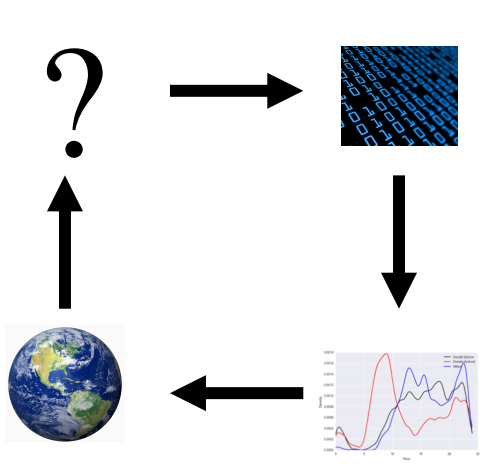


# Data 100

## Lecture 7: EDA & *Visualization*

# Exploratory Data Analysis (EDA)

*"Getting to know the data"*

A process of transforming, visualizing, and summarizing data to:

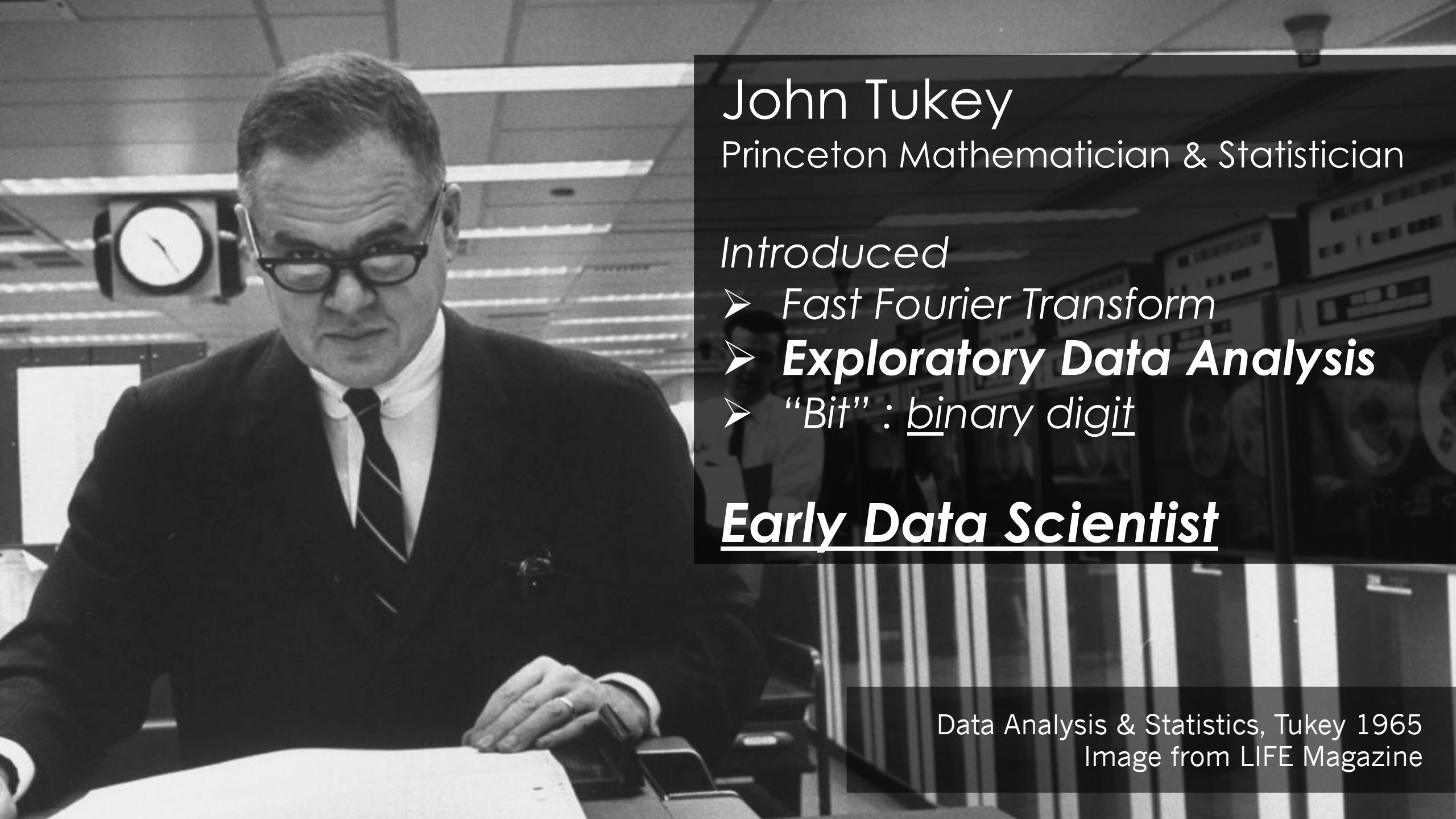
- Build/confirm understanding of the data
- Identify and address potential issues in data
- Inform the subsequent analysis
- Discover *potential* relationships

- **EDA is an open-ended analysis**
  - Be willing to find something surprising

# Exploratory Data Analysis (EDA)

*"Getting to know the data"*

- We used EDA with the $CO_2$ data and DAWN data to check the quality of the data.
- We also use EDA to help prepare for formal modeling.
- We also use EDA to confirm our modeling was reasonable
- Plots can uncover features, distributions, and relationships that can't be detected from numerical summaries



# John Tukey

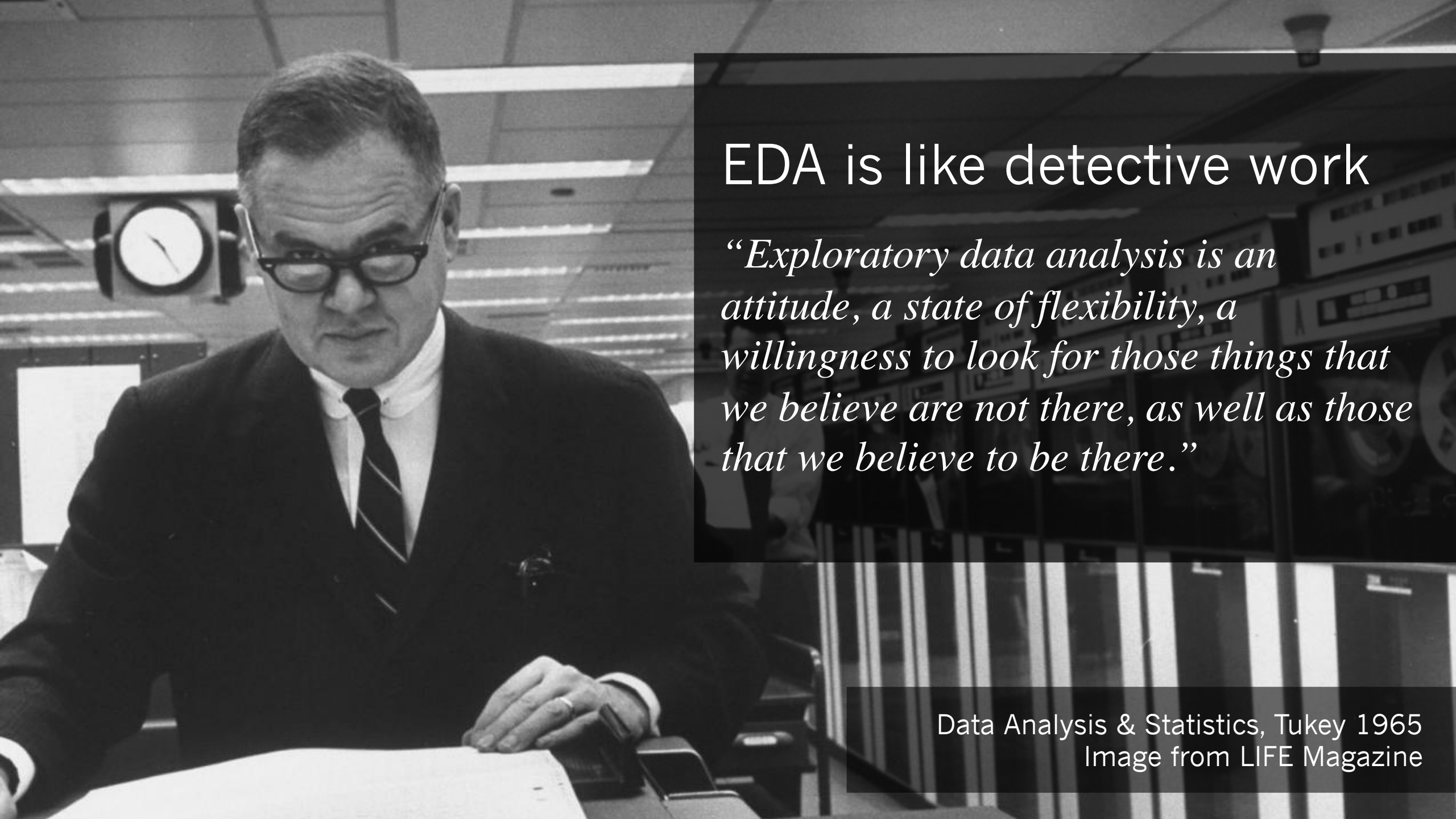
Princeton Mathematician & Statistician

*Introduced*

- *Fast Fourier Transform*
- ***Exploratory Data Analysis***
- *"Bit" : binary digit*

## ***Early Data Scientist***

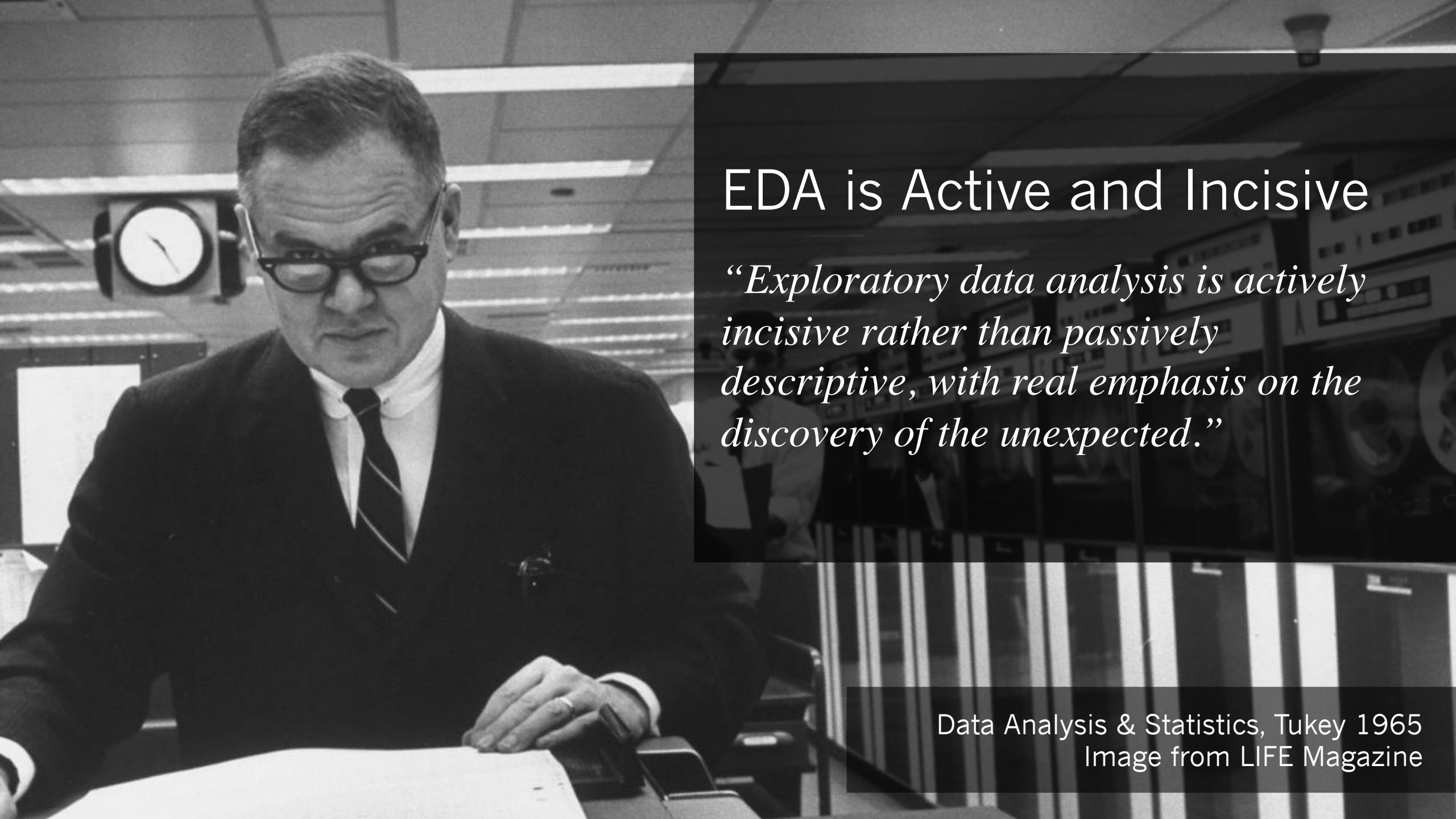
Data Analysis & Statistics, Tukey 1965
Image from LIFE Magazine



# EDA is like detective work

*"Exploratory data analysis is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those that we believe to be there."*

Data Analysis & Statistics, Tukey 1965
Image from LIFE Magazine



# EDA is Active and Incisive

*"Exploratory data analysis is actively incisive rather than passively descriptive, with real emphasis on the discovery of the unexpected."*

Data Analysis & Statistics, Tukey 1965
Image from LIFE Magazine

# The Variable Represents

# The Variable Represents

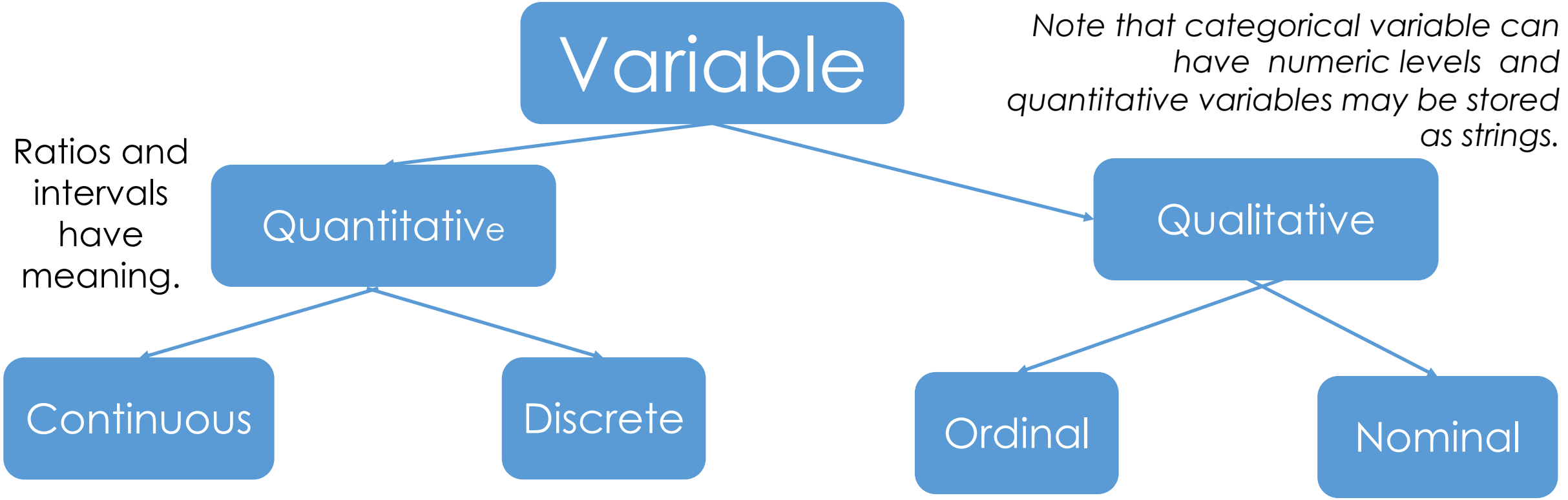
Urban Dictionary:

Go and be a good example to the others of your group or in your position

Huh?

*A Variable represents a feature*

It is distinct from it's coding in a data file or data frame. It is more than a column in a table.




*Note that categorical variable can have numeric levels and quantitative variables may be stored as strings.*

Ratios and intervals have meaning.

Could be measured to arbitrary precision.

**Examples:**
- Price
- Temperature

Finite possible values

**Examples:**
- Number of siblings
- Yrs of education

Categories w/ levels but no consistent meaning to difference

**Examples:**
- Preferences
- Level of education

Categories w/ no specific ordering.

**Examples:**
- Political Affiliation
- CalD number

| | Quantitative Continuous | Quantitative Discrete | Qualitative Nominal | Qualitive Ordinal |
|---|---|---|---|---|
| $CO_2$ level | | | | |
| Number of siblings | | | | |
| GPA | | | | |
| Income bracket | | | | |
| Race | | | | |
| Number of years of education | | | | |
| Yelp Rating | | | | |
| Lane of traffic (left, middle, right) | | | | |

# Basic Plots

Match Variable Type to Plot Type

# Basic Visualizations

- How to choose the “right” one(s)
- How to read them –
  - Distributions
  - Relationships

# Kaiser Study

- Oakland Kaiser mothers
- 1960s
- Measure the babies weight (in ounces) at birth
- All babies:
  - Male
  - Single births (no twins, etc.)
  - Survived 28 days

**Data provenance:**

Mothers who use Kaiser

Starts as administrative dataset, expanded into study with new data

Selection mechanism **not** random

# Information collected on mother's and their babies

- Birth weight (ounces)
- Gestation (weeks)
- Parity - total number of previous pregnancies
- Mother's height and weight
- Mother's smoking status
- Mother's age, race, education level, income level
- Father's information and more…

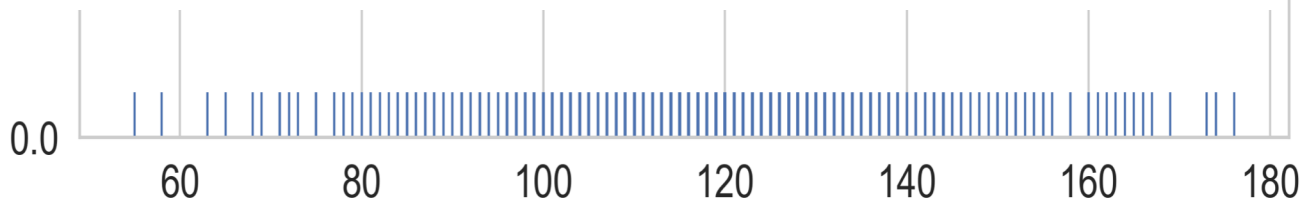
# One Variable

What is the Distribution of the values of the variable?

