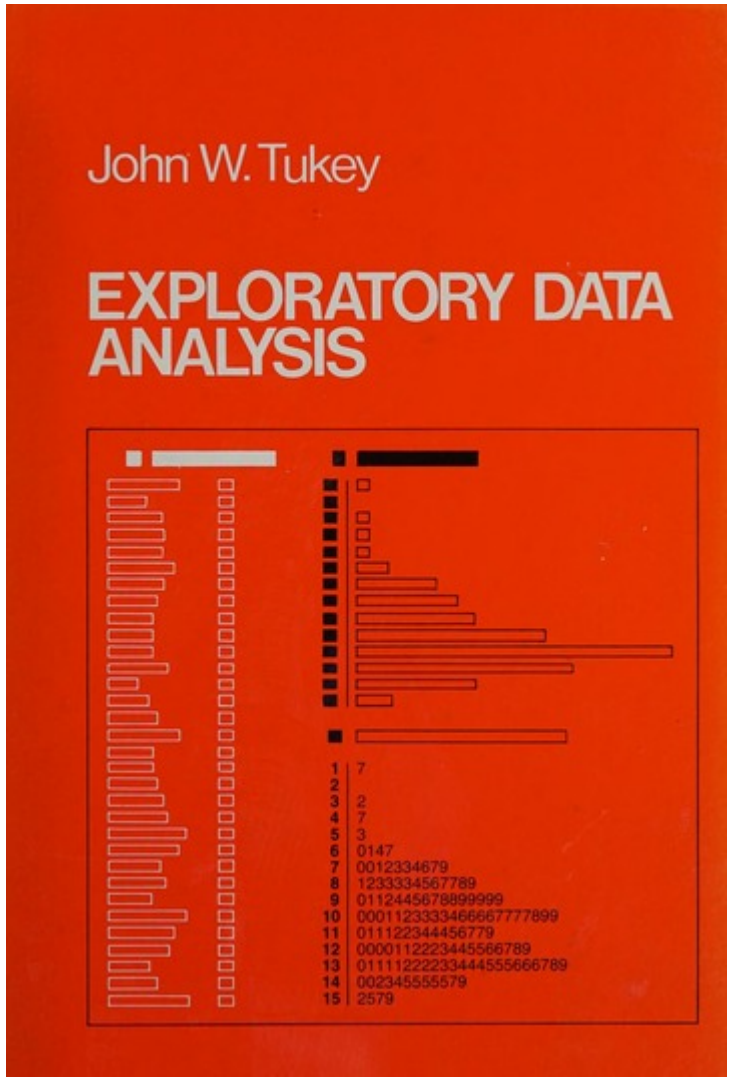
Exploratory Data Analysis

Julian Faraway



- New computer workstations enabled interactive use
- Explore data and make conclusions without formal modelling
- Precursor to data science
- Led to the S statistical programming language which led to R

Published 1977

Purpose of EDA

- Familiarisation with data suggesting suitable models
- Sometimes EDA is enough - formal modelling not necessary
- Avoidance of error - don’t make stupid mistakes

Where do models come from?

Theory

Scientific and/or engineering knowledge leads to development of a proposed model.

- No data or data assimilation (Applied Maths)
- Model fitting using data (Statistics)
- Who needs a model! (ML/AI)

Data

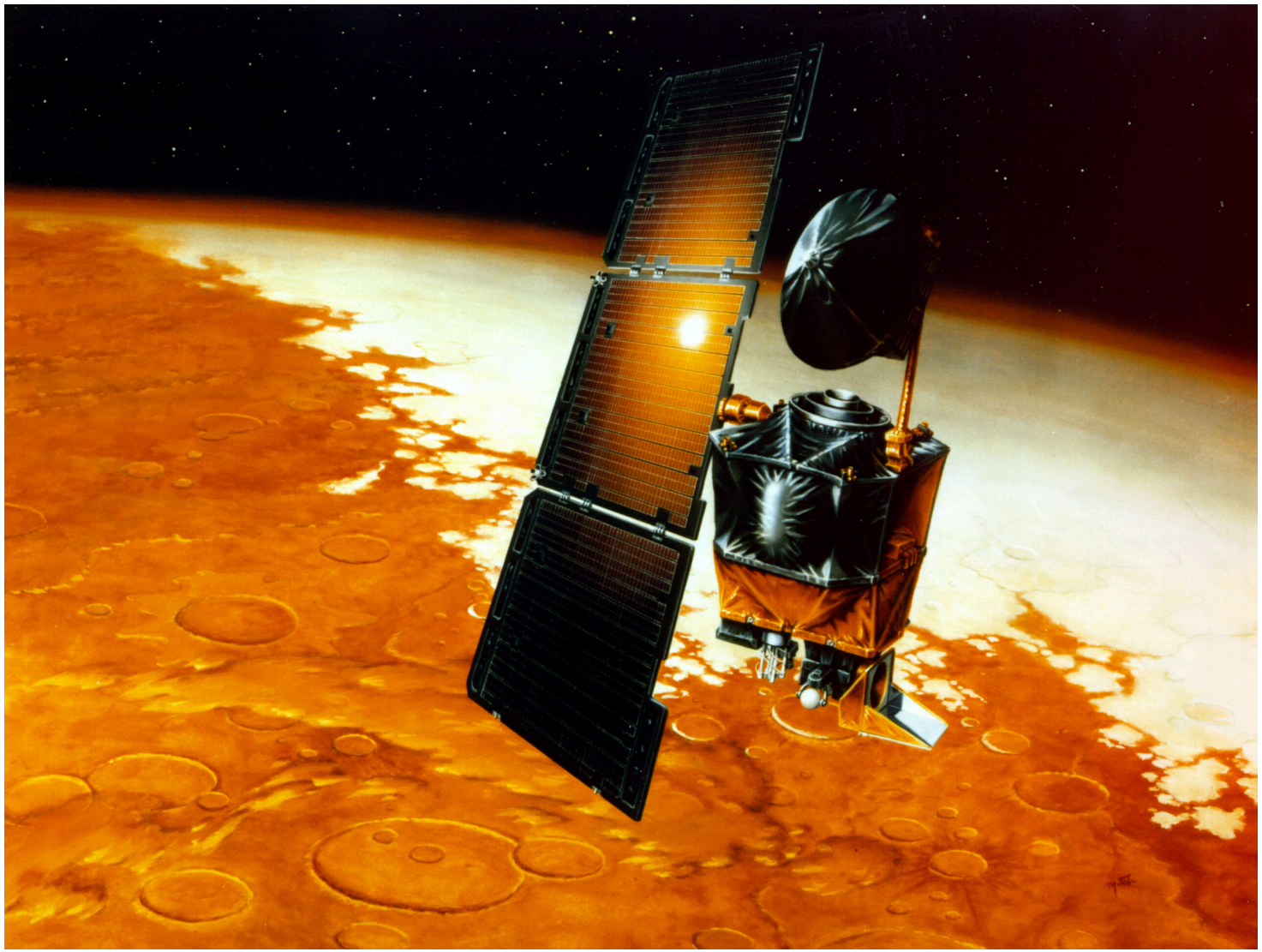
Empirical development of model form based on examination of the data.

EDA techniques help suggest suitable form for the model.

Look before you leap

Don’t fit a model without looking at the data

What could possibly go wrong?



Mars Orbiter Crash in 1999 due to unit mismatch between metric and imperial

Avoidance of mistakes

- Projects involve a large number of decisions, assumptions, properties etc. that need to be checked
- Individual checks are typically straightforward but a single error can ruin everything.

> I got it mostly right and my error was trivial - I should get $(100 - \epsilon)\%$

> Only stupid, lazy people make mistakes. Just pay proper attention and you won't get it wrong.

> I'm a big-brained mathematician fitting incredibly sophisticated models. If something goes wrong, it's all the fault of the people who collected or supplied the data.

SURGICAL SAFETY CHECKLIST

Patient Name:

Procedure:

Date:

Notes:

Before induction of anesthesia

SIGN IN

- ❑ **Patient has confirmed:**
  - Identity
  - Site
  - Procedure
  - Consent
- ❑ **Site marked** ❑ **Not applicable**
- ❑ **Anesthesia safety check completed**
- ❑ **Pulse Oximeter on patient and functioning**

Does patient have a Known allergy?

❑ NO ❑ YES

Difficult airway/aspiration risk?

❑ NO ❑ YES, and equipment/assistance available

Risk of >500ml bood loss (7ml/kg in children)?

❑ NO ❑ YES, and adequate intravenous access and fluids planned

Before skin incision

TIME OUT

- ❑ **Confirm all team members have introduced themselves by name and role**
- ❑ **Surgeon, Anesthesia Professional and Nurse verbally confirm:**
  - Patient
  - Site
  - Procedure

Anticipated critical events:

- ❑ **Surgeon reviews:** What are the critical or unexpected steps, operative duration, anticipated blood loss?
- ❑ **Anesthesia team reviews:** Are there any patient-specific concerns?
- ❑ **Nursing team reviews:** Has sterility (including indicator results) been confirmed? Are there equipment issues or any concerns?

Has antibiotic Prophylaxis been given within the last 60 minutes?

❑ YES ❑ Not applicable

Is essential imaging displayed?

❑ YES ❑ Not applicable

Before patient leaves operating room

SIGN OUT

Nurse verbally confirms with the team:

- ❑ **The name of the procedure recorded**
- ❑ **That instrument, sponge, and needle counts are correct** (or not applicable)
- ❑ **How the specimen is labelled** (including patient name)
- ❑ **Whether there are any equipment problems to be addressed**
- ❑ **Surgeon, Anesthesia Professional and Nurse review the key concerns for recovery and management of this patient**

Healthcare Inspirations

Introducing the Data

- Format, CSV, Fixed format, binary, Excel XLSX, others
- Privacy and security concerns
- Very large datasets and databases
- Sharing with others, shared file space, github
- Ethical considerations
- Metadata, codebook
- Archive original

Date Data

- North America has π day as 3/14
- Europe calls it 14/3, or 14th March or ….
- ISO date format is 2024-03-14 (promote and use this!)
- Use custom software for dates/times (never roll your own)

Excel for data analysis

Better not.

The London Whale

A $6.2billion dollar loss at JP Morgan 2012

> Following that decision, further errors were discovered in the Basel II.5 model, including, most significantly, an operational error in the calculation of the relative changes in hazard rates and correlation estimates. Specifically, after subtracting the old rate from the new rate, the spreadsheet divided by their sum instead of their average, as the modeler had intended.

Coding in Excel is opaque and difficult to check.

One of many such stories

Excel for Data Management

Excel: Why using Microsoft's tool caused Covid-19 results to be lost

5 October 2020

By Leo Kelion
Technology desk editor

The badly thought-out use of Microsoft's Excel software was the reason nearly 16,000 coronavirus cases went unreported in England.

SCIENCE / TECH / MICROSOFT

Scientists rename human genes to stop Microsoft Excel from misreading them as dates

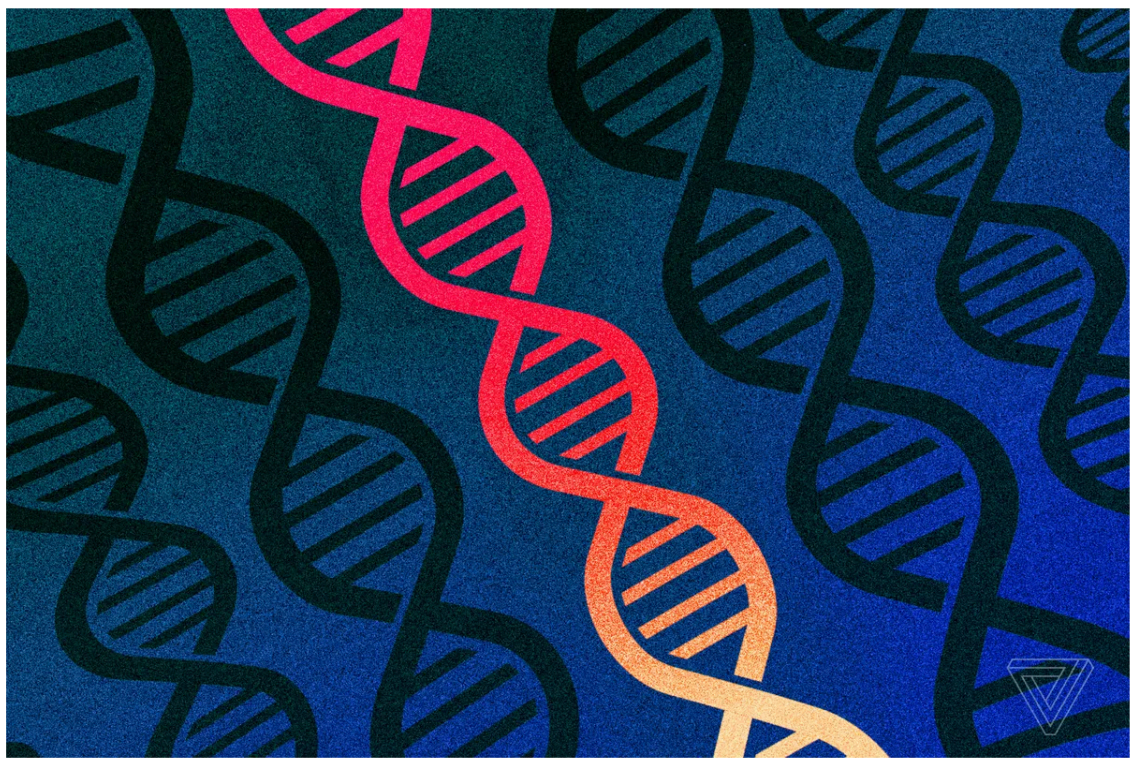


Illustration by Alex Castro / The Verge

/ Sometimes it's easier to rewrite genetics than update Excel

By James Vincent, a senior reporter who has covered AI, robotics, and more for eight years at The Verge.

Aug 6, 2020, 1:44 PM GMT+1

0 Comments (0 New)

Accept the reality of Excel

- People will keep using Excel whether we like it or not.
- Collaborators are not all experts in data analysis - Excel is a rational choice.
- Excel is widely available while better tools are not.
- Understand enough Excel to reduce the chance of errors.
- Google sheets or other alternatives are not better.
- *Data Organization in Spreadsheets* by Karl Broman & Kara Woo (2017)

Editing data

In order of preference:

1. Read the data into R, Python or other software first. There is a wide range of add-on packages for different data formats. Now edit the data within the software. Easier to keep a record of changes. Easier if the original data is updated.
2. Data won't read in. Use other tools such as `awk`, `perl`, `python` or unix command line tools to programmatically edit the data. Easier to keep a record, easier to deal with updates, easier if there's a lot of data.
3. Manual editing is the last choice. Never use Word. Keep records of what you changed.

- Feedback to data owners may be appreciated.
- Don't refer to *cleaning the data* as this may be insulting.

