Graphics

Data Types & Plot Types

Why is graphics in this course?

- Good graphics today requires the computer
- Visualization enters every step of the data analysis cycle
  - Data cleaning – are there anomalies?
  - Exploration
  - Model checking
  - Reporting results
- Plots can uncover structure in data that can't be detected with numerical summaries
- Important communication skill

Keep in Mind

- Meta Data: The source of information and the selection process for the observations
- Are these data representative of the population that you are trying to generalize to?
- What is a clear and informative way to present the data so that insights are readily discernable?

Know your data types

The appropriate graphical techniques depend on the kind of data that you are working with

- Quantitative
  - continuous – e.g., height, weight
  - discrete – numeric data with few values, e.g., number of children in family
- Qualitative
  - ordered – categories with an order but no meaningful distance between, e.g., number of stars for a movie rating
  - nominal - categories have no meaningful order, e.g., gender, race

Data Type can depend on

- Units of measurement
- What constitutes a record in the data
- These concepts are connected

What type of data is handedness?

A. Quantitative

B. Qualitative – nominal

C. Qualitative – ordinal

D. Possibly A or B

E. Possibly A or C

What type of data is income?

A. Quantitative

B. Qualitative – nominal

C. Qualitative – ordinal

D. Possibly A or B

E. Possibly A or C

Individual report the activities performed with left hand (write, eat, bat, sweep, etc.) and these are counted

What type of data is handedness?

A. Quantitative – discrete
B. Quantitative – contin
C. Qualitative – nominal
D. Qualitative – ordinal

Family income reported in a survey, choose from brackets, e.g. < $30,000, $30,000 - $45,000, etc

What type of data is income?

A. Quantitative – discrete
B. Quantitative - contin
C. Qualitative – nominal
D. Qualitative – ordinal

Consider sex as reported in the DAWN survey

What type of data?

A. Quantitative

B. Qualitative – nominal

C. Qualitative – ordinal

Consider sex as reported in World Bank Data on Countries

What type of data?

A. Quantitative

B. Qualitative – nominal

C. Qualitative – ordinal

Different Plots for Different Data Types

load(url("http://www.stat.berkeley.edu/users/nolan/data/babiesLab133.rda"))

Kaiser Study

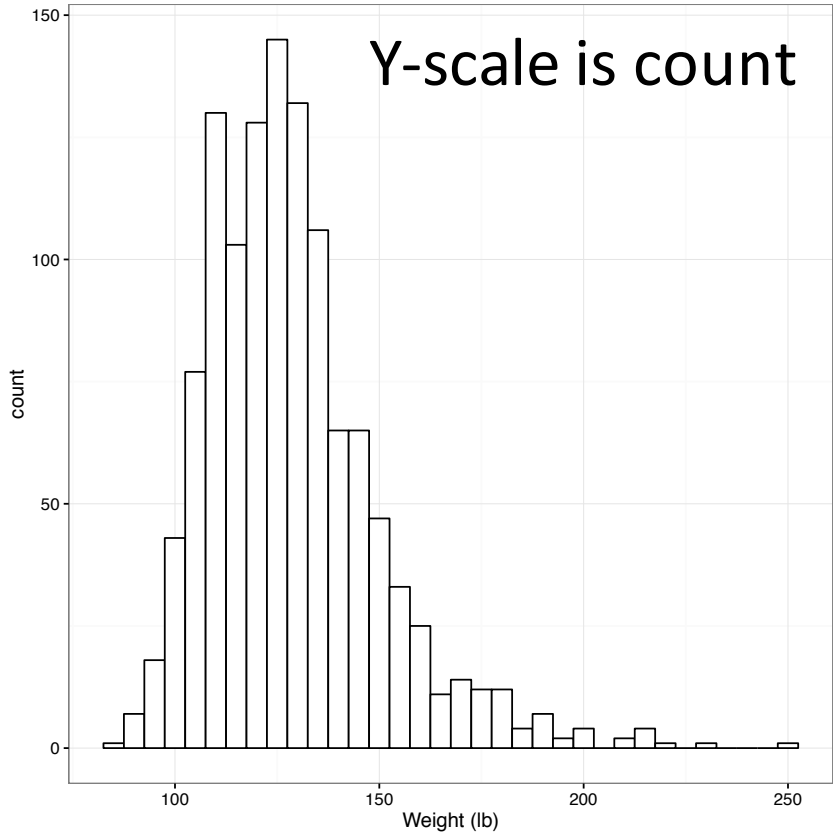
- Oakland Kaiser mothers
- 1960s
- Measure the babies weight (in ounces) at birth
- All babies:
  - Male
  - Single births (no twins, etc.)
  - Survived 28 days

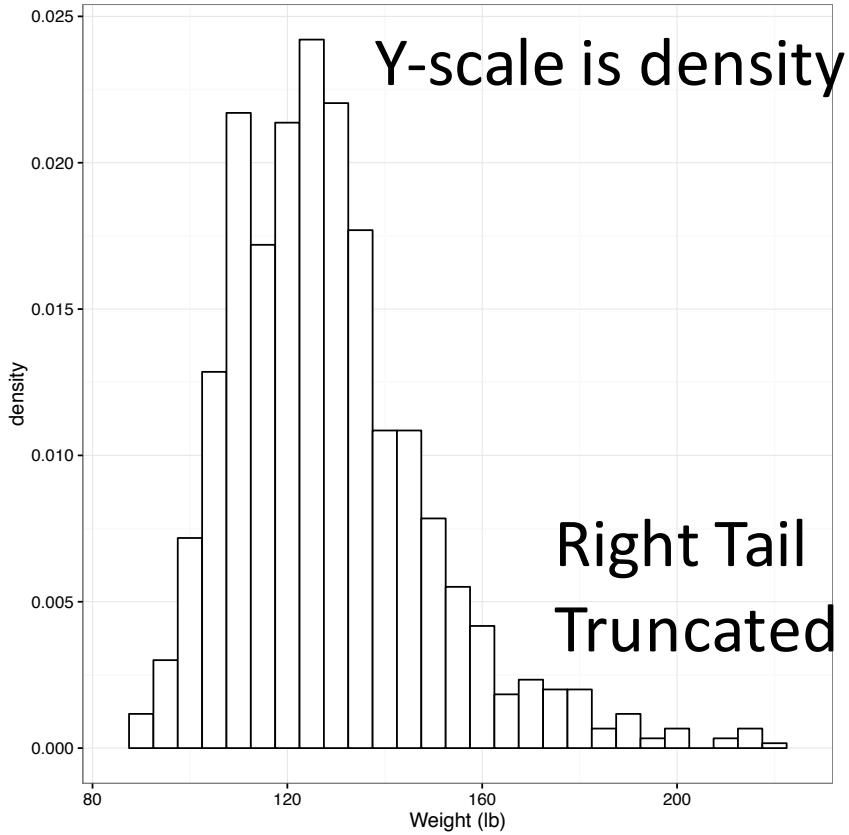
Information collected on mother's and their babies

- 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
- Father's information and more…

