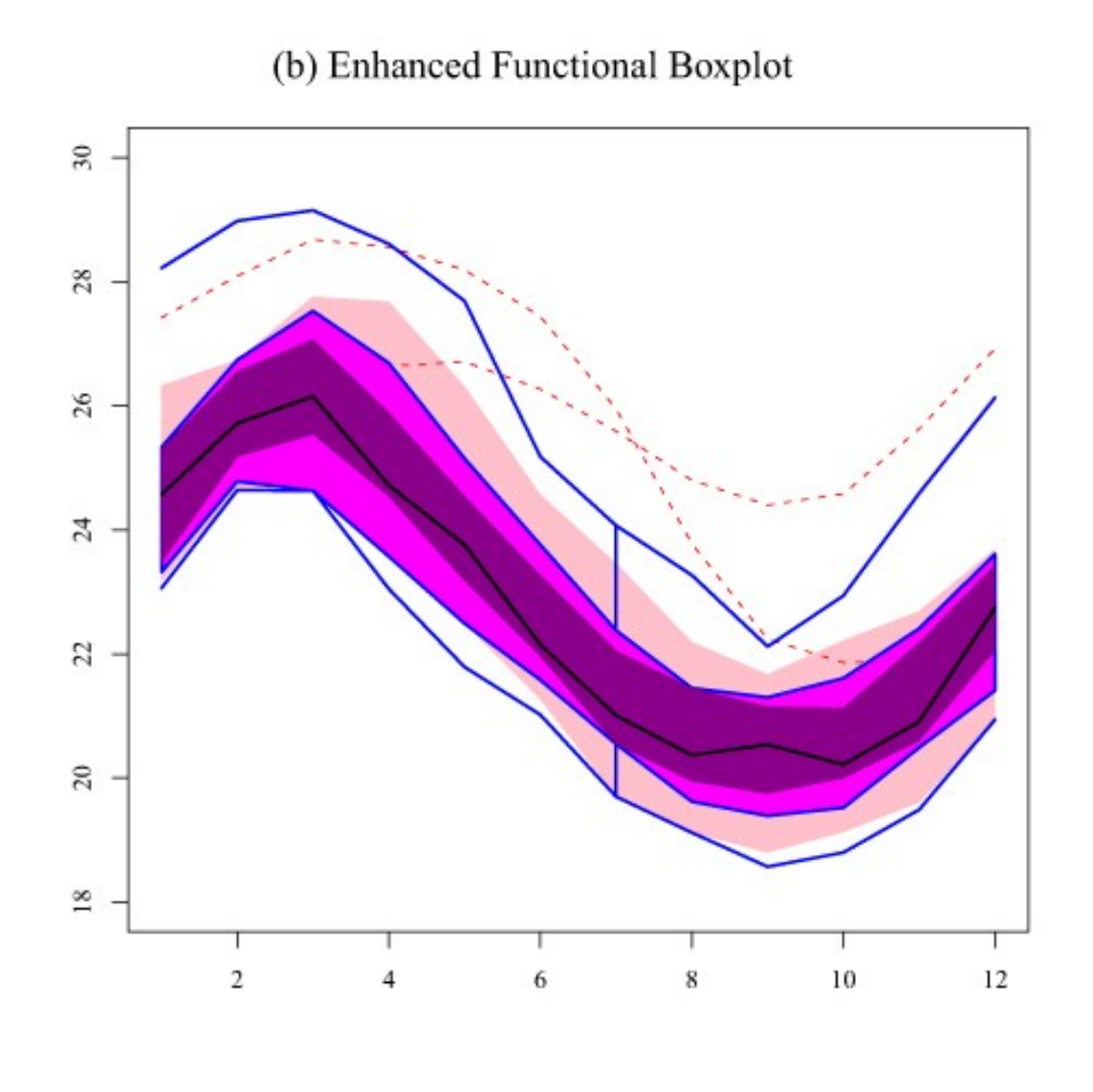


Data of monthly sea surface temperatures Pacific Ocean from 1951 to 2007.



A black curve representing the median curve. The red dashed curves are the outlier candidates detected by the 1.5 times the 50% central region rule.

Functional Box Plot



The enhanced functional boxplot of SST with dark magenta denoting the 25% central region, magenta representing the 50% central region and pink indicating the 75% central region.

Yours to Explore

- Bagplot: https://en.wikipedia.org/wiki/Bagplot
- Candlestick chart: https://en.wikipedia.org/wiki/Candlestick_chart
- Kagi Chart: https://en.wikipedia.org/wiki/Kagi_chart
- Fan Chart: https://en.wikipedia.org/wiki/Fan_chart_(statistics)

... there are still a lot out there!

Class Activity

- How can you improve the readability of these charts?

