Introduction to R

Antonia Santos

Program

PART I

Introduction to R and base R programming

PART II

Data manipulation

PART III

Data visualisation

PART IV

Introduction to modelling in R

Bibliography

- R Manual (`https://cran.r-project.org/doc/manuals/R-intro.html`)
- R for Data Science (2e) (`https://r4ds.hadley.nz`)
- Fundamentals of Data Visualisation (`https://clauswilke.com/dataviz/`)

Part III

Data visualisation

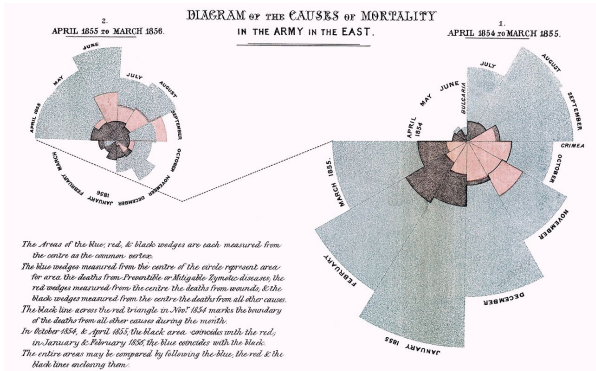
1. PRINCIPLES OF DATA VISUALISATION
2. PLOTTING IN BASE R
3. PLOTTING IN GGPLOT2

PRINCIPLES OF DATA VISUALISATION

- What is data visualisation?
- What are the benefits?
- A practical example: A graph tells a story.

PRINCIPLES OF DATA VISUALISATION

Florence Nightingale's Rose Diagram



PRINCIPLES OF DATA VISUALISATION

Florence Nightingale's Rose Diagram

Why does this chart tell a story?

- Context: It highlights a real problem (preventable deaths) within a specific context (the Crimean War).
- Clarity: The visualisation is simple and easy to understand, even for those without a background in statistics.
- Impact: The plot led to tangible changes (improvements in sanitary conditions).

PRINCIPLES OF DATA VISUALISATION

Fundamental Principles

- Clarity: Visualisations should be clear and easy to interpret.
- Accuracy: Represent the data correctly.
- Efficiency: Maximise information with minimal elements.
- Aesthetics: Make the chart visually appealing.
- Relevance: Only visualise data that is relevant to the message or story you are trying to tell.
- Choosing the right graph: Make sure you select an appropriate graph type for the data and insight you want to pass on.

PRINCIPLES OF DATA VISUALISATION

What to avoid?

- Confusing or cluttered charts.
- Inappropriate choice of chart type.
- Incorrect use of colours and scales.

PRINCIPLES OF DATA VISUALISATION

Misleading and bad visualisations

- Lets say we want to report on whether the number of crimes has increased over two selected years.

PRINCIPLES OF DATA VISUALISATION

Misleading and bad visualisations

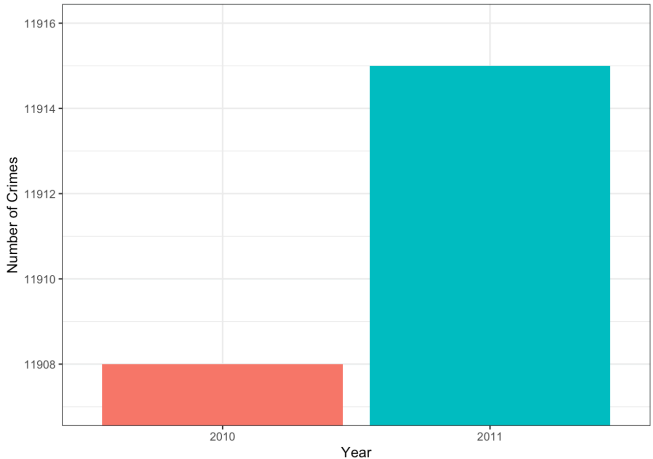


Figure: Example 1.

PRINCIPLES OF DATA VISUALISATION

Misleading and bad visualisations

- From the previous plot, it seems that the crime rate has jumped significantly from 2010 to 2011. However, can you notice anything suspicious about the plot?

