Predictive Health Assessment: Leveraging Machine Learning to Gauge Lifestyle Impacts

Our Dataset: EFFECTS OF SMOKING AND DRINKING ON HEALTH

Modified dataset to 100,000 records for usability purpose: MODIFIED DATA

```
library(readxl)
data <- read_excel("data.xlsx")
View(data)
summary(data)
```

```
##      sex                 age              height          weight      
##  Length:100000      Min.   :20.00   Min.   :130.0   Min.   : 25.00  
##  Class :character   1st Qu.:35.00   1st Qu.:155.0   1st Qu.: 55.00  
##  Mode  :character   Median :45.00   Median :160.0   Median : 60.00  
##                     Mean   :47.65   Mean   :162.2   Mean   : 63.35  
##                     3rd Qu.:60.00   3rd Qu.:170.0   3rd Qu.: 70.00  
##                     Max.   :85.00   Max.   :190.0   Max.   :140.00  
##    waistline        sight_left      sight_right          hear       
##  Min.   : 35.00   Min.   :0.1000   Min.   :0.1000   Min.   :0.0000  
##  1st Qu.: 74.20   1st Qu.:0.7000   1st Qu.:0.7000   1st Qu.:1.0000  
##  Median : 81.00   Median :1.0000   Median :1.0000   Median :1.0000  
##  Mean   : 81.28   Mean   :0.9811   Mean   :0.9787   Mean   :0.9702  
##  3rd Qu.: 88.00   3rd Qu.:1.2000   3rd Qu.:1.2000   3rd Qu.:1.0000  
##  Max.   :999.00   Max.   :9.9000   Max.   :9.9000   Max.   :1.0000  
##       SBP              DBP              BLDS         tot_chole     
##  Min.   : 72.0   Min.   : 39.00   Min.   : 34.0   Min.   :  58.0  
##  1st Qu.:112.0   1st Qu.: 70.00   1st Qu.: 88.0   1st Qu.: 169.0  
##  Median :120.0   Median : 76.00   Median : 96.0   Median : 193.0  
##  Mean   :122.4   Mean   : 76.04   Mean   :100.5   Mean   : 195.5  
##  3rd Qu.:131.0   3rd Qu.: 82.00   3rd Qu.:105.0   3rd Qu.: 219.0  
##  Max.   :220.0   Max.   :160.00   Max.   :784.0   Max.   :2067.0  
##    HDL_chole        LDL_chole       triglyceride      hemoglobin    
##  Min.   :  2.00   Min.   :   1.0   Min.   :   7.0   Min.   : 3.90  
##  1st Qu.: 46.00   1st Qu.:  89.0   1st Qu.:  74.0   1st Qu.:13.20  
##  Median : 55.00   Median : 111.0   Median : 106.0   Median :14.30  
##  Mean   : 56.91   Mean   : 113.1   Mean   : 132.2   Mean   :14.23  
##  3rd Qu.: 66.00   3rd Qu.: 135.0   3rd Qu.: 159.0   3rd Qu.:15.40  
##  Max.   :636.00   Max.   :2111.0   Max.   :5236.0   Max.   :23.30  
##  urine_protein   serum_creatinine     SGOT_AST          SGOT_ALT      
##  Min.   :1.000   Min.   : 0.1000   Min.   :   1.00   Min.   :   1.00  
##  1st Qu.:1.000   1st Qu.: 0.7000   1st Qu.:  19.00   1st Qu.:  15.00  
##  Median :1.000   Median : 0.8000   Median :  23.00   Median :  20.00  
##  Mean   :1.095   Mean   : 0.8605   Mean   :  25.94   Mean   :  25.79  
##  3rd Qu.:1.000   3rd Qu.: 1.0000   3rd Qu.:  28.00   3rd Qu.:  30.00  
##  Max.   :6.000   Max.   :68.0000   Max.   :3440.00   Max.   :3517.00
```

```
##    gamma_GTP      SMK_stat_type_cd      DRK_YN
##  Min.   :  2.00   Length:100000      Length:100000
##  1st Qu.: 16.00   Class :character   Class :character
##  Median : 23.00   Mode  :character   Mode  :character
##  Mean   : 37.06
##  3rd Qu.: 40.00
##  Max.   :999.00
```

```
str(data)
```

```
## tibble [100,000 × 23] (S3: tbl_df/tbl/data.frame)
##  $ sex             : chr [1:100000] "Female" "Female" "Male" "Male" ...
##  $ age             : num [1:100000] 55 70 55 50 50 40 70 40 55 45 ...
##  $ height          : num [1:100000] 160 150 170 170 165 165 155 180 150
155 ...
##  $ weight          : num [1:100000] 65 55 80 60 70 55 75 75 50 55 ...
##  $ waistline       : num [1:100000] 98 82 90 73 86 68 103 78 67 76 ...
##  $ sight_left      : num [1:100000] 1.2 0.7 0.9 1.2 0.9 0.1 0.7 1 0.8 1
...
##  $ sight_right     : num [1:100000] 1.5 0.8 1 1.5 0.5 0.8 0.8 1 1 1 ...
##  $ hear            : num [1:100000] 1 1 1 1 1 1 1 1 1 1 ...
##  $ SBP             : num [1:100000] 139 118 116 123 115 94 130 111 98 120
...
##  $ DBP             : num [1:100000] 81 76 79 80 84 56 80 65 66 60 ...
##  $ BLDS            : num [1:100000] 96 95 96 84 110 89 97 93 100 97 ...
##  $ tot_chole       : num [1:100000] 151 267 191 211 137 145 207 177 207
163 ...
##  $ HDL_chole       : num [1:100000] 60 55 57 46 54 50 48 47 98 51 ...
##  $ LDL_chole       : num [1:100000] 75 194 109 145 38 80 113 108 97 100
...
##  $ triglyceride    : num [1:100000] 80 92 250 96 223 72 227 109 59 59 ...
##  $ hemoglobin      : num [1:100000] 13.3 11.4 14 15.6 14.8 12.4 14 15 13.5
15.6 ...
##  $ urine_protein   : num [1:100000] 1 1 1 1 1 1 1 1 1 1 ...
##  $ serum_creatinine: num [1:100000] 0.7 0.8 1 1.2 1.1 0.6 1.1 1 0.5 0.6
...
##  $ SGOT_AST        : num [1:100000] 34 20 26 24 39 20 21 18 17 32 ...
##  $ SGOT_ALT        : num [1:100000] 28 9 28 21 68 19 33 14 15 39 ...
##  $ gamma_GTP       : num [1:100000] 33 11 65 26 56 13 34 20 14 105 ...
##  $ SMK_stat_type_cd: chr [1:100000] "Low" "Low" "High" "High" ...
##  $ DRK_YN          : chr [1:100000] "N" "N" "Y" "Y" ...
```

```
colSums(is.na(data))
```

```
##              sex              age           height           weight
##                0                0                0                0
##        waistline       sight_left      sight_right             hear
##                0                0                0                0
##              SBP              DBP             BLDS        tot_chole
##                0                0                0                0
##        HDL_chole        LDL_chole     triglyceride       hemoglobin
```

```
##                0                0                0                0
##    urine_protein serum_creatinine         SGOT_AST         SGOT_ALT
##                0                0                0                0
##        gamma_GTP SMK_stat_type_cd           DRK_YN
##                0                0                0
```

