Exploratory Data Analysis

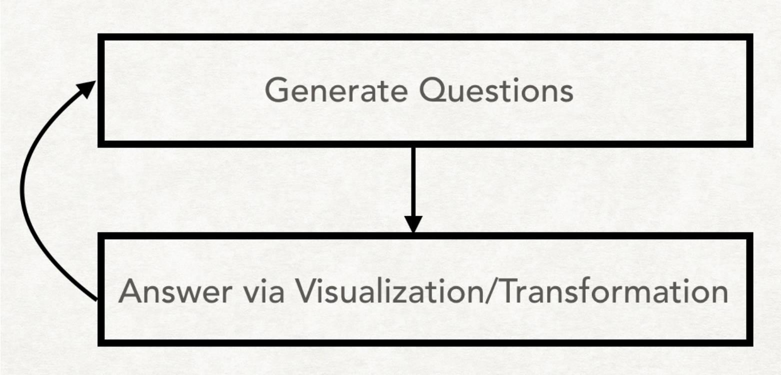
36-290 – Statistical Research Methodology

Week 3 Tuesday – Fall 2021

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

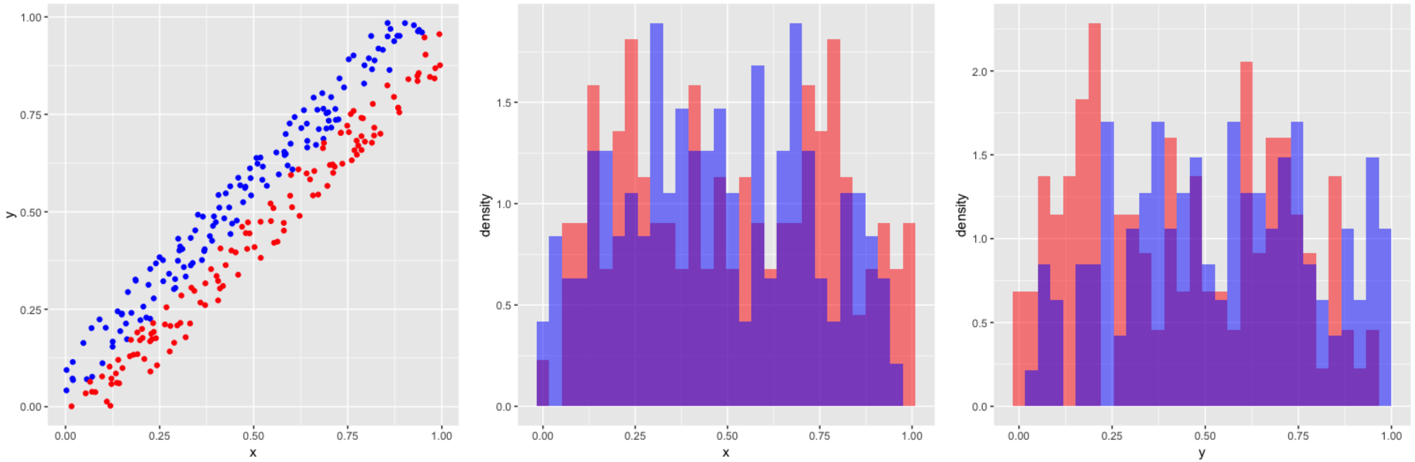
The basic questions on the previous slide motivate the statement of an important point.

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*.

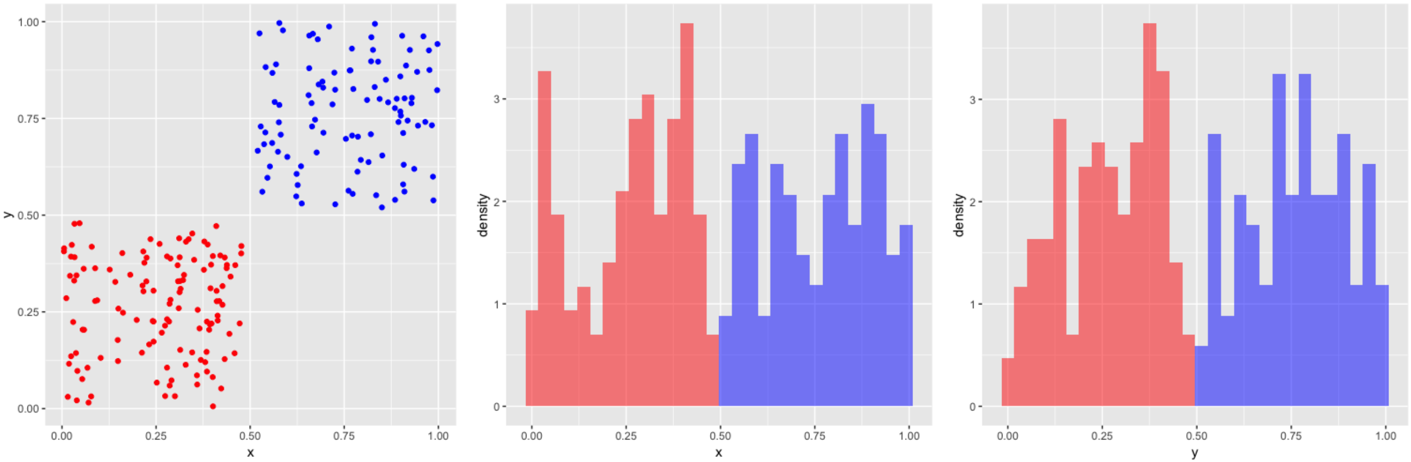
In other words: **EDA itself is not a replacement for statistical learning**.

