Data Types, Data Visualization, and Basic Analytics Techniques

Dr. Paul McNicholas

November 4, 2019

Winter School on Computational Data Science and Optimization

Introduction

- This winter school will aim to provide a general introduction to data science, and machine learning techniques.

- To use these approaches effectively in practice, we encourage you to take appropriate courses and programs.

- The slides, R code, and other materials related to this winter school can be found in the GitHub repository at the following address: https://github.com/gallaump/Fields-Winter-School-on-Computational-Data-Science-and-Optimization

Data Science

- What is data science anyway?
- That is a good question and I will not attempt to answer it today.
- Some words of wisdom:

  "...we must help the student to recognise the computer for what it is — a sophisticated tool, not a substitute for thought." (Barnett, 1999)[a]

[a] Barnett, V. (1999), *Comparative Statistical Inference*, 3rd edition, Chichester: England.

More Words of Wisdom

- From Tukey and Wilk (1966)[a]:

> Nothing — not the careful logic of mathematics, not statistical models and theories, not the awesome arithmetic power of modern computers — nothing can substitute here for the flexibility of the informed human mind. . . Accrdingly, both analysis approaches and techniques need to be structured so as to facilitate human involvement and intervention.

[a]Tukey, J. and Wilk, M. (1966), 'Data Analysis and Statistics: An Expository Overview', AFIPS, International Workshop on Managing Requirements Knowledge, pp. 695–709.

A Brief Note on Data Types

- Data come in many different types.
- Some Examples:
  - binary: yes/no, 0/1.
  - nominal: eye colour, race, blood type.
  - ordinal: income bracket, education level.
  - counts: number of children, number of earthquakes in a year.
  - continuous: weight, height, length.
  - more complex types: three-way and multiway data (images, video clips), clickstream data, text.

Introduction to Data Visualization

- People talk about "looking at data", "looking at plots", etc.
- More than just looking at data, we want to be able to develop hypotheses or draw tentative conclusions using graphs.
- This is sometimes called graphical data analysis (GDA) — in a *bona fide* GDA, a lot of graphs would be used.
- First, we will look at some data sets; then we will look at R code.
- Many of the figures we will look at are based on examples in Unwin (2015)[a].

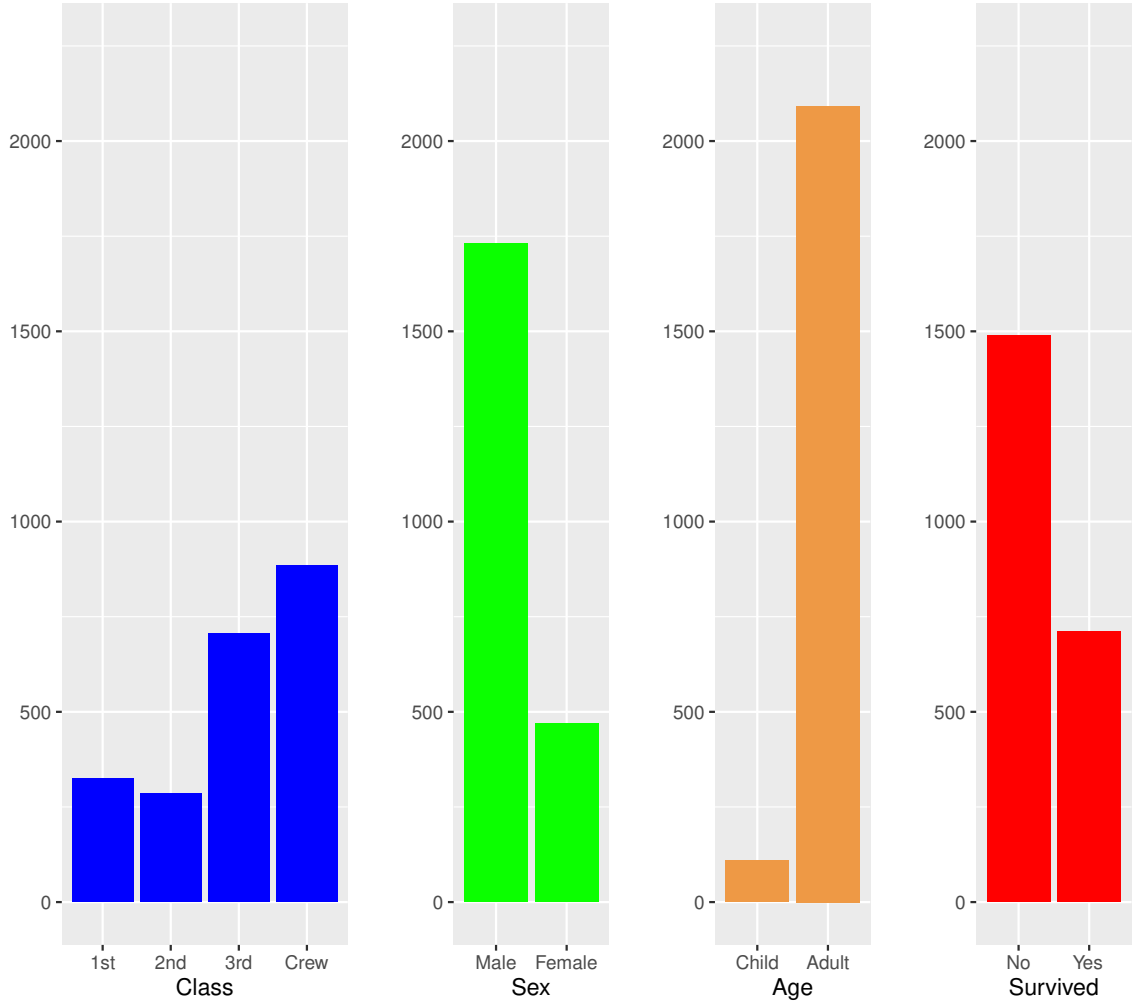
[a] Unwin, A. (2015) *Graphical Data Analysis with R*. Boca Raton: Chapman & Hall/CRC Press.

The Titanic Data

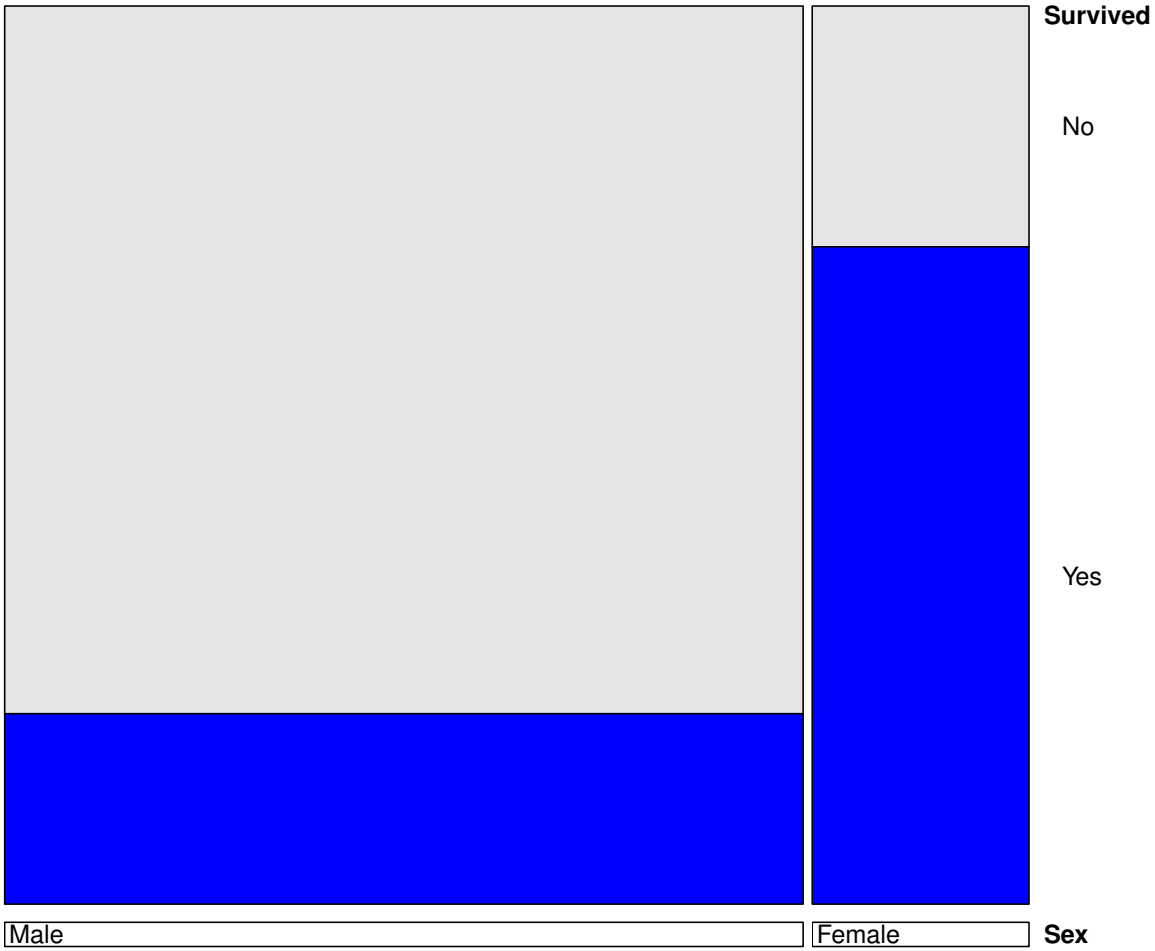
- The `Titanic` data contains four variables for 2,201 passengers.

- The variables are:
  - Class (1st, 2nd, 3rd, crew)
  - Sex (male, female)
  - Age (child, adult)
  - Survived (no, yes)

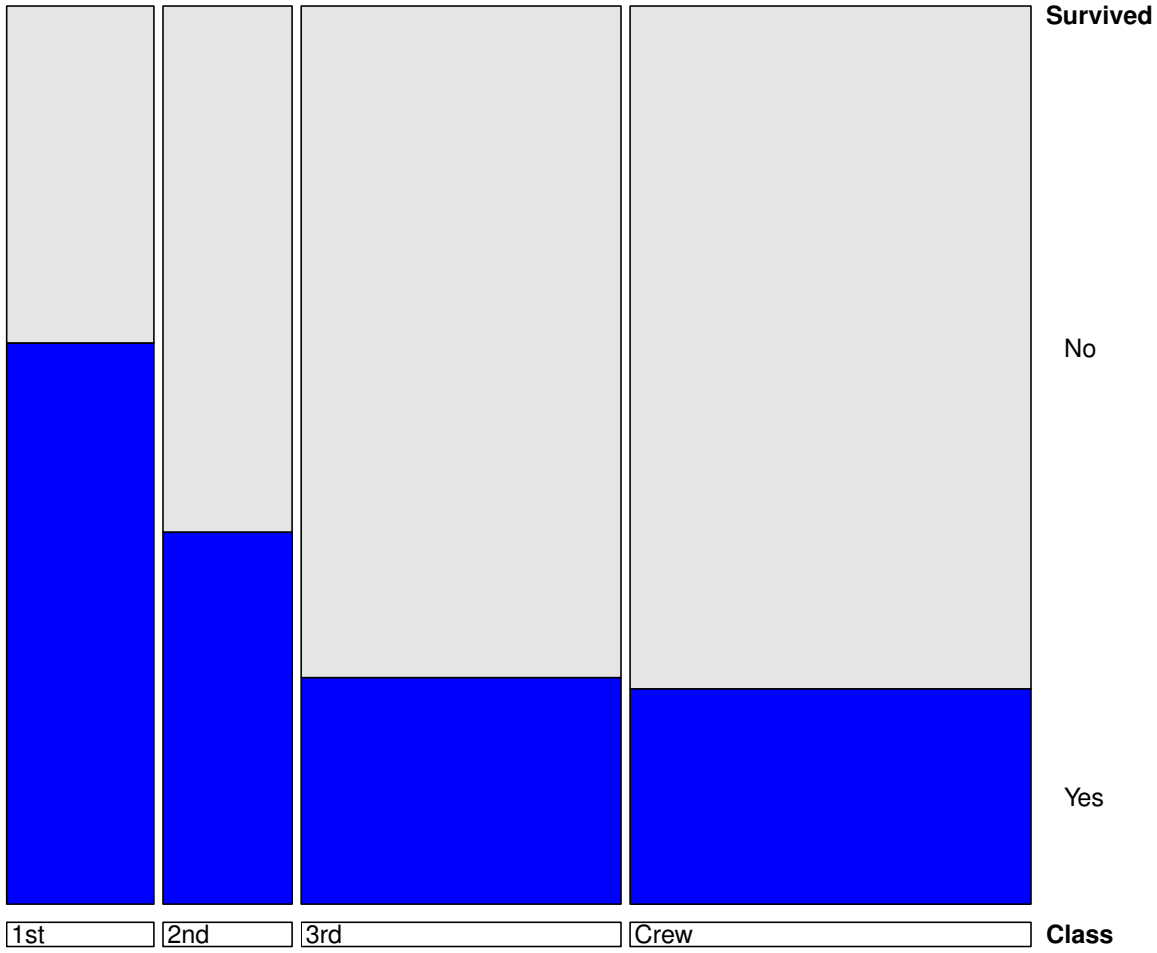
Simple Grid


2000
1500
1000
500
0
1st
2nd
3rd
Crew
Class
Male
Female
Sex
Child
Adult
Age
No
Yes
Survived