– EXPLORATORY DATA ANALYSIS

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(dplyr)
```

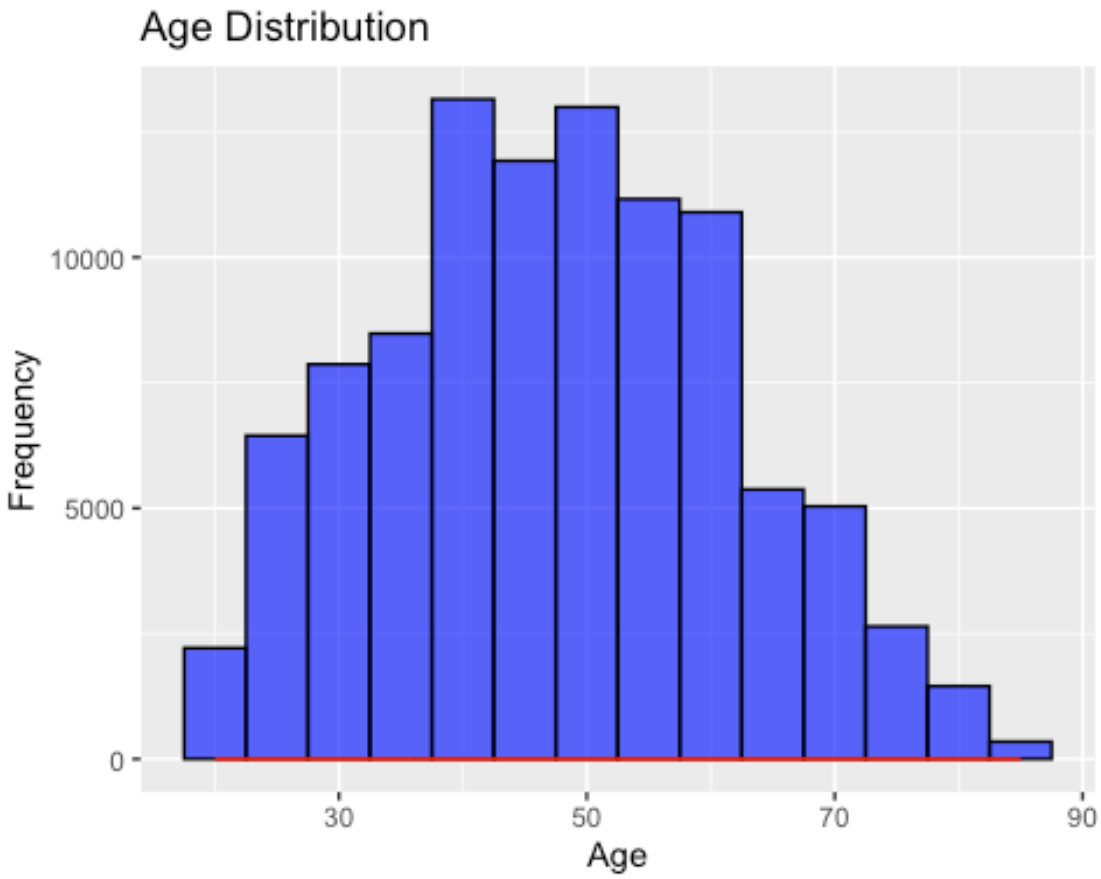
```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

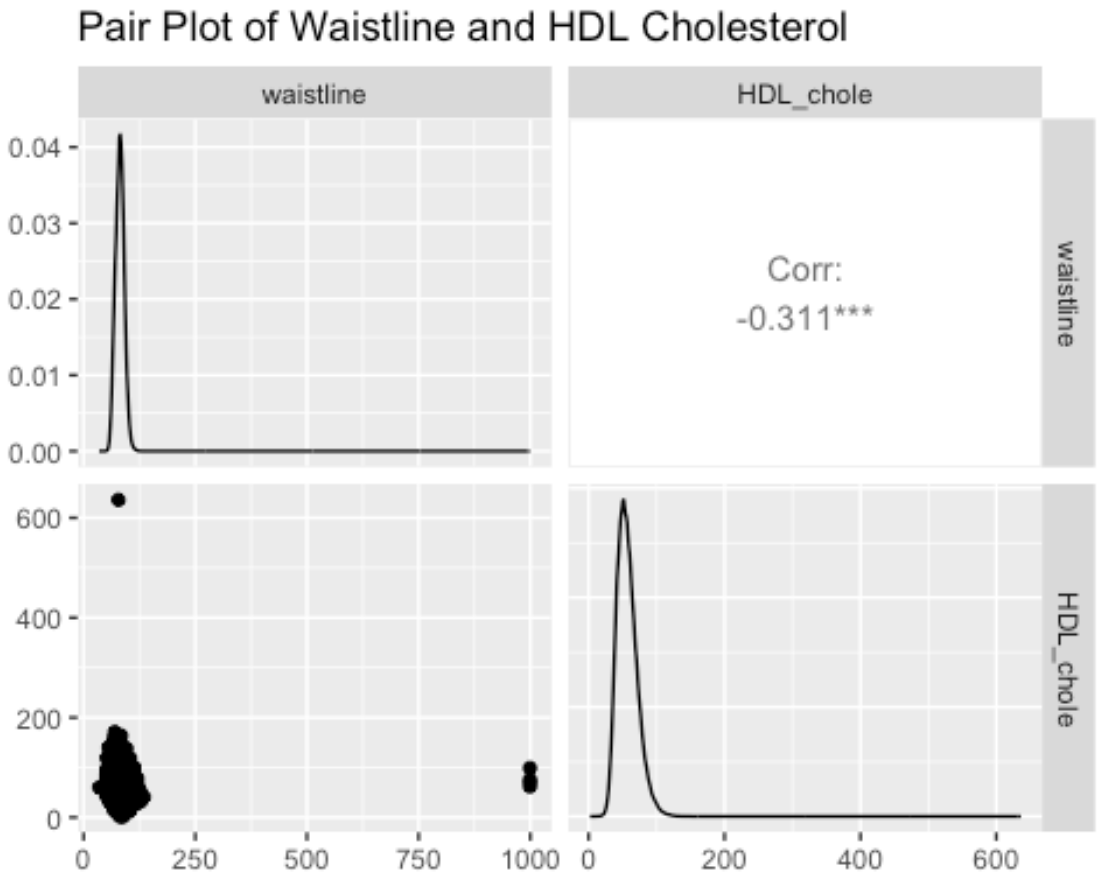
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
# Distribution of Ages
ggplot(data, aes(x=age)) +
  geom_histogram(binwidth=5, fill="blue", color="black", alpha=0.7) +
  geom_density(aes(y=..density.. * 5), color="red") +
  labs(title="Age Distribution", x="Age", y="Frequency")
```

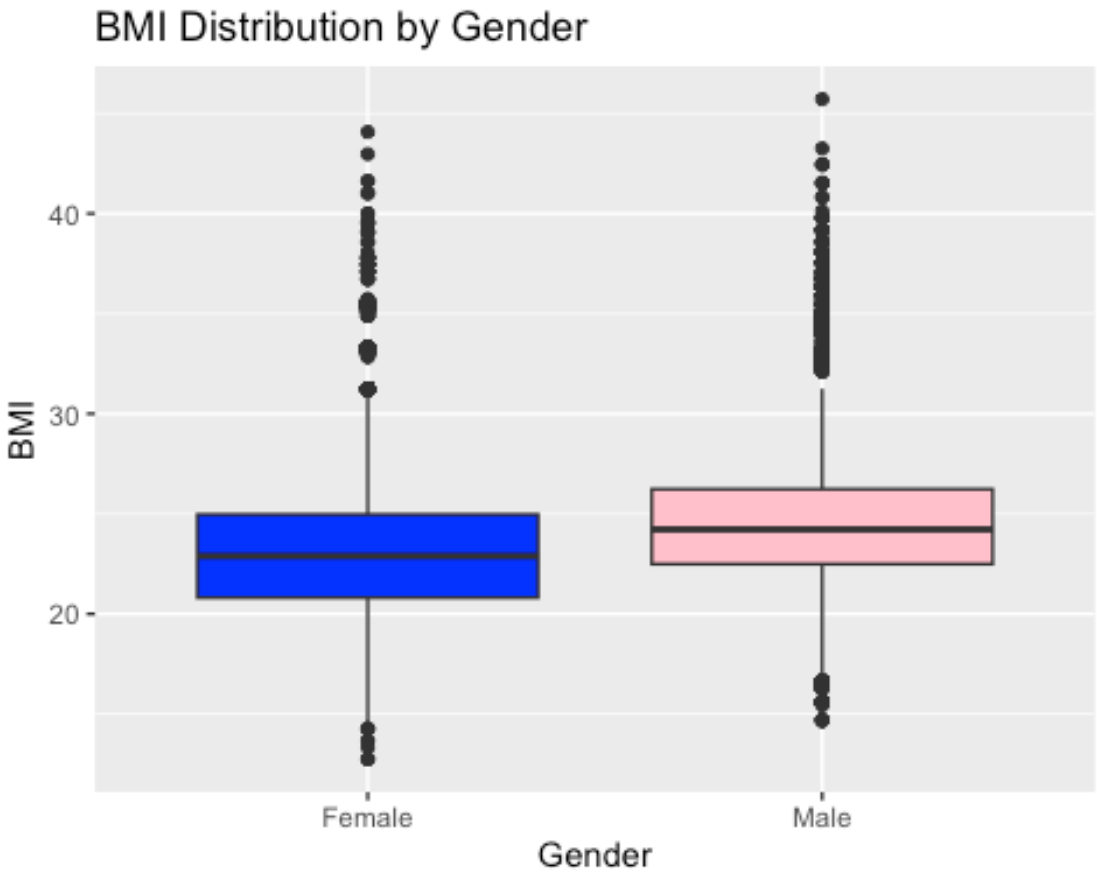
```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2
3.4.0.
## ℹ Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



