Veri Biliminde R Uygulamalari Odev

Ceyda Murat

2024-01-18

Contents

Veri setine erişim linki: https://archive.ics.uci.edu/dataset/109/wine

"Wine" veri seti, üç farklı sınıfa ait üzüm şaraplarından elde edilen kimyasal bileşenleri içerir.Bu veri setinin özellikleri şunlardır:

- **Wine:** Her bir şarap örneğinin sınıfını belirten bir değişkeni ifade eder. Bu değişken, şarap örneklerinin sınıflarını temsil eden kategorik bir değişkendir. Üç farklı sınıfa ait şarap örneklerini içerir
- **Alcohol:** Şaraptaki alkol oranını ölçen sayısal bir özellik.
- **Malic Acid:** Şaraptaki elma asidi miktarını ölçen sayısal bir özellik.
- **Ash:** Şaraptaki kül miktarını ölçen sayısal bir özellik.
- **Alcalinity(Acl):** Şaraptaki külün alkalinitesini ölçen sayısal bir özellik.
- **Magnesium(Mg):** Şaraptaki magnezyum miktarını ölçen sayısal bir özellik.
- **Phenols:** Şaraptaki toplam fenol miktarını ölçen sayısal bir özellik.
- **Flavanoids:** Şaraptaki flavanoid miktarını ölçen sayısal bir özellik.

- **Nonflavanoid Phenols:** Şaraptaki nonflavanoid fenol miktarını ölçen sayısal bir özellik.
- **Proanthocyanins:** Şaraptaki proantosiyandin miktarını ölçen sayısal bir özellik.
- **Color Intensity:** Şaraptaki renk yoğunluğunu ölçen sayısal bir özellik.
- **Hue:** Şaraptaki renk tonunu ölçen sayısal bir özellik.
- **OD:** Şarabın 280/315 oranındaki optik yoğunluğunu ölçen sayısal bir özellik.
- **Proline:** Şaraptaki prolin miktarını ölçen sayısal bir özellik.

1 Veri setinin detaylı incelenmesi ve özet halinde açıklanması

```r
library(dplyr)
library(tidyverse)
wine_data = read.csv("wine.csv",header = T, sep=",")
wine_data = as_tibble(wine_data)
head(wine_data)
```

```
## # A tibble: 6 x 14
##    Wine Alcohol Malic.acid   Ash   Acl    Mg Phenols Flavanoids
##   <int>   <dbl>      <dbl> <dbl> <dbl> <int>   <dbl>      <dbl>
## 1     1    14.2       1.71  2.43  15.6   127    2.8        3.06
## 2     1    13.2       1.78  2.14  11.2   100    2.65       2.76
## 3     1    13.2       2.36  2.67  18.6   101    2.8        3.24
## 4     1    14.4       1.95  2.5   16.8   113    3.85       3.49
## 5     1    13.2       2.59  2.87  21     118    2.8        2.69
## 6     1    14.2       1.76  2.45  15.2   112    3.27       3.39
## # i 6 more variables: Nonflavanoid.phenols <dbl>, Proanth <dbl>,
## #   Color.int <dbl>, Hue <dbl>, OD <dbl>, Proline <int>
```

```r
glimpse(wine_data)
```

```
## Rows: 178
## Columns: 14
## $ Wine                 <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ Alcohol              <dbl> 14.23, 13.20, 13.16, 14.37, 13.24, 14.20, 14.39, ~
## $ Malic.acid           <dbl> 1.71, 1.78, 2.36, 1.95, 2.59, 1.76, 1.87, 2.15, 1~
## $ Ash                  <dbl> 2.43, 2.14, 2.67, 2.50, 2.87, 2.45, 2.45, 2.61, 2~
## $ Acl                  <dbl> 15.6, 11.2, 18.6, 16.8, 21.0, 15.2, 14.6, 17.6, 1~
## $ Mg                   <int> 127, 100, 101, 113, 118, 112, 96, 121, 97, 98, 10~
## $ Phenols              <dbl> 2.80, 2.65, 2.80, 3.85, 2.80, 3.27, 2.50, 2.60, 2~
## $ Flavanoids           <dbl> 3.06, 2.76, 3.24, 3.49, 2.69, 3.39, 2.52, 2.51, 2~
## $ Nonflavanoid.phenols <dbl> 0.28, 0.26, 0.30, 0.24, 0.39, 0.34, 0.30, 0.31, 0~
## $ Proanth              <dbl> 2.29, 1.28, 2.81, 2.18, 1.82, 1.97, 1.98, 1.25, 1~
## $ Color.int            <dbl> 5.64, 4.38, 5.68, 7.80, 4.32, 6.75, 5.25, 5.05, 5~
## $ Hue                  <dbl> 1.04, 1.05, 1.03, 0.86, 1.04, 1.05, 1.02, 1.06, 1~
## $ OD                   <dbl> 3.92, 3.40, 3.17, 3.45, 2.93, 2.85, 3.58, 3.58, 2~
## $ Proline              <int> 1065, 1050, 1185, 1480, 735, 1450, 1290, 1295, 10~
```

```r
class(wine_data)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

2 Veri Ön İşleme

2.1 Veri öz nitelikleri

2.1.1 Seçilen veri setinde analiz için kullanılacak özelliklerin belirlenmesi

```
features = select(wine_data, Alcohol:Proline)
features
```

```
## # A tibble: 178 x 13
##    Alcohol Malic.acid   Ash   Acl    Mg Phenols Flavanoids Nonflavanoid.phenols
##      <dbl>      <dbl> <dbl> <dbl> <int>   <dbl>      <dbl>                <dbl>
##  1    14.2       1.71  2.43  15.6   127    2.8        3.06                 0.28
##  2    13.2       1.78  2.14  11.2   100    2.65       2.76                 0.26
##  3    13.2       2.36  2.67  18.6   101    2.8        3.24                 0.3
##  4    14.4       1.95  2.5   16.8   113    3.85       3.49                 0.24
##  5    13.2       2.59  2.87  21     118    2.8        2.69                 0.39
##  6    14.2       1.76  2.45  15.2   112    3.27       3.39                 0.34
##  7    14.4       1.87  2.45  14.6    96    2.5        2.52                 0.3
##  8    14.1       2.15  2.61  17.6   121    2.6        2.51                 0.31
##  9    14.8       1.64  2.17  14      97    2.8        2.98                 0.29
## 10    13.9       1.35  2.27  16      98    2.98       3.15                 0.22
## # i 168 more rows
## # i 5 more variables: Proanth <dbl>, Color.int <dbl>, Hue <dbl>, OD <dbl>,
## #   Proline <int>
```