One Quantitative Variable

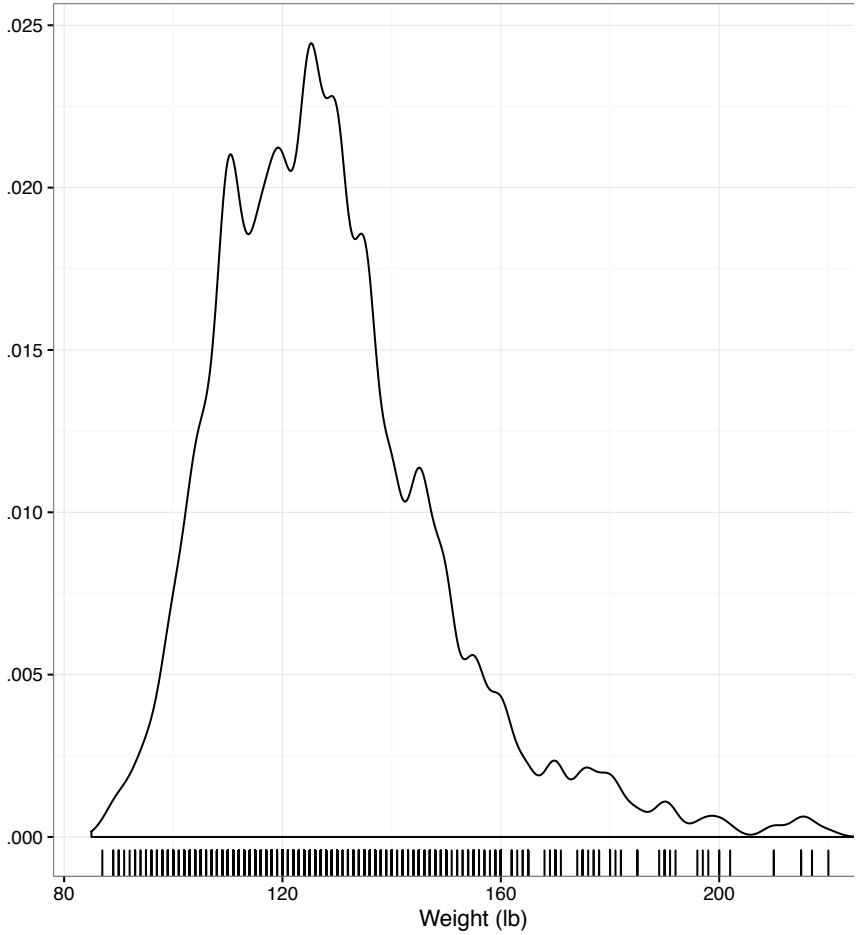
Histogram – Mother's weight





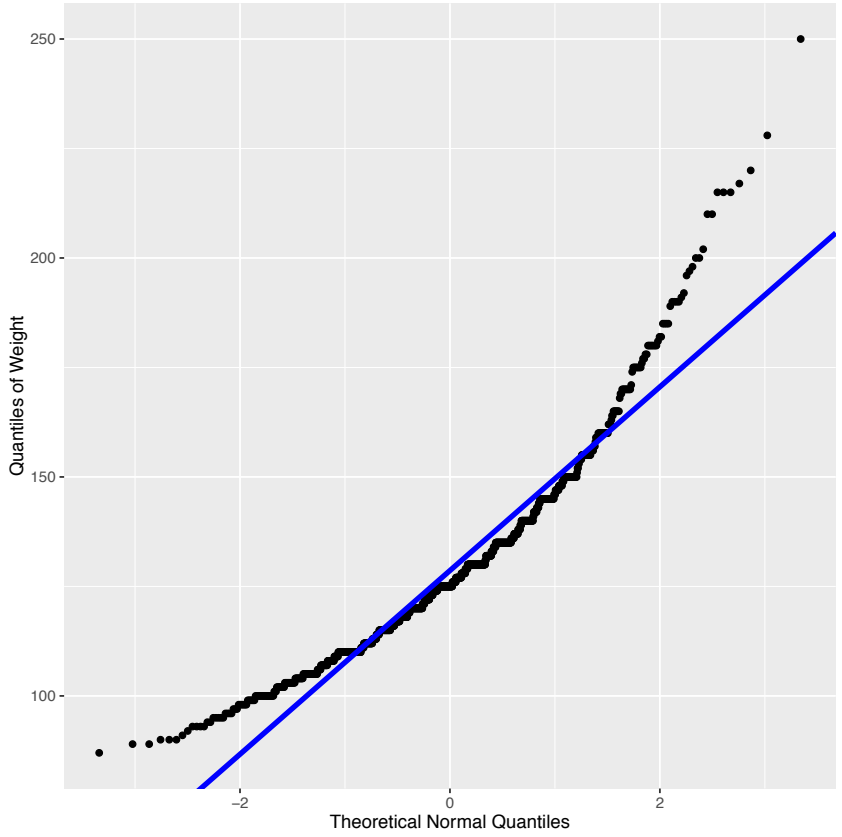
What is the difference between these 2 histograms?

Density curve



- Band width is small so see fluctuations in individual values
- Rug plot thickness matches these little peaks

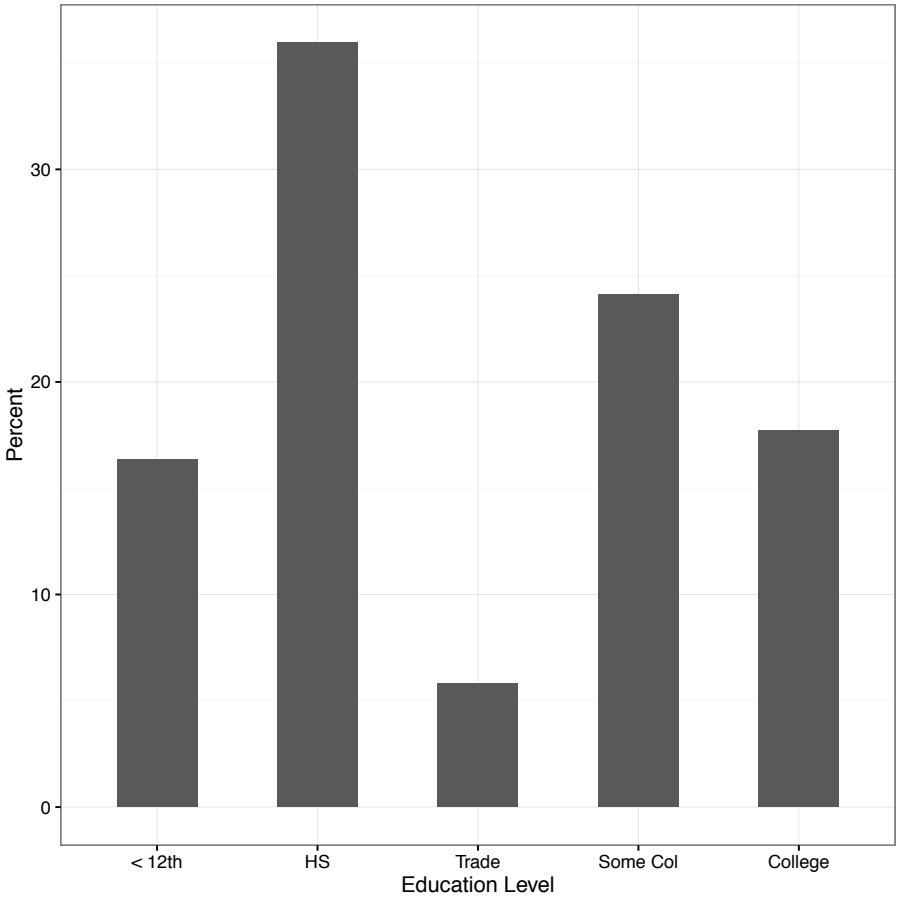
Quantile Plot

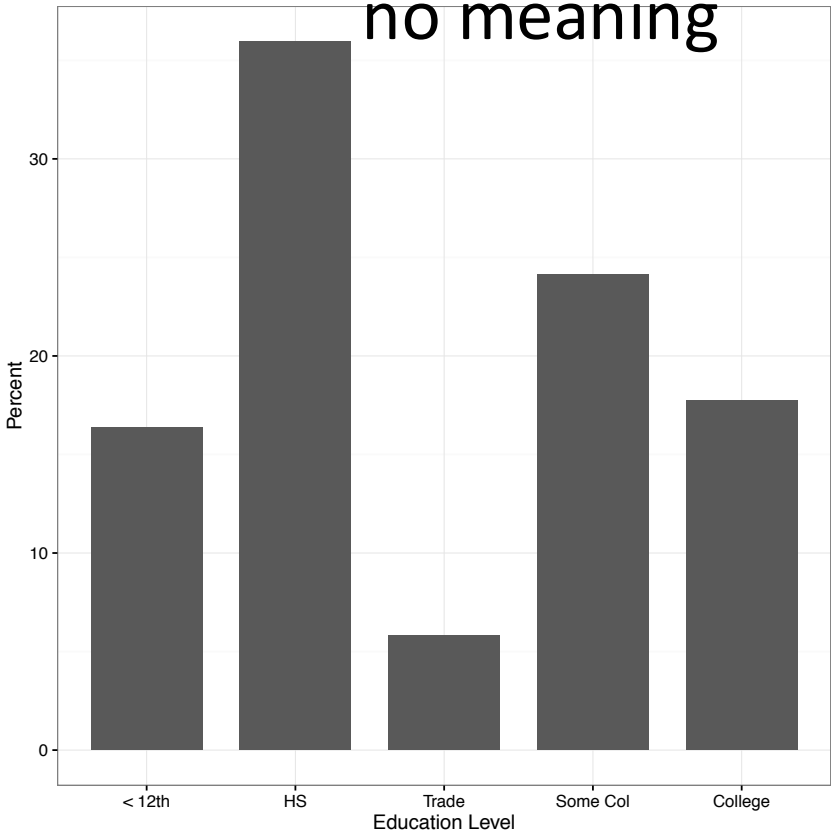


- Compare the distribution to a theoretical one
- Upward curve for small values indicates a short left tail
- Upward curve for large values indicates a long right tail

