빅데이터를이용한통계그래픽스 HW #3

1678226 유찬미 (Chanmi Yoo)

2019년 10월 16일

###### \* R에 내장되어 있는 iris 자료를 이용하시오. (변수들의 설명은 help 를 참고할 것)

# Load Packages  
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.4.4

## -- Attaching packages ------------------------------------ tidyverse 1.2.1 --

## √ ggplot2 3.0.0 √ purrr 0.2.5  
## √ tibble 1.4.2 √ dplyr 0.7.6  
## √ tidyr 0.8.1 √ stringr 1.3.1  
## √ readr 1.2.1 √ forcats 0.4.0

## Warning: package 'ggplot2' was built under R version 3.4.4

## Warning: package 'tibble' was built under R version 3.4.4

## Warning: package 'tidyr' was built under R version 3.4.4

## Warning: package 'readr' was built under R version 3.4.4

## Warning: package 'purrr' was built under R version 3.4.4

## Warning: package 'dplyr' was built under R version 3.4.4

## Warning: package 'stringr' was built under R version 3.4.4

## Warning: package 'forcats' was built under R version 3.4.4

## -- Conflicts --------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

# Load Data  
data(iris)  
# 변수 설명 확인   
help(iris)

## starting httpd help server ...

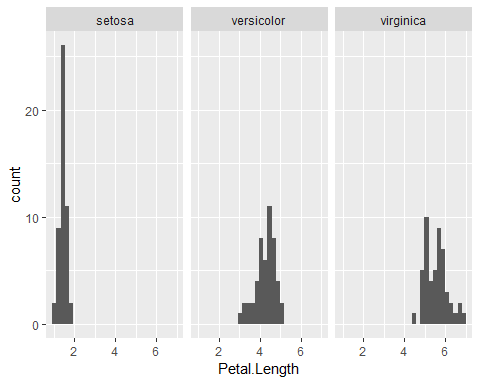
## done

continuous variables: Sepal.Length, Sepal.Width, Petal.Length, Petal.Width  
categorical variables: Species

###### 1. iris자료에서 Species 별로 Petal.Length의 분포를 알아보려고 한다. 이를 위한 다양한 그림을 그리고 설명하시오.

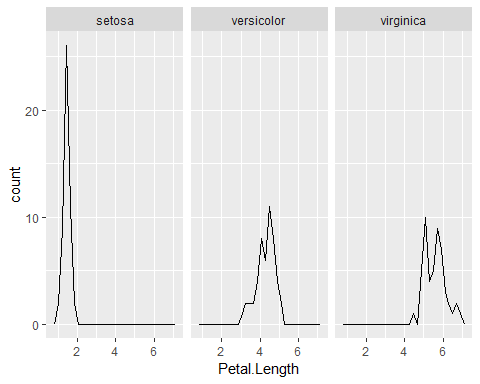
ggplot(iris, aes(Petal.Length)) +  
 geom\_histogram() +  
 facet\_wrap(~Species)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

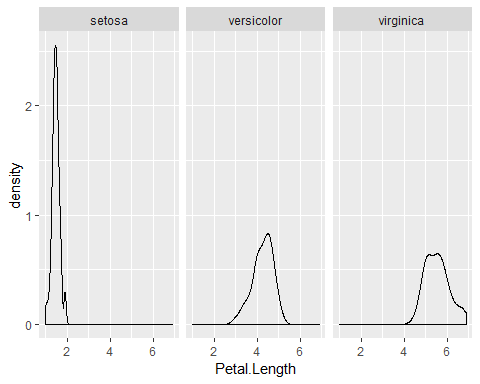


ggplot(iris, aes(Petal.Length)) +  
 geom\_freqpoly() +  
 facet\_wrap(~Species)

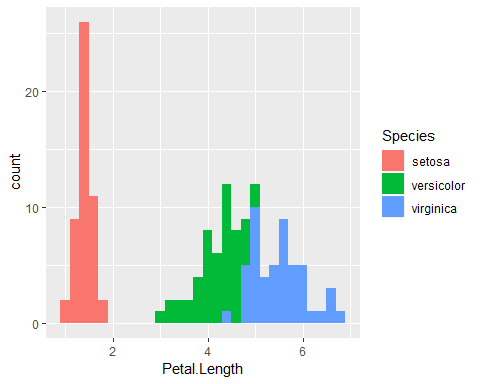
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



ggplot(iris, aes(Petal.Length)) +  
 geom\_density() +  
 facet\_wrap(~Species)