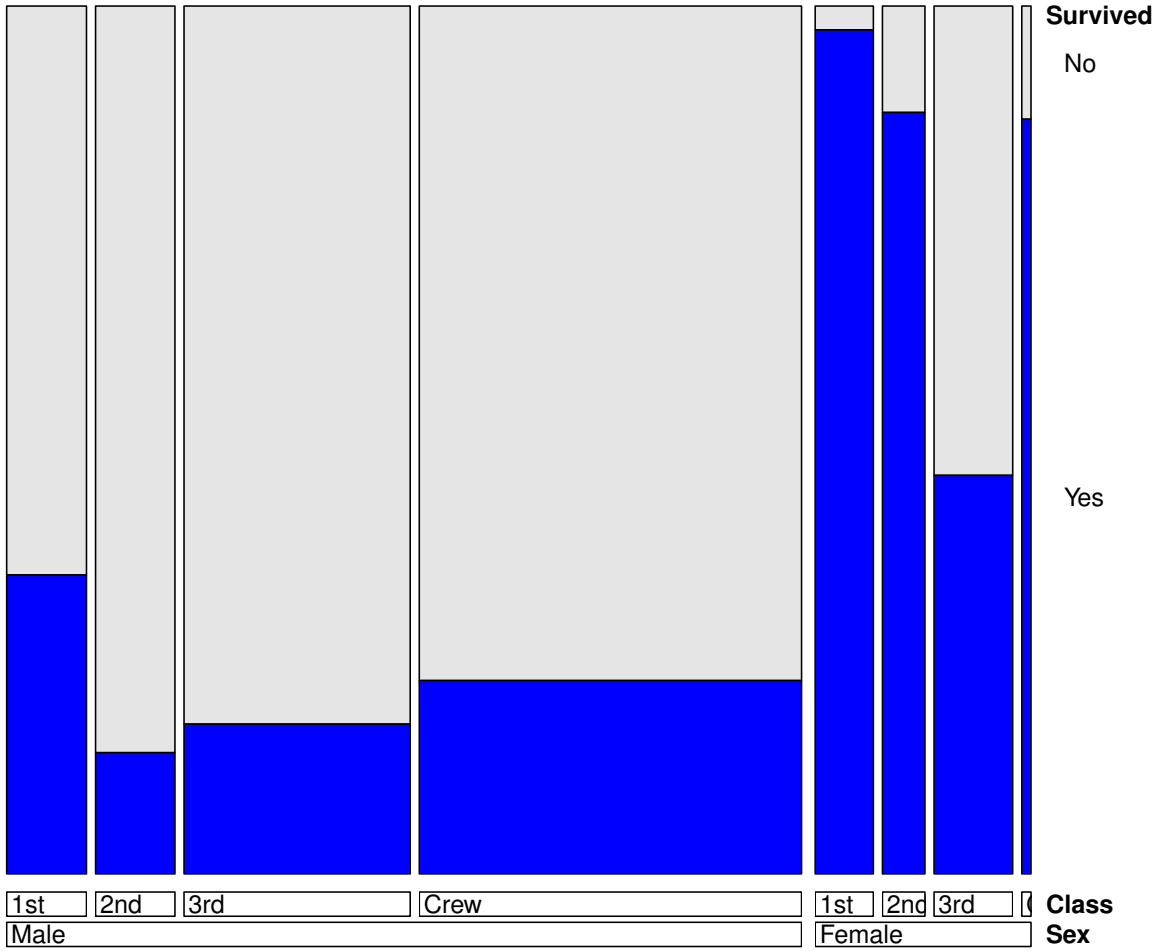
Doubledecker Plot 1



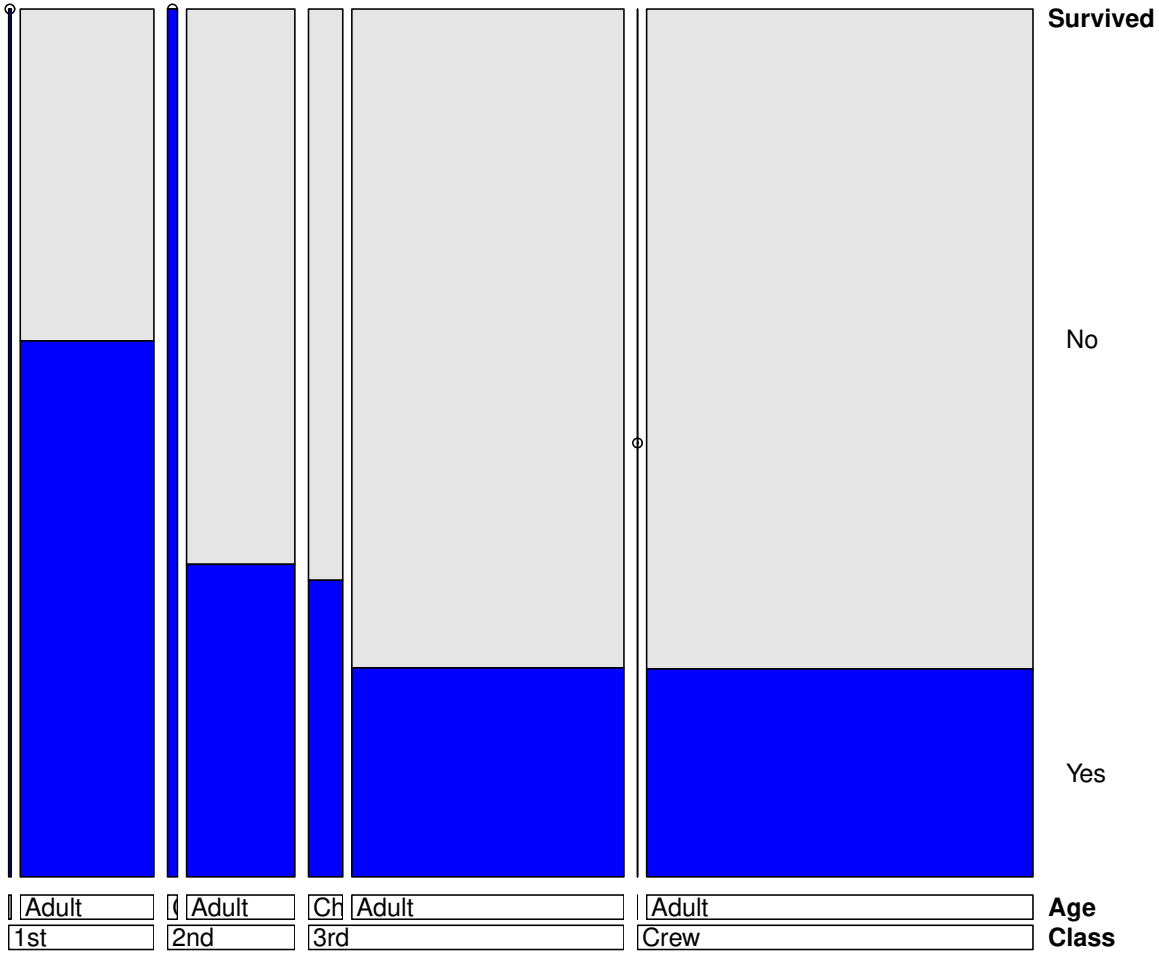
Doubledecker Plot 2



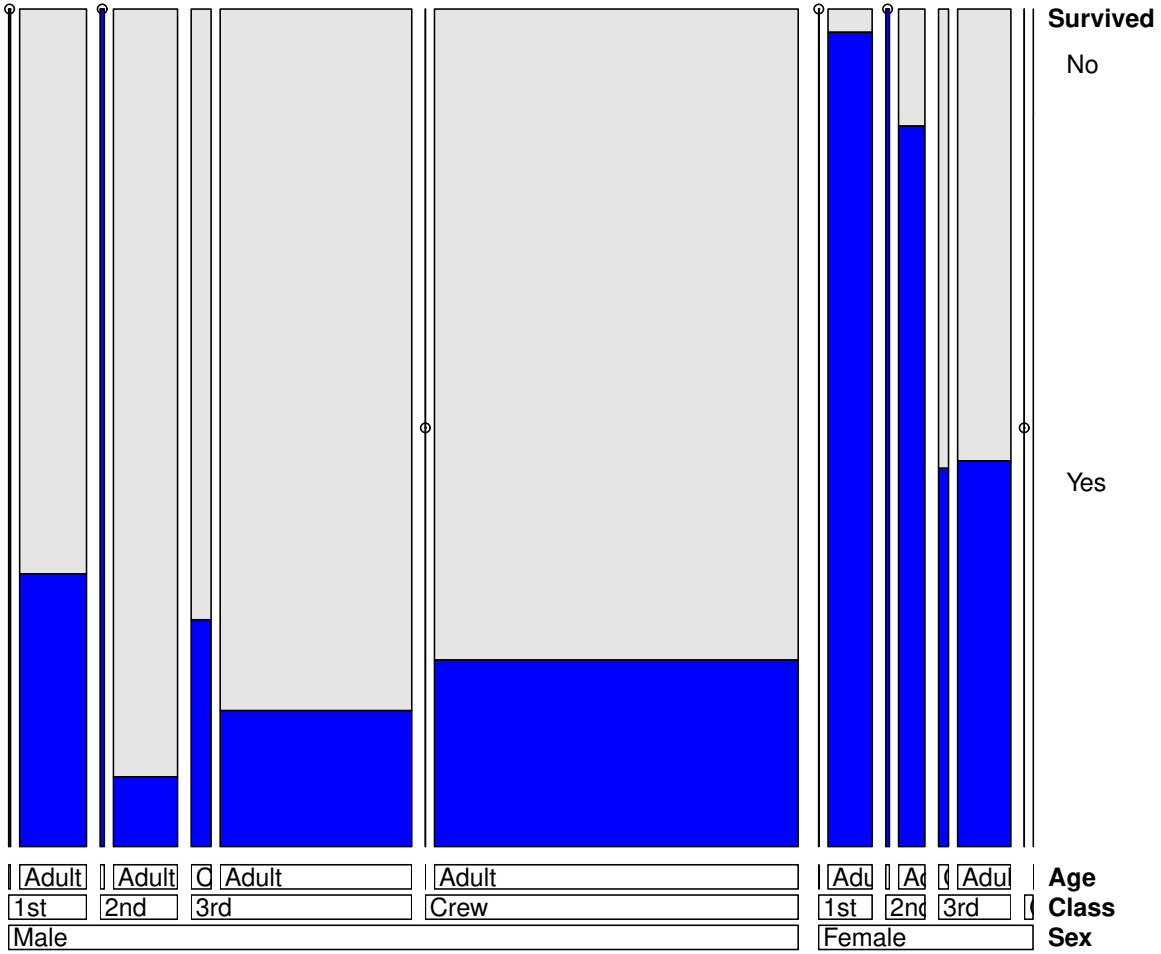
Doubledecker Plot 3



Doubledecker Plot 4



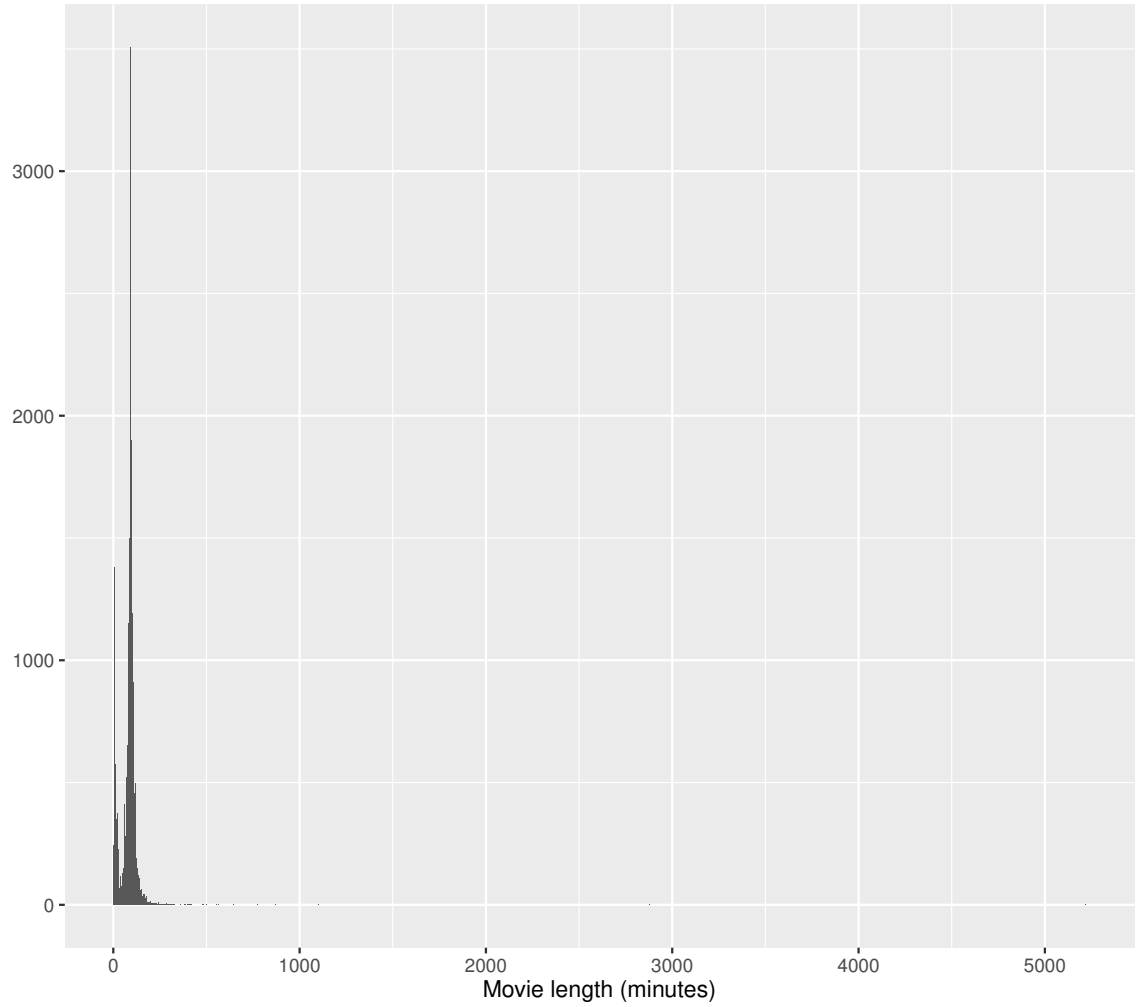
Doubledecker Plot 5



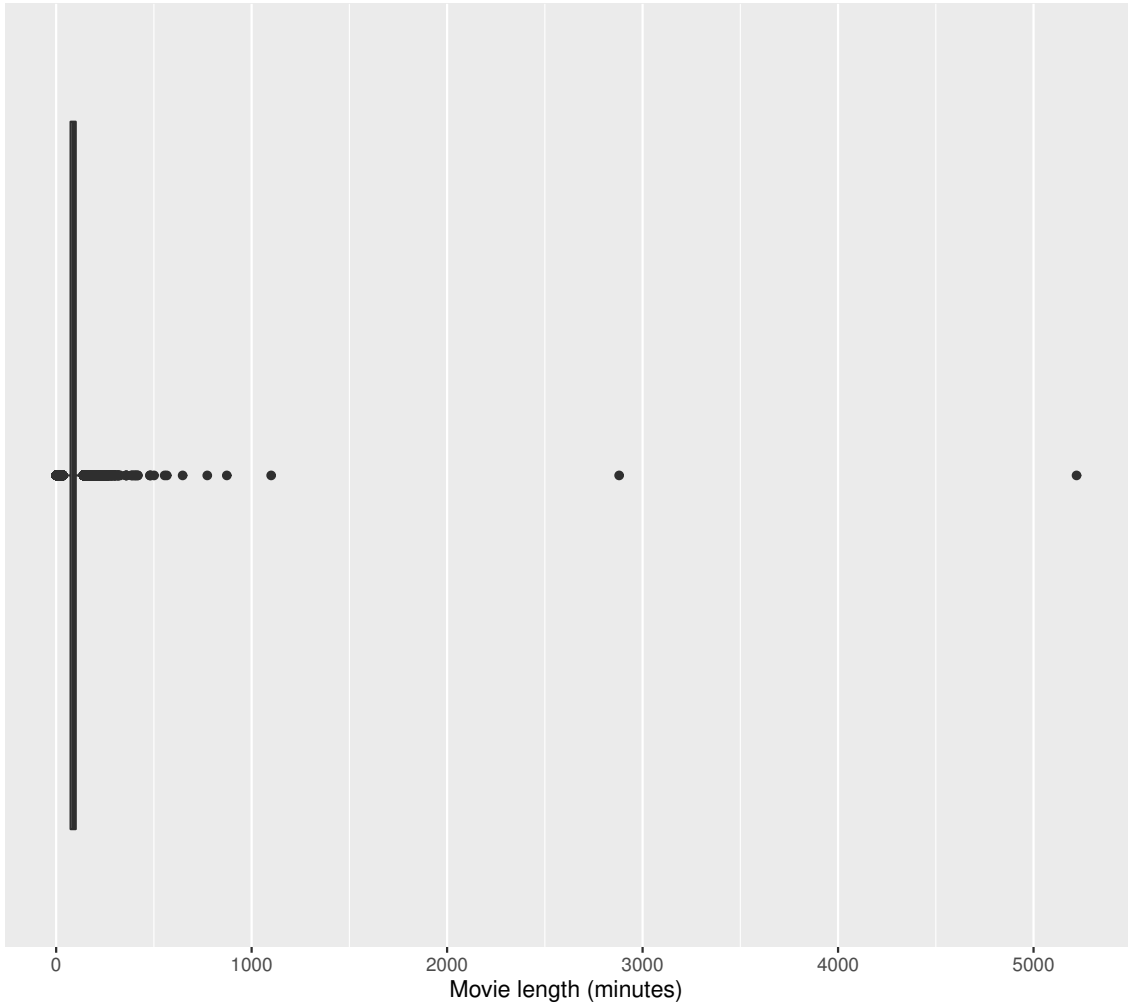
Movies Data

- When the Clint Eastwood movie Sully came out, it received some critical acclaim, including praise for its length.
- The movie comes in at around 96 minutes.
- How long is a movie anyway?
- The `movies` data contain 24 measurements on 28,819 movies.
- One of the measurements concerns length.

