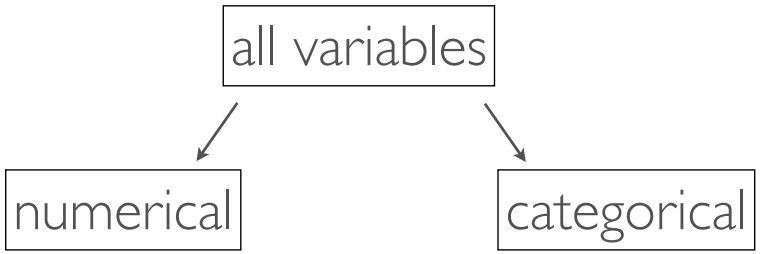
Data and Data Summaries

Sta 111

Colin Rundel

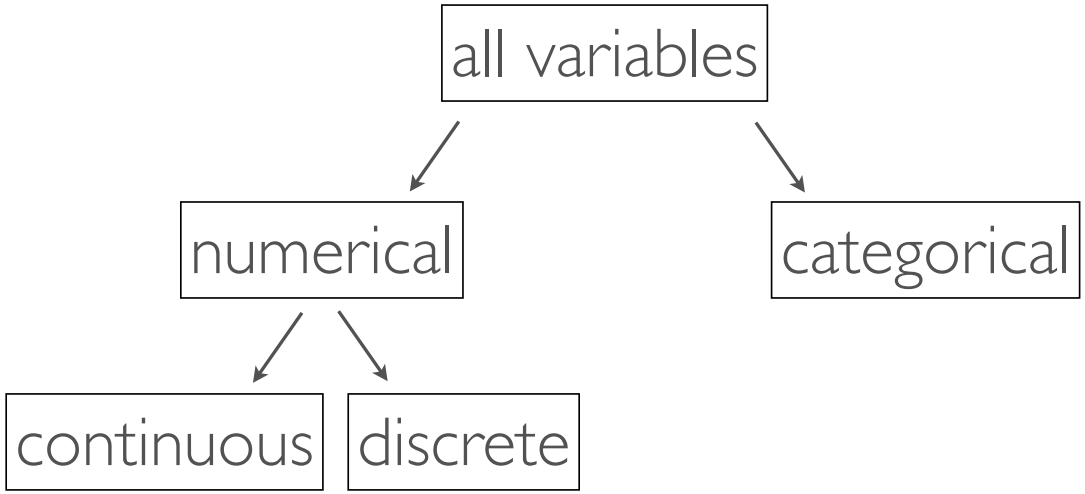
May 20, 2014

Data



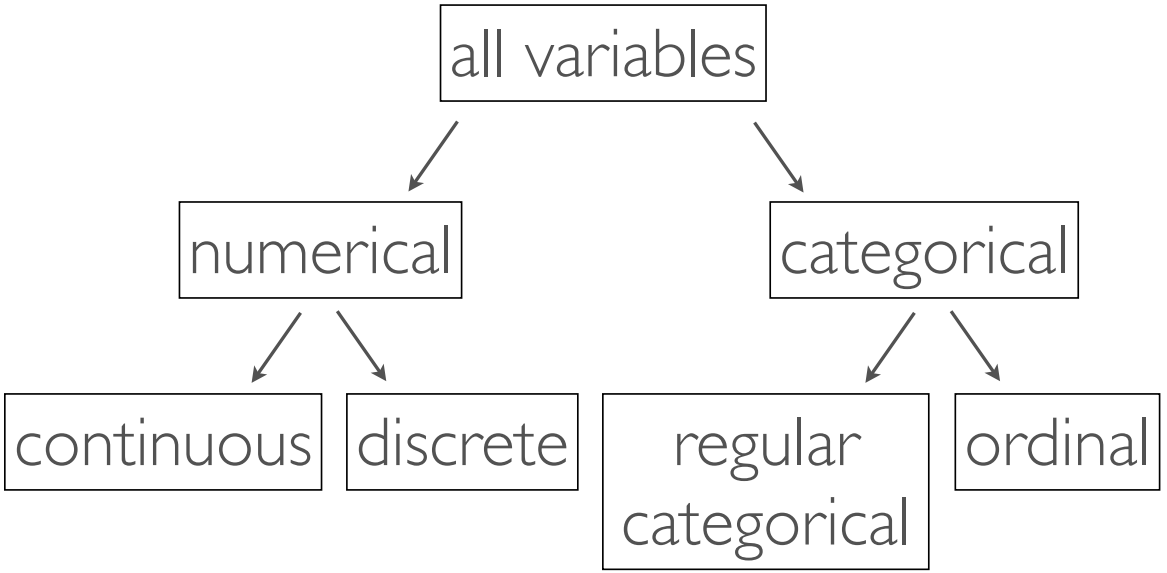
- *Numerical (quantitative)* - takes on a numerical values
  - Ask yourself - is it sensible to add, subtract, or calculate an average of these values?
- *Categorical (qualitative)* - takes on one of a set of distinct categories
  - Ask yourself - are there only certain values (or categories) possible?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