Bar plot - Education Level

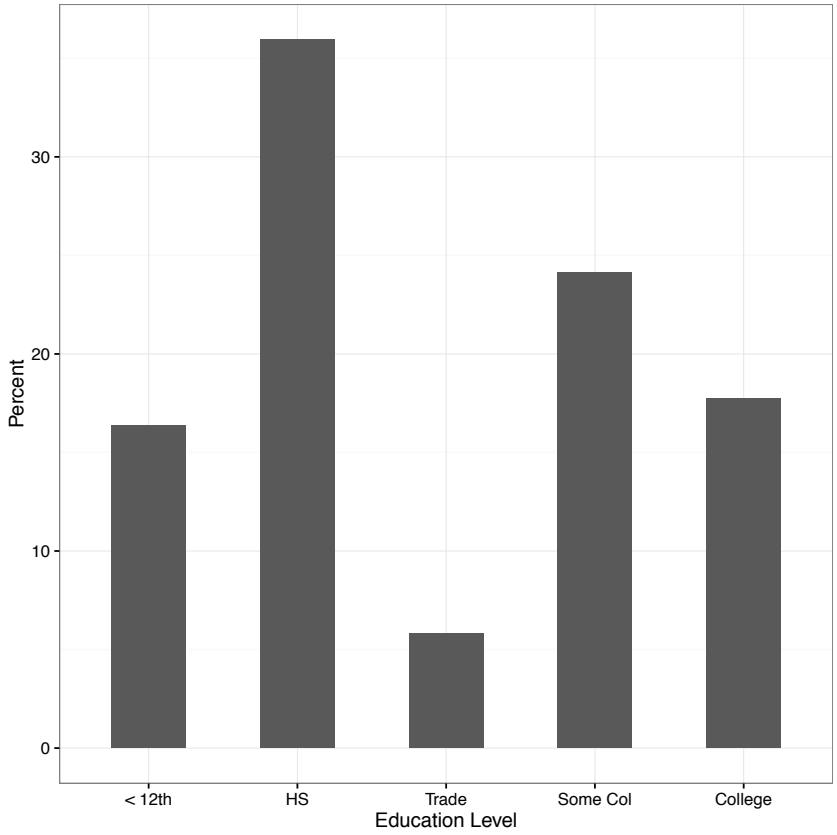
Bar Width – has no meaning

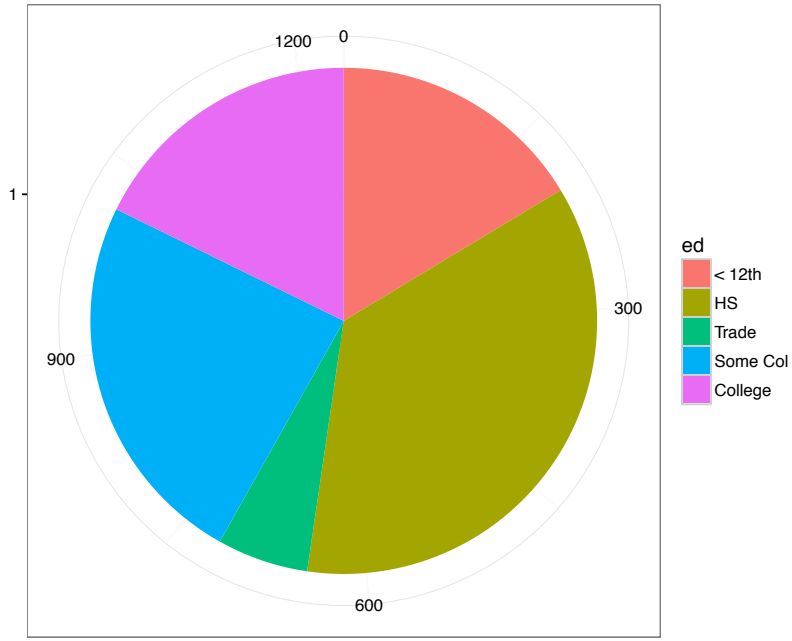




What's the difference between these 2 plots?

Pie chart - Education Level

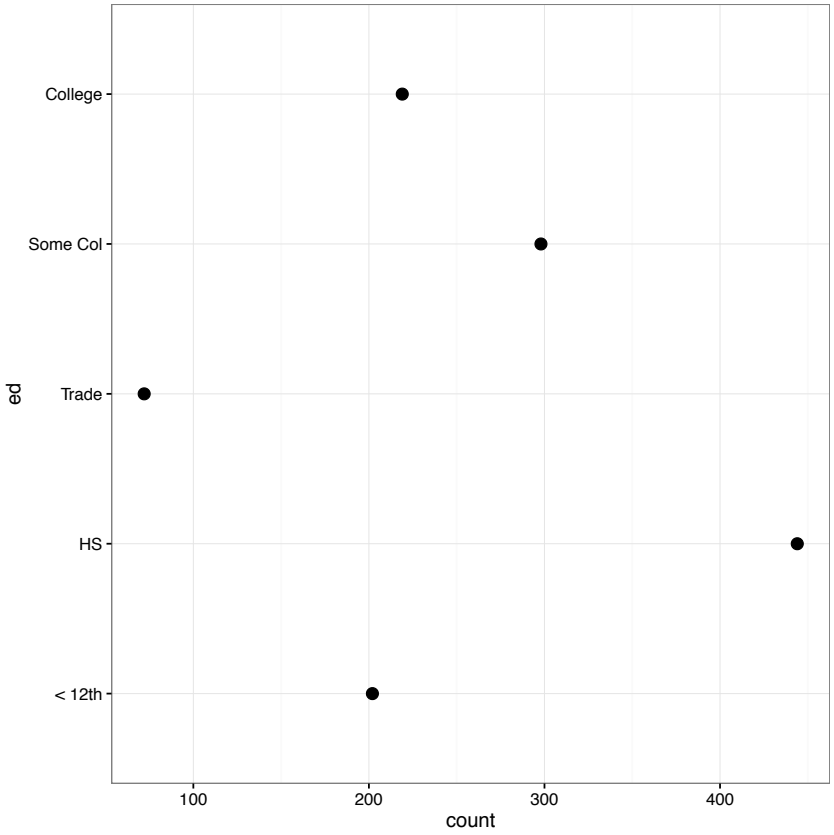




Easier to compare heights of bars than angles

Dot Chart - Education Level

- Width of bars in a bar plot have no meaning
- Dot plot (aka Cleveland) focus on comparison of the values



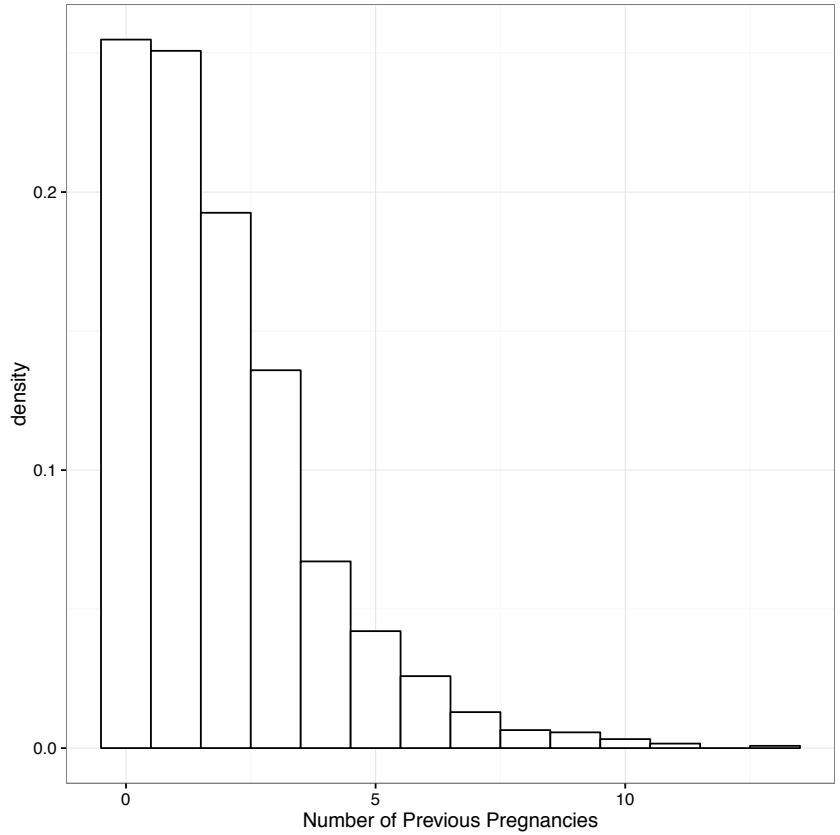
Discrete Quantitative Variable can sometimes look like a Qualitative Variable