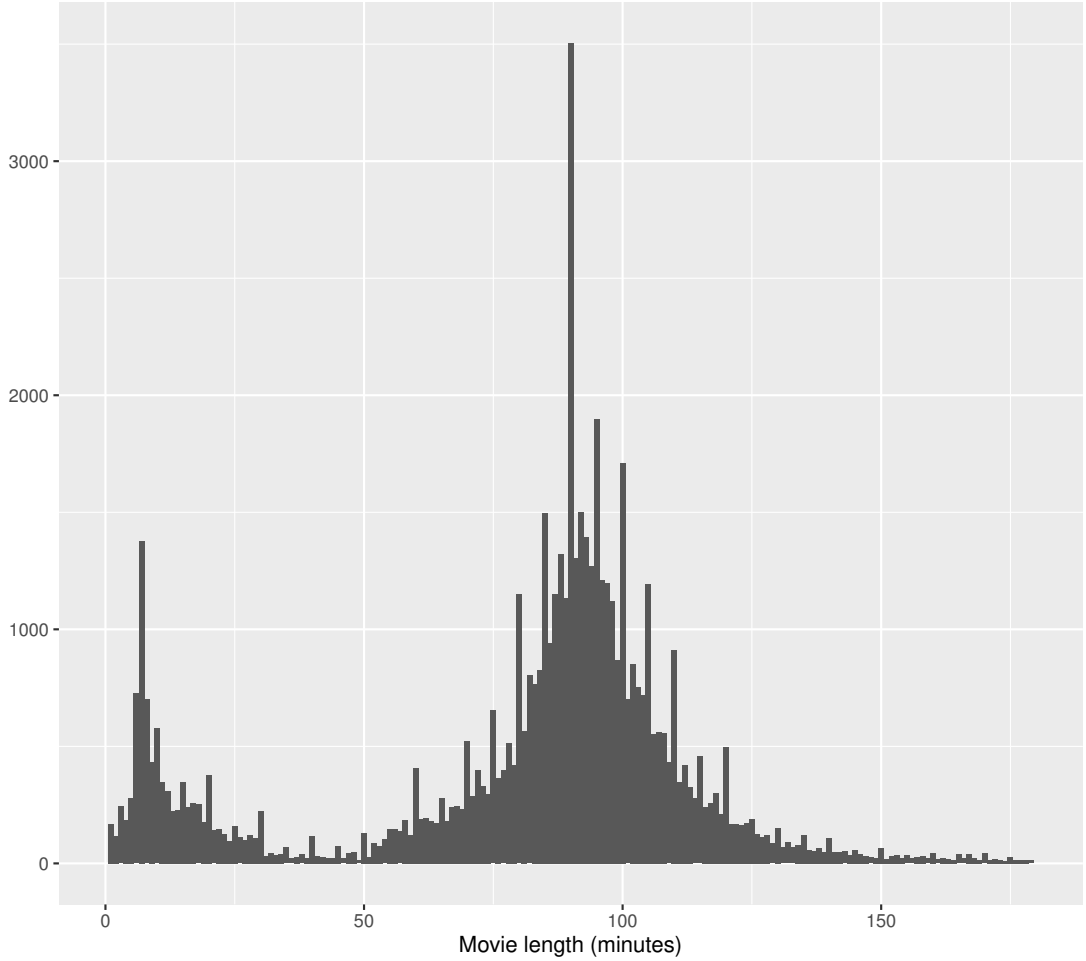
Simple Histogram



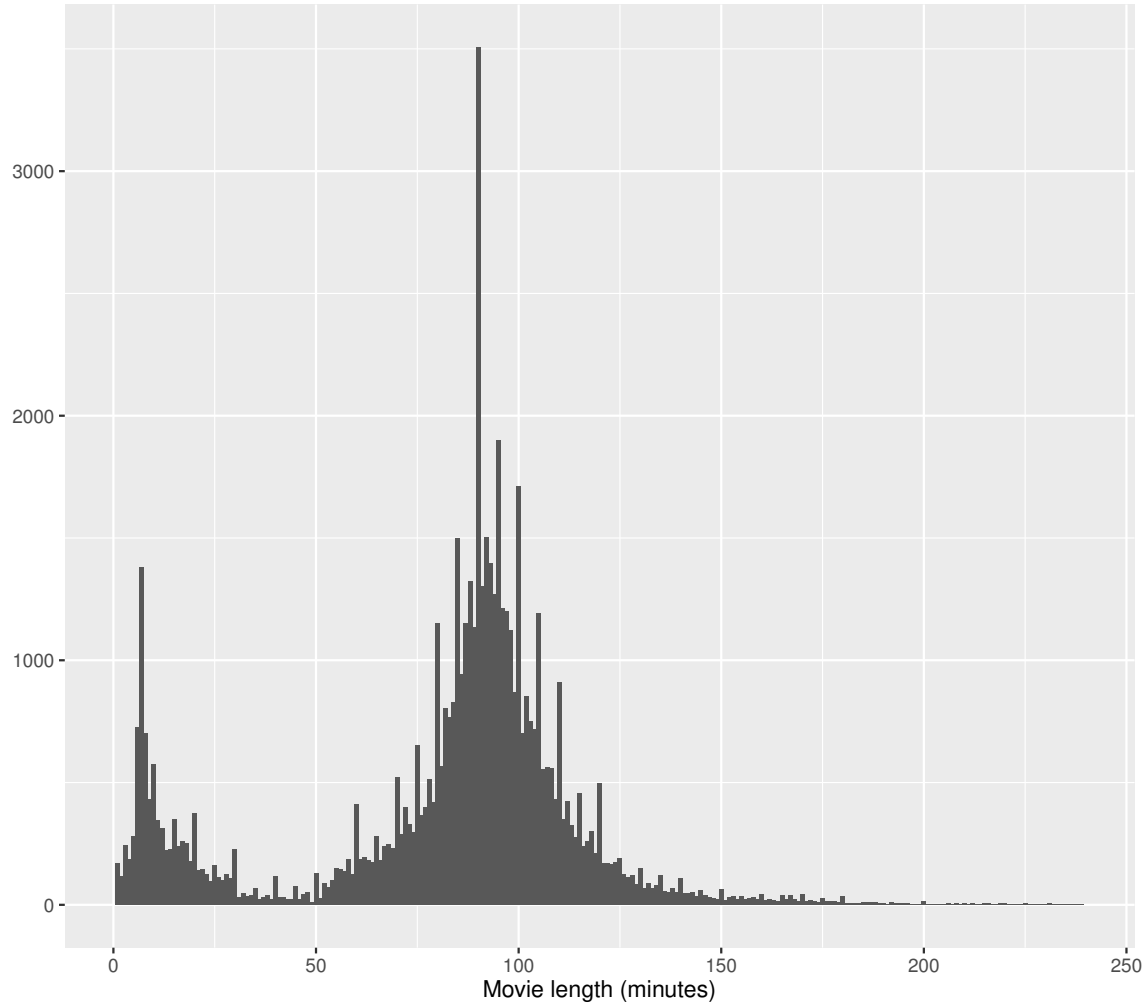
Alternative Plot



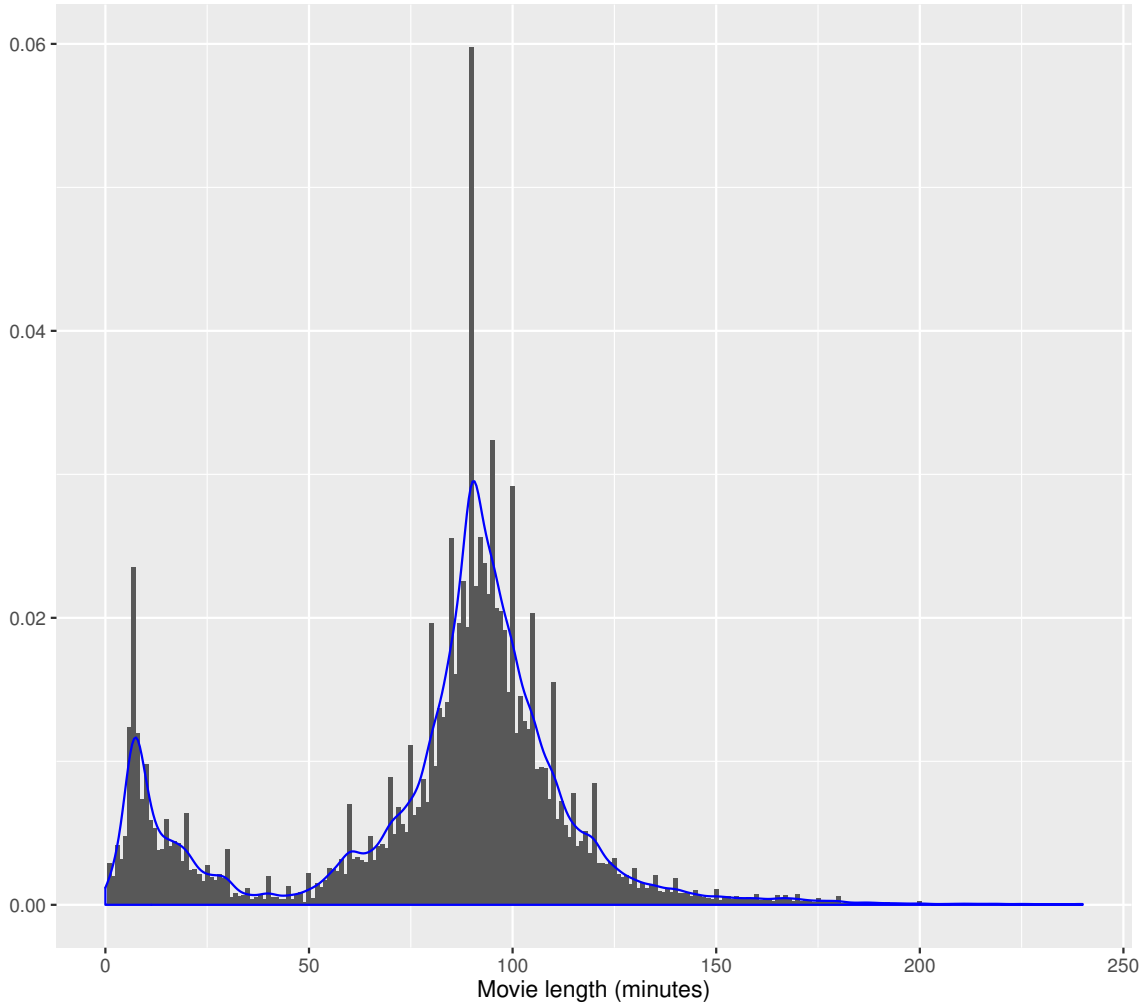
Histogram: Up to 3 Hours



Histogram: Up to 4 Hours



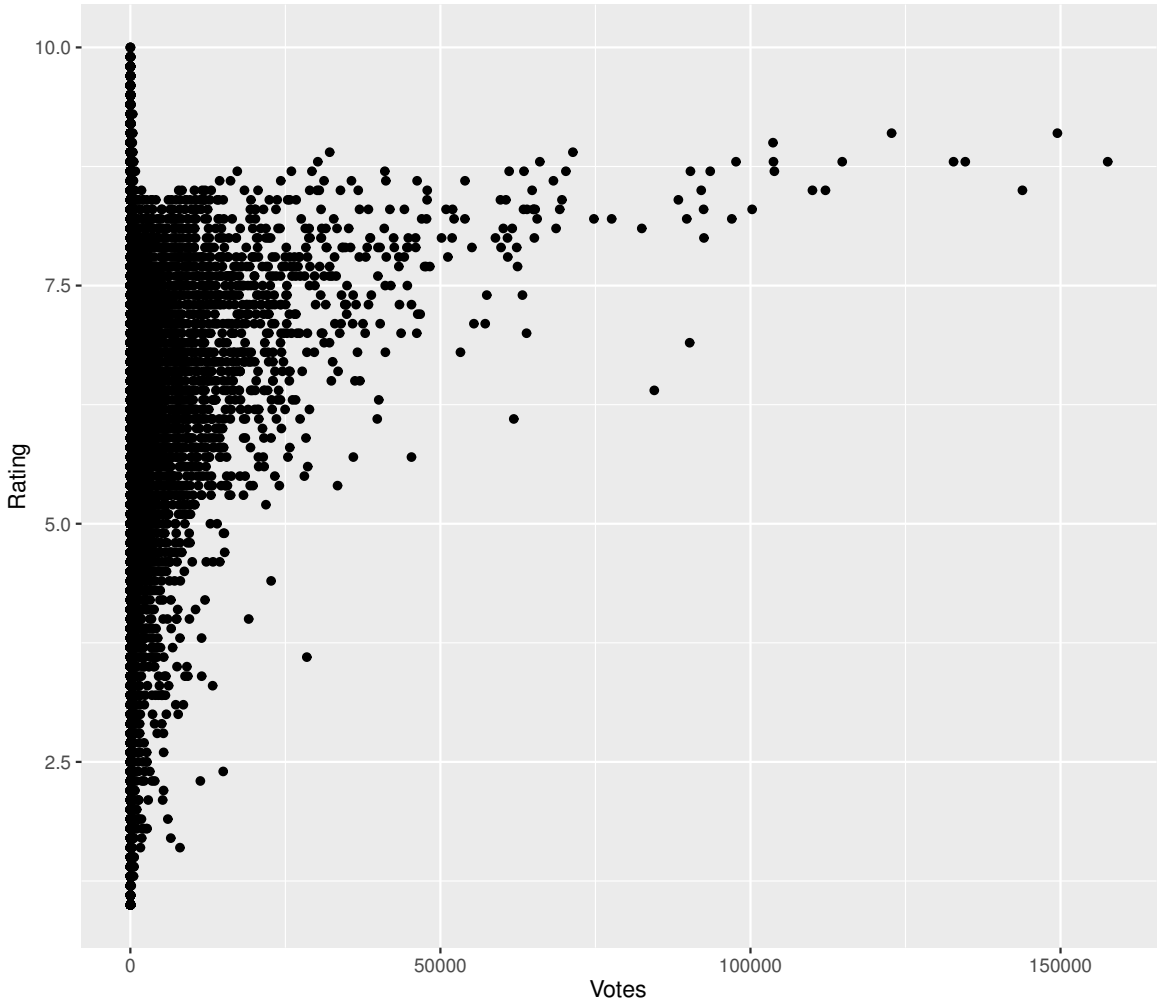
Histogram: With Density



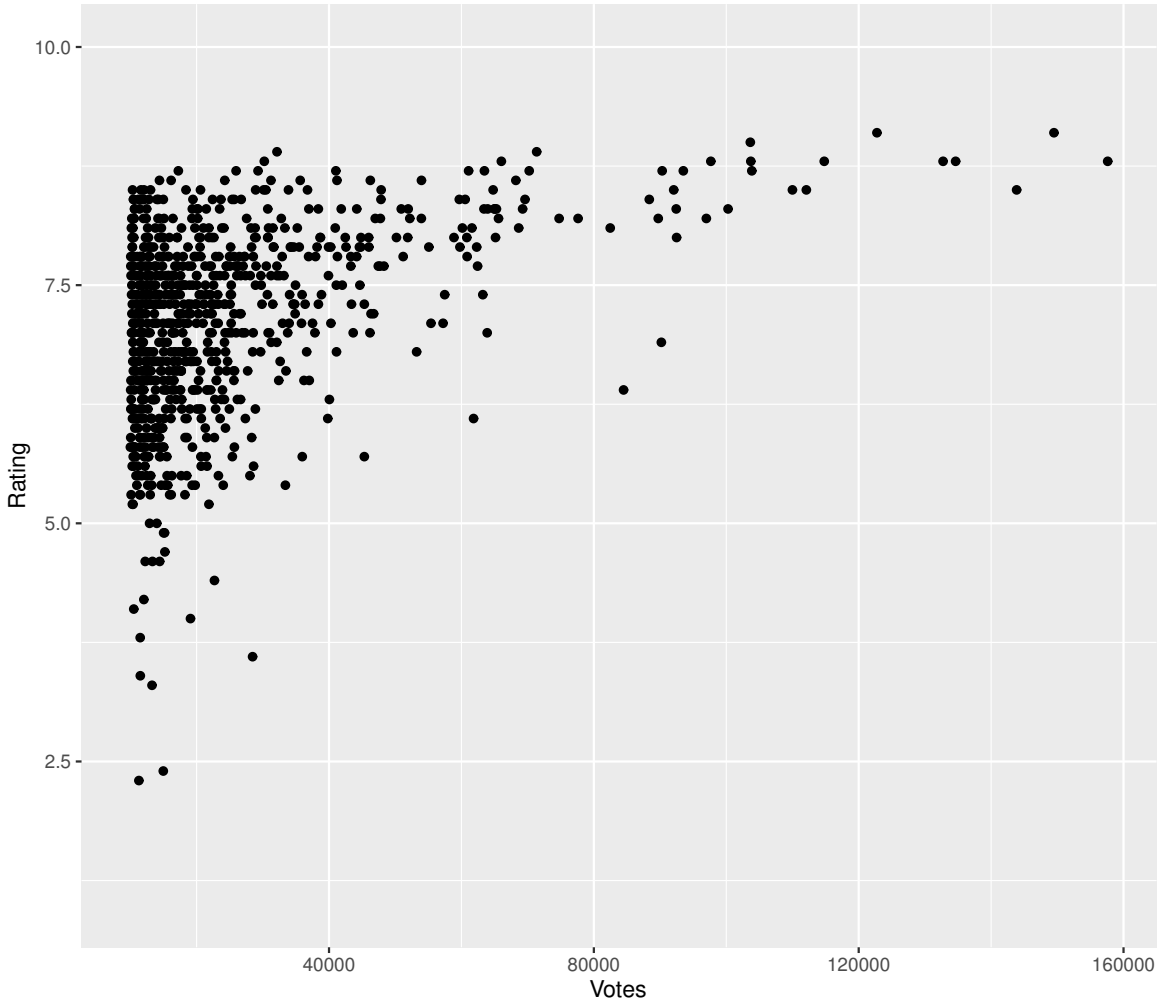
Movie Data Comments

- The first couple of plots gave a sense of outliers.
- The other plots illustrate a few points.
- For one, there is clear bimodality.
- There is also strong evidence to suggest rounding in places.
- Aside from length, there are other interesting aspects of the `movies` data...

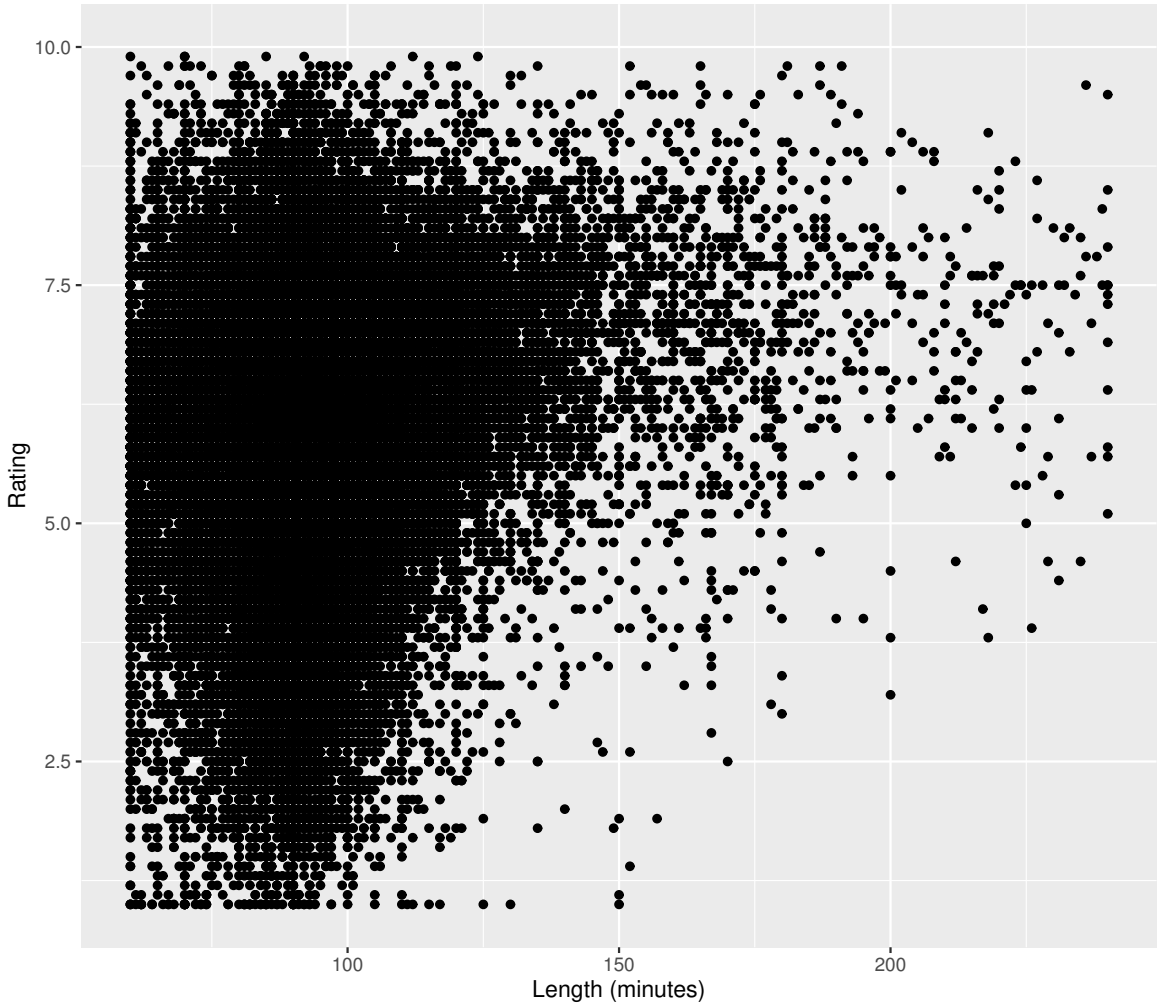
Votes vs. Rating



Votes vs. Rating (>10,000 Votes)



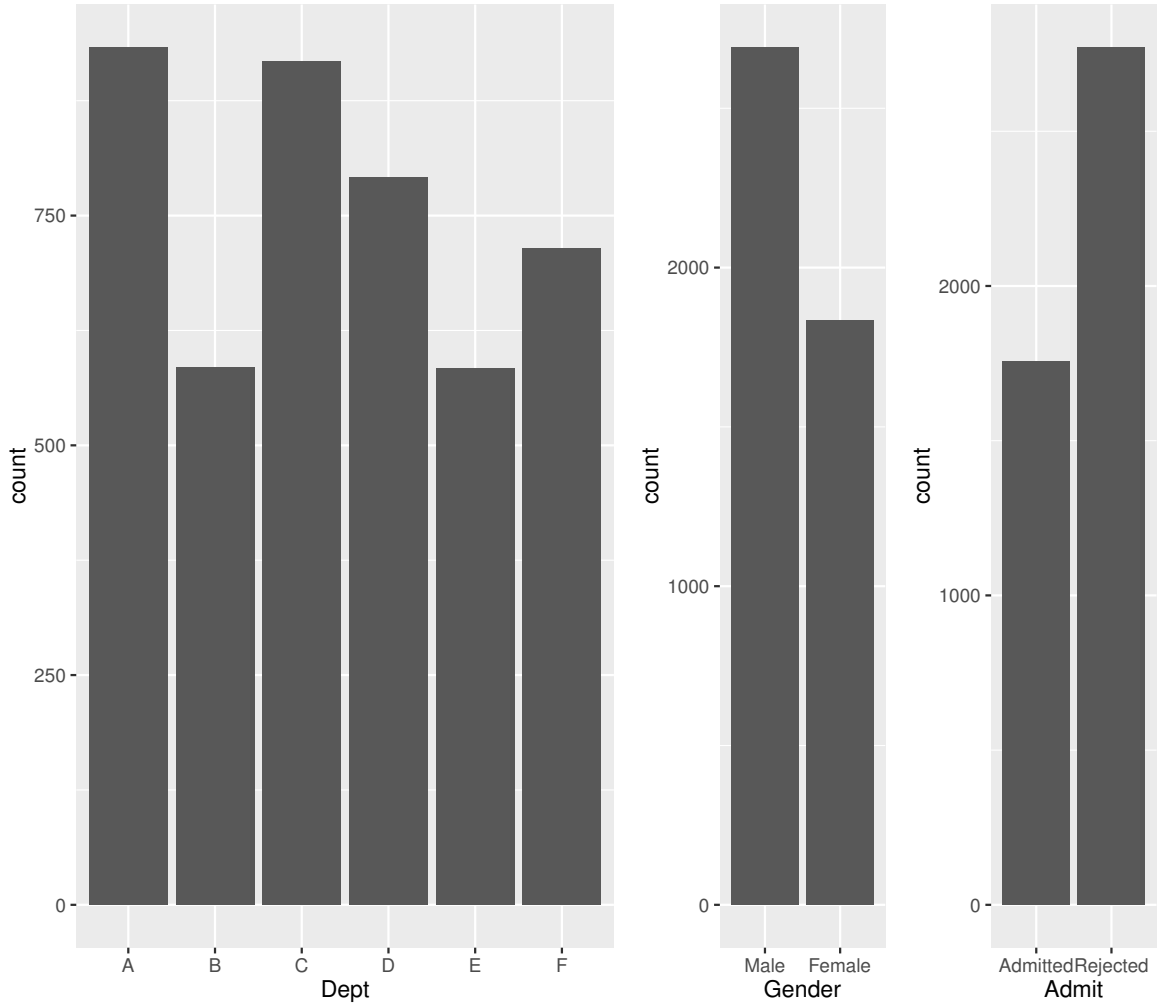
Length vs. Rating (1–4 Hours)