Numerical Data



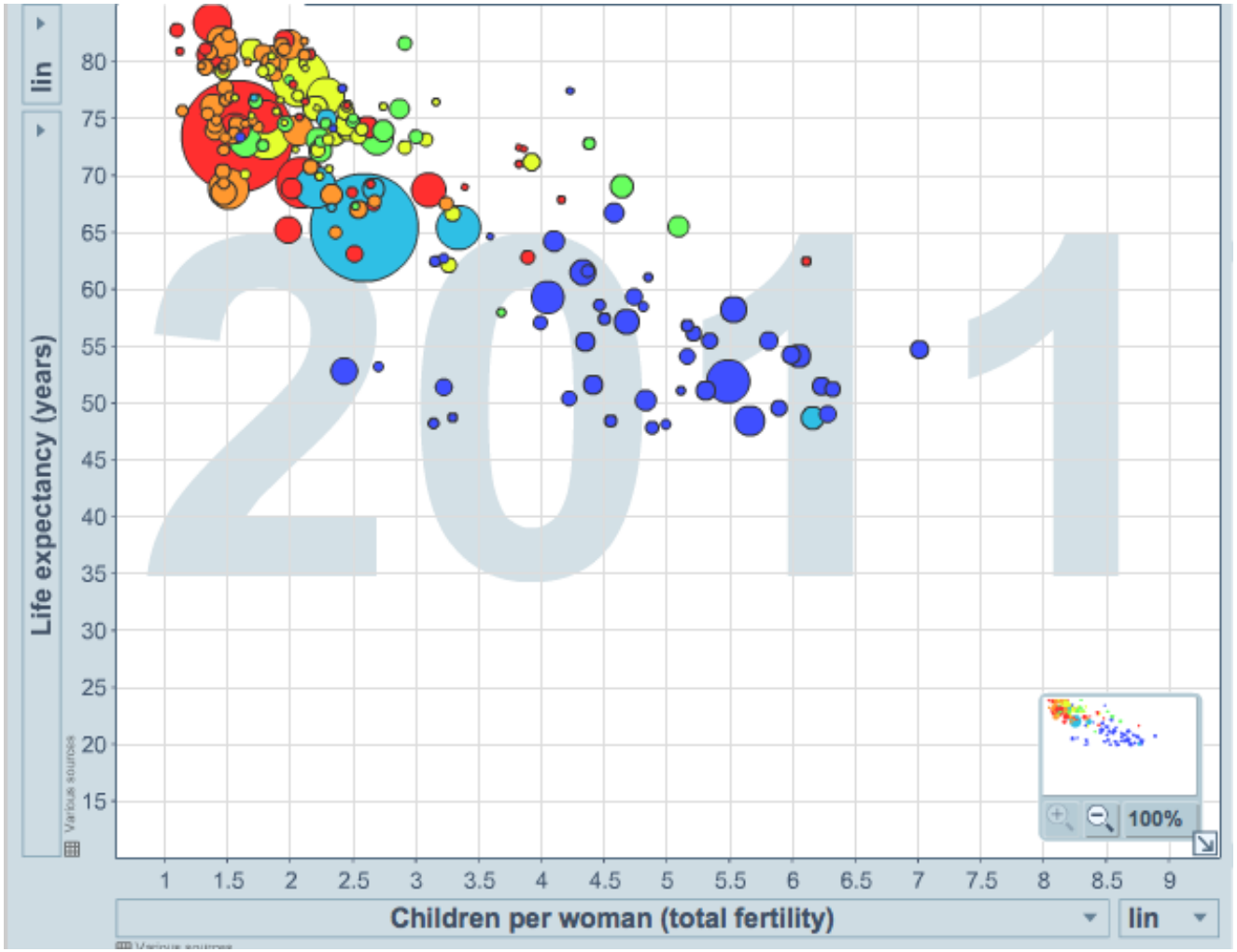
- *Continuous* - data that is measured, any numerical (decimal) value
- *Discrete* - data that is counted, only whole non-negative numbers

