Data Science Practice

STATS 369 Coursebook: Week 2

Lecture 4

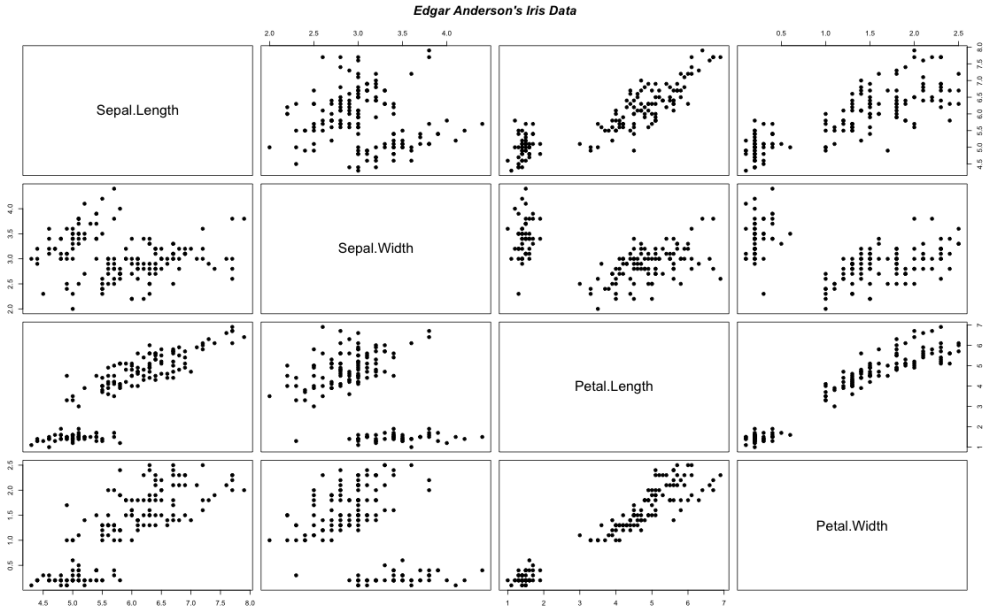
Plan for this week

- **[L04] Data Visualisation with `{ggplot2}`**
  - Motivation
  - `{ggplot2}` package
    - Aesthetic attributes
    - Geometric objects
    - Facets
- **[L05] Examples**
- **[L06] General Comments**
  - Which (common) plot to use?
  - What to pay attention to?
  - Further comments

Motivation

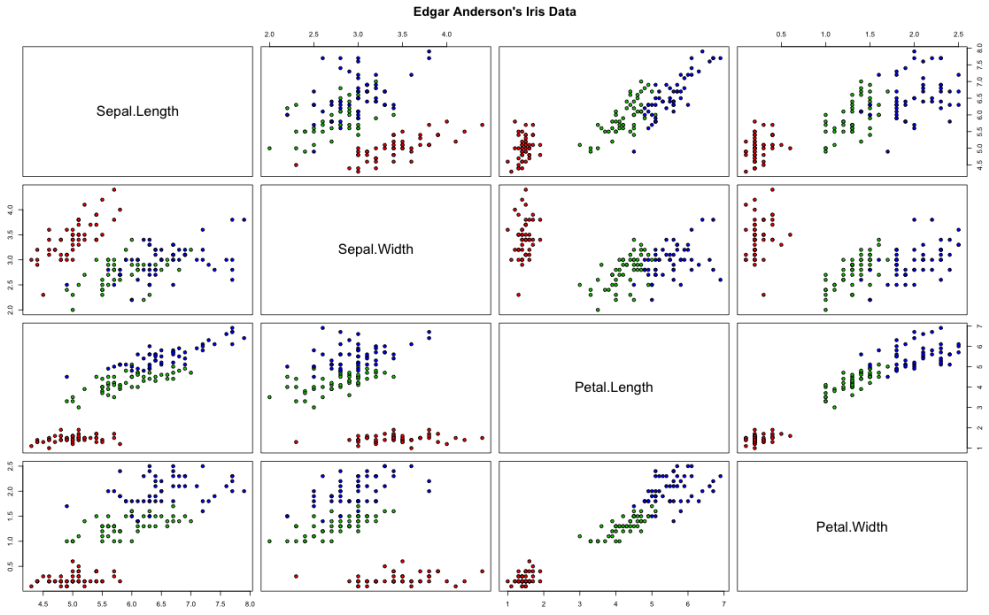
Why graphs are important?

- *First impression* matters -- it is visually stimulating.
- *Efficiency in exploring the data* -- Always visualise your data sets before creating any models!
- *Effective communication* -- 'A picture is worth a thousand words'.
- Sometimes, summary statistics are just not enough -- see examples here.



Why graphs are important?

- *First impression* matters -- it is visually stimulating.
- *Efficiency in exploring the data* -- Always visualise your data sets before creating any models!
- *Effective communication* -- 'A picture is worth a thousand words'.
- Sometimes, summary statistics are just not enough -- see examples here.



ggplot

The name `ggplot` comes from the book *The Grammar of Graphics* by Leland Wilkinson (2005) (ref: ISBN 978-0-387-98774-3). A grammar of graphics is a framework that allows a structured and layered approach to construct graphics.