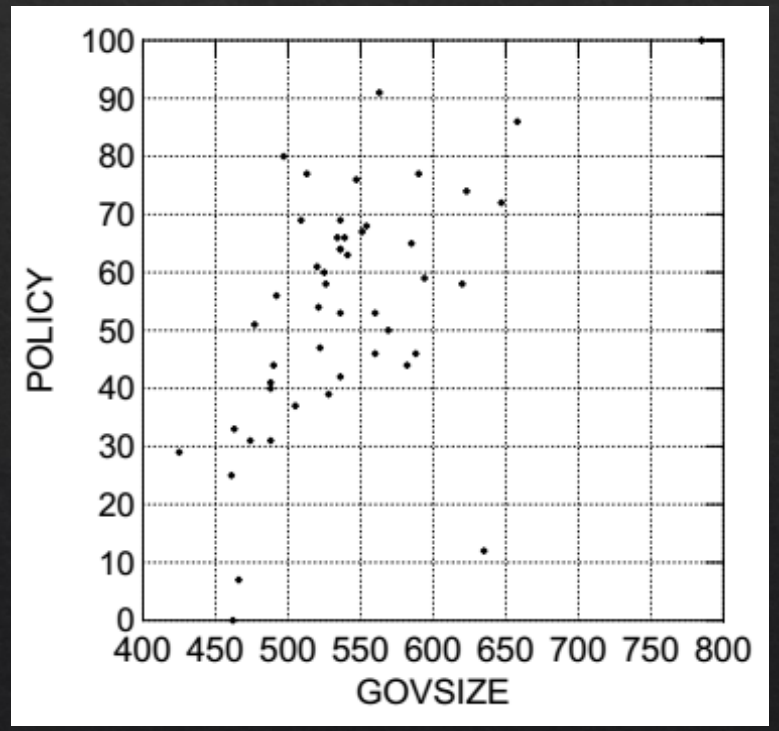


Chart 1

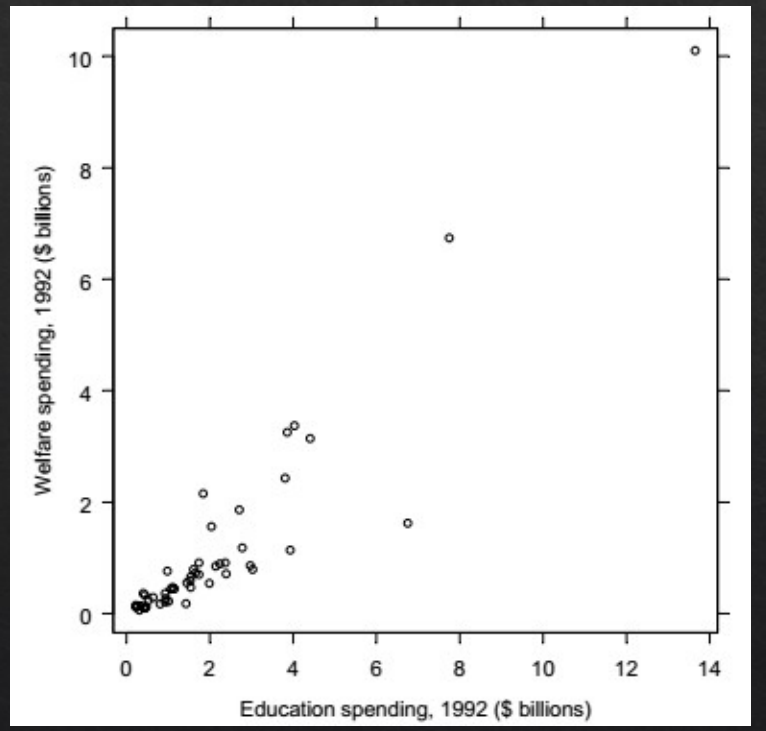
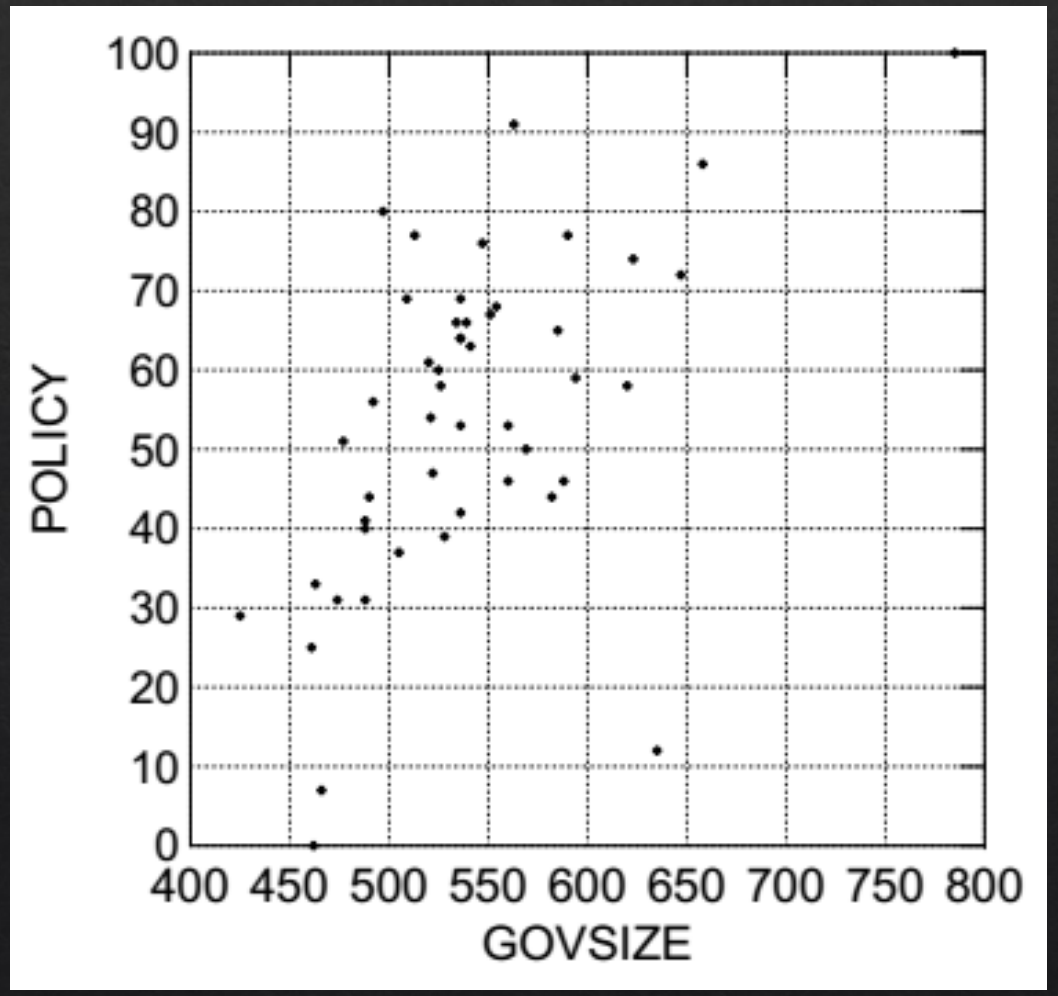
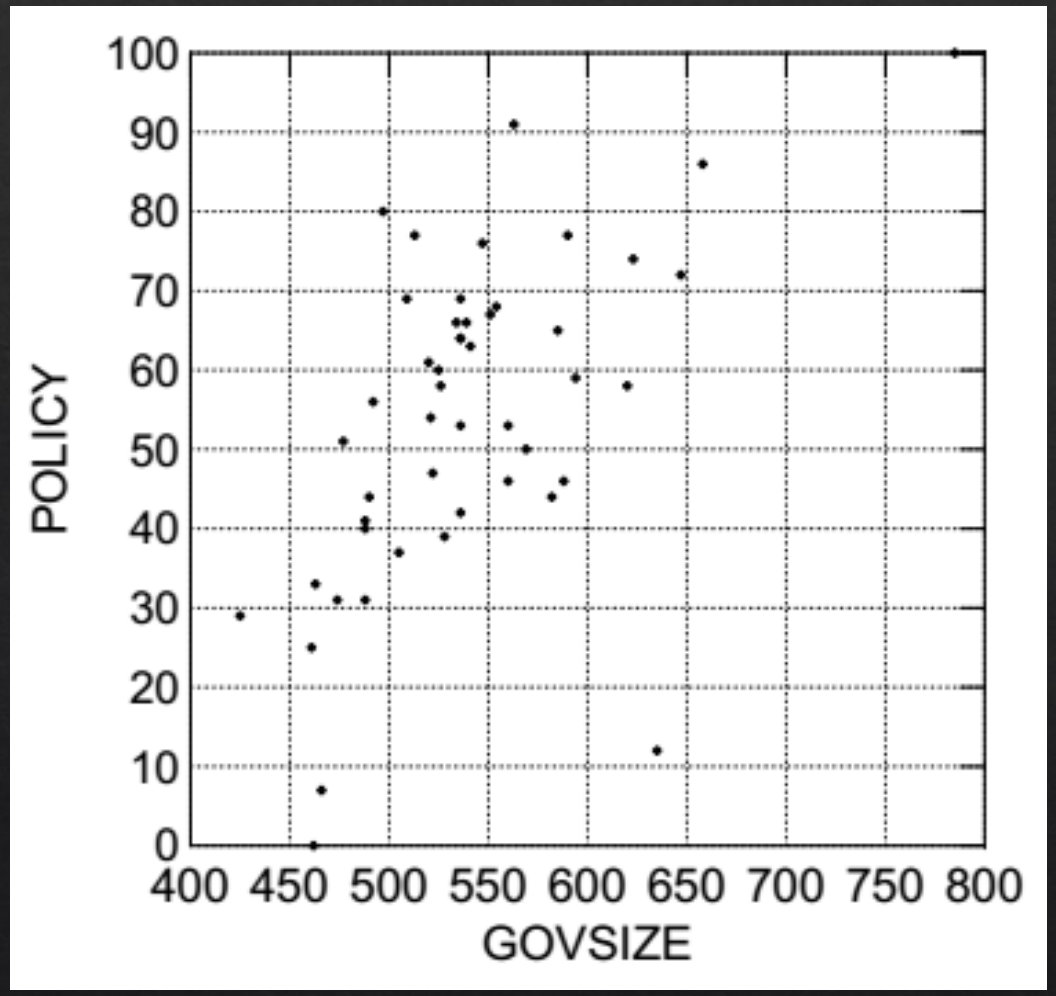


