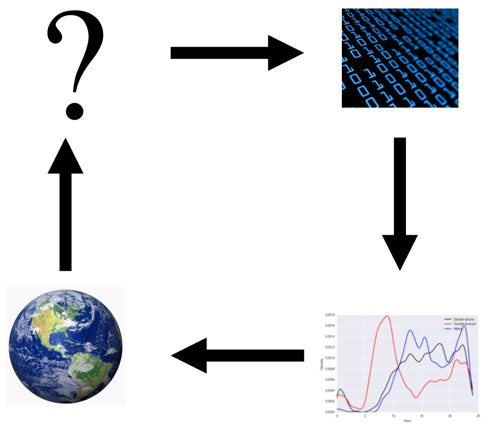


Data 100

Lecture 7: EDA & *Visualization*

Exploratory Data Analysis (EDA)

"Getting to know the data"

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

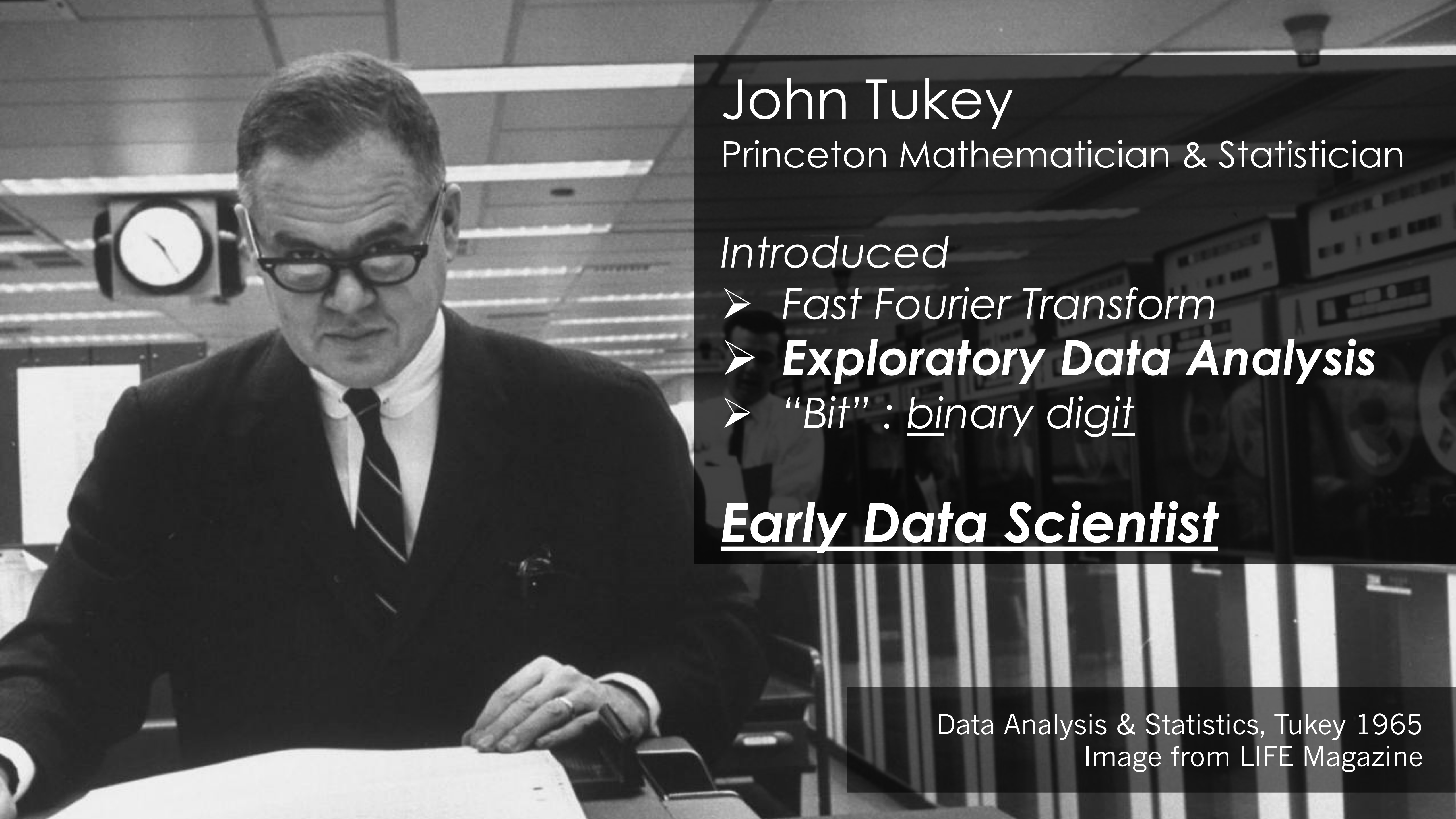
- Build/confirm understanding of the data
- Identify and address potential issues in data
- Inform the subsequent analysis
- Discover *potential* relationships

- **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

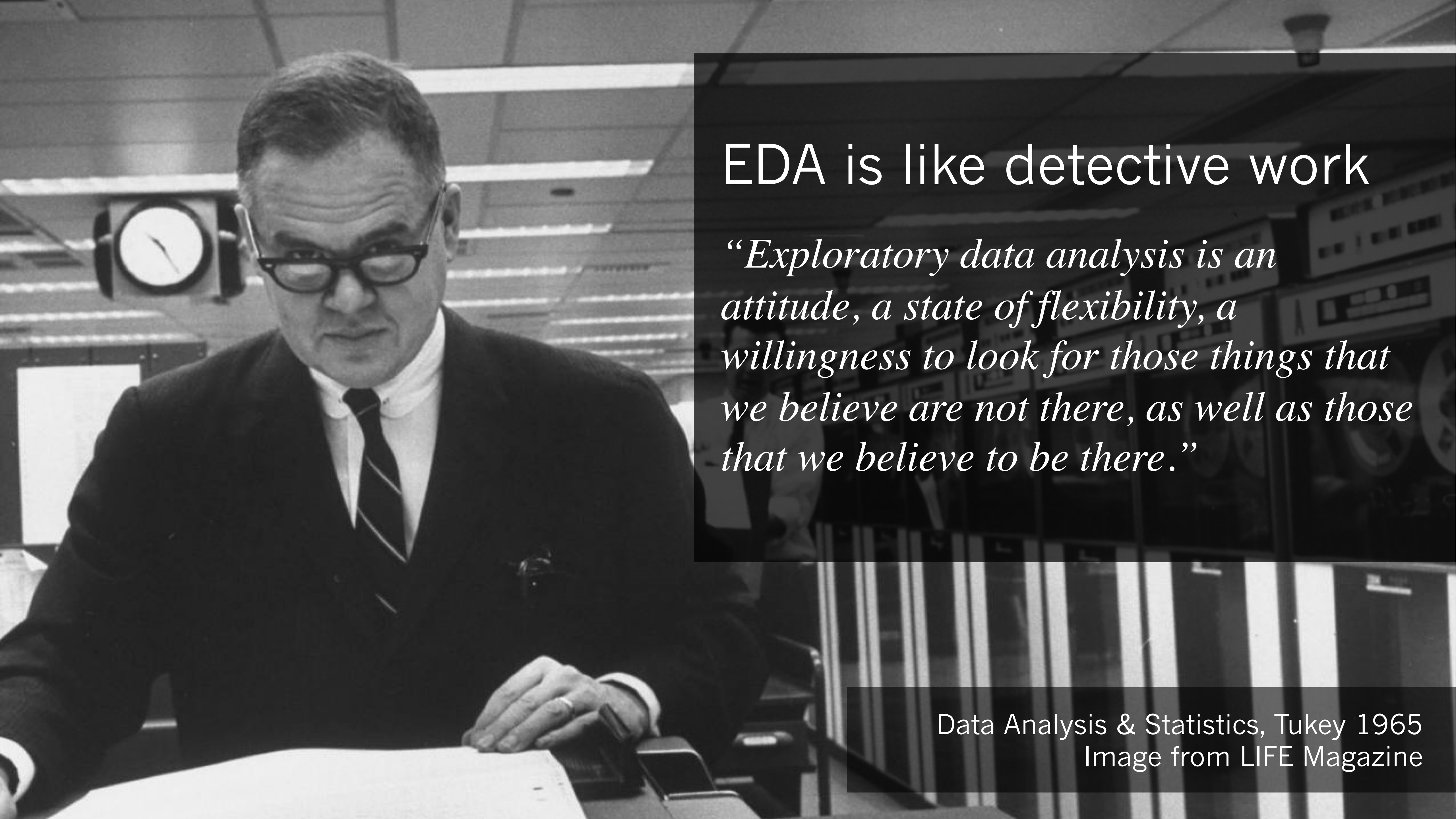
Princeton Mathematician & Statistician

Introduced

- *Fast Fourier Transform*
- ***Exploratory Data Analysis***
- *"Bit" : <u>bi</u>nary dig<u>it</u>*

<u>Early Data Scientist</u>

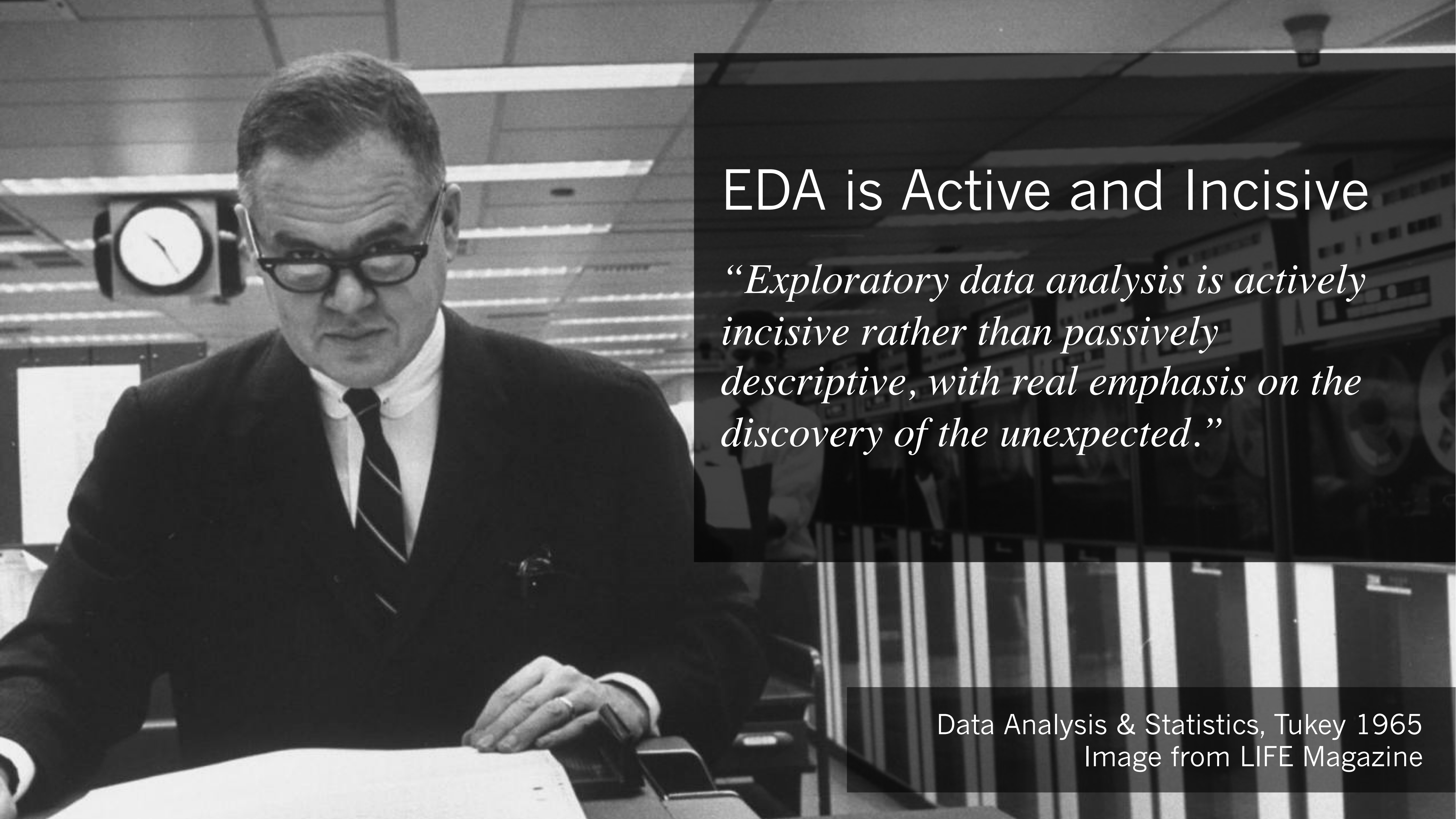
Data Analysis & Statistics, Tukey 1965
Image from LIFE Magazine



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

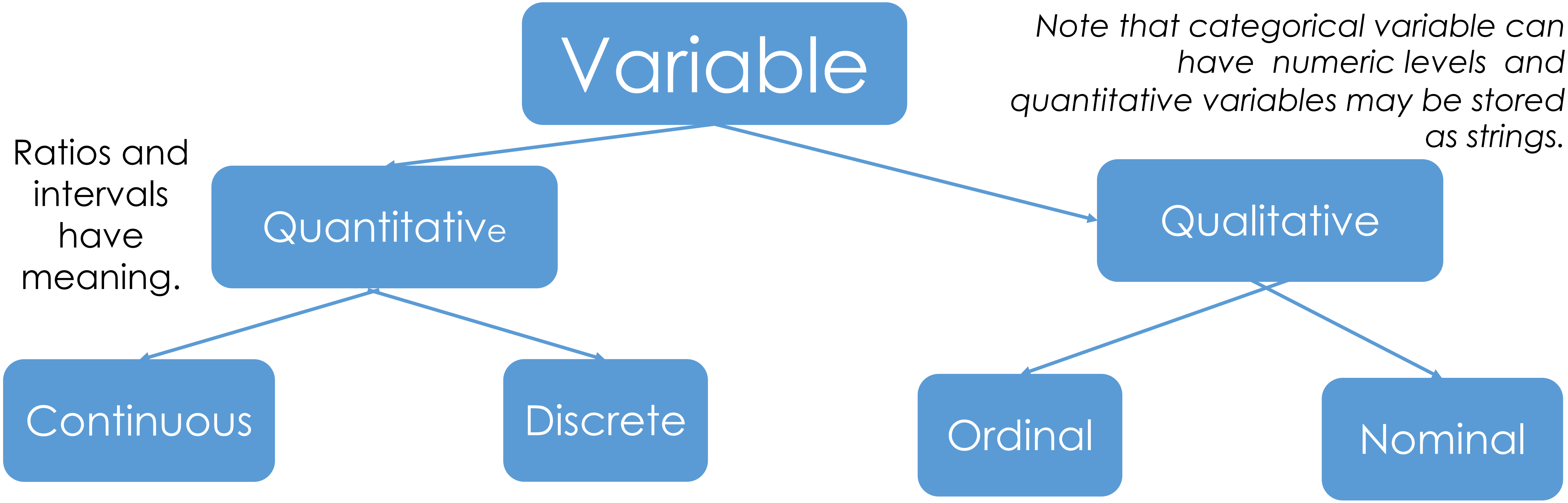
Urban Dictionary:

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.

Continuous: Could be measured to arbitrary precision.

Examples:
- Price
- Temperature

Discrete: Finite possible values

Examples:
- Number of siblings
- Yrs of education

Ordinal: Categories w/ levels but no consistent meaning to difference

Examples:
- Preferences
- Level of education

Nominal: Categories w/ no specific ordering.

Examples:
- Political Affiliation
- CalD number

	Quantitative Continuous	Quantitative Discrete	Qualitative Nominal	Qualitive Ordinal
CO_2 level	X			
Number of siblings		X		
GPA	X			
Income bracket				X
Race			X	
Number of years of education		X		
Yelp Rating				X
Lane of traffic (left, middle, right)			X	
Letter GRADE in 100				X

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

Qualitative

Qualitative Ordinal

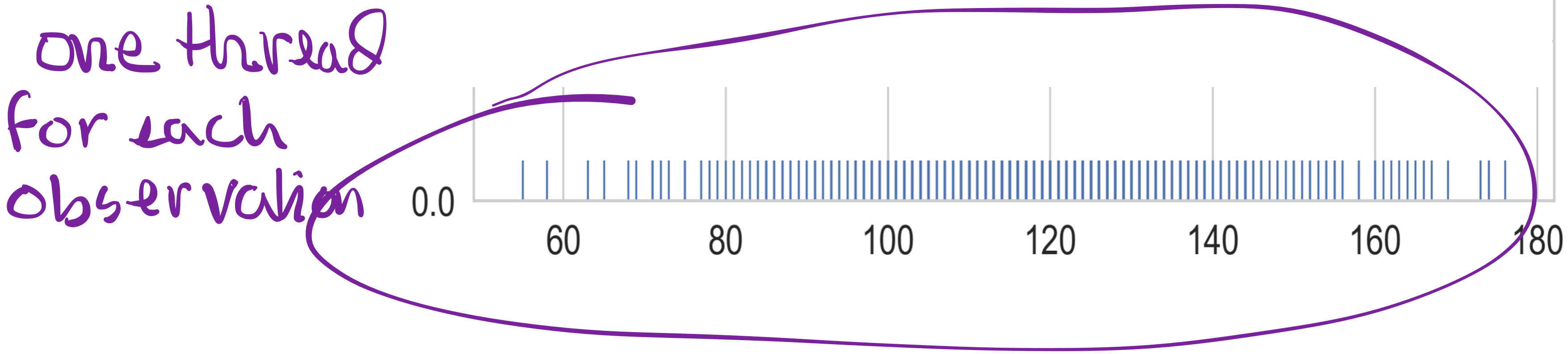
One Variable

What is the Distribution of the values of the variable?

