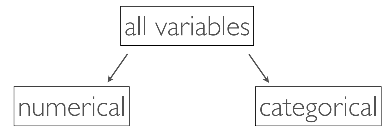
Lecture 2 - Data and Data Summaries

Sta102 / BME 102

Colin Rundel

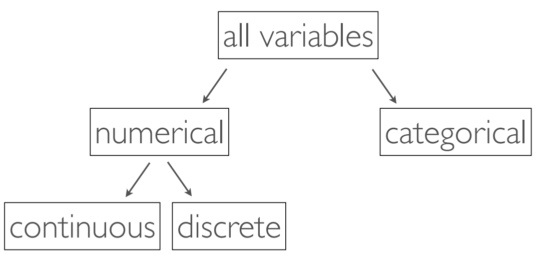
January 14, 2015

Data



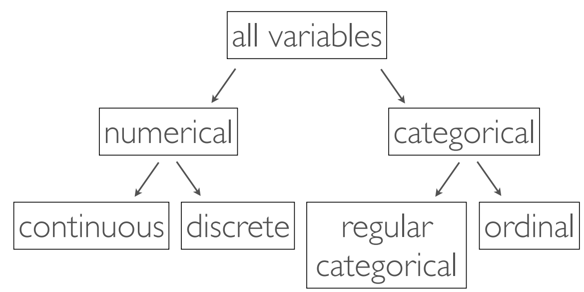
- *Numerical (quantitative)* - takes on a numerical values
  - Ask yourself - is it sensible to add, subtract, or calculate an average of these values?
- *Categorical (qualitative)* - takes on one of a set of distinct categories
  - Ask yourself - are there only certain values (or categories) possible? Even if the categories can be identified with numbers, check if it would be sensible to do arithmetic operations with these values.

Numerical Data



- *Continuous* - data that is measured, any numerical (decimal) value
- *Discrete* - data that is counted, only whole non-negative numbers

Categorical Data



- *Ordinal* - data where the categories have a natural order
- *Regular categorical* - categories do *not* have a natural order

Example - Class Survey

Students in an introductory statistics course were asked the following questions as part of a class survey:

1. What is your gender?
2. Are you introverted or extraverted?
3. On average, how much sleep do you get per night?
4. When do you go to bed: 8pm-10pm, 10pm-12am, 12am-2am, later than 2am?
5. How many countries have you visited?
6. On a scale of 1 (very little) - 5 (a lot), how much do you dread this semester?

What type of data is each variable?

Representing Data - Class Survey

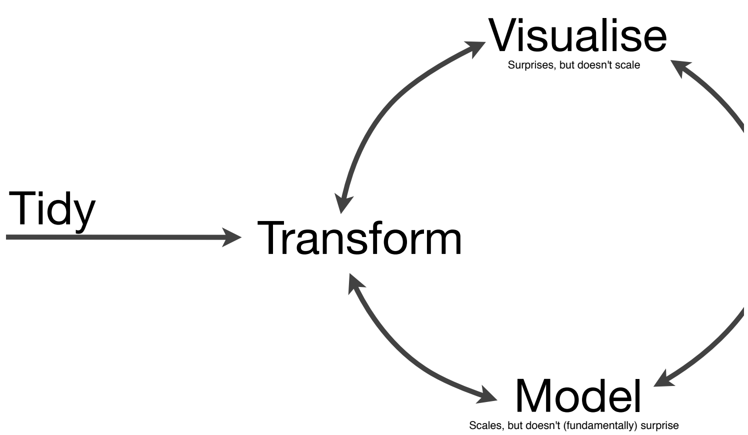
We use a *data matrix (data frame)* to represent responses from this survey.