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

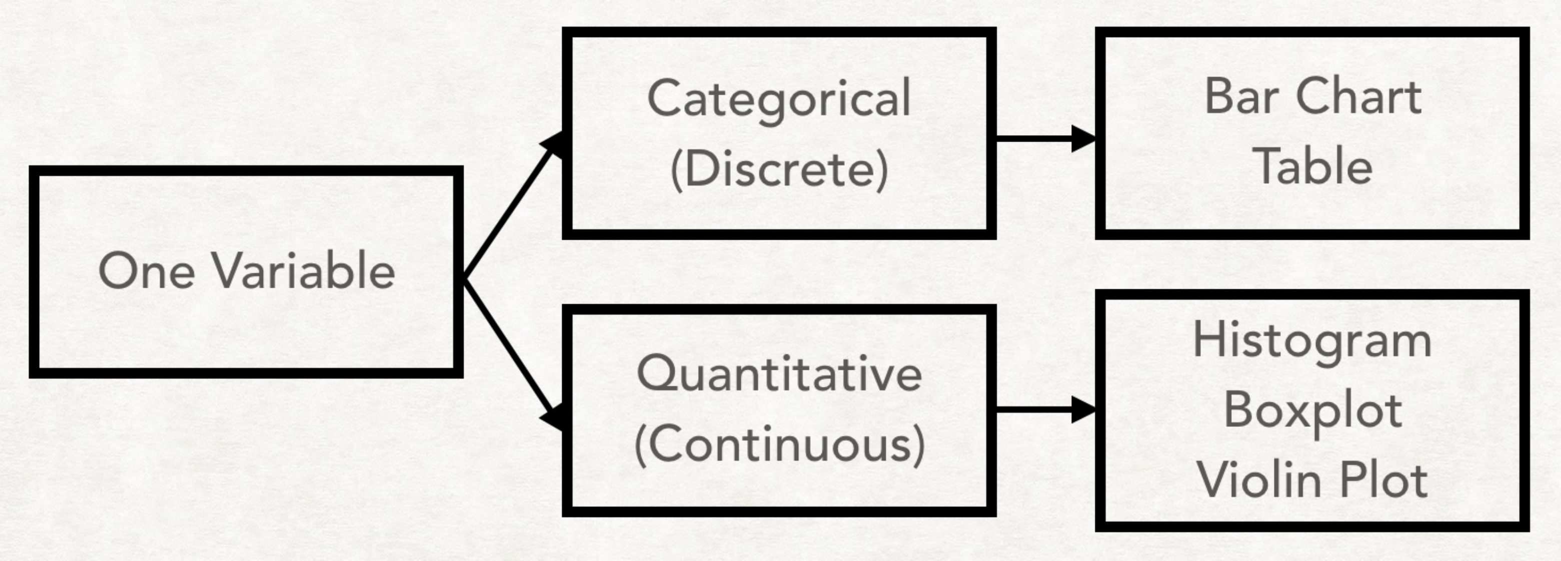
The Limits of Exploratory Data Analysis



The Limits of Exploratory Data Analysis



EDA: Single Variable



One Variable

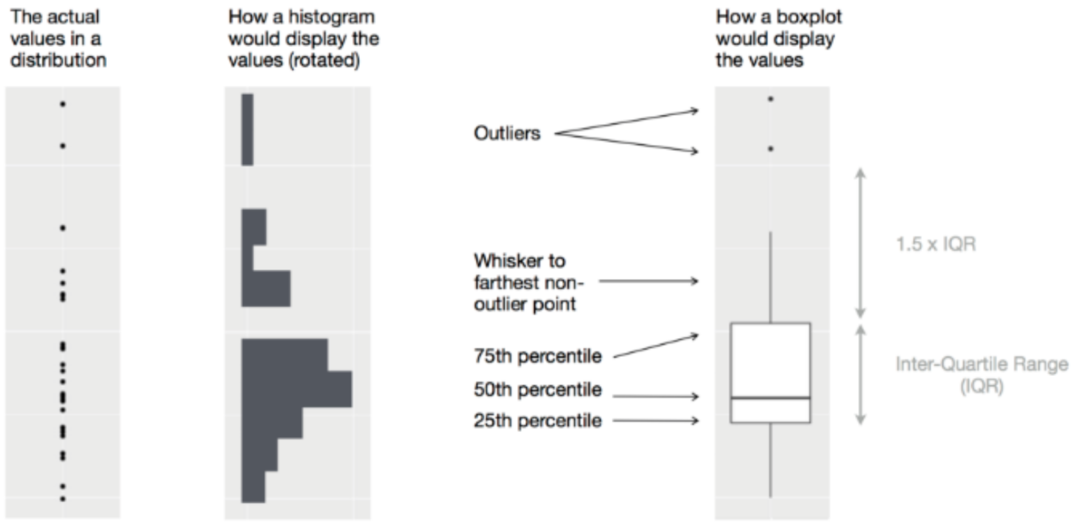
Categorical (Discrete)

Quantitative (Continuous)

Bar Chart
Table

Histogram
Boxplot
Violin Plot