```
summary(features)
```

```
##     Alcohol        Malic.acid         Ash             Acl
##  Min.   :11.03   Min.   :0.740   Min.   :1.360   Min.   :10.60
##  1st Qu.:12.36   1st Qu.:1.603   1st Qu.:2.210   1st Qu.:17.20
##  Median :13.05   Median :1.865   Median :2.360   Median :19.50
##  Mean   :13.00   Mean   :2.336   Mean   :2.367   Mean   :19.49
##  3rd Qu.:13.68   3rd Qu.:3.083   3rd Qu.:2.558   3rd Qu.:21.50
##  Max.   :14.83   Max.   :5.800   Max.   :3.230   Max.   :30.00
##        Mg            Phenols        Flavanoids    Nonflavanoid.phenols
##  Min.   : 70.00   Min.   :0.980   Min.   :0.340   Min.   :0.1300
##  1st Qu.: 88.00   1st Qu.:1.742   1st Qu.:1.205   1st Qu.:0.2700
##  Median : 98.00   Median :2.355   Median :2.135   Median :0.3400
##  Mean   : 99.74   Mean   :2.295   Mean   :2.029   Mean   :0.3619
##  3rd Qu.:107.00   3rd Qu.:2.800   3rd Qu.:2.875   3rd Qu.:0.4375
##  Max.   :162.00   Max.   :3.880   Max.   :5.080   Max.   :0.6600
##     Proanth        Color.int          Hue               OD
##  Min.   :0.410   Min.   : 1.280   Min.   :0.4800   Min.   :1.270
##  1st Qu.:1.250   1st Qu.: 3.220   1st Qu.:0.7825   1st Qu.:1.938
##  Median :1.555   Median : 4.690   Median :0.9650   Median :2.780
##  Mean   :1.591   Mean   : 5.058   Mean   :0.9574   Mean   :2.612
##  3rd Qu.:1.950   3rd Qu.: 6.200   3rd Qu.:1.1200   3rd Qu.:3.170
##  Max.   :3.580   Max.   :13.000   Max.   :1.7100   Max.   :4.000
##     Proline
##  Min.   : 278.0
```

```
##   1st Qu.: 500.5
##   Median : 673.5
##   Mean   : 746.9
##   3rd Qu.: 985.0
##   Max.   :1680.0
```

```
correlation_matrix = cor(features)
head(correlation_matrix)
```

```
##                Alcohol  Malic.acid       Ash         Acl          Mg    Phenols
## Alcohol     1.00000000  0.09439694 0.2115446 -0.31023514  0.27079823  0.2891011
## Malic.acid  0.09439694  1.00000000 0.1640455  0.28850040 -0.05457510 -0.3351670
## Ash         0.21154460  0.16404547 1.0000000  0.44336719  0.28658669  0.1289795
## Acl        -0.31023514  0.28850040 0.4433672  1.00000000 -0.08333309 -0.3211133
## Mg          0.27079823 -0.05457510 0.2865867 -0.08333309  1.00000000  0.2144012
## Phenols     0.28910112 -0.33516700 0.1289795 -0.32111332  0.21440123  1.0000000
##             Flavanoids Nonflavanoid.phenols      Proanth   Color.int         Hue
## Alcohol      0.2368149           -0.1559295  0.136697912  0.54636420 -0.07174720
## Malic.acid  -0.4110066            0.2929771 -0.220746187  0.24898534 -0.56129569
## Ash          0.1150773            0.1862304  0.009651935  0.25888726 -0.07466689
## Acl         -0.3513699            0.3619217 -0.197326836  0.01873198 -0.27395522
## Mg           0.1957838           -0.2562940  0.236440610  0.19995001  0.05539820
## Phenols      0.8645635           -0.4499353  0.612413084 -0.05513642  0.43368134
##                     OD    Proline
## Alcohol     0.072343187  0.6437200
## Malic.acid -0.368710428 -0.1920106
## Ash         0.003911231  0.2236263
## Acl        -0.276768549 -0.4405969
## Mg          0.066003936  0.3933508
## Phenols     0.699949365  0.4981149
```

2.2 Değişken seçimi ve dönüşüm işlemleri

```
#Seçilen sayısal değişkenler gather fonksiyonu ile uzun formatlı hale getirildi.
(long_data = wine_data %>% keep(is.numeric) %>% gather())
```

```
## # A tibble: 2,492 x 2
##      key   value
##      <chr> <dbl>
##   1 Wine       1
##   2 Wine       1
##   3 Wine       1
##   4 Wine       1
##   5 Wine       1
##   6 Wine       1
##   7 Wine       1
##   8 Wine       1
##   9 Wine       1
##  10 Wine       1
## # i 2,482 more rows
```

2.3 dplyr paketi ile temel işlemler(veri seçme ve filtreleme)

```
filter(wine_data, Alcohol > 13 & Phenols > 2)
```

```
## # A tibble: 66 x 14
##     Wine Alcohol Malic.acid   Ash   Acl    Mg Phenols Flavanoids
##    <int>   <dbl>      <dbl> <dbl> <dbl> <int>   <dbl>      <dbl>
##  1     1    14.2       1.71  2.43  15.6   127    2.8        3.06
##  2     1    13.2       1.78  2.14  11.2   100    2.65       2.76
##  3     1    13.2       2.36  2.67  18.6   101    2.8        3.24
##  4     1    14.4       1.95  2.5   16.8   113    3.85       3.49
##  5     1    13.2       2.59  2.87  21     118    2.8        2.69
##  6     1    14.2       1.76  2.45  15.2   112    3.27       3.39
##  7     1    14.4       1.87  2.45  14.6    96    2.5        2.52
##  8     1    14.1       2.15  2.61  17.6   121    2.6        2.51
##  9     1    14.8       1.64  2.17  14      97    2.8        2.98
## 10     1    13.9       1.35  2.27  16      98    2.98       3.15
## # i 56 more rows
## # i 6 more variables: Nonflavanoid.phenols <dbl>, Proanth <dbl>,
## #   Color.int <dbl>, Hue <dbl>, OD <dbl>, Proline <int>
```

```
wine_data %>%
  group_by(Wine) %>%
  summarise(count = n())
```

```
## # A tibble: 3 x 2
##    Wine count
##   <int> <int>
## 1     1    59
## 2     2    71
## 3     3    48
```

```
grouped_data <- wine_data %>%
  group_by(Wine) %>%
  summarise(mean_Alcohol = mean(Alcohol), mean_Color_Int = mean(Color.int))
print(grouped_data)
```

```
## # A tibble: 3 x 3
##    Wine mean_Alcohol mean_Color_Int
##   <int>        <dbl>          <dbl>
## 1     1         13.7           5.53
## 2     2         12.3           3.09
## 3     3         13.2           7.40
```

3 Veri Manipülasyonu

3.1 Veri setinin özelliklerinin analize hazır hale getirilmesi(reshaping data)

```
normalized_data = scale(wine_data[, 2:ncol(wine_data)])
head(normalized_data)
```