Quantitative – Continuous

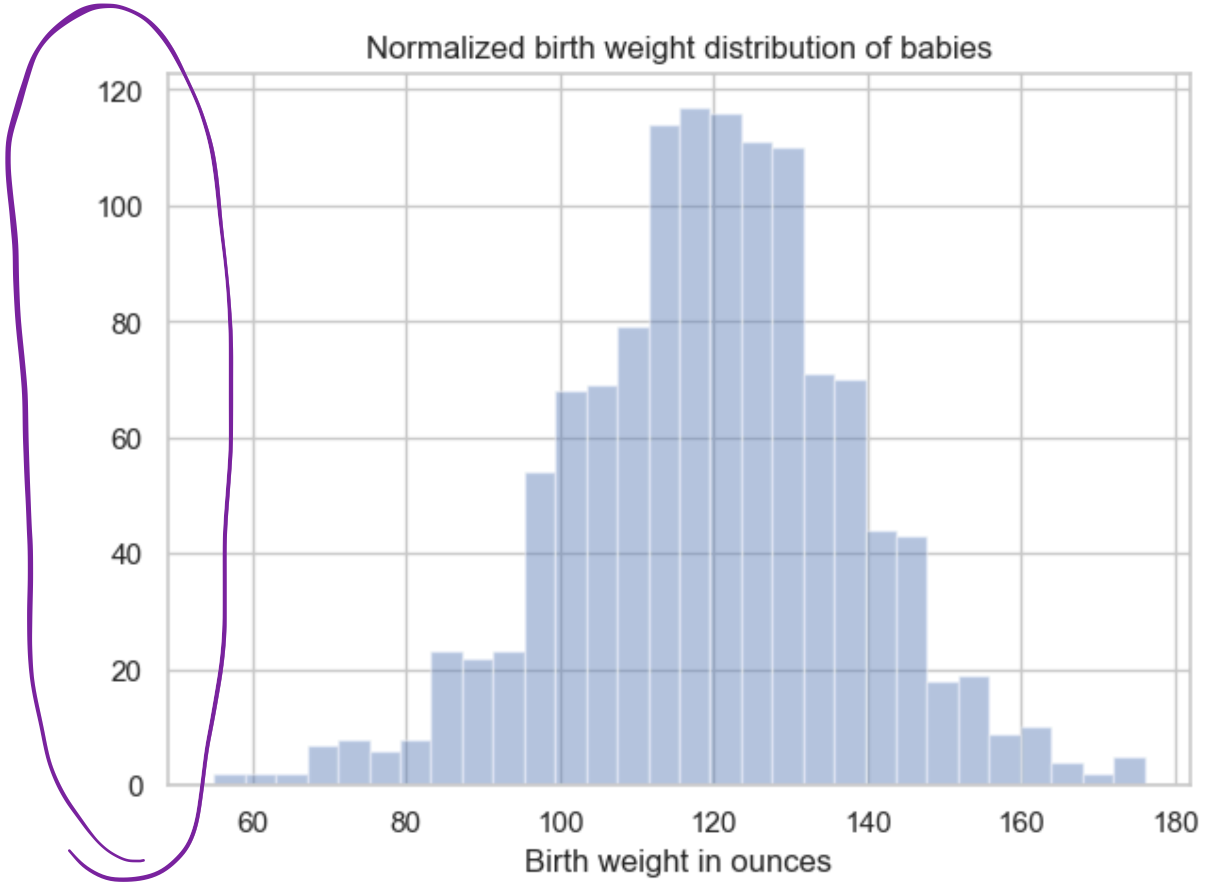
- Birthweight
- The most basic visual representation of one quantitative variable is the *rug plot*

Hard to see much of the distribution with this rug plot



Birthweight

Histogram

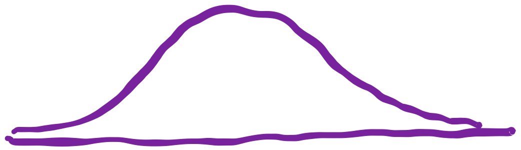


With the histogram we hide the details of individual observations and view the general features of the distribution.

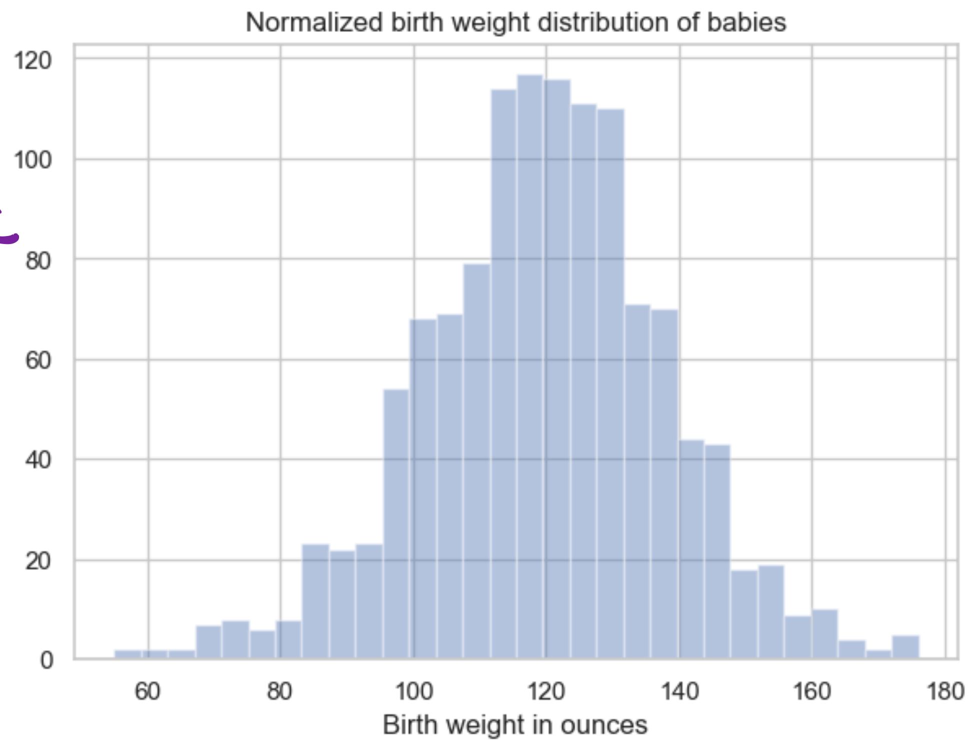
How would we describe the distribution of birth weight?

Distribution Features

Normal Tails



- Modes
  - Number 1
  - Location near 120 oz
  - Size main mode
- Symmetry very symmetric
  - Symmetric
  - Skewed left or right slight left skew
- Tails
  - Long, short, "normal"
- Gaps
- Outliers

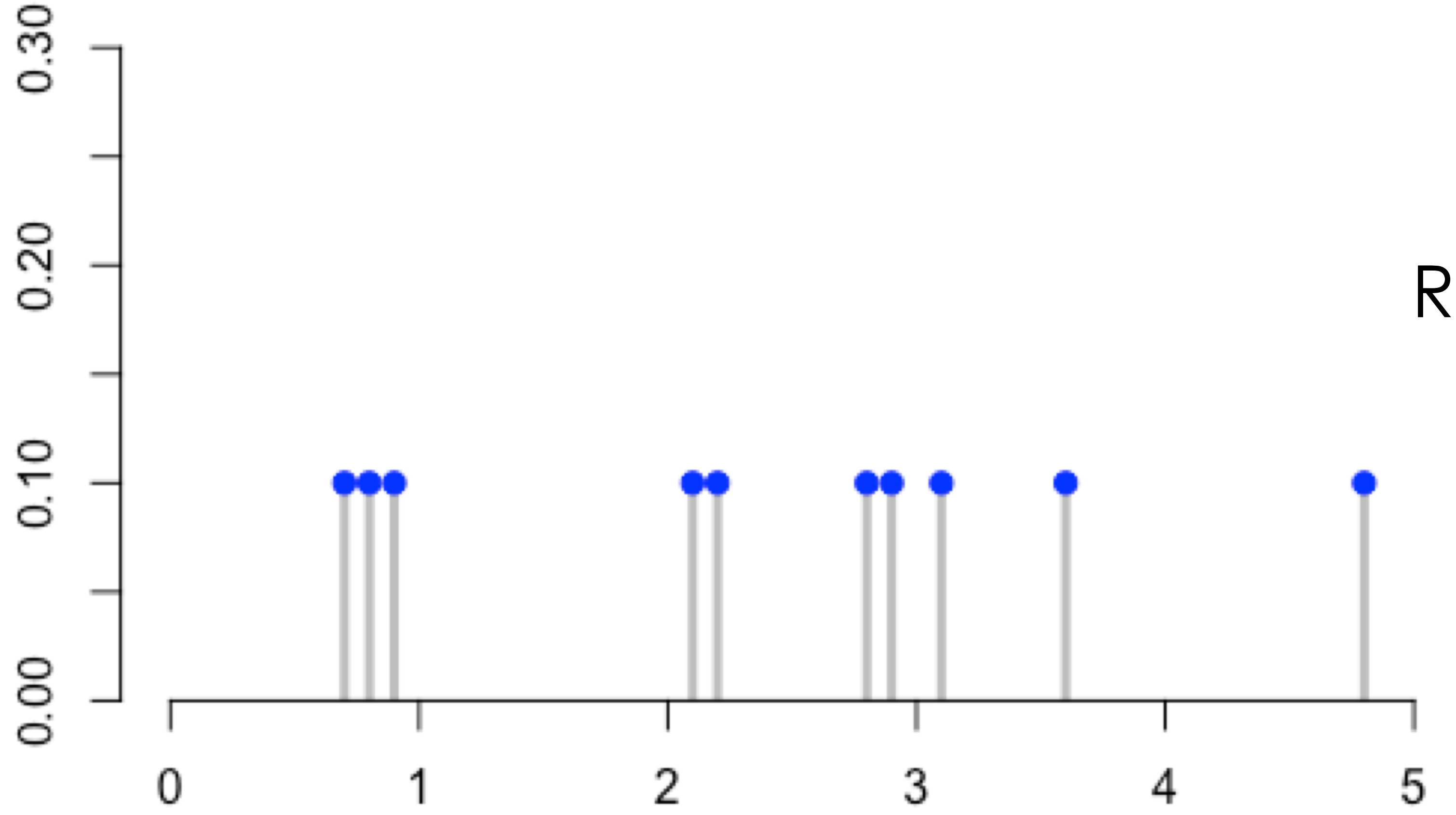


Distributions & Smoothing

A Small Dataset

10 values

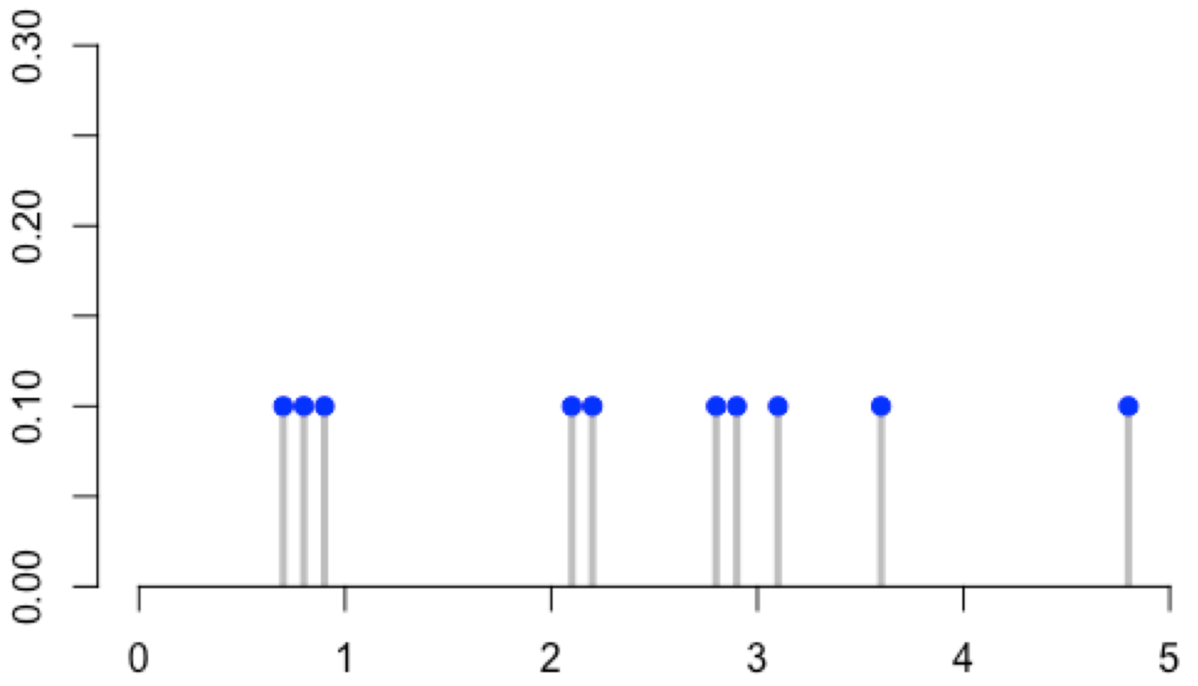
0.7, 0.8, 0.9, 2.1, 2.2, 2.8, 2.9, 3.1, 3.6, 4.8



Rug Plot

shows the location of each value

We want to smooth these rug threads



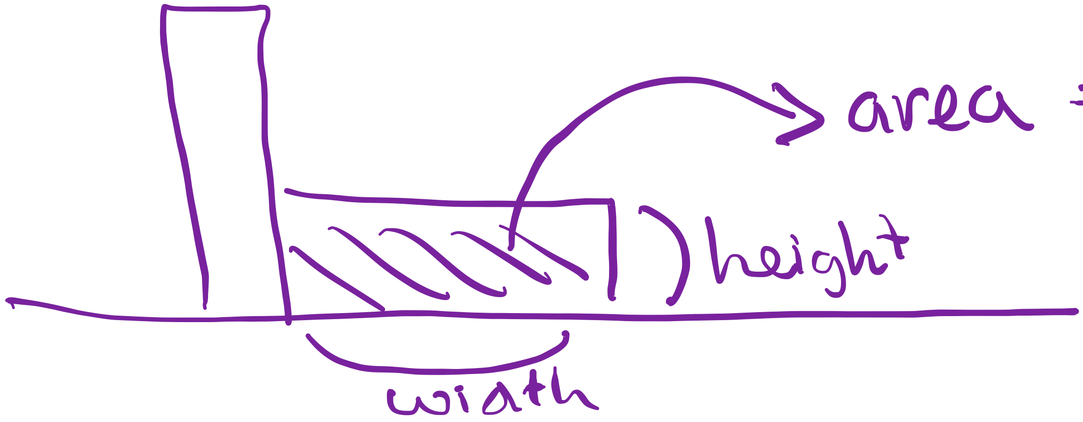
BECAUSE

- 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: x-units km   y-units %/km



area = x km wide × y %/km high

= x × y %

Area in %

cancel km units

Example - One large bin from 0 to 5

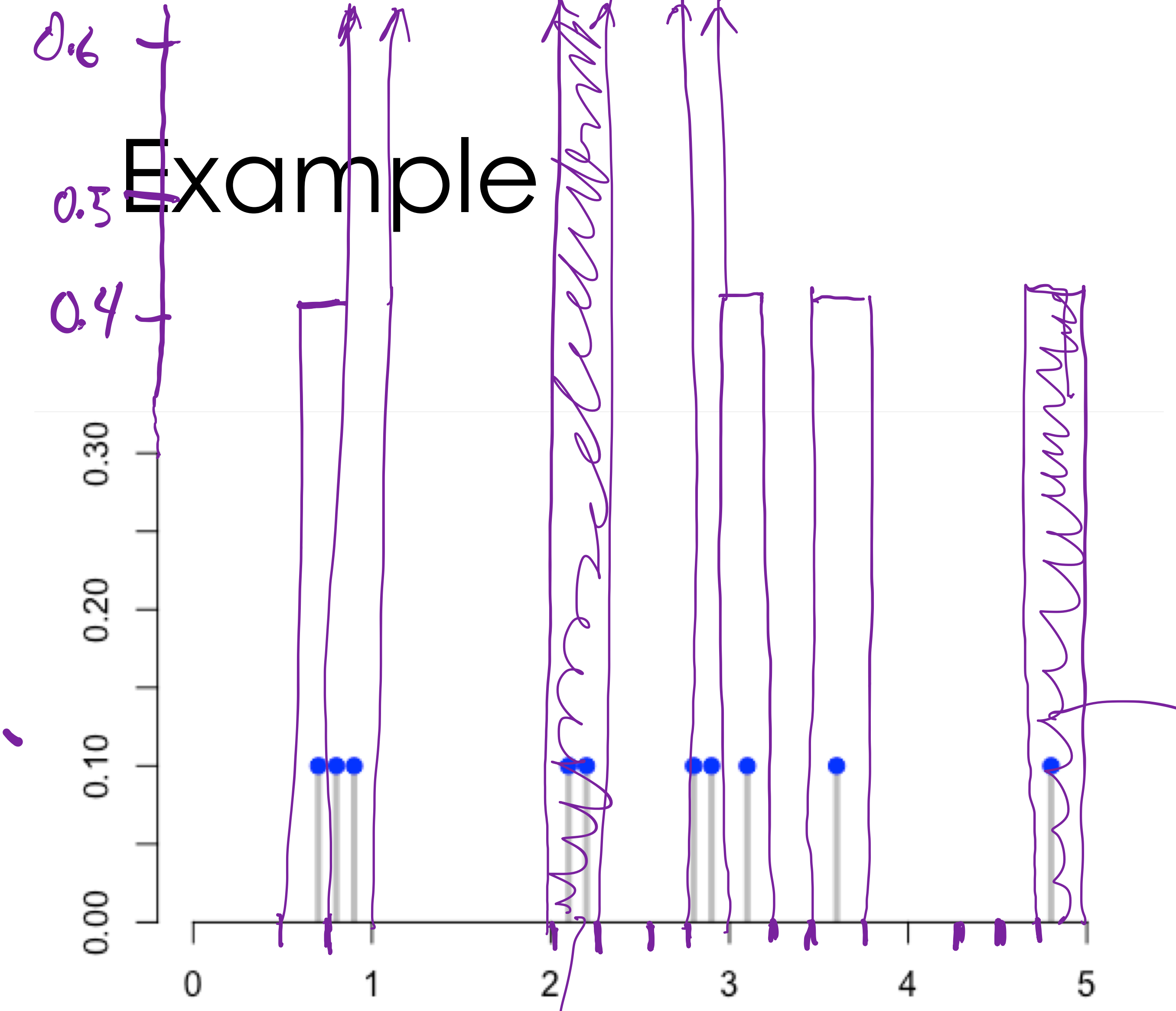
one 'mode'
no/short tails

The 10 points are spread evenly across one large bin

→ Area = 5 km × 0.2 frac/km
= 1 fraction
all observations

Not Very Informative Distribution

Example



Bin width 1/4 km

With these narrow bins, the histogram is little more than a rug plot

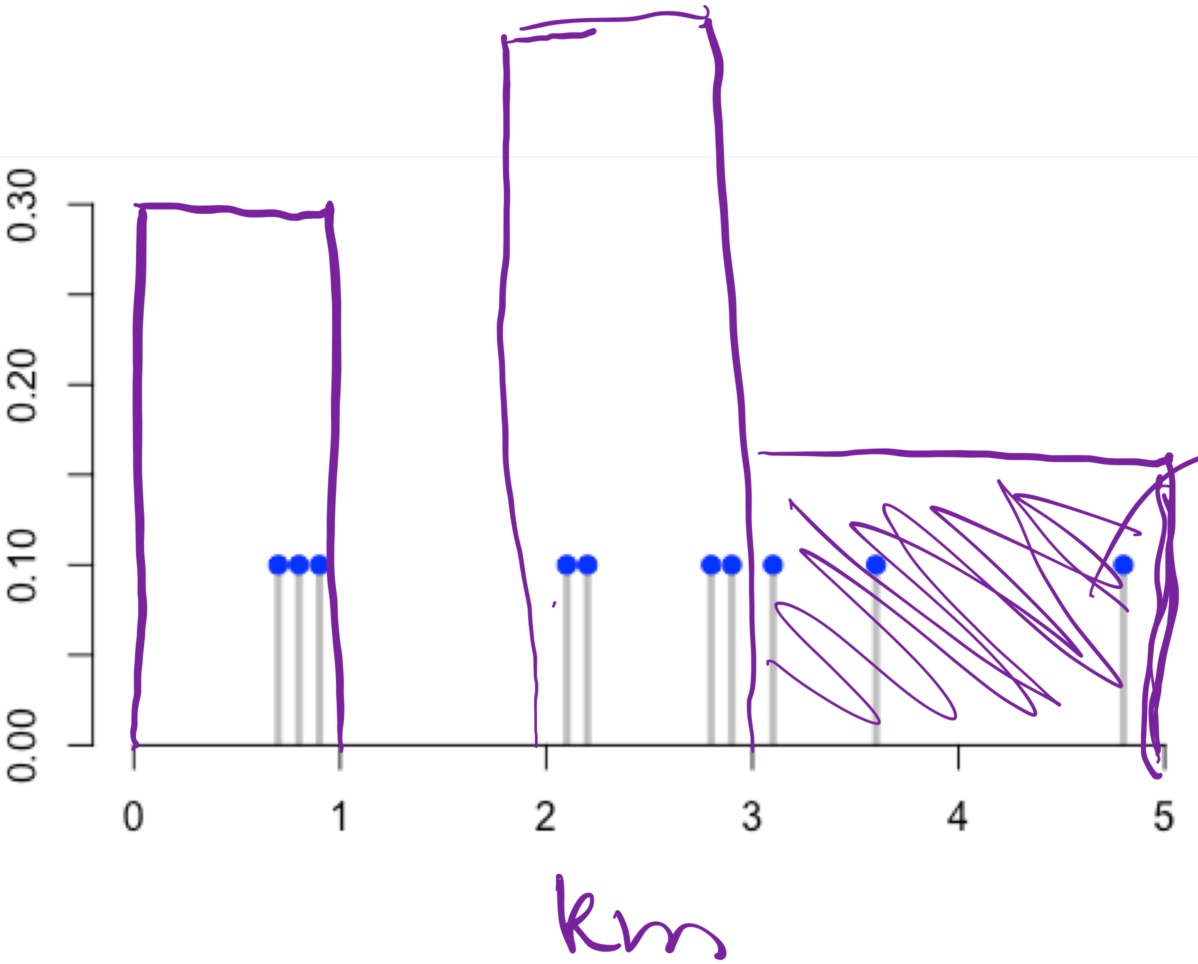
Area = $\frac{1}{4}$ km × 0.4 frac/km

= 0.1 fraction of sample

Area = $\frac{1}{4}$ × 0.8 = 0.2 (2 observations)