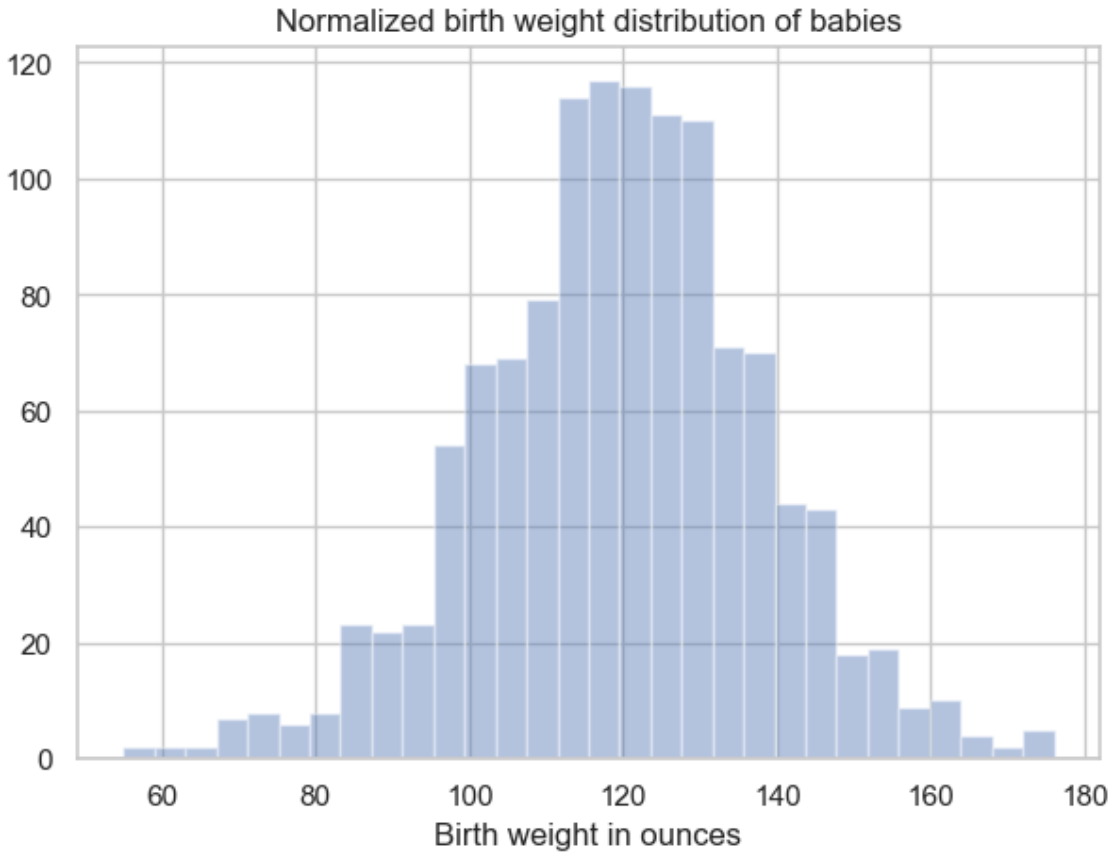
# Quantitative – Continuous

- Birthweight
- The most basic visual representation of one quantitative variable is the *rug plot*

Hard to see much of the distribution with this rug plot



# Birthweight



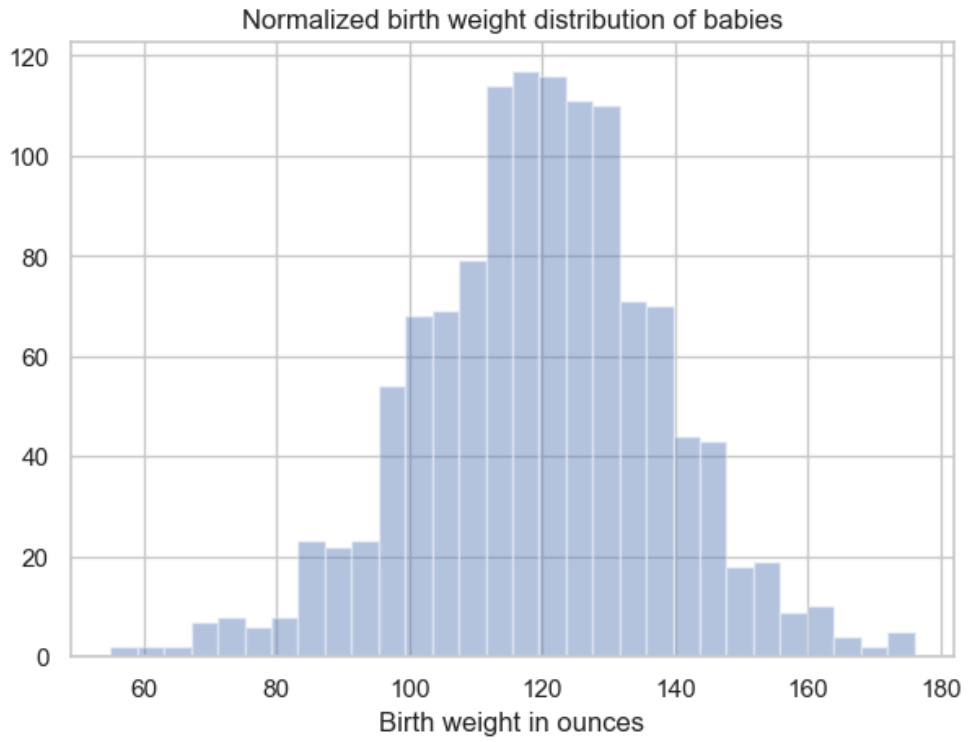

Histogram

With the histogram we hide the details of individual observations and view the general features of the distribution.

How would we describe the distribution of birth weight?

# Distribution Features

- Modes
  - Number
  - Location
  - Size
- Symmetry
  - Symmetric
  - Skewed left or right
- Tails
  - Long, short, “normal”
- Gaps
- Outliers

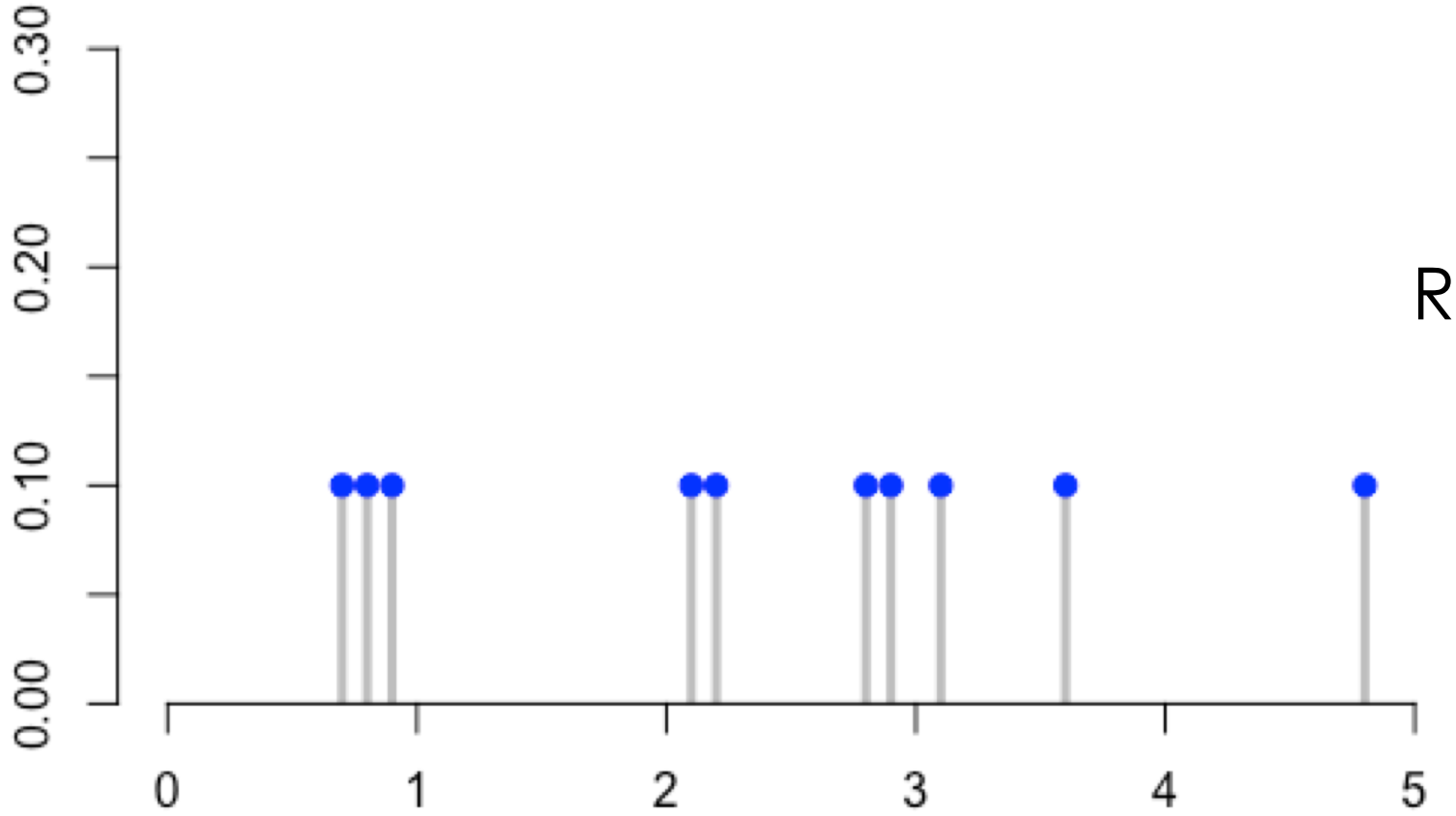


# Distributions & Smoothing

# A Small Dataset

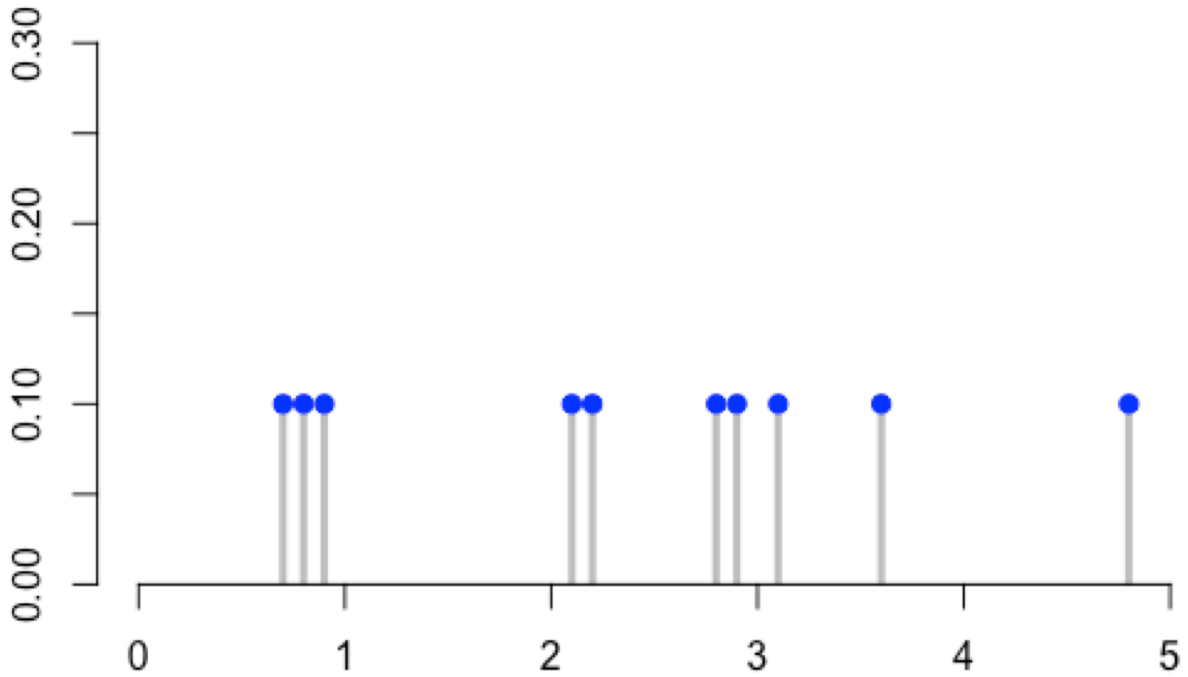
10 values

0.7, 0.8, 0.9, 2.1, 2.2, 2.8, 2.9, 3.1, 3.6, 4.8




Rug Plot

# We want to smooth these rug threads




BECAUSE

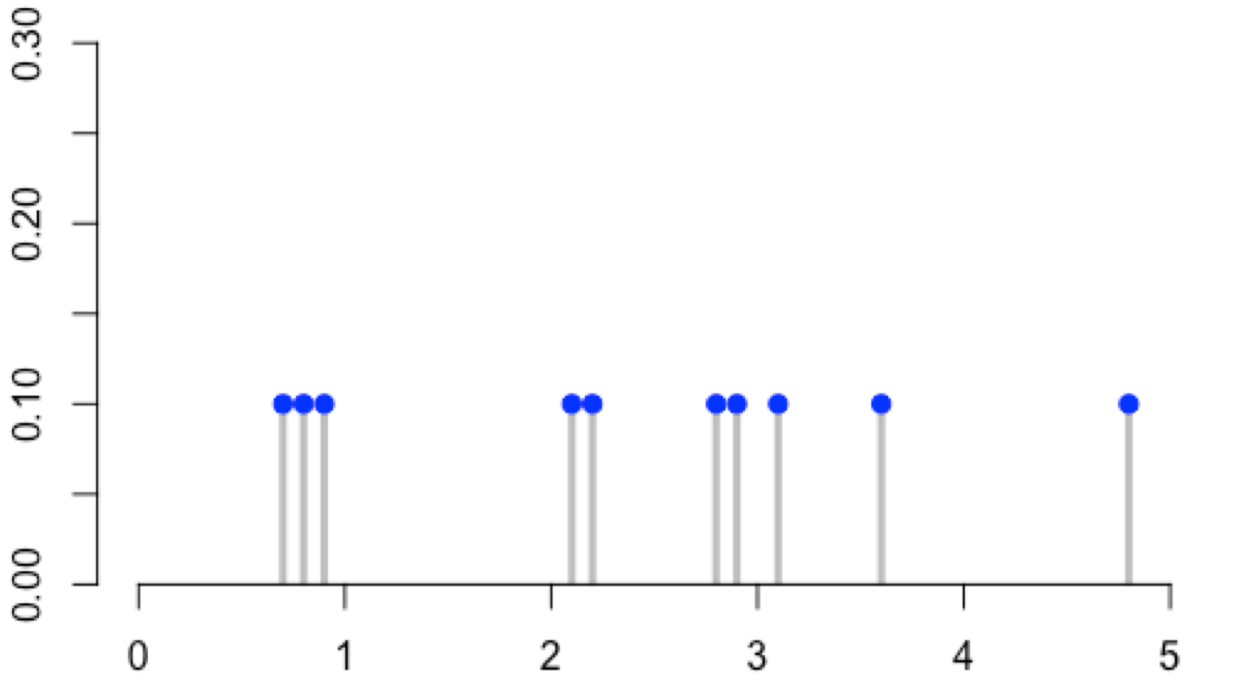
- this is a sample and we believe that other values near the ones we observed are reasonable
- we want to focus on general structure rather than individual observations

# Important Properties of Histograms

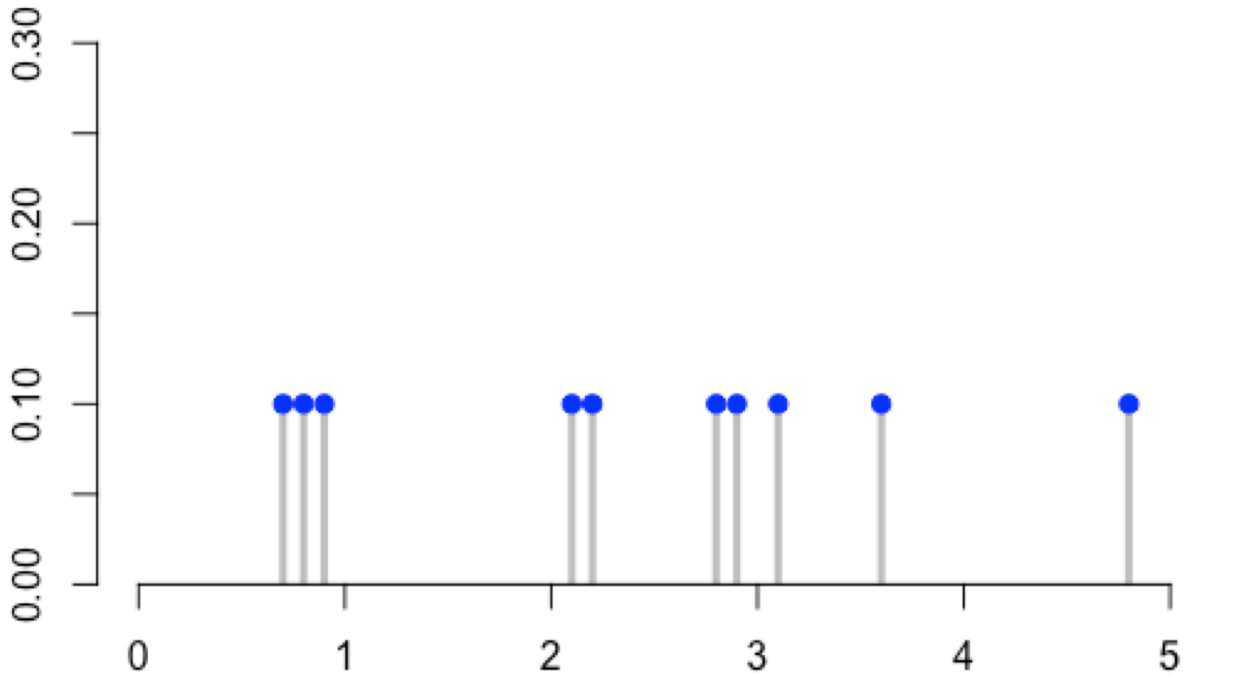
- Total Area of the bars = 100% (or 1)
- Units on the y-axis are percent/x-unit
- Area of a bar = percentage of values in that bar

unit matching:

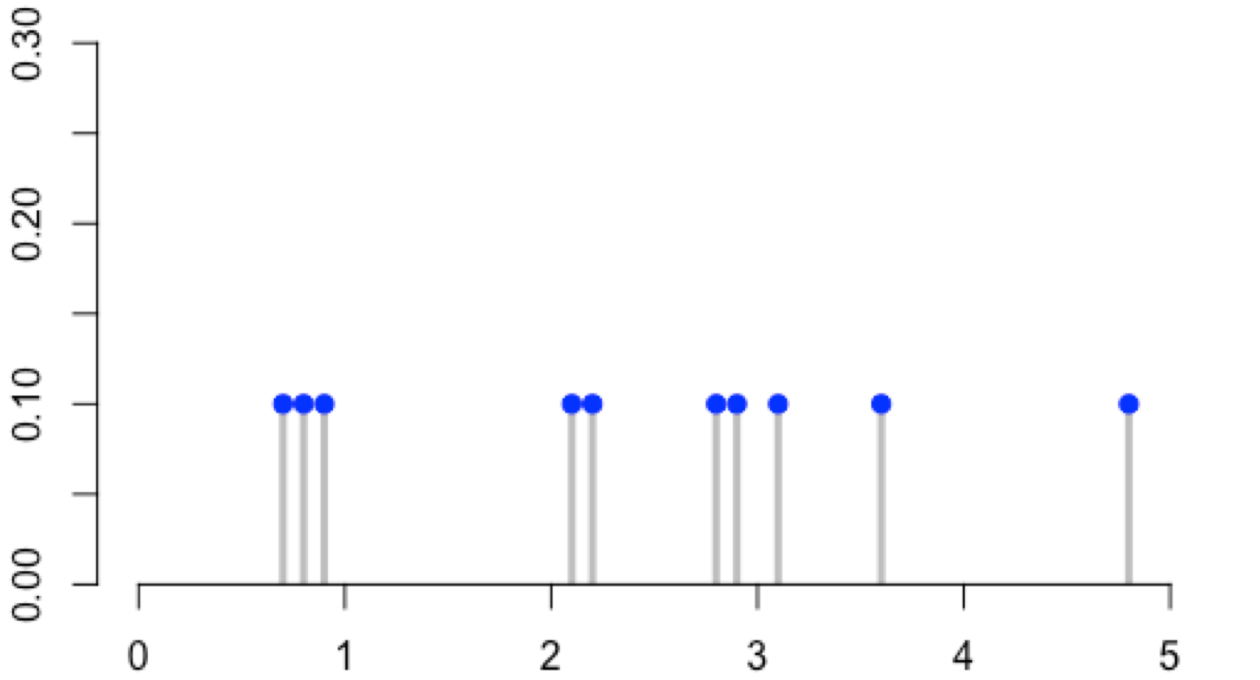
# Example



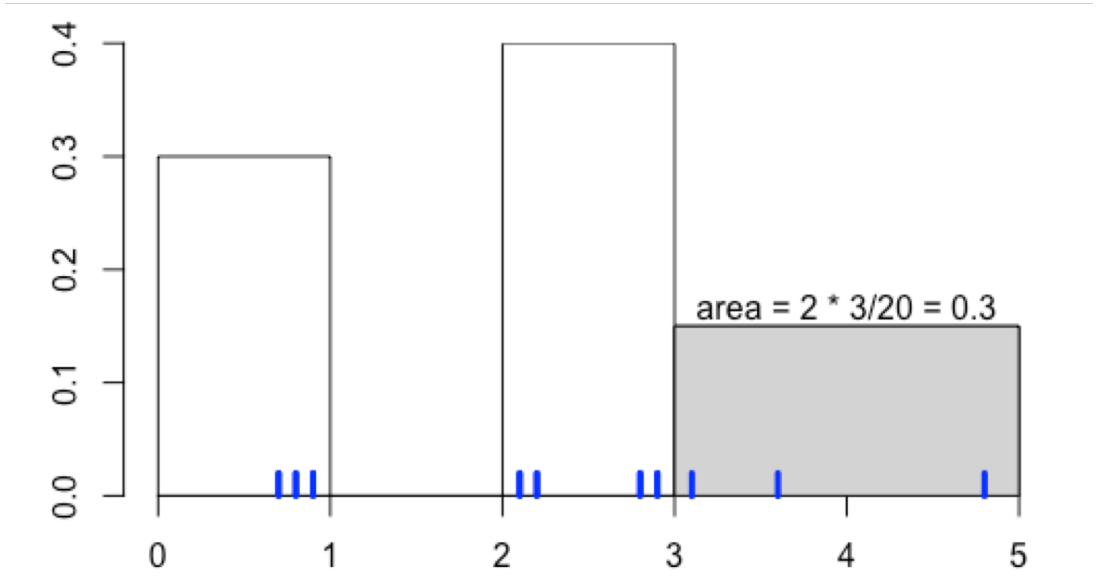
# Example



# Example



# A histogram smooths

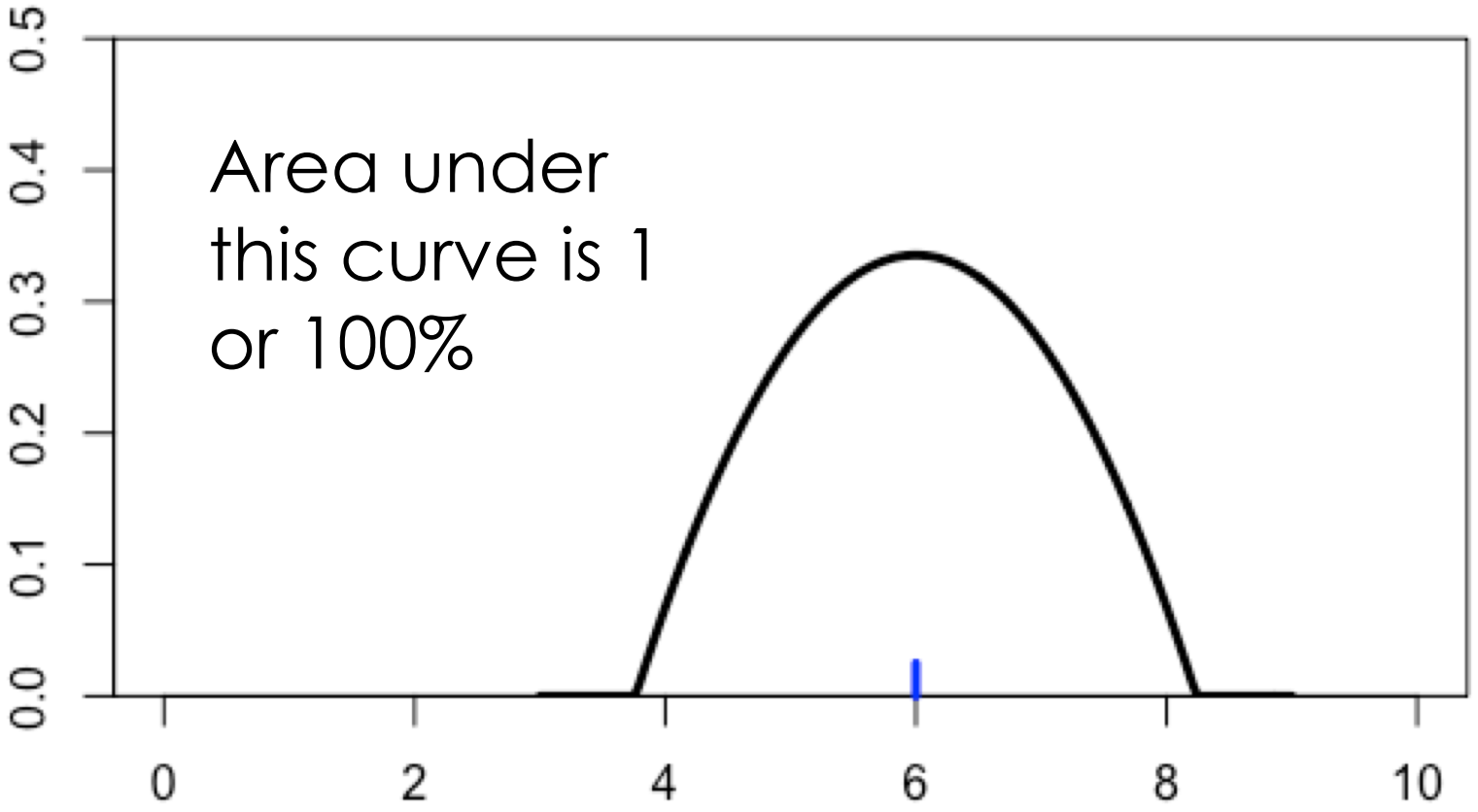



We want to smooth out these points because:

