Exploratory Data Analysis (EDA)

Overview

- **Exploratory Data Analysis (EDA)**
- **Data Exploration: Descriptive Statistics**
- **Data Exploration: Visualizations**

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA)

Why?

- EDA encompasses the "*explore* data" part of the data science process
- EDA is crucial but often overlooked:
  - If your data is bad, your results will be bad
  - Conversely, understanding your data well can help you create smart, appropriate models

Exploratory Data Analysis (EDA)

What?

1. Store data in data structure(s) that will be convenient for exploring/processing (Memory is fast. Storage is slow)
2. Clean/format the data so that:
    - Each row represents a single object/observation/entry
    - Each column represents an attribute/property/feature of that entry
    - Values are numeric whenever possible
    - Columns contain atomic properties that cannot be further decomposed

Exploratory Data Analysis (EDA)

What? (continued)

3. Explore **global** properties: use histograms, scatter plots, and aggregation functions to summarize the data
4. Explore **group** properties: group like-items together to compare subsets of the data (are the comparison results reasonable/expected?)

This process transforms your data into a format which is easier to work with, gives you a basic overview of the data's properties, and likely generates several questions for you to follow-up in subsequent analysis.

Exploratory Data Analysis (EDA)

Purposes of EDA:

- Maximize insight into a dataset
- Uncover underlying structure
- Detect outliers
- Test underlying assumptions
- Develop parsimonious models

Data Exploration: Descriptive Statistics

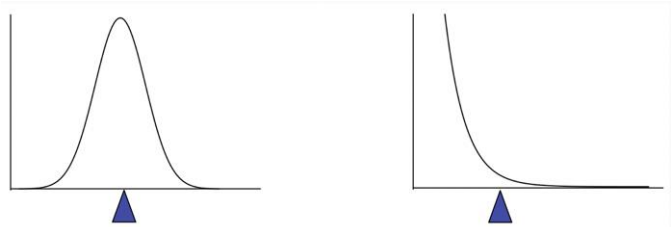
Basics of Sampling

- Population versus sample:
- A **population** is the entire set of objects or events under study. Population can be hypothetical “all students” or all students in this class.
- A **sample** is a “representative” subset of the objects or events under study. Needed because it’s impossible or intractable to obtain or compute with population data.
- Biases in samples:
- **Selection bias**: some subjects or records are more likely to be selected
- Volunteer/**nonresponse bias**: subjects or records who are not easily available are not represented

Sample mean

- The **mean** of a set of *n* observations of a variable is denoted $\bar{x}$ and is defined as:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



- The mean describes what a “typical” sample value looks like, or where is the “center” of the distribution of the data.
- Key theme: there is always uncertainty involved when calculating a sample mean to estimate a population mean.

Sample median

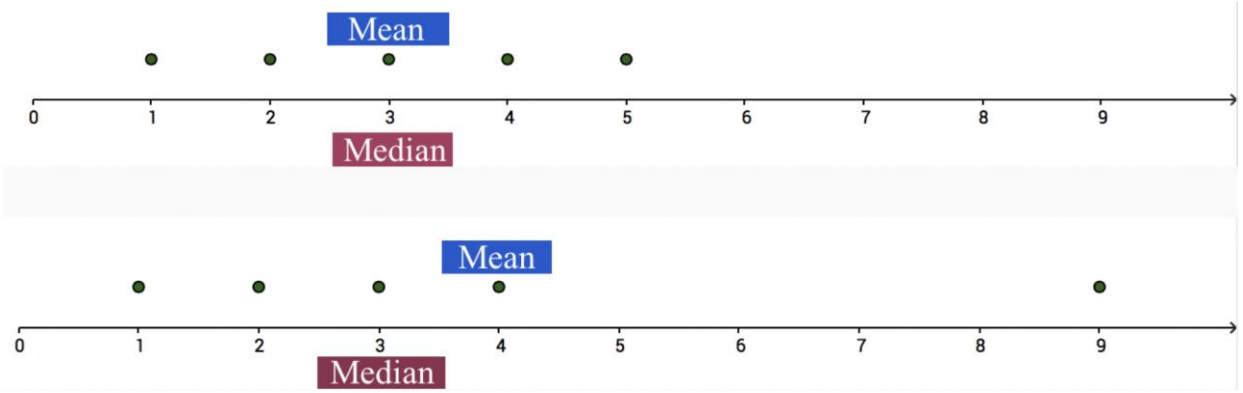
- The **median** of a set of n number of observations in a sample, ordered by value, of a variable is is defined by

$$\text{Median} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{x_{n/2} + x_{(n+1)/2}}{2} & \text{if } n \text{ is even} \end{cases}$$

- Example (already in order):
  Ages: 17, 19, 21, 22, 23, 23, 23, 38
- Median = (22+23)/2 = 22.5
- The median also describes what a typical observation looks like, or where is the center of the distribution of the sample of observations.

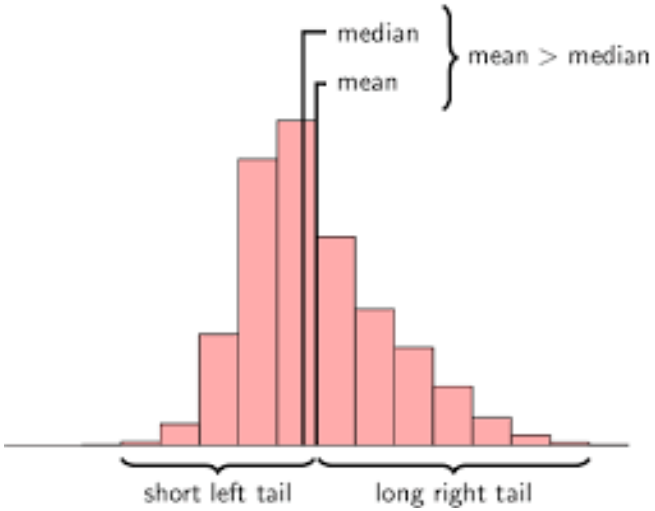
Mean vs. Median

- The mean is sensitive to extreme values (**outliers**)



Mean, median, and skewness

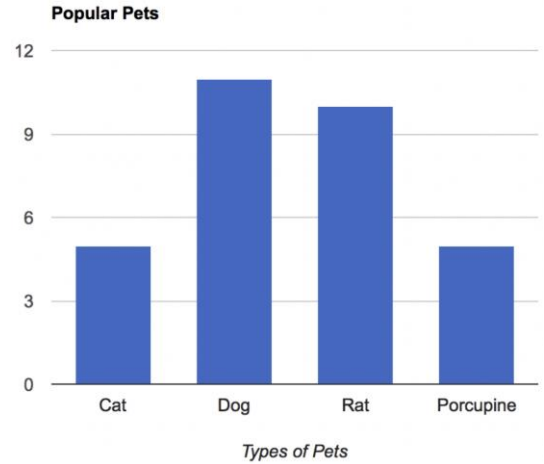
- The mean is sensitive to outliers:



- The above distribution is called **right-skewed** since the mean is greater than the median. Note: **skewness** often "follows the longer tail".

Regarding Categorical Variables...

- For categorical variables, neither mean or median make sense. Why?



-
- The mode might be a better way to find the most “representative” value.