Example

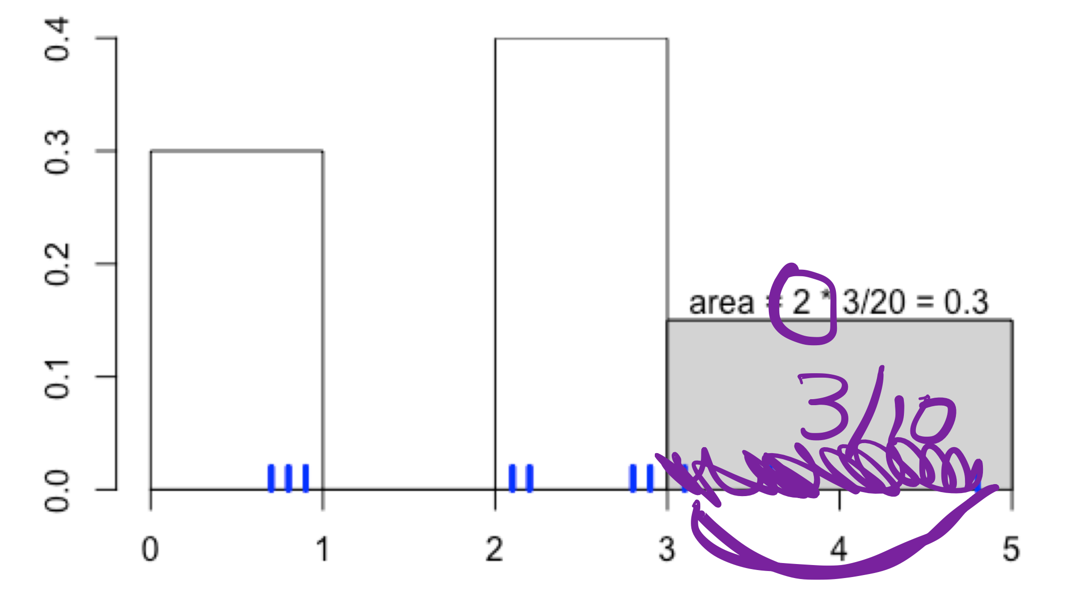
Bins can be different widths



Area = 2 km × 0.15/km

= 0.3

A histogram smooths

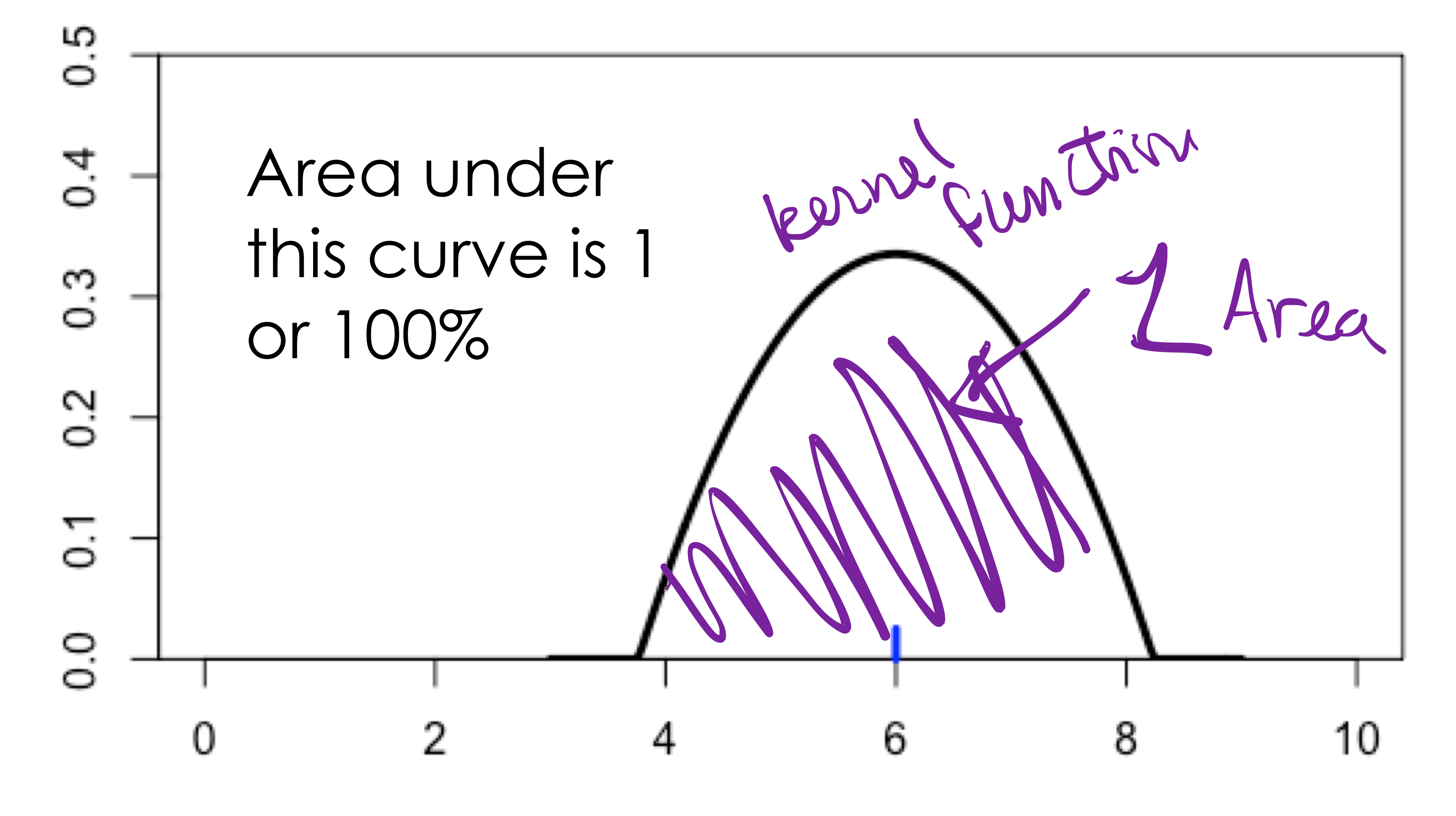


We want to smooth out these points because:

- 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

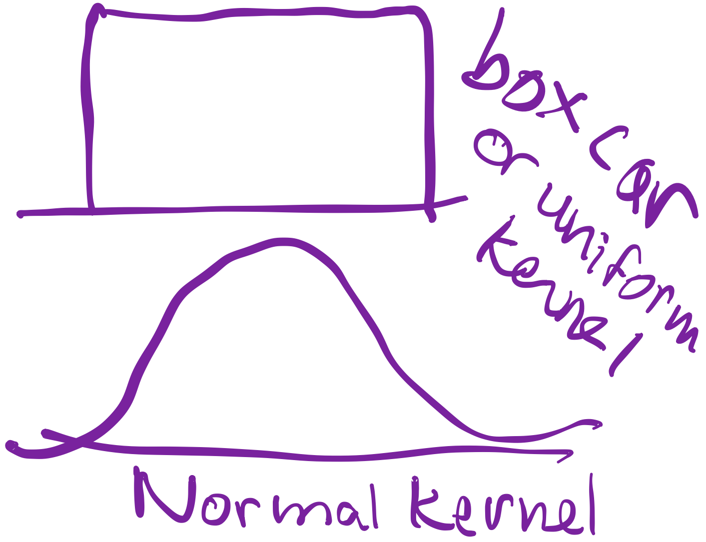
The values 3.1, 3.6, and 4.8 have their proportion (3/10) spread over the bin [3,5] That is, without the rug, we can't tell where the points are in the bin

Kernel Density Estimate: Alternative Smoother

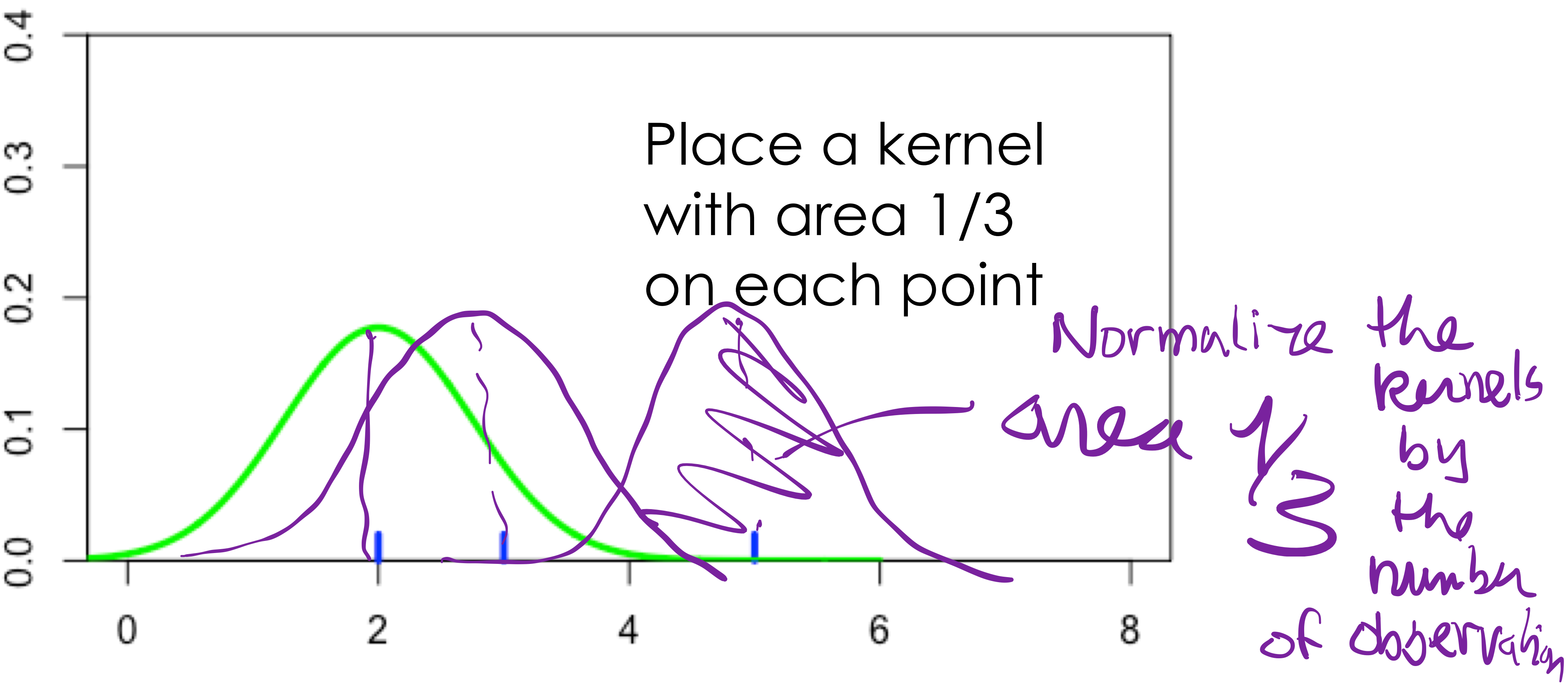


Consider one point

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



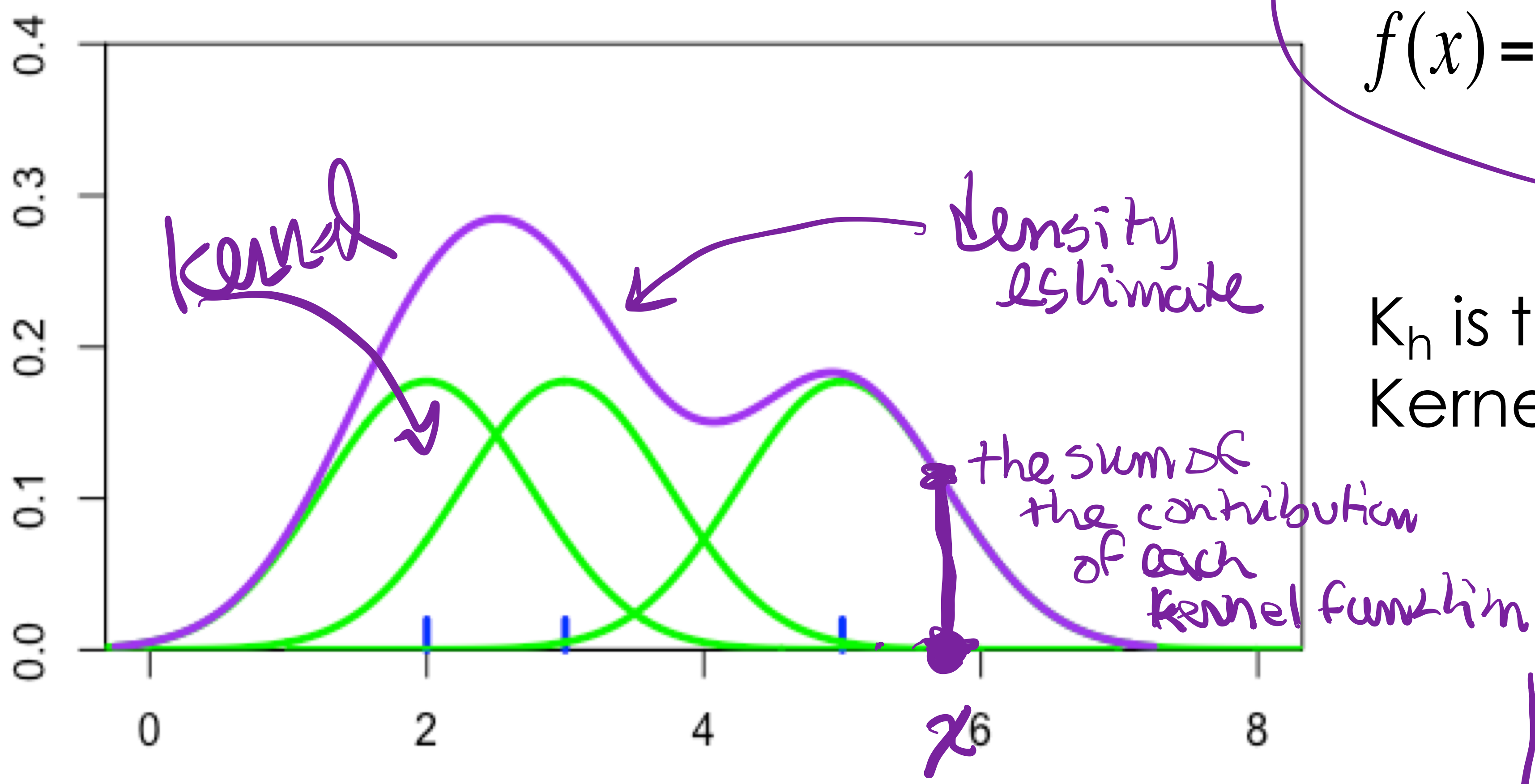
3 points –
each represents 1/3 of the data



KDE – 3 points

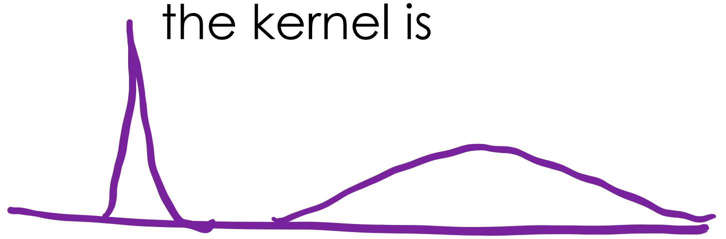
Sum the 3 kernels at each point to get the density curve

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



K_h is the green Kernel function

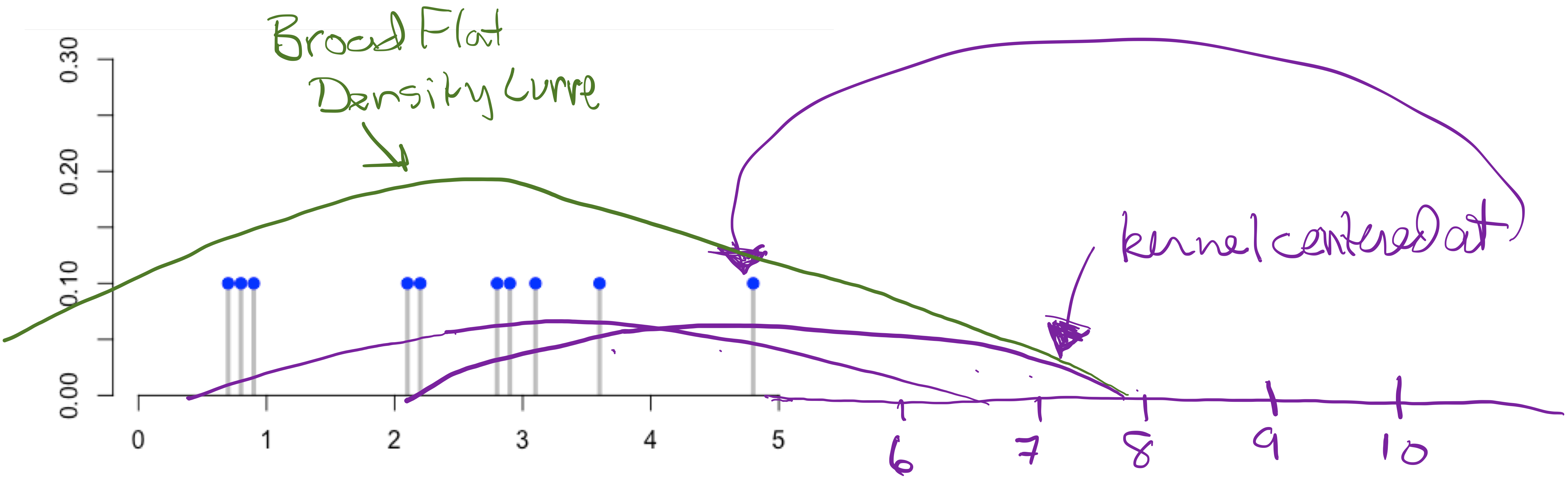
h refers to how peaked /spread the kernel is



