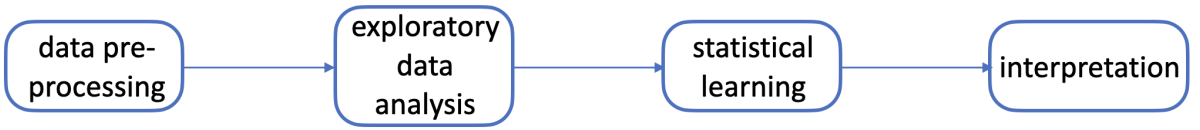
Exploratory Data Analysis

36-600

Fall 2021

The Canonical Analysis Workflow

[From Week 1]



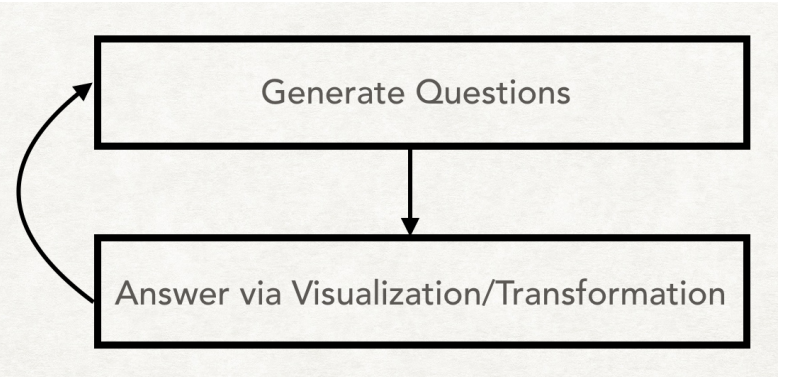
- *Data Pre-Processing*: the act of extracting analyzable data (e.g., a structured data table) from unstructured sources (e.g., images, audio files, text, etc.), as well as the act of editing the data table to mitigate missing data.
- *Exploratory Data Analysis*: the act of visualizing the observed data--via, e.g., histograms, scatter plots, box plots, etc., etc.--so as to build intuition about them. No statistical modeling is involved. EDA is not a substitute for statistical modeling, due to its implicit dimensional reduction!
- *Statistical Learning*: the attempt to find meaningful structures in the data or to uncover relationships between elements of the dataset. More on this below.
- *Interpretation*: what did you discover through your analysis?

While our primary focus in this course is **statistical learning**, today we will discuss EDA.

What is Exploratory Data Analysis?

The book R for Data Science claims that exploratory data analysis, or EDA, is a "state of mind." More usefully, it states that "[y]our goal during EDA is to develop an understanding of your data."

The EDA "cycle" looks something like the following:



The basic questions that one can ask are the following:

- What type of variation do the variables exhibit?
- What type of covariation do pairs of variables exhibit?

The Limits of Exploratory Data Analysis

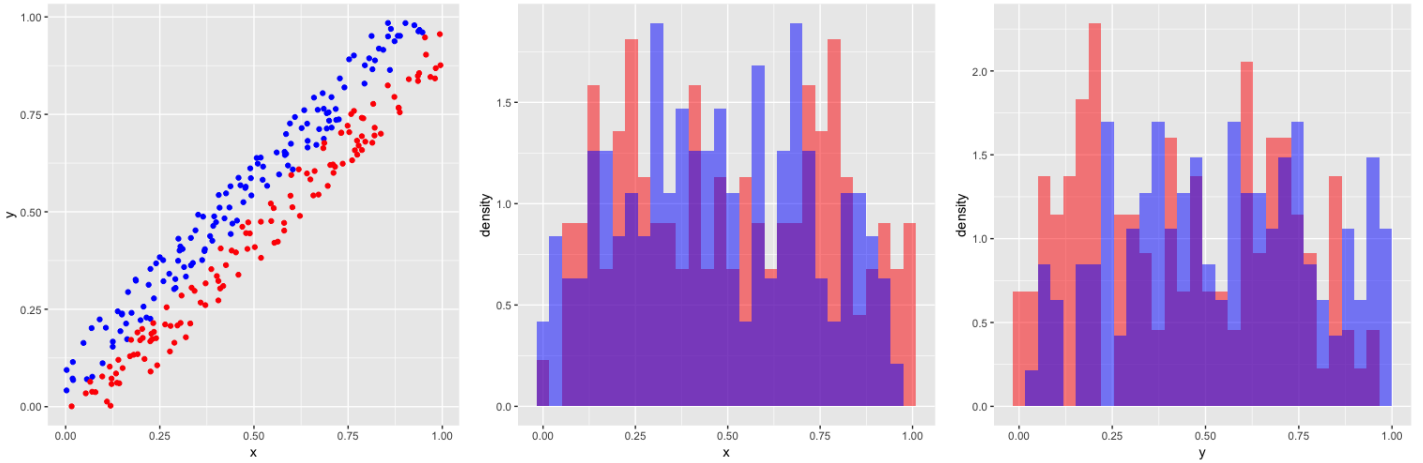
The basic questions on the previous slide motivate the statement of an important point.

