


통계패키지활용 자료분석(2024년 겨울학기)			
담당 교수 : 김 태 수			
강좌 번호	100961-31001	본인의 과제 자체 평가	9/10 점
과제명 : 제 2차 자료시각화 with ggplot2			

이름	이지원
	

제 출 일	2024년 01월 14일
학 과	산업공학과 ITM전공
학 번	22102009

목차

1. R의 자료 iris를 이용한 대한 데이터 시각화 실습

- (1) iris dataset에 대한 설명
- (2) 실습목표
- (3) iris dataset 살펴보기
- (4) iris dataset 시각화
- (5) 결론

2. R의 자료 diamond를 이용한 대한 데이터 시각화 실습

- (1) diamond dataset에 대한 설명
- (2) 실습목표
- (3) diamond dataset 살펴보기
- (4) diamond dataset 시각화
- (5) 결론

1. R의 자료 iris를 이용한 대한 데이터 시각화 실습

(1) iris dataset에 대한 설명

- R에 기본적으로 내장된 데이터셋인 iris는 붓꽃(iris)에 대한 데이터셋이다.
- setosa, versicolor, virginica 이 3가지의 품종으로 분류되어 있다.
- 포함된 변수는 아래와 같다.
 - Sepal.Length: 꽃받침의 길이
 - Sepal.Width: 꽃받침의 너비
 - Petal.Length: 꽃잎의 길이
 - Petal.Width: 꽃잎의 너비
 - Species: 붓꽃의 종류
- 자료의 총 개수는 150개이다.

(2) 실습 목표

- iris 데이터를 기반으로 각 품종별로 나타난 특징을 시각화하여 분석할 수 있다.
- 붓꽃의 종류에 따른 4가지 데이터(꽃받침의 길이, 꽃받침의 너비, 꽃잎의 길이, 꽃잎의 너비)를 시각적으로 표현하는 그래프를 생성할 수 있다.
- 붓꽃의 종류에 따른 4가지 데이터의 상관성을 명확하게 전달할 수 있는 시각화 과정을 수행할 수 있다.

(3) iris dataset 살펴보기

- iris의 데이터 구조

```
> str(iris)
'data.frame': 150 obs. of 5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species      : Factor w/ 3 levels "setosa","versicolor",...: 1 1 1 1 1 1 1 1 1 1 ...
```

- iris 데이터 요약

```
> summary(iris)
  Sepal.Length   Sepal.Width   Petal.Length   Petal.Width      Species
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

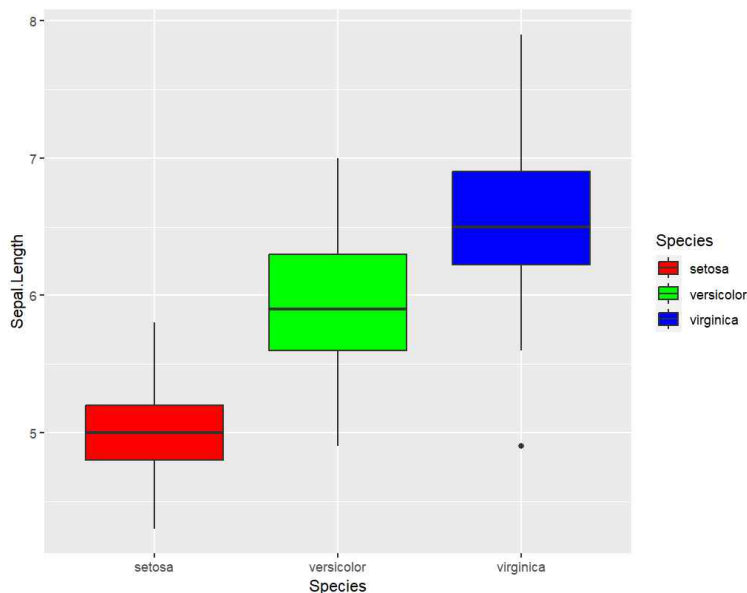
- 품종별 Sepal.Length, Sepal.Width, Petal.Length, Petal.Width에 대한 통계량 요약