- this is a sample and we believe that other values near the ones we observed are reasonable
- we want to focus on general structure rather than individual observations

The values 3.1, 3.6, and 4.8 have their proportion (3/10) spread over the bin [3,5] That is, without the rug, we can't tell where the points are in the bin

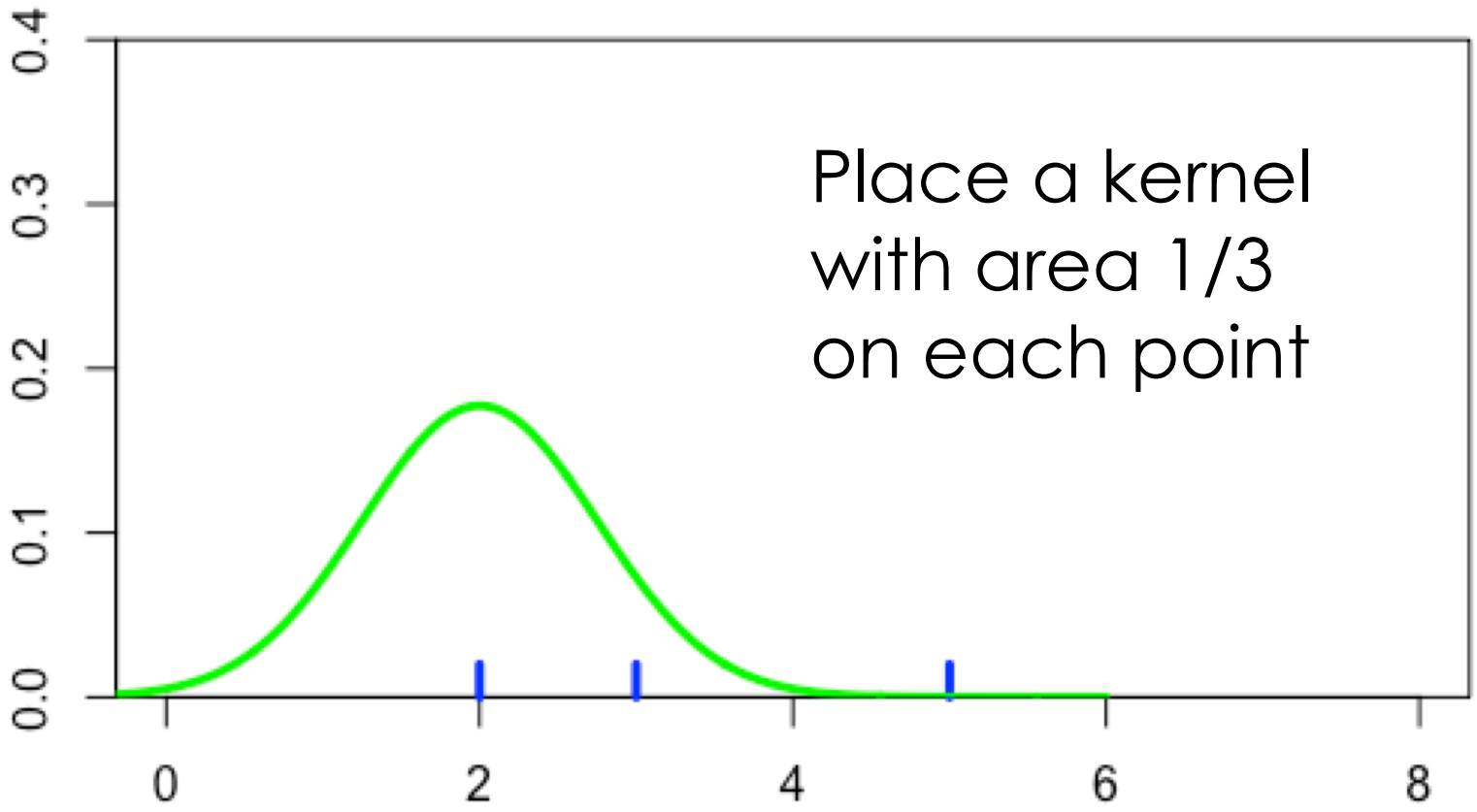
# Kernel Density Estimate: Alternative Smoother




Consider one point

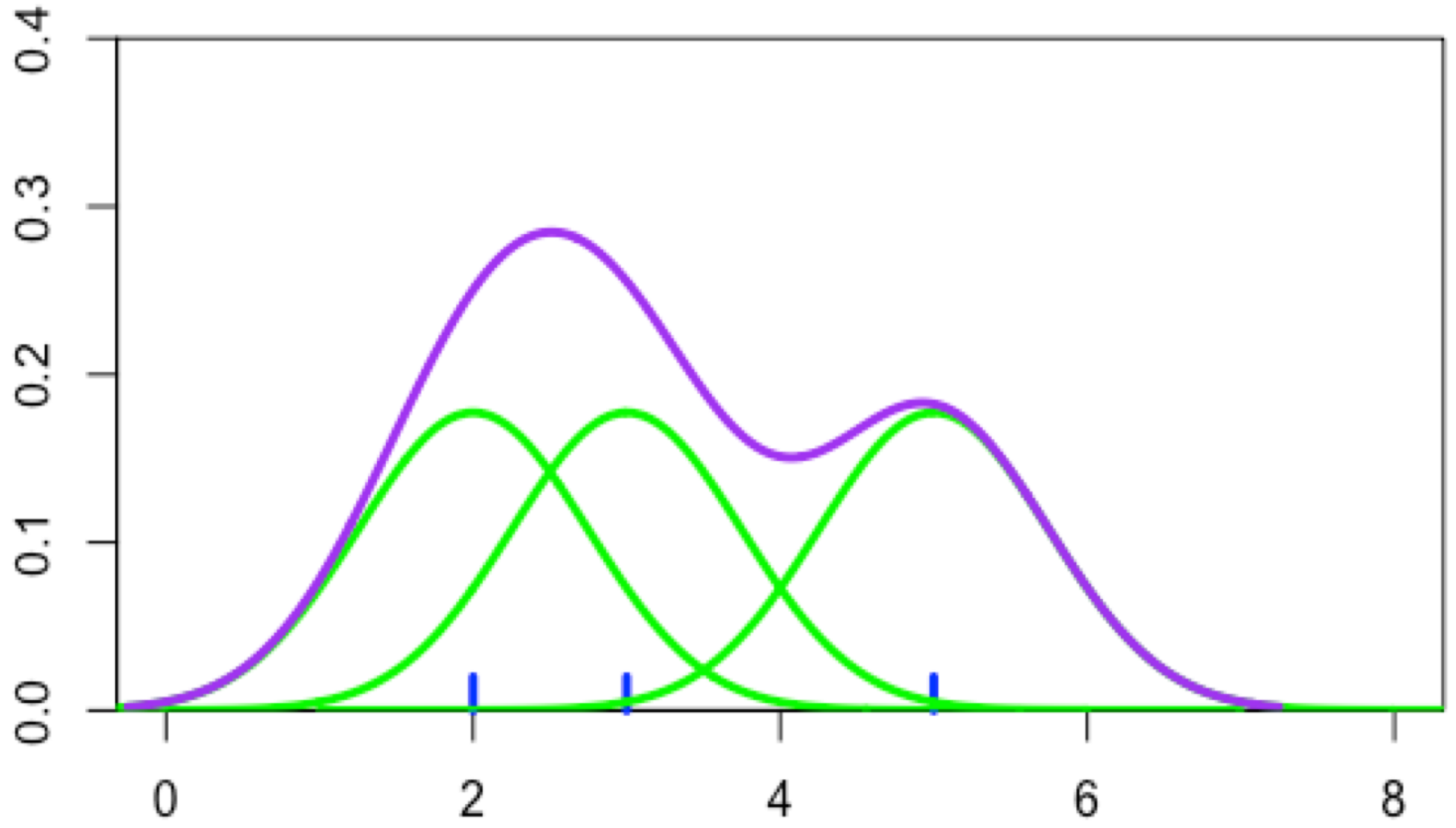
Smooth with a kernel function, rather than in a histogram bin

# 3 points –
# each represents 1/3 of the data



# KDE – 3 points

Sum the 3 kernels at each point to get the density curve

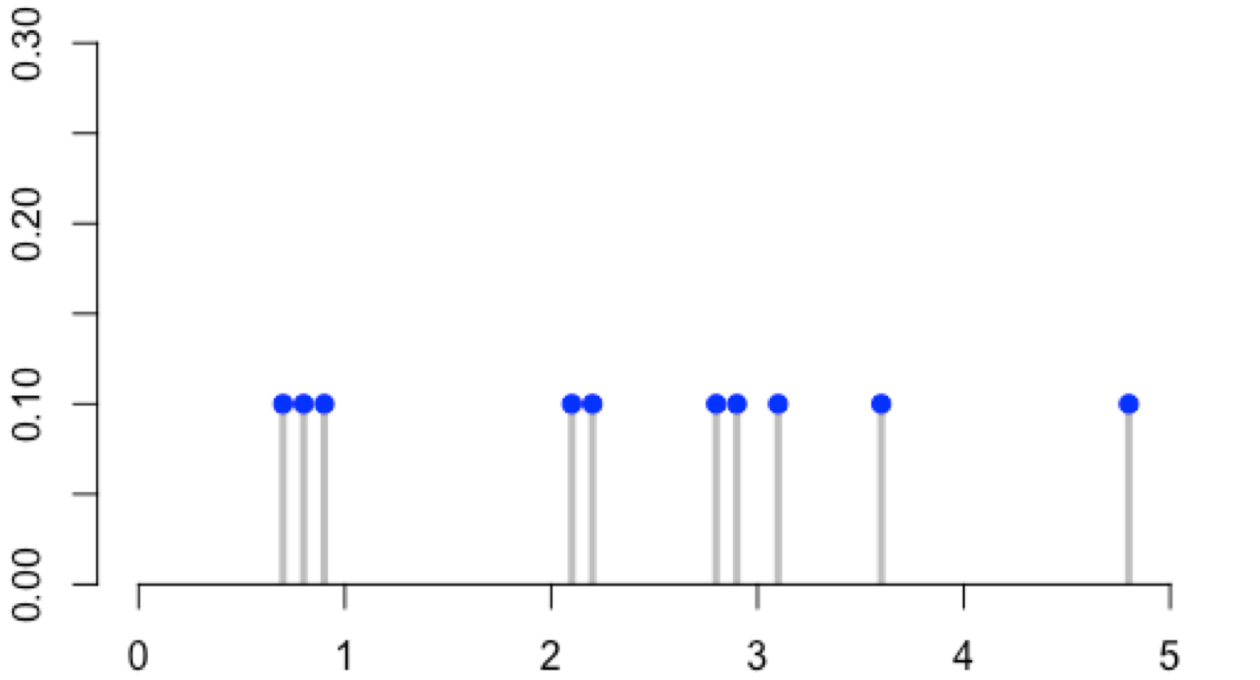



$$f(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i)$$

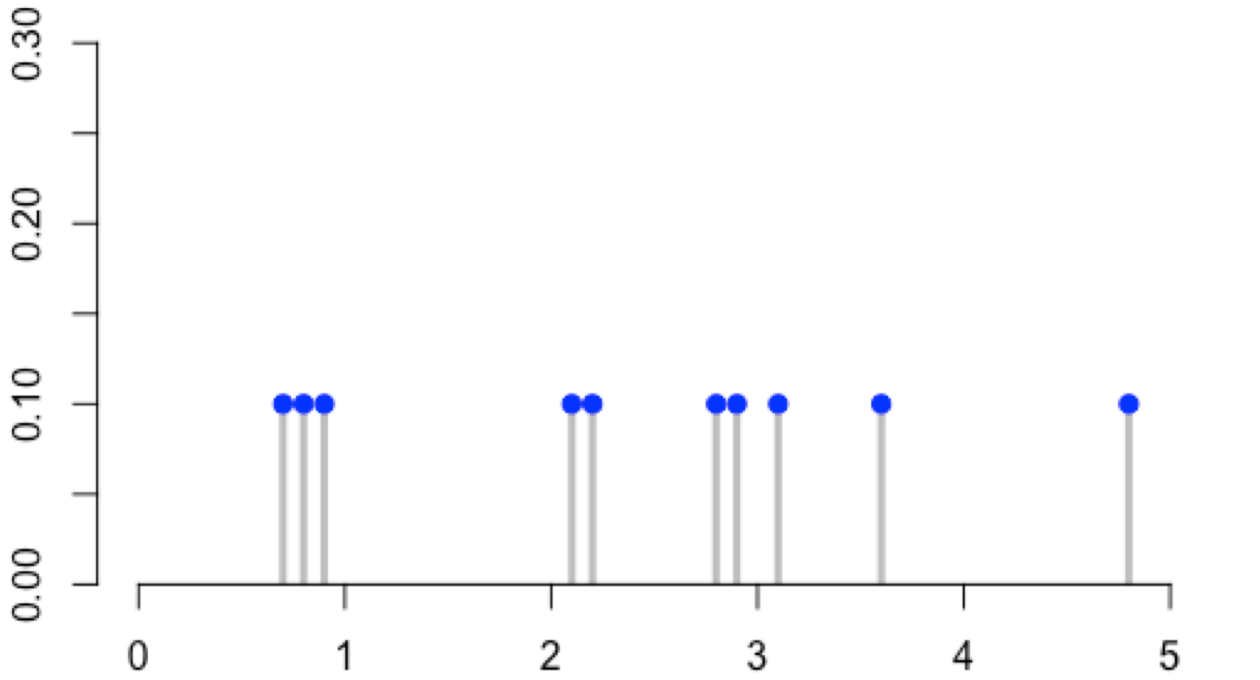
$K_h$ is the green Kernel function

h refers to how peaked /spread the kernel is

# Example

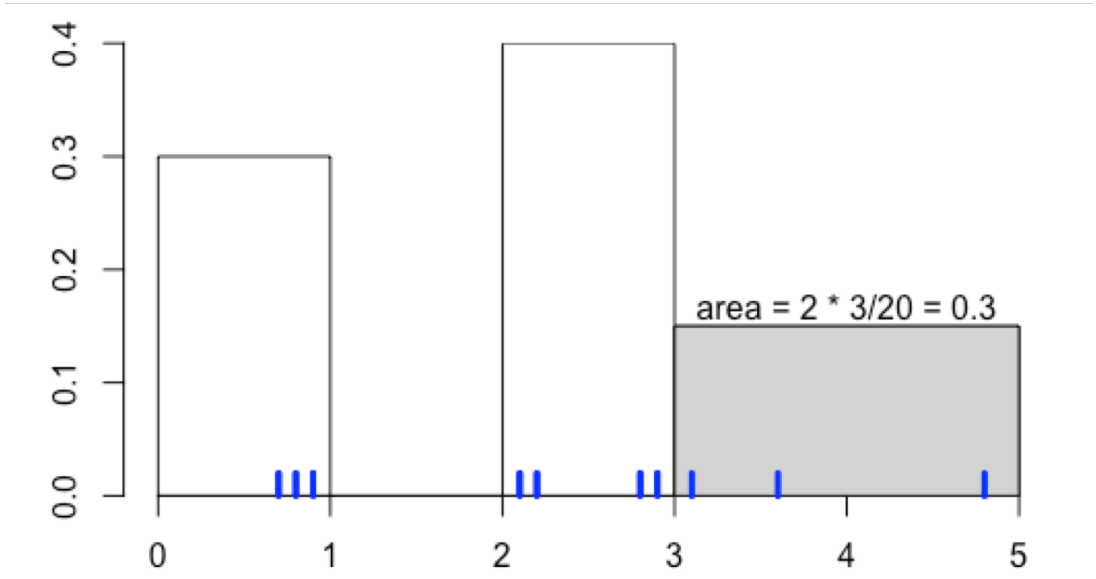


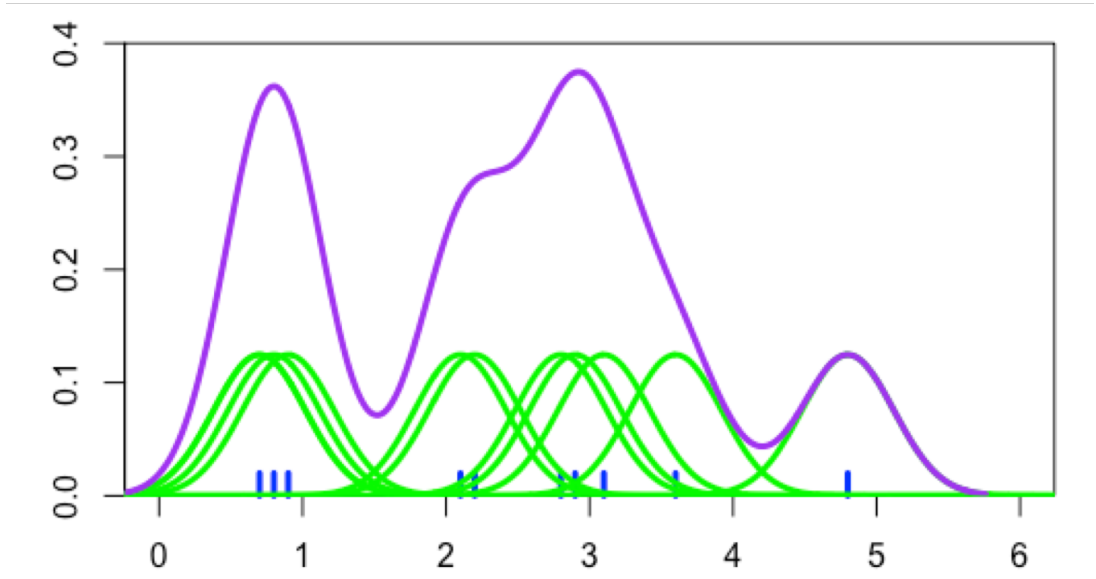
# Example



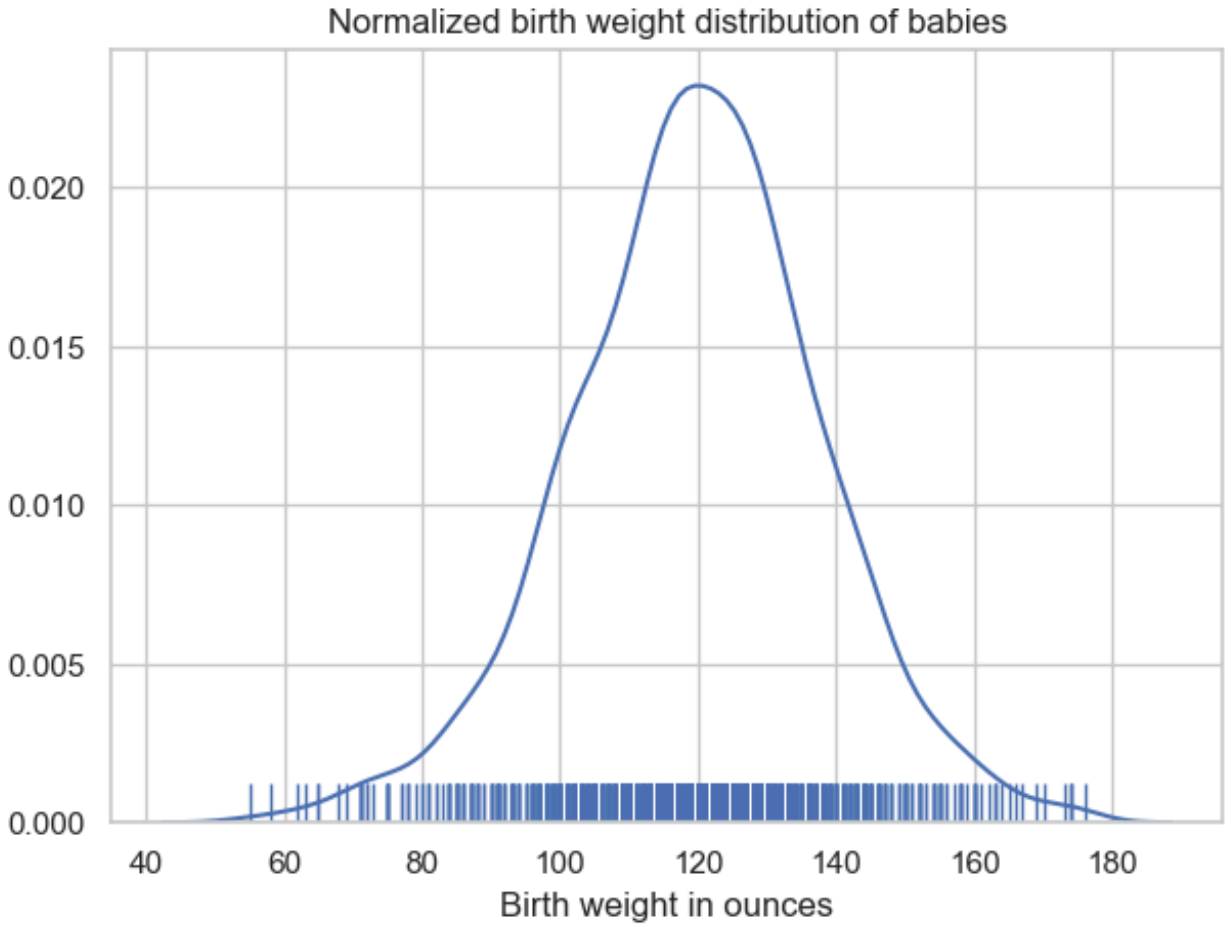
# Compare the Histogram and the KDE

0.7, 0.8, 0.9, 2.1, 2.2, 2.8, 2.9, 3.1, 3.6, 4.8



# Birthweight – Density Curve




How would we describe the distribution of birth weight?