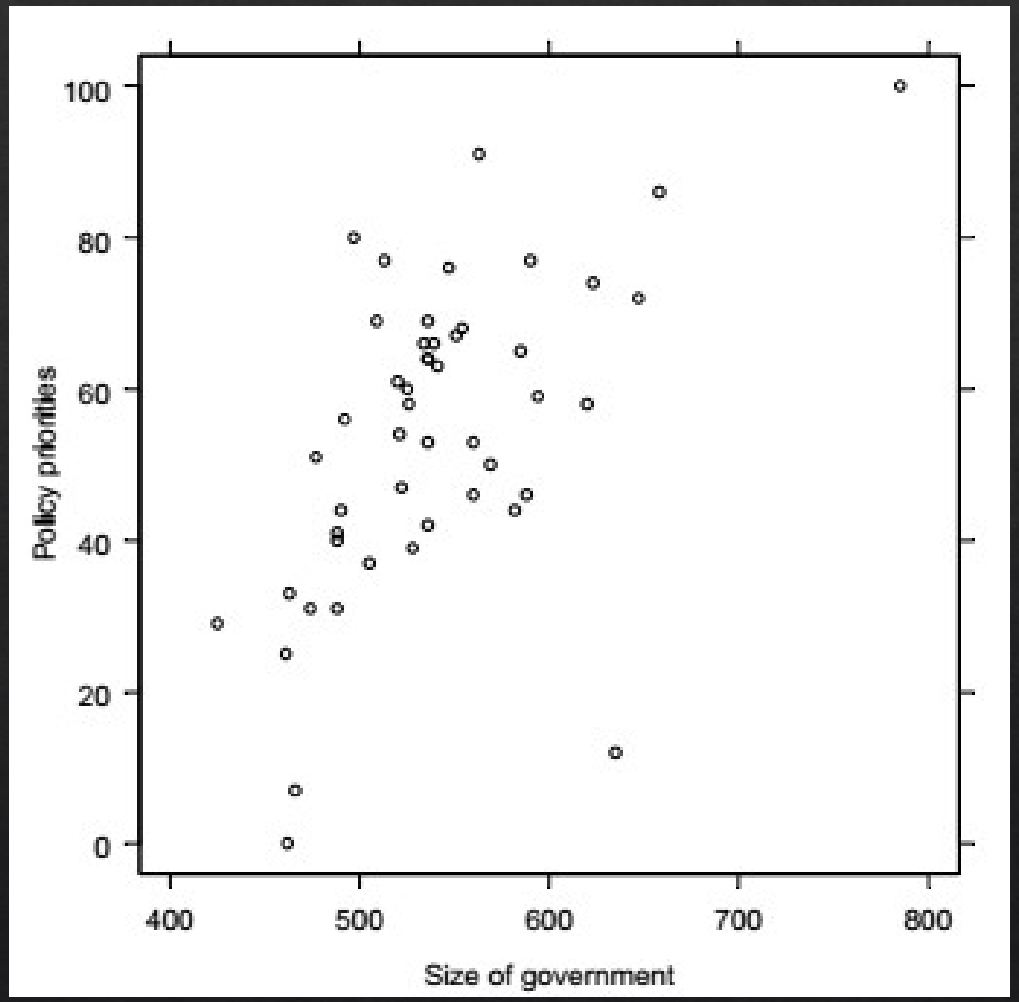
Chart 2

Poorly Constructed Scatterplot

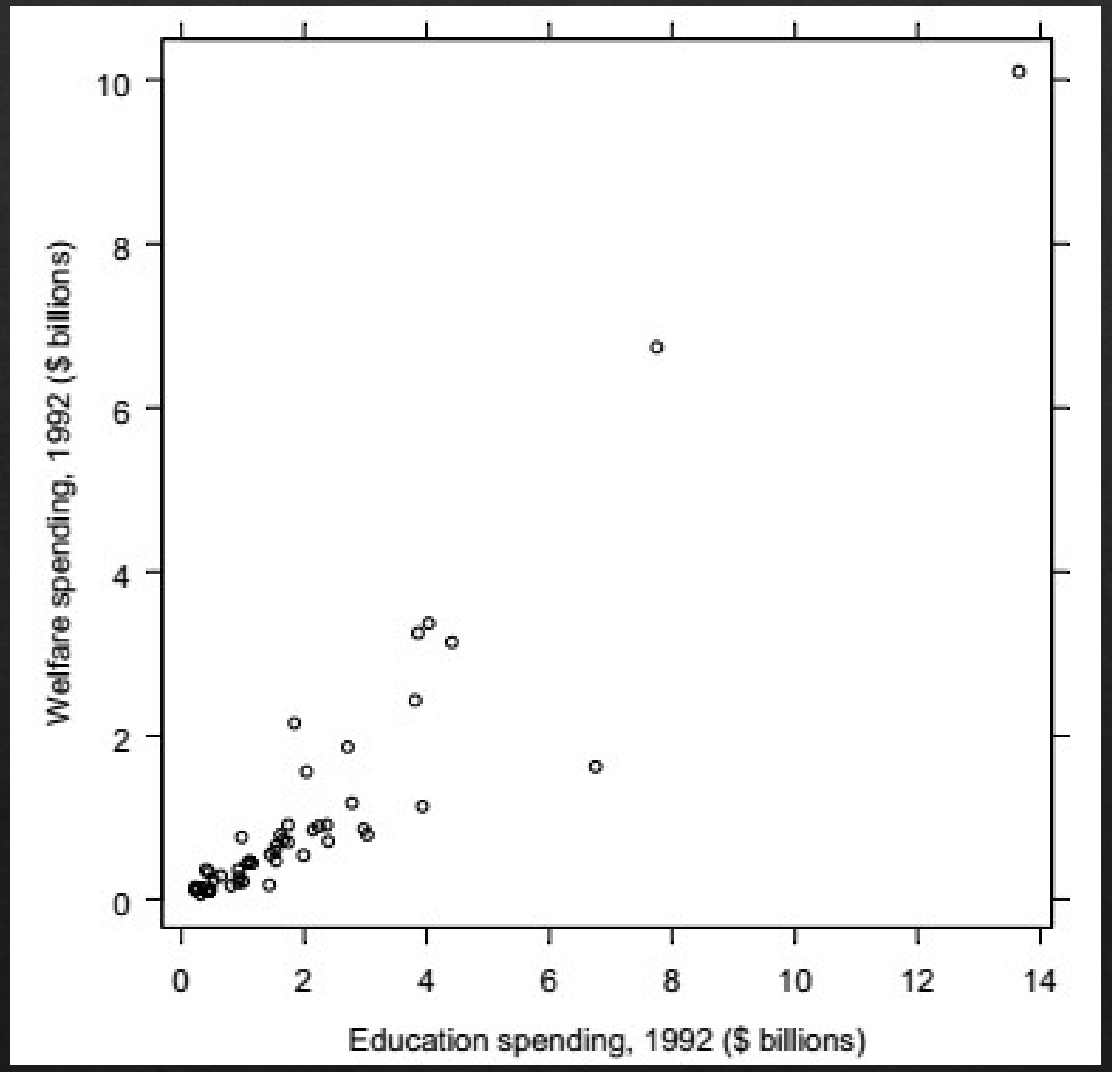


