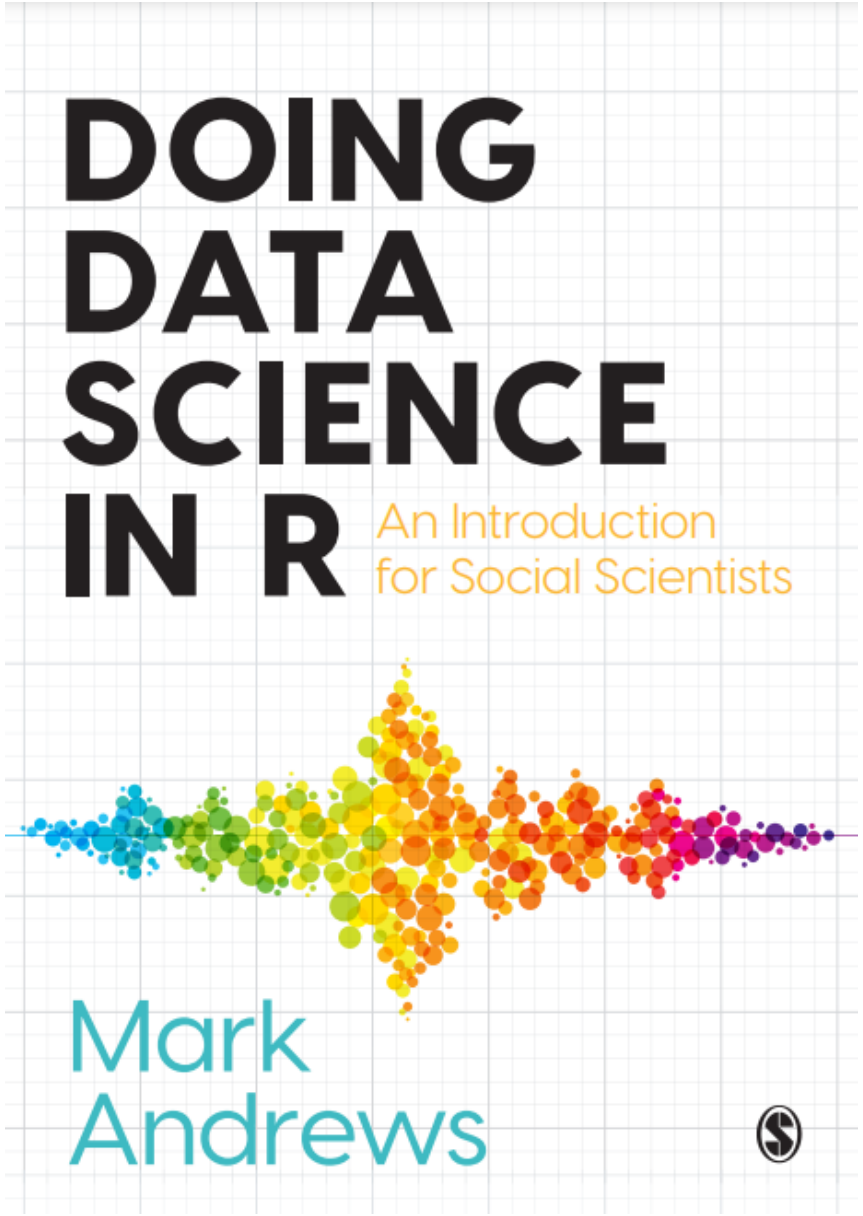

Doing Data Science in R: An Introduction for Social Scientists

Chapter 5

Exploratory Data Analysis

Introduction

John Tukey (1977) describes exploratory data analysis as detective work.

Exploratory data analysis is a vital first step prior to confirmatory data analysis and data modelling generally

Univariate data

Univariate data is data concerning a single variable.

Conditionally univariate analysis is also a major part of exploratory data analysis

Types of univariate data

rStevens (1946) defined four (univariate) data types that are characterized by whether their so-called level of measurement is nominal, ordinal, interval, or ratio.

Continuous data

Continuous data represents the values or observations of variable that can take any value in a continuous metric space such as the real line or some interval thereof

Categorical data

Categorical data is where each value takes one of a finite number of values that are categorically distinct and so are not ordered

Ordinal data

Ordinal data represents values of a variable that can be ordered but have no natural or uncontroversial sense of distance between them.

Count data

Count data is tallies of the number of times something has happened or some value of a variable has occurred.

Characterizing univariate distributions

We can describe any univariate distribution in terms of three major features: location, spread, and shape

Measures of central tendency

Three commonly used measures of central tendency are the arithmetic mean, the median, and the mode.

Arithmetic mean

The arithmetic mean, or usually known as simply the mean, is defined as follows:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i .$$

Median

The median of a finite sample is defined as the middle point in the sorted list of its values

```
median(rt_data)
#> [1] 517
```

Mode

The sample mode is the value with the highest frequency

$$\text{mode} = \underset{x}{\operatorname{argmax}} \, \mathrm{P}(X = x),$$

Robust measures of central tendency

The trimmed mean removes a certain percentage of values from each extreme of the distribution before calculating the mean as normal.

The following code, for example, removes 10% of values from the high and low extremes of rt_data:

```
mean(rt_data, trim = 0.10)
#> [1] 489.6429
```

Measures of dispersion

The standard measure of the dispersion of distribution is the variance or standard deviation.

For a continuous random variable X the variance is defined as

variance = (x x) (X x) , dx 2 P

For a discrete random variable X it is defined as

variance = (x x) (X x)

Trimmed and winsorized estimates of variance and standard deviation

The variance and standard deviation are doubly or triply susceptible to outliers.

We may trim or winsorize the values before calculating the variance or standard deviation

Median absolute deviation

As the name implies, it is the median of the absolute differences of all values from the median

$$\mathrm{MAD} = \mathrm{median}(|x_i - m|).$$

Range estimates of dispersion

By far the simplest measure of the dispersion of a set of values is the range, which is the difference between the maximum and minimum values.

Measure of skewness

Skewness is a measure of the asymmetry of a distribution of numbers

The skewness is the third standardized moment defined as follows:

$$\text{skew} = \frac{\left\langle (X-\mu)^3 \right\rangle}{\left\langle (X-\mu)^2 \right\rangle^{3/2}} = \frac{\left\langle (X-\mu)^3 \right\rangle}{\sigma^3}.$$

Trimmed and winsorized skewness

The measure of sample skewness just given is highly sensitive to outliers.

This is so because it is based on sample means and standard deviations, and also because it involves cubic functions.

Quantile skewness

If the median is closer to the lower quantile than the corresponding upper quantile, the distribution is right-tailed and so there is a positive skew.

If it is closer to the upper quantile than the corresponding lower one, the distribution is left-tailed and there is a negative skew.

Nonparametric skewness

The following function is known as the nonparametric skewness measure

$$\text{skew} = \frac{\bar{x} - m}{s},$$

Measures of kurtosis

Kurtosis is better understood as relating to the heaviness of a distribution's tails.

In a random variable X, kurtosis is defined as the fourth standardized moment

Quantile-based measures of kurtosis

We can use quantiles to allow us to calculate robust estimates of kurtosis.

One simple procedure is to calculate the ratio of the 95% (or 99%, etc.) inner quantile interval to the interquartile range.

Graphical exploration of univariate distributions

There are valuable graphical methods for exploring univariate distributions.

These ought to be seen as complementary to the quantitative methods described above

Graphical exploration of univariate distributions

Stem-and-leaf plots

Histograms

Boxplots

Q-Q and P-P plots