```
> iris %>% group_by(Species) %>% summarise(mean = mean(Sepal.Length), sd = sd(Sepal.Length), cnt = n())
# A tibble: 3 x 4
  Species    mean    sd    cnt
  <fct>    <dbl> <dbl> <int>
1 setosa    5.01  0.352    50
2 versicolor 5.94  0.516    50
3 virginica 6.59  0.636    50
> iris %>% group_by(Species) %>% summarise(mean = mean(Sepal.Width), sd = sd(Sepal.Width), cnt = n())
# A tibble: 3 x 4
  Species    mean    sd    cnt
  <fct>    <dbl> <dbl> <int>
1 setosa    3.43  0.379    50
2 versicolor 2.77  0.314    50
3 virginica 2.97  0.322    50
> iris %>% group_by(Species) %>% summarise(mean = mean(Petal.Length), sd = sd(Petal.Length), cnt = n())
# A tibble: 3 x 4
  Species    mean    sd    cnt
  <fct>    <dbl> <dbl> <int>
1 setosa    1.46  0.174    50
2 versicolor 4.26  0.470    50
3 virginica 5.55  0.552    50
> iris %>% group_by(Species) %>% summarise(mean = mean(Petal.Width), sd = sd(Petal.Width), cnt = n())
# A tibble: 3 x 4
  Species    mean    sd    cnt
  <fct>    <dbl> <dbl> <int>
1 setosa    0.246  0.105    50
2 versicolor 1.33  0.198    50
3 virginica 2.03  0.275    50
```

(3) iris dataset 시각화

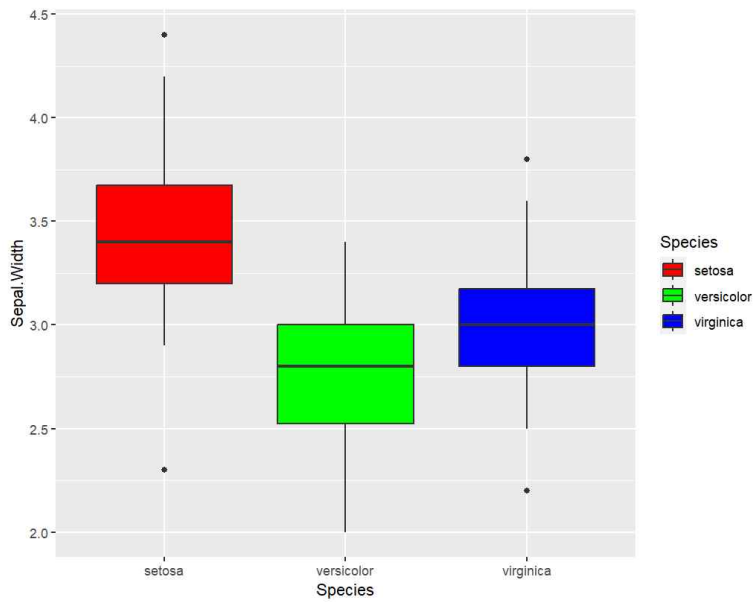
- 품종에 따른 Sepal.Length(꽃받침 길이) 분포 시각화

```
> graphSL <- ggplot(iris, aes(x = Species, y = Sepal.Length, fill = Species)) +
+   geom_boxplot() +
+   scale_fill_manual(values = c("red", "green", "blue"))
> graphSL
```



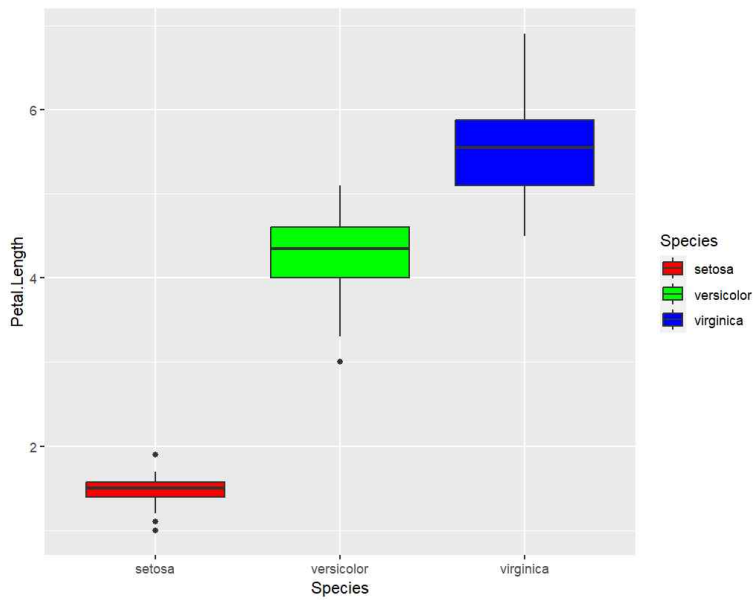
- 품종에 따른 Sepal.Width(꽃받침 길이) 분포 시각화