PRINCIPLES OF DATA VISUALISATION

Misleading and bad visualisations

- From the previous plot, it seems that the crime rate has jumped significantly from 2010 to 2011. However, can you notice anything suspicious about the plot?
- Now, if we set the scale to start at zero we get:

PRINCIPLES OF DATA VISUALISATION

Misleading and bad visualisations

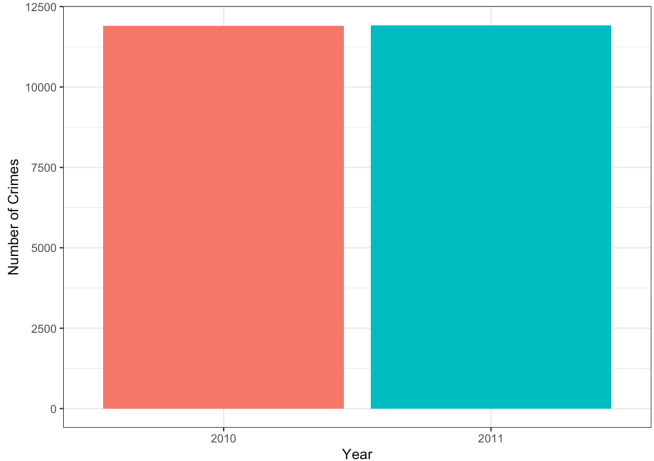


Figure: Example 1.

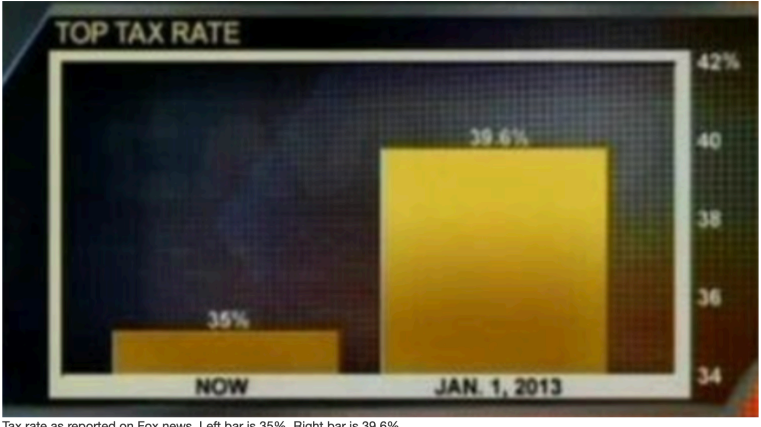
PRINCIPLES OF DATA VISUALISATION

Misleading and bad visualisations

- Now we can see that in reality, the crime rate has only increased marginally. This is a common tactic used in politics and the news when reporting. For example:

PRINCIPLES OF DATA VISUALISATION

Misleading and bad visualisations



Tax rate as reported on Fox news. Left bar is 35%. Right bar is 39.6%

Figure: Example 2.

PRINCIPLES OF DATA VISUALISATION

Misleading and bad visualisations

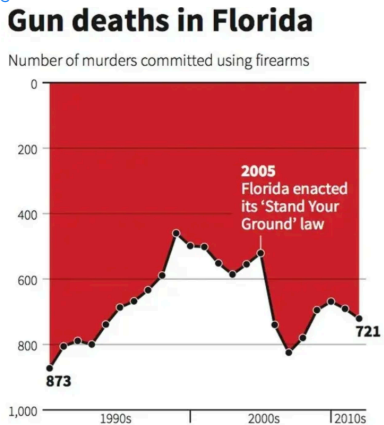


Figure: Example 3.

PRINCIPLES OF DATA VISUALISATION

Not All Bad

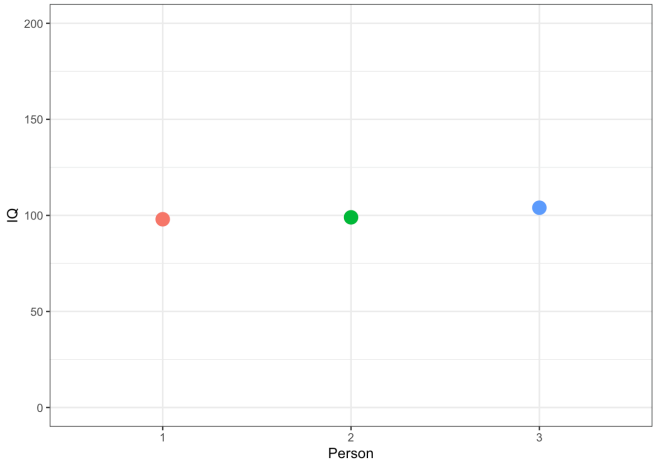


Figure: Example 4.

