R for Data Science (I): Exploration

Alex Sanchez, Miriam Mota, Ricardo Gonzalo and Mireia Ferrer

Statistics and Bioinformatics Unit. Vall d'Hebron Institut de Recerca

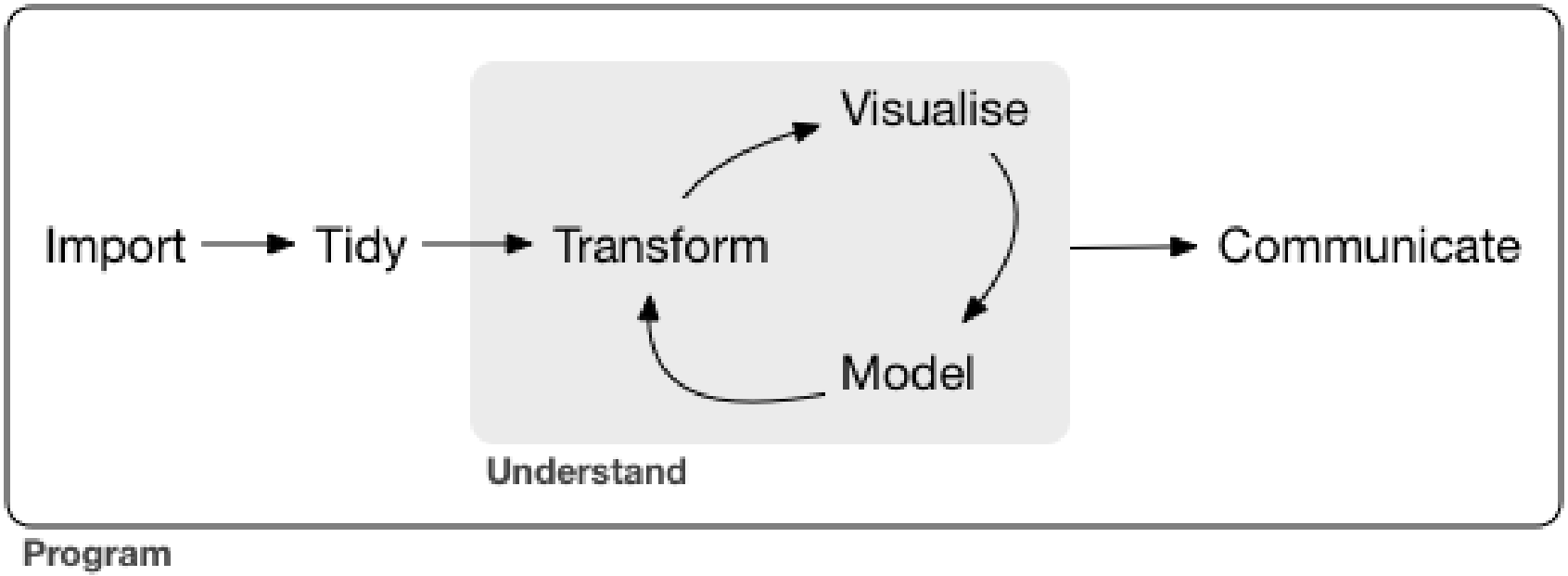
Readme

- License: Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License http://creativecommons.org/licenses/by-nc-sa/4.0/
- You are free to:
  - **Share** : copy and redistribute the material
  - **Adapt** : rebuild and transform the material
- Under the following conditions:
  - **Attribution** : You must give appropriate credit, provide a link to the license, and indicate if changes were made.
  - **NonCommercial** : You may not use this work for commercial purposes.
  - **Share Alike** : If you remix, transform, or build upon this work, you must distribute your contributions under the same license to this one.

Outline: Data Exploration

- The Data Science Approach in R
- Data Visualization
- Data Transformation
- Exploratory Data Analysis

Recall: The Data Science Approach in R



Data Visualization

Introduction

> *"The simple graph has brought more information to the data analyst's mind than any other device."*
>
> — *John Tukey*

We consider three components of visualization:

1. Aesthetics
2. Facetting
3. Geoms

Aesthetic mappings

Datasets in packages

- Packages include example datasets
- Datasets can be loaded by loading the package

```
library(ggplot2)
data(package="ggplot2")
```