- Columns represent *variables*
- Rows represent *observations (cases)*

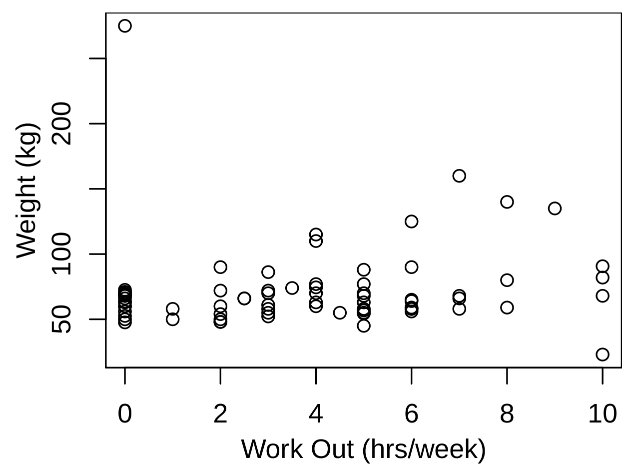
student	gender	intro_extra	sleep	bedtime	countries	dread
1	male	extravert	8	10-12	13	3
2	female	extravert	8	8-10	7	2
3	female	introvert	5	12-2	1	4
4	female	extravert	6.5	12-2	0	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
86	male	extravert	7	12-2	5	3

Visualization and Statistics

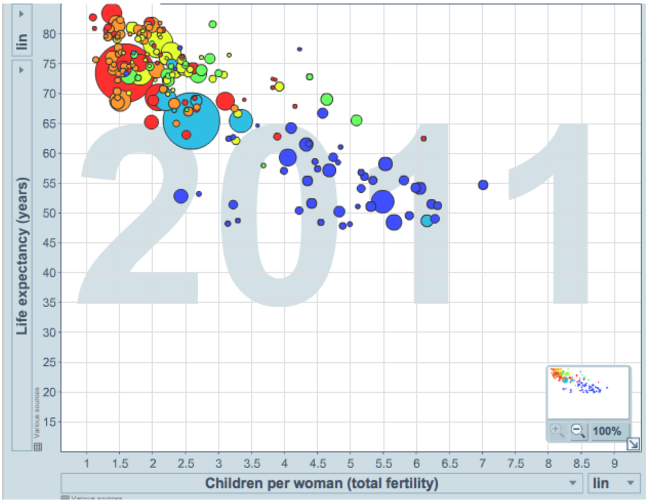
Hadley Wickham's Data Cycle



Scatterplots



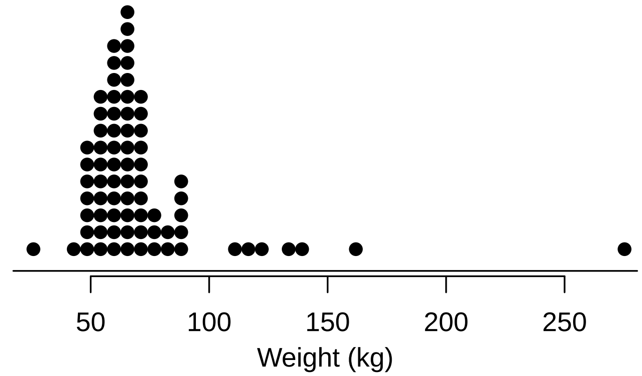
Scatterplots (Fancy)



http://www.gapminder.org/world

Dot plots

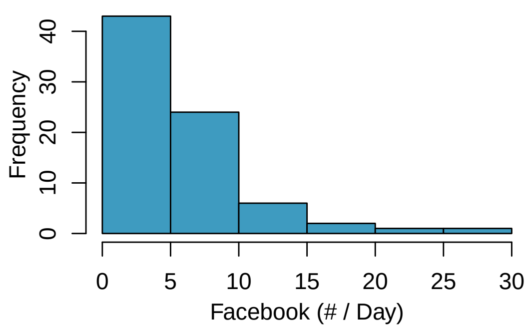
Useful for visualizing a single numerical variable, especially useful when individual values are of interest.



Do you see anything out of the ordinary?