Measures of Spread: Range

- The spread of a sample of observations measures how well the mean or median describes the sample.
- One way to measure spread of a sample of observations is via the **range**.
-
- Range = Maximum Value - Minimum Value

Measures of Spread: Variance

- The (sample) **variance**, denoted s^2, measures how much on average the sample values deviate from the mean:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} |x_i - \bar{x}|^2$$

- Note: the term $|x_i - \bar{x}|$ measures the amount by which each x_i deviates from the mean $\bar{x}$. Squaring these deviations means that s^2 is sensitive to extreme values (outliers).
- Note: s^2 doesn't have the same units as the x_i :(
- What does a variance of 1,008 mean? Or 0.0001?

Measures of Spread: Standard Deviation

- The (sample) **standard deviation**, denoted *s*, is the square root of
- the variance

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} |x_i - \bar{x}|^2}$$

- Note: *s* does have the same units as the x_i. Phew!

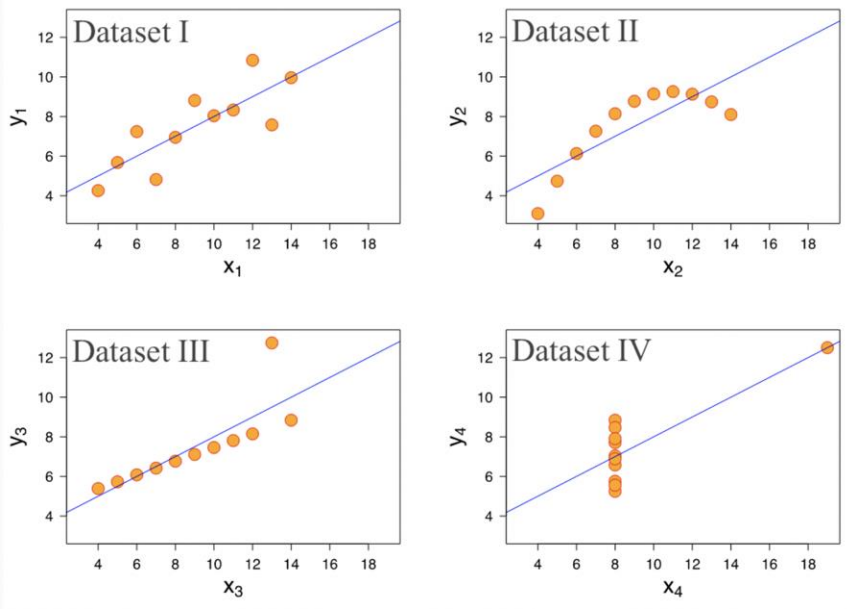
Data Exploration: Visualizations

Four Data Sets

- The following four data sets have identical simple summary statistics.

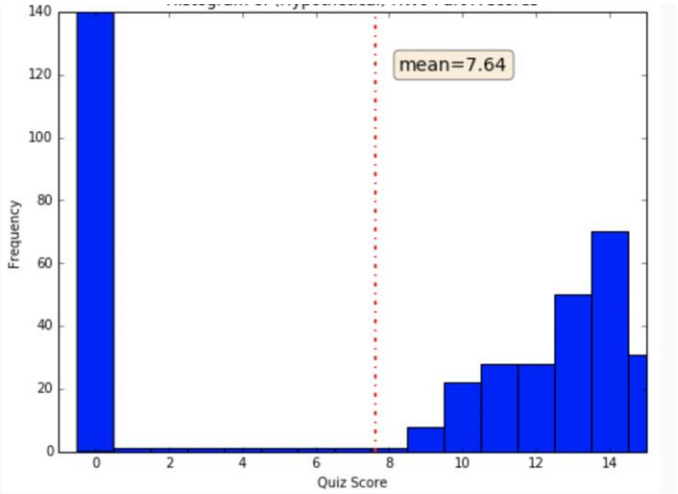
	Dataset I		Dataset II		Dataset III		Dataset IV	
	x	y	x	y	x	y	x	y
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Sum:	99.00	82.51	99.00	82.51	99.00	82.51	99.00	82.51
Avg:	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Std:	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03

- Summary statistics clearly don't tell the story of how they differ. But a picture can be worth a thousand words:



More Visualization Motivation

- If I tell you that the average score for Homework 0 was: 7.64/15 = 50.9% last year, what does that suggest?



- And what does the graph suggest?

More Visualization Motivation

- Visualizations help us to analyze and explore the data. They help to:

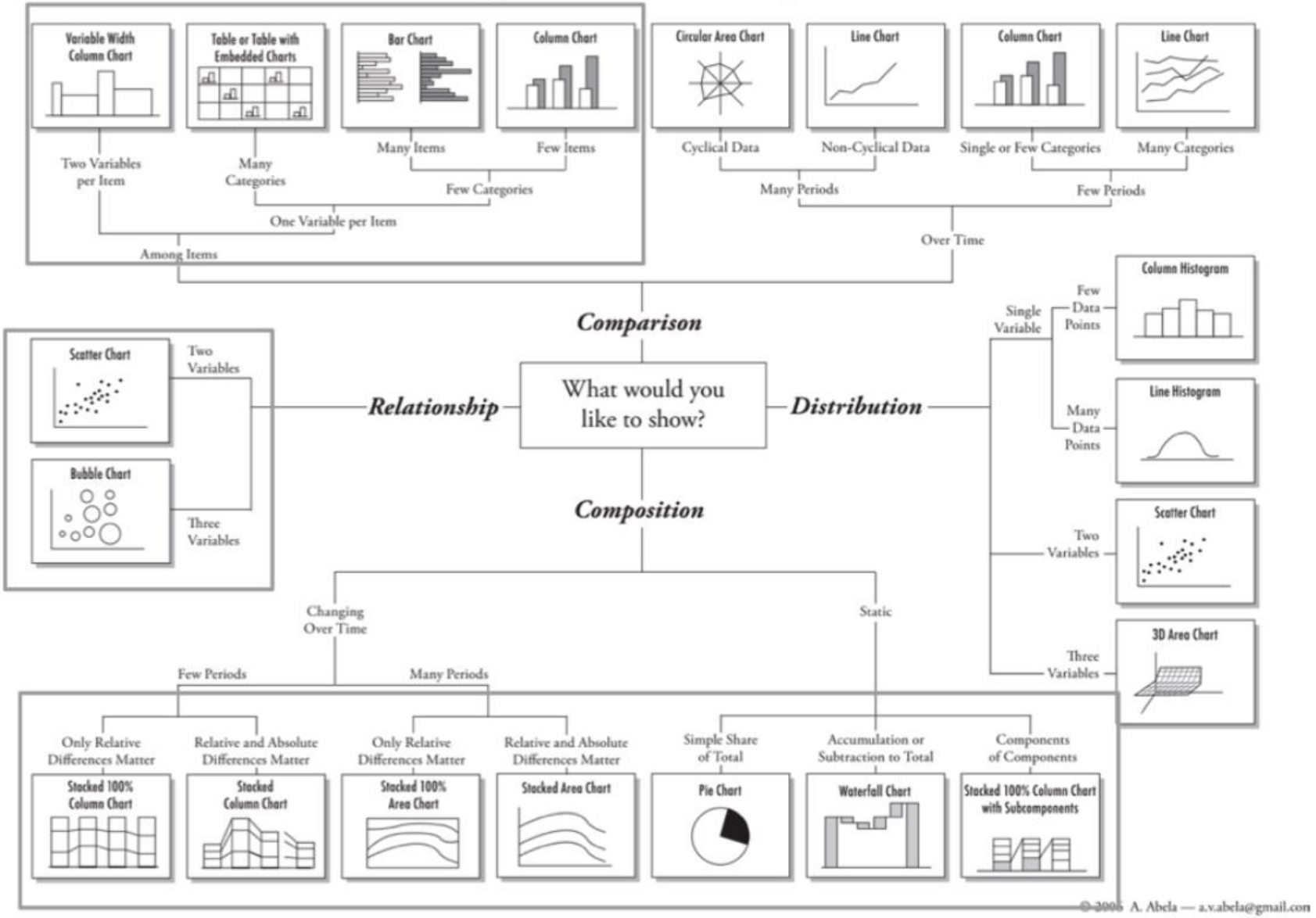
- Identify hidden patterns and trends
- Formulate/test hypotheses
- Communicate any modeling results
  - Present information and ideas succinctly
  - Provide evidence and support
  - Influence and persuade
- Determine the next step in analysis/modeling

Types of Visualizations

- What do you want your visualization to show about your data?

