

# Exploratory Data Analysis

Daniel Lawson University of Bristol

Lecture 01.2 (v2.0.0)

# Signposting

This Lecture on Exploratory Data Analysis is split into two short parts:

- ▶ Slides covering the (few) abstract notions
- ▶ An RStudio session covering the details

# Intended Learning Outcomes

- ▶ ILO1 Be able to **use and apply basic machine learning** tools.
- ▶ ILO2 Be able to **make and report appropriate inferences** from the results of applying basic tools to data.

# Dataset and getting started

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

Should we at least find out what the range of each variable is?

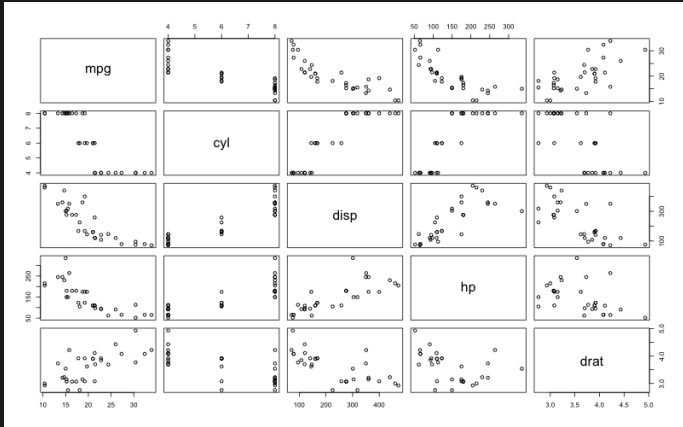
```
> apply(mtcars,2,range)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
[1,]	10.4	4	71.1	52	2.76	1.513	14.5	0	0	3	1
[2,]	33.9	8	472.0	335	4.93	5.424	22.9	1	1	5	8

(luckily for us, the data are all numeric!)

## Initial plot

```
> pairs(mtcars[,1:5])
```



# Summaries of distributions

- ▶ Important **positional summaries**:
  - ▶ Mean (`mean(x)`)
  - ▶ Median (`median(x)`)
  - ▶ Weighted Mean (`weighted.mean(x,w)`)
- ▶ Important additional summaries:
  - ▶ Sample variance (`var(x)`)
  - ▶ Sample standard deviation (s.d.) (`sd(x)`)
  - ▶ Quantiles  
(`quantile(x, probs=c(0.05,0.25,0.5,0.75,0.95))`)

# Summary and boxplots

The *five number summary* shows:  $(\min, Q_1, Q_2, Q_3, \max)$

- ▶ **Outliers:**

- ▶ can be defined with respect to the Normal distribution.
- ▶ Define the interquartile range  $IQR = Q_3 - Q_1$ .
- ▶ **outliers** as those observations at least  $3/2IQR$  above  $Q_3$  or below  $Q_1$ .
- ▶ This is just a heuristic for exploratory data analysis.

## Summary and boxplots (2)

```
> summary(mtcars[,1:5])
```

mpg	cyl	disp	hp
Min. :10.40	Min. :4.000	Min. : 71.1	Min. : 52
1st Qu.:15.43	1st Qu.:4.000	1st Qu.:120.8	1st Qu.: 91
Median :19.20	Median :6.000	Median :196.3	Median :123
Mean :20.09	Mean :6.188	Mean :230.7	Mean :146.7
3rd Qu.:22.80	3rd Qu.:8.000	3rd Qu.:326.0	3rd Qu.:181
Max. :33.90	Max. :8.000	Max. :472.0	Max. :335



# Standardization

- ▶ **Standardized variables**  $z_i$  are commonly defined from data  $x_i$  using the **sample mean**  $\bar{x}$  and the **sample s.d.**  $\hat{s}_x$ :

$$z_i = \frac{x_i - \bar{x}}{\hat{s}_x}$$

- ▶ The standardized variables have mean 0 and s.d. 1.
- ▶  $z_i$  is also called the standard score, z-value, z-score, and the normal score.
- ▶ An individual z-score  $z_i$  gives the number of standard deviations an observation  $x_i$  is from the mean.
- ▶ The standardized score has no units.

# Can you guess the output of:

```
> summary(scale(mtcars))
```

# Standardization against a reference

- ▶ In machine learning, we often use a **training** set, and a **test** set. It is essential that **both are standardized against the training data**:

$$z_i = \frac{x_i - \bar{x}_{train}}{\hat{s}_{train}}$$

- ▶ Test data may **not have** mean (close to) 0 and s.d. (close to) 1.

# Types of Data

- ▶ **Quantitative Variables**

- ▶ Quantitative variables are those for which arithmetic operations like addition and differences make sense.
- ▶ Another name for quantitative variables is **features**.

- ▶ **Categorical Variables**

- ▶ Categorical variables partition the individuals into classes.
- ▶ Other names for categorical variables are levels or **factors**.

# Further Types of Data

- ▶ Later we'll cover more complex data types, including:
  - ▶ relational tables
  - ▶ graphs
  - ▶ images
  - ▶ text
- ▶ This basic Exploratory Data Analysis still applies then, but to summaries:
  - ▶ Counts of nodes, edges
  - ▶ Tree depths
  - ▶ corpus size
  - ▶ etc

## Categorical variables: Table

The most straightforward summary for categorical variables is to count them.