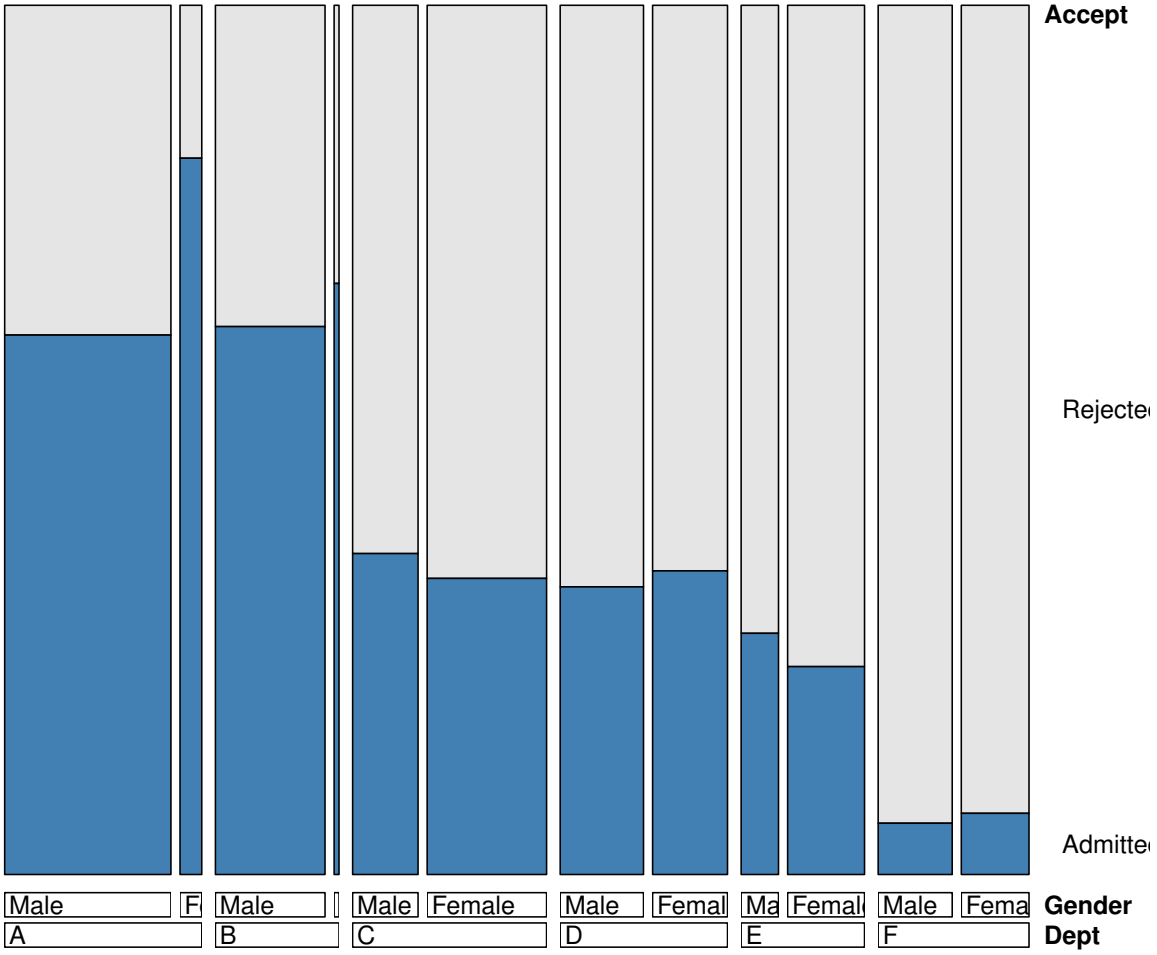
Berkeley Admissions Data

- Applicants to graduate school at Berkeley, for the six largest departments, in 1973.
- Four variables: admission decision, gender (sex), department, and count.
- Famous for illustrating Simpson's paradox:
  - 1,198 of 2,691 male applicants (44.5%) were admitted, and
  - 557 of 1,835 female applicants (30.4%) were admitted.
- Is this *ipso facto* evidence of gender bias?

Berkeley Data: Bar Charts



Berkeley Data: Doubledecker Plot



Berkeley Data Comments

- Looking at data across all departments, Bickel et al. (1975)[a] conclude that "if the data are properly pooled. . . there is a small but statistically significant bias in favour of women."

- Bickel et al. (1975) did not have access to software like R but they nevertheless give an interesting graphical representation of the data.

- Now we will move from one famous data set to another.

[a]Bickel, P.J., E.A. Hammel and J.W. O'Connell (1975). 'Sex bias in graduate admissions: Data from Berkeley'. *Science* **187**(4175) 398–404.

Geyser Data

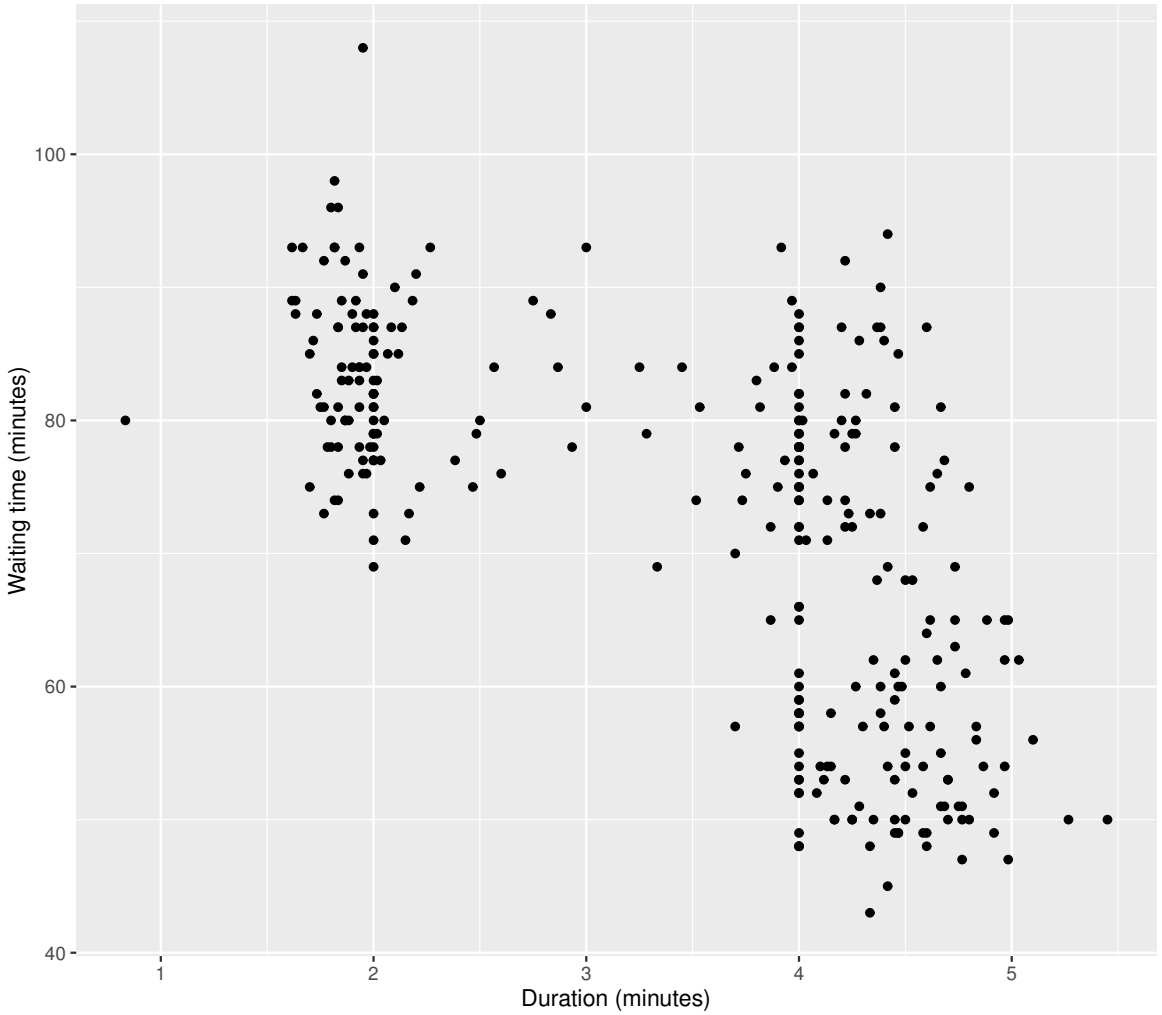
- The `geyser` data set contains a total of 299 observations of eruption duration (in minutes) and waiting time (in minutes, for this eruption) for the Old Faithful geyser.

- Available as `geyser` for the `MASS` package in R.

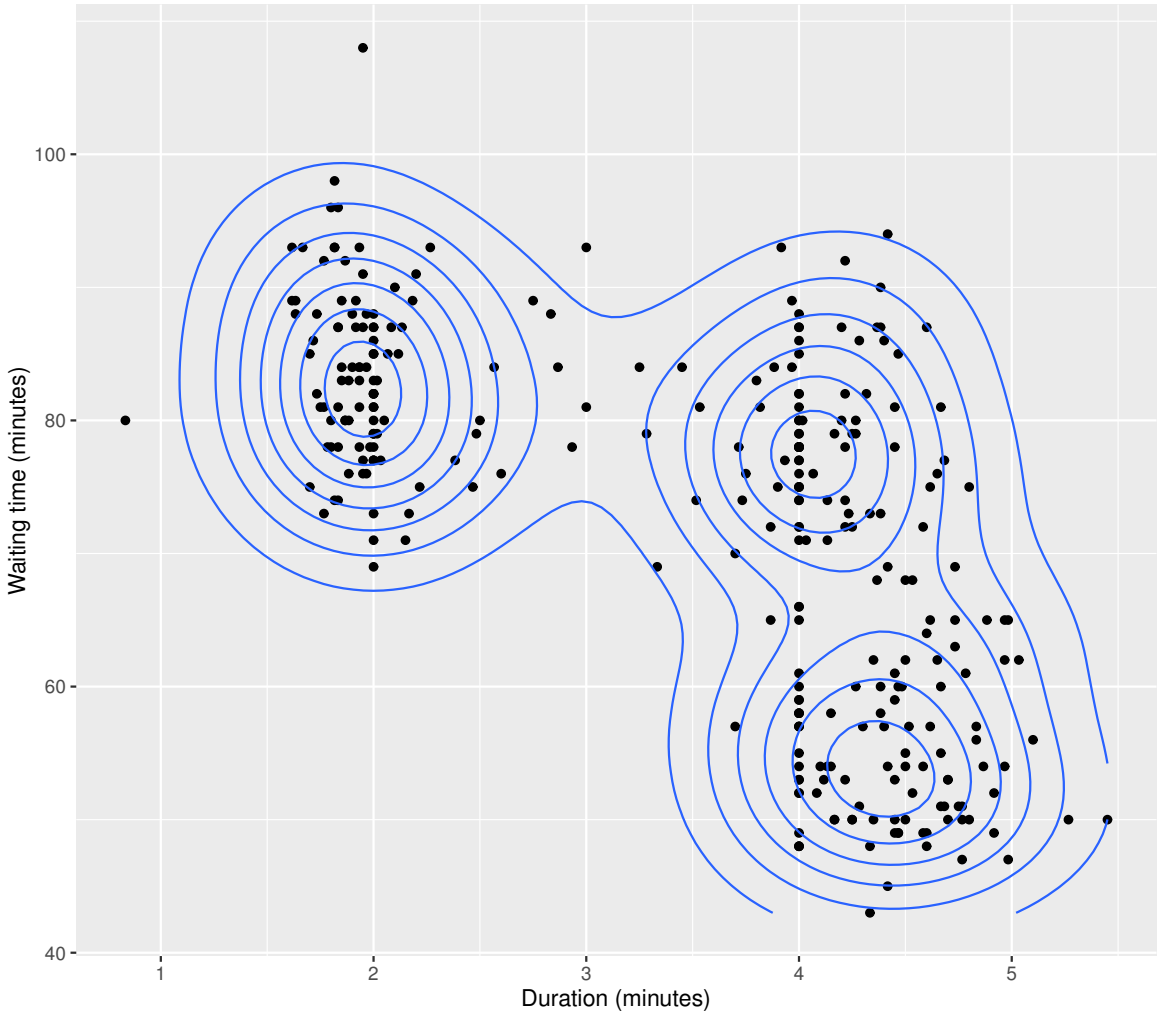
- Studied in detail by Azzalini and Bowman (1990)[a].

[a] Azzalini, A. and A.W. Bowman (1990). 'A look at some data on the Old Faithful geyser'. *Applied Statistics* **39**, 357–365.

Geyser Data: Scatter Plot



Geyser Data: With Contours



Geyser Data Comments

- A famous data set.
- Very clear evidence to suggest rounding at 2 and 4 minutes; perhaps also at 3 minutes.
- Some outliers are also present.
- We will look at a more (graphically) difficult scatter plot with contours in a minute.
- But first, we will play around a bit in R.

Bivariate Gaussian Distrobutions

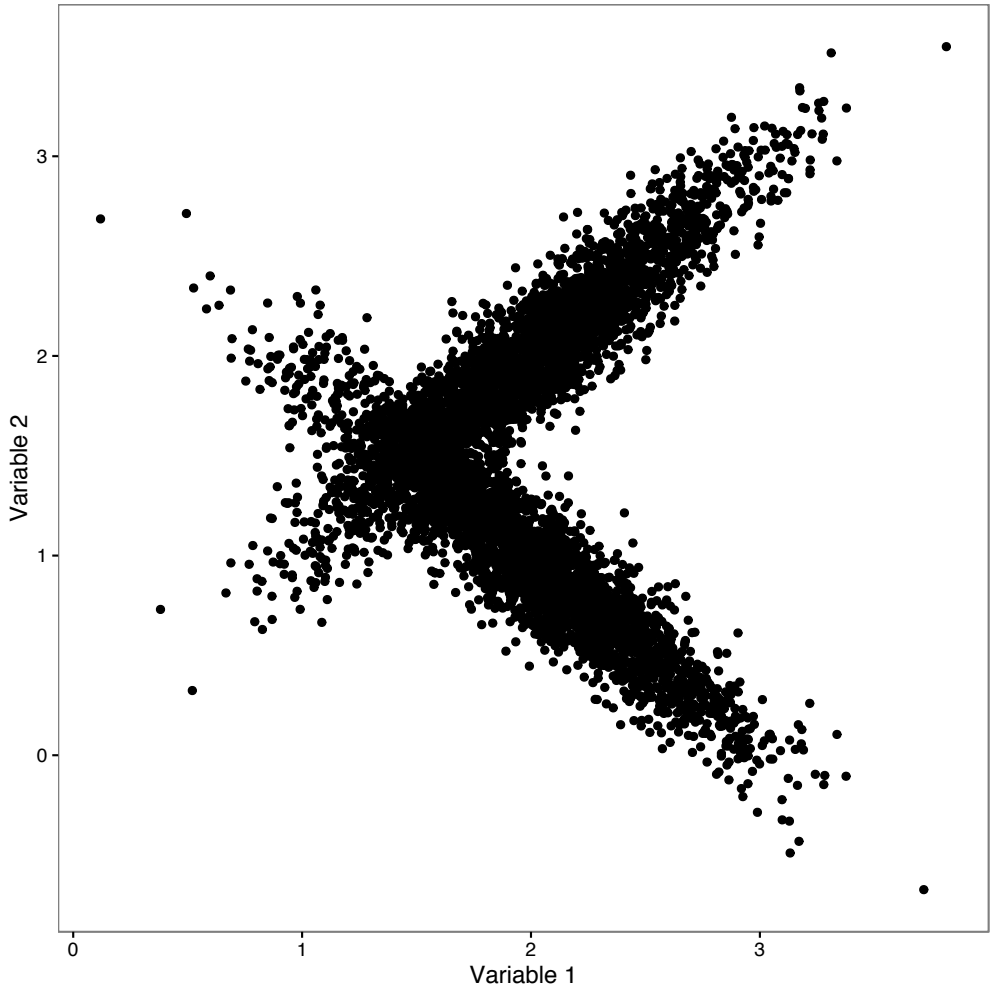
- McNicholas (2016) uses the example of two overlapping bivariate Gaussian densities to illustrate why the number of clusters should not be taken equal to the number of modes.

- These data contain 6,000 observations, each simulated from one of two bivariate Gaussian distributions.

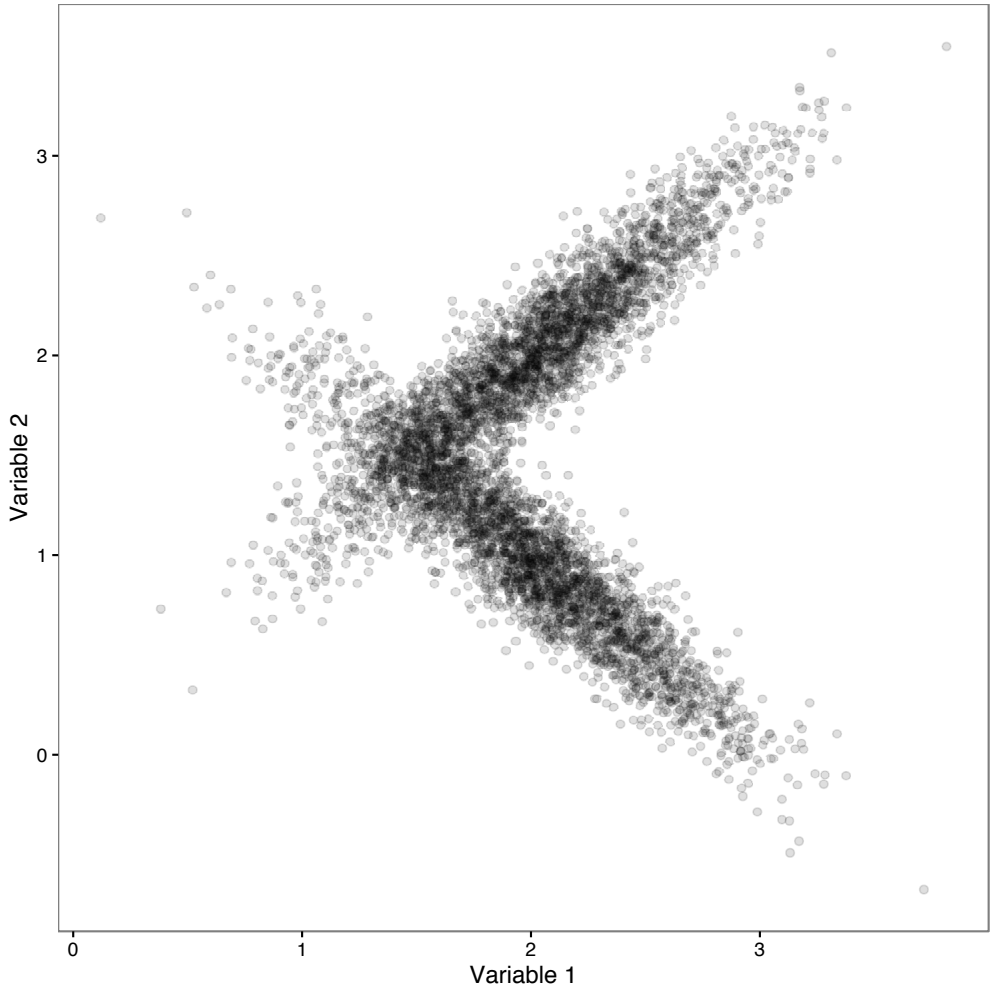
- This presents a data visualization challenge — a standard scatter plot looks like a blob.

