Data Visualization

Prof. Sharon McNicholas

STATS 780/CSE 780

Introduction

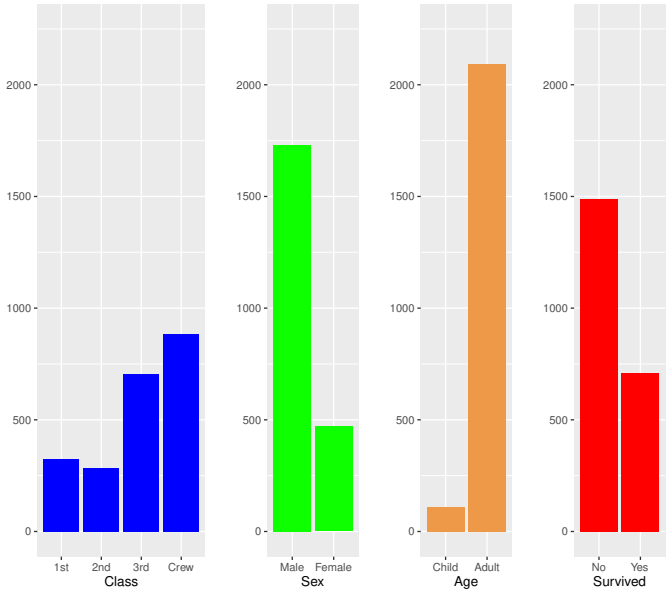
- People talk about "looking at data", "looking at plots", etc.
- More than just looking at data, we want to be able to develop hypotheses or draw tentative conclusions using graphs.
- This is sometimes called graphical data analysis (GDA) — in a *bona fide* GDA, a lot of graphs would be used.
- First, 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.

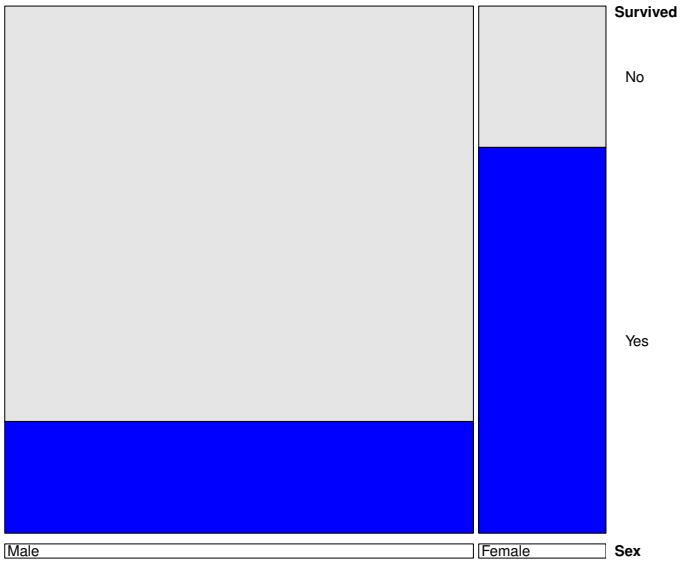
The Titanic Data Revisited

- Last class, we used association rules to analyze the Titanic data.
- Recall that `Titanic` contains four variables for 2,201 passengers.
- The variables are:
  - Class (1st, 2nd, 3rd, crew)
  - Sex (male, female)
  - Age (child, adult)
  - Survived (no, yes)