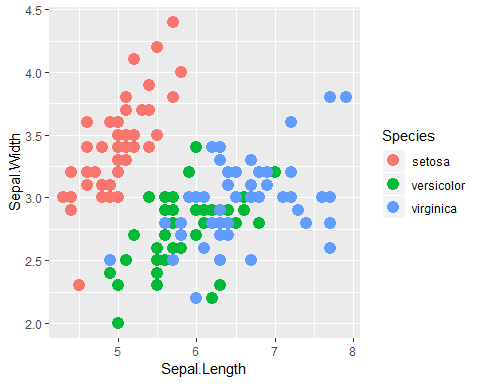
ggplot(iris, aes(Petal.Length, fill=Species)) +  
 geom\_histogram(binwidth = 0.2)



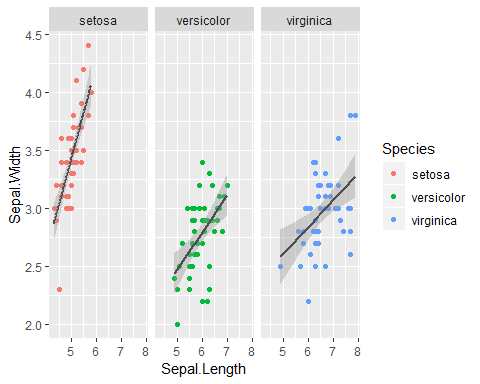
[설명] 위의 그림을 통해 iris 자료의 Species는 setosa, versicolor, virginica 총 3가지 종류가 있으며, Species별로 꽃받침 길이가 다르다는 것을 알 수 있다.  
1) setosa는 짧은 편(1 정도)  
2) versicolor는 중간 길이(4 정도)  
3) virginica는 긴 편(5.5 정도)

###### 2. Iris 자료에서 Sepal.Length와 Sepal.Width의 관계가 Species에 따라 어떻게 달라지는지를 알아보려고 한다. 이를 위한 그림을 그리고 설명하시오.

ggplot(iris, aes(Sepal.Length, Sepal.Width, color=Species)) +  
 geom\_point(size=4)



ggplot(iris, aes(Sepal.Length, Sepal.Width)) +  
 geom\_point(aes(color=Species)) +  
 geom\_smooth(method="lm", color = "grey30") +  
 facet\_wrap(~Species)



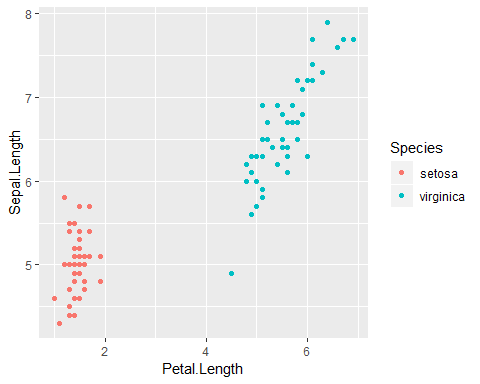
[설명] Species별로 꽃받침 길이와 꽃받침 너비는 다른 관계 양상을 보인다.  
1) setosa의 경우, outlier가 하나 있지만 꽃받침의 길이와 너비가 양의 선형 관계에 있음을 알 수 있다(직선 모양에 가까우며 표준편차가 작음).  
2) versicolor와 virginica의 경우, 꽃받침의 길이와 너비가 선형적인 관계에 있지 않은 것으로 보인다.

###### 3. Species 중 “versicolor”인 자료를 VERSICOLOR에 저장하시오. 그리고 “versicolor” 를 제외한 자료를 SETVIR에 저장하시오.

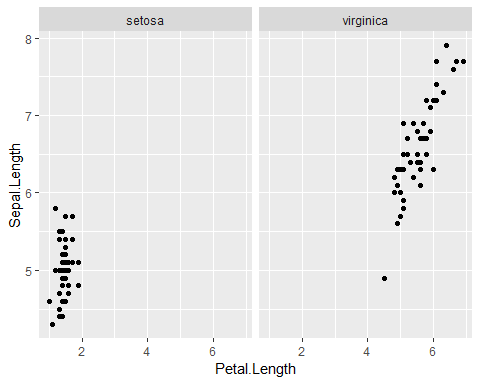
VERSICOLOR <- subset(iris, iris$Species == "versicolor")  
SETVIR <- iris[iris$Species != "versicolor", ]

###### 4. SETVIR자료를 이용하여 Species별로 Petal.Length와 Sepal.Length의 관계를 비교하려고 한다. 이를 위한 그림을 그리고 설명하시오.

ggplot(SETVIR, aes(Petal.Length, Sepal.Length, color = Species)) +  
 geom\_point()