Possible Improvement

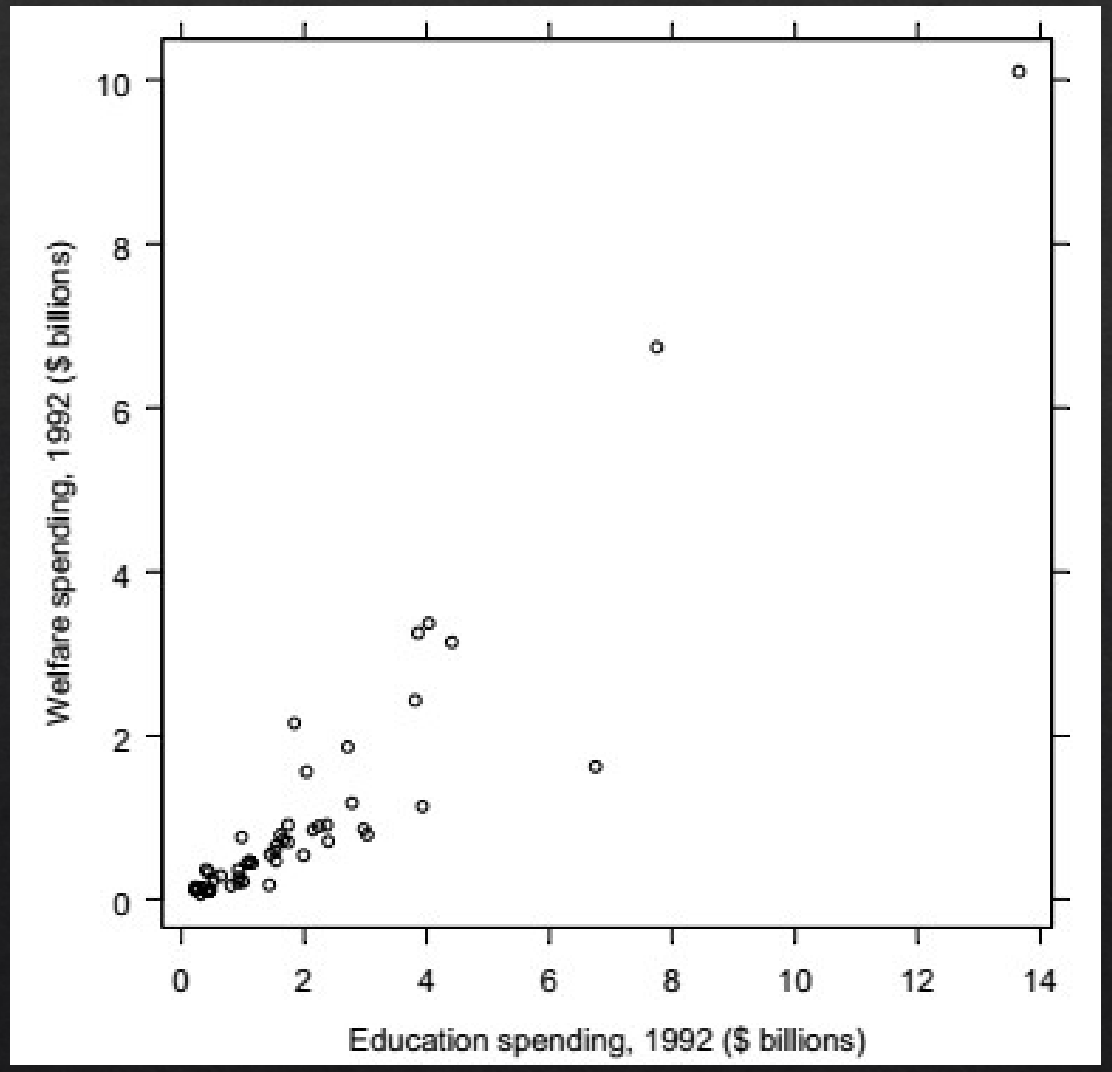


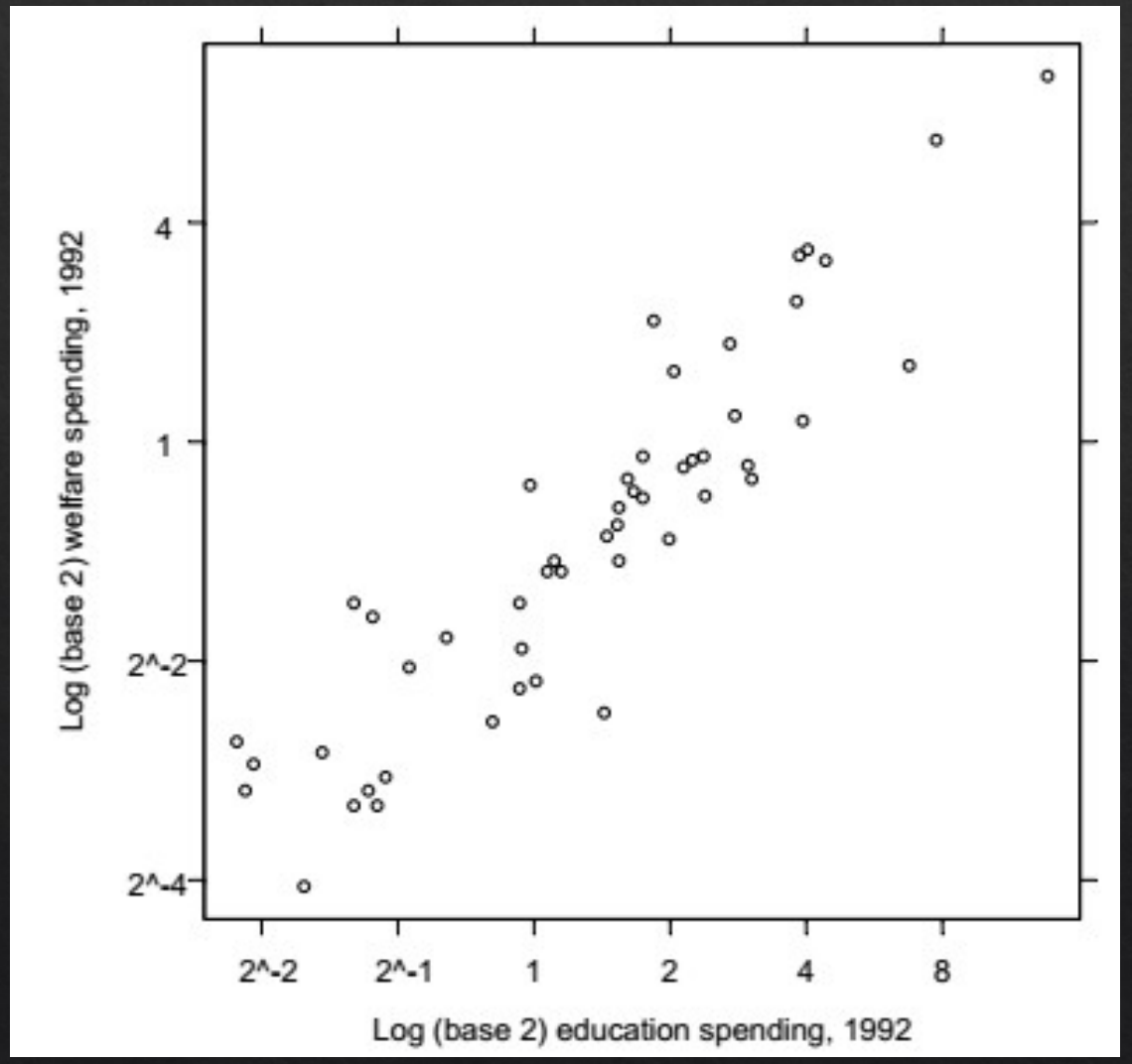


Poorly Constructed Scatterplot