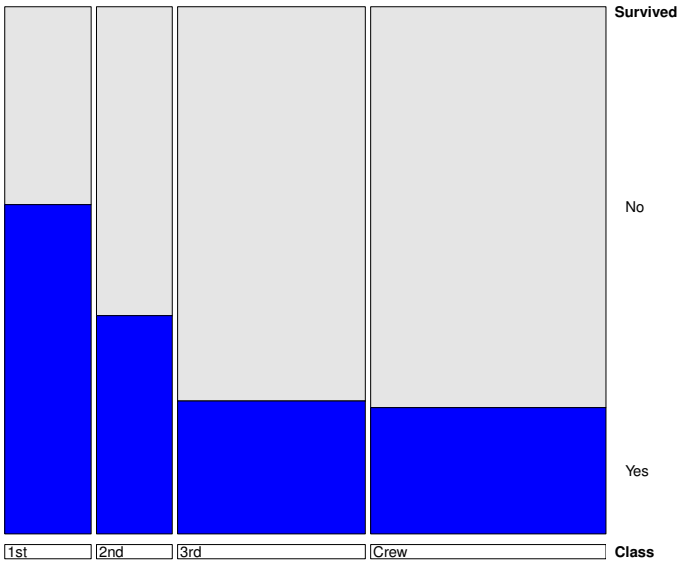
Simple Grid



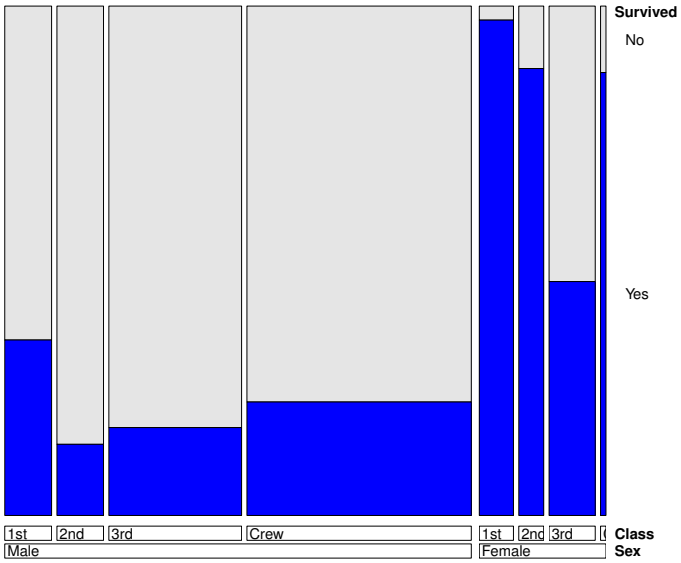
Doubledecker Plot 1



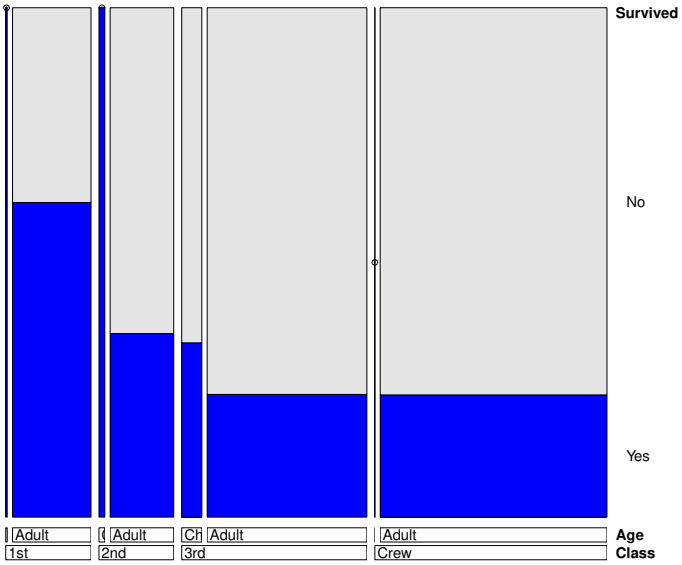
Doubledecker Plot 2



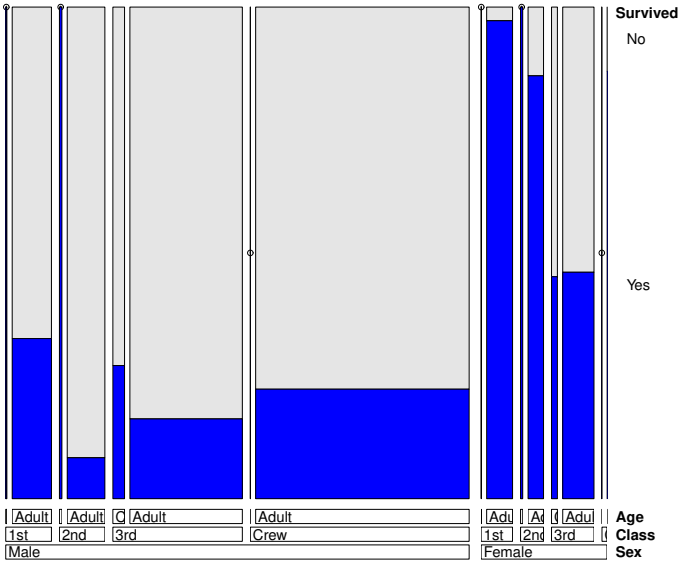
Doubledecker Plot 3



Doubledecker Plot 4



Doubledecker Plot 5



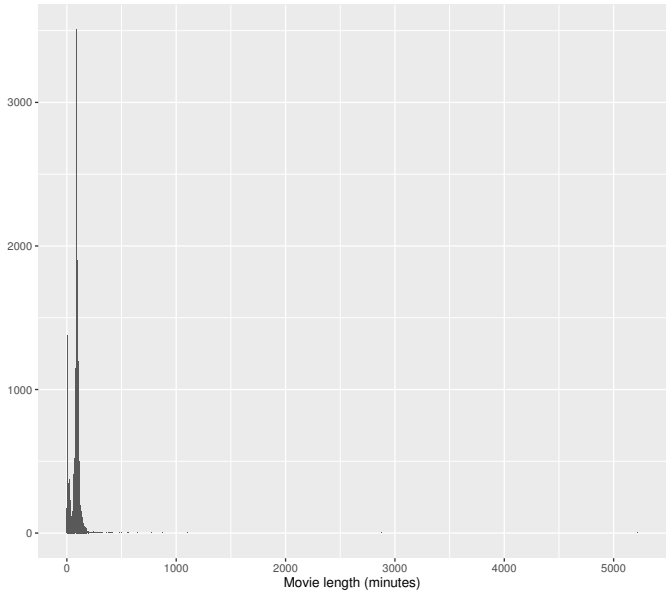
The Titanic Data Comments

- The grid plot gave a sense of how the numbers of passengers brake down.
