STA286 Lecture 02

Neil Montgomery

Last edited: 2017-01-11 09:49

numerical summaries of dataset variables — definitions first with examples after

sample measures of “location”

The dataset is often called the “sample” (no matter where the data came from).

sample measures of “location”

The dataset is often called the “sample” (no matter where the data came from).

For a particular numerical variable in the sample with observations:

$$\{x_1, x_2, \ldots, x_n\}$$

the *sample average* is just the arithmetic mean:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

sample measures of "location"

The dataset is often called the "sample" (no matter where the data came from).

For a particular numerical variable in the sample with observations:

$$\{x_1, x_2, \ldots, x_n\}$$

the *sample average* is just the arithmetic mean:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Could be sensitive to extreme observations.

sample medians, sample percentiles

Order the observations:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

A number that divides the observations into two groups is called a *sample median*. For example:

$$\tilde{x} = \begin{cases} x_{((n+1)/2)} & : n \text{ odd} \\ \left(x_{(n/2)} + x_{(n/2+1)}\right)/2 & : n \text{ even} \end{cases},$$

which is harder to write out than it is to understand.

sample medians, sample percentiles

Order the observations:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

A number that divides the observations into two groups is called a *sample median*. For example:

$$\tilde{x} = \begin{cases} x_{((n+1)/2)} & : n \text{ odd} \\ \left(x_{(n/2)} + x_{(n/2+1)}\right)/2 & : n \text{ even} \end{cases},$$

which is harder to write out than it is to understand.

A *sample p^{th} percentile* has $p\%$ of the data below or equal to it. Special cases include (sample...): quartiles, quintiles, deciles, and indeed the median itself.

sample measures of variation of a numerical variable

Very (too?) simple measure: *sample range* which is just $x_{(n)} - x_{(1)}$.

sample measures of variation of a numerical variable

Very (too?) simple measure: *sample range* which is just $x_{(n)} - x_{(1)}$.

More common to consider the set of deviations from the sample mean:

$$x_i - \overline{x}$$

Adding them up just gives 0, so instead consider positive functions such as:

$$|x_i - \overline{x}| \qquad \text{or} \qquad (x_i - \overline{x})^2$$

sample measures of variation of a numerical variable

Very (too?) simple measure: *sample range* which is just $x_{(n)} - x_{(1)}$.

More common to consider the set of deviations from the sample mean:

$$x_i - \overline{x}$$

Adding them up just gives 0, so instead consider positive functions such as:

$$|x_i - \overline{x}| \qquad \text{or} \qquad (x_i - \overline{x})^2$$

Summing up over all the observations gives the *sum of absolute deviations* (aka SAD) and the *sample variance* respectively. Notation and formula:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$$

sample standard deviation

s^2 is essentially the average squared deviation. (More on $n-1$ later in the course.)

The sample variance is good for theory but has an inconvenient unit. More practical is the *sample standard deviation*:

$$s = \sqrt{s^2}$$

numerical summaries for categorical variables

The oil readings data had one categorical variable, the `Ident` variable which is just a serial number. I added a fake one `TakenBy` for illustration.

```
## # A tibble: 5 × 17
##    Ident       Date WorkingAge   TakenBy    Fe    Al    Cu    Cr
##   <fctr>     <date>      <dbl>    <fctr> <dbl> <dbl> <dbl> <dbl>
## 1 448576 1999-05-10        243 EMPL_0917    13     5    14     1
## 2 448576 1999-07-26        569 EMPL_0917    18     6    25     1
## 3 448576 1999-09-29        830 EMPL_9375    26     6    35     1
## 4 448576 1999-10-08        862 EMPL_0917    15     9    14     1
## 5 448576 1999-11-02        946 EMPL_9375    14     4    19     1
## # ... with 9 more variables: Si <dbl>, Pb <dbl>, Ph <dbl>, Ca <dbl>,
## #   Zn <dbl>, Mg <dbl>, Mo <dbl>, Sn <dbl>, Na <dbl>
```

tables of counts (or proportions)