Unimodal

Main mode at 120 oz

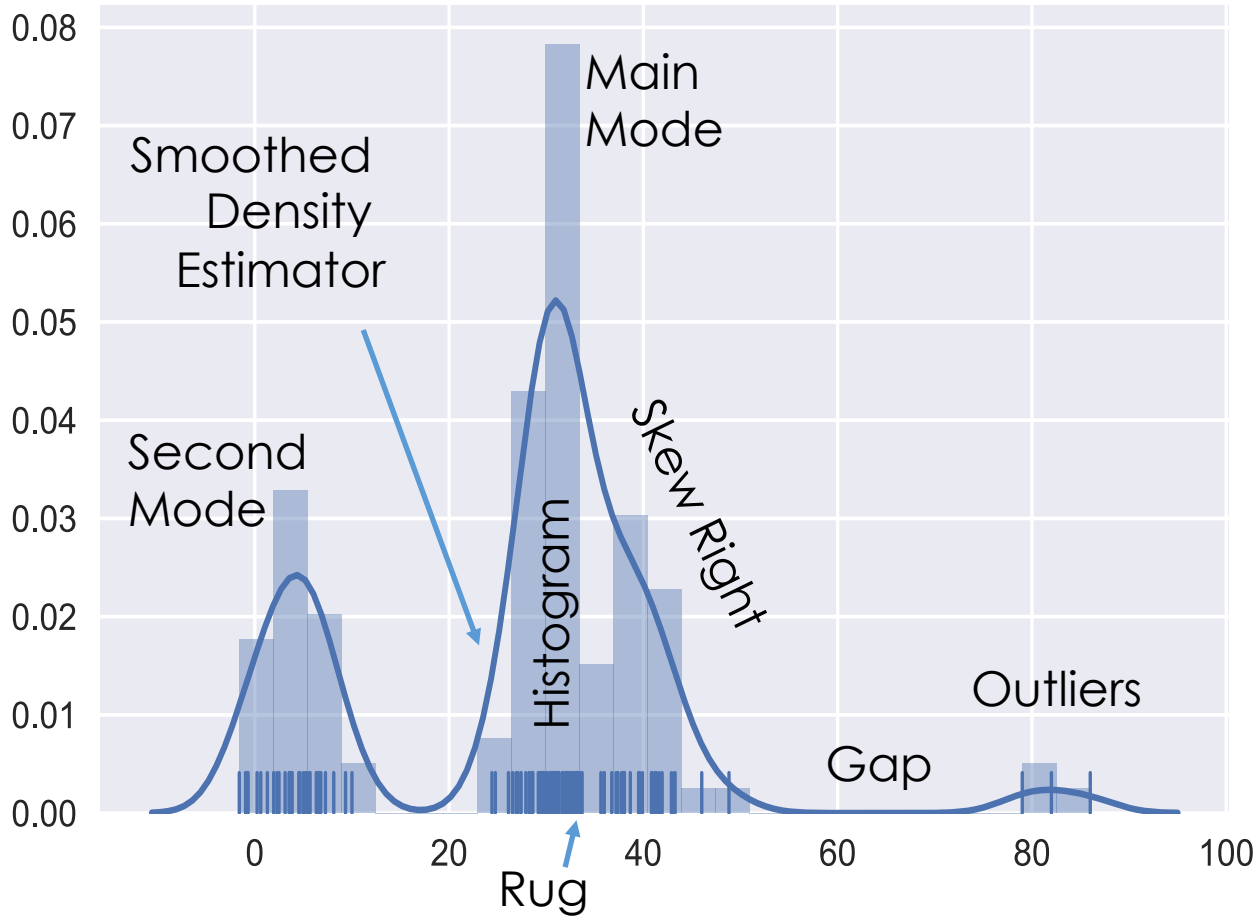
Slight left skew

Tails about normal

# Histograms and Density Curves

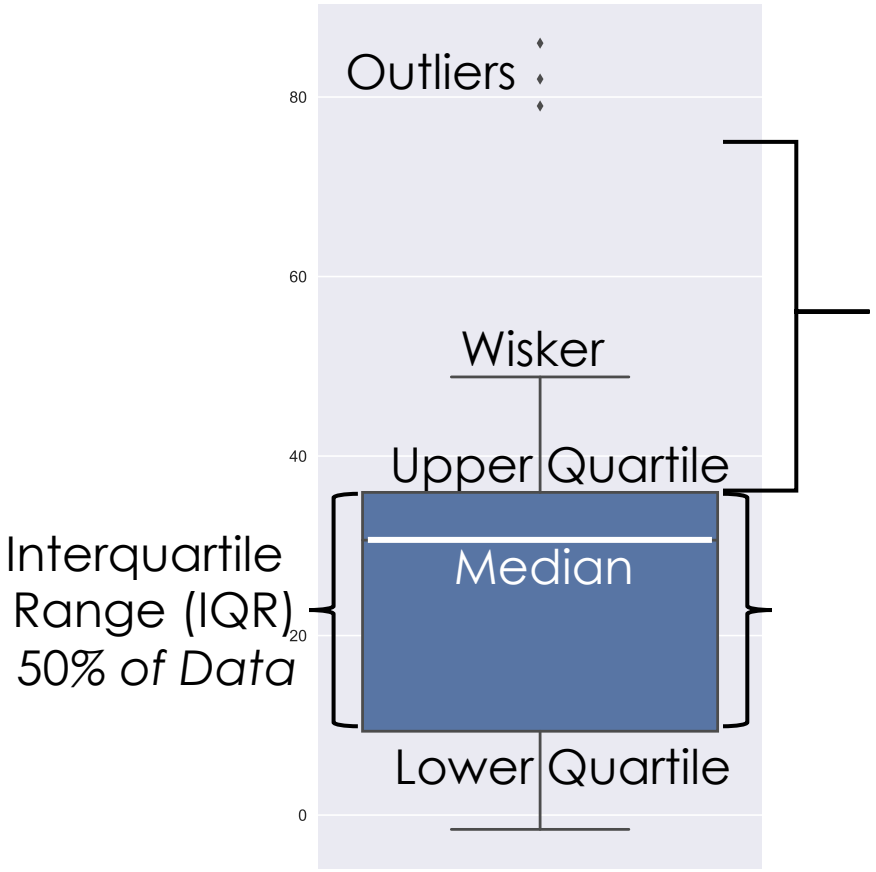
Describes distribution of data – relative prevalence of values

- Histogram
  - relative frequency of values
  - Tradeoff of bin sizes
- Rug Plot
  - Shows the actual data locations
- Smoothed density estimator
  - Tradeoff of "bandwidth" parameter (more on this later)



# Box Plot

- Useful for summarizing distributions and comparing multiple distributions




Outliers are more than 1.5 * IQR away from lower and upper quartiles.

Visualization of summary statistics

Can lose a lot of features, such as…?

# Our Data 0.7, 0.8, 0.9, 2.1, 2.2, 2.8, 2.9, 3.1, 3.6, 4.8

- Median
- Lower Quartile
- IQR
- Hinge

# Our Data 0.7, 0.8, 0.9, 2.1, 2.2, 2.8, 2.9, 3.1, 3.6, 4.8

- Median
- Lower Quartile
- IQR
- Hinge

# Quartiles from Tukey's "depth"

- Depth of the Median = (n + 1)/2
  - Count in from top or bottom of ordered set of values
  - If depth has a half then average the two values on either side

- Depth of Quartile = (round(m) + 1)/2
  - Round the median depth down to nearest integer
  - Count in from bottom to get the LQ and from the top to get the UQ
  - If depth has a half in it then average the two values on either side

# Percentile – Need a more general def

- The $P^{th}$ percentile of a set of data is:

**Smallest** value that has **at least** P% of the data **at or below it**

0.7, 0.8, 0.9, 2.1, 2.2, 2.8, 2.9, 3.1, 3.6, 4.8

$10^{th}$%tile =

$15^{th}$%tile =

$90^{th}$%tile =

$83^{rd}$%tile =

$60^{th}$%tile =

$66^{th}$%tile =

# Percentile – with weighted data

- The $P^{th}$ percentile of a set of data is:

**Smallest** value that has **at least** P% of the data **at or below it**

| 5. | 5. | 5. | 5. | 5. | 5. | 20. | 20. | 20. | 20. | 50. | 50. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ½ | ½ | ½ | ½ | ½ | ½ | 1 | 1 | 1 | 1 | ½ | ½ |

$50^{th}$%tile =

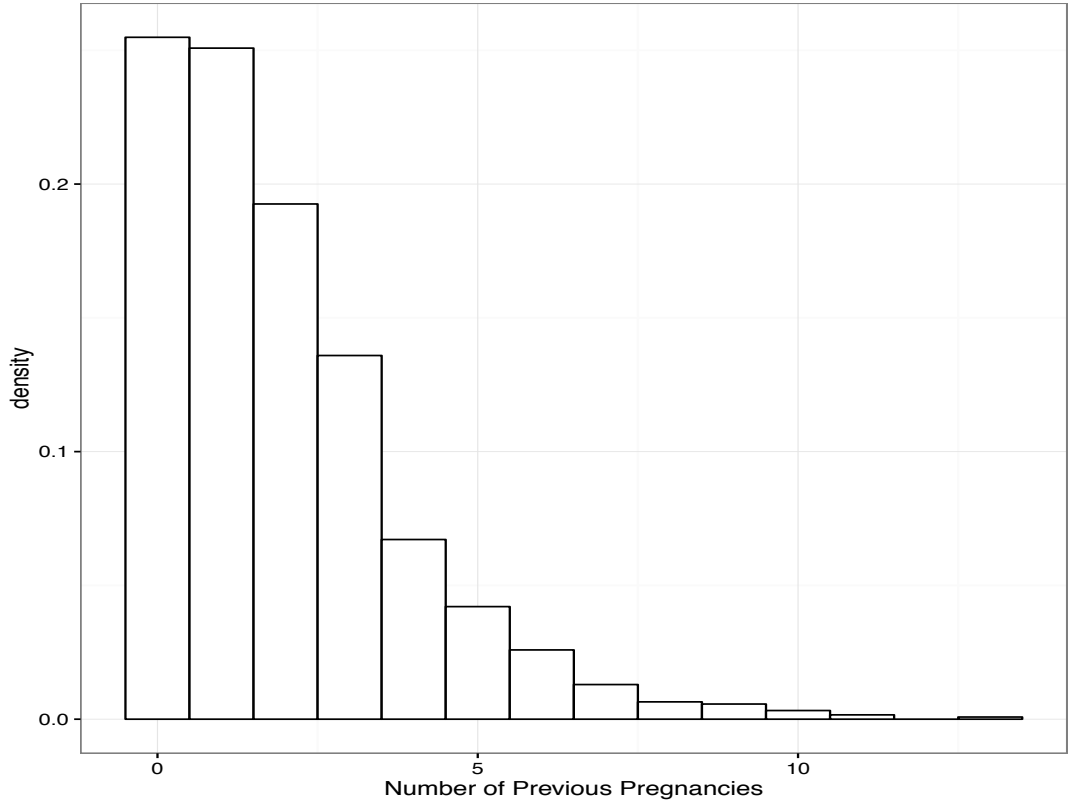
$75^{th}$%tile =

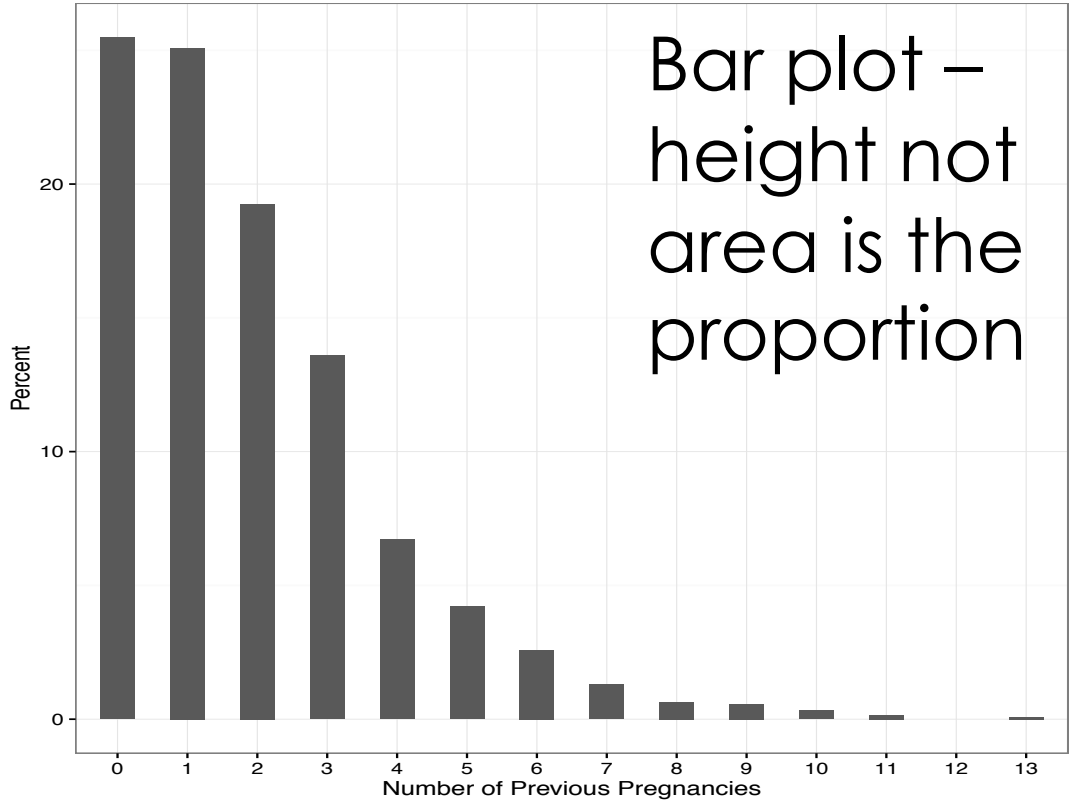
# Quantitative Discrete

We look for the same features

- Symmetry and skew
- Modes (number, location, and size)
- Tails (long, short, normal)
- Gaps
- Outliers

# Discrete Quantitative # of Siblings




Bar plot – height not area is the proportion

What's the difference between these 2 plots?

# Qualitative

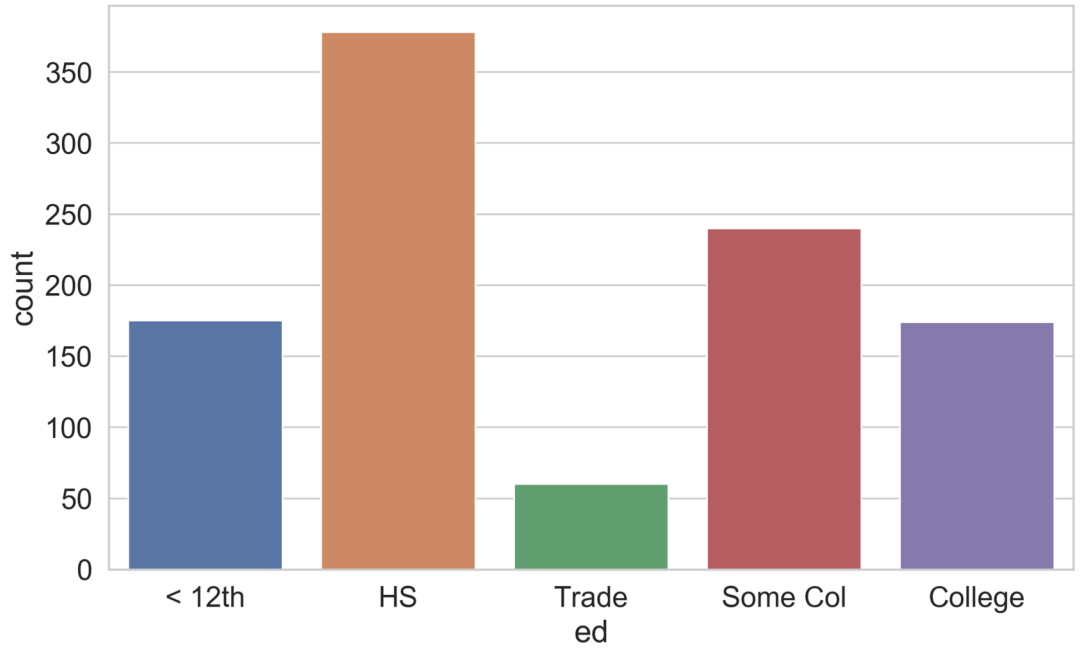
We look at the relative size of groups

- Equally distributed
- Symmetry, Modes, Tails and Gaps don't make sense
- Do most fall in one group?

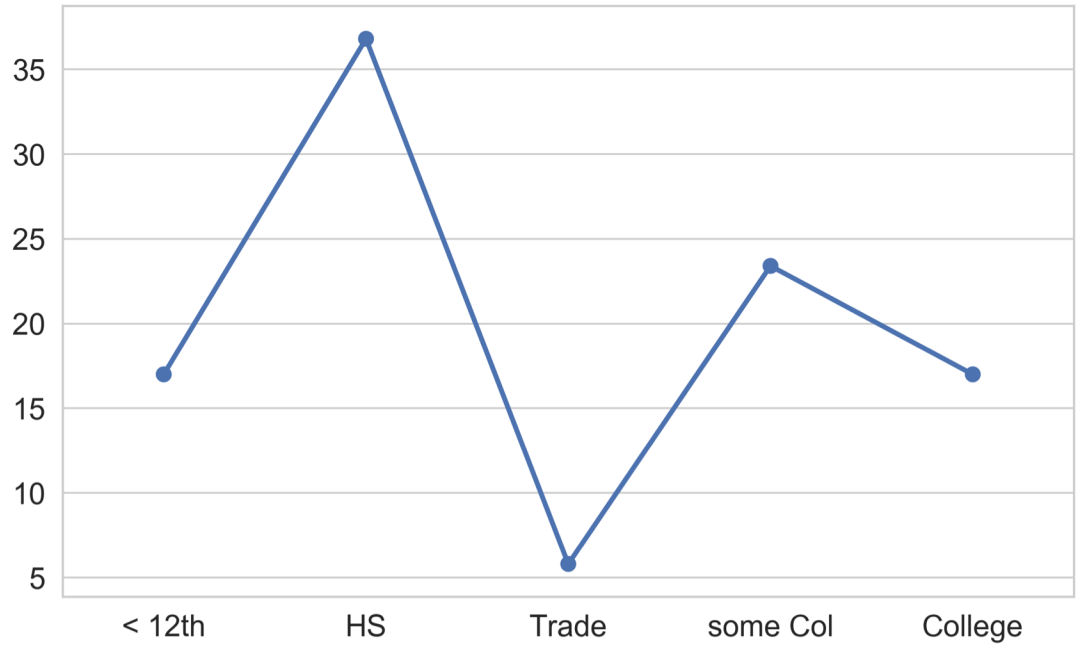
Answers have implications in building prediction models

# Qualitative Variable Education level

Bar Width – has no meaning




Dot plot focuses on comparison of the values




Why do we not reorder the bars according from shortest to tallest?