Reproducibility

- *Reproducibility crisis in Science*
- Repeating the experiment is good but not our job
- Reproducible calculation is our responsibility.
  - Be nice to future you
  - Be ready to defend your conclusions
  - Work well with others

Reproducibility tools

- R Markdown
- Quarto (R and Python)
- Jupyter Notebook (Python and R)
- Github

Numerical Summaries

What to look for:

- Scale of variables. Units as expected?
- Minimums and Maximums. Reasonable values?
- Mean/Median. About what you expected?
- Missing values may be apparent.
- *Table 1* in medical statistics papers is usually a nice version of this information. May be more generally appreciated.

```r
data(pima, package="faraway")
summary(pima)
```

```
    pregnant         glucose       diastolic        triceps          insulin     
 Min.   : 0.00   Min.   :  0   Min.   :  0.0   Min.   : 0.0   Min.   :  0.0  
 1st Qu.: 1.00   1st Qu.: 99   1st Qu.: 62.0   1st Qu.: 0.0   1st Qu.:  0.0  
 Median : 3.00   Median :117   Median : 72.0   Median :23.0   Median : 30.5  
 Mean   : 3.85   Mean   :121   Mean   : 69.1   Mean   :20.5   Mean   : 79.8  
 3rd Qu.: 6.00   3rd Qu.:140   3rd Qu.: 80.0   3rd Qu.:32.0   3rd Qu.:127.2  
 Max.   :17.00   Max.   :199   Max.   :122.0   Max.   :99.0   Max.   :846.0  
      bmi           diabetes          age             test       
 Min.   : 0.0   Min.   :0.078   Min.   :21.0   Min.   :0.000  
 1st Qu.:27.3   1st Qu.:0.244   1st Qu.:24.0   1st Qu.:0.000  
 Median :32.0   Median :0.372   Median :29.0   Median :0.000  
 Mean   :32.0   Mean   :0.472   Mean   :33.2   Mean   :0.349  
 3rd Qu.:36.6   3rd Qu.:0.626   3rd Qu.:41.0   3rd Qu.:1.000  
 Max.   :67.1   Max.   :2.420   Max.   :81.0   Max.   :1.000  
```

Discrete data

```
1  imdb = read.csv("data/movie_metadata.csv")
2  xtabs(~ country, imdb)
```

```
country
                    Afghanistan             Argentina
              5               1                     4
          Aruba         Australia               Bahamas
              1              55                     1
        Belgium          Brazil               Bulgaria
              4               8                     1
       Cambodia         Cameroon                Canada
              1               1                   126
          Chile           China              Colombia
              1              30                     1
Czech Republic         Denmark   Dominican Republic
              3              11                     1
          Egypt         Finland                France
              1               1                   154
```

Correlation

```r
data(nels88,package="faraway")
head(nels88)
```

```
     sex  race   ses paredu math
1 Female White -0.13     hs   48
2   Male White -0.39     hs   48
3   Male White -0.80     hs   53
4   Male White -0.72     hs   42
5 Female White -0.74     hs   43
6 Female White -0.58     hs   57
```

```
1 xtabs(~ race + paredu, nels88)
```

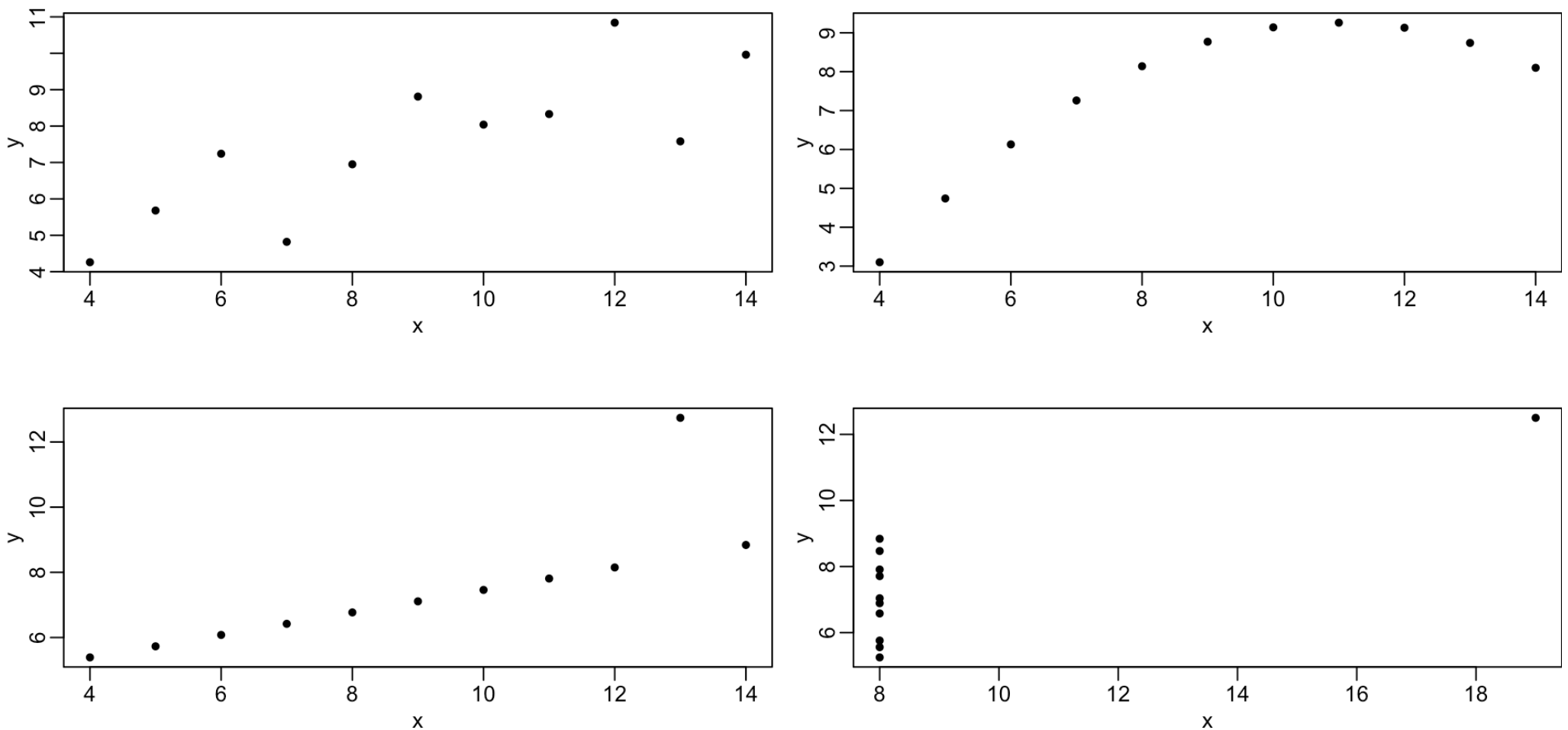
```
          paredu
race       ba college hs lesshs ma phd
  White    23      60 37     14 31  24
  Asian     2       2  0      1  2   1
  Black     2      13  9     15  1   0
  Hispanic  1       8  4      9  0   1
```

```
1 with(nels88, cor(ses,math))
```

```
[1] 0.62115
```

```r
par(mfrow=c(2,2))
for(i in 1:4) plot(anscombe[,i],anscombe[,i+4],xlab="x",ylab="y")
```

```r
par(mfrow=c(1,1))
```



Anscombe Quartet - all correlations are 0.6, all means and SDs the same

Graphical Summaries

Working plots

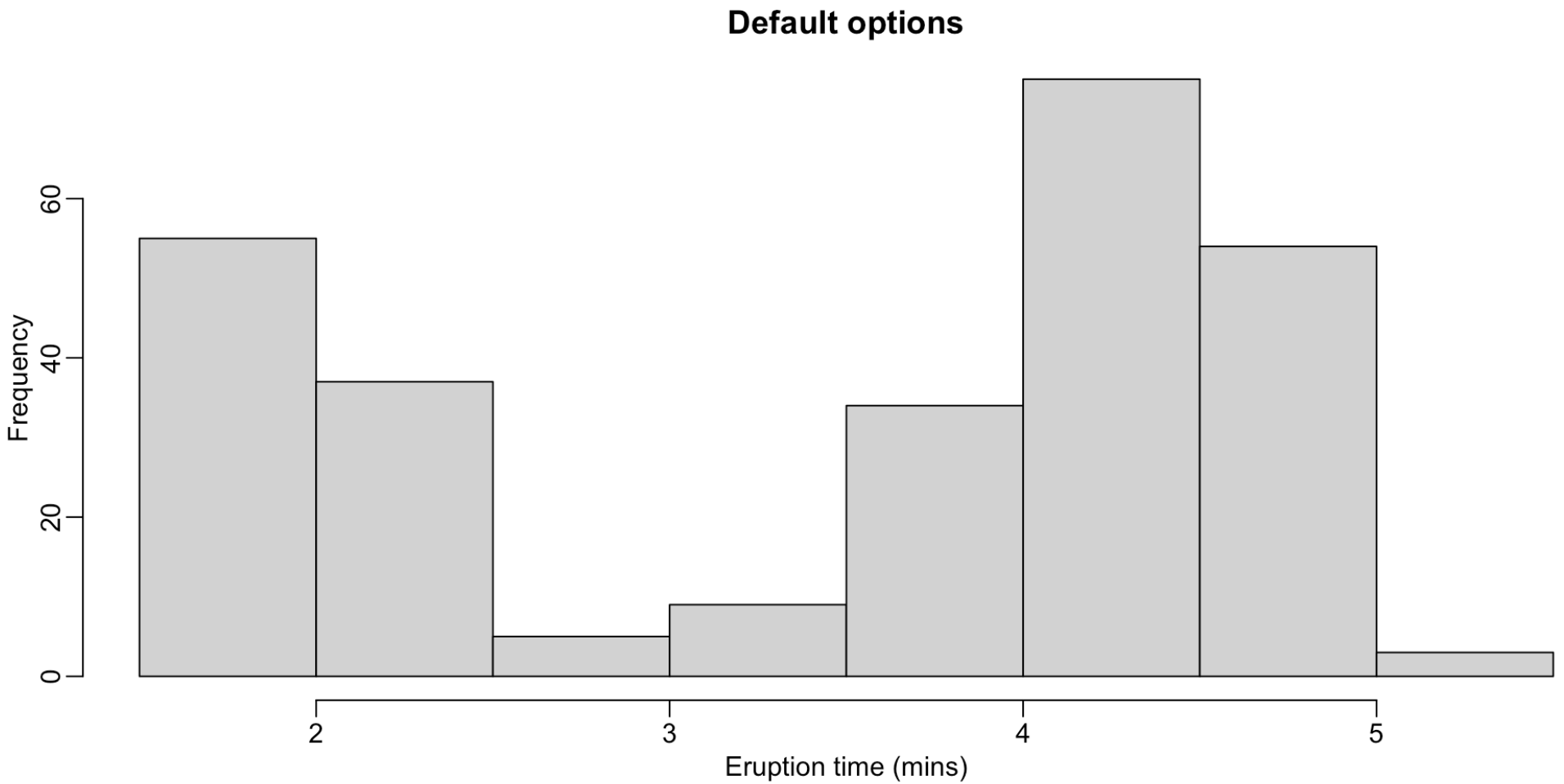
- Plots constructed for the immediate use of you or your collaborators
- Need to be quick and easy to construct
- Need enough polish for understanding but no more
- Might be large in number and temporary in value

Presentation plots

- Plots constructed for explaining your conclusions to others
- May take considerable effort to construct
- Highly valuable and important component of your output
- Only a few because you want the audience to focus

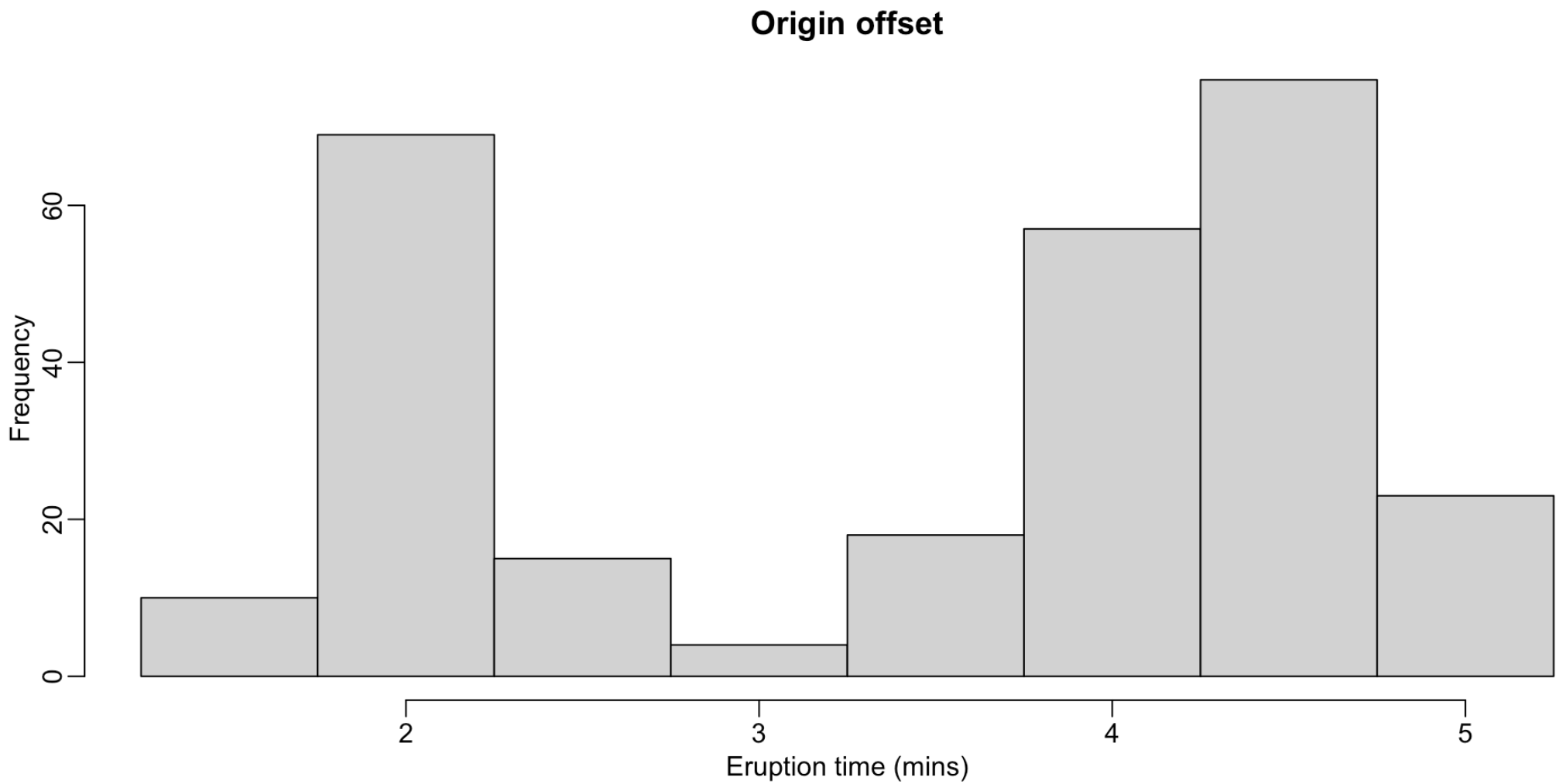
Single continuous variables

```
1 hist(faithful$eruptions,main="Default options",  xlab="Eruption time (mins)")
```



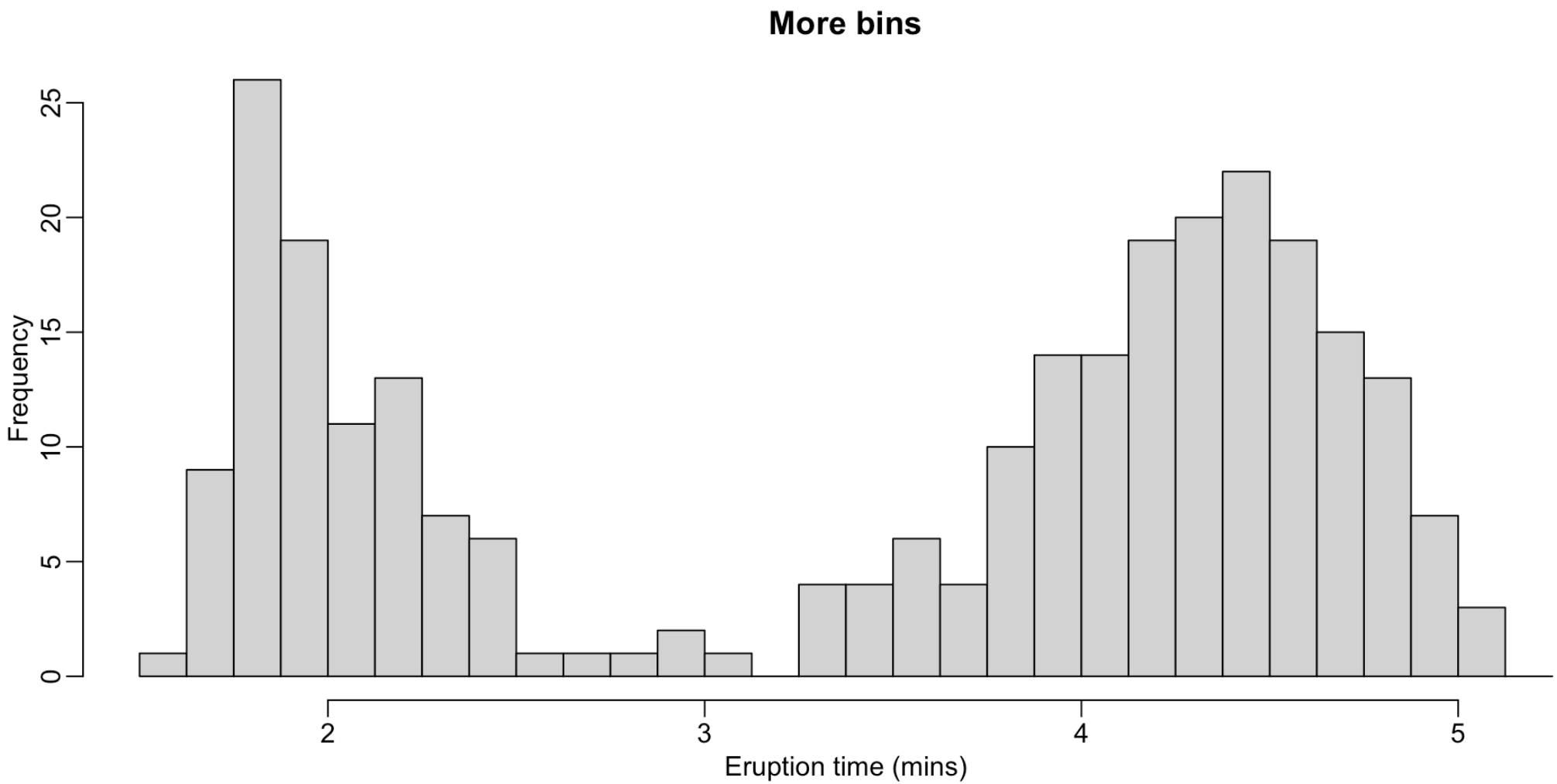
Histogram is the default but there is some choice

```
hist(faithful$eruptions,main="Origin offset",xlab="Eruption time (mins)",
     breaks=1:9/2+0.75)
```