Components of a graph

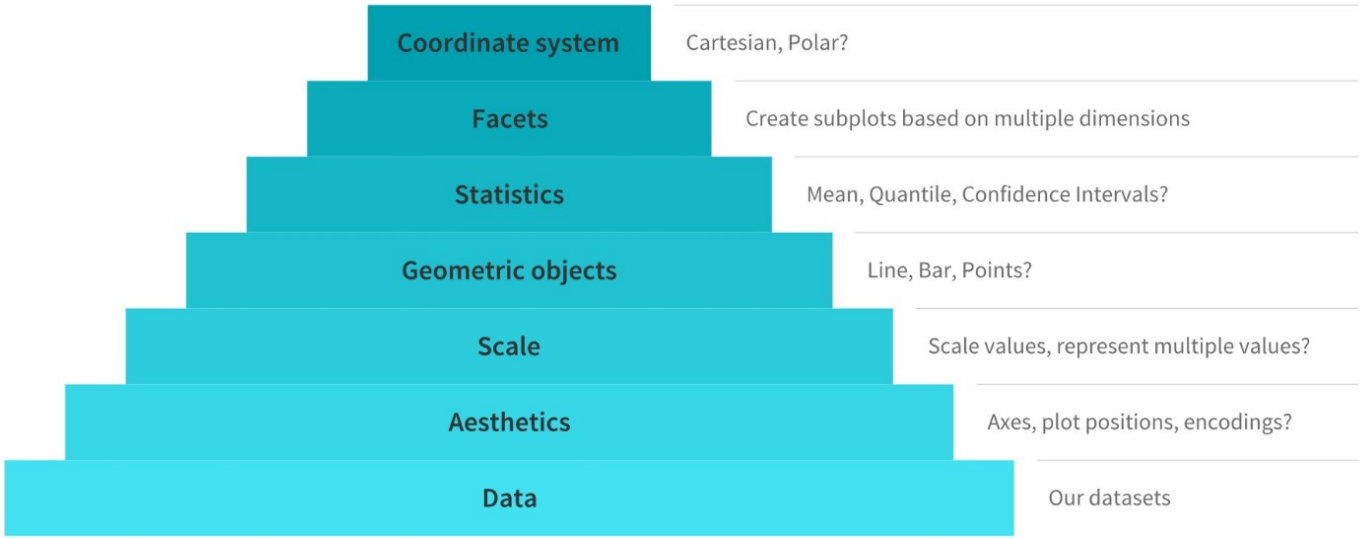


Image source

Data Visualisation with `{ggplot2}`

The {ggplot2} package

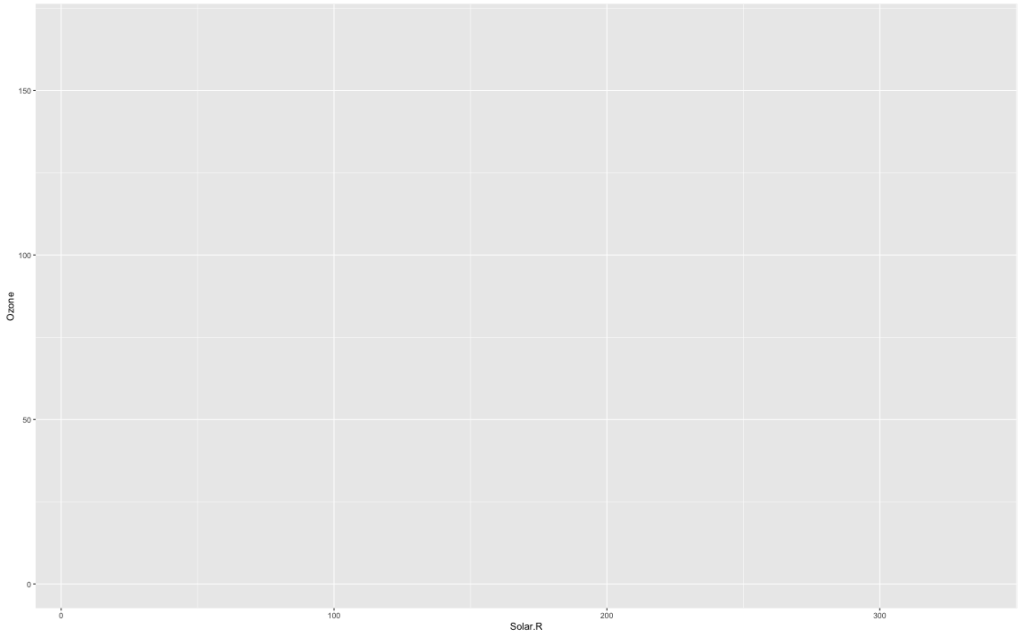
An R visualisation package developed by Hadley Wickham (2007) that adapts and implements the concept of *ggplot*.

> 'ggplot2 (Wickham 2009) builds on Wilkinson's grammar by focussing on the primacy of layers and adapting it for use in R. In brief, the grammar tells us that a graphic maps the data to the **aesthetic attributes** (colour, shape, size) of **geometric objects** (points, lines, bars). The plot may also include **statistical transformations** of the data and information about the plot's **coordinate system**. **Facetting** can be used to plot for different subsets of the data. The combination of these independent components are what make up a graphic.' -- Hadley Wickham, *ggplot2*

Check out the 'R Graph Gallery' here.

Aesthetic attributes

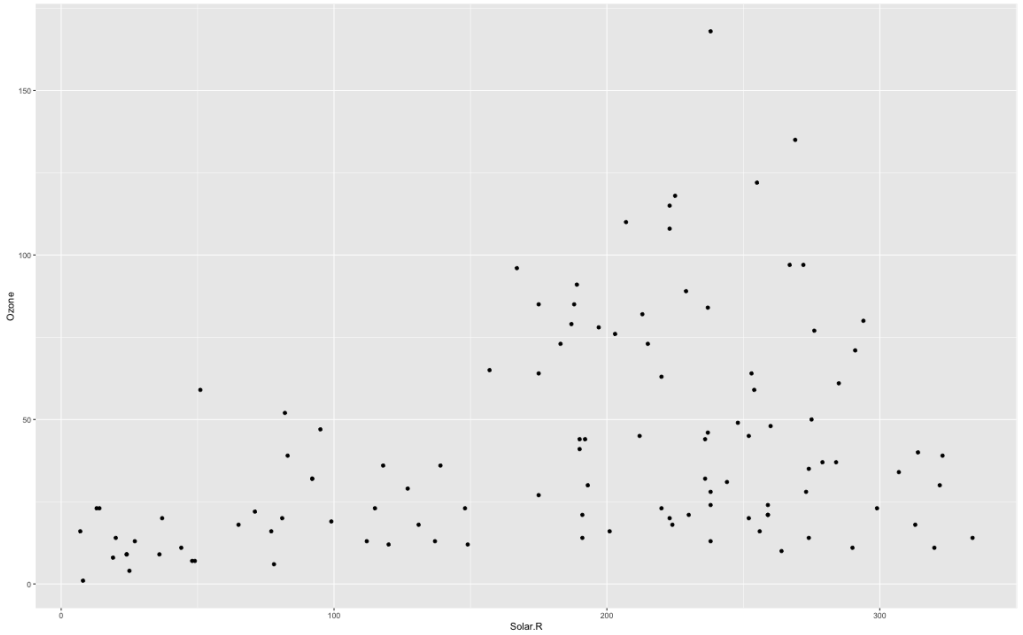
What attributes will be mapped onto the x-axis and y-axis?



Scatter plot

Now we can actually add some points

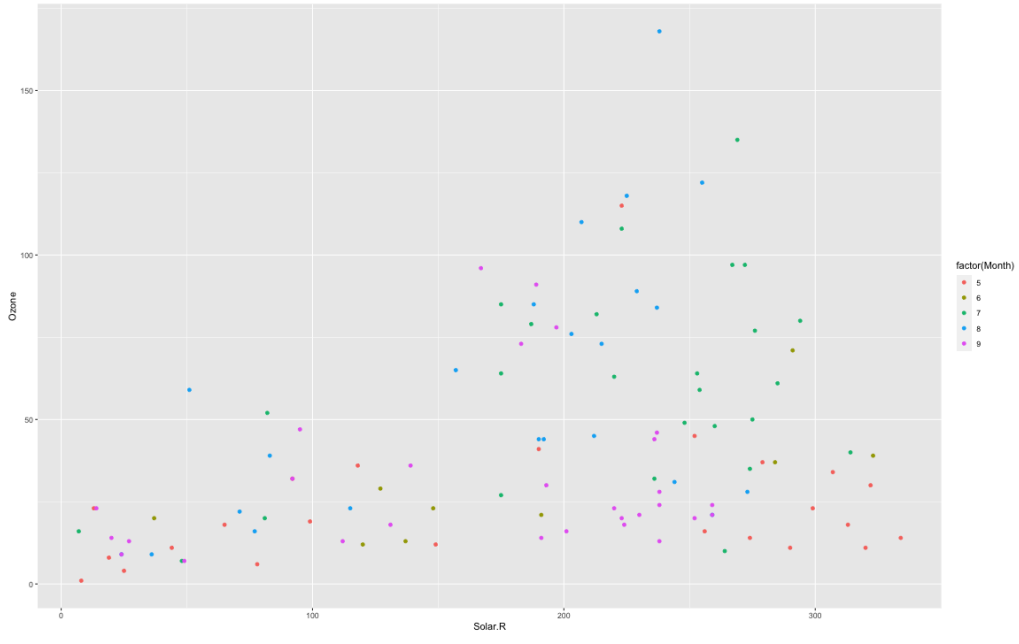
```
airquality %>%
  ggplot(aes(x = Solar.R, y = Ozone)) +
  geom_point()
```



Colours

How about adding the colour to points based on another (factor) variable.