- McNicholas (2016) uses semi-transparent points to overcome this challenge.

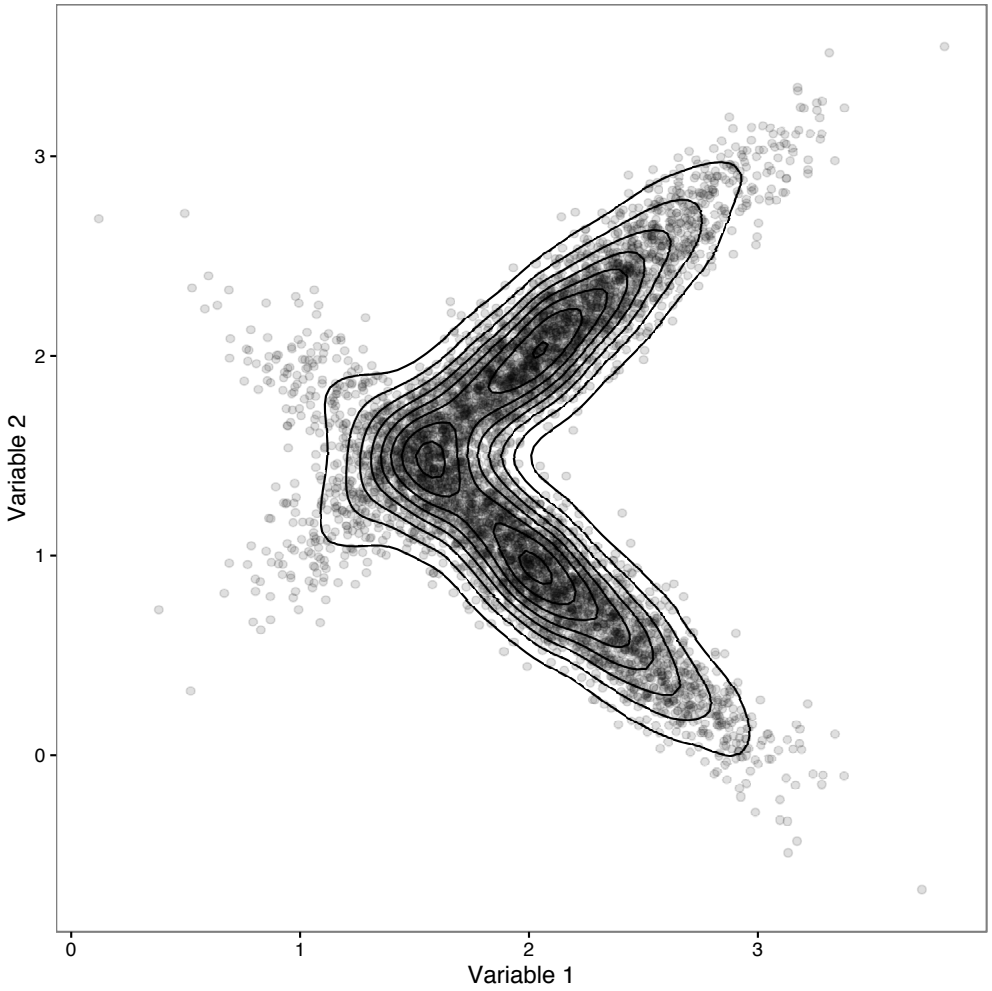
Scatter Plot



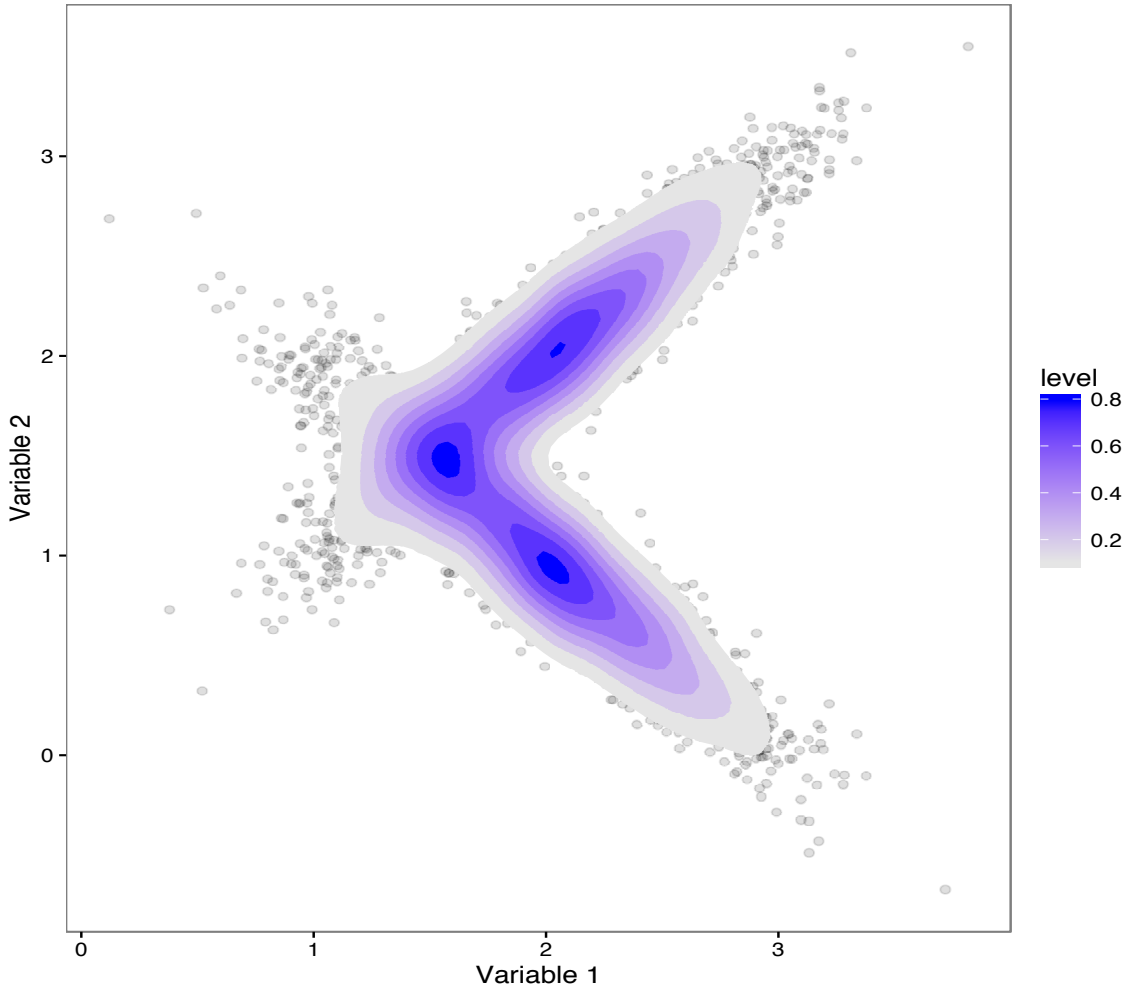
With Semi-Transparent Points



With Countours



With Shaded Contours



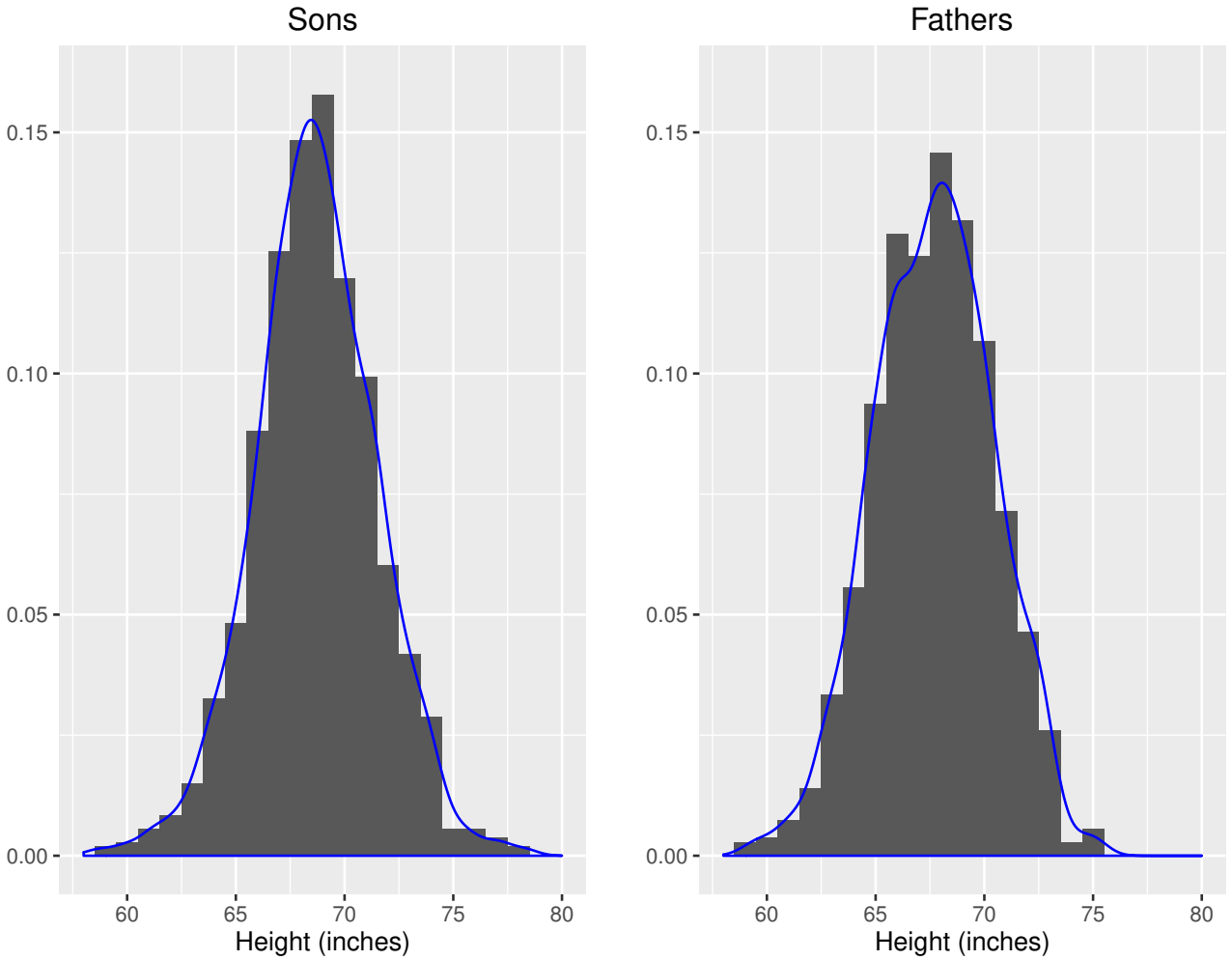
Comments

- Using semi-transparent points can be very convenient.
- In this example, we were able to see areas of high-density on the scatter plot (rather than a blob).
- Let's look at a very famous data set...

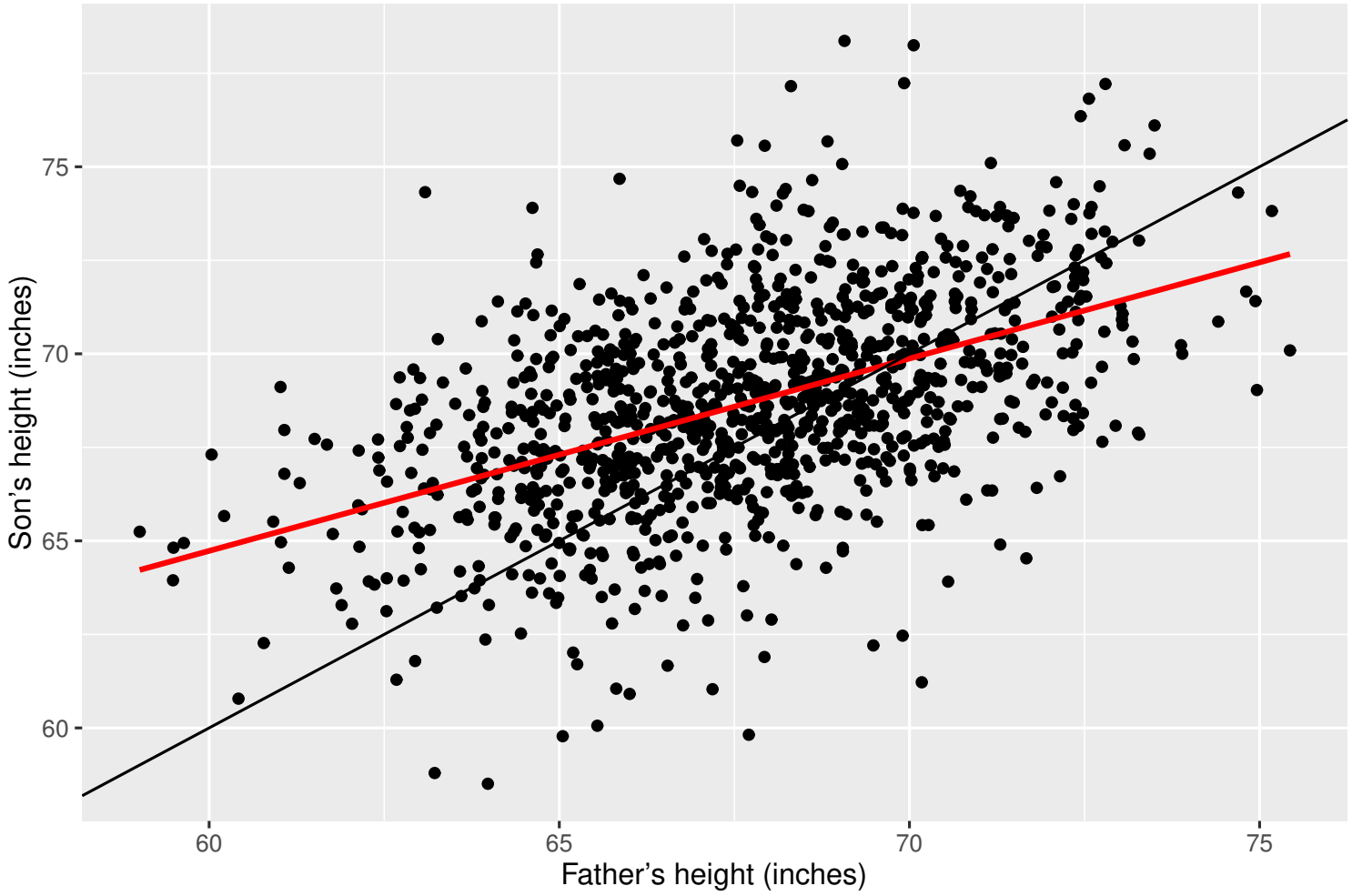
Fathers & Sons

- Not the Cat Stephens song!
- Height for 1,078 fathers and sons (in inches).
- Very famous example used by Pearson.
- One of the fundamental examples of regression — perhaps the fundamental example.