```
table(mtcars[, "gear"])  
## from ?mtcars :  
# gear  Number of forward gears
```

Var1	Freq
3	15
4	12
5	5

# Two-way Table

Relationships between two categorical variables can be shown through a **two-way table** or **contingency table** (also known as cross tabulation):

```
table(mtcars[,c("vs", "gear")])  
#   vs      Engine (0 = V-shaped, 1 = straight)
```

	3	4	5
0	12	2	4
1	3	10	1

# Types of plot

Some essential plots include<sup>1</sup>:

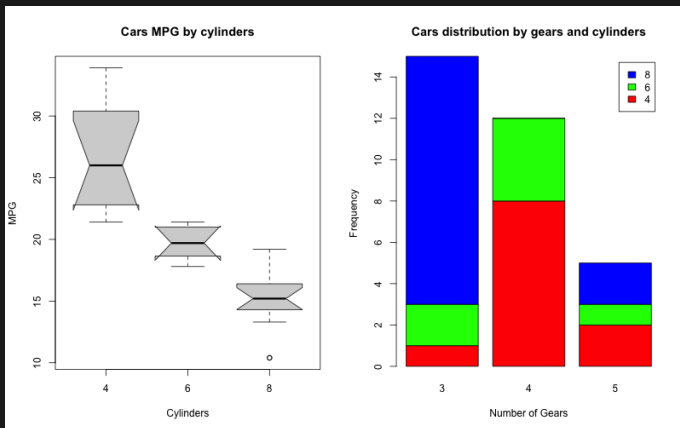
- ▶ Bar Chart
  - ▶ Segmented Bar Chart
- ▶ Heatmap
  - ▶ Highlight table
- ▶ Histograms
  - ▶ Kernel Density estimates
- ▶ Cumulative Distribution Functions

---

<sup>1</sup>Know what these are **for**. Applies to all plot we use in the course.

# Boxplot example

```
combined = table(mtcars$cyl, mtcars$gear)
boxplot(mpg~cyl,data=mtcars,notch=TRUE,...)
barplot(combined,...)
```





# Empirical Cumulative Distribution Function

- ▶ The **empirical cumulative distribution** function:

$$F_X(x) = Pr(X \leq x),$$

- ▶ is, for a continuous RV:

$$F_X(x) = \int_{-\infty}^x f_X(t) dt$$

- ▶ where  $f_X(t)$  is the density function of the Random Variable  $X$ .
- ▶ For a discrete RV

$$F_X(x) = \sum_{x_i \leq x} x_i$$

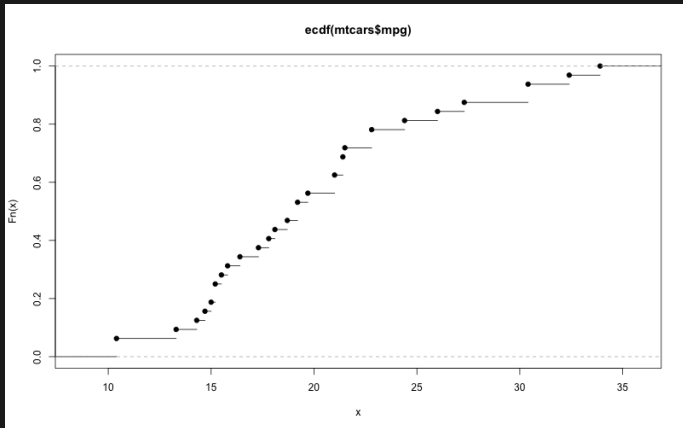
# Empirical Cumulative Distribution Function

To create a graph of the empirical cumulative distribution function:

- ▶ **Sort the observations** from smallest to largest
- ▶ Next **match these up** with the integral multiples of the  $1$  over the number of observations
- ▶ Display it with the correct **type of line**.

# ECDF