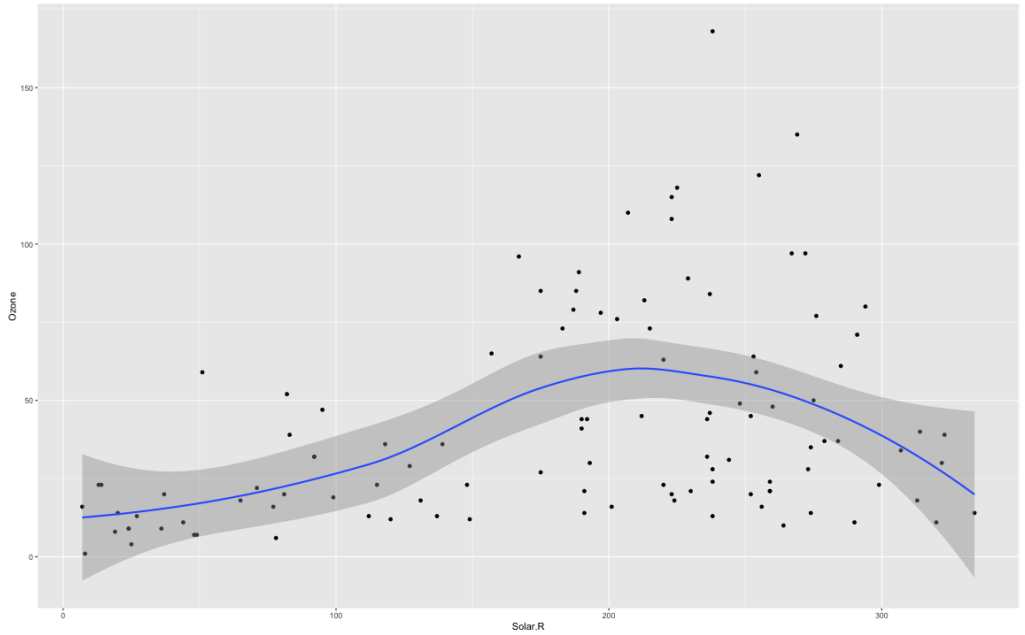
```
airquality %>%
  ggplot(aes(x = Solar.R, y = Ozone,
             color = factor(Month))) +
  geom_point()
```



Geometric objects

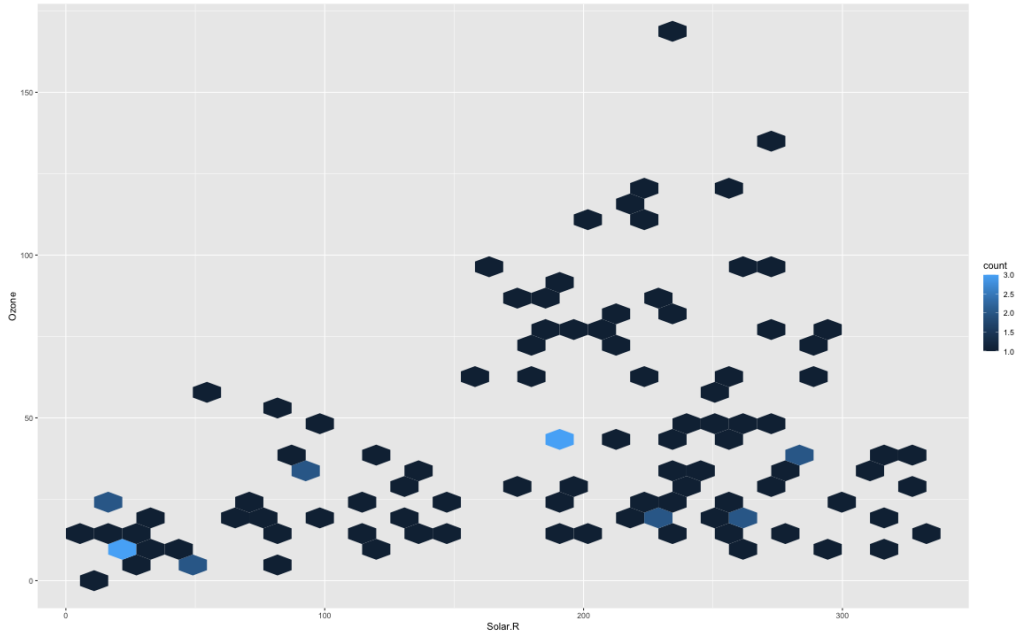
We have seen points, what else?

```
airquality %>%
  ggplot(aes(x = Solar.R, y = Ozone)) +
  geom_point() +
  geom_smooth()
```



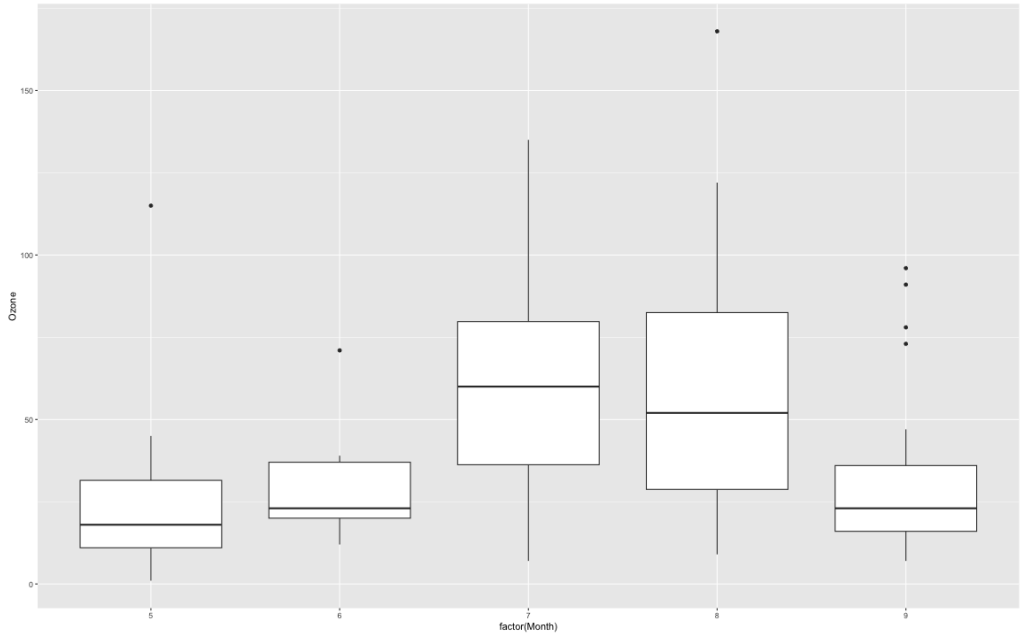
Geometric objects

```
airquality %>%
  ggplot(aes(x = Solar.R, y = Ozone)) +
  geom_hex()
```