# Pairs of Variables

Combinations:

Both qualitative,

One qualitative and one Quantitative,

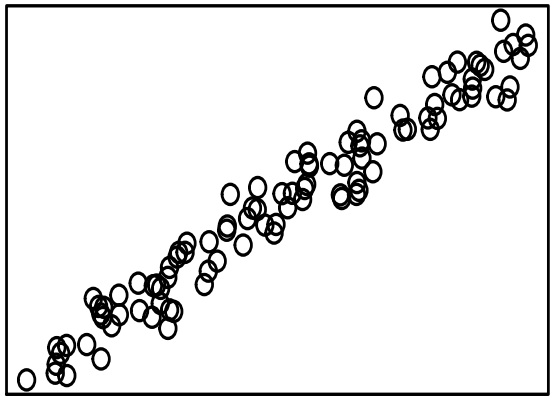
Both Qualitative

# Plotting Pairs of Quantitative Variables

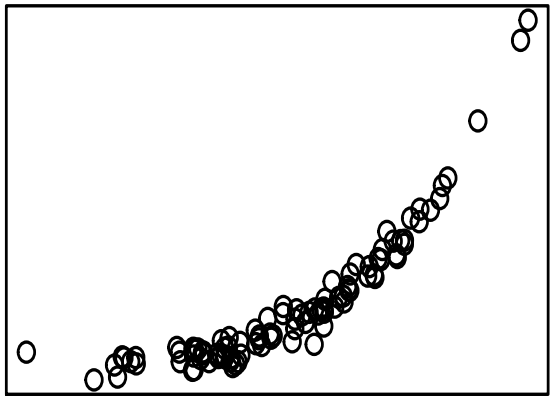
- Scatter plot uncovers form of relationship between 2 variables
- Linear relationships are particularly simple to interpret
- Simple and elegant statistical theory for linear relationships
- Models are typically approximations, choose a simpler model over a complex one