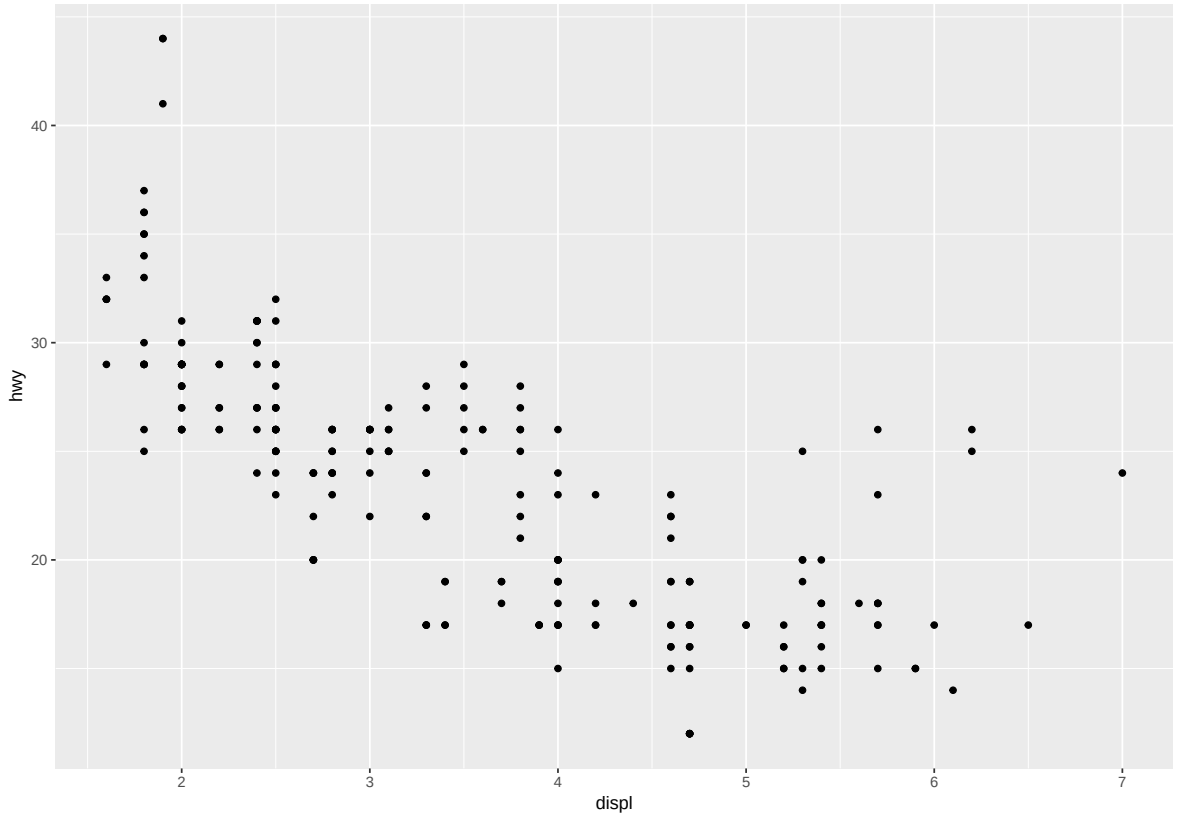
The mpg dataset

- The dataset mpg is included in package ggplot2
- Obtain some information about it.

```
?mpg
head(mpg)
str(mpg)
summary(mpg)
dim(mpg)
```

Scatterplot basics

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy))
```



Additional information

- Plots can be enhanced by displaying additional information
- **aesthetics** displays it using distinct shapes, colors or sizes.
- **faceting** breaks displays into multiple smaller displays for different subsets.

Improving plots

For better plot "add" the information to the call

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy,
                           color = class))

ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy,
                           alpha = class))

ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy,
                           shape = class))
```

Your turn now

- Experiment with colour, size, and shape aesthetics.
- What's the difference between discrete or continuous variables?
- What happens when you combine multiple aesthetics?

In summary

	Discrete	Continuous
Colour	Rainbow	Gradient
Size	Disrete size steps	Linear mapping radius-value
Shape	Different shape each	Doesn't work