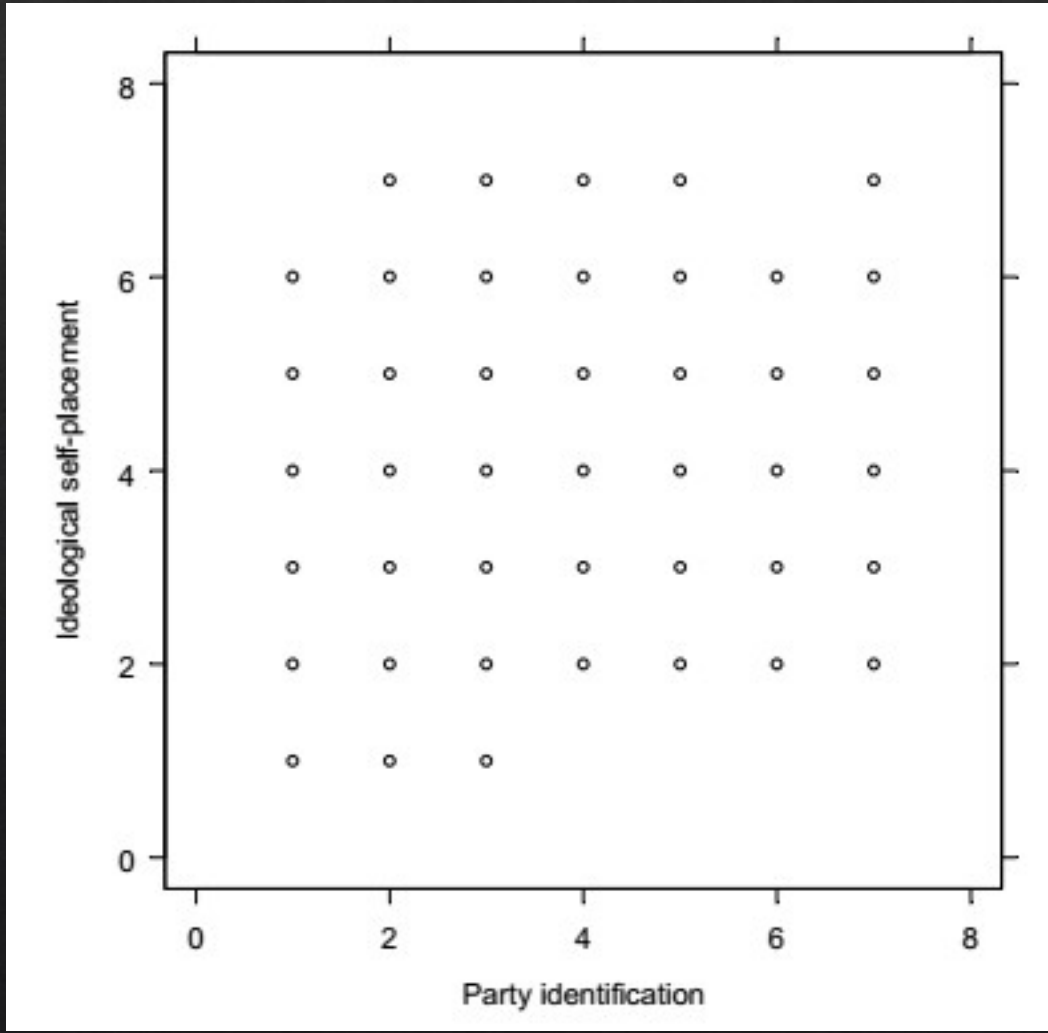
Possible Improvement



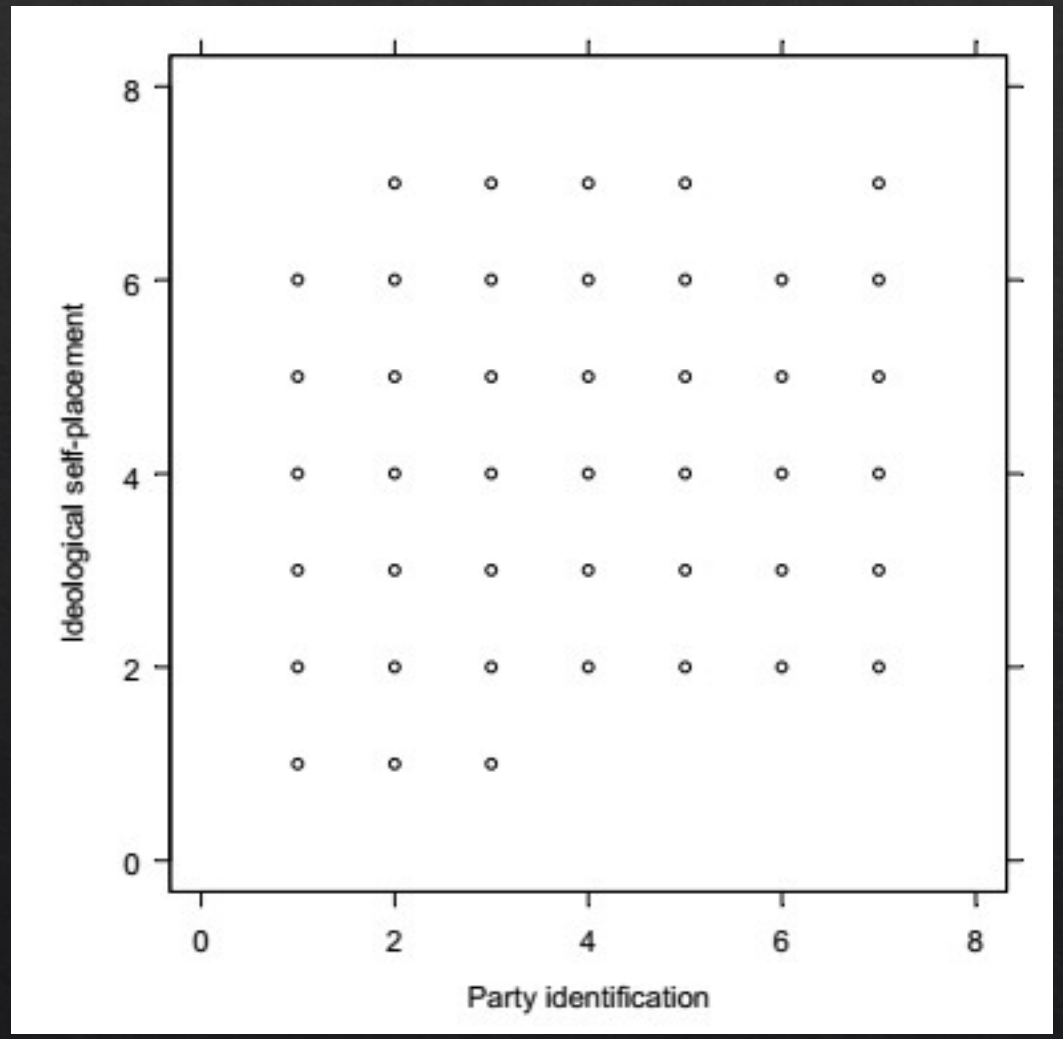