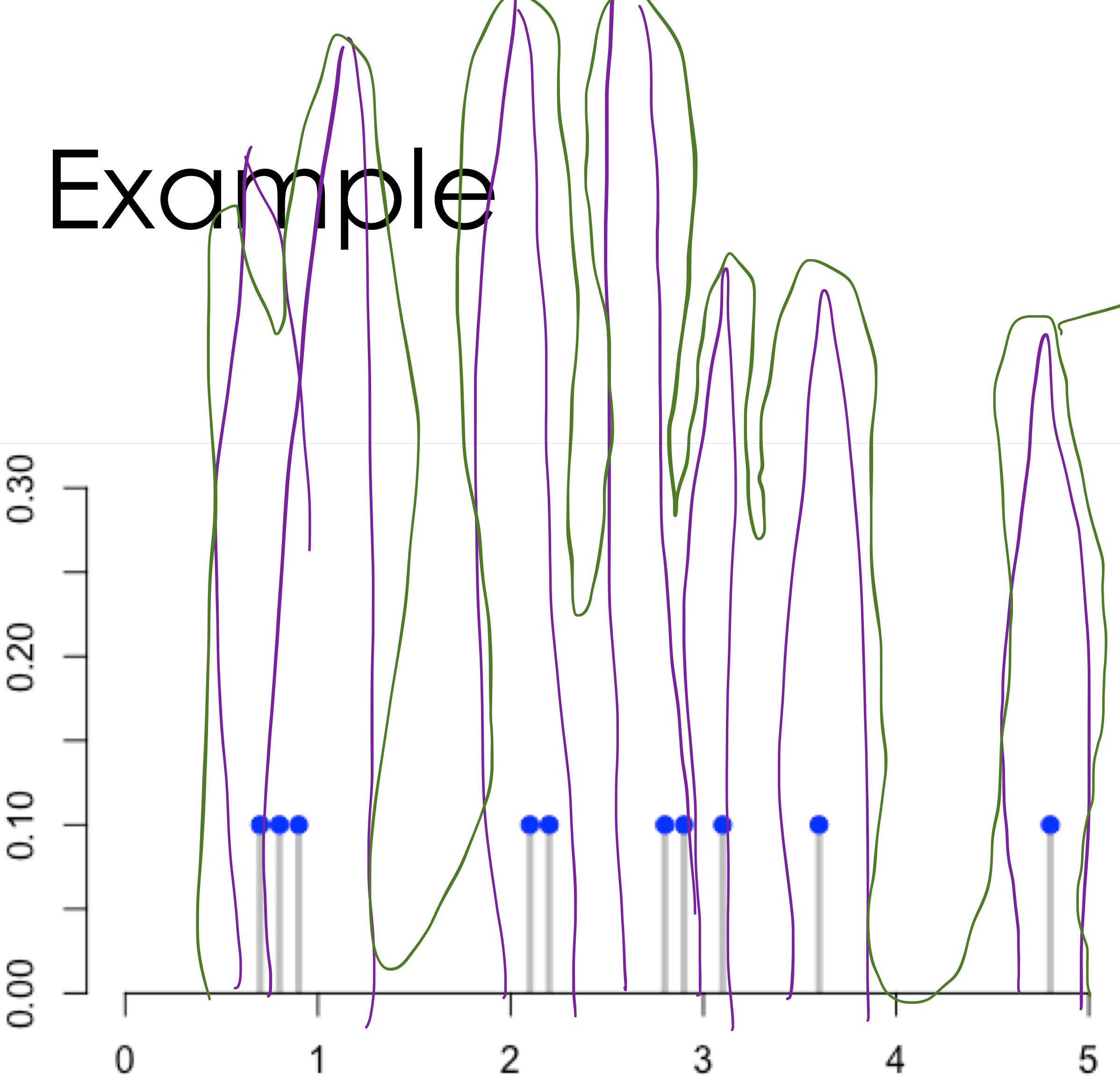
Example

Flat kernel

Density Curve is akin to 1 bar histogram



Example

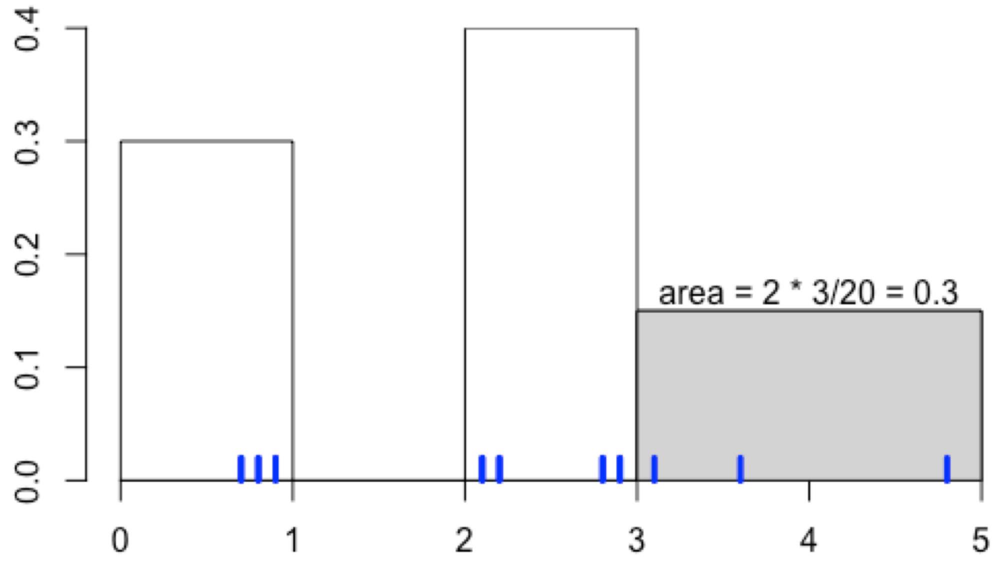


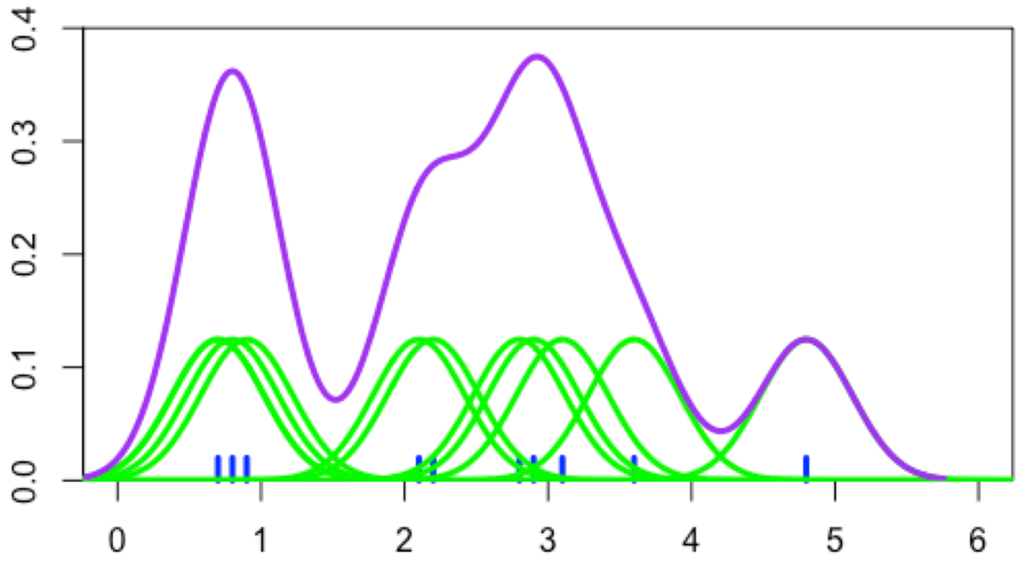
Density Curve is too grainy
It's like a rug plot

The software chooses a kernel bandwidth for you, but you can also specify your own.

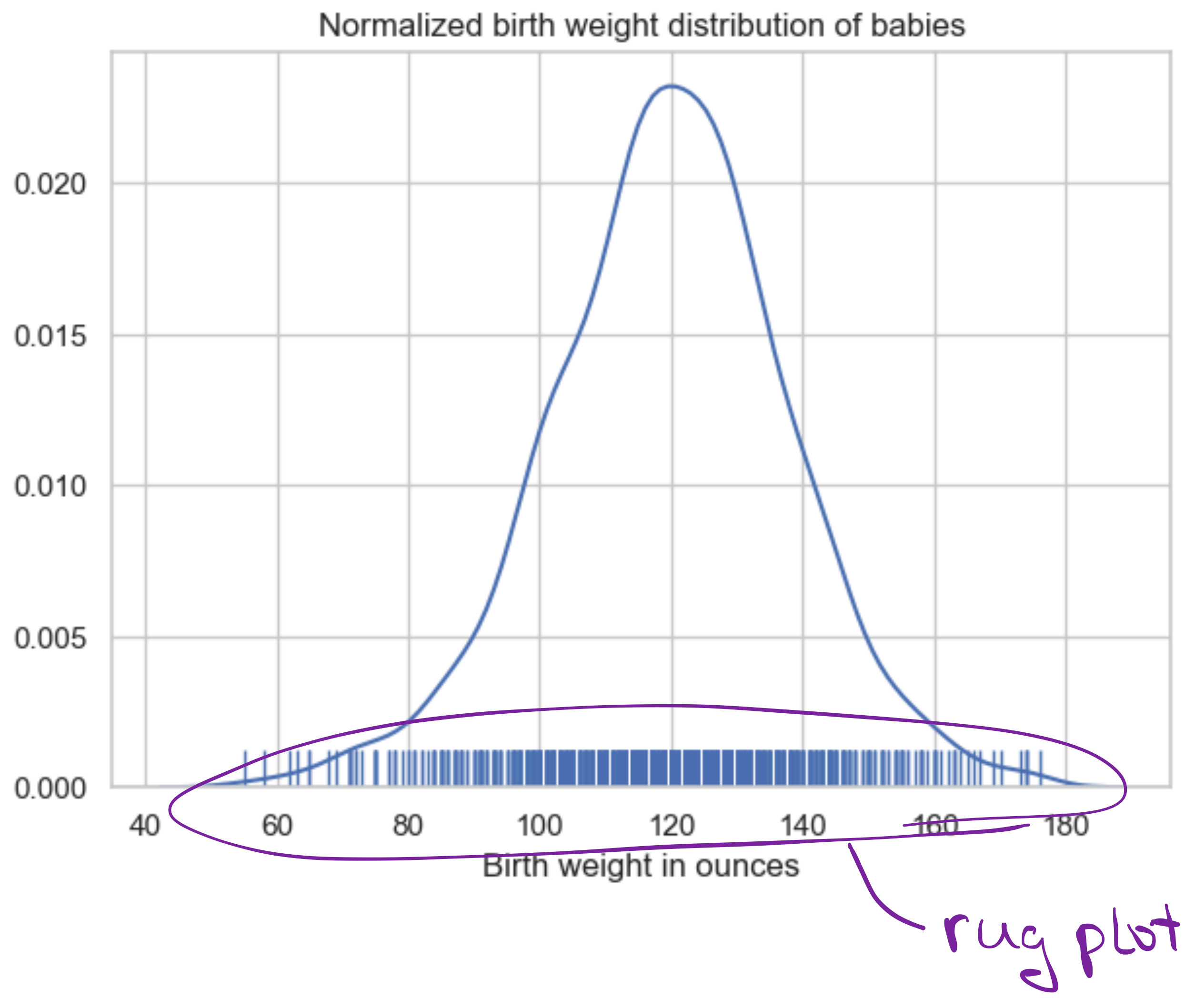
Compare the Histogram and the KDE

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

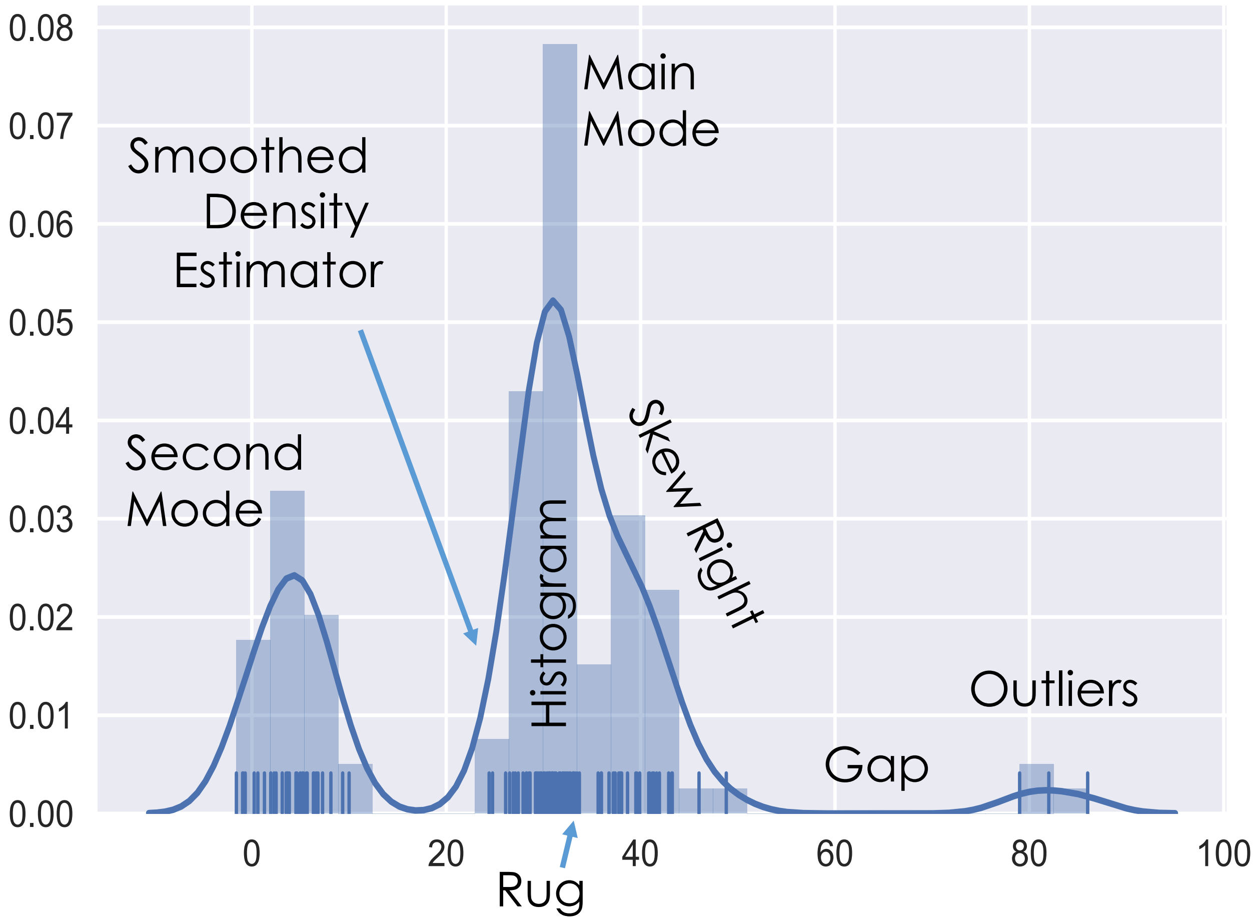
Slight left skew

Tails about normal

Histograms and Density Curves

Describes distribution of data – relative prevalence of values

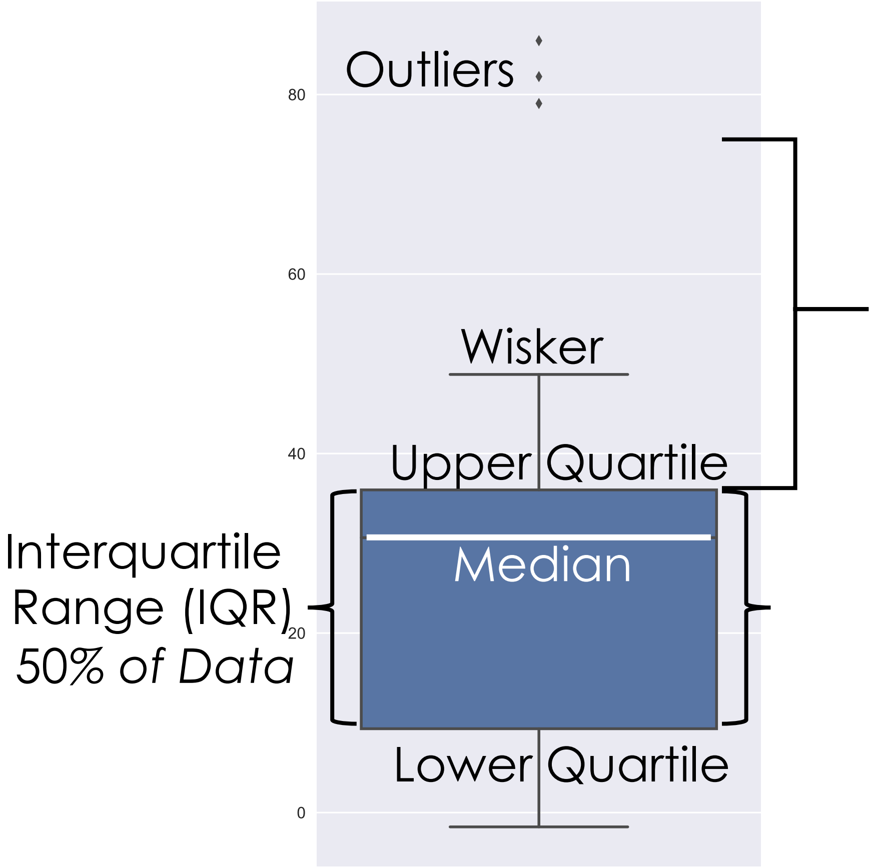
- Histogram
  - relative frequency of values
  - Tradeoff of bin sizes
- Rug Plot
  - Shows the actual data locations
- Smoothed density estimator
  - Tradeoff of "bandwidth" parameter (more on this later)



Box Plot

Tukey invented the box plot
AKA box and whisker 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…?