Geometric objects

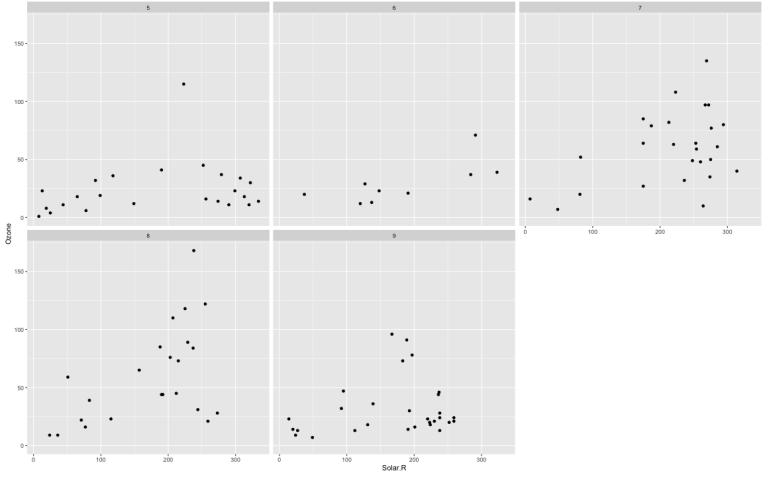
```
airquality %>%
  ggplot(aes(x = factor(Month),
             y = Ozone)) +
  geom_boxplot()
```



Facets

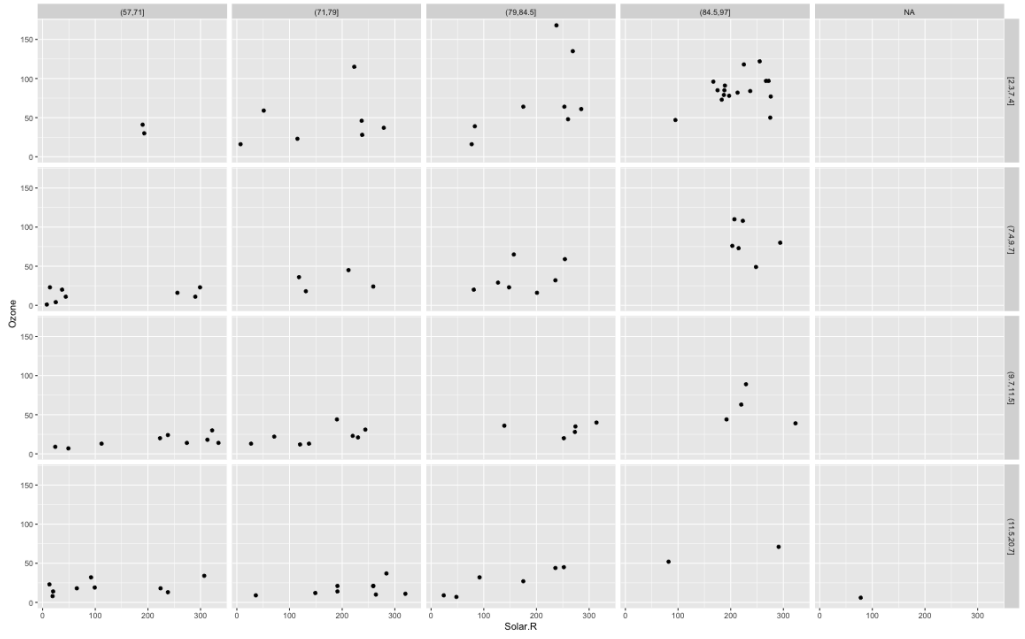
Dividing data sets into sub groups and plot separately for each group. It is useful when when the relationship is beyond 2D -- you can explore relationship between two variables conditioned on other variable(s). There are two common types of faceting in R, `facet_grid` and `facet_wrap`.

```
airquality %>%
  ggplot(aes(x = Solar.R, y = Ozone)) +
  geom_point() +
  facet_wrap(~Month, nrow = 2)
```



Incorporating data processing

```
airquality %>% na.omit() %>%
  mutate(TempGp = cut(Temp,
           breaks = quantile(Temp, (0:4)/4),
           nc = TRUE)) %>%   # use %>%
  mutate(WindGp = cut(Wind,
           breaks = quantile(Wind, (0:4)/4),
           inc = TRUE)) %>%
  ggplot(aes(x = Solar.R, y = Ozone)) +
  geom_point() + # use +
  facet_grid(WindGp ~ TempGp)
```



The NZ Vehicle Registration Data

Car registration open data

NZ vehicle registration open data provides a snapshot of the currently registered fleets in NZ. The 2019 dataset has 206,099 rows with 34 columns.

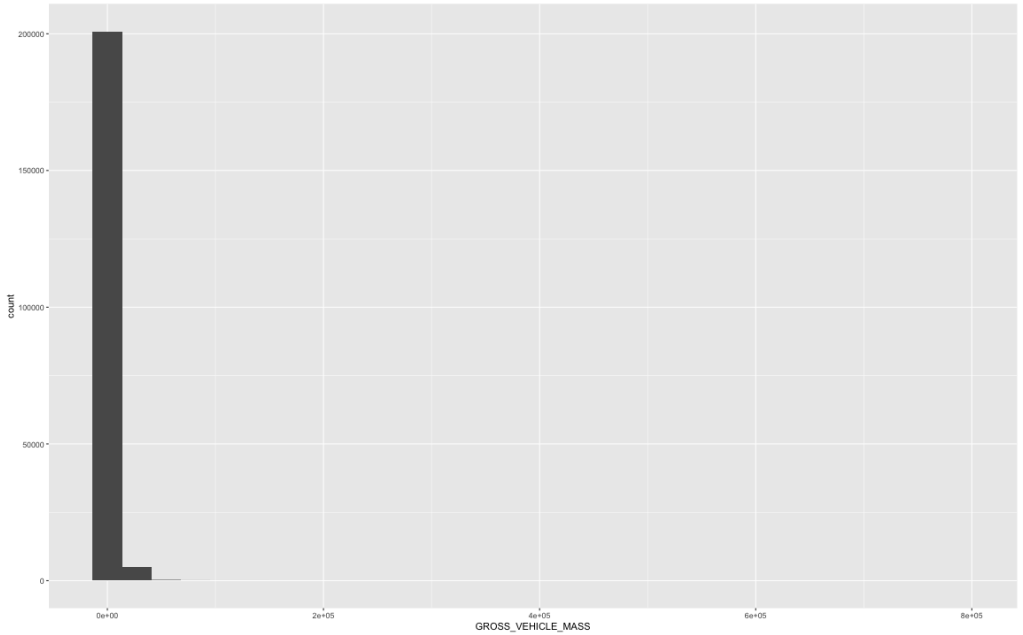
```
cars.df <- read_csv("datasets/VehicleYear-2019.csv")
# glimpse(cars.df)
dim(cars.df)  # dimension of the data frame
head(names(cars.df), 20)  # some column names of the cars.df
```

```
## [1] 206099     34
```

```
##  [1] "ALTERNATIVE_MOTIVE_POWER"    "BASIC_COLOUR"
##  [3] "BODY_TYPE"                   "CC_RATING"
##  [5] "CHASSIS7"                    "CLASS"
##  [7] "ENGINE_NUMBER"               "FIRST_NZ_REGISTRATION_YEAR"
##  [9] "FIRST_NZ_REGISTRATION_MONTH" "GROSS_VEHICLE_MASS"
## [11] "HEIGHT"                      "IMPORT_STATUS"
## [13] "INDUSTRY_CLASS"              "INDUSTRY_MODEL_CODE"
## [15] "MAKE"                        "MODEL"
## [17] "MOTIVE_POWER"                "MVMA_MODEL_CODE"
## [19] "NUMBER_OF_AXLES"             "NUMBER_OF_SEATS"
```