A categorical variable could also be called a *factor* variable with *levels*, and to tabulate the frequency of each level is the way to summarize.

```
## # A tibble: 25 × 3
##      Ident     n proportion
##     <fctr> <int>      <dbl>
## 1   448572    31 0.05065359
## 2   448574    31 0.05065359
## 3   448576    36 0.05882353
## 4   448577    29 0.04738562
## 5   448578    34 0.05555556
## 6   448579    36 0.05882353
## 7   448580    28 0.04575163
## 8   448581    31 0.05065359
## 9   448582    41 0.06699346
## 10  448583    42 0.06862745
## # ... with 15 more rows
```

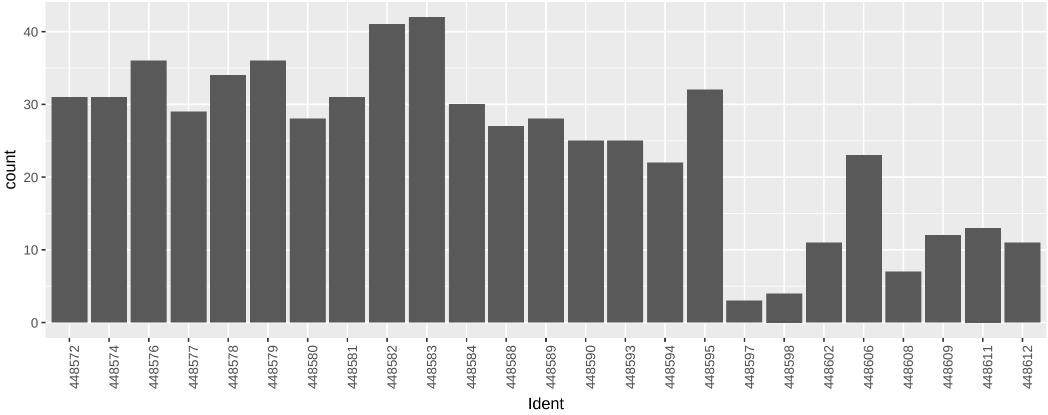
two-way classification with Ident and TakenBy

```
##            Ident
## TakenBy     448572 448574 448576 448577 448578 448579 448580 448581 4485
##   EMPL_0592     18     16      0      0     12      0      0      7
##   EMPL_0917      0      0     18     11      0     22     10      0
##   EMPL_2095      8      8      0      0      8      0      0      7
##   EMPL_4925      0      0     10      9      0      6     10      0
##   EMPL_9134      5      7      0      0     14      0      0     17
##   EMPL_9375      0      0      8      9      0      8      8      0
##            Ident
## TakenBy     448583 448584 448588 448589 448590 448593 448594 448595 4485
##   EMPL_0592      0      0      0     10      0     10      0      0
##   EMPL_0917     24      9     11      0     13      0     10     18
##   EMPL_2095      0      0      0     12      0      7      0      0
##   EMPL_4925     10     11     11      0      7      0     10     10
##   EMPL_9134      0      0      0      6      0      8      0      0
##   EMPL_9375      8     10      5      0      5      0      2      4
##            Ident
```