Modes
Gaps

Our Data

Count in →

1	2	3	4	5	6				
				6	5	4	3	2	1
0.7,	0.8,	0.9,	2.1,	2.2,	2.8,	2.9,	3.1,	3.6,	4.8

← Count in

- ➢ Median
- ➢ Lower Quartile
- ➢ IQR
- ➢ Hinge

Tukey's short cut for finding these values

n = # obs = 10

median is the $\frac{n+1}{2}$ smallest or largest obs

$= \frac{10+1}{2} = 5.5$ average the 5th & 6th

median $= \frac{2.2 + 2.8}{2} = 2.5$

Our Data

1 2 3 3 2 1

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

To find the quartiles, we take the count-down value for the median, drop the ½ (if it was it), and add one & divide by 2, e.g.,

$5.5 \rightarrow \frac{5+1}{2} = 3$

LQ is 3 in from bottom

UQ is 3 in from top

IQR is UQ−LQ

= 3.1 − 0.9 = 2.2

Hinge is 1.5 IQR = 1.5 × 2.2 = 3.3

Any value more than 3.3 away from LQ/UQ is an outlier

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

Lower Quartile = 25th percentile

Percentile – Need a more general def

Notice the percentile will always correspond to a data point

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

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

0.75 2.5 3.7

0.6 0.7, 0.8, 0.9, 2.1, 2.2, 2.8, 2.9, 3.1, 3.6, 4.8

20% 90% of the data is at or below 3.6

10^{th}%tile = 0.7 90^{th}%tile = 3.6 60^{th}%tile = 2.8

15^{th}%tile = 0.8 83^{rd}%tile = 3.6 66^{th}%tile = 2.9

Any value below 3.6 will not have 83% below it

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**

sorted wages

sorted wages	5.	5.	5.	5.	5.	5.	20.	20.	20.	20.	50.	50.
corresponding weights	½	½	½	½	½	½	1	1	1	1	½	½

sum to 8

50^{th}%tile = 20

4/8 = 50% is at or below 20

75^{th}%tile = 20

7/8 = 87.5% is at or below 20

3/8 = 37.5 is at or below 5

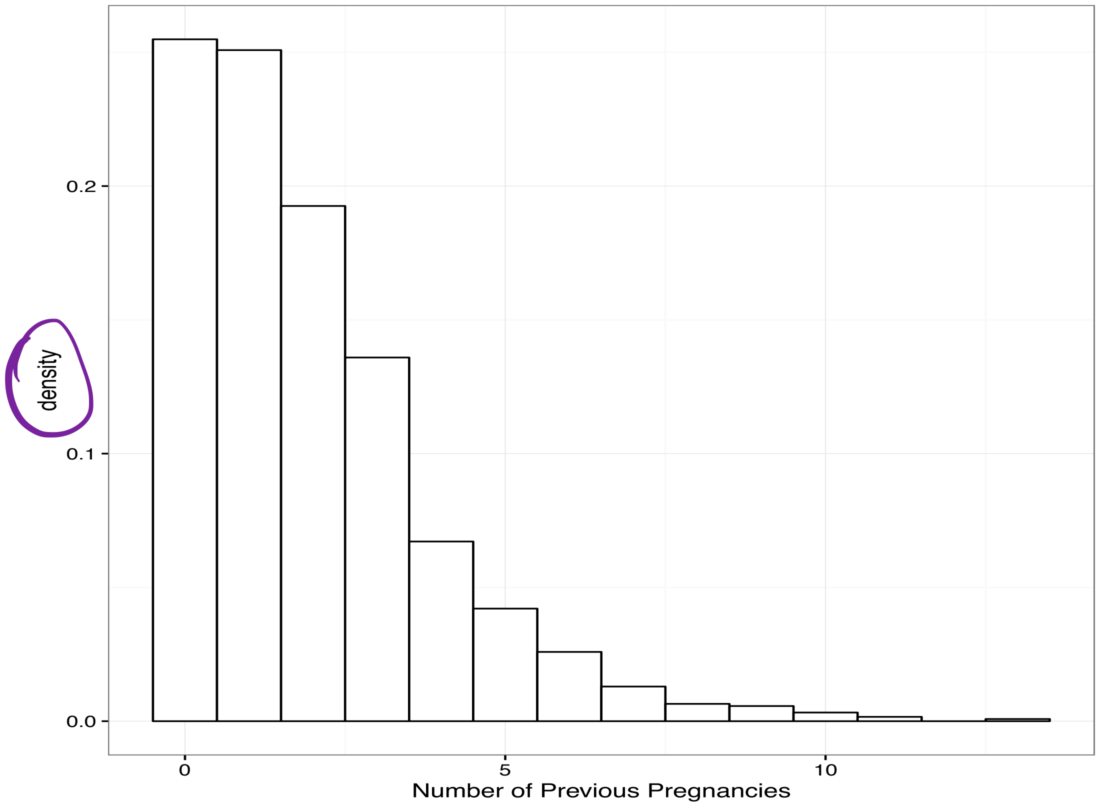
> nothing in between

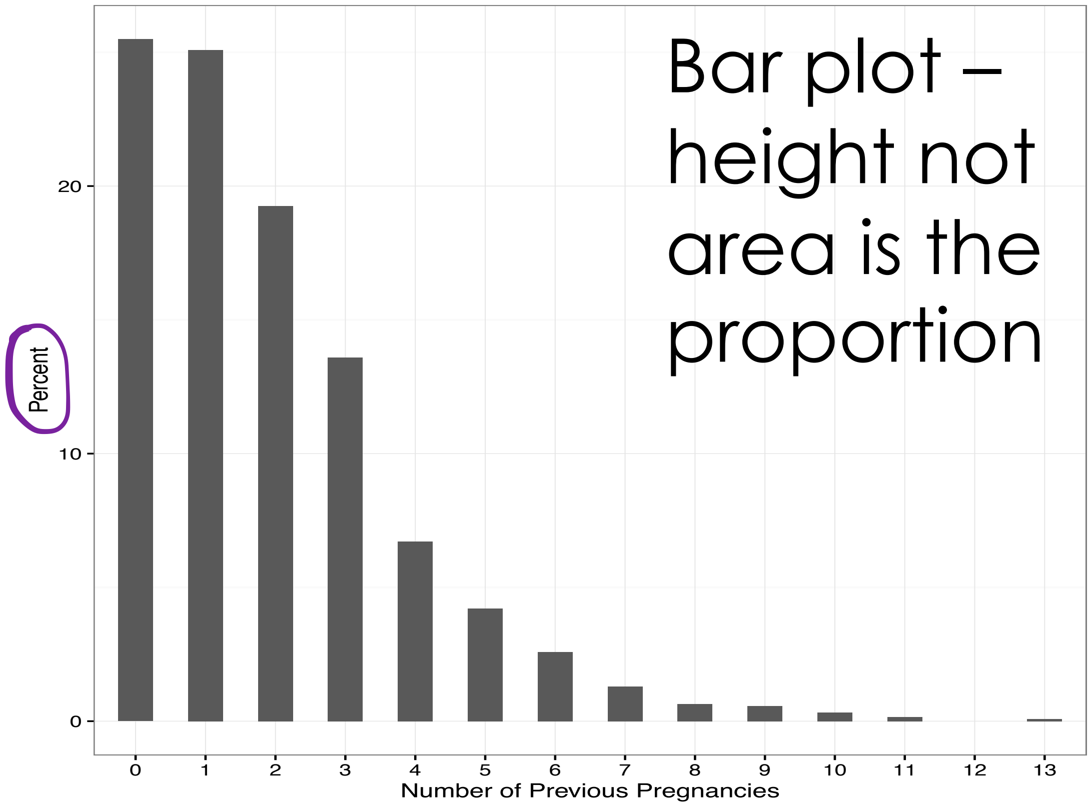
Quantitative Discrete

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

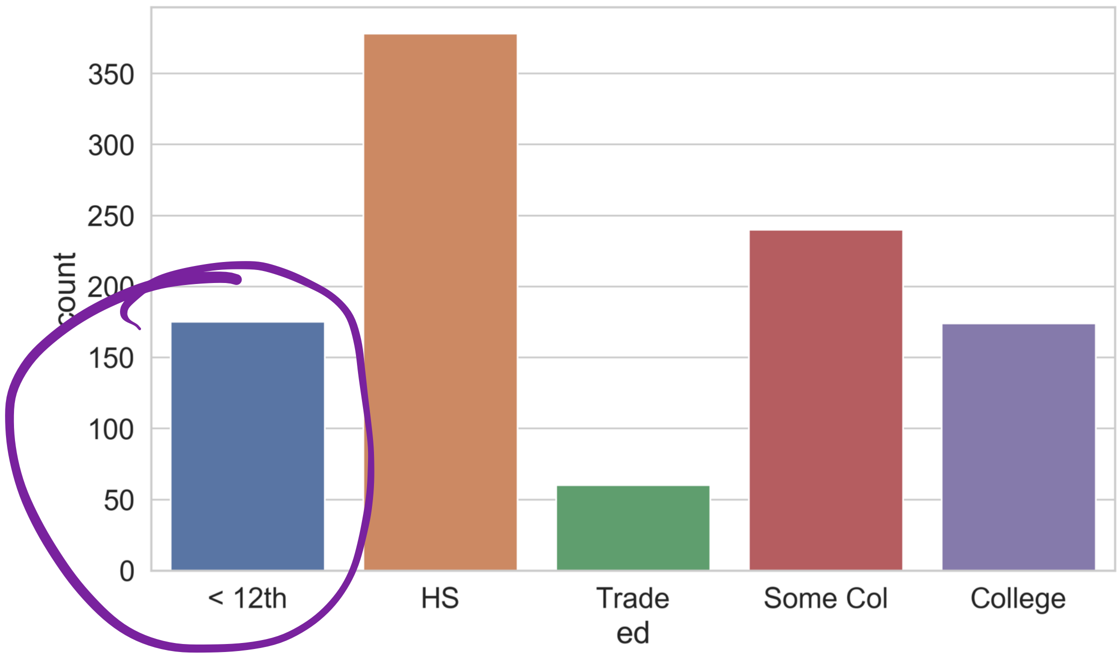
We look at the relative size of groups

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

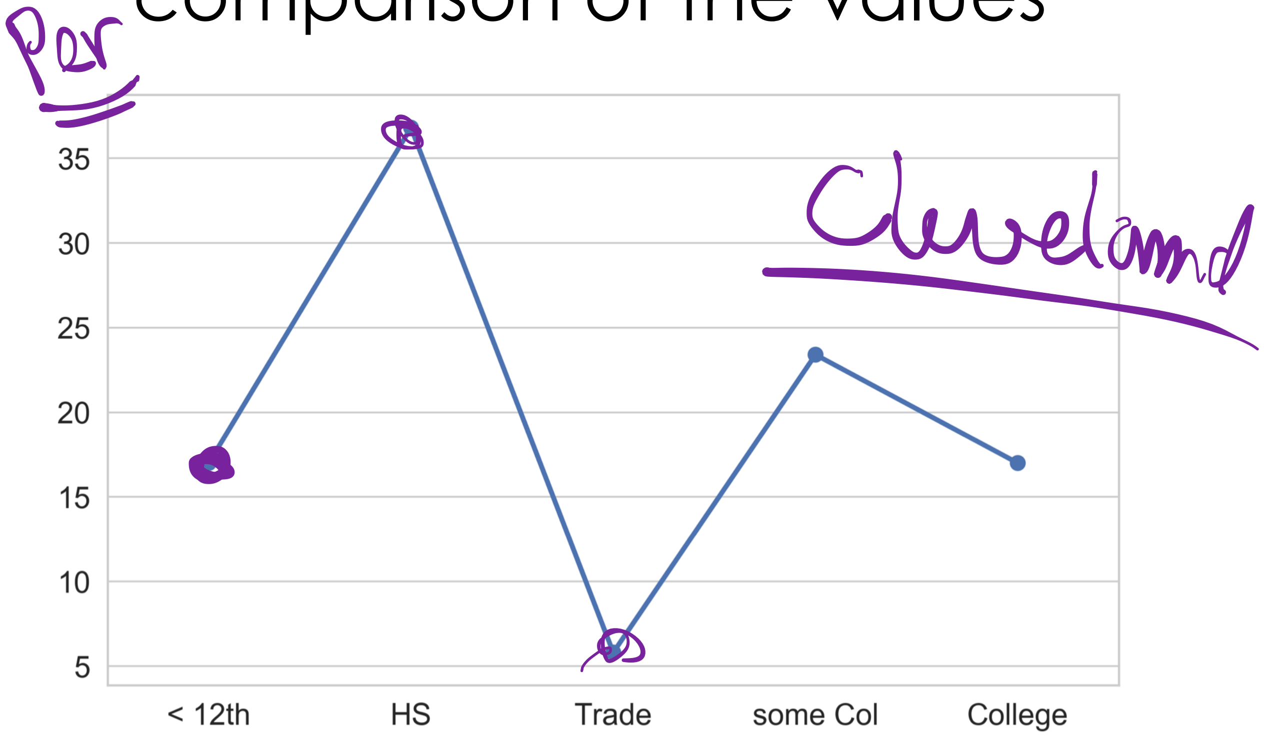
Answers have implications in building prediction models

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

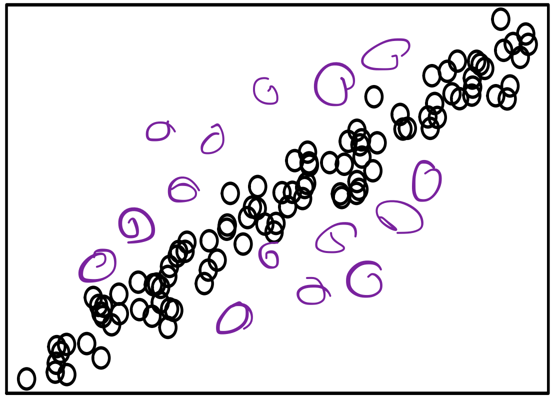
- Scatter plot uncovers form of relationship between 2 variables
- Linear relationships are particularly simple to interpret
- Simple and elegant statistical theory for linear relationships
- Models are typically approximations, choose a simpler model over a complex one

Common Relationships

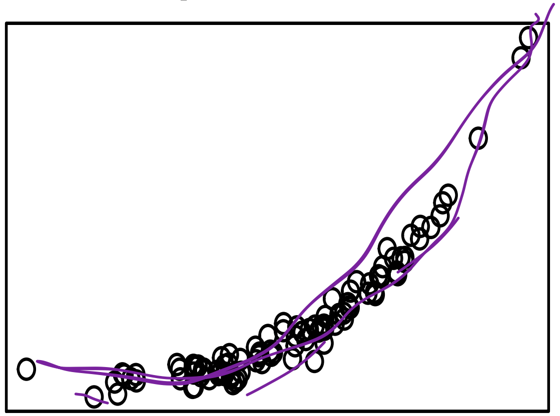
Ideal

Can also have more spread

simple linear



simple nonlinear

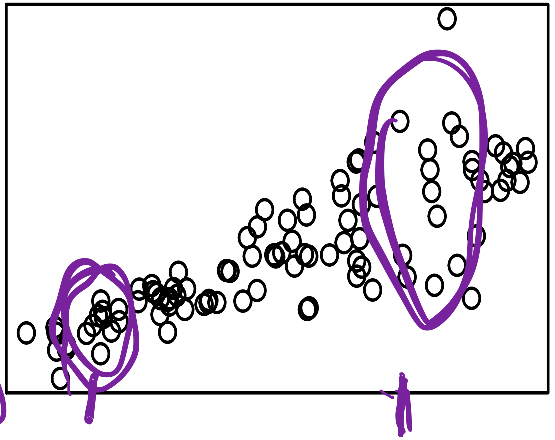


Good too

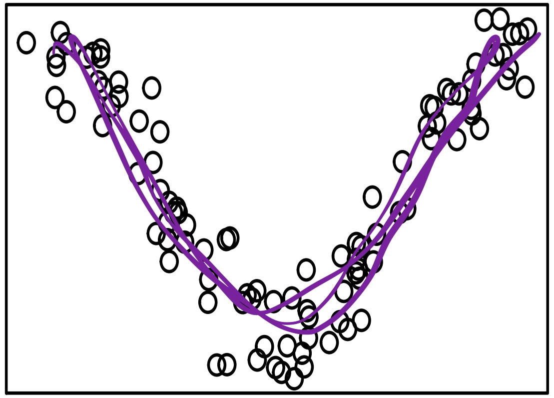
Typically we transform to a linear rel.

