Point: 단계별 데이터 이름을 다르게 지정하여 각 단계의 데이터를 보존하도록 함

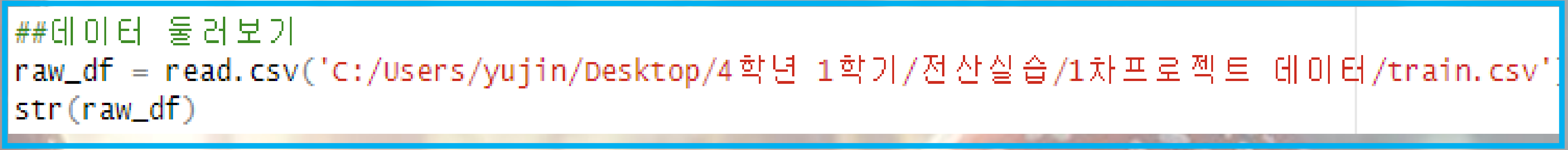
"독성 버섯의 이진예측"

분석 목표: 버섯 특성 데이터를 기반으로 독성 여부를 예측하고 중요한 변수를 식별

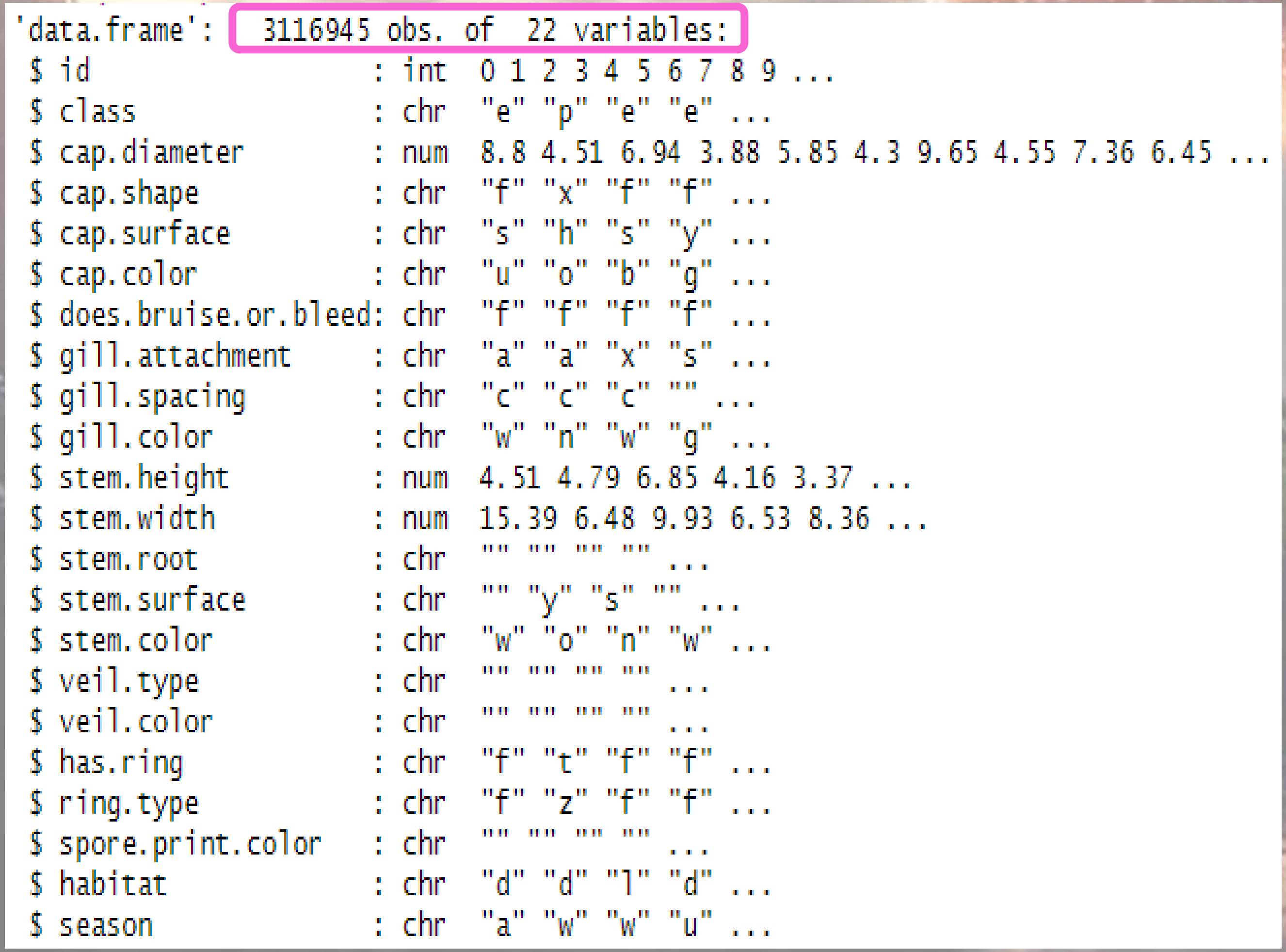
데이터:
https://www.kaggle.com/competitions/playground-series-s4e8
"Binary Prediction of Poisonous Mushrooms"

202110456
강유진

1.데이터 로드



```
##데이터 둘러보기
raw_df = read.csv('C:/Users/yujin/Desktop/4학년 1학기/전산실습/1차프로젝트 데이터/train.csv'
str(raw_df)
```



```
'data.frame':	3116945 obs. of  22 variables:
 $ id                  : int  0 1 2 3 4 5 6 7 8 9 ...
 $ class               : chr  "e" "p" "e" "e" ...
 $ cap.diameter        : num  8.8 4.51 6.94 3.88 5.85 4.3 9.65 4.55 7.36 6.45 ...
 $ cap.shape           : chr  "f" "x" "f" "f" ...
 $ cap.surface         : chr  "s" "h" "s" "y" ...
 $ cap.color           : chr  "u" "o" "b" "g" ...
 $ does.bruise.or.bleed: chr  "f" "f" "f" "f" ...
 $ gill.attachment     : chr  "a" "a" "x" "s" ...
 $ gill.spacing        : chr  "c" "c" "c" "" ...
 $ gill.color          : chr  "w" "n" "w" "g" ...
 $ stem.height         : num  4.51 4.79 6.85 4.16 3.37 ...
 $ stem.width          : num  15.39 6.48 9.93 6.53 8.36 ...
 $ stem.root           : chr  "" "" "" "" ...
 $ stem.surface        : chr  "" "y" "s" "" ...
 $ stem.color          : chr  "w" "o" "n" "w" ...
 $ veil.type           : chr  "" "" "" "" ...
 $ veil.color          : chr  "" "" "" "" ...
 $ has.ring            : chr  "f" "t" "f" "f" ...
 $ ring.type           : chr  "f" "z" "f" "f" ...
 $ spore.print.color   : chr  "" "" "" "" ...
 $ habitat             : chr  "d" "d" "l" "d" ...
 $ season              : chr  "a" "w" "w" "u" ...
```

컬럼별의미

id: 각 버섯의 고유 식별자.

class: 독성 여부, 'e' (식용 가능), 'p' (독성).

cap-diameter: 버섯의 갓 지름 (숫자형, cm).

cap-shape: 갓 모양 (예: f, x, p 등).

cap-surface: 갓 표면의 질감 (예: s, y, h 등).

cap-color: 갓의 색상 (예: u, o, n 등).

does-bruise-or-bleed: 멍이 들거나 진물이 나오는지 (예: f, t).

gill-attachment: 주름이 자라는 방식 (예: a, s).

gill-spacing: 주름 간격 (예: c, w 등).

gill-color: 주름의 색상 (예: w, g 등).

stem-height: 줄기의 높이 (숫자형, cm).

stem-width: 줄기의 너비 (숫자형, cm).

stem-root: 줄기의 뿌리 형태.

stem-surface: 줄기의 표면 질감.

stem-color: 줄기의 색상.

veil-type: 줄기를 감싸는 얇은 막의 유형.

veil-color: 막의 색상.

has-ring: 줄기 주변의 고리가 있는지 여부.

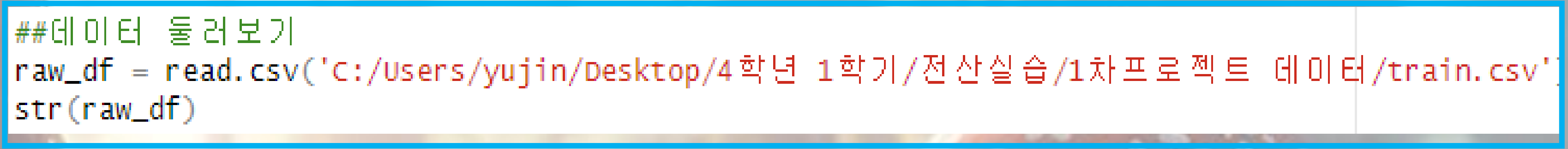
ring-type: 고리의 종류.

spore-print-color: 포자 색상.

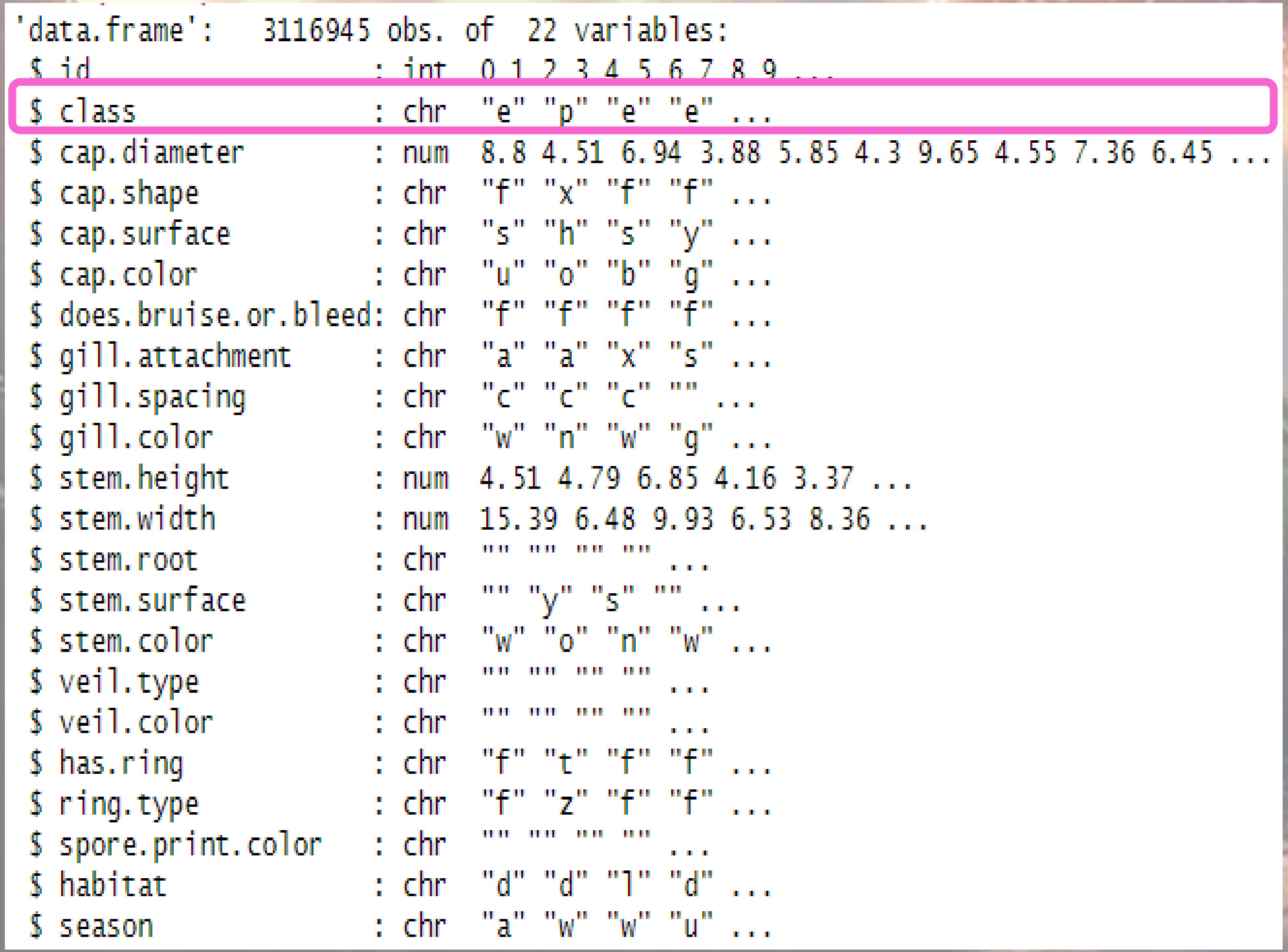
habitat: 서식지 유형 (예: d, l, g 등).

season: 계절 (예: a, w 등).