PRINCIPLES OF DATA VISUALISATION

Not All Bad

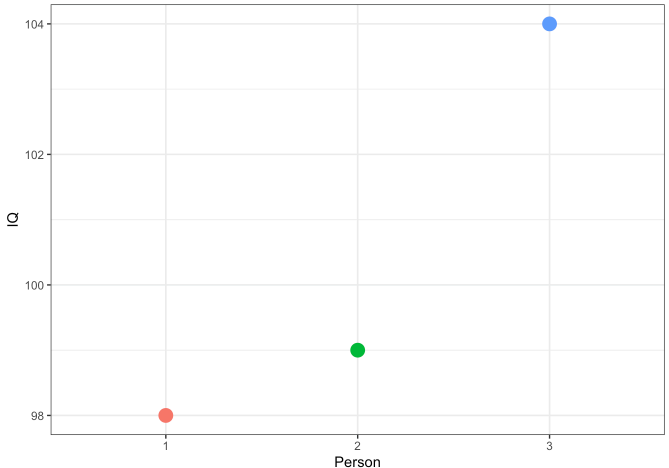


Figure: Example 4.

PRINCIPLES OF DATA VISUALISATION

Common visualisation types.

- Scatter plots: For visualising the relationship between two continuous variables.
- Bar plots: For comparing categorical data.
- Histograms: For displaying the distribution of a numeric variable.
- Line plots: For trends over time or continuous data.

DATA VISUALISATION

R base vs ggplot2

DATASETS

Diamonds Dataset

- Contains prices and attributes of over 50,000 diamonds.
- Variables:
  - carat (numeric): Weight of the diamond.
  - cut (factor): Quality of the cut (Fair, Good, Very Good, Premium, Ideal).
  - color (factor): Diamond color (D to J).
  - clarity (factor): Clarity measurement (I1, SI2, SI1, VS2, VS1, VVS2, VVS1, IF).
  - depth (numeric): Total depth percentage.
  - table (numeric): Width of the top relative to the widest point.
  - price (numeric): Price in USD.
  - x, y, z (numeric): Dimensions of the diamond.
- How to load it in R:

```
data("diamonds")
```

DATASETS

Airquality Dataset

- Daily air quality measurements in New York (May to September 1973).
- Variables:
  - Ozone (numeric): Ozone concentration (ppb).
  - Solar.R (numeric): Solar radiation (Langley).
  - Wind (numeric): Wind speed (mph).
  - Temp (numeric): Temperature (F).
  - Month (integer): Month (1 = January, 12 = December).
  - Day (integer): Day of the month.
- How to load it in R:

```
data("airquality")
```

DATASETS

Iris Dataset

- Measurements of 150 iris flowers from three species.
- Variables:
  - Sepal.Length (numeric): Sepal length (cm).
  - Sepal.Width (numeric): Sepal width (cm).
  - Petal.Length (numeric): Petal length (cm).
  - Petal.Width (numeric): Petal width (cm).
  - Species (factor): Species of iris (setosa, versicolor, virginica).
- How to load it in R:

```
data("iris")
```

DATASETS

mtcars Dataset

- Data on 32 cars from the 1974 Motor Trend magazine.
- Variables:
  - mpg (numeric): Miles per gallon (fuel efficiency).
  - cyl (numeric): Number of cylinders.
  - disp (numeric): Displacement (cubic inches).
  - hp (numeric): Horsepower.
  - drat (numeric): Rear axle ratio.
  - wt (numeric): Weight (1000 lbs).
  - qsec (numeric): Quarter-mile time (seconds).
  - vs (numeric): Engine type (0 = V-shaped, 1 = straight).
  - am (numeric): Transmission type (0 = automatic, 1 = manual).
  - gear (numeric): Number of forward gears.
  - carb (numeric): Number of carburetors.
- How to load it in R:

```
data("mtcars")
```