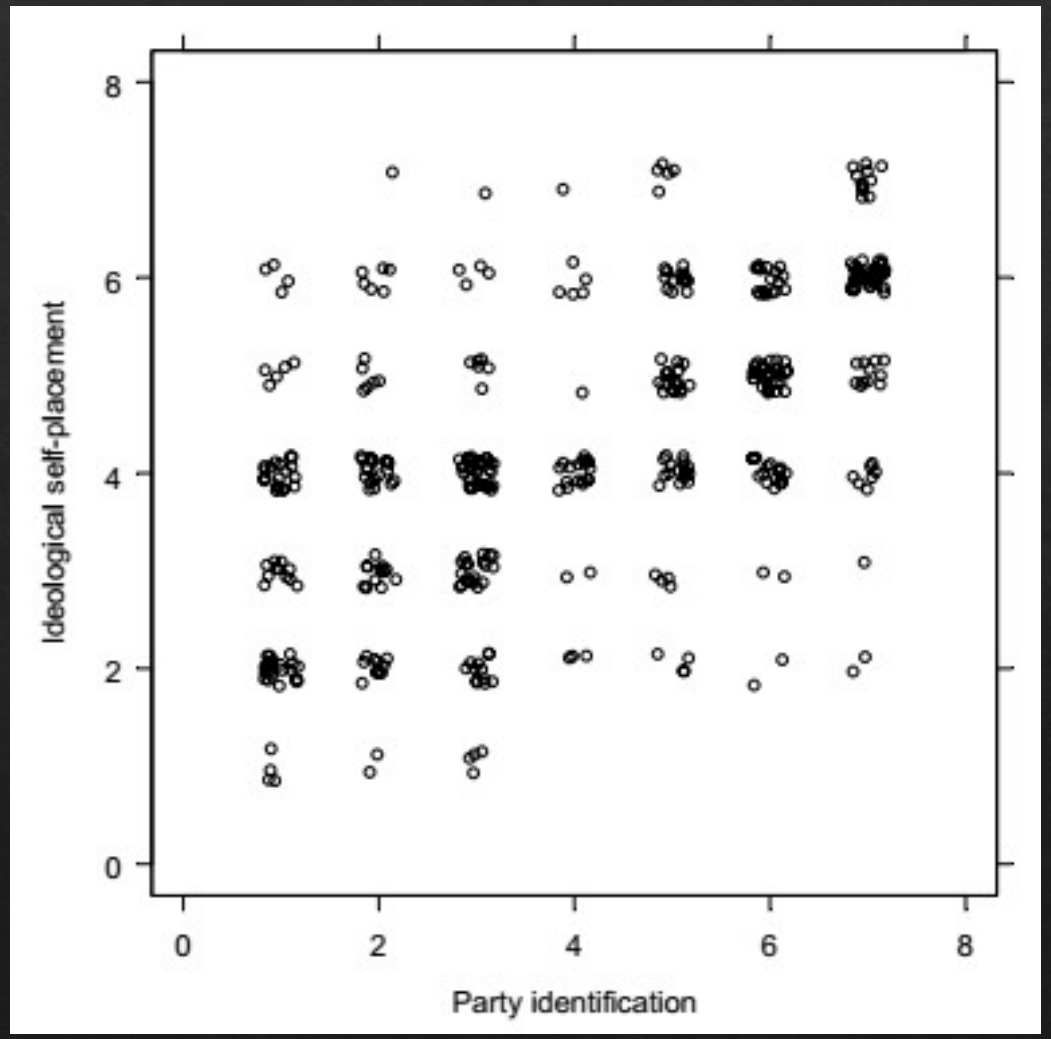
Transforming variable values in order to improve visual resolution in a scatterplot