If the number of predictors variables is larger than 3, one cannot visualize the native space; one can only visualize *projections* of that space. If those projections yield useful information, great! But if they do not: one should not give up, *because information may have been lost in projection*.

In other words: **EDA itself is not a replacement for statistical learning**.

It is a means by which to build intuition prior to "turning the learning crank."

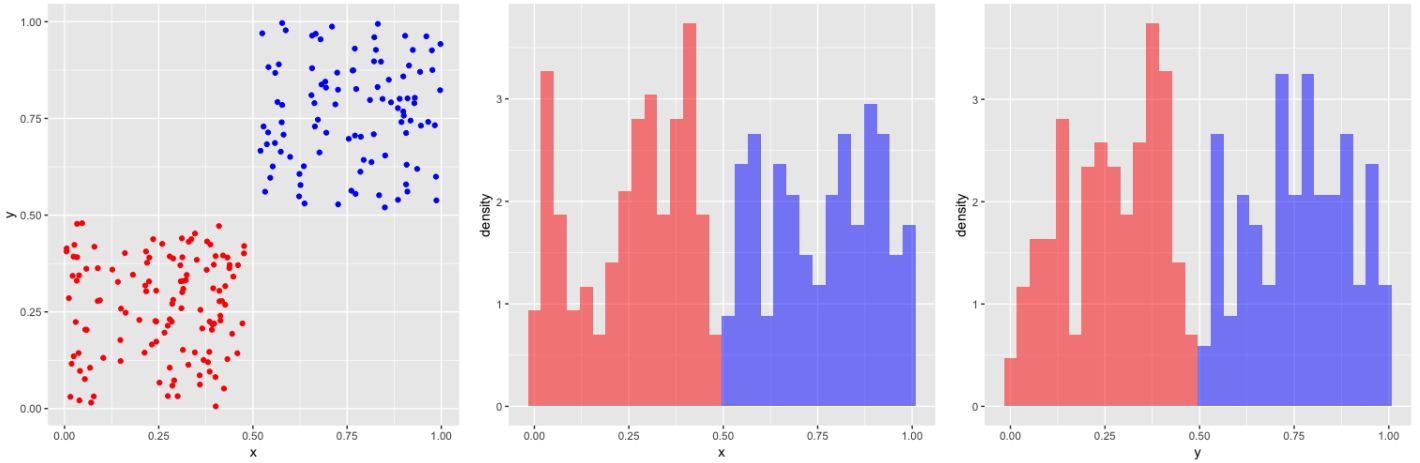
The Limits of Exploratory Data Analysis



A "linelander" (a being who can only visualize in one dimension) might look at the overlaid histograms for data projected to the x-axis and then to the y-axis and conclude that statistical learning will be hopeless. However, the act of projection to a lower dimension obscures the true structure of the data in the native two-dimensional space (where learning itself will occur).

- EDA: visualization of *projected* data.
- Statistical learning: modeling of data in their *native space*.

The Limits of Exploratory Data Analysis



Here, the linelander sees definite structure in the data, even in projection. The linelander thus knows that a classification model learned in the native space will do a good job in predicting classes. How good? You still have to learn the models and generate the metrics to know. But obviously better than if you were to (a) randomly assign classes to data (the trained monkey scenario), or (b) assign one class to *all* the data (the lazy analyst scenario).

If by eye you see clear associations between, e.g., at least some predictor variables and the response variable, then you **know** that statistical learning will uncover something meaningful.