```
> graphSW <- ggplot(iris, aes(x = Species, y = Sepal.Width, fill = Species)) +
+   geom_boxplot() +
+   scale_fill_manual(values = c("red", "green", "blue"))
> graphSW
```



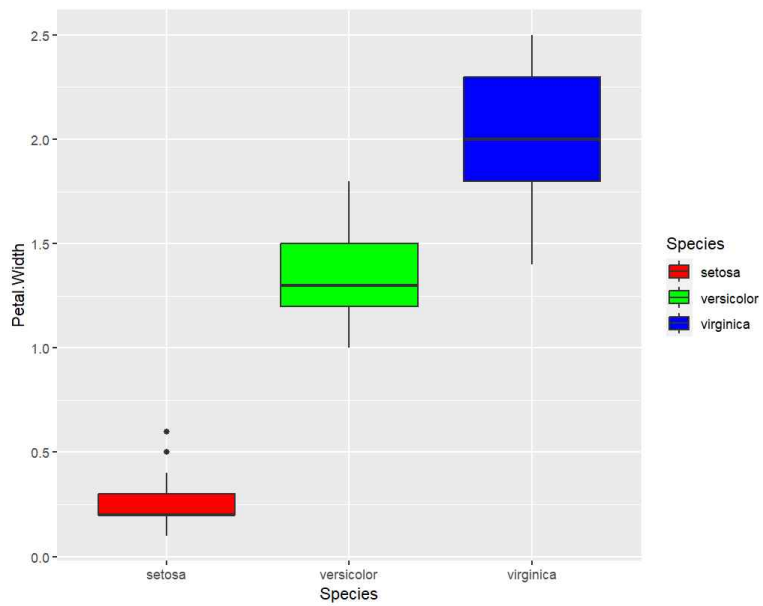
- 품종에 따른 Petal.Length(꽃잎 길이) 분포 시각화

```
> graphPL <- ggplot(iris, aes(x = Species, y = Petal.Length, fill = Species)) +
+   geom_boxplot() +
+   scale_fill_manual(values = c("red", "green", "blue"))
> graphPL
```



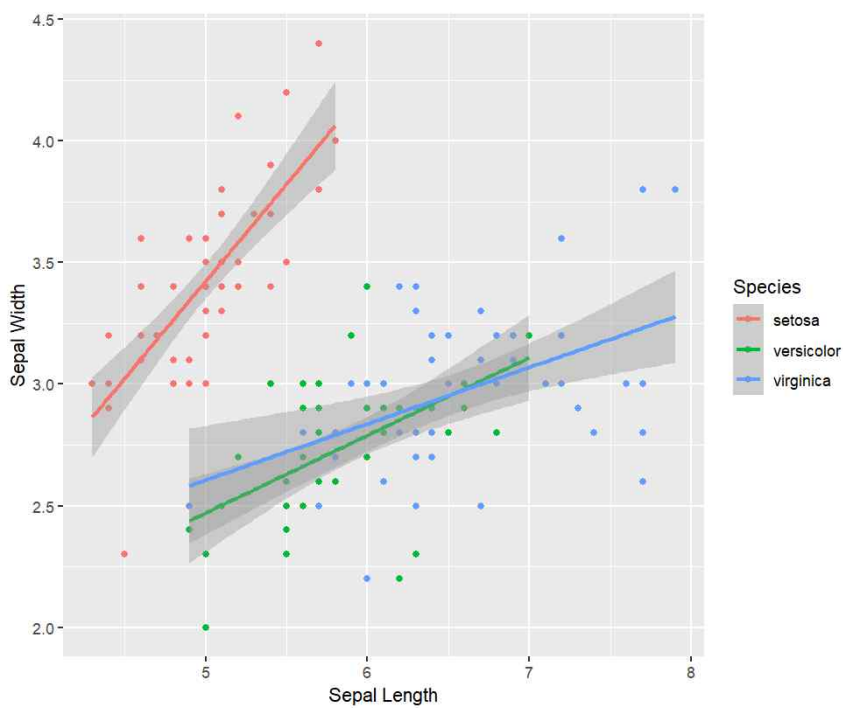
- 품종에 따른 Petal.Width(꽃잎 너비) 분포 시각화