Poorly Constructed Scatterplot



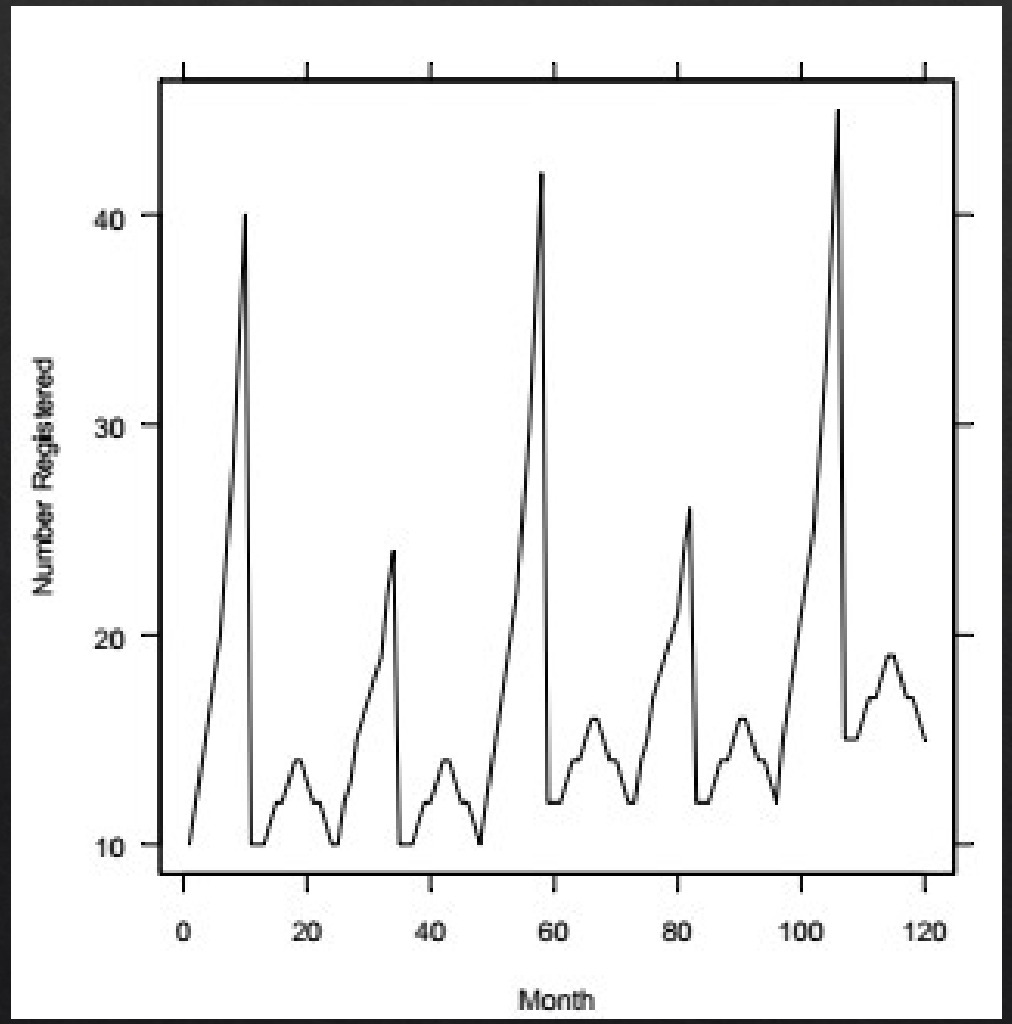
Possible Improvement



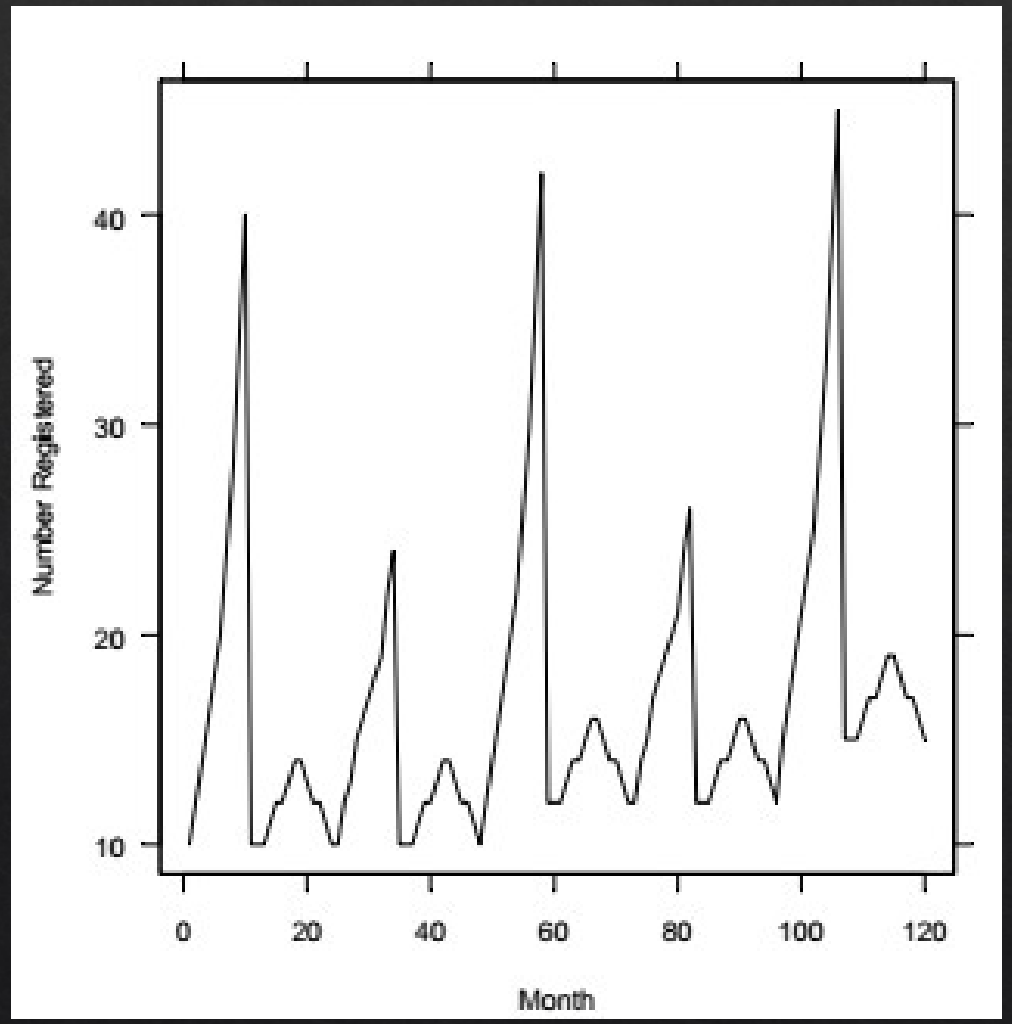


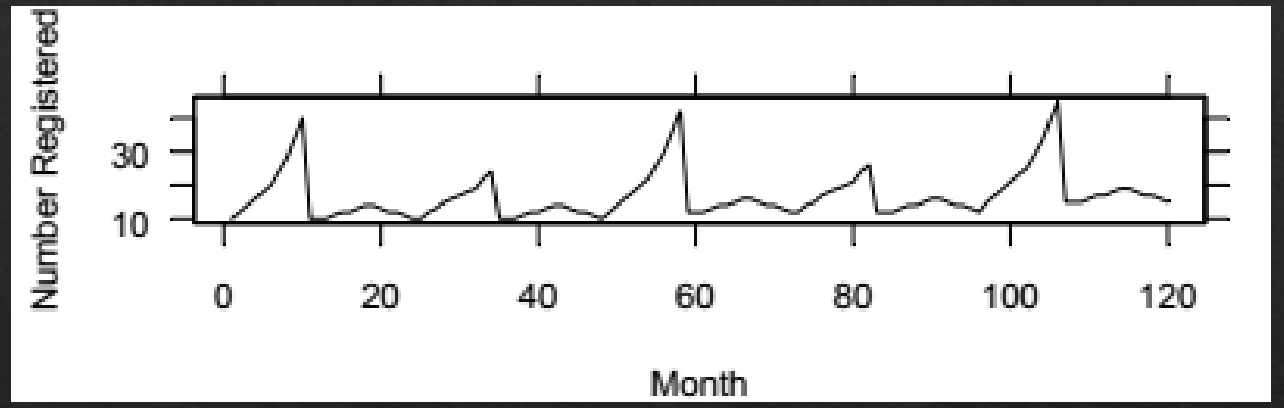
Add a little random noise to the data in order to see the cloud more clearly

Poorly Constructed Scatterplot



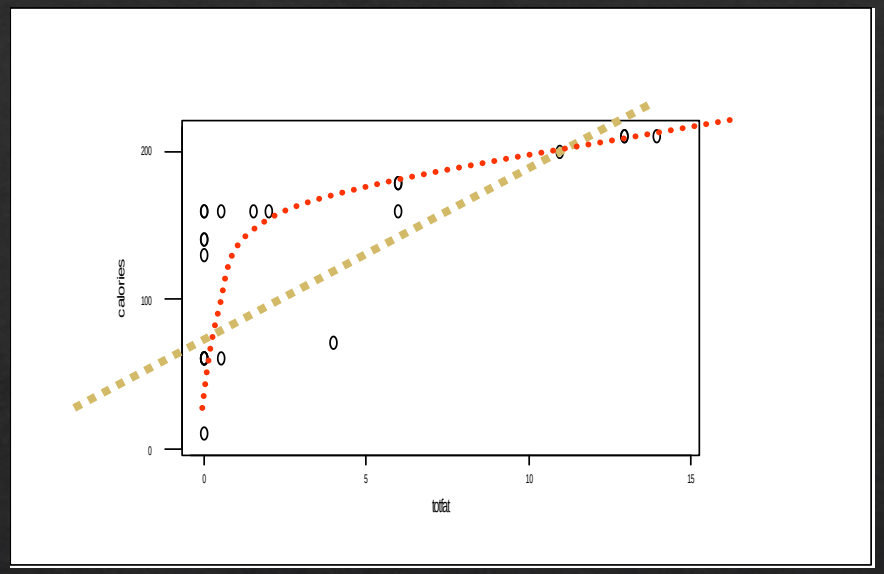
Possible Improvement



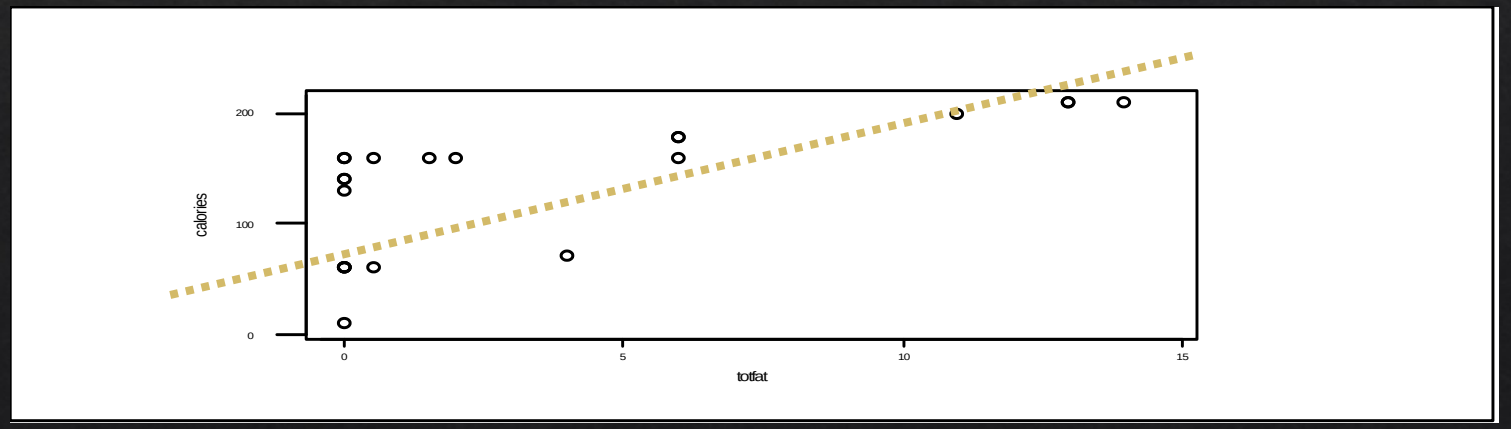


Control aspect ratio of the bounding box for better visual perception.

Careful!



Changing the relative lengths of the axes can change how the relationship is perceived.



General Guidelines

- Use axes on all four sides to enclose the plotting region
- Clear axis labels
- Rectangular grid lines within the plotting region are usually unnecessary
- Tick marks should point outward, rather than inward and relatively few tick marks should be used on each axis
- Data rectangle should be slightly smaller than the scale rectangle.
- If necessary, transform data values (alternatively, used transformed scales on coordinate axes) so plotted points fill up as much of the data rectangle as possible

Learning Goals

- Learning Basic Statistical Plots
- Understanding how to interpret Box-Plots
- Evaluating the effectiveness of a chart or describing possible improvements
- Given a data set, planning quick plots to get insight from the data
- Being able to choose charts that best suit the context