Data Analytics for Social Science
Data inspection and visualisation

Johan A. Elkink

School of Politics & International Relations
University College Dublin

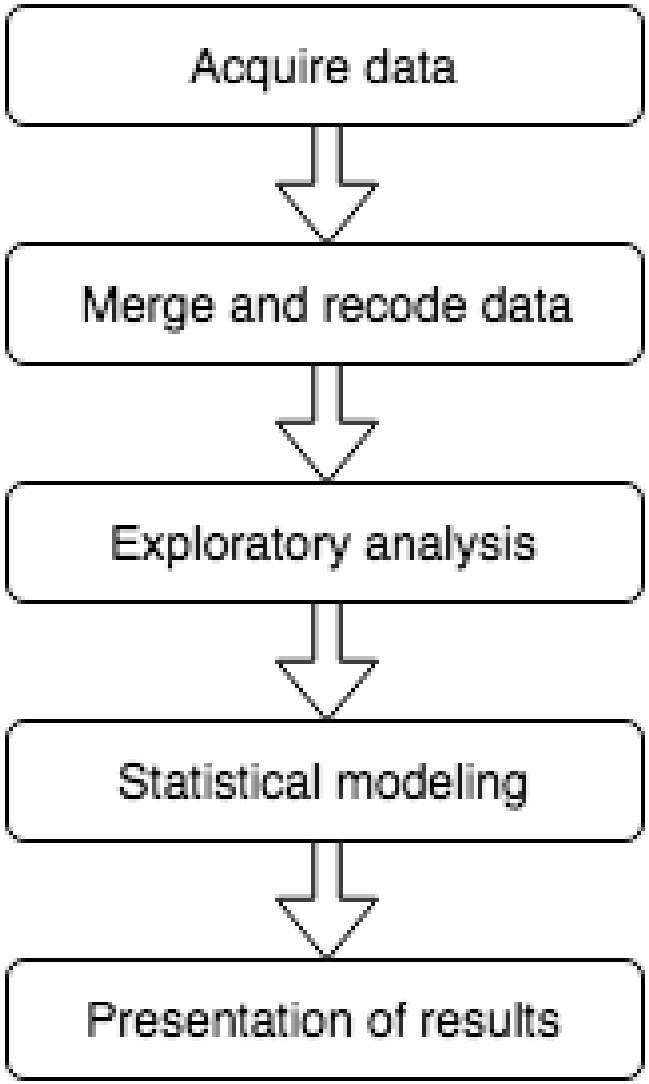
31 January 2017

Data analysis process

R at the cross-section of social science analysis and data science

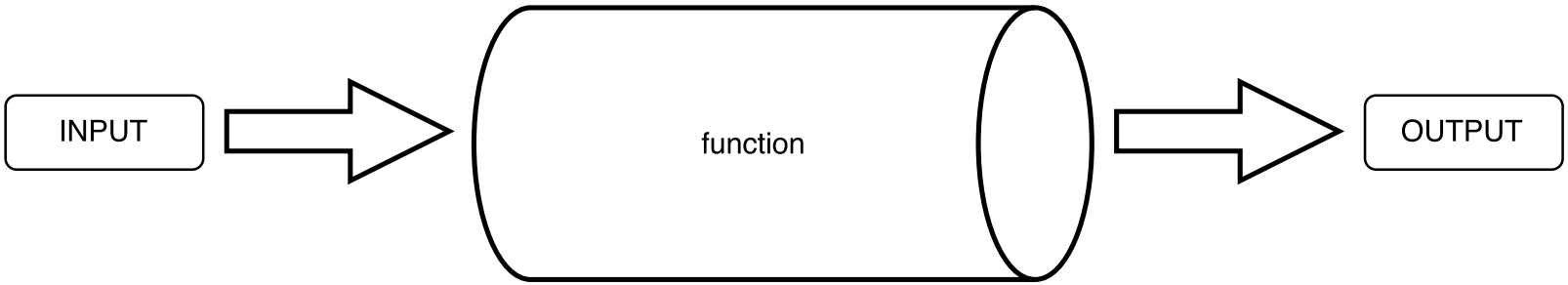
R can be extended using packages (video)

Data to be found in many formats, or entered using Excel (video)



Outline

Functions



Using functions

The below code creates a new function called "myMean", which takes one parameter "x" as input, and uses existing functions "sum" and "length" to calculate and return the mean (of "x").

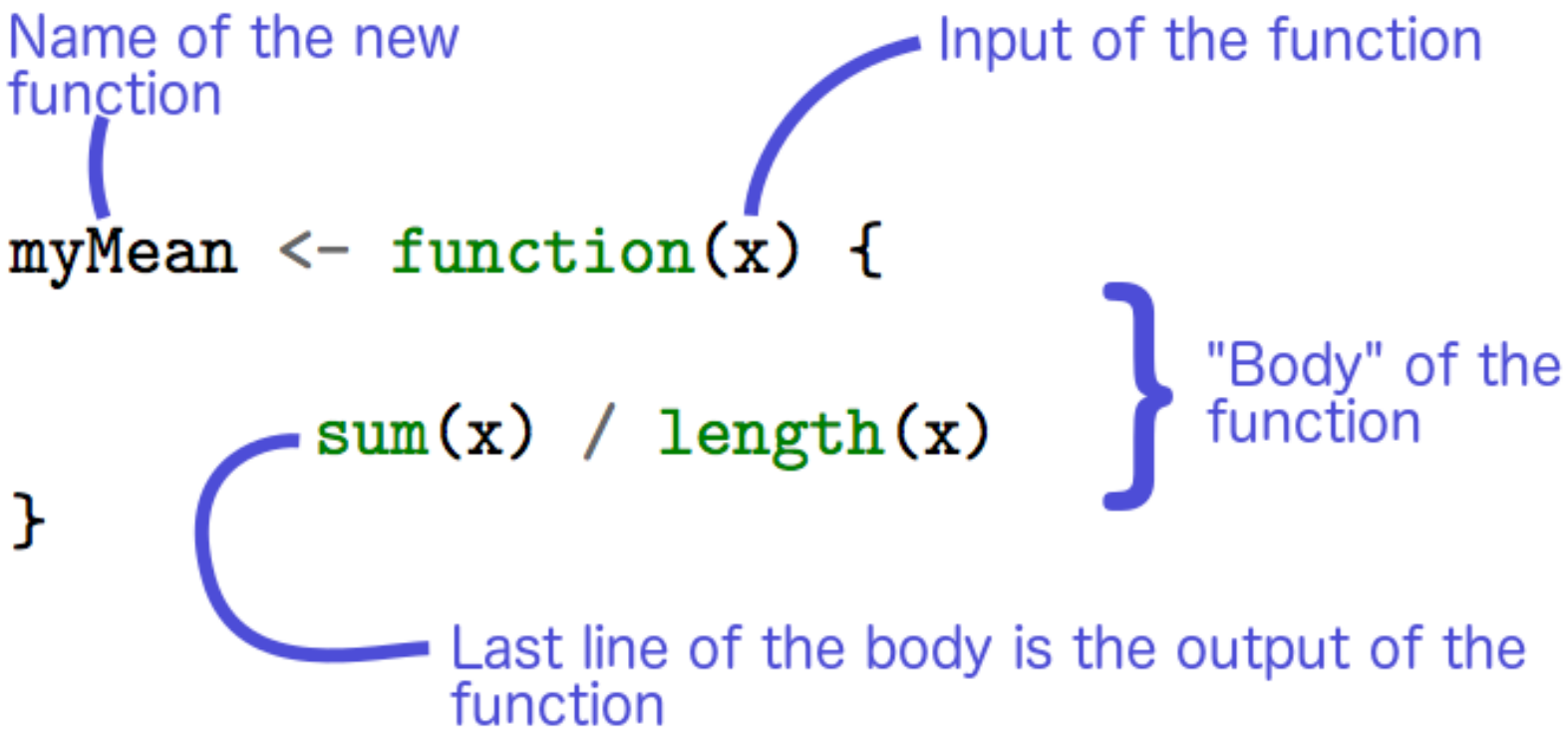
```
myMean <- function(x) {

        sum(x) / length(x)
}

a <- c(1, 4, 3, 5)
myMean(a)
```