- The doubledecker plots gave similar information to the association rules we saw last week.
- The fifth doubledecker plot gave the clearest breakdown.
- I think doubledecker plots are a nice compliment to an association rule analysis; however, they are not a method of visualizing association rules *per se*.

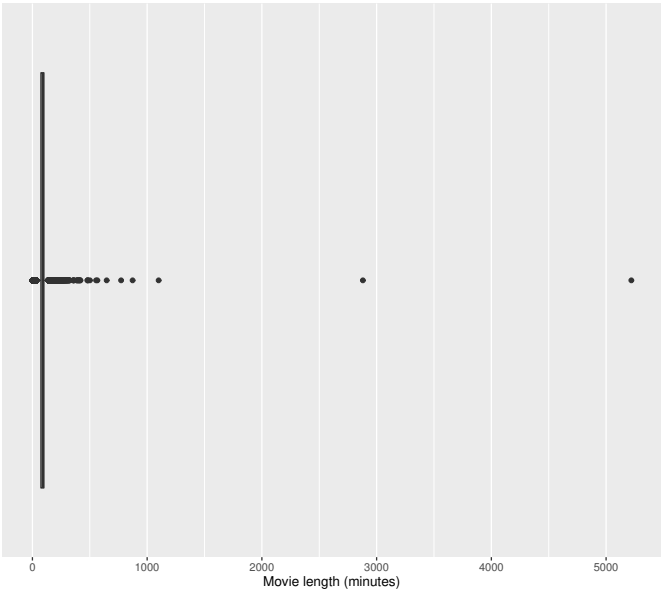
Movies Data

- The Clint Eastwood movie Sully has 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.