1.데이터 로드

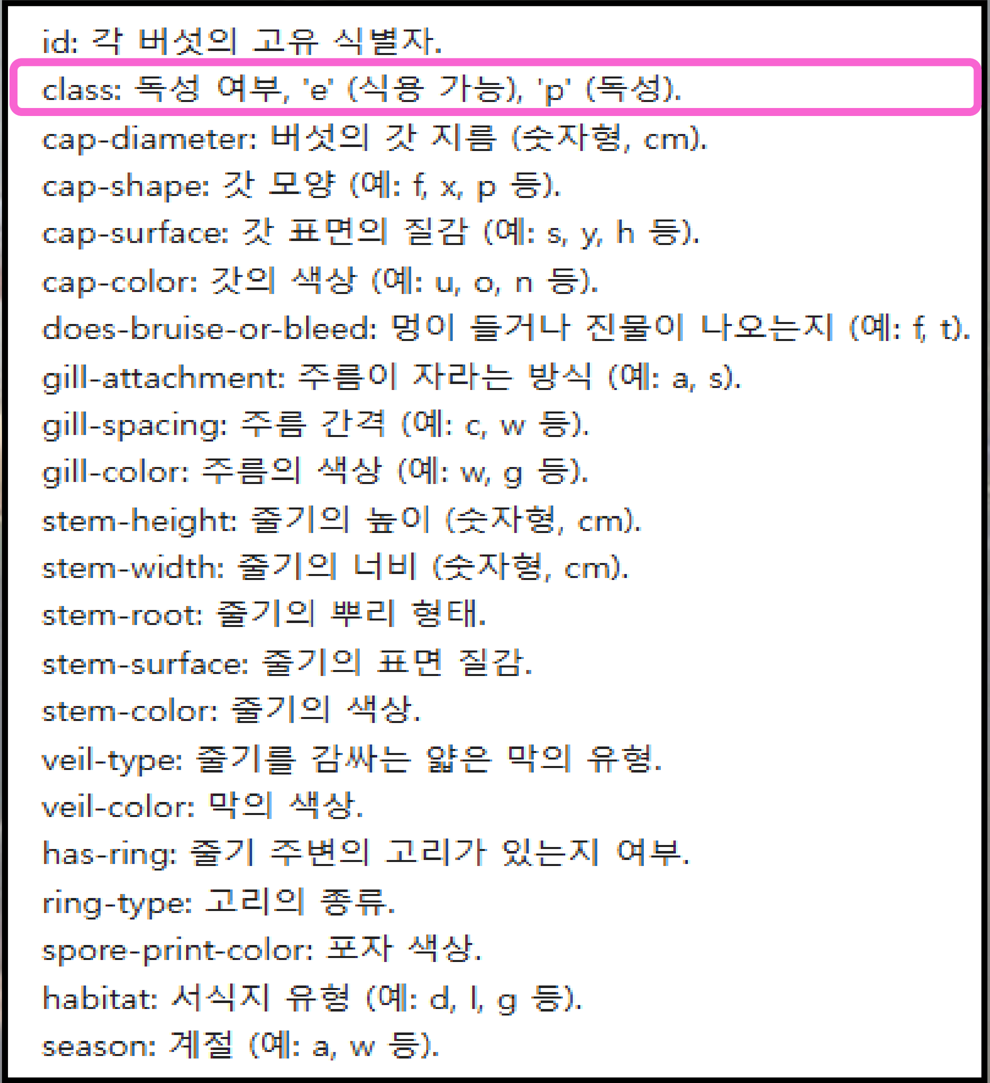


```
##데이터 둘러보기
raw_df = read.csv('C:/Users/yujin/Desktop/4학년 1학기/전산실습/1차프로젝트 데이터/train.csv'
str(raw_df)
```



```
'data.frame':	3116945 obs. of  22 variables:
 $ id                  : int  0 1 2 3 4 5 6 7 8 9 ...
 $ class               : chr  "e" "p" "e" "e" ...
 $ cap.diameter        : num  8.8 4.51 6.94 3.88 5.85 4.3 9.65 4.55 7.36 6.45 ...
 $ cap.shape           : chr  "f" "x" "f" "f" ...
 $ cap.surface         : chr  "s" "h" "s" "y" ...
 $ cap.color           : chr  "u" "o" "b" "g" ...
 $ does.bruise.or.bleed: chr  "f" "f" "f" "f" ...
 $ gill.attachment     : chr  "a" "a" "x" "s" ...
 $ gill.spacing        : chr  "c" "c" "c" "" ...
 $ gill.color          : chr  "w" "n" "w" "g" ...
 $ stem.height         : num  4.51 4.79 6.85 4.16 3.37 ...
 $ stem.width          : num  15.39 6.48 9.93 6.53 8.36 ...
 $ stem.root           : chr  "" "" "" "" ...
 $ stem.surface        : chr  "" "y" "s" "" ...
 $ stem.color          : chr  "w" "o" "n" "w" ...
 $ veil.type           : chr  "" "" "" "" ...
 $ veil.color          : chr  "" "" "" "" ...
 $ has.ring            : chr  "f" "t" "f" "f" ...
 $ ring.type           : chr  "f" "z" "f" "f" ...
 $ spore.print.color   : chr  "" "" "" "" ...
 $ habitat             : chr  "d" "d" "l" "d" ...
 $ season              : chr  "a" "w" "w" "u" ...
```

컬럼별의미



id: 각 버섯의 고유 식별자.

class: 독성 여부, 'e' (식용 가능), 'p' (독성).

cap-diameter: 버섯의 갓 지름 (숫자형, cm).

cap-shape: 갓 모양 (예: f, x, p 등).

cap-surface: 갓 표면의 질감 (예: s, y, h 등).

cap-color: 갓의 색상 (예: u, o, n 등).

does-bruise-or-bleed: 멍이 들거나 진물이 나오는지 (예: f, t).

gill-attachment: 주름이 자라는 방식 (예: a, s).

gill-spacing: 주름 간격 (예: c, w 등).

gill-color: 주름의 색상 (예: w, g 등).

stem-height: 줄기의 높이 (숫자형, cm).

stem-width: 줄기의 너비 (숫자형, cm).

stem-root: 줄기의 뿌리 형태.

stem-surface: 줄기의 표면 질감.

stem-color: 줄기의 색상.

veil-type: 줄기를 감싸는 얇은 막의 유형.

veil-color: 막의 색상.

has-ring: 줄기 주변의 고리가 있는지 여부.

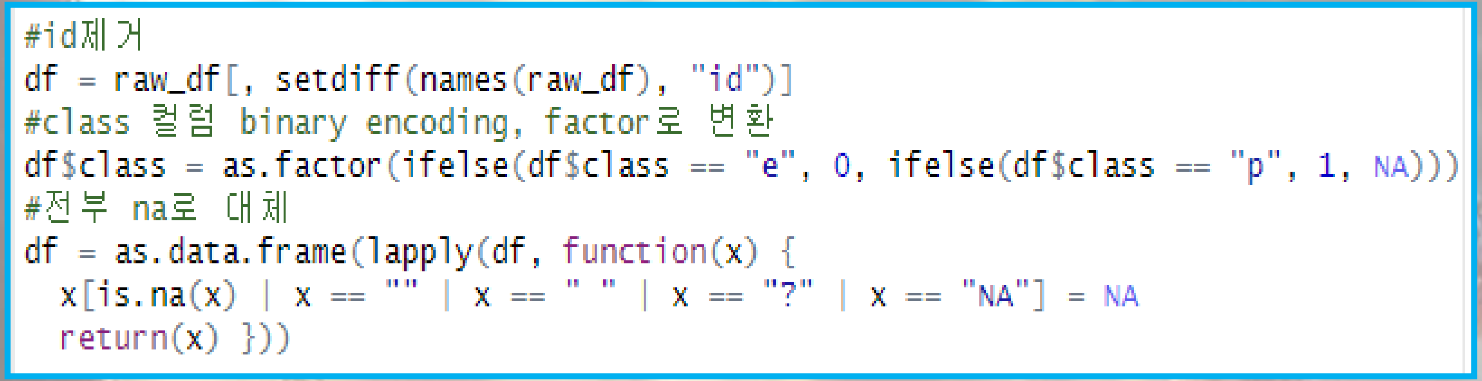
ring-type: 고리의 종류.

spore-print-color: 포자 색상.

habitat: 서식지 유형 (예: d, l, g 등).

season: 계절 (예: a, w 등).

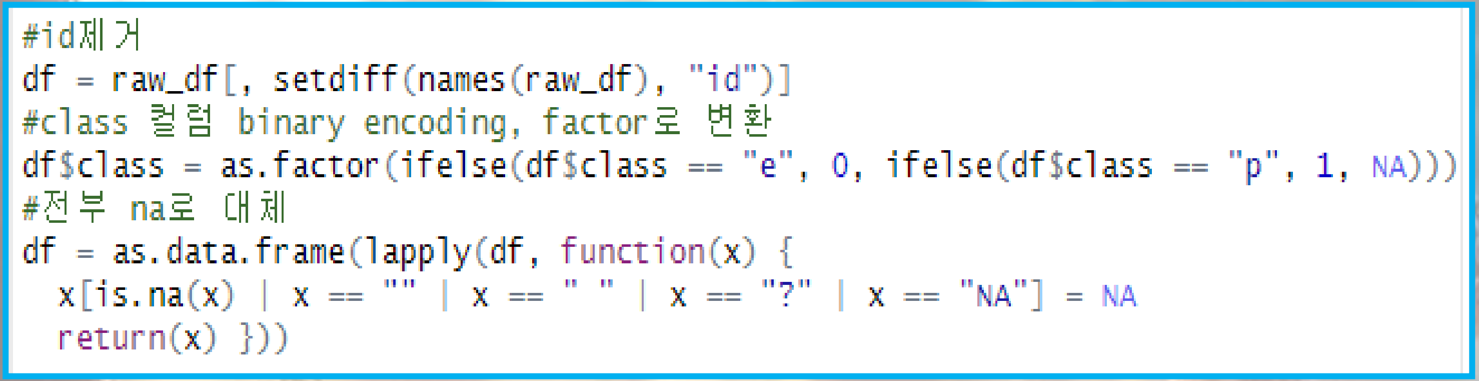
2.데이터 정제



```r
#id제거
df = raw_df[, setdiff(names(raw_df), "id")]
#class 컬럼 binary encoding, factor로 변환
df$class = as.factor(ifelse(df$class == "e", 0, ifelse(df$class == "p", 1, NA)))
#전부 na로 대체
df = as.data.frame(lapply(df, function(x) {
  x[is.na(x) | x == "" | x == " " | x == "?" | x == "NA"] = NA
  return(x) }))
```

독성 여부 예측 목적 -> P를 1로 설정

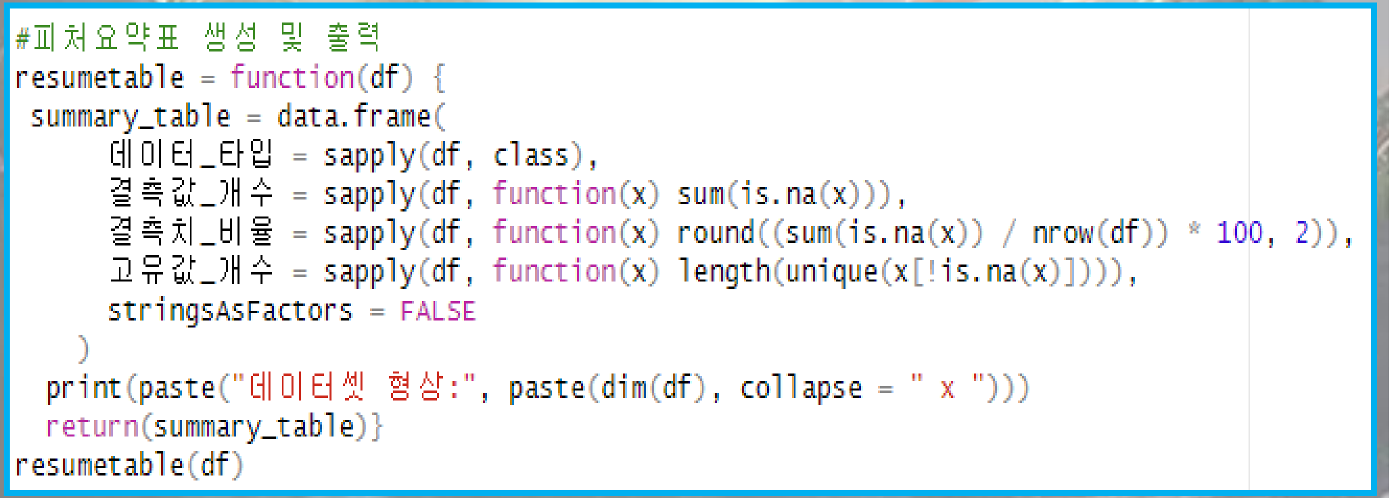
2.데이터 정제



```r
#id제거
df = raw_df[, setdiff(names(raw_df), "id")]
#class 컬럼 binary encoding, factor로 변환
df$class = as.factor(ifelse(df$class == "e", 0, ifelse(df$class == "p", 1, NA)))
#전부 na로 대체
df = as.data.frame(lapply(df, function(x) {
  x[is.na(x) | x == "" | x == " " | x == "?" | x == "NA"] = NA
  return(x) }))
```