EDA: A Starting Point

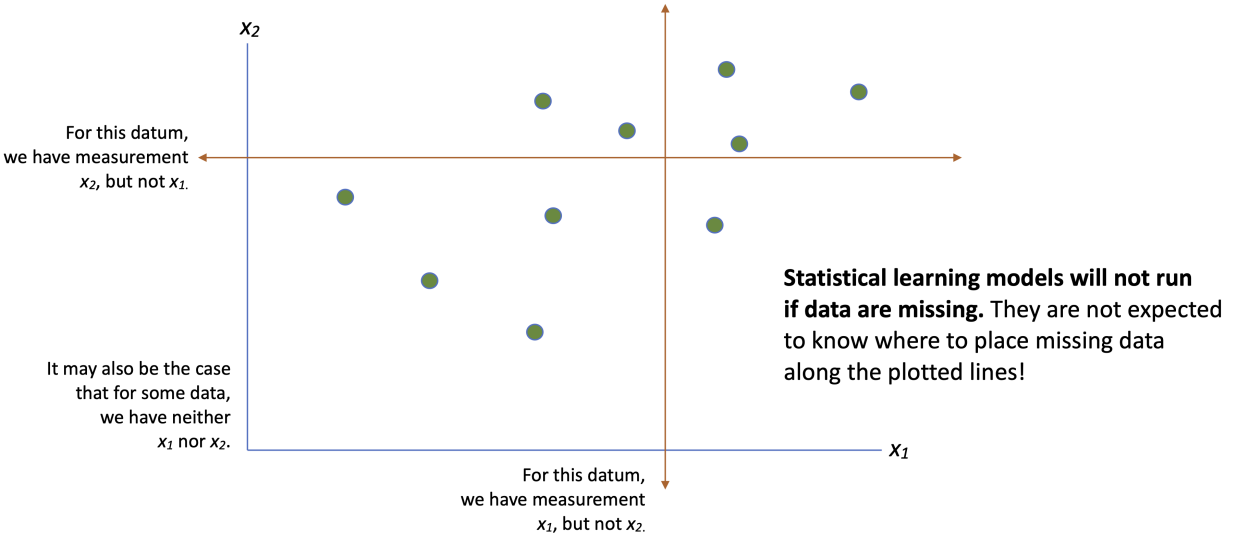
```
df <- read.csv("http://www.stat.cmu.edu/~pfreeman/GalaxyMass.csv",stringsAsFactors=TRUE)
```

A summary is often a good place to start, as it will identify amounts of missing data (assuming they have been identified and marked as NA) and give you a sense of how the data are distributed:

```
summary(df)
```

```
##     field          Gini              M20                C               A                  size              n
##  COSMOS:905   Min.   :0.03829   Min.   :-2.2154   Min.   :1.093   Min.   :-4.72887   Min.   :0.1173   Min.   :0.200
##  EGS   :750   1st Qu.:0.41397   1st Qu.:-1.7181   1st Qu.:2.665   1st Qu.: 0.08763   1st Qu.:0.5100   1st Qu.:0.68
##  GOODSN:464   Median :0.45774   Median :-1.5802   Median :3.008   Median : 0.12729   Median :0.6838   Median :1.28
##  GOODSS:588   Mean   :0.44359   Mean   :-1.5135   Mean   :3.023   Mean   : 0.14008   Mean   :0.7005   Mean   :1.93
##  UDS   :749   3rd Qu.:0.48947   3rd Qu.:-1.3431   3rd Qu.:3.402   3rd Qu.: 0.17834   3rd Qu.:0.8637   3rd Qu.:2.47
##                Max.   :0.85754   Max.   :-0.4342   Max.   :4.922   Max.   : 0.79319   Max.   :2.8572   Max.   :8.00
##        q                z.mode           mass
##  Min.   :0.003971   Min.   :0.010   Min.   : 8.285
##  1st Qu.:0.398650   1st Qu.:1.610   1st Qu.: 9.777
##  Median :0.556750   Median :1.760   Median :10.061
##  Mean   :0.559674   Mean   :1.767   Mean   :10.137
##  3rd Qu.:0.724725   3rd Qu.:1.920   3rd Qu.:10.457
##  Max.   :0.998900   Max.   :6.720   Max.   :11.697
```