# Common Relationships

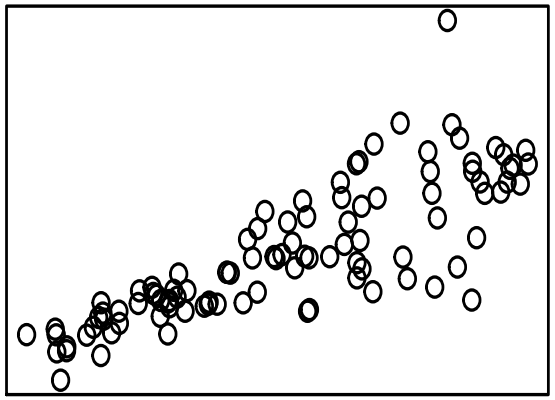
simple linear



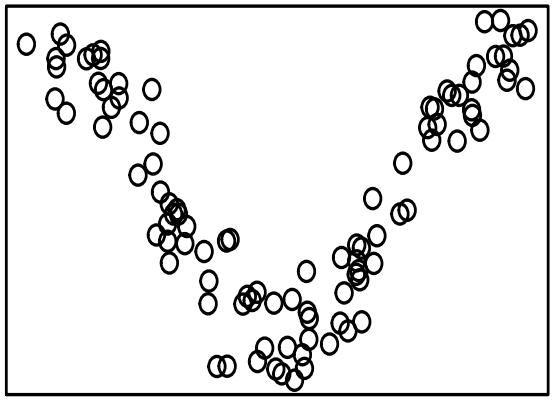
simple nonlinear

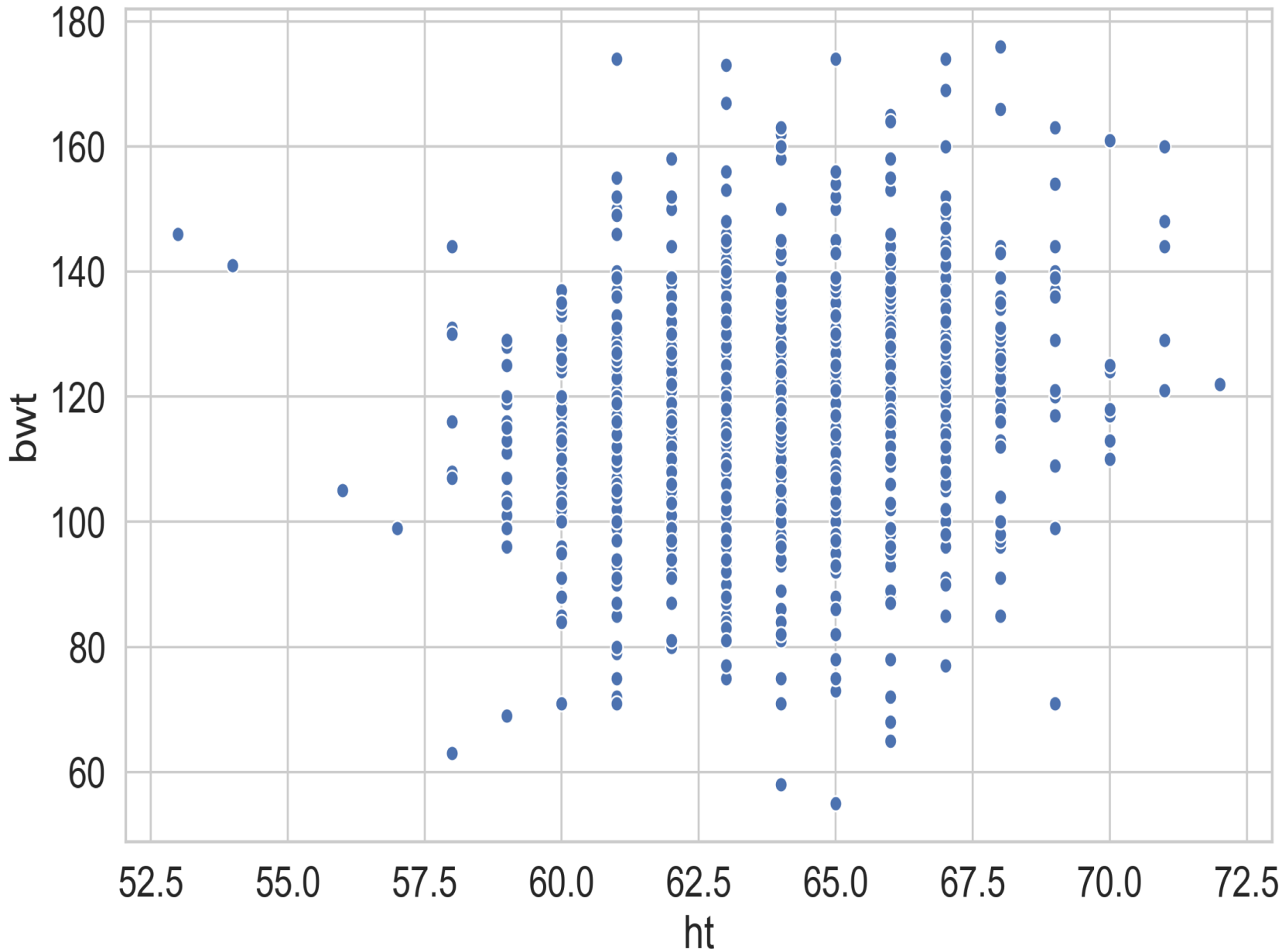


unequal spread

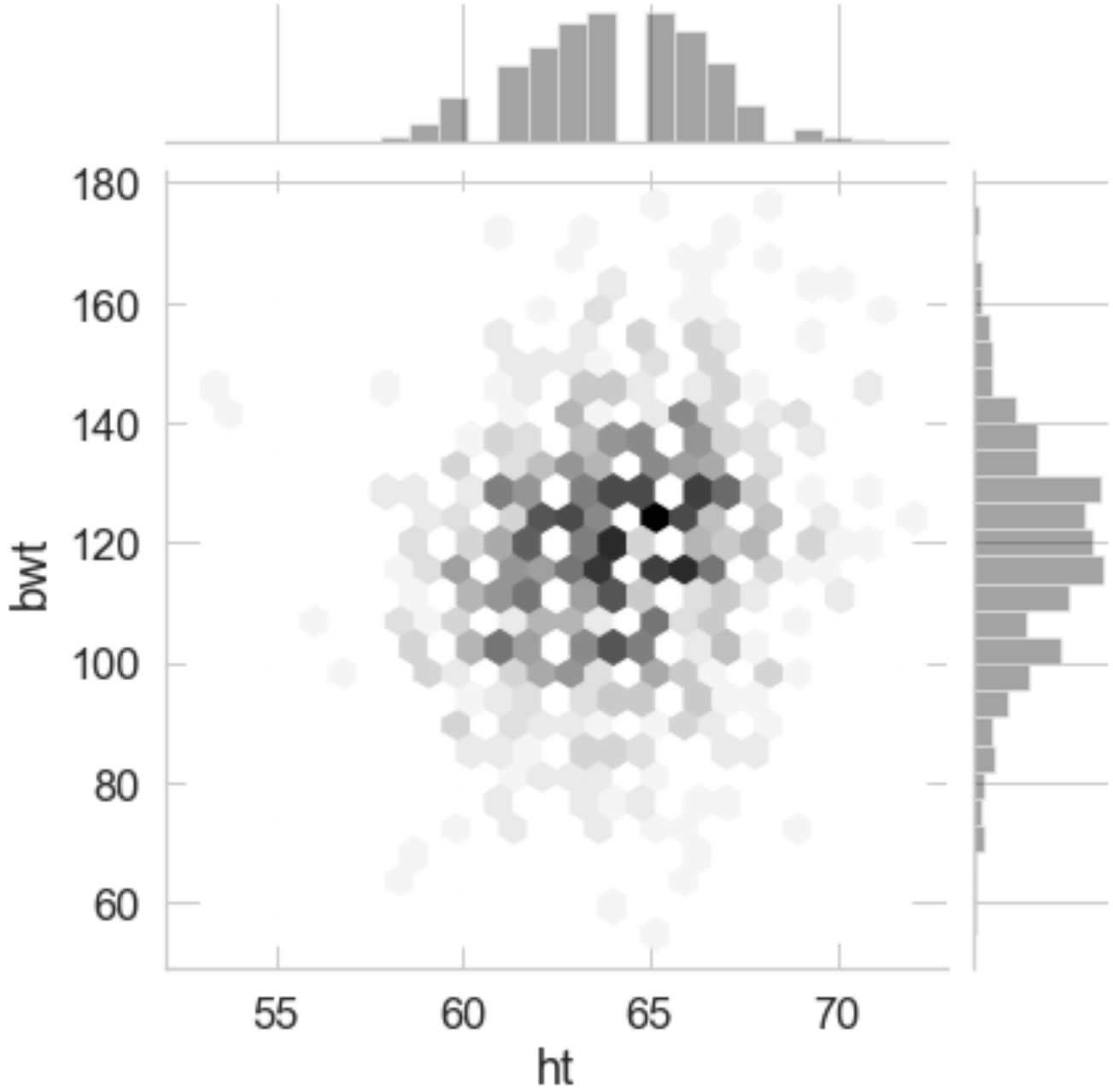


complex nonlinear

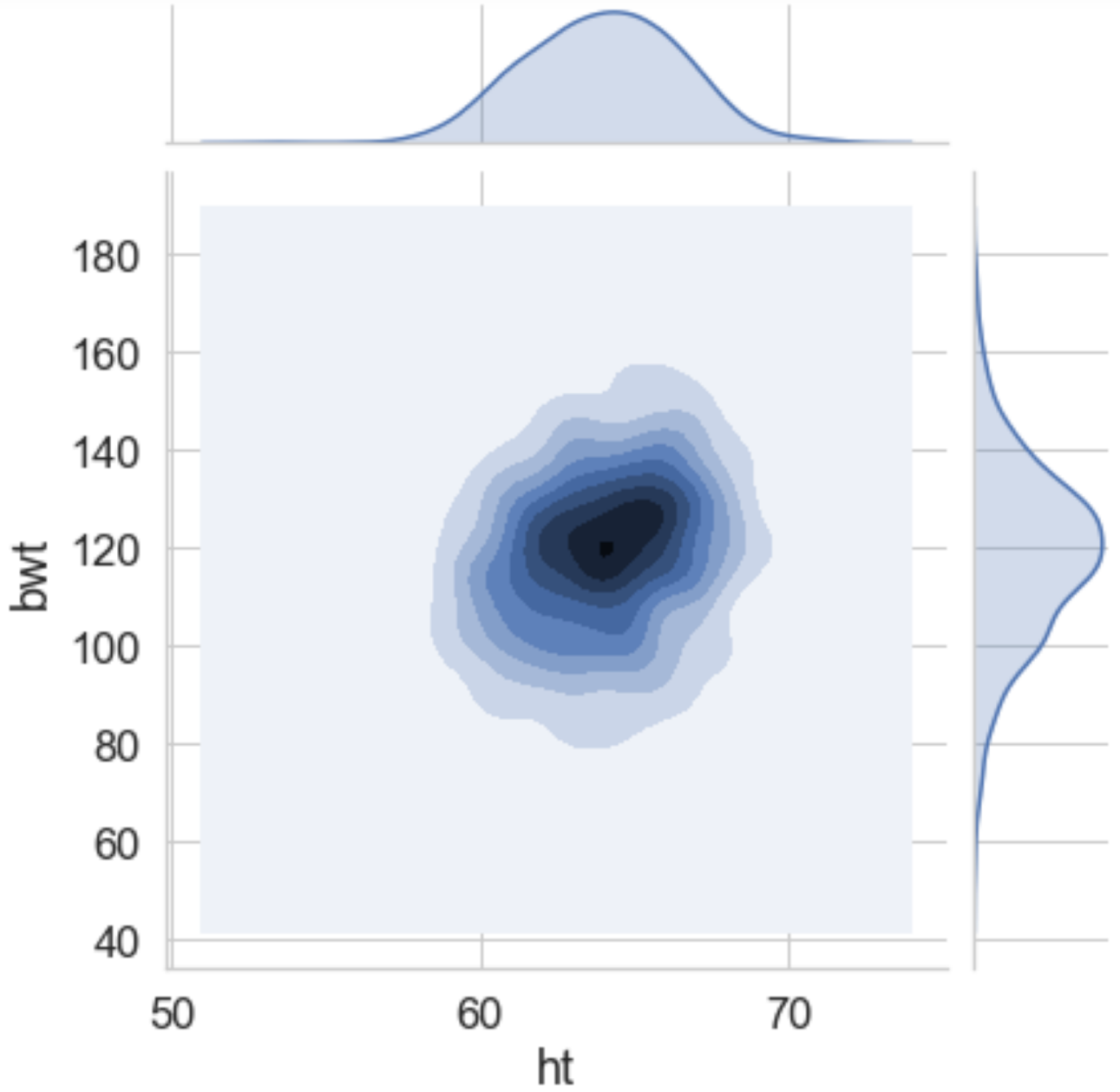



180
160
140
120
100
80
60
bwt
52.5
55.0
57.5
60.0
62.5
65.0
67.5
70.0
72.5
ht

# Hex Bin

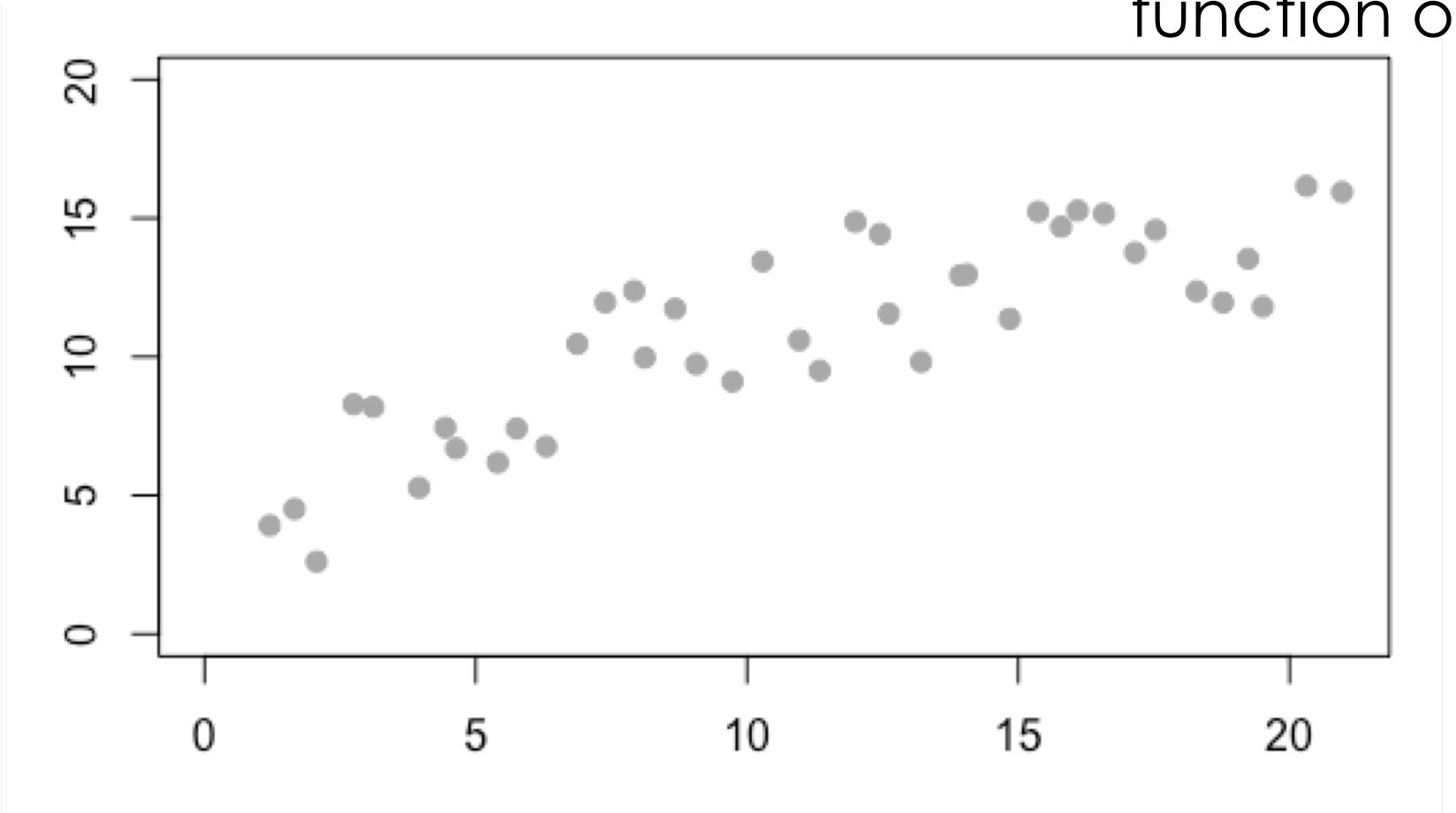


# Smooth Contour



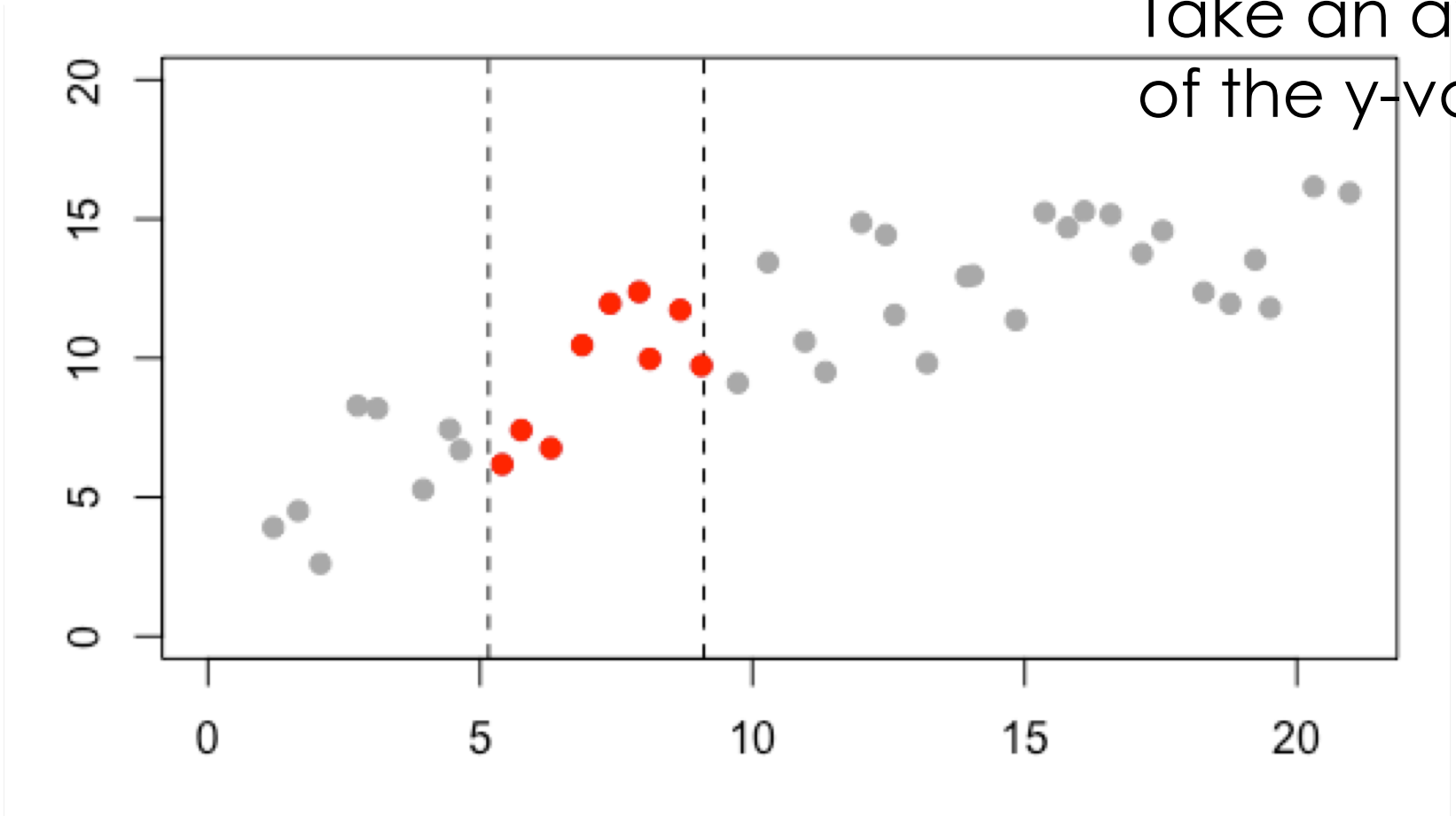
# Smoothing Scatter plots

Now we want to smooth the y-values as a function of x



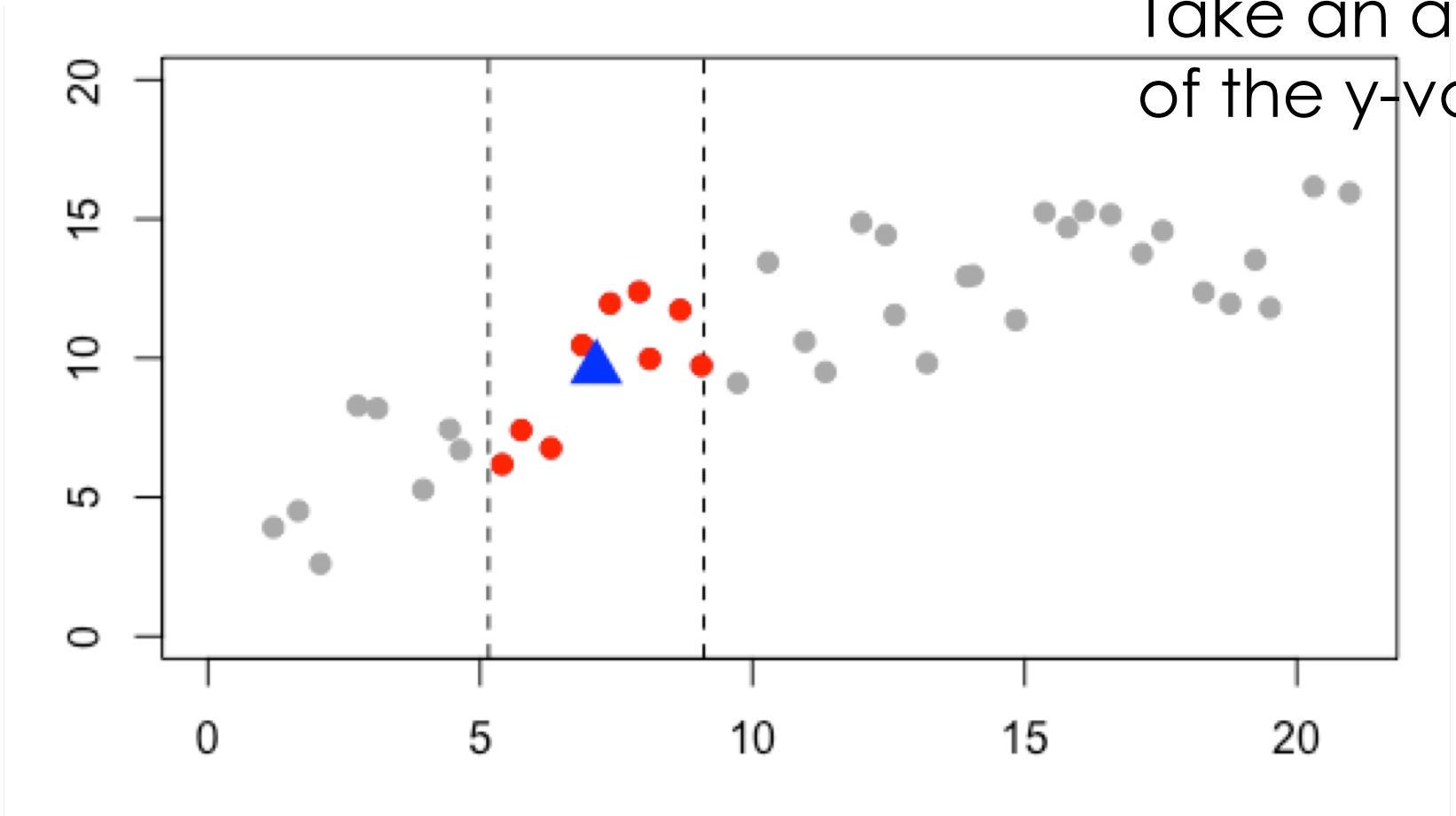
# Smoothing Scatter plots

For an x-value consider all of the x's near it
Take an average of the y-values



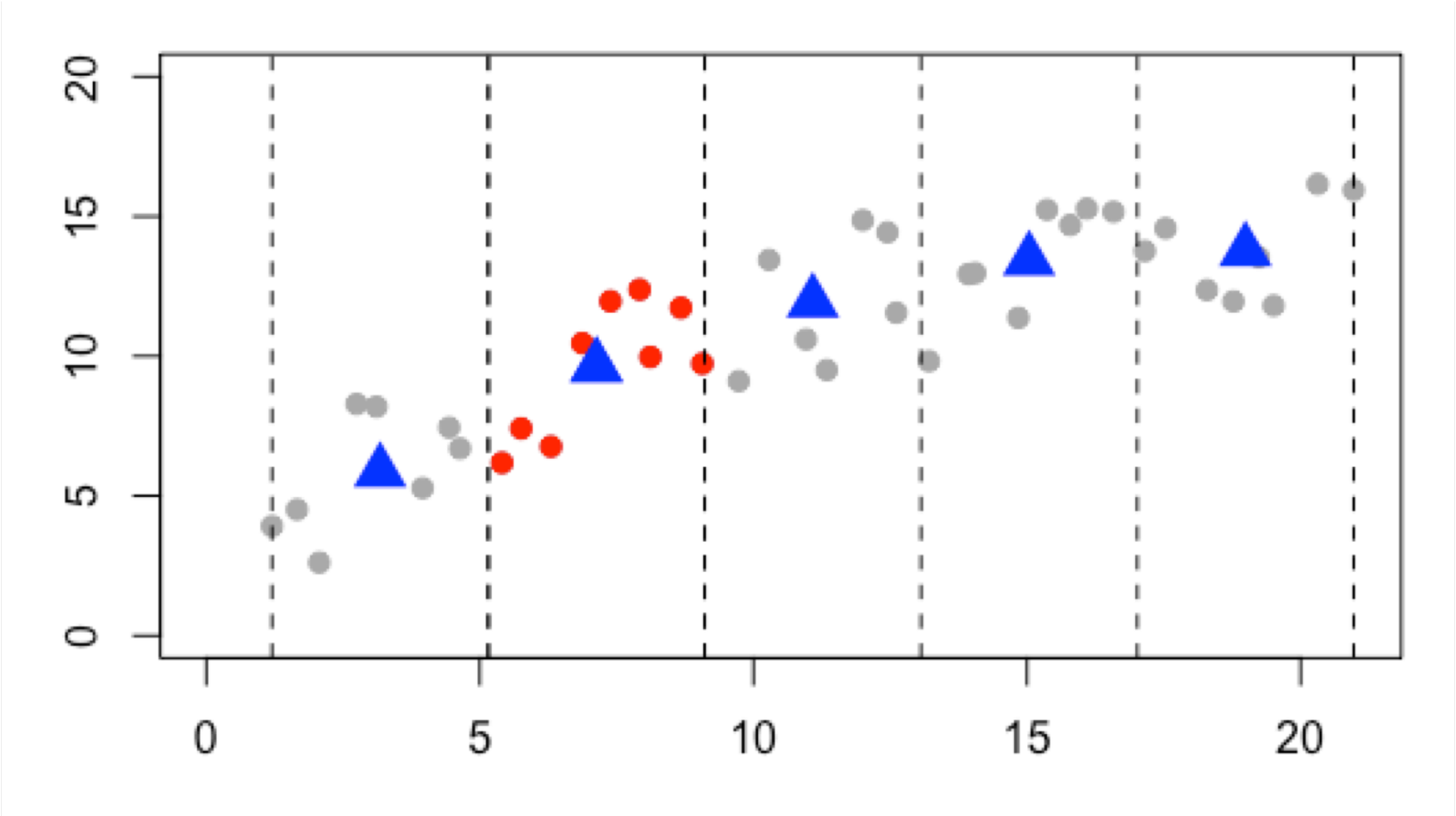
# Smoothing Scatter plots

For an x-value consider all of the x's near it
Take an average of the y-values



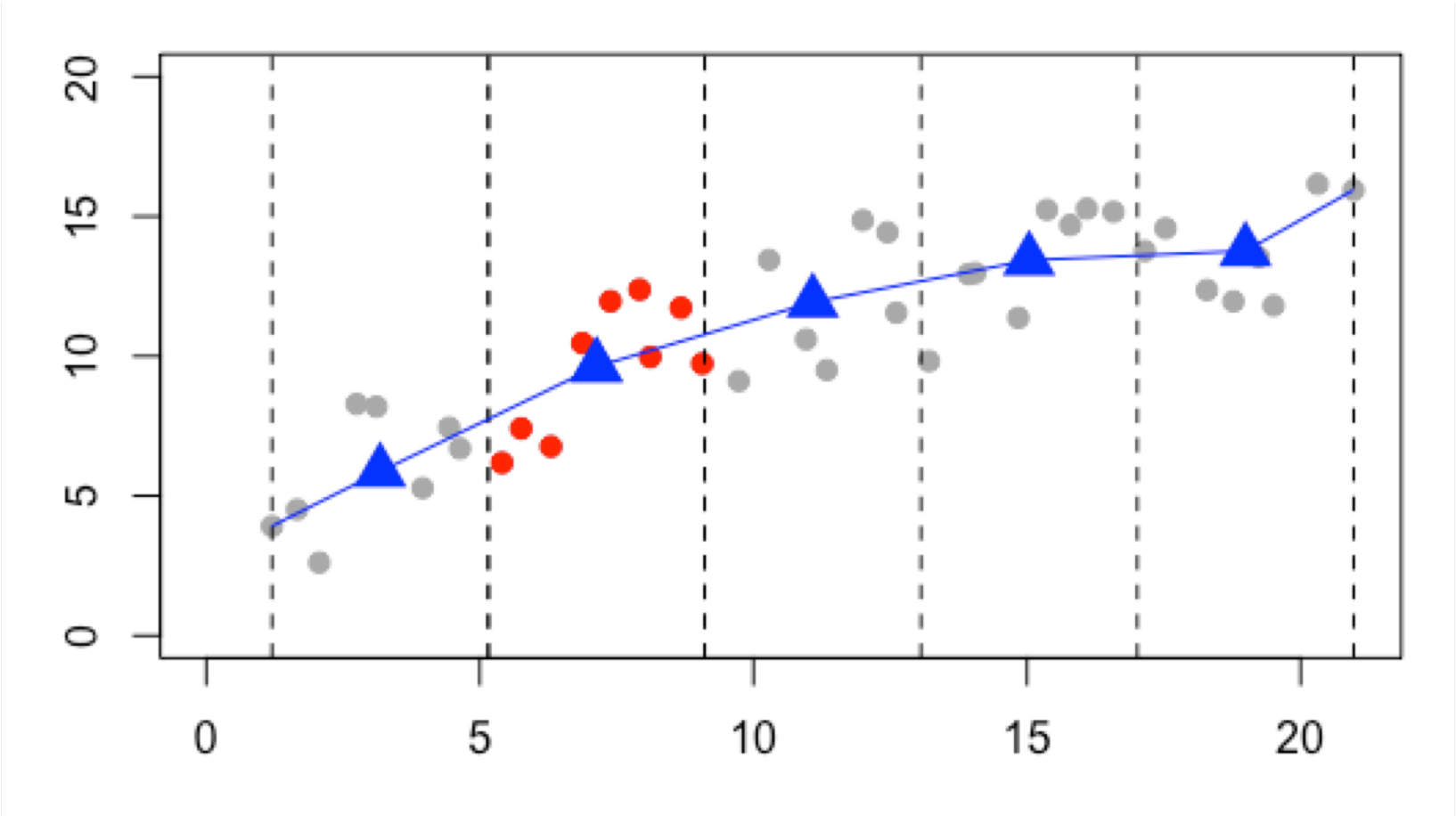
# Smoothing Scatter plots

Create bins for all x
Average y-values
in each bin

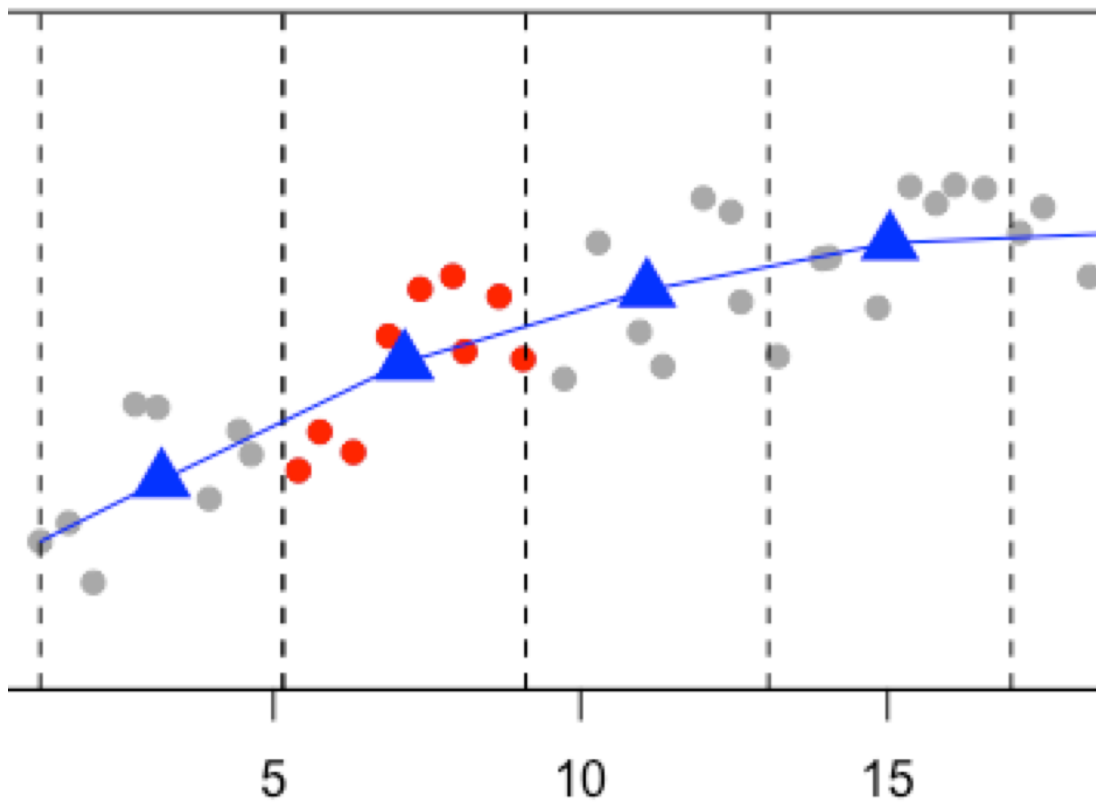


# Smoothing Scatter plots

These averages sketch out a curve



# Smoothing Scatter plots



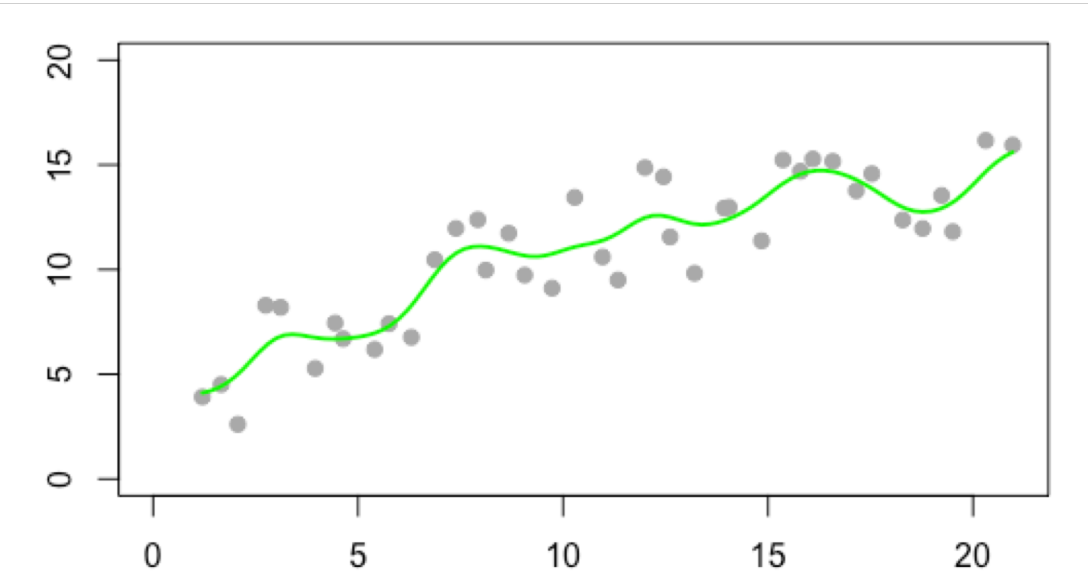

Rather than a simple average in fixed bins

We use kernels positioned on the $x_i$ to determine the weights to place on the $y_i$ in the average

$$g(x) = \sum_{i=1}^{n} \frac{K_h(x - x_i) y_i}{\sum K_h(x - x_i)}$$

The denominator ensures the weights sum to 1

# Smoothing Scatter plots




Rather than a simple average

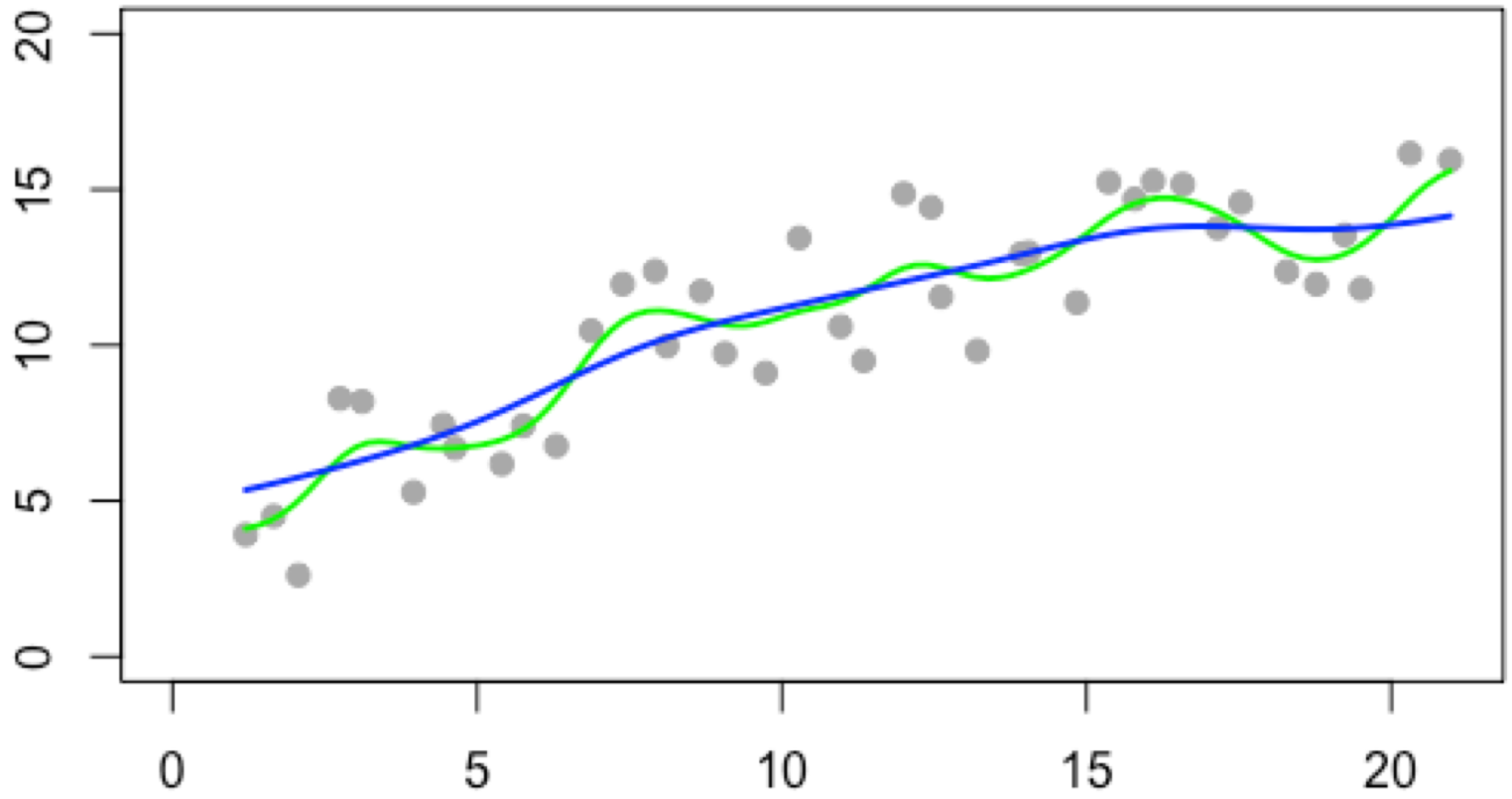
We use kernels positioned on the $x_i$ to determine the weights to place on the $y_i$ in the average

$$g(x) = \sum_{i=1}^{n} \frac{K_h(x - x_i) y_i}{\sum K_h(x - x_i)}$$

The denominator ensures the weights sum to 1

# Smoothing Scatter plots

$$g(x) = \sum_{i=1}^{n} \frac{K_h(x - x_i) y_i}{\sum K_h(x - x_i)}$$




For each x, we find g(x) by a weighted average of the $y_i$

The $y_i$ are weighted according to the kernel function.
So $x_i$ far from x do not contribute much to g(x)