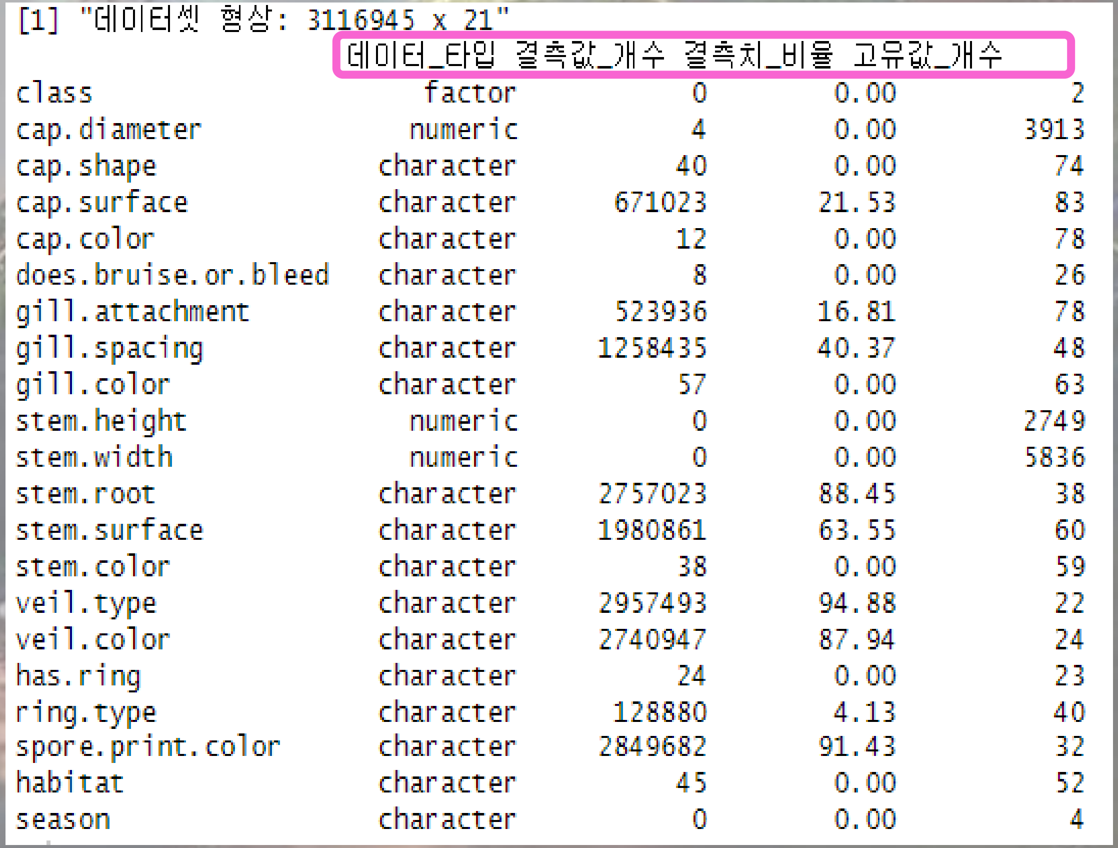
독성 여부 예측 목적 -> P를 1로 설정

2.1 결측치 처리



```r
#피처요약표 생성 및 출력
resumetable = function(df) {
 summary_table = data.frame(
      데이터_타입 = sapply(df, class),
      결측값_개수 = sapply(df, function(x) sum(is.na(x))),
      결측치_비율 = sapply(df, function(x) round((sum(is.na(x)) / nrow(df)) * 100, 2)),
      고유값_개수 = sapply(df, function(x) length(unique(x[!is.na(x)]))),
      stringsAsFactors = FALSE
    )
  print(paste("데이터셋 형상:", paste(dim(df), collapse = " x ")))
  return(summary_table)}
resumetable(df)
```

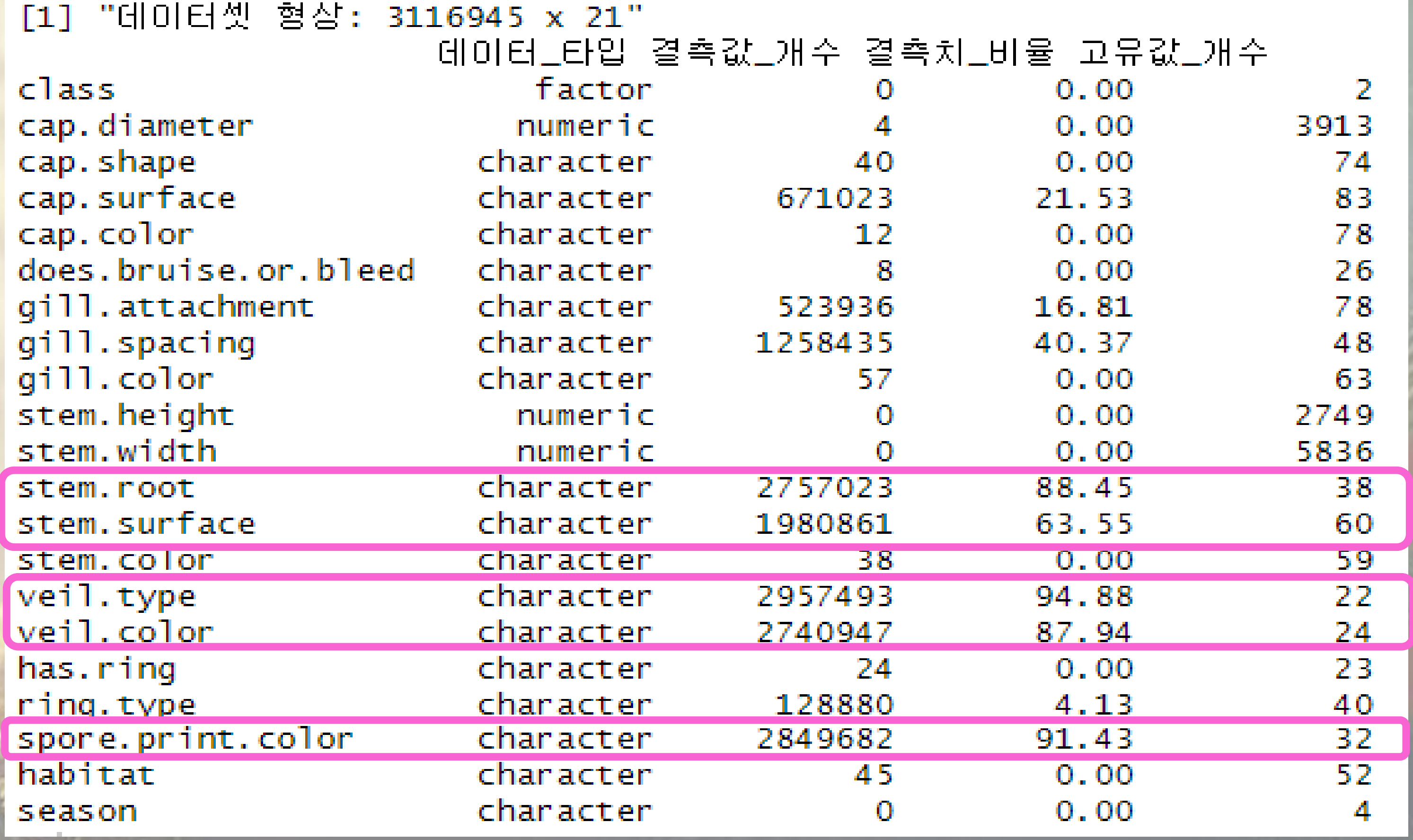
피처요약표



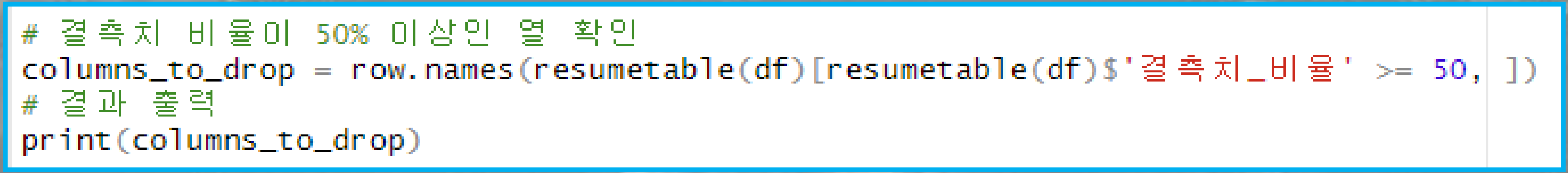
```
[1] "데이터셋 형상: 3116945 x 21"
                     데이터_타입 결측값_개수 결측치_비율 고유값_개수
class                     factor           0        0.00           2
cap.diameter             numeric           4        0.00        3913
cap.shape              character          40        0.00          74
cap.surface            character      671023       21.53          83
cap.color              character          12        0.00          78
does.bruise.or.bleed   character           8        0.00          26
gill.attachment        character      523936       16.81          78
gill.spacing           character     1258435       40.37          48
gill.color             character          57        0.00          63
stem.height              numeric           0        0.00        2749
stem.width               numeric           0        0.00        5836
stem.root              character     2757023       88.45          38
stem.surface           character     1980861       63.55          60
stem.color             character          38        0.00          59
veil.type              character     2957493       94.88          22
veil.color             character     2740947       87.94          24
has.ring               character          24        0.00          23
ring.type              character      128880        4.13          40
spore.print.color      character     2849682       91.43          32
habitat                character          45        0.00          52
season                 character           0        0.00           4
```

2.1 결측치 처리

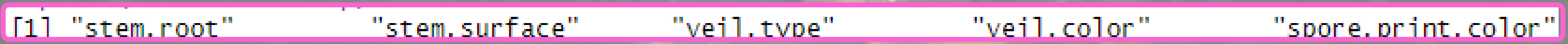
결측치제거



```
[1] "데이터셋 형상: 3116945 x 21"
                       데이터_타입 결측값_개수 결측치_비율 고유값_개수
class                     factor           0        0.00           2
cap.diameter             numeric           4        0.00        3913
cap.shape              character          40        0.00          74
cap.surface            character      671023       21.53          83
cap.color              character          12        0.00          78
does.bruise.or.bleed   character           8        0.00          26
gill.attachment        character      523936       16.81          78
gill.spacing           character     1258435       40.37          48
gill.color             character          57        0.00          63
stem.height              numeric           0        0.00        2749
stem.width               numeric           0        0.00        5836
stem.root              character     2757023       88.45          38
stem.surface           character     1980861       63.55          60
stem.color             character          38        0.00          59
veil.type              character     2957493       94.88          22
veil.color             character     2740947       87.94          24
has.ring               character          24        0.00          23
ring.type              character      128880        4.13          40
spore.print.color      character     2849682       91.43          32
habitat                character          45        0.00          52
season                 character           0        0.00           4
```

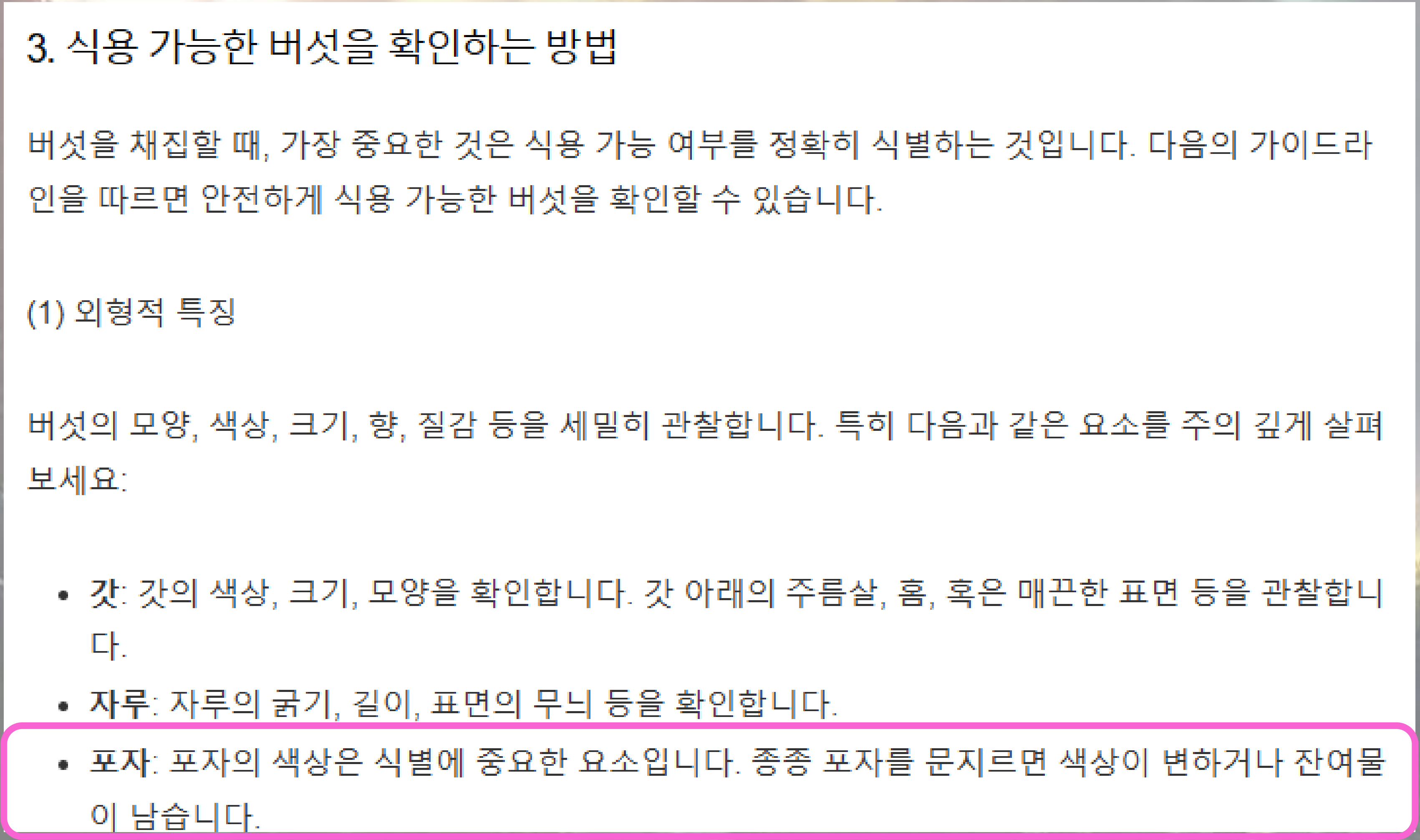


```
# 결측치 비율이 50% 이상인 열 확인
columns_to_drop = row.names(resumetable(df)[resumetable(df)$'결측치_비율' >= 50, ])
# 결과 출력
print(columns_to_drop)
```



```
[1] "stem.root"          "stem.surface"       "veil.type"          "veil.color"          "spore.print.color"
```

