iris 데이터셋으로 데이터 시각화하기

R의 데이터셋중에 iris라는 데이터셋이 있다. 꽃의 종류에 따라 꽃잎(petal)의 길이와 넓이, 꽃받침(sepal)의 길이와 넓이에 대한 150개의 자료이다.

먼저, 데이터를 불러왔다

```
data("iris")
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          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          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

data.frame의 형태로 만들어져있어서 가공은 쉬운 편이었다. 일단 setosa, versicolor, verginica의 종류별 꽃잎과 꽃받침 값에 대한 평균을 구해보기로 했다.

```
library(dplyr)
mean_iris = iris %>%
  group_by(Species) %>%
  summarise(Sepal.length.mean = mean(Sepal.Length), Sepal.width.mean = mean(Sepal.Width), Petal.length.m

mean_iris
```

```
## # A tibble: 3 x 5
##      Species Sepal.length.mean Sepal.width.mean Petal.length.mean
##       <fctr>             <dbl>            <dbl>             <dbl>
## 1     setosa             5.006            3.428             1.462
## 2 versicolor             5.936            2.770             4.260
## 3  virginica             6.588            2.974             5.552
## # ... with 1 more variables: Petal.width.mean <dbl>
```

- 먼저 summarise에서 일일히 4개의 변수로 만들었는데, 짧게 줄일 수 있는 방법이 있을 것 같다. '질문사항'
- verginica라는 종의 꽃잎, 꽃받침 수치가 모두 높은 것을 알 수 있다. 큰 꽃인듯 하다
- setosa라는 종의 꽃잎이 너무 작아 구글에 검색해보니 해초같은 것이었다.
- 하지만 이 결과를 시각적으로 보여줄 수 있도록 box-plot을 사용하기로 했다.

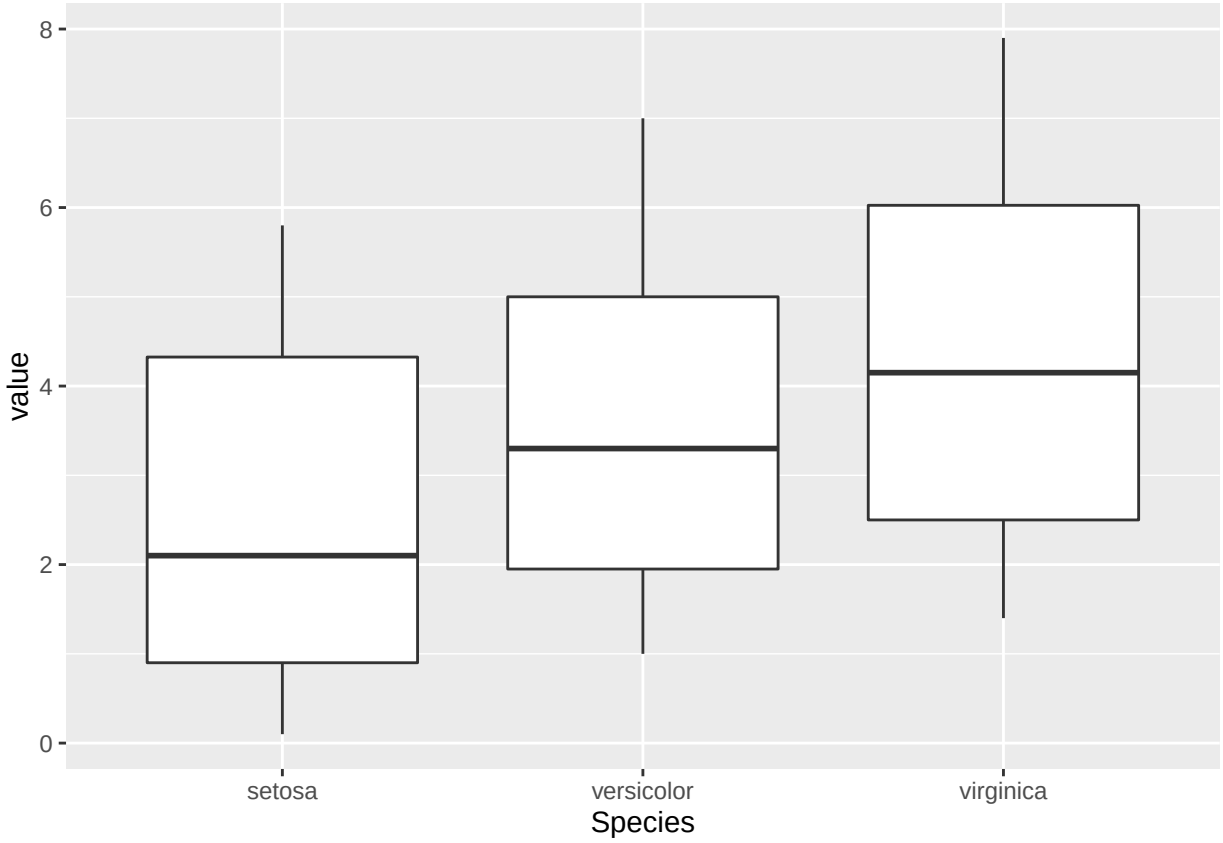
먼저, 데이터를 시각화 가능한 형태로 변환하기 위해 reshape2 라이브러리를 사용하기로 했다. id.vars가 적용되지 않아 한 20분정도 머리를 싸맸는데, *Species*에 따옴표를 붙이지 않아 객체로 인식한 탓에 작동하지 않았다.. 하..