graphical summaries

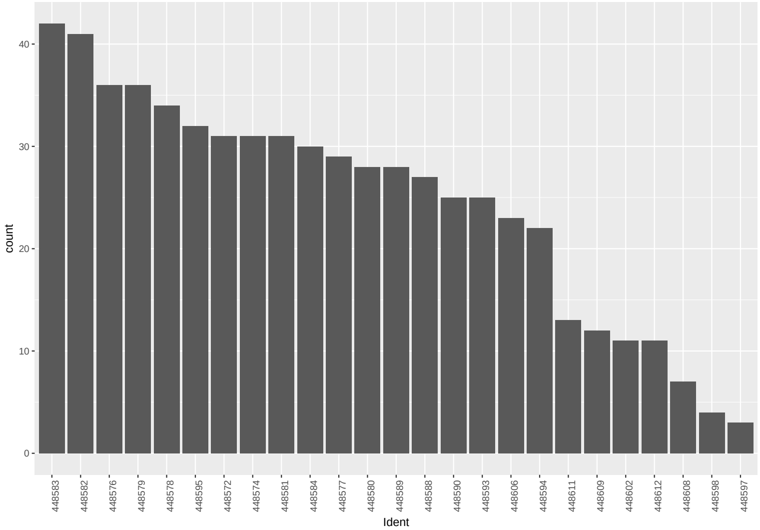
barchart

A barchart is a table of counts, in graphical form.

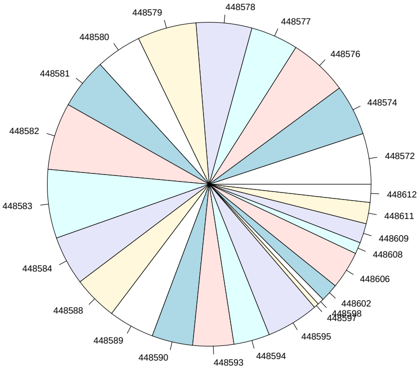


“Pareto” chart

Ordered by count.



piecharts are problematic

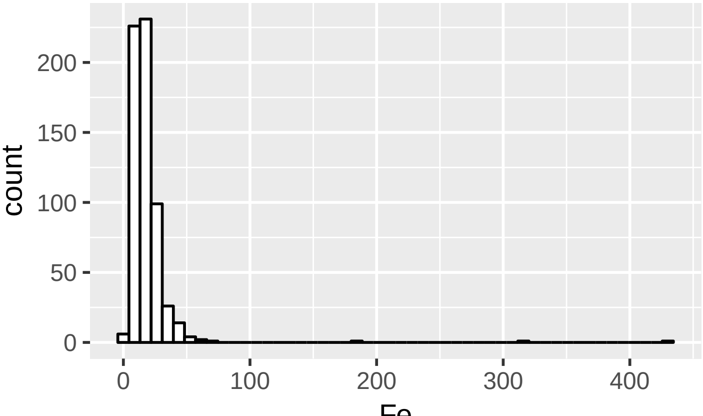


histograms

A histogram is a special case of a barchart.

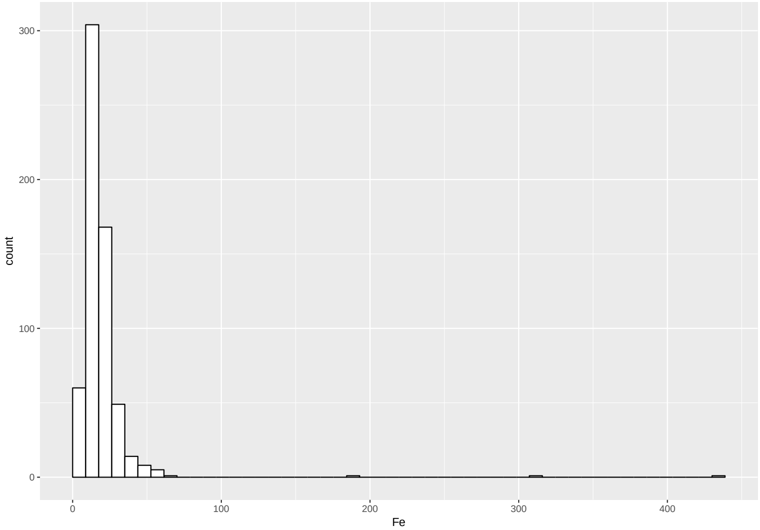
A numerical variable is split into classes and a barchart is made from the table of counts of obvservations within each class.

Histograms are done by the computer. Always play around with the number of classes.