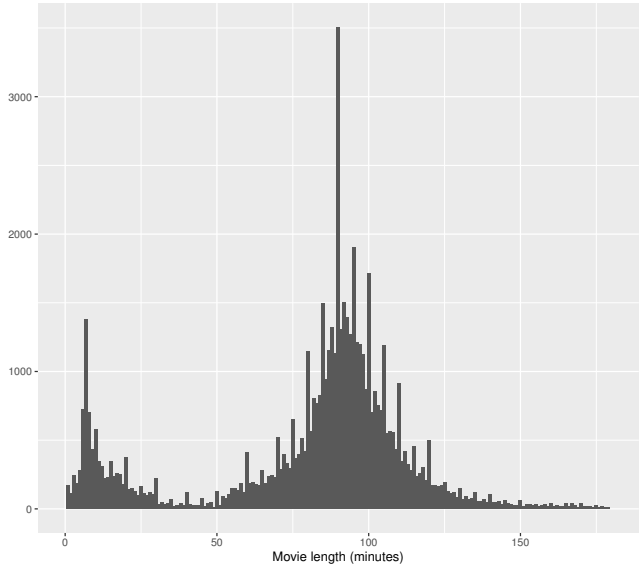
Simple Histogram



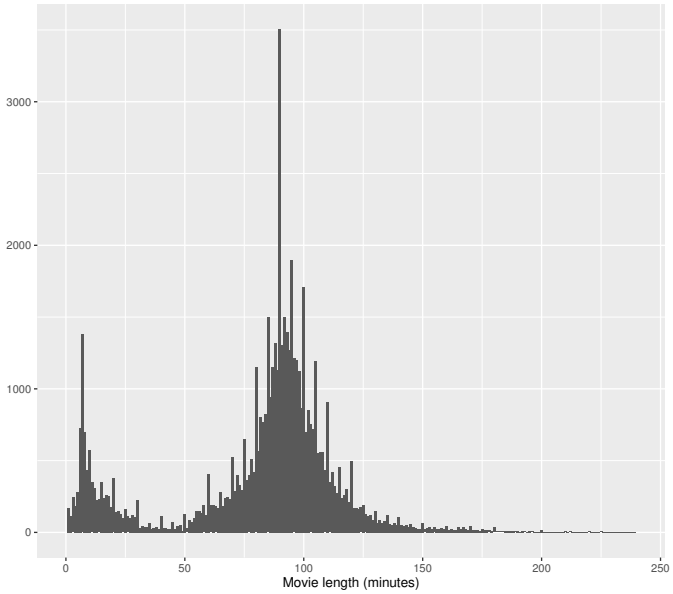
Alternative Plot



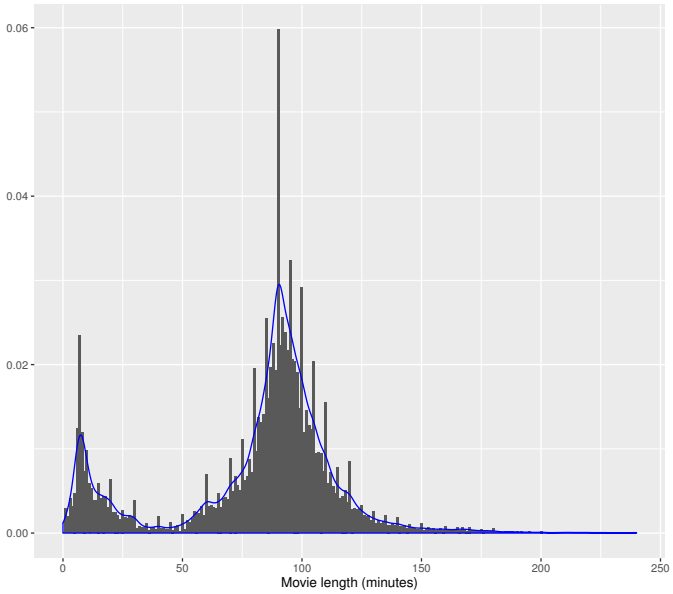
Histogram: Up to 3 Hours



Histogram: Up to 4 Hours



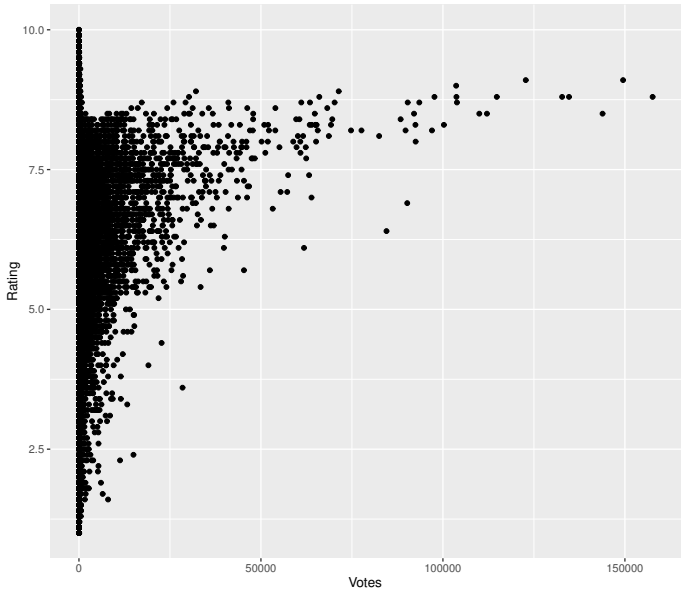
Histogram: With Density