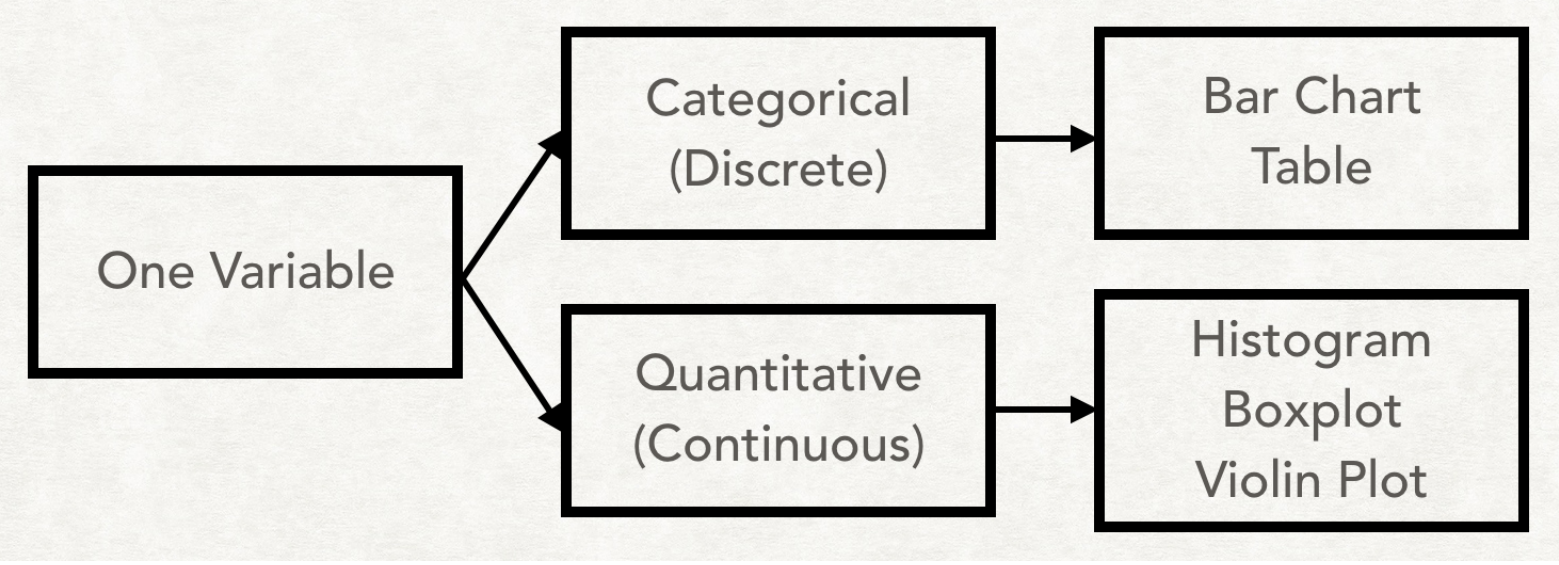
EDA: What Happens of Data Are Missing?



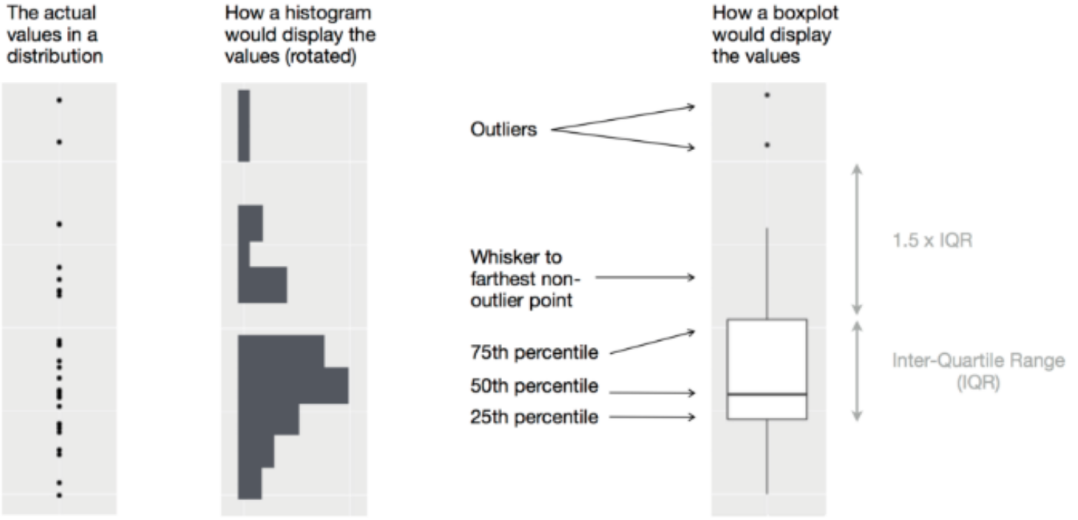
You will have to remove data from your sample, or do data imputation (kinda dangerous and beyond the scope of this class).

