CREATING SCATTERPLOTS

- so far - univariate analysis (looking at one variable at a time)
- now - look at associations between variables

scatterplot - show relationship bw 2 quantitative variables

INSTALL AND LOAD PACKAGES

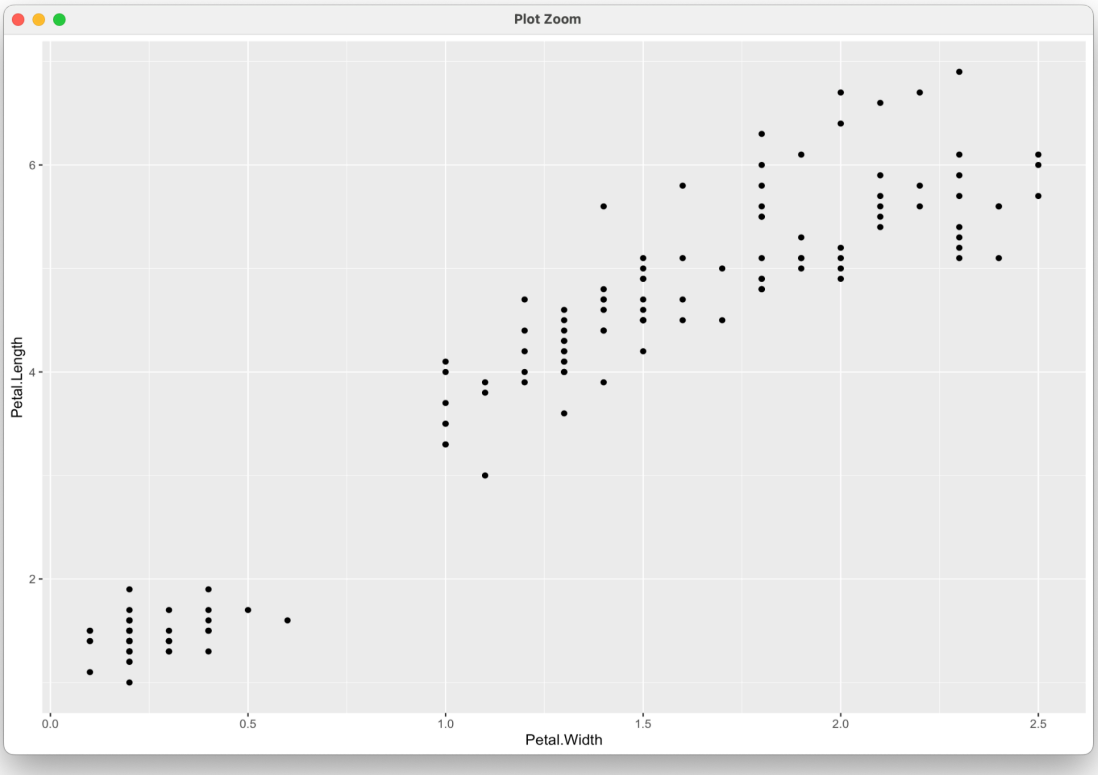
```
pacman::p_load(datasets, pacman, rio, tidyverse)
```

QPLOT

Basic scatterplot:

```
qplot(Petal.Width, Petal,Length, data = iris)
# (1st variable, 2nd variable, source)
```