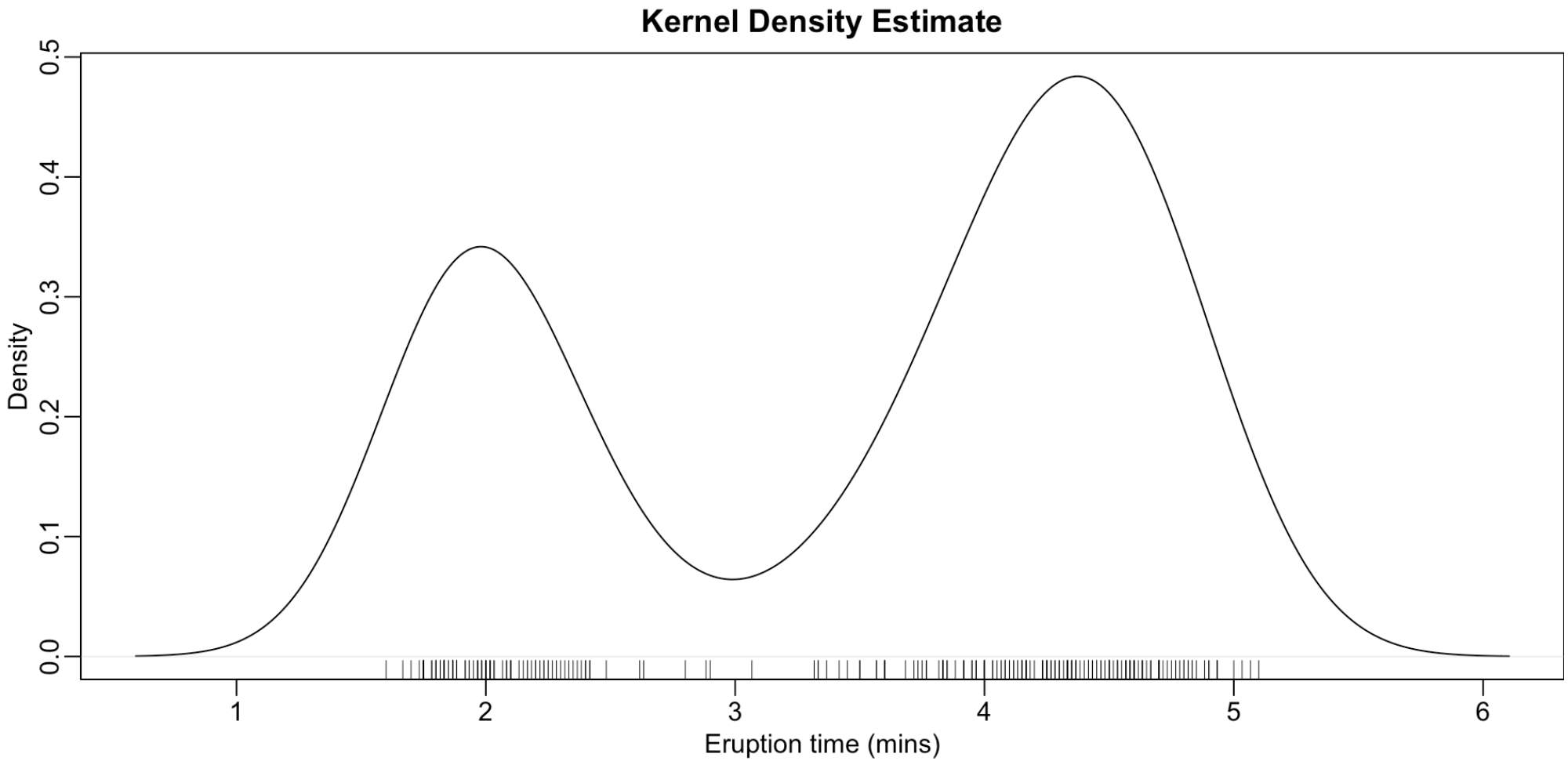
Shifted breakpoints

```
hist(faithful$eruptions,  main="More bins",  xlab="Eruption time (mins)",
      breaks=4:34/8+1)
```



More breakpoints

```r
plot(density(faithful$eruptions),
  type="l",xlab="Eruption time (mins)",
  main="Kernel Density Estimate")
rug(faithful$eruptions)
```



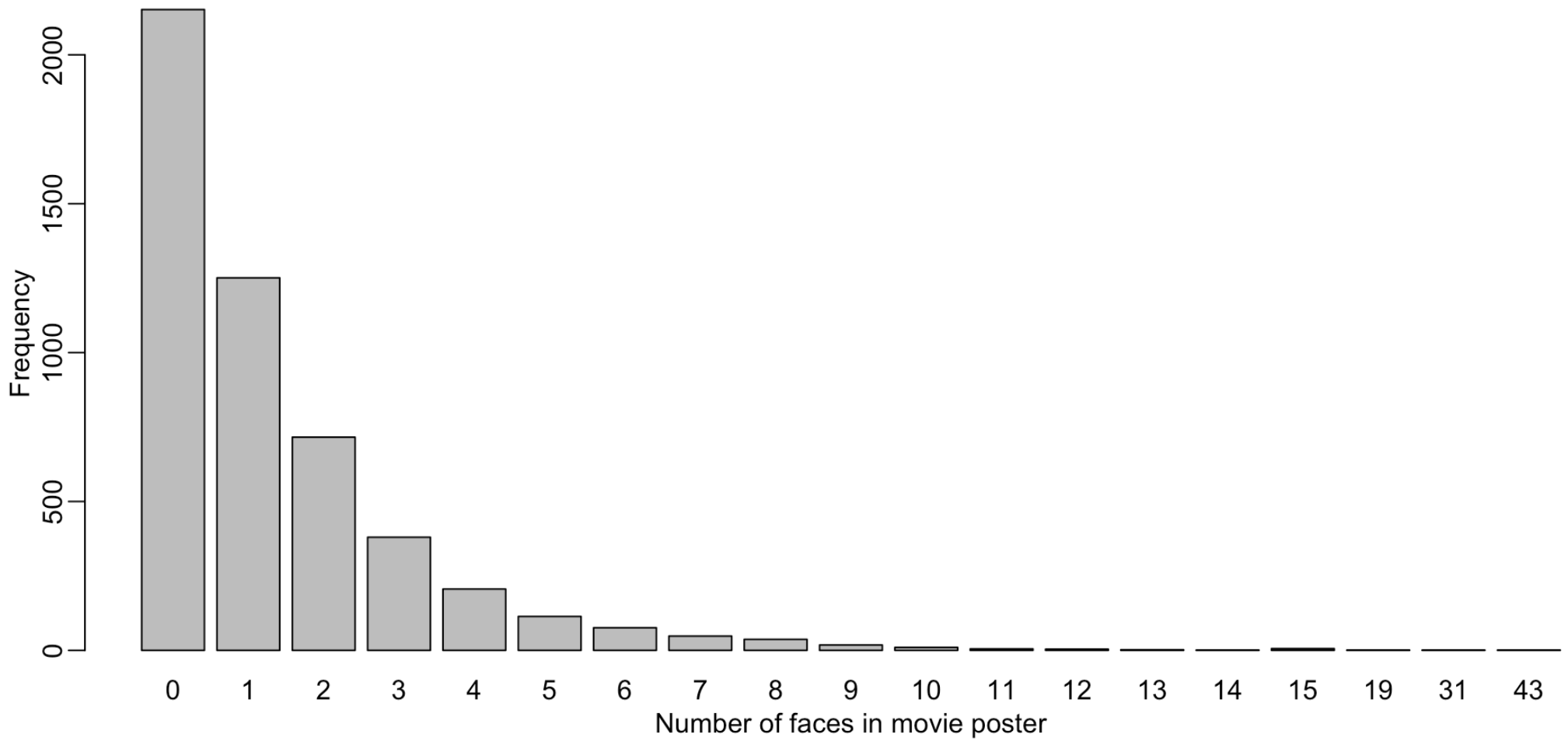
May require more explanation. Problems with finite domains

Discrete 1D data

```
(mpt = table(imdb$facenumber_in_poster))
```

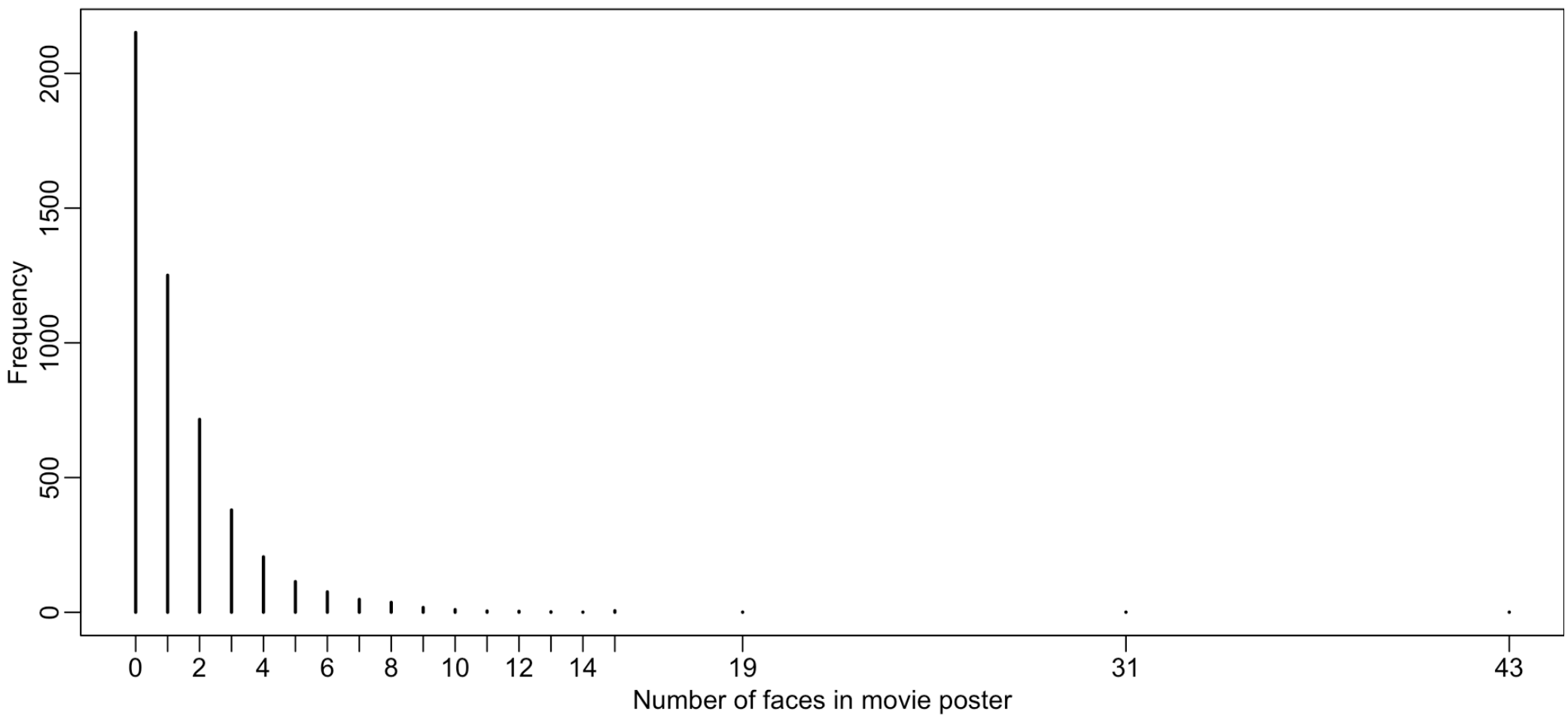
```
   0    1    2    3    4    5    6    7    8    9   10   11   12   13   14
2152 1251  716  380  206  114   76   48   37   18   10    5    4    2    1
  15   19   31   43
   6    1    1    1
```

```
1 barplot(mpt, ylab="Frequency",xlab="Number of faces in movie poster")
```



Mind the (missing) gaps

```r
plot(mpt, ylab="Frequency",xlab="Number of faces in movie poster")
```