Still linear but we need to take care when modeling

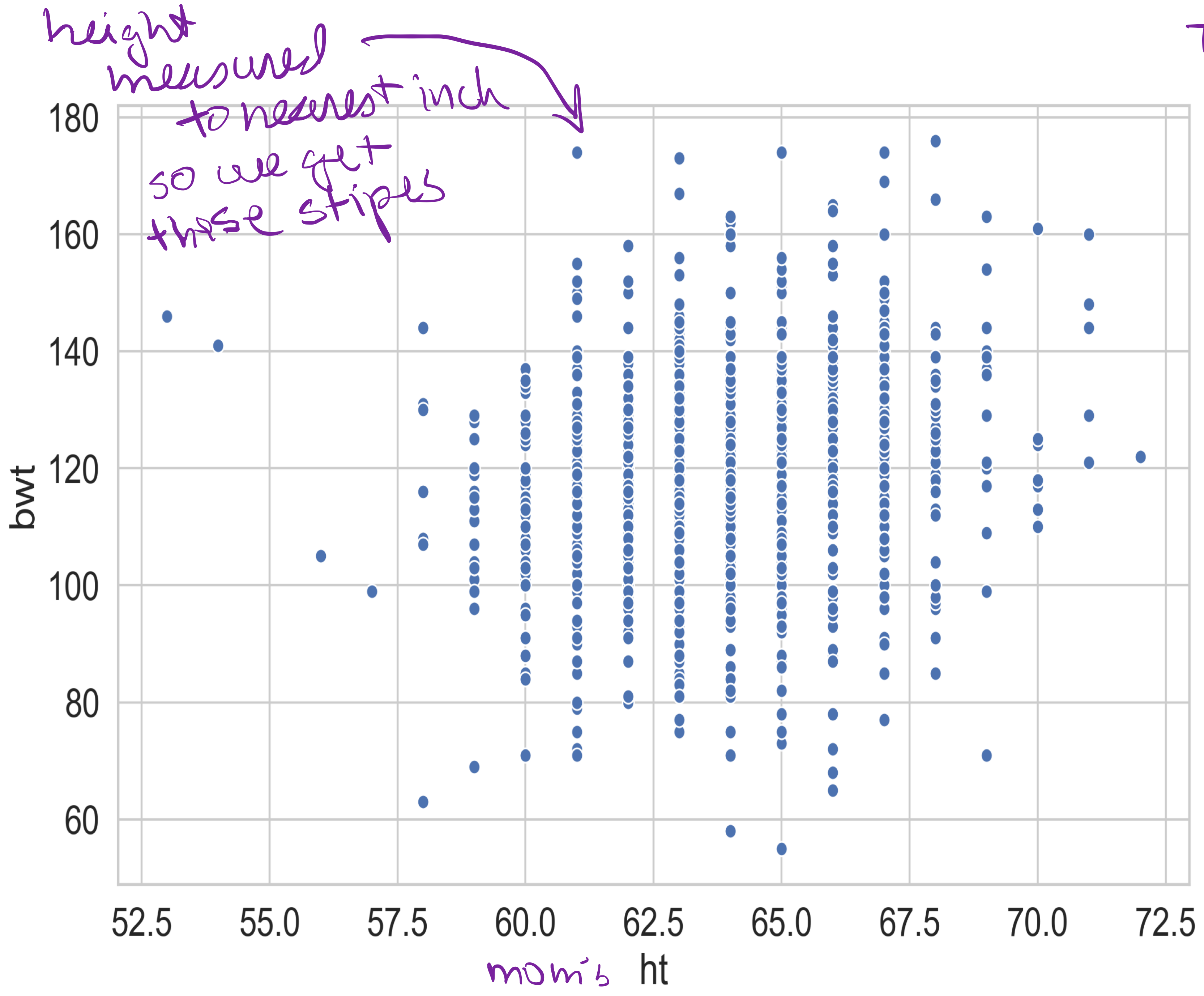
unequal spread



complex nonlinear



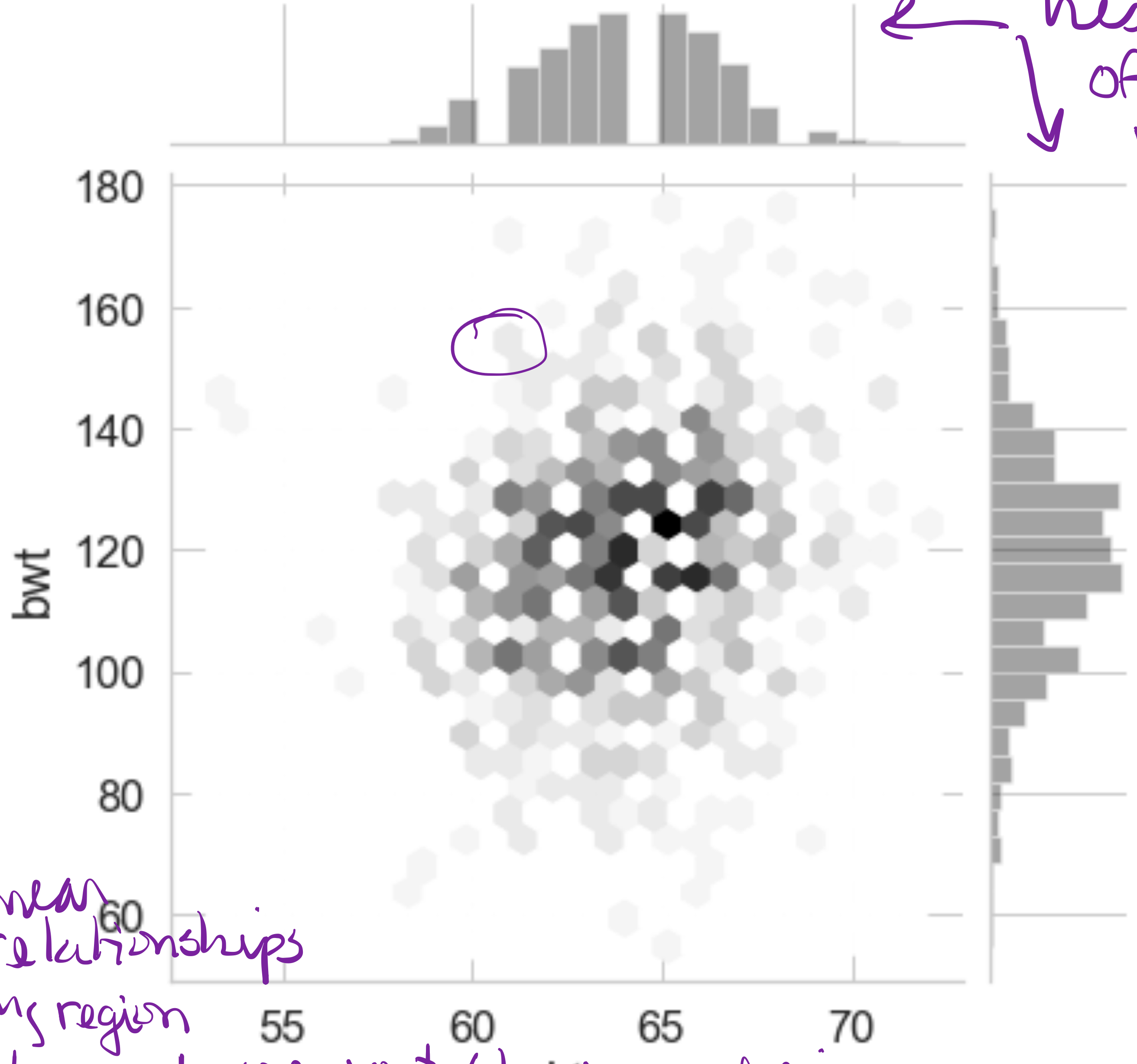
Very difficult to work with



The scatter plot is a 2-d rug plot

Hex Bin

Shading corresponds to density of points in the cell



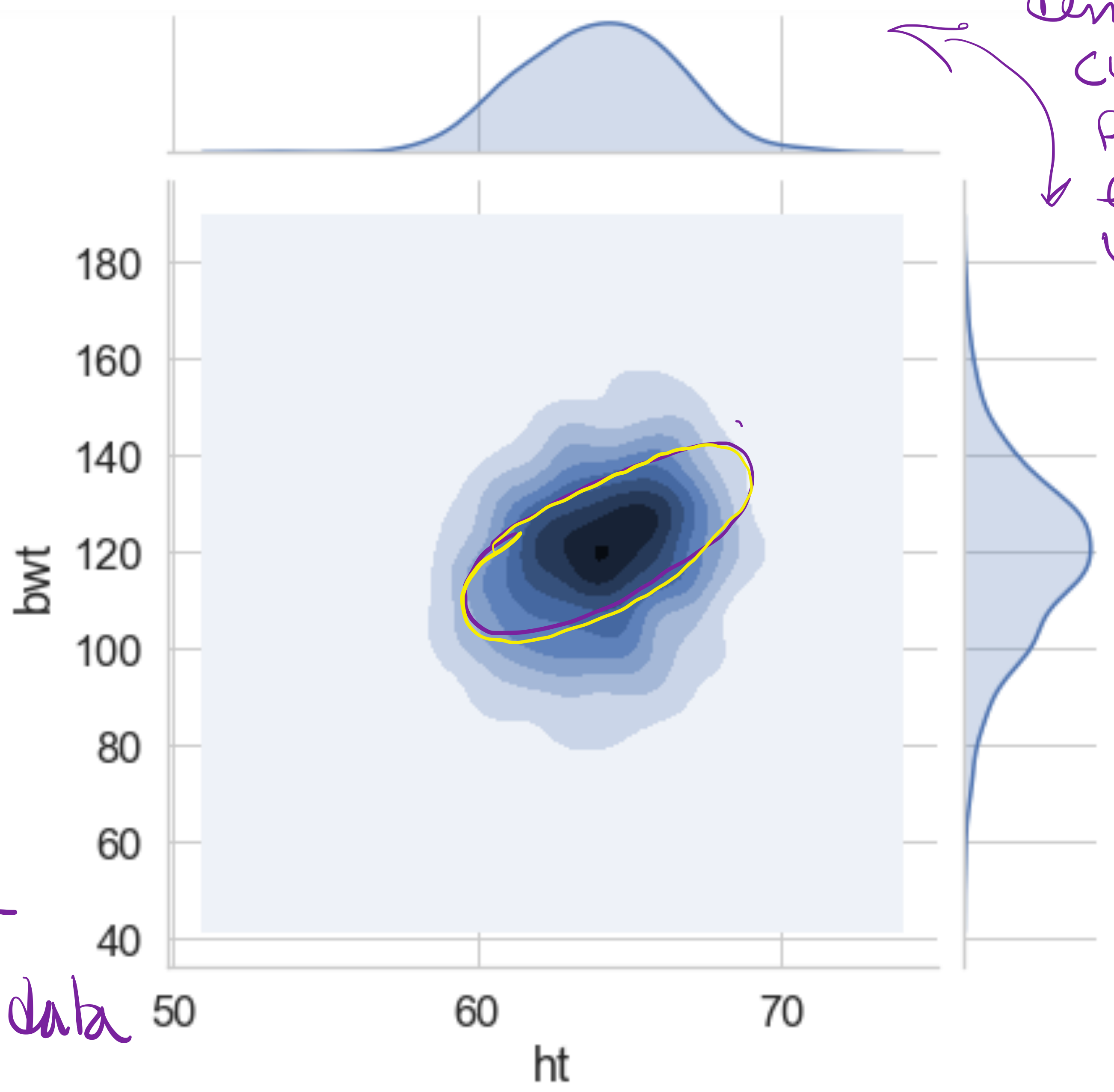
← histogram of each variable

Why hexagons –
- Easier to see elliptical/linear relationships
- More efficient for covering region
- Visual bias of squares – drawn to see vert & horizontal lines

Smooth Contour

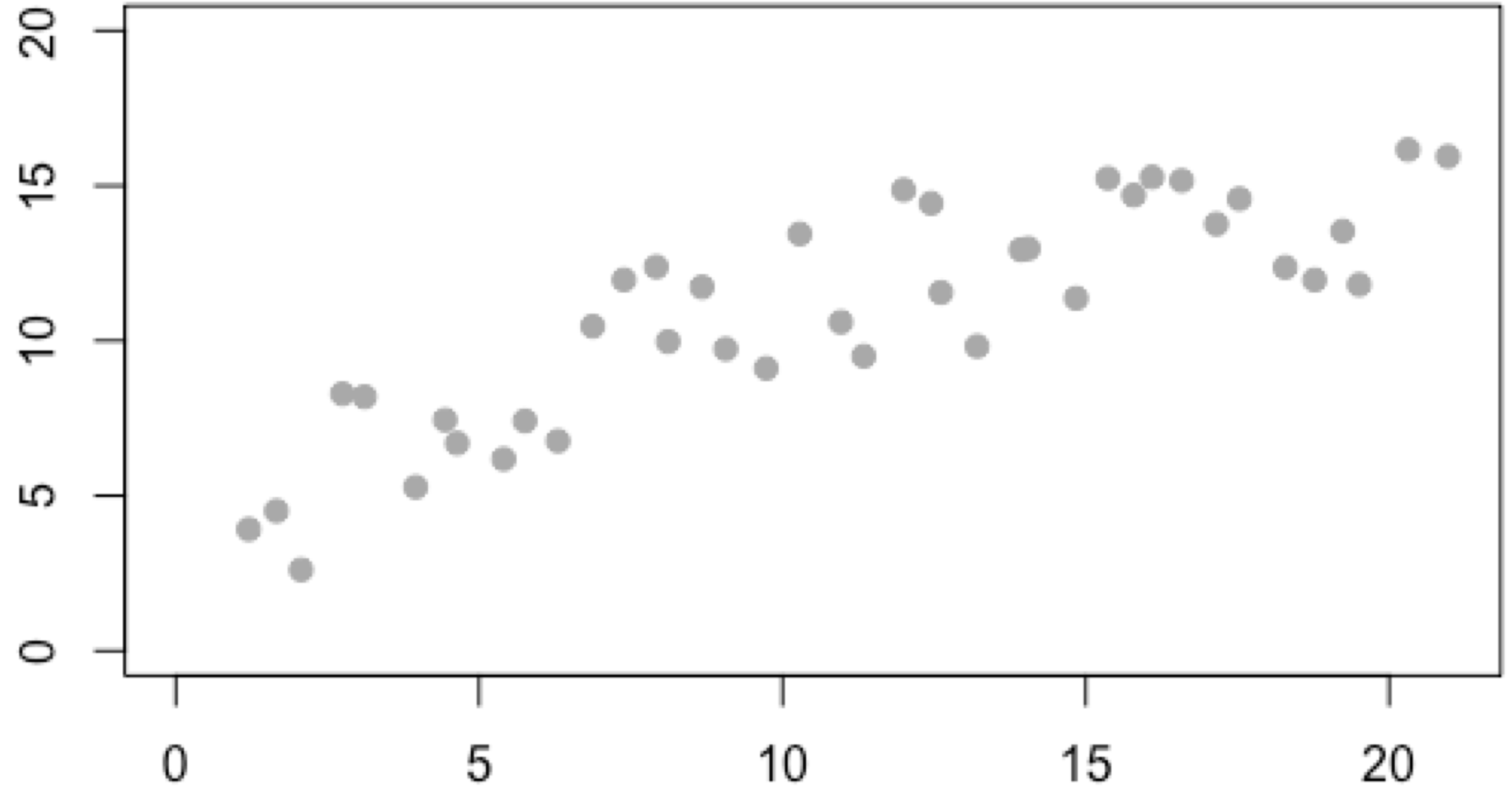
kde
in 2-d

Contours of
the 3-d
density
smooth of
our 2-d data



Smoothing Scatter plots

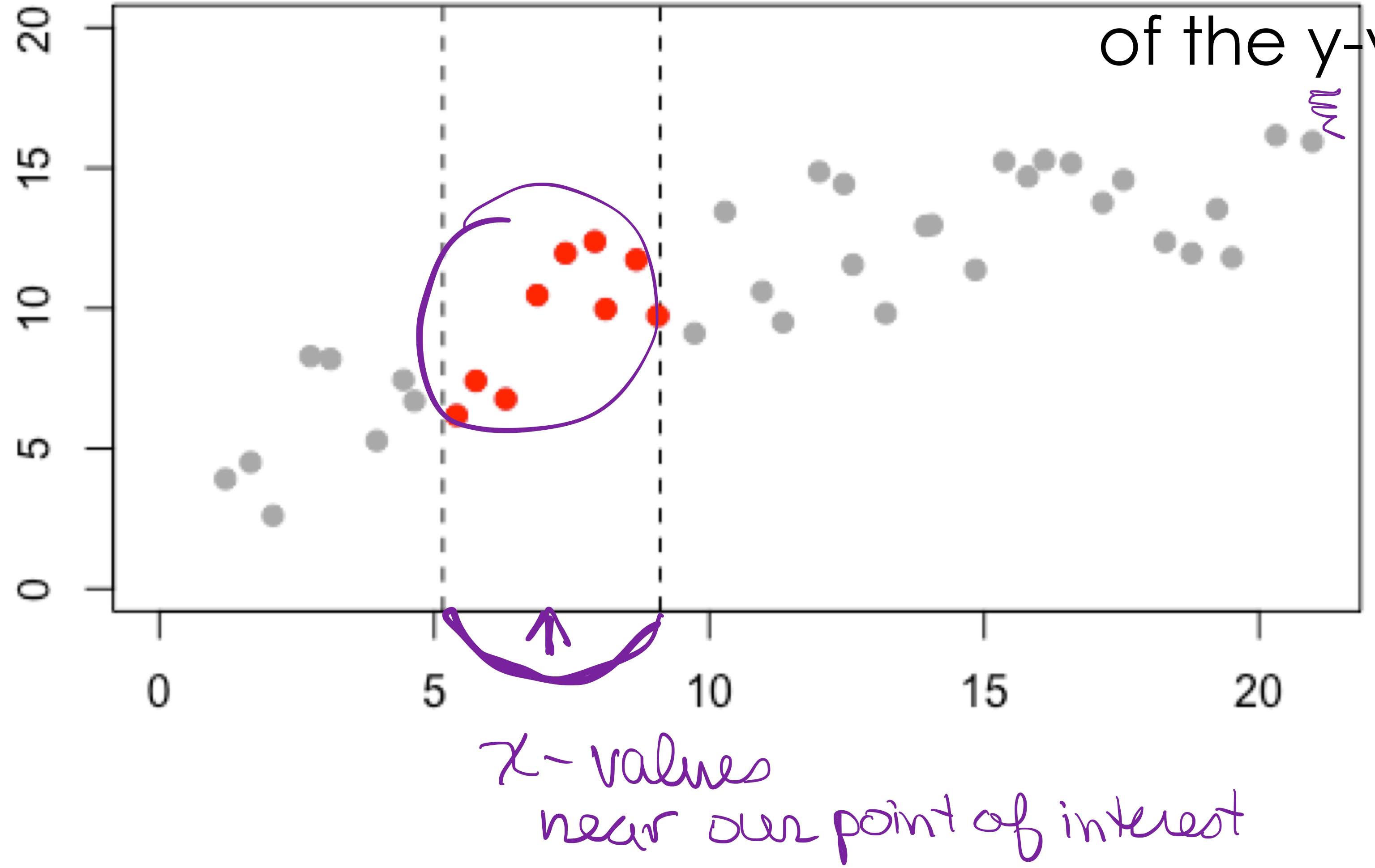
Now we want to smooth the y-values as a function of x



Smoothing Scatter plots

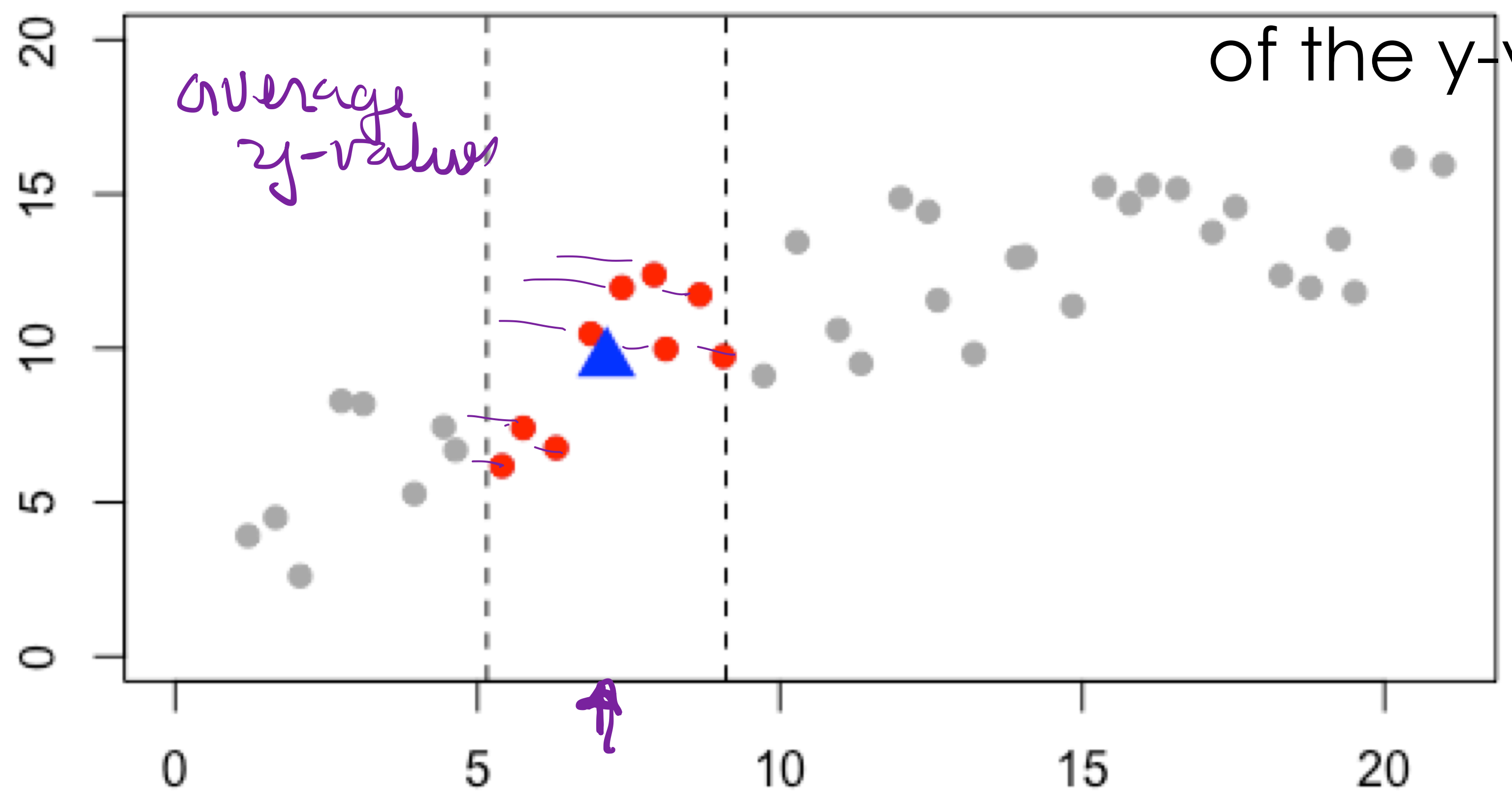
For an x-value consider all of the x's near it