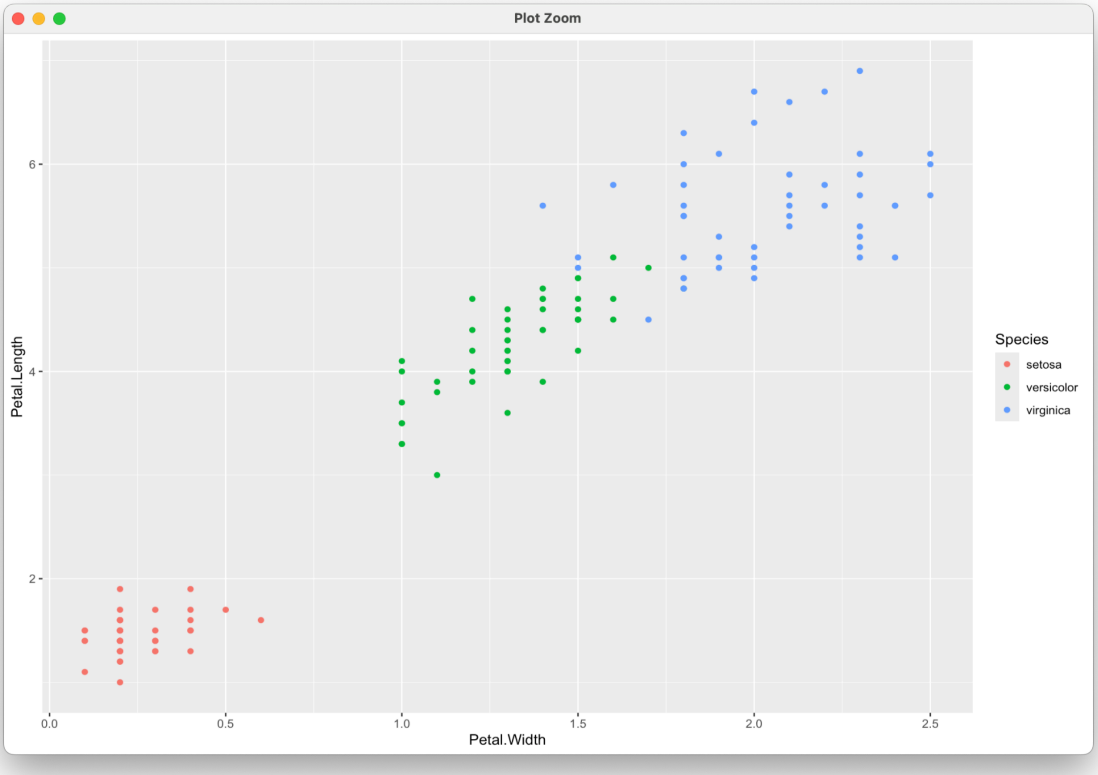
- result: scatterplot — with petal width on bottom , petal length on side, AND dot for each data-point (150 dots)
- something funny going on - bc gap near bottom , but otherwise a nice linear association



Colour by species:

```
qplot(Petal.Width,
    Petal.Length,
    color = Species,
    data = iris)
```