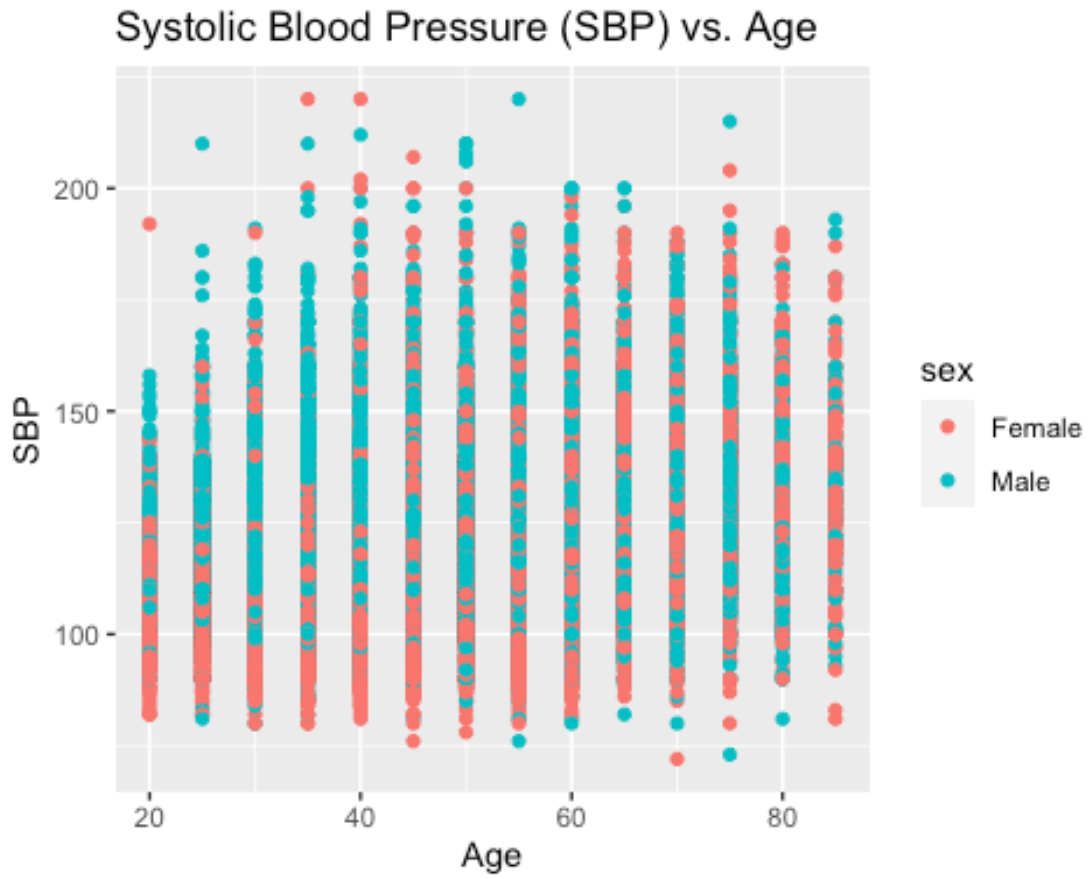

```r
# Pair Plot
library(GGally)

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

# Select a subset of columns for the pair plot to avoid overcrowding
subset_data <- data %>% select( waistline, HDL_chole)

# Create the pair plot
ggpairs(subset_data, title = "Pair Plot of Waistline and HDL Cholesterol")
```



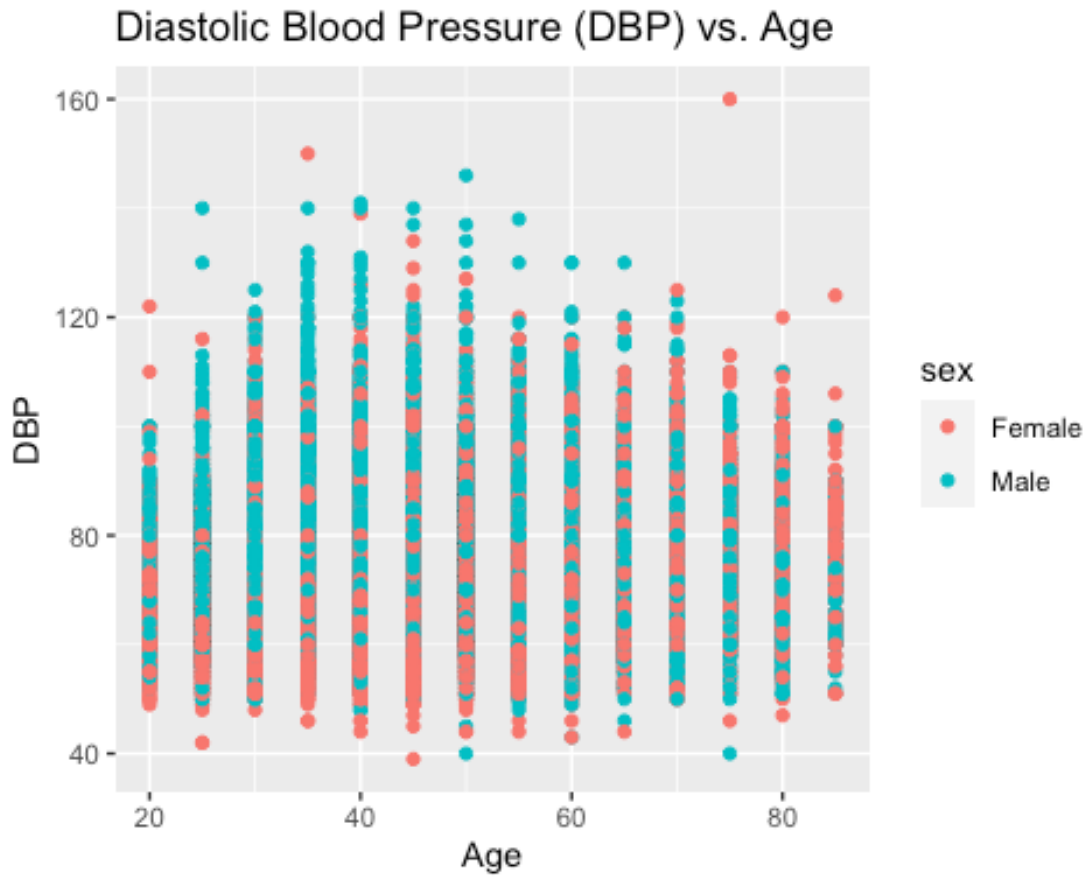

```r
data <- data %>%
  mutate(BMI = weight / (height / 100)^2)

# BMI Distribution by Gender
ggplot(data, aes(x=sex, y=BMI)) +
  geom_boxplot(fill=c("blue", "pink")) +
  labs(title="BMI Distribution by Gender", x="Gender", y="BMI")
```