2.1 결측치 처리



3. 식용 가능한 버섯을 확인하는 방법

버섯을 채집할 때, 가장 중요한 것은 식용 가능 여부를 정확히 식별하는 것입니다. 다음의 가이드라인을 따르면 안전하게 식용 가능한 버섯을 확인할 수 있습니다.

(1) 외형적 특징

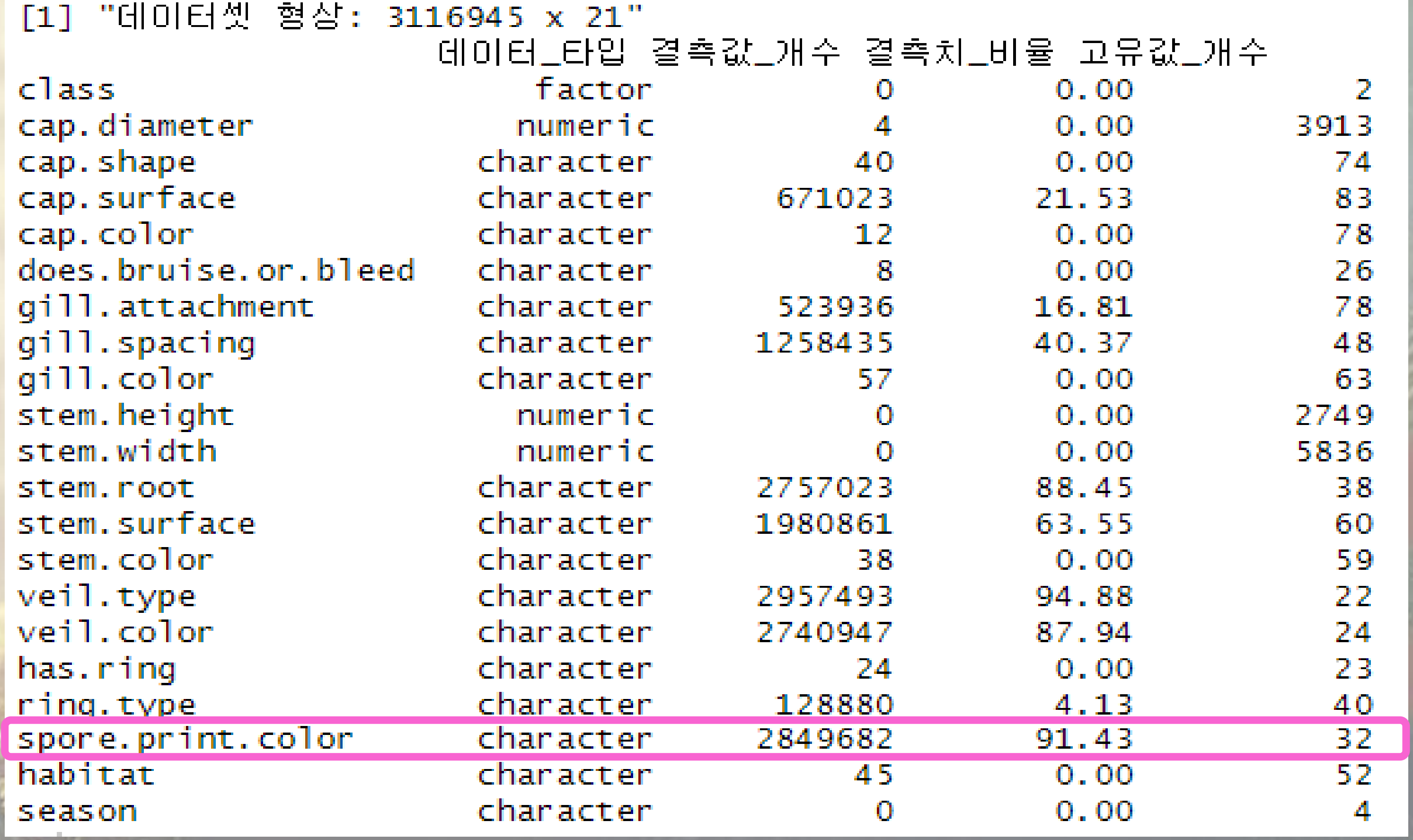
버섯의 모양, 색상, 크기, 향, 질감 등을 세밀히 관찰합니다. 특히 다음과 같은 요소를 주의 깊게 살펴보세요:

- **갓**: 갓의 색상, 크기, 모양을 확인합니다. 갓 아래의 주름살, 홈, 혹은 매끈한 표면 등을 관찰합니다.
- **자루**: 자루의 굵기, 길이, 표면의 무늬 등을 확인합니다.
- **포자**: 포자의 색상은 식별에 중요한 요소입니다. 종종 포자를 문지르면 색상이 변하거나 잔여물이 남습니다.

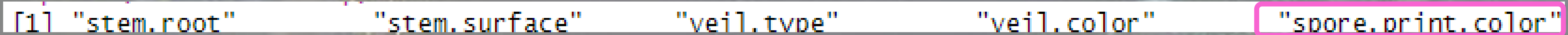
출처: https://teaestar.tistory.com/entry/%EB%B2%84%EC%84%AF-%EC%97%B0%EA%B5%AC%EA%B0%80%EC%9D%98-%EC%84%B8%EA%B3%84-%EB%B2%84%EC%84%AF-%EC%A2%85%EB%A5%98%EC%99%80-%EC%8B%9D%EC%9A%A9-%EA%B0%80%EB%8A%A5%ED%95%9C-%EB%B2%84%EC%84%AF-%ED%99%95%EC%9D%B8-%EB%B0%A9%EB%B2%95

2.1 결측치 처리

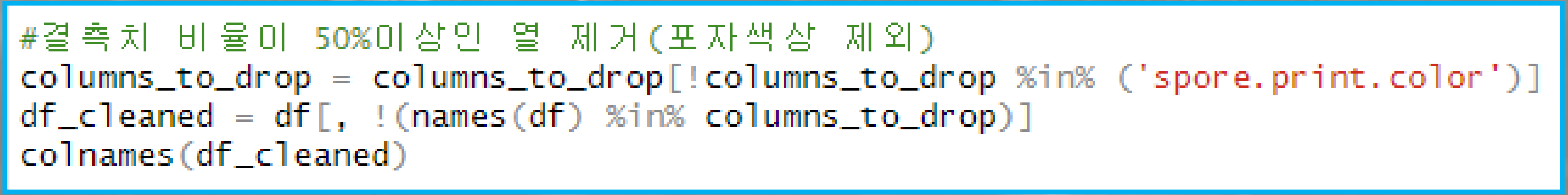
결측치제거



```
[1] "데이터셋 형상: 3116945 x 21"
                      데이터_타입 결측값_개수 결측치_비율 고유값_개수
class                      factor           0        0.00           2
cap.diameter              numeric           4        0.00        3913
cap.shape               character          40        0.00          74
cap.surface             character      671023       21.53          83
cap.color               character          12        0.00          78
does.bruise.or.bleed    character           8        0.00          26
gill.attachment         character      523936       16.81          78
gill.spacing            character     1258435       40.37          48
gill.color              character          57        0.00          63
stem.height               numeric           0        0.00        2749
stem.width                numeric           0        0.00        5836
stem.root               character     2757023       88.45          38
stem.surface            character     1980861       63.55          60
stem.color              character          38        0.00          59
veil.type               character     2957493       94.88          22
veil.color              character     2740947       87.94          24
has.ring                character          24        0.00          23
ring.type               character      128880        4.13          40
spore.print.color       character     2849682       91.43          32
habitat                 character          45        0.00          52
season                  character           0        0.00           4
```



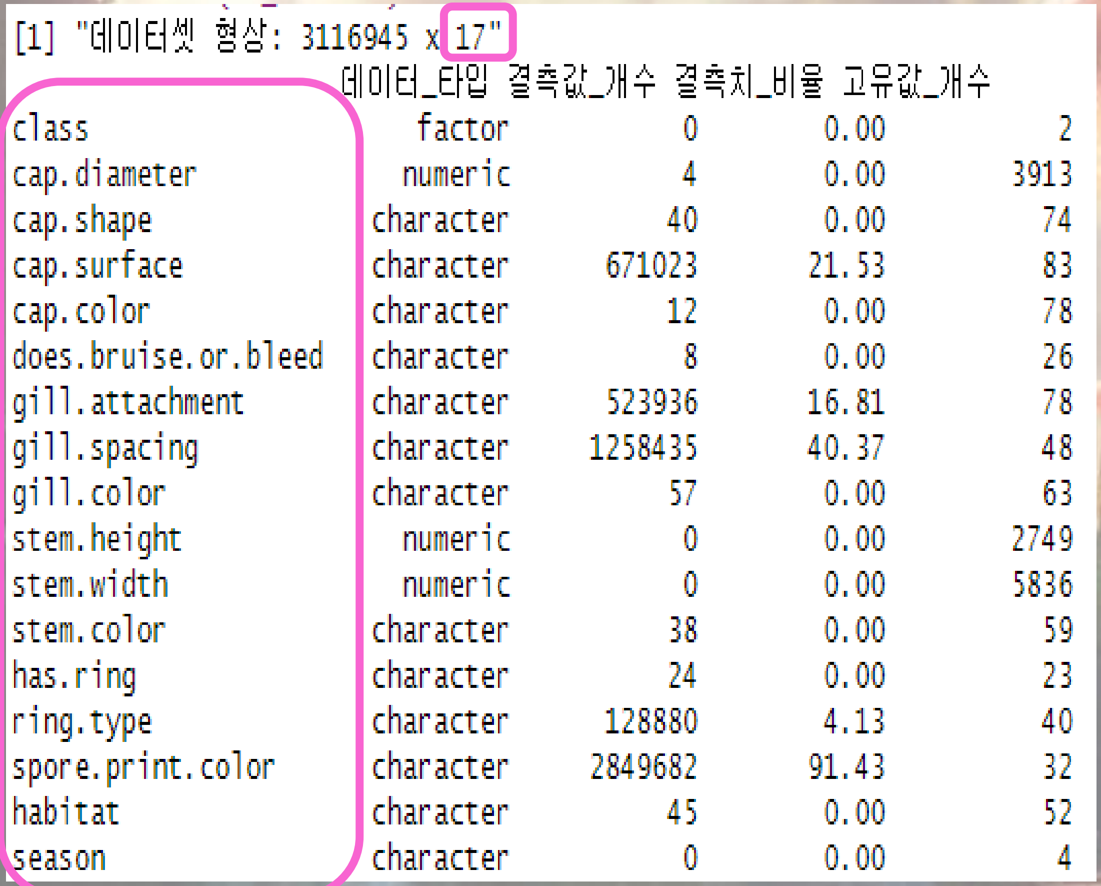
```
[1] "stem.root"          "stem.surface"       "veil.type"          "veil.color"         "spore.print.color"
```



```
#결측치 비율이 50%이상인 열 제거(포자색상 제외)
columns_to_drop = columns_to_drop[!columns_to_drop %in% ('spore.print.color')]
df_cleaned = df[, !(names(df) %in% columns_to_drop)]
colnames(df_cleaned)
```

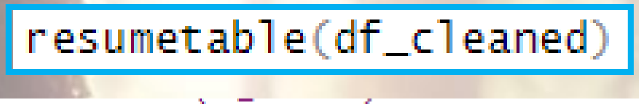
2.1 결측치 처리

```
resumetable(df_cleaned)
```