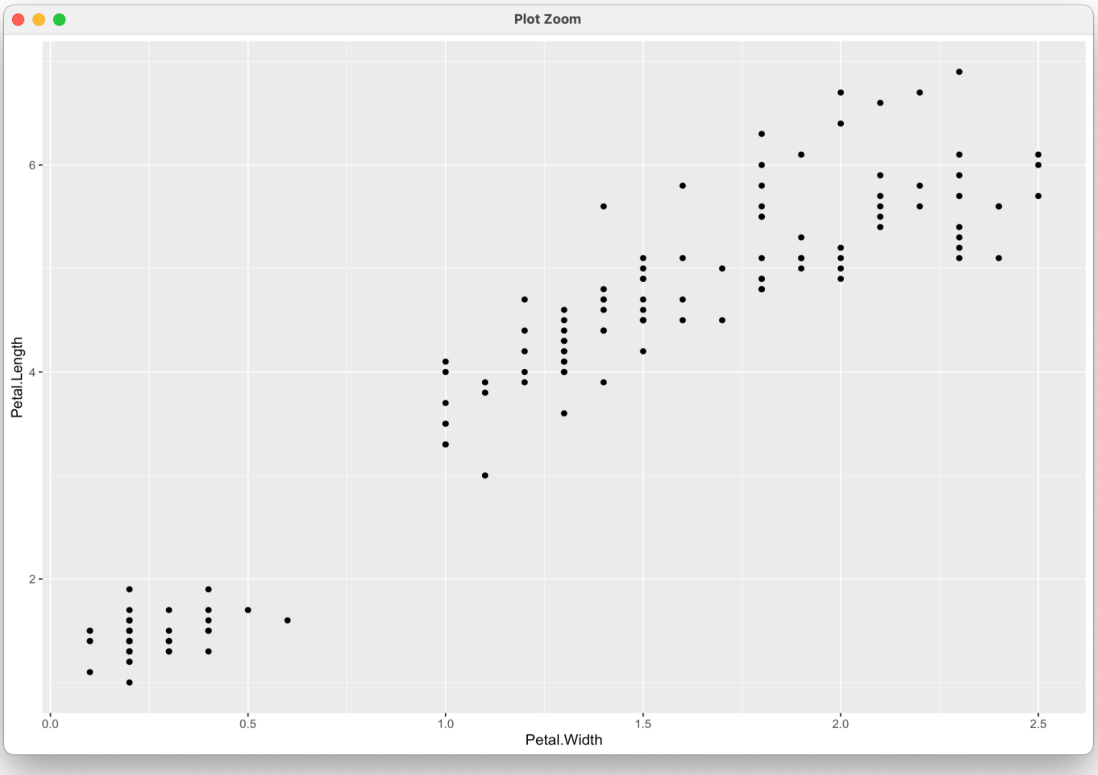
- result: see the 3 groups with one on bottom, and other 2 groups on top
- conclusion: important differences between species, even though general pattern is consistent



GGPLOT2

Basic scatterplot:

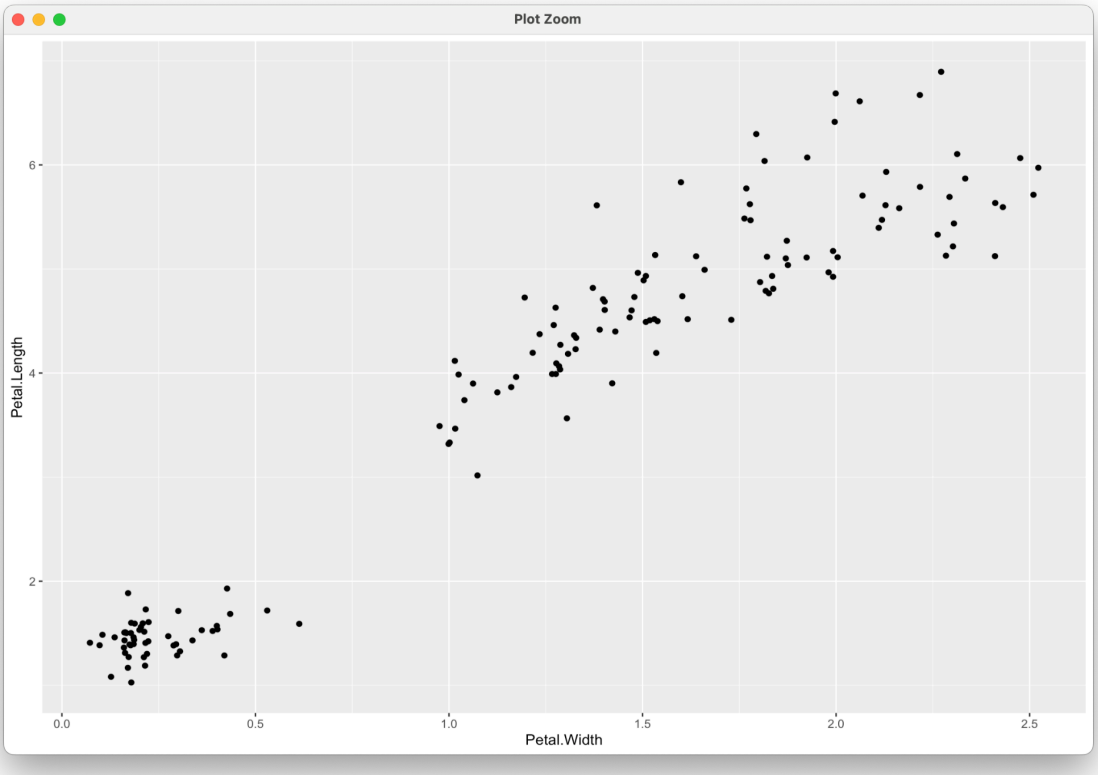
```
ggplot(iris,
    aes(Petal.Width, Petal.Length)) +
    geom_point()          # dot plot
```