The code then creates a vector "a" with 4 values,
$a = \begin{bmatrix} 1 & 4 & 3 & 5 \end{bmatrix}$
and then uses the function "myMean" to calculate the mean.

Defining functions



Calling functions

```
a <- c(1, 4, 3, 5)
```

This code calls function "c" (concatenate), passes as **input** to the function the numbers 1, 4, 3 and 5, and saves the **output** of the function to object "a".

```
myMean(a)
```

This code calls function "myMean", passes as **input** to the function the object "a", and does not save the output, so R will just **print** the output on the screen.

Outline

Measurement

"**Measurement** is the process of determining and recording which of the possible traits of a variable an individual case exhibits or possesses."

"A **case** is an entity that displays or possesses the traits of a given variable."

"A **population** is the set of all cases of interest. A **sample** is a subset of the **population**."

(Argyrous, 1997, 3–4)

Levels of measurement

Categorical	Nominal	categories
	Ordinal	... in particular order
Scale	Interval	... with meaningful distance
	Ratio	... with meaningful zero

(Argyrous, 1997, 11)

Levels of measurement

Categorical	**Nominal**	categories
	Ordinal	... in particular order
Scale	Interval	... with meaningful distance
	Ratio	... with meaningful zero

Example: UN membership (country), committee membership (MP)

(Argyrous, 1997, 11)

Levels of measurement

Categorical	Nominal	categories
	Ordinal	... in particular order
Scale	Interval	... with meaningful distance
	Ratio	... with meaningful zero

Example: education (voter, in level), Likert scale

(Argyrous, 1997, 11)

Levels of measurement

Categorical	Nominal	categories
	Ordinal	... in particular order
Scale	**Interval**	... with meaningful distance
	Ratio	... with meaningful zero

Example: left-right orientation (party, voters)

(Argyrous, 1997, 11)

Levels of measurement

Categorical	Nominal	categories
	Ordinal	... in particular order
Scale	Interval	... with meaningful distance
	Ratio	... with meaningful zero

Example: geographical distance, education (voter, in years), GDP per capita (country).

(Argyrous, 1997, 11)

Levels of measurement

Categorical	Nominal	categories
	Ordinal	... in particular order
Scale	Interval	... with meaningful distance
	Ratio	... with meaningful zero

"A **discrete variable** is measured by a unit that cannot be subdivided. It has a countable number of values."

"A **continuous variable** is measured by units that can be subdivided infinitely. It can take any value in a line interval."

(Argyrous, 1997, 11)

Levels of measurement: Why?

The level of measurement of a variable determines:

- What plots are appropriate to visualise the data.
- What tables can reasonably be produced.
- What statistics can be computed to summarize the data.
- What statistical analysis can be performed with the data.

Outline

Tables

Frequency table: A table of one variable, showing the number or proportion of cases in each category.

For categorical variables only.

Numerical equivalent of the barplots or piechart.

Cross table: A table of two or more variables, showing the number or proportion of cases in each combination of categories. Also: **contingency table** or **pivot table**.

For categorical variables only.

Similar to barplots by category, or stacked barplots.

Example: cross table

Table 4. Preference for freedom and equality in the US and Canada controlling for race, % (fictional data)

	White		Racial minorities	
	US	Canada	US	Canada
Freedom	75	60	60	58
Equality	25	40	40	42
Total, %	100	100	100	100