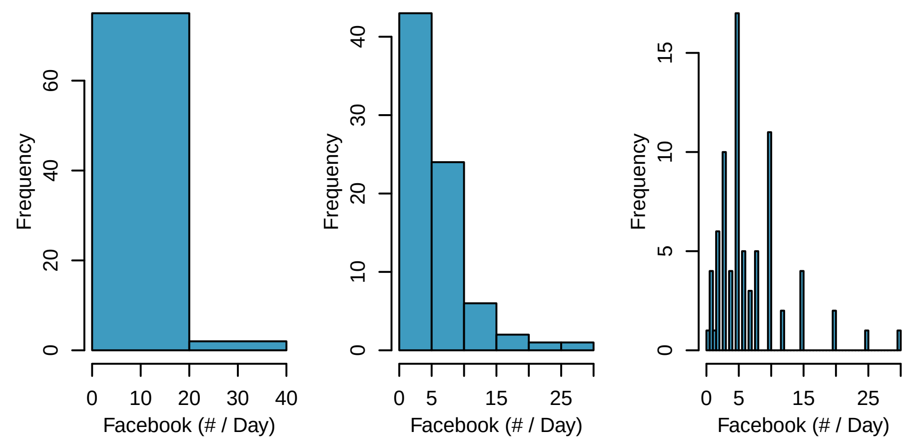
Histograms

- Preferable when sample size is large but hides finer details like individual observations.
- Histograms provide a view of the data's *density*, higher bars represent where the data are more common.
- Histograms are especially useful for describing the *shape* of the distribution.



Bin width

The chosen *bin width* can alter the story the histogram is telling.



Which histogram is the most useful? Why.

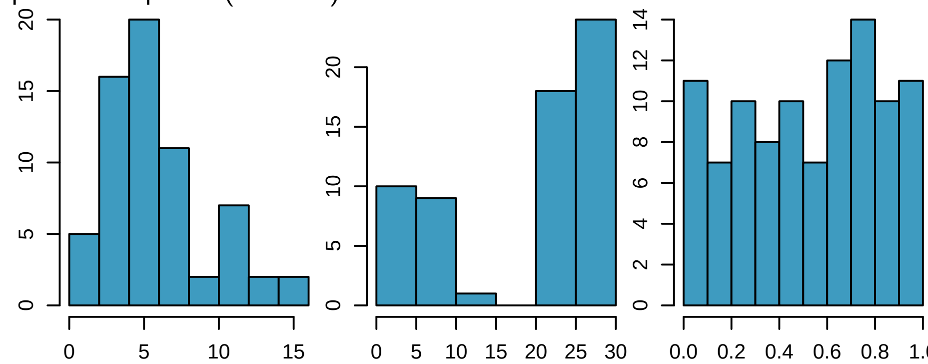
Describing Distributions

We would like a standardized way to describe the shape of a distribution. There are several critical features that we need to describe:

- Center
- Spread
- Modality (peaks)
- Skewness (asymmetry)
- Kurtosis (peakedness)

Modality

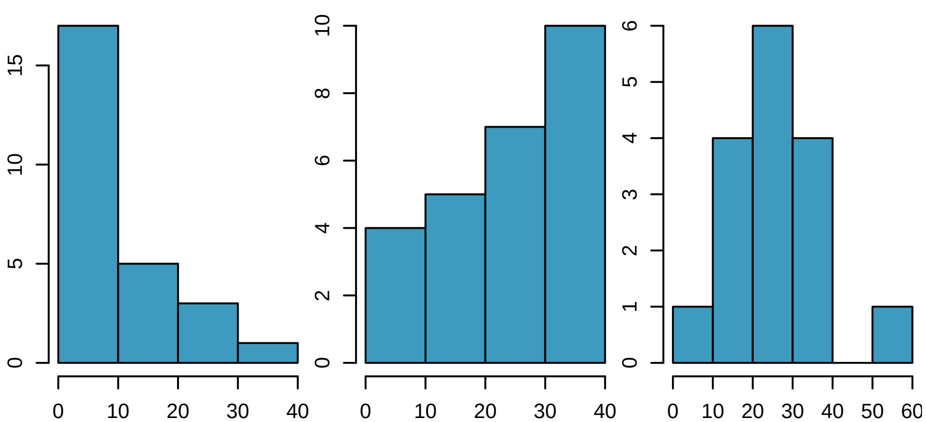
This describes the pattern of the peaks in peaks in the histogram - a single prominent peak (*unimodal*), several (*bimodal/multimodal*), or no prominent peaks (*uniform*)?