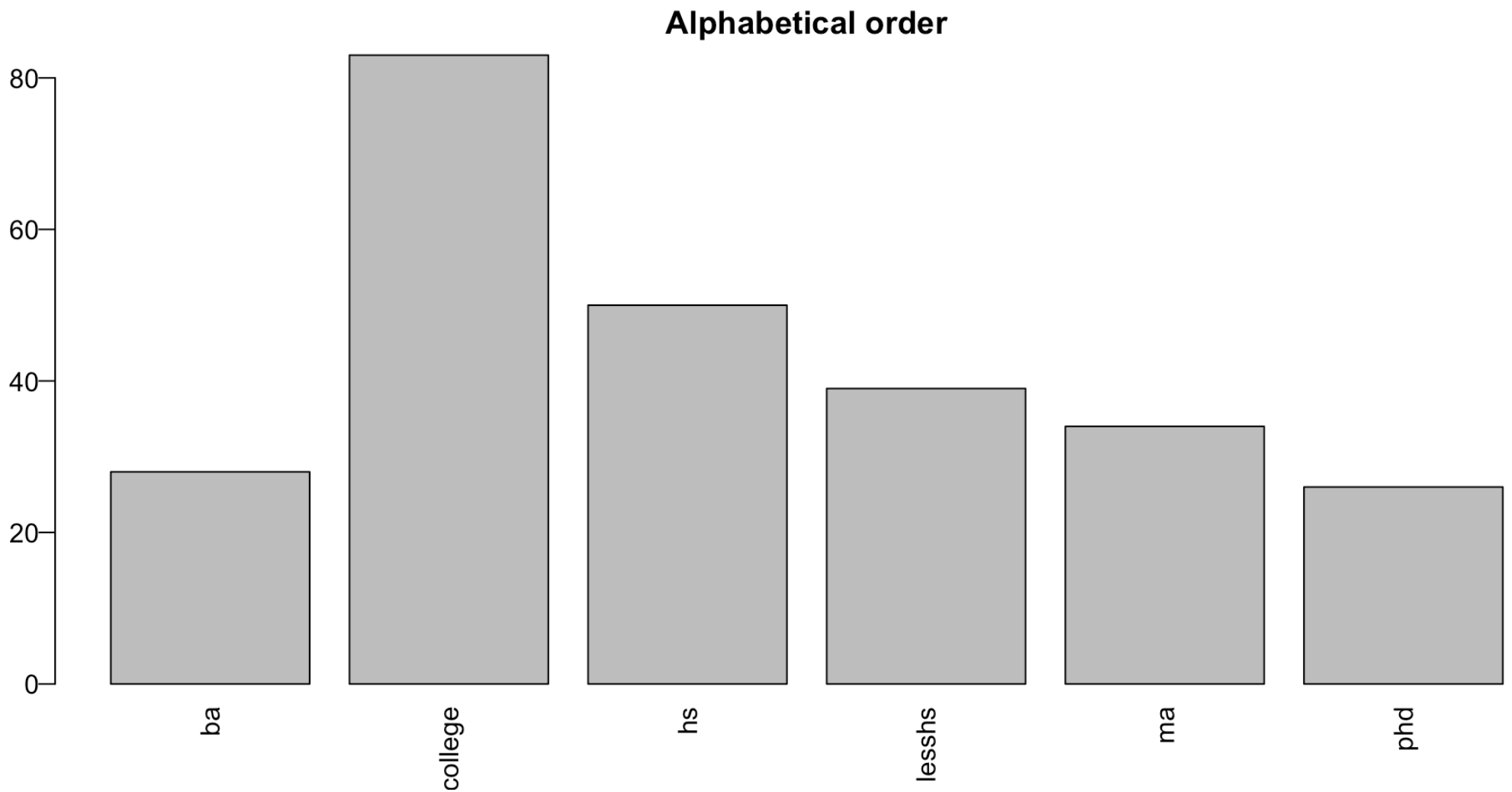


Default plot() does the job

Categorical 1D data

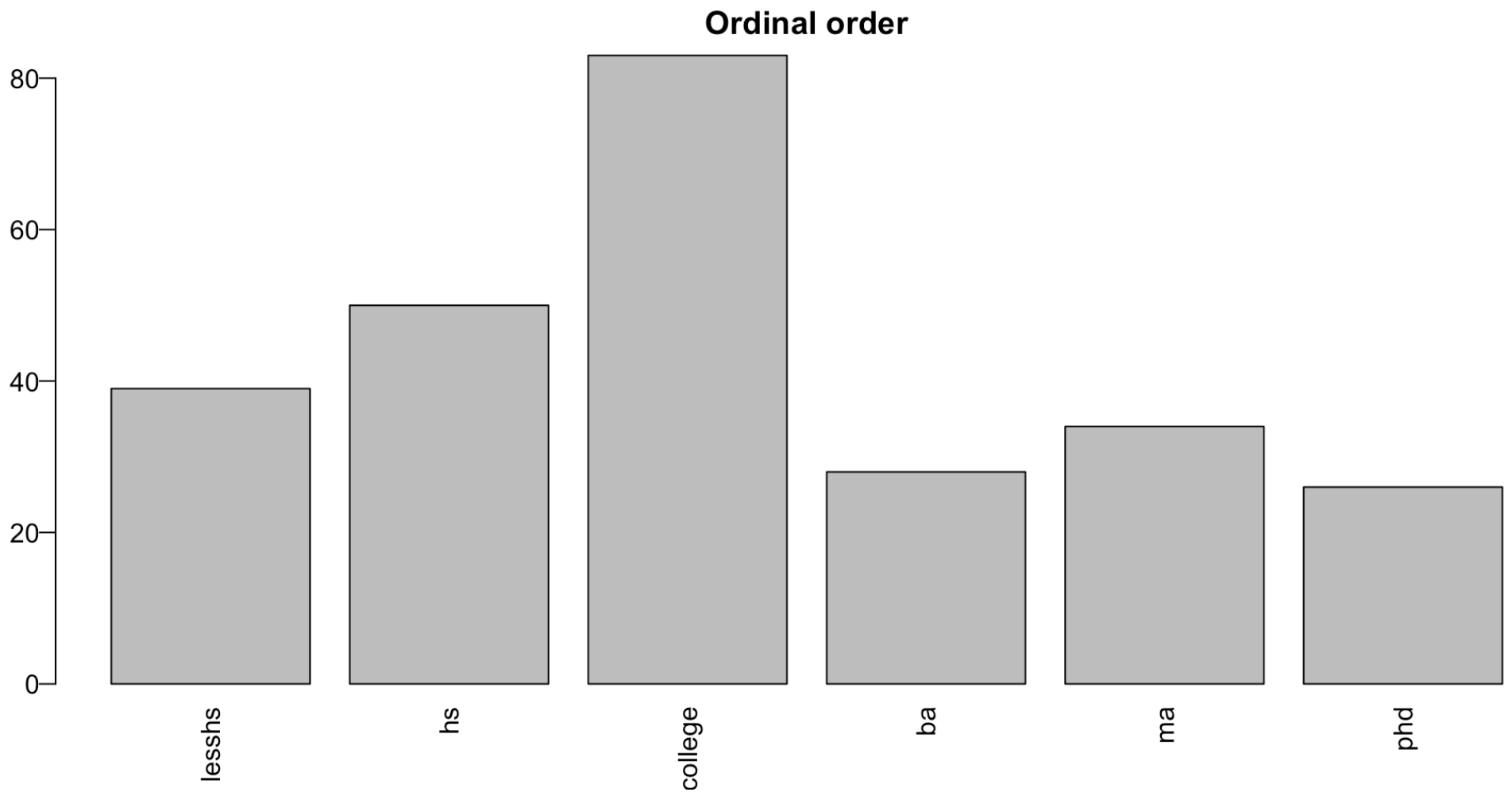
```
paredu
     ba college       hs  lesshs       ma      phd
     28      83       50      39       34       26
```

```r
barplot(edt,main="Alphabetical order",las=2)
```



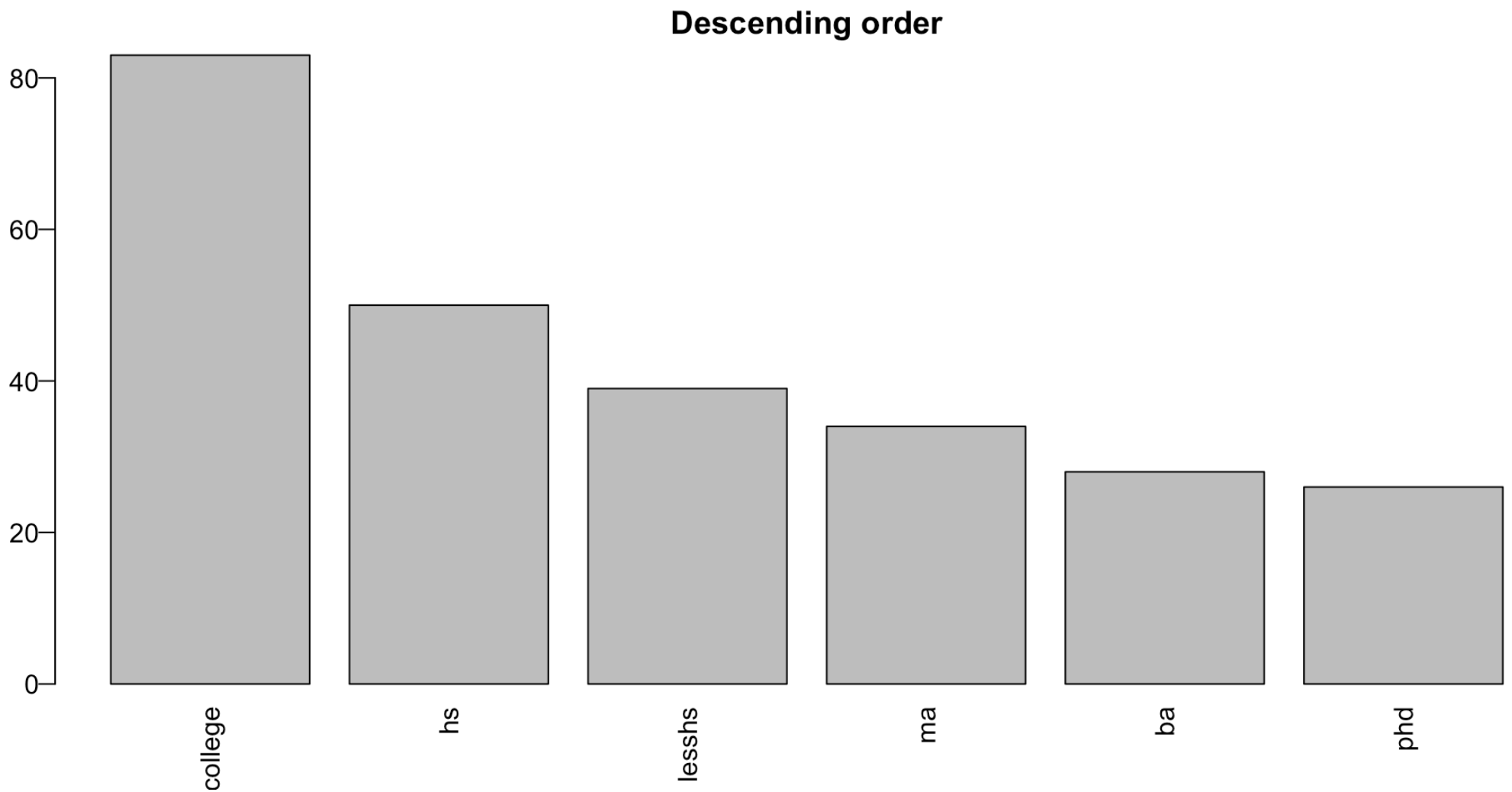
Order is not correct

```r
nels88$paredu=factor(nels88$paredu,
                     levels=c("lesshs","hs","college","ba","ma","phd"))
edt = xtabs(~ paredu, nels88)
barplot(edt,main="Ordinal order",las=2)
```



Natural order

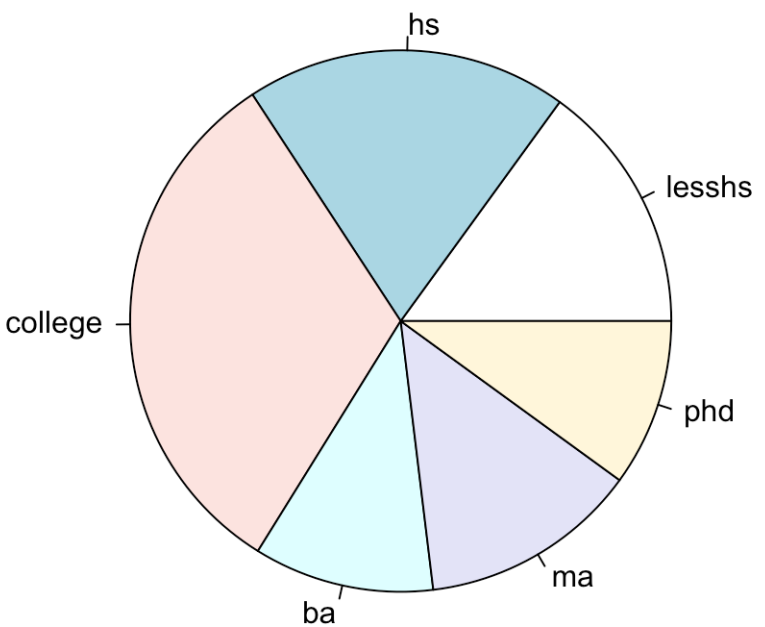
```r
barplot(edt[order(edt, decreasing = TRUE)],main="Descending order",las=2)
```



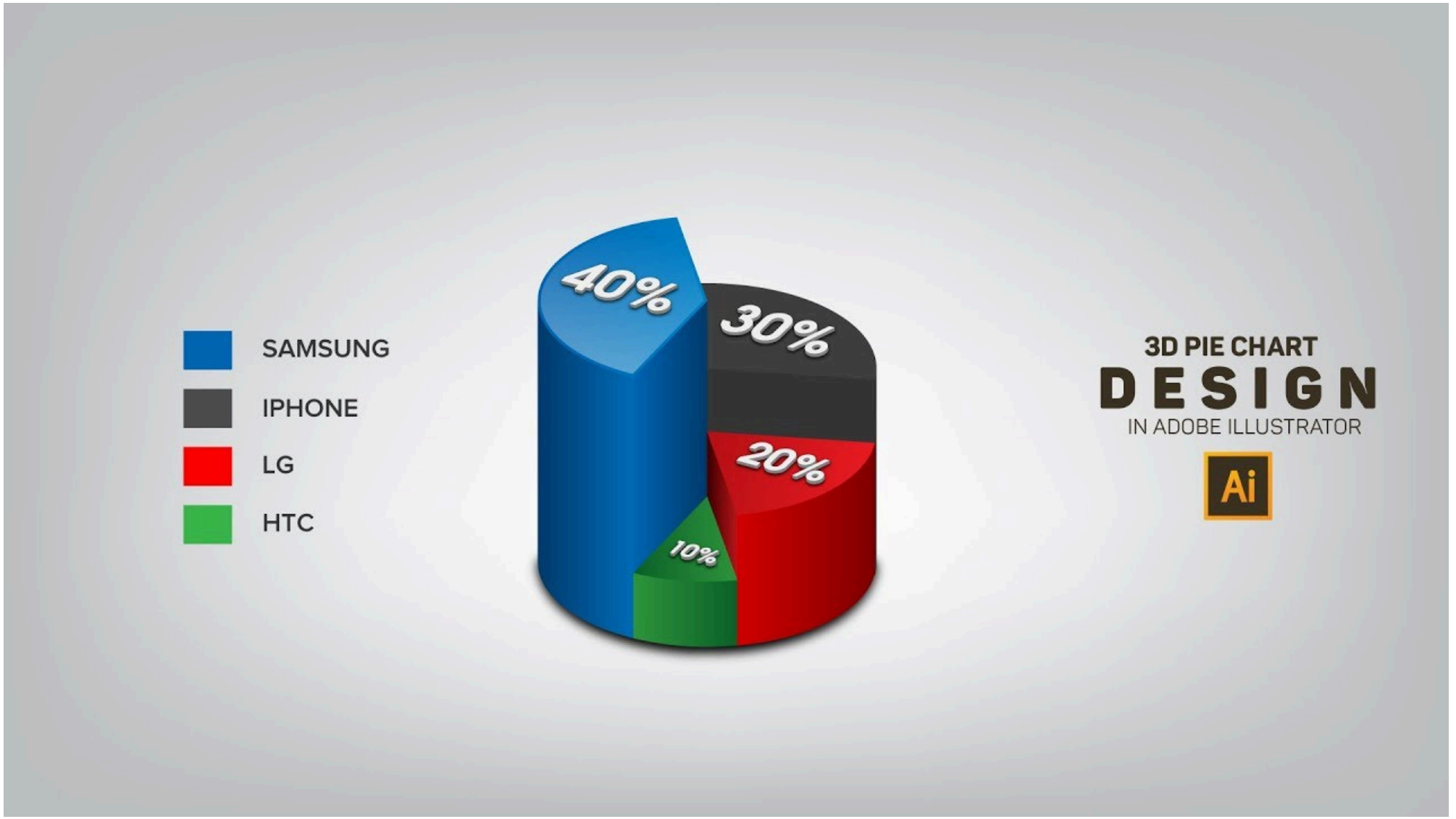
Pareto chart, see also 80/20 rule

Pie charts

```
1 pie(edt)
```



Pie charts are not recommended because psychologists have found that we are much better at judging linear quantities than angles or areas.



Never, ever, do this.