# Local Smoothing

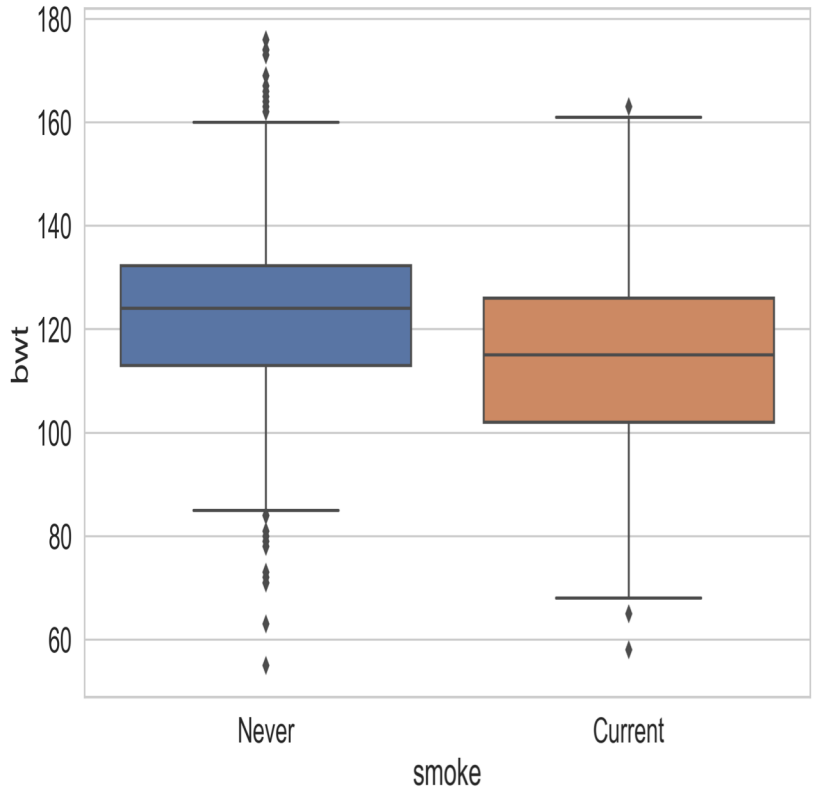
- Moving window
- Smooth/Average y values in the window
- Many different approaches for doing this:
  - kernel methods (what we just showed),
  - cubic splines, thin plate splines,
  - Locally weighted smooth scatterplot (lowess)

Allows us to see shape of the relationship between y and x

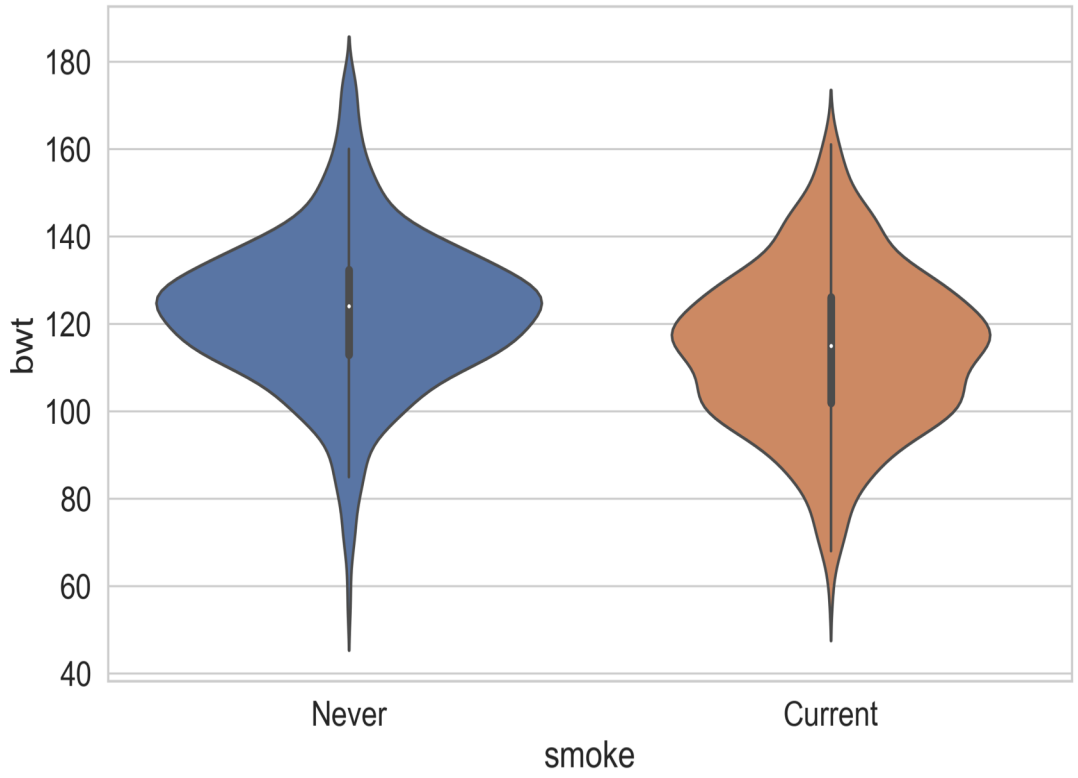
# Mix Quantitative & Qualitative

# Mix Quantitative & Qualitative

Side-by-side Boxplots

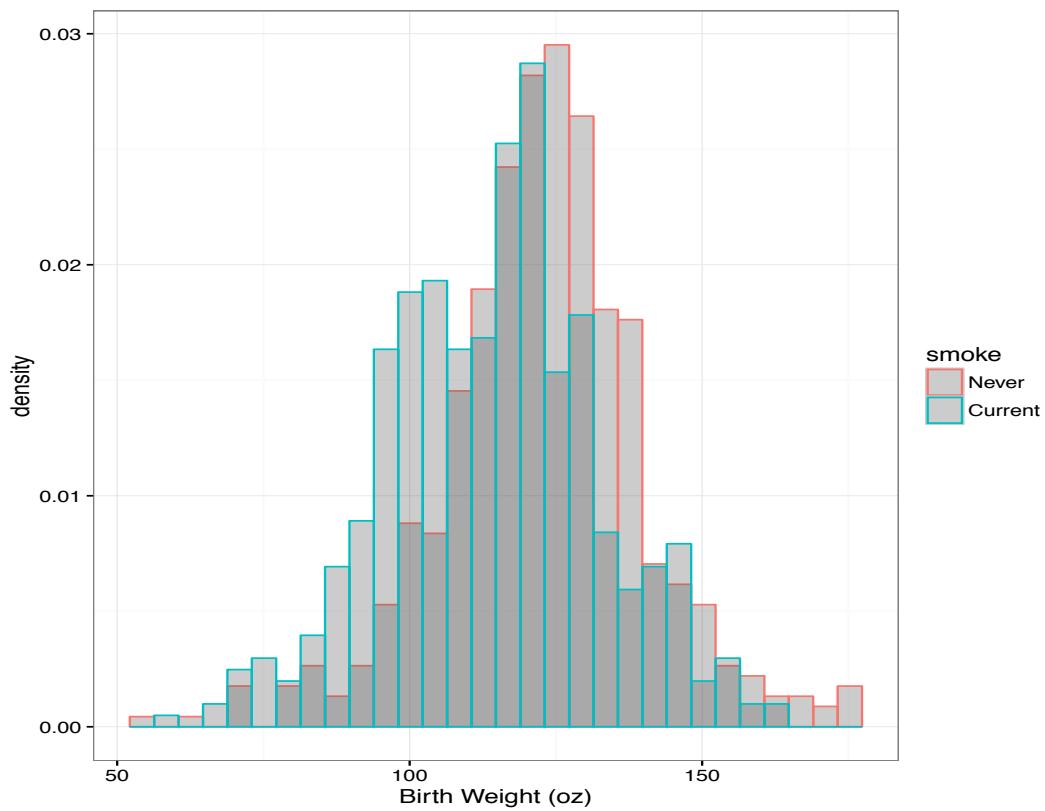


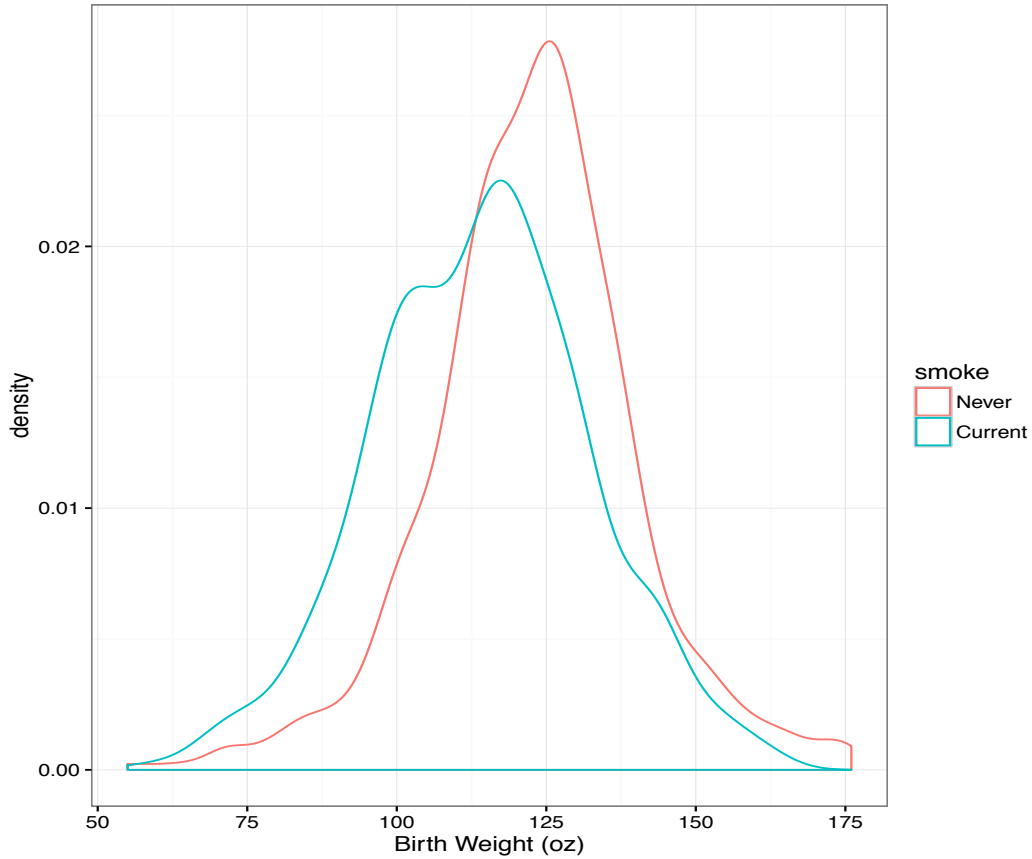

Side-by-side violin plots



# Mix of Qualitative and Quantitative

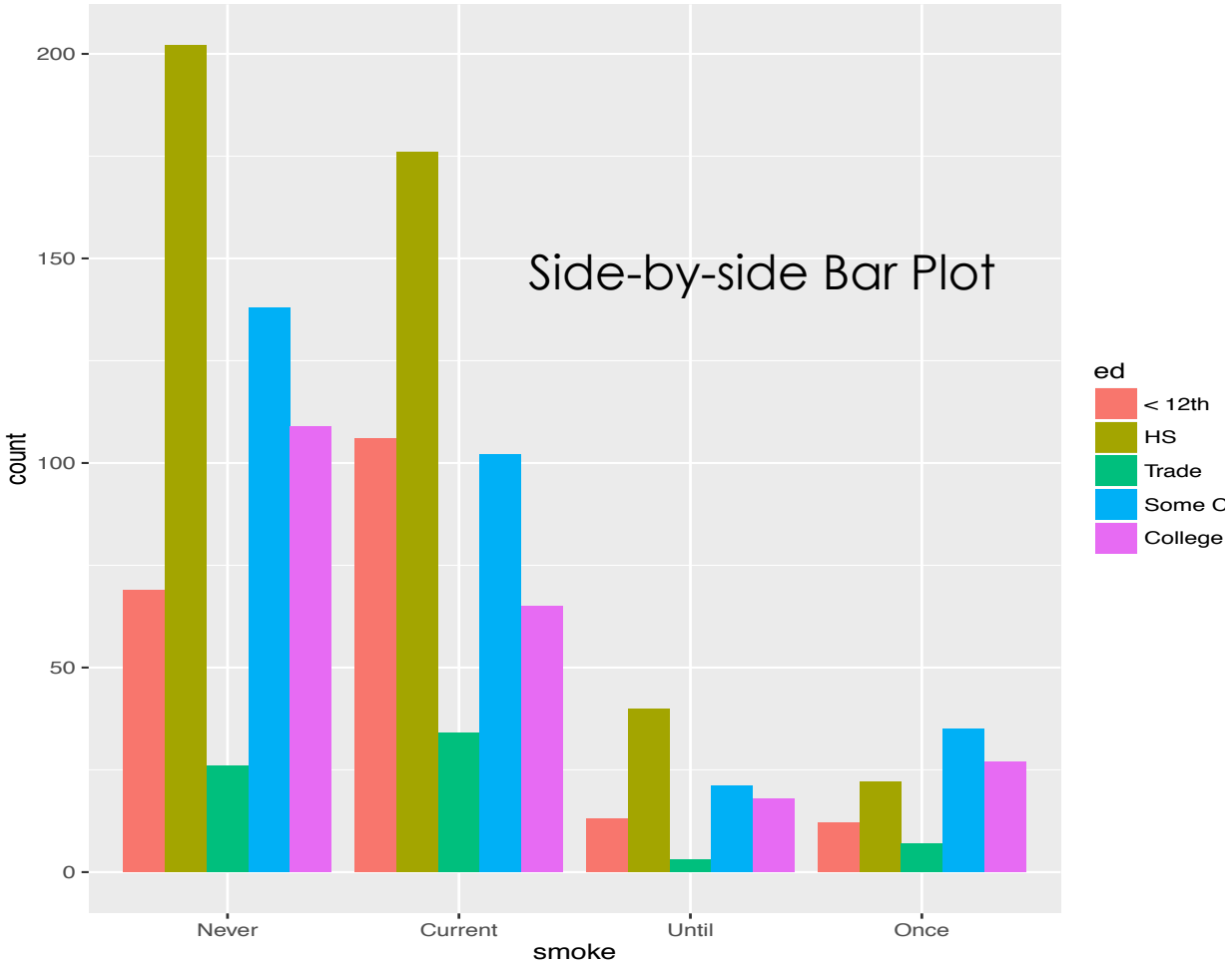
## Overlaid bars/curves

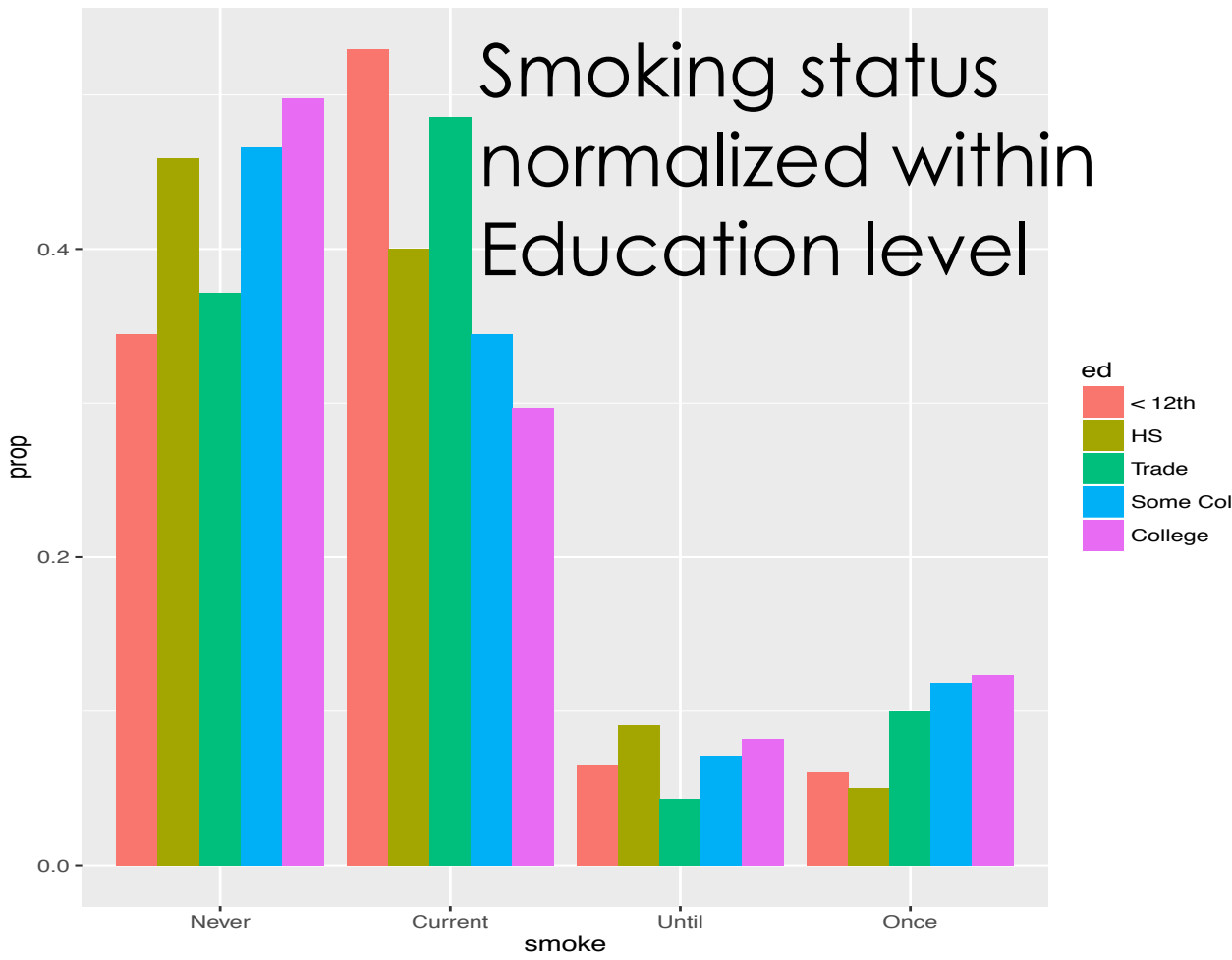


# Two Qualitative Variables

# Pairs of Qualitative Variables

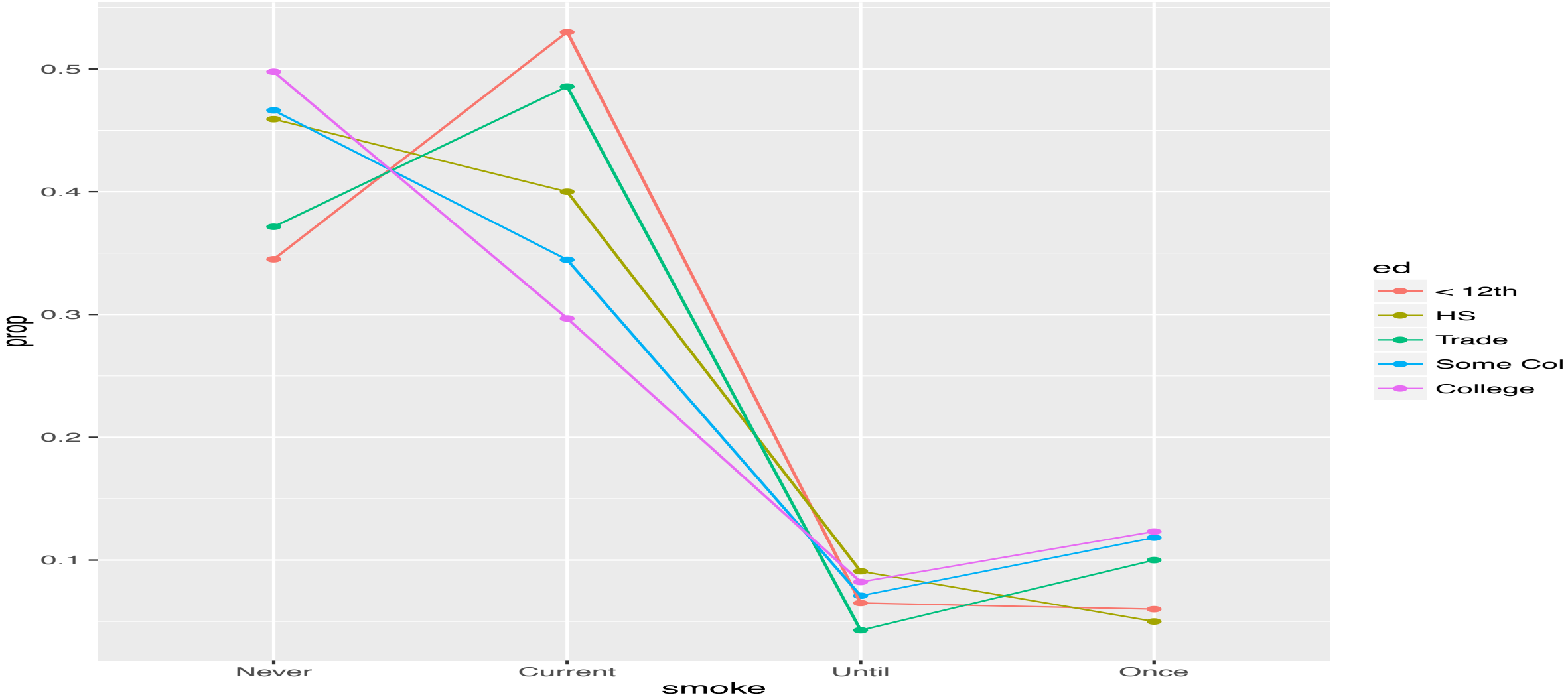



What's the difference between these 2 plots?