Note: In order to determine modality, it's best to step back and imagine a smooth curve (limp spaghetti) over the histogram.

Skewness

Histograms are said to be skewed towards the direction with the longer tail. A histogram can be *right skewed*, *left skewed*, or *symmetric*.



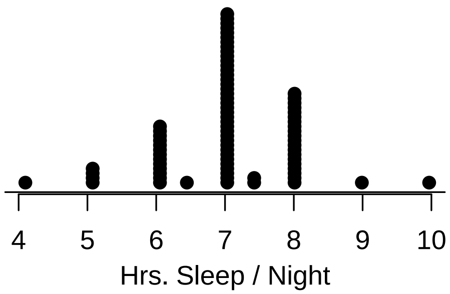
Thinking about distributions

How are of the following variables likely to be distributed?

1. weights of adult males
2. salaries of a random sample of workers in North Carolina
3. exam scores in Sta 102?
4. birthdays of classmates (day of the month)
5. weights of adults

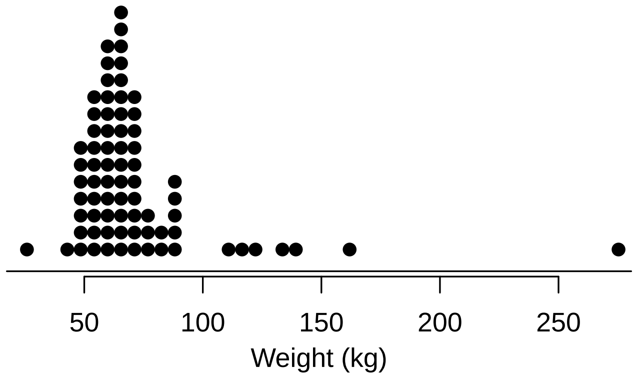
Guess the center

What would you guess is the average numer of hours students sleep per night?



Guess the center, cont.

What would you guess is the average weight of students?



Mean

- *Sample mean* ($\bar{x}$) - Arithmetic average of values in sample.

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + x_3 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- *Population mean* (μ) - Computed the same way but it is often not possible to calculate μ since population data is rarely available.

$$\mu = \frac{1}{N}(x_1 + x_2 + x_3 + \cdots + x_N) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- The sample mean is a *sample statistics*, or a *point estimate* of the population mean. This estimate may not be perfect, but if the sample is good (representative of the population) it is usually a good guess.

Variance

- *Sample Variance* - Average* deviance from the sample mean.

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

- *Population Variance* - Average deviance from the population mean.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

- The sample variance is a *sample statistics*, or a *point estimate* of the population variance.
- Variance is a measure of spread, the wider the distribution the larger the variance will be

Square Deviance?

Why do we use the squared deviation in the calculation of variance?

Standard deviation

- *Sample Standard Deviarton*

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

- *Population SD*

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

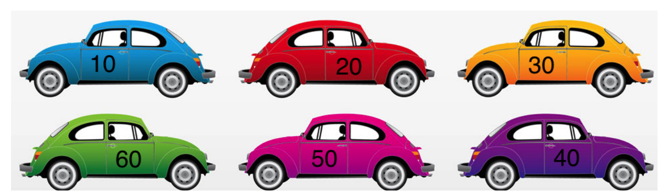
- Just like Variance, SD is a measure of spread, the wider the distribution the larger the SD will be.
- We often prefer SD to Variance because it has a more natural interpretation. Variance is measured in square units while the SD is in same units as the observed data.

