R による探索的データ分析入門 - 可視化

発電基盤開発課高津一誠

2018 年 *10* 月 *12* 日

1　可視化から学ぶ

前回お話したように、データプロセスの全体をサポートできるツールが R です。[*1]

では、どこから勉強を始めるべきでしょうか？

- プログラムミング言語だから文法から？
- 他の言語との違いを確認すべき？
- サポートするデータ形式から学ぶ？

そうではないと思います。

R を使ってデータ分析を論理的に行えるようになることが、このコースの目的です。ですから、データ分析プロセスのそれぞれを具体例を元に勉強するのがよいはずです。その中でも、可視化は効果が高く分かりやすいことから、学び始めるのに最も適していると思います。

2　割当（mapping）と階層 (layer)

R に[*2]おける可視化は、論理的に行えるようになっています。論理的とは、簡潔で一貫性のある方法でということです。Excel でのグラフ作成を反例として、考えていきましょう。

Excel で散布図を作るときは、データ範囲とグラフ種類を選べば一発です。これはデータ配置に依存していて、Excel の想定する「X に対応する列が一番左」というルールに合致すれば簡単ですが、そうでないと 1 つづつ系列を編集することになり急に非効率になります。

では、グラフ作成とは何をしているのでしょうか？それはデータのグラフ要素への割当と、グラフ要素を階層的に組み合わせることです。

それでは、R のグラフ (ggplot2) を例に割当と階層を見ていきましょう

データは R に[*3]組み込まれたデータを使います。前回と同じで以下のとおりです。[*4]

[*1] データ分析をすべて R を使って行うべき、ということではありません。Excel の方がやりやすいこと、計測器に付属する分析ツールや解析ツールのポスト処理を使う方が効率的なことも、もちろんあります。R を使った方が有効なときに使ってください。その判断は、これから勉強していくうちにできるようになります。

[*2] より正確には tidyverse パッケージにというべきですが、それについては後日お話します。

[*3] より正確には ggplot2 パッケージに。

[*4] 生態学のデータで、アヤメの花弁（petal）とガク（sepal）の長さと幅を、3 つの種について 50 個体ずつ計測したもの。

```
## # A tibble: 150 x 5
##    Sepal.Length Sepal.Width Petal.Length Petal.Width Species
##           <dbl>       <dbl>        <dbl>       <dbl> <fct>
##  1          5.1         3.5          1.4         0.2 setosa
##  2          4.9         3            1.4         0.2 setosa
##  3          4.7         3.2          1.3         0.2 setosa
##  4          4.6         3.1          1.5         0.2 setosa
##  5          5           3.6          1.4         0.2 setosa
##  6          5.4         3.9          1.7         0.4 setosa
##  7          4.6         3.4          1.4         0.3 setosa
##  8          5           3.4          1.5         0.2 setosa
##  9          4.4         2.9          1.4         0.2 setosa
## 10          4.9         3.1          1.5         0.1 setosa
## # ... with 140 more rows
```

このデータのガクの幅と長さの散布図を種ごとに色を変えて表示したいなら、割当は以下のようになります。

表1: 割当の一覧

データ列 (変数)	グラフ要素
Sepal.Length	X 値
Sepal.Width	Y 値
Species	色

この割当を R では以下のように書きます。**%>%** はデータを次の処理に渡す演算子です。(この詳細についても後日お話します。)

```
iris %>% ggplot(aes(x = Sepal.Width, y = Sepal.Length, color = Species))
```