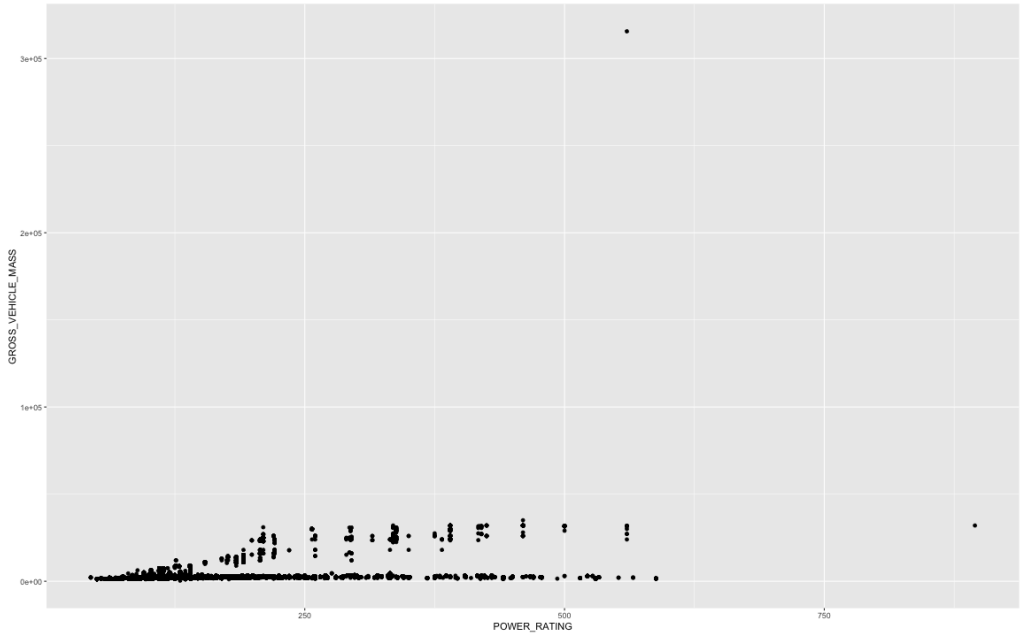
Distribution of car weight

```
cars.df %>%
  ggplot(aes(x = GROSS_VEHICLE_MASS)) +
  geom_histogram()
```



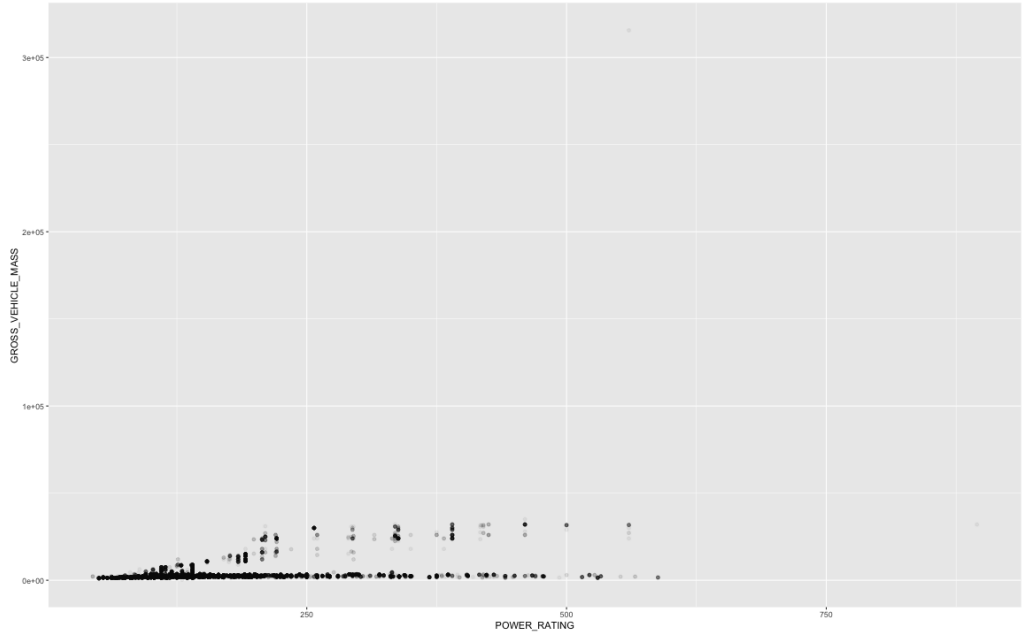
Filter zero weight cars

```
cars.df %>%
  filter(GROSS_VEHICLE_MASS > 0,
         POWER_RATING > 0) %>%
  ggplot(aes(x = POWER_RATING,
             y = GROSS_VEHICLE_MASS)) +
  geom_point()
```