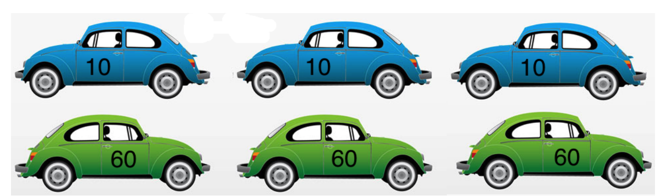
Diversity vs Variability

Which group of cars has a more diverse set of colors?

Diversity vs Variability (cont.)

Which group of cars has a more variable mileage?





Diversity vs Variability (cont.)

Group 1:

$$\bar{x} = (10 + 20 + 30 + 40 + 50 + 60)/6 = 35$$
$$s^2 = \frac{1}{6-1}\left((10-35)^2 + (20-35)^2 + \ldots + (60-35)^2\right) = 350$$

Group 2:

$$\bar{x} = (10 + 10 + 10 + 60 + 60 + 60)/6 = 35$$
$$s^2 = \frac{1}{6-1}\left((10-35)^2 + (10-35)^2 + \ldots + (60-35)^2\right) = 750$$

Median, Quartiles, and IQR

- The *median* is the value that splits the data in half when ordered in ascending order, i.e. 50^{th} *percentile*.

$$0, 1, 2, 3, 4$$

If there are an even number of observations, then the median is the average of the two values in the middle.

$$0, 1, \underline{2, 3}, 4, 5 \rightarrow \frac{2+3}{2} = 2.5$$

- The 25^{th} percentile is called the first quartile, *Q1*.
- The 75^{th} percentile is called the third quartile, *Q3*.
- The range spaned by the middle 50% of the is the *interquartile range*, or the *IQR*.

Robust statistics

The median and IQR are examples of what are known as robust statistics - because they are less affected by skewness and outliers than statistics like mean and SD.

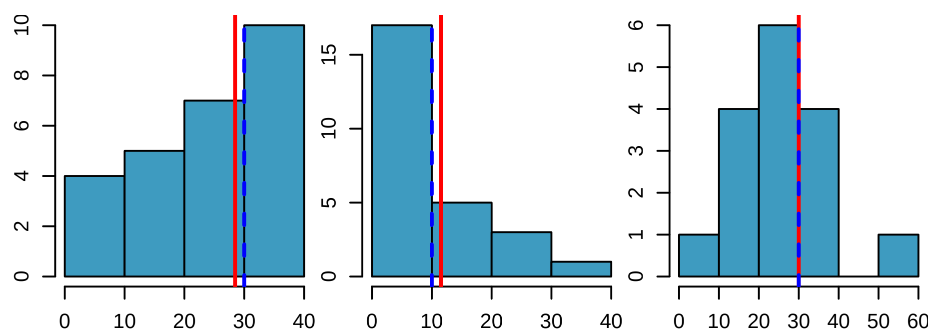
As such:

- for skewed distributions it is more appropriate to use median and IQR to describe the center and spread
- for symmetric distributions it is more appropriate to use the mean and SD to describe the center and spread

If you were searching for a house and are price conscious, should you be more interested in the mean or median house price when considering a particular neighborhood?

Mean vs. median

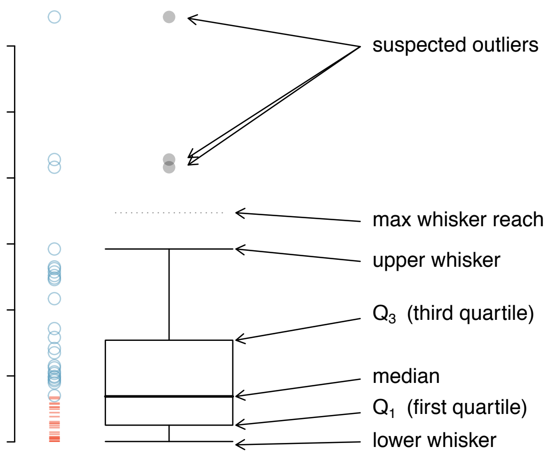
- If the distribution is symmetric, center is the mean
  - Symmetric: mean = median
- If the distribution is skewed or has outliers center is the median
  - Right-skewed: mean > median
  - Left-skewed: mean < median



red solid - mean, black dashed - median

Box plot

A *box plot* visualizes the median, the quartiles, and suspected outliers.