```
> graphPW <- ggplot(iris, aes(x = Species, y = Petal.Width, fill = Species)) +
+   geom_boxplot() +
+   scale_fill_manual(values = c("red", "green", "blue"))
> graphPW
```



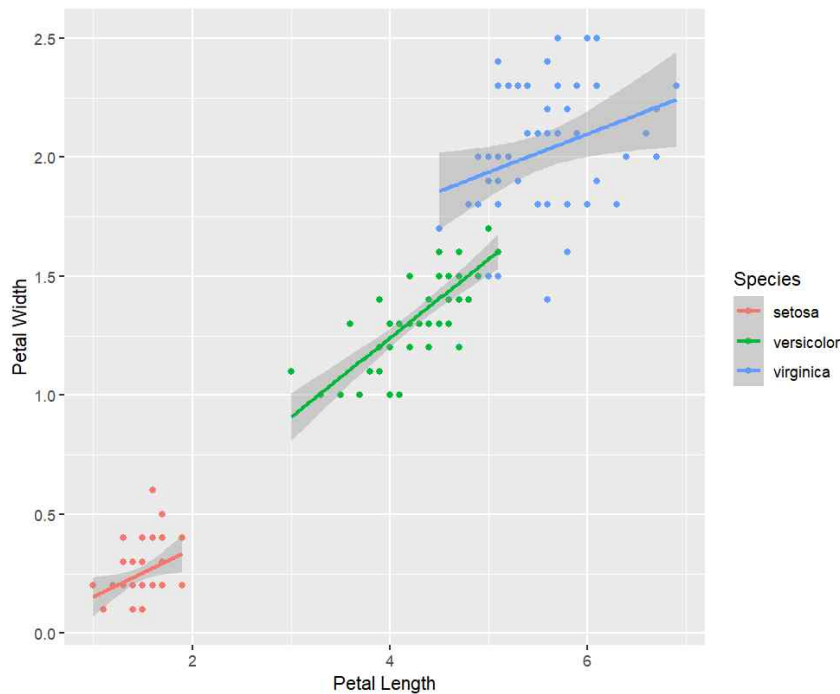
- Sepal.Length와 Sepal.Width 상관관계 시각화

```
> slsw <- ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width, group = Species,
color = species)) +
+   geom_point() +
+   geom_smooth(method = "lm") +
+   labs(x = "Sepal Length", y = "Sepal Width")
> slsw
```



- Petal.Length와 Petal.Width 상관관계 시각화

```
> plpw <- ggplot(iris, aes(x = Petal.Length, y = Petal.Width, group = Species,  
color = species)) +  
+   geom_point() +  
+   geom_smooth(method = "lm") +  
+   labs(x = "Petal Length", y = "Petal Width")
```



(5) 결론

- 종에 따른 꽃받침 길이와 너비, 꽃잎 길이와 너비 비교

- 꽃받침의 길이는 virginica > versicolor > setosa 순으로 크다.
- 꽃받침의 너비는 setosa > virginica > versicolor 순으로 크다.
- 꽃잎의 길이는 virginica > versicolor > setosa 순으로 크다.
- 꽃잎의 너비는 virginica > versicolor > setosa 순으로 크다.

- 꽃받침의 길이와 너비의 상관관계

- 꽃받침의 길이(Sepal.Length)가 증가할수록 꽃받침의 너비(Sepal.Width)가 증가하는 양의 상관관계를 가진다.

- virginica > versicolor > setosa 순으로 꽃받침의 길이와 너비 사이의 더 큰 상관관계를 가진다.

- 꽃잎의 길이와 너비의 상관관계

- 꽃잎의 길이(Petal.Length)가 증가할수록 꽃잎의 너비(Petal.Width)가 증가하는 양의 상관관계를 가진다.

- versicolor > setosa > virginica 순으로 꽃받침의 길이와 너비 사이의 더 큰 상관관계를 가진다.

2. R의 자료 diamonds를 이용한 대한 데이터 시각화 실습

(1) diamonds dataset에 대한 설명

- R에 기본적으로 내장된 데이터셋인 diamonds는 다이아몬드의 속성과 가격에 대한 정보를 포함하고 있다.
- 포함된 변수는 아래와 같다.
 - carat: 다이아몬드의 무게를 나타내는 연속형 변수
 - cut: 다이아몬드의 절단 품질을 나타내는 범주형 변수
 - color: 다이아몬드의 색상을 나타내는 범주형 변수
 - clarity: 다이아몬드의 투명도를 나타내는 범주형 변수
 - depth: 다이아몬드의 깊이를 나타내는 연속형 변수.
 - table: 다이아몬드의 상단 표면의 너비를 밑바탕으로 한 백분율을 나타내는 연속형 변수
 - price: 다이아몬드의 가격을 나타내는 연속형 변수.
 - x, y, z: 다이아몬드의 길이, 너비, 높이를 나타내는 연속형 변수.
- 자료의 총 개수는 53940개이다.

(2) 실습 목표

- diamonds 데이터를 기반으로 ggplot2를 이용한 시각화 작업을 통해 원하는 결론을 도출할 수 있다.

(3) diamonds dataset 살펴보기

- diamonds 의 데이터 구조