Categorical Data



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

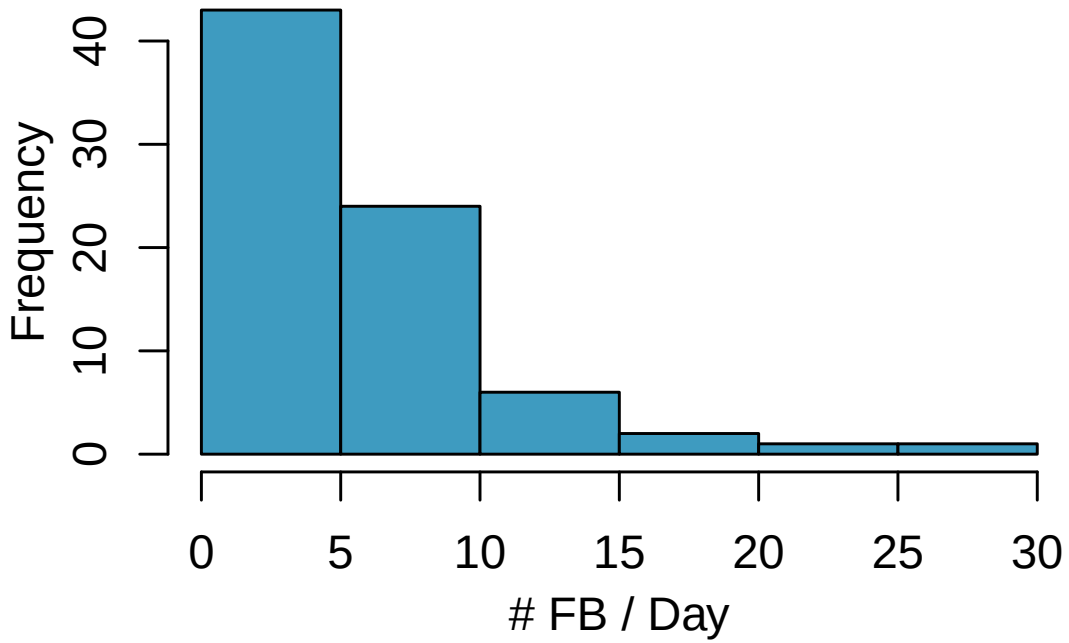
Scatterplots



http://www.gapminder.org/world

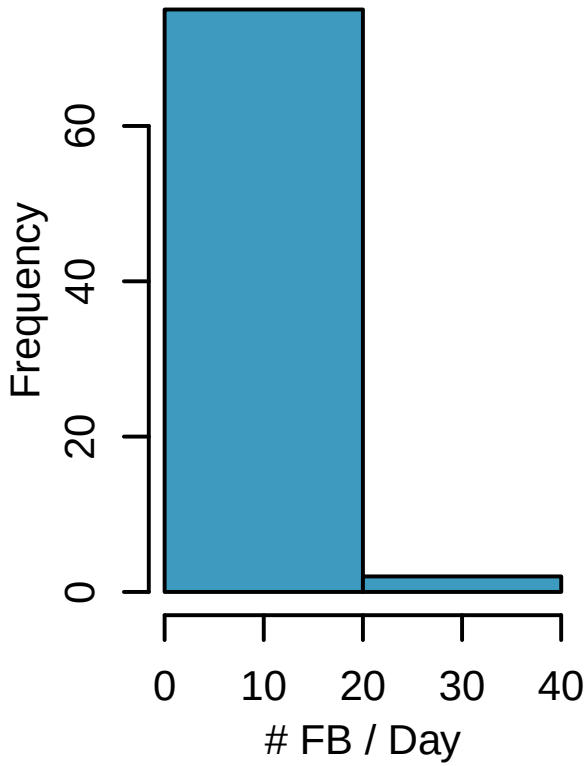
Histograms

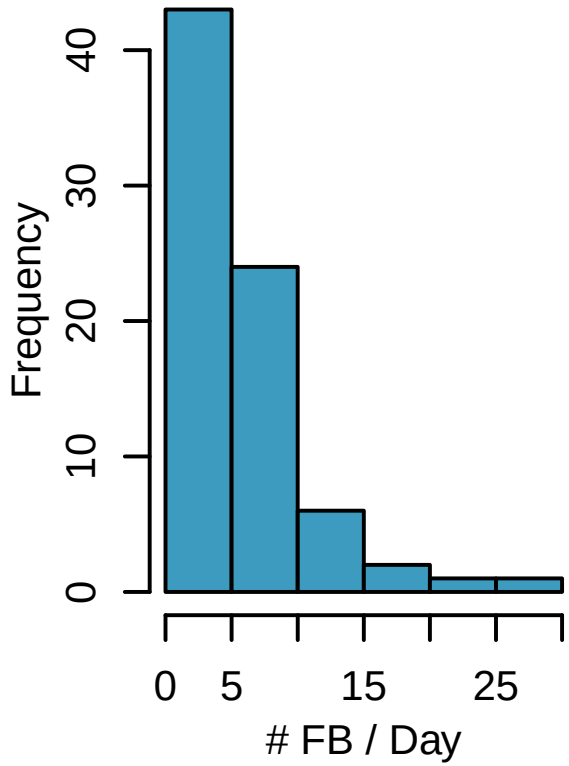
- Histograms provide a view of the data's *density* / *shape*, higher bars represent where the data are more common.