Box plot - Example

Resting Pulses:

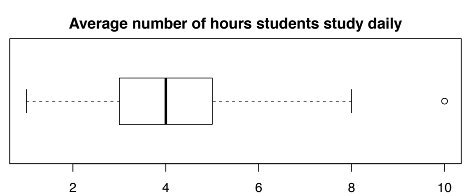
(62, 64, 68, 70, 70, 74, 74, 76, 76, 78, 78, 80)

Steps:

1. Calculate median, Q1, Q3, IQR, min, and max
2. Calculate upper and lower fences (Q1 - 1.5 IQR, Q3 + 1.5 IQR)
3. Find the location of the upper and lower wiskers
4. Consider data points outside whiskers as potential outliers

Reading a boxplot

Which of the following are false about the distribution of average number of hours students study daily.




```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.000   3.000   4.000   3.821   5.000  10.000
```

a) There are no students who don't study at all.
b) The IQR is 2 hours.
c) 75% of the students study $\geq$ 5 hours daily, on average.
d) 25% of the students study $\leq$ 3 hours, on average.

Tables and Contingency tables

For a single categorical variable we can always summarize it by showing the # of counts for each category. If we are interested in looking at a relationship between two categorical variables we need to construct a contigency table (cross tabulation).

Belief in a higher power:

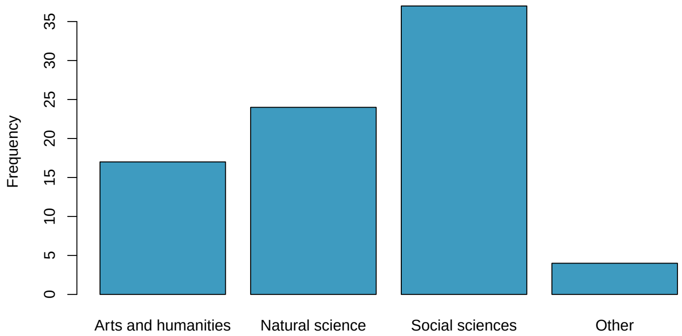
No	Somewhat	Yes
22	23	36

Gender:

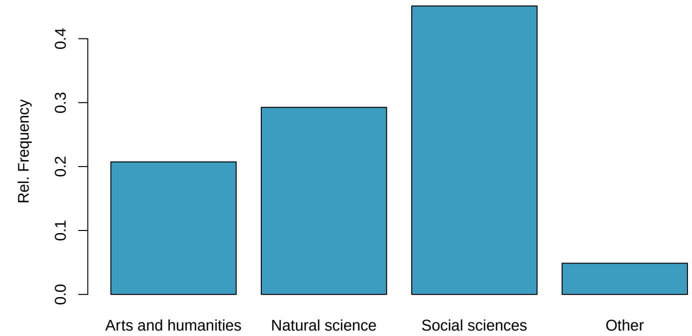
Female	Male
57	25

	Female	Male
No	14	8
Somewhat	16	7
Yes	26	10

Barplots - *Absolute (Counts)* vs Relative (Proportions)



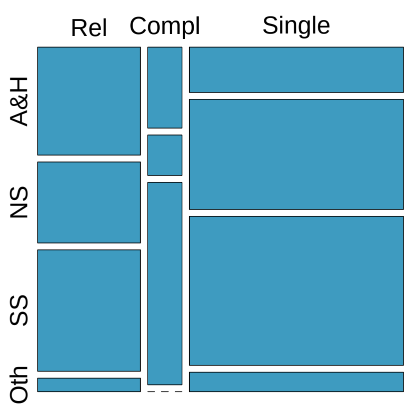
Barplots - Absolute (Counts) vs *Relative (Proportions)*