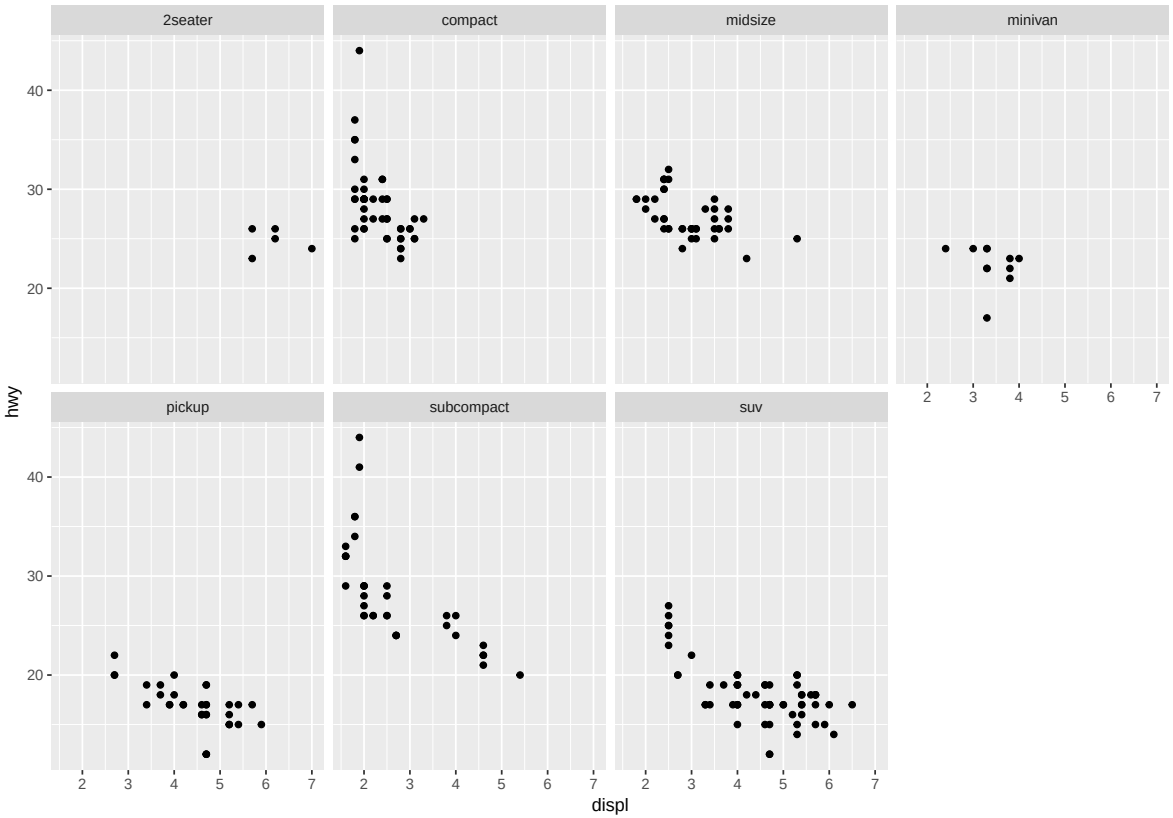
Facets

Faceting

- Break the visualization in many small plots -
- Each (sub)plot reflects one of multiple conditions defined by one or more (categorical) variables.
- Useful for exploring conditional relationships or for when there are many data.