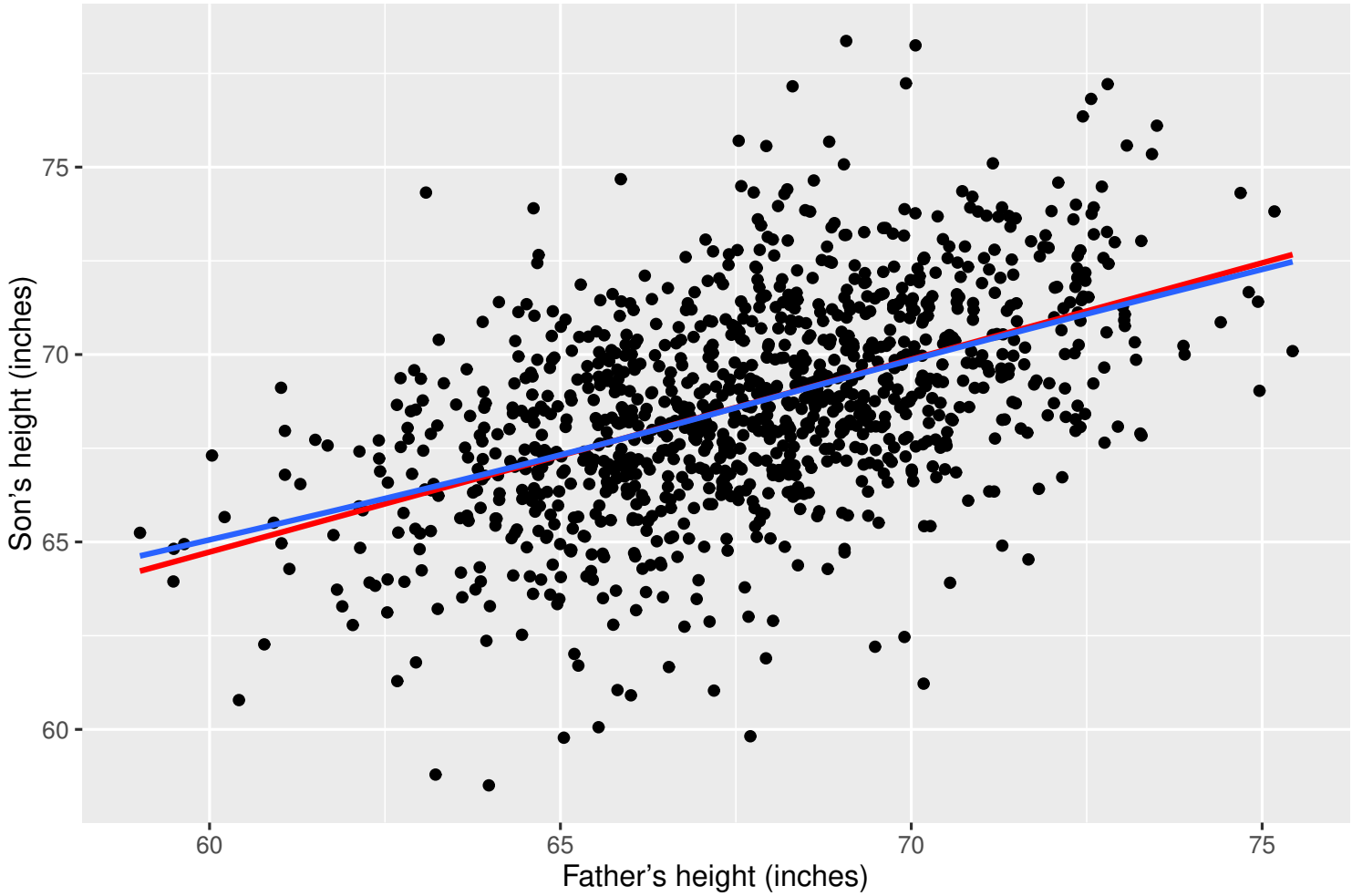
Fathers & Sons: Histograms



Fathers & Sons: Scatter + Line



Fathers & Sons: Line + Smooth Line

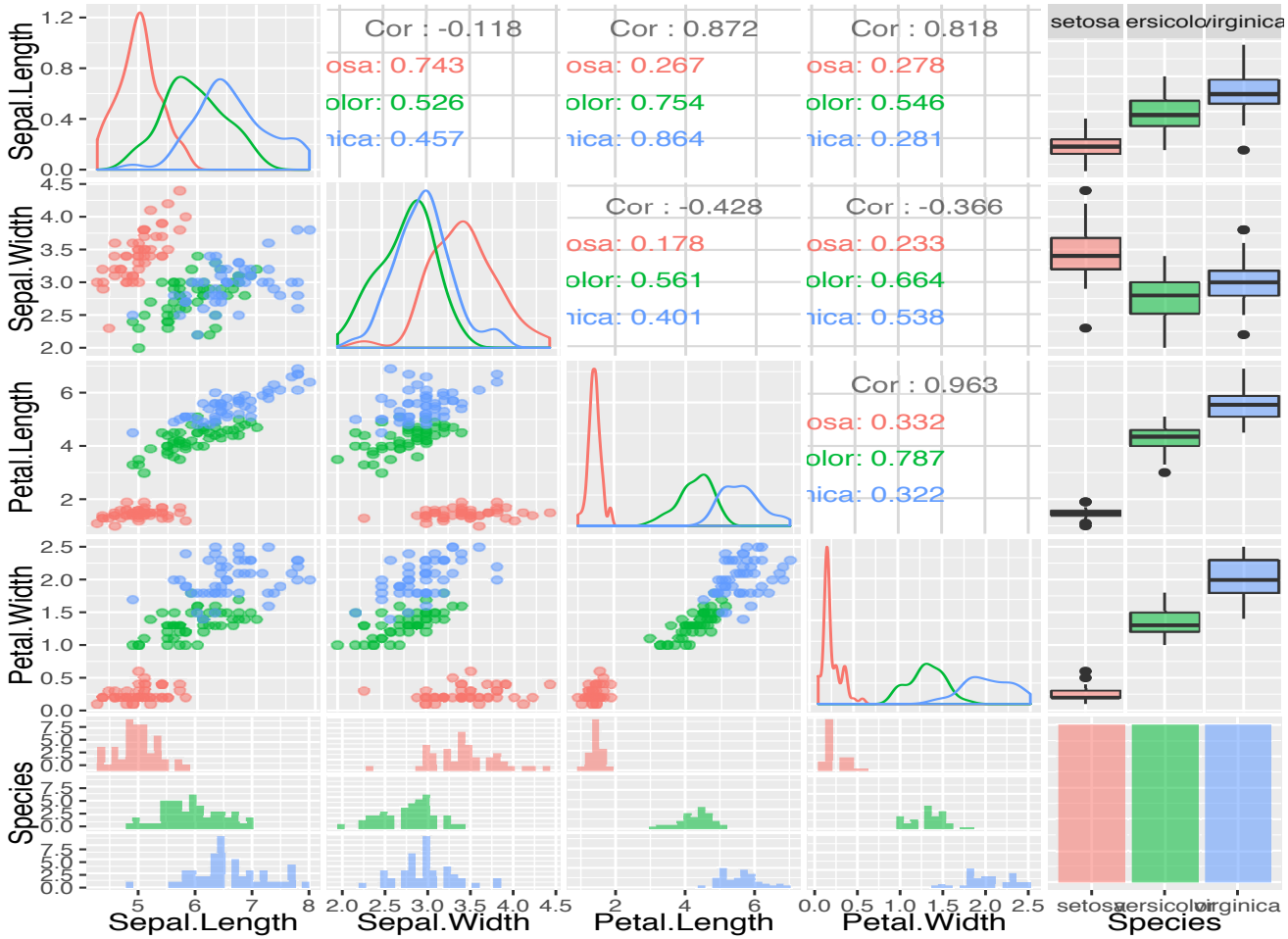


Iris Data

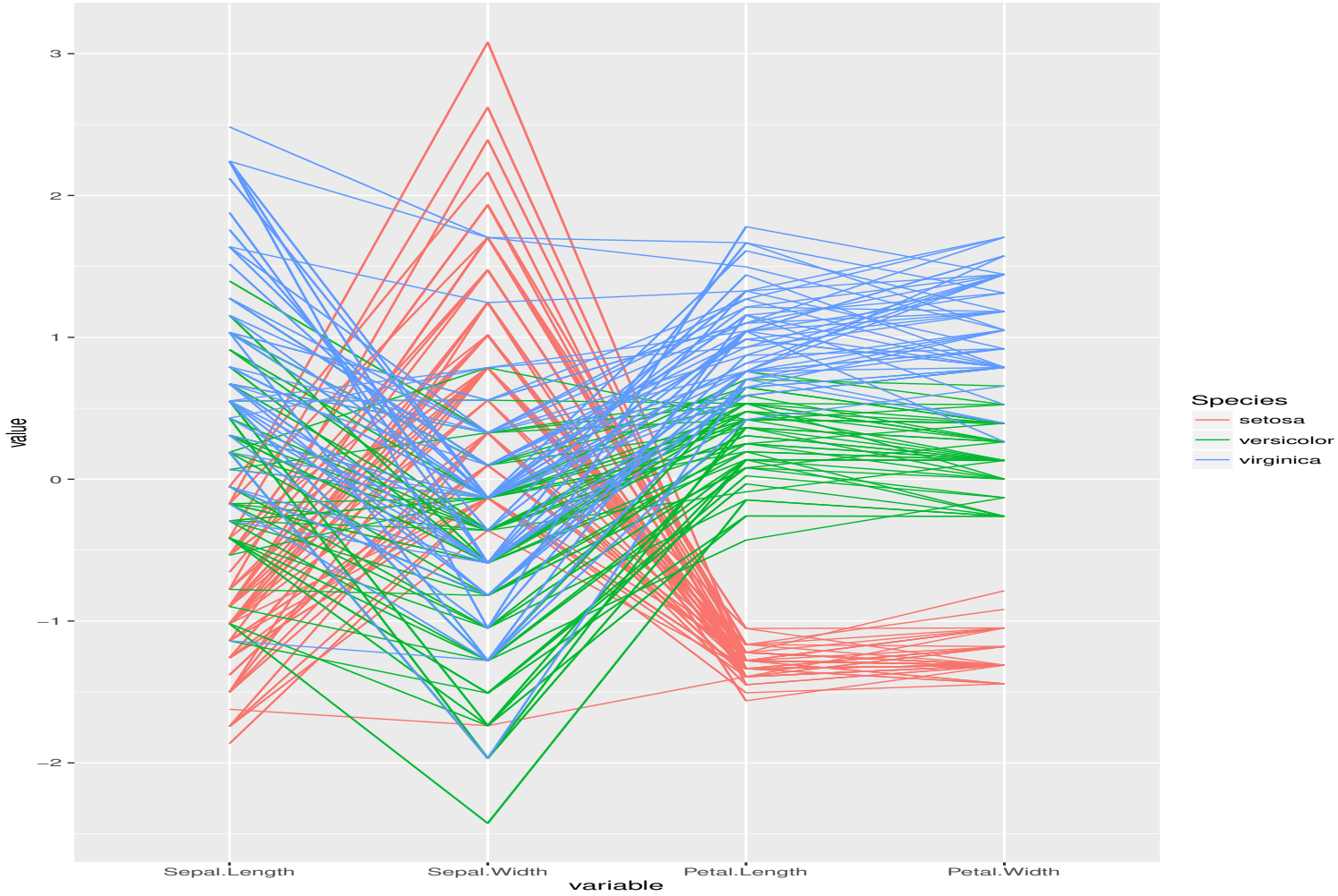
- Very famous data collected by Anderson (1935)[a]
- Measurements (in cm) of four variables — sepal length and width as well as petal length and width — for 50 flowers from each of 3 species of iris.
- The species are Iris setosa, versicolor, and virginica.
- There are 50 flowers of each species.

[a] Anderson, Edgar (1935). The irises of the Gaspe Peninsula, Bulletin of the American Iris Society **59**, 2–5.

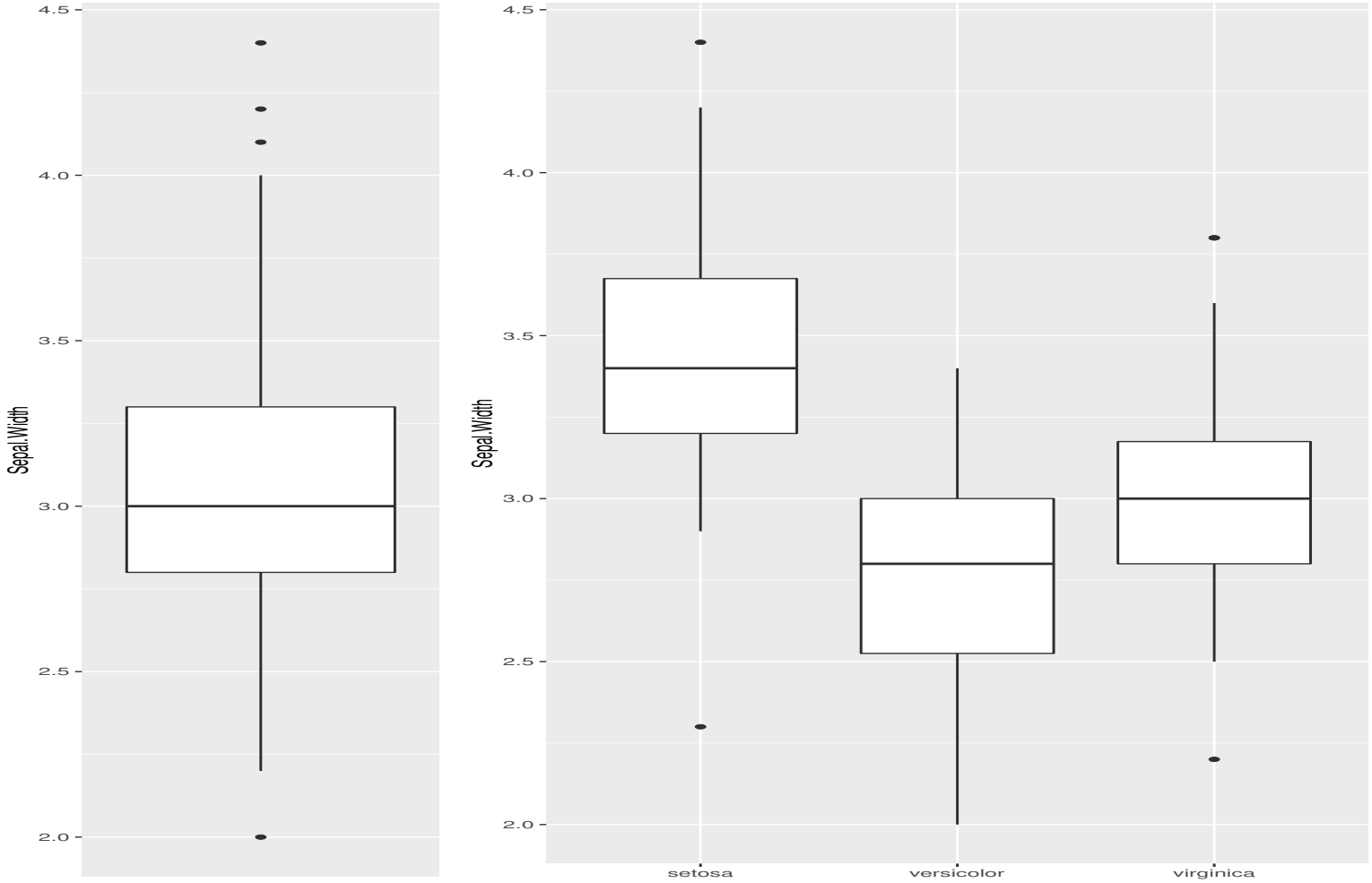
Iris: Pairs+


Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Species
Cor : -0.118
Cor : 0.872
Cor : 0.818
Cor : -0.428
Cor : -0.366
Cor : 0.963
setosa
versicolor
virginica

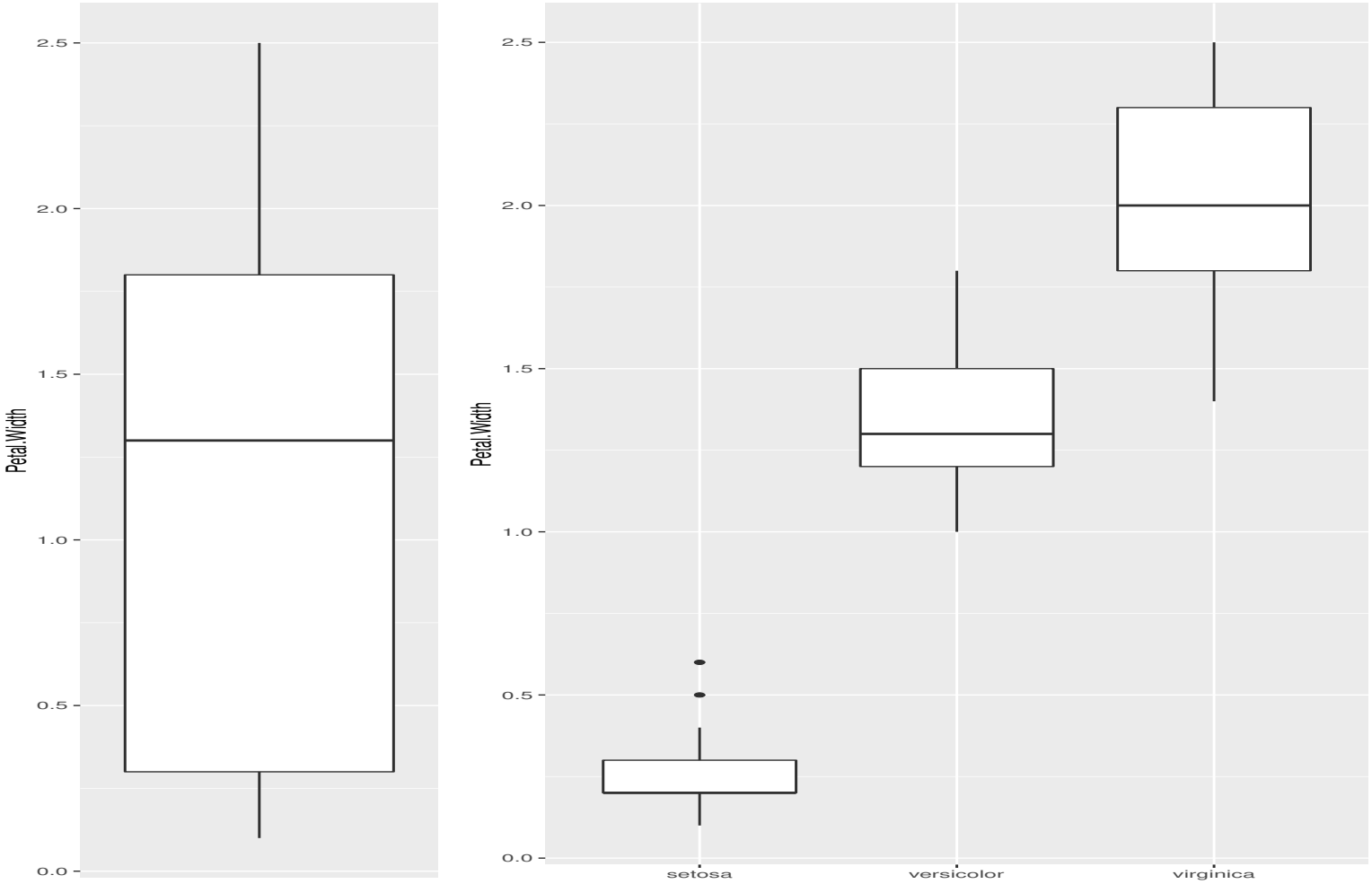
Iris: Parallel Coordinates



Iris: Some Box Plots



Iris: Some More Box Plots



Body Data

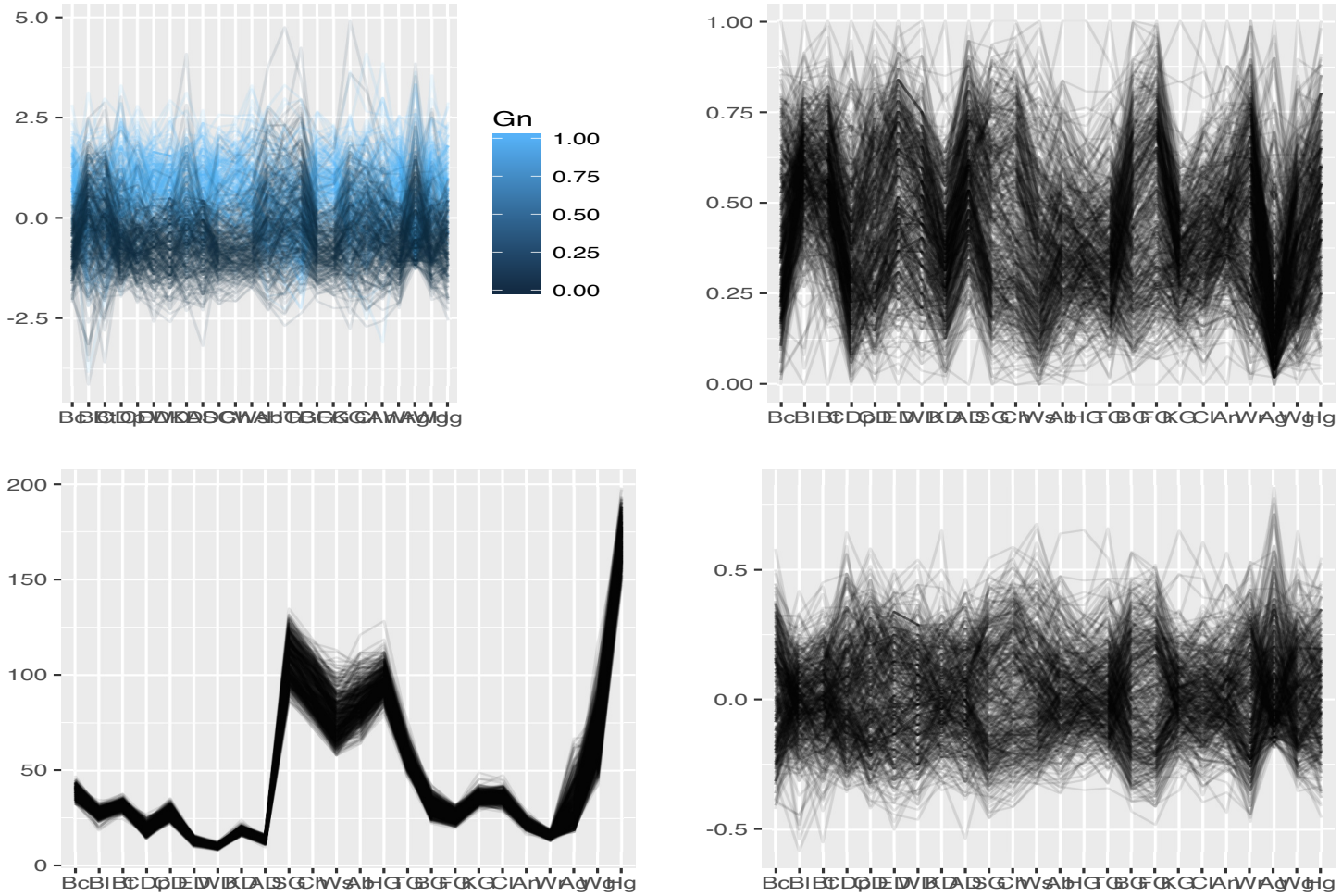
- The `body` dataset from `gclus` contains 21 body dimension measurements as well as age (in years), weight (in kg), height (in cm), and gender (binary) on 507 individuals.