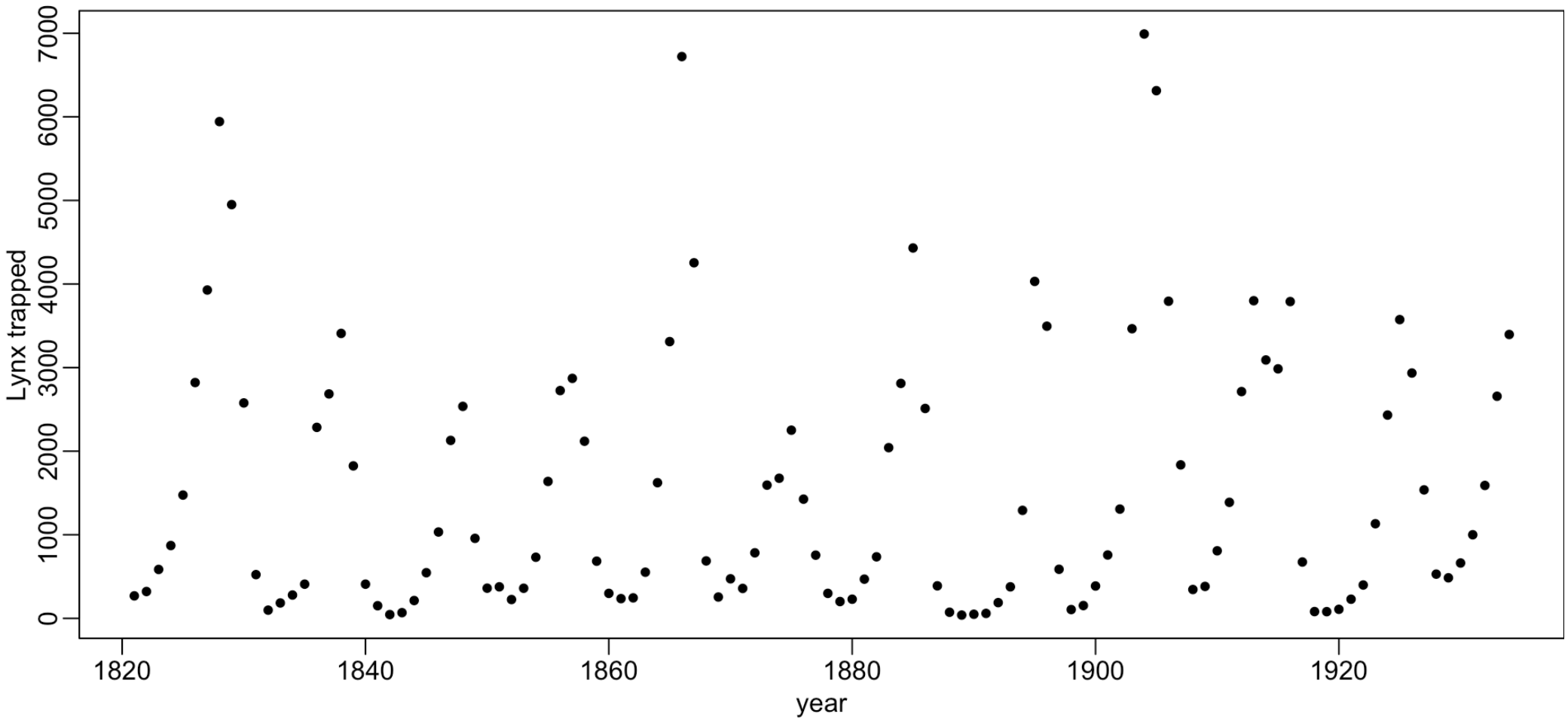
Time Series

```
1 lynx
```

```
Time Series:
Start = 1821
End = 1934
Frequency = 1
  [1]  269  321  585  871 1475 2821 3928 5943 4950 2577  523   98  184  279
 [15]  409 2285 2685 3409 1824  409  151   45   68  213  546 1033 2129 2536
 [29]  957  361  377  225  360  731 1638 2725 2871 2119  684  299  236  245
 [43]  552 1623 3311 6721 4254  687  255  473  358  784 1594 1676 2251 1426
 [57]  756  299  201  229  469  736 2042 2811 4431 2511  389   73   39   49
 [71]   59  188  377 1292 4031 3495  587  105  153  387  758 1307 3465 6991
 [85] 6313 3794 1836  345  382  808 1388 2713 3800 3091 2985 3790  674   81
 [99]   80  108  229  399 1132 2432 3574 2935 1537  529  485  662 1000 1590
[113] 2657 3396
```

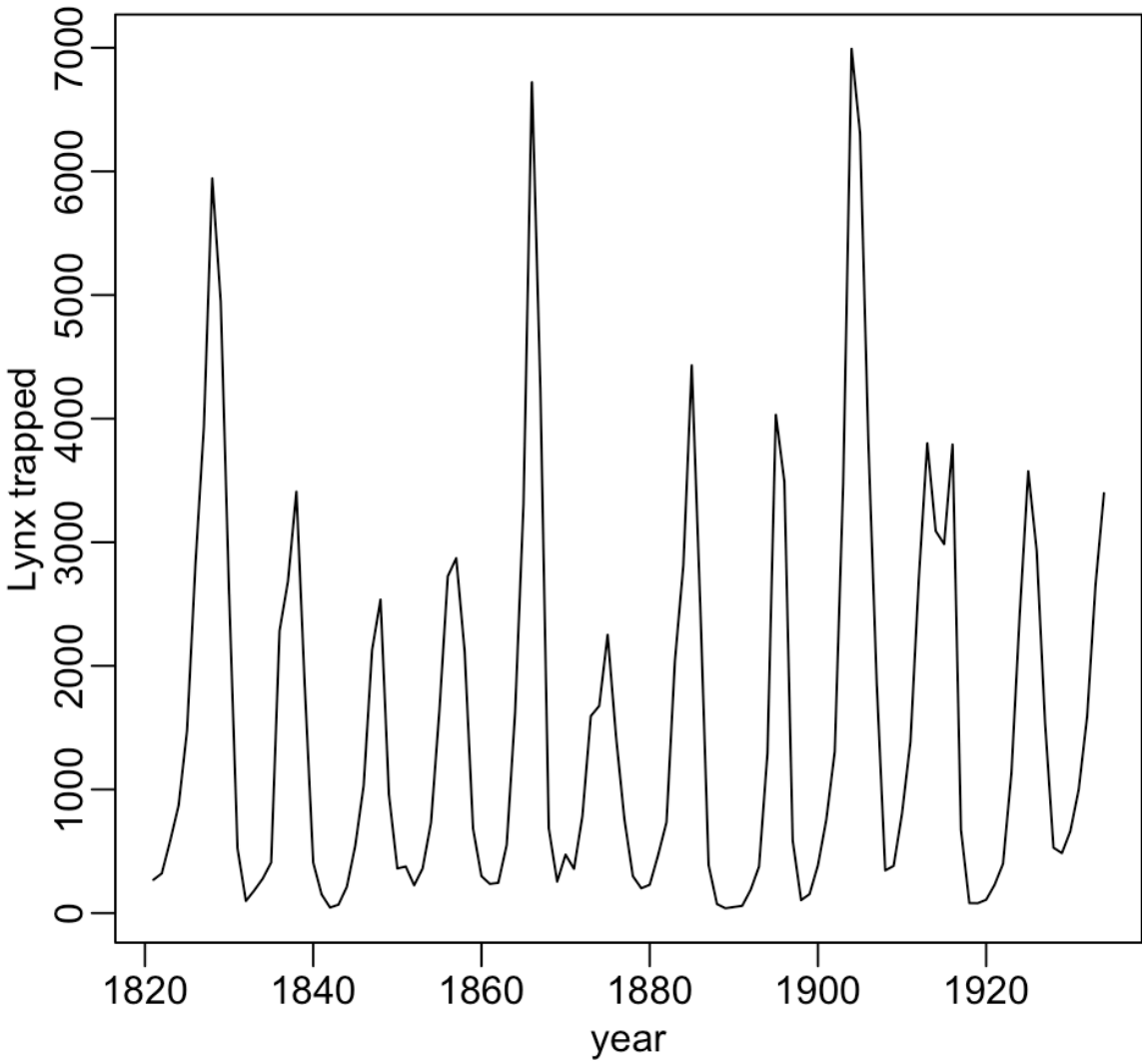
```
1 yr = 1821:1934
```

```r
plot(yr, lynx, xlab="year",ylab="Lynx trapped")
```



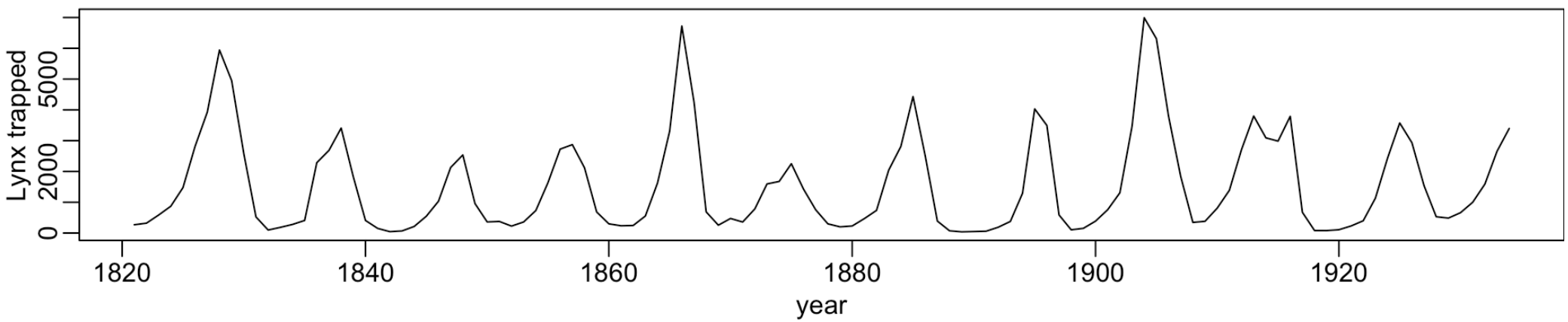
Don’t use points

```r
plot(yr, lynx, type="l", xlab="year",ylab="Lynx trapped")
```

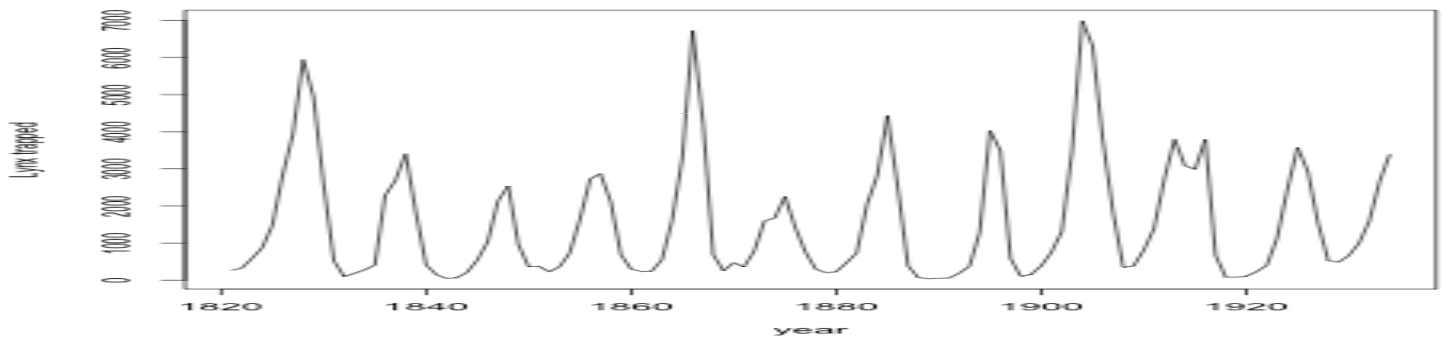


Lines are better

```r
plot(yr, lynx, type="l", xlab="year",ylab="Lynx trapped")
```



Adjust aspect ratio to get 45 degree slopes

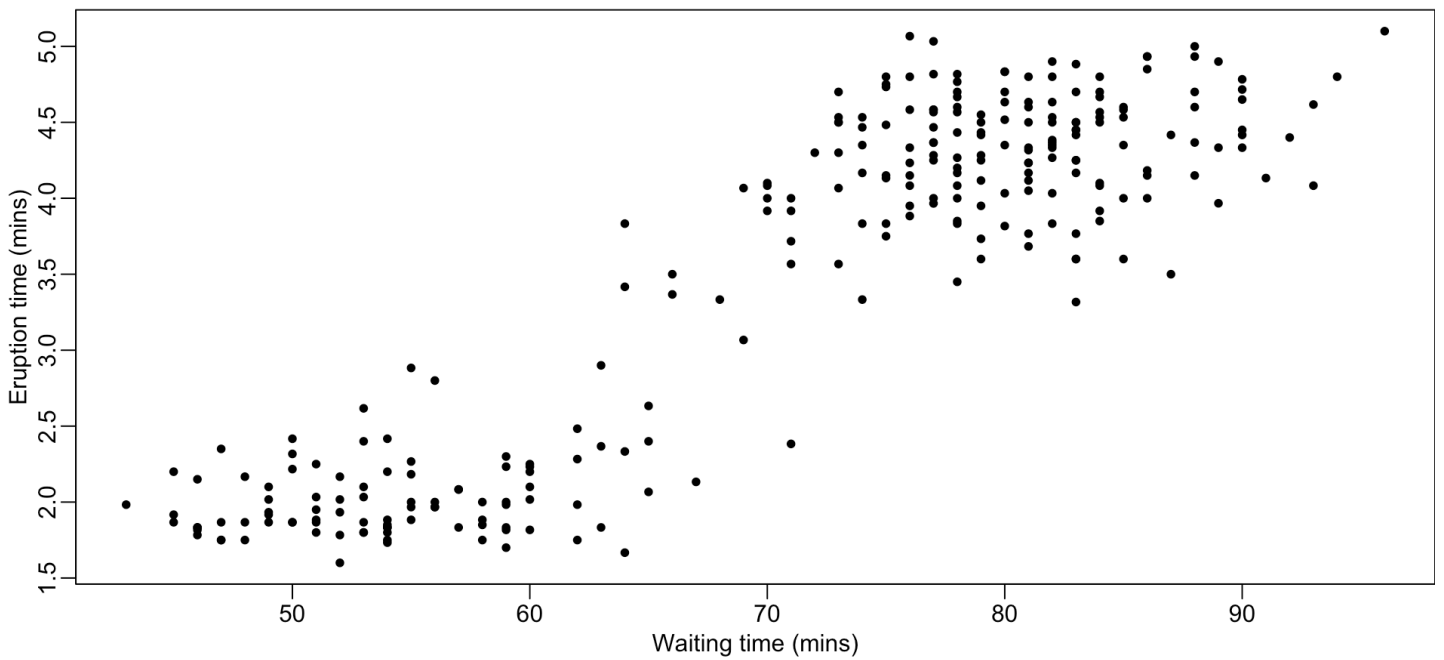


How to make an ugly graphic:

1. Use a bitmapped format for the original plot
2. Change the aspect ratio from the original

2D Scatterplots

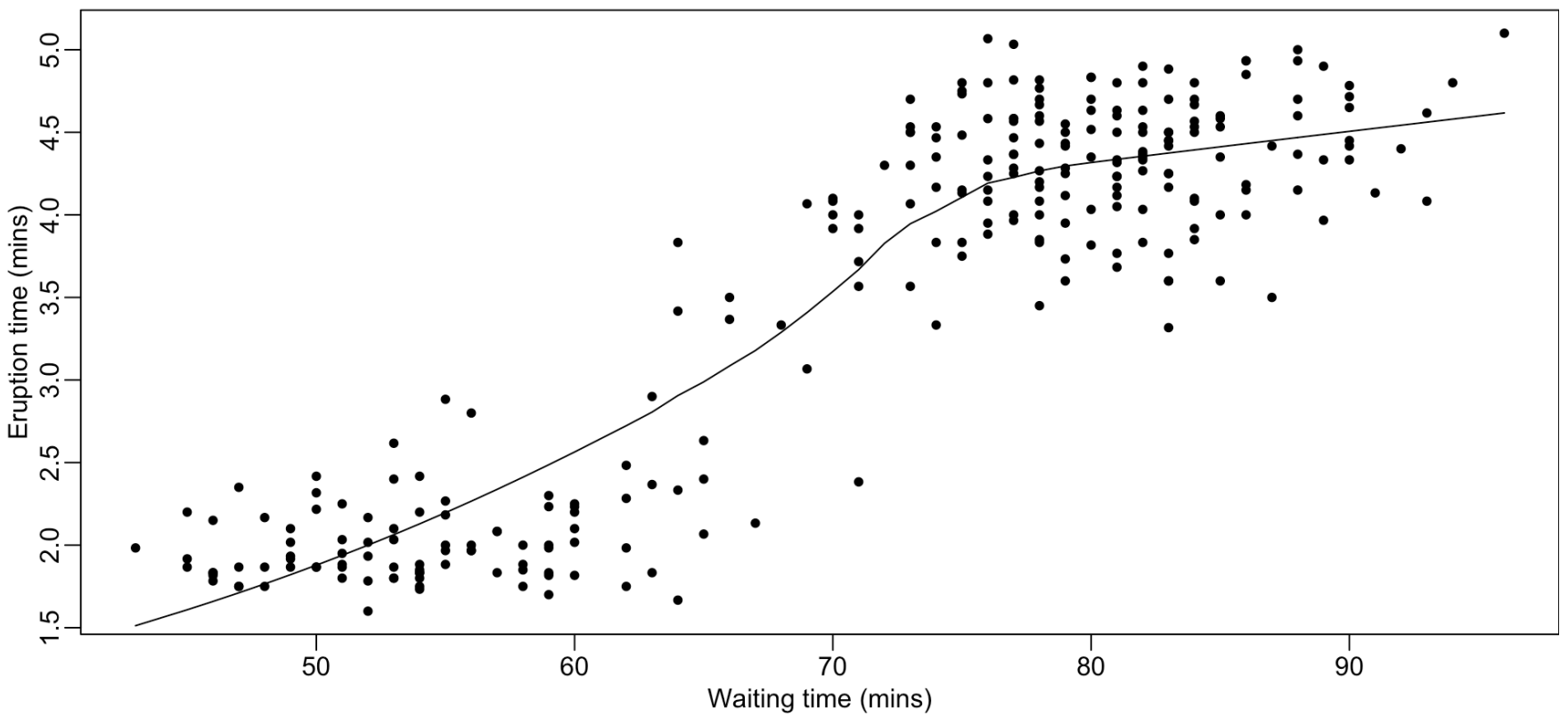
```r
plot(eruptions ~ waiting, faithful,
     xlab="Waiting time (mins)", ylab="Eruption time (mins)")
```