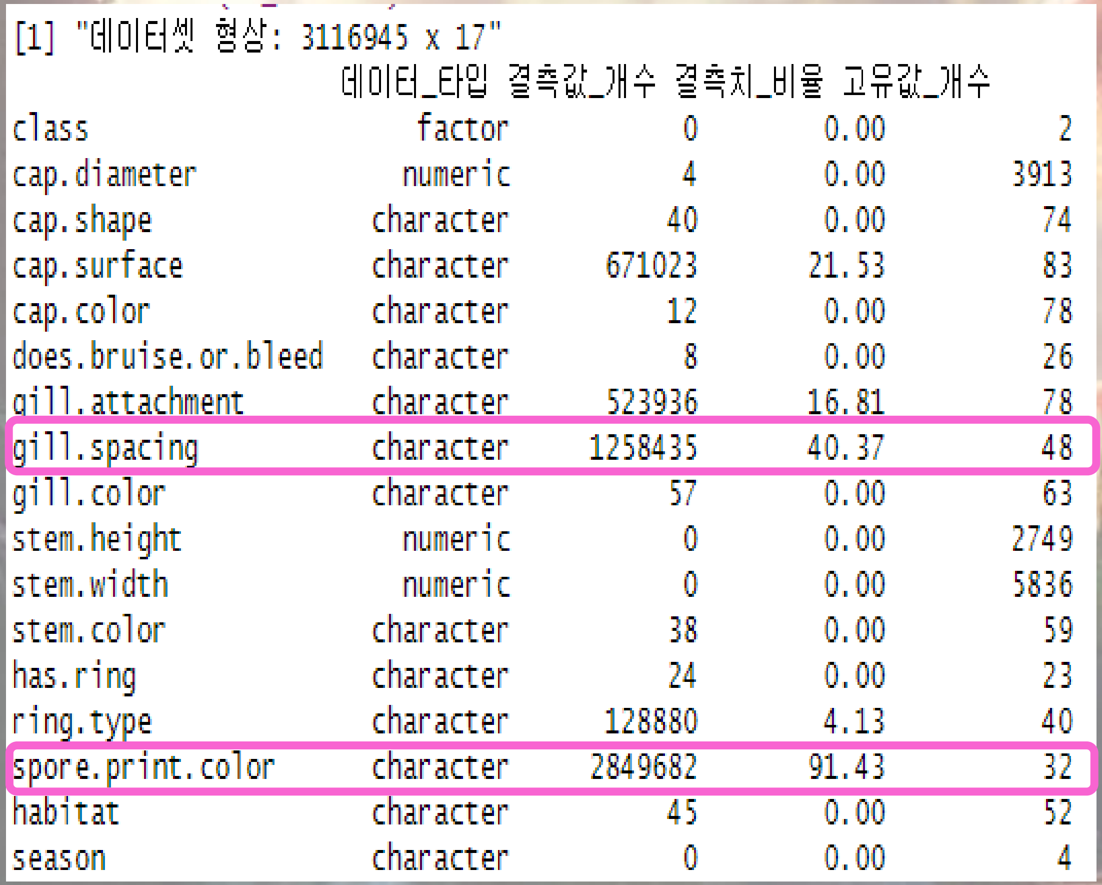


```
[1] "데이터셋 형상: 3116945 x 17"
                  데이터_타입 결측값_개수 결측치_비율 고유값_개수
class                  factor           0        0.00           2
cap.diameter          numeric           4        0.00        3913
cap.shape           character          40        0.00          74
cap.surface         character      671023       21.53          83
cap.color           character          12        0.00          78
does.bruise.or.bleed character          8        0.00          26
gill.attachment     character      523936       16.81          78
gill.spacing        character     1258435       40.37          48
gill.color          character          57        0.00          63
stem.height           numeric           0        0.00        2749
stem.width            numeric           0        0.00        5836
stem.color          character          38        0.00          59
has.ring            character          24        0.00          23
ring.type           character      128880        4.13          40
spore.print.color   character     2849682       91.43          32
habitat             character          45        0.00          52
season              character           0        0.00           4
```

2.1 결측치 처리

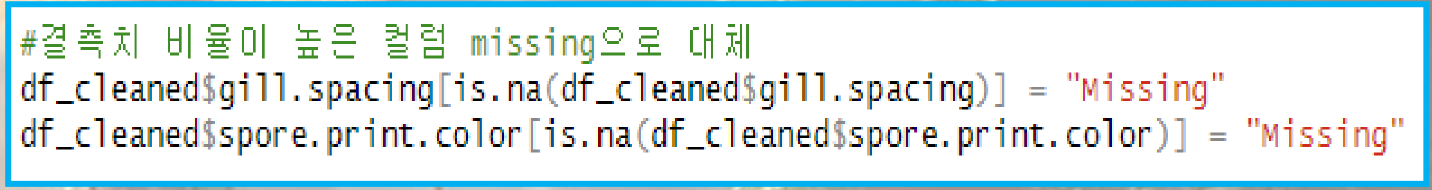


```
resumetable(df_cleaned)
```



```
[1] "데이터셋 형상: 3116945 x 17"
                      데이터_타입 결측값_개수 결측치_비율 고유값_개수
class                      factor          0        0.00           2
cap.diameter              numeric          4        0.00        3913
cap.shape               character         40        0.00          74
cap.surface             character     671023       21.53          83
cap.color               character         12        0.00          78
does.bruise.or.bleed    character          8        0.00          26
gill.attachment         character     523936       16.81          78
gill.spacing            character    1258435       40.37          48
gill.color              character         57        0.00          63
stem.height               numeric          0        0.00        2749
stem.width                numeric          0        0.00        5836
stem.color              character         38        0.00          59
has.ring                character         24        0.00          23
ring.type               character     128880        4.13          40
spore.print.color       character    2849682       91.43          32
habitat                 character         45        0.00          52
season                  character          0        0.00           4
```

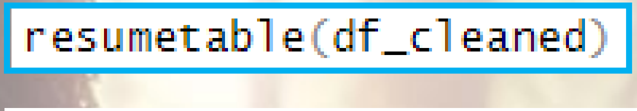
결측치채우기



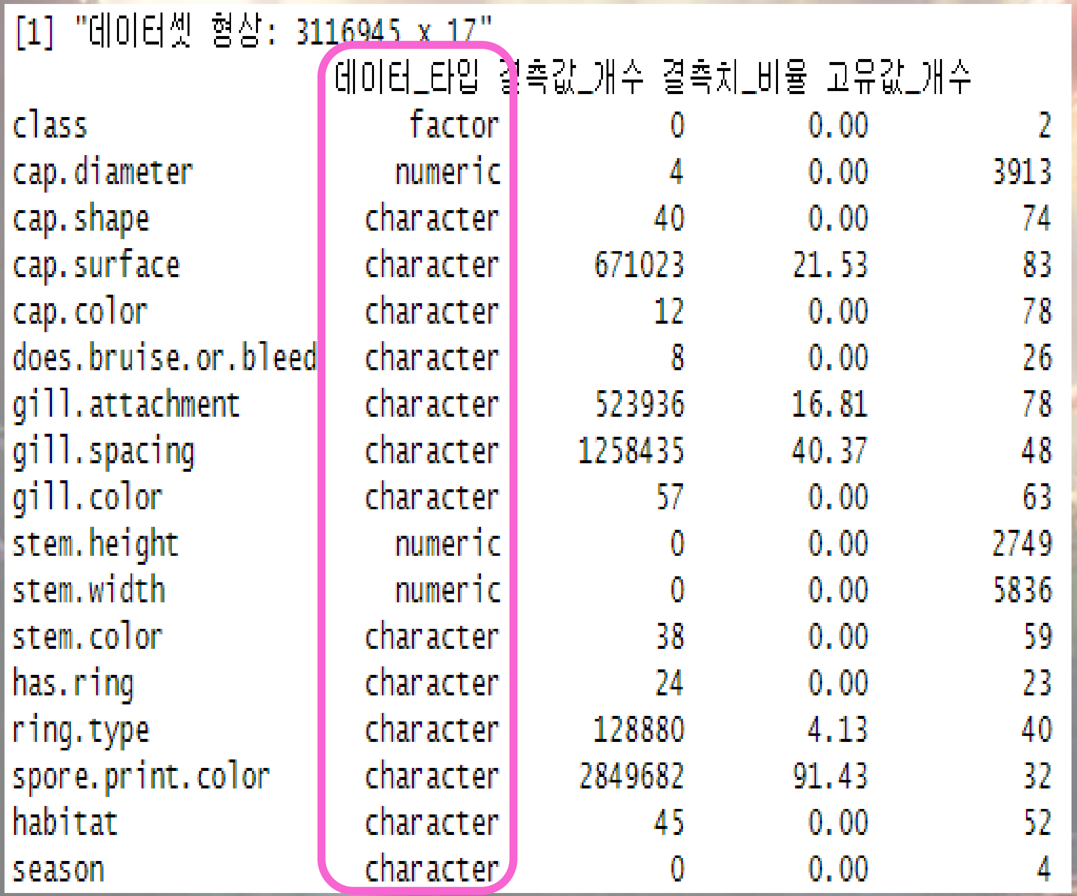
```
#결측치 비율이 높은 컬럼 missing으로 대체
df_cleaned$gill.spacing[is.na(df_cleaned$gill.spacing)] = "Missing"
df_cleaned$spore.print.color[is.na(df_cleaned$spore.print.color)] = "Missing"
```

2.1 결측치 처리

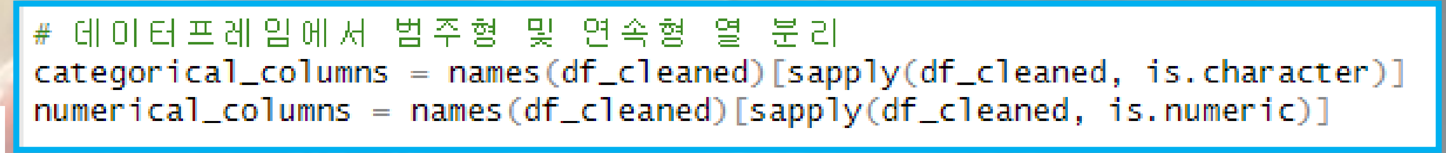
결측치채우기



```
resumetable(df_cleaned)
```

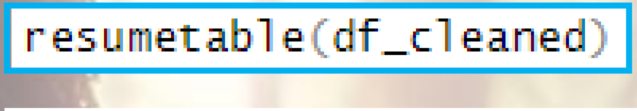


```
[1] "데이터셋 형상: 3116945 x 17"
                     데이터_타입 결측값_개수 결측치_비율 고유값_개수
class                     factor           0        0.00           2
cap.diameter             numeric           4        0.00        3913
cap.shape              character          40        0.00          74
cap.surface            character      671023       21.53          83
cap.color              character          12        0.00          78
does.bruise.or.bleed   character           8        0.00          26
gill.attachment        character      523936       16.81          78
gill.spacing           character     1258435       40.37          48
gill.color             character          57        0.00          63
stem.height              numeric           0        0.00        2749
stem.width               numeric           0        0.00        5836
stem.color             character          38        0.00          59
has.ring               character          24        0.00          23
ring.type              character      128880        4.13          40
spore.print.color      character     2849682       91.43          32
habitat                character          45        0.00          52
season                 character           0        0.00           4
```

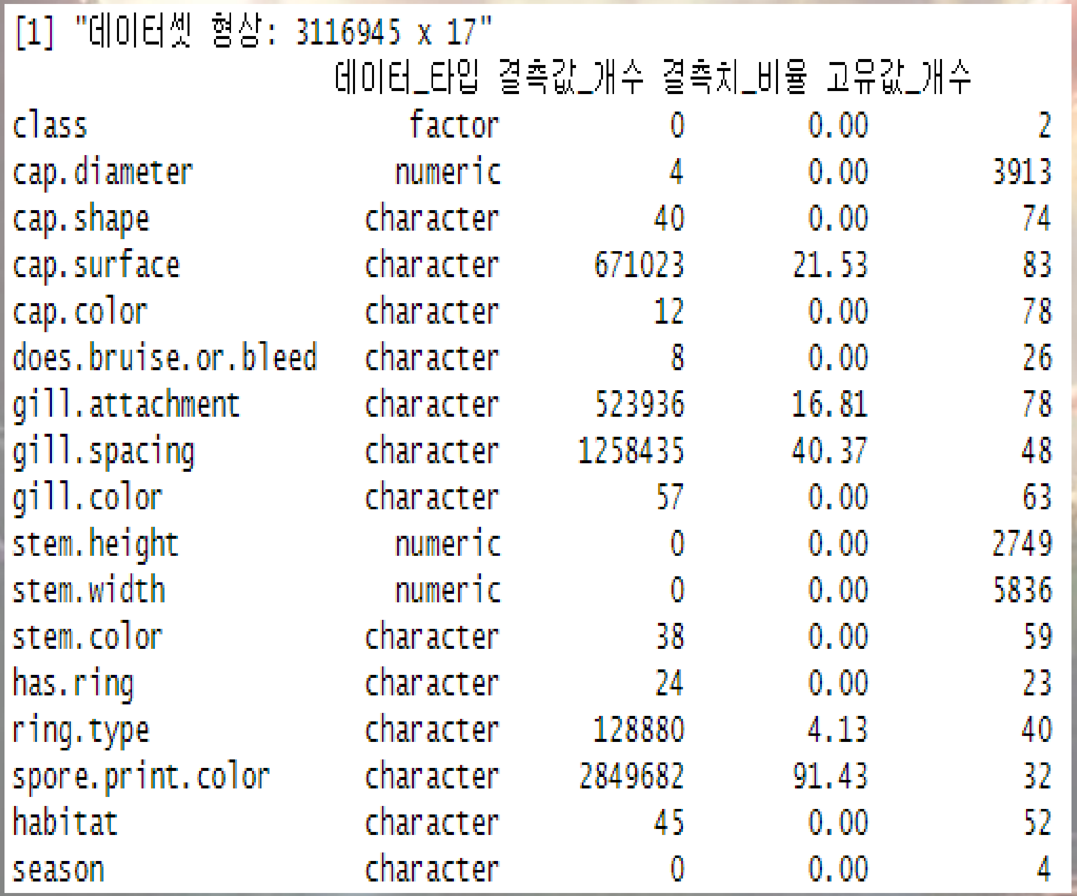


```
# 데이터프레임에서 범주형 및 연속형 열 분리
categorical_columns = names(df_cleaned)[sapply(df_cleaned, is.character)]
numerical_columns = names(df_cleaned)[sapply(df_cleaned, is.numeric)]
```