slides by Ivan Katchanovski, 2008

Outline

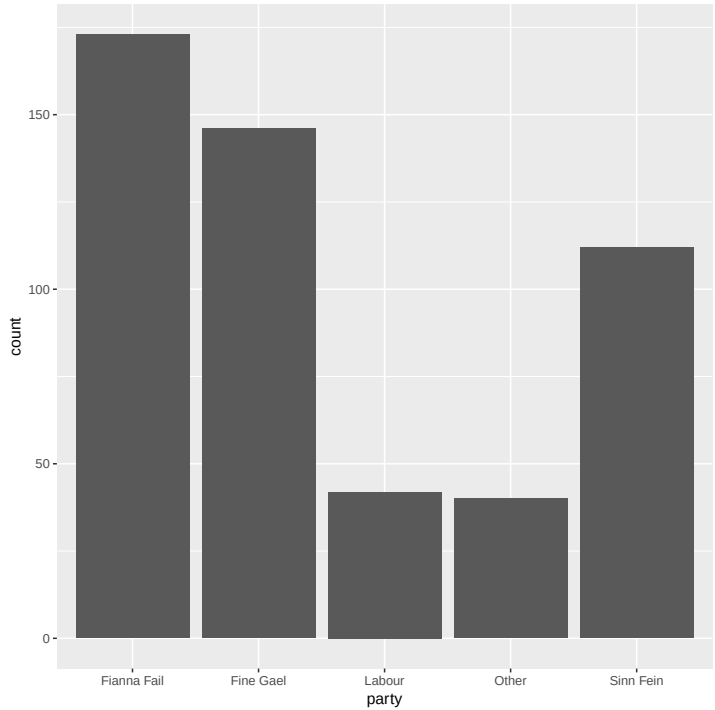
Main types

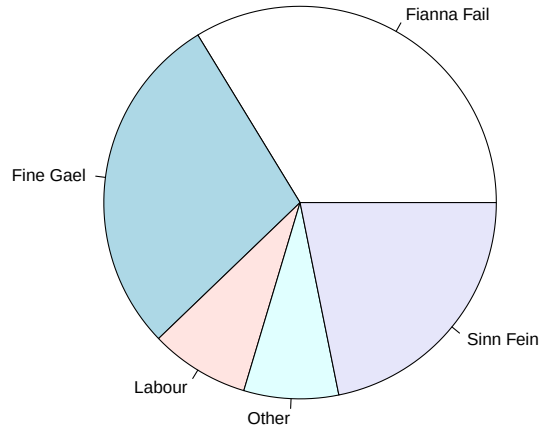
univariate	
Categorical	pie-charts
	barplots
Scale	histogram
	density plot
	boxplot
multivariate	
Scale by scale	scatterplot
Scale by categorical	boxplots

Categorical variables

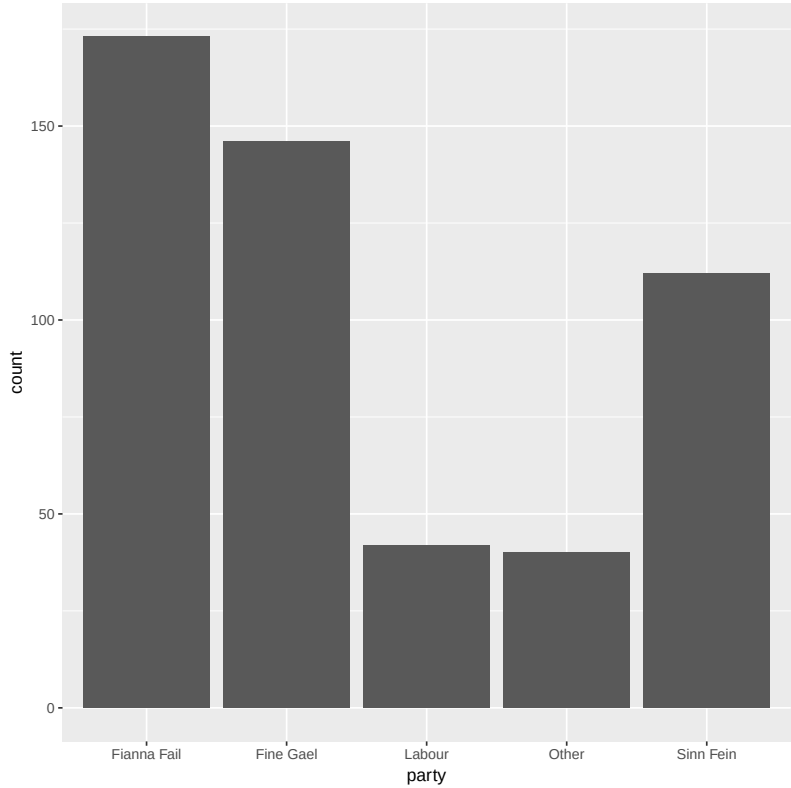
For categorical variables, it is often useful to look at the number of cases or the proportion of cases in a particular category.

Barplots and **pie charts** are useful for this.





Barplot



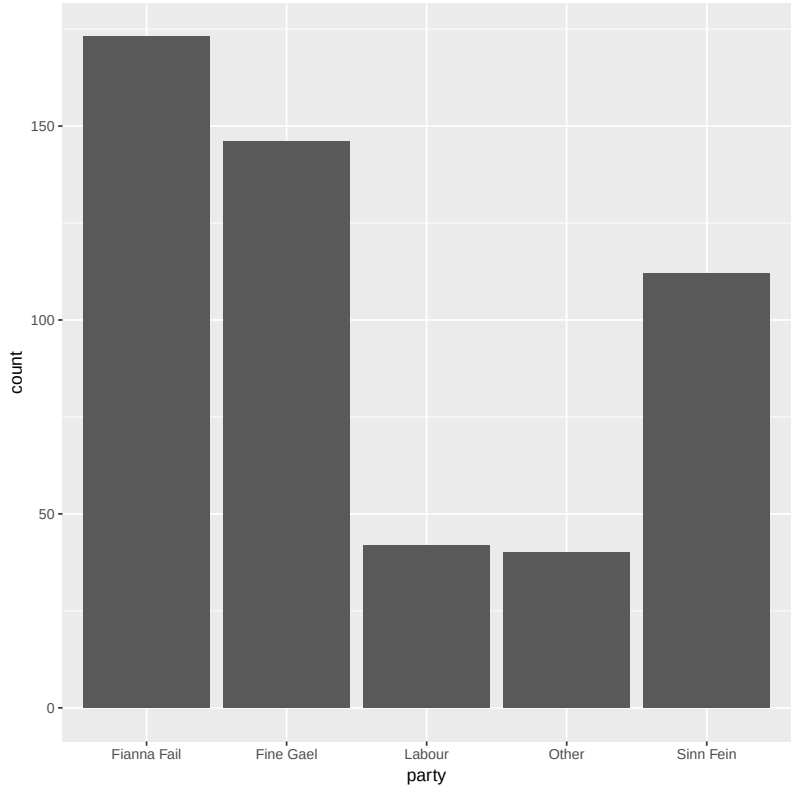

```
ggplot(ieVoters, aes(x = party)) + geom_bar()
```

Grammar of Graphics

"In brief, the grammar tells us that a statistical graphic is a **mapping** from data to **aesthetic** attributes (colour, shape, size) of **geometric** objects (points, lines, bars)" (Wickham, 2015, 5).