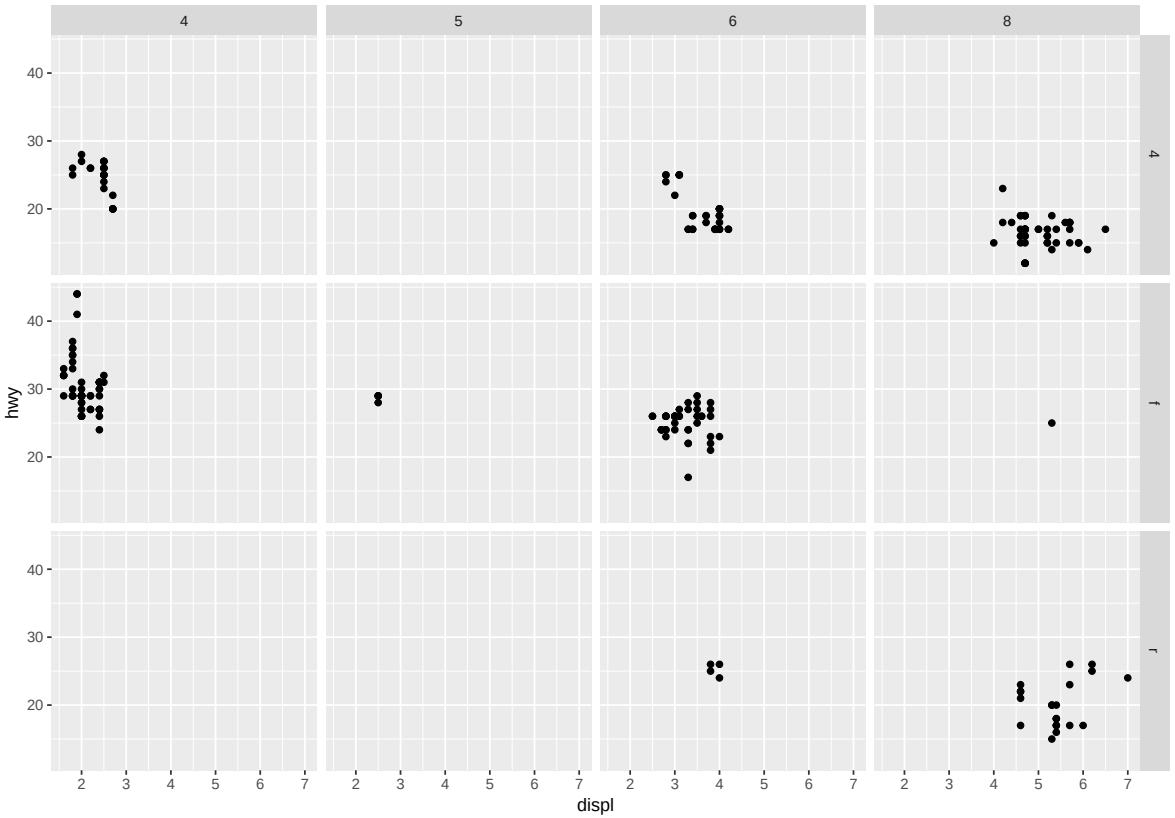
facet_wrap: split plots by one variable

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy))
+ facet_wrap(~ class, nrow = 2)
```



facet_grid: split plots by two variables

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy))
+ facet_grid(drv ~ cyl)
```



Your turn

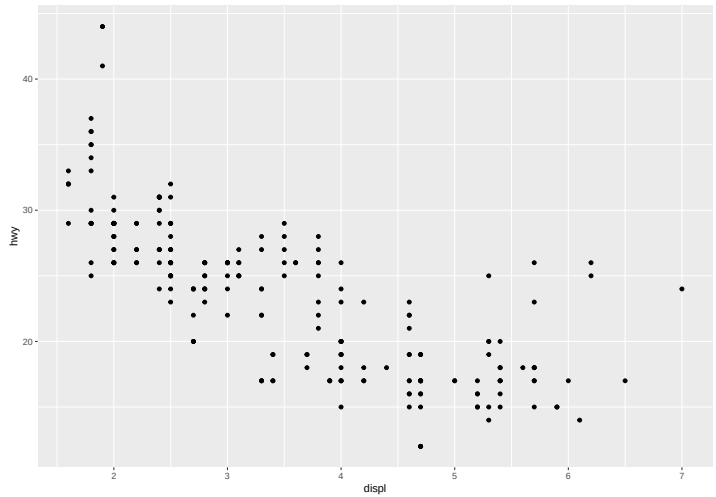
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy))
+ facet_grid(. ~ cyl)

ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy))
+ facet_grid(drv ~ .)
```

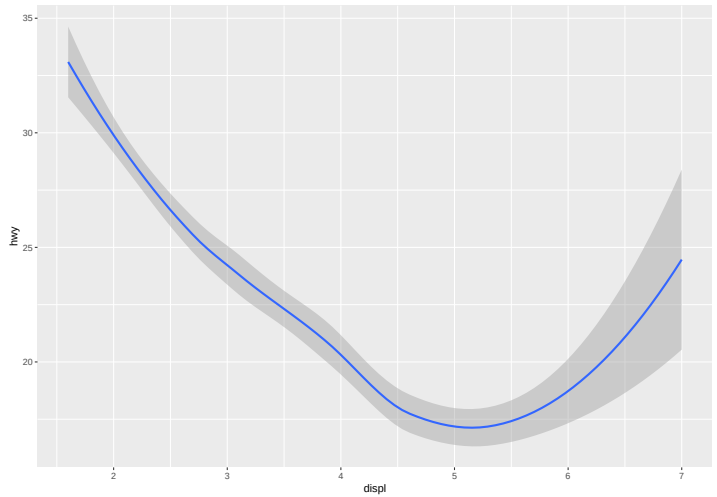
Geometric Objects “Geoms”

What are "geom"

- A `geom` is the geometrical object that a plot uses to represent data.
- For example,
  - Bar charts use bar geoms,
  - Line charts use line geoms,
  - Boxplots use boxplot geoms
  - Scatterplots use the point geom !

Applying `geoms`: How are these plots similar?





Using `geoms`

- Both plots contain the same x variable and the same y variable,
- both describe the same data.
- Each plot uses a different visual object to represent the data.
- In ggplot2 syntax, we say that they use different **geoms**.

Changing geoms

- To change the geom in your plot, change the geom function that you add to ggplot()

```
# left
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy))

# right
ggplot(data = mpg) +
  geom_smooth(mapping = aes(x = displ, y = hwy))
```