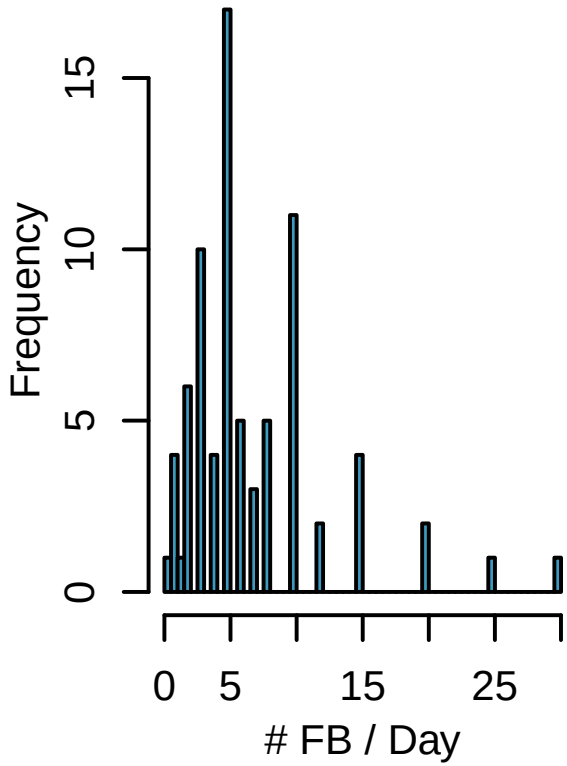


Histogram - Bin width

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

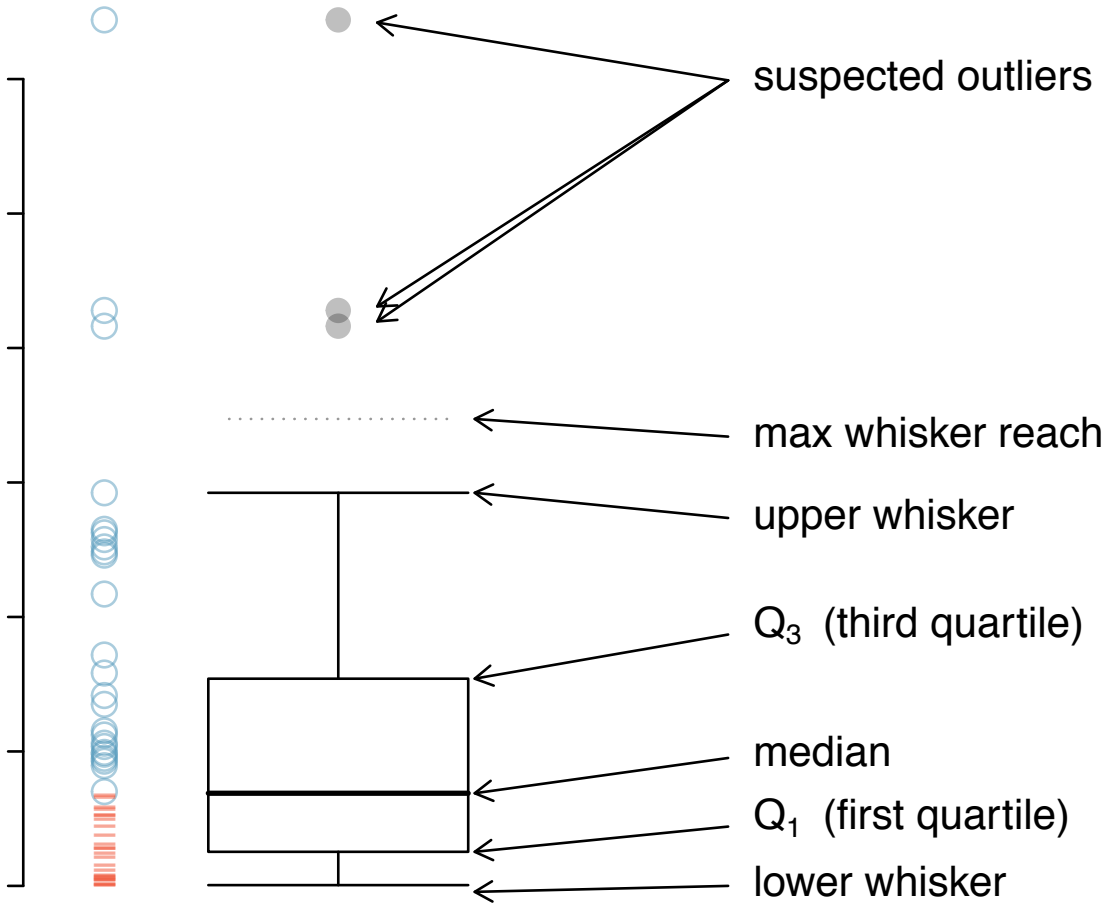






Box plot

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



Box plot - Example

Resting Pulse

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

Box plot - Example

Resting Pulse

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

Steps:

1. Calculate median, Q1, Q3, IQR, min, and max
   - $med = 74$
   - $Q1 = 69$
   - $Q2 = 77$
   - $IQR = 77 - 69 = 8$
   - $min = 62$
   - $max = 80$
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

Box plot - Example

Resting Pulse

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

Steps:

1. Calculate median, Q1, Q3, IQR, min, and max
   - $med = 74$
   - $Q1 = 69$
   - $Q2 = 77$
   - $IQR = 77 - 69 = 8$
   - $min = 62$
   - $max = 80$
2. Calculate upper and lower fences (Q1 - 1.5 IQR, Q3 + 1.5 IQR)
   - $F_L = Q1 - 1.5\ IQR = 69\text{-}12 = 57$
   - $F_U = Q3 + 1.5\ IQR = 77{+}12 = 89$
3. Find the location of the upper and lower wiskers
   - $W_L = 62$
   - $W_U = 80$
4. Consider data points outside whiskers as potential outliers

Tables and Contingency tables

We might be interested in looking at if there is a relationship between religion belief in God and gender, in which case we need to summarize both variables:

No	Somewhat	Yes
22	23	36

Female	Male
57	25

but this is not enough alone.

Tables and Contingency tables

We might be interested in looking at if there is a relationship between religion belief in God and gender, in which case we need to summarize both variables:

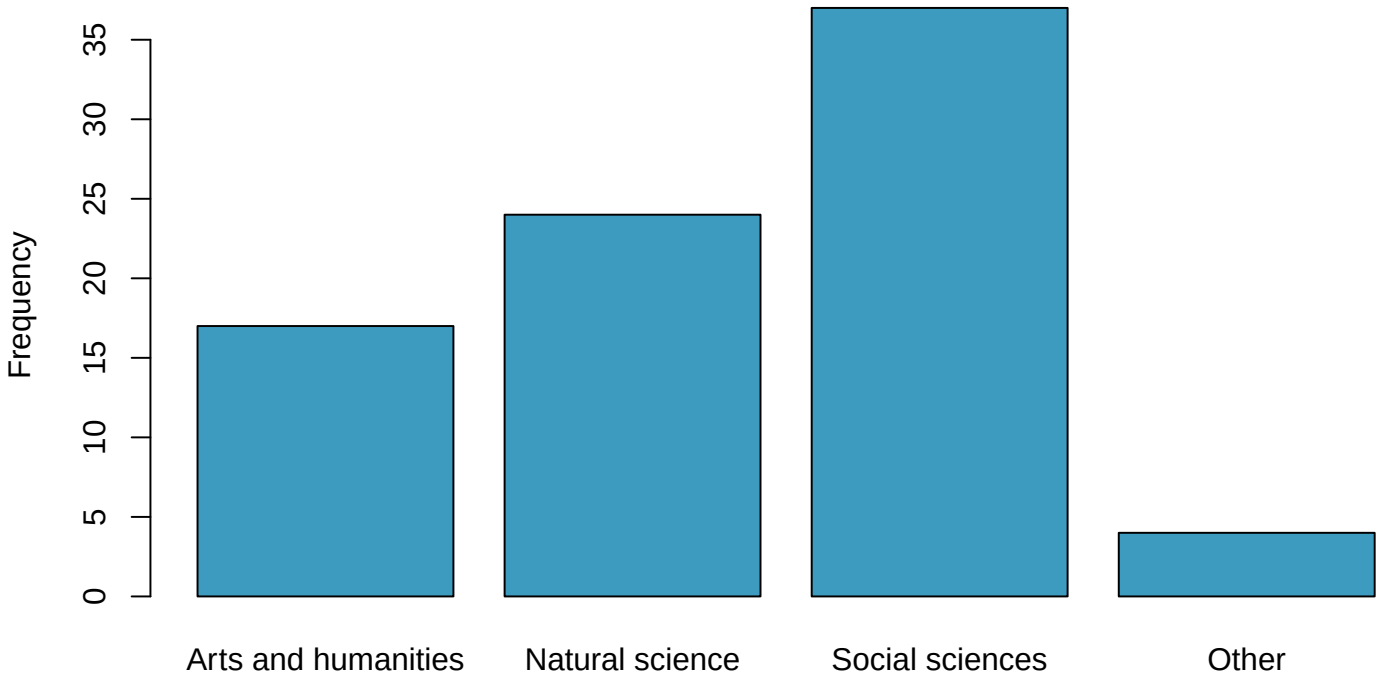
No	Somewhat	Yes
22	23	36

Female	Male
57	25

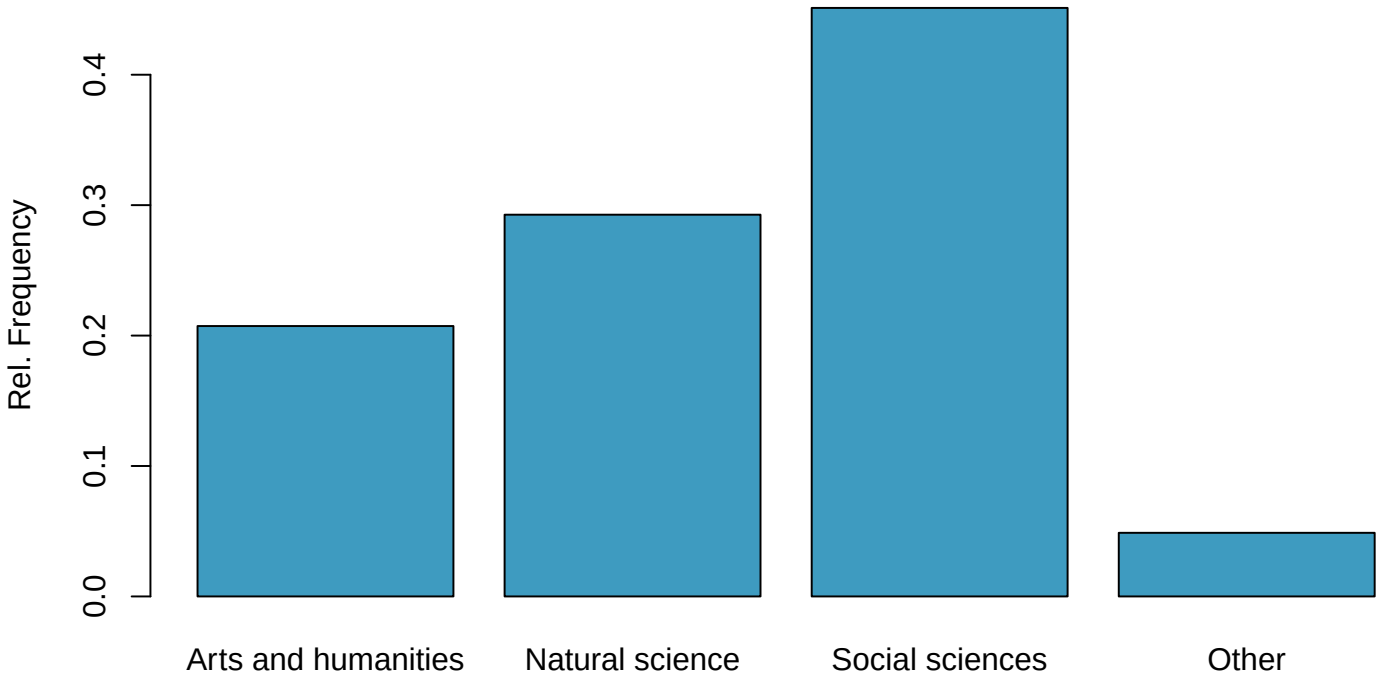
but this is not enough alone.