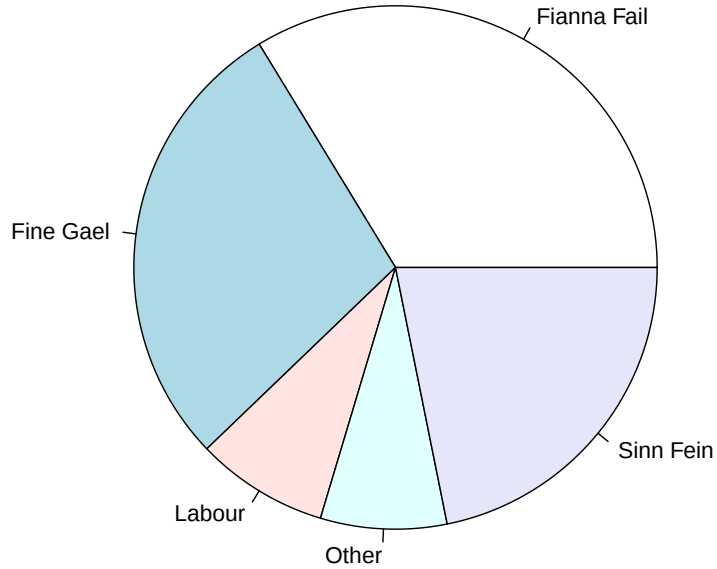
slides by Roger Peng

Barplot




```
ggplot(ieVoters, aes(x = party)) + geom_bar()
```

Pie chart



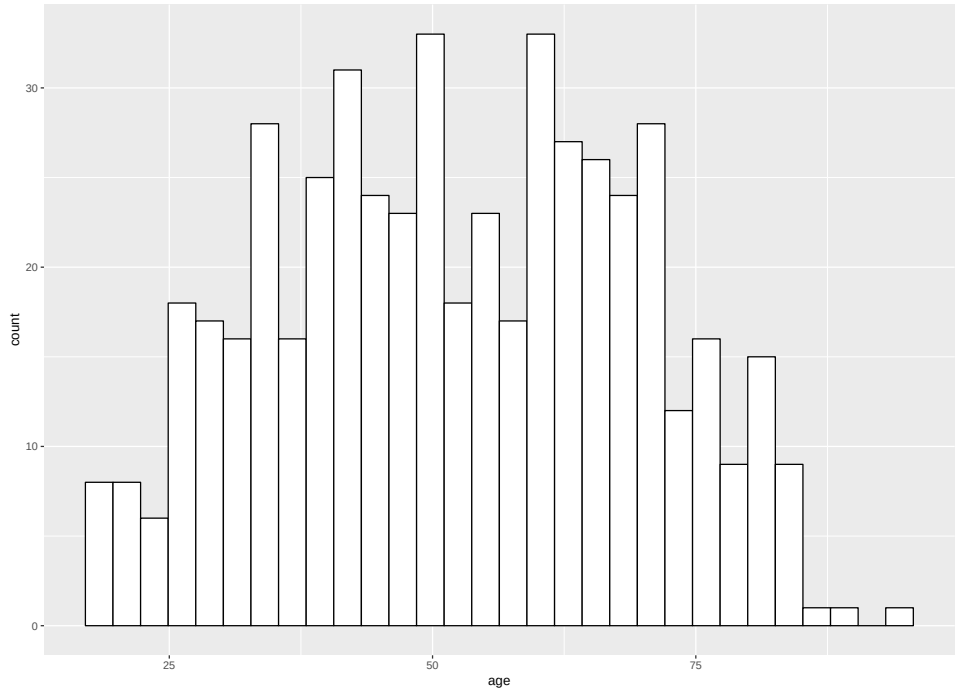

```
pie(table(ieVoters$party))
```

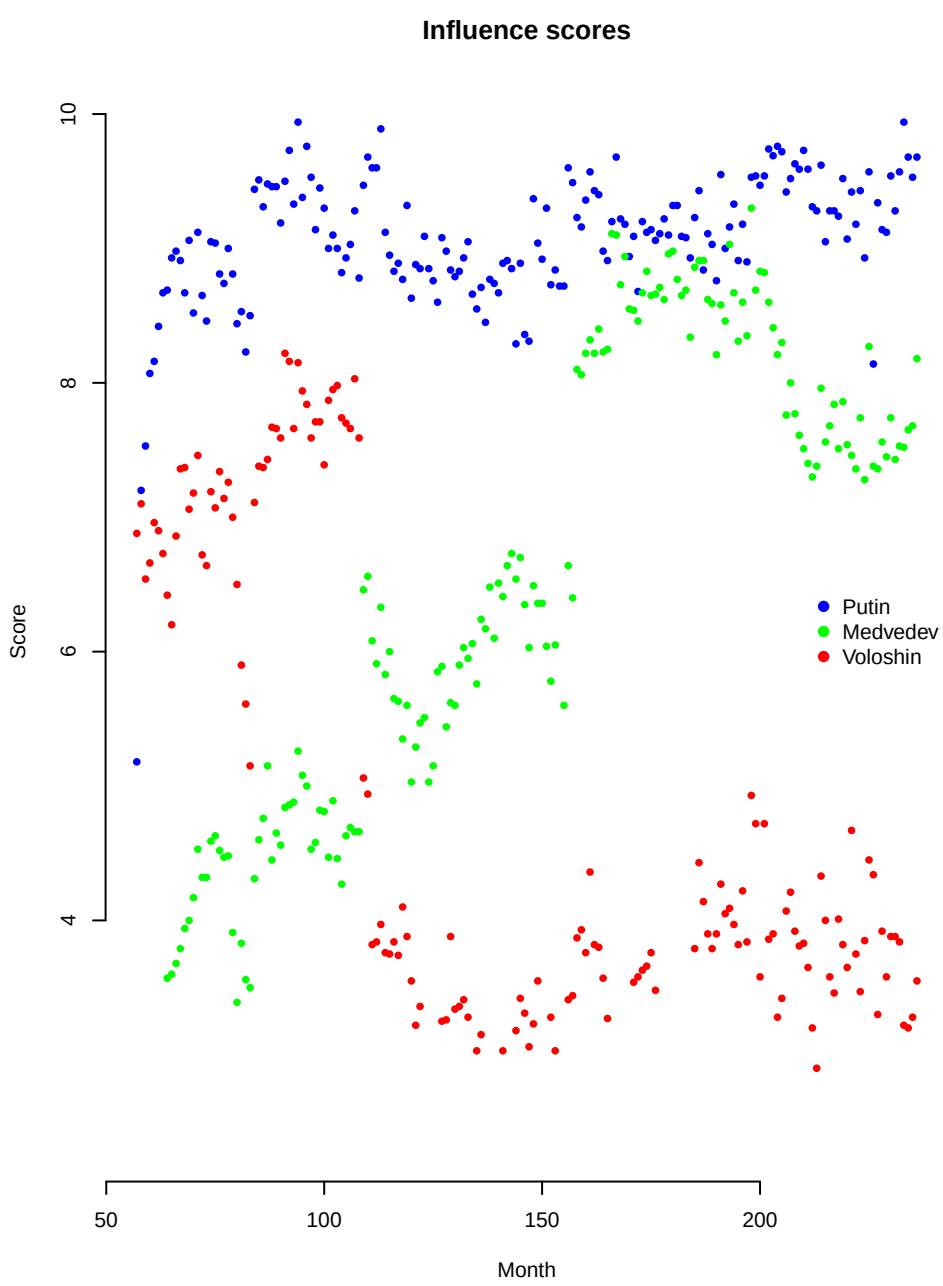
Histogram

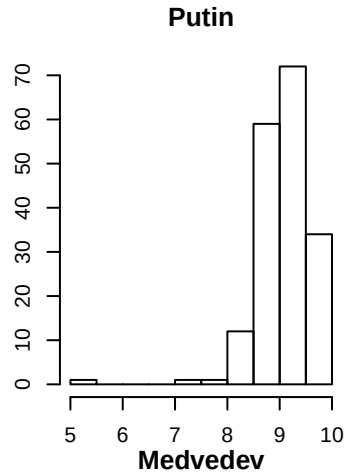
For continuous (or scale) variables, we often want to get an idea of the **distribution** of values.