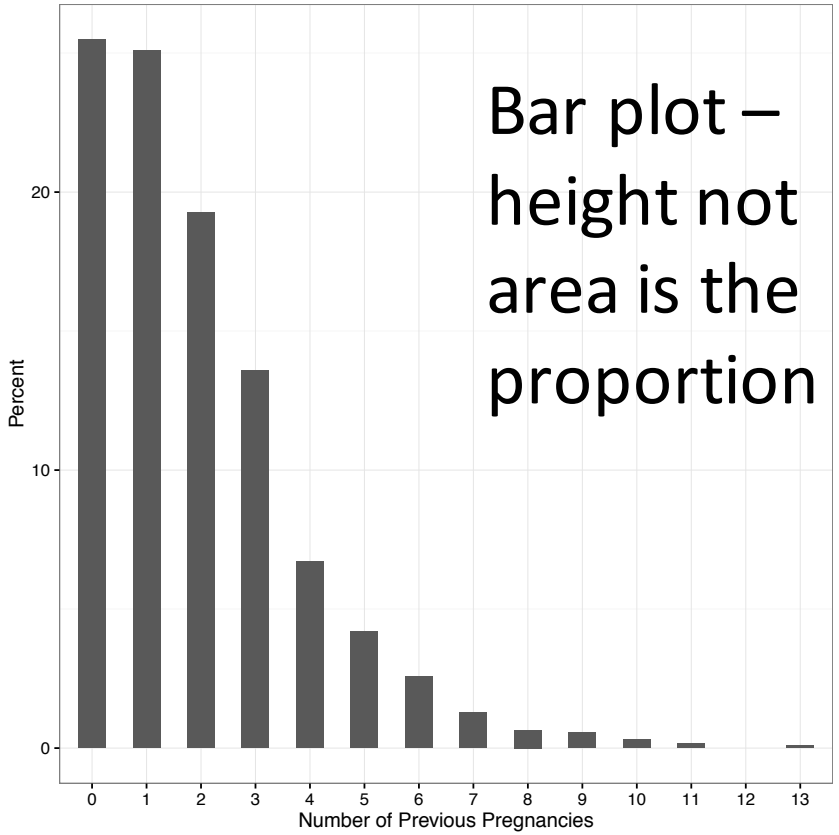
Parity: Number of siblings

- This quantitative variable is different from birth weight – there are only a few possible values, i.e., it's not possible to have 2.3 siblings, and it's highly unlikely to have 17

```
> table(infants$parity)
  0   1   2   3   4   5   6   7   8   9  10  11  13
315 310 238 168  83  52  32  16   8   7   4   2   1
```

Number of Previous Pregnancies





Bar plot – height not area is the proportion

What's the difference between these 2 plots?

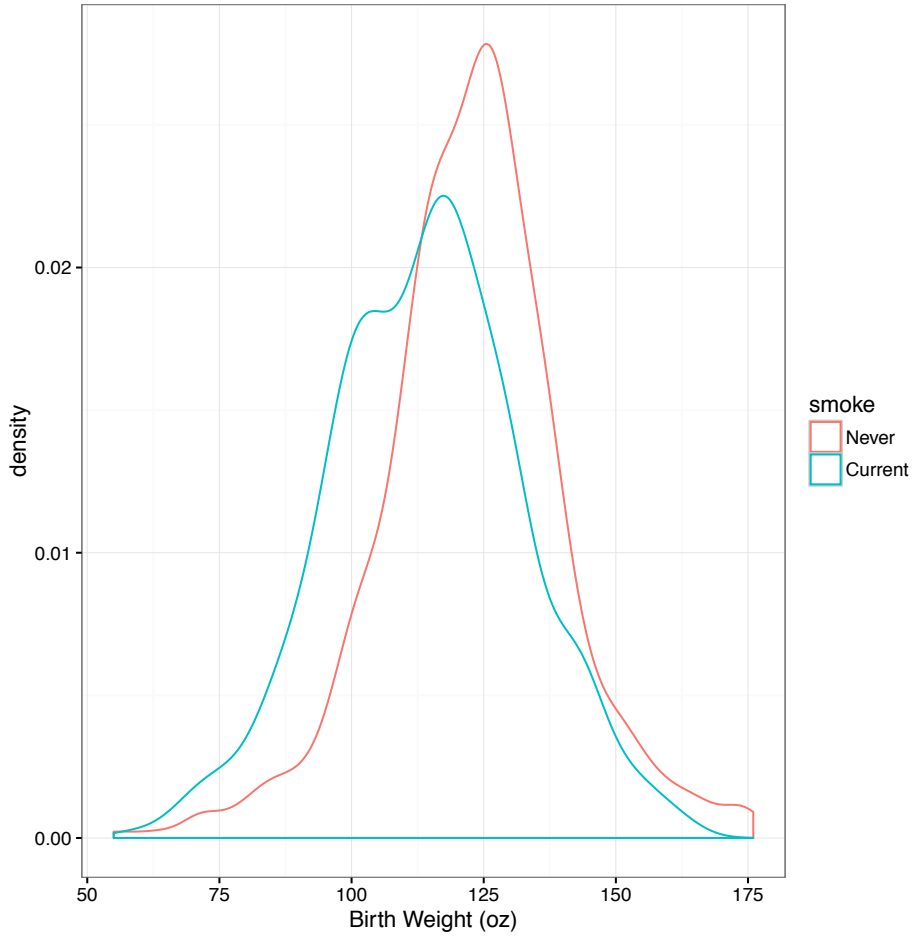
Method of Comparison

- Often, we not only want to better understand a distribution, but we want to compare the distribution for subgroups or to compare against another population or standard

- How do you think the birth weight distribution might vary with smoking status?

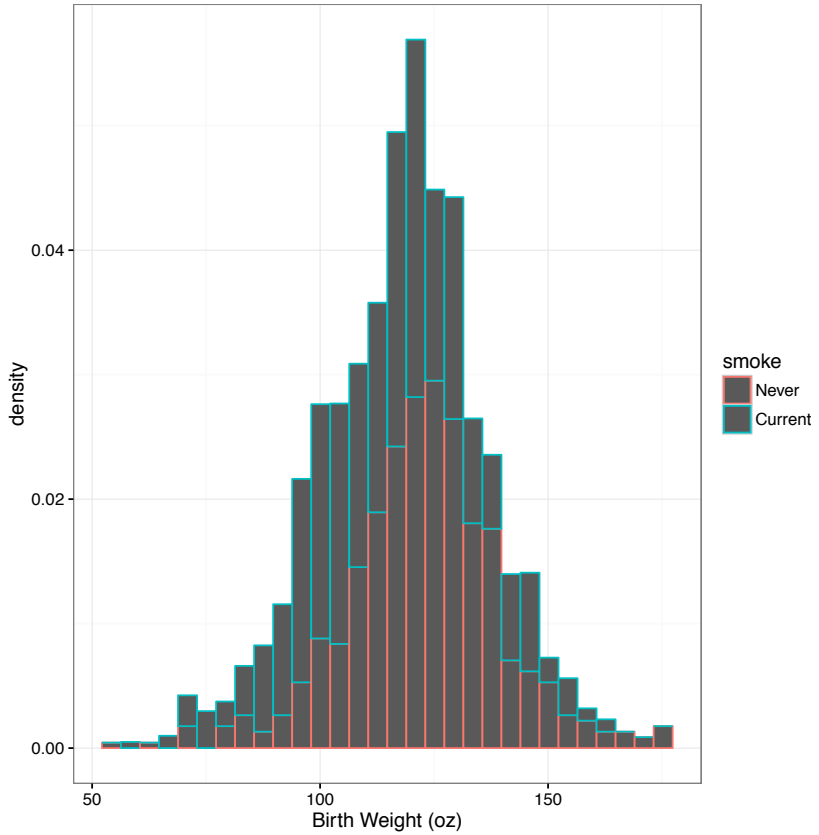
One Quantitative Variable and One Qualitative Variable