- **Distribution:** how a variable or variables in the dataset distribute over a range of possible values.
- **Relationship:** how the values of multiple variables in the dataset relate
- **Composition:** how the dataset breaks down into subgroups
- **Comparison:** how trends in multiple variable or datasets compare

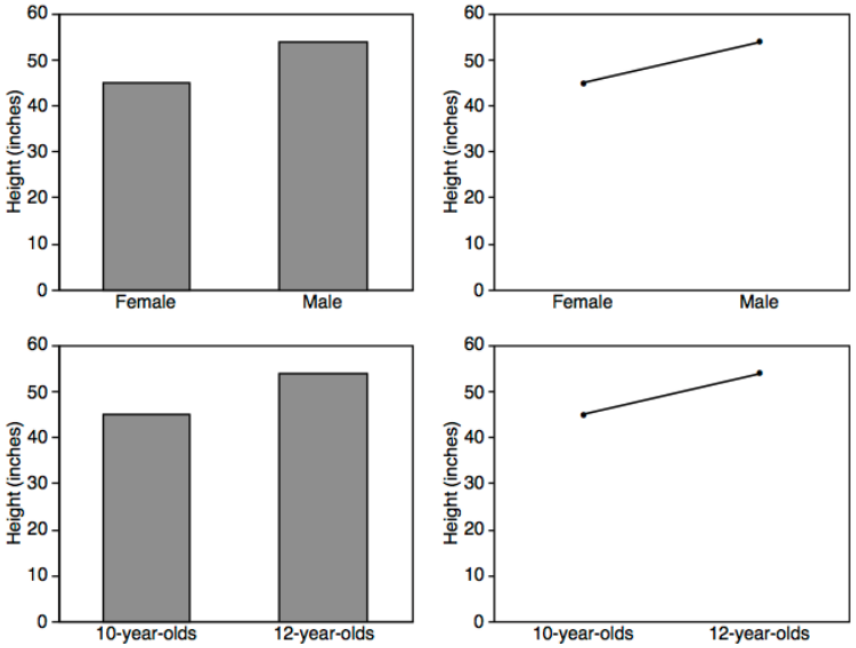
Chart Suggestions—A Thought-Starter



http://extremepresentation.typepad.com/blog/files/choosing_a_good_chart.pdf

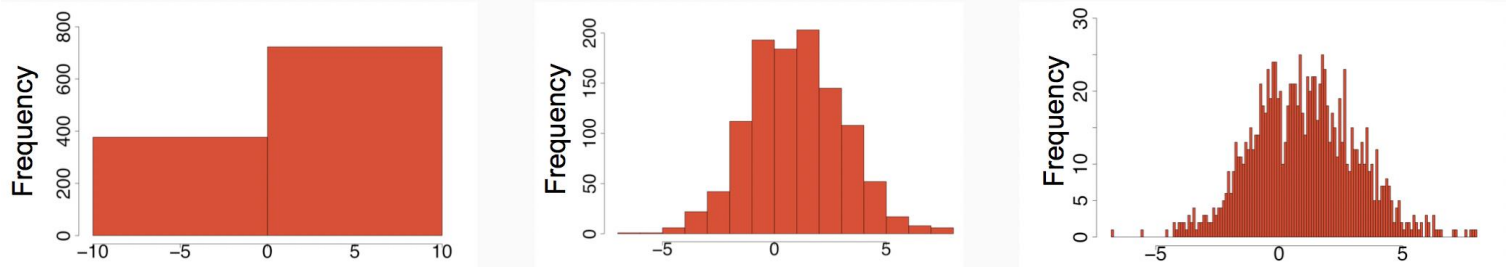
Displays: comparisons

Bars vs. Lines



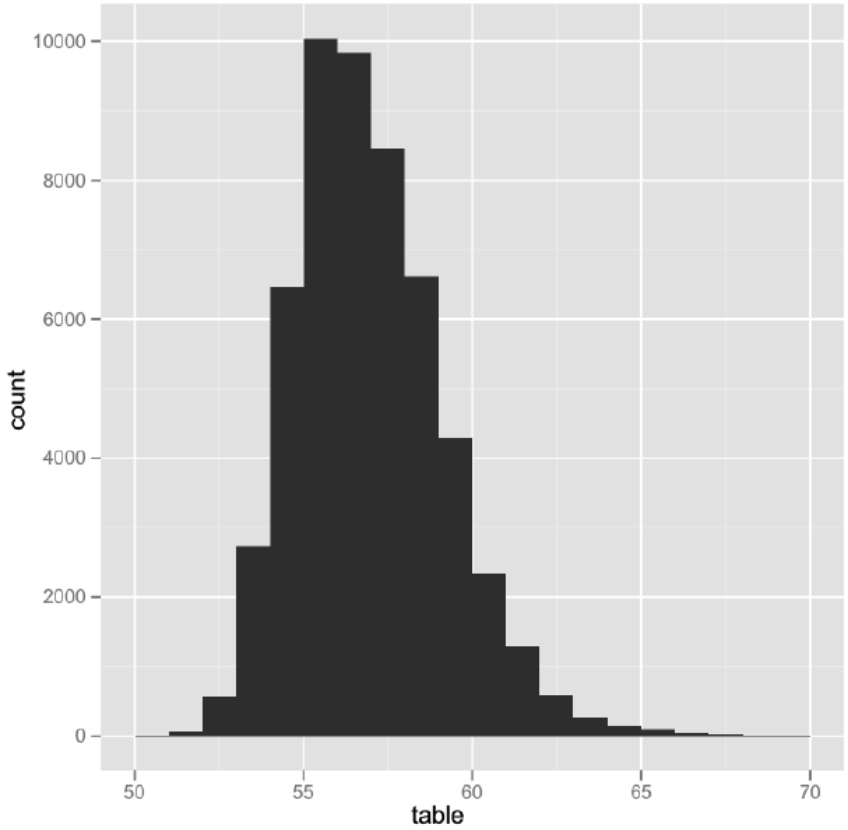
Histograms to visualize distribution

- A **histogram** is a way to visualize how 1-dimensional data is distributed across certain values.