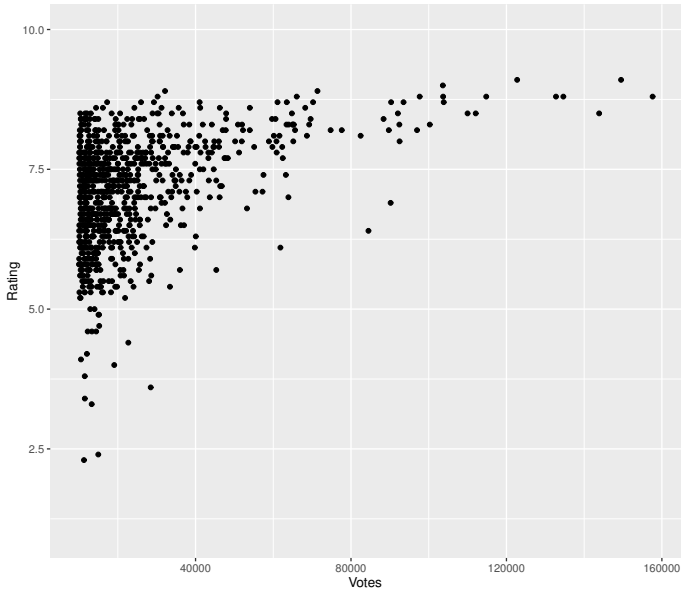
Movie Data Comments

- 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...

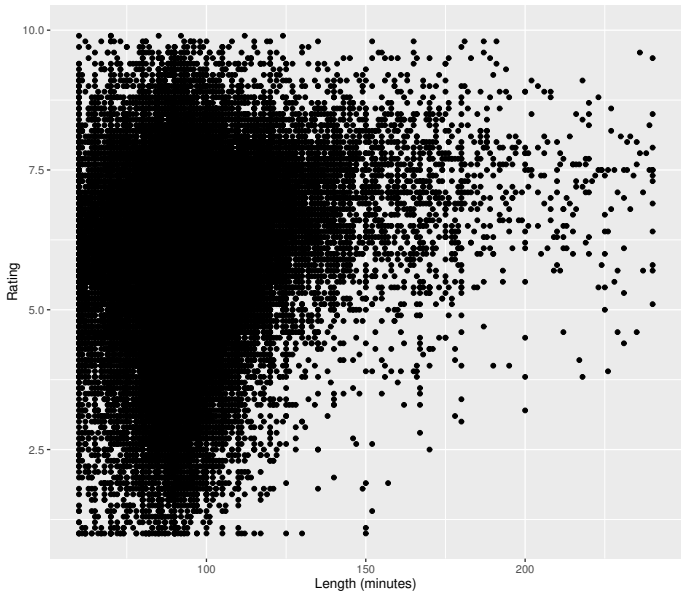
Votes vs. Rating



Votes vs. Rating (>10,000 Votes)



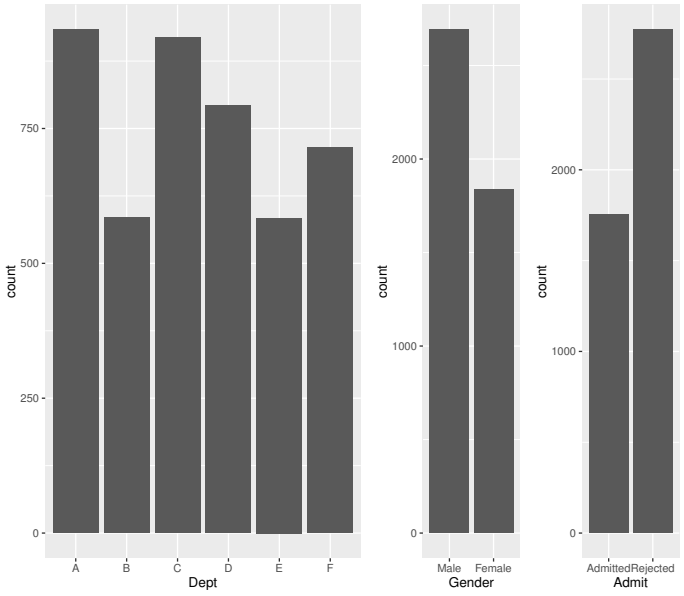
Length vs. Rating (1–4 Hours)