Super-posed Density Plots – one per level

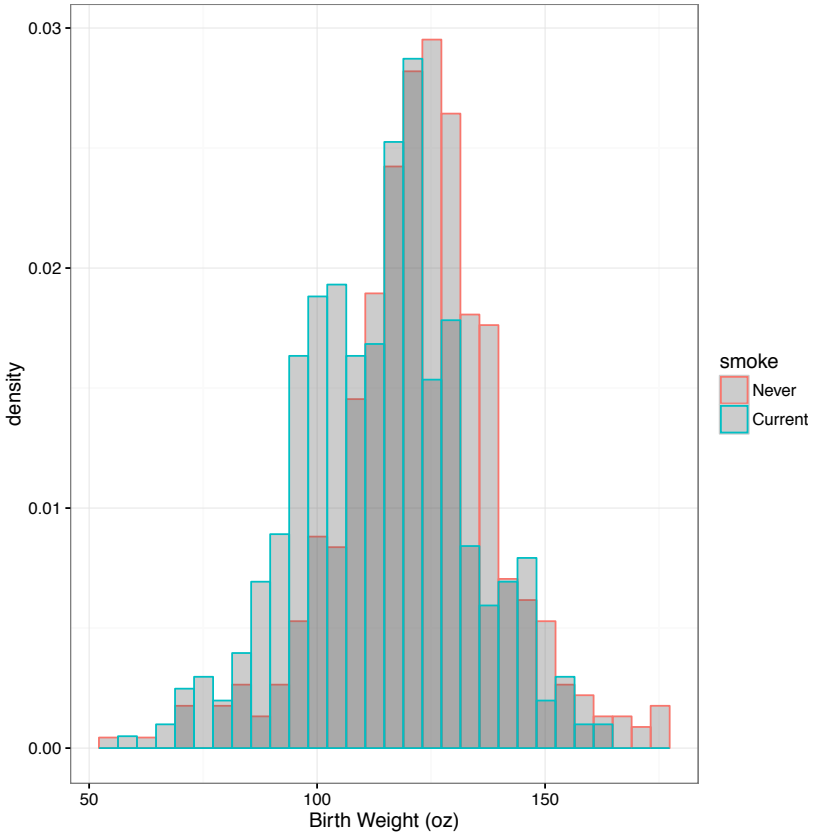


Multiple histograms on 1 plot

Stacked Bars

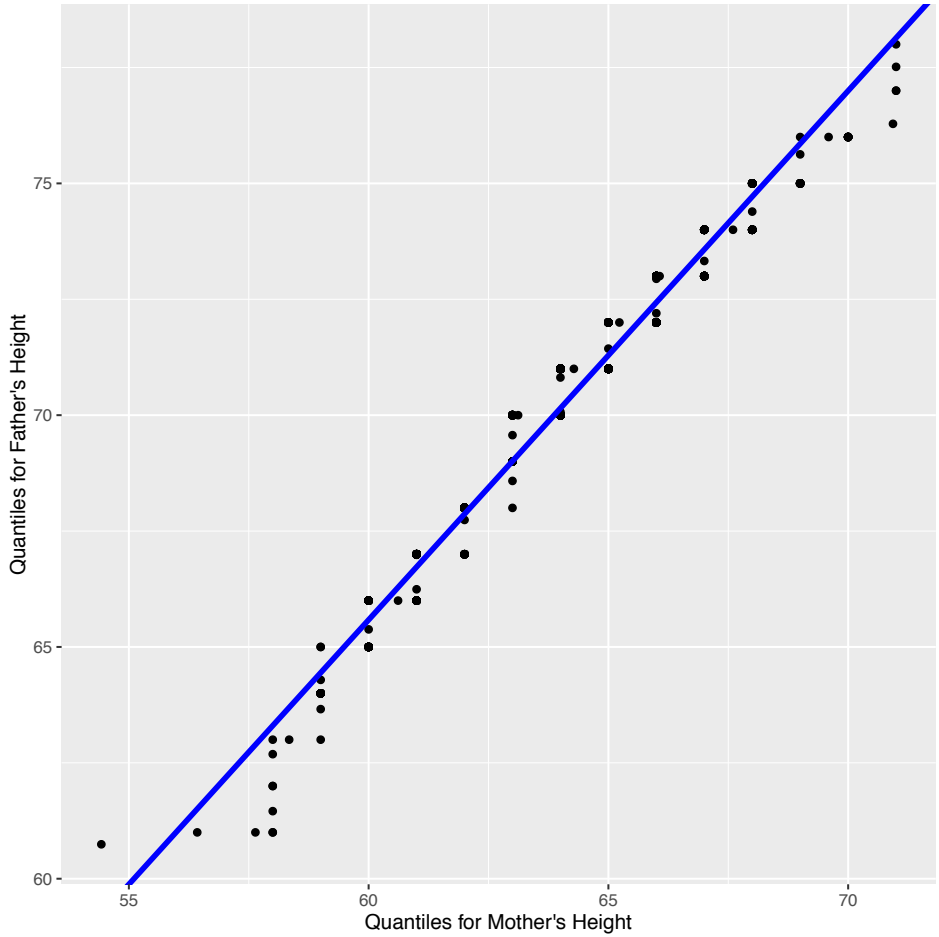


Overlaid bars

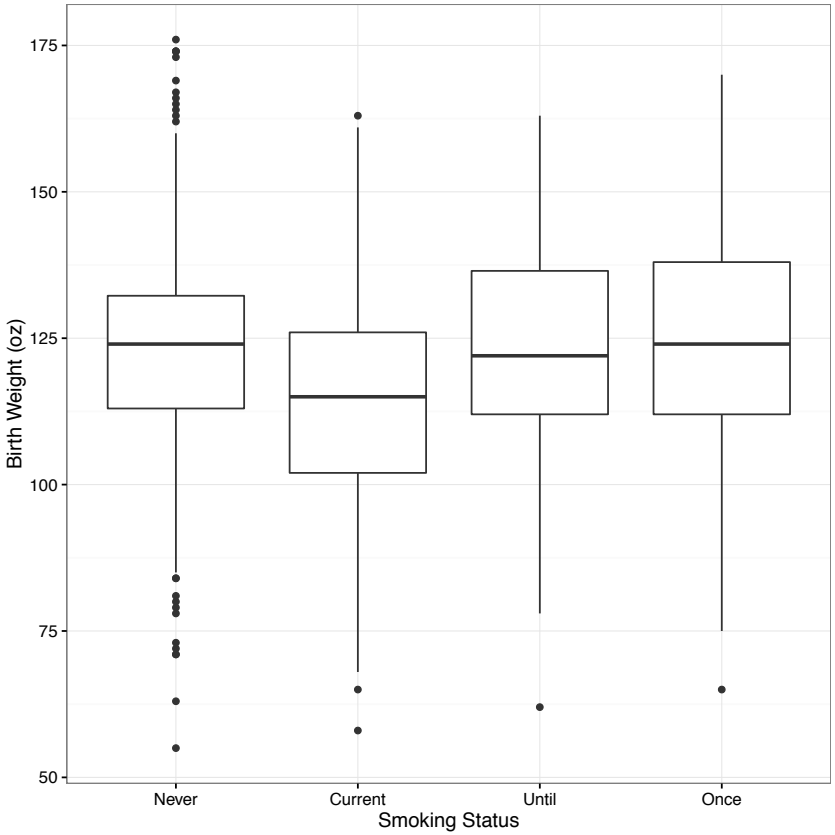


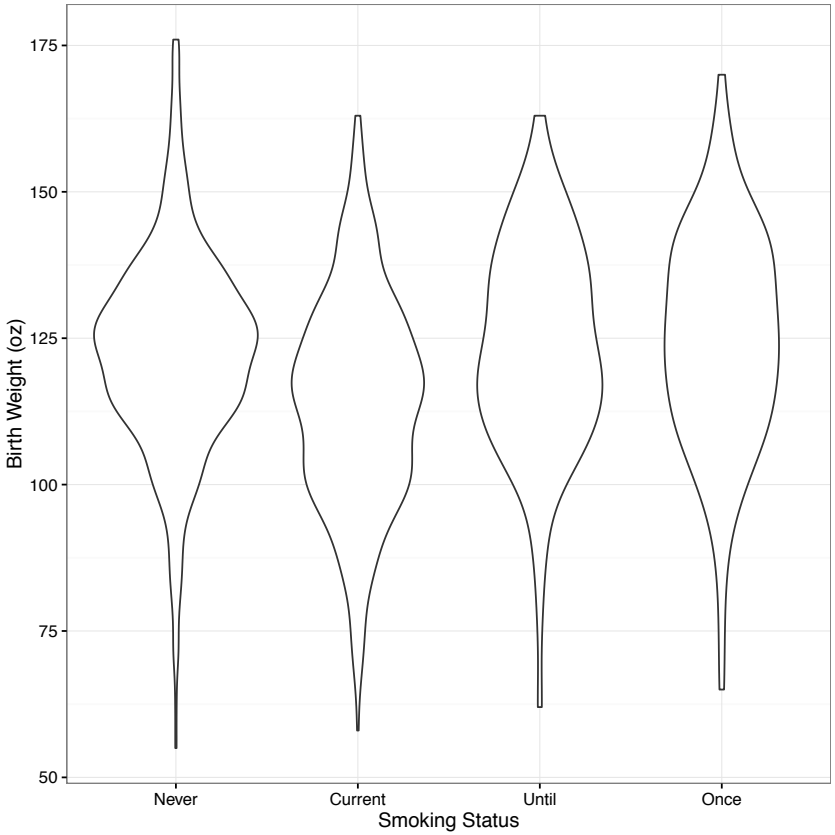
What the difference between these 2 plots?