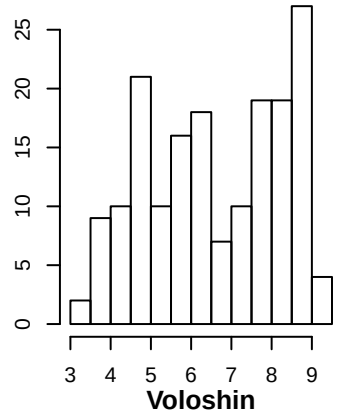
Histograms are useful to get an impression.

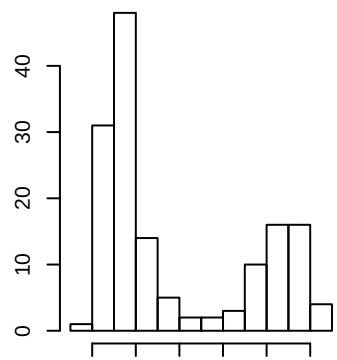
- bin the data using equal-distance cut-off points
- then produce a barplot of the number in each bin.





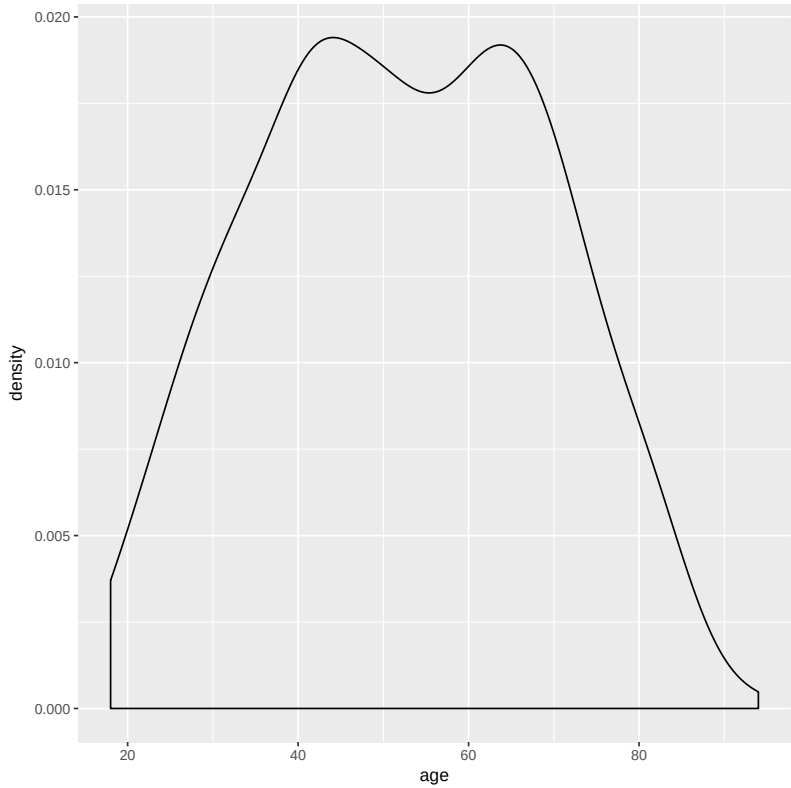






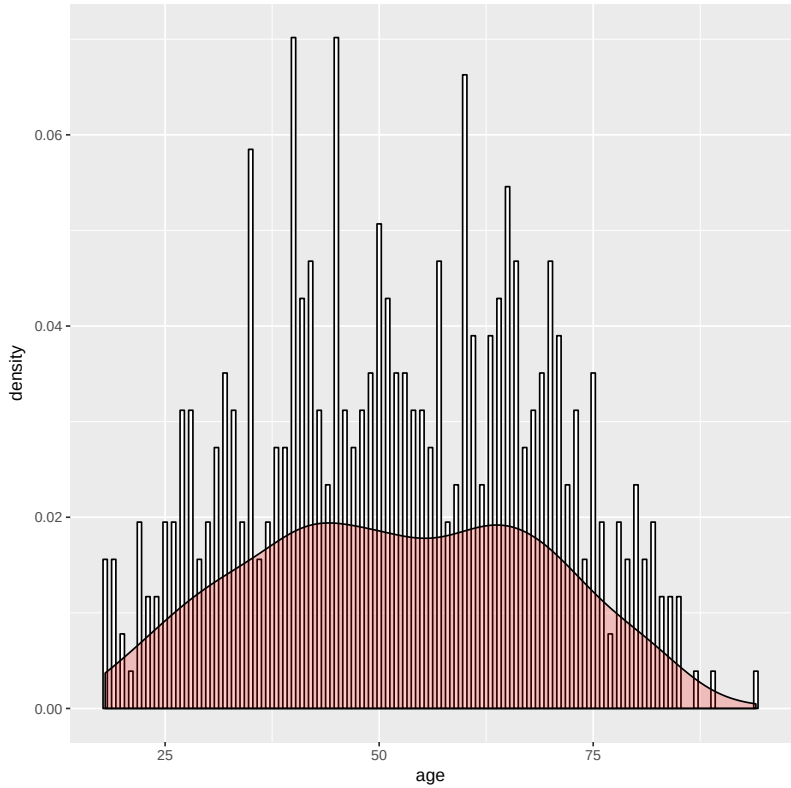
Density plot

A density plot is a smoothened version of a histogram, based on a non-parametric estimation of the shape of the distribution.