```
> str(diamonds)
tibble [53,940 × 10] (S3: tbl_df/tbl/data.frame)
 $ carat   : num [1:53940] 0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
 $ cut     : Ord.factor w/ 5 levels "Fair"<"Good"<...: 5 4 2 4 2 3 3 3 1 3 ...
 $ color   : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<...: 2 2 2 6 7 7 6 5 2 5 ...
 $ clarity : Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<...: 2 3 5 4 2 6 7 3 4 5 ...
 $ depth   : num [1:53940] 61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
 $ table   : num [1:53940] 55 61 65 58 58 57 57 55 61 61 ...
 $ price   : int [1:53940] 326 326 327 334 335 336 336 337 337 338 ...
 $ x       : num [1:53940] 3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
 $ y       : num [1:53940] 3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
 $ z       : num [1:53940] 2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
```

- diamonds 데이터 요약


```
> summary(diamonds)
```

carat		cut	color	clarity	depth
Min.	:0.2000	Fair : 1610	D: 6775	SI1 :13065	Min. :43.00
1st Qu.	:0.4000	Good : 4906	E: 9797	VS2 :12258	1st Qu.:61.00
Median	:0.7000	Very Good:12082	F: 9542	SI2 : 9194	Median :61.80
Mean	:0.7979	Premium :13791	G:11292	VS1 : 8171	Mean :61.75
3rd Qu.	:1.0400	Ideal :21551	H: 8304	VVS2 : 5066	3rd Qu.:62.50
Max.	:5.0100		I: 5422	VVS1 : 3655	Max. :79.00
			J: 2808	(Other): 2531	

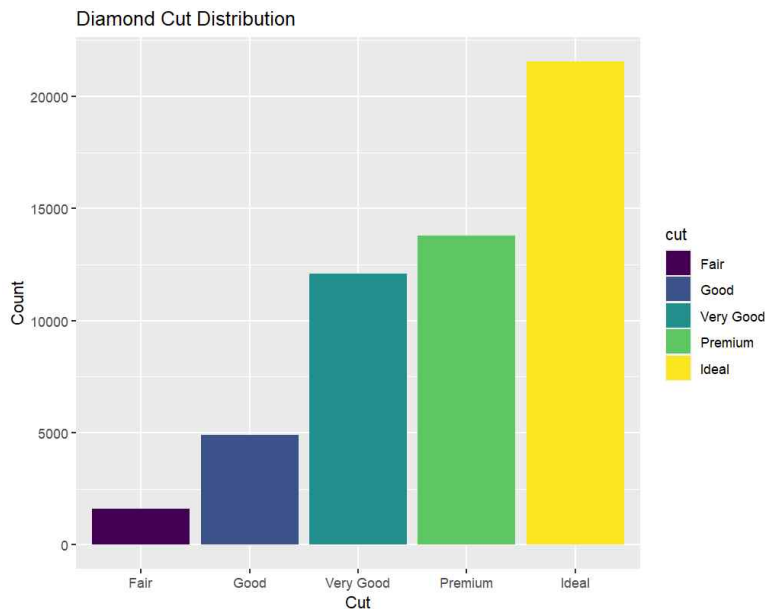
table	price	x	y	z
Min. :43.00	Min. : 326	Min. : 0.000	Min. : 0.000	Min. : 0.000
1st Qu.:56.00	1st Qu.: 950	1st Qu.: 4.710	1st Qu.: 4.720	1st Qu.: 2.910
Median :57.00	Median : 2401	Median : 5.700	Median : 5.710	Median : 3.530
Mean :57.46	Mean : 3933	Mean : 5.731	Mean : 5.735	Mean : 3.539
3rd Qu.:59.00	3rd Qu.: 5324	3rd Qu.: 6.540	3rd Qu.: 6.540	3rd Qu.: 4.040
Max. :95.00	Max. :18823	Max. :10.740	Max. :58.900	Max. :31.800

(3) diamonds dataset 시각화

- 다이아몬드 등급 분포

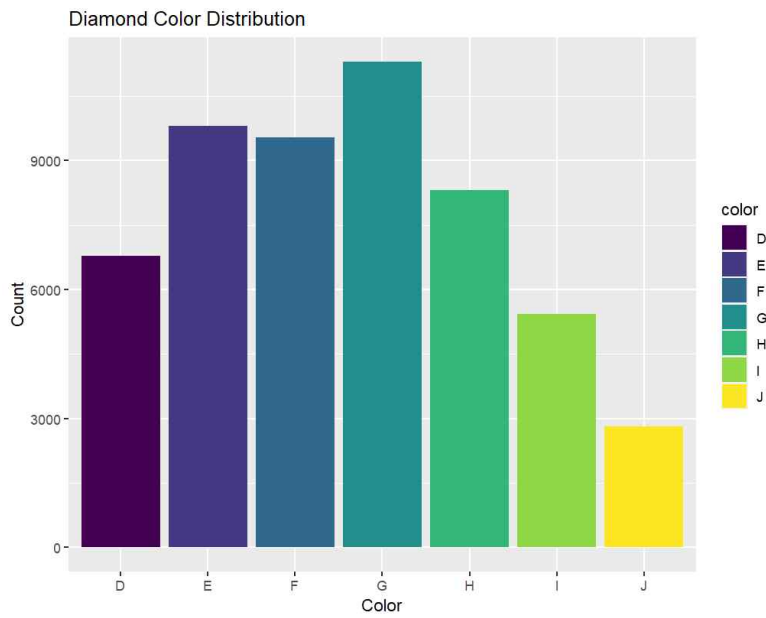
- cut 기준

```
> Cut <- ggplot(diamonds, aes(x = cut, fill = cut)) +
+   geom_bar() +
+   labs(title = "Diamond Cut Distribution", x = "Cut", y = "Count")
>
> Cut
```



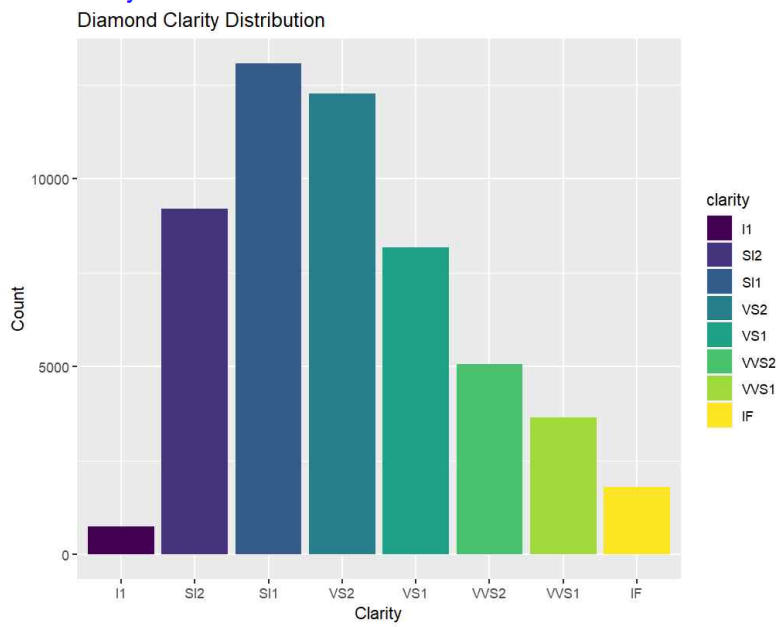
- Color 기준