Female	Male
14	8
16	7
26	10

Barplots - Absolute vs Relative

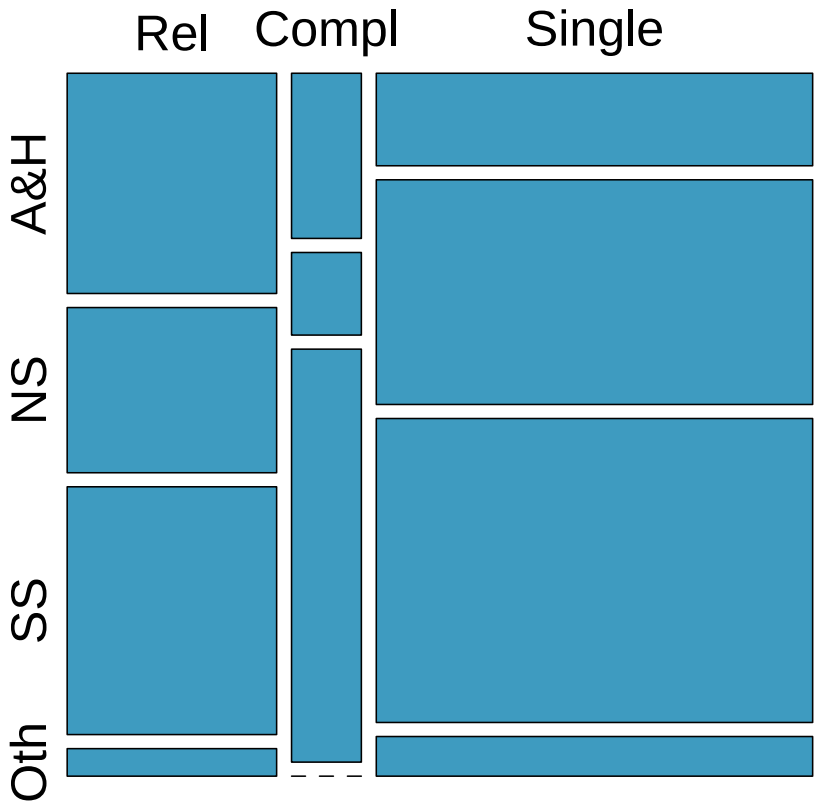


Barplots - Absolute vs Relative