Scatterplot, jittered:

```
ggplot(iris,
    aes(Petal.Width, Petal.Length)) +
    geom_jitter()          # jitter the points so not on top of each other
```

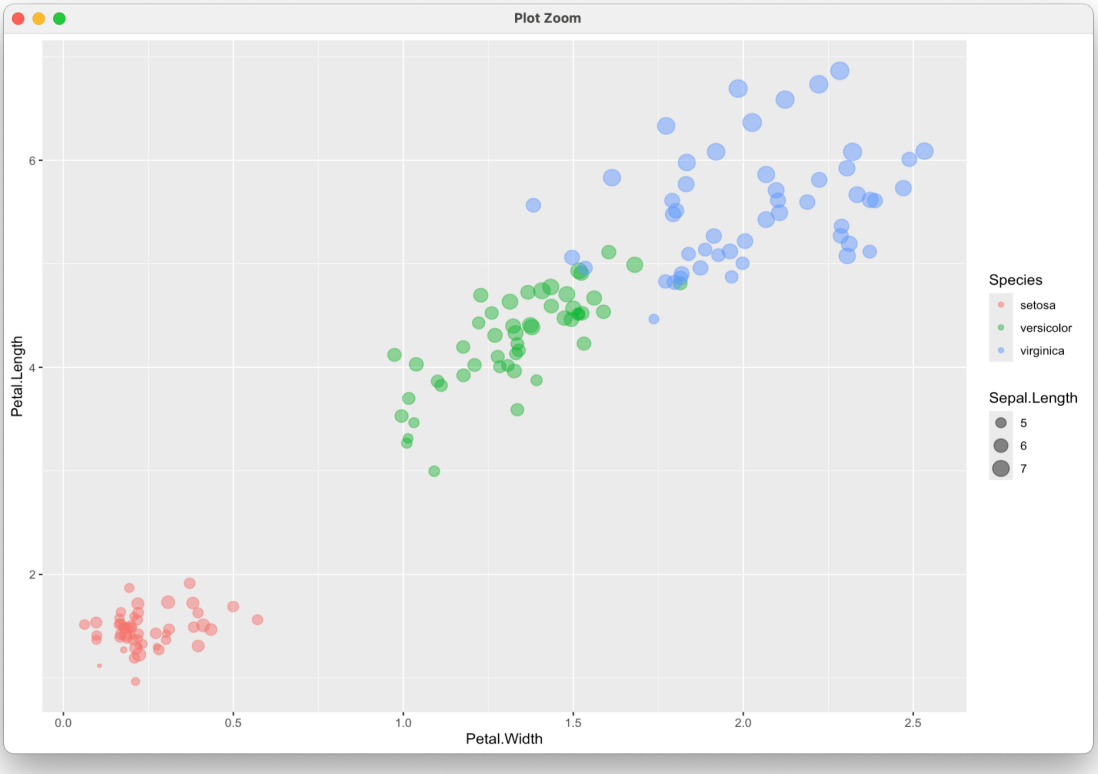
- result: helps see overall pattern better



Scatterplot, jittered, variable size, coloured by species:

```
ggplot(iris,
    aes(Petal.Width, Petal.Length,
        size = Sepal.Length,   # change size of points depending on length of sepal
        color = Species)) +
    geom_jitter(alpha = .5)   # somewhat transparent jitter points
```