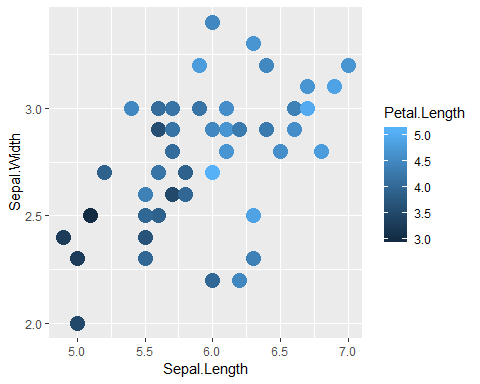
ggplot(SETVIR, aes(Petal.Length, Sepal.Length)) +  
 geom\_point() +  
 facet\_wrap(~Species)



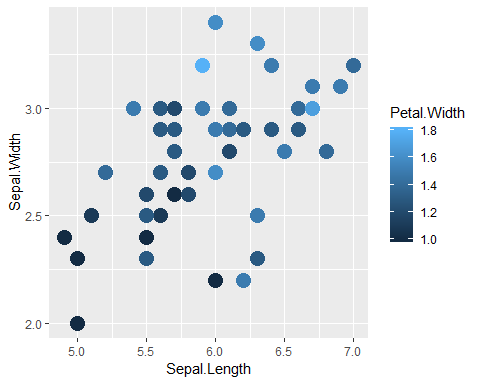
[설명] SERVIR 자료에는 Species가 setosa와 virginica, 2종류가 있다.  
1) setosa는 꽃잎 길이와 꽃받침 길이가 각각 2 이하, 6이하로 짧으며, 두 변수 간 어떠한 선형 관계도 찾기 어렵다.  
2) 한편 versicolor는 꽃잎 길이와 꽃받침 길이가 모두 길며, 두 변수 간 양적 선형 관계가 있음을 알 수 있다.

###### 5. VERSICOLOR 자료에서 네 변수간의 관계를 비교하려고 한다. 이를 위한 그림을 그리고 설명하시오.

par(mfrow=c(2,1))  
  
ggplot(VERSICOLOR, aes(Sepal.Length, Sepal.Width)) +  
 geom\_point(aes(color = Petal.Length), size=5)



ggplot(VERSICOLOR, aes(Sepal.Length, Sepal.Width)) +  
 geom\_point(aes(color = Petal.Width), size=5)



library(GGally)

## Warning: package 'GGally' was built under R version 3.4.4

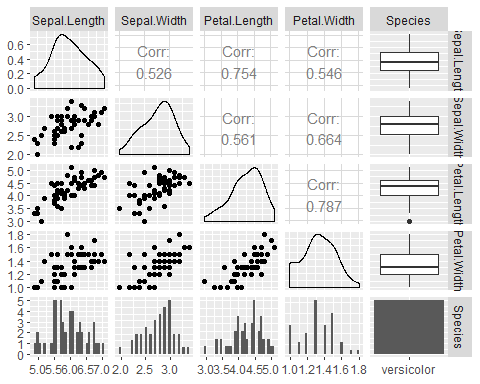
##   
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':  
##   
## nasa

ggpairs(VERSICOLOR)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



[설명]  
1) VERSICOLOR 자료에서 꽃받침 길이-꽃받침 너비는 양의 선형 관계를 보이며, 이때 꽃잎 길이는 비례하는 경향이 있지만 꽃잎 너비는 특별한 패턴을 보인다고 보기 어렵다.  
2) 산점도 행렬을 살펴보면, 4가지 변수들의 관계에서 꽃받침 길이-꽃잎 길이(0.754), 꽃받침 너비-꽃잎 너비(0.664), 꽃잎 길이-꽃잎 너비(0.787)에서 양의 선형 관계가 나타남을 알 수 있다.

###### \* R에 내장되어 있는 Theoph 자료는 12명의 사람으로부터 시간별로 측정한 약물의 혈중농도에 관한 자료이다.

###### - Subject : 12명의 ID

###### - Wt : 몸무게

###### - Dose : 복용양

###### - Time : 시간

###### - conc : 혈중농도

###### 6. Theoph 자료를 이용하여 12명의 시간별 혈중농도 곡선을 그리시오.

data(Theoph)  
  
ggplot(Theoph, aes(Time, conc, color=Subject)) +  
 geom\_line(aes(group = Subject))