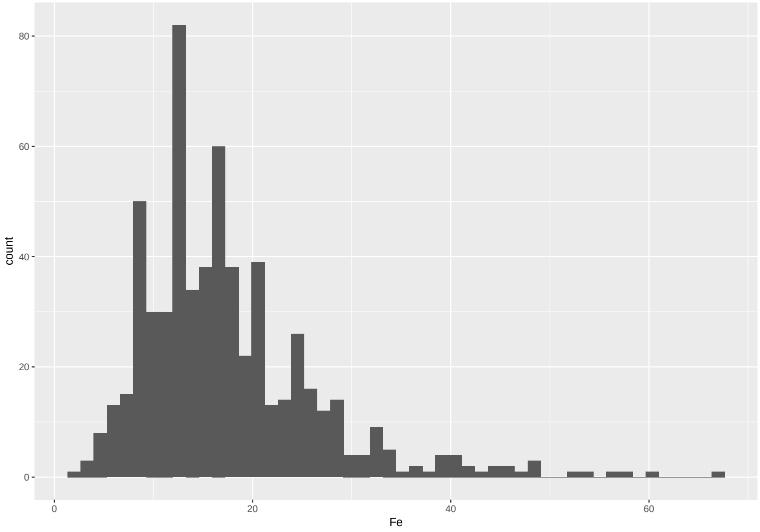


histograms are hard to implement!

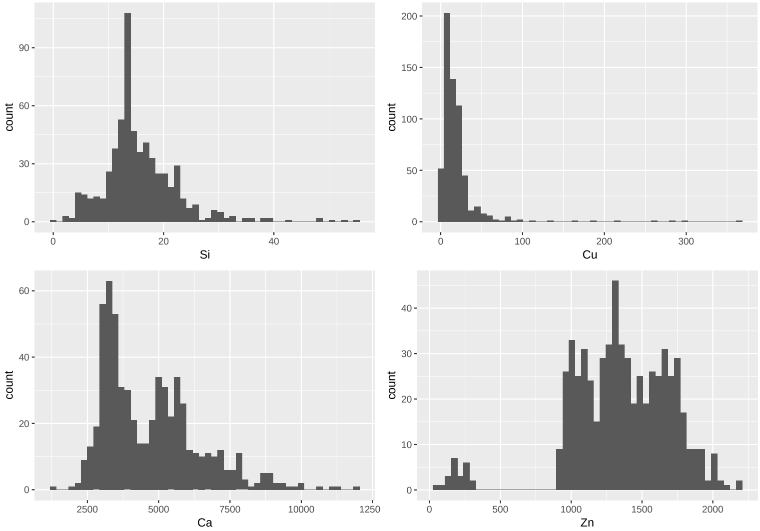
Better picture around 0. Possibly not important for EDA?