2.1 결측치 처리

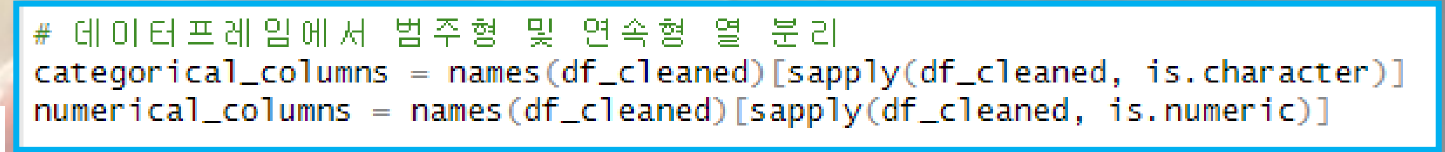


```r
resumetable(df_cleaned)
```

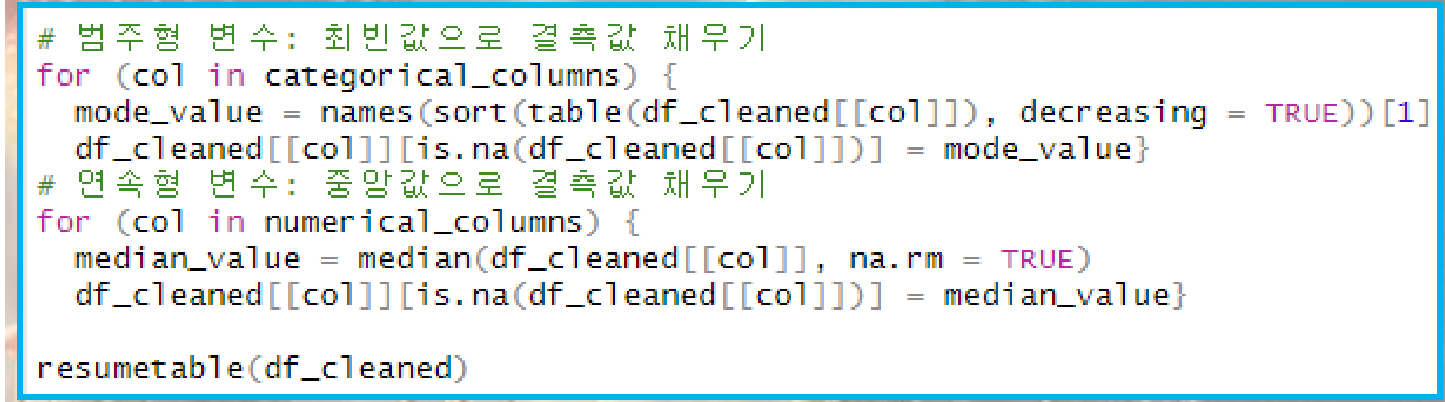


```
[1] "데이터셋 형상: 3116945 x 17"
                    데이터_타입 결측값_개수 결측치_비율 고유값_개수
class                    factor           0        0.00           2
cap.diameter            numeric           4        0.00        3913
cap.shape             character          40        0.00          74
cap.surface           character      671023       21.53          83
cap.color             character          12        0.00          78
does.bruise.or.bleed  character           8        0.00          26
gill.attachment       character      523936       16.81          78
gill.spacing          character     1258435       40.37          48
gill.color            character          57        0.00          63
stem.height             numeric           0        0.00        2749
stem.width              numeric           0        0.00        5836
stem.color            character          38        0.00          59
has.ring              character          24        0.00          23
ring.type             character      128880        4.13          40
spore.print.color     character     2849682       91.43          32
habitat               character          45        0.00          52
season                character           0        0.00           4
```

결측치채우기

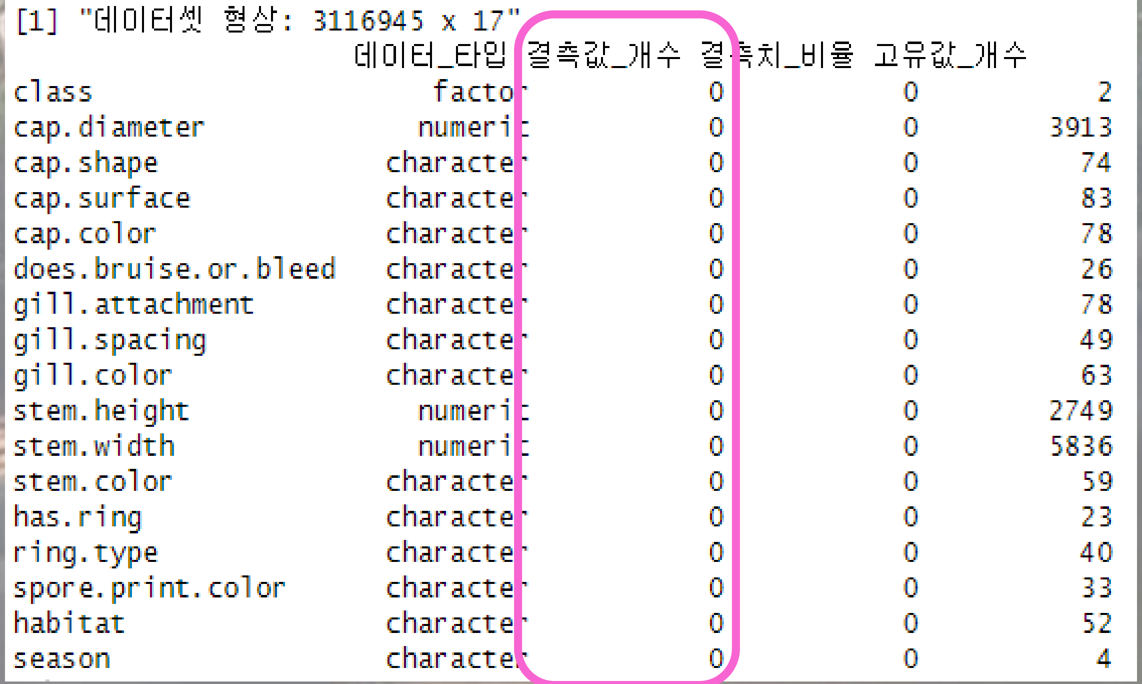


```r
# 데이터프레임에서 범주형 및 연속형 열 분리
categorical_columns = names(df_cleaned)[sapply(df_cleaned, is.character)]
numerical_columns = names(df_cleaned)[sapply(df_cleaned, is.numeric)]
```



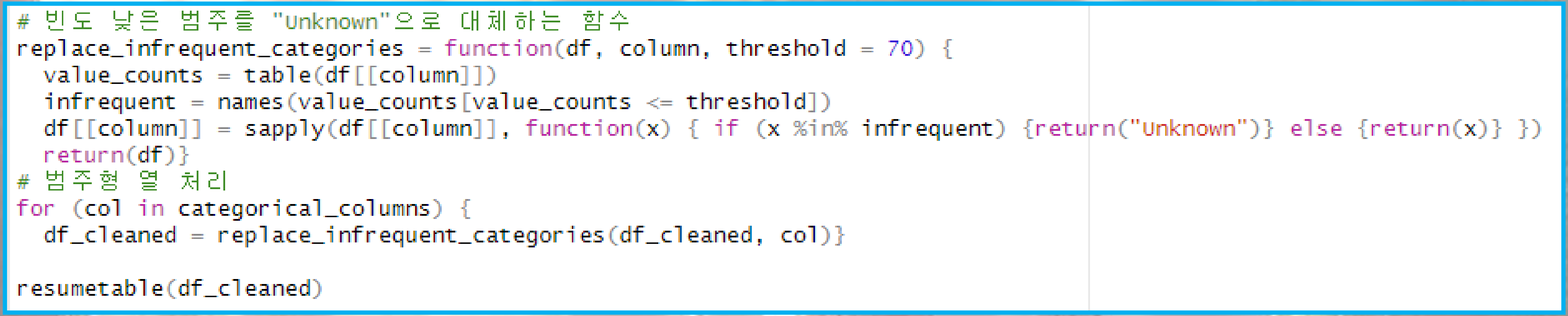
```r
# 범주형 변수: 최빈값으로 결측값 채우기
for (col in categorical_columns) {
  mode_value = names(sort(table(df_cleaned[[col]]), decreasing = TRUE))[1]
  df_cleaned[[col]][is.na(df_cleaned[[col]])] = mode_value}
# 연속형 변수: 중앙값으로 결측값 채우기
for (col in numerical_columns) {
  median_value = median(df_cleaned[[col]], na.rm = TRUE)
  df_cleaned[[col]][is.na(df_cleaned[[col]])] = median_value}

resumetable(df_cleaned)
```



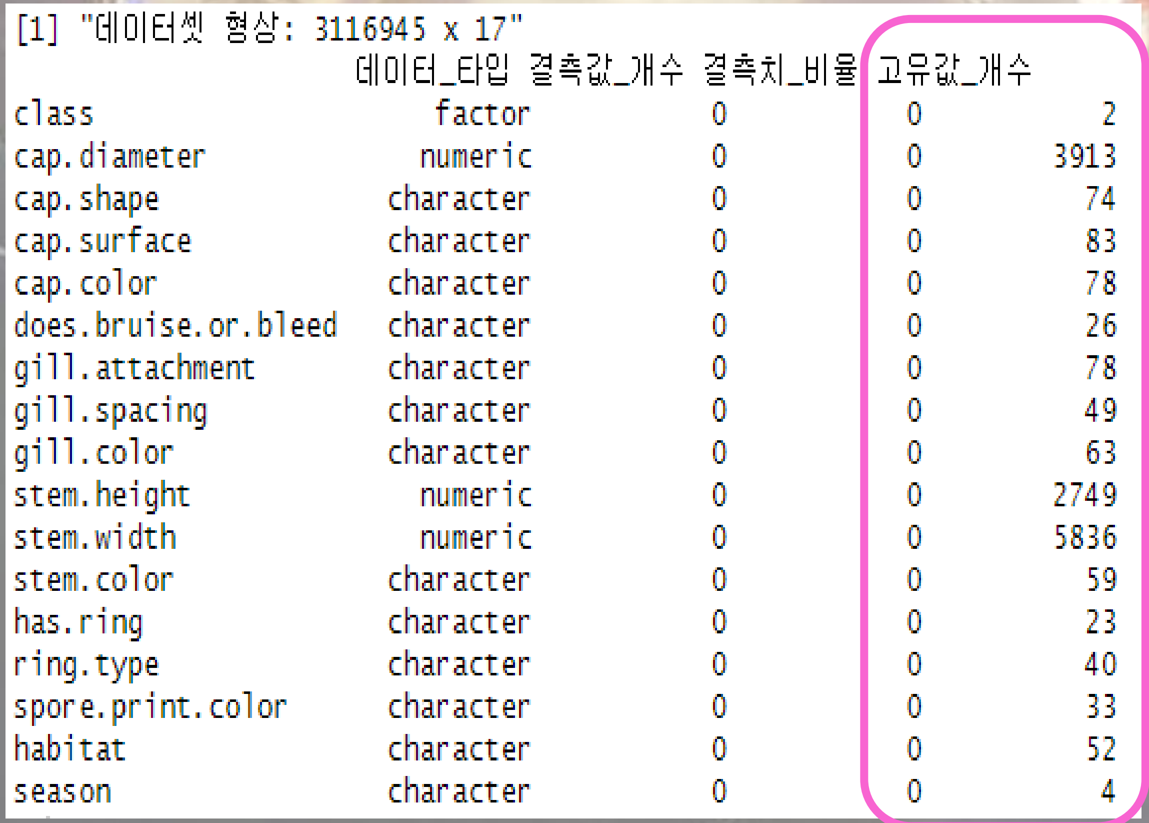
```
[1] "데이터셋 형상: 3116945 x 17"
                     데이터_타입 결측값_개수 결측치_비율 고유값_개수
class                     factor           0           0           2
cap.diameter             numeric           0           0        3913
cap.shape              character           0           0          74
cap.surface            character           0           0          83
cap.color              character           0           0          78
does.bruise.or.bleed   character           0           0          26
gill.attachment        character           0           0          78
gill.spacing           character           0           0          49
gill.color             character           0           0          63
stem.height              numeric           0           0        2749
stem.width               numeric           0           0        5836
stem.color             character           0           0          59
has.ring               character           0           0          23
ring.type              character           0           0          40
spore.print.color      character           0           0          33
habitat                character           0           0          52
season                 character           0           0           4
```