```
##        Alcohol  Malic.acid        Ash        Acl          Mg   Phenols
## [1,] 1.5143408 -0.56066822  0.2313998 -1.1663032 1.90852151 0.8067217
## [2,] 0.2455968 -0.49800856 -0.8256672 -2.4838405 0.01809398 0.5670481
## [3,] 0.1963252  0.02117152  1.1062139 -0.2679823 0.08810981 0.8067217
## [4,] 1.6867914 -0.34583508  0.4865539 -0.8069748 0.92829983 2.4844372
## [5,] 0.2948684  0.22705328  1.8352256  0.4506745 1.27837900 0.8067217
## [6,] 1.4773871 -0.51591132  0.3043010 -1.2860793 0.85828399 1.5576991
##      Flavanoids Nonflavanoid.phenols    Proanth  Color.int        Hue        OD
## [1,]  1.0319081           -0.6577078  1.2214385  0.2510088  0.3611585 1.8427215
## [2,]  0.7315653           -0.8184106 -0.5431887 -0.2924962  0.4049085 1.1103172
## [3,]  1.2121137           -0.4970050  2.1299594  0.2682629  0.3174085 0.7863692
## [4,]  1.4623994           -0.9791134  1.0292513  1.1827317 -0.4263410 1.1807407
## [5,]  0.6614853            0.2261576  0.4002753 -0.3183774  0.3611585 0.4483365
## [6,]  1.3622851           -0.1755994  0.6623487  0.7298108  0.4049085 0.3356589
##          Proline
## [1,]  1.01015939
## [2,]  0.96252635
## [3,]  1.39122370
## [4,]  2.32800680
## [5,] -0.03776747
## [6,]  2.23274072
```

3.2 Eksik veri ve aykırı değerlerin tespiti

```
missing_values = wine_data %>%
  summarise_all(~ sum(is.na(.)))
missing_values
```

```
## # A tibble: 1 x 14
##    Wine Alcohol Malic.acid   Ash   Acl    Mg Phenols Flavanoids
##   <int>   <int>      <int> <int> <int> <int>   <int>      <int>
## 1     0       0          0     0     0     0       0          0
## # i 6 more variables: Nonflavanoid.phenols <int>, Proanth <int>,
## #   Color.int <int>, Hue <int>, OD <int>, Proline <int>
```

```
outliers = wine_data %>%
  filter_all(all_vars(!is.na(.) & (. < quantile(., 0.25) - 1.5 * IQR(.) | . > quantile(., 0.75) + 1.5 *
outliers
```

```
## # A tibble: 0 x 14
## # i 14 variables: Wine <int>, Alcohol <dbl>, Malic.acid <dbl>, Ash <dbl>,
## #   Acl <dbl>, Mg <int>, Phenols <dbl>, Flavanoids <dbl>,
## #   Nonflavanoid.phenols <dbl>, Proanth <dbl>, Color.int <dbl>, Hue <dbl>,
## #   OD <dbl>, Proline <int>
```

3.3 Eksik verilerin tamamlanması ya da analiz dışı bırakılması

Eksik veri bulunmamıştır.

3.4 Veri normalizasyonu ya da standardizasyonu

```r
normalize_et = function(x) {
  return((x - min(x)) / (max(x) - min(x)))
}

veri_setini_normalize_et = function(veri_seti) {
  normalize_edilmis_set = as.data.frame(lapply(veri_seti, function(col) {
    if (is.numeric(col)) {
      return(normalize_et(col))
    } else {
      return(col)
    }
  }))
  return(normalize_edilmis_set)
}

normalize_data = veri_setini_normalize_et(wine_data)
head(normalize_data)
```

```
##   Wine   Alcohol Malic.acid       Ash        Acl        Mg   Phenols Flavanoids
## 1    0 0.8421053  0.1916996 0.5721925 0.25773196 0.6195652 0.6275862  0.5738397
## 2    0 0.5710526  0.2055336 0.4171123 0.03092784 0.3260870 0.5758621  0.5105485
## 3    0 0.5605263  0.3201581 0.7005348 0.41237113 0.3369565 0.6275862  0.6118143
## 4    0 0.8789474  0.2391304 0.6096257 0.31958763 0.4673913 0.9896552  0.6645570
## 5    0 0.5815789  0.3656126 0.8074866 0.53608247 0.5217391 0.6275862  0.4957806
## 6    0 0.8342105  0.2015810 0.5828877 0.23711340 0.4565217 0.7896552  0.6434599
##   Nonflavanoid.phenols   Proanth Color.int       Hue        OD   Proline
## 1            0.2830189 0.5930599 0.3720137 0.4552846 0.9706960 0.5613409
## 2            0.2452830 0.2744479 0.2645051 0.4634146 0.7802198 0.5506419
## 3            0.3207547 0.7570978 0.3754266 0.4471545 0.6959707 0.6469330
## 4            0.2075472 0.5583596 0.5563140 0.3089431 0.7985348 0.8573466
## 5            0.4905660 0.4447950 0.2593857 0.4552846 0.6080586 0.3259629
## 6            0.3962264 0.4921136 0.4667235 0.4634146 0.5787546 0.8359486
```

```r
standardize_et = function(x) {
  return((x - mean(x)) / sd(x))
}

veri_setini_standardize_et = function(veri_seti) {
  standardize_edilmis_set = as.data.frame(lapply(veri_seti, function(col) {
    if (is.numeric(col)) {
      return(standardize_et(col))
    } else {
      return(col)
    }
  }))
  return(standardize_edilmis_set)
}

standardize_data = veri_setini_standardize_et(wine_data)
head(standardize_data)
```

```
##        Wine   Alcohol  Malic.acid        Ash        Acl         Mg   Phenols
## 1 -1.210529 1.5143408 -0.56066822  0.2313998 -1.1663032 1.90852151 0.8067217
## 2 -1.210529 0.2455968 -0.49800856 -0.8256672 -2.4838405 0.01809398 0.5670481
## 3 -1.210529 0.1963252  0.02117152  1.1062139 -0.2679823 0.08810981 0.8067217
## 4 -1.210529 1.6867914 -0.34583508  0.4865539 -0.8069748 0.92829983 2.4844372
## 5 -1.210529 0.2948684  0.22705328  1.8352256  0.4506745 1.27837900 0.8067217
## 6 -1.210529 1.4773871 -0.51591132  0.3043010 -1.2860793 0.85828399 1.5576991
##   Flavanoids Nonflavanoid.phenols    Proanth  Color.int        Hue        OD
## 1  1.0319081           -0.6577078  1.2214385  0.2510088  0.3611585 1.8427215
## 2  0.7315653           -0.8184106 -0.5431887 -0.2924962  0.4049085 1.1103172
## 3  1.2121137           -0.4970050  2.1299594  0.2682629  0.3174085 0.7863692
## 4  1.4623994           -0.9791134  1.0292513  1.1827317 -0.4263410 1.1807407
## 5  0.6614853            0.2261576  0.4002753 -0.3183774  0.3611585 0.4483365
## 6  1.3622851           -0.1755994  0.6623487  0.7298108  0.4049085 0.3356589
##       Proline
## 1  1.01015939
## 2  0.96252635
## 3  1.39122370
## 4  2.32800680
## 5 -0.03776747
## 6  2.23274072
```

3.5 Veri seçme ve filtreleme işlemlerinin gerçekleştirilmesi

```
Alcohol_Category = cut(wine_data$Alcohol, breaks = c(0, 12, 14, 16), labels = c("Low", "Medium", "High")
```

3.6 Yeni hesaplamaların veri setine dâhil edilmesi