```
ecdf(mtcars$mpg)
```



# Cumulative Distribution Function for categorical data

- ▶ Categorical data have a **natural ordering** too: by frequency. This allows the creation of key concepts such as  $P(X < x)$ .
- ▶ It is often useful to establish natural orderings, which may exist in other settings.
- ▶ One example is ordinal data.

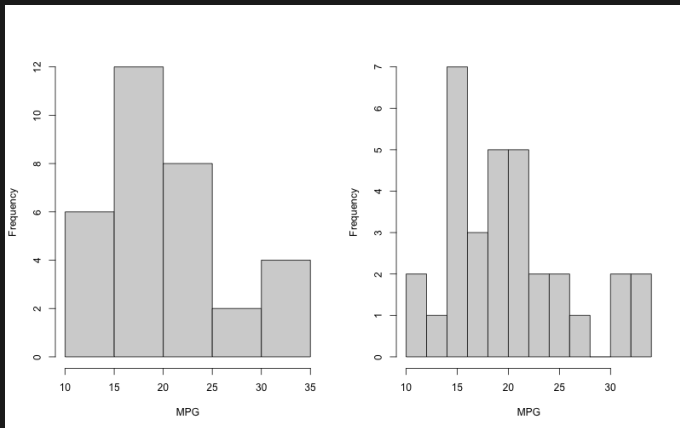
# Survival Function

- ▶ It is sometimes more convenient to work with the **fraction of samples that are larger than some value**.
- ▶ The **survival function**  $S_X$  is trivially related to the ECDF:

$$S_X(x) = Pr(X > x) = 1 - F_X(x)$$

# Histograms

- ▶ Histograms are a common visual representation of a quantitative variable. Histograms visual the data using **rectangles of area** to display frequencies and proportions.
- ▶ **It is critical that bins are comparable.** Many comparisons are impossible if bins are poorly chosen.



# Scatterplots

- ▶ **Scatterplots** show the relationship for **pairs of observations**.
- ▶ The values of the first variable

$$\{x_1, \dots, x_n\}$$

are often assumed known.

- ▶ They are often called **explanatory**, predictor, or descriptor variables, and are displayed on the horizontal axis.
- ▶ The values of the second variable

$$\{y_1, \dots, y_n\}$$

are viewed as observations with input  $\{x_1, \dots, x_n\}$ .

- ▶ Called the **response** variable, they are displayed on the vertical axis.

# Interpretation

Interpret plots considering:

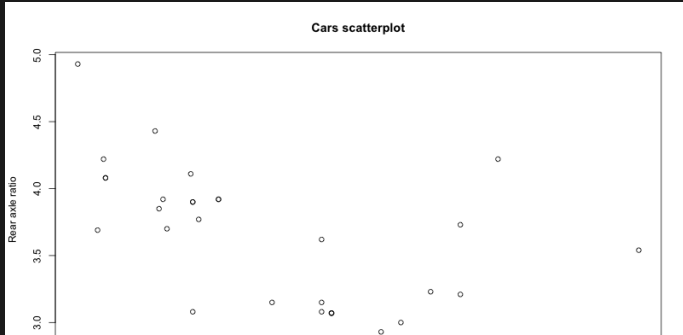
- ▶ the overall **pattern**
- ▶ the **center**
- ▶ the **spread**
- ▶ the **shape** (symmetry, skewness, peaks)
- ▶ and **deviations** from the pattern
- ▶ **outliers**
- ▶ **gaps**



# Scatterplots

In describing a scatterplot, take into consideration

- ▶ positive or negative association/**trend**
- ▶ **intercept**
- ▶ **clusters**
- ▶ the **form**, for example,
  - ▶ linear
  - ▶ curved relationships
  - ▶ (uni/multi)modal conditional distributions
- ▶ magnitude of the **noise**



## Further reading

- ▶ **R for Data Science** by Hadley Wickham and Garrett Grolemund is an excellent resource!
- ▶ It uses R tidyverse. You don't have to, but look into it.
- ▶ EDA is an **art** not a science. There is no **right** way to do it.
- ▶ You should be proactive in exploring solutions that others use and keep experimenting to find a better way to represent the data.

# Reflection

By the end of the course, you should:

- ▶ Be able to describe basic tools of EDA
- ▶ Be able to suggest appropriate EDA for a wide variety of data
- ▶ Be able to spot mistakes in an analysis from EDA plots
- ▶ Have practical experience to draw on to go beyond simple examples
- ▶ **However**, EDA is not proscriptive. Only general ideas are essential.

# Signposting

- ▶ The Workshop Lecture 1.3.1 demonstrate these features.
- ▶ There are further workshops on background: working with RStudio, setting up a Data Science environment with GitHub, and understanding the Assessments.
- ▶ There are text notes and links in the Coursebook.
- ▶ Block 02 covers **Regression and correlations** where we say something more rigorous about the relationship between variables.