# Interaction/Factor Plot

Smoking status normalized within Education level



# Univariate Graphical Displays

| Type | Plot |
| --- | --- |
| Numeric – | few observations<br>Histogram, Density curve<br>Box plot, Violin plot<br>Normal quantile plot<br>Few Observations - Rug plot, Dot plot<br>Caution if discrete: density curves and box plots may be misleading |
| Categorical –<br>Counts of categories | Dot chart<br>Bar chart<br>Pie chart (avoid!)<br>Caution if ordinal –order of bars, dots, etc. should reflect category order |

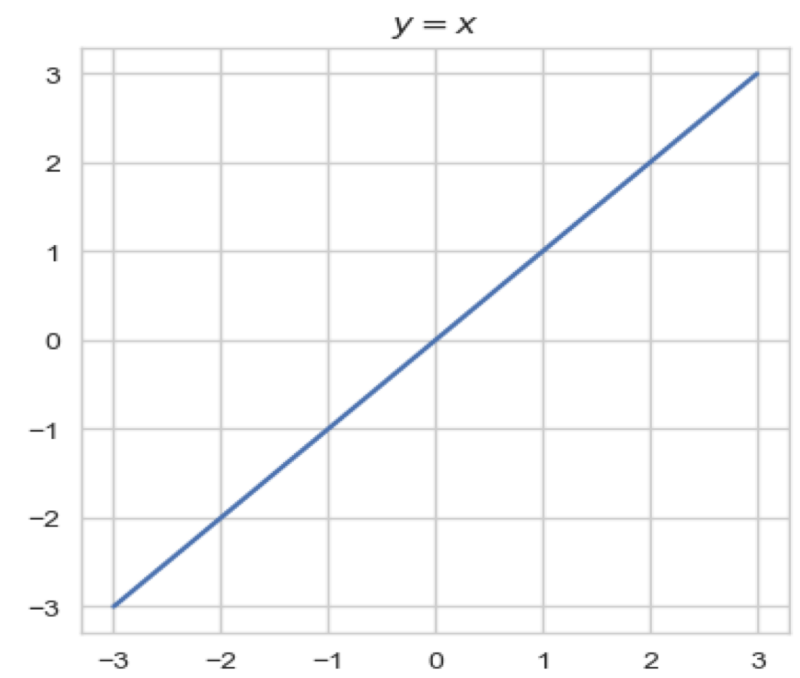
# Bivariate Graphical Displays

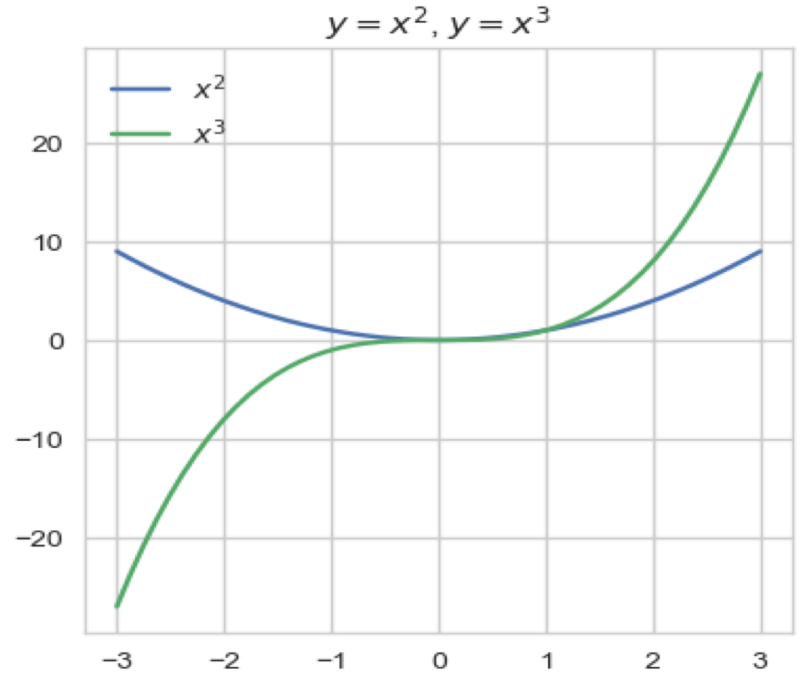
| | Numeric | Categorical |
|---|---|---|
| Numeric | Scatter plot<br>Smooth scatter<br>Contour plot<br>Smooth lines and curves | Multiple histograms,<br>density curves,<br>Avoid jiggling! |
| Categorical | | Side-by-side bar plot<br>Overlaid Lines plot<br>Side-by-side dot chart<br>Mosaic plot<br>Avoid stacking! |

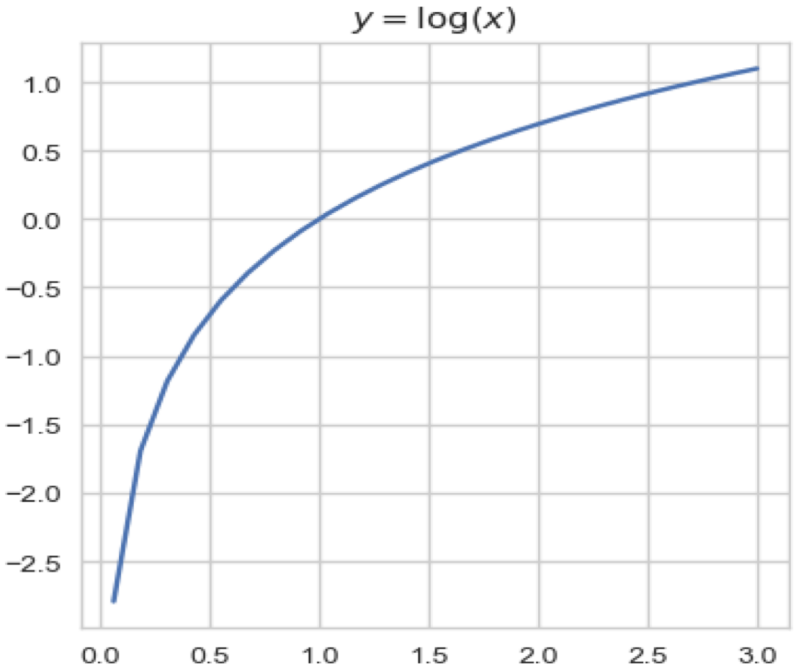
# Transformations

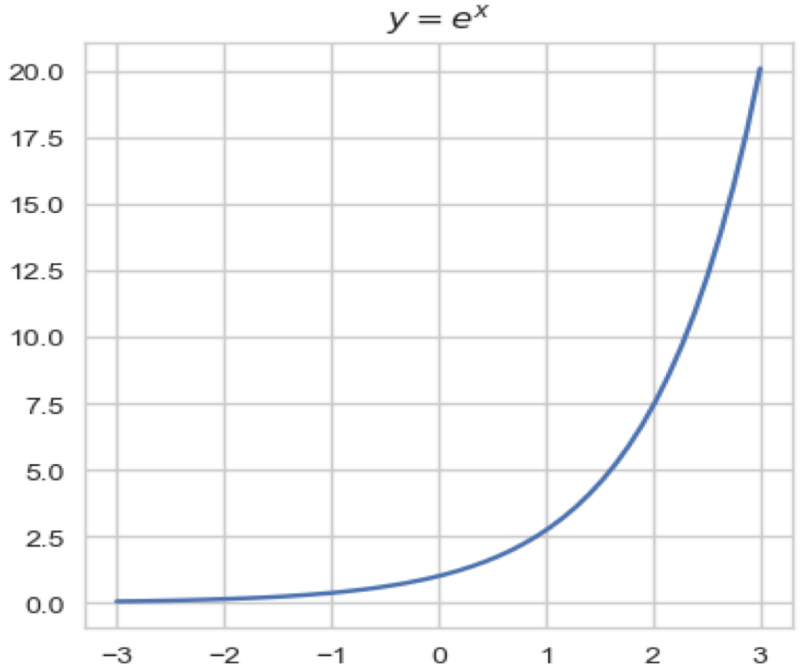
Scatter plots

# Basic Functional Relations



# Log transformation: Swiss army knife

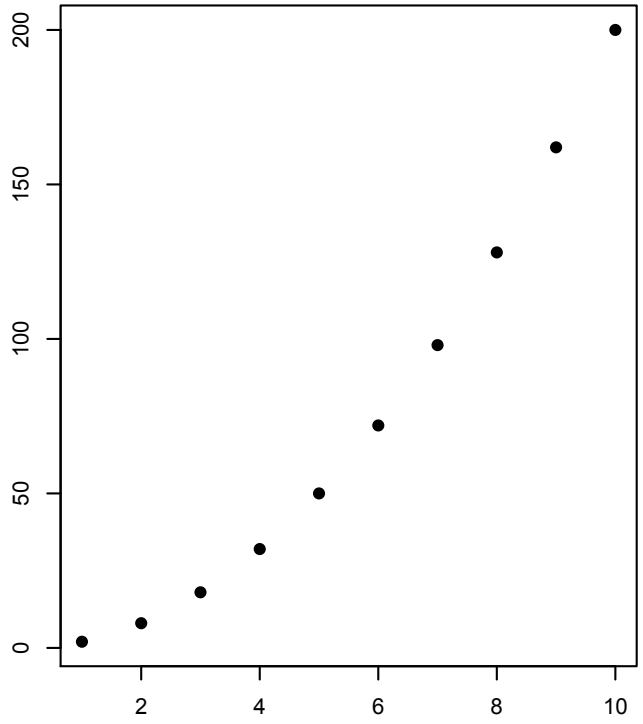
$$y = a^x \rightarrow \log(y) = x \log(a)$$

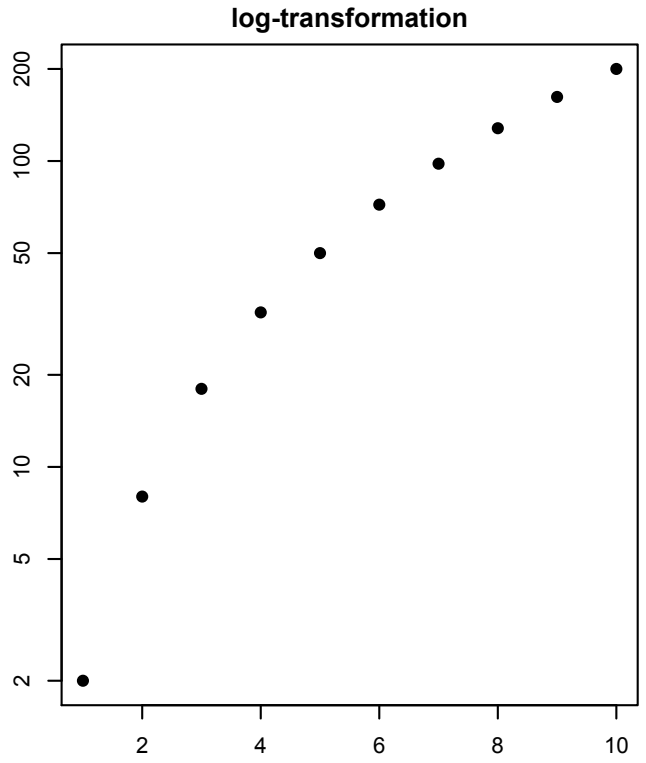
$$y = ax^k \rightarrow \log(y) = \log(a) + k \log(x)$$

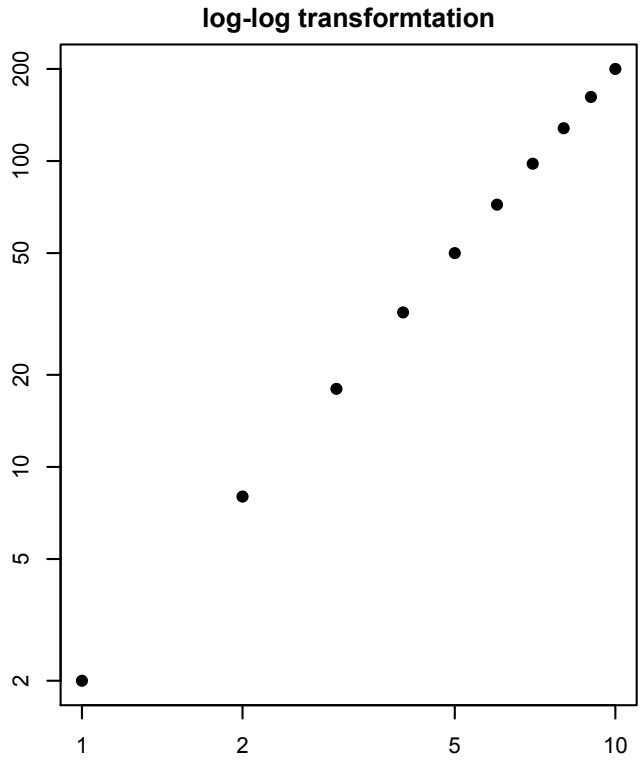
# Why Straighten Relationships?

- Easier to uncover the form of the relationship if we can transform it to linear relationship; we see what transformation used to make it linear
- Linear relationships are particularly simple to interpret & fit
- Choose a transformation that's simple and easily interpreted in the context of the problem, e.g., a power of 2, 3, ½, 0 (log), -1

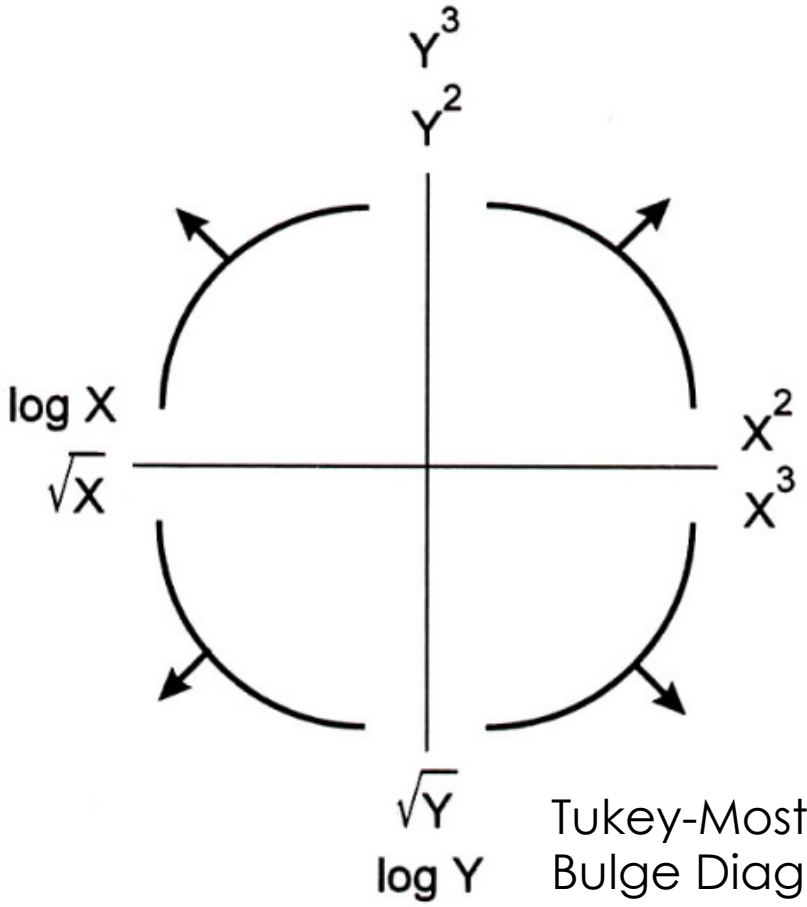
# Straighten Relationships with Transformations



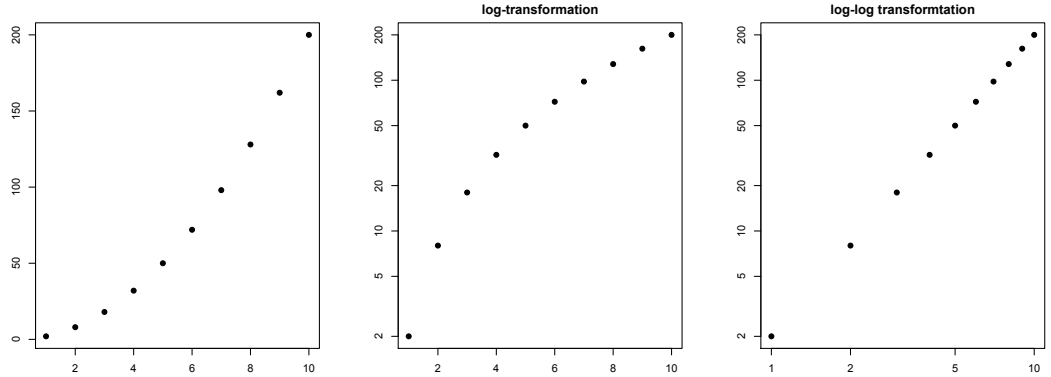



# Selecting a Transformation




Tukey-Mosteller
Bulge Diagram



# Transformation of one variable

Log transformation tends to be effective when max/min > 5

Ratio of hinges can help select a transformation

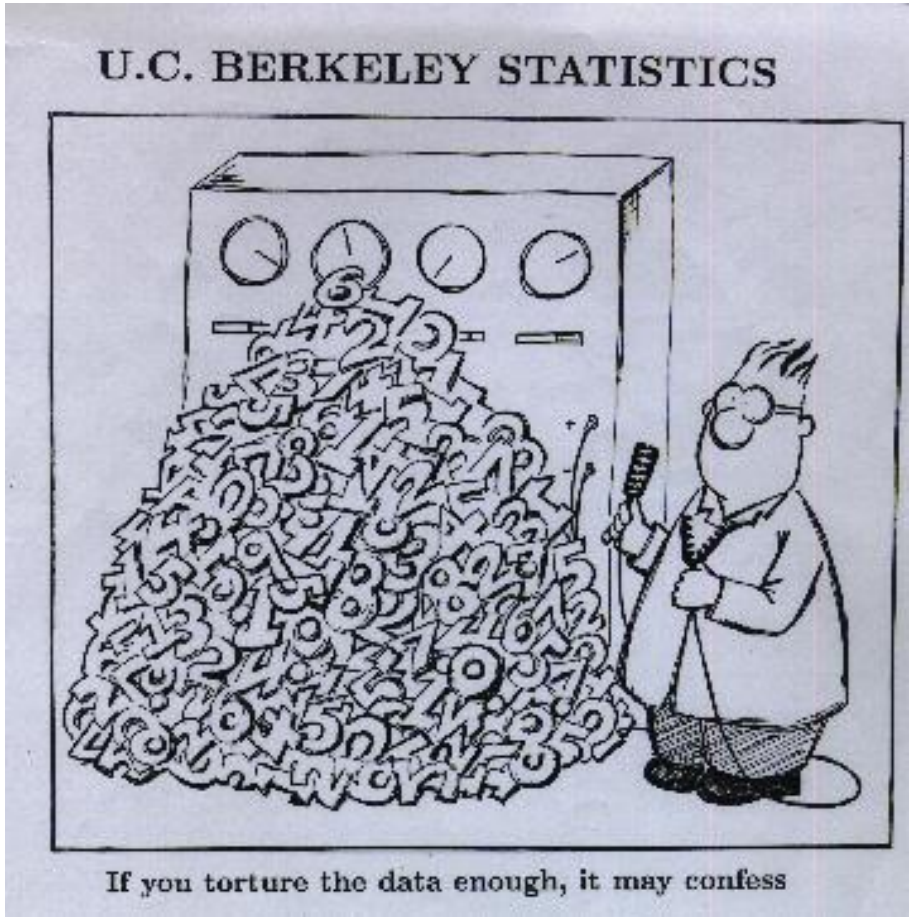
$$\frac{Upper\ Quartile - Median}{Median - Lower\ Quartile} \approx 1$$

When the distribution is symmetric

# Caution about EDA

With enough data, if you look hard enough you will find something ***"interesting"***

Important to differentiate **inferential conclusions** about world from **exploratory analysis of data**



# Take care with EDA

- EDA can provide valuable insights about the data and data collection process

BUT

- Be cautious about drawing/reporting conclusions
  - Recognize that EDA biases your view
  - Be careful about sharing plots or hypothesis without additional validation …
- Have a lot of data? Apply EDA to sample of the data before conducting formal analysis.