```
wine_data = wine_data %>%
  mutate(Alcohol_Category = cut(Alcohol, breaks = c(0, 12, 14, 16), labels = c("Low", "Medium", "High"))

head(wine_data)
```

```
## # A tibble: 6 x 15
##    Wine Alcohol Malic.acid   Ash   Acl    Mg Phenols Flavanoids
##   <int>   <dbl>      <dbl> <dbl> <dbl> <int>   <dbl>      <dbl>
## 1     1    14.2       1.71  2.43  15.6   127    2.8        3.06
## 2     1    13.2       1.78  2.14  11.2   100    2.65       2.76
## 3     1    13.2       2.36  2.67  18.6   101    2.8        3.24
## 4     1    14.4       1.95  2.5   16.8   113    3.85       3.49
## 5     1    13.2       2.59  2.87  21     118    2.8        2.69
## 6     1    14.2       1.76  2.45  15.2   112    3.27       3.39
## # i 7 more variables: Nonflavanoid.phenols <dbl>, Proanth <dbl>,
## #   Color.int <dbl>, Hue <dbl>, OD <dbl>, Proline <int>, Alcohol_Category <fct>
```

```
wine_data %>%
  group_by(Alcohol_Category) %>%
  summarise(count = n())
```

```
## # A tibble: 3 x 2
##   Alcohol_Category count
##   <fct>            <int>
## 1 Low                 22
## 2 Medium             134
## 3 High                22
```

3.7 Temel istatistiklerin hesaplanması

```r
summary(wine_data)
```

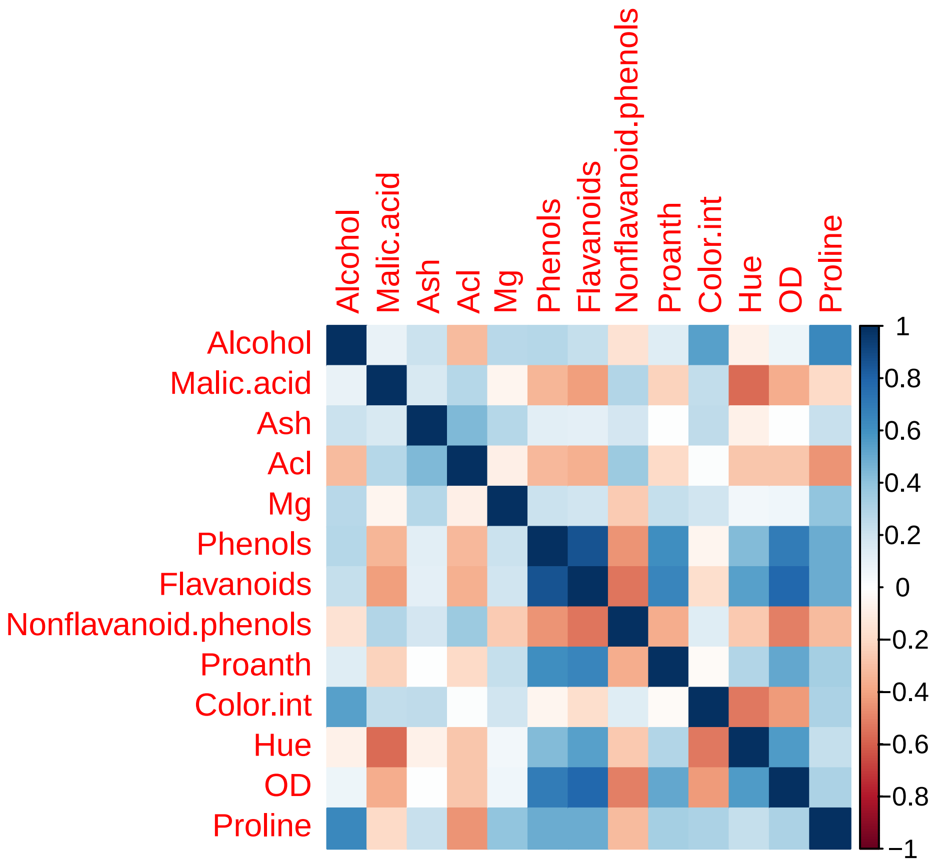
```
##       Wine          Alcohol        Malic.acid         Ash       
##  Min.   :1.000   Min.   :11.03   Min.   :0.740   Min.   :1.360  
##  1st Qu.:1.000   1st Qu.:12.36   1st Qu.:1.603   1st Qu.:2.210  
##  Median :2.000   Median :13.05   Median :1.865   Median :2.360  
##  Mean   :1.938   Mean   :13.00   Mean   :2.336   Mean   :2.367  
##  3rd Qu.:3.000   3rd Qu.:13.68   3rd Qu.:3.083   3rd Qu.:2.558  
##  Max.   :3.000   Max.   :14.83   Max.   :5.800   Max.   :3.230  
##       Acl             Mg            Phenols        Flavanoids   
##  Min.   :10.60   Min.   : 70.00   Min.   :0.980   Min.   :0.340  
##  1st Qu.:17.20   1st Qu.: 88.00   1st Qu.:1.742   1st Qu.:1.205  
##  Median :19.50   Median : 98.00   Median :2.355   Median :2.135  
##  Mean   :19.49   Mean   : 99.74   Mean   :2.295   Mean   :2.029  
##  3rd Qu.:21.50   3rd Qu.:107.00   3rd Qu.:2.800   3rd Qu.:2.875  
##  Max.   :30.00   Max.   :162.00   Max.   :3.880   Max.   :5.080  
##  Nonflavanoid.phenols    Proanth        Color.int           Hue        
##  Min.   :0.1300       Min.   :0.410   Min.   : 1.280   Min.   :0.4800  
##  1st Qu.:0.2700       1st Qu.:1.250   1st Qu.: 3.220   1st Qu.:0.7825  
##  Median :0.3400       Median :1.555   Median : 4.690   Median :0.9650  
##  Mean   :0.3619       Mean   :1.591   Mean   : 5.058   Mean   :0.9574  
##  3rd Qu.:0.4375       3rd Qu.:1.950   3rd Qu.: 6.200   3rd Qu.:1.1200  
##  Max.   :0.6600       Max.   :3.580   Max.   :13.000   Max.   :1.7100  
##        OD           Proline       Alcohol_Category
##  Min.   :1.270   Min.   : 278.0   Low   : 22      
##  1st Qu.:1.938   1st Qu.: 500.5   Medium:134      
##  Median :2.780   Median : 673.5   High  : 22      
##  Mean   :2.612   Mean   : 746.9                   
##  3rd Qu.:3.170   3rd Qu.: 985.0                   
##  Max.   :4.000   Max.   :1680.0                   
```