- These break down as 247 men and 260 women.

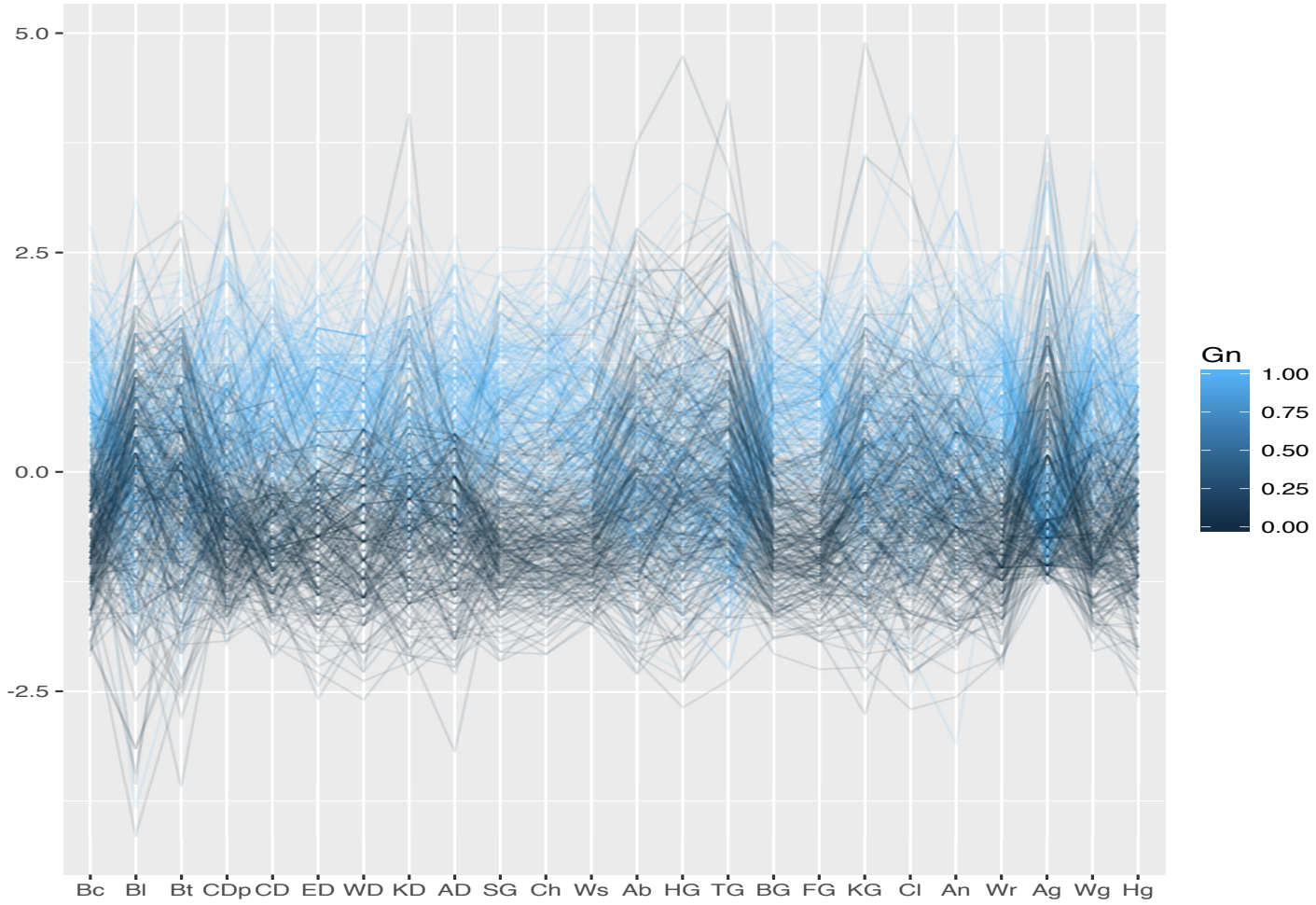
- Mostly, these are people in their twenties and thirties, with some older men and women.

- All exercise several hours a week.

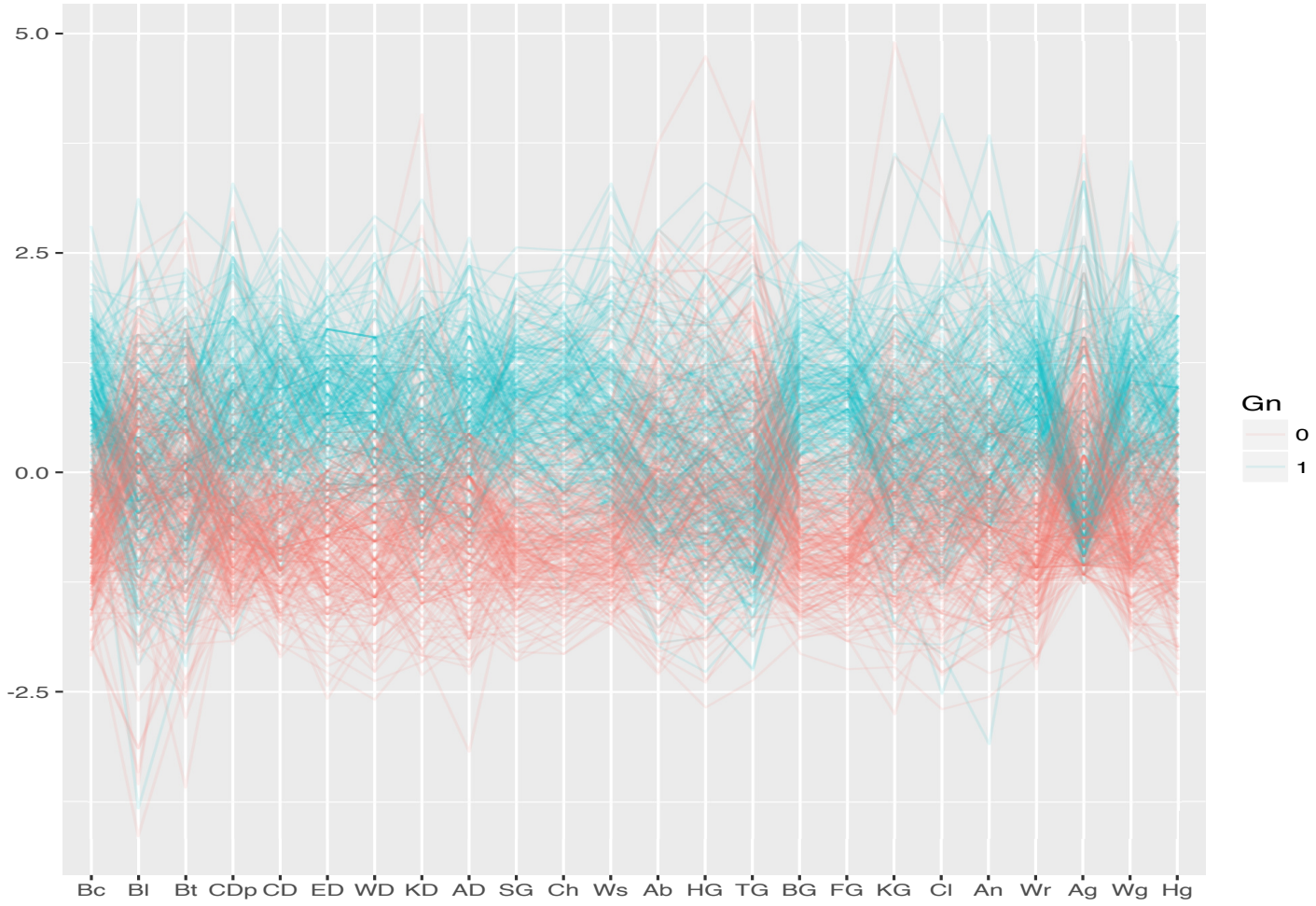
Body Data: Four Parallel Coord.



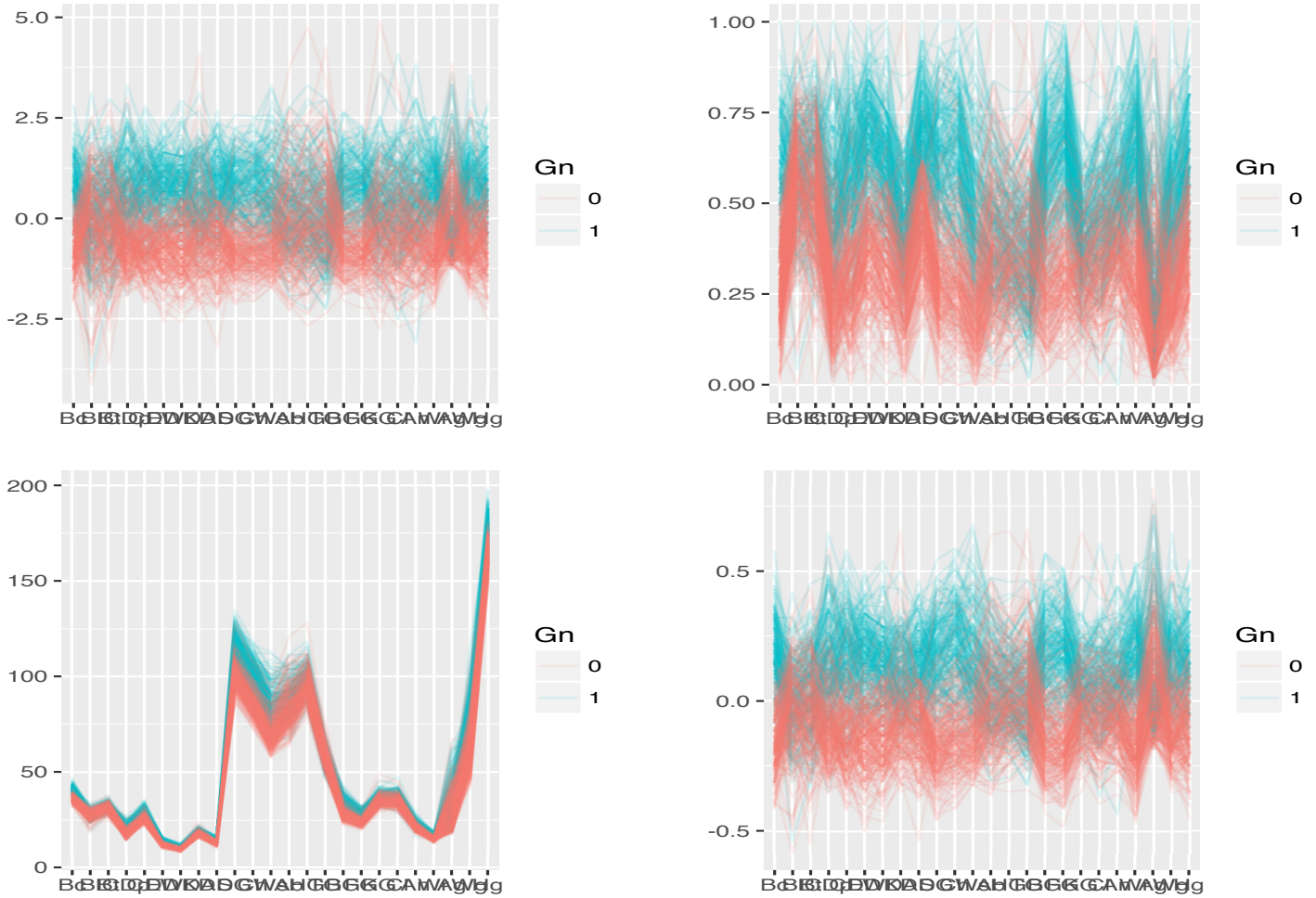
Body Data: Top-Left by Gender (??)



Body Data: Top-Left by Gender



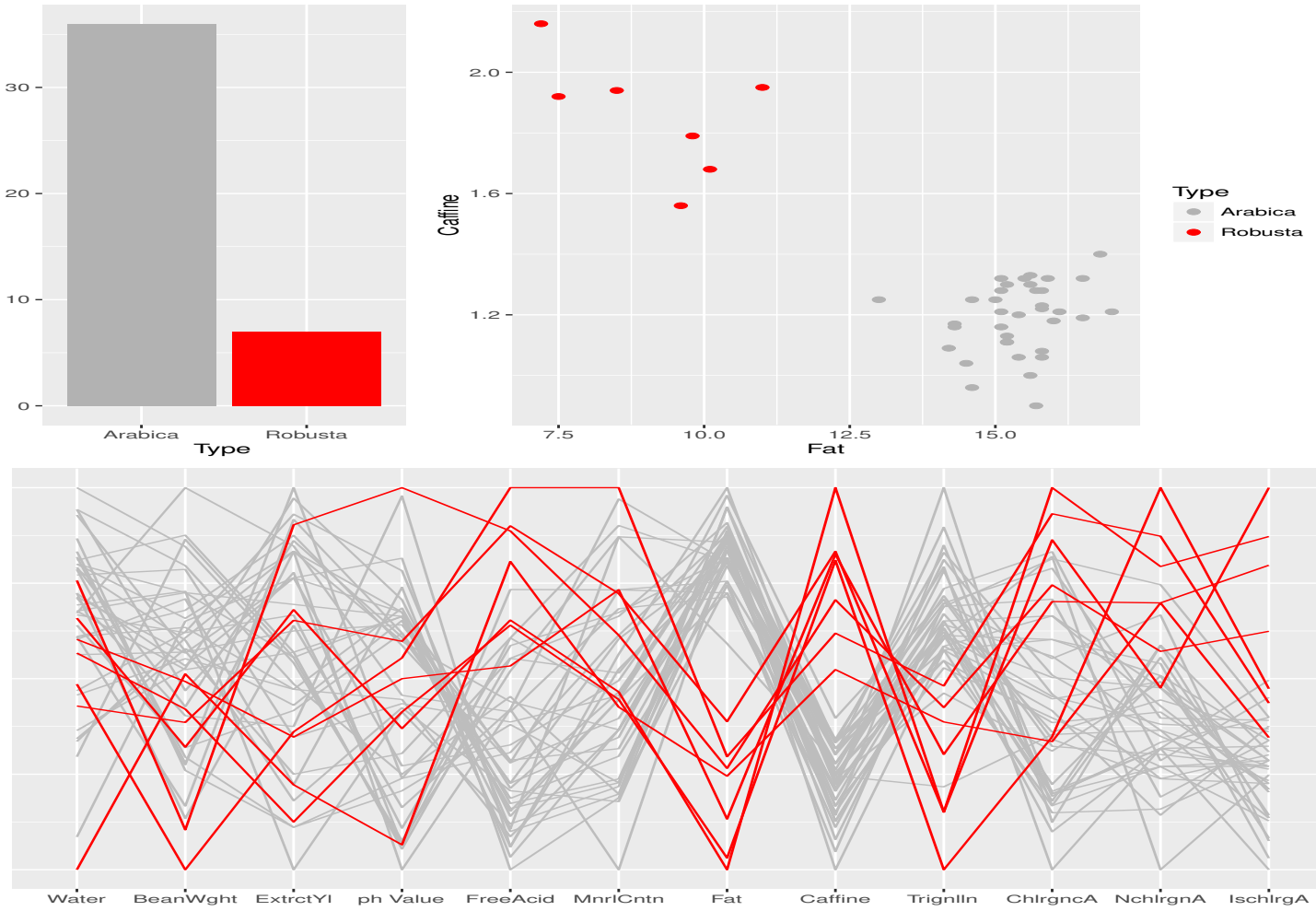
Body Data: All by Gender



Coffee Data

- Data on the chemical composition of coffee samples collected from around the world.
- A total of 43 samples from 29 countries.
- Each sample is either of the Arabica or Robusta variety.
- Twelve chemical constituents available in the `pgmm` package.

Coffee Data: Summary Panel



Comments

- We have seen of data visualization techniques for both continuous and categorical data types.
- We have learned, *inter alia*, that scale is very important for parallel coordinate plots (and in general).
- For further reading, see Unwin (2015).
- We will now look at basic, put important analytic technique for categorical data.

Introduction

- Consider a (large) data set comprising binary variables.
- One example is a transaction data base (or a market basket).
- A market basket analysis can be carried out.
- Or, more generally, an association rule analysis.

What is an Association Rule?

- Used to discover relationships in transaction databases.
- Transaction databases contain, exclusively, binary variables.
- Although formally introduced by Agrawal *et al*. (1993), many of the ideas behind association rules can be seen in the literature at least as far back as Yule (1903).
- No underlying statistical model is assumed and no hypotheses are formally proposed.

Definition of an Association Rule

- Given a non-empty set, *I*, an association rule is a statement of the form $A \Rightarrow B$, where $A, B \subset I$ such that $A \neq \emptyset, B \neq \emptyset$, and $A \cap B = \emptyset$.

- The set A is called the antecedent of the rule, the set B is called the consequent of the rule, and *I* is called the itemset. Association rules are generated over a large set of transactions, denoted $\tau_1, \ldots, \tau_n$.

- An association rule is deemed interesting if the items involved occur together often and there is evidence to suggest that one of the sets might in some sense lead to the presence of the other set.

- Association rules are commonly characterized by mathematical notions called support, confidence, and lift.

Functions of Association Rules

- Functions by which associations are traditionally characterised:

  Support : $s(A \Rightarrow B) = P(A, B)$.

  Confidence : $c(A \Rightarrow B) = P(B \mid A) = \dfrac{P(A, B)}{P(A)}$.

  Lift : $L(A \Rightarrow B) = \dfrac{c(A \Rightarrow B)}{P(B)} = \dfrac{P(B \mid A)}{P(B)} = \dfrac{P(A, B)}{P(A)P(B)}$.

- A variety of other functions have also been introduced.

Mining Rules

- Association rules are commonly generated using the apriori algorithm or a variant thereof (Agrawal & Skirant 1994, Borgelt & Kruse 2002, Bolgelt 2003).
- An implementation is available in the `arules` package in R.
- The algorithm requires a minimum support threshold, a minimum confidence threshold, and maximum rule length.
- Often, this will result in many association rules being generated.
- Pruning is then used to remove rules that are not "interesting".

Interestingness

- Confidence is popular as a measure of interestingness.
- Confidence, combined with support, is also popular.
- Lift is another approach that has some intuitive appeal.
- There are also other options, e.g., Gray and Orlowska's interestingness

Gray and Orlowska's Interestingness

- An example of another function is Gray and Orlowska's interestingness (Gray and Orlowska, 1998).

- Gray and Orlowska's is defined as follows:

$$\text{Int}(A \Rightarrow B; K, M) = \left[\left(\frac{P(A,B)}{P(A)P(B)} \right)^K - 1 \right] [P(A)P(B)]^M.$$

- Presents a compromise between the distance of lift (to the power of K) from one and the respective magnitudes of $P(A)$ and $P(B)$ (to the power of M); K and M can be viewed as weights.