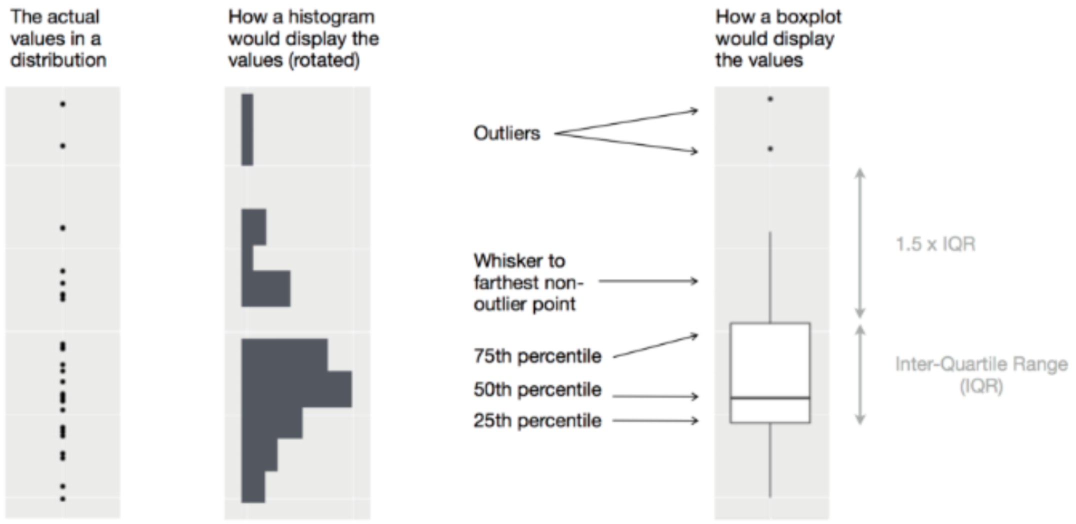
EDA: Histograms


The actual values in a distribution
How a histogram would display the values (rotated)
How a boxplot would display the values
Outliers
Whisker to farthest non-outlier point
75th percentile
50th percentile
25th percentile
1.5 x IQR
Inter-Quartile Range (IQR)

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.