2.2 고유값 처리

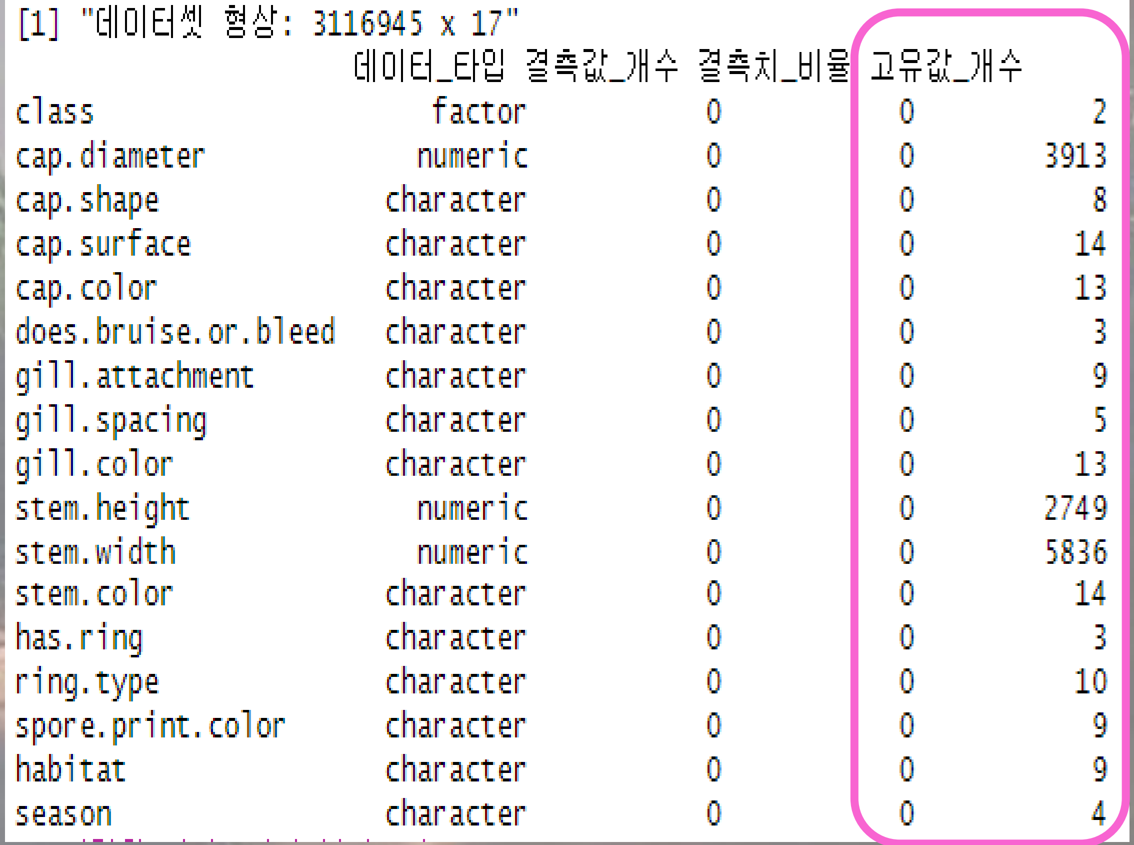


```
# 빈도 낮은 범주를 "Unknown"으로 대체하는 함수
replace_infrequent_categories = function(df, column, threshold = 70) {
  value_counts = table(df[[column]])
  infrequent = names(value_counts[value_counts <= threshold])
  df[[column]] = sapply(df[[column]], function(x) { if (x %in% infrequent) {return("Unknown")} else {return(x)} })
  return(df)}
# 범주형 열 처리
for (col in categorical_columns) {
  df_cleaned = replace_infrequent_categories(df_cleaned, col)}

resumetable(df_cleaned)
```

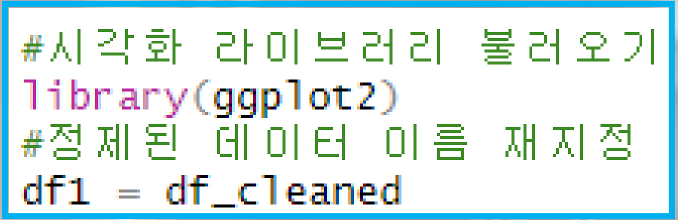


```
[1] "데이터셋 형상: 3116945 x 17"
                    데이터_타입 결측값_개수 결측치_비율 고유값_개수
class                    factor          0          0           2
cap.diameter            numeric          0          0        3913
cap.shape             character          0          0          74
cap.surface           character          0          0          83
cap.color             character          0          0          78
does.bruise.or.bleed  character          0          0          26
gill.attachment       character          0          0          78
gill.spacing          character          0          0          49
gill.color            character          0          0          63
stem.height             numeric          0          0        2749
stem.width              numeric          0          0        5836
stem.color            character          0          0          59
has.ring              character          0          0          23
ring.type             character          0          0          40
spore.print.color     character          0          0          33
habitat               character          0          0          52
season                character          0          0           4
```

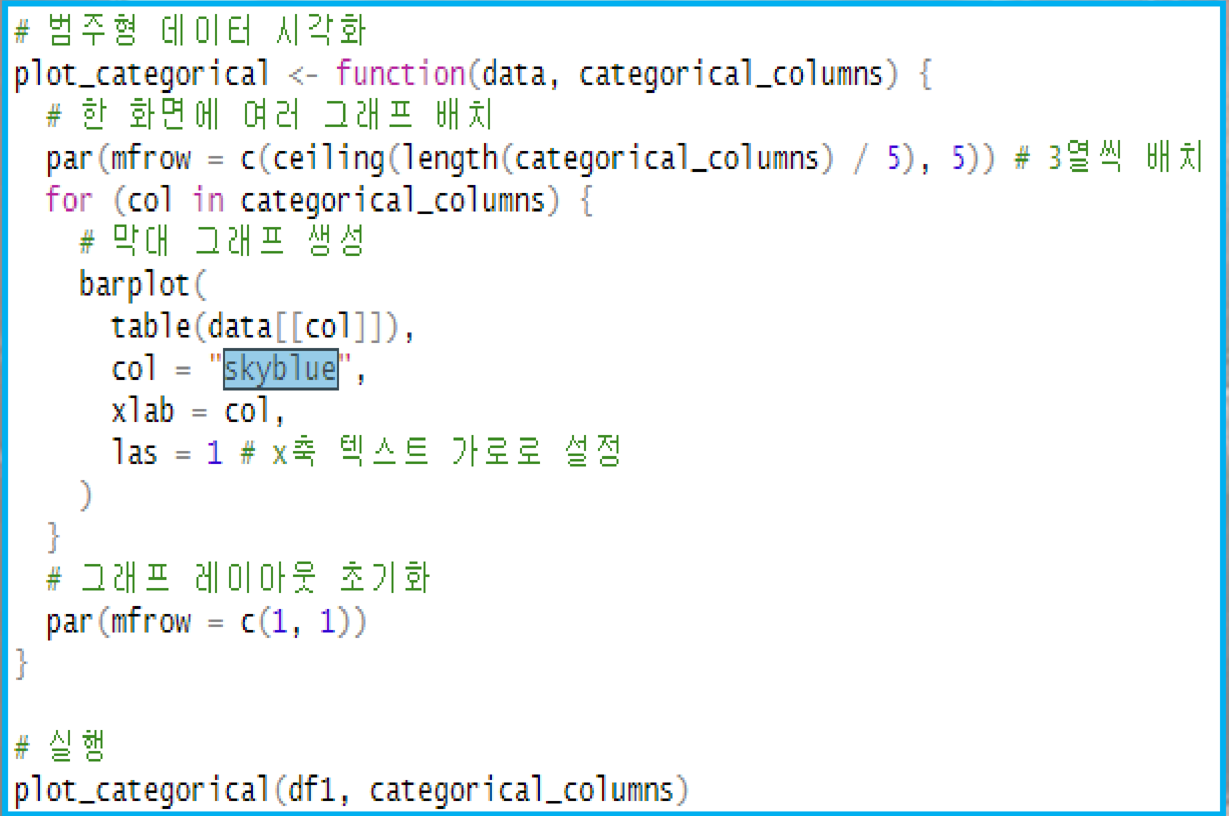


```
[1] "데이터셋 형상: 3116945 x 17"
                    데이터_타입 결측값_개수 결측치_비율 고유값_개수
class                    factor          0          0           2
cap.diameter            numeric          0          0        3913
cap.shape             character          0          0           8
cap.surface           character          0          0          14
cap.color             character          0          0          13
does.bruise.or.bleed  character          0          0           3
gill.attachment       character          0          0           9
gill.spacing          character          0          0           5
gill.color            character          0          0          13
stem.height             numeric          0          0        2749
stem.width              numeric          0          0        5836
stem.color            character          0          0          14
has.ring              character          0          0           3
ring.type             character          0          0          10
spore.print.color     character          0          0           9
habitat               character          0          0           9
season                character          0          0           4
```

3.정제된 데이터 시각화

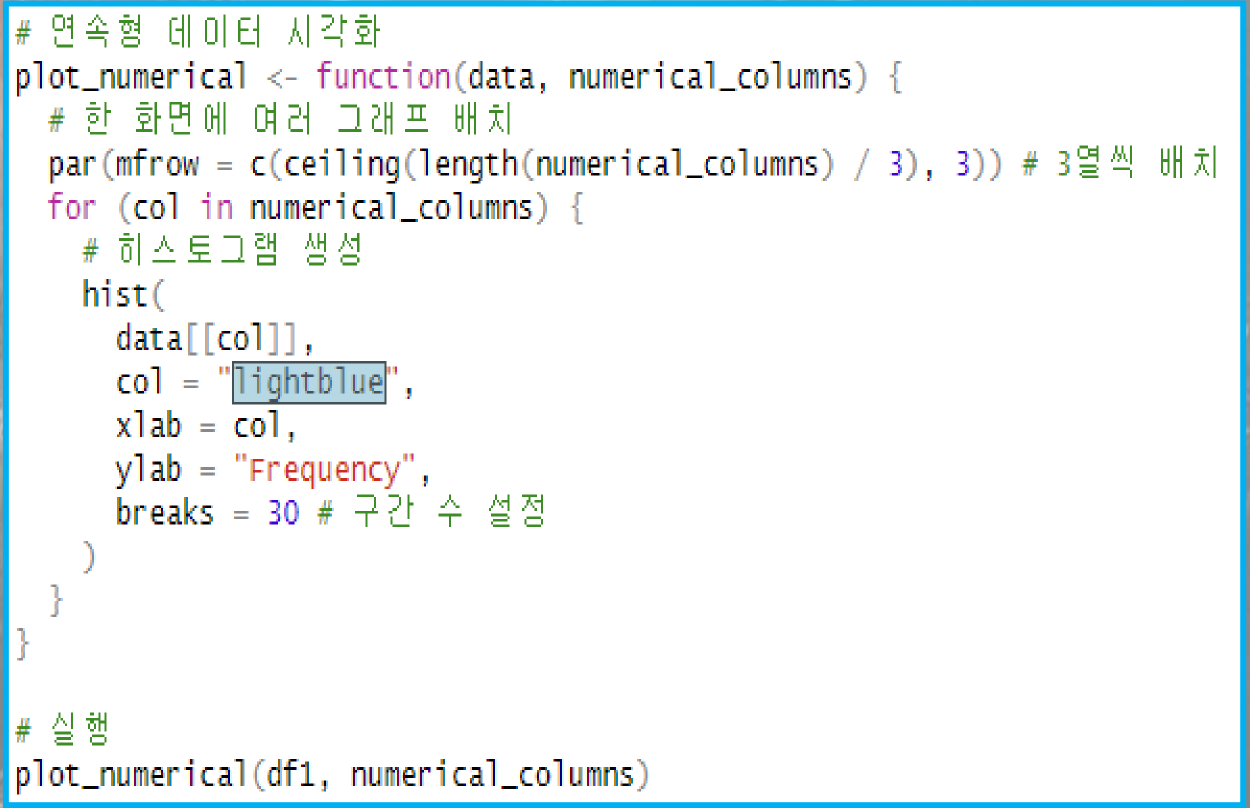


```r
#시각화 라이브러리 불러오기
library(ggplot2)
#정제된 데이터 이름 재지정
df1 = df_cleaned
```



```r
# 범주형 데이터 시각화
plot_categorical <- function(data, categorical_columns) {
  # 한 화면에 여러 그래프 배치
  par(mfrow = c(ceiling(length(categorical_columns) / 5), 5)) # 3열씩 배치
  for (col in categorical_columns) {
    # 막대 그래프 생성
    barplot(
      table(data[[col]]),
      col = "skyblue",
      xlab = col,
      las = 1 # x축 텍스트 가로로 설정
    )
  }
  # 그래프 레이아웃 초기화
  par(mfrow = c(1, 1))
}

# 실행
plot_categorical(df1, categorical_columns)
```