Quantile – Quantile Plot

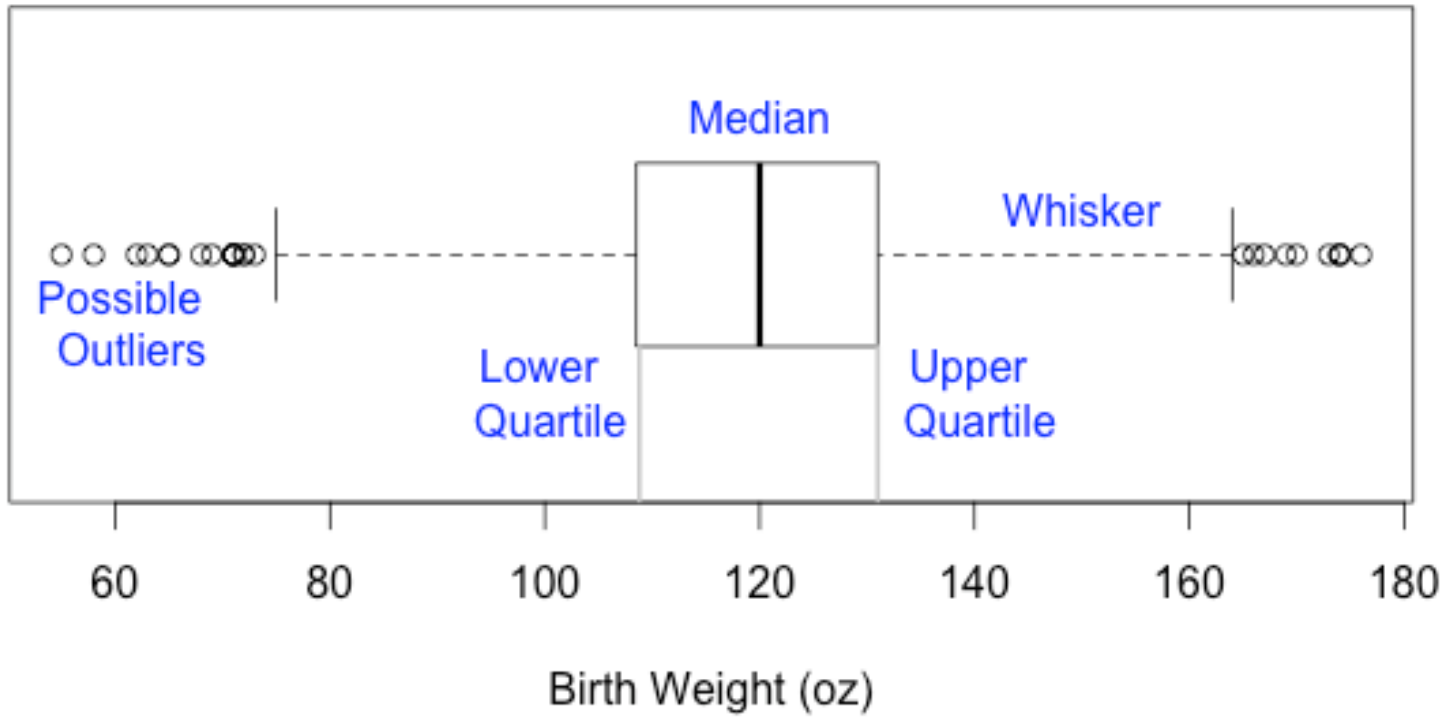


Side-by-side Boxplots & Violin Plots



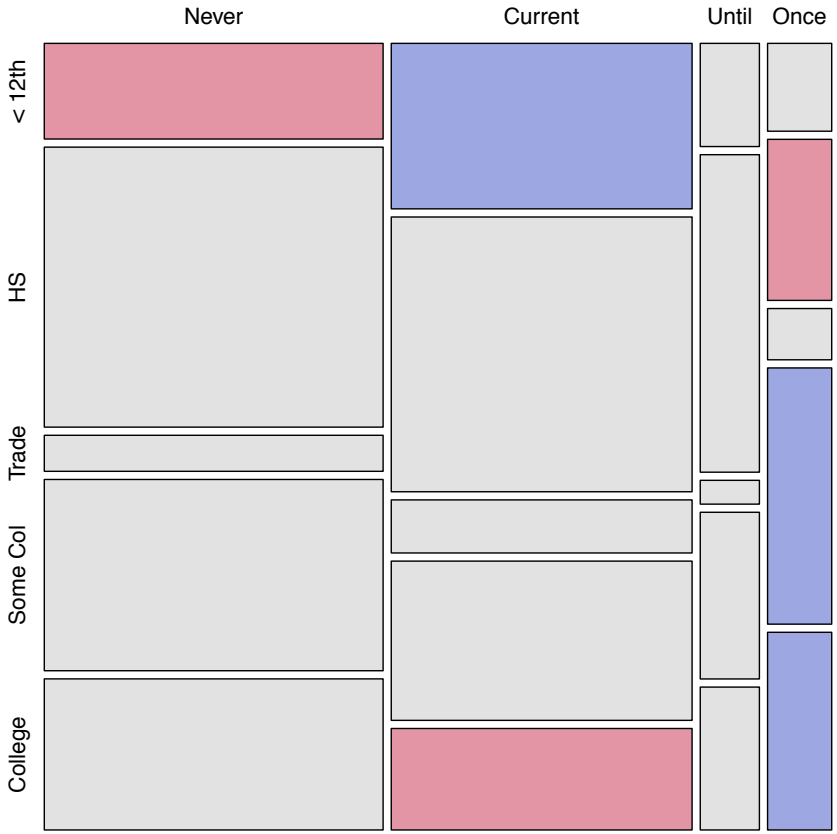


Boxplot Definition



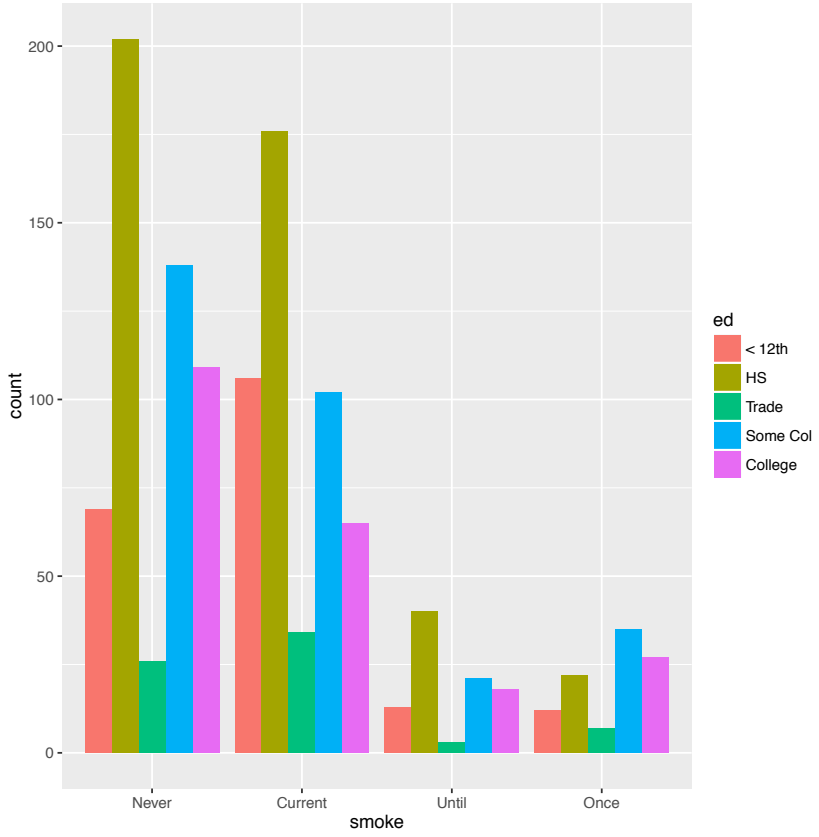
Two Qualitative Variables