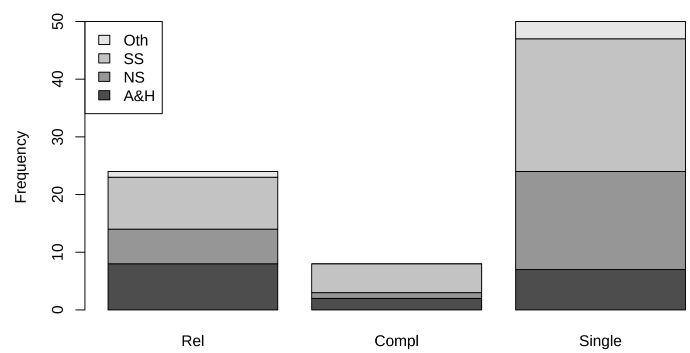
Mosaic plots

Is there a relationship between major and relationship status?

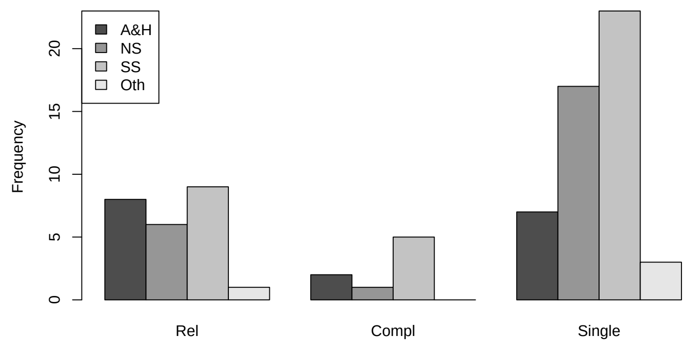
	Rel	Compl	Single
A&H	8	2	7
NS	6	1	17
SS	9	5	23
Oth	1	0	3



Bivariate Barplots - *Stacked* vs Juxtaposed

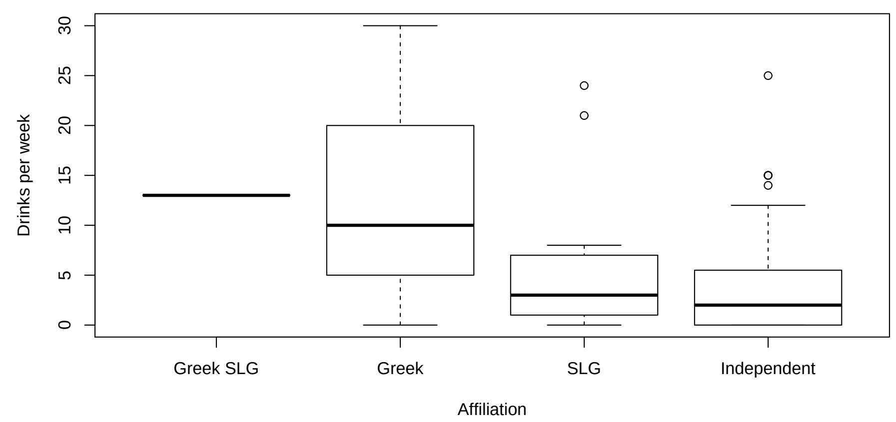


Bivariate Barplots - Stacked vs *Juxtaposed*



Side-by-side box plot

How does number of drinks consumed per week vary by affiliation?



Visualization Summary

- Single numeric - dot plot, box plot, histogram
- Single categorical - bar plot (or a table)
- Two numeric - scatter plot
- Two categorical - mosaic plot, stacked or side-by-side bar plot
- Numeric and categorical - side-by-side box plot

Tufte's Principles:

1. Above all else show data.
2. Maximize the data-ink ratio.
3. Erase non-data-ink.
4. Erase redundant data-ink.
5. Revise and edit