Think about: axis spacing and labelling, plotting character, aspect ratio, margin spacing

Smoothed line

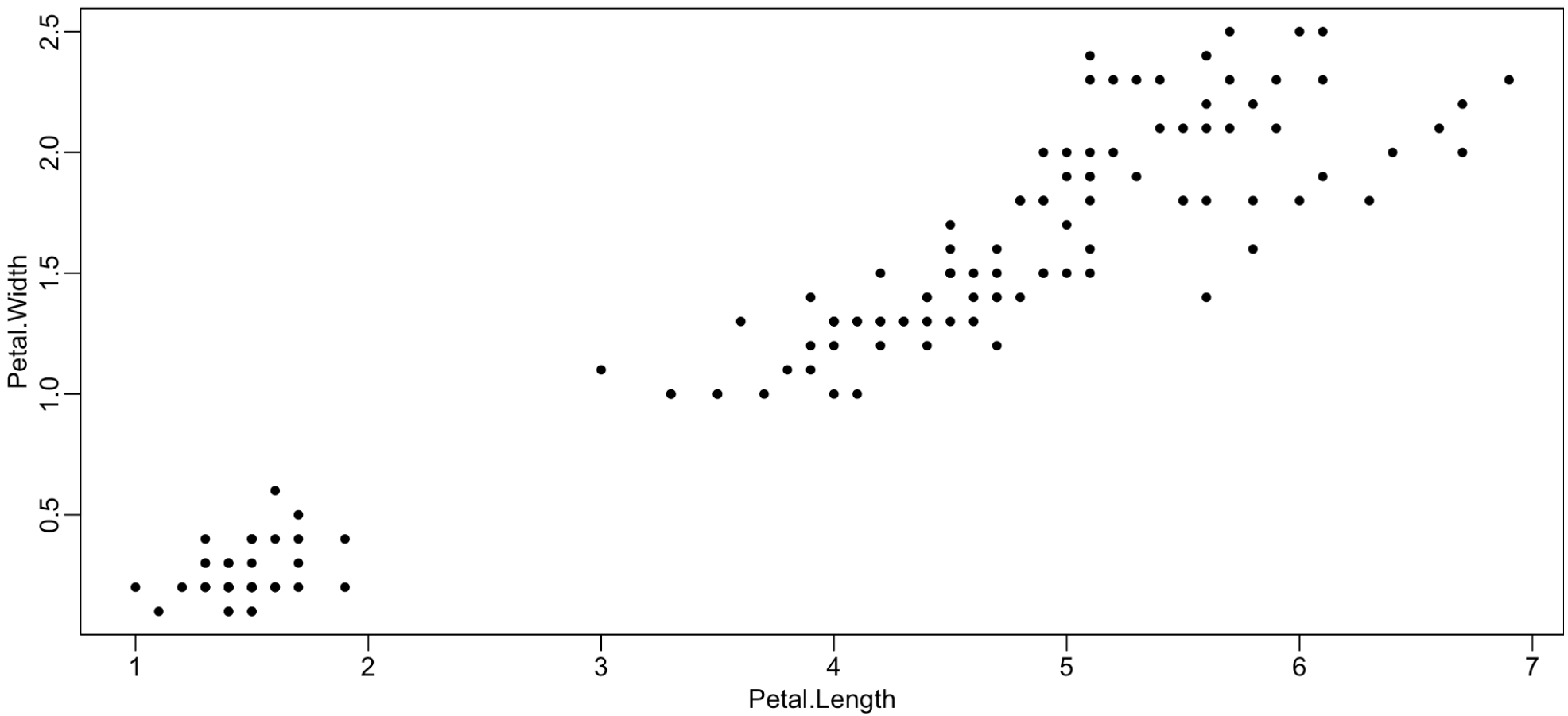
```
plot(eruptions ~ waiting, faithful,
     xlab="Waiting time (mins)", ylab="Eruption time (mins)")
with(faithful,lines(lowess(waiting,eruptions)))
```



Sometimes adding a smoothed line can reveal structure

Overplotting

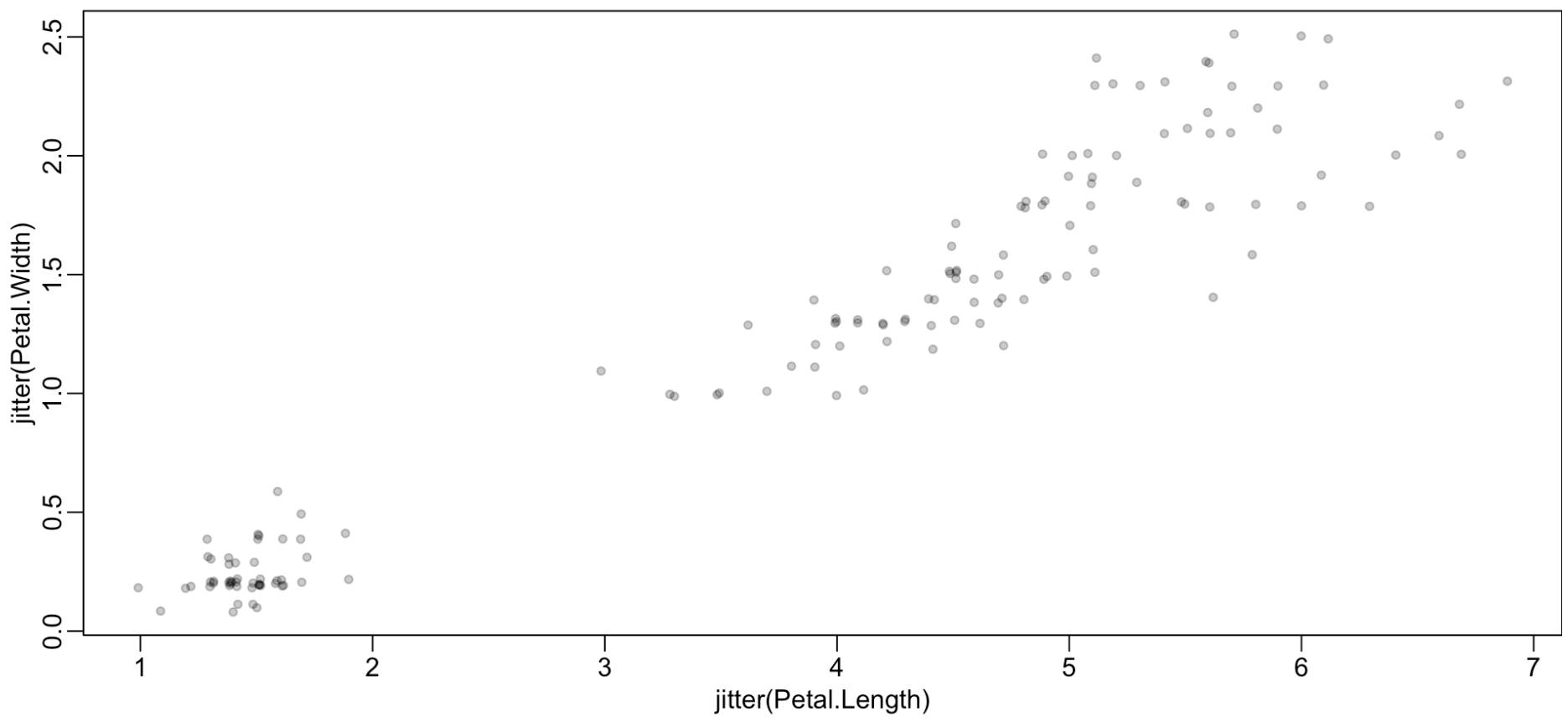
```
1 plot(Petal.Width ~ Petal.Length, iris)
```



Overplotted points may not be apparent

Jittering

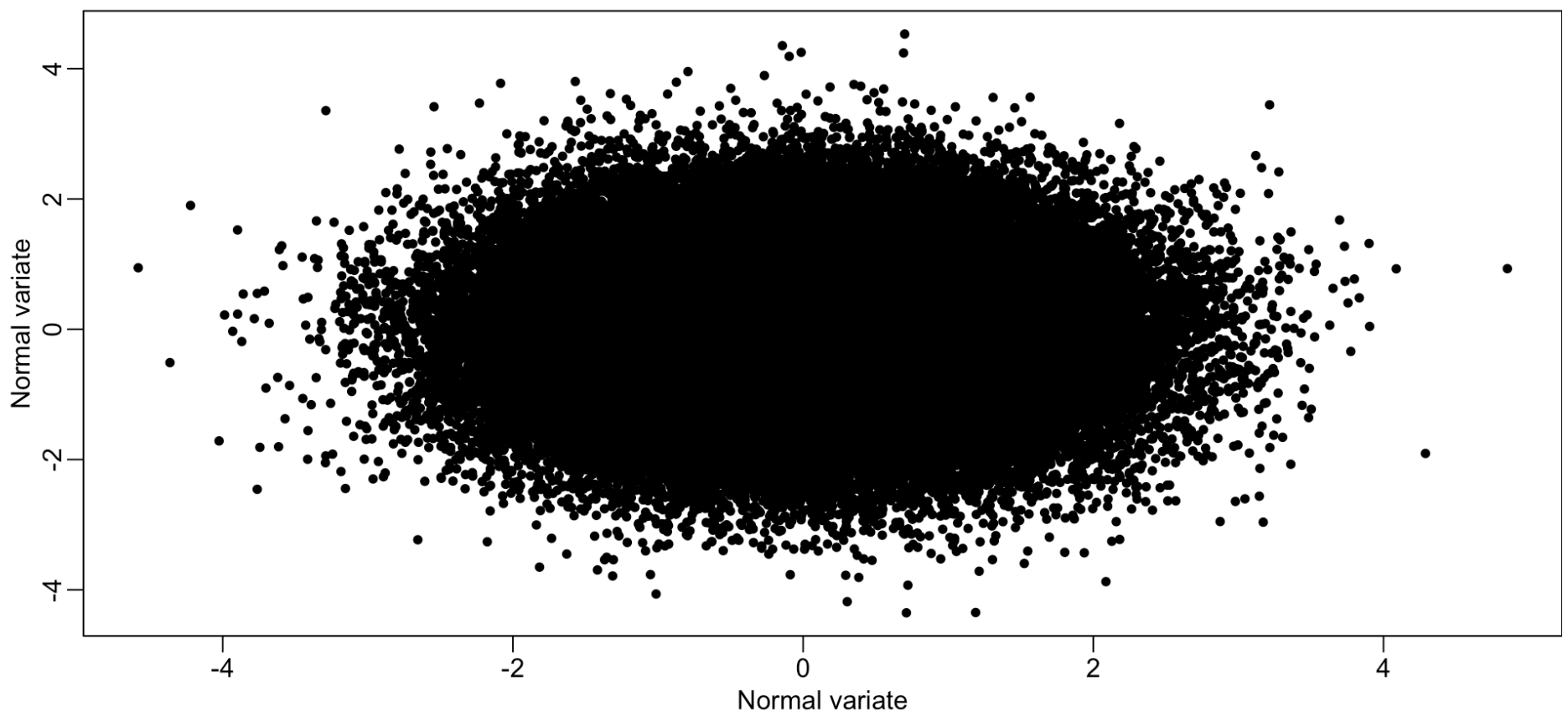
```
1 plot(jitter(Petal.Width) ~ jitter(Petal.Length), iris,
2      col = rgb(0,0,0,0.25))
```



Alpha transparency is also helpful

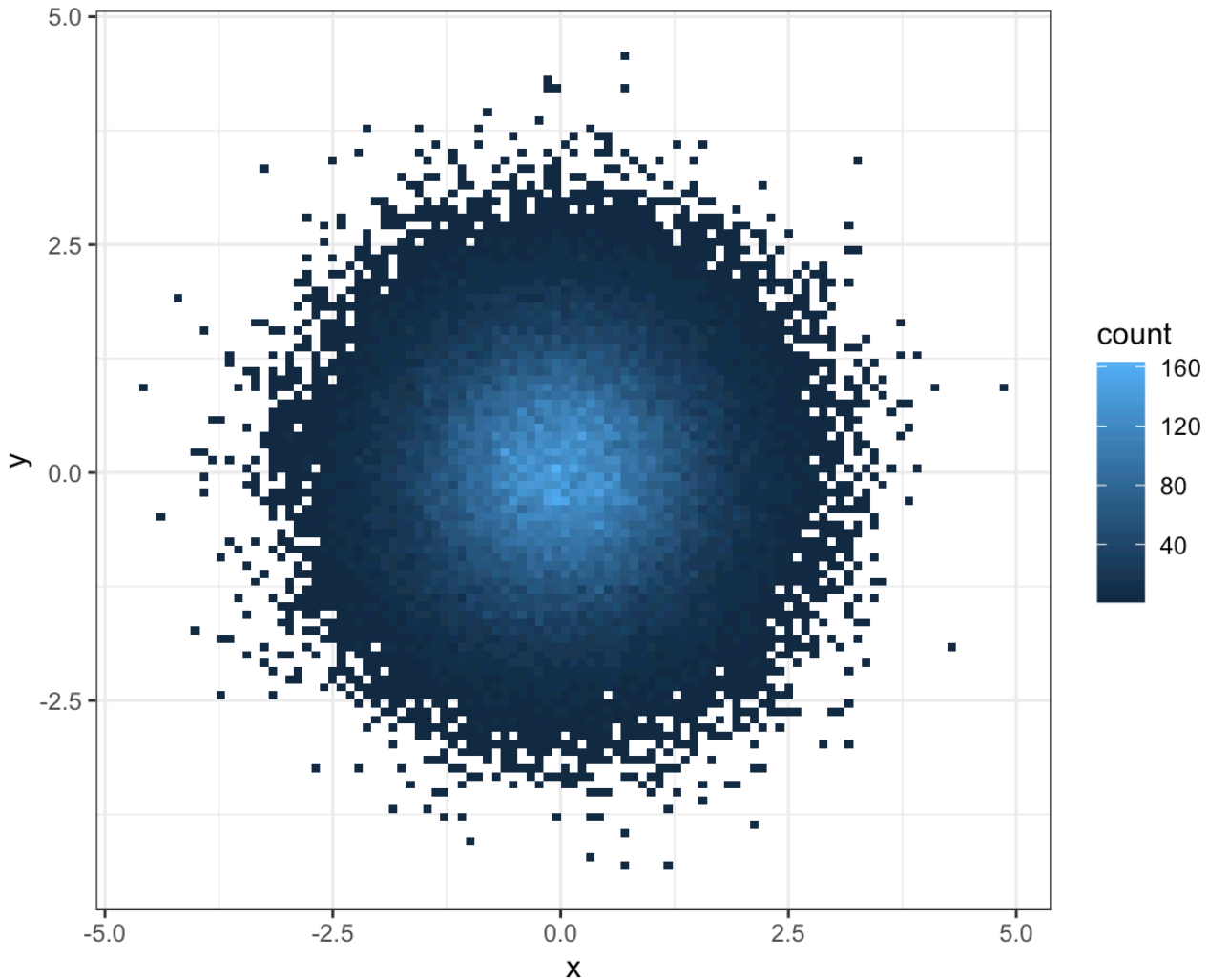
High density plots

```
1 bdf = data.frame(x=rnorm(1e5),y=rnorm(1e5))
2 plot(y ~ x, bdf,xlab="Normal variate",ylab="Normal variate")
```