- result: size of dots indicates sepal length (a third measurement)
    - sepal length is bigger on top of plot than on bottom (which makes sense)

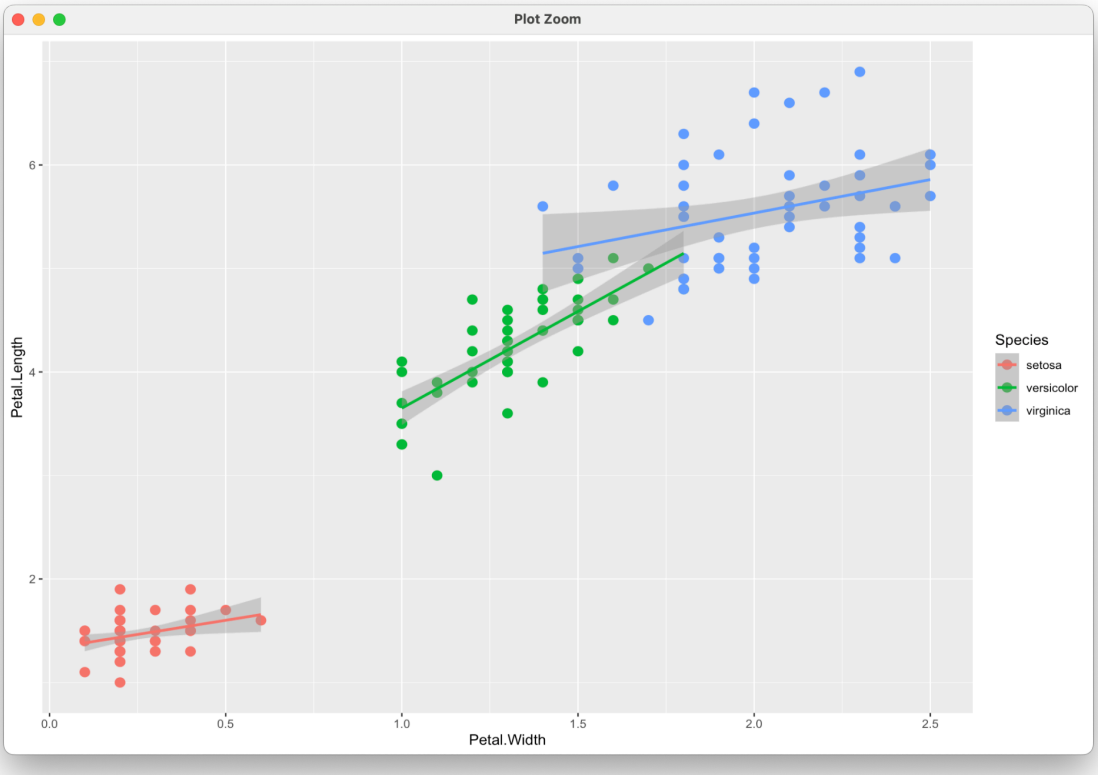


Scatterplot, coloured by species, fit line: (regression line)

- done separately for each of the categories

```
ggplot(iris,
    aes(Petal.Width, Petal.Length,
       color = Species)) +
    geom_point(size = 3) +
    geom_smooth(method = lm)   # lm = linear model
```

- result: draws separate regression line for each species - and standard error for each of the groups independently
    - basic uphill - a little stronger uphill for the green versicolor



Scatterplot, coloured by species, fit line, density:

```
ggplot(iris,
    aes(Petal.Width, Petal.Length,
        color = Species)) +
    geom_point(size = 3) +
    geom_smooth(method = lm) +
    geom_density2d(alpha = .5) +      # add density 2d
      theme(legend.position = "bottom")
```

- add density 2d - make look like maps with circles drawn around them. indicate something akin to confidence intervals, but is really a map to indicate how bunched up the data are.
- result: Looks like a topographic map around the dots. Look at the density of data points.
    - density close together for seritosa - bc points close together