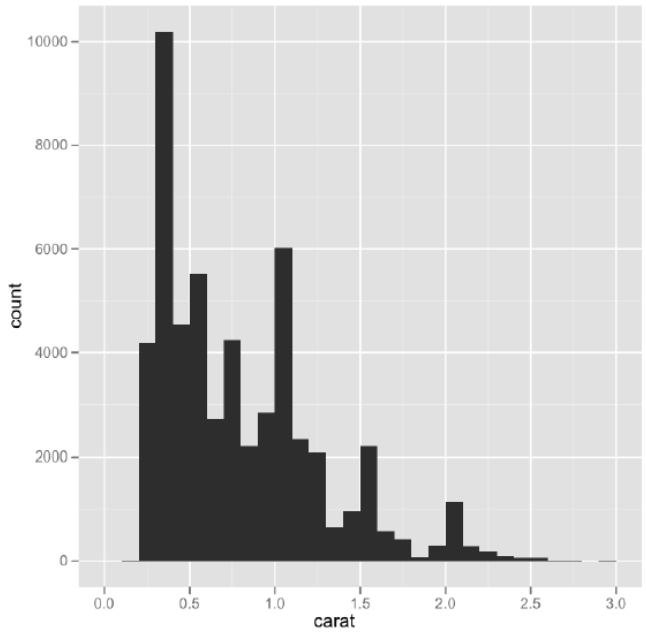
- Note: Trends in histograms are sensitive to number of bins.

Displays: distributions

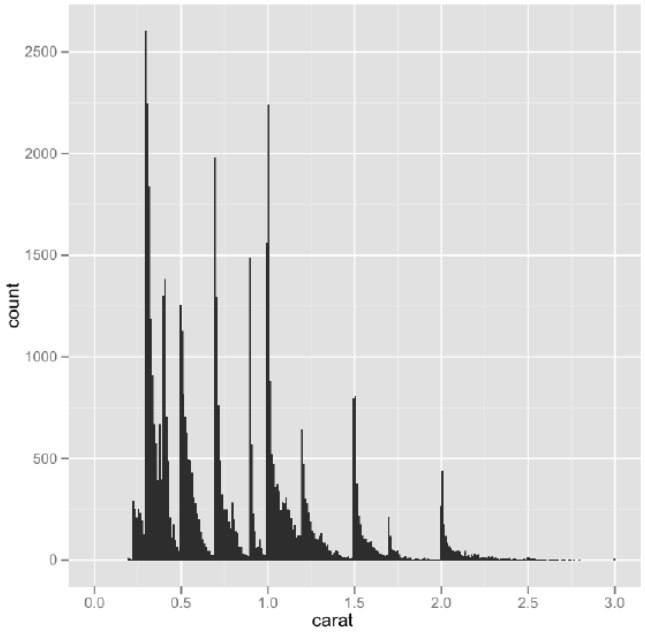


Displays: distributions

Bin Width

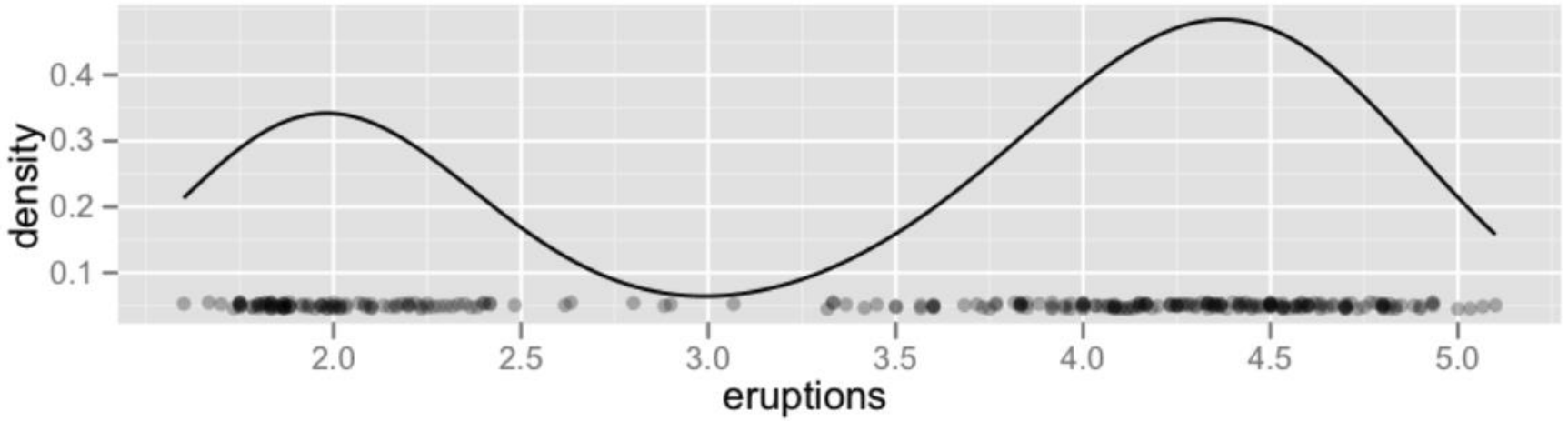


binwidth = 0.1



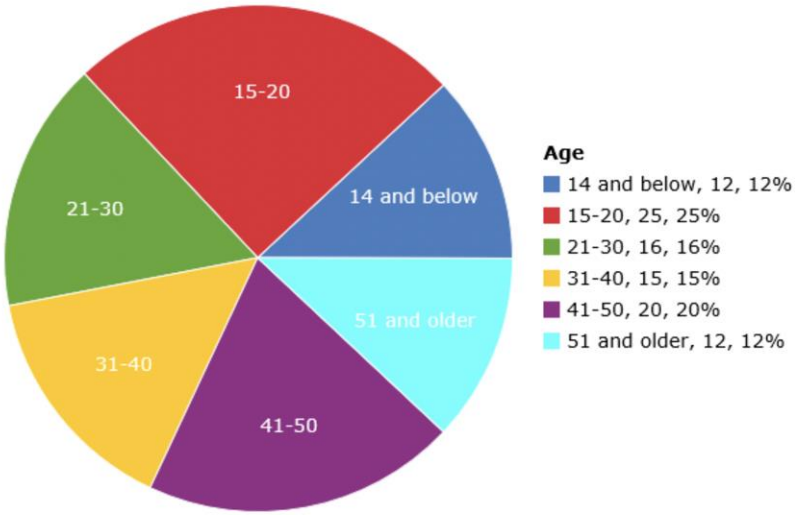
binwidth = 0.01

Displays: density plots



Pie chart for a categorical variable?

- A **pie chart** is a way to visualize the static composition (aka, distribution) of a variable (or single group).



- Pie charts are often frowned upon (and bar charts are used instead). Why?