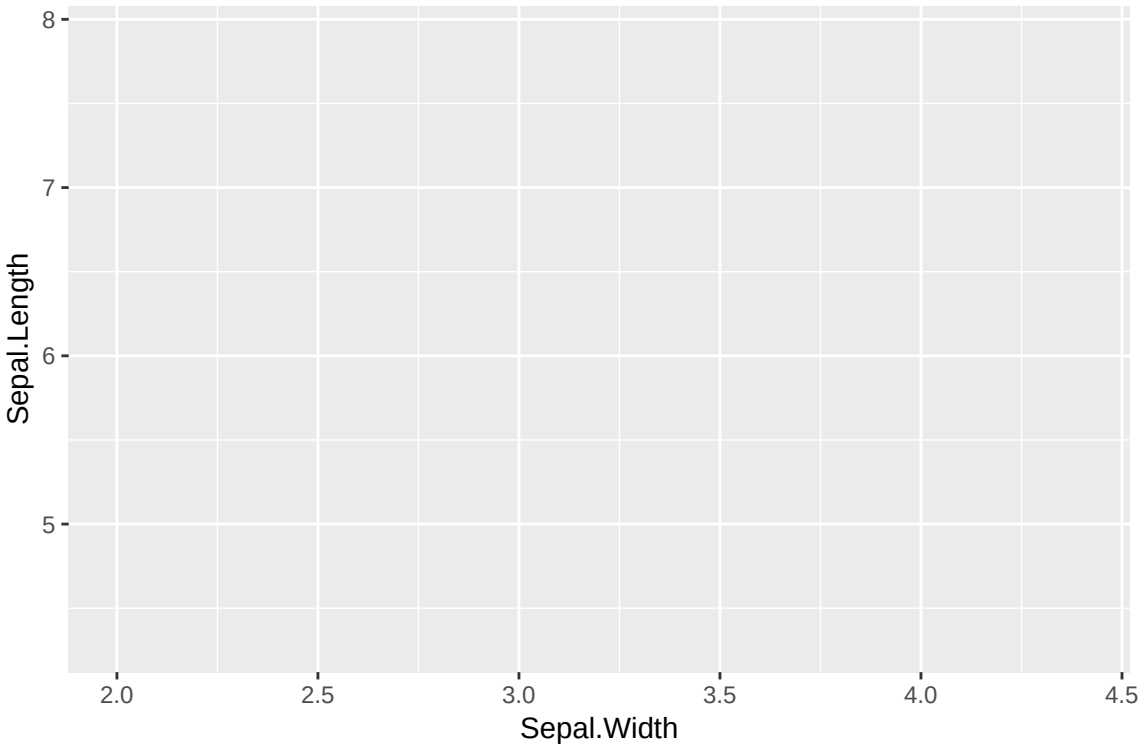

図1　割当を行っただけのグラフ

するとなにもプロットされていないワクが表示されましたが、X と Y のレンジは指定されています。これは与えられたデータに合っており、データ割当だけ行った状態で描けるものが描かれています。

次に階層を追加します。散布図なので、点を描く階層を以下のように追加します。

```
iris %>% ggplot(aes(x = Sepal.Width, y = Sepal.Length, color = Species)) +
  geom_point()
```

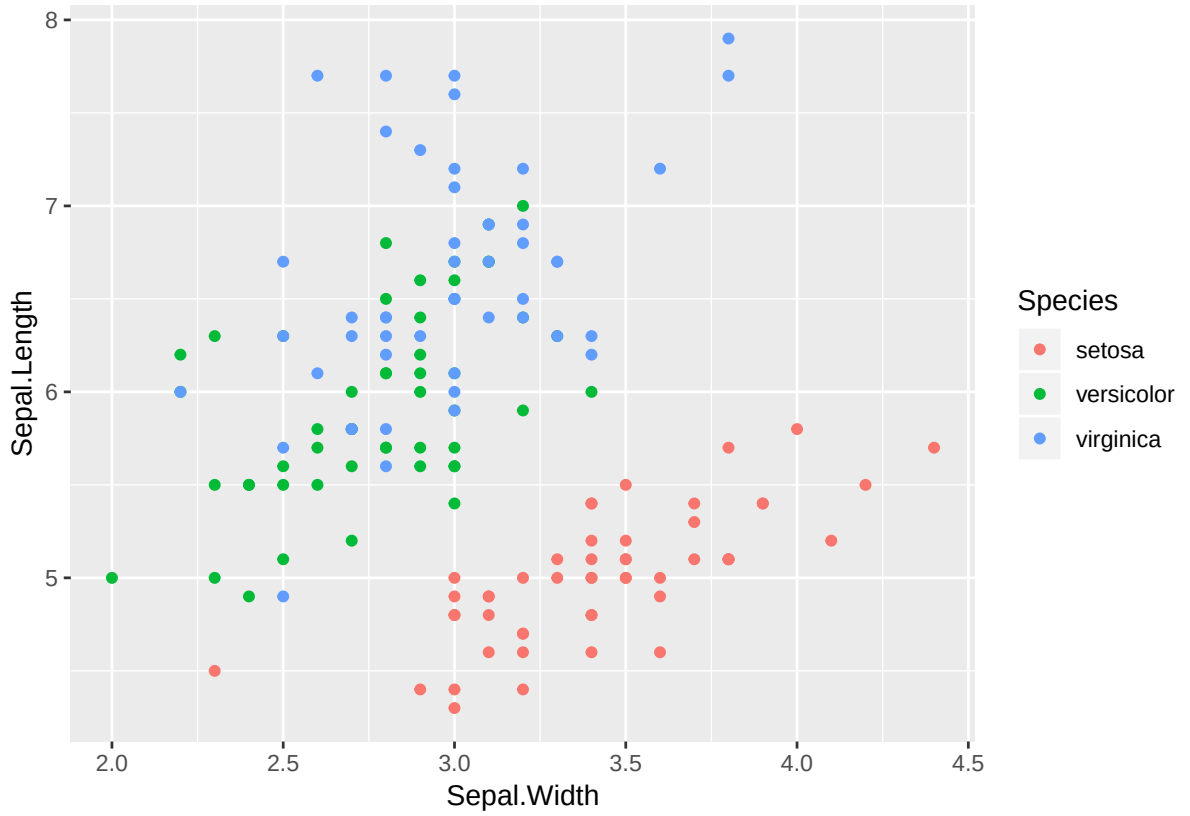

図2　点の階層を追加した

オーバラップしていて分かりにくいので種ごとにサブグラフに分けてみましょう。これも階層として以下のように指定できます。

```
iris %>% ggplot(aes(x = Sepal.Width, y = Sepal.Length, color = Species)) +
  geom_point() +
  facet_wrap(~ Species)
```

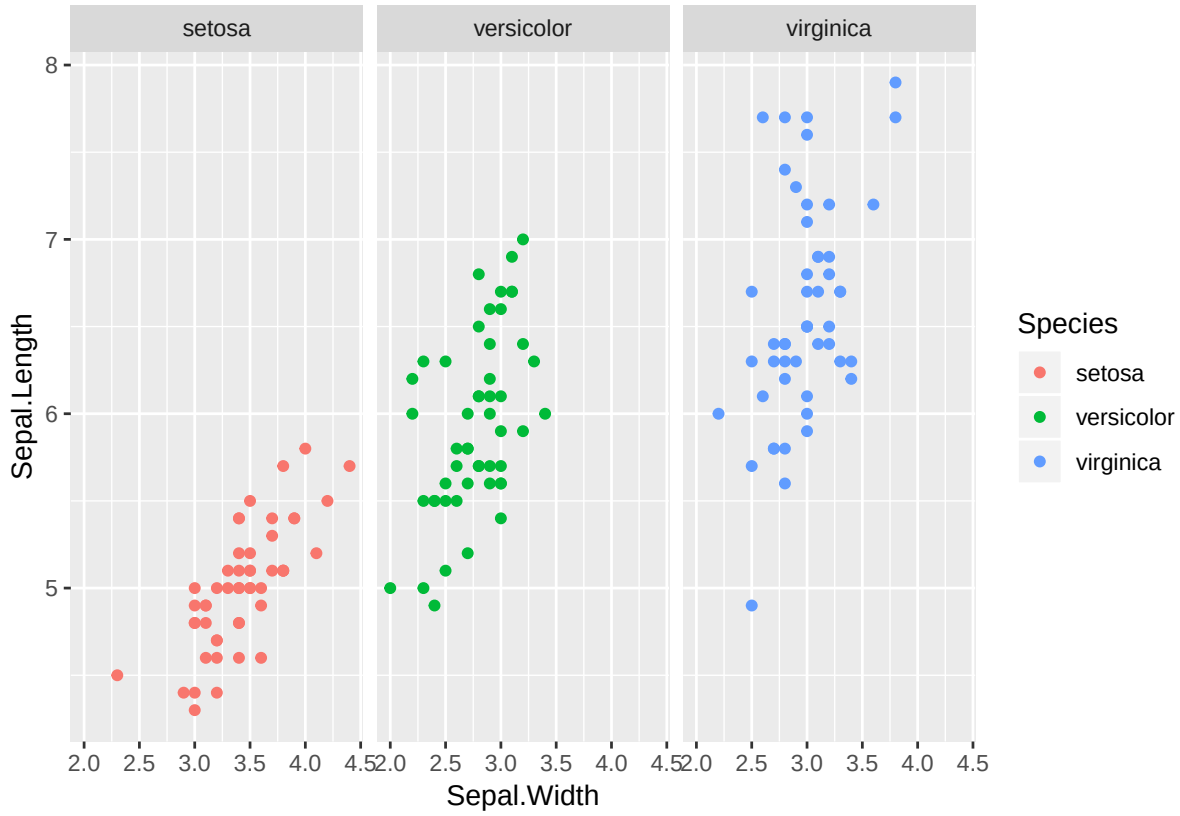

図3　サブグラフの階層を追加した

さらに、近似直線も追加してみましょう。

```
iris %>% ggplot(aes(x = Sepal.Width, y = Sepal.Length, color = Species)) +
  geom_point() +
  geom_smooth(method = "lm") +
  facet_wrap(~ Species)
```

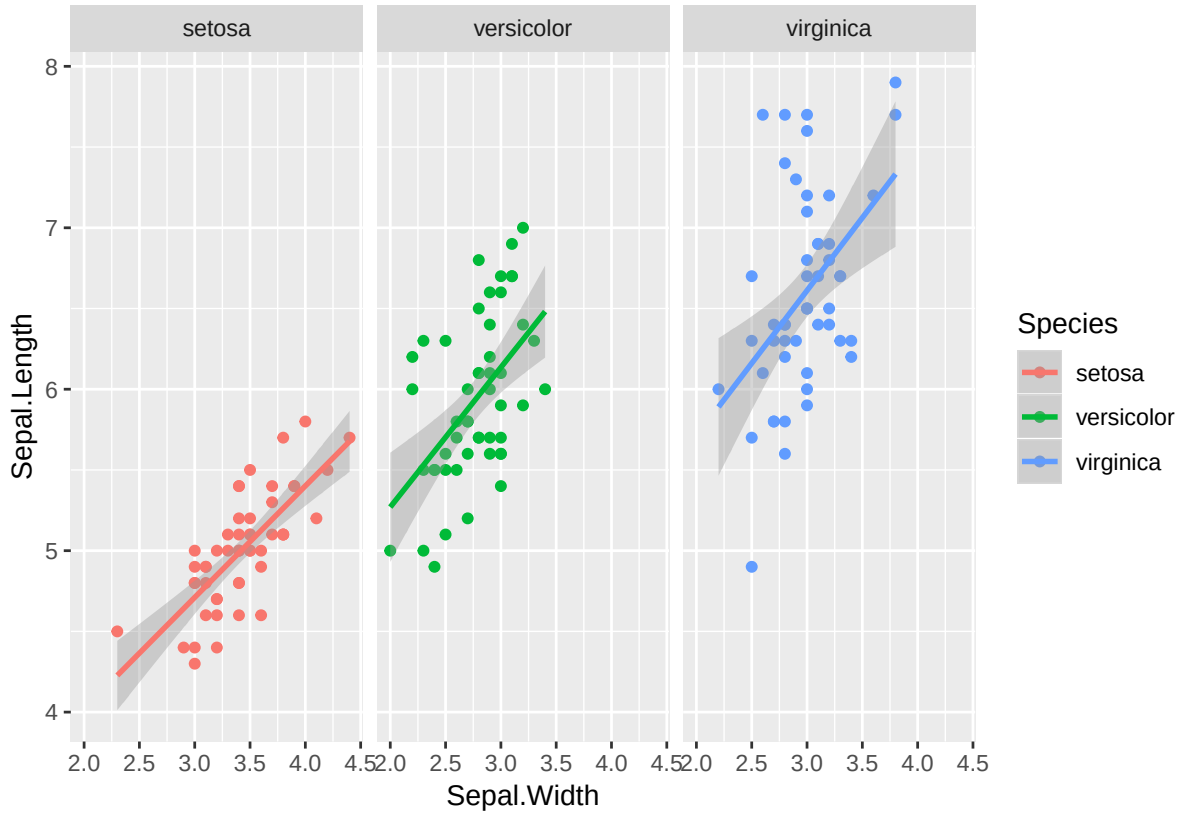

図4　近似直線も追加した

このようにグラフ要素を 1 つづつ重ねてグラフを作成していきます。

花弁のデータもプロットしたいときは、割当を上書きした階層を追加する方法で書けますが、もっと論理的な方法をデータ整形の回に勉強しましょう。[*5]

```
iris %>% ggplot(aes(x = Sepal.Width, y = Sepal.Length, color = Species)) +
  geom_point(aes(shape = "Sepal")) +
  geom_point(aes(x = Petal.Width, y = Petal.Length, shape = "Petal")) +
  facet_wrap(~ Species)
```

[*5] これはよくない例だということを覚えておいてください。また、ここでは更に点のマーカーの形への割当も追加していますが、これも次回説明します。

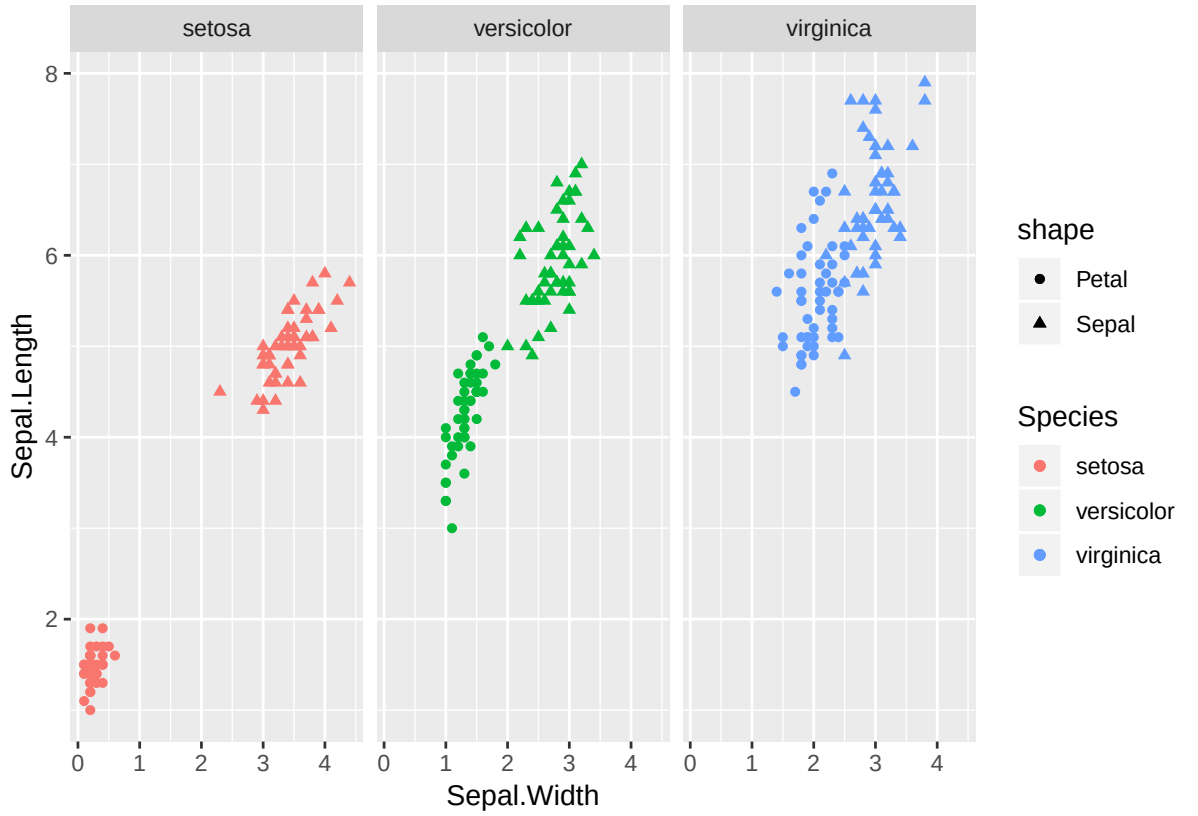

図5　よくない例

グラフ要素には色々な種類があり、ヒストグラムのように Excel だとあらかじめ集計（度数分布表の作成）が必要なものも簡単に作成できます。

```
iris %>% ggplot(aes(x = Sepal.Width)) +
  geom_histogram(binwidth = 0.2) +
  facet_wrap(~ Species)
```

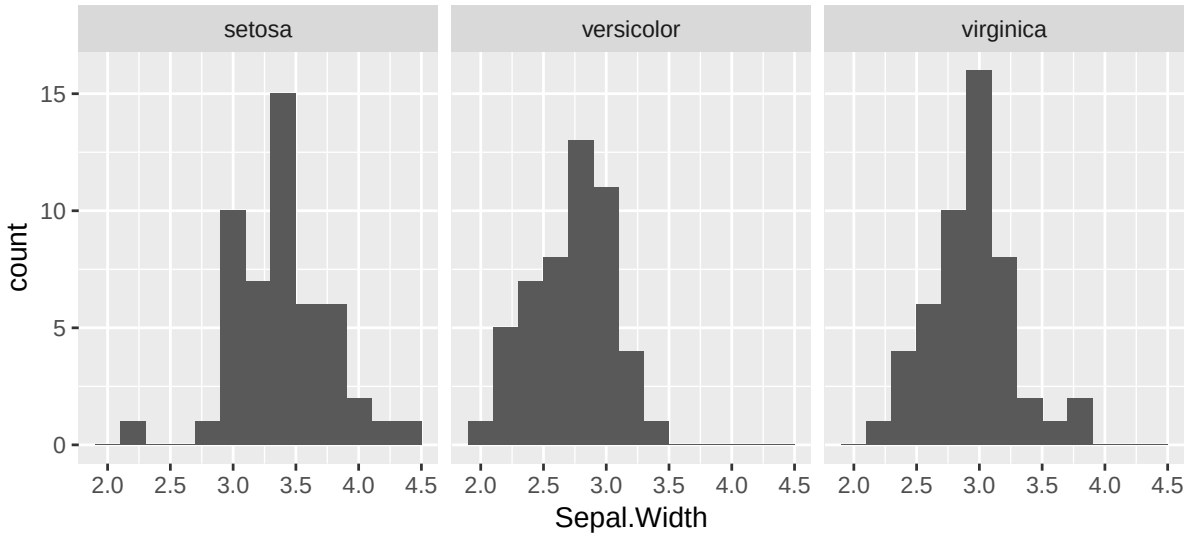

図6　ヒストグラムの階層を追加した

3　次回

次回は、様々なグラフ要素について勉強します。普段使っているグラフの種類にどんなものがあるか、お聞きしながら進めていきたいと思います。

4　R インストール

R を使うために、以下の 3 つのプログラムをインストールしてください。インストーラは以下にあります。

`CoCoDe共有\インストール\プログラミング\統計`

インストールの際の注意点を以下にまとめておきます。

- インストールするときは、インストーラをローカルにコピーしてから起動してください。
- インストール先は C ドライブ直下にしてください。（エラーが起きることがあります。）
- R 本体のみ最新版はインストールしないでください。（他は最新で OK です。）

表2: インストールツール一覧

ツール	説明	現在の対象バージョン
R	R の本体	3.4.4
RStudio	開発環境	1.1.456
Rtoos	サポートツール	35