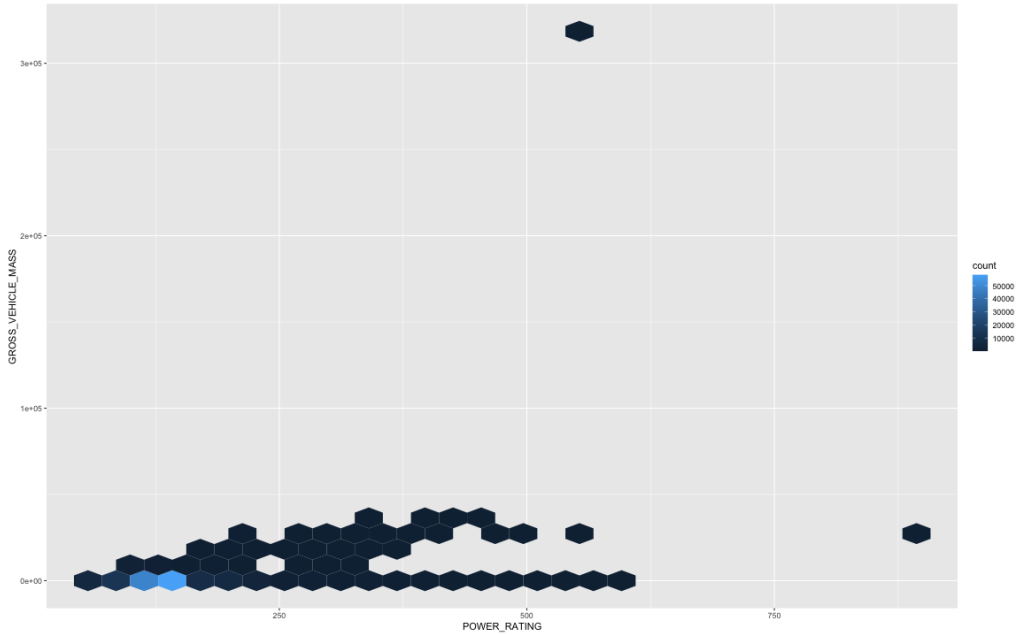
Better scatter plot

```
cars.df %>%
  filter(GROSS_VEHICLE_MASS > 0,
         POWER_RATING > 0) %>%
  ggplot(aes(x = POWER_RATING,
             y = GROSS_VEHICLE_MASS)) +
   geom_point(alpha = 0.05)
```



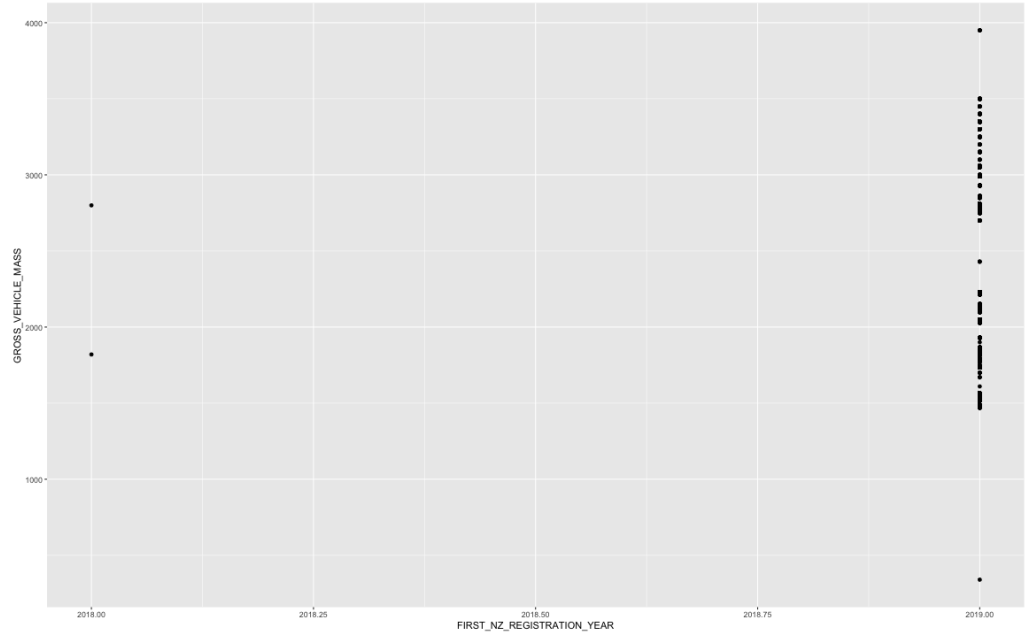
Try hex(bin) plot

```
cars.df %>%
  filter(GROSS_VEHICLE_MASS > 0,
         POWER_RATING > 0) %>%
  ggplot(aes(x = POWER_RATING,
             y = GROSS_VEHICLE_MASS)) +
   geom_hex()
```



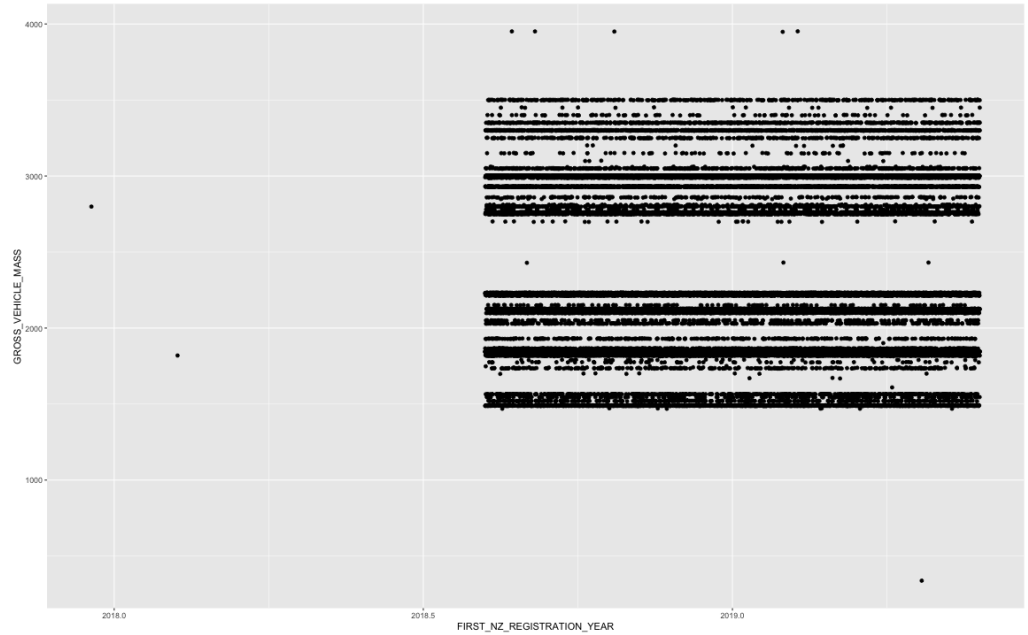
Further exploration

```
cars.df %>%
  filter(GROSS_VEHICLE_MASS > 0,
         POWER_RATING > 0,
         MAKE == 'TOYOTA') %>%
  ggplot(aes(x = FIRST_NZ_REGISTRATION_YEAR,
             y = GROSS_VEHICLE_MASS)) +
  geom_point()
```



Add 'jitter'