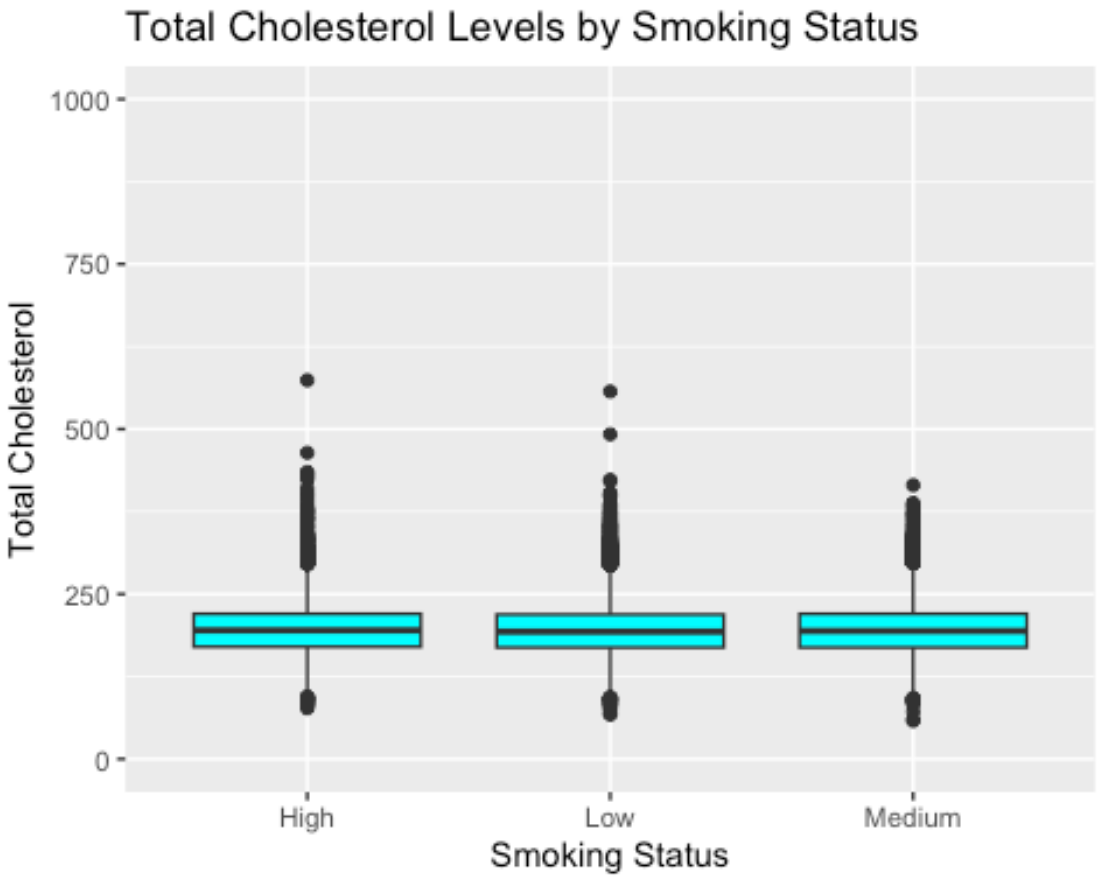

```r
# SBP vs Age
ggplot(data, aes(x=age, y=SBP, color=sex)) +
  geom_point() +
  labs(title="Systolic Blood Pressure (SBP) vs. Age", x="Age", y="SBP")
```



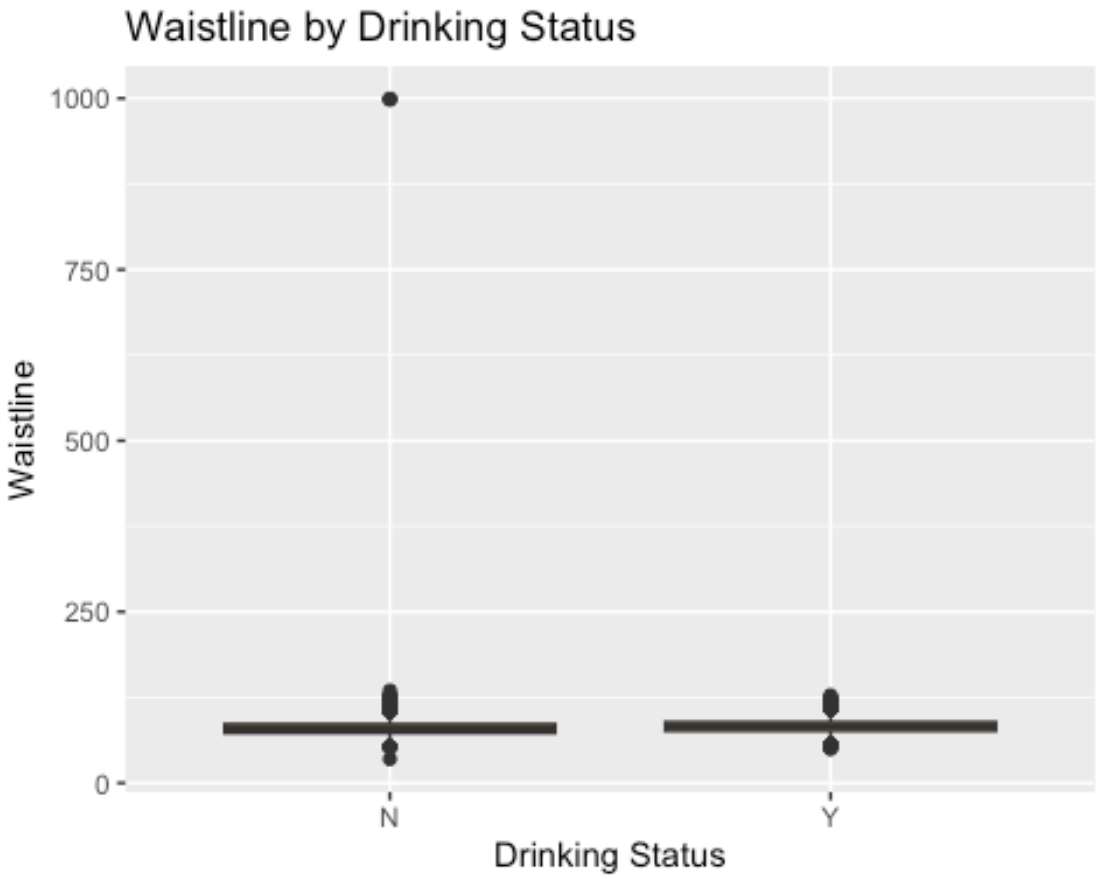

```r
# DBP vs Age
ggplot(data, aes(x=age, y=DBP, color=sex)) +
  geom_point() +
  labs(title="Diastolic Blood Pressure (DBP) vs. Age", x="Age", y="DBP")
```



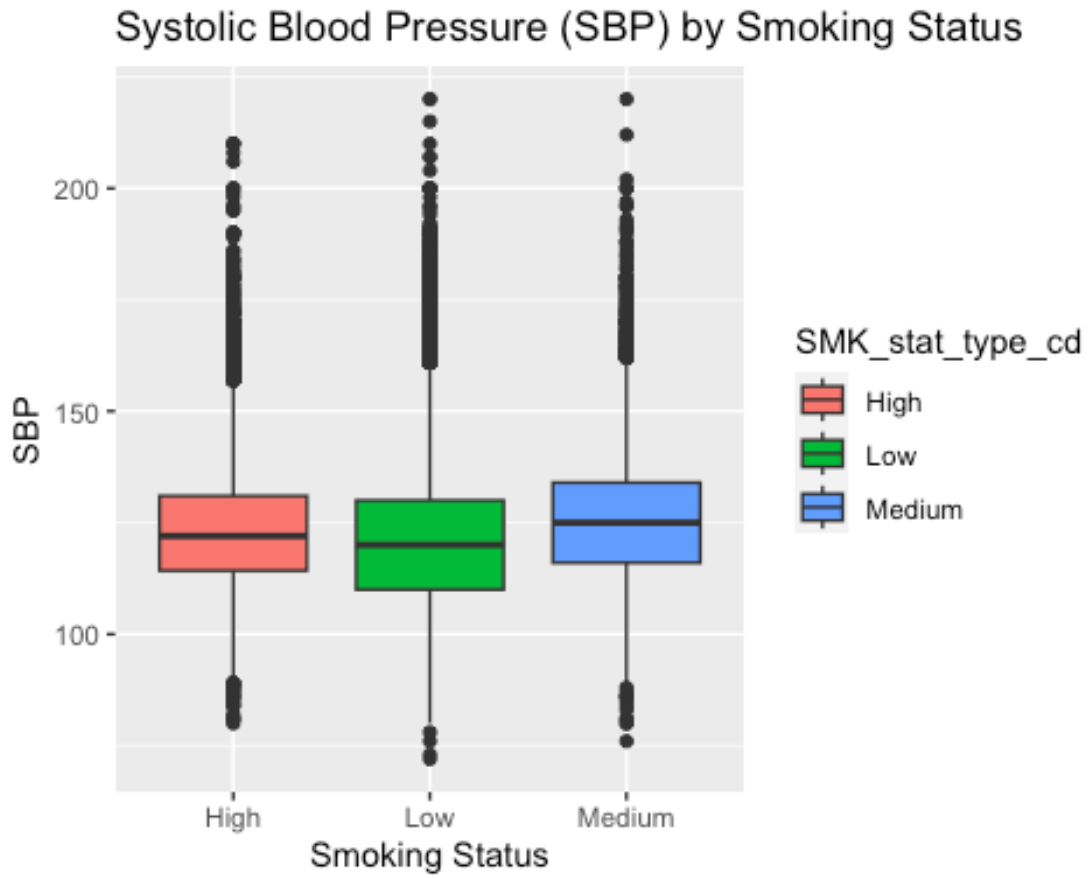

```
# Boxplot of Cholesterol Levels by Smoking Status with y-axis scaled
ggplot(data, aes(x=SMK_stat_type_cd, y=tot_chole)) +
  geom_boxplot(fill="cyan") +
  labs(title="Total Cholesterol Levels by Smoking Status", x="Smoking
Status", y="Total Cholesterol") +
  ylim(0, 1000)

## Warning: Removed 1 rows containing non-finite values (`stat_boxplot()`).
```