```r
library(psych)
describe(wine_data)
```

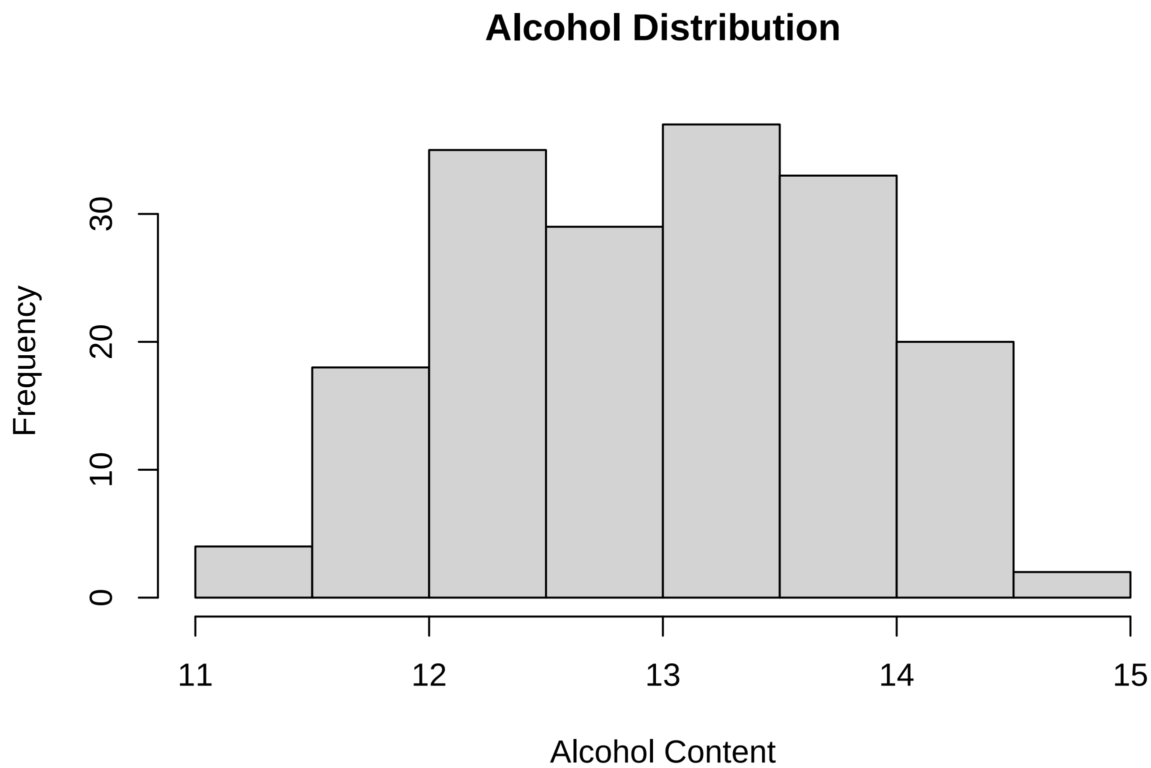
```
##                         vars   n   mean    sd median trimmed   mad   min
## Wine                       1 178   1.94  0.78   2.00    1.92  1.48  1.00
## Alcohol                    2 178  13.00  0.81  13.05   13.01  1.01 11.03
## Malic.acid                 3 178   2.34  1.12   1.87    2.21  0.77  0.74
## Ash                        4 178   2.37  0.27   2.36    2.37  0.24  1.36
## Acl                        5 178  19.49  3.34  19.50   19.42  3.04 10.60
## Mg                         6 178  99.74 14.28  98.00   98.44 14.83 70.00
## Phenols                    7 178   2.30  0.63   2.36    2.29  0.75  0.98
```

```
## Flavanoids             8 178   2.03   1.00   2.13    2.02   1.24   0.34
## Nonflavanoid.phenols   9 178   0.36   0.12   0.34    0.36   0.13   0.13
## Proanth               10 178   1.59   0.57   1.56    1.56   0.56   0.41
## Color.int             11 178   5.06   2.32   4.69    4.83   2.24   1.28
## Hue                   12 178   0.96   0.23   0.96    0.96   0.24   0.48
## OD                    13 178   2.61   0.71   2.78    2.63   0.77   1.27
## Proline               14 178 746.89 314.91 673.50  719.30 300.23 278.00
## Alcohol_Category*     15 178   2.00   0.50   2.00    2.00   0.00   1.00
##                          max   range  skew kurtosis    se
## Wine                    3.00    2.00  0.11    -1.34  0.06
## Alcohol                14.83    3.80 -0.05    -0.89  0.06
## Malic.acid              5.80    5.06  1.02     0.22  0.08
## Ash                     3.23    1.87 -0.17     1.03  0.02
## Acl                    30.00   19.40  0.21     0.40  0.25
## Mg                    162.00   92.00  1.08     1.96  1.07
## Phenols                 3.88    2.90  0.09    -0.87  0.05
## Flavanoids              5.08    4.74  0.02    -0.91  0.07
## Nonflavanoid.phenols    0.66    0.53  0.44    -0.68  0.01
## Proanth                 3.58    3.17  0.51     0.47  0.04
## Color.int              13.00   11.72  0.85     0.30  0.17
## Hue                     1.71    1.23  0.02    -0.40  0.02
## OD                      4.00    2.73 -0.30    -1.11  0.05
## Proline              1680.00 1402.00  0.75    -0.31 23.60
## Alcohol_Category*       3.00    2.00  0.00     1.00  0.04
```

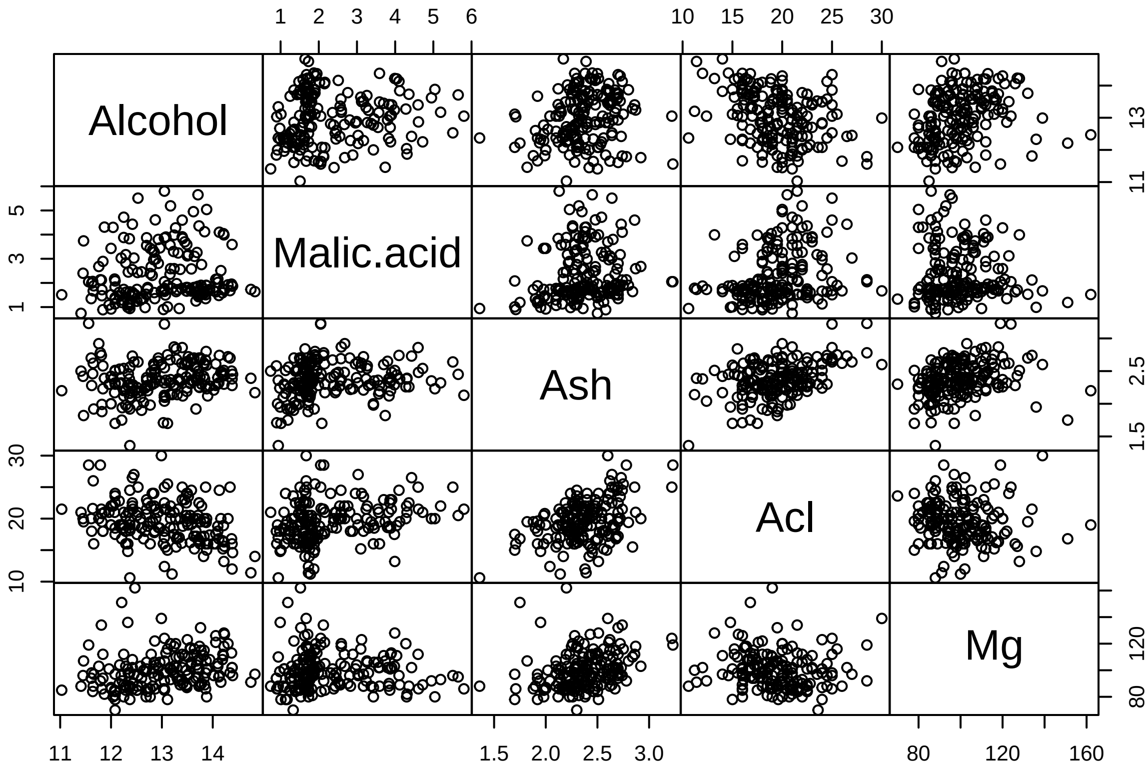
```r
library(corrplot)
corrplot(correlation_matrix, method = "color")
```