Take an average of the y-values



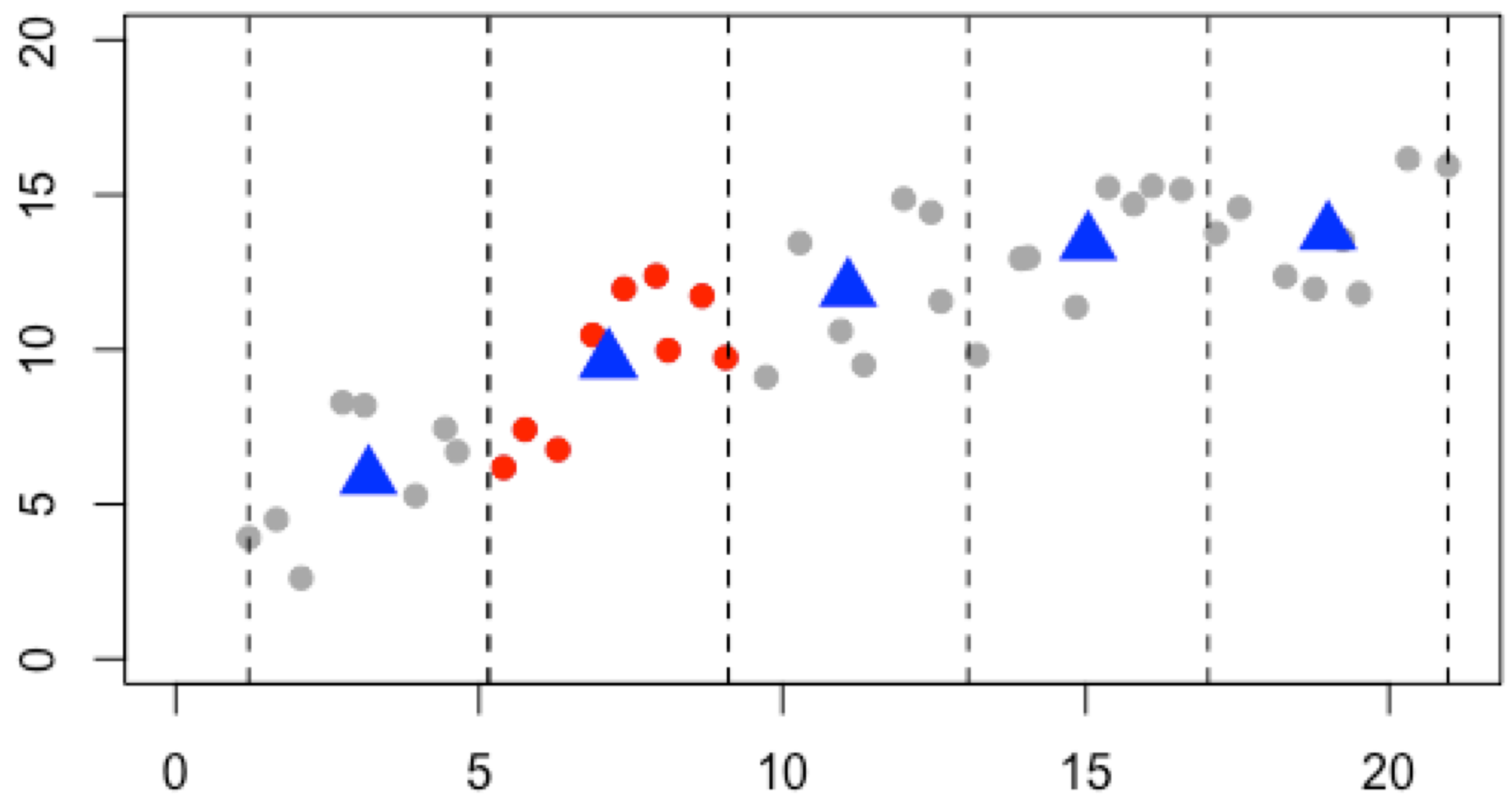
Smoothing Scatter plots

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



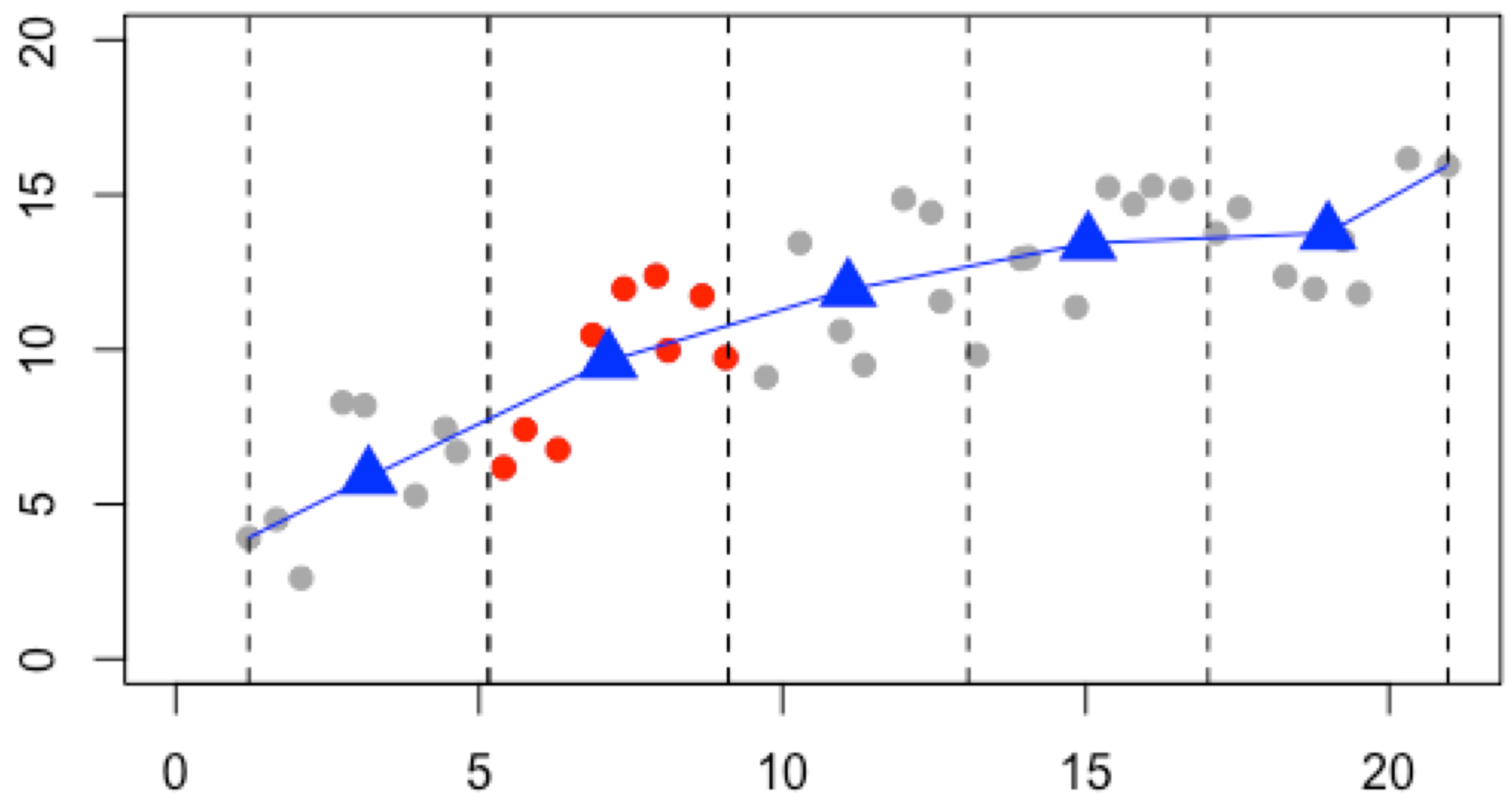
Smoothing Scatter plots

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



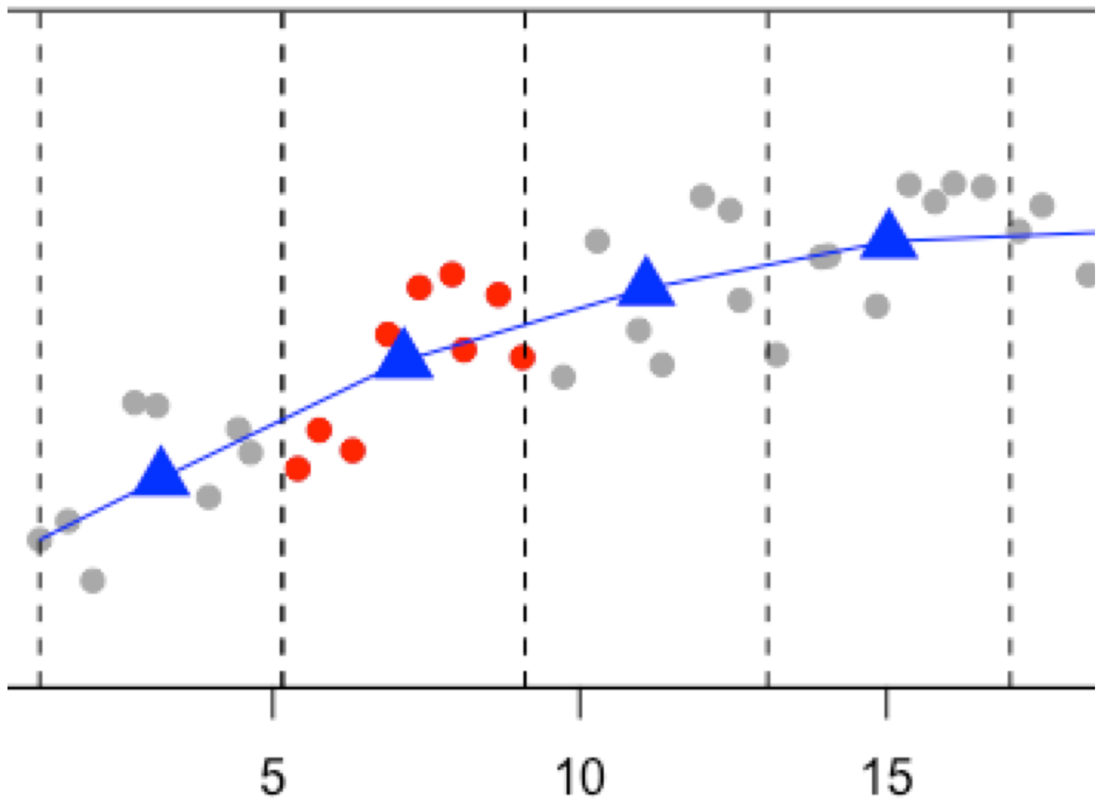
Smoothing Scatter plots

These averages sketch out a curve



Connected these blue

Smoothing Scatter plots



Like with histograms we average based on distance from x

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

distance from x

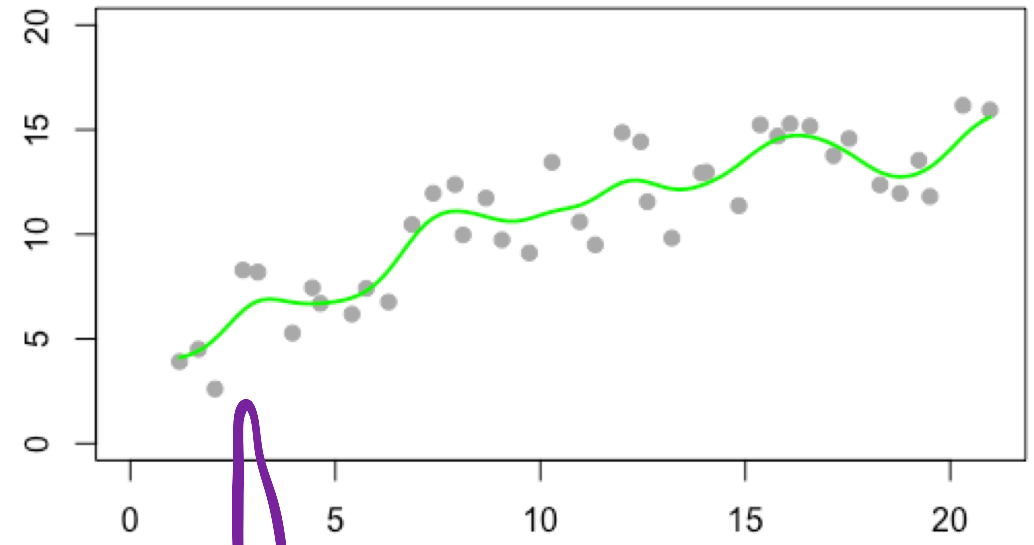
$$g(x) = \sum_{i=1}^{n} \frac{K_h(x-x_i)\,y_i}{\sum K_h(x-x_i)}$$

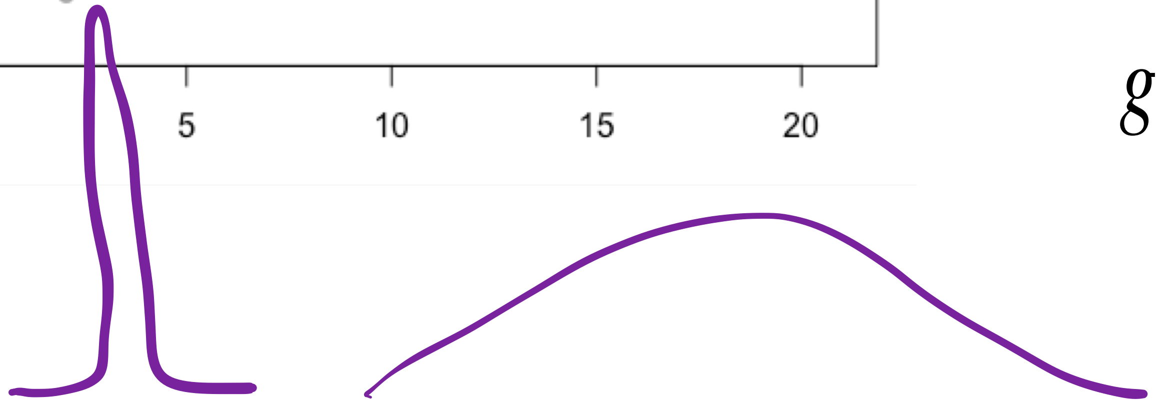
We average the y-value

The denominator ensures the weights sum to 1

Smoothing Scatter plots

small/narrow kernels give wiggly results





Rather than a simple average

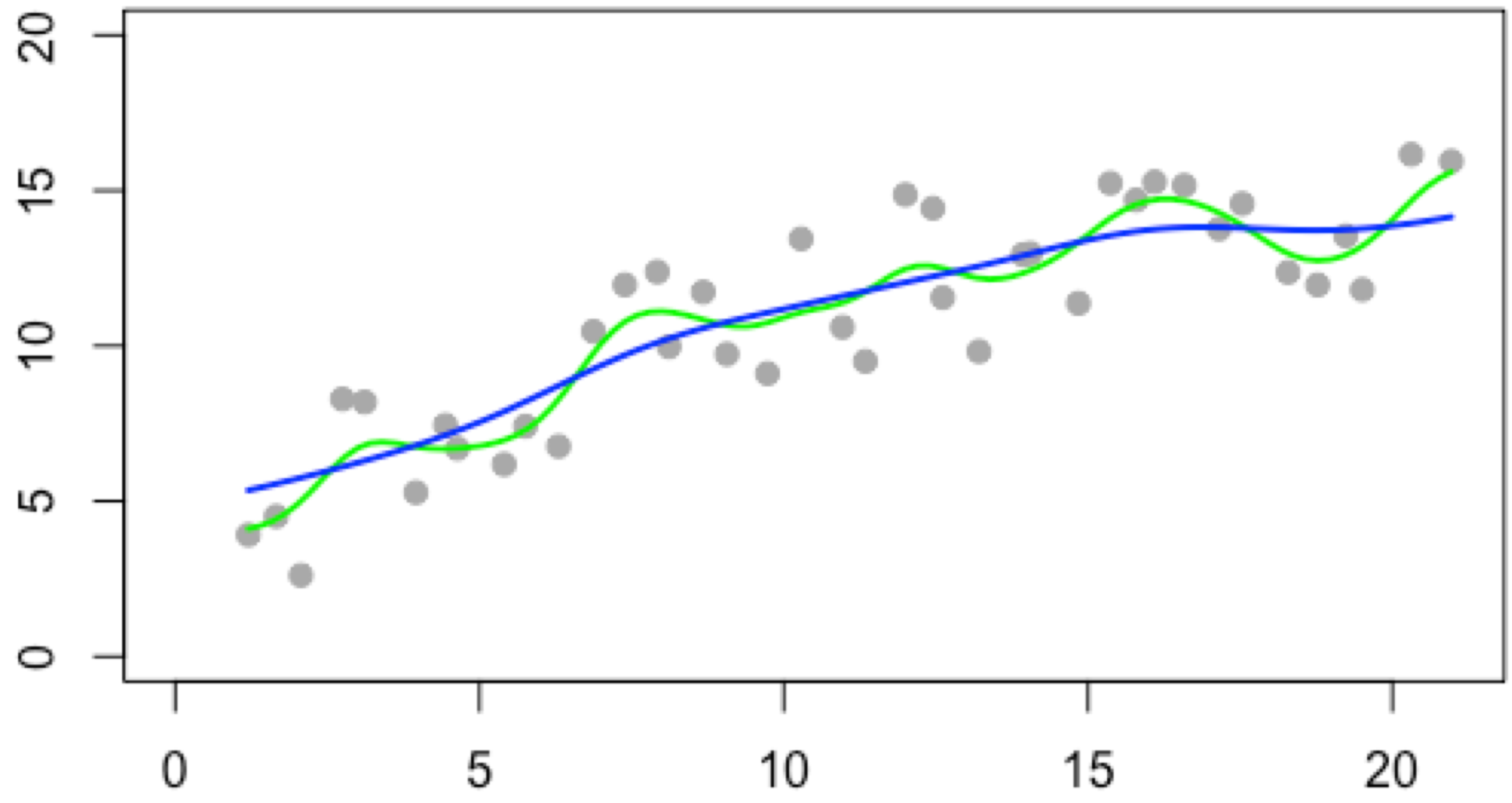
We use kernels positioned on the x_i to determine the weights to place on the y_i in the average

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

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

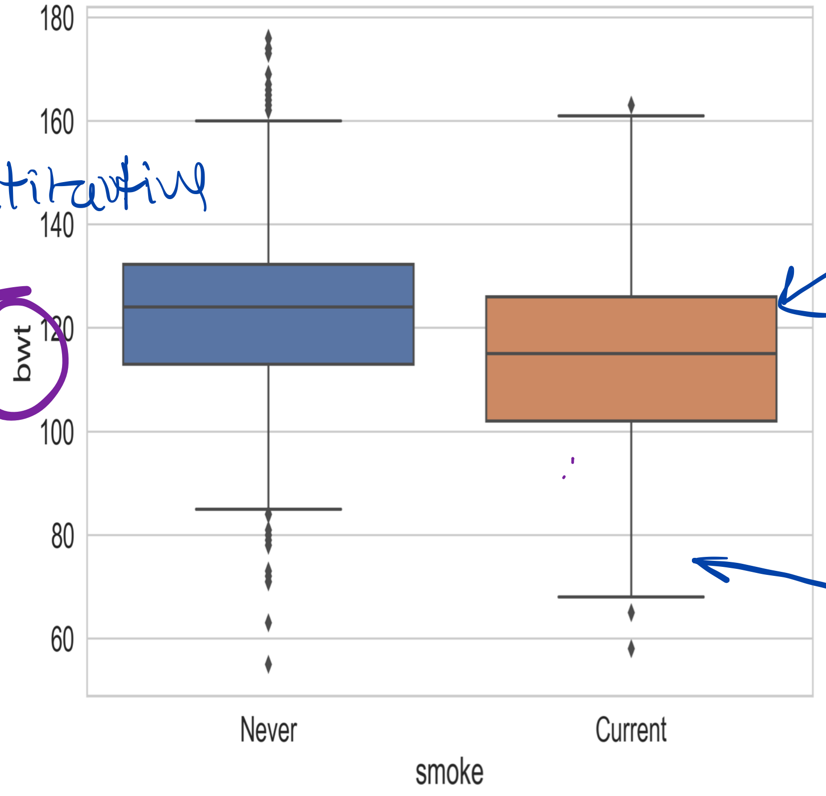
- Moving window
- Smooth/Average y values in the window
- Many different approaches for doing this:
  - kernel methods (what we just showed),
  - cubic splines, thin plate splines,
  - Locally weighted smooth scatterplot (lowess)