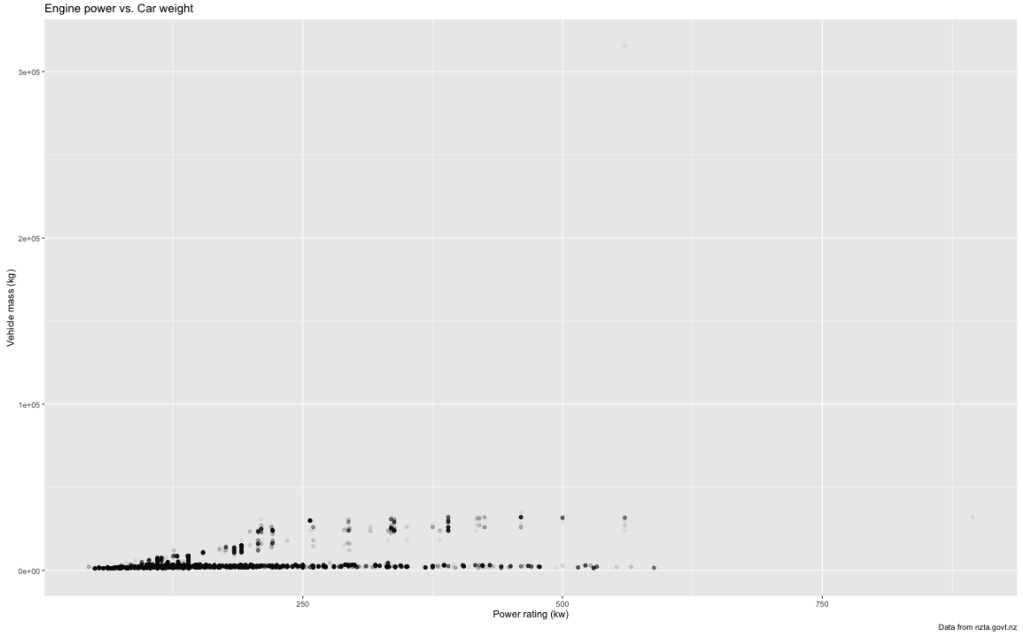
```
cars.df %>%
  filter(GROSS_VEHICLE_MASS > 0,
         POWER_RATING > 0,
         MAKE == 'TOYOTA') %>%
  ggplot(aes(x = FIRST_NZ_REGISTRATION_YEAR,
             y = GROSS_VEHICLE_MASS)) +
   geom_jitter()
```



Scales, labels and theme

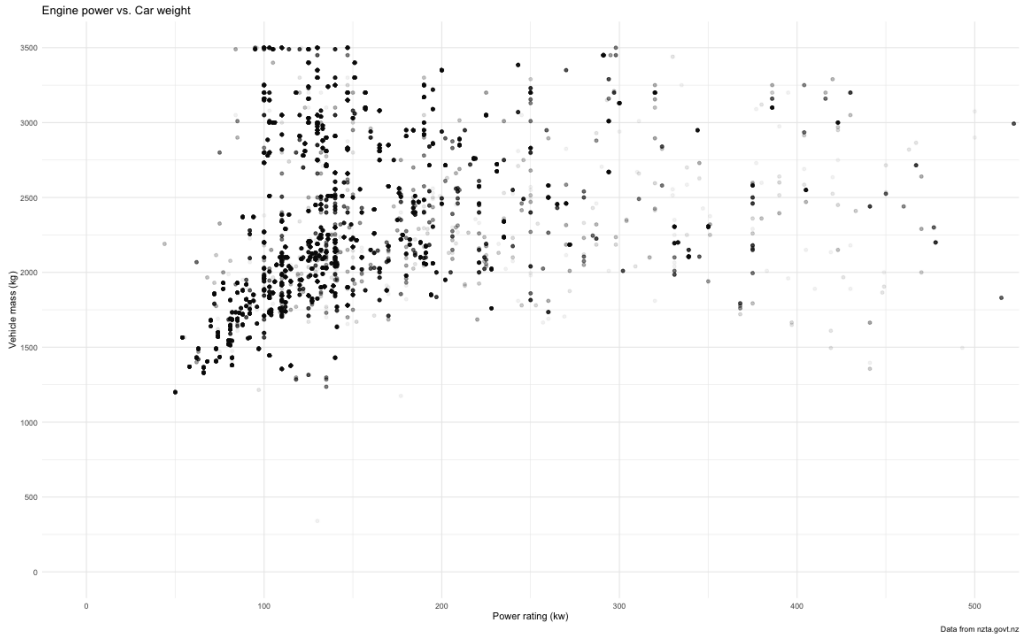
If you are making plots for others, it is a good idea to make them clear and readable.

```
p <- cars.df %>%
  filter(GROSS_VEHICLE_MASS > 0,
         POWER_RATING > 0) %>%
  ggplot(aes(x = POWER_RATING,
             y = GROSS_VEHICLE_MASS)) +
  geom_point(alpha = 0.05) +
  labs(title = "Engine power vs. Car weight",
       x = "Power rating (kw)",
       y = 'Vehicle mass (kg)',
       caption = 'Data from nzta.govt.nz')
p
```



Get the scale and coordinate right, and a different theme

```
p +
  scale_y_continuous(
    limits = c(0,3500),
    breaks = seq(0,3500,by=500)) +
  coord_cartesian(xlim = c(0, 500)) +
  theme_minimal()
```



Which (common) plot to use?

Single Variable (univariate)

Type	(Common) Plot to use	Features to Pay Attention to
Quantitative (e.g. a variable of measurement)	dot plot/stript chart, histogram, density plot, box plot	shape, peaks, center, variability, outliers.
Qualitative (e.g. count of a grouping variable)	bar plot, pie chart, table of counts	majority/minority group, gaps in group counts.

Which (common) plot to use?

Two Variables (bivariate)

Type	(Common) Plot to use	Features to Pay Attention to
Quantitative vs. Qualitative	side-by-side histogram/density/box plot	compare shapes, centers, variability; outliers from individual group.
Quantitative vs. Quantitative	scatter plot, line plot	shapes, peaks, center, variability, outliers, correlation, grouping of observations, seasonal variation (for time series).
Qualitative vs. Qualitative	faceted bar plot, 2-way table of counts, pie chart (?)	compare group counts, distributions and gaps.

NB: Plotting for 2+ variables can often be achieved by 'reducing' it to some variations of bi-variate plots.

General Plotting Advice

- Use colors, shapes etc, but keep things **balanced**.
- Keep the focus -- produce a plot with clear message in mind.
- Be aware of scales, labels and Hierarchy.
- Leave some white space.
- ...