```r
# Histogram grafiği
hist(wine_data$Alcohol, main = "Alcohol Distribution", xlab = "Alcohol Content")
```




```r
#Dağılım Grafiği
pairs(wine_data[, 2:6], gap = 0.01)
```



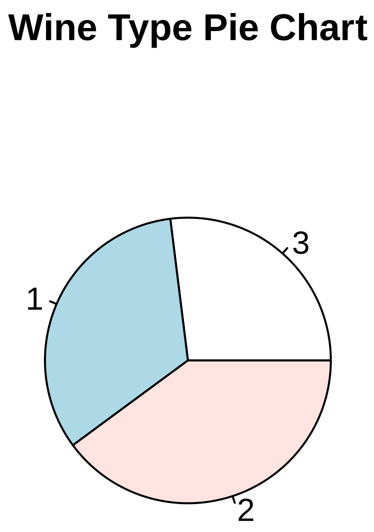
4 Keşifçi ve Açıklayıcı Veri Analizi

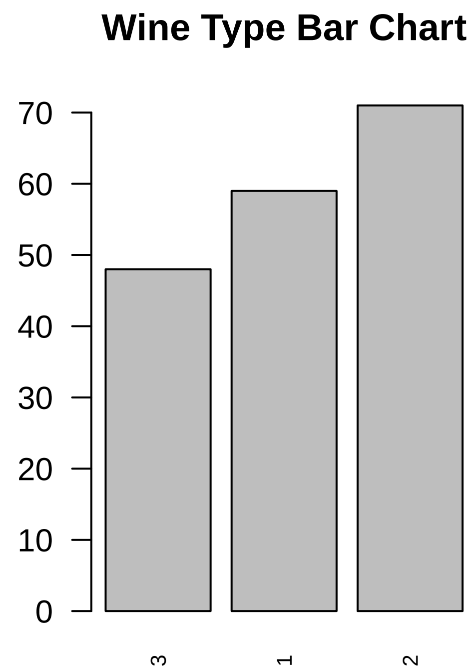
4.1 ggplot2 paketi ile uygun özelliklere ait veri görselleştirmenin gerçekleştirilmesi

```
par(wine_data, mfrow = c(1,2))
```

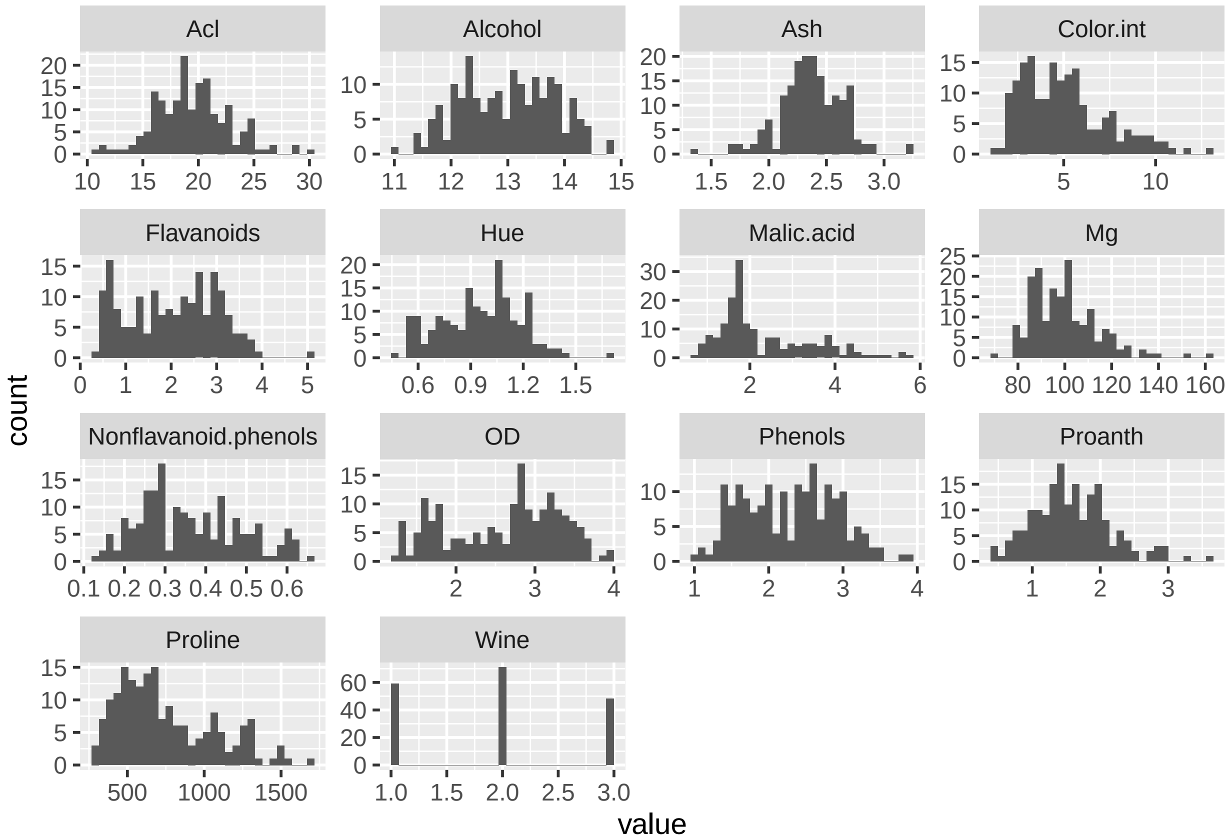
```
## Warning in par(wine_data, mfrow = c(1, 2)): argument 1 does not name a
## graphical parameter
```