histogram without those really big values

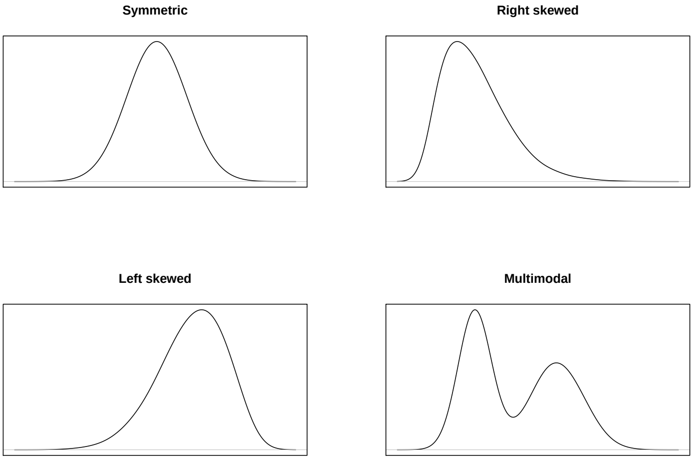


a few more ppm histograms



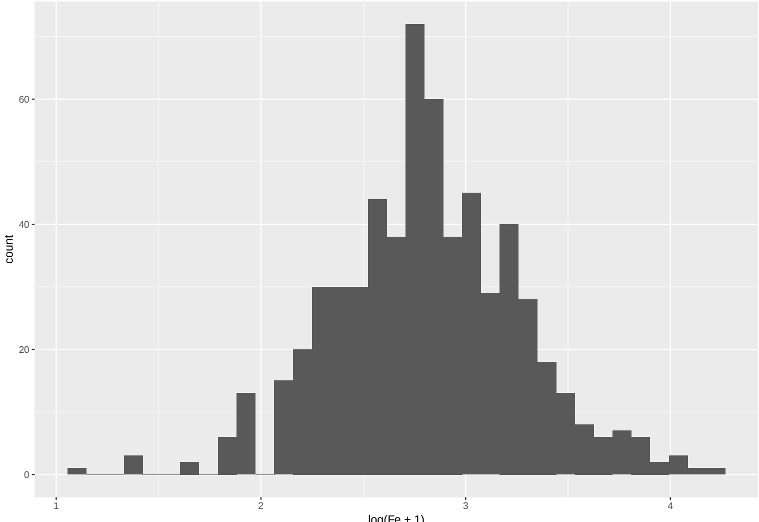
“shapes” of “distributions”

To use a histogram, *glance* at it and look for any of the following (without getting fooled by plot artefacts):



transforming variables

Apply log or square root to a variable will change the shape of the empirical distribution, e.g. transform right-skewed to symmetric.



boxplots

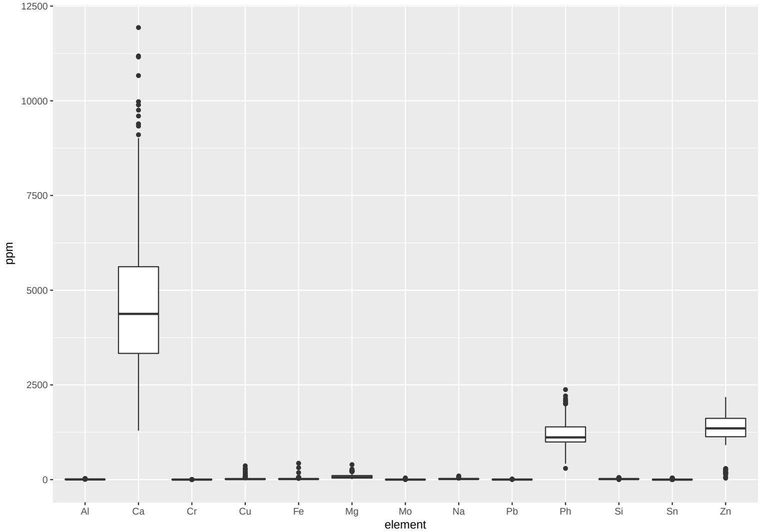
A special plot of these (or similar) five numbers:

min $\qquad$ 25^{th} percentile $\qquad$ median $\qquad$ 75^{th} percentile $\qquad$ max

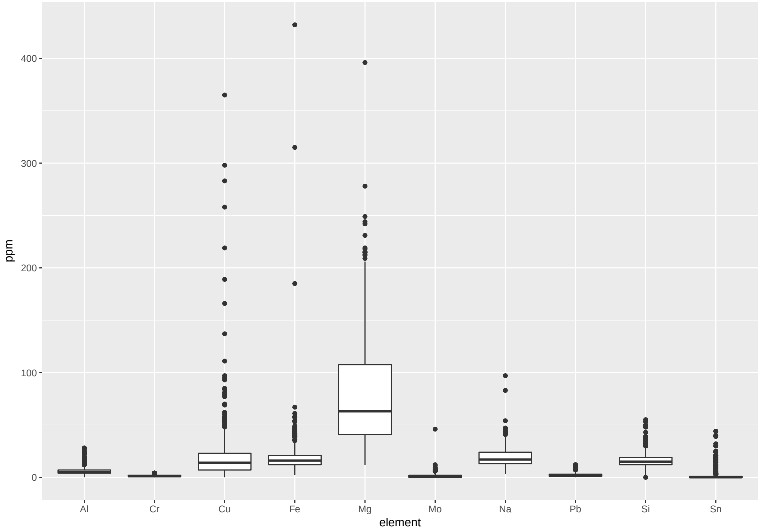
is called a *boxplot*. Often the extreme values are shown individually (see documentation for the (irrelevant) details.)