```
> Color <- ggplot(diamonds, aes(x = color, fill = color)) +
+   geom_bar() +
+   labs(title = "Diamond Color Distribution", x = "Color", y = "Count")
> Color
```



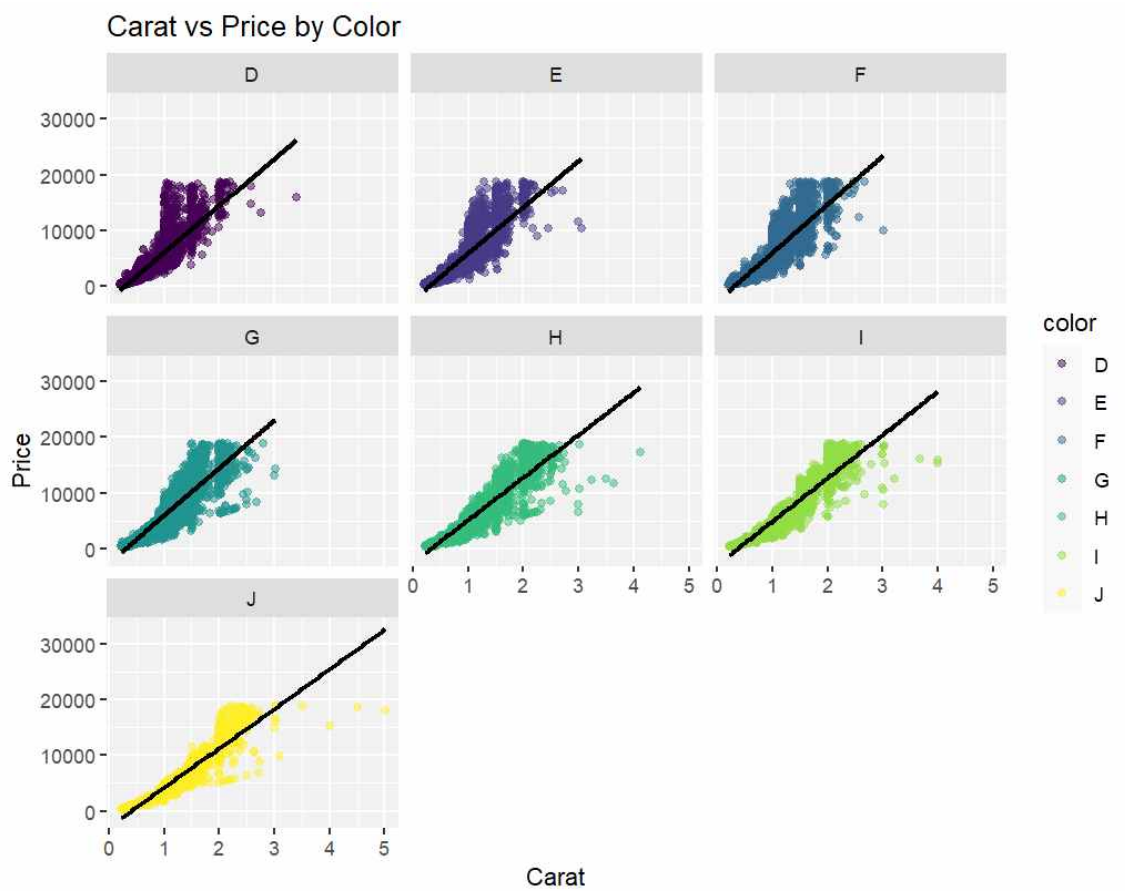
- Clarity 기준

```
> Clarity <- ggplot(diamonds, aes(x = clarity, fill = clarity)) +
+   geom_bar() +
+   labs(title = "Diamond Clarity Distribution", x = "clarity", y = "Count")
> Clarity
```



- 다이아몬드 색상별 Carat에 따른 가격분포

```
> CpC <- ggplot(diamonds, aes(x = carat, y = price, color = color)) +
+   geom_point(alpha = 0.5) +
+   geom_smooth(method = "lm", color = "black") +
+   labs(title = "Carat vs Price by Color", x = "Carat", y = "Price") +
+   facet_wrap(~color)
> CpC
```



(5) 결론.

- 다이아몬드 등급 분포
 - Cut 기준 ideal > premium > very good > good > fair 순으로 많다.
 - Color 기준 G > E > F > H > I > D > I 순으로 많다.
 - Clarity 기준