Big blob - can’t see structure. Would use a lot of ink

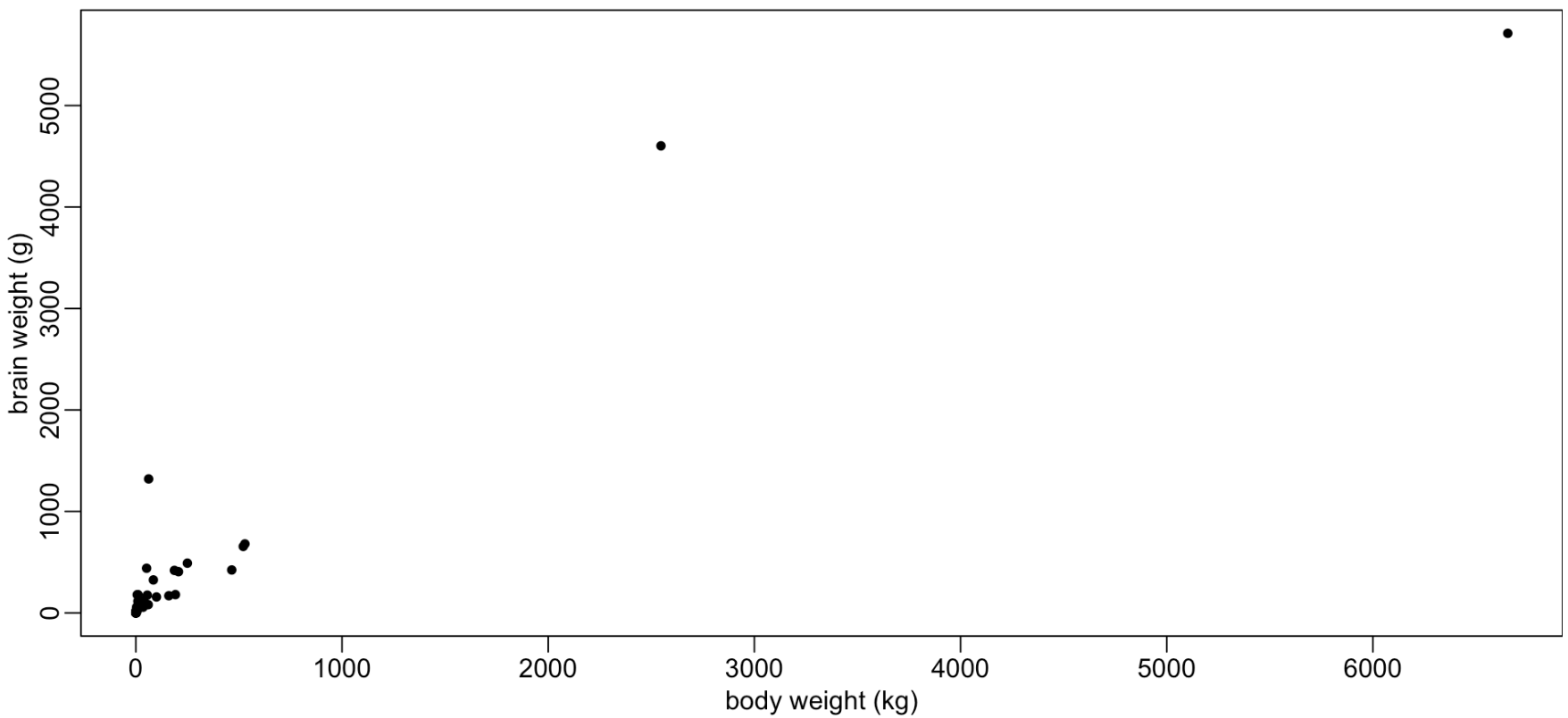
```r
library(ggplot2)
ggplot(bdf,aes(x,y)) + geom_bin2d(bins=100)+ coord_fixed()
```



2D binned histogram

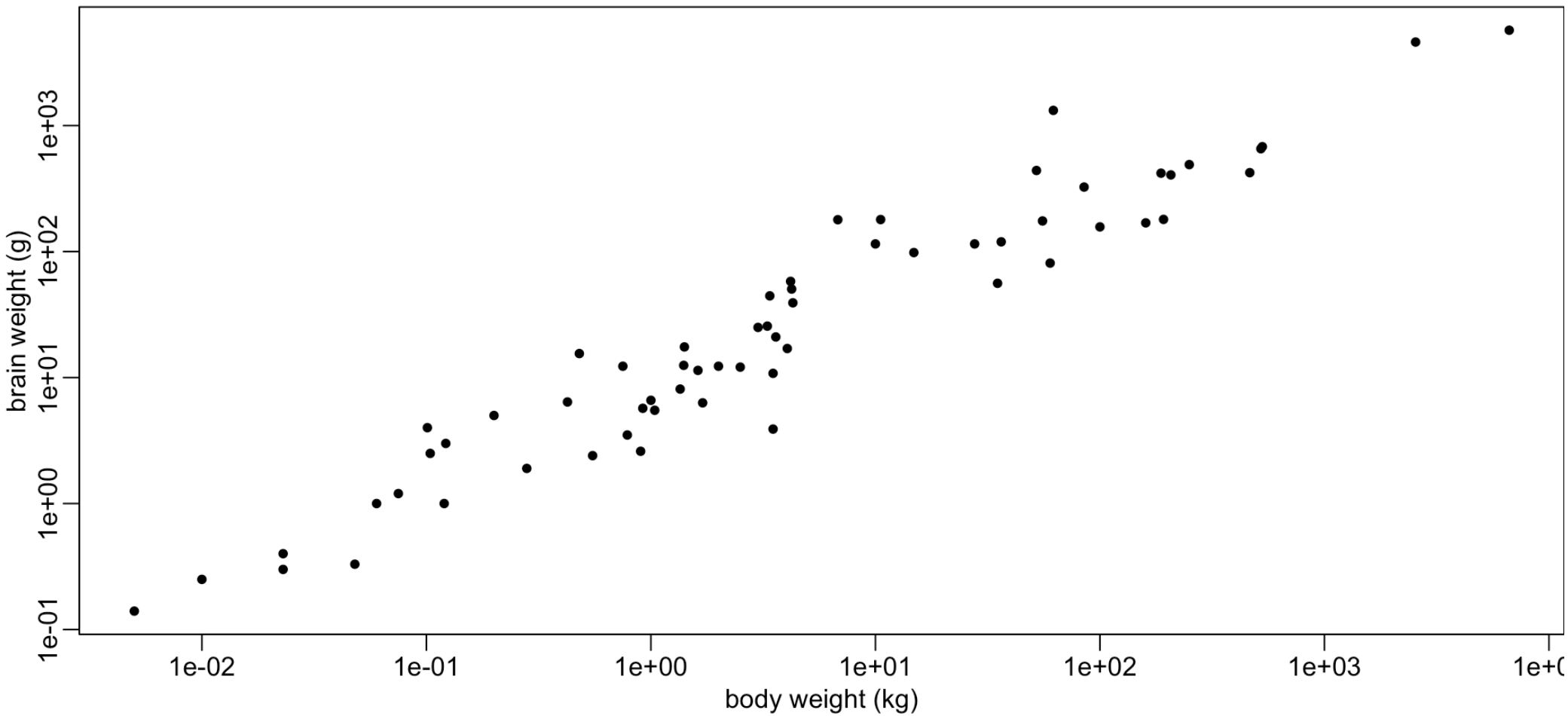
Log Scales

```
1 data(mammalsleep,package="faraway")
2 plot(brain ~ body, mammalsleep,
3      xlab="body weight (kg)",ylab="brain weight (g)")
```

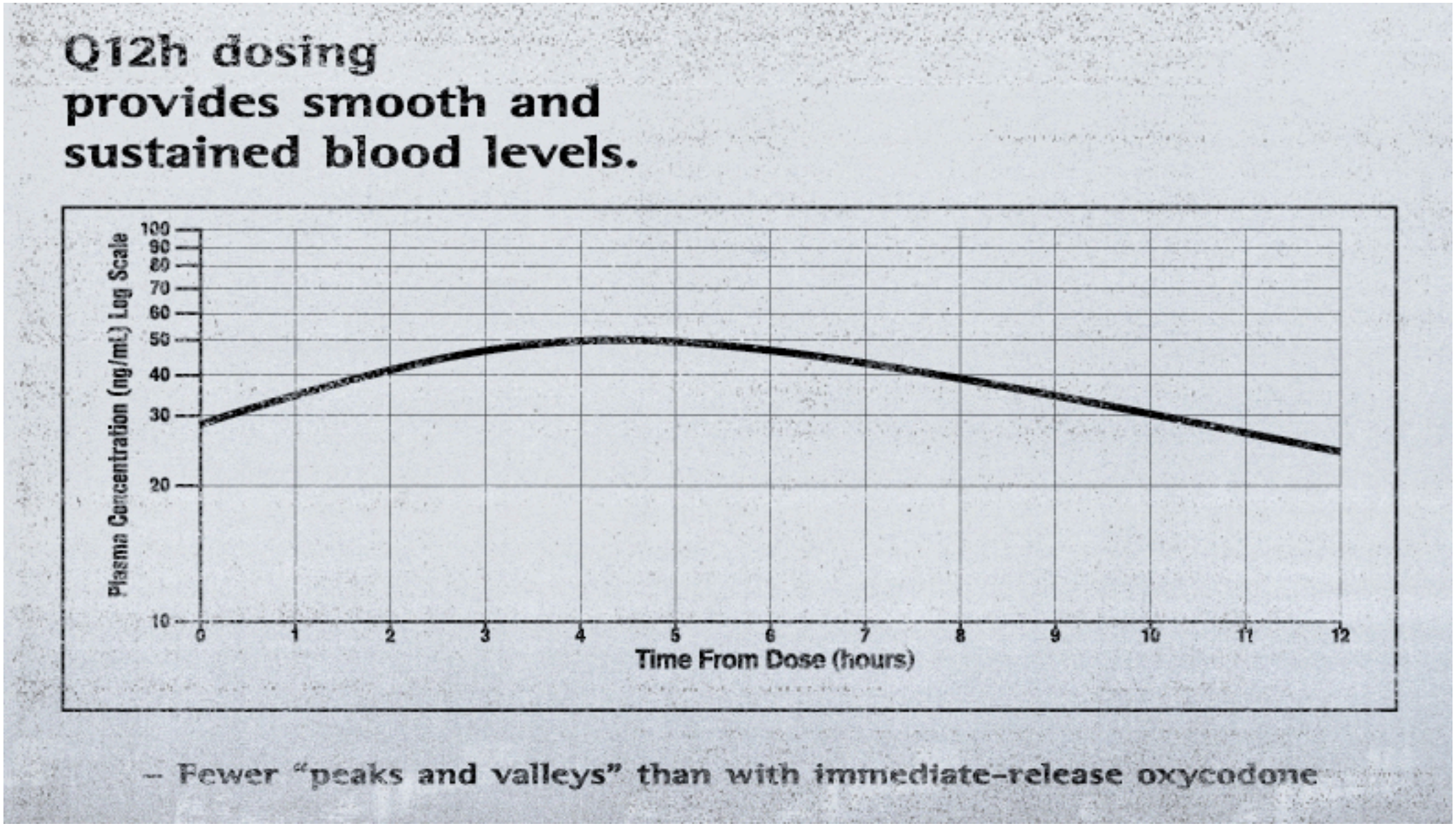


Data ranges over orders of magnitude

```r
plot(brain ~ body, mammalsleep, log="xy",
     xlab="body weight (kg)",ylab="brain weight (g)")
```



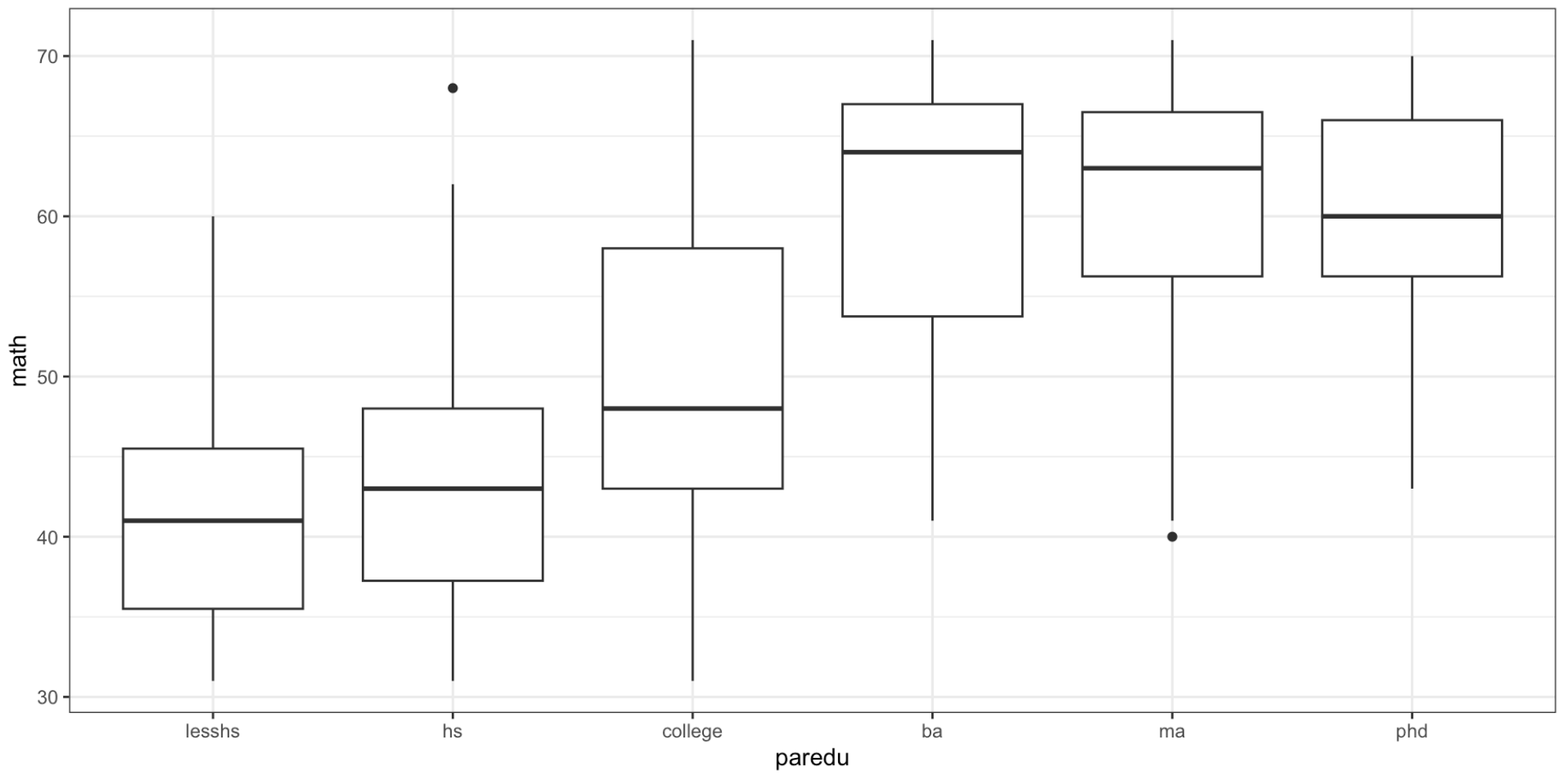
True relationship is revealed



Log scales are not always good

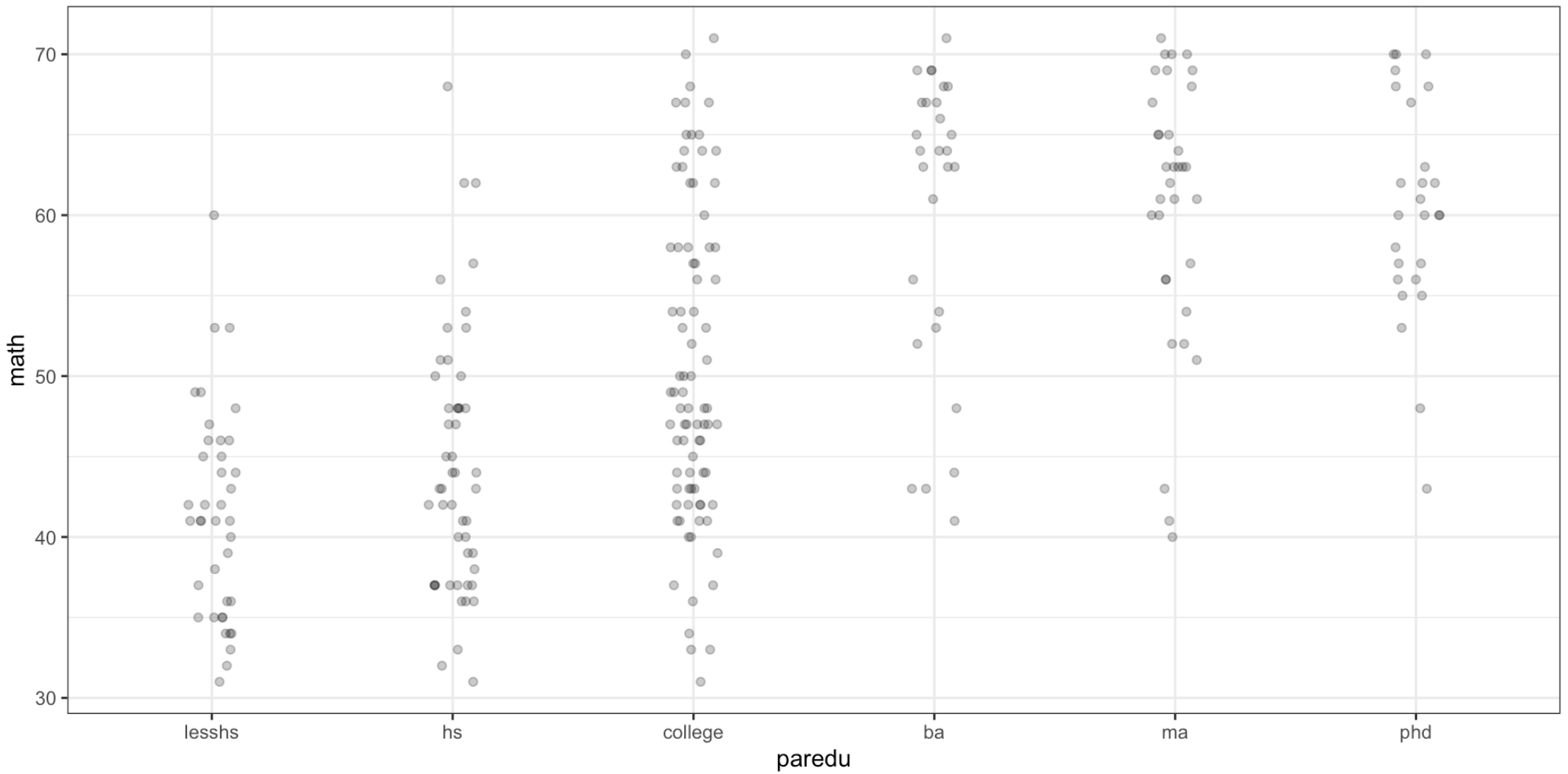
One discrete and one continuous

```
1 ggplot(nels88, aes(paredu, math)) + geom_boxplot()
```