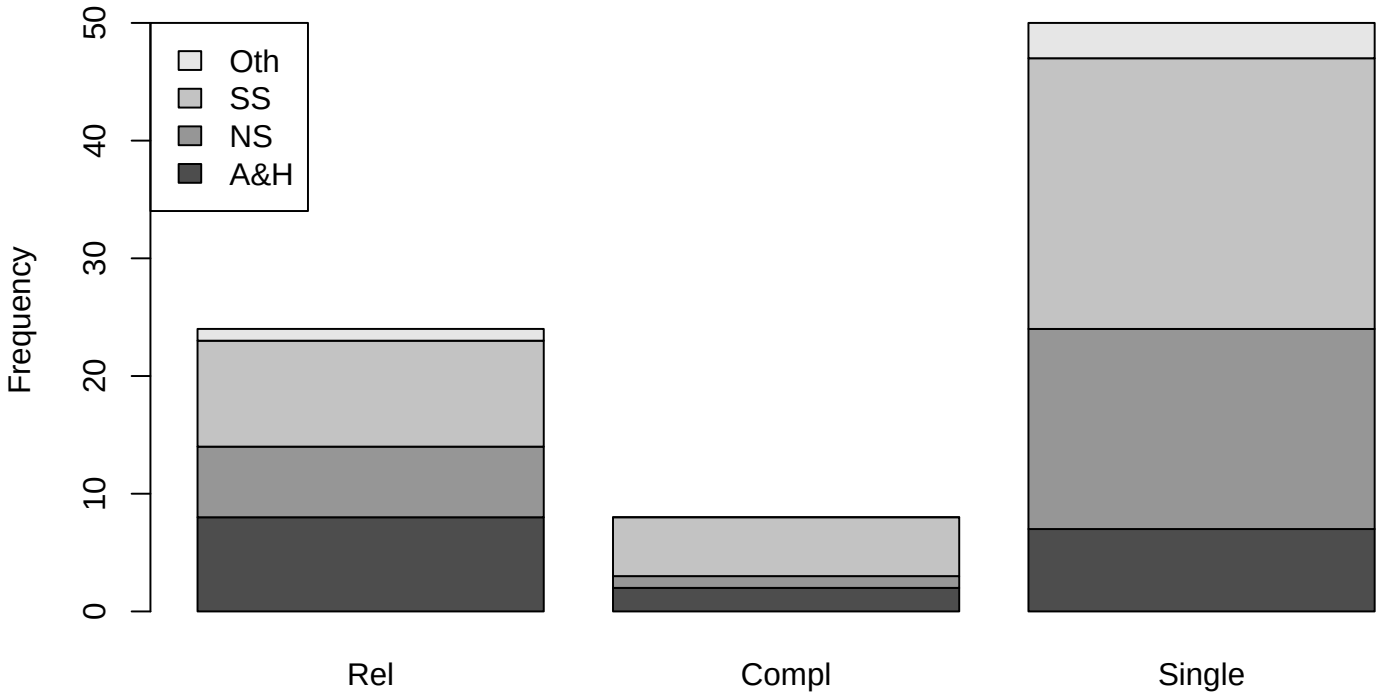


Mosaic plots

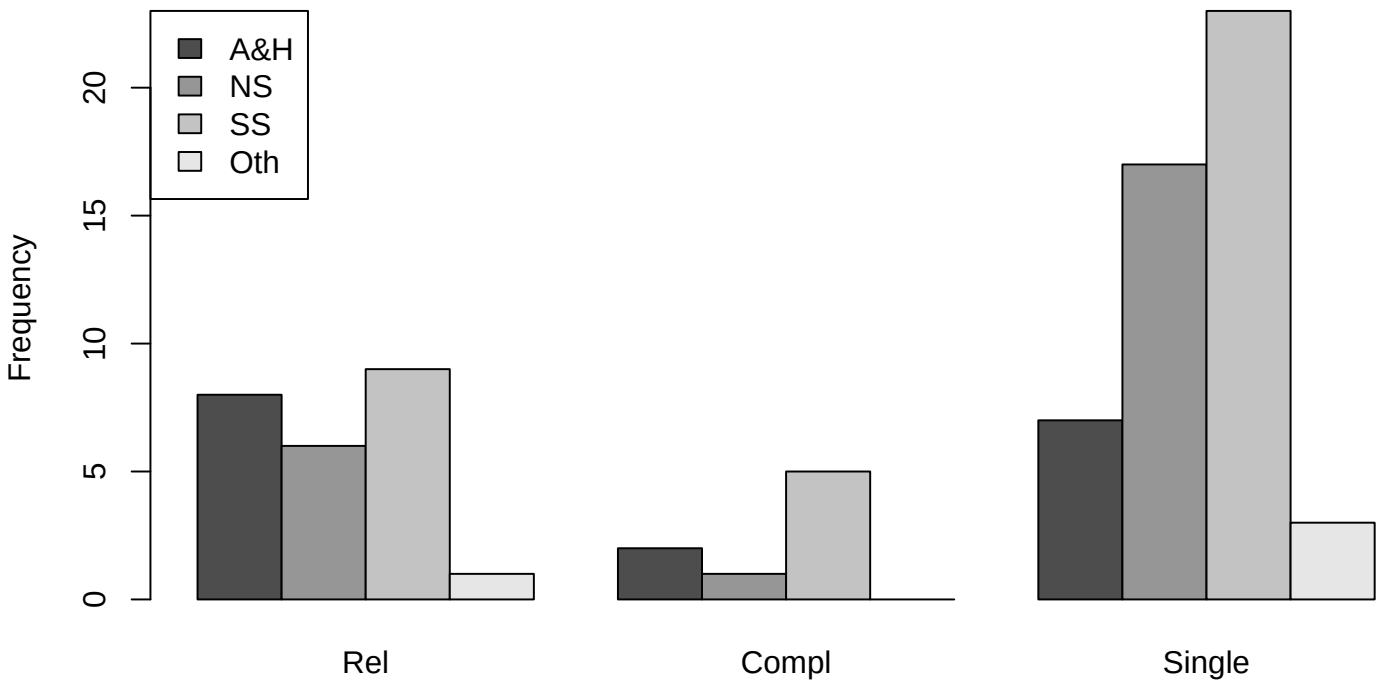
Rel	Compl	Single
8	2	7
6	1	17
9	5	23
1	0	3