PLOTTING IN BASE R

Histograms

- A histogram helps visualise the distribution of a continuous variable.
- Let's create a histogram for the price of diamonds.

```
hist(diamonds$price,
main = "Histogram of Diamond Prices",
xlab = "Price",
col = "orange",
border = "black")
```

PLOTTING IN BASE R

Bar plots

- A bar plot visualises the frequency of categories in a factor variable.
- Let's create a bar plot for the cut variable, which represents the quality of the diamond's cut.

```
barplot(table(diamonds$cut),
main = "Bar Plot of Diamond Cut",
xlab = "Cut",
ylab = "Frequency",
col = "lightblue")
```

PLOTTING IN BASE R

Scatter Plot

- A scatter plot shows the relationship between two continuous variables.
- Let's create a scatter plot between carat (diamond size) and price.

```
plot(diamonds$carat, diamonds$price,
main = "Scatter Plot of Carat vs Price",
xlab = "Carat",
col = "blue",
pch = 19)
```

PLOTTING IN BASE R

Line Plots

- A line plot is ideal to use when you want to show trends over time, compare multiple series, and display relationships between variables.
- Let's plot the average price of diamonds by carat size.

```
x = plot(avg_price$carat, y = avg_price$price,
type = "l",
main = "Line Plot of Average Price by Carat",
xlab = "Carat",
ylab = "Average Price",
col = "blue",
lwd = 2)
```

PLOTTING IN BASE R

Box Plots

- A box plot is ideal to use when you want to summarise the distribution of data, identify outliers, and understand data variability.
- Let's create a boxplot.

```
boxplot(price   cut, data = diamonds,
main = "Boxplot of Diamond Prices by Cut",
xlab = "Cut",
ylab = "Price (USD)",
col = "lightblue",
border = "black")
```

PLOTTING IN BASE R

Your turn.

Question 1:

- *Create a histogram to visualize the distribution of temperature (Temp) in the airquality dataset.*
- *Customize the plot with appropriate colors, titles, and labels.*
- *Comment on what the histogram reveals about the distribution of temperatures.*

PLOTTING IN BASE R

Your turn.

Question 2:

- *Use the dplyr package to calculate the average ozone concentration (Ozone) by month (Month) in the airquality dataset.*
- *Create a bar plot to visualize the average ozone concentration for each month.*
- *Customize the plot with appropriate colors, titles, and labels.*
- *Comment on which months have the highest and lowest average ozone concentrations.*

PLOTTING IN BASE R

Your turn.

Question 3:

- *Create a scatter plot to explore the relationship between wind speed (Wind) and ozone concentration (Ozone) in the airquality dataset.*
- *Customize the plot with appropriate colors, titles, and labels.*
- *Comment on the observed relationship.*

PLOTTING IN BASE R

Your turn.

Question 4:

- *Create a scatter plot to explore the relationship between wind speed (Wind) and ozone concentration (Ozone) in the airquality dataset.*
- *Customize the plot with appropriate colors, titles, and labels.*
- *Comment on the observed relationship.*

PLOTTING IN BASE R

Your turn.

Question 5:

- *Use the airquality dataset to create a boxplot that compares the distribution of Ozone levels (Ozone) across different months (Month).*
- *Customize the boxplot to include:*
  - *A title: "Distribution of Ozone Levels by Month"*
  - *Axis labels: "Month" (x-axis) and "Ozone*
  - *Concentration (ppb)" (y-axis)*
  - *Different colors for each month's boxplot.*
- *Interpret the boxplot:*
  - *Which month has the highest median Ozone level?*
  - *Are there any outliers in the data? If so, in which months do they occur?*

PLOTTING IN GGPLOT2

Philosophy of ggplot2

- Grammar of graphics:
  - ggplot2 is based on the Grammar of Graphics, a systematic way to describe and build visualisations.
  - It breaks down plots into layers and components, making it highly flexible and consistent.
- Layered approach:
  - Plots are built step-by-step by adding layers (e.g., data, aesthetics, geometries, scales, themes).
  - Each layer can be modified independently, allowing for complex and customisable visualisations.