There are no simple heuristics in how to go about removing data. For instance, if you have six columns of data and a sample size of 100, and 30 of the data in column 6 are missing, do you remove the 30 rows with missing data, or the sixth column? (What if the sixth column has statistically informative data?) The usual rule applies: if you have missing data in your own life, talk to me or the TA and we can try to consult.

EDA: Single Variable



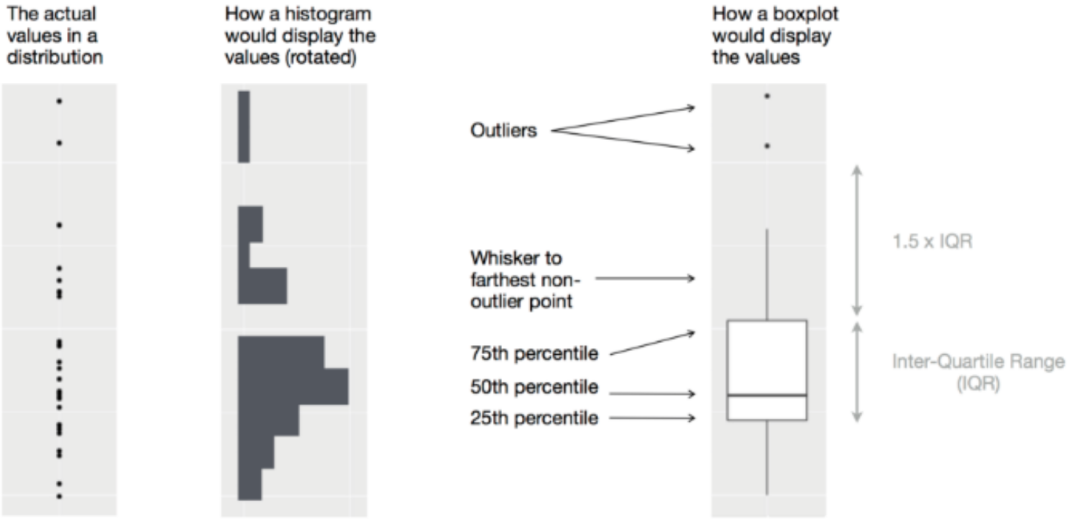
EDA: Histograms



Some of the features of histograms to keep in mind:

- They exhibit a nonparametric estimate of the shapes of underlying data distributions.
- They are sensitive to bin width: too few bins, and noise and localized features are smoothed out; too many bins, and noise obscures the underlying distribution.
- Unlike boxplots, histograms do not provide statistics and do not identify outliers by rule.