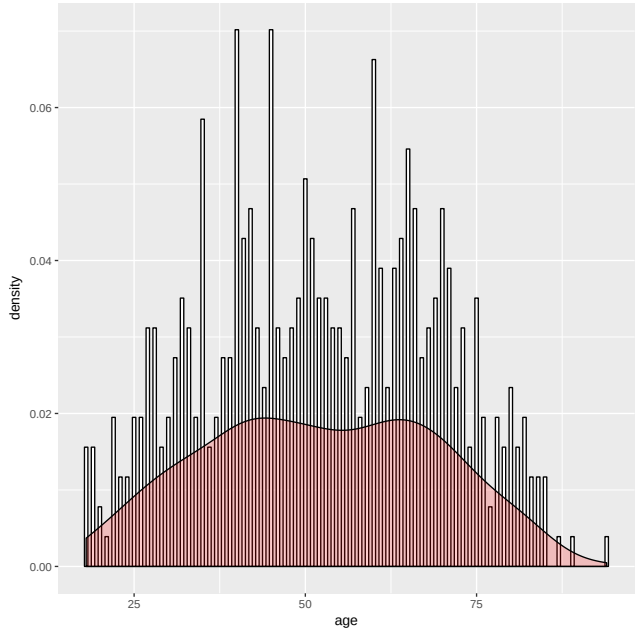


Density plot

A density plot is a smoothened version of a histogram, based on a non-parametric estimation of the shape of the distribution.



Density plot




```
ggplot(ieVoters, aes(x=age)) +
  geom_histogram(aes(y=..density..),
                 binwidth=.5,
                 colour="black", fill="white") +
  geom_density(alpha=.2, fill="red")
```

Distributions

For graphs of distributions (histogram, density plot, boxplot, etc.) you want to get an impression of:

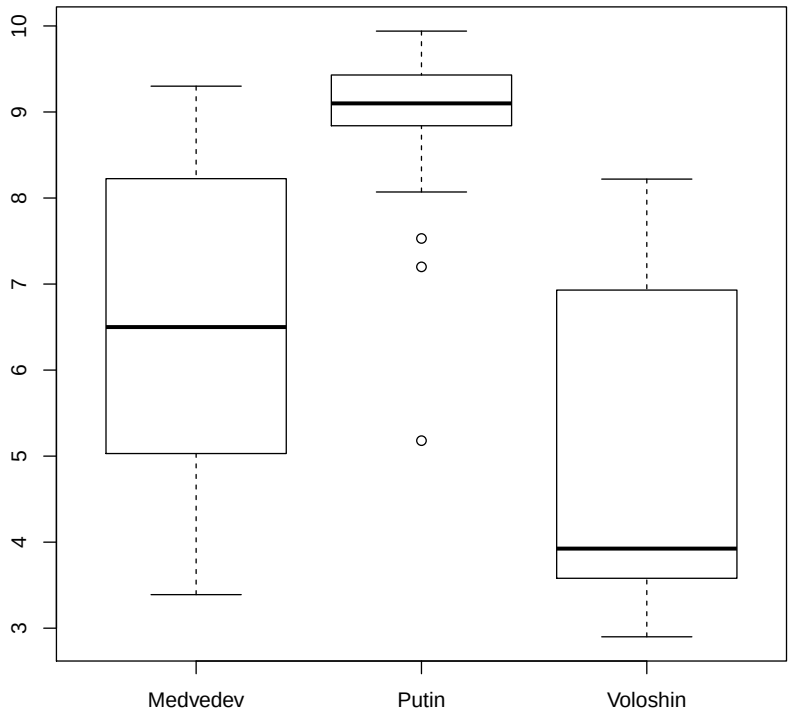
- the **shape** of the distribution;
- the **center** and **spread** of the distribution;
- the presence of **outliers**.

(Moore, 2003, 12)

Boxplot

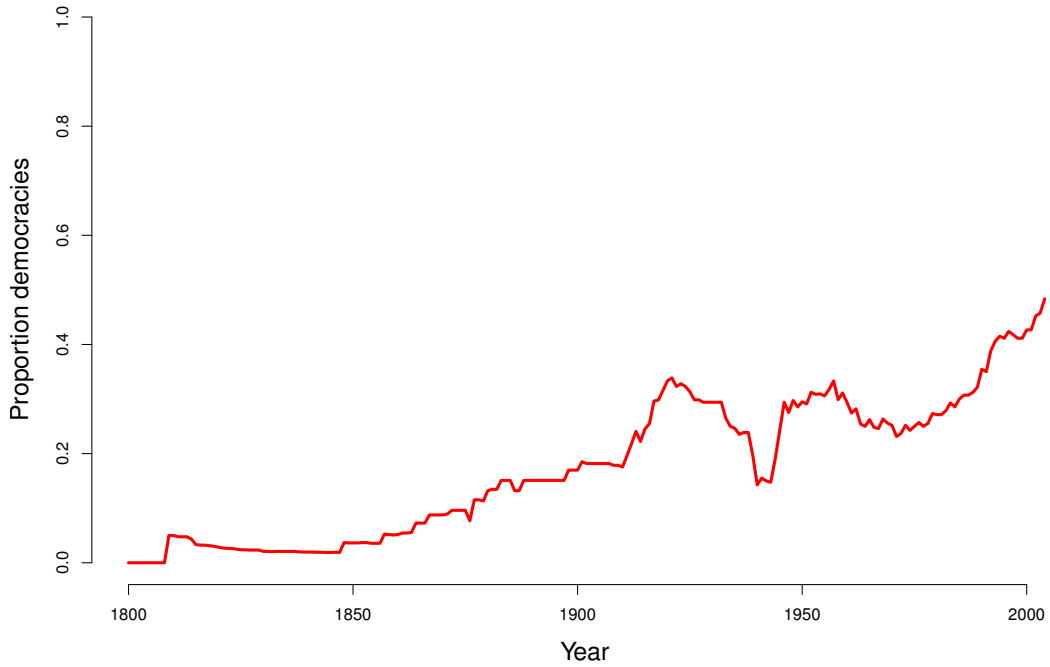
Another way of looking at the distribution of a continuous variable is to find out where the lowest 25% are located, where the lowest 50% are located, and where the top 25% are located.

A plot that shows this is the **boxplot**.



Time plot

When data is measured over time, another useful plot is a time plot, to see trends over time.