Displays: hands-on exercise

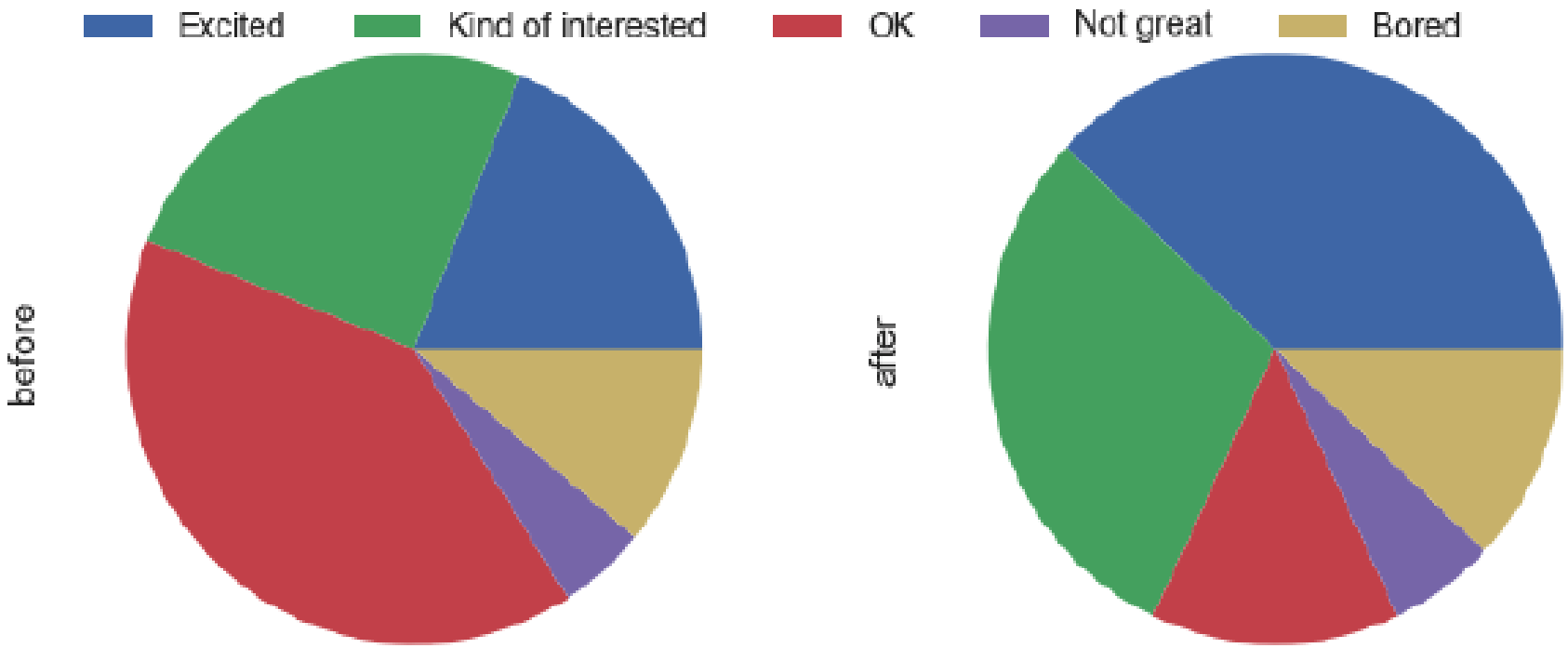
How do you feel about doing science?

Table

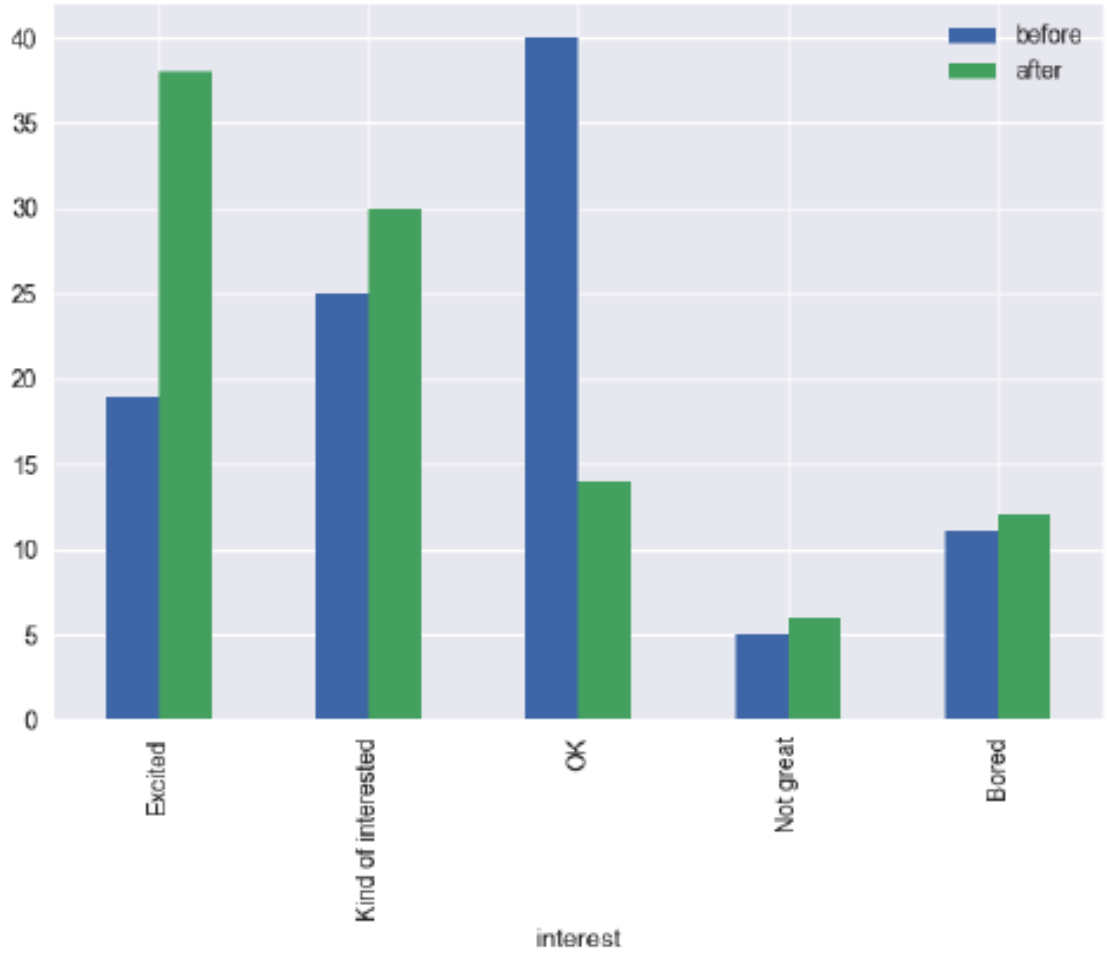
Interest	Before	After
Excited	19	38
Kind of interested	25	30
OK	40	14
Not great	5	6
Bored	11	12

