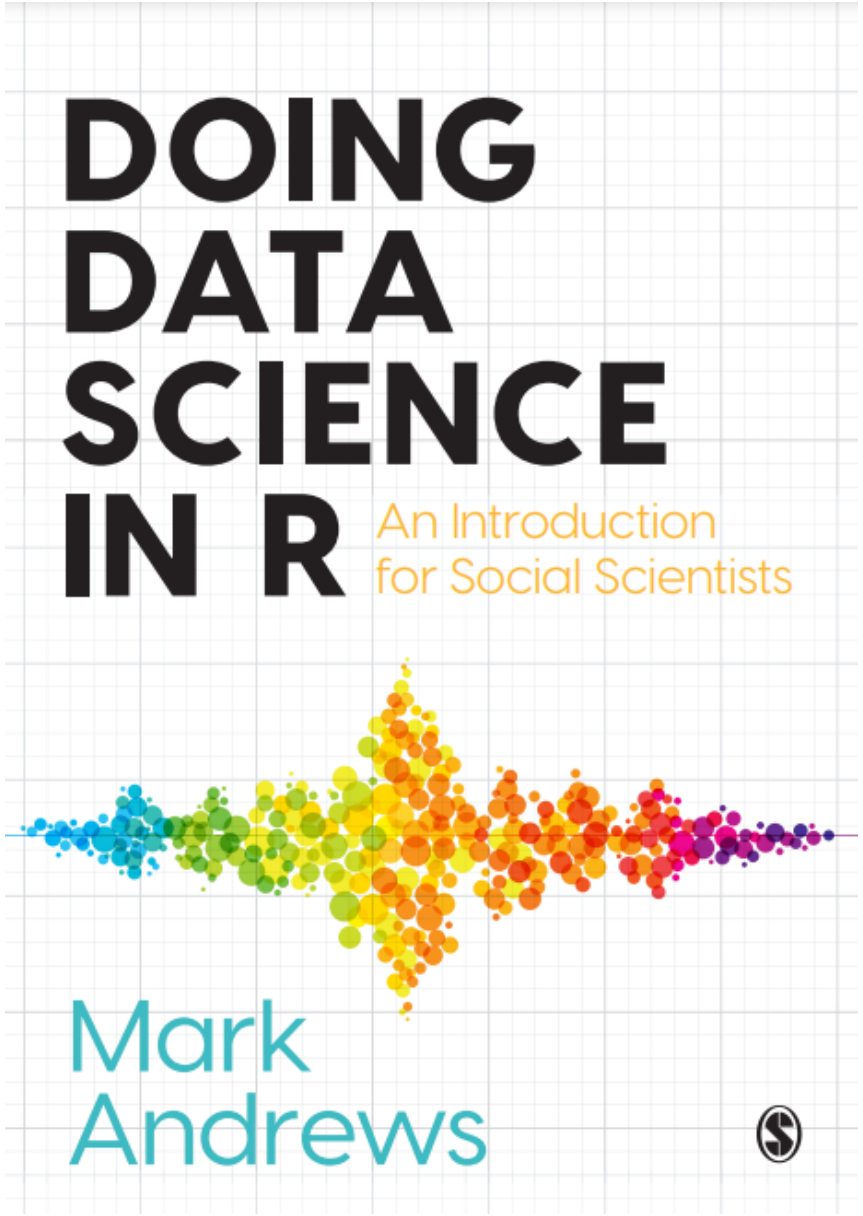


Doing Data Science in R: An Introduction for Social Scientists

Chapter 4

Data Visualisation

Introduction

Data visualization is a major part of data analysis.

Visualization allows us explore data and find patterns that would easily be missed were we to rely only on numerical summary statistics

Plotting in R with ggplot

In R, there are two major sets of tools for visualization. These are usually known as the base R and the ggplot plotting systems.

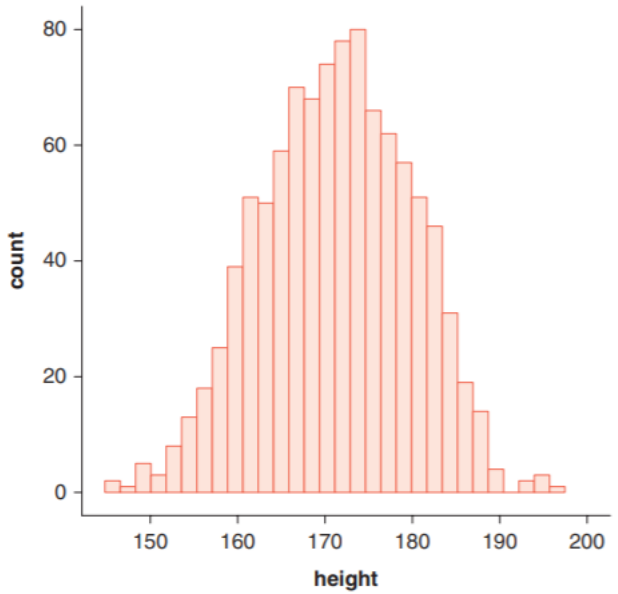
While the base R plotting system is powerful and not to be dismissed or seen as obsolete, here will exclusively use ggplot.