Boxplots are the most common choice

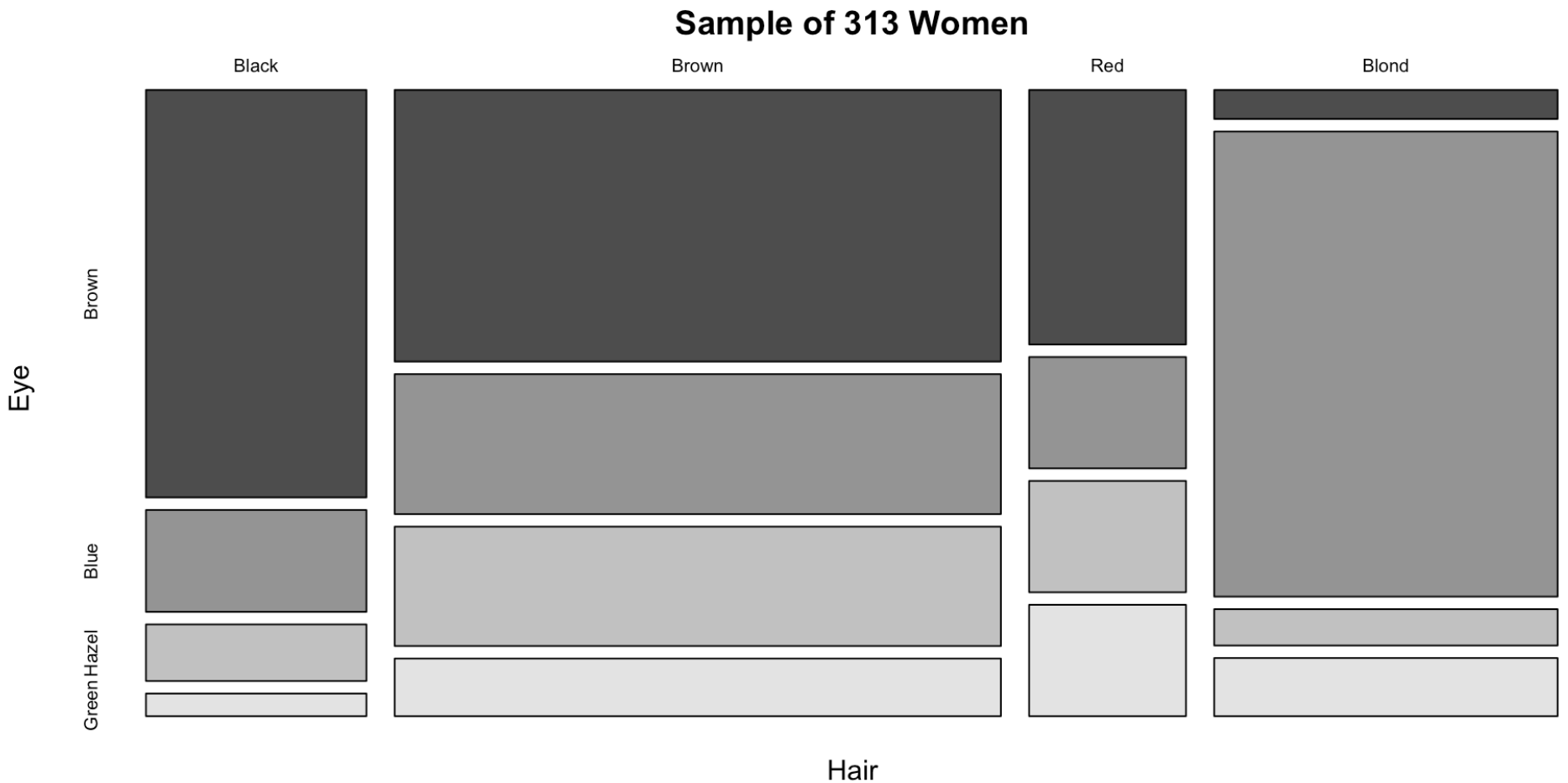
```
1 ggplot(nels88, aes(paredu, math)) + geom_jitter(width=0.1,height=0,alpha=0.
```



Sometimes plotting points with some jittering is better

Two discrete

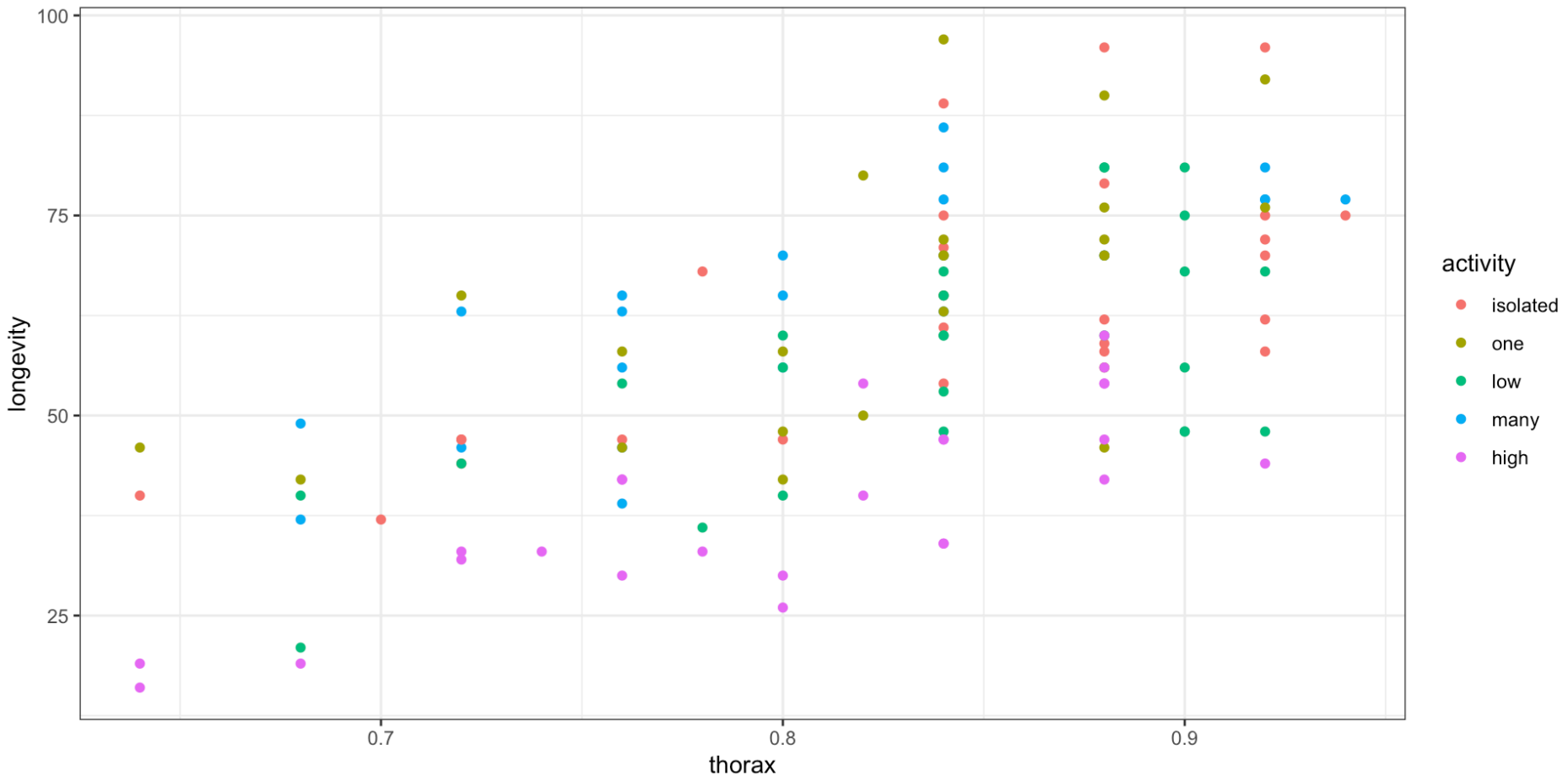
```
1 mosaicplot(HairEyeColor[,,2],color = TRUE,main="Sample of 313 Women")
```



Works for relative few categories

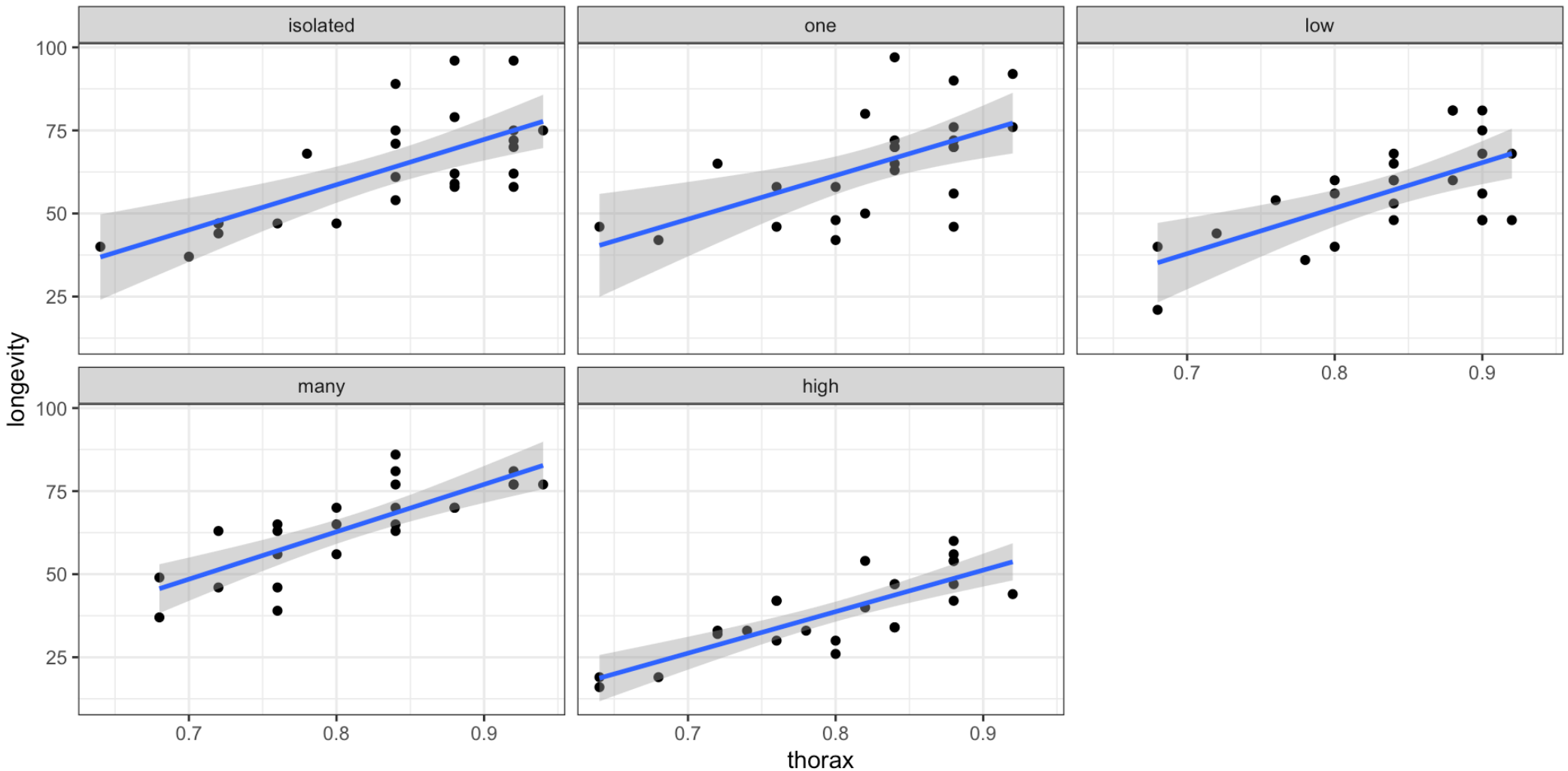
Two continuous and one categorical

```
1 data(fruitfly,package="faraway")
2 ggplot(fruitfly, aes(thorax,longevity,color=activity)) + geom_point()
```



Use a different plotting character

```
ggplot(fruitfly, aes(thorax,longevity)) + geom_point() +
  facet_wrap(~ activity) + geom_smooth(method="lm")
```



Use facets

Missing Data

Possibly the most annoying and time consuming problem with data

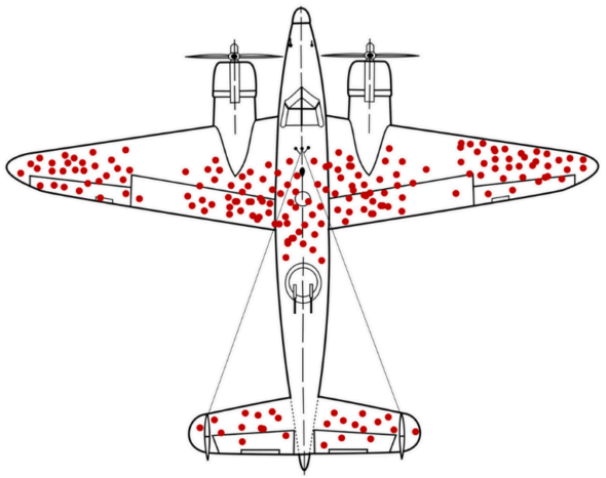
Coding of Missing Data

- NA in R
- NaN means “Not a Number” i.e. 0/0 is different
- None in Python (also has NaN)
- Excel often uses empty cell (may cause problems)
- Other codes such as -999 are used (sometimes without explanation)
- Maybe an explanation for why missing

Reasons for missingness

- Statisticians distinguish missingness for reasons:
  - not related to the data
  - related to the seen data
  - related to what *would have been observed*
- Analysis will depend on this assumption.

Missing cases or Unseen data



WW2 bomber

- Wald was asked: *Where to reinforce the bomber?*
- Ans: At locations with less damage.
- Sampling bias (Survivor bias)
- Internet meme (Wald never made such a plot)

Imputation of missing values

- Replacing missing values with estimated values
- Often uses mean or most common value for categorical data
- Time series might use
  - LOCF (last observation carried forward)
  - Interpolation
- Imputation reduces variability and can lead to bias
- Multiple imputation preferred to single imputation
- Don’t impute now, retain the missing values and leave the decision to the modelling stage.

Outliers

- Extreme values in a variable (no fixed definition)
- Numerical and graphical summaries should help detect.
- Can have a serious effect on the model estimation.
- Errors vs. actually happened
- Multivariate or clusters of outliers harder to find
- Less of a problem for larger datasets.
- Identify but delay action

Practice

Wide diversity in

- types of data
- problems encountered
- helpful tools

Experience is key - this cannot be automated.