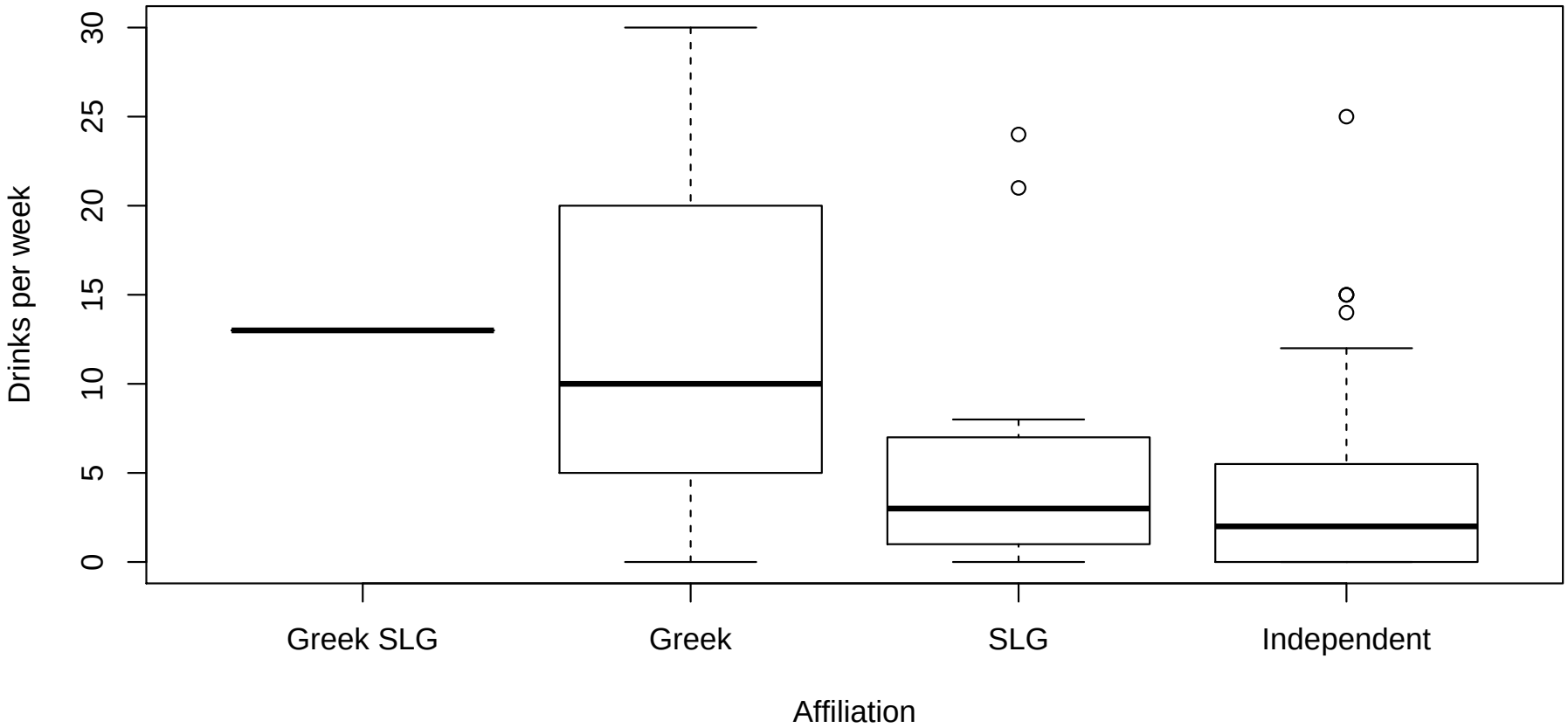
Bivariate Barplots - Stacked vs Juxtaposed



Bivariate Barplots - Stacked vs Juxtaposed



Side-by-side box plot



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