Data courtesy of Cole Nussbaumer

Displays: exercise options

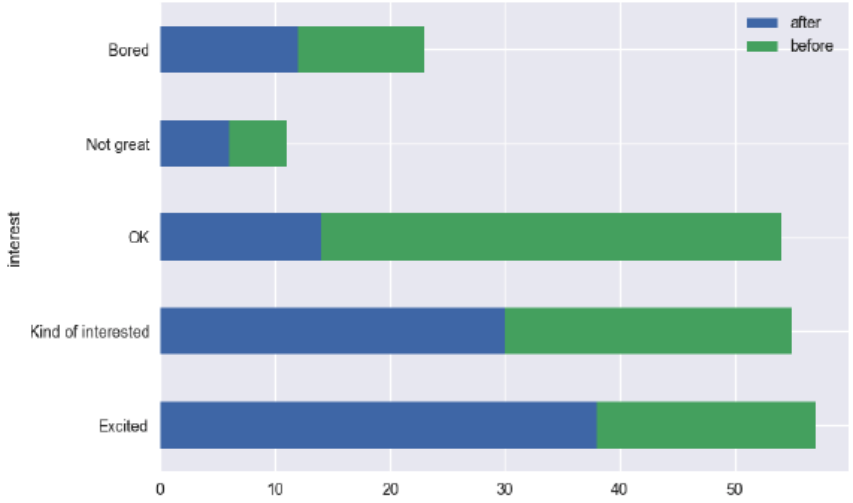


Displays: exercise options



Displays: exercise options

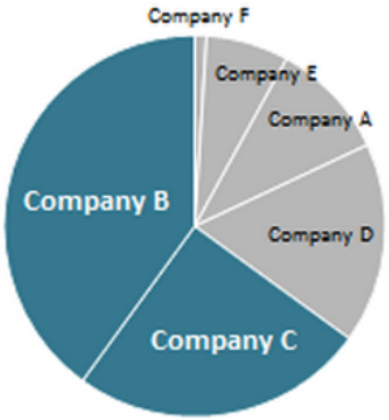
Stacked bar, not very useful

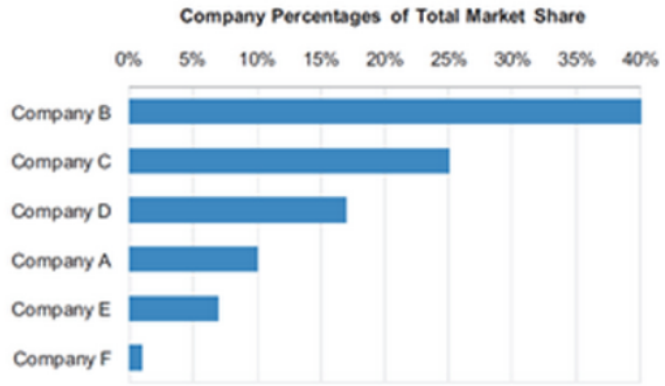


Displays: perceptual effectiveness

Pie vs. Bar Charts

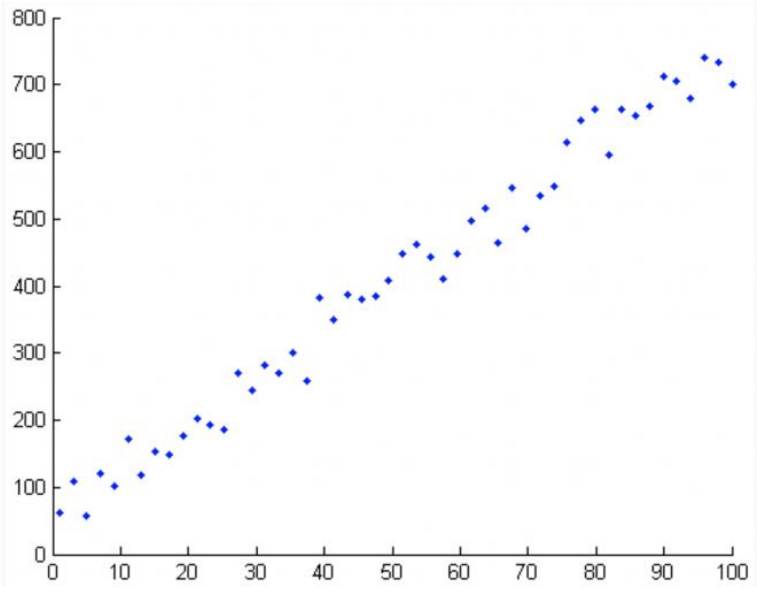
65% of the market is controlled by companies B and C





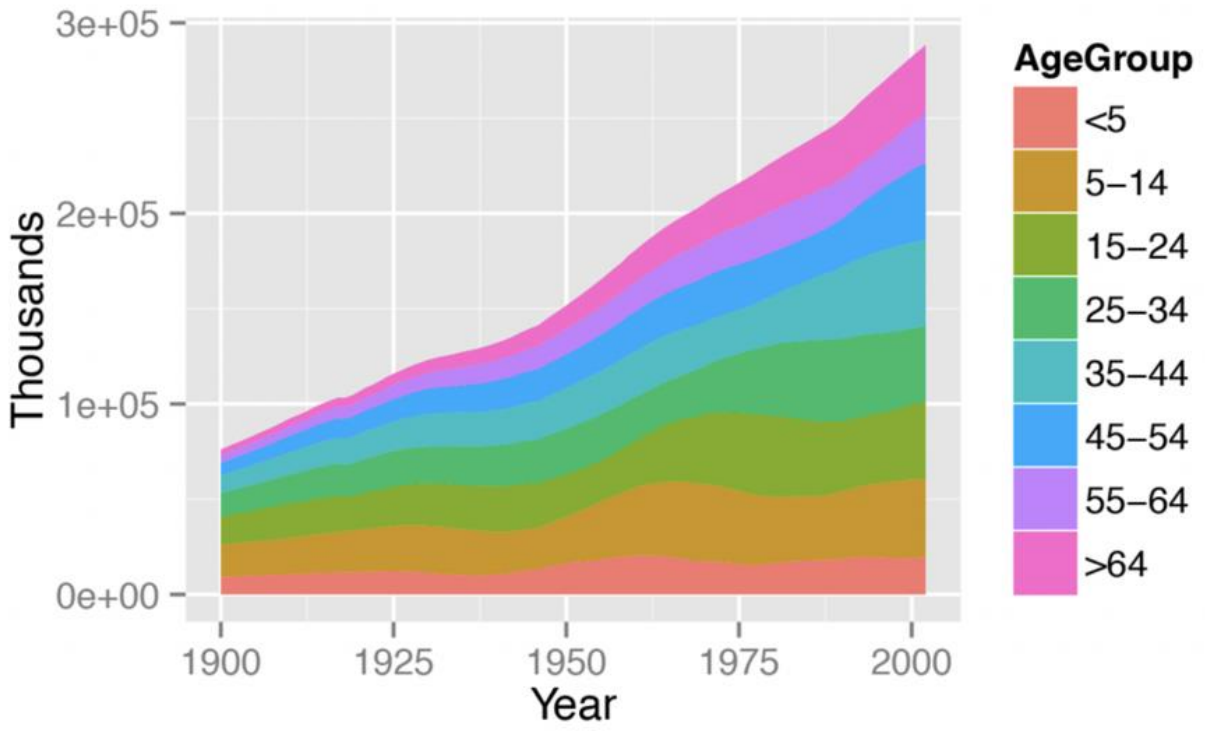
Scatter plots to visualize relationships

- A **scatter plot** is a way to visualize the relationship between two different attributes of multi-dimensional data.