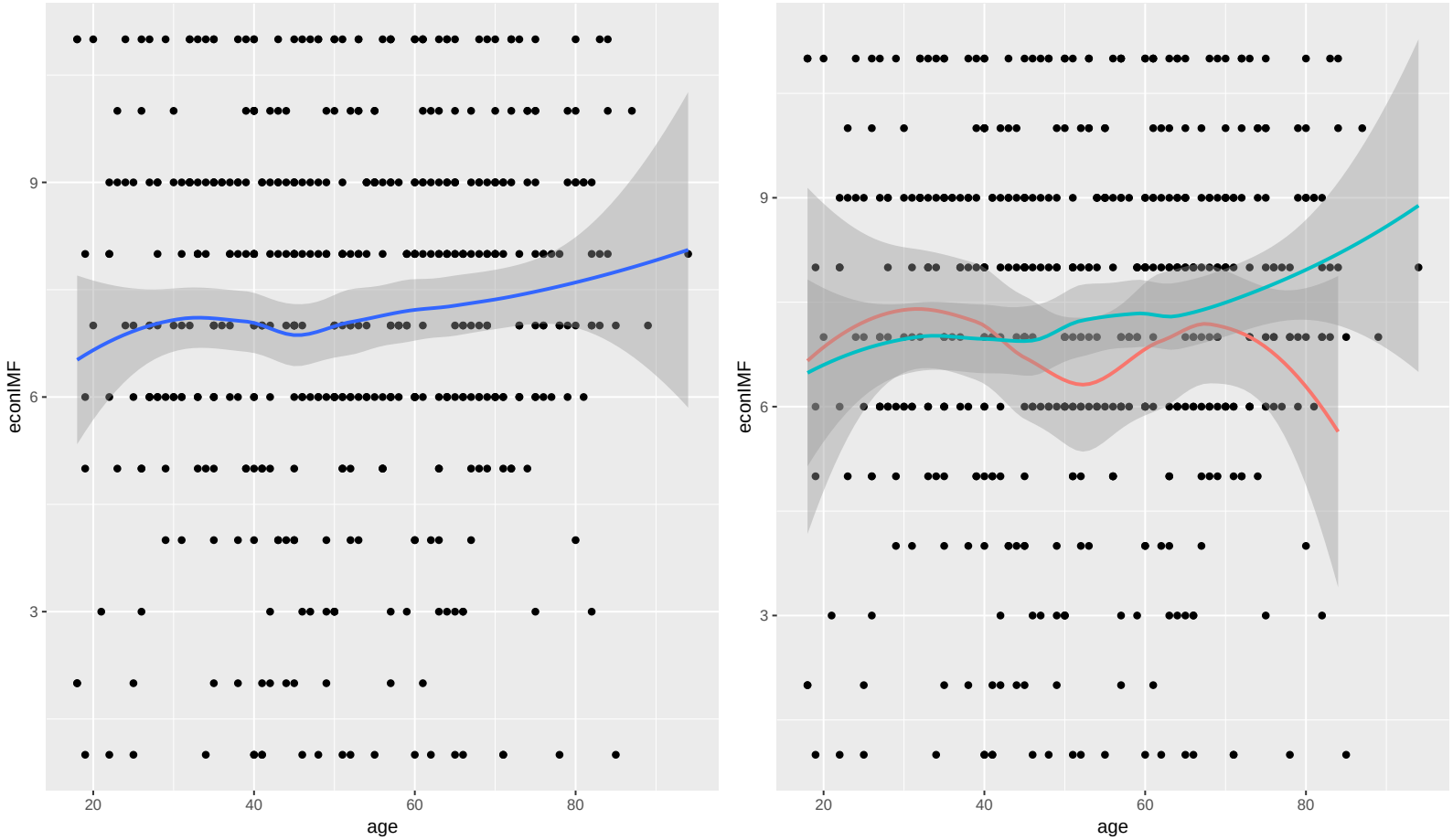
Polity IV (Marshall and Jaggers, 2002)

Basic principles

- Show comparisons

slides by Roger Peng, based on Tufte (2006)

Comparisons



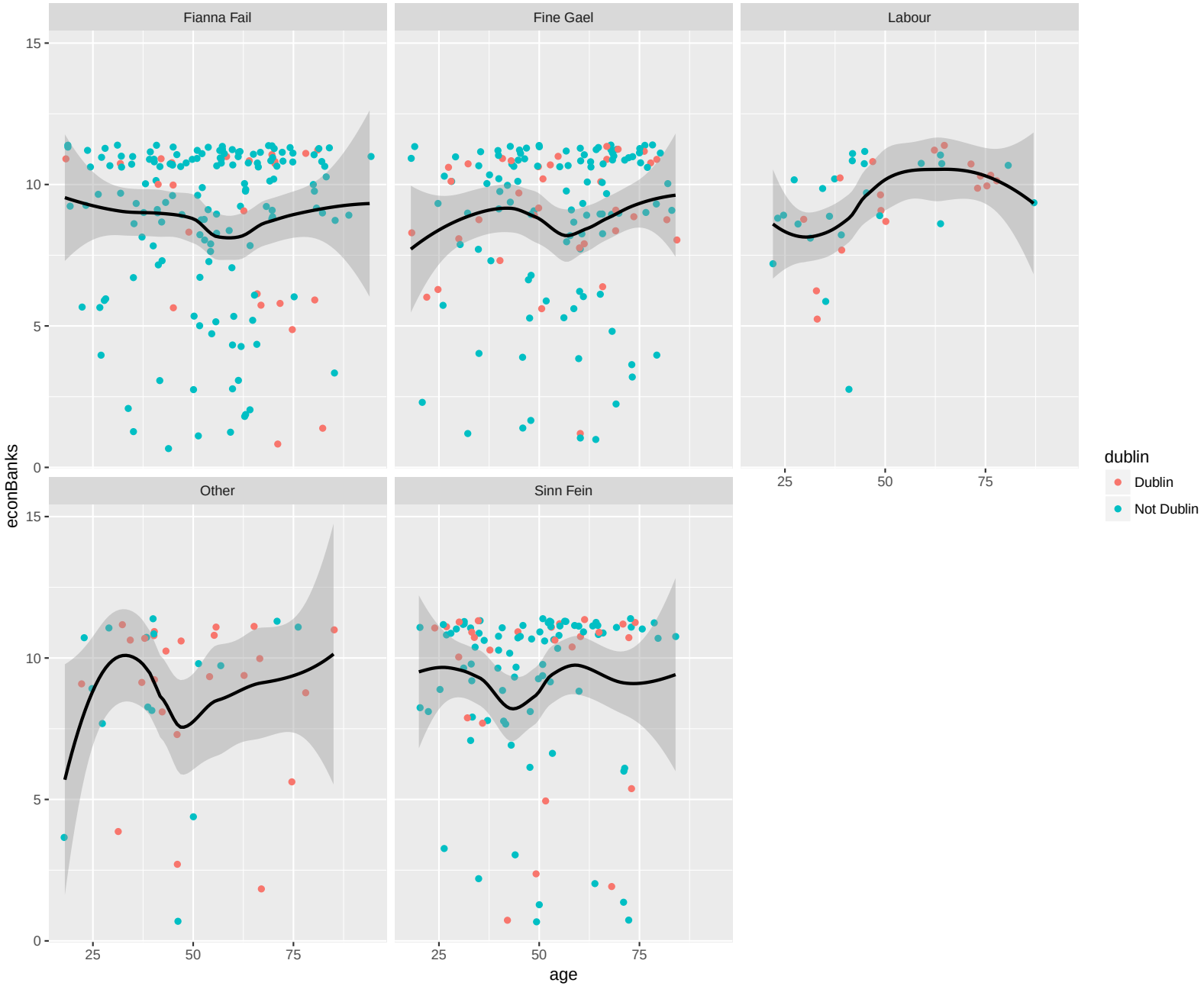
On the right, with different lines for Dublin (red) vs others (green), tells more of a story.