```
tbl = sort(table(wine_data$Wine))
pie(tbl)
title("Wine Type Pie Chart")
barplot(tbl, las = 2, cex.names = 0.7)
title("Wine Type Bar Chart")
```





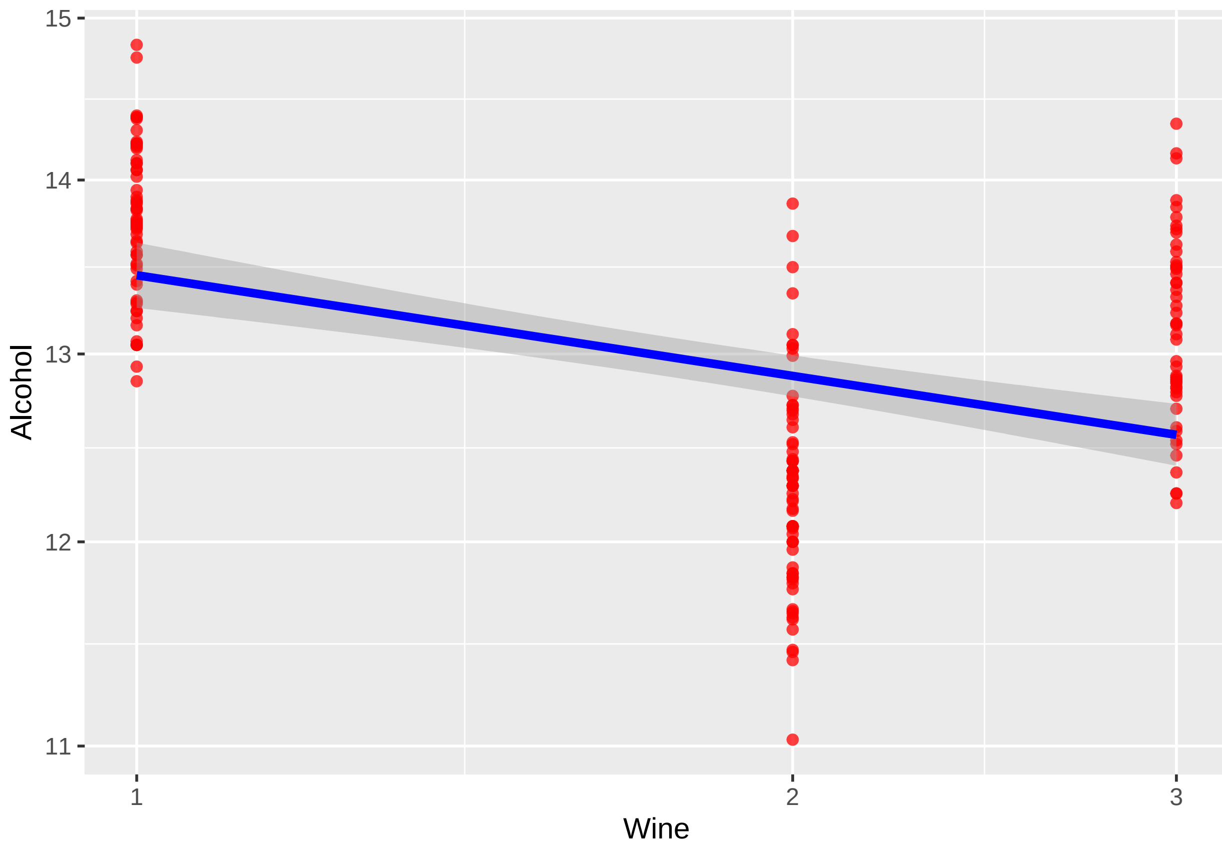

```r
#Her bir sayısal değişkenin histogramı
long_data %>% ggplot(aes(value)) +
  facet_wrap(~ key, scales = "free") + geom_histogram(bins = 30)
```



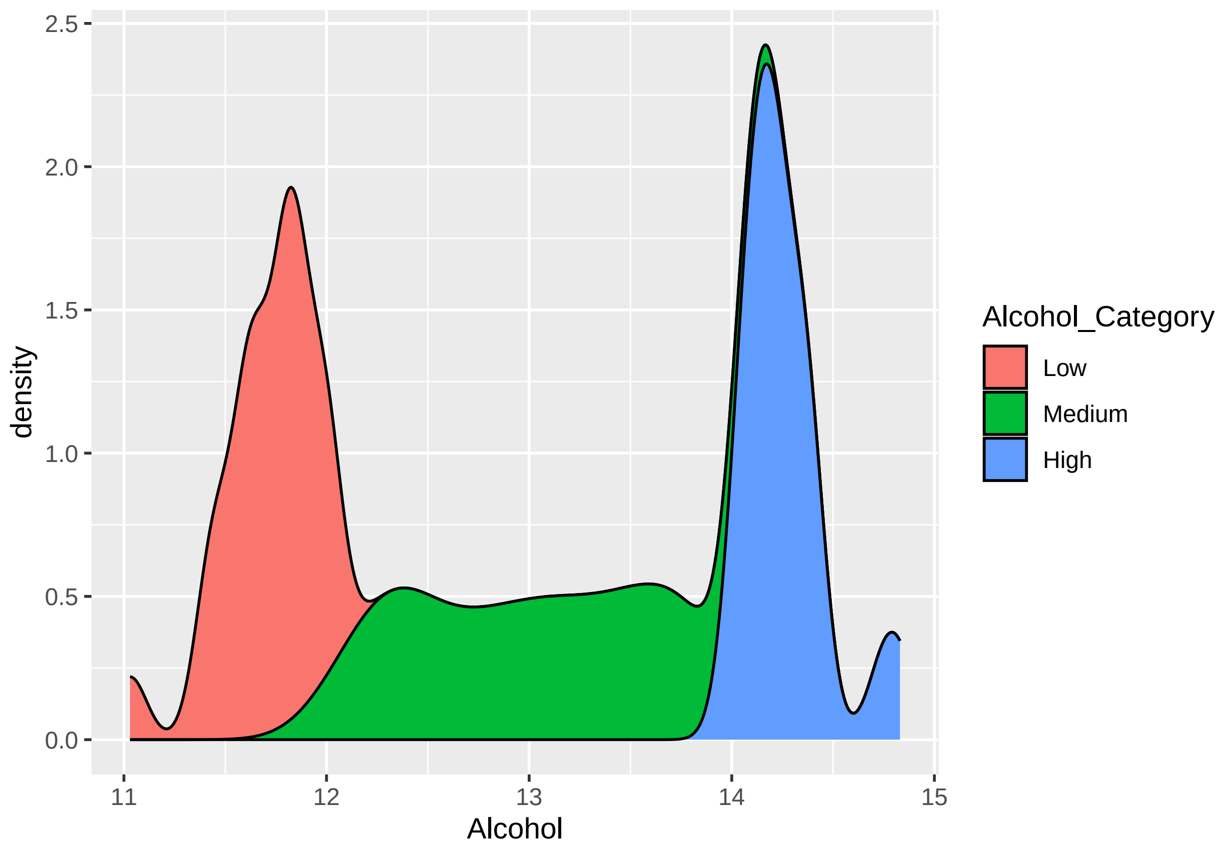

```
ggplot(wine_data, aes(x = Wine, y = Alcohol)) +
  geom_point(alpha = 0.75, col = "red") +
  scale_x_log10() +
  scale_y_log10() +
  stat_smooth(method = "lm", se = T, col = "blue", size = 1.5)
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```




```
ggplot(data = wine_data, aes(Alcohol)) + geom_density(aes(fill = Alcohol_Category), position = "stack")
```