Histograms, density plots, barplots, etc.

Histograms and related visualization methods are simple but still highly effective tools to visualize the distribution of values of continuous variables.

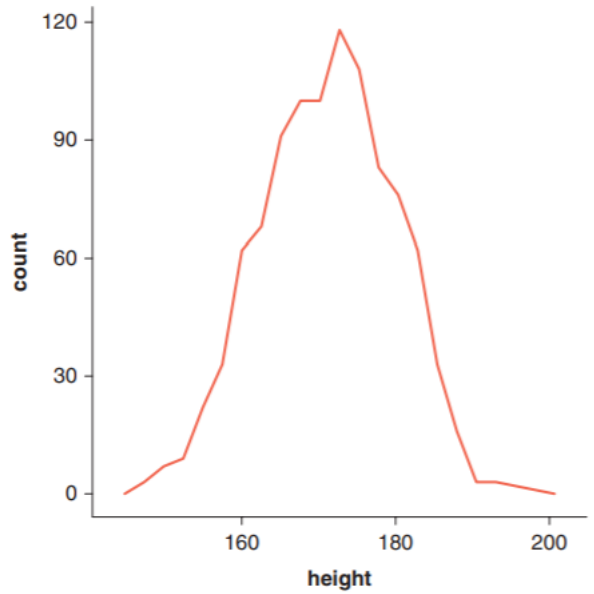
Histograms

Histograms are one of the simplest and generally most useful ways of visualizing distributions of the values of individual variables.



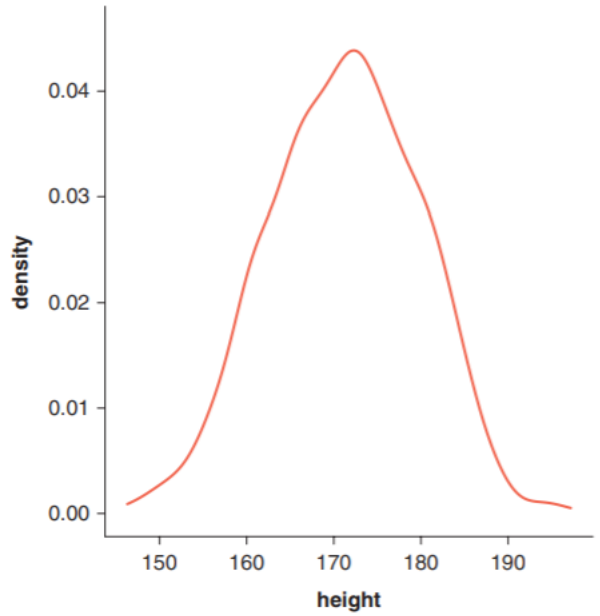
Frequency polygons

A frequency polygon is similar to a histogram but instead of using bars to display the number of values in each bin, it uses connected lines



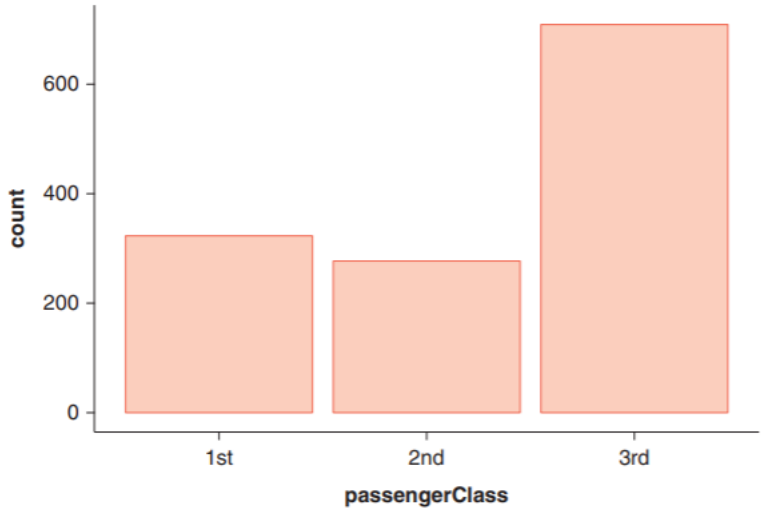
Density plots

Density plots use kernel density estimation to estimate a probability density over the variable