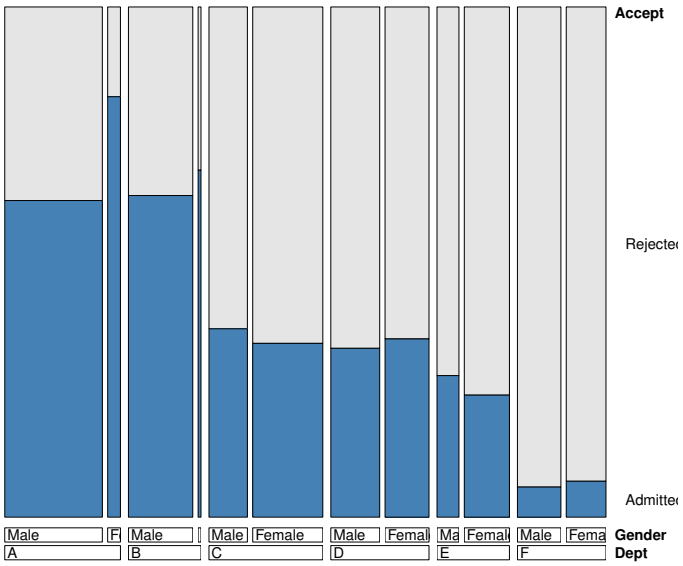
Berkeley Admissions Data

- 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

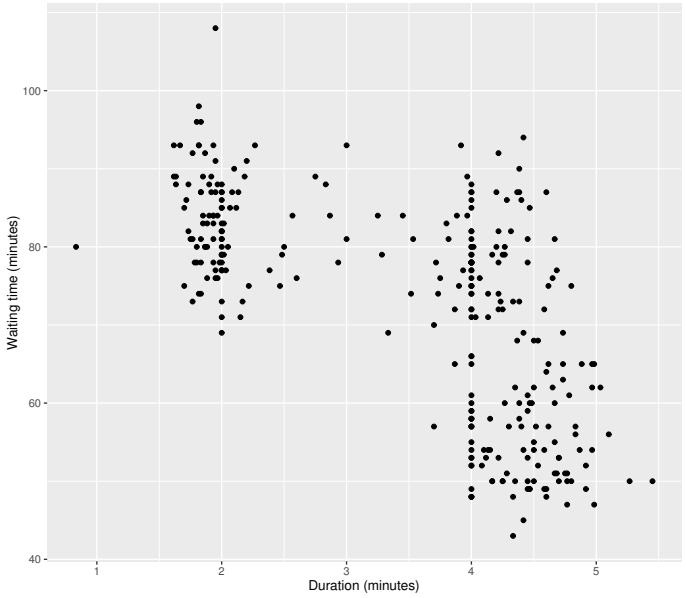
- The `geyser` data set contains a total of 299 observations of eruption duration (in minutes) and waiting time (in minutes, for this eruption) for the Old Faithful geyser.

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

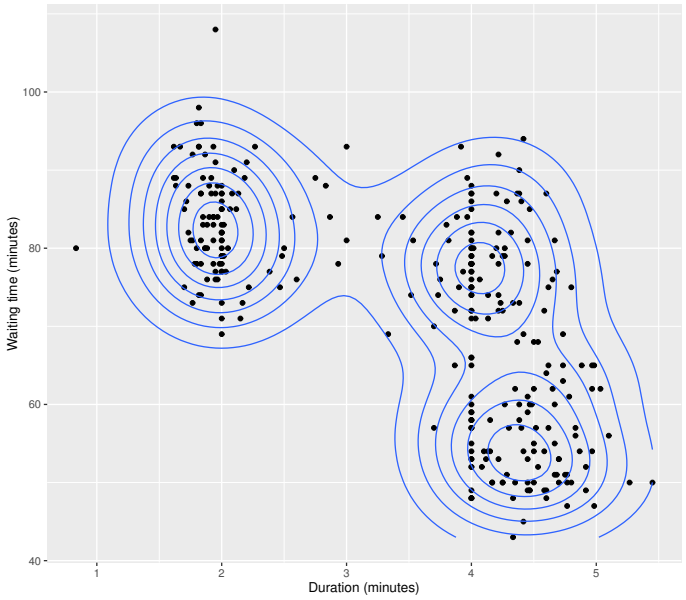
- Studied in detail by Azzalini and Bowman (1990)[a].

[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.
- At beginning of the next "lecture", we will look at a more (graphically) difficult scatter plot with contours.
- But first, we will play around a bit in R.