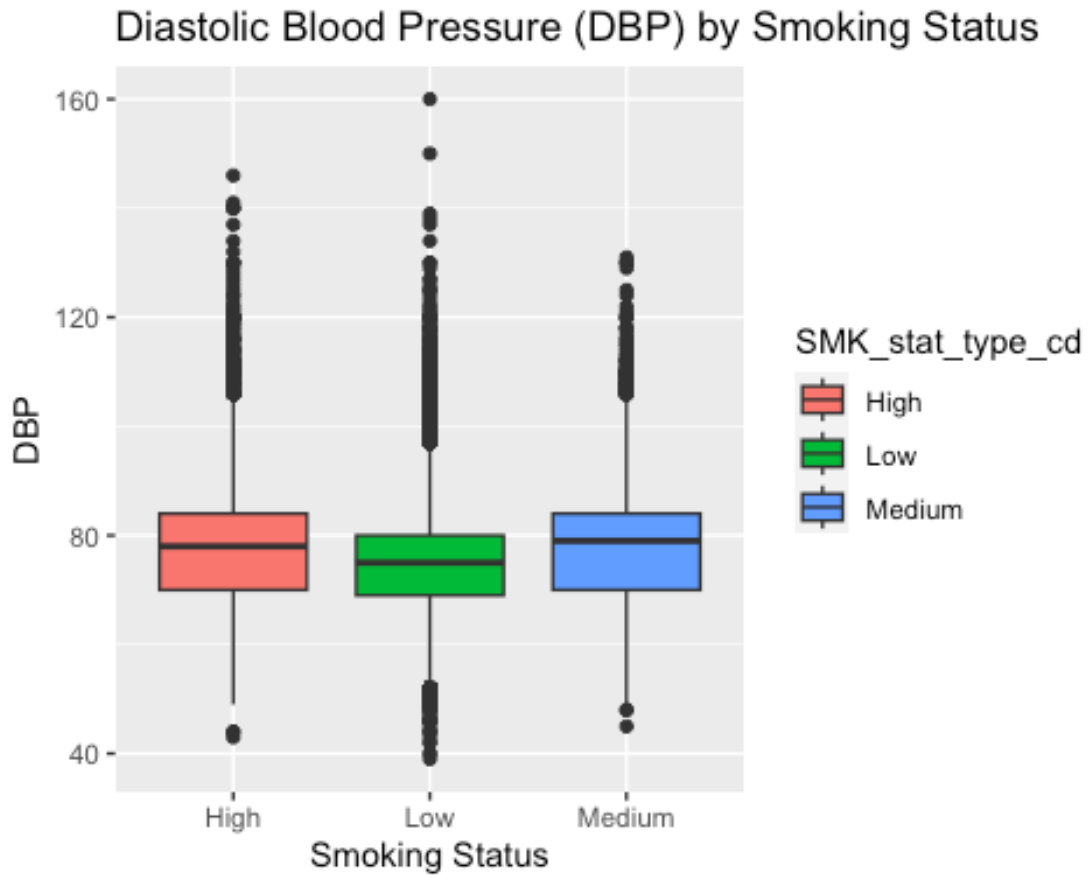

```
# Waistline by Drinking Status
ggplot(data, aes(x=DRK_YN, y=waistline)) +
  geom_boxplot(fill="orange") +
  labs(title="Waistline by Drinking Status", x="Drinking Status",
y="Waistline")
```




```
# SBP by Smoking Status
ggplot(data, aes(x=SMK_stat_type_cd, y=SBP, fill=SMK_stat_type_cd)) +
  geom_boxplot() +
  labs(title="Systolic Blood Pressure (SBP) by Smoking Status", x="Smoking
Status", y="SBP")
```




```r
# DBP by Smoking Status
ggplot(data, aes(x=SMK_stat_type_cd, y=DBP, fill=SMK_stat_type_cd)) +
  geom_boxplot() +
  labs(title="Diastolic Blood Pressure (DBP) by Smoking Status", x="Smoking
Status", y="DBP")
```



FEATURE ENGINEERING

```
##LABEL ENCODING
library(dplyr)
data <- data %>% mutate(SMK_stat_type_cd = recode(SMK_stat_type_cd, "Low"=1,
"Medium"=2,"High"=3))
head(data)
```

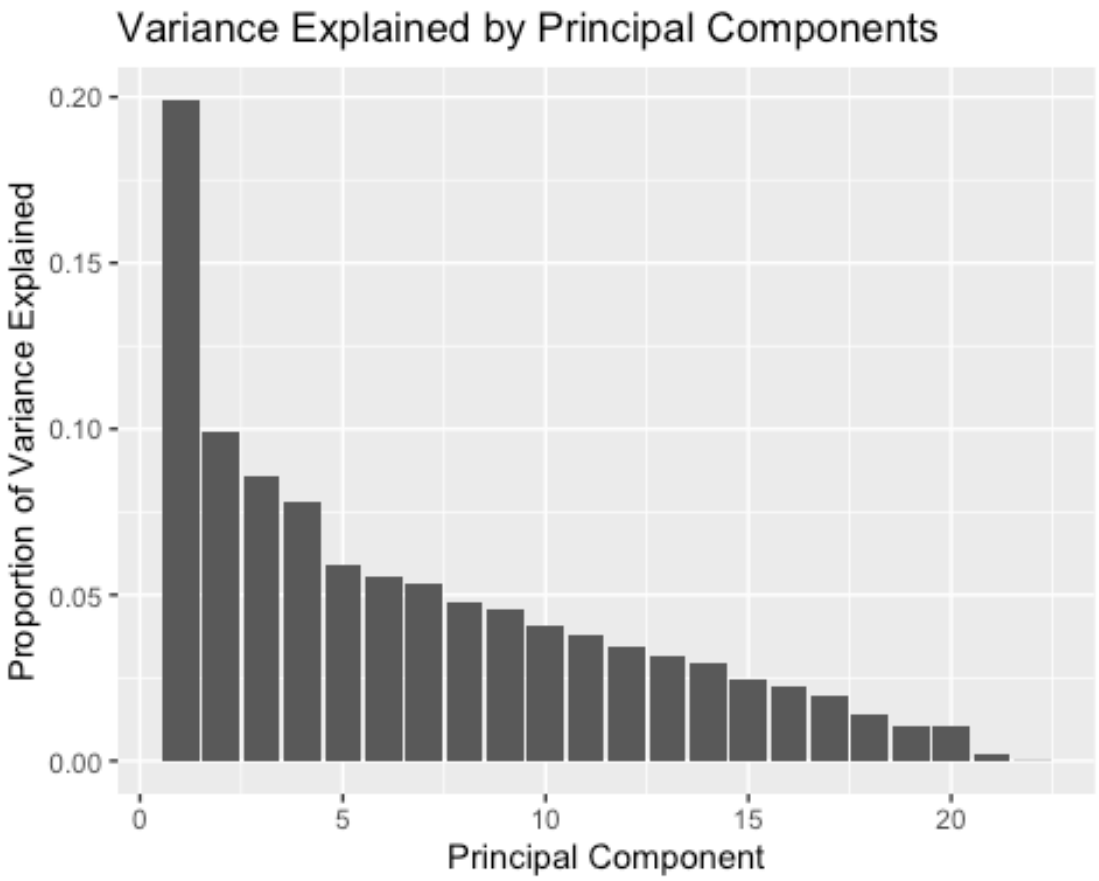
```
## # A tibble: 6 × 24
##   sex      age height weight waistline sight_left sight_right  hear   SBP
DBP
##   <chr>  <dbl>  <dbl>  <dbl>     <dbl>      <dbl>       <dbl> <dbl> <dbl>
<dbl>
## 1 Female    55    160     65        98        1.2         1.5     1   139
81
## 2 Female    70    150     55        82        0.7         0.8     1   118
76
## 3 Male      55    170     80        90        0.9         1       1   116
79
## 4 Male      50    170     60        73        1.2         1.5     1   123
80
## 5 Male      50    165     70        86        0.9         0.5     1   115
84
## 6 Female    40    165     55        68        0.1         0.8     1    94
```

```
56
## # ℹ 14 more variables: BLDS <dbl>, tot_chole <dbl>, HDL_chole <dbl>,
## #   LDL_chole <dbl>, triglyceride <dbl>, hemoglobin <dbl>, urine_protein
<dbl>,
## #   serum_creatinine <dbl>, SGOT_AST <dbl>, SGOT_ALT <dbl>, gamma_GTP
<dbl>,
## #   SMK_stat_type_cd <dbl>, DRK_YN <chr>, BMI <dbl>
```