Barplots

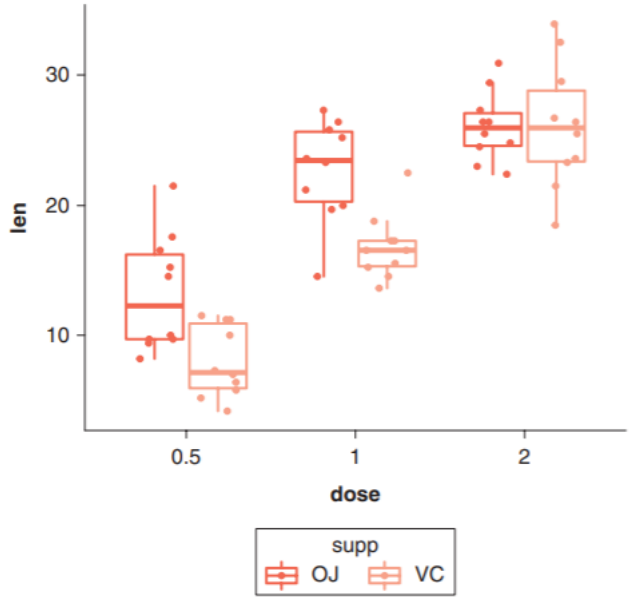
For each value of the discrete variable, the barplot displays the number of observed instances of that value in the data



Tukey boxplots

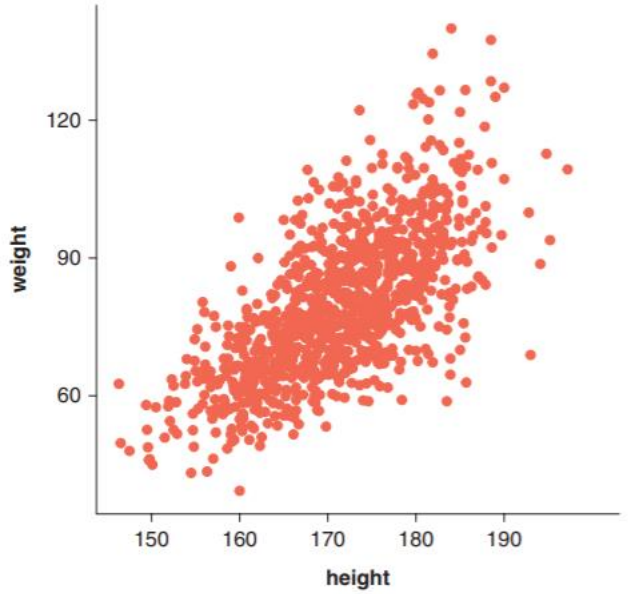
One subtype of boxplot is the Tukey boxplot (Tukey, 1977).

These are in fact the most common subtype and are the default type implemented in ggplot2 using the geom_ boxplot function.



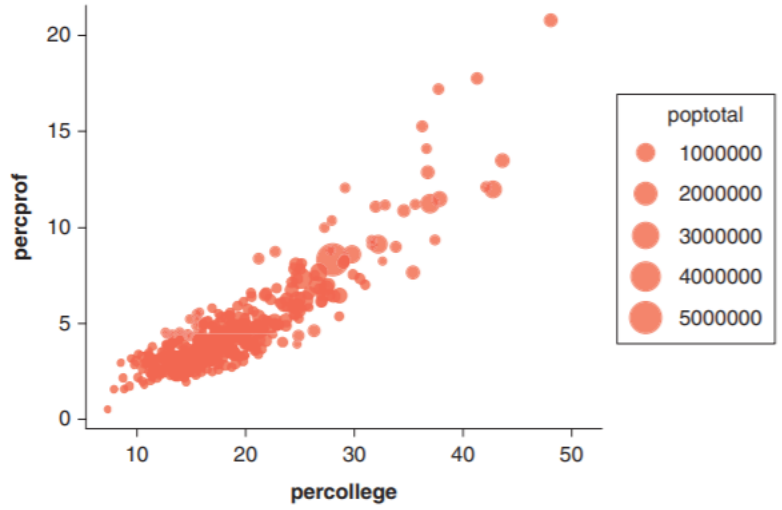
Scatterplots

We've already seen some simple scatterplots. Here, we'll provide more in-depth coverage using the weight_df data frame.



Bubbleplots

Bubbleplots are scatterplots where the size of the point is determined by the value of a third variable



Facet plots

Facet plots allow us to produce multiple related subplots, where each subplot displays some subset of the data.