Best as *side-by-side* boxplots with more than one varaible on the same scale.

boxplot example - I

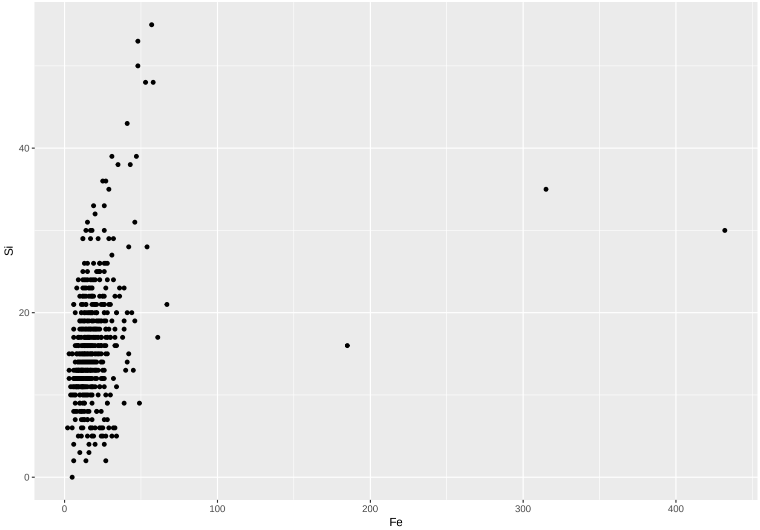


boxplot example - II

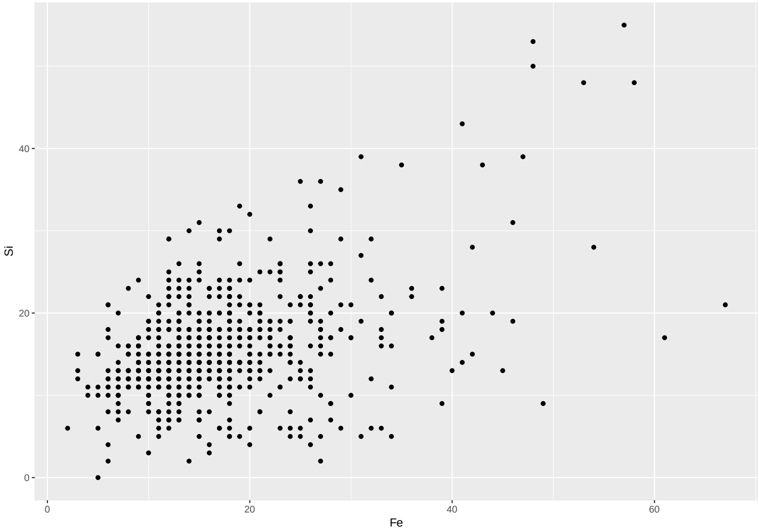


scatterplot

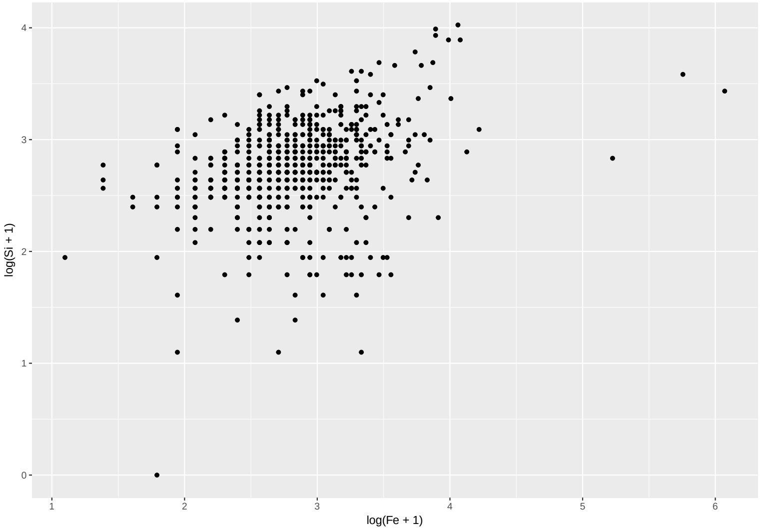
A graphic for two numerical variables, e.g. `Fe` and `Si`