- It is symmetric in the sense that $\text{Int}(A \Rightarrow B; K, M) = \text{Int}(B \Rightarrow A; K, M)$ — because lift is symmetric in the sense that $L(A \Rightarrow B) = L(B \Rightarrow A)$.

Gray and Orlowska's Int. Contd.

- McNicholas (2007) gives an argument for setting $K = M$, based on:

$$\begin{aligned}
\text{Int}(A \Rightarrow B; K, M) &= \left[\left(\frac{P(A,B)}{P(A)P(B)}\right)^K - 1\right](P(A).P(B))^M \\
&= \left[\frac{\left(\frac{P(A,B)}{P(A)}\right)^K - P(B)^K}{P(B)^K}\right](P(A).P(B))^M \\
&= \left[P(B \mid A)^K - P(B)^K\right] P(A)^M P(B)^{M-K} \\
&= \left[c(A \Rightarrow B)^K - P(B)^K\right] P(A)^M P(B)^{M-K}.
\end{aligned}$$

- Now, for $K = M$:

$$\text{Int}(A \Rightarrow B; K) = \left[c(A \Rightarrow B)^K - P(B)^K\right] P(A)^K.$$

Lift

- Recall that
$$L(A \Rightarrow B) = \frac{P(A, B)}{P(A)P(B)}.$$
- The lift seems appealing as a measure of interestingness.
- An interesting rule has lift "far" from 1, but the lift is not symmetric about 1.
- One solution is to consider $\log L(A \Rightarrow B)$.
- A better solution would be to find the upper and lower bounds of lift.

Bounds in $P(A)$ and $P(B)$

- The range of values that the lift of an association rule $A \Rightarrow B$ can take is restricted by the respective values of $P(A)$ and $P(B)$;

$$\frac{\max\{P(A)+P(B)-1, 1/n\}}{P(A)P(B)} \leq L(A \Rightarrow B) \leq \frac{1}{\max\{P(A), P(B)\}}, \quad (1)$$

  where n is the number of transactions.

- These bounds that are almost identical to those derived by Fréchet (1951) and could be used to standardize the lift.

- What if minimum thresholds for support and confidence were used in the mining process?

Considering Minimum Support

- Suppose the minimum support threshold is s.
- We have
$$\frac{4s}{(1+s)^2} \leq L(A \Rightarrow B) \leq \frac{1}{s}.$$
- Setting $s = 1/n$ gives the bound if no support threshold is used;
$$\frac{4n}{(n+1)^2} \leq L(A \Rightarrow B) \leq n.$$
- These quantities, alone, are useless for standardizing lift with a view to ranking association rules.

Adding Minimum Confidence

- Suppose the minimum confidence threshold is c.

- Considering this and the other bounds, we have

$$\max\left\{\frac{P(A)+P(B)-1}{P(A)P(B)}, \frac{4s}{(1+s)^2}, \frac{s}{P(A)P(B)}, \frac{c}{P(B)}\right\} \leq L(A \Rightarrow B) \leq \frac{1}{\max\{P(A), P(B)\}}. \quad (2)$$

- McNicholas et al. (2008) use (2) to standardize the lift.

Standardized Lift

- Denote:

$$\upsilon = \frac{1}{\max\{P(A), P(B)\}},$$

$$\lambda = \max\left\{\frac{P(A)+P(B)-1}{P(A)P(B)}, \frac{4s}{(1+s)^2}, \frac{s}{P(A)P(B)}, \frac{c}{P(B)}\right\}.$$

- The standardized lift is given by

$$\mathcal{L}(A \Rightarrow B) = \frac{L(A \Rightarrow B) - \lambda}{\upsilon - \lambda}.$$

German Social Life Feeling Data

- Consider data taken from a study of German "social life feelings" that appeared in Schuessler (1982) and Krebs & Schuessler (1987).

- These data have been analyzed many many times, including by Bartholomew & Schuessler (1991), Bartholomew (1991), Bartholomew *et al.* (1997), de Menezes & Bartholomew (1996) and Bartholomew & Knott (1999).

- The data used herein represent the answers given by a sample of 1,490 Germans to five questions.

- McNicholas et al. (2008) use these data in a paper on association rules.

The Questions

1. Anyone can raise his standard of living if he is willing to work at it.

2. Our country has too many poor people who can do little to raise their standard of living.

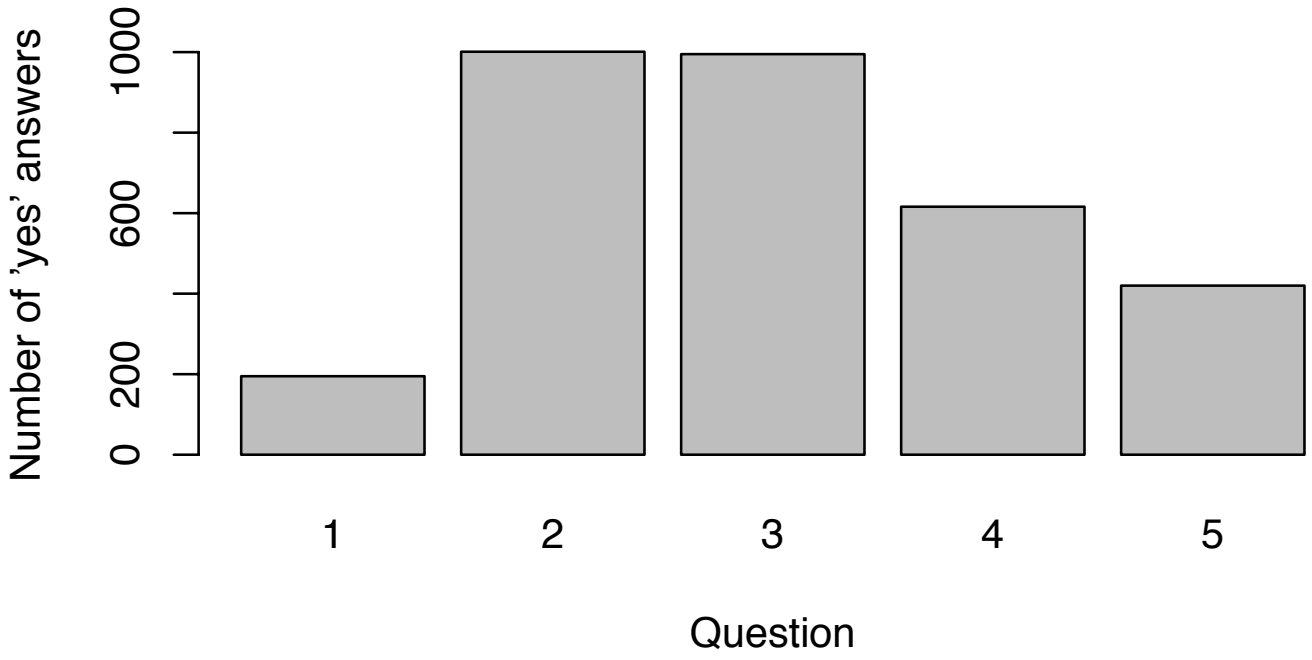
3. Individuals are poor because of the lack of effort on their part.

4. Poor people could improve their lot if they tried.

5. Most people have a good deal of freedom in deciding how to live.

The Answers

- Overall, there were 3,227 "yes" answers and 4,223 "no" answers.



- Only questions 2 and 3 had more than 50% "yes" answers.

Negations?

- These data raise an interesting point: the fact that "yes" is coded "1" and "no" is coded "0" can be viewed as arbitrary.

- Further, had the questions been worded differently, the "1"s and "0"s could have been be flipped in some or all of the questions.

- The term "negation" can be used to denote the absence of an item from a transaction.

Negations & German Social Data

- The "1"s were coded y1, y2, y3, y4 and y5, respectively, while the "0"s, or negations, were coded n1, n2, n3, n4 and n5, respectively.

- Association rules were generated using the `arules` package in R with minimum support set at 20% and minimum confidence at 80%.

- This approach led to the generation of 38 association rules.

- We will look at the R code later...

Negations & German Social Data

- The "best" rule, ranked by standardized lift was $\{\text{n3,n4}\} \Rightarrow \{\text{n1}\}$.

- 95.1% of those who did not agree that people were poor because of lack of effort or that poor people could improve their lot if they tried, also did not agree that people could raise their standard of living if they were willing to work at it.

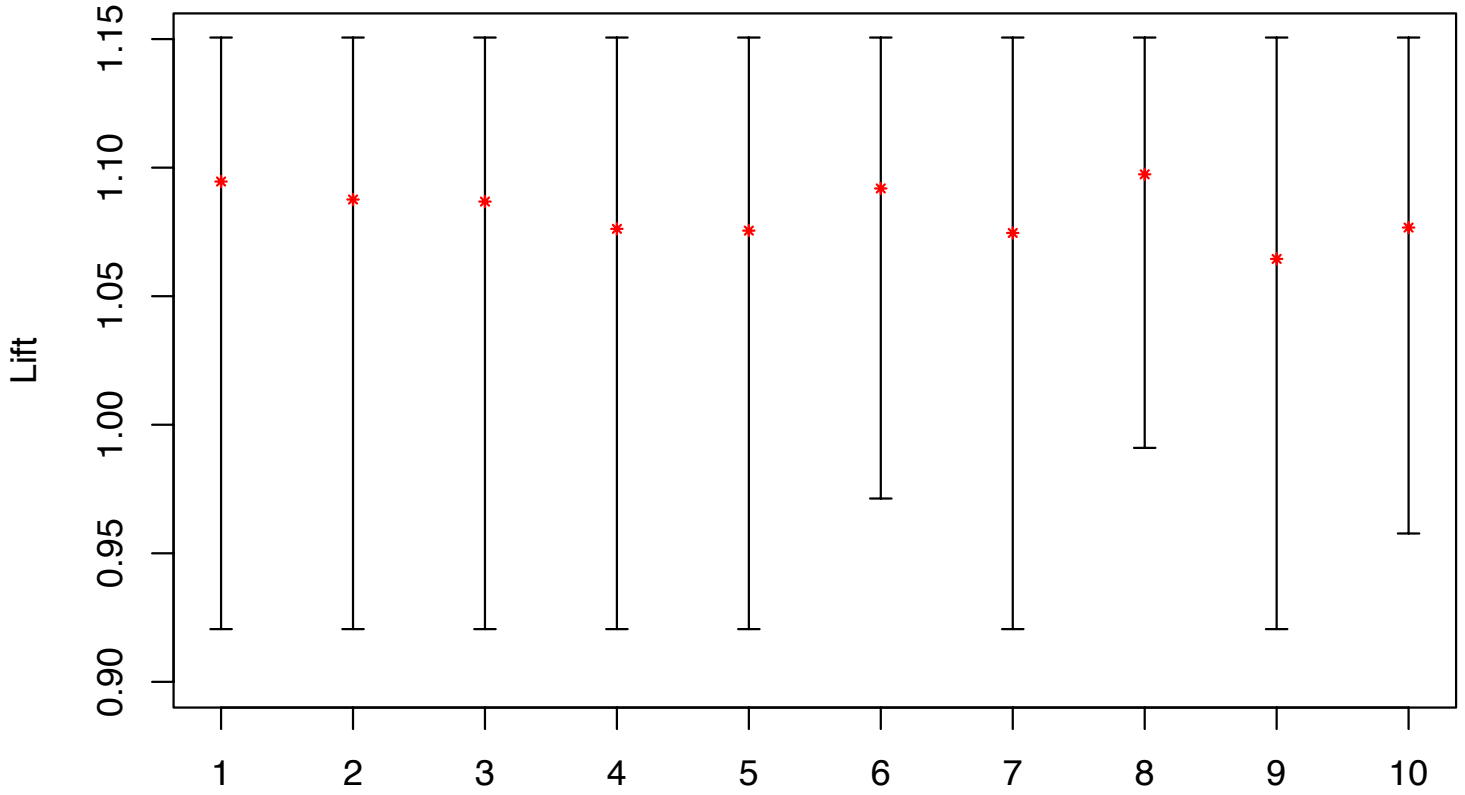
- The ranking of the 38 rules by $\mathcal{L}$ is different than the ranking by either confidence or lift.

German Social Data: Top 10 Rules

Id.	Rule	Supp.	Conf.	Lift	$\mathcal{L}$	LB	UB
1	{n3,n4} ⇒ {n1}	0.262	0.951	1.095	0.757	0.920	1.151
2	{n3,n5} ⇒ {n1}	0.255	0.945	1.088	0.726	0.920	1.151
3	{n4,n5} ⇒ {n1}	0.446	0.945	1.087	0.723	0.920	1.151
4	{n3} ⇒ {n1}	0.311	0.935	1.076	0.677	0.920	1.151
5	{n4} ⇒ {n1}	0.548	0.935	1.076	0.674	0.920	1.151
6	{n2,n4} ⇒ {n1}	0.225	0.949	1.092	0.673	0.971	1.151
7	{n4,n5,y2} ⇒ {n1}	0.256	0.934	1.075	0.670	0.920	1.151
8	{n3,n4,n5} ⇒ {n1}	0.221	0.954	1.097	0.667	0.991	1.151
9	{n4,y2} ⇒ {n1}	0.323	0.925	1.064	0.626	0.920	1.151
10	{n4,n5,y3} ⇒ {n1}	0.225	0.936	1.077	0.617	0.958	1.151

Visualization of Std. Lift

- Standardizing the lift shows that rules with the higher lift are not necessarily better rules.



Redundant Rules

- Looking at the top 10 rules for the German social data, raises an interesting question.
- Are some of the rules redundant.
- That is, are there rules with "extra" elements?
- Let's take a look.

German Data: Redundant Rules (in Top 10)

Id.	Rule	Supp.	Conf.	Lift	$\mathcal{L}$	LB	UB
1	{n3,n4} ⇒ {n1}	0.262	0.951	1.095	0.757	0.920	1.151
2	{n3,n5} ⇒ {n1}	0.255	0.945	1.088	0.726	0.920	1.151
3	{n4,n5} ⇒ {n1}	0.446	0.945	1.087	0.723	0.920	1.151
4	{n3} ⇒ {n1}	0.311	0.935	1.076	0.677	0.920	1.151
5	{n4} ⇒ {n1}	0.548	0.935	1.076	0.674	0.920	1.151
6	{n2,n4} ⇒ {n1}	0.225	0.949	1.092	0.673	0.971	1.151
7	{n4,n5,y2} ⇒ {n1}	0.256	0.934	1.075	0.670	0.920	1.151
8	{n3,n4,n5} ⇒ {n1}	0.221	0.954	1.097	0.667	0.991	1.151
9	{n4,y2} ⇒ {n1}	0.323	0.925	1.064	0.626	0.920	1.151
10	{n4,n5,y3} ⇒ {n1}	0.225	0.936	1.077	0.617	0.958	1.151

Why Include Negations?

- Negations were used to facilitate a full analysis of the German social data.
- In general, it will often be the case that the absence of items from the antecedent and the consequent parts of an association rule may of interest.
- The absence of items from the antecedent part can be related to the presence or absence of items from the consequent part and *vice versa*.
- McNicholas et al. (2008) discuss the history of negations and negative association rules.

How Many More Rules?

- Including negations will lead to the generation of more rules, but how many more?

- Without negations, Hipp et al. (2002) calculate the number of rules that could be generated as:

$$3^n - 2\,(2^n) + 1.$$

- When negations are included, McNicholas et al. (2008) calculate the number as:

$$5^n - 2\,(3^n) + 1.$$

Derivation of $5^n - 2\left(3^n\right) + 1$

- If A and B contain a total of m items then the number of rules involving these m items and their negations is given by

$$\left[\sum_{r=1}^{m-1} {}^{m}C_r\right] 2^m = \left[\sum_{r=0}^{m} {}^{m}C_r - 2\right] 2^m = \left(2^m - 2\right) 2^m = 2^{2m} - 2^{m+1}.$$

- Therefore, from an itemset of size n there are $2^{2n} - 2^{n+1}$ rules of length n,
  ${}^{n}C_{n-1}\left[2^{2(n-1)} - 2^{(n-1)+1}\right]$ rules of length $n-1$,
  ${}^{n}C_{n-2}\left[2^{2(n-2)} - 2^{(n-2)+1}\right]$ rules of length $n-2$ and so on.

- It follows that the total number of rules that can be generated from these n items and their negations is given by

$$\left(2^{2n} - 2^{n+1}\right) + {}^{n}C_{n-1}\left[2^{2(n-1)} - 2^{(n-1)+1}\right] + \cdots + {}^{n}C_{2}\left[2^{2(2)} - 2^{2+1}\right].$$

Derivation of $5^n - 2\left(3^n\right) + 1$.

- Now, this can be expressed as $\sum_{i=2}^{n} {}^nC_i \left(2^{2i} - 2^{i+1}\right)$.
- From the binomial theorem, $(1+x)^n = \sum_{i=0}^{n} {}^nC_i x^i$.
- So we can write,

$$\sum_{i=2}^{n} {}^nC_i \left(2^{2i} - 2^{i+1}\right) = \sum_{i=2}^{n} {}^nC_i 2^{2i} - \sum_{i=2}^{n} {}^nC_i 2^{i+1}$$

$$= \sum_{i=2}^{n} {}^nC_i 4^i - 2\sum_{i=2}^{n} {}^nC_i 2^i$$

$$= \left[\sum_{i=0}^{n} {}^nC_i 4^i - ({}^nC_0 + {}^nC_1(4))\right] - 2\left[\sum_{i=0}^{n} {}^nC_i 2^i - ({}^nC_0 + {}^nC_1(2))\right]$$

$$= [5^n - (1+4n)] - 2\left[3^n - (1+2n)\right] = 5^n - 2\left(3^n\right) + 1.$$

How Many More Rules?

- So, the number of 'extra' rules that can be mined when negations are included is given by

$$5^n - 2\left(3^n\right) + 1 - \left[3^n - 2\left(2^n\right) + 1\right] = 5^n - 3^{n+1} + 2^{n+1}. \tag{3}$$

- Now, the proportion of rules that contain negations is given by

$$\frac{5^n - 3^{n+1} + 2^{n+1}}{5^n - 2\left(3^n\right) + 1}. \tag{4}$$

- Therefore, when an itemset of just 20 items is considered, it follows that 99.996% of potential association rules involve negations.

How Many More Rules?

- Of course, the number of rules given by (3) may be unrealistic because a minimum of 2^n transactions would be required in order that all potential rules may exist.

- Also, it is unlikely in most practical applications that a transaction involving all, or even most, of the items will occur.

- However, (4) does provide a useful estimate for the proportion of potential rules that will contain negations.

Comments

- Association rules have been introduced.
- We saw one example, on the German social life feeling data.
- In this example, negations were introduced to facilitate the mining process.
- A general argument was given for the inclusion of negations in the association rule mining process, including a quantification of the amount of rules that can be mined when negations are included.
- Next, we will mine some rules (in R), starting with another look at the German social life feeling data.
- Full bibliographical details for the key references related to association rules are given by McNicholas et al. (2008)[a].

[a] McNicholas, P.D., Murphy, T.B. and O'Regan, M. (2008), 'Standardising the lift of an association rule', *Computational Statistics and Data Analysis* **52**(10), 4712–4721.