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

Mix Quantitative & Qualitative

Mix Quantitative & Qualitative

Side-by-side Boxplots



quantitative on y-axis

can't see modes

Can see tails and assymmetry

Side-by-side violin plots

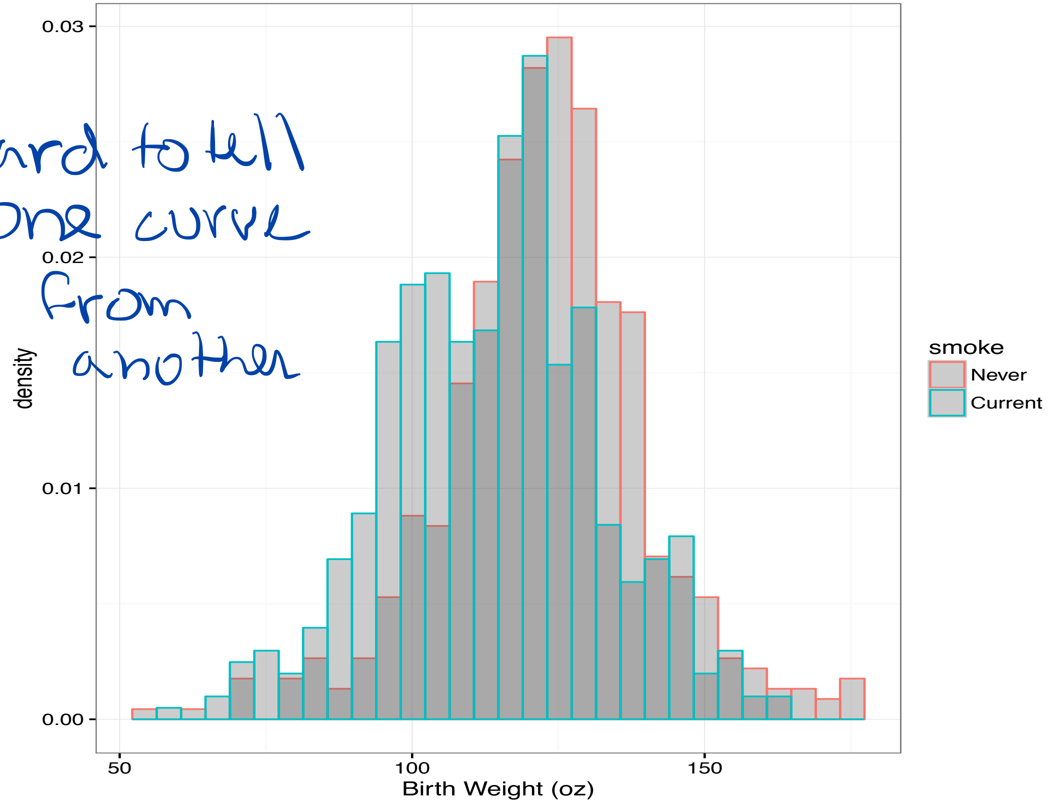
Density on its side

Can put Qualitative Variable on the x-axis

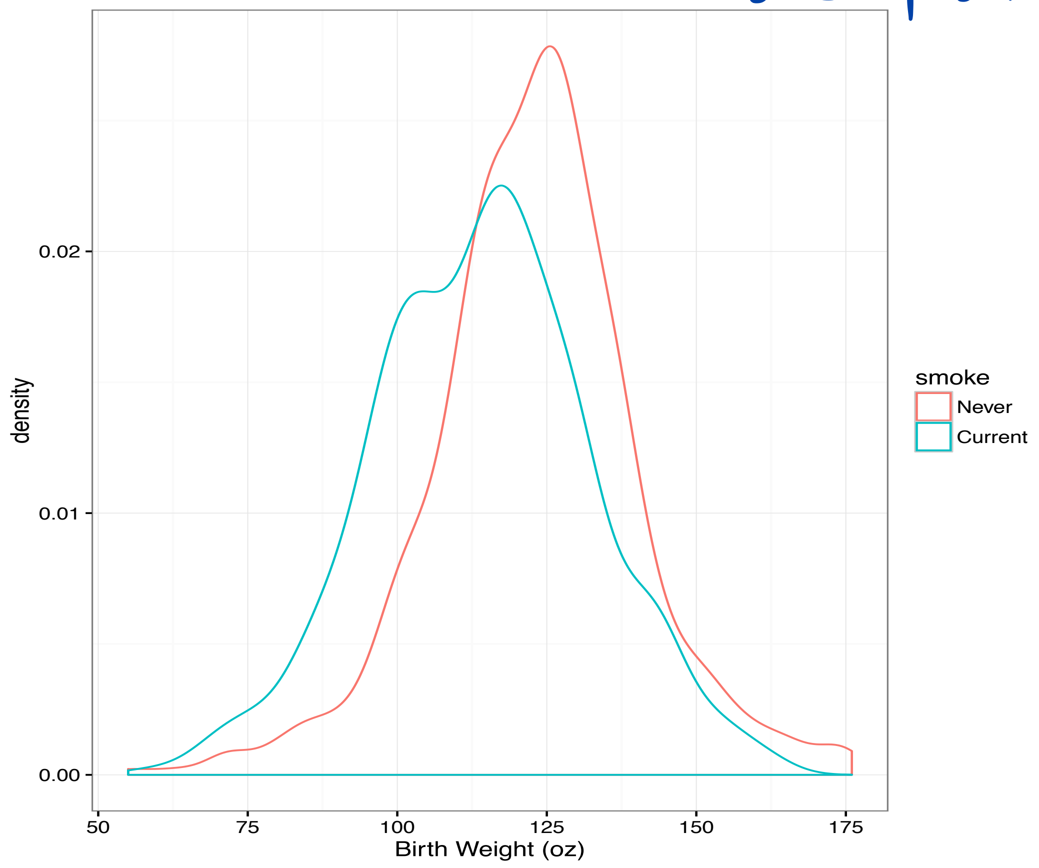
Mix of Qualitative and Quantitative

Overlaid bars/curves

Hard to tell one curve from another



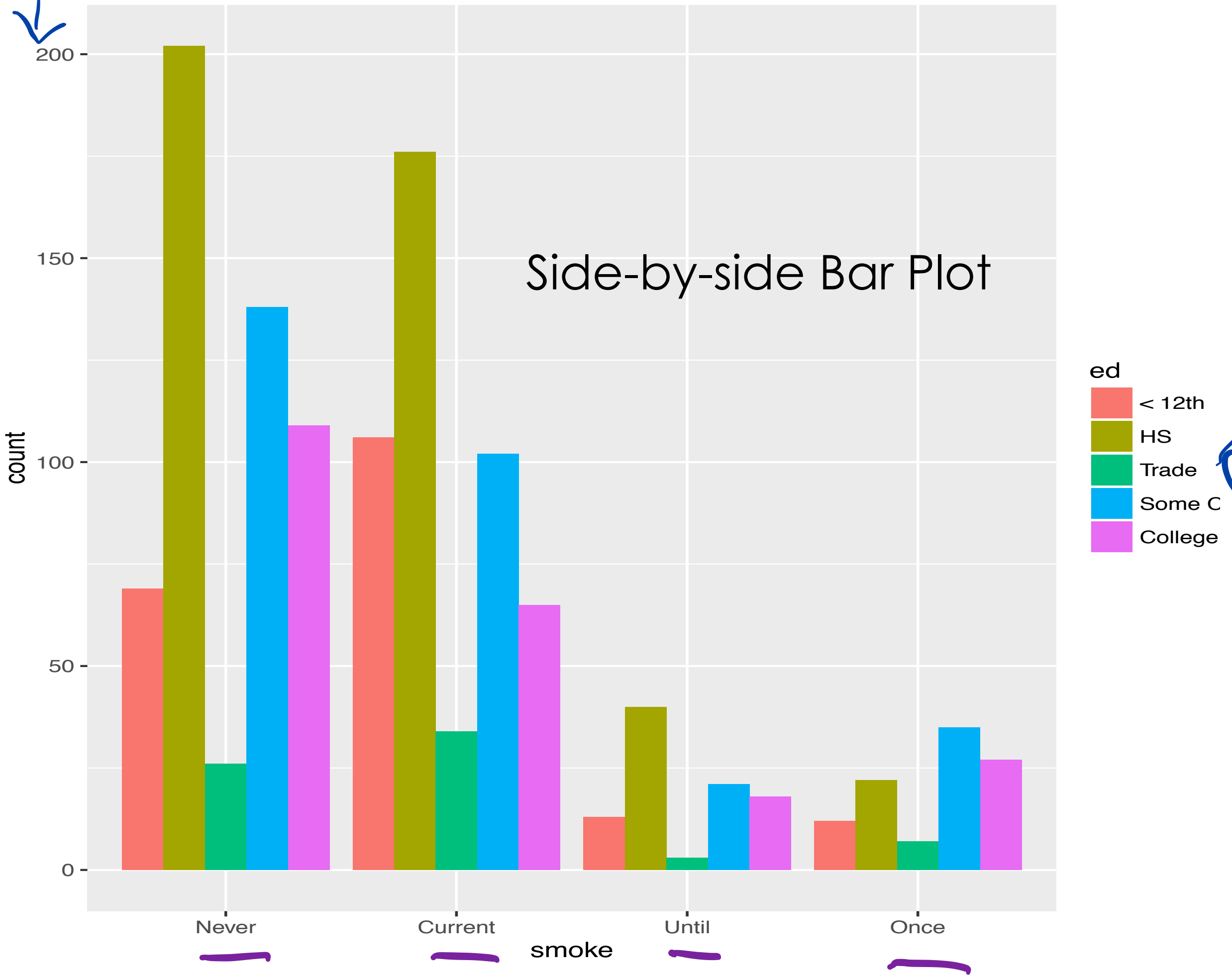
Much preferred

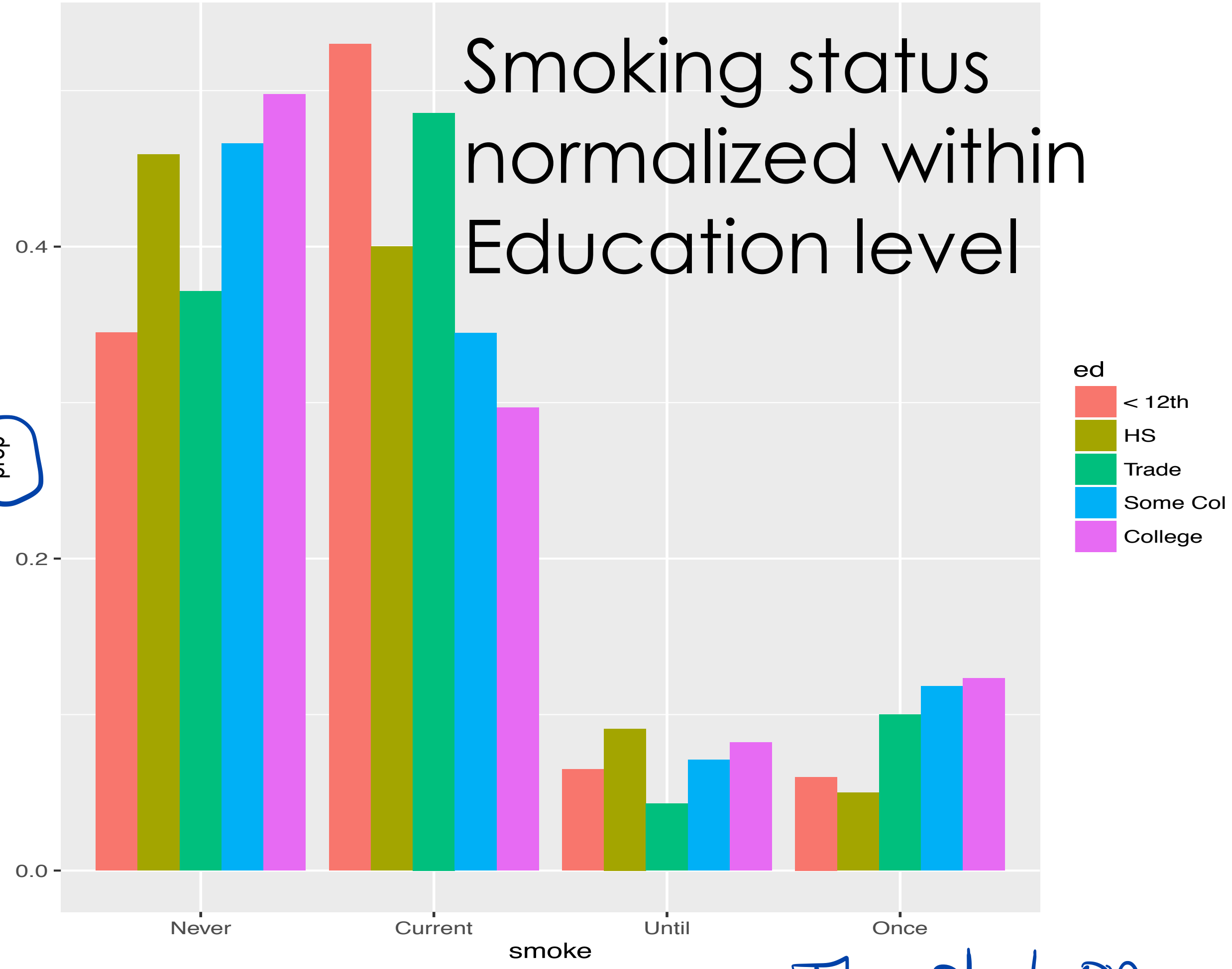


Two Qualitative Variables

Pairs of Qualitative Variables

should be proportion



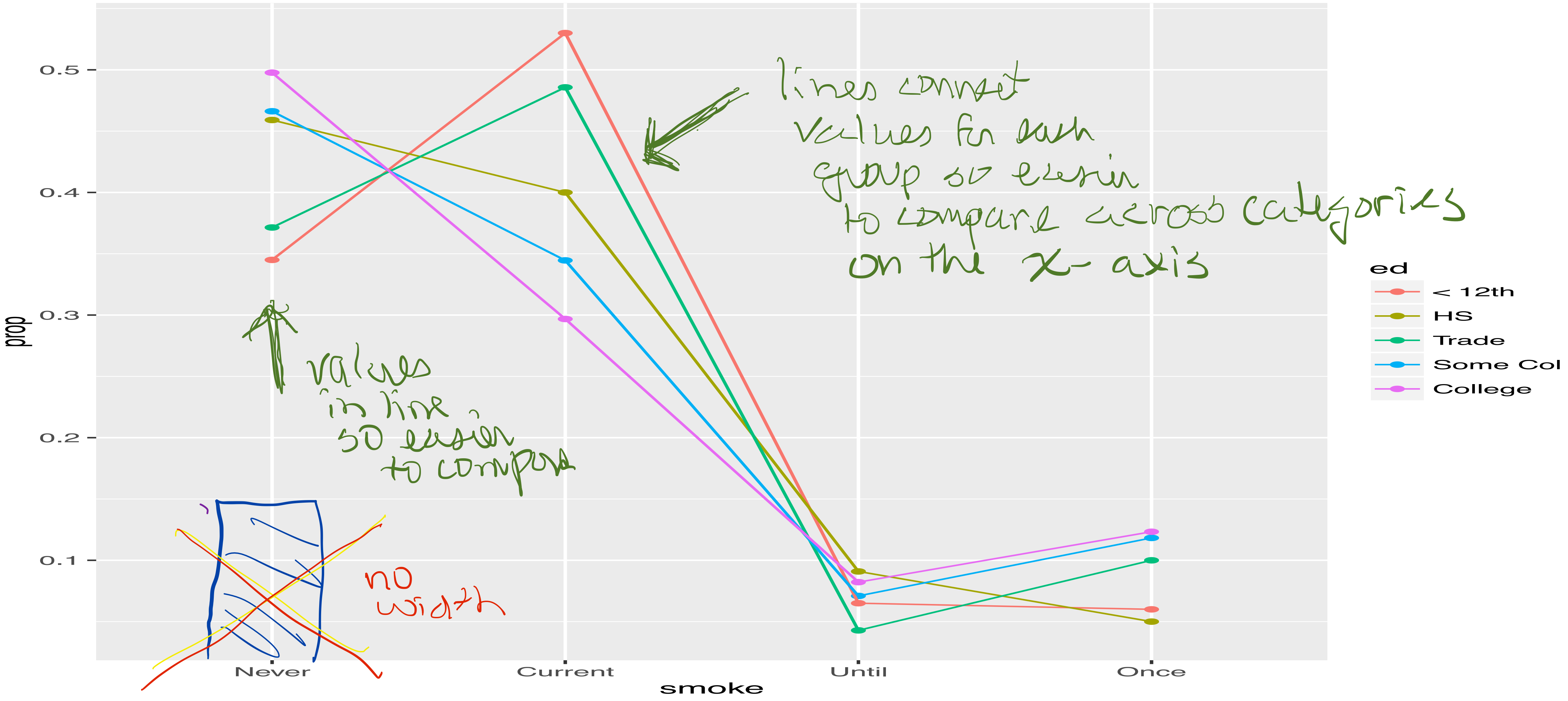


What's the difference between these 2 plots?

The plot on the right is normalized

Interaction/Factor Plot

Smoking status normalized within Education level



Univariate Graphical Displays

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

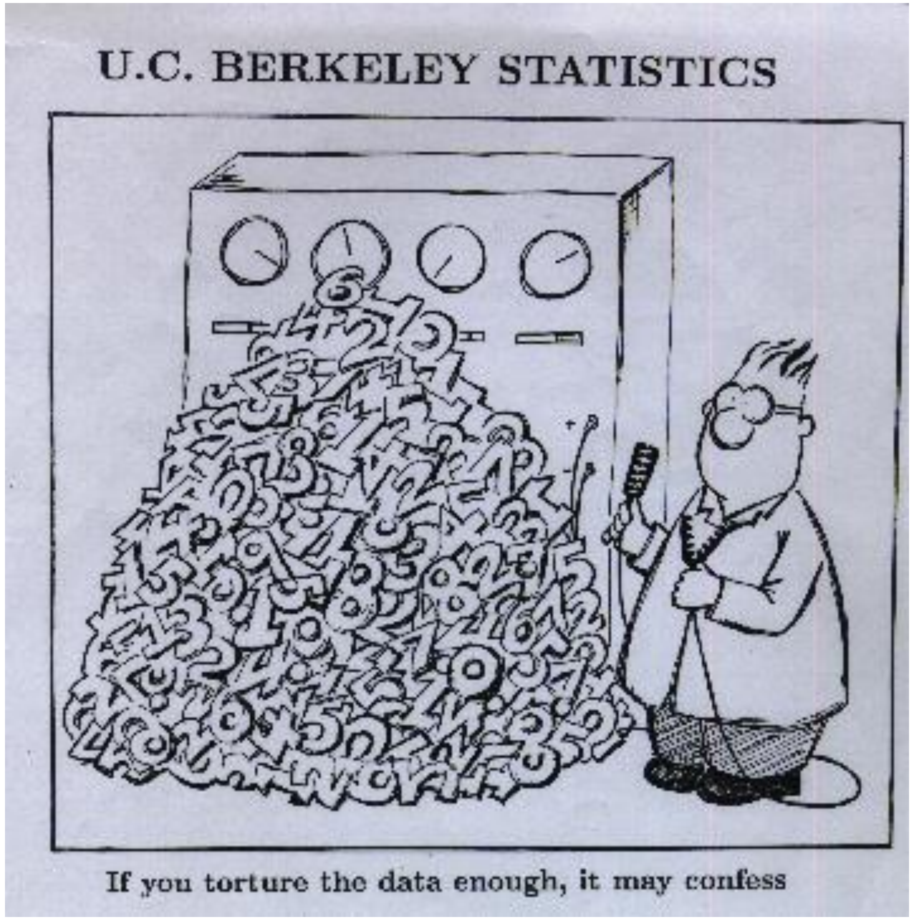
Bivariate Graphical Displays

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

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.