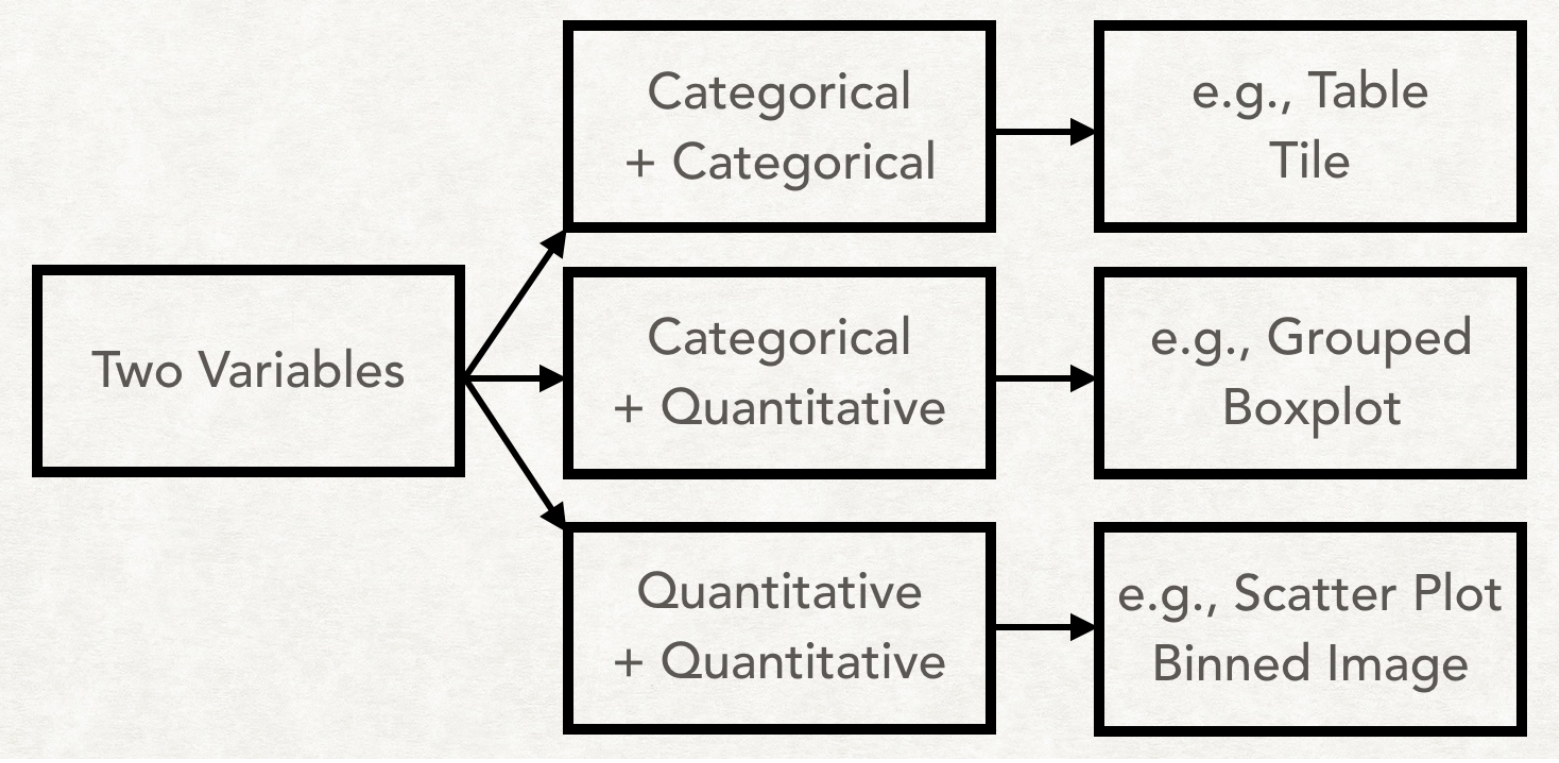
EDA: Boxplots



Some of the features of boxplots to keep in mind:

- They do not exhibit estimates of the shapes of underlying data distributions.
- They provide data statistics, such as the sample median.
- They identify outliers by rule. *However*: the default rule is "insufficient" for large datasets, i.e., if your sample size (the number of rows in your data frame) is thousands or tens of thousands or larger, then you will always see supposed outliers that really aren't. So beware. (For instance, 99.73% of data are sampled with 3 standard deviations of the mean for a normal distribution. If you have 100 data, you will maybe see at most one 3 σ "outlier". If you have 10,000 data, you will see approximately 27. The bigger the dataset, the more "outliers" there will appear to be.)

EDA: Two Variables



A feature of scatter plots to keep in mind: they show the locations of samples from a bivariate distribution, but unlike histograms they do not estimate that distribution itself. (For that you might go beyond typical workaday EDA and utilize kernel density estimation, which we'll cover elsewhere.) Scatter plots *do*, however, indicate the level of covariance between two variables.

EDA: Scatter Plots

IMPORTANT Here are some tips of the trade to keep in mind when constructing scatter plots:

- If your sample size is $\gtrsim 10^4$, randomly sample ~ 1000 points for plotting. Otherwise your computer may become very unhappy.
- To improve interpretability, mitigate the effect of point overlap by both reducing the size of points and altering point transparency.
- Don't let outliers "drive the bus." Change the plot limits manually if necessary to zoom in on the bulk of the points. This also improves interpretability. (It's easier to interpret what you can see more clearly.)

EDA: Questions to Ponder

- Could the pattern you observe arise by coincidence?
- Is the observed pattern consistent, or does it change? (For instance, does variable y vary linearly with x, but only for some values of x and not for others?)
- If you observe an association between variables, can you think of confounding variables that might cause the association? (Because *association is not causation*. Say it again. Say it many times. Good.)
- How strong is the association between variables?
- Does it appear that the assumptions that underlie some methods of statistical learning (e.g., constant variance in typical linear regression modeling) hold for your data?
- Etc.

But remember: even if you see no apparent associations, they may still exist in the data's native space! Again: *EDA is not a replacement for statistical learning*.