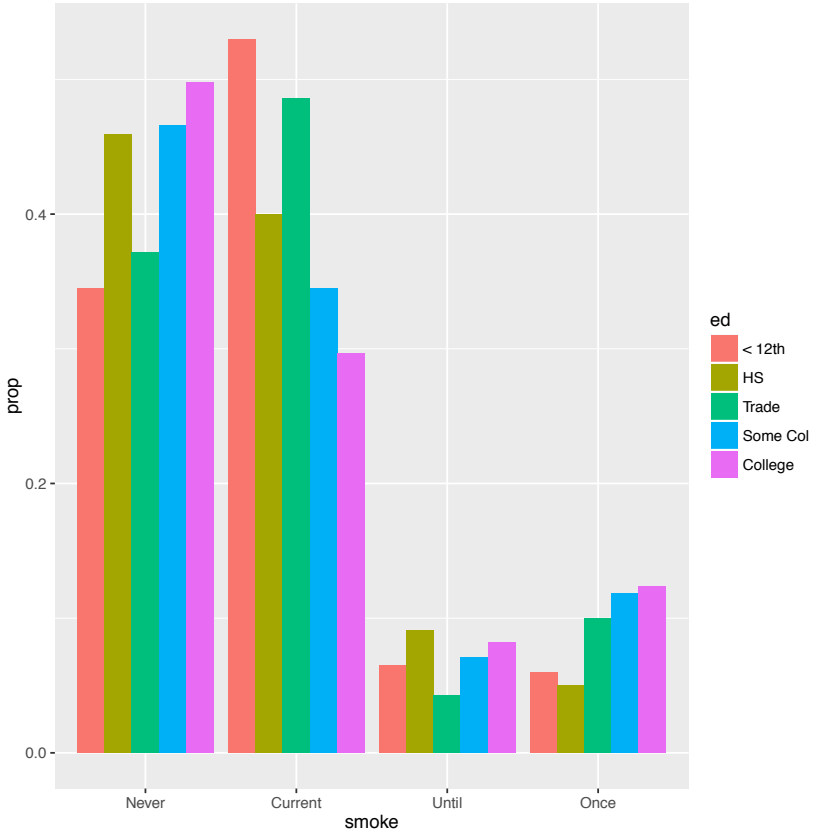
Mosaic Plot - Education and Income



Side-by-side Bar Plot

Smoking status normalized within Education level

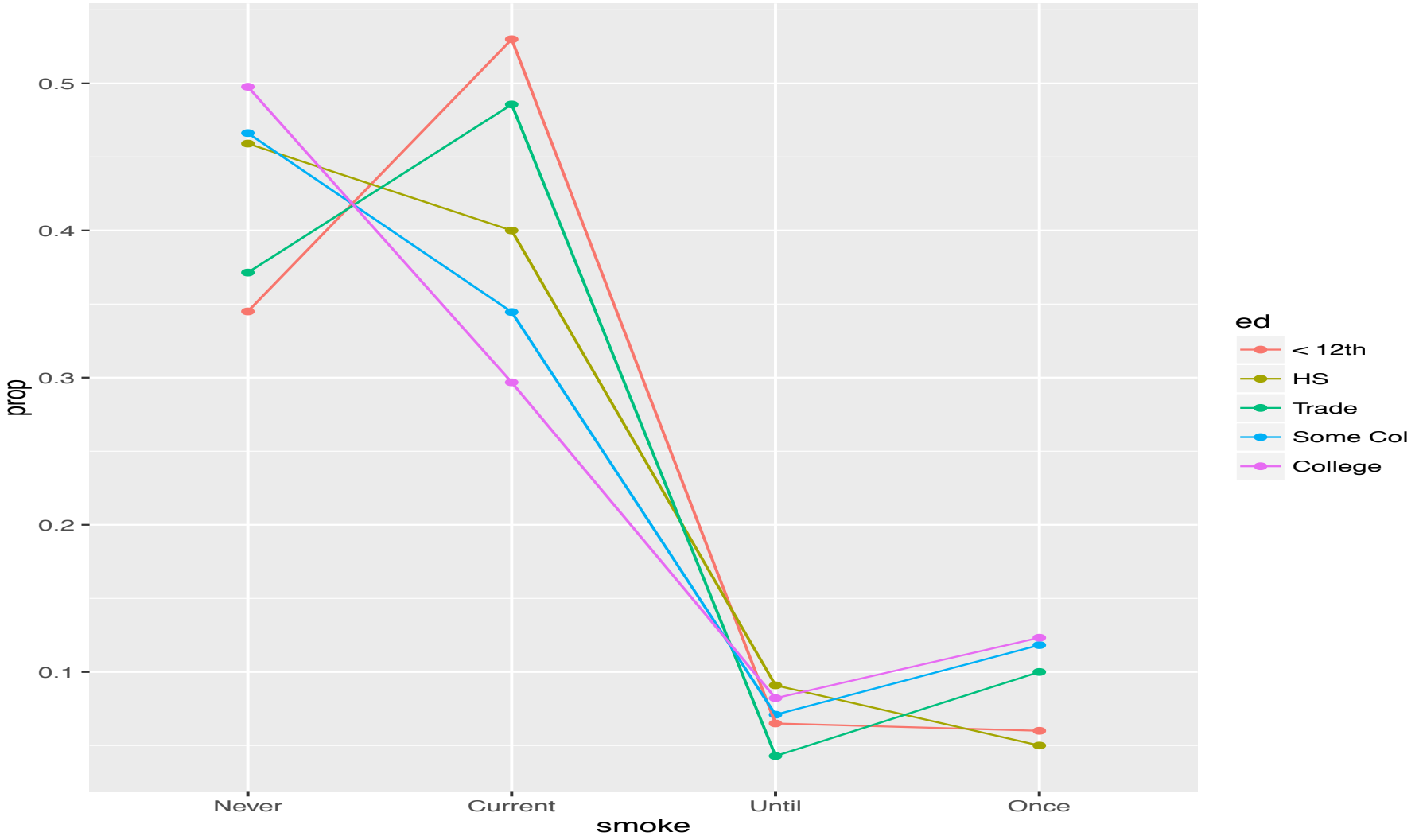




What's the difference between these 2 plots?