Fe vs Si without the "outliers"

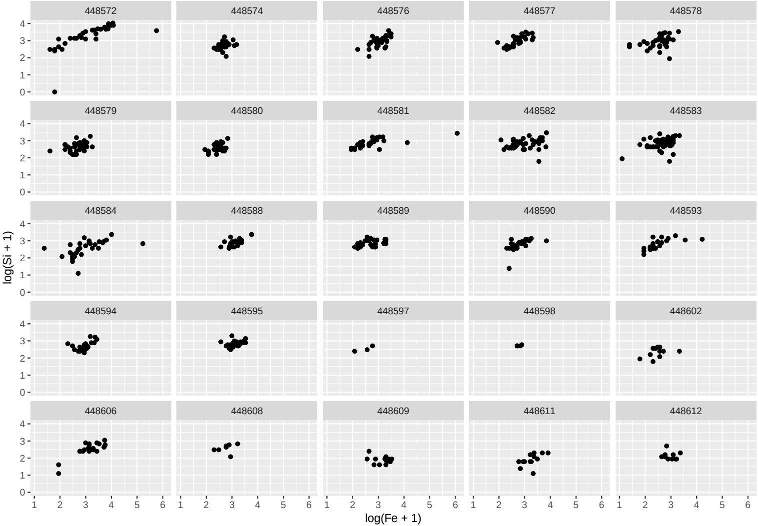


alternatively, on a log-log scale

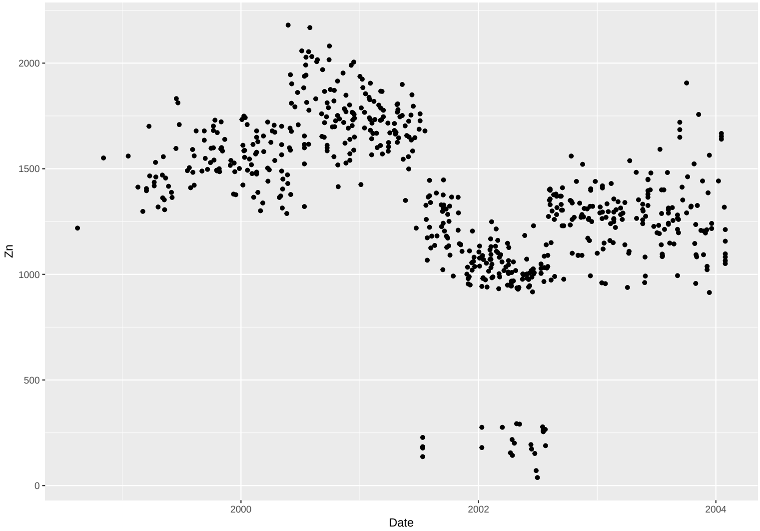


“small multiples” through faceting

A powerful exploratory tool is to make a grid of small plots on subsets of the data.



what about that “Date” variable. . . (!)



Fe versus Date, facet by Ident