MACHINE LEARNING ALGORITHMS

KNN

```
library(class)
```

```
## Warning: package 'class' was built under R version 4.3.2
```

```
set.seed(123)
index = sample(1:nrow(wine_data), 0.7 * nrow(wine_data))
train_data = wine_data[index, ]
test_data = wine_data[-index, ]
```

```
k <- 3
knn_model = knn(train = train_data[, 2:ncol(normalize_data)],
                test = test_data[, 2:ncol(normalize_data)],
                cl = train_data$Wine,
                k = k)
```

```
# Confusion matrix
conf_matrix = table(Actual = test_data$Wine, Predicted = knn_model)
conf_matrix
```

```
##          Predicted
## Actual   1  2  3
```

```
##      1 17  0  2
##      2  1 14  9
##      3  0  1 10
```

```
# Accuracy değeri
accuracy = sum(diag(conf_matrix)) / sum(conf_matrix)
cat("Accuracy:", accuracy, "\n")
```

```
## Accuracy: 0.7592593
```

Logistic Regression

```
set.seed(123)
index = sample(1:nrow(wine_data), 0.7 * nrow(wine_data))
train_data = wine_data[index, ]
test_data = wine_data[-index, ]
```

```
glm_model = glm(as.factor(Wine) ~ ., data = train_data, family = "binomial")
```

```
glm_predictions = predict(glm_model, test_data, type = "response")
```

```
glm_predictions = ifelse(glm_predictions > 0.5, "Class_2", "Class_1")
```

```
# Confusion matrix
conf_matrix_glm = table(Actual = test_data$Wine, Predicted = glm_predictions)
conf_matrix_glm
```

```
##       Predicted
## Actual Class_1 Class_2
##      1      19       0
##      2       0      24
##      3       0      11
```

```
# Accuracy değeri
accuracy_glm = sum(diag(conf_matrix_glm)) / sum(conf_matrix_glm)
cat("Logistic Regression Accuracy:", accuracy_glm, "\n")
```

```
## Logistic Regression Accuracy: 0.7962963
```

DECISION TREE

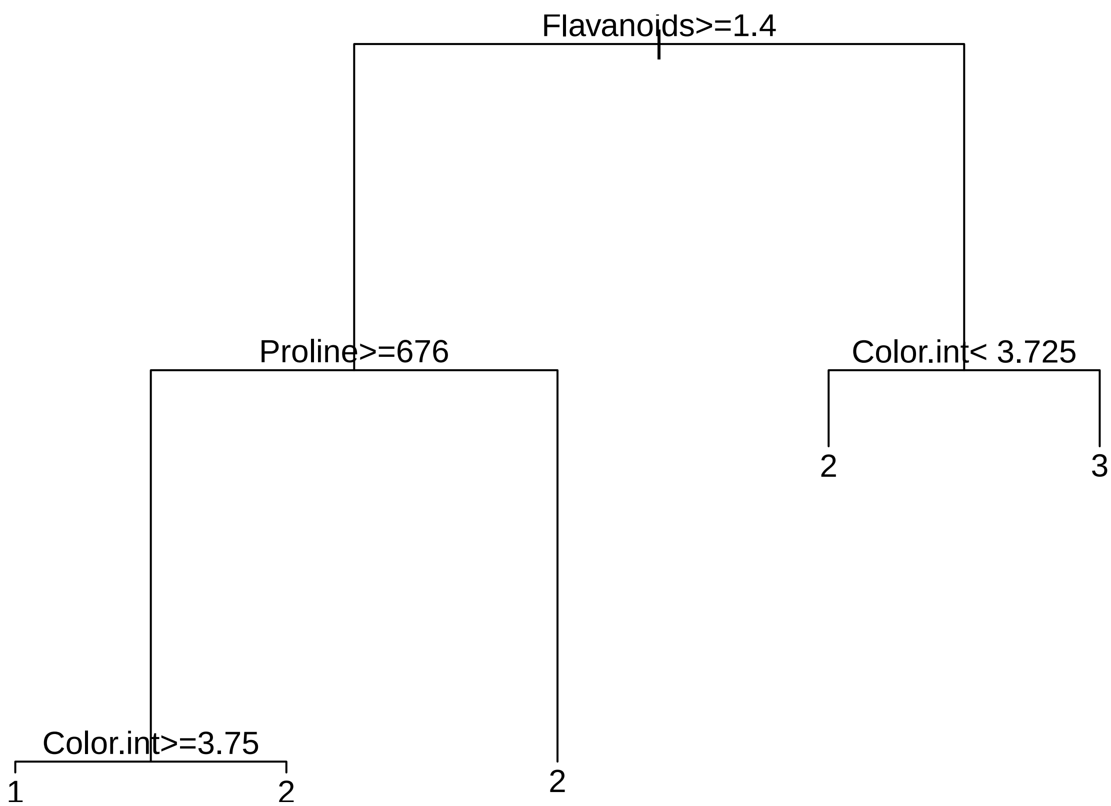
```
library(rpart)
```

```
## Warning: package 'rpart' was built under R version 4.3.2
```

```
set.seed(123)
index = sample(1:nrow(wine_data), 0.7 * nrow(wine_data))
train_data = wine_data[index, ]
test_data = wine_data[-index, ]
```

```r
tree_model <- rpart(as.factor(Wine) ~ ., data = train_data, method = "class")
```

```r
par(mar = c(1, 1, 1, 1))
plot(tree_model)
text(tree_model)
```




```r
tree_predictions = predict(tree_model, test_data, type = "class")
```

```r
# Confusion matrix
conf_matrix_tree = table(Actual = test_data$Wine, Predicted = tree_predictions)
conf_matrix_tree
```

```
##        Predicted
## Actual  1  2  3
##       1 19  0  0
##       2  2 22  0
##       3  0  1 10
```

```r
# Accuracy değeri
accuracy_tree = sum(diag(conf_matrix_tree)) / sum(conf_matrix_tree)
cat("Decision Tree Accuracy:", accuracy_tree, "\n")
```

```
## Decision Tree Accuracy: 0.9444444
```

SVM

```r
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 4.3.2
```

```r
wine_data$Wine = as.factor(wine_data$Wine)
```

```r
set.seed(123)
indices = sample(1:nrow(wine_data), 0.7 * nrow(wine_data))
train_data = wine_data[indices, ]
test_data = wine_data[-indices, ]
```

```r
svm_model = svm(Wine ~ ., data = train_data, kernel = "linear")
predictions = predict(svm_model, newdata = test_data)
```

```r
# Confusion matrix
conf_matrix_svm = table(Actual = test_data$Wine, Predicted = predictions)
conf_matrix_svm
```

```
##       Predicted
## Actual  1  2  3
##      1 19  0  0
##      2  0 24  0
##      3  0  1 10
```

```r
# Accuracy değeri
accuracy = sum(predictions == test_data$Wine) / nrow(test_data)
cat("SVM Accuracy:", accuracy, "\n")
```

```
## SVM Accuracy: 0.9814815
```