Basic EDA workflow for environmental assessment

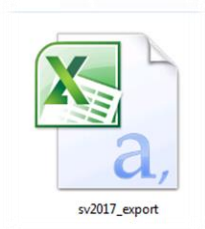
Clean data

Can be done in Excel or R

- single heading row
- no blank rows or columns
- censor unreliable values
- consistent missing value codes and factor levels
- ideally, save with metadata
- *etc.*

Import data

From csv, Excel, online, shapefile, etc.




```
data <-
      read.csv(...)
# and so on...
```

Check the data in R as well, using `str(...)` and/or `summary(...)`

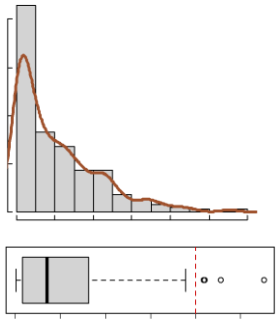
Check variables

Distributions

- normality
- modality
- transform if necessary
- re-check transformed variables (esp. for modality)

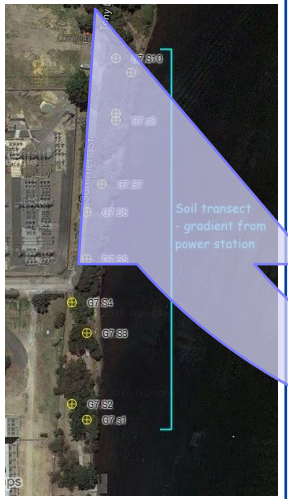
Extremes

- identify unusual values
- identify values over guidelines



New variables

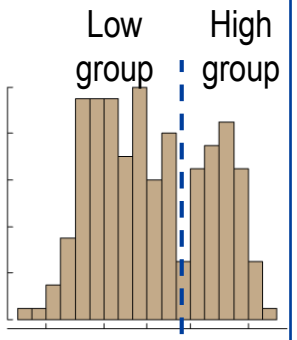
- (including transformed variables)
- distance (from upstream, from source, *etc.*) [use coordinates and Pythagoras]



New factors

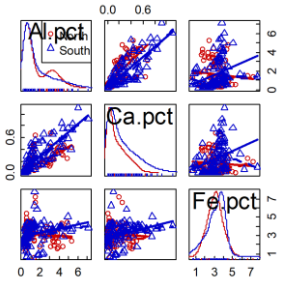
Based on:

- clearly bimodal or multimodal variables
- near/far, *etc.*



Relationships

- bivariate plots *e.g.* scatter plot matrices
- separate by groupings
- correlation matrices
- outliers in variables



Pearson's r

	Al.pct	Ca.pct	Fe.pct
Al.pct	1.00	0.77	0.16
Ca.pct	0.77	1.00	0.18
Fe.pct	0.16	0.18	1.00

P

	Al.pct	Ca.pct	Fe.pct
Al.pct		0.0000	0.0232
Ca.pct	0.0000		0.0095
Fe.pct	0.0232	0.0095	

Means comparisons

Between groups representing:

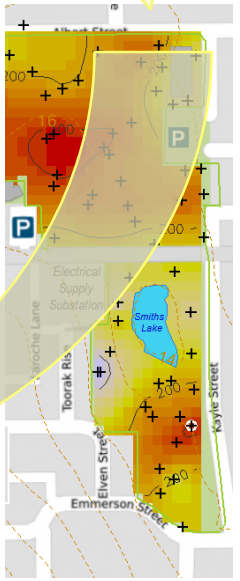
- sampling design (*e.g.* strata)
- any new factors

```
t.test(...)
aov(...)
oneway.test(...)

wilcox.test(...)
kruskal.test(...)

library(effsize)
  cohen.d(...)
```

Spatial patterns

Are these:

- related to sampling design?
- related to conceptual site model?
- related to any new factors?



EXPLORATORY DATA ANALYSIS