General Plotting Advice

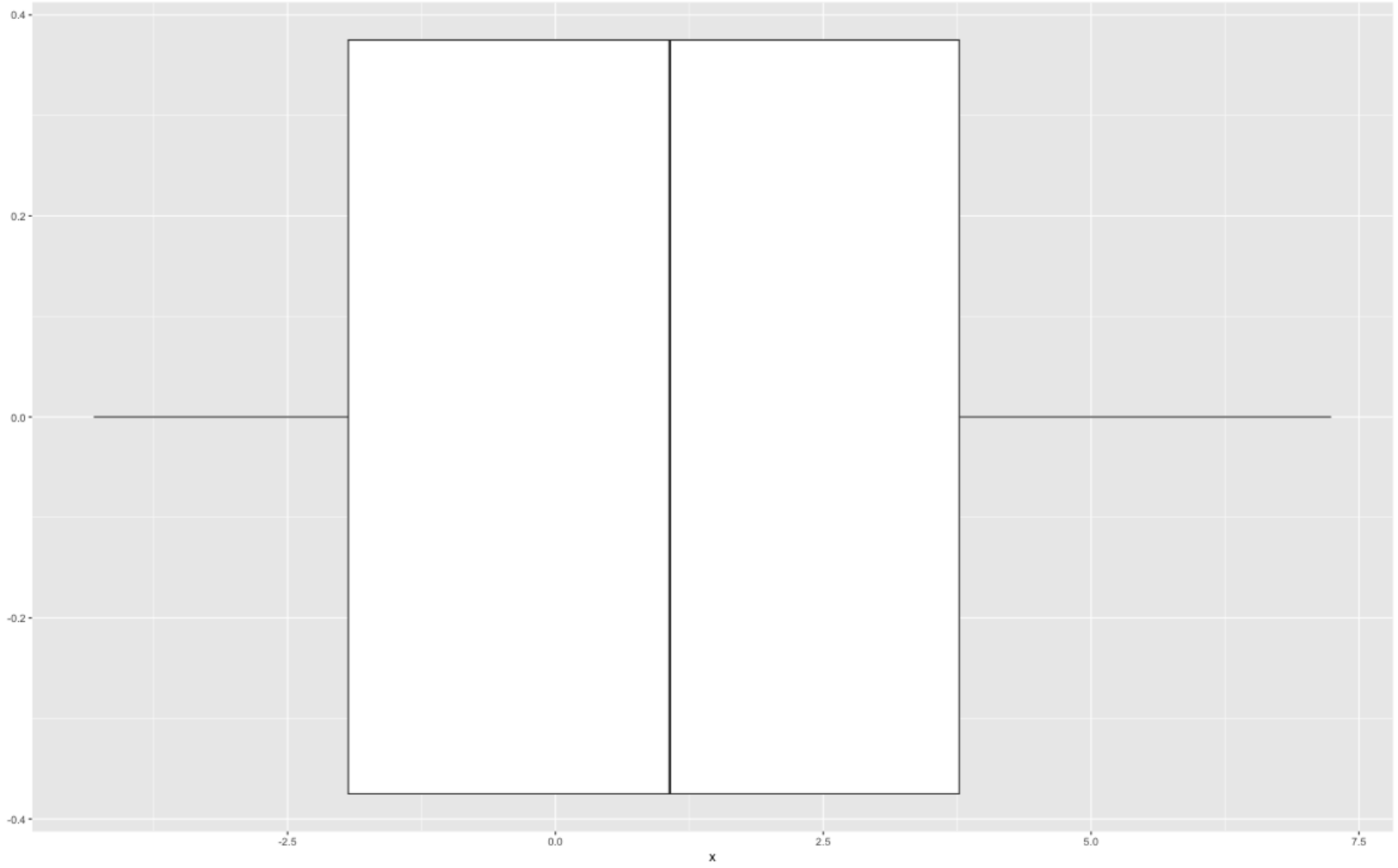
- Avoid pie charts (?!)

> "Avoid pie-charts. Especially 3d pie-charts. Especially 3d pie-charts with exploding wedges. I promise all my students an instant fail if I ever see anything so appalling." - Rob J Hyndman, from "Twenty rules for good graphics"

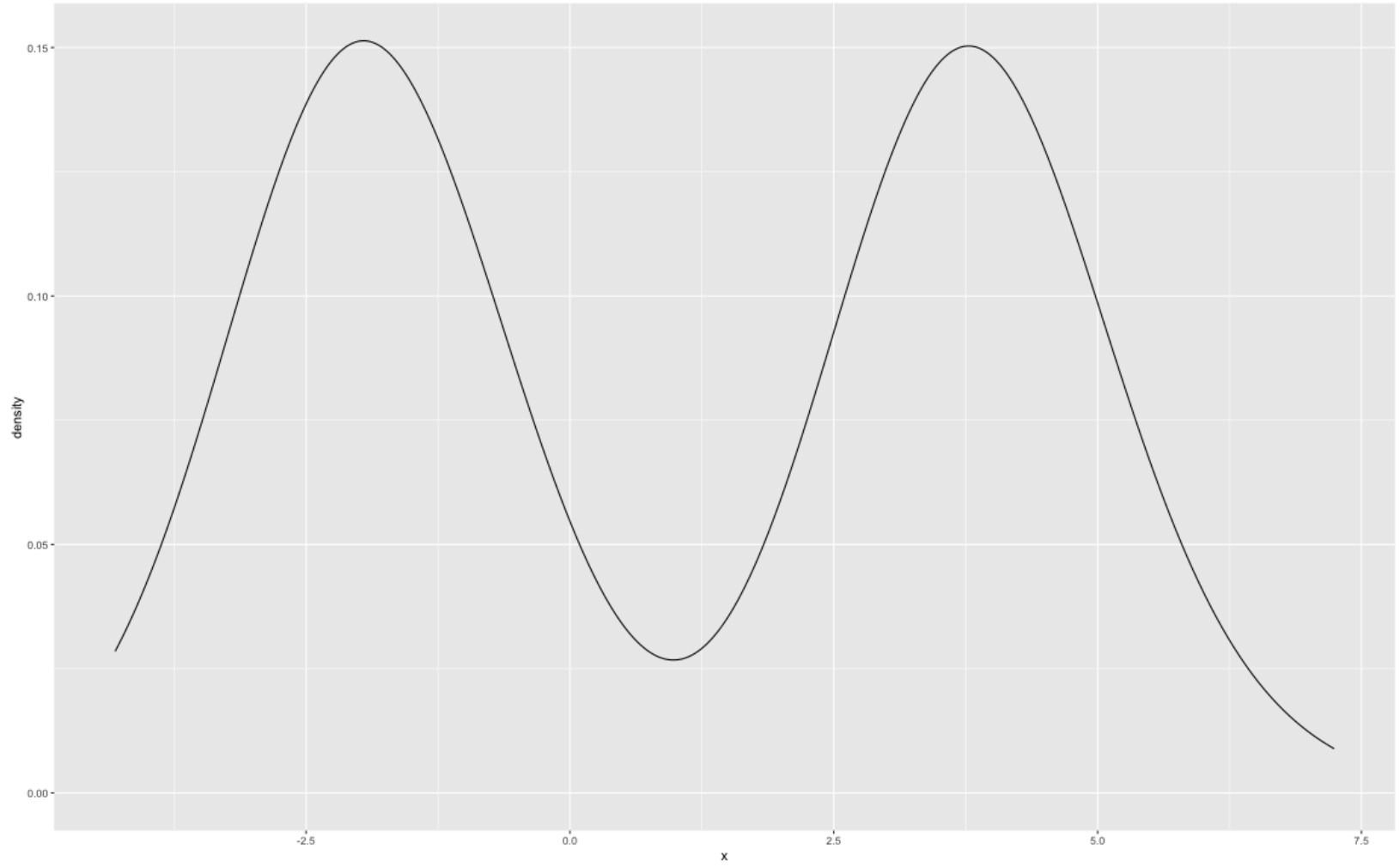
- Sometimes, it would be helpful to produce multiple (types of) plots for the same data to reveal the real pattern.

  For example...

...For example, what can you see from the boxplot below?



But if we look at the density plot...



Charts and accessibility

While charts are very much a visual medium, we can improve accessibility of our charts by including 'alternative text', often known as 'alt text'.

📚 To read

Cesal., A. (2020). *Writing Alt Text for Data Visualization*. Nightingale https://nightingaledvs.com/writing-alt-text-for-data-visualization/