Basic principles

- Show comparisons
- Show causality, mechanisms
- Show multivariate data patterns

slides by Roger Peng, based on Tufte (2006)

Multivariate patterns

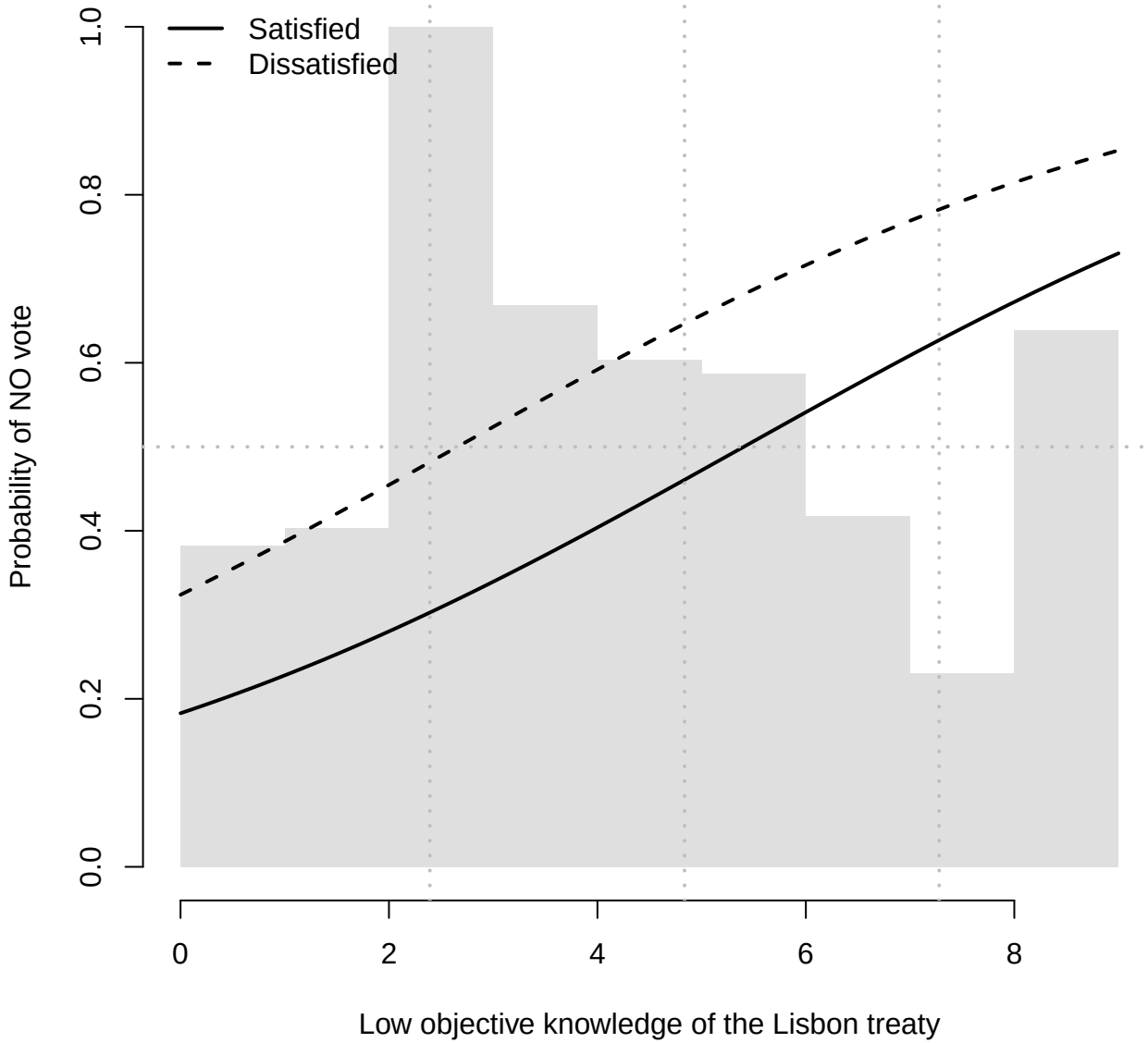


Basic principles

- Show comparisons
- Show causality, mechanisms
- Show multivariate data patterns
- Integrate multiple modes of evidence

slides by Roger Peng, based on Tufte (2006)

Example: knowledge and voting



Basic principles

- Show comparisons
- Show causality, mechanisms
- Show multivariate data patterns
- Integrate multiple modes of evidence
- Describe the data, sources, measurement scales, etc.
- In the end it stands or falls by the quality and relevance of the data

slides by Roger Peng, based on Tufte (2006)

Argyrous, George. 1997. *Statistics for social research*. Basingstoke: MacMillan.

Baturo, Alexander and Johan A. Elkink. 2014. "Office or Officeholder? Regime Deinstitutionalisation and Sources of Individual Political Influence." *Journal of Politics* 76(3).

Baturo, Alexander and Johan A. Elkink. 2015. "Dynamics of Regime Personalization and Patron-Client Networks in Russia, 1999–2014." *Post-Soviet Affairs* 32(1):75–98.

Marshall, M.G. and K. Jaggers. 2002. "Polity IV project: political regime characteristics and transitions, 1800-2002.".
URL: *http://www.bsos.umd.edu/cidcm/polity/*

Moore, David S. 2003. *The basic practice of statistics*. 3rd ed. New York: W.H. Freeman.

Tufte, Edward R. 2006. *Beautiful Evidence*. Graphics Press.

Wickham, Hadley. 2015. *ggplot2: Elegant Graphics for Data Analysis*. Springer.
URL: *http://ms.mcmaster.ca/∼bolker/misc/ggplot2-book.pdf*