-PCA

```
##PCA
numeric_data <- data %>% select_if(is.numeric)
scaled_data <- scale(numeric_data)
pca_result <- prcomp(scaled_data, center = TRUE, scale. = TRUE)
summary(pca_result)

## Importance of components:
##                           PC1     PC2     PC3     PC4    PC5     PC6
PC7
## Standard deviation     2.093 1.47789 1.37202 1.30750 1.1393 1.10462
1.08315
## Proportion of Variance 0.199 0.09928 0.08557 0.07771 0.0590 0.05546
0.05333
## Cumulative Proportion  0.199 0.29831 0.38387 0.46158 0.5206 0.57604
0.62937
##                            PC8    PC9    PC10    PC11    PC12    PC13
PC14
## Standard deviation     1.02131 1.0005 0.94448 0.91013 0.86660 0.82972
0.80752
## Proportion of Variance 0.04741 0.0455 0.04055 0.03765 0.03414 0.03129
0.02964
## Cumulative Proportion  0.67678 0.7223 0.76283 0.80048 0.83462 0.86591
0.89555
##                           PC15    PC16    PC17    PC18    PC19    PC20
PC21
## Standard deviation     0.73794 0.70794 0.66230 0.55870 0.48588 0.47313
0.1936
## Proportion of Variance 0.02475 0.02278 0.01994 0.01419 0.01073 0.01018
0.0017
## Cumulative Proportion  0.92030 0.94308 0.96302 0.97721 0.98794 0.99812
0.9998
##                           PC22
## Standard deviation     0.06311
## Proportion of Variance 0.00018
## Cumulative Proportion  1.00000

var_explained <- pca_result$sdev^2 / sum(pca_result$sdev^2)
var_df <- data.frame(Principal_Component = seq_along(var_explained),
Variance_Explained = var_explained)
library(ggplot2)
ggplot(var_df, aes(x = Principal_Component, y = Variance_Explained)) +
```

```
       geom_bar(stat = "identity") +
       labs(x = "Principal Component", y = "Proportion of Variance
Explained") +
       ggtitle("Variance Explained by Principal Components")
```




```
pca_data <- as.data.frame(pca_result$x)
head(pca_data)
```

```
##         PC1        PC2        PC3        PC4        PC5        PC6
PC7
## 1  0.334769  0.5585918 -1.6593431 -0.2325317  0.7129040 -1.9124306
0.50829717
## 2 -1.505633  2.5116221  2.3189104 -0.8594319 -0.9777495 -0.5446020 -
0.42199654
## 3  2.111858 -0.5441907 -0.2809482 -0.2957125 -0.6503949  0.3420870 -
0.03708558
## 4  0.435371 -1.2166387  0.9989026 -0.2087235  1.0507645  1.1605375 -
0.56216518
## 5  1.504504 -0.2554503 -2.5515983  1.2208192 -0.2712053  0.2942404
0.49566547
## 6 -3.041373 -1.3863132 -1.1871538  0.7422570 -2.0010363  0.4925191
0.23528544
##           PC8           PC9        PC10       PC11        PC12       PC13
```

```
## 1  0.19787847 -0.0198825379 -0.02473317 -0.2430759  0.20576780 -0.4932482
## 2  0.18839566  0.0003797996 -0.65481357 -0.3078360 -0.01696747 -0.7586539
## 3 -0.37623369 -0.0604351580  0.02787854 -0.1553400  0.96789986 -0.8240059
## 4  0.02410999 -0.9634908701 -1.20139510  0.1246366 -0.81922196 -0.9517683
## 5 -0.27401088  0.1963260682 -0.38464676 -0.2304123  0.34906705  0.5581742
## 6 -0.41083797  0.2902257417 -0.09662243  0.4685546 -0.20762739  0.2471961
##           PC14        PC15        PC16        PC17        PC18
PC19
## 1 -0.277163690  0.08628736 -0.20107016 -0.60072744  0.74065101 -
0.345979137
## 2  0.008755368 -0.35585555  0.71830793 -0.49722208 -0.07505047
0.464690993
## 3  0.028036295  0.72755431  0.84152148 -0.43512577 -0.50673437
0.692188523
## 4 -0.195571095 -0.10467415  0.37334066  0.05390136 -0.16477320
0.197136743
## 5  0.368950101  0.85006976 -0.07121871  0.21008456  0.06196679
0.869772998
## 6 -0.833019956 -0.63562590  0.09759674 -0.55008942 -0.31683534
0.004508841
##           PC20          PC21          PC22
## 1  0.211800079 -0.0613194605 -0.0198380064
## 2  0.003609334  0.0078762240 -0.0006623977
## 3 -0.178998662 -0.3338541305  0.0356816073
## 4  0.002507702 -0.0005400293 -0.0577307400
## 5 -0.769752386  0.0451763317 -0.0186488551
## 6 -0.137476769 -0.0131734259 -0.0370335148
```

– The PCA components are as above, and the number of principal directions can be chosen according to the required dimensions. – Our Data set does not require PCA, but feature engineering such as ONE HOT coding and LABEL Coding is required.

```
## ONE-HOT ENCODING
one_hot_encoded_data <- as.data.frame(model.matrix(~ DRK_YN - 1, data =
data))
data <- cbind(data, one_hot_encoded_data)
data <- select(data, -DRK_YN)
head(data)

##      sex age height weight waistline sight_left sight_right hear SBP DBP
BLDS
## 1 Female  55    160     65        98        1.2         1.5    1 139  81
96
## 2 Female  70    150     55        82        0.7         0.8    1 118  76
95
## 3   Male  55    170     80        90        0.9         1.0    1 116  79
96
## 4   Male  50    170     60        73        1.2         1.5    1 123  80
84
## 5   Male  50    165     70        86        0.9         0.5    1 115  84
```

```
110
## 6 Female  40    165     55        68        0.1         0.8    1  94  56
89
##   tot_chole HDL_chole LDL_chole triglyceride hemoglobin urine_protein
## 1       151        60        75           80       13.3             1
## 2       267        55       194           92       11.4             1
## 3       191        57       109          250       14.0             1
## 4       211        46       145           96       15.6             1
## 5       137        54        38          223       14.8             1
## 6       145        50        80           72       12.4             1
##   serum_creatinine SGOT_AST SGOT_ALT gamma_GTP SMK_stat_type_cd      BMI
## 1              0.7       34       28        33                1 25.39062
## 2              0.8       20        9        11                1 24.44444
## 3              1.0       26       28        65                3 27.68166
## 4              1.2       24       21        26                3 20.76125
## 5              1.1       39       68        56                2 25.71166
## 6              0.6       20       19        13                1 20.20202
##   DRK_YNN DRK_YNY
## 1       1       0
## 2       1       0
## 3       0       1
## 4       0       1
## 5       0       1
## 6       1       0
```