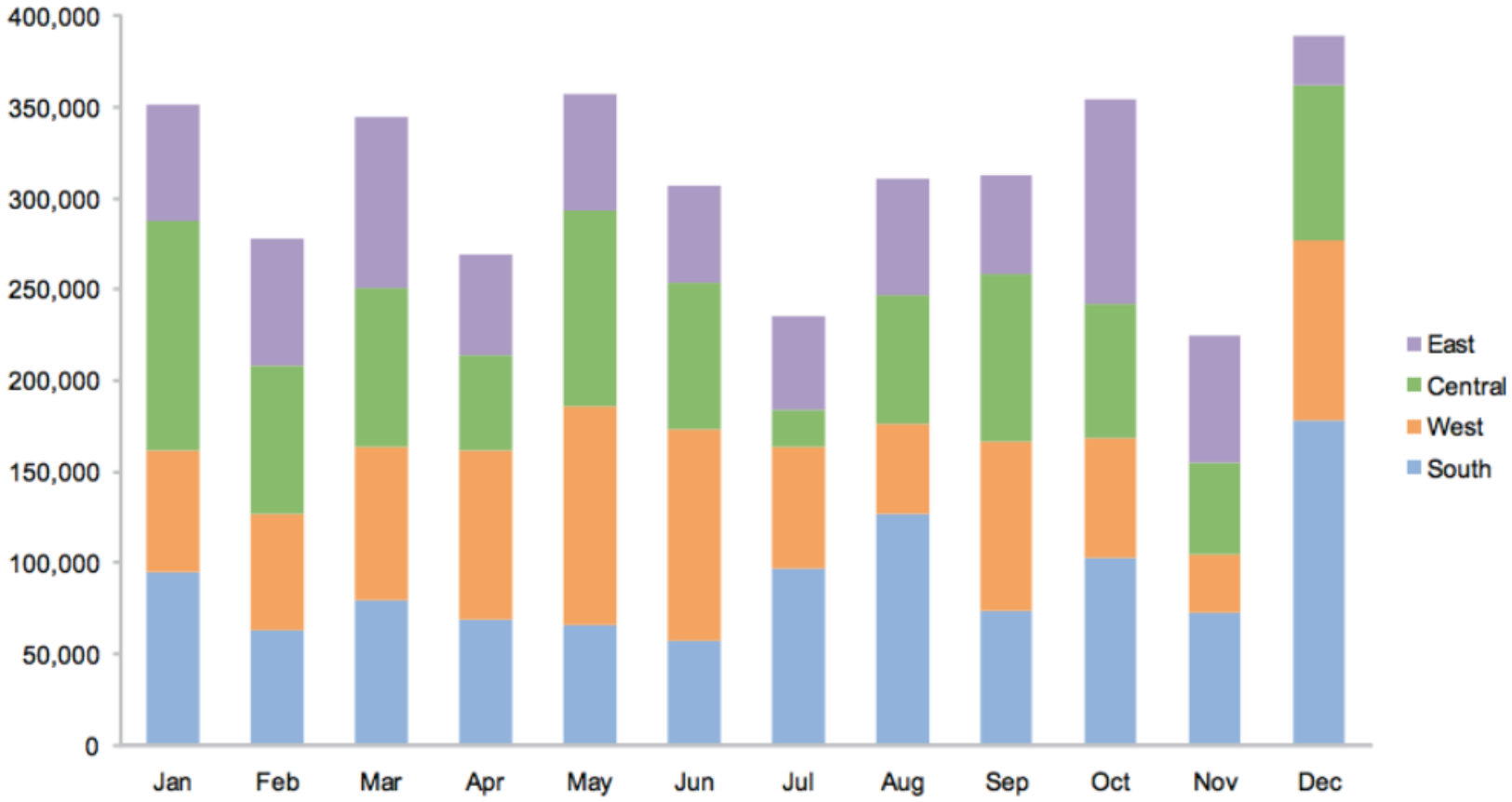


Stacked area graph to show trend over time

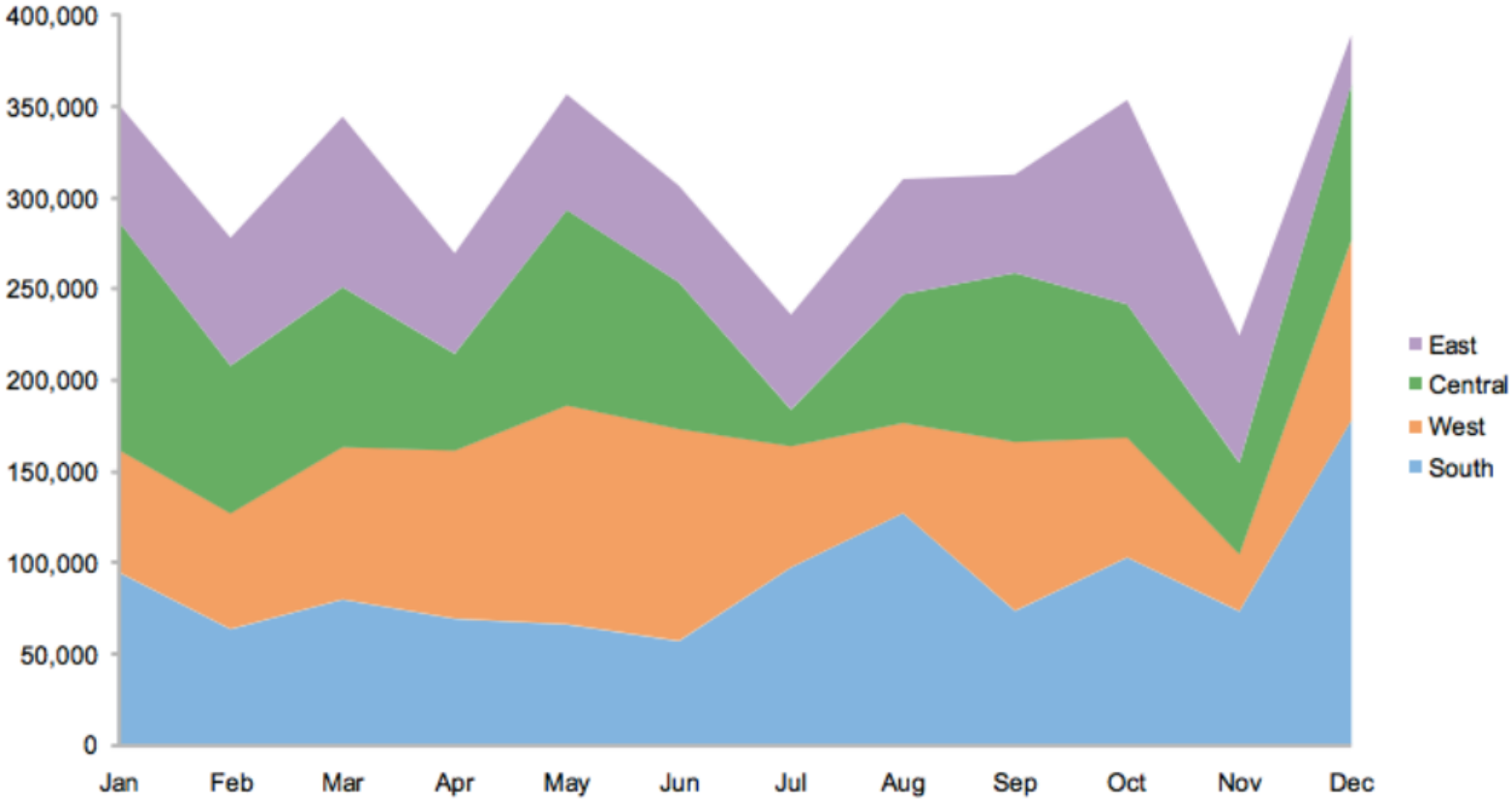
- A **stacked area graph** is a way to visualize the composition of a group as it changes over time (or some other quantitative variable). This shows the relationship of a categorical variable (AgeGroup) to a quantitative variable (year).