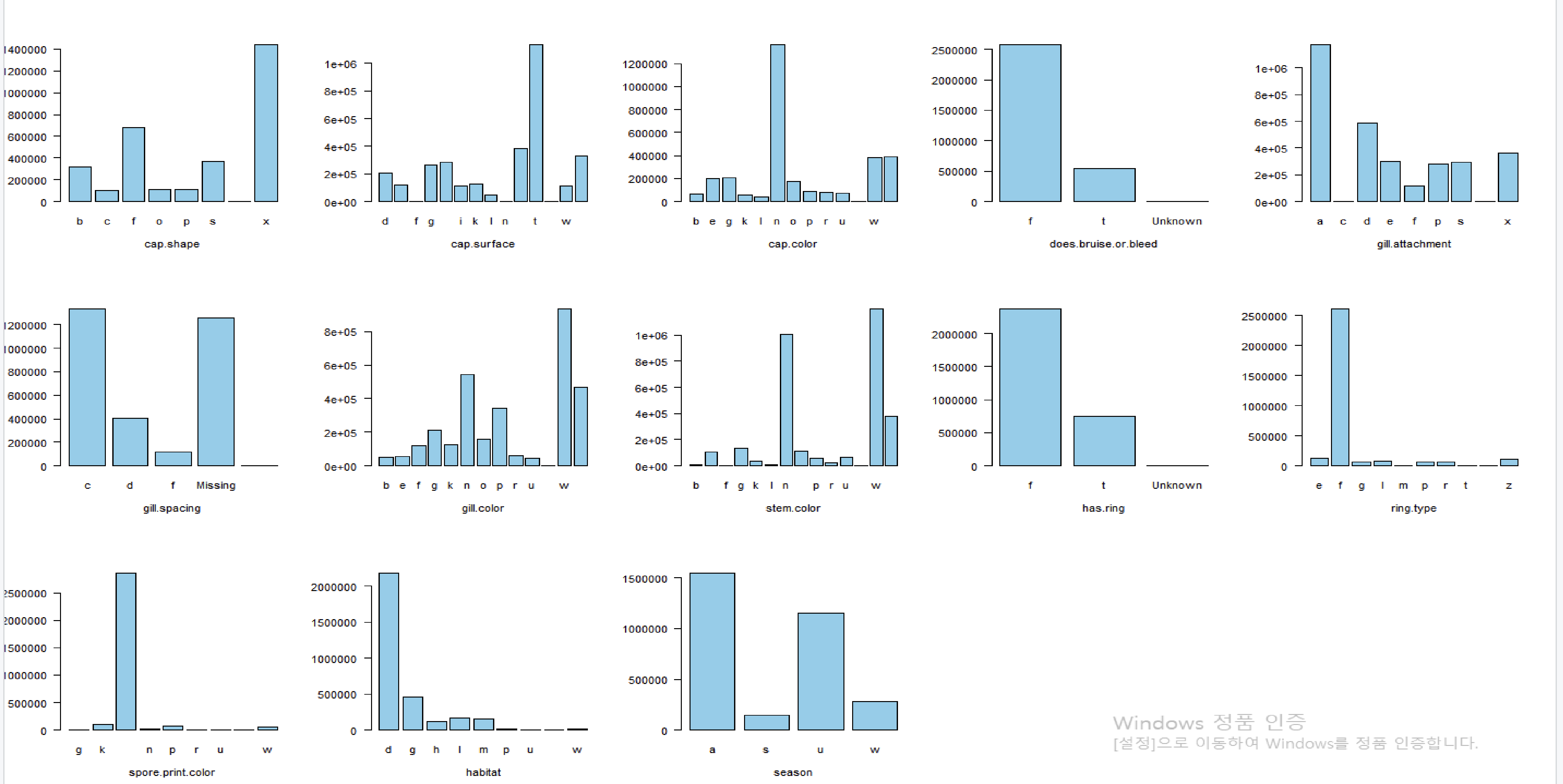
```r
# 연속형 데이터 시각화
plot_numerical <- function(data, numerical_columns) {
  # 한 화면에 여러 그래프 배치
  par(mfrow = c(ceiling(length(numerical_columns) / 3), 3)) # 3열씩 배치
  for (col in numerical_columns) {
    # 히스토그램 생성
    hist(
      data[[col]],
      col = "lightblue",
      xlab = col,
      ylab = "Frequency",
      breaks = 30 # 구간 수 설정
    )
  }
}

# 실행
plot_numerical(df1, numerical_columns)
```

고유값 별 개수(빈도수)

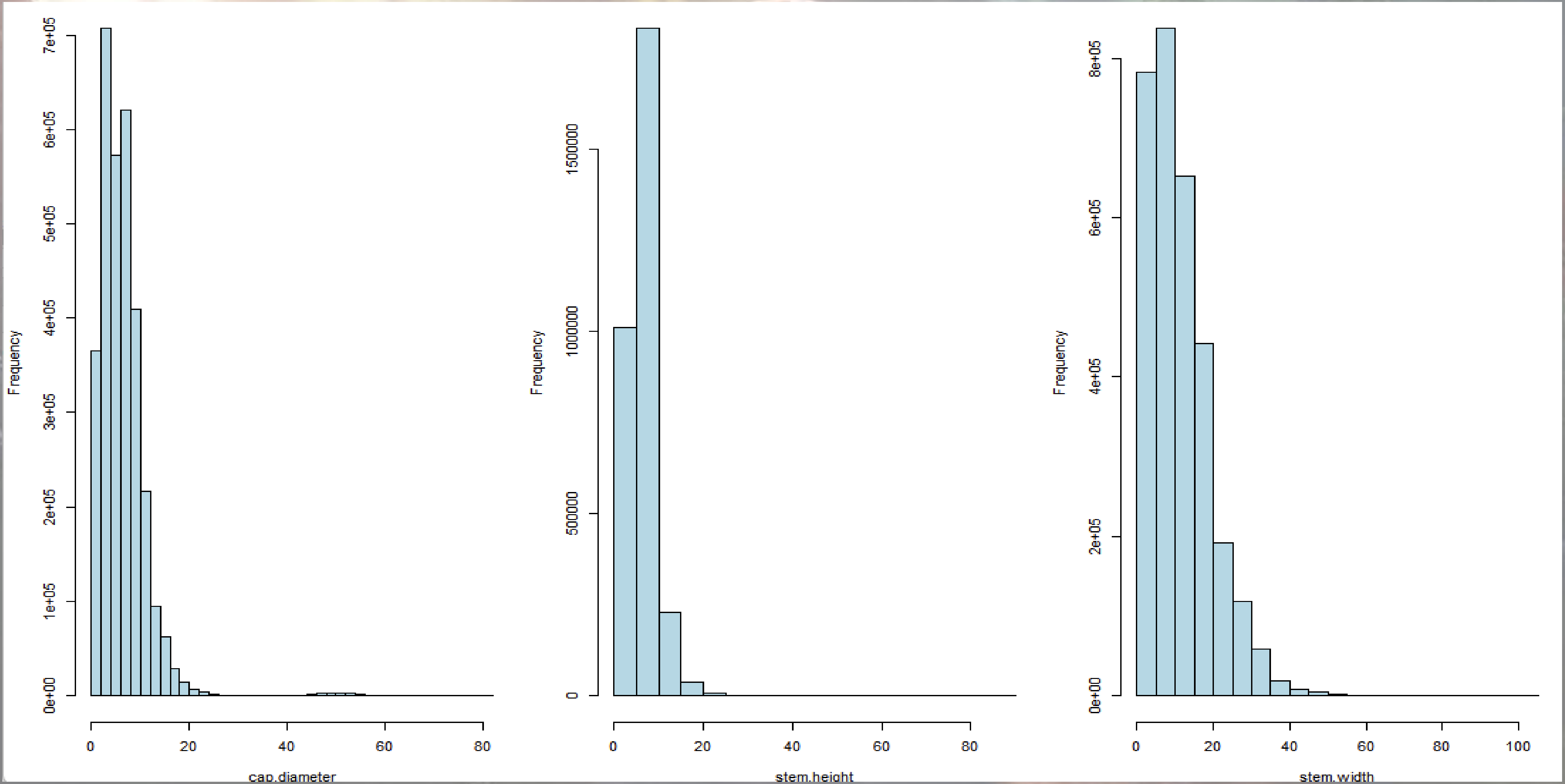
3.정제된 데이터 시각화

막대 그래프



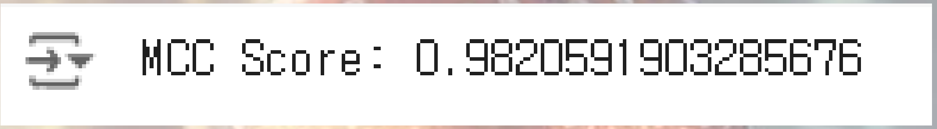
3.정제된 데이터 시각화

히스토그램

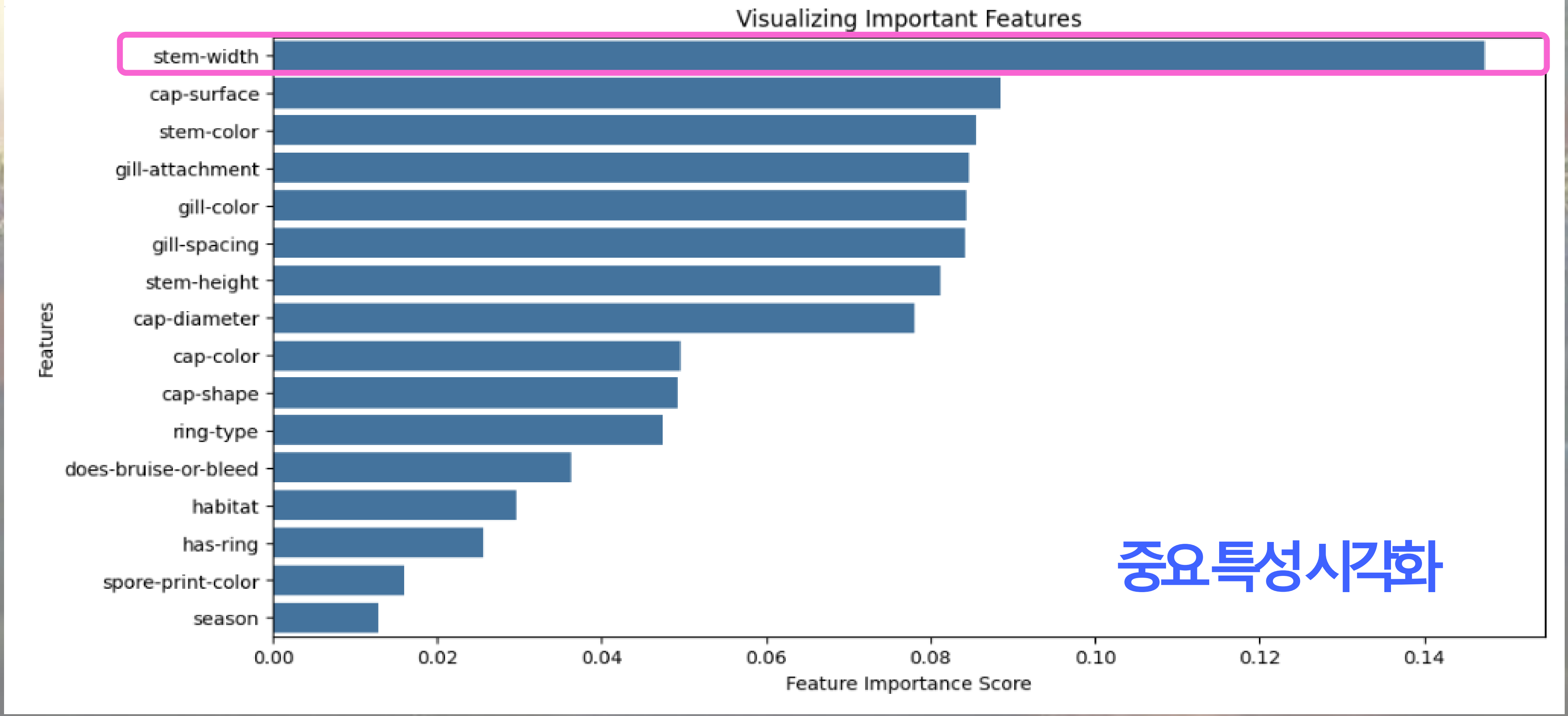


4. 모델링 - RandomForest

R에서는 메모리 부족 문제-> colab에서 진행



```
MCC Score: 0.9820591903285676
```



중요특성시각화

감사합니다!!