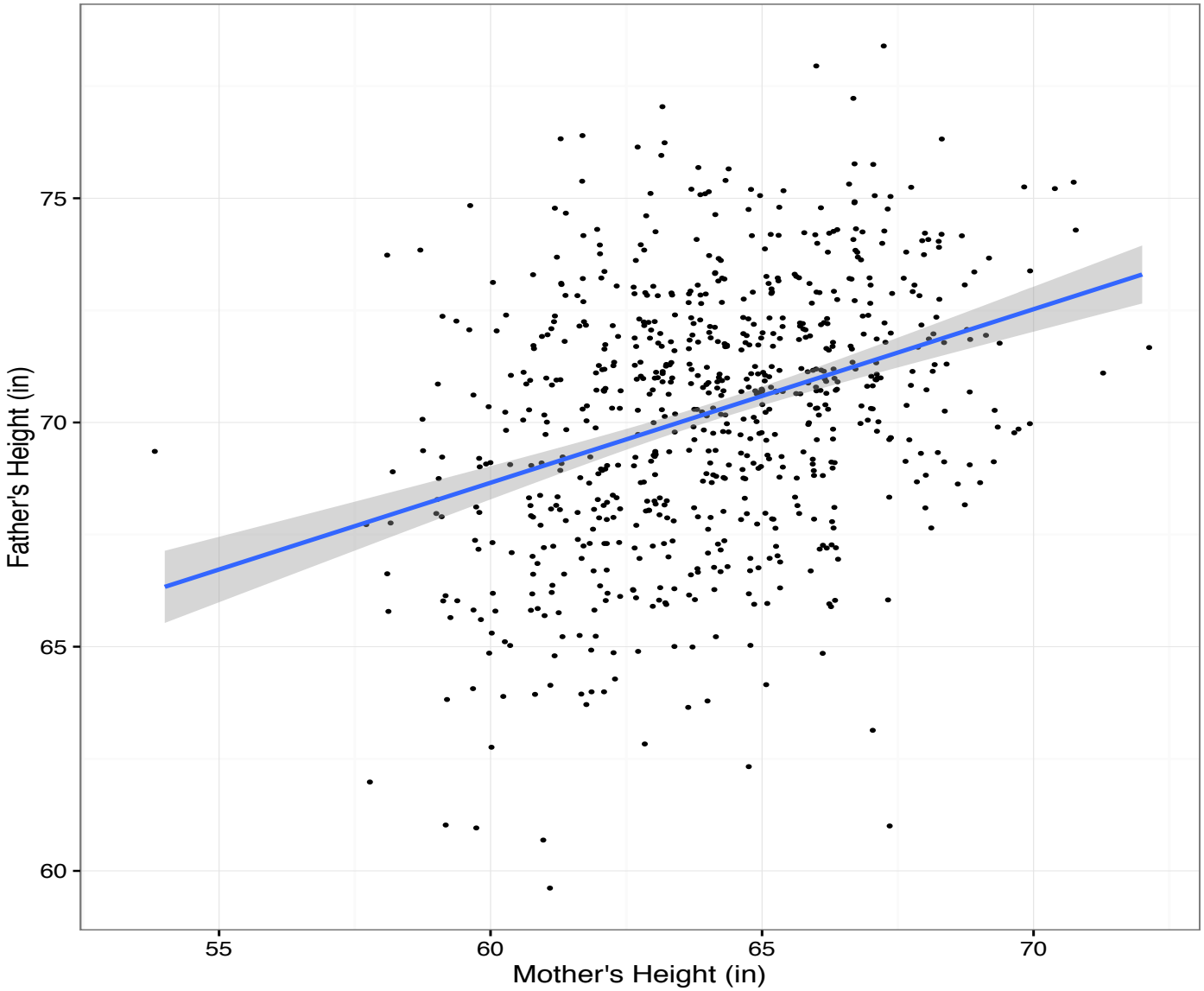
Interaction Plot

Smoking status normalized within Education level



Two Quantitative Variables

Scatter Plot and Smooths



This smooth is a linear fit

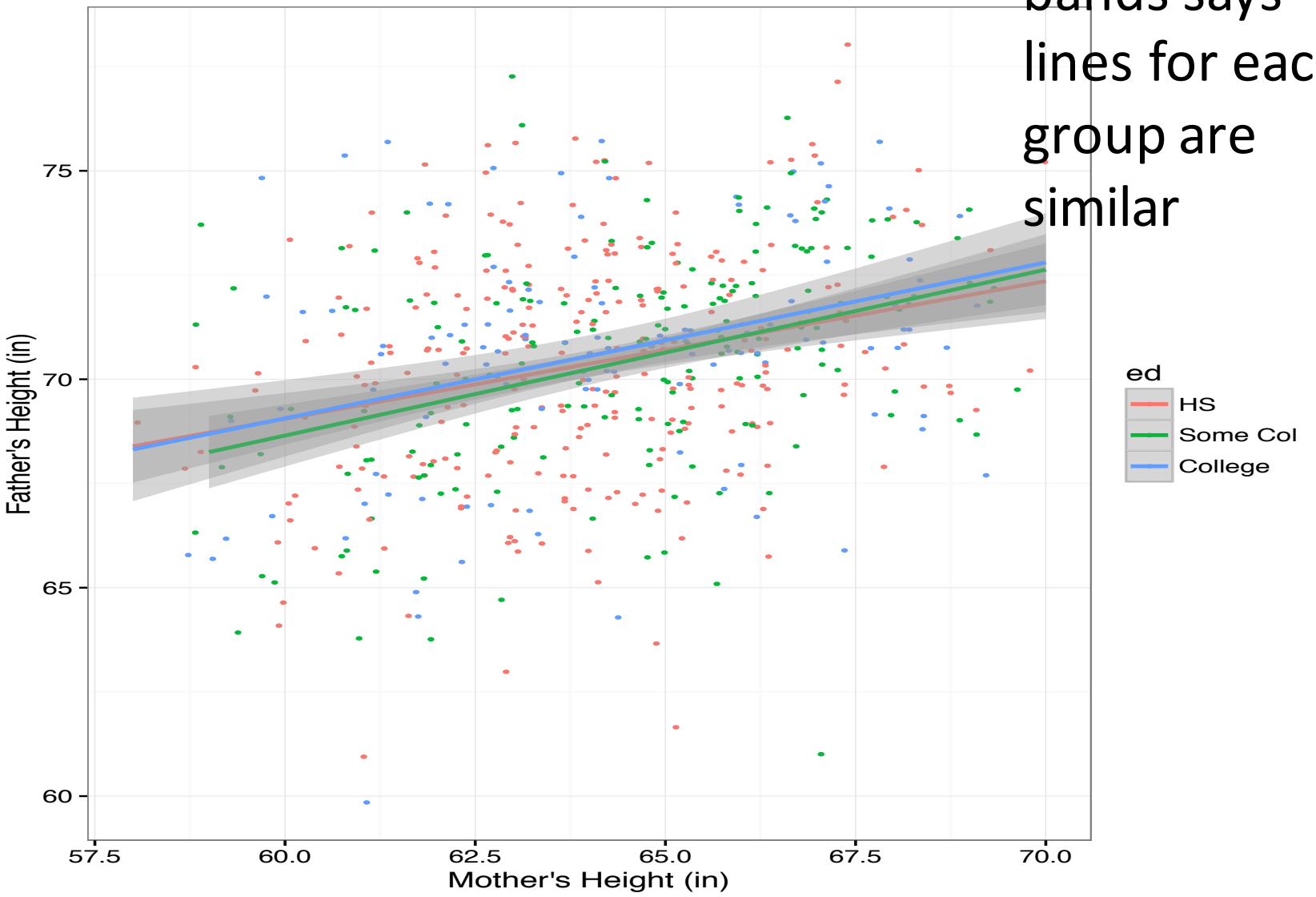
Gray bands give a sense of accuracy of linear fit

Relationships between more than 2 variables

- Qualitative information can be conveyed in plots through color, plotting symbol, juxtaposed panels

2 Quant + 1 Qual

Overlap in Gray bands says lines for each group are similar



Summary of graph relationships between two variables

- Two Qualitative variables
  - Mosaic plot, side-by-side barplots (watch normalization), interaction plot
- One Quantitative and one Qualitative
  - Side-by-side boxplots, violin plots, dotcharts, super-posed density curves, qq-plot
- Two Quantitative variables
  - Scatter plot, line plot (time), smooths