–NORMALIZATION

```
##DATA NORMALIZATION
numeric_columns <- names(data)[sapply(data, is.numeric)]
numeric_columns <- setdiff(numeric_columns, c("DRK_YNN", "DRK_YNY",
"SMK_stat_type_cd"))
data[numeric_columns] <- scale(data[numeric_columns])
head(data)

##      sex        age     height     weight  waistline sight_left
sight_right
## 1 Female  0.5185096 -0.2421566  0.1316500  1.4881283  0.3590636
0.85029337
## 2 Female  1.5761529 -1.3206853 -0.6644916  0.0643521 -0.4612257 -
0.29140562
## 3   Male  0.5185096  0.8363720  1.3258623  0.7762402 -0.1331100
0.03479409
## 4   Male  0.1659619  0.8363720 -0.2664208 -0.7365220  0.3590636
0.85029337
## 5   Male  0.1659619  0.2971077  0.5297207  0.4202961 -0.1331100 -
0.78070519
## 6 Female -0.5391337  0.2971077 -0.6644916 -1.1814520 -1.4455729 -
0.29140562
##        hear         SBP          DBP       BLDS  tot_chole    HDL_chole
## 1 0.1753479  1.13878308  0.502311630 -0.1839433 -1.1569709  0.205533328
```

```
## 2 0.1753479 -0.30053079 -0.004193866 -0.2248220  1.8556800 -0.126932975
## 3 0.1753479 -0.43760830  0.299709432 -0.1839433 -0.1181258  0.006053546
## 4 0.1753479  0.04216299  0.401010531 -0.6744882  0.4012968 -0.725372320
## 5 0.1753479 -0.50614706  0.806214928  0.3883591 -1.5205667 -0.193426235
## 6 0.1753479 -1.94546093 -2.030215849 -0.4700945 -1.3127977 -0.459399278
##    LDL_chole triglyceride hemoglobin urine_protein serum_creatinine
## 1 -1.0520476   -0.5137342 -0.5899076    -0.2154037       -0.4007897
## 2  2.2377911   -0.3957481 -1.7897345    -0.2154037       -0.1510967
## 3 -0.1120937    1.1577352 -0.1478660    -0.2154037        0.3482893
## 4  0.8831517   -0.3564194  0.8625146    -0.2154037        0.8476753
## 5 -2.0749386    0.8922665  0.3573243    -0.2154037        0.5979823
## 6 -0.9138190   -0.5923915 -1.1582466    -0.2154037       -0.6504827
##       SGOT_AST   SGOT_ALT   gamma_GTP SMK_stat_type_cd        BMI DRK_YNN
## 1  0.411391666  0.0811721 -0.08165343                1  0.4121251       1
## 2 -0.303239400 -0.6178864 -0.52407886                1  0.1437685       1
## 3  0.003031057  0.0811721  0.56187446                3  1.0619106       0
## 4 -0.099059096 -0.1763758 -0.22242516                3 -0.9008635       0
## 5  0.666617047  1.5528742  0.38088224                2  0.5031779       0
## 6 -0.303239400 -0.2499609 -0.48385837                1 -1.0594715       1
##   DRK_YNY
## 1       0
## 2       0
## 3       1
## 4       1
## 5       1
## 6       0
```