PLOTTING IN GGPLOT2

Basic Components of ggplot2

- Data: The dataset you want to visualise; Passed as the first argument to ggplot().
- Aesthetics (aes): Maps variables in the data to visual properties (e.g., x-axis, y-axis, color, size, shape).
- Geometries (geom_*): Defines the type of plot (e.g., scatter plot, bar plot, line plot).
- Scales: Control how variables are mapped to aesthetics (e.g., color scales, axis scales).
- Facets: Splits the data into subsets and creates multiple plots (small multiples).
- Themes: Controls the non-data elements of the plot (e.g., background, fonts, grid lines).
- Labels and Annotations: Adds titles, axis labels, and annotations to the plot.

PLOTTING IN GGPLOT2

Basic Components of ggplot2

- Consistency.
- Flexibility.
- Automatic Legends.
- Publication-Quality Plots.
- Faceting.
- Active Community.

PLOTTING IN GGPLOT2

Scatter Plot

- Let's create a scatter plot of carat vs price, similar to the base R example, but using ggplot2.

```
ggplot(data = diamonds, aes(x = carat, y = price)) +
geom_point(color = "blue") +
labs(title = "Scatter Plot of Carat vs Price", x = "Carat", y = "Price")
```

PLOTTING IN GGPLOT2

Your turn.

Question 6:

- *Using the iris dataset, create a scatter plot to analyse the relationship between Sepal.Length and Sepal.Width. Colour the points by Species.*

PLOTTING IN GGPLOT2

Bar Plot

- Let's recreate the bar plot of the cut variable using ggplot2.

```
ggplot(data = diamonds, aes(x = cut)) +
geom_bar(fill = "lightblue") +
labs(title = "Bar Plot of Diamond Cut", x = "Cut", y = "Frequency")
```

PLOTTING IN GGPLOT2

Your turn.

Question 7: *Using the iris dataset, create a bar plot to display the average Petal.Length for each species. Colour the bars by Species.*

PLOTTING IN GGPLOT2

Histogram

- Let's create a histogram of price using ggplot2.

```
ggplot(data = diamonds, aes(x = price)) +
geom_histogram(binwidth = 1000, fill = "orange", color = "black") +
labs(title = "Histogram of Diamond Prices", x = "Price", y = "Count")
```

PLOTTING IN GGPLOT2

Your turn.

Question 8: *Using the iris dataset, create a histogram to visualise the distribution of Sepal.Length. Use different colours to represent each Species in the dataset.*

PLOTTING IN GGPLOT2

Line plots

- Let's create a line plot using the data airquality

```
ggplot(data = avg_price, aes(x = carat, y = price)) + geom_line() + labs(title = "Average Price by Carat (Faceted by Cut)", x = "Carat", y = "Average Price (USD)")
```

PLOTTING IN GGPLOT2

Your turn.

Question 9: *Using the airquality dataset, create a line plot to visualise the trend of Ozone levels over the days of the month. Separate the lines by Month so that each month's trend is clearly visible.*

PLOTTING IN GGPLOT2

Box plots

- Let’s create a line plot of price using ggplot2.

```
ggplot(airquality, aes(x = Month, y = Ozone, color = factor(Month))) +
geom_boxplot() +
labs( x = ”Month”, y = ”Ozone Levels (ppb)”, color = ”Month” )
```

PLOTTING IN GGPLOT2

Your turn.

Question 10: *Using the iris dataset, create a box plot to compare the distribution of Petal.Width across the three Species. Ensure the plot includes:*

- *Different colours for each species.*
- *Proper axis labels and a descriptive title.*

PLOTTING IN GGPLOT2

Customising Plots

- Adding titles, axis labels, and captions;
- Adjusting themes;
- Modifying scales;
- Faceting for subplots;
- Adding annotations;

PLOTTING IN GGPLOT2

Your turn

- Use the dplyr package to summarise the data:
  - Calculate the mean Ozone levels and mean Wind speed for each Month.
  - Include the number of observations (n) in each month.
- Create a scatter plot using ggplot2 to visualise the relationship between Wind and Ozone:
  - Plot Wind on the x-axis and Ozone on the y-axis.
  - Use different colours for each Month to distinguish them.
  - Add a regression line.
- Use facet_wrap to create individual scatter plots for each month to better visualise monthly trends.

Thank you!