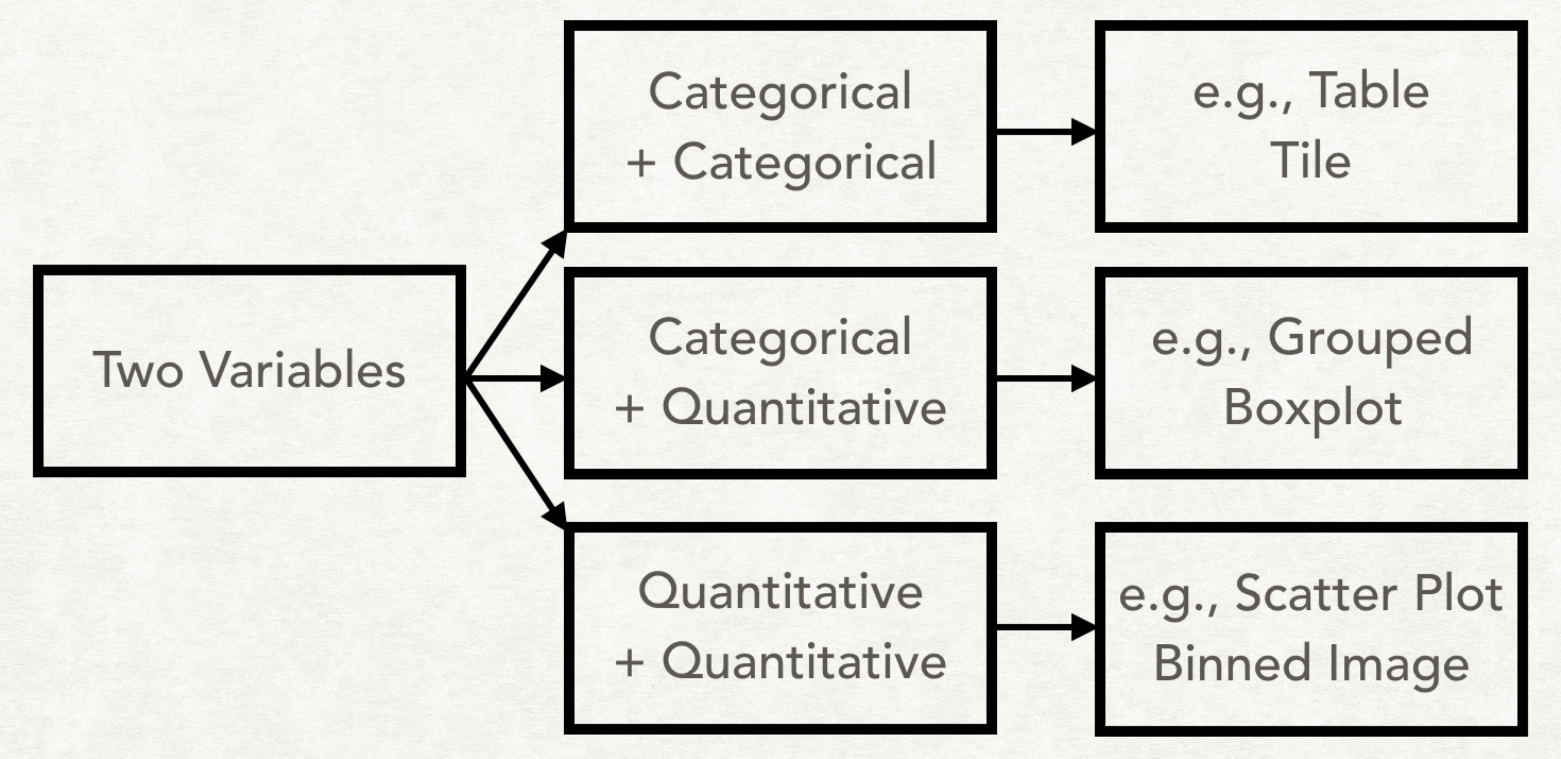
EDA: Boxplots


The actual values in a distribution
How a histogram would display the values (rotated)
How a boxplot would display the values
Outliers
Whisker to farthest non-outlier point
75th percentile
50th percentile
25th percentile
1.5 x IQR
Inter-Quartile Range (IQR)

Some of the features of boxplots to keep in mind:

- They do not exhibit estimates of the shapes of underlying data distributions. (Use violin plots.)
- They provide data statistics, such as the sample median.
- They identify outliers by rule.

EDA: Two Variables


Two Variables
Categorical + Categorical
e.g., Table Tile
Categorical + Quantitative
e.g., Grouped Boxplot
Quantitative + Quantitative
e.g., Scatter Plot Binned Image

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: Covariance and Correlation

Covariance:

- A metric that quantifies the *linear* association between a pair of variables.

Correlation:

- A "normalized" form of covariance with values between -1 (the value of variable y decreases absolutely as variable x increases in value) and +1 (y strictly increases with x).

Note that "uncorrelated" and "independent" are *not* synonymous. Uncorrelated means that there is no linear association between the variables; independent means that there is no association between the variables, period.

EDA: Scatter Plots

Some tips of the trade to keep in mind when constructing scatter plots:

- If your sample size is $\gtrsim 10^4$, randomly sample $\sim$ 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! *EDA is not a replacement for statistical learning*. Say it again. Say it many times. Good.