– IMPLEMENTING KNN

```r
#K-MEANS
train_indices <- sample(1:nrow(data), size = 0.7 * nrow(data))
training_set <- data[train_indices, ]
testing_set <- data[-train_indices, ]
columns_for_clustering <- setdiff(names(training_set), c("healthindex",
"DRK_YNN", "DRK_YNY", "SMK_stat_type_cd","sex"))
kmeans_result <- kmeans(training_set[, columns_for_clustering], centers = 3,
nstart = 10)
```

```
## Warning: Quick-TRANSfer stage steps exceeded maximum (= 3500000)

## Warning: Quick-TRANSfer stage steps exceeded maximum (= 3500000)

## Warning: Quick-TRANSfer stage steps exceeded maximum (= 3500000)
```

```r
training_set$healthindex <- kmeans_result$cluster


# Function to assign clusters based on the centroids
assign_cluster <- function(row, centroids) {
  distances <- apply(centroids, 1, function(centroid) sum((row -
```

```
centroid)^2))
  return(which.min(distances))
}

# Ensure we only use numeric columns for the distance calculation
testing_numeric <- testing_set[, columns_for_clustering]
centroids <- kmeans_result$centers

# Assign clusters to the testing set
testing_set$healthindex <- apply(testing_numeric, 1, assign_cluster,
centroids = centroids)

# Verify the new column in the testing set
head(testing_set)
```

```
##       sex        age     height     weight  waistline sight_left
sight_right
## 5    Male  0.1659619  0.2971077  0.5297207  0.4202961 -0.1331100  -
0.7807052
## 6  Female -0.5391337  0.2971077 -0.6644916 -1.1814520 -1.4455729  -
0.2914056
## 11 Female  1.2236051 -1.8599496 -1.8587039 -0.4695640 -0.7893415  -
0.7807052
## 14   Male -0.5391337  0.2971077 -1.4606331 -1.5373961  0.8512372
0.8502934
## 15   Male  1.9287006  0.2971077 -0.2664208 -0.4250710 -1.4455729  -
0.7807052
## 20 Female -0.5391337 -1.3206853 -0.6644916 -0.2915919  0.3590636
0.8502934
##         hear        SBP        DBP        BLDS  tot_chole  HDL_chole
LDL_chole
## 5  0.1753479 -0.5061471  0.8062149  0.38835907 -1.5205667 -0.1934262 -
2.0749386
## 6  0.1753479 -1.9454609 -2.0302158 -0.47009448 -1.3127977 -0.4593993 -
0.9138190
## 11 0.1753479 -1.3286121 -1.0172049 -0.22482204 -0.9751730 -0.9913454 -
0.5544249
## 14 0.1753479  0.6590118 -0.9159038 -0.06130707 -0.9232308  1.6683851 -
1.4667331
## 15 0.1753479 -0.9859184 -0.8146027 -0.38833700 -0.3518659  1.2029322 -
0.3609050
## 20 0.1753479 -0.5061471 -1.0172049 -0.30657952  0.3233834  1.1364390
0.2473005
##    triglyceride hemoglobin urine_protein serum_creatinine    SGOT_AST
## 5     0.8922665  0.3573243    -0.2154037       0.59798228  0.66661705
## 6    -0.5923915 -1.1582466    -0.2154037      -0.65048273 -0.30323940
## 11   -0.1892725 -1.9160321    -0.2154037      -0.15109673 -0.15010417
## 14   -0.4350768 -0.4004612    -0.2154037       0.34828928 -0.30323940
## 15   -0.9758463 -0.1478660    -0.2154037       0.34828928 -0.04801402
## 20   -0.7300420 -2.4212224    -0.2154037       0.09859628 -0.60950986
```

```
##       SGOT_ALT  gamma_GTP SMK_stat_type_cd        BMI DRK_YNN DRK_YNY
## 5   1.55287418  0.3808822                 2  0.5031779       0       1
## 6  -0.24996087 -0.4838584                 1 -1.0594715       1       0
## 11 -0.54430129 -0.6045198                 1 -1.3933074       1       0
## 14 -0.36033853 -0.2425354                 3 -2.1012377       1       0
## 15  0.04437954 -0.2023149                 2 -0.5385883       1       0
## 20 -0.58109384 -0.3229764                 1  0.1437685       0       1
##    healthindex
## 5            1
## 6            3
## 11           3
## 14           3
## 15           3
## 20           3

# Compare the distribution of clusters in the training and testing sets
table(training_set$healthindex)

##
##     1     2     3
## 24887 21425 23688

table(testing_set$healthindex)

##
##     1     2     3
## 10563  9376 10061

# Visualize the clusters
ggplot(training_set, aes(x = training_set[,columns_for_clustering][,1], y =
training_set[,columns_for_clustering][,2], color = as.factor(healthindex))) +
  geom_point() +
  labs(title = "Training Set Clusters", color = "Cluster")
```


```
ggplot(testing_set, aes(x = testing_set[,columns_for_clustering][,1], y =
testing_set[,columns_for_clustering][,2], color = as.factor(healthindex))) +
  geom_point() +
  labs(title = "Testing Set Clusters", color = "Cluster")
```


```
head(training_set)
```

```
##          sex        age     height     weight  waistline  sight_left
## 11661 Female  1.2236051 -0.2421566 -0.6644916  0.0643521  0.03094788
## 23501 Female -1.5967769 -0.2421566 -0.6644916 -0.8255080  0.35906361
## 14525   Male  0.1659619  1.3756363  0.9277915  0.6872542 -0.29716786
## 16807 Female -0.5391337  0.2971077 -1.0625623 -1.5373961  0.03094788
## 56713 Female  0.5185096 -0.7814210 -0.2664208  0.1533381 -0.13310999
## 70120   Male  0.1659619  1.3756363  0.9277915  0.7139500  0.35906361
##       sight_right      hear        SBP         DBP       BLDS  tot_chole
## 11661 -0.29140562 0.1753479  0.3163180  0.09710723  0.5109953 -0.4038082
## 23501  0.03479409 0.1753479 -0.9859184 -1.62501145 -0.3474583  0.3233834
## 14525 -0.45450547 0.1753479 -0.3690695  0.19840833  0.4292378  1.4920842
## 16807 -0.12830576 0.1753479  0.2477793  1.21141932 -0.6744882 -0.4557504
## 56713 -0.61760533 0.1753479  0.5219343  1.21141932 -0.1839433  0.5830947
## 70120  0.03479409 0.1753479  0.7960893  1.00881713  1.1650551  0.1675566
##        HDL_chole  LDL_chole triglyceride  hemoglobin urine_protein
## 11661  0.8039727 -0.7755905  -0.02212553 -0.71620513    -0.2154037
## 23501  2.6657840 -0.2779679  -0.90702107 -1.41084179    -0.2154037
## 14525 -0.3929060  0.9937345   1.28555342  1.49400243    -0.2154037
## 16807  1.0699457 -0.5267792  -0.77920283 -2.35807360    -0.2154037
## 56713  0.9369592  0.5790489  -0.65138458 -0.40046119    -0.2154037
## 70120 -0.7253723  0.6896317  -0.42524461 -0.08471726    -0.2154037
```

```
##       serum_creatinine    SGOT_AST    SGOT_ALT   gamma_GTP
SMK_stat_type_cd
## 11661       -0.4007897  0.25825644  0.04437954 -0.10176368
1
## 23501       -0.1510967 -0.60950986 -0.54430129 -0.52407886
1
## 14525        0.3482893  0.15616629  1.03777845  3.43763973
3
## 16807       -0.4007897 -0.40532955 -0.61788639 -0.32297639
1
## 56713       -0.4007897 -0.04801402 -0.24996087 -0.38330713
1
## 70120        0.5979823 -0.09905910  0.22834231 -0.08165343
2
##               BMI DRK_YNN DRK_YNY healthindex
## 11661 -0.6957689       0       1           2
## 23501 -0.6957689       0       1           3
## 14525  0.1566311       0       1           1
## 16807 -1.5803546       0       1           3
## 56713  0.2939574       1       0           2
## 70120  0.1566311       0       1           1
```

```
head(testing_set)
```

```
##       sex        age     height     weight  waistline sight_left
sight_right
## 5    Male  0.1659619  0.2971077  0.5297207  0.4202961 -0.1331100  -
0.7807052
## 6  Female -0.5391337  0.2971077 -0.6644916 -1.1814520 -1.4455729  -
0.2914056
## 11 Female  1.2236051 -1.8599496 -1.8587039 -0.4695640 -0.7893415  -
0.7807052
## 14   Male -0.5391337  0.2971077 -1.4606331 -1.5373961  0.8512372
0.8502934
## 15   Male  1.9287006  0.2971077 -0.2664208 -0.4250710 -1.4455729  -
0.7807052
## 20 Female -0.5391337 -1.3206853 -0.6644916 -0.2915919  0.3590636
0.8502934
##         hear        SBP        DBP        BLDS  tot_chole  HDL_chole
LDL_chole
## 5  0.1753479 -0.5061471  0.8062149  0.38835907 -1.5205667 -0.1934262 -
2.0749386
## 6  0.1753479 -1.9454609 -2.0302158 -0.47009448 -1.3127977 -0.4593993 -
0.9138190
## 11 0.1753479 -1.3286121 -1.0172049 -0.22482204 -0.9751730 -0.9913454 -
0.5544249
## 14 0.1753479  0.6590118 -0.9159038 -0.06130707 -0.9232308  1.6683851 -
1.4667331
## 15 0.1753479 -0.9859184 -0.8146027 -0.38833700 -0.3518659  1.2029322 -
0.3609050
```

```
## 20 0.1753479 -0.5061471 -1.0172049 -0.30657952  0.3233834  1.1364390
0.2473005
##    triglyceride hemoglobin urine_protein serum_creatinine    SGOT_AST
## 5     0.8922665  0.3573243    -0.2154037       0.59798228  0.66661705
## 6    -0.5923915 -1.1582466    -0.2154037      -0.65048273 -0.30323940
## 11   -0.1892725 -1.9160321    -0.2154037      -0.15109673 -0.15010417
## 14   -0.4350768 -0.4004612    -0.2154037       0.34828928 -0.30323940
## 15   -0.9758463 -0.1478660    -0.2154037       0.34828928 -0.04801402
## 20   -0.7300420 -2.4212224    -0.2154037       0.09859628 -0.60950986
##       SGOT_ALT  gamma_GTP SMK_stat_type_cd        BMI DRK_YNN DRK_YNY
## 5   1.55287418  0.3808822                2  0.5031779       0       1
## 6  -0.24996087 -0.4838584                1 -1.0594715       1       0
## 11 -0.54430129 -0.6045198                1 -1.3933074       1       0
## 14 -0.36033853 -0.2425354                3 -2.1012377       1       0
## 15  0.04437954 -0.2023149                2 -0.5385883       1       0
## 20 -0.58109384 -0.3229764                1  0.1437685       0       1
##    healthindex
## 5            1
## 6            3
## 11           3
## 14           3
## 15           3
## 20           3
```

```
##CONVERT 1,2,3 TO DESCRIPTIVE HEALTH INDEX
   convert_health_risk <- function(index) {
  if (index == 1) {
    return("low health risk")
  } else if (index == 2) {
    return("medium health risk")
  } else if (index == 3) {
    return("high health risk")
  }
}


    training_set$healthindex <- sapply(training_set$healthindex,
convert_health_risk)
   testing_set$healthindex <- sapply(testing_set$healthindex,
convert_health_risk)
head(testing_set)
```

```
##       sex        age     height     weight  waistline sight_left
sight_right
## 5    Male  0.1659619  0.2971077  0.5297207  0.4202961 -0.1331100  -
0.7807052
## 6  Female -0.5391337  0.2971077 -0.6644916 -1.1814520 -1.4455729  -
0.2914056
## 11 Female  1.2236051 -1.8599496 -1.8587039 -0.4695640 -0.7893415  -
0.7807052
```

```
## 14   Male -0.5391337  0.2971077 -1.4606331 -1.5373961  0.8512372
0.8502934
## 15   Male  1.9287006  0.2971077 -0.2664208 -0.4250710 -1.4455729  -
0.7807052
## 20 Female -0.5391337 -1.3206853 -0.6644916 -0.2915919  0.3590636
0.8502934
##         hear        SBP        DBP        BLDS  tot_chole  HDL_chole
LDL_chole
## 5  0.1753479 -0.5061471  0.8062149  0.38835907 -1.5205667 -0.1934262 -
2.0749386
## 6  0.1753479 -1.9454609 -2.0302158 -0.47009448 -1.3127977 -0.4593993 -
0.9138190
## 11 0.1753479 -1.3286121 -1.0172049 -0.22482204 -0.9751730 -0.9913454 -
0.5544249
## 14 0.1753479  0.6590118 -0.9159038 -0.06130707 -0.9232308  1.6683851 -
1.4667331
## 15 0.1753479 -0.9859184 -0.8146027 -0.38833700 -0.3518659  1.2029322 -
0.3609050
## 20 0.1753479 -0.5061471 -1.0172049 -0.30657952  0.3233834  1.1364390
0.2473005
##    triglyceride hemoglobin urine_protein serum_creatinine    SGOT_AST
## 5     0.8922665  0.3573243    -0.2154037       0.59798228  0.66661705
## 6    -0.5923915 -1.1582466    -0.2154037      -0.65048273 -0.30323940
## 11   -0.1892725 -1.9160321    -0.2154037      -0.15109673 -0.15010417
## 14   -0.4350768 -0.4004612    -0.2154037       0.34828928 -0.30323940
## 15   -0.9758463 -0.1478660    -0.2154037       0.34828928 -0.04801402
## 20   -0.7300420 -2.4212224    -0.2154037       0.09859628 -0.60950986
##       SGOT_ALT  gamma_GTP SMK_stat_type_cd        BMI DRK_YNN DRK_YNY
## 5   1.55287418  0.3808822                2  0.5031779       0       1
## 6  -0.24996087 -0.4838584                1 -1.0594715       1       0
## 11 -0.54430129 -0.6045198                1 -1.3933074       1       0
## 14 -0.36033853 -0.2425354                3 -2.1012377       1       0
## 15  0.04437954 -0.2023149                2 -0.5385883       1       0
## 20 -0.58109384 -0.3229764                1  0.1437685       0       1
##         healthindex
## 5   low health risk
## 6  high health risk
## 11 high health risk
## 14 high health risk
## 15 high health risk
## 20 high health risk
```