Displays: proportions

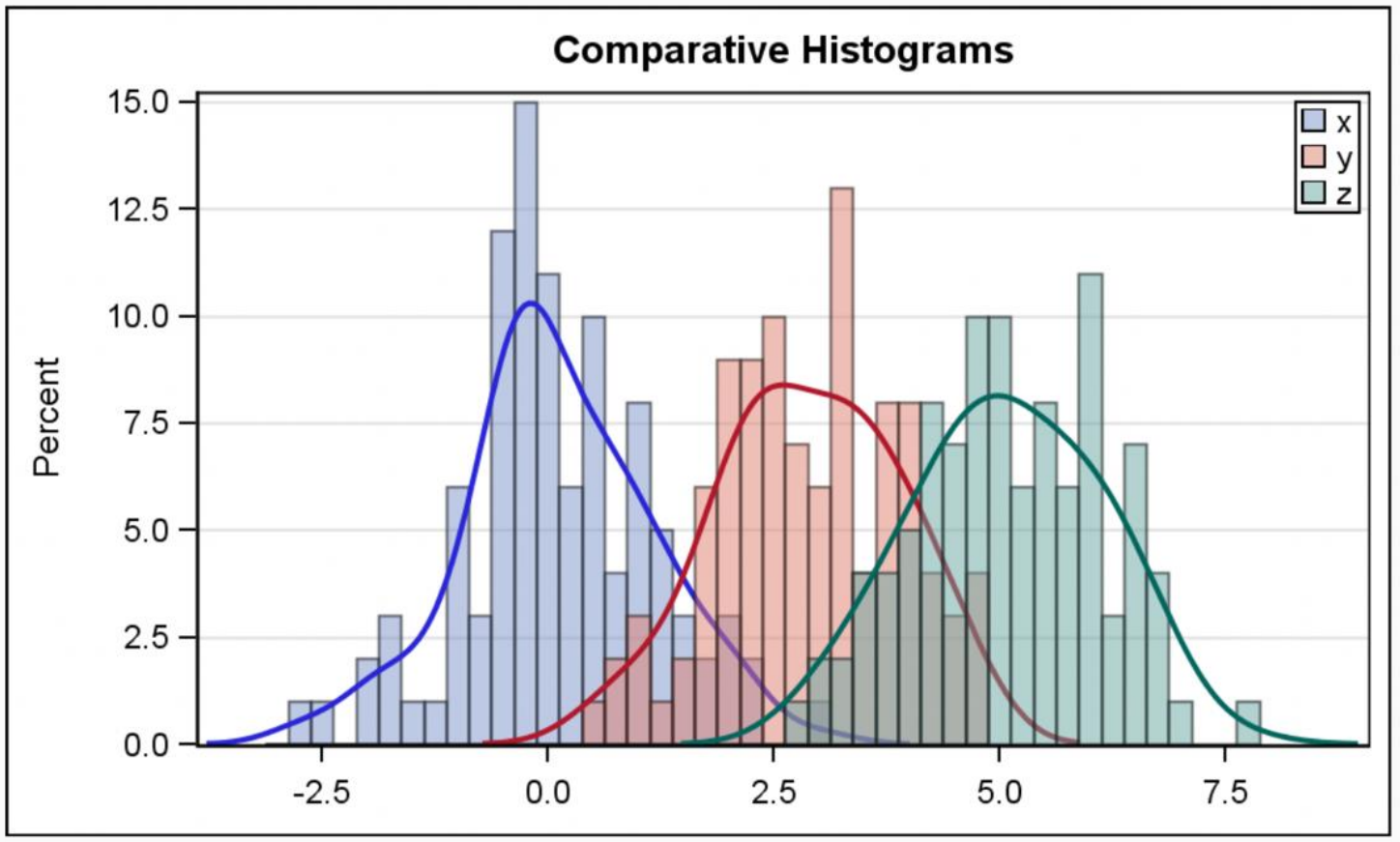


Displays: proportions



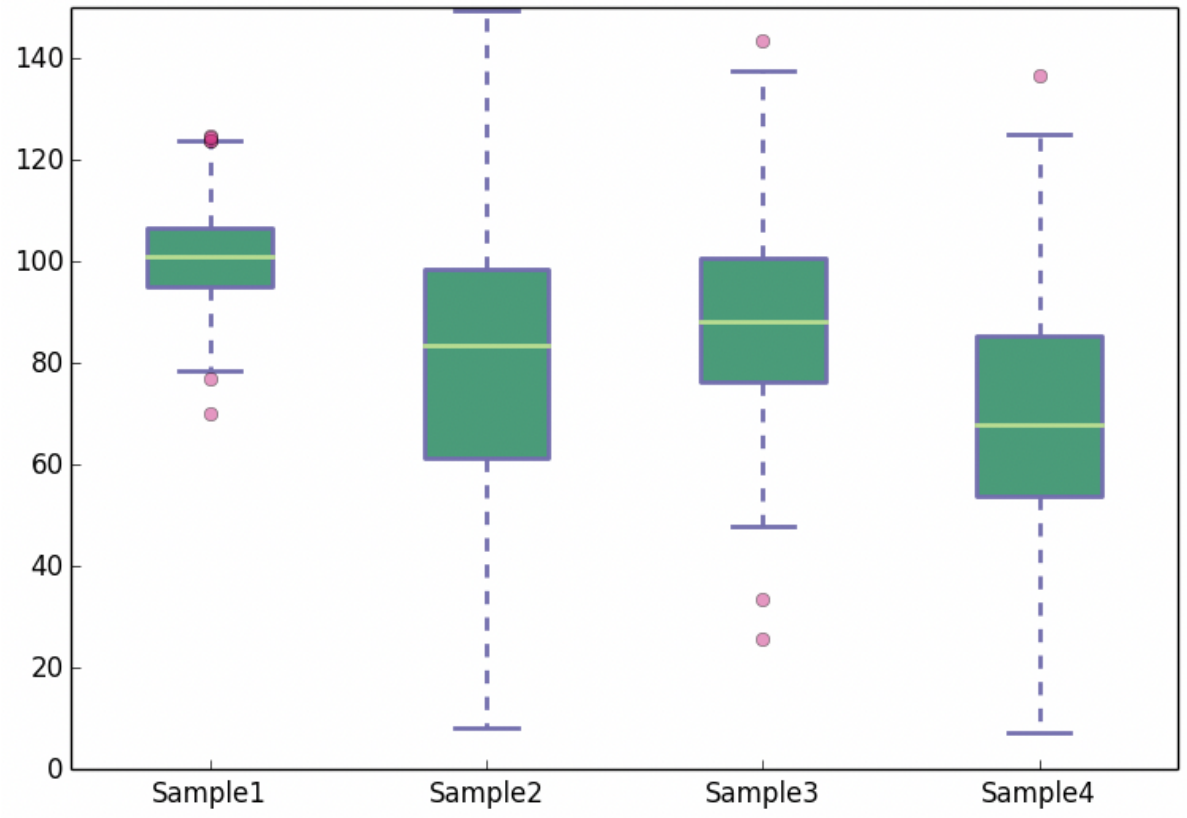
Multiple histograms

- Plotting **multiple histograms (**and **kernel density estimates** of the distribution, here) on the same axes is a way to visualize how different variables compare (or how a variable differs over specific groups).



Boxplots

- A **boxplot** is a simplified visualization to compare a quantitative variable across groups. It highlights the range, quartiles, median and any outliers present in a data set.



Communication

Key Considerations

- Who is your **audience**
- What questions are you answering?
- Why should the audience care?
- What are your major insights and surprises?
- What change do you want to affect?

Communication

Don't Make Them Think!

- Your audience does not want to spend cognitive effort on things you know and can just show them
- Lead them through the major steps of your story
- Point out interesting key facts and insights using captions and annotations

Communication

Final Takeaways

- How you choose to display your data greatly influences how people interpret the data
- Humans are visual, *emotional* creations; make graphs that don't make others feel confused, insulted, etc
- Your graphs should illicit good feelings and effectively convey your narrative