```
library(reshape2)
long_iris = melt(iris, id.vars = "Species")
head(long_iris)
```

```
##   Species     variable value
## 1  setosa Sepal.Length   5.1
## 2  setosa Sepal.Length   4.9
## 3  setosa Sepal.Length   4.7
## 4  setosa Sepal.Length   4.6
## 5  setosa Sepal.Length   5.0
## 6  setosa Sepal.Length   5.4
```

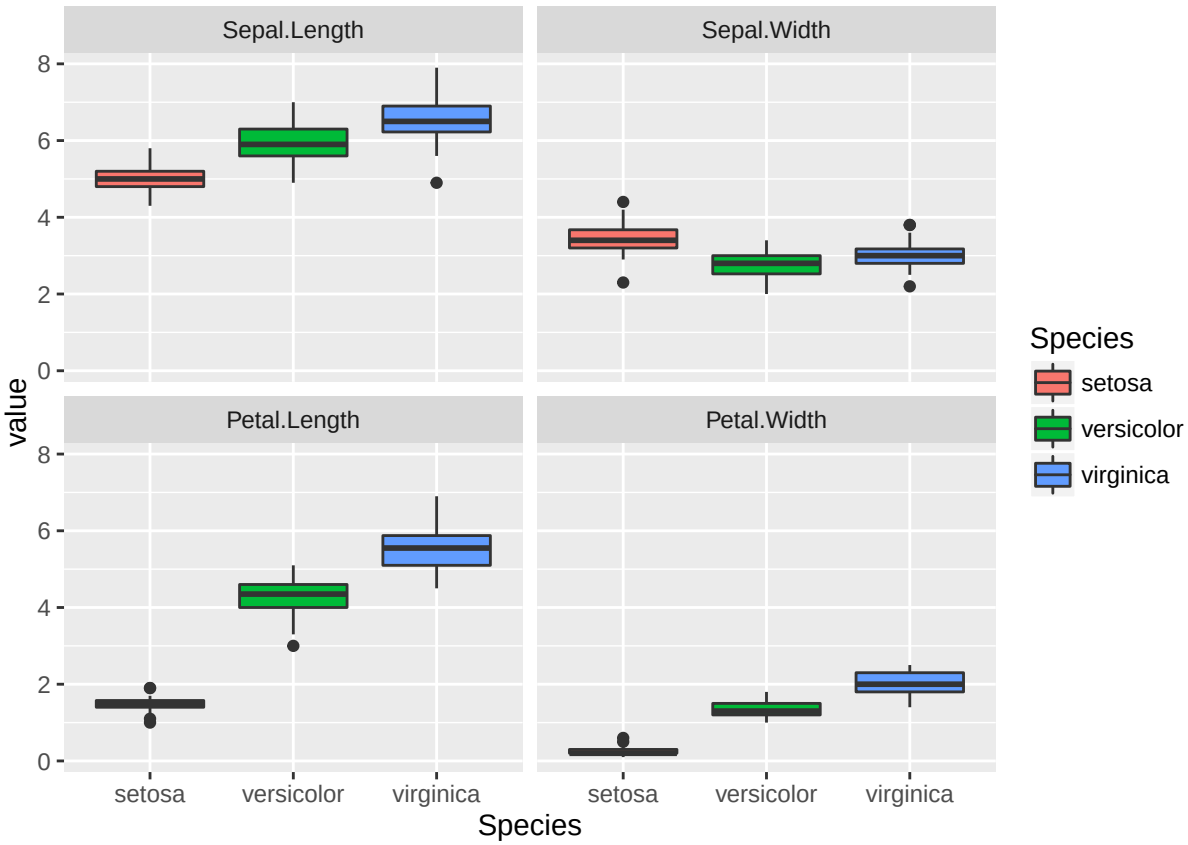
이제 시각화를 해보자. `bar_plot`은 너무 흔하고, `scatter_plot`으로 해보려고 했더니, 데이터 자체가 scatter할 수 있는 것이 아니라는 것을 깨달았다. x축이 수치가 아니라 **종** 또는 **꽃잎,꽃받침**으로 구분되는 변수라서 `box-plot`으로 결정했다.

```
library(ggplot2)
ggplot(long_iris, aes(Species, value)) +
  geom_boxplot()
```



이렇게보니 꽃잎과 꽃받침의 구분이 되지 않은채 모든 값들이 표현되어 의미없는 데이터가 되어버렸다. 그래서 `facet-warp`를 통해 4개로 구분하였다.

```
ggplot(long_iris, aes(Species, value, fill = Species)) +
  geom_boxplot() +
  facet_wrap(~ variable, nrow = 2 )
```



데이터를 2줄로 구분하였고, 종 별로 색깔을 넣어서 밋밋함을 조금이나마 해소해보았다. 색깔을 다양하게 주는 방법은 추가적으로 공부해보아야할 것 같고, box-plot의 의미